HKU STAT6010/STAT7610 Study Notes

Ka Long (Leo) Chiu*

Last Updated: 2025-09-11

Contents

---

*email ✉: leockl@connect.hku.hk; personal website 🔗: https://leochiukl.github.io

1 Measure Theory I: Measurability of Sets

1.0.1 A thorough and rigorous study of probability theory *requires* measure theory. While you may have learnt probability without ever using measure theory before (typically the case for a first course in probability), many concepts and terminologies you have seen are in fact measure theoretic. For example, an *event* is a *measurable* set, and a *random variable* is a *measurable* function. For now, you may not find the need of introducing measure theory in the study of probability. But as we shall see, there are indeed some pathological situations that have to be handled via measure theory.

1.0.2 As you know from your first course in probability, *events* are sets, so let us study/review some preliminary concepts in set theory that will be helpful for developing measure theoretic probability theory.

1.1 Set Theory Preliminaries

1.1.1 Informally, a *set* can be seen as a collection of distinct objects or *elements*.[1] While it appears to be intuitive that such collection can be arbitrary and unrestricted, some paradoxes can actually result if no restriction is imposed, e.g., Russell's paradox. Such issues lead to the development of *Zermelo-Fraenkel* (ZF) set theory, which imposes a series of axioms that restrict what such "collection" can be.[2]

On top of the axioms in the ZF set theory, another commonly imposed axiom is the *axiom of choice*, which suggests that for all collections $\mathcal{A}$ of non-empty sets, there is a *choice function* f such that $f(A) \in A$ for all $A \in \mathcal{A}$. Intuitively, this axiom means that it is possible to "choose"/"pick" an element from each set in any fixed collection (finite or countably infinite or even uncountably infinite).

The ZF set theory together with the axiom of choice is abbreviated as *ZFC*, and we shall always work with ZFC (which is typically the case when you encounter any mathematics involving sets).

1.1.2 Set terminologies. In the following, we will review some terminologies about sets, which should be very familiar to you. Here, we let Ω be a nonempty *universal* set (meaning that every set below is a *subset* of it).

Name	Notation	Definition
A is a subset of Ω, or Ω is a superset of A	$A \subseteq \Omega$, or $\Omega \supseteq A$	For any $\omega \in A$, we have $\omega \in \Omega$.
(Absolute) complement	A^c	$\{\omega \in \Omega : \omega \notin A\}$
Set difference	$A \setminus B$	$A \cap B^c$
Intersection (I: index set)	$\bigcap_{i \in I} A_i$	$\{\omega \in \Omega : \omega \in A_i \text{ for all } i \in I\}$
Union	$\bigcup_{i \in I} A_i$	$\{\omega \in \Omega : \omega \in A_i \text{ for some } i \in I\}$
Disjoint union	$\biguplus_{i \in I} A_i$	meaning the same as $\bigcup_{i \in I} A_i$, with the emphasis on the pairwise disjointness, i.e., $A_i \cap A_j = \varnothing$ for all $i, j \in I$ with $i \neq j$.

[Note: The concept of complement depends on what the universal set is. In some occasions, we may use the somewhat awkward notation A^{c_Ω} to stress that the universal set being considered is Ω for this complement.]

1.1.3 Basic set properties. You should be very familiar to the following properties of sets (the sets below are arbitrary):

- *Associativity of union and intersection:* $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$.
- *Commutativity of union and intersection:* $A \cup B = B \cup A$ and $A \cap B = B \cap A$.
- *De Morgan's laws:* $(\bigcup_{i \in I} A_i)^c = \bigcap_{i \in I} A_i^c$ and $(\bigcap_{i \in I} A_i)^c = \bigcup_{i \in I} A_i^c$. More generally, for any $B \subseteq \Omega$, we have $B \setminus \bigcup_{i \in I} A_i = \bigcap_{i \in I} B \setminus A_i$ and $B \setminus \bigcap_{i \in I} A_i = \bigcup_{i \in I} B \setminus A_i$.

---

[1]Sometimes the terms "set", "collection", and "family" are used interchangeably.
[2]Here we shall not delve into the details. If you are interested, you can take a look at its Wikipedia page.

Make sure you are able to *prove* these! (An useful way to prove a set equality $S = T$ is to prove (i) $S \subseteq T$ and (ii) $T \subseteq S$.)

1.1.4 **Further set terminologies.** For the following set theoretic terminologies, you may not have encountered them before, so perhaps you should pay more attention here:

Name	Notation	Definition
Infimum (set)	$\inf_{k \geq n} A_k$	$\bigcap_{k \geq n} A_k$
Supremum (set)	$\sup_{k \geq n} A_k$	$\bigcup_{k \geq n} A_k$
Limit inferior (set)	$\liminf_{n \to \infty} A_n$	$\bigcup_{n=1}^{\infty} \bigcap_{k \geq n} A_k = \sup_{n \geq 1} \inf_{k \geq n} A_k$
Limit superior (set)	$\limsup_{n \to \infty} A_n$	$\bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k = \inf_{n \geq 1} \sup_{k \geq n} A_k$
Limit of A_n	$\lim_{n \to \infty} A_n = A$ or $A_n \to A$	$\liminf_{n \to \infty} A_n = \limsup_{n \to \infty} A_n = A$
$\{A_n\}_{n \in \mathbb{N}}$ is **increasing**	$A_n \nearrow$	$A_1 \subseteq A_2 \subseteq \cdots$
$\{A_n\}_{n \in \mathbb{N}}$ is **decreasing**	$A_n \searrow$	$A_1 \supseteq A_2 \supseteq \cdots$

1.1.5 **Further set properties.** Here we will discuss some results about the less familiar set terminologies from [1.1.4]:

(a) *(Interpreting limit inferior and limit superior)*

    i. $\liminf_{n \to \infty} A_n = \{\omega \in \Omega : \omega \in A_n$ for all but finitely many $n\} =: \{\omega \in A_n$ abfm$\}$.

    ii. $\limsup_{n \to \infty} A_n = \{\omega \in \Omega : \omega \in A_n$ for infinitely many $n\} = \{\omega \in \Omega : \omega \in A_n$ infinitely often$\} =: \{\omega \in A_n$ io$\}$.

Proof.

    i. Note that
$$\liminf_{n \to \infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k \geq n} A_k = \{\omega \in \Omega : \exists n \in \mathbb{N} \text{ s.t. } \omega \in A_k \ \forall k \geq n\}$$
and "$\exists n \in \mathbb{N}$ s.t. $\omega \in A_k \ \forall k \geq n$" just means "$\omega \in A_n$ for all but finitely many n" in words.

    ii. Similar to above, and we can interpret "$\forall n \in \mathbb{N} \ \exists k \geq n$ s.t. $\omega \in A_k$" as "$\omega \in A_n$ for infinitely many n". $\qquad\square$

(b) *(Relating limit inferior and limit superior)*

    i. $\liminf_{n \to \infty} A_n \subseteq \limsup_{n \to \infty} A_n$.

    ii. $(\liminf_{n \to \infty} A_n)^c = \limsup_{n \to \infty} A_n^c$.

    iii. $(\limsup_{n \to \infty} A_n)^c = \liminf_{n \to \infty} A_n^c$.

Proof.

    i. Because "$\omega \in A_n$ for all but finitely many n" is just a special case of "$\omega \in A_n$ for infinitely many n".

    ii. Apply De Morgan's laws twice:
$$\left(\liminf_{n \to \infty} A_n\right)^c = \left(\bigcup_{n=1}^{\infty} \bigcap_{k \geq n} A_k\right)^c \overset{\text{(DM)}}{=} \bigcap_{n=1}^{\infty} \left(\bigcap_{k \geq n} A_k\right)^c \overset{\text{(DM)}}{=} \bigcap_{n=1}^{\infty} \left(\bigcup_{k \geq n} A_k^c\right) = \limsup_{n \to \infty} A_n^c.$$

    iii. Again apply De Morgan's laws twice:
$$\left(\limsup_{n \to \infty} A_n\right)^c = \left(\bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k\right)^c \overset{\text{(DM)}}{=} \bigcup_{n=1}^{\infty} \left(\bigcup_{k \geq n} A_k\right)^c \overset{\text{(DM)}}{=} \bigcup_{n=1}^{\infty} \left(\bigcap_{k \geq n} A_k^c\right) = \liminf_{n \to \infty} A_n^c.$$

$\qquad\square$

(c) *(Limits of monotone sequences of sets)*

    i. If $A_n \nearrow$, then $\lim_{n\to\infty} A_n$ exists and equals $\bigcup_{k=1}^{\infty} A_k$.
    [Note: Setting $A_n := \bigcup_{i=1}^{n} B_i$, since $A_n \nearrow$, we can get the intuitively appealing equality $\lim_{n\to\infty} \bigcup_{i=1}^{n} B_i = \bigcup_{i=1}^{\infty} B_i$, as $\bigcup_{i=1}^{\infty} B_i = \bigcup_{k=1}^{\infty}\left(\bigcup_{i=1}^{k} B_k\right)$.]

    ii. If $A_n \searrow$, then $\lim_{n\to\infty} A_n$ exists and equals $\bigcap_{k=1}^{\infty} A_k$.

    iii. For any collection $\{A_n\} \subseteq \mathcal{P}(\Omega)$, $\liminf_{n\to\infty} A_n = \lim_{n\to\infty}(\inf_{k\geq n} A_k)$ and $\limsup_{n\to\infty} A_n = \lim_{n\to\infty}(\sup_{k\geq n} A_k)$. [Note: This explains the rationale behind the notations "$\liminf_{n\to\infty} A_n$" and "$\limsup_{n\to\infty} A_n$".]

Proof.

    i. Assuming $A_n \nearrow$, we have $A_k \subseteq \bigcap_{i=k}^{\infty} A_i$. Hence,

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \subseteq \boxed{\bigcup_{k=1}^{\infty} A_k} \subseteq \bigcup_{n=1}^{\infty} \bigcap_{i=k}^{\infty} A_i = \liminf_{n\to\infty} A_n \subseteq \limsup_{n\to\infty} A_n.$$

    This forces $\liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n = \bigcup_{k=1}^{\infty} A_k$, as desired.

    ii. Let $B_n = A_n^c$ for every n, then $B_n \nearrow$. Applying (i), we get

$$\liminf_{n\to\infty} B_n = \limsup_{n\to\infty} B_n = \bigcup_{k=1}^{\infty} B_k.$$

    By [1.1.5]b, we have $\liminf_{n\to\infty} B_n = (\limsup_{n\to\infty} A_n)^c$ and $\limsup_{n\to\infty} B_n = (\liminf_{n\to\infty} A_n)^c$. It then follows that

$$\liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n = \left(\bigcup_{k=1}^{\infty} B_k\right)^c = \bigcap_{k=1}^{\infty} A_k^c.$$

    iii. Note that $\inf_{k\geq n} A_k \nearrow$ and $\sup_{k\geq n} A_k \searrow$. Hence,

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k = \bigcup_{n=1}^{\infty} \inf_{k\geq n} A_k \overset{(i)}{=} \lim_{n\to\infty}\left(\inf_{k\geq n} A_k\right)$$

    and

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k = \bigcap_{n=1}^{\infty} \sup_{k\geq n} A_k \overset{(ii)}{=} \lim_{n\to\infty}\left(\sup_{k\geq n} A_k\right).$$

$\square$

In view of [1.1.5]c, we sometimes write $A_n \nearrow A$ ($A_n \searrow A$) to mean A_n is increasing (decreasing) and $\lim_{n\to\infty} A_n = A$, i.e., $\bigcup_{k=1}^{\infty} A_k = A$ ($\bigcap_{k=1}^{\infty} A_k = A$).

1.1.6 Equivalence relations. Concepts related to equivalence relation have been discussed in MATH2012; here we will go through some key points:

- An *equivalence relation* is a relation $\sim$ on a set A that is (i) reflexive: $a \sim a$, (ii) symmetric: $a \sim b \implies b \sim a$, and (iii) transitive: $a \sim b$ and $b \sim c \implies a \sim c$. (where a, b, c are all arbitrary elements in A).

- The *equivalence class* of $a \in A$ under an equivalence relation $\sim$ on A is $[a] := \{x \in A : x \sim a\}$, i.e., the set of all elements in A that are "equivalent" to a under $\sim$.

- Property: Two equivalence classes are either identical or disjoint, and $[a] = [b]$ iff $a \sim b$.

- Property: The set of all distinct equivalence classes, $P = \{[a] : a \in A\}$, forms a partition of A (i.e., the classes in P are pairwise disjoint and their union is A).

1.1.7 **Indicator functions.** You should have learnt what an indicator function is in your first probability course. This function continues to be of great use here, so let us review it a bit.

- The *indicator function* of $A \subseteq \Omega$ is given by

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise.} \end{cases}$$

- Property: $\mathbf{1}_A \leq \mathbf{1}_B$ (i.e., $\mathbf{1}_A(\omega) \leq \mathbf{1}_B(\omega)$ for all $\omega \in \Omega$) iff $A \subseteq B$.

[Note: In some cases, we would use the notation $\mathbf{1}_{\{\text{statement}\}}$ to denote an indicator variable that is 1 if the statement is true, and 0 if it is not. For example, $\mathbf{1}_{\{x \geq 0\}} = 1$ if $x \geq 0$ and $\mathbf{1}_{\{x \geq 0\}} = 0$ otherwise. While it has a similar notation and carries similar meaning as the indicator function, it may not be a function of ω anymore; in this example, the indicator $\mathbf{1}_{\{x \geq 0\}}$ becomes a function of x.]

Indicator function can be applied for describing $\limsup_{n \to \infty} A_n$ and $\liminf_{n \to \infty} A_n$:

$$\limsup_{n \to \infty} A_n = \left\{ \omega \in \Omega : \sum_{n=1}^{\infty} \mathbf{1}_{A_n}(\omega) = \infty \right\}, \quad \liminf_{n \to \infty} A_n = \left\{ \omega \in \Omega : \sum_{n=1}^{\infty} \mathbf{1}_{A_n^c}(\omega) < \infty \right\}.$$

This is because when $\omega \in A_n$ infinitely often, infinitely many $\mathbf{1}_{A_n}(\omega)$ would equal one; when $\omega \in A_n$ for all but finitely many n, finitely many $\mathbf{1}_{A_n^c}(\omega)$ would equal one.

1.1.8 **Images and preimages.** A pair of concepts that will frequently appear in our discussion of measure theory is *image* and *preimage*, which is again covered in MATH2012.

Let Ω and Ω' be two sets, and $X : \Omega \to \Omega'$ be a function. The *image* of $A \subseteq \Omega$ under X is $X(A) := \{X(\omega) : \omega \in A\}$, and the *preimage* of $A' \subseteq \Omega'$ under is $X^{-1}(A') := \{\omega \in \Omega : X(\omega) \in A'\}$.
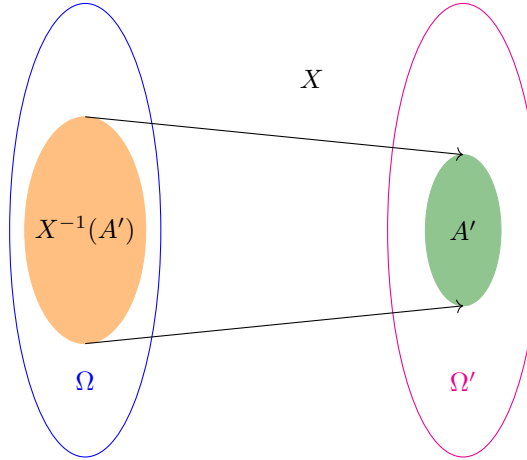
Let $\mathcal{A}'$ be a family of sets in Ω' (i.e., a subset of $\mathcal{P}(\Omega')$). Then, the notation $X^{-1}(\mathcal{A}')$ refers to the set of all preimages of sets in $\mathcal{A}$, i.e., $\{X^{-1}(A') : A' \in \mathcal{A}'\}$.

1.1.9 **Properties of preimages.** Let A' be any subset of Ω', and $\{A_i'\}_{i \in I}$ be any collection of subsets of Ω'.

(a) *(Preservation of complementation)* $\left(X^{-1}(A')\right)^c = X^{-1}(A'^c)$.

(b) *(Preservation of union)* $X^{-1}(\bigcup_{i \in I} A_i') = \bigcup_{i \in I} X^{-1}(A_i')$.

(c) *(Preservation of intersection)* $X^{-1}(\bigcap_{i \in I} A_i') = \bigcap_{i \in I} X^{-1}(A_i')$.

(d) *(Monotonicity)* Let $\mathcal{A}', \mathcal{B}' \subseteq \mathcal{P}(\Omega')$. Then, $\mathcal{A}' \subseteq \mathcal{B}' \implies X^{-1}(\mathcal{A}') \subseteq X^{-1}(\mathcal{B}')$.

Proof. We demonstrate the proof for (b) and (d) here and leave the rest as exercises.

(b) "⊆": Fix any $\omega \in X^{-1}(\bigcup_{i \in I} A_i')$. By definition, $X(\omega) \in \bigcup_{i \in I} A_i'$, thus there exists some $i \in I$ such that $X(\omega) \in A_i'$, or $\omega \in X^{-1}(A_i')$. This means $\omega \in \bigcup_{i \in I} X^{-1}(A_i')$.

"⊇": Highly similar to "⊆"; just work backward.

(d) Assume $\mathcal{A}' \subseteq \mathcal{B}'$. Now, fix any $A \in X^{-1}(\mathcal{A}')$. By definition, we have $A = X^{-1}(A')$ for some $A' \in \mathcal{A}'$. Since $\mathcal{A}' \subseteq \mathcal{B}'$, we also have $A' \in \mathcal{B}'$, meaning that $A \in \{X^{-1}(A') : A' \in \mathcal{B}'\} = X^{-1}(\mathcal{B}')$.

$\square$

1.2 Non-Measurable Sets

1.2.1 A critical concept in measure theory is *measurable set*. Roughly speaking, it refers to any set whose "volume" (or "length", or "area", or *probability* later...) can be measured. A natural and important question is then, are there really sets that *cannot* be measured? The answer is yes, and actually the existence of *non-measurable* sets leads to many interesting and fruitful developments in measure theory.

1.2.2 Before discussing non-measurable sets, let us first consider what is meant by "measure" intuitively[3]. First, we can view a measure as a function λ that assigns a value ("volume") to a set. For simplicity, let us focus on the case where the universal set is $\Omega = \mathbb{R}$ here.

A "reasonable measure" $\lambda : \mathcal{F} \to [0, \infty]$[4] should satisfy the following:

(1) *(Assigning to an interval its length)* $\lambda((a, b]) = b - a$ for any $a, b \in \mathbb{R}$ with $a \le b$. Here, $(a, b]$ can be replaced by (a, b), $[a, b)$, or $[a, b]$.

(2) *(Invariant under translation, rotations, and reflections)* λ assigns the same value to *congruent* sets (i.e., one can be obtained from another using just translations, rotations, and reflections).

(3) *(σ-additivity)* Given any pairwise disjoint collection $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F}$, we have $\lambda(\biguplus_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \lambda(A_i)$. [Note: While (finite) *additivity* appears to be a more natural requirement (obtained by changing $\mathbb{N} \to \{1, \ldots, n\}$ and $\infty \to n$ above), the "σ-" part is necessary for working with limiting argument.
On the other hand, requiring *uncountable* additivity would be way too strong since in such case we would have, for all $A \in \mathcal{F}$,

$$\lambda(A) = \lambda\left(\bigcup_{x \in A} \{x\}\right) \overset{\text{(uncountable add.)}}{=} \sum_{x \in A} \lambda(\{x\}) := \sup_{B \subseteq A,\ B \text{ finite}} \sum_{x \in B} \underbrace{\lambda(\{x\})}_{\lambda([x,x])=x-x=0} = 0,$$

which means that *every* set has a "zero volume"!]

Based on these requirements, we can deduce:

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

(b) *(Additivity)* For any $\{A_i\}_{i=1}^{n} \subseteq \mathcal{P}(\mathbb{R})$, we have $\lambda(\biguplus_{i=1}^{n} A_i) = \sum_{i=1}^{n} \lambda(A_i)$.

(c) *(Monotonicity)* For any $A \subseteq B$, we have $\lambda(A) \le \lambda(B)$.

Proof.

(a) $\lambda(\varnothing) = \lambda((a, a]) = a - a = 0$.

(b) $\lambda(\biguplus_{i=1}^{n} A_i) = \lambda(\biguplus_{i=1}^{n} A_i \uplus \varnothing \uplus \varnothing \uplus \cdots) \overset{\text{(σ-add.)}}{=} \sum_{i=1}^{n} \lambda(A_i) + \sum_{i=n+1}^{\infty} 0 = \sum_{i=1}^{n} \lambda(A_i)$.

(c) $\lambda(B) = \lambda(B \uplus (B \setminus A)) = \lambda(A) + \underbrace{\lambda(B \setminus A)}_{\ge 0} \ge \lambda(A)$.

$\square$

---

[3] We will define *measure* more formally later.

[4] The domain $\mathcal{F}$ is a family of sets in $\Omega = \mathbb{R}$. We choose the codomain as $[0, \infty]$ since it makes sense for a "volume" to be always nonnegative, and we sometimes want to assign "infinite volume".

1.2.3   Now we are ready to start discussing non-measurable sets. The idea is that, if every set was measurable and there were not non-measurable sets, then we could simply set $\mathcal{F}$ as $\mathcal{P}(\mathbb{R})$, allowing every subset of $\mathbb{R}$ to be measured by λ (being "measurable" in this sense). Now, the issue is that, setting $\mathcal{F} = \mathcal{P}(\mathbb{R})$ would actually lead to a contradiction 🔺 as suggested by the following theorem:

Theorem 1.2.a (Vitali's theorem). There is no λ defined on $\mathcal{P}(\mathbb{R})$ satisfying (1)-(3) in [1.2.2].

Proof. The proof is by explicitly constructing a set V (known as **Vitali set**) living in $\mathcal{P}(\mathbb{R})$ such that for every λ on $\mathcal{P}(\mathbb{R})$ satisfying (a)-(c), any choice of value for $\lambda(V)$ would lead to a contradiction.

Defining equivalence relation. We start with the interval $[0, 1]$ and define an equivalence relation $\sim$ on $[0, 1]$ by $x \sim y$ if $x - y \in \mathbb{Q}$, or in words, two values in $[0, 1]$ are "equivalent" (under $\sim$) if they have a rational difference. (Check $\sim$ is indeed an equivalence relation!)

Constructing Vitali set V. Then, the *Vitali set V* is constructed by picking exactly one element from each distinct equivalence class under $\sim$, and collecting them together (possible under the axiom of choice!). Symbolically, we can write:

$$V = \{v \in [0, 1] : \forall x \in [0, 1] \; \exists! v \in [x]\}.$$

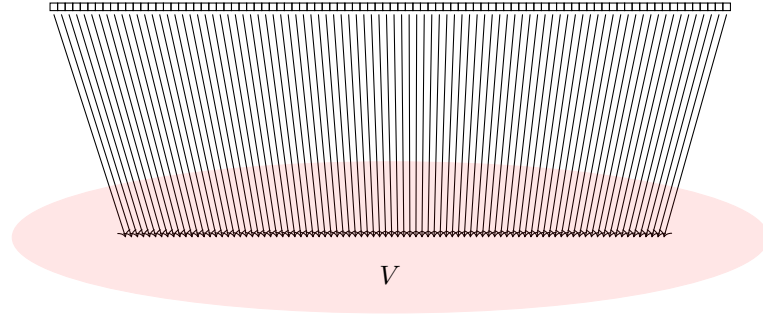[Note: Although the condition "$\forall x \in [0, 1] \; \exists! v \in [x]$" does not only consider distinct classes (it involves all possible equivalence classes indeed), it does not matter because for identical equivalence classes, the condition would be satisfied by the same unique element v picked.]

Distinct equivalence classes

V

Shifting Vitali set V by rationals. Next, we enumerate $\mathbb{Q} \cap [-1, 1]$ in a *diagonal* way (used for proving the well-known result that $\mathbb{Q}$ is countable) with repeated values or values lying outside $[-1, 1]$ being skipped, to form an infinite sequence $\{q_k\}_{k \in \mathbb{N}}$ of *distinct* rational numbers in $[-1, 1]$.

The rational numbers in $\{q_k\}$ are then used to "shift" the Vitali set V to yield $V_k := V + q_k := \{v + q_k : v \in V\}$ for every $k \in \mathbb{N}$.

Proving two auxiliary claims.

Claim 1: $[0, 1] \subseteq \bigcup_{k=1}^{\infty} V_k \subseteq [-1, 2]$.

Proof. By construction, $V_k \subseteq [-1, 2]$ for every $k \in \mathbb{N}$, thus $\bigcup_{k=1}^{\infty} V_k \subseteq [-1, 2]$. Next, fix any $x \in [0, 1]$. By construction of V, we have $x \in [v]$ for some $v \in V$, which means that for some $k \in \mathbb{N}$, $x - v = q_k$ or $x = v + q_k \in V_k$. This shows that $[0, 1] \subseteq \bigcup_{k=1}^{\infty} V_k$. □

Claim 2: V_k's are pairwise disjoint.

Proof. Suppose not, then there is $x \in V_k \cap V_j$ for some $k \neq j$. Then, we can write $x = v_k + q_k$ and $x = v_j + q_j$ for some $v_k, v_j \in V$ and *distinct* $q_k, q_j \in \mathbb{Q} \cap [-1, 1]$. This implies that $v_k = x - q_k \neq x - q_j = v_j$. However, we have $v_k - x = -q_k \in \mathbb{Q}$ and $v_j - x = -q_j \in \mathbb{Q}$, thus $v_k \sim x$ and $x \sim v_j$, which implies $v_k \sim v_j$ by transitivity (i.e., v_k and v_j are in the same equivalence class). This means V contains two *distinct* elements in the *same* equivalence class, contradiction. □

Showing the contradiction. With the two claims established, we can write $[0,1] \subseteq \biguplus_{k=1}^{\infty} V_k \subseteq [-1,2]$. Then, by monotonicity of λ, we have

$$1 = \lambda([0,1]) \leq \lambda\left(\biguplus_{k=1}^{\infty} V_k \right) \leq \lambda([-1,2]) = 3.$$

Applying σ-additivity on $\lambda(\biguplus_{k=1}^{\infty} V_k)$ gives $1 \leq \sum_{k=1}^{\infty} \lambda(V_k) \leq 3$. Note that $\lambda(V_k) = \lambda(V)$ by the invariance under translation, so $\sum_{k=1}^{\infty} \lambda(V_k) = \sum_{k=1}^{\infty} \lambda(V)$. But then this sum is either 0 (when $\lambda(V) = 0$) or ∞ (when $\lambda(V) > 0$). Neither is between 1 and 3, contradiction. $\qquad\square$

The Vitali set V in the proof is an example of *non-measurable* set. Assigning a "volume" to V would break the mathematics. This suggests that the power set $\mathcal{P}(\mathbb{R})$ is "too large" and contains some "pathological" sets like the Vitali set. This calls for the need to define λ on only a selected collection of sets, which can be constructed using the concepts to be introduced in Section 1.3.

1.2.4 Another interesting example of non-measurable set is given by the *Banach-Tarski paradox*.

Theorem 1.2.b (Banach-Tarski paradox)**.** Let $d \geq 3$ be an integer, and $A, B \subseteq \mathbb{R}^d$ be any bounded sets with nonempty interior. Then, for some $k \in \mathbb{N}$, there exist pairwise disjoint collections $\{A_i\}_{i=1}^{k}$ and $\{B_i\}_{i=1}^{k}$ such that $A = \biguplus_{i=1}^{k} A_i$ and $B = \biguplus_{i=1}^{k} B_i$ where A_i and B_i are congruent for every $i = 1, \ldots, k$.

Proof. It is quite technical and hence omitted. $\qquad\square$

The paradoxical feature of Theorem 1.2.b is that while there is no limitation on how the sizes/"volumes" of A and B can differ, e.g., A can be a very small "pea •" and B can be the very large "sun ☀", each of them can always be decomposed into finitely many pieces where every pair of pieces is *congruent* ("same in size")⚠. Informally, we may say that *"a pea • can be chopped up and reassembled into the sun ☀"*. What went wrong here?

The main issue is again related to non-measurable sets. We claim that at least one set in the collections $\{A_i\}_{i=1}^{k}$ and $\{B_i\}_{i=1}^{k}$ is non-measurable. Suppose not, then additivity would imply that $\lambda(A) = \lambda(\biguplus_{i=1}^{k} A_i) = \sum_{i=1}^{k} \lambda(A_i) \overset{\text{(invariance)}}{=} \sum_{i=1}^{k} \lambda(B_i) = \lambda(\biguplus_{i=1}^{k} B_i) = \lambda(B)$. But we can select A and B that get different assigned values by λ. Contradiction.

1.3 Systems of Sets

1.3.1 *Systems of sets* are utilized for constructing families of "selected" sets on which a "reasonable measure" λ can be defined consistently without having any issue (those sets are *measurable sets*; see [2.1.1] for the formal definition). There are multiple systems of sets here, but they all share a common theme, which is about constructing a family $\mathcal{A}$ of sets that is *closed* under certain set operations, i.e., performing these operations on sets in $\mathcal{A}$ would not yield something outside $\mathcal{A}$ — being "stable" in some sense. Intuitively, we are interested in this kind of properties because they can make the families of sets "rich enough", in the sense that the families contain "sufficiently many well-behaved sets".

The systems of sets to be discussed here are: (i) semiring, (ii) ring, (iii) algebra, and (iv) ⭐ σ-algebra (important concept for probability theory!).

1.3.2 **Definitions.**

- A family $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is a **semiring** on Ω if:
  (1) $\varnothing \in \mathcal{A}$.
  (2) *("Stable" under set differences)* $A, B \in \mathcal{A} \implies A \setminus B = \biguplus_{i=1}^{n} A_i$ for some pairwise disjoint $A_1, \ldots, A_n \in \mathcal{A}$, where $n \in \mathbb{N}$.
  (3) *(Closed under intersections)* $A, B \in \mathcal{A} \implies A \cap B \in \mathcal{A}$.

  [Note: The "stability" under set differences suggests that, while taking set differences may yield something outside $\mathcal{A}$, we can still express the result as a disjoint union of finitely many sets in $\mathcal{A}$ (so still "stable" to a certain degree).]

9

- A family $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is a **ring** on Ω if:
  1. $\varnothing \in \mathcal{A}$.
  2. *(Closed under set differences)* $A, B \in \mathcal{A} \implies A \setminus B \in \mathcal{A}$.
  3. *(Closed under unions)* $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A}$.
- A family $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is an **algebra** (or **field**) on Ω if:
  1. $\varnothing \in \mathcal{A}$.
  2. *(Closed under complementations)* $A \in \mathcal{A} \implies A^c = \Omega \setminus A \in \mathcal{A}$.
  3. *(Closed under unions)* $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A}$.
- ⭐ A family $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is a **σ-algebra** (or **σ-field**) on Ω if:
  1. $\varnothing \in \mathcal{A}$.
  2. *(Closed under complementations)* $A \in \mathcal{A} \implies A^c = \Omega \setminus A \in \mathcal{A}$.
  3. *(Closed under countable unions)* $A_1, A_2, \ldots \in \mathcal{A} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

  [Note: A σ-algebra is actually also closed under countable *intersections*. To show this, we can use De Morgan laws, and the closedness under countable unions and complementations:

  $$A_1, A_2, \ldots \in \mathcal{A} \implies \bigcap_{i=1}^{\infty} A_i \overset{\text{(DM)}}{=} \left(\bigcup_{i=1}^{\infty} A_i^c \right)^c \in \mathcal{A}.$$

  ]

1.3.3 Relationships between systems of sets.
It turns out that these four systems are imposing stronger requirements in the order of introduction, resulting in this chain of relationships: σ-algebra $\subsetneq$ algebra $\subsetneq$ ring $\subsetneq$ semiring, or in words, every σ-algebra is an algebra; every algebra is a ring; etc., with the inclusion being strict, i.e., some algebra is not σ-algebra; some ring is not algebra; etc.

Proposition 1.3.a (Relationships between systems of sets)**.**

(a) σ-algebra $\subsetneq$ algebra $\subsetneq$ ring $\subsetneq$ semiring.

(b) An algebra on a *finite* set Ω is a σ-algebra.

(c) A semiring that is also closed under unions is a ring.

Proof.

(a)
- σ-algebra $\subsetneq$ algebra
  Inclusion: Fix any σ-algebra $\mathcal{A}$ on Ω, and any $A, B \in \mathcal{A}$. Let $A_1 = A$, $A_2 = B$, and $A_3 = A_4 = \cdots = \varnothing$. Applying closedness under countable unions, we get $A \cup B = \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$. This means $\mathcal{A}$ is closed under unions, thus is an algebra.
  Strictness: Take $\Omega = (0, 1]$ and $\mathcal{A} = \{ \biguplus_{i=1}^{n} (a_i, b_i] : 0 \leq a_i \leq b_i \leq 1, n \in \mathbb{N} \}$, i.e., the family of all possible finite disjoint unions of intervals of the form $(a, b]$ lying in $\Omega = (0, 1]$.
  It can be directly checked that $\mathcal{A}$ is an algebra.

$$(a_1, b_1] \uplus (a_2, b_2]$$

$$\big((a_1, b_1] \uplus (a_2, b_2] \big)^c$$
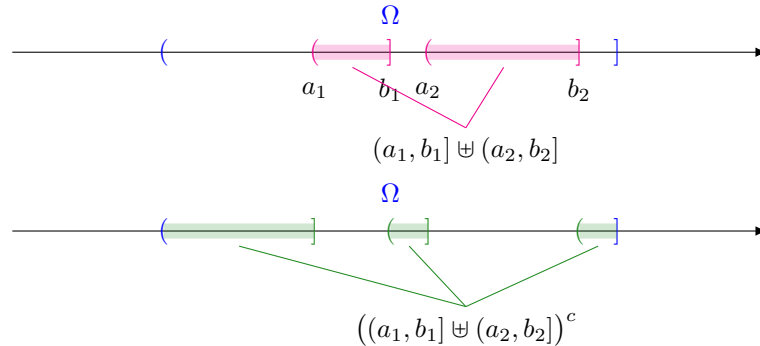
However, $\mathcal{A}$ is not a σ-algebra because we have $(0,1) = \bigcup_{n=1}^{\infty}(0, 1-1/n]$ with $(0, 1-1/n] \in \mathcal{A}$ for every n, but $(0,1) \notin \mathcal{A}$ as finite disjoint union of left-open and right-closed intervals must still be left-open and right-closed.

- algebra $\subsetneq$ ring
  Inclusion: Fix any algebra $\mathcal{A}$ on Ω, and any $A, B \in \mathcal{A}$. Applying closedness under complementations and unions, we have $A \setminus B = A \cap B^c \overset{(DM)}{=} (A^c \cup B)^c \in \mathcal{A}$. This shows $\mathcal{A}$ is closed under set differences, and hence is a ring.
  Strictness: Take $\Omega = \{1\}$ and $\mathcal{A} = \{\varnothing\}$. It is not hard to see that $\mathcal{A}$ is a ring, but $\mathcal{A}$ is not an algebra since $\varnothing^c = \Omega = \{1\} \notin \mathcal{A}$.

- ring $\subsetneq$ semiring
  Inclusion: Fix any ring $\mathcal{A}$ on Ω, and any $A, B \in \mathcal{A}$. Applying closedness under set differences, we get $A \cap B \overset{(DM)}{=} A \cap (A \cap B^c)^c = A \setminus (A \setminus B) \in \mathcal{A}$, and also $A \setminus B = \underbrace{(A \setminus B)}_{\in \mathcal{A}} \uplus \underbrace{\varnothing}_{\in \mathcal{A}}$.
  Strictness: Take $\Omega = \{1, 2\}$ and $\mathcal{A} = \{\varnothing, \{1\}, \{2\}\}$. It is straightforward to check that $\mathcal{A}$ is a semiring, but $\mathcal{A}$ is not a ring as $\{1\} \cup \{2\} = \{1, 2\} \notin \mathcal{A}$.

(b) Because every countably infinite union of sets in $\mathcal{A}$ can always be expressed as a finite union of sets in $\mathcal{A}$ in such case, by removing redundancies.

(c) Assuming a semiring $\mathcal{A}$ is closed under unions, by induction we have $A_1, \ldots, A_n \in \mathcal{A} \implies \bigcup_{i=1}^{n} A_i \in \mathcal{A}$ for all $n \in \mathbb{N}$. By the "stability" under set differences we have $A, B \in \mathcal{A} \implies A \setminus B = \biguplus_{i=1}^{n} A_i$, where $A_1, \ldots, A_n \in \mathcal{A}$ are pairwise disjoint. But then by the closedness under union, we have $\biguplus_{i=1}^{n} A_i \in \mathcal{A}$. Thus, $\mathcal{A}$ is closed under set differences, hence is a ring.

$\square$

1.3.4 Due to the importance of σ-algebra in probability theory, let us try to understand it better through some examples in the following. We start with the two simplest ones:

- The **trivial σ-algebra** on Ω is $\mathcal{F} = \{\varnothing, \Omega\}$. [Note: It is the *smallest* σ-algebra on Ω, i.e., every σ-algebra on Ω is a superset of the trivial σ-algebra. This is because containing $\varnothing$ and being closed under complementations would force a σ-algebra to at least contain $\varnothing$ and Ω.]

- The power set $\mathcal{F} = \mathcal{P}(\Omega)$ is a σ-algebra on Ω.
  Remarks:
    - It is a σ-algebra on Ω because complement of subset of Ω is still a subset of Ω, and countable union of subsets of Ω is still a subset of Ω.
    - It is the *largest* σ-algebra on Ω, i.e., every σ-algebra on Ω is a subset of $\mathcal{P}(\Omega)$ (which follows from definition).

1.3.5 **Constructing a σ-algebra from one set.** Start with a set $A \subseteq \Omega$. Then consider the following two sets that partition Ω:

$$A \qquad A^c$$

By putting zero/one/two of them into an union, we can get 4 combinations:

- *zero:* $\varnothing$
- *one:* A, A^c
- *two:* $A \cup A^c = \Omega$

These four sets form a σ-algebra: $\{\varnothing, A, A^c, \Omega\}$.

1.3.6 **Constructing a σ-algebra from two sets.** Here we start with two sets $A, B \subseteq \Omega$, where $A \nsubseteq B$, $B \nsubseteq A$, $A \cap B \neq \varnothing$, and $A \cup B \neq \Omega$. (These conditions are for ensuring the σ-algebra constructed below has no repeated members.)

Then consider the following $2^2 = 4$ sets that partition Ω: $A \cap B$, $A^c \cap B$, $A \cap B^c$, and $A^c \cap B^c$.

	A	A^c
B^c	$A \cap B^c$	$A^c \cap B^c$
B	$A \cap B$	$A^c \cap B$

By putting zero/one/two/three/four of them into an union, we can get $\binom{4}{0} + \binom{4}{1} + \binom{4}{2} + \binom{4}{3} + \binom{4}{4} = (1+1)^4 = 16$ combinations:

- *zero:* $\varnothing$
- *one:* $A \cap B$, $A^c \cap B$, $A \cap B^c$, $A^c \cap B^c$
- *two:*
  - $B = (A \cap B) \cup (A^c \cap B)$
  - $A = (A \cap B) \cup (A \cap B^c)$
  - $(A \cap B) \cup (A^c \cap B^c)$
  - $(A^c \cap B) \cup (A \cap B^c)$
  - $A^c = (A^c \cap B) \cup (A^c \cap B^c)$
  - $B^c = (A \cap B^c) \cup (A^c \cap B^c)$
- *three:* (just complement of each of the four sets in the partition indeed)
  - $A^c \cup B^c = (A \cap B)^c$
  - $A \cup B^c = (A^c \cap B)^c$
  - $A^c \cup B = (A \cap B^c)^c$
  - $A \cup B = (A^c \cap B^c)^c$
- *four:* Ω

These 16 sets form a σ-algebra.

[Note: In general, constructing a σ-algebra from n sets in the way above would yield 2^{2^n} ($\neq 4^n = 2^{2n}$) sets. This number grows *very* quickly as n rises; for instance, constructing from 4 sets would already yield 65536 sets! This tells us that it is almost impossible to "visualize" a σ-algebra in general (unfortunately!).]

1.3.7 The **countable-cocountable σ-algebra** on Ω is $\mathcal{F} = \{A \subseteq \Omega : A \text{ is countable or } A^c \text{ is countable}\}$. Let us prove that it is indeed a σ-algebra.

Proof. Since $\varnothing$ is countable, we have $\varnothing \in \mathcal{F}$.

Now fix any $A \in \mathcal{F}$.

- *Case 1: A is countable.* Then, $(A^c)^c = A$ is countable.
- *Case 2: A is uncountable.* Then, A^c has to be countable (or else A would not be in $\mathcal{F}$!).

Hence, we have A^c is countable or $(A^c)^c$ is countable, meaning that $A^c \in \mathcal{F}$.

Next, fix any $A_1, A_2, \ldots \in \mathcal{F}$.

- *Case 1: All A_i's are countable.* Then, enumerating all the elements in a diagonal way (e.g., see https://math.stackexchange.com/a/603499), we can deduce that $\bigcup_{i=1}^{\infty} A_i$ is also countable, hence $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

- *Case 2:* A_k *is uncountable for some* $k \in \mathbb{N}$. Then A_k^c has to be countable. Since $(\bigcup_{i=1}^{\infty} A_i)^c = \bigcap_{i=1}^{\infty} A_i^c \subseteq A_k^c$, the complement $(\bigcup_{i=1}^{\infty} A_i)^c$ is forced to be countable, thus $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

$\square$

1.3.8 Let $\mathcal{F}$ be an σ-algebra on Ω and $\Omega' \subseteq \Omega$. The **trace σ-algebra** of Ω' in $\mathcal{F}$ is $\mathcal{F}' = \mathcal{F}|_{\Omega'} := \{A \cap \Omega' : A \in \mathcal{F}\}$. Let us verify that the trace σ-algebra is indeed an σ-algebra below.

Proof. Firstly, we have $\varnothing = \varnothing \cap \Omega' \in \mathcal{F}'$.

Next, fix any $A' \in \mathcal{F}'$. By definition, we can write $A' = A \cap \Omega'$ for some $A \in \mathcal{F}$. Then, emphasizing universal sets in the complement notations, we have

$$(A')^{c_{\Omega'}} = (A \cap \Omega')^{c_{\Omega'}} \stackrel{\text{(DM)}}{=} A^{c_{\Omega'}} \cup \Omega'^{c_{\Omega'}} = A^{c_{\Omega'}} = \underbrace{A^{c_\Omega}}_{\in \mathcal{F}} \cap \Omega' \in \mathcal{F}'.$$

Finally, fix any $A_1', A_2', \ldots \in \mathcal{F}$. Then, for all $i \in \mathbb{N}$, we have $A_i' = A_i \cap \Omega'$ for some $A_i \in \mathcal{F}$. Hence,

$$\bigcup_{i=1}^{\infty} A_i' = \bigcup_{i=1}^{\infty} (A_i \cap \Omega') \stackrel{\text{(distributive)}}{=} \underbrace{\left(\bigcup_{i=1}^{\infty} A_i \right)}_{\in \mathcal{F}} \cap \Omega' \in \mathcal{F}'.$$

$\square$

1.3.9 Let $X : \Omega \to \Omega'$ be a function, and $\mathcal{F}'$ be a σ-algebra on Ω'. Then, the **σ-algebra generated by X** or **preimage σ-algebra** is

$$\mathcal{F} = \sigma(X) := X^{-1}(\mathcal{F}') = \{X^{-1}(A') : A' \in \mathcal{F}'\},$$

which is indeed a σ-algebra.

Proof. First, we have $\varnothing = X^{-1}(\underbrace{\varnothing}_{\in \mathcal{F}'}) \in \mathcal{F}$.

Now fix any $A \in \mathcal{F}$. Then $A = X^{-1}(A')$ for some $A' \in \mathcal{F}'$. Hence,

$$A^c = (X^{-1}(A'))^c \stackrel{\text{(preserv. comp.)}}{=} X^{-1}(\underbrace{(A')^c}_{\in \mathcal{F}'}) \in \mathcal{F}.$$

Next, fix any $A_1, A_2, \ldots \in \mathcal{F}$. Then for all $i \in \mathbb{N}$, $A_i = X^{-1}(A_i')$ for some $A_i' \in \mathcal{F}'$. Thus,

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} X^{-1}(A_i') \stackrel{\text{(preserv. union)}}{=} X^{-1}\left(\underbrace{\bigcup_{i=1}^{\infty} A_i'}_{\in \mathcal{F}'} \right) \in \mathcal{F}.$$

$\square$

1.3.10 The last example of σ-algebra here is about an important concept in the theory of stochastic process, called *filtration*. An example of filtration is an increasing sequence $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ of σ-algebras on Ω. It is useful for modelling *information* accrued over time.

Suppose we model an experiment of tossing a coin (countably) infinitely many times by setting the sample space as

$$\Omega = \{0, 1\}^\infty := \{\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots) : \omega_1, \omega_2, \ldots \in \{0, 1\}\}$$

with "0" and "1" being labels for "tails' and "heads" respectively, say. [Note: More formally, the notation "$(\omega_1, \omega_2, \ldots)$" actually refers to a *sequence* $\{\omega_n\}_{n \in \mathbb{N}}$, which is in turn defined as a *function* $\omega : \mathbb{N} \to \{0, 1\}$ with $\omega(n) =: \omega_n$ for each $n \in \mathbb{N}$. Thus Ω is indeed a set of functions to be precise.]

For each $n \in \mathbb{N}$, let

$$\mathcal{F}_n = \{\{\boldsymbol{\omega} \in \Omega : (\omega_1, \ldots, \omega_n) \in A\} : A \subseteq \{0, 1\}^n\},$$

the set of all events whose *occurrence can be decided after the first n tosses*. For example, the event "2nd toss is heads", $\{\boldsymbol{\omega} \in \Omega : \text{2nd toss is heads}\} = \{\boldsymbol{\omega} \in \Omega : \omega_2 = 1\}$, belongs to $\mathcal{F}_2$ but not $\mathcal{F}_1$. The set $\mathcal{F}_n$ can be viewed as describing the information we have about the true outcome $\boldsymbol{\omega} \in \Omega$ (how "precise" we can specify the true $\boldsymbol{\omega} \in \Omega$) after the first n tosses. For instance, the set $\mathcal{F}_1 = \{\varnothing, \{\boldsymbol{\omega} \in \Omega : \omega_1 = 0\}, \{\boldsymbol{\omega} \in \Omega : \omega_1 = 1\}, \Omega\}$ suggests that after the first toss, we can only decide whether the true $\boldsymbol{\omega}$ belongs to each of the sets in $\mathcal{F}_1$.

The property that $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ corresponds to the fact that "more information accrues over time". This increasing sequence is indeed a filtration as well:

Claim 1: $\mathcal{F}_n$ is a σ-algebra on Ω for each $n \in \mathbb{N}$.

Proof. Fix any $n \in \mathbb{N}$.

First, by taking $A = \varnothing$, we see that $\varnothing \in \mathcal{F}_n$. [Intuition 💡: We always know that the impossible event $\varnothing$ never occurs, so we can readily decide its occurrence regardless of the number of tosses made.]

Fix any $B \in \mathcal{F}_n$. Then we can write $B = \{\boldsymbol{\omega} \in \Omega : (\omega_1, \ldots, \omega_n) \in A\}$ for some $A \subseteq \{0,1\}^n$. Thus, $B^c = \Omega \setminus B = \{\boldsymbol{\omega} \in \Omega : (\omega_1, \ldots, \omega_n) \in A^c\}$ with $A^c = \{0,1\}^n \setminus A \subseteq \{0,1\}^n$, which means that $B^c \in \mathcal{F}_n$.

After that, fix any $B_1, B_2, \ldots \in \mathcal{F}_n$. For any $i \in \mathbb{N}$, we can write $B_i = \{\boldsymbol{\omega} \in \Omega : (\omega_1, \ldots, \omega_n) \in A_i\}$ for some $A_i \subseteq \{0,1\}^n$. Then, we have

$$\bigcup_{i=1}^{\infty} B_i = \{\boldsymbol{\omega} \in \Omega : (\omega_1, \ldots, \omega_n) \in A_i \text{ for some } i \in \mathbb{N}\} = \left\{\boldsymbol{\omega} \in \Omega : (\omega_1, \ldots, \omega_n) \in \underbrace{\bigcup_{i=1}^{\infty} A_i}_{\subseteq \{0,1\}^n} \right\} \in \mathcal{F}_n.$$

$\square$

Another set of interest is the union of all those "information sets" $\mathcal{F}_n$'s: $\mathcal{F} = \bigcup_{n=1}^{\infty} \mathcal{F}_n$, containing all "information" obtainable from finite number of coin tosses. While each $\mathcal{F}_n$ is a σ-algebra, the union $\mathcal{F}$ is *not* a σ-algebra, and is only an algebra.

Claim 2: $\mathcal{F}$ is an algebra on Ω, but not a σ-algebra on Ω.

Proof. Since $\mathcal{F}_1$ is a σ-algebra, we have $\varnothing \in \mathcal{F}_1 \subseteq \mathcal{F}$.

Next, fix any $A \in \mathcal{F}$. By definition of $\mathcal{F}$, we know $A \in \mathcal{F}_n$ for some $n \in \mathbb{N}$. By closedness under complementations, we have $A^c \in \mathcal{F}_n \subseteq \mathcal{F}$.

Now, fix any $A_1, A_2, \ldots, A_m \in \mathcal{F}$. Then, for every $i = 1, \ldots, m$, we have $A_i \in \mathcal{F}_{n_i}$ for some $n_i \in \mathbb{N}$. Since $\mathcal{F}_n \nearrow$, letting $n_{\max} = \max\{n_1, \ldots, n_m\} \in \mathbb{N}$, we can write $A_i \in \mathcal{F}_{n_{\max}}$ for every $i = 1, \ldots, m$. As $\mathcal{F}_{n_{\max}}$ is a σ-algebra, we have $\bigcup_{i=1}^{m} A_i \in \mathcal{F}_{n_{\max}} \subseteq \mathcal{F}$. This shows $\mathcal{F}$ is an algebra on Ω. [Note: Food for thought: Why does this argument not work for countable union?]

Now we proceed to show that $\mathcal{F}$ is not a σ-algebra on Ω. First let $A_i = \{\boldsymbol{\omega} \in \Omega : i\text{th toss is heads}\} = \{\boldsymbol{\omega} \in \Omega : \omega_i = 1\}$ for every $i \in \mathbb{N}$. Using the A_i's, we will show the failure of closedness under countable intersections for $\mathcal{F}$, which implies that $\mathcal{F}$ is not a σ-algebra.

First of all, note that $A_i \in \mathcal{F}_i \subseteq \mathcal{F}$ for every $i \in \mathbb{N}$. Then, consider the event "all tosses are heads", $\bigcap_{i=1}^{\infty} A_i = \{\boldsymbol{\omega} \in \Omega : \omega_i = 1 \text{ for all } i \in \mathbb{N}\}$. Since the occurrence of this event *cannot* be decided after any finite number of tosses, we have $\bigcap_{i=1}^{\infty} A_i \notin \mathcal{F}_n$ for all $n \in \mathbb{N}$, thus $\bigcap_{i=1}^{\infty} A_i \notin \mathcal{F}$. $\square$

1.4   More on σ-Algebra

1.4.1 **σ-algebras generated by families of sets.** Actually the σ-algebras mentioned in [1.3.5] and [1.3.6] are examples of σ-algebra *generated by a family of sets*. The one-set example is the σ-algebra generated by the family $\mathcal{A} = \{A\}$, and the two-set example is the σ-algebra generated by the family $\mathcal{A} = \{A, B\}$.

In general, let $\mathcal{A}$ be a family of subsets of Ω, and let $\mathcal{F}_{\mathcal{A}}$ be the set of all σ-algebras on Ω that contain $\mathcal{A}$, namely $\{\mathcal{F} : \mathcal{F} \text{ is an } \sigma\text{-algebra on } \Omega, \mathcal{F} \supseteq \mathcal{A}\}$, which is nonempty since it at least contains $\mathcal{P}(\Omega)$. Then the **σ-algebra generated by $\mathcal{A}$** is given by the intersection of all those σ-algebras, i.e., $\sigma(\mathcal{A}) := \bigcap_{\mathcal{F} \in \mathcal{F}_{\mathcal{A}}} \mathcal{F}$.

1.4.2 **Interpretation of $\sigma(\mathcal{A})$.** The family $\sigma(\mathcal{A})$ is the smallest/minimal σ-algebra on Ω containing $\mathcal{A}$, i.e., (i) $\sigma(\mathcal{A})$ is a σ-algebra on Ω that contains $\mathcal{A}$, and (ii) for any σ-algebra $\mathcal{F}'$ on Ω with $\mathcal{A} \subseteq \mathcal{F}'$, we have $\sigma(\mathcal{A}) \subseteq \mathcal{F}'$. [Note: ⭐ The property (ii) is often quite handy.]

Proof. $\sigma(\mathcal{A})$ **is a σ-algebra on Ω.** First, we have $\varnothing \in \mathcal{F}$ for every σ-algebra $\mathcal{F} \in \mathcal{F}_{\mathcal{A}}$, thus $\varnothing \in \bigcap_{\mathcal{F} \in \mathcal{F}_{\mathcal{A}}} = \sigma(\mathcal{A})$.

Now fix any $A \in \sigma(\mathcal{A}) = \bigcap_{\mathcal{F} \in \mathcal{F}_{\mathcal{A}}} \mathcal{F}$, so $A \in \mathcal{F}$ for all $\mathcal{F} \in \mathcal{F}_{\mathcal{A}}$. As $\mathcal{F}$ is a σ-algebra, we have $A^c \in \mathcal{F}$ for all $\mathcal{F} \in \mathcal{F}_{\mathcal{A}}$, thus $A^c \in \bigcap_{\mathcal{F} \in \mathcal{F}_{\mathcal{A}}} \mathcal{F} = \sigma(\mathcal{A})$.

Next, fix any $A_1, A_2, \ldots \in \sigma(\mathcal{A})$. Then for all $\mathcal{F} \in \mathcal{F}_{\mathcal{A}}$, we have $A_1, A_2, \ldots \in \mathcal{F}$, which implies $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$. Therefore, $\bigcup_{i=1}^{\infty} A_i \in \sigma(\mathcal{A})$.

$\sigma(\mathcal{A})$ **contains $\mathcal{A}$.** By definition, $\mathcal{A} \subseteq \mathcal{F}$ for all $\mathcal{F} \in \mathcal{F}_{\mathcal{A}}$, thus $\mathcal{A} \subseteq \bigcap_{\mathcal{F} \in \mathcal{F}_{\mathcal{A}}} \mathcal{F} = \sigma(\mathcal{A})$.

Smallest. For any σ-algebra $\mathcal{F}' \supseteq \mathcal{A}$, we have $\mathcal{F}' \in \mathcal{F}_{\mathcal{A}}$, thus $\sigma(\mathcal{A}) = \bigcap_{\mathcal{F} \in \mathcal{F}_{\mathcal{A}}} \mathcal{F} \subseteq \mathcal{F}'$. $\qquad\square$

1.4.3 **σ-algebras generated by partitions.** While we know that the σ-algebra generated by $\mathcal{A}$ can be obtained by finding $\bigcap_{\mathcal{F} \in \mathcal{F}_{\mathcal{A}}} \mathcal{F}$, it is usually quite hard to find what that intersection is, as $\mathcal{F}_{\mathcal{A}}$ can be rather large! In practice, it is more helpful to use the following shortcut for deducing $\sigma(\mathcal{A})$, by using a *partition* as the generator for the σ-algebra.

Lemma 1.4.a. Let $\mathcal{A} = \{A_i\}_{i \in \mathbb{N}}$ be a partition of Ω. Then, $\sigma(\mathcal{A})$ is the set of all possible disjoint unions of sets in $\mathcal{A}$, i.e., $\mathcal{F} := \{\biguplus_{i \in I} A_i : A_i \in \mathcal{A} \text{ for all } i \in I \text{ and all } I \subseteq \mathbb{N}\}$.

Proof. We first show that $\mathcal{F}$ is a σ-algebra on Ω:

- With I set as $\varnothing$, we know $\varnothing \in \mathcal{F}$.
- Fix any $A \in \mathcal{F}$. Then $A = \biguplus_{i \in I} A_i$ for some $I \subseteq \mathbb{N}$. Then $A^c = \Omega \setminus \biguplus_{i \in I} A_i = \biguplus_{i \in \mathbb{N} \setminus I} A_i \in \mathcal{F}$ as $\mathbb{N} \setminus I \subseteq \mathbb{N}$ still.
- Fix any $B_1, B_2, \ldots \in \mathcal{F}$. Then $B_k = \biguplus_{i \in I_k} A_i$ for some $I_k \subseteq \mathbb{N}$, for every $k \in \mathbb{N}$. Hence,

$$\bigcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{\infty} \bigcup_{i \in I_k} A_i = \bigcup_{i \in I_1} A_i \cup \bigcup_{i \in I_2} A_i \cup \cdots = \bigcup_{i \in I_1 \cup I_2 \cup \cdots} A_i \in \mathcal{F}$$

  as $I_1 \cup I_2 \cup \cdots \subseteq \mathbb{N}$.

After that, we show the two subset inclusions.

- $\sigma(\mathcal{A}) \subseteq \mathcal{F}$: For any $i \in \mathbb{N}$, we have $A_i \in \mathcal{F}$ by setting $I = \{i\}$. Hence $\mathcal{A} \subseteq \mathcal{F}$. As $\mathcal{F}$ is a σ-algebra, $\sigma(\mathcal{A}) \subseteq \mathcal{F}$ by the minimality of $\sigma(\mathcal{A})$.
- $\mathcal{F} \subseteq \sigma(\mathcal{A})$: Fix any $A \in \mathcal{F}$. Then $A = \biguplus_{i \in I} A_i$ for some $I \subseteq \mathbb{N}$ with $A_i \in \mathcal{A}$ for all $i \in I$. As $\mathcal{A} \subseteq \sigma(\mathcal{A})$, we have $A_i \in \sigma(\mathcal{A})$ for all $i \in I$. Finally, since $\sigma(\mathcal{A})$ is a σ-algebra, applying the closedness under countable unions gives $A = \biguplus_{i \in I} A_i \in \sigma(\mathcal{A})$.

$\qquad\square$

With Lemma 1.4.a, we can obtain the following shortcut for finding $\sigma(\mathcal{A})$:

(1) Find a partition $\mathcal{A}^*$ of Ω such that every set $A \in \mathcal{A}$ is a disjoint union of sets in $\mathcal{A}^*$, by taking (countable) intersections/complementations on the sets in $\mathcal{A}$.

(2) $\sigma(\mathcal{A})$ is equal to $\sigma(\mathcal{A}^*)$, which can be obtained conveniently by Lemma 1.4.a.

Indeed, the σ-algebras in [1.3.5] and [1.3.6] are constructed via this method.

Justification. First, by construction, $\mathcal{A} \subseteq \{\biguplus_{i \in I} A_i^* : A_i^* \in \mathcal{A}^* \text{ for all } i \in I \text{ and any } I \subseteq \mathbb{N}\} = \sigma(\mathcal{B})$. As $\sigma(\mathcal{A}^*)$ is a σ-algebra, we have $\sigma(\mathcal{A}) \subseteq \sigma(\mathcal{A}^*)$.

Next, by closedness under (countable) intersections/complementations, we have $\mathcal{A}^* \subseteq \sigma(\mathcal{A})$. As $\sigma(\mathcal{A})$ is a σ-algebra, we have $\sigma(\mathcal{A}^*) \subseteq \sigma(\mathcal{A})$. This shows $\sigma(\mathcal{A}) = \sigma(\mathcal{A}^*)$.

1.4.4 σ-algebra generated by unions and collections of functions.

- *Generated by unions:* Let $\mathcal{F}_i$ be a σ-algebra on Ω for all $i \in I$. Then the σ-algebra generated by their union $\bigcup_{i \in I} \mathcal{F}_i$ has a special notation: $\sigma(\mathcal{F}_i : i \in I) := \sigma(\bigcup_{i \in I} \mathcal{F}_i)$.
  [Note: Generally, the union $\bigcup_{i \in I} \mathcal{F}_i$ of σ-algebras is *not* a σ-algebra. For instance, let $\Omega = \{0, 1, 2\}$. Consider the two σ-algebras: $\mathcal{F}_1 = \{\varnothing, \{0\}, \{1, 2\}, \Omega\}$ and $\mathcal{F}_2 = \{\varnothing, \{1\}, \{0, 2\}, \Omega\}$ (verify that they are indeed σ-algebras!). The union $\mathcal{F}_1 \cup \mathcal{F}_2 = \{\varnothing, \{0\}, \{1\}, \{0, 2\}, \{1, 2\}, \Omega\}$ is not a σ-algebra since $\{0\}, \{1\} \in \mathcal{F}_1 \cup \mathcal{F}_2$ but $\{0, 1\} = \{0\} \cup \{1\} \notin \mathcal{F}_1 \cup \mathcal{F}_2$.]

- *Generated by collections of functions:* For every $i \in I$, let $X_i : \Omega \to \Omega_i$ be a function with a σ-algebra $\mathcal{F}_i$ on Ω_i. Then, the **σ-algebra generated by** $\{X_i\}_{i \in I}$ (the collection of all the functions X_i's) is defined by the σ-algebra generated by the unions of $\sigma(X_i)$'s: $\sigma(X_i : i \in I) := \sigma(\bigcup_{i \in I} \sigma(X_i)) = \sigma(\bigcup_{i \in I} X_i^{-1}(\mathcal{F}_i))$.

[Note: In the notations $\sigma(\mathcal{F}_i : i \in I)$ and $\sigma(X_i : i \in I)$, we may instead list out the $\mathcal{F}_i$'s/X_i's in the parenthesis if I is countable. For example, we can write $\sigma(\mathcal{F}_1, \mathcal{F}_2)$ instead of $\sigma(\mathcal{F}_i : i \in \{1, 2\})$.]

1.4.5 σ-algebra generated by union of two σ-algebras.
For the special case of σ-algebra generated by union of *two* σ-algebras, it has an alternative characterization. Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be σ-algebras on Ω. Then, the σ-algebra generated by their union is the σ-algebra generated by the family of all two-set intersections (not unions ⚠) across them, i.e.,

$$\boxed{\sigma(\mathcal{F}_1, \mathcal{F}_2) = \sigma(\{A_1 \cap A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\})}.$$

Proof. Let $\mathcal{A} = \{A_1 \cap A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}$.

- "$\subseteq$": Fix any $A \in \mathcal{A}$, which can then be written as $A = A_1 \cap A_2$ for some $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$. Since $\mathcal{F}_1, \mathcal{F}_2 \subseteq \mathcal{F}_1 \cup \mathcal{F}_2 \subseteq \sigma(\mathcal{F}_1, \mathcal{F}_2)$, we have $A_1, A_2 \in \sigma(\mathcal{F}_1, \mathcal{F}_2)$. Thus, $A = A_1 \cap A_2 \in \sigma(\mathcal{F}_1, \mathcal{F}_2)$. This means $\mathcal{A} \subseteq \sigma(\mathcal{F}_1, \mathcal{F}_2)$, and hence $\sigma(\mathcal{A}) \subseteq \sigma(\mathcal{F}_1, \mathcal{F}_2)$, as $\sigma(\mathcal{F}_1, \mathcal{F}_2)$ is a σ-algebra.

- "$\supseteq$": For every $A_1 \in \mathcal{F}_1$, we can write $A_1 = A_1 \cap \overbrace{\Omega}^{\in \mathcal{F}_2} \in \mathcal{A}$, hence $\mathcal{F}_1 \subseteq \mathcal{A}$. Similarly, we have $\mathcal{F}_2 \subseteq \mathcal{A}$. Thus, $\mathcal{F}_1 \cup \mathcal{F}_2 \subseteq \mathcal{A} \subseteq \sigma(\mathcal{A})$, which implies that $\sigma(\mathcal{F}_1, \mathcal{F}_2) = \sigma(\mathcal{F}_1 \cup \mathcal{F}_2) \subseteq \sigma(\mathcal{A})$ as $\sigma(\mathcal{A})$ is a σ-algebra.

$\square$

1.4.6 Product spaces.
A *product space* can be seen as a generalization to *Cartesian product*. For every $i \in I$, let Ω_i be a set (which may be interpreted as a sample space). Then, the **product space** Ω of the collection $\{\Omega_i\}_{i \in I}$ is

$$\Omega = \prod_{i \in I} \Omega_i := \left\{ \omega : I \to \bigcup_{i \in I} \Omega_i : \omega(i) \in \Omega_i \text{ for all } i \in I \right\},$$

a set of functions. This kind of set has appeared earlier in [1.3.10]; there we have

$$\Omega = \prod_{i=1}^{\infty} \Omega_i = \{\omega : \mathbb{N} \to \{0, 1\} : \omega(i) \in \{0, 1\} \text{ for all } i \in \mathbb{N}\} = \{\boldsymbol{\omega} = (\omega_1, \omega_2, \dots) : \omega_1, \omega_2, \dots \in \{0, 1\}\},$$

where $I = \mathbb{N}$ and $\Omega_i = \{0, 1\}$ for every $i \in \mathbb{N}$. [Note: Cartesian product of n sets $\Omega_1, \dots, \Omega_n$ can be regarded as a product space by identifying each vector $(\omega_1, \dots, \omega_n) \in \Omega_1 \times \cdots \times \Omega_n$ as a function $\omega : \{1, \dots, n\} \to \bigcup_{i=1}^{n} \Omega_i$, with $\omega(i) = \omega_i$ for each $i = 1, \dots, n$.]

1.4.7 σ-algebras on product spaces.
Let $\mathcal{F}_i$ be a σ-algebra on Ω_i for all $i \in I$. Naturally, one may guess that the product space $\prod_{i \in I} \mathcal{F}_i$ would be a σ-algebra on $\prod_{i \in I} \Omega_i$. However, this is unfortunately *not* the case in general, as the following example suggests.

Counter-example. Take $\Omega_1 = \Omega_2 = \{0,1\}$ with $\Omega = \Omega_1 \times \Omega_2 = \{(0,0),(0,1),(1,0),(1,1)\}$. Let $\mathcal{F}_i = \mathcal{P}(\Omega_i) = \{\varnothing, \{0\}, \{1\}, \{0,1\}\}$ be a σ-algebra on Ω_i for every $i = 1,2$, and let $\mathcal{F} := \mathcal{F}_1 \times \mathcal{F}_2$. Now consider $A_1 = \{0,1\} \times \{0,1\} = \Omega_1 \times \Omega_2$ and $A_2 = \{1\} \times \{1\}$ in $\mathcal{F}$. However, $A_1 \setminus A_2 = \{(0,0),(0,1),(1,0)\}$ does *not* belong to $\mathcal{F}$ as it cannot be expressed as $A_1' \times A_2'$ with $A_1' \in \mathcal{F}_1$ and $A_2' \in \mathcal{F}_2$.

Instead of the product space $\prod_{i \in I} \mathcal{F}_i$, we can construct a σ-algebra on $\prod_{i \in I} \Omega_i$ by using the *product σ-algebra*, which is generated by the projection mappings π_i's. Let us first define formally the concept of *projection*. For each $i \in I$, let $\pi_i : \Omega \to \Omega_i$ denote the **projection onto the ith coordinate**, i.e., $\pi_i(\omega) = \omega_i$ for all $\omega \in \Omega$, where ω_i is the "ith coordinate" of ω (that is, the value $\omega(i)$ in the function interpretation above).

Then, the **product σ-algebra** on Ω is

$$\bigotimes_{i \in I} \mathcal{F}_i := \sigma(\pi_i : i \in I) = \sigma\left(\bigcup_{i \in I} \sigma(\pi_i)\right) = \sigma\left(\bigcup_{i \in I} \pi_i^{-1}(\mathcal{F}_i)\right).$$

1.4.8 **Interpretation of product σ-algebra with countable I.** While the above definition of product σ-algebra may not seem to be natural, we do have a more intuitive interpretation when I is countable, which suggests why we have the name "*product σ-algebra*".

Proposition 1.4.b. If I is countable, then the product σ-algebra is $\bigotimes_{i \in I} \mathcal{F}_i = \sigma(\{\prod_{i \in I} A_i : A_i \in \mathcal{F}_i \ \forall i \in I\})$.

Proof.

- "$\subseteq$": For any $i \in I$,

$$\begin{aligned}
\pi_i^{-1}(\mathcal{F}_i) &= \{\pi_i^{-1}(A_i) : A_i \in \mathcal{F}_i\} \\
&= \{\omega \in \Omega : \omega(i) \in A_i, \ \omega(k) \in \Omega_k \text{ for all } k \neq i, \text{ and } A_i \in \mathcal{F}_i\} \\
&= \prod_{i \in I} A_i^* : A_i \in \mathcal{F}_i \qquad (A_i^* = A_i \text{ and } A_k^* = \Omega_k \text{ for all } k \neq i) \\
&\subseteq \prod_{i \in I} A_i : A_i \in \mathcal{F}_i \ \forall i \in I \\
&\subseteq \sigma\left(\prod_{i \in I} A_i : A_i \in \mathcal{F}_i \ \forall i \in I\right)
\end{aligned}$$

  As this holds for all $i \in I$, it follows that $\bigcup_{i \in I} \pi_i^{-1}(\mathcal{F}_i) \subseteq \sigma(\{\prod_{i \in I} A_i : A_i \in \mathcal{F}_i \ \forall i \in I\})$. Thus, $\sigma(\bigcup_{i \in I} \pi_i^{-1}(\mathcal{F}_i)) \subseteq \sigma(\{\prod_{i \in I} A_i : A_i \in \mathcal{F}_i \ \forall i \in I\})$.

- "$\supseteq$": Note first that

$$\prod_{i \in I} A_i = \{\omega \in \Omega : \omega(i) \in A_i \ \forall i \in I\} = \bigcap_{i \in I} \{\omega \in \Omega : \omega(i) \in A_i \text{ and } \omega(k) \in \Omega_k \text{ for all } k \neq i\} = \bigcap_{i \in I} \pi_i^{-1}(A_i).$$

  Now, as $\pi_i^{-1}(A_i) \in \pi^{-1}(\mathcal{F}_i) \subseteq \bigcup_{i \in I} \pi^{-1}(\mathcal{F}_i) \subseteq \sigma(\bigcup_{i \in I} \pi^{-1}(\mathcal{F}_i)) = \bigotimes_{i \in I} \mathcal{F}_i$ for every $i \in I$, by closedness under countable intersections (I needs to be countable here!) we have $\prod_{i \in I} A_i = \bigcap_{i \in I} \pi_i^{-1}(A_i) \in \bigotimes_{i \in I} \mathcal{F}_i$. Therefore, $\{\prod_{i \in I} A_i : A_i \in \mathcal{F}_i\} \subseteq \bigotimes_{i \in I} \mathcal{F}_i$.

$\square$

1.4.9 **Dynkin system and π-system.** To verify that a family is a σ-algebra, showing closedness under countable unions is often not so convenient as there can be many combinations to be checked. Here, we will introduce some new systems of sets that allow us to do that more conveniently, namely *Dynkin system* and π-system.

- *Dynkin system:* Its definition only has one difference from σ-algebra: Instead of requiring closedness under *countable unions*, it just requires closedness under *countable disjoint unions*, which is often easier to check.

  A family $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ is a **Dynkin system** (or λ-**system**) on Ω if:

  (1) $\varnothing \in \mathcal{D}$.

  (2) *(Closed under complementations)* $A \in \mathcal{D} \implies A^c = \Omega \setminus A \in \mathcal{D}$.

  (3) *(Closed under countable disjoint unions)* $A_1, A_2, \ldots \in \mathcal{D}$ pairwise disjoint $\implies \biguplus_{i=1}^{\infty} A_i \in \mathcal{D}$.

- π-*system:* It is more of an auxiliary system to the Dynkin system: It turns out that a Dynkin system that is also a π-system is equivalent to a σ-algebra (see Proposition 1.4.c).

  A family $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is a π-**system** on Ω if $A, B \in \mathcal{A} \implies A \cap B \in \mathcal{A}$ *(closed under intersections)*.

1.4.10 **Properties of Dynkin system.** The following result provides some alternative characterizations of Dynkin system and also suggests how it relates with σ-algebra.

Proposition 1.4.c (Properties of Dynkin systems).

(a) Let:

- (1') be "$\Omega \in \mathcal{D}$",
- (2') be "$A, B \in \mathcal{D}$ with $A \subseteq B \implies B \setminus A \in \mathcal{D}$",
- (3') be "$A_1, A_2, \ldots \in \mathcal{D}$ with $A_i \nearrow \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{D}$" *(closedness under countable unions of increasing sets)*.

Then the conditions (1), (2), and (3) for Dynkin system are equivalent to the conditions (1'), (2'), and (3').

(b) σ-algebra $\subsetneq$ Dynkin system.

(c) If $\mathcal{D}$ is a Dynkin system and *also a π-system*, then $\mathcal{D}$ is a σ-algebra.

Proof.

(a) - (1), (2), (3) $\implies$ (1'), (2'), (3'):
  Firstly, we have $\Omega = \varnothing^c \in \mathcal{D}$ by (1) and (2). This shows (1').

  Fix any $A, B \in \mathcal{D}$ with $A \subseteq B$. Then, $B \setminus A = B \cap A^c \overset{\text{(DM)}}{=} (B^c \uplus A)^c$. (Here A and B^c are disjoint since $A \subseteq B$.) Now, using the conditions (2) and (3), we have $(B^c \uplus A)^c \in \mathcal{D}$. This shows (2').

  Next, fix any $A_1, A_2, \ldots \in \mathcal{D}$ with $A_i \nearrow$. Let $A_0 := \varnothing \overset{(1)}{\in} \mathcal{D}$ and $B_n := A_n \setminus A_{n-1} \overset{(2')}{\in} \mathcal{D}$ for every $n \in \mathbb{N}$. Then we have $\bigcup_{n=1}^{\infty} A_n = \lim_{N \to \infty} \bigcup_{n=1}^{N} A_n = \lim_{N \to \infty} \biguplus_{n=1}^{N} B_n = \biguplus_{n=1}^{\infty} B_n \overset{(3)}{\in} \mathcal{D}$. This shows (3').
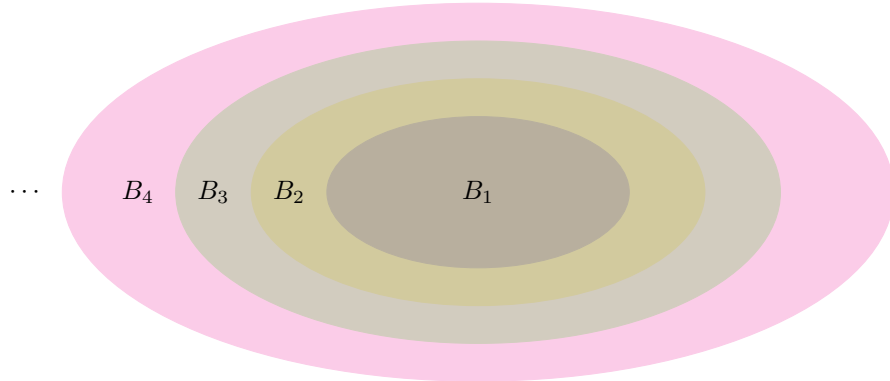
- (1'), (2'), (3') $\implies$ (1), (2), (3):
  Firstly, we have $\varnothing = \Omega \setminus \Omega \in \mathcal{D}$ by (1') and (2'). This shows (1).

18

For any $A \in \mathcal{D}$, we have $A^c = \Omega \setminus A \overset{(2')}{\in} \mathcal{D}$ as $A \subseteq \Omega$. This shows (2).

Claim: $A_1, A_2 \in \mathcal{D}$ pairwise disjoint $\implies A_1 \uplus A_2 \in \mathcal{D}$.

Proof. Since $A_1 \cap A_2 = \varnothing$, we have $A_2 \subseteq \Omega \setminus A_1$. By (1') and (2'), we know $\Omega \setminus A_1 \in \mathcal{D}$. Hence, by (2') again, we have $(\Omega \setminus A_1) \setminus A_2 = \Omega \setminus (A_1 \uplus A_2) \in \mathcal{D}$. Using (2') once more, this implies $A_1 \uplus A_2 = \Omega \setminus (\Omega \setminus (A_1 \uplus A_2)) \in \mathcal{D}$. $\qquad\square$

Fix any $A_1, A_2, \ldots \in \mathcal{D}$ that are pairwise disjoint. Let $B_n := \uplus_{i=1}^{n} A_i$ for every $n \in \mathbb{N}$. Then $B_n \in \mathcal{D}$ by the claim and induction, and $B_n \nearrow$. Thus $\uplus_{n=1}^{\infty} A_n = \lim_{N \to \infty} \uplus_{n=1}^{N} A_n = \lim_{N \to \infty} \bigcup_{n=1}^{N} B_n = \bigcup_{n=1}^{\infty} B_n \overset{(3')}{\in} \mathcal{D}$.

(b) **Inclusion:** Fix any σ-algebra $\mathcal{A}$ and any $A_1, A_2, \ldots \in \mathcal{A}$ that are pairwise disjoint. Applying the closedness under countable unions for σ-algebra, we have $\uplus_{i=1}^{\infty} A_i \in \mathcal{A}$.

Strictness: Take $\Omega = \{0, 1, 2, 3\}$, $A = \{0, 1\}$, and $B = \{1, 2\}$. Then the family $\mathcal{D} = \{\varnothing, A, B, A^c, B^c, \Omega\}$ is a Dynkin system (note that A and B are not pairwise disjoint), but is not a σ-algebra since $A \cup B = \{0, 1, 2\} \notin \mathcal{D}$.

(c) Fix any $A_1, A_2, \ldots \in \mathcal{D}$. Let $B_1 := A_1 \in \mathcal{D}$ and, for all $n = 2, 3, \ldots$, $B_n := A_n \setminus \bigcup_{i=1}^{n-1} A_i = A_n \cap (\bigcup_{i=1}^{n-1} A_i)^c \overset{(\text{DM})}{=} A_n \cap \bigcap_{i=1}^{n-1} A_i^c \overset{((2),\ \pi\text{-system})}{\in} \mathcal{D}$. Note that B_n's are pairwise disjoint and $\bigcup_{n=1}^{N} A_n = \uplus_{n=1}^{N} B_n$ for all $N \in \mathbb{N}$. Hence,

$$\bigcup_{n=1}^{\infty} A_i = \lim_{N \to \infty} \bigcup_{n=1}^{N} A_n = \lim_{N \to \infty} \uplus_{n=1}^{N} B_n = \uplus_{n=1}^{\infty} B_n \overset{(3)}{\in} \mathcal{D},$$

thus $\mathcal{D}$ is a σ-algebra.

$\qquad\square$

[Note: In the part (c) of the proof, the technique of constructing such pairwise disjoint B_n's out of general $A_1, A_2, \ldots$ is known as the **disjointification**, which is an useful tactic for measure theoretic proof.]

Because every σ-algebra is a Dynkin system, we indeed already know many examples of Dynkin system.

1.4.11 **Dynkin systems generated by families of sets.** Like σ-algebra, given a family $\mathcal{A}$ of subsets of Ω and letting $\mathcal{D}_{\mathcal{A}} = \{\mathcal{D} : \mathcal{D} \text{ is a Dynkin system on } \Omega, \ \mathcal{D} \supseteq \mathcal{A}\}$, the **Dynkin system generated by $\mathcal{A}$** is defined by $\delta(\mathcal{A}) := \bigcap_{\mathcal{D} \in \mathcal{D}_{\mathcal{A}}} \mathcal{D}$.

Interpretation. The family $\delta(\mathcal{A})$ carries a similar interpretation as before: $\delta(\mathcal{A})$ is the smallest Dynkin system on Ω containing $\mathcal{A}$, i.e., (i) $\delta(\mathcal{A})$ is a Dynkin system on Ω that contains $\mathcal{A}$, and (ii) for all Dynkin system $\mathcal{D}'$ on Ω with $\mathcal{A} \subseteq \mathcal{D}'$, we have $\delta(\mathcal{A}) \subseteq \mathcal{D}'$.

Proof. Analogous to the proof in [1.4.2]. $\qquad\square$

1.4.12 **Relationship between $\delta(\mathcal{A})$ and $\sigma(\mathcal{A})$.** The two families $\delta(\mathcal{A})$ and $\sigma(\mathcal{A})$ can be related by the following important theorem, which gives us a sufficient condition for them to be equal.

Theorem 1.4.d (Dynkin's π-λ theorem). If $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ is a Dynkin system containing a π-*system* $\mathcal{A}$, then (i) $\sigma(\mathcal{A}) \subseteq \mathcal{D}$, which particularly implies (ii) $\delta(\mathcal{A}) = \sigma(\mathcal{A})$.

An important strategy used in the proof is called the *principle of good sets*. It is also used in proving some other important theorems in measure theory. This principle is usually utilized for showing that a certain property holds for all elements of a σ-algebra $\mathcal{F}$.

Principle of good sets. Let $\mathcal{G}$ be the family of all "good" sets in $\mathcal{F}$, which can be written as $\mathcal{G} = \{A \in \mathcal{F} : A \text{ satisfies a certain property}\}$. If $\mathcal{G}$ is a σ-algebra that contains a generator $\mathcal{A}$ for $\mathcal{F}$, then we have $\mathcal{F} = \mathcal{G}$, meaning that every set in $\mathcal{F}$ is "good". [Note: This is because we have $\mathcal{F} = \sigma(\mathcal{A}) \subseteq \mathcal{G}$, and $\mathcal{G} \subseteq \mathcal{F}$ by construction.]

Remarks:

- The principle of good sets can be adapted for the case where $\mathcal{F}$ is a Dynkin system instead, by changing "σ-algebra" $\to$ "Dynkin system" and $\sigma(\mathcal{A}) \to \delta(\mathcal{A})$. Such adapted version is used in the proof here.

- After establishing the Dynkin's π-λ theorem below, due to (i) in the theorem, we can apply the principle of good sets by showing $\mathcal{G}$ is a Dynkin system that contains a π-system $\mathcal{A}$ instead, but we would only be able to deduce that every set in $\sigma(\mathcal{A})$ (not necessarily $\mathcal{F}$ in this case) is "good".

Proof. We first prove (i).

Defining "good". Fix any $B \in \delta(\mathcal{A})$. We call a set $A \in \delta(\mathcal{A})$ "good" if it is still in $\delta(\mathcal{A})$ after intersecting with B. [Intuition 💡: It is related with the closedness under intersections for a π-system.] Then, the family of all "good" sets in $\delta(\mathcal{A})$ is

$$\mathcal{D}_B := \{A \in \delta(\mathcal{A}) : A \cap B \in \delta(\mathcal{A})\}.$$

Showing that $\mathcal{D}_B$ is a Dynkin system.

(1) Since $\varnothing \in \delta(\mathcal{A})$ and $\varnothing \cap B = \varnothing \in \delta(\mathcal{A})$, we have $\varnothing \in \mathcal{D}_B$.

(2) Fix any $A \in \mathcal{D}_B$. Then $A \in \delta(\mathcal{A})$ and $A \cap B \in \delta(\mathcal{A})$. Thus $A^c \in \delta(\mathcal{A})$ and $A^c \cap B = B \setminus (A \cap B) \overset{(2')}{\in} \delta(\mathcal{A})$ as $A \cap B \subseteq B$. Hence $A^c \in \mathcal{D}_B$.

(3) Fix any $A_1, A_2, \ldots \in \mathcal{D}_B$ that are pairwise disjoint. Then for each $i \in \mathbb{N}$, $A_i \in \delta(\mathcal{A})$ and $A_i \cap B \in \delta(\mathcal{A})$. By closedness under countable disjoint unions for $\delta(\mathcal{A})$, we have $\biguplus_{i=1}^{\infty} A_i \in \delta(\mathcal{A})$. Also, $(\biguplus_{i=1}^{\infty} A_i) \cap B = \biguplus_{i=1}^{\infty} \underbrace{(A_i \cap B)}_{\in \delta(\mathcal{A})} \in \delta(\mathcal{A})$. Thus $\biguplus_{i=1}^{\infty} A_i \in \mathcal{D}_B$.

Showing that $\mathcal{D}_B$ contains a generator $\mathcal{A}$ for the Dynkin system $\delta(\mathcal{A})$.

Lemma: For any $A' \in \delta(\mathcal{A})$ and $B' \in \mathcal{A}$, we have $A' \cap B' \in \delta(\mathcal{A})$.

Proof. Fix any $B' \in \mathcal{A}$. Then for all $A \in \mathcal{A}$, since $\mathcal{A}$ is a π-system, we have $A \cap B \in \mathcal{A} \subseteq \delta(\mathcal{A})$, which means $A \in \mathcal{D}_{B'}$. Thus, $\mathcal{A} \subseteq \mathcal{D}_{B'}$, which implies $\delta(\mathcal{A}) \subseteq \mathcal{D}_{B'}$ as $\mathcal{D}_{B'}$ is a Dynkin system.

Having the inclusion $\delta(\mathcal{A}) \subseteq \mathcal{D}_{B'} \; \forall B' \in \mathcal{A}$, we know that for all $A' \in \delta(\mathcal{A})$ and $B' \in \mathcal{A}$, we have $A' \in \mathcal{D}_{B'}$, so $A' \cap B' \in \delta(\mathcal{A})$. $\square$

With the fixed $B \in \delta(\mathcal{A})$, we want to show that $\mathcal{A} \subseteq \mathcal{D}_B$. By the lemma with $A' = B$ and $B' = A$, for all $A \in \mathcal{A} \subseteq \delta(\mathcal{A})$, we have $A \cap B \in \delta(\mathcal{A})$, which means $A \in \mathcal{D}_B$. Thus, $\mathcal{A} \subseteq \mathcal{D}_B$ as desired.

Applying the principle of good sets to show that $\delta(\mathcal{A})$ is a π-system. By the principle of good sets, every set A in $\delta(\mathcal{A})$ is "good", i.e, $A \cap B \in \delta(\mathcal{A})$, given any fixed $B \in \delta(\mathcal{A})$. As this holds for all $B \in \delta(\mathcal{A})$, we know that $A, B \in \delta(\mathcal{A}) \implies A \cap B \in \delta(\mathcal{A})$, implying that $\delta(\mathcal{A})$ is a π-system.

Proving (i). Since $\delta(\mathcal{A})$ is a Dynkin system that is also a π-system, $\delta(\mathcal{A})$ is a σ-algebra (containing $\mathcal{A}$). This shows $\sigma(\mathcal{A}) \subseteq \delta(\mathcal{A})$. Also, since $\mathcal{D}$ is a Dynkin system containing $\mathcal{A}$, we have $\delta(\mathcal{A}) \subseteq \mathcal{D}$. Hence, $\sigma(\mathcal{A}) \subseteq \mathcal{D}$.

Next, we use (i) to prove (ii). Firstly, (i) implies that $\sigma(\mathcal{A}) \subseteq \delta(\mathcal{A})$ as $\delta(\mathcal{A})$ is a Dynkin system containing $\mathcal{A}$. Also, since every σ-algebra is a Dynkin system, $\sigma(\mathcal{A})$ is a Dynkin system containing $\mathcal{A}$, thus $\delta(\mathcal{A}) \subseteq \sigma(\mathcal{A})$. This completes the proof. $\square$

1.4.13 **Borel σ-algebra.** Borel σ-algebra is a σ-algebra generated by *topology*, which is a family of open sets. Before giving the formal definition, let us introduce/review what a topology is. A **topology** on a set Ω is a family of subsets $\mathcal{T} \subseteq \mathcal{P}(\Omega)$ such that:

(1) $\varnothing, \Omega \in \mathcal{T}$.

(2) *(Closed under arbitrary unions)* $A_i \in \mathcal{T} \; \forall i \in I \implies \bigcup_{i \in I} A_i \in \mathcal{T}$.

(3) *(Closed under finite intersections)* $A_1, \ldots, A_n \in \mathcal{T} \implies \bigcap_{i=1}^{n} A_i \in \mathcal{T}$.
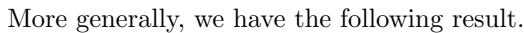
[Intuition 💡: The closedness under arbitrary unions and finite intersections originate from the property of openness and closedness for a *metric space*: arbitrary union of open sets is open, and finite intersection of open sets is open.]

The pair $(\Omega, \mathcal{T})$ is called a **topological space**. Every set in $\mathcal{T}$ is said to be **open** in Ω, and the complement of an open set in Ω is said to be **closed** in Ω.[5] [Note: From this, we can see that the concepts of openness and closedness are *relative to* the topology $\mathcal{T}$.]

By default, when we talk about the case where $\Omega = \mathbb{R}^d$, the corresponding topology $\mathcal{T}$ is supposed to be the "usual topology", i.e., the one that is induced by the Euclidean metric, under which our usual notion of "open set" and "closed set" (the concepts for *metric space*; see e.g., MATH3401) agrees with the openness and closedness under the "usual topology". In general, a metric space can always induce a corresponding topological space. But here we will mostly focus on the case where $\Omega = \mathbb{R}^d$ when discussing Borel σ-algebra.

Let $(\Omega, \mathcal{T})$ be a topological space. Then the **Borel σ-algebra** on Ω is given by $\mathcal{B}(\Omega) := \sigma(\mathcal{T}) = \sigma(\{O : O \text{ is open in } \Omega\}$. Every set in $\mathcal{B}(\Omega)$ is called a **Borel set**.

From this definition, we can see that Borel σ-algebra is a rather large family as it contains all open sets, closed sets, and countable unions/intersections of these.

1.4.14 **Generators of Borel σ-algebras.** While Borel σ-algebra is generated by all open sets, it may be a bit difficult to understand intuitively what the generating process looks like. Fortunately, in the special case where $\Omega = \mathbb{R}$ or $\Omega = \mathbb{R}^d$, some alternative and perhaps more intuitive generators are available.

First we consider the case where $\Omega = \mathbb{R}$. To derive an alternative generator, we need the following lemma.

Lemma 1.4.e. Every open set in $\mathbb{R}$ is a countable disjoint union of open intervals.

Proof. Fix any open $O \subseteq \mathbb{R}$. For any $x \in O$, let $\mathcal{J}_x := \{I : I \text{ is an open interval in } O \text{ and } x \in I\}$ and $I_x := \bigcup_{I \in \mathcal{J}_x} I$ be the open interval of maximal length containing x (in O).

If $x, y \in O$, then we have either $I_x = I_y$ or $I_x \cap I_y = \varnothing$ by construction. By the axiom of choice, we select I_x for every $x \in O$. Collecting them with redundancies removed, we get a set $\mathcal{I} = \{I_x : x \in O\}$ of all distinct and pairwise disjoint intervals of maximal length.

Then, we have $O = \bigcup_{x \in O} I_x = \biguplus_{I \in \mathcal{I}} I$. The latter disjoint union is countable since we can define an injection f from $\mathcal{I}$ to $\mathbb{Q}$ (countable) by $f(I) = q_I \in \mathbb{Q}$ where q_I is any rational number in I (which exists due to the density of $\mathbb{Q}$ on $\mathbb{R}$). (The injectivity follows from the pairwise disjointness of intervals in $\mathcal{I}$.) $\qquad\square$

Having Lemma 1.4.e, we can derive another generator for Borel σ-algebra on $\mathbb{R}$, which is the family of all *open intervals*. It turns out that, while the family of all open intervals appears to be much smaller than the family $\mathcal{T}$ of all open sets in $\mathbb{R}$, it can still generate the Borel σ-algebra!

Proposition 1.4.f ($\mathcal{B}(\mathbb{R})$ is generated by all open intervals)**.** We have $\mathcal{B}(\mathbb{R}) = \sigma(\{(a, b) : a < b\})$.

[Note: Sometimes the shorthand notation $\mathcal{B}$ is used instead of $\mathcal{B}(\mathbb{R})$.]

Proof. "$\supseteq$": Note that $\{(a, b) : a < b\} \subseteq \{O : O \text{ open in } \mathbb{R}\} = \mathcal{T} \subseteq \sigma(\mathcal{T}) = \mathcal{B}(\mathbb{R})$. Thus, by the minimality we have $\sigma(\{(a, b) : a < b\}) \subseteq \mathcal{B}(\mathbb{R})$.

"$\subseteq$": By Lemma 1.4.e, every open $O \subseteq \mathbb{R}$ is a countable disjoint union of open intervals, hence is in $\sigma(\{(a, b) : a < b\})$. It follows that $\mathcal{T} = \{O : O \text{ is open in } \mathbb{R}\} \subseteq \sigma(\{(a, b) : a < b\})$, thus $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{T}) \subseteq \sigma(\{(a, b) : a < b\})$ by the minimality. $\qquad\square$

More generally, we have the following result.

---

[5]When we talk about a set is open/closed in Ω, such set is understood to be a subset of Ω.

Proposition 1.4.g (Generators of $\mathcal{B}(\mathbb{R}^d)$)**.** We have

$$\mathcal{B}(\mathbb{R}^d) = \sigma(\{(\boldsymbol{a}, \boldsymbol{b}) : \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d, \boldsymbol{a} < \boldsymbol{b}\}) = \sigma(\{[\boldsymbol{a}, \boldsymbol{b}] : \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d, \boldsymbol{a} < \boldsymbol{b}\})$$
$$= \sigma(\{(\boldsymbol{a}, \boldsymbol{b}] : \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d, \boldsymbol{a} < \boldsymbol{b}\}) = \sigma(\{[\boldsymbol{a}, \boldsymbol{b}) : \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d, \boldsymbol{a} < \boldsymbol{b}\})$$
$$= \sigma(\{(-\boldsymbol{\infty}, \boldsymbol{b}) : \boldsymbol{b} \in \mathbb{R}^d\}) = \sigma(\{(\boldsymbol{a}, \boldsymbol{\infty}) : \boldsymbol{a} \in \mathbb{R}^d\})$$
$$= \sigma(\{(-\boldsymbol{\infty}, \boldsymbol{b}] : \boldsymbol{b} \in \mathbb{R}^d\}) = \sigma(\{[\boldsymbol{a}, \boldsymbol{\infty}) : \boldsymbol{a} \in \mathbb{R}^d\}).$$

[Note: Here, the interval and inequality notations carry componentwise meaning: $\boldsymbol{a} < \boldsymbol{b}$ means $a_i < b_i$ for all $i = 1, \ldots, d$, $(\boldsymbol{a}, \boldsymbol{b}) = (a_1, b_1) \times \cdots \times (a_d, b_d)$, etc., where $\boldsymbol{a} = (a_1, \ldots, a_d)$ and $\boldsymbol{b} = (b_1, \ldots, b_d)$.]

Proof. Omitted. $\qquad\square$

1.4.15 Borel σ-algebras on $\mathbb{R}^d$, $\bar{\mathbb{R}}^d$, and their subsets.

(a) The Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$ is the product σ-algebra $\bigotimes_{i=1}^d \mathcal{B}(\mathbb{R})$.

(b) Later in Section 5, we will encounter the case $\Omega = \bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\} =: [-\infty, \infty]$, the extended real number system. An order topology can be equipped on $\bar{\mathbb{R}}$ to form a topological space. Then it can be shown that $A \subseteq \bar{\mathbb{R}}$ is open iff A is a countable union of members in $\{(a, b) : a, b \in \mathbb{R}\} \cup \{[-\infty, b) : b \in \mathbb{R}\} \cup \{(a, \infty] : a \in \mathbb{R}\}$.
Consequently, we have $\mathcal{B}(\bar{\mathbb{R}}) = \{B \cup E : B \in \mathcal{B}(\mathbb{R}), E \subseteq \{-\infty, \infty\}\}$. From this we can get, e.g., $\mathcal{B}(\bar{\mathbb{R}}) = \sigma(\{(a, \infty] : a \in \mathbb{R}\})$ and $\mathcal{B}(\bar{\mathbb{R}}) = \sigma(\{[-\infty, b] : b \in \mathbb{R}\})$.

(c) The Borel σ-algebra $\mathcal{B}(\bar{\mathbb{R}}^d)$ is the product σ-algebra $\bigotimes_{i=1}^d \mathcal{B}(\bar{\mathbb{R}})$.

(d) Borel σ-algebra on $\Omega \subseteq \mathbb{R}^d$ or $\Omega \subseteq \bar{\mathbb{R}}^d$ can be obtained by the respective *trace σ-algebra*. For instance, if $\Omega \subseteq \mathbb{R}^d$, then $\mathcal{B}(\Omega) = \mathcal{B}(\mathbb{R}^d)|_\Omega = \{B \cap \Omega : B \in \mathcal{B}(\mathbb{R}^d)\}$.

2 Measure Theory II: Measures

2.0.1 After learning more about different systems of sets used for describing *measurable sets* in Section 1, we will shift our focus to the main object of interest in measure theory, namely the *measure* itself. Here we will study its properties and constructions. Particularly, in Section 2.5 we will formalize some probabilistic notions that you may have encountered before.

2.1 Measures

2.1.1 **Basic terminologies.** Let $\mathcal{F}$ be a σ-algebra on Ω. Then the pair $(\Omega, \mathcal{F})$ is a **measurable space** and every set in $\mathcal{F}$ is called **measurable set**. A **measure** μ on $(\Omega, \mathcal{F})$ is a function such that:

(1) *(Nonnegativity)* μ is a function from $\mathcal{F}$ (domain) to $[0, \infty]$ (codomain).

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

(3) *(σ-additivity)* $A_1, A_2, \ldots \in \mathcal{F}$ pairwise disjoint $\implies \mu(\biguplus_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

Remarks:

- Informally, σ-additivity allows us to "pull $\biguplus$ out of μ and make it $\sum$" (explaining why there is a "+" in the notation "$\biguplus$").

- *(Finite additivity)* By taking $A_{n+1} = A_{n+2} = \cdots = \varnothing$, we can show also the *finite additivity*: $A_1, \ldots, A_n \in \mathcal{F}$ pairwise disjoint $\implies \mu(\biguplus_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i)$.

The triplet $(\Omega, \mathcal{F}, \mu)$ is called a **measure space** (not to be confused with "measurable space" above ⚠). If we can write $\Omega = \bigcup_{i=1}^{\infty} A_i$ with $A_1, A_2, \ldots \in \mathcal{F}$ where $\mu(A_i) < \infty$ for each $i \in \mathbb{N}$, then μ is called a **σ-finite**. If we have $\mu(\Omega) < \infty$, then μ is called **finite**. A measure μ on $\mathcal{F} = \mathcal{B}(\mathbb{R}^d)$ is said to be a **Borel measure on $\mathbb{R}^d$**.

2.1.2 **Examples of measures.**

(a) Let $(\Omega, \mathcal{F})$ be a measurable space. Then, we can define a measure that (i) assigns zero measure to every set in $\mathcal{F}$, or (ii) assigns infinite measure to every nonempty set in $\mathcal{F}$, and zero measure to the empty set. Symbolically, we write:

(i) $\mu : \mathcal{F} \to [0, \infty]$ with $\mu(A) = 0 \ \forall A \in \mathcal{F}$.

(ii) $\mu : \mathcal{F} \to [0, \infty]$ with $\mu(A) = \infty \cdot \mathbf{1}_{\{A \neq \varnothing\}} \ \forall A \in \mathcal{F}$ (with the convention that $\infty \cdot 0 = 0$).

(Verify that they are indeed valid measures on $\mathcal{F}$!) These measures are called **trivial measures**.

(b) Let Ω be an uncountable set and consider the *countable-cocountable σ-algebra* on it: $\mathcal{F} = \{A \subseteq \Omega : A \text{ is countable or } A^c \text{ is countable}\}$. Then the function $\mu : \mathcal{F} \to [0, \infty]$ defined by $\mu(A) = \mathbf{1}_{\{A \text{ is uncountable}\}} \ \forall A \in \mathcal{F}$ is a measure on $\mathcal{F}$.

Proof.

(1) Nonnegativity is immediate.

(2) $\mu(\varnothing) = 0$ as $\varnothing$ is countable.

(3) Fix any pairwise disjoint $A_1, A_2, \ldots \in \mathcal{F}$.

- *Case 1: A_i is countable for all $i \in \mathbb{N}$.* Then the union $\biguplus_{i=1}^{\infty} A_i$ is countable, hence $\mu(\biguplus_{i=1}^{\infty} A_i) = 0$. On the other hand, note that for all $i \in \mathbb{N}$, $\mu(A_i) = 0$ as A_i is countable. Thus $\sum_{i=1}^{\infty} \mu(A_i) = 0$.

- *Case 2: A_j is uncountable for some $j \in \mathbb{N}$.* Then since $A_j \subseteq \biguplus_{i=1}^{\infty} A_i$, the union $\biguplus_{i=1}^{\infty} A_i$ is uncountable. Hence $\mu(\biguplus_{i=1}^{\infty} A_i) = 1$. On the other hand, note that A_j^c has to be countable (or else $A_j \notin \mathcal{F}$, contradiction). Now, for all $i \neq k$, we have $A_i \subseteq A_j^c$ (due to the pairwise disjointness), so A_i must be countable, meaning that $\mu(A_i) = 0 \ \forall i \neq k$. We also have $\mu(A_j) = 1$, thus $\sum_{i=1}^{\infty} \mu(A_i) = 1$.

$\square$

(c) Let Ω be a countable set and consider the largest σ-algebra $\mathcal{F} = \mathcal{P}(\Omega)$. Then for every function $f : \Omega \to [0, \infty]$, we can define a measure $\mu : \mathcal{F} \to [0, \infty]$ by $\mu(A) = \sum_{\omega \in A} f(\omega) \ \forall A \in \mathcal{F}$.

Proof.

(1) Codomain of μ can be set as $[0, \infty]$ as the codomain of f is $[0, \infty]$.

(2) We have $\mu(\varnothing) = \sum_{\omega \in \varnothing} f(\omega) = 0$ as an empty sum is always zero.

(3) Fix any pairwise disjoint $A_1, A_2, \ldots \in \mathcal{F}$. Then

$$\mu\left(\biguplus_{i=1}^{\infty} A_i\right) = \sum_{\omega \in \biguplus_{i=1}^{\infty} A_i} f(\omega) = \sum_{i=1}^{\infty} \sum_{\omega \in A_i} f(\omega) = \sum_{i=1}^{\infty} \mu(A_i).$$

$\square$

Special cases:

- If $f \equiv 1$, then we have $\mu(A) = |A| \ \forall A \in \mathcal{F}$. Such measure is said to be the **counting measure**.

- If we define f by $f(\omega) = \mathbf{1}_{\{\omega = \widetilde{\omega}\}} \ \forall \omega \in \Omega$ where $\widetilde{\omega}$ is a fixed point in Ω, then we have $\mu(A) = \mathbf{1}_{\{\widetilde{\omega} \in A\}}$. Such measure is said to be the **Dirac measure** or the **point/unit mass** of $\widetilde{\omega}$.

2.1.3 Properties of measures.
Now let us deduce some basic properties of measures based on the definition. Since probability measure is indeed just a special kind of measure, some properties here should be familiar to you.

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. In the following, let A, B, A_i's, and B_i's be sets in $\mathcal{F}$.

(a) *(Union + intersection = individual sum)* $\mu(A \cup B) + \mu(A \cap B) = \mu(A) + \mu(B)$. If μ is finite, then we further have $\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$.

(b) *(Monotonicity)* $A \subseteq B \implies \mu(A) \leq \mu(B)$.

(c) *(Subtractivity)* If $A \subseteq B$ and $\mu(A) < \infty$, then $\mu(B \setminus A) = \mu(B) - \mu(A)$.

(d) *(σ-subadditivity)* $\mu(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$. [Note: By taking $A_{n+1} = A_{n+2} = \cdots = \varnothing$, we can get the *finite subadditivity*: $\mu(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} \mu(A_i)$.]

(e) *(Inclusion-exclusion principle)* For every integer $n \geq 2$, let $S_{i,n} := \sum_{I \subseteq \{1,\ldots,n\}:|I|=i} \mu(\bigcap_{j \in I} A_j)$ (summing over all combinations of i indices) for all $i = 1, \ldots, n$. Suppose μ is finite. Then

$$\mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} (-1)^{i-1} S_{i,n}.$$

(f) *(Continuity from below)* If $A_n \nearrow$, then $\mu(\lim_{n \to \infty} A_n) = \lim_{n \to \infty} \mu(A_n)$.

(g) *(Continuity from above)* If $A_n \searrow$ and $\mu(A_1) < \infty$, then $\mu(\lim_{n \to \infty} A_n) = \lim_{n \to \infty} \mu(A_n)$.

(h) *(Continuity)* Suppose that $\lim_{n \to \infty} A_n$ exists. If μ is finite, then $\mu(\lim_{n \to \infty} A_n) = \lim_{n \to \infty} \mu(A_n)$.

(i) *(Law of total measure)* If $\{B_i : i \in \mathbb{N}\}$ is a partition of Ω, then $\mu(A) = \sum_{i=1}^{\infty} \mu(A \cap B_i)$. [Note: If $\{B_1, \ldots, B_n\}$ is a partition of Ω, we can apply this result by setting $B_{n+1} = B_{n+2} = \cdots = \varnothing$, which gives $\mu(A) = \sum_{i=1}^{n} \mu(A \cap B_i)$.]

Proof.

(a) Write $A \cup B = A \uplus (B \setminus A)$ and $B = (A \cap B) \uplus (B \setminus A)$. By finite additivity,

$$\mu(A \cup B) = \mu(A) + \mu(B \setminus A), \tag{1}$$

$$\mu(A \cap B) + \mu(B \setminus A) = \mu(B). \tag{2}$$

Adding them together, we get

$$\mu(A \cup B) + \mu(A \cap B) + \mu(B \setminus A) = \mu(A) + \mu(B \setminus A) + \mu(B). \tag{3}$$

24

If $\mu(B\backslash A) < \infty$, subtracting both sides of Equation (3) by $\mu(B\backslash A)$ gives the result. If $\mu(B\backslash A) = \infty$, then (i) $\mu(A \cup B) = \infty$ by Equation (1), and (ii) $\mu(B) = \infty$ by Equation (2). So the result is essentially saying "$\infty = \infty$", which is true (the "∞" here is referring to the element in the extended real number system). This shows the first part.

For the second part, assuming μ is finite, we have $\mu(A \cap B) < \infty$. Thus, subtracting both sides of Equation (3) by $\mu(A \cap B)$ gives the desired result.

(b) Since $A \subseteq B$, we have $A \cap B = A$. Also, $B = (A \cap B) \uplus (B \setminus A)$. Thus,

$$\mu(B) = \mu(A \cap B) + \mu(B \setminus A) = \mu(A) + \underbrace{\mu(B \setminus A)}_{\geq 0} \geq \mu(A).$$

(c) As $A \subseteq B$, by Equation (2) we have $\mu(A) + \mu(B \setminus A) = \mu(B)$. Subtracting both sides by $\mu(A) < \infty$ then gives the desired result.

(d) Let $B_1 := A_1$ and $B_n := A_n \setminus \bigcup_{i=1}^{n-1} A_i$ for every integer $n \geq 2$. Then by construction, (i) B_n's are disjoint, (ii) $B_n \subseteq A_n$ for every $n \in \mathbb{N}$, and (iii) $\bigcup_{i=1}^{n} A_i = \biguplus_{i=1}^{n} B_i$ for every $n \in \mathbb{N}$.

By (iii), we have $\bigcup_{i=1}^{\infty} A_i = \lim_{N \to \infty} \bigcup_{i=1}^{N} A_i = \lim_{N \to \infty} \biguplus_{i=1}^{N} B_i = \biguplus_{i=1}^{\infty} B_i$. Thus,

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \mu\left(\biguplus_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mu(B_i) \overset{\text{((ii), monotonicity)}}{\leq} \sum_{i=1}^{\infty} \mu(A_i).$$

(e) We prove by induction. Firstly, the base case with $n = 2$ follows from [2.1.3]a as we can write $\mu(A_1) + \mu(A_2) - \mu(A_1 \cap A_2) = \sum_{i=1}^{2} (-1)^{i-1} S_{i,2}$.

Now, assume for induction that the result holds when $n = k$ for an integer $k \geq 2$. Then, consider:

$$\mu\left(\bigcup_{i=1}^{n+1} A_i\right) = \mu\left(\left(\bigcup_{i=1}^{n} A_i\right) \cup A_{n+1}\right)$$

$$\overset{([2.1.3]a)}{=} \mu\left(\bigcup_{i=1}^{n} A_i\right) + \mu(A_{n+1}) - \mu\left(\left(\bigcup_{i=1}^{n} A_i\right) \cap A_{n+1}\right)$$

$$= \underbrace{\sum_{i=1}^{n} (-1)^{i-1} S_{i,n}}_{S_{1,n} + \sum_{i=2}^{n} (-1)^{i-1} S_{i,n}} + \mu(A_{n+1}) - \mu\left(\bigcup_{i=1}^{n} (A_i \cap A_{n+1})\right)$$

$$= \sum_{i=2}^{n} (-1)^{i-1} S_{i,n} + S_{1,n+1} - \sum_{i=1}^{n} (-1)^{i-1} \sum_{I \subseteq \{1,\ldots,n\}: |I|=i} \mu\left(\left(\bigcap_{j \in I} A_j\right) \cap A_{n+1}\right)$$

$$= S_{1,n+1} + \sum_{i=2}^{n+1} (-1)^{j-1} S_{j,n+1}$$

$$= \sum_{i=1}^{n+1} (-1)^{j-1} S_{i,n+1}.$$

So the results hold when $n = k + 1$, completing the proof by induction.

(f) Let $B_1 := A_1$ and $B_n := A_n \setminus A_{n-1}$ for every integer $n \geq 2$. By construction, B_n's are disjoint, and $A_n = \bigcup_{i=1}^{n} A_i = \biguplus_{i=1}^{n} B_i$ for every $n \in \mathbb{N}$. We then have $\bigcup_{i=1}^{\infty} A_i = \lim_{n \to \infty} \bigcup_{i=1}^{n} A_i = \lim_{n \to \infty} \biguplus_{i=1}^{n} B_i = \biguplus_{i=1}^{\infty} B_i$, and thus

$$\mu\left(\lim_{n \to \infty} A_n\right) \overset{[1.1.5]c}{=} \mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \mu\left(\biguplus_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mu(B_i) = \lim_{n \to \infty} \sum_{i=1}^{n} \mu(B_i)$$

$$= \lim_{n \to \infty} \mu\left(\biguplus_{i=1}^{n} B_i\right) = \lim_{n \to \infty} \mu(A_n).$$

(g) Let $B_n = A_1 \setminus A_n = A_1 \cap A_n^c$ for every $n \in \mathbb{N}$. Since $A_n \searrow$, we have $B_n \nearrow$. Note also that $\lim_{n\to\infty} B_n = \bigcup_{i=1}^\infty B_i = \bigcup_{i=1}^\infty (A_1 \cap A_i^c) = A_1 \cap \bigcup_{i=1}^\infty A_i^c \stackrel{\text{(DM)}}{=} A_1 \cap (\bigcap_{i=1}^\infty A_i)^c = A_1 \setminus \bigcap_{i=1}^\infty A_i$. Since we have $\mu(A_1) < \infty$, by subtractivity we have

$$\mu(A_1) - \mu\left(\bigcap_{i=1}^\infty A_i\right) = \mu\left(A_1 \setminus \bigcap_{i=1}^\infty A_i\right) = \mu\left(\lim_{n\to\infty} B_n\right) \stackrel{(B_n \nearrow)}{=} \lim_{n\to\infty} \mu(B_n)$$

$$= \lim_{n\to\infty} \mu(A_1 \setminus A_n) \underset{(A_n \subseteq A_1)}{\overset{(\mu(A_n) \leq \mu(A_1) < \infty)}{=}} \lim_{n\to\infty} (\mu(A_1) - \mu(A_n)) = \mu(A_1) - \lim_{n\to\infty} \mu(A_n).$$

Subtracting both sides by $\mu(A_1) < \infty$ implies the desired result, as $\mu(\bigcap_{i=1}^\infty A_i) = \mu(\lim_{n\to\infty} A_n)$ when $A_n \searrow$.

(h) Let $A = \lim_{n\to\infty} A_n = \liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n$. Now consider:

$$\mu(A) = \mu\left(\liminf_{n\to\infty} A_n\right) \stackrel{[1.1.5]c}{=} \mu\left(\lim_{n\to\infty} \inf_{k\geq n} A_k\right)$$

$$\underset{\text{(continuity from below)}}{\overset{(\inf_{k\geq n} A_k \nearrow)}{=}} \lim_{n\to\infty} \mu\left(\inf_{k\geq n} A_k\right) \underset{\text{(monotonicity)}}{\overset{(\inf_{k\geq n} A_k \subseteq A_n)}{\leq}} \lim_{n\to\infty} \mu(A_n) = \liminf_{n\to\infty} \mu(A_n)$$

and, similarly,

$$\mu(A) = \mu\left(\limsup_{n\to\infty} A_n\right) \stackrel{[1.1.5]c}{=} \mu\left(\lim_{n\to\infty} \sup_{k\geq n} A_k\right)$$

$$\underset{\text{(continuity from above)}}{\overset{(\sup_{k\geq n} A_k \searrow)}{=}} \lim_{n\to\infty} \mu\left(\sup_{k\geq n} A_k\right) \underset{\text{(monotonicity)}}{\overset{(\sup_{k\geq n} A_k \supseteq A_n)}{\geq}} \lim_{n\to\infty} \mu(A_n) = \limsup_{n\to\infty} \mu(A_n).$$

This implies that

$$\limsup_{n\to\infty} \mu(A_n) \geq \liminf_{n\to\infty} \mu(A_n) \geq \mu(A) \geq \limsup_{n\to\infty} \mu(A_n) \geq \limsup_{n\to\infty} \mu(A_n),$$

and hence $\mu(A) = \limsup_{n\to\infty} \mu(A_n) = \liminf_{n\to\infty} \mu(A_n) = \lim_{n\to\infty} \mu(A_n)$.

(i) Note that $\mu(A) = \mu(A \cap \Omega) = \mu(A \cap \biguplus_{i=1}^\infty B_i) = \mu(\biguplus_{i=1}^\infty (A \cap B_i)) = \sum_{i=1}^\infty \mu(A \cap B_i)$.

$\square$

2.1.4 Uniqueness of measures.
After exploring some basic properties of measures, we are going to study a more technical aspect of measure, namely *uniqueness*. To be more precise, we are going to show that under certain conditions, fixing the measures assigned for all sets in a π-*system* $\mathcal{A}$ (just a "small" number of sets) would already be enough for uniquely determining the measure on $\sigma(\mathcal{A})$ (a "large" number of sets).

Proposition 2.1.a (Uniqueness of measures). Let $(\Omega, \mathcal{F})$ be a measurable space, μ and ν be measures on $\mathcal{F}$, and $\mathcal{A}$ be a π-system such that $\bigcup_{i=1}^\infty A_i = \Omega$ for some $A_1, A_2, \ldots \in \mathcal{A}$ with $\mu(A_i) < \infty$ for all $i \in \mathbb{N}$ (σ-*finiteness on* $\mathcal{A}$). If $\mu|_{\mathcal{A}} = \nu|_{\mathcal{A}}$, then $\mu|_{\sigma(\mathcal{A})} = \nu|_{\sigma(\mathcal{A})}$. [Note: Particularly, if we also have $\sigma(\mathcal{A}) = \mathcal{F}$, then $\mu = \nu$, establishing the uniqueness of measure (on the whole domain).]

Proof. In this proof we will again use the helpful *principle of good sets*.

Defining "good". Fix any $B \in \mathcal{A}$ with $\mu(B) < \infty$. Let the family of all "good" sets in $\sigma(\mathcal{A})$ be $\mathcal{D}_B := \{A \in \sigma(\mathcal{A}) : \mu(A \cap B) = \nu(A \cap B)\}$.

Showing that $\mathcal{D}_B$ is a Dynkin system.

(1) $\varnothing \in \mathcal{D}_B$ since $\mu(\varnothing \cap B) = \mu(\varnothing) = 0 = \nu(\varnothing) = \nu(\varnothing \cap B)$, with $\varnothing \in \sigma(\mathcal{A})$.

(2) Fix any $A \in \mathcal{D}_B \subseteq \sigma(\mathcal{A})$. By closedness under complementations, $A^c \in \sigma(\mathcal{A})$. Then it remains to show $\mu(A^c \cap B) = \nu(A^c \cap B)$. Consider

$$\mu(A^c \cap B) = \mu(B \setminus (A \cap B)) \stackrel{(\mu(A \cap B) \leq \mu(B) < \infty)}{=} \mu(B) - \mu(A \cap B)$$
$$\stackrel{(B \in \mathcal{A}, A \in \mathcal{D}_B)}{=} \nu(B) - \nu(A \cap B) \stackrel{(\nu(B) = \mu(B) < \infty)}{=} \nu(A^c \cap B).$$

This implies $A^c \in \mathcal{D}_B$.

(3) Fix any pairwise disjoint $A_1, A_2, \ldots \in \mathcal{D}_B$. Similarly, by closedness under countable unions, $\biguplus_{i=1}^{\infty} A_i \in \sigma(\mathcal{A})$, so it remains to show that $\mu((\biguplus_{i=1}^{\infty} A_i) \cap B) = \nu((\biguplus_{i=1}^{\infty} A_i) \cap B)$. Consider

$$\mu\left(\left(\biguplus_{i=1}^{\infty} A_i\right) \cap B\right) = \mu\left(\biguplus_{i=1}^{\infty} (A_i \cap B)\right) = \sum_{i=1}^{\infty} \mu(A_i \cap B)$$
$$\stackrel{(A_i \in \mathcal{D}_B \ \forall i \in \mathbb{N})}{=} \sum_{i=1}^{\infty} \nu(A_i \cap B) \stackrel{\text{(same argument as above)}}{=} \nu\left(\left(\biguplus_{i=1}^{\infty} A_i\right) \cap B\right).$$

This implies $\biguplus_{i=1}^{\infty} A_i \in \mathcal{D}_B$.

Showing that $\mathcal{D}_B$ contains $\mathcal{A}$ to apply the principle of good sets. Since $\mathcal{A}$ is a π-system by assumption, for all $A \in \mathcal{A} \subseteq \sigma(\mathcal{A})$, we have $A \cap B \in \mathcal{A}$, thus $\mu(A \cap B) = \nu(A \cap B)$, which implies $A \in \mathcal{D}_B$. Hence, we have $\mathcal{A} \subseteq \mathcal{D}_B$. By Dynkin's π-λ theorem, we conclude that $\sigma(\mathcal{A}) \subseteq \mathcal{D}_B$. So every set in $\sigma(\mathcal{A})$ is "good". In other words, given any $B \in \mathcal{A}$ with $\mu(B) < \infty$, we have $\mu(A \cap B) = \nu(A \cap B)$ for all $A \in \sigma(\mathcal{A})$.

Constructing an exhausting sequence of sets. From the σ-finiteness on $\mathcal{A}$, there are $A_1, A_2, \ldots \in \mathcal{A} \subseteq \sigma(\mathcal{A})$ with $\mu(A_i) < \infty$ for all $i \in \mathbb{N}$ and $\bigcup_{i=1}^{\infty} A_i = \Omega$. By modifying $A_n \to \widetilde{A}_n := A_n \setminus \bigcup_{i=1}^{n-1} A_i \in \sigma(\mathcal{A})$ for every integer $n \geq 2$ if necessary, we may assume A_i's are pairwise disjoint. Then let $A'_n := \biguplus_{i=1}^{n} A_i \in \sigma(\mathcal{A})$ for every $n \in \mathbb{N}$, which satisfies $A'_n \nearrow \Omega$. [Note: Such A'_ns form an *exhausting sequence* of sets.]

Showing that $\mu(A) = \nu(A) \ \forall A \in \sigma(\mathcal{A})$ by considering intersections of sets. Note that $A_i \in \mathcal{A}$ with $\mu(A_i) < \infty$ for all $i \in \mathbb{N}$. Thus, for all $A \in \sigma(\mathcal{A})$, we have

$$\mu(A \cap A'_n) = \mu\left(\biguplus_{i=1}^{n} (A \cap A_i)\right) = \sum_{i=1}^{n} \mu(A \cap A_i)$$
$$\stackrel{\text{(prev. result with } B = A_i)}{=} \sum_{i=1}^{n} \nu(A \cap A_i) \stackrel{\text{(same argument as before)}}{=} \nu(A \cap A'_n).$$

Since $A'_n \nearrow \Omega$, we have $(A \cap A_n) \nearrow A \cap \Omega$ also. Therefore, for all $A \in \sigma(\mathcal{A})$,

$$\mu(A) = \mu(A \cap \Omega) \stackrel{\text{(cont. from below)}}{=} \lim_{n \to \infty} \mu(A \cap A'_n) \stackrel{\text{(above)}}{=} \lim_{n \to \infty} \nu(A \cap A'_n) = \nu(A \cap \Omega) = \nu(A).$$

$\square$

2.1.5 **Product measures.** We have seen the concept of *product σ-algebra* previously. As a measure is defined on a σ-algebra, a natural question is then whether there is also a "product"-type concept for measure. It turns out that we do have such kind of measure, and we will study it here.

Let $\mathcal{F}_j$ be a σ-algebra on a set Ω_j for all $j = 1, \ldots, d$. Then the *product* space $\prod_{i=1}^{d} \Omega_j$ can be equipped with the *product σ-algebra* $\bigotimes_{j=1}^{d} \mathcal{F}_j = \sigma\left(\left\{\prod_{i=1}^{d} A_j : A_j \in \mathcal{F}_j \ \forall j = 1, \ldots, d\right\}\right)$ (the equality holds due to Proposition 1.4.b).

The product measure is then defined in a special way that utilizes the uniqueness result from Proposition 2.1.a. First suppose that μ_j is a σ-finite measure on $\mathcal{F}_j$ for all $j = 1, \ldots, d$ (for the applicability of Proposition 2.1.a). Then, the **product measure** on $\bigotimes_{j=1}^{d} \mathcal{F}_j$, denoted by $\prod_{i=1}^{d} \mu_j$, is the unique

27

measure μ on $\bigotimes_{j=1}^d \mathcal{F}_j$ satisfying that $\mu(\prod_{j=1}^d A_j) = \prod_{j=1}^d \mu_j(A_j)$ for all $A_1 \in \mathcal{F}_1, \ldots, A_d \in \mathcal{F}_d$. [Note: The uniqueness follows by applying Proposition 2.1.a with $\mathcal{A} = \{\prod_{j=1}^d A_j : A_j \in \mathcal{F}_j \ \forall j = 1, \ldots, d\}$, which is a π-system (check!).]

Let us verify that such μ is a valid measure as suggested above.

Proof.

(1) Codomain of the product measure can be set as $[0, \infty]$ since the codomain of μ_j is $[0, \infty]$ for all $j = 1, \ldots, d$.

(2) Writing $\varnothing = \prod_{j=1}^d A_j$, we know $A_{j^*} = \varnothing$ for some $j^* = 1, \ldots, d$. Thus we have $\mu(\varnothing) = \prod_{j=1}^d \mu_j(A_j) \overset{(\mu_{j^*}(A_{j^*}) = \mu_{j^*}(\varnothing) = 0)}{=} 0$.

(3) Fix any pairwise disjoint $A_1, A_2, \ldots \in \bigotimes_{j=1}^d \mathcal{F}_j$. Then for all $i = 1, \ldots, n$, we can write $A_i = \prod_{j=1}^d A_{ij}$ where $A_{ij} \in \mathcal{F}_j \ \forall j = 1, \ldots, d$. Since $\biguplus_{i=1}^\infty A_i = \biguplus_{i=1}^\infty \prod_{j=1}^d A_{ij} = \prod_{j=1}^d \biguplus_{i=1}^\infty A_{ij}$, we have

$$\mu\left(\biguplus_{i=1}^\infty A_i\right) = \mu\left(\prod_{j=1}^d \biguplus_{i=1}^\infty A_{ij}\right) = \prod_{j=1}^d \mu_j\left(\biguplus_{i=1}^\infty A_{ij}\right) = \prod_{j=1}^d \sum_{i=1}^\infty \mu_j(A_{ij})$$

$$\overset{\text{(distributivity)}}{=} \sum_{i=1}^\infty \prod_{j=1}^d \mu_j(A_{ij}) = \sum_{i=1}^\infty \mu\left(\prod_{j=1}^d A_{ij}\right) = \sum_{i=1}^\infty \mu(A_i).$$

$\square$

2.2 Null Sets

2.2.1 Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Every set $N \in \mathcal{F}$ with $\mu(N) = 0$ is called a (μ-)**null set**. If a statement holds for all $\omega \in \Omega \setminus N$ where N is a null set, then it is said to hold (μ-)**almost everywhere** (a.e.), or (μ-)**almost surely** (a.s.) when μ is a probability measure. If all subsets of every null set are in $\mathcal{F}$, then μ is called a **complete measure**.

The reason why we have the term "almost" here is that a null set may *not* be an empty set, although it has measure zero. For instance, with $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$, and μ being the Lebesgue measure (corresponding to our usual notion of "length", to be formally constructed in Section 2.4), any singleton $\{x\} \subseteq \mathbb{R}$ would have a zero measure: $\mu(\{x\}) = \mu([x, x]) = x - x = 0$, but it is certainly nonempty.

However, whether a statement holds **everywhere**/**surely** (i.e., holds for all $\omega \in \Omega$) or *almost* everywhere/surely often does not really matter, because we usually do not care much about the points inside a null set, which would not impact the measure anyway.

2.2.2 **Property of null sets.** The main property of null sets is that countable union of null sets is still a null set: Taking *countable* union does not lead to qualitative changes.

Lemma 2.2.a. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Every countable union of null sets in $\mathcal{F}$ is a null set in $\mathcal{F}$.

Proof. Fix any null sets $N_1, N_2, \ldots \in \mathcal{F}$. Then by nonnegativity and σ-subadditivity we have $0 \leq \mu(\bigcup_{i=1}^\infty N_i) \leq \sum_{i=1}^\infty \mu(N_i) = 0$, which forces $\mu(\bigcup_{i=1}^\infty N_i) = 0$. $\square$

2.2.3 **Completion of measures.** With the help of Lemma 2.2.a, we can prove the main result in Section 2.2, which is about extending a non-complete measure to a complete one *(completion)*. It turns out that completion is always possible, and is also *unique*.

Theorem 2.2.b. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $\mathcal{N} = \{N \in \mathcal{F} : \mu(N) = 0\}$ be the set of all null sets in $\mathcal{F}$. Then:

(a) $\bar{\mathcal{F}} := \{A \cup N' : A \in \mathcal{F}, N' \subseteq N$ for some $N \in \mathcal{N}\}$ is a σ-algebra on Ω. [Note: It is said to be the **completion of** $\mathcal{F}$.]

(b) Consider the function $\bar{\mu} : \bar{\mathcal{F}} \to [0, \infty]$ with $\bar{\mu}(A \cup N') = \mu(A)$ for all $A \in \mathcal{F}$ and all $N' \subseteq N$ with $N \in \mathcal{N}$. It uniquely extends μ to a complete measure on $\bar{\mathcal{F}}$.

Proof.

(a) (1) We can write $\varnothing = \underbrace{\varnothing}_{\in \mathcal{F}} \cup \underbrace{\varnothing}_{\subseteq \varnothing \in \mathcal{N}}$, so $\varnothing \in \bar{\mathcal{F}}$.

(2) Fix any $\bar{A} \in \bar{\mathcal{F}}$. Then we have $\bar{A} = A \cup N'$ for some $A \in \mathcal{F}$ and $N' \subseteq N$ with $N \in \mathcal{N}$. By changing $N' \to N' \setminus A$ and $N \to N \setminus A$ if necessary (we still have $N' \setminus A \subseteq N \setminus A \in \mathcal{N}$), we may assume that $A \cap N = \varnothing$, which implies that $A \cap N' = \varnothing$ and $A \cap N^c = A$. Also note that $N' = N \cap N'$ as $N' \subseteq N$. Therefore,

$$\bar{A} = A \cup N' = \left[(A \cap N^c) \cup \underbrace{(A \cap N')}_{\varnothing} \right] \cup \left[\underbrace{(N \cap N^c)}_{\varnothing} \cup (N \cap N') \right]$$

$$\overset{\text{(distributivity)}}{=} [A \cap (N^c \cup N')] \cup [N \cap (N^c \cup N')]$$

$$\overset{\text{(distributivity)}}{=} (A \cup N) \cap (N^c \cup N'),$$

which implies that

$$(\bar{A})^c = [(A \cup N) \cap (N^c \cup N')]^c \overset{\text{(DM)}}{=} \overbrace{(A \cup N)^c}^{\in \mathcal{F} \text{ as } A \in \mathcal{F} \text{ and } N \in \mathcal{N} \subseteq \mathcal{F}} \cup \underbrace{(N \setminus N')}_{\subseteq N \in \mathcal{N}} \in \bar{\mathcal{F}}.$$

(3) Fix any $\bar{A}_1, \bar{A}_2, \ldots \in \bar{\mathcal{F}}$. For all $i \in \mathbb{N}$, write $\bar{A}_i = A_i \cup N_i'$ for some $A_i \in \mathcal{F}$ and $N_i' \subseteq N_i$ with $N_i \in \mathcal{N}$. Then, we have $\bigcup_{i=1}^{\infty} \bar{A}_i = \bigcup_{i=1}^{\infty} (A_i \cup N_i') = (\bigcup_{i=1}^{\infty} A_i) \cup (\bigcup_{i=1}^{\infty} N_i') \in \bar{\mathcal{F}}$, since $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ and $\bigcup_{i=1}^{\infty} N_i' \subseteq \bigcup_{i=1}^{\infty} N_i \overset{\text{(Lemma 2.2.a)}}{\in} \mathcal{N}$

(b) **Well-definedness.** For all $A_1', A_2' \in \bar{\mathcal{F}}$ with $A_1' = A_2'$, we can write $A_1 \cup N_1' = A_1' = A_2' = A_2 \cup N_2'$. Then, from $A_1 \subseteq A_1 \cup N_1' = A_2 \cup N_2'$, by monotonicity we have $\mu(A_1) \leq \mu(A_2) + 0 = \mu(A_2)$. Interchanging the roles of A_1 and A_2, we have $\mu(A_2) \leq \mu(A_1)$. Hence, $\mu(A_1) = \mu(A_2)$, and so $\bar{\mu}(A_1') = \mu(A_1) = \mu(A_2) = \bar{\mu}(A_2')$, establishing the well-definedness (the same input is always mapped to the same output regardless of how the input is represented).

Showing that $\bar{\mu}$ is a complete measure on $\bar{\mathcal{F}}$. First we show that $\bar{\mu}$ is a measure on $\bar{\mathcal{F}}$.

(1) The codomain of $\bar{\mu}$ can be set to be $[0, \infty]$ since the codomain of μ is $[0, \infty]$ and $\bar{\mu}$ is defined through the values taken by μ.

(2) We can write $\varnothing = \underbrace{\varnothing}_{\in \mathcal{F}} \cup \underbrace{\varnothing}_{\subseteq \varnothing \in \mathcal{N}}$, so $\bar{\mu}(\varnothing) = \mu(\varnothing) = 0$.

(3) Fix any pairwise disjoint $\bar{A}_1, \bar{A}_2, \ldots \in \bar{\mathcal{F}}$. For all $i \in \mathbb{N}$, write $\bar{A}_i = A_i \cup N_i'$ for some $A_i \in \mathcal{F}$ and $N_i' \subseteq N_i$ with $N_i \in \mathcal{N}$. Note that the pairwise disjointness of $\bar{A}_i$'s forces A_i's to be pairwise disjoint also, because $\varnothing \subseteq A_i \cap A_j \subseteq \bar{A}_i \cap \bar{A}_j = \varnothing \ \forall i \neq j$. Thus,

$$\bar{\mu}\left(\biguplus_{i=1}^{\infty} \bar{A}_i \right) = \bar{\mu}\left(\biguplus_{i=1}^{\infty} (A_i \cup N_i') \right) = \bar{\mu}\left(\biguplus_{i=1}^{\infty} A_i \cup \bigcup_{i=1}^{\infty} N_i' \right)$$

$$\overset{(\bigcup_{i=1}^{\infty} N_i' \subseteq \bigcup_{i=1}^{\infty} N_i \in \mathcal{N})}{=} \mu\left(\biguplus_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mu(A_i) = \sum_{i=1}^{\infty} \bar{\mu}(A_i \cup N_i') = \sum_{i=1}^{\infty} \bar{\mu}(\bar{A}_i).$$

We also see that $\bar{\mu}$ is complete since $\bar{\mathcal{F}}$ contains all subsets of every null set.

Showing that $\bar{\mu}$ extends μ. For all $A \in \mathcal{F}$, we can take $N' = \varnothing$ to get $\bar{\mu}(A) = \mu(A)$.

Uniqueness. Suppose there is another complete measure $\bar{\nu}$ on $\mathcal{F}$ with $\bar{\nu}(A \cup N') = \mu(A)$ for all $A \in \mathcal{F}$ and all $N' \subseteq N$ with $N \in \mathcal{N}$. Then we have, for all $A \in \mathcal{F}$ and all $N' \subseteq N$ with $N \in \mathcal{N}$,

$$\bar{\nu}(A \cup N') \overset{\overset{\text{(finite subadditivity)}}{}}{\leq} \bar{\nu}(A) + \underbrace{\bar{\nu}(N')}_{\bar{\nu}(N' \cup N') = \mu(N') = 0} = \bar{\nu}(A) \overset{\overset{\text{(extend)}}{}}{=} \mu(A) \overset{\overset{\text{(extend)}}{}}{=} \bar{\mu}(A) \overset{\overset{\text{(monotonicity)}}{}}{\leq} \bar{\mu}(A \cup N'),$$

and we can similarly get $\bar{\mu}(A \cup N') \leq \bar{\nu}(A \cup N')$, which implies $\bar{\mu}(A \cup N') = \bar{\nu}(A \cup N')$.

$\square$

2.3 Construction of Measures

2.3.1 While we have seen some simple examples of measure in [2.1.2], in general it is not so straightforward to construct a measure. For example, the *Lebesgue measure* on $\mathbb{R}^d$ (capturing our usual notion of "volumes" of geometrical objects), while being intuitive, requires quite a lot of work to be constructed (see Section 2.4).

One important theoretical result underlying the construction of measures is the *Carathéodory extension theorem*, which justifies a construction approach that starts from a "simple version" of measure and extends there to construct a measure. To be more precise, the idea is to start from a "premeasure" μ_0 on a simple system of sets $\mathcal{A}$ and then extend μ_0 through an "outer measure" μ^* to a measure μ on $\sigma(\mathcal{A})$. There are different versions of Carathéodory extension theorem, depending on the type of $\mathcal{A}$. Here, we will study the version where $\mathcal{A}$ is a *semiring*.

2.3.2 **Premeasures and outer measures.** We start by introducing some preliminary terms: premeasures and outer measures. Let Ω be a set.

- *Premeasure:* Let $\mathcal{A}$ be a semiring on Ω. A **premeasure** μ_0 on $\mathcal{A}$ is a function such that:
  (1) *(Nonnegativity)* μ_0 is a function from $\mathcal{A}$ to $[0, \infty]$.
  (2) $\mu_0(\varnothing) = 0$.
  (3) *(σ-additivity)* $A_1, A_2, \ldots \in \mathcal{A}$ pairwise disjoint and $\biguplus_{i=1}^{\infty} A_i \in \mathcal{A} \implies \mu_0(\biguplus_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu_0(A_i)$.
      Remarks:
      – The condition $\biguplus_{i=1}^{\infty} A_i \in \mathcal{A}$ is needed as $\mathcal{A}$ is just a semiring, which may not be closed under countable unions.
      – *(Finite additivity)* Like before, by setting $A_{n+1} = A_{n+2} = \cdots = \varnothing$, we can show finite additivity: $A_1, A_2, \ldots, A_n \in \mathcal{A}$ pairwise disjoint and $\biguplus_{i=1}^{n} A_i \in \mathcal{A} \implies \mu_0(\biguplus_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu_0(A_i)$

  If $\Omega = \bigcup_{i=1}^{\infty} A_i$ for some $A_1, A_2, \ldots \in \mathcal{A}$ with $\mu_0(A_i) < \infty \; \forall i \in \mathbb{N}$, then μ_0 is called σ**-finite**.

- *Outer measure:* An **outer measure** is a function $\mu^* : \mathcal{P}(\Omega) \to [0, \infty]$ satisfying
  (1) $\mu^*(\varnothing) = 0$.
  (2) *(Monotonicity)* $A \subseteq B \implies \mu^*(A) \leq \mu^*(B)$
  (3) *(σ-subadditivity)* $\mu^*(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu^*(A_i)$. [Note: Like above, it implies *finite subadditivity* by setting $A_{n+1} = A_{n+2} = \cdots = \varnothing$: $\mu^*(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} \mu^*(A_i)$.]

  It is called <u>out</u>er measure since it is used for *approximating* "volumes"/measures from "<u>out</u>side"/"above"; the inequalities appearing in the definition are helpful for the approximation.

2.3.3 The following lemma will be used for proving the Carathéodory extension theorem.

Lemma 2.3.a (Representing set differences in terms of disjoint union)**.** Let $\mathcal{A}$ be a semiring and $A, A_1, \ldots, A_n \in \mathcal{A}$. Then $A \setminus \biguplus_{i=1}^{n} A_i$ is a finite disjoint union of sets in $\mathcal{A}$. Symbolically, $A \setminus \biguplus_{i=1}^{n} A_i = \biguplus_{j=1}^{m} B_j$ for some pairwise disjoint $B_1, \ldots, B_m \in \mathcal{A}$.

Proof. We prove by induction. For $n = 1$, we have $A \setminus A_1 = \biguplus_{j=1}^m B_j$ by definition of semiring ("stable" under set differences).

Now suppose the case $n = k$ holds for a $k \in \mathbb{N}$. Then consider:

$$A \setminus \biguplus_{i=1}^{k+1} A_i = \left(A \setminus \biguplus_{i=1}^k A_i \right) \setminus A_{k+1} = \left(\biguplus_{j=1}^m B_j \right) \setminus A_{k+1} = \biguplus_{j=1}^m (B_j \setminus A_{k+1})$$

$$\overset{\text{("stability" under set differences)}}{=} \biguplus_{j=1}^m \bigcup_{\ell=1}^{m_j} \underbrace{B_{j\ell}}_{\in \mathcal{A}},$$

which is a finite disjoint union of sets in $\mathcal{A}$, so the case $n = k + 1$ holds, completing the proof by induction. $\qquad\square$

2.3.4 **Carathéodory extension theorem.** Now we are ready to prove the Carathéodory extension theorem. While the statement is short, the proof turns out to be very lengthy!

Theorem 2.3.b (Carathéodory extension theorem). Let $\mathcal{A}$ be a semiring on Ω and μ_0 be a σ-finite premeasure on $\mathcal{A}$. Then μ_0 can be uniquely extended to a σ-finite measure μ on $\sigma(\mathcal{A})$.

Proof. **Showing that μ_0 induces an outer measure μ^*.** Let $\mu^* : \mathcal{P}(\Omega) \to [0, \infty]$ be a function defined by

$$\mu^*(A) := \inf_{\substack{A_1, A_2, \ldots \in \mathcal{A}: \\ \bigcup_{i=1}^\infty A_i \supseteq A}} \sum_{i=1}^\infty \mu_0(A_i),$$

the infimum of the sums of premeasures $\sum_{i=1}^\infty \mu_0(A_i)$ (serving as approximations to the "volume" of A) over all possible $A_1, A_2, \ldots, \in \mathcal{A}$ with $\bigcup_{i=1}^\infty A_i \supseteq A$, i.e., over all possible *covers* $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{A}$ of A. [Note: We have $\inf \varnothing := \infty$.]
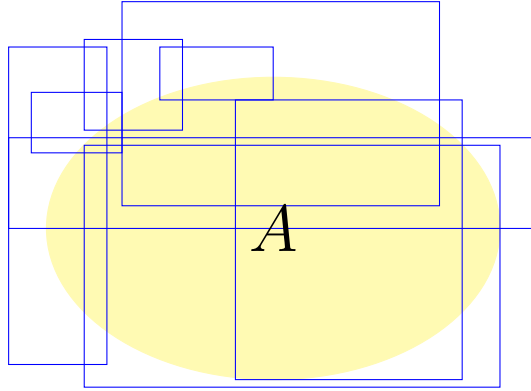
Two extreme cases:

- If no such cover $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{A}$ exists, then $\mu^*(A) = \infty$ *(largest possible sum of premeasures)*.
- If there is a cover $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{A}$ of A such that $\mu_0(A_i) = 0 \; \forall i \in \mathbb{N}$, then $\mu^*(A) = 0$ *(smallest possible sum of premeasures)*.

Now we show that μ^* is an outer measure.

(1) By setting $A_i = \varnothing \in \mathcal{A} \; \forall i \in \mathbb{N}$, we get a cover of $\varnothing$ such that $\mu_0(A_i) = \mu_0(\varnothing) = 0 \; \forall i \in \mathbb{N}$, thus $\mu^*(\varnothing) = 0$.

(2) Fix any $A, B \subseteq \Omega$ with $A \subseteq B$. Since every cover $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{A}$ of B must also cover A (but not vice versa), we conclude that $\mu^*(A) \leq \mu^*(B)$ as we are taking infimum over a larger number of covers for the former.

(3) Fix any $A_1, A_2, \ldots \in \mathcal{P}(\Omega)$. In case $\mu^*(A_i) = \infty$ for some $i \in \mathbb{N}$, the σ-subadditivity would just be saying "$\infty \leq \infty$", which trivially holds. Thus, assume henceforth that $\mu^*(A_i) < \infty$ for all $i \in \mathbb{N}$.

Fix any $i \in \mathbb{N}$. By definition of the infimum, for all $\varepsilon > 0$, there exists a cover $\{A_{ik}\}_{k \in \mathbb{N}}$ of A_i such that $\sum_{k=1}^{\infty} \mu_0(A_{ik}) \leq \mu^*(A_i) + \varepsilon/2^i$. Collecting these covers for all $i \in \mathbb{N}$ together yields a cover of $\bigcup_{i=1}^{\infty} A_i$ since $\bigcup_{i=1}^{\infty} \bigcup_{k=1}^{\infty} A_{ik} \supseteq \bigcup_{i=1}^{\infty} A_i$. Then consider:

$$\mu^*\left(\bigcup_{i=1}^{\infty} A_i\right) \overset{\text{(infimum)}}{\leq} \underbrace{\sum_{i=1}^{\infty}\sum_{k=1}^{\infty} \mu_0(A_{ik})}_{\text{sum of premeasures for } \textit{one} \text{ cover}} \leq \sum_{i=1}^{\infty}(\mu^*(A_i) + \varepsilon/2^i) = \varepsilon + \sum_{i=1}^{\infty} \mu^*(A_i)$$

As this holds for all $\varepsilon > 0$, we conclude that $\mu^*(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu^*(A_i)$.

Goal: Showing that μ^* extends μ_0.

Idea: We start with a *semiring* $\mathcal{A}$ on which μ_0 is defined. To show the desired extension, we would need to first show the monotonicity and σ-subadditivity of μ_0. However, the semiring $\mathcal{A}$ itself does not have enough "structure" for us to do that, thus we need to use a somewhat indirect approach which considers a *ring* $\mathcal{A}_\uplus$ *generated by* $\mathcal{A}$, and prove those two properties on the ring $\mathcal{A}_\uplus \supseteq \mathcal{A}$, which would then imply the properties hold on $\mathcal{A}$ as well.

Showing that $\mathcal{A}_\uplus := \{\uplus_{i=1}^n A_i : A_i \in \mathcal{A}, n \in \mathbb{N}\}$ is a ring *(ring generated by $\mathcal{A}$)*. It is clear that $\mathcal{A}_\uplus \supseteq \mathcal{A}$ by considering one-set unions.

Note that $\mathcal{A}_\uplus$ is a π-system since for all $A, B \in \mathcal{A}_\uplus$, we can write $A = \bigcup_{i=1}^n A_i$ and $B = \uplus_{j=1}^m B_j$, so $A \cap B = \bigcup_{i=1}^n \bigcup_{j=1}^m \underbrace{(A_i \cap B_j)}_{\in \mathcal{A} \text{ due to closedness under intersections}}$ is a finite disjoint union of sets in $\mathcal{A}$, thus $A \cap B \in \mathcal{A}_\uplus$.

Now we show that $\mathcal{A}_\uplus$ is a ring.

(1) Since $\varnothing \in \mathcal{A}$, we have $\varnothing \in \mathcal{A}_\uplus$ (take $n = 1$ and $A_1 = \varnothing$).

(2) Fix any $A, B \in \mathcal{A}_\uplus$. Then write $A = \uplus_{i=1}^n A_i$ and $B = \uplus_{j=1}^m B_j$. Hence,

$$A \setminus B = \left(\biguplus_{i=1}^n A_i\right) \setminus \left(\biguplus_{j=1}^m B_j\right) = \biguplus_{i=1}^n \left(A_i \setminus \left(\biguplus_{j=1}^m B_j\right)\right) \overset{\text{(Lemma 2.3.a)}}{=} \biguplus_{i=1}^n \biguplus_{k=1}^p \underbrace{C_{ik}}_{\in \mathcal{A}},$$

which is a finite disjoint union of sets in $\mathcal{A}$, thus belongs to $\mathcal{A}_\uplus$.

(3) Fix any $A, B \in \mathcal{A}_\uplus$. Then from above we know $A \setminus B \in \mathcal{A}_\uplus$. Also, $A \cap B \in \mathcal{A}_\uplus$ as $\mathcal{A}_\uplus$ is a π-system. So both $A \setminus B$ and $A \cap B$ are finite disjoint unions of sets in $\mathcal{A}$, hence $A \cup B = (A \setminus B) \uplus (A \cap B)$ is a finite disjoint union of sets in $\mathcal{A}$, which belongs to $\mathcal{A}_\uplus$.

Extending μ_0 from $\mathcal{A}$ to $\mathcal{A}_\uplus$. To perform the extension, we define $\mu_0(\uplus_{i=1}^n A_i) := \sum_{i=1}^n \mu_0(A_i)$ for all $A_1, \ldots, A_n \in \mathcal{A}$. (By considering union of one set, we see that the measures assigned for sets in $\mathcal{A}$ remain unchanged, aligning with the extension property.) Now we will show that such extended μ_0 is well-defined (which would then imply uniqueness).

Suppose we have $\uplus_{i=1}^n A_i = \uplus_{j=1}^m B_j$ where $A_1, \ldots, A_n, B_1, \ldots, B_m \in \mathcal{A}$. Then we have $A_i \overset{(A_i \subseteq \uplus_{i=1}^n A_i)}{=} A_i \cap \uplus_{i=1}^n A_i = A_i \cap \uplus_{j=1}^m B_j = \uplus_{j=1}^m(A_i \cap B_j)$. Thus, by finite additivity of μ_0 on $\mathcal{A}$, we have $\mu_0(A_i) = \sum_{j=1}^m \mu_0(A_i \cap B_j)$. Therefore,

$$\sum_{i=1}^n \mu_0(A_i) = \sum_{i=1}^n \sum_{j=1}^m \mu_0(A_i \cap B_j) = \sum_{j=1}^m \sum_{i=1}^n \mu_0(B_j \cap A_i) \overset{\text{(similar argument)}}{=} \sum_{j=1}^m \mu_0(B_j),$$

establishing the well-definedness.

Showing that μ_0 is a premeasure on $\mathcal{A}_\uplus$. As μ_0 is already a premeasure on $\mathcal{A}$, it suffices to show σ-additivity of μ_0 on $\mathcal{A}_\uplus$ (which is a ring, hence also a semiring).

Fix any pairwise disjoint $A_1, A_2, \ldots \in \mathcal{A}_\uplus$ with $\uplus_{i=1}^\infty A_i \in \mathcal{A}_\uplus$ (don't miss ⚠). Then, we can express each A_i as a disjoint union of sets in $\mathcal{A}$. Symbolically, we can write $A_i = \uplus_{j=1}^{n_i} A_{ij} \ \forall i \in \mathbb{N}$. As A_i's are also pairwise disjoint themselves, we know the A_{ij}'s (across both i and j) are pairwise disjoint.

Note that we can write $\uplus_{i=1}^\infty A_i = \uplus_{i=1}^\infty \uplus_{j=1}^{n_i} A_{ij}$, which is a countable disjoint union of sets in $\mathcal{A}$ (those A_{ij}'s). Hence, we have

$$\mu_0\left(\biguplus_{i=1}^\infty A_i\right) \overset{(\sigma\text{-add. on } \mathcal{A})}{=} \sum_{i=1}^\infty \sum_{j=1}^{n_i} \mu_0(A_{ij}) \overset{(\text{extension of } \mu_0)}{=} \sum_{i=1}^\infty \mu_0\left(\biguplus_{j=1}^{n_i} A_{ij}\right) = \sum_{i=1}^\infty \mu_0(A_i).$$

Showing the monotonicity and σ-subadditivity of μ_0 on $\mathcal{A}_\uplus$. Techniques similar to the ones used for proving the monotonicity and σ-subadditivity of a general measure in [2.1.3] can be utilized here.

- *Monotonicity:* Fix any $A, B \in \mathcal{A}_\uplus$ with $A \subseteq B$. Note that $A = A \cap B \overset{(\text{ring} \Rightarrow \text{semiring})}{\in} \mathcal{A}_\uplus$, and $B \setminus A \in \mathcal{A}_\uplus$ by closedness under set differences. Hence, writing $B = (A \cap B) \uplus (B \setminus A) \in \mathcal{A}_\uplus$ gives

$$\mu(B) \overset{(\text{finite additivity})}{=} \mu(A \cap B) + \mu(B \setminus A) = \mu(A) + \underbrace{\mu(B \setminus A)}_{\geq 0} \geq \mu(A).$$

- *σ-subadditivity:* Fix any $A_1, A_2, \ldots \in \mathcal{A}_\uplus$ with $\bigcup_{i=1}^\infty A_i \in \mathcal{A}_\uplus$. Let $B_1 := A_1 \in \mathcal{A}_\uplus$ and $B_n := A_n \setminus \bigcup_{i=1}^{n-1} A_i \in \mathcal{A}_\uplus$ for every integer $n \geq 2$. Then by construction, (i) B_n's are disjoint, (ii) $B_n \subseteq A_n$ for every $n \in \mathbb{N}$, and (iii) $\bigcup_{i=1}^n A_i = \uplus_{i=1}^n B_i$ for every $n \in \mathbb{N}$.

  By (iii), we have $\bigcup_{i=1}^\infty A_i = \lim_{N \to \infty} \bigcup_{i=1}^N A_i = \lim_{N \to \infty} \uplus_{i=1}^N B_i = \uplus_{i=1}^\infty B_i$. Thus,

$$\mu\left(\bigcup_{i=1}^\infty A_i\right) = \mu\left(\biguplus_{i=1}^\infty B_i\right) = \sum_{i=1}^\infty \mu(B_i) \overset{((\text{ii}), \text{ monotonicity})}{\leq} \sum_{i=1}^\infty \mu(A_i).$$

Showing that μ^* extends μ_0 on $\mathcal{A}$. Fix any $A \in \mathcal{A}$ and any cover $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{A}$ of A. Then, we have $\bigcup_{i=1}^\infty A_i \supseteq A$, which implies $\bigcup_{i=1}^\infty (A_i \cap A) = (\bigcup_{i=1}^\infty A_i) \cap A = A$. Thus,

$$\mu_0(A) = \mu_0\left(\bigcup_{i=1}^\infty (A_i \cap A)\right) \overset{(\sigma\text{-subadditivity})}{\leq} \sum_{i=1}^\infty \mu_0(A_i \cap A) = \lim_{n \to \infty} \sum_{i=1}^n \mu_0(A_i \cap A)$$

$$\overset{(\text{monotonicity, } A_i \cap A \subseteq A)}{\leq} \lim_{n \to \infty} \sum_{i=1}^n \mu_0(A_i) = \sum_{i=1}^\infty \mu_0(A_i).$$

As this holds for all covers $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{A}$ of A, $\mu_0(A)$ is a lower bound for the sums of premeasures of the covers. Since the infimum is the greatest lower bound, we have $\mu_0(A) \leq \mu^*(A)$.

On the other hand, note that $\{A, \varnothing, \varnothing, \ldots\} \subseteq \mathcal{A}$ is a cover of A, and $\mu_0(A)$ is the sum of premeasures for this cover. So, $\mu^*(A) \overset{(\text{infimum})}{\leq} \mu_0(A)$. Hence we have $\mu^*(A) = \mu_0(A) \ \forall A \in \mathcal{A}$, meaning that μ^* extends μ_0 on $\mathcal{A}$.

Goal: Applying the principle of good sets to show that every set in $\sigma(\mathcal{A})$ is Carathéodory-measurable

Let $\mathcal{A}^* := \{A \subseteq \Omega : \mu^*(B) = \mu^*(B \cap A) + \mu^*(B \cap A^c) \ \forall B \subseteq \Omega\}$ be the family of all **Carathéodory-measurable sets**, or "good" sets. To apply the principle of good sets, we will show that (i) $\mathcal{A} \subseteq \mathcal{A}^*$ and (ii) $\mathcal{A}^*$ is a σ-algebra, which would then imply that every set in $\sigma(\mathcal{A})$ is "good" (Carathéodory-measurable).

Showing that $\mathcal{A} \subseteq \mathcal{A}^*$. Fix any $A \in \mathcal{A}$ and any $B \subseteq \Omega$. Using a property of the infimum, for all $\varepsilon > 0$, there is a cover $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{A}$ of B such that $\sum_{i=1}^{\infty} \mu_0(A_i) \leq \mu^*(B) + \varepsilon$. Let $E_i := A_i \cap A \overset{\text{(semiring)}}{\in} \mathcal{A}$ for every $i \in \mathbb{N}$. Then, for every $i \in \mathbb{N}$, we have $A_i \setminus A = A_i \setminus E_i \overset{\text{(semiring)}}{=} \biguplus_{j=1}^{n_i} B_{ij}$ for some pairwise disjoint $B_{i1}, \ldots, B_{in_i} \in \mathcal{A}$.

Thus, we have:

- $A_i = E_i \uplus (A_i \setminus E_i) = E_i \uplus \bigcup_{j=1}^{n_i} B_{ij}$.

- $B \cap A \overset{\text{(cover)}}{\subseteq} \left(\bigcup_{i=1}^{\infty} A_i \right) \cap A = \bigcup_{i=1}^{\infty} (A_i \cap A) = \bigcup_{i=1}^{\infty} E_i$.

- $B \setminus A \overset{\text{(cover)}}{\subseteq} \left(\bigcup_{i=1}^{\infty} A_i \right) \setminus A = \bigcup_{i=1}^{\infty} (A_i \setminus A) = \bigcup_{i=1}^{\infty} \biguplus_{j=1}^{n_i} B_{ij}$.
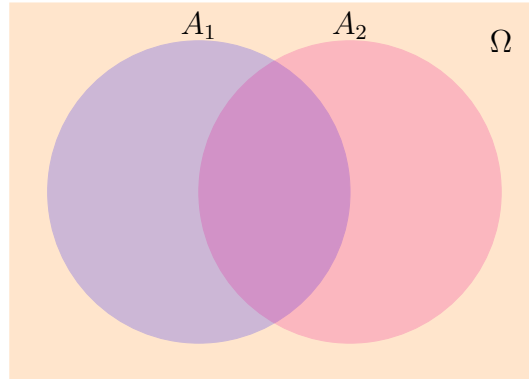
Hence,

$$
\mu^*(B \cap A) + \mu^*(B \setminus A) \overset{\text{(monotonicity)}}{\leq} \mu^* \left(\bigcup_{i=1}^{\infty} E_i \right) + \mu^* \left(\bigcup_{i=1}^{\infty} \biguplus_{j=1}^{n_i} B_{ij} \right)
$$

$$
\overset{\text{(σ-subadditivity)}}{\leq} \sum_{i=1}^{\infty} \mu^*(E_i) + \sum_{i=1}^{\infty} \sum_{j=1}^{n_i} \mu^*(B_{ij})
$$

$$
\overset{\text{(μ^* extends μ on $\mathcal{A}$)}}{=} \sum_{i=1}^{\infty} \mu_0(E_i) + \sum_{i=1}^{\infty} \sum_{j=1}^{n_i} \mu_0(B_{ij})
$$

$$
= \sum_{i=1}^{\infty} \left(\mu_0(E_i) + \sum_{j=1}^{n_i} \mu_0(B_{ij}) \right) = \sum_{i=1}^{\infty} \mu \left(E_i \uplus \bigcup_{j=1}^{n_i} B_{ij} \right) = \sum_{i=1}^{\infty} \mu_0(A_i) \leq \mu^*(B) + \varepsilon
$$

for all $\varepsilon > 0$, which implies $\mu^*(B \cap A) + \mu^*(B \setminus A) \leq \mu^*(B)$. On the other hand, by finite subadditivity we have $\mu^*(B) \leq \mu^*(B \cap A) + \mu^*(B \setminus A)$. So $\mu^*(B) = \mu^*(B \cap A) + \mu^*(B \setminus A)$. As this holds for all $B \subseteq \Omega$, we have $A \in \mathcal{A}^*$. This shows $\mathcal{A} \subseteq \mathcal{A}^*$.

Showing that $\mathcal{A}^*$ is a σ-algebra on Ω. Due to Proposition 1.4.c, we will show that $\mathcal{A}^*$ is a σ-algebra on Ω by showing it is both (i) π-system and (ii) Dynkin system.

- *π-system:* Fix any $A_1, A_2 \in \mathcal{A}^*$ and any $B \subseteq \Omega$. Observe that $(A_1 \cap A_2)^c = (A_1^c \cap A_2) \uplus (A_1 \cap A_2^c) \uplus (A_1^c \cap A_2^c)$.

Hence, we have

$$\mu^*(B \cap (A_1 \cap A_2)) + \mu^*(B \cap (A_1 \cap A_2)^c)$$
$$= \mu^*(B \cap A_1 \cap A_2) + \mu^*(B \cap ((A_1^c \cap A_2) \uplus (A_1 \cap A_2^c) \uplus (A_1^c \cap A_2^c)))$$
$$\overset{\text{(finite subadditivity)}}{\leq} \mu^*(B \cap A_1 \cap A_2) + \mu^*(B \cap A_1^c \cap A_2) + \mu^*(B \cap A_1 \cap A_2^c) + \mu^*(B \cap A_1^c \cap A_2^c)$$
$$= \mu^*(B \cap A_2 \cap A_1) + \mu^*(B \cap A_2 \cap A_1^c)$$
$$+ \mu^*(B \cap A_2^c \cap A_1) + \mu^*(B \cap A_2^c \cap A_1^c) \overset{(A_1 \in \mathcal{A}^*)}{\underset{(B \cap A_2, \, B \cap A_2^c \subseteq \Omega)}{=}} \mu^*(B \cap A_2) + \mu^*(B \cap A_2^c) = \mu^*(B).$$

On the other hand, by finite subadditivity we have $\mu^*(B \cap (A_1 \cap A_2)) + \mu^*(B \cap (A_1 \cap A_2)^c) \geq \mu(B)$. So the equality holds and $A_1 \cap A_2 \in \mathcal{A}^*$.

- *Dynkin system:*

  (1) We have shown that $\varnothing \in \mathcal{A}^*$ above.

  (2) From the definition of $\mathcal{A}^*$, we can clearly see that $A \in \mathcal{A}^*$ iff $A^c \in \mathcal{A}^*$.

  (3) Fix any pairwise disjoint $A_1, A_2, \ldots \in \mathcal{A}^*$. Let $\widetilde{A}_n := \biguplus_{i=1}^n A_i$ for every $n \in \mathbb{N}$ and $\widetilde{A}_\infty := \biguplus_{i=1}^\infty A_i$. By De Morgan's law, we have $\widetilde{A}_n = (\bigcap_{i=1}^n A_i^c)^c \in \mathcal{A}^*$, since $\mathcal{A}^*$ is closed under complements (proven above) and intersections (as it is shown to be a π-system).
  Now, note that for all $B \subseteq \Omega$, we have $\forall n \in \mathbb{N}$,

  $$\mu^*(B \cap \widetilde{A}_{n+1}) \overset{(\widetilde{A}_n \in \mathcal{A}^*)}{=} \mu^*((B \cap \widetilde{A}_{n+1}) \cap \widetilde{A}_n) + \mu^*((B \cap \widetilde{A}_{n+1}) \cap (\widetilde{A}_n)^c) = \mu^*(B \cap \widetilde{A}_n) + \mu^*(B \cap A_{n+1})$$

  Noting that $\widetilde{A}_1 = A_1$, this recursive relationship yields $\mu^*(B \cap \widetilde{A}_n) = \sum_{i=1}^n \mu^*(B \cap A_i) \; \forall n \in \mathbb{N}$. Then consider, for all $n \in \mathbb{N}$,

  $$\mu^*((B \cap \widetilde{A}_\infty) \cup (B \cap \widetilde{A}_\infty^c)) = \mu^*(B) \overset{(\widetilde{A}_n \in \mathcal{A}^*)}{=} \mu^*(B \cap \widetilde{A}_n) + \mu^*(B \cap \widetilde{A}_n^c)$$
  $$\overset{\text{(monotonicity)}}{\geq} \mu^*(B \cap \widetilde{A}_n) + \mu^*(B \cap \widetilde{A}_\infty^c)$$
  $$= \sum_{i=1}^n \mu^*(B \cap A_i) + \mu^*(B \cap \widetilde{A}_\infty^c).$$

  As this holds for all $n \in \mathbb{N}$, taking limit $n \to \infty$ gives

  $$\mu^*((B \cap \widetilde{A}_\infty) \cup (B \cap \widetilde{A}_\infty^c)) \geq \sum_{i=1}^\infty \mu^*(B \cap A_i) + \mu^*(B \cap \widetilde{A}_\infty^c) \qquad (4)$$
  $$\overset{\text{(σ-subadditivity)}}{\geq} \mu^*\left(\biguplus_{i=1}^\infty (B \cap A_i)\right) + \mu^*(B \cap \widetilde{A}_\infty^c)$$
  $$\overset{\text{(distributivity)}}{=} \mu^*(B \cap \widetilde{A}_\infty) + \mu^*(B \cap \widetilde{A}_\infty^c).$$

  On the other hand, by finite subadditivity we have

  $$\mu^*((B \cap \widetilde{A}_\infty) \cup (B \cap \widetilde{A}_\infty^c)) \leq \mu^*(B \cap \widetilde{A}_\infty) + \mu^*(B \cap \widetilde{A}_\infty^c).$$

  Hence the equality holds and $\widetilde{A}_\infty \in \mathcal{A}^*$.

Goal: Extending uniquely μ_0 to a σ-finite measure μ on $\sigma(\mathcal{A})$

Showing that μ^* is a measure on the σ-algebra $\mathcal{A}^*$.

(1) We can view the outer measure μ^* as a function on $\mathcal{A}^* \subseteq \mathcal{P}(\Omega)$. Also its codomain is $[0, \infty]$ by the definition of outer measure.

(2) By the definition of outer measure, we have $\mu^*(\varnothing) = 0$.

(3) Setting $B = \widetilde{A}_\infty$ in (4) gives $\mu^*(\widetilde{A}_\infty) \geq \sum_{i=1}^\infty \mu^*(A_i)$. On the other hand, by σ-subadditivity we have $\mu^*(\widetilde{A}_\infty) \leq \sum_{i=1}^\infty \mu^*(A_i)$. Thus the equality holds, implying the σ-additivity on $\mathcal{A}^*$.

Extending μ_0 to a measure μ. By the principle of good sets, we have previously shown that $\sigma(\mathcal{A}) \subseteq \mathcal{A}^*$. Thus we can define the restriction $\mu := \mu^*|_{\sigma(\mathcal{A})}$, which is a measure on $\sigma(\mathcal{A})$. (The underlying measure space is $(\Omega, \sigma(\mathcal{A}), \mu)$.) Also, note that $\mu|_{\mathcal{A}} \overset{(\mathcal{A} \subseteq \sigma(\mathcal{A}))}{=} \mu^*|_{\mathcal{A}} \overset{(\mu^* \text{ extends } \mu_0)}{=} \mu_0|_{\mathcal{A}}$, so μ extends μ_0.

Showing the σ-finiteness of μ and the uniqueness of such measure.

- *σ-finiteness:* By assumption, μ_0 is σ-finite on $\mathcal{A}$, i.e., $\Omega = \bigcup_{i=1}^\infty A_i$ for some $A_1, A_2, \ldots \in \mathcal{A} \subseteq \sigma(\mathcal{A})$ with $\mu_0(A_i) < \infty \; \forall i \in \mathbb{N}$. Since $\mu|_{\mathcal{A}} = \mu_0|_{\mathcal{A}}$, we have also $\mu(A_i) < \infty \; \forall i \in \mathbb{N}$, thus μ is σ-finite.

- *Uniqueness:* Note that the semiring $\mathcal{A}$ is also a π-system, and that the *σ-finiteness on $\mathcal{A}$* is satisfied. Thus, by Proposition 2.1.a, another σ-finite measure ν on $\sigma(\mathcal{A})$ that extends μ_0 must coincide with μ, establishing the uniqueness.

$\square$

2.3.5 By Theorem 2.2.b, a measure can always be made complete, including the extension μ here. Hence, by replacing it with the completed version if necessary, we may assume that the measure μ extended from μ_0 in Theorem 2.3.b is complete; this will be assumed henceforth.

2.4 Borel Measures on $\mathbb{R}^d$

2.4.1 One main application of Theorem 2.3.b is to construct *Borel measures* on $\mathbb{R}^d$, which include the frequently used *Lebesgue measure* on $\mathbb{R}^d$ (corresponding to our usual notion of "volume") as a special case.

2.4.2 **Preliminary concepts.** Let us first go through some concepts to be used in the construction. A function $F : \mathbb{R}^d \to \mathbb{R}$ is:

- **right-continuous** if $F(\boldsymbol{x}) = \lim_{\boldsymbol{h} \to \boldsymbol{0}^+} F(\boldsymbol{x} + \boldsymbol{h}) =: F(\boldsymbol{x}+)$ for all $\boldsymbol{x} \in \mathbb{R}^d$.
  [Note: "$F(\boldsymbol{x}) = \lim_{\boldsymbol{h} \to \boldsymbol{0}^+} F(\boldsymbol{x} + \boldsymbol{h})$" means that for every sequence $\{\boldsymbol{h}_n\} \to \boldsymbol{0}$ with $\boldsymbol{h}_n \geq \boldsymbol{0}$ for each $n \in \mathbb{N}$, we have $\lim_{\boldsymbol{h}_n \to \boldsymbol{0}} F(\boldsymbol{x} + \boldsymbol{h}_n) = F(\boldsymbol{x})$.]

- **grounded** if $\lim_{x_j \to -\infty} F(\boldsymbol{x}) = 0$ (with $\boldsymbol{x} = (x_1, \ldots, x_n)$) for all $j = 1, \ldots, d$.

- **d-increasing** if $\Delta_{(\boldsymbol{a},\boldsymbol{b}]}F \geq 0$ for all $\boldsymbol{a} \leq \boldsymbol{b}$, where $\Delta_{(\boldsymbol{a},\boldsymbol{b}]}F$ is the F-**volume** (over $(\boldsymbol{a},\boldsymbol{b}]$):

$$\Delta_{(\boldsymbol{a},\boldsymbol{b}]}F := \sum_{\boldsymbol{i} \in \{0,1\}^d} (-1)^{i_1 + \cdots + i_d} F\big(a_1^{i_1} b_1^{1-i_1}, \ldots, a_d^{i_d} b_d^{1-i_d}\big)$$

$$\overset{\text{(in words)}}{=} \sum_{\boldsymbol{i} \in \{0,1\}^d} (\text{alternating } +/- \text{ sign}) \, F\left(\begin{cases} b_1 & \text{if } i_1 = 0 \\ a_1 & \text{if } i_1 = 1 \end{cases}, \ldots, \begin{cases} b_d & \text{if } i_d = 0 \\ a_d & \text{if } i_d = 1 \end{cases}\right),$$

with $\boldsymbol{i} = (i_1, \ldots, i_d)$, $\boldsymbol{a} = (a_1, \ldots, a_d)$, and $\boldsymbol{b} = (b_1, \ldots, b_d)$.

Examples:

- *(d = 1)* $\Delta_{(a_1,b_1]}F = F(b_1) - F(a_1)$. [Note: In this case, *d-increasing* coincides with our usual notion of *increasing*.]

- *(d = 2)* $\Delta_{(a_1,b_1] \times (a_2,b_2]}F = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2)$.
  Note that:
  * $\Delta^{(2)}_{(a_2,b_2]} \Delta^{(1)}_{(a_1,b_1]} F(x_1, x_2) = \Delta^{(1)}_{(a_1,b_1]} F(x_1, b_2) - \Delta^{(1)}_{(a_1,b_1]} F(x_1, a_2) = F(b_1, b_2) - F(a_1, b_2) - (F(b_1, a_2) - F(a_1, a_2)) = \Delta_{(a_1,b_1] \times (a_2,b_2]}F$.
  * $\Delta^{(1)}_{(a_1,b_1]} \Delta^{(2)}_{(a_2,b_2]} F(x_1, x_2) = \Delta^{(2)}_{(a_2,b_2]} F(b_1, x_2) - \Delta^{(2)}_{(a_2,b_2]} F(a_1, x_2) = F(b_1, b_2) - F(b_1, a_2) - (F(a_1, b_2) - F(a_1, a_2)) = \Delta_{(a_1,b_1] \times (a_2,b_2]}F$.

[Note: The notation $\Delta_{(a,b]}^{(i)}F(x_1, \ldots, x_n)$ refers to the *first-order difference* with ith input of F substituted, i.e., $F(x_1, \ldots, x_{i-1}, b, x_{i+1}, \ldots, x_n) - F(x_1, \ldots, x_{i-1}, a, x_{i+1}, \ldots, x_n)$. Note that the ith input in the parenthesis would not affect the first-order difference (as it will get substituted anyway), but the other inputs matter as they would remain the same in the resulting expressions. For convenience in notations, sometimes we will drop the superscript "(i)" when the meaning is clear from context.]

Before proceeding further, let us also introduce some special notations that will be useful later on. Let $J \subseteq \{1, \ldots, d\}$. Then we have the following notations for vectors:

- $\boldsymbol{x}_J := (x_j)_{j \in J}$.
- $\boldsymbol{x}_{J^c} := (x_j)_{j \notin J}$.
- For all $x \in \mathbb{R}$, $_d x := (x, \ldots, x) \in \mathbb{R}^d$.
- For all $\boldsymbol{x} = (x_1, \ldots, x_d), \boldsymbol{y} = (y_1, \ldots, y_d) \in \mathbb{R}^d$:

  - $\boldsymbol{x}_{J \leftarrow \boldsymbol{y}_J} := \begin{cases} x_j & \text{if } j \notin J, \\ y_j & \text{if } j \in J. \end{cases}$

  - *Special case:* If $J = \{j\}$, we may simply write $\boldsymbol{x}_{j \leftarrow y_j}$ instead of $\boldsymbol{x}_{J \leftarrow \boldsymbol{y}_J}$. More explicitly, we can write $\boldsymbol{x}_{j \leftarrow y_j} = (x_1, \ldots, x_{j-1}, y_j, x_{j+1}, \ldots, x_d)$.

2.4.3 **Properties of F-volumes.** The definition of the F-volume above may not be intuitive. So to understand it better, we will study some of its properties. Let $F : \mathbb{R}^d \to \mathbb{R}$ be a function.

(a) *(Permutation of first-order differences)* We have $\Delta_{(\boldsymbol{a},\boldsymbol{b}]}F = \Delta_{(a_d,b_d]} \cdots \Delta_{(a_1,b_1]}F(x_1, \ldots, x_d)$ (or just $\Delta_{(a_d,b_d]} \cdots \Delta_{(a_1,b_1]}F)$, and the order of "$\Delta$"s can be freely changed.

(b) *(Monotonicity)* If F is d-increasing, then $(\boldsymbol{a}, \boldsymbol{b}] \subseteq (\boldsymbol{\alpha}, \boldsymbol{\beta}] \implies \Delta_{(\boldsymbol{a},\boldsymbol{b}]}F \leq \Delta_{(\boldsymbol{\alpha},\boldsymbol{\beta}]}F$.

(c) *(Product of differences)* If $F(\boldsymbol{x}) = \prod_{j=1}^d F_j(x_j)$, then $\Delta_{(\boldsymbol{a},\boldsymbol{b}]}F = \prod_{j=1}^d (F_j(b_j) - F_j(a_j))$.

(d) If F is grounded, then $\lim_{\boldsymbol{a} \to -\infty} \Delta_{(\boldsymbol{a},\boldsymbol{x}]}F =: \Delta_{(-\infty,\boldsymbol{x}]}F = F(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

(e) *(Nonnegativity)* If F is grounded and d-increasing, and J is strictly between $\varnothing$ and $\{1, \ldots, d\}$ (i.e, $\varnothing \subsetneq J \subsetneq \{1, \ldots, d\}$), then $\Delta_{(\boldsymbol{a}_J,\boldsymbol{b}_J]}F(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}_{J^c} \in \mathbb{R}^{|J^c|}$. [Note: Particularly, by setting $J = \{j\}$ for each $j = 1, \ldots, d$, this result suggests that F is componentwise increasing under such conditions.]

Proof.

(a) The possibility of freely reordering the first-order differences follows from the commutativity of addition (see the $d = 2$ example above for an illustration). So henceforth we will just prove $\Delta_{(\boldsymbol{a},\boldsymbol{b}]}F = \Delta_{(a_d,b_d]} \cdots \Delta_{(a_1,b_1]}F$ for all $d \in \mathbb{N}$.

The case $d = 1$ holds trivially as both sides are referring to the same thing. Now assume for induction that the case $d = k$ holds for a $k \in \mathbb{N}$. Then,

$$
\begin{aligned}
\Delta_{(\boldsymbol{a},\boldsymbol{b}]}F &= \sum_{\boldsymbol{i} \in \{0,1\}^d} (-1)^{i_1 + \cdots + i_d} F(a_1^{i_1} b_1^{1-i_1}, \ldots, a_d^{i_d} b_d^{1-i_d}) \\
&= \sum_{\boldsymbol{i} \in \{0,1\}^d : i_d = 0} (-1)^{i_1 + \cdots + i_d} F(a_1^{i_1} b_1^{1-i_1}, \ldots, a_{d-1}^{i_{d-1}} b_{d-1}^{1-i_{d-1}}, b_d) \\
&\quad - \sum_{\boldsymbol{i} \in \{0,1\}^d : i_d = 1} (-1)^{i_1 + \cdots + i_d} F(a_1^{i_1} b_1^{1-i_1}, \ldots, a_{d-1}^{i_{d-1}} b_{d-1}^{1-i_{d-1}}, a_d) \\
&= \Delta_{(a_{d-1},b_{d-1}]} \cdots \Delta_{(a_1,b_1]}F(x_1, \ldots, x_{d-1}, b_d) \\
&\quad - \Delta_{(a_{d-1},b_{d-1}]} \cdots \Delta_{(a_1,b_1]}F(x_1, \ldots, x_{d-1}, a_d) \quad \text{(inductive hypothesis)} \\
&= \Delta_{(a_d,b_d]} \Delta_{(a_{d-1},b_{d-1}]} \cdots \Delta_{(a_1,b_1]}F(x_1, \ldots, x_d),
\end{aligned}
$$

so the case $d = k + 1$ holds, completing the proof by induction.

(b) For all $j = 1, \ldots, d$ with $a_j \leq b_j$, we have

$$\Delta_{(a_j, b_j]} \Delta_{(\boldsymbol{a}_{\{j\}^c}, \boldsymbol{b}_{\{j\}^c}]} F(x_1, \ldots, x_j, \ldots, x_d) \overset{[2.4.3]\mathrm{a}}{=} \Delta_{(\boldsymbol{a}, \boldsymbol{b}]} F \overset{(d\text{-increasing})}{\geq} 0,$$

which implies that

$$\Delta_{(\boldsymbol{a}_{\{j\}^c}, \boldsymbol{b}_{\{j\}^c}]} F(x_1, \ldots, b_j, \ldots, x_d) - \Delta_{(\boldsymbol{a}_{\{j\}^c}, \boldsymbol{b}_{\{j\}^c}]} F(x_1, \ldots, a_j, \ldots, x_d) \geq 0$$

for all $a_j \leq b_j$. Thus, the function $x_j \mapsto \Delta_{(\boldsymbol{a}_{\{j\}^c}, \boldsymbol{b}_{\{j\}^c}]} F(x_1, \ldots, x_j, \ldots, x_d)$ is increasing for every $j = 1, \ldots, d$. Particularly, we have

$$\Delta_{(\boldsymbol{a}_{\{j\}^c}, \boldsymbol{b}_{\{j\}^c}]} F(x_1, \ldots, \beta_j, \ldots, x_d) - \Delta_{(\boldsymbol{a}_{\{j\}^c}, \boldsymbol{b}_{\{j\}^c}]} F(x_1, \ldots, \alpha_j, \ldots, x_d)$$
$$\geq \Delta_{(\boldsymbol{a}_{\{j\}^c}, \boldsymbol{b}_{\{j\}^c}]} F(x_1, \ldots, b_j, \ldots, x_d) - \Delta_{(\boldsymbol{a}_{\{j\}^c}, \boldsymbol{b}_{\{j\}^c}]} F(x_1, \ldots, a_j, \ldots, x_d)$$

if $\alpha_j \leq a_j \leq b_j \leq \beta_j$, for every $j = 1, \ldots, d$. Now, since $(\boldsymbol{a}, \boldsymbol{b}] \subseteq (\boldsymbol{\alpha}, \boldsymbol{\beta}]$, applying this iteratively gives

$$\Delta_{(\boldsymbol{a}, \boldsymbol{b}]} F \overset{[2.4.3]\mathrm{a}}{=} \Delta_{(a_d, b_d]} \cdots \Delta_{(a_1, b_1]} F \leq \Delta_{(\alpha_d, \beta_d]} \Delta_{(a_{d-1}, b_d]} \cdots \Delta_{(a_1, b_1]} F$$
$$\overset{[2.4.3]\mathrm{a}}{=} \Delta_{(a_{d-1}, b_d]} \Delta_{(\alpha_d, \beta_d]} \cdots \Delta_{(a_1, b_1]} F \leq \Delta_{(\alpha_{d-1}, \beta_d]} \Delta_{(\alpha_d, \beta_d]} \cdots \Delta_{(a_1, b_1]} F$$
$$\leq \cdots \leq \Delta_{(\boldsymbol{\alpha}, \boldsymbol{\beta}]} F.$$

(c) We use an iterative argument:

$$\Delta_{(\boldsymbol{a}, \boldsymbol{b}]} F \overset{[2.4.3]\mathrm{a}}{=} \Delta_{(a_d, b_d]} \cdots \Delta_{(a_2, b_2]} \Delta_{(a_1, b_1]} F(x_1, \ldots, x_d)$$
$$= \Delta_{(a_d, b_d]} \cdots \Delta_{(a_2, b_2]} \Delta_{(a_1, b_1]} F(x_1) F(x_2) \cdots F(x_d)$$
$$= \Delta_{(a_d, b_d]} \cdots \Delta_{(a_2, b_2]} (F_1(b_1) - F_1(a_1)) F(x_2) \cdots F(x_d)$$
$$= (F_1(b_1) - F_1(a_1)) \Delta_{(a_d, b_d]} \cdots \Delta_{(a_2, b_2]} F(x_2) \cdots F(x_d)$$
$$= \cdots = \prod_{j=1}^{d} (F_j(b_j) - F_j(a_j)).$$

(d) For all $\boldsymbol{x} \in \mathbb{R}^d$, we have

$$\Delta_{(-\infty, \boldsymbol{x}]} F = \lim_{\boldsymbol{a} \to -\infty} \Delta_{(\boldsymbol{a}, \boldsymbol{x}]} F = \lim_{\boldsymbol{a} \to -\infty} \sum_{\boldsymbol{i} \in \{0,1\}^d} (-1)^{i_1 + \cdots + i_d} F\big(a_1^{i_1} x_1^{1-i_1}, \ldots, a_d^{i_d} x_d^{1-i_d}\big)$$
$$= \sum_{\boldsymbol{i} \in \{0,1\}^d} (-1)^{i_1 + \cdots + i_d} \lim_{\boldsymbol{a} \to -\infty} F\big(a_1^{i_1} x_1^{1-i_1}, \ldots, a_d^{i_d} x_d^{1-i_d}\big). \tag{5}$$

We now analyze the limit $\lim_{\boldsymbol{a} \to -\infty} F\big(a_1^{i_1} x_1^{1-i_1}, \ldots, a_d^{i_d} x_d^{1-i_d}\big)$ for every $\boldsymbol{i} = (i_1, \ldots, i_d) \in \{0,1\}^d$:

- *Case 1: $i_j = 1$ for some $j = 1, \ldots, d$.* Then, the groundedness of F forces the limit to be zero.
- *Case 2: $i_j = 0$ for all $j = 1, \ldots, d$.* Then the limit is just $\lim_{\boldsymbol{a} \to -\infty} F(x_1, \ldots, x_d) = F(\boldsymbol{x})$.

Therefore, the sum in (5) equals $F(\boldsymbol{x})$, as desired.

(e) Fix any $\boldsymbol{x}_{J^c} \in \mathbb{R}^{|J^c|}$. Then take $\boldsymbol{a}^* = \boldsymbol{a}_{J^c \leftarrow -\infty}$ and $\boldsymbol{b}^* = \boldsymbol{b}_{J^c \leftarrow \boldsymbol{x}_{J^c}}$ (note that $\boldsymbol{a}^* \leq \boldsymbol{b}^*$), and we have

$$\Delta_{(\boldsymbol{a}_J, \boldsymbol{b}_J]} F(\boldsymbol{x}) \overset{[2.4.3]\mathrm{d}}{=} \Delta_{(-\infty, \boldsymbol{x}]} \Delta_{(\boldsymbol{a}_J, \boldsymbol{b}_J]} F(\boldsymbol{x}) \overset{[2.4.3]\mathrm{a}}{=} \Delta_{(\boldsymbol{a}^*, \boldsymbol{b}^*]} F \overset{(d\text{-increasing})}{\geq} 0.$$

$\square$

2.4.4 Constructing Borel measures on $\mathbb{R}^d$.

We now have enough ingredients to construct Borel measures on $\mathbb{R}^d$, which is the main goal of Section 2.4.

Theorem 2.4.a (Construction of Borel measures on $\mathbb{R}^d$)**.** If $F : \mathbb{R}^d \to \mathbb{R}$ is d-increasing and right-continuous, then there exists a unique Borel measure λ_F such that $\lambda_F((\boldsymbol{a}, \boldsymbol{b}]) = \Delta_{(\boldsymbol{a}, \boldsymbol{b}]} F$ for all $\boldsymbol{a} \leq \boldsymbol{b}$.

Proof. **Showing that $\mathcal{A} := \{(\boldsymbol{a}, \boldsymbol{b}] : -\infty < \boldsymbol{a} \leq \boldsymbol{b} < \infty\}$ is a semiring on $\mathbb{R}^d$.**

(1) Note that $\varnothing = (\boldsymbol{a}, \boldsymbol{a}] \in \mathcal{A}$.

(2) Fix any $(\boldsymbol{a}_1, \boldsymbol{b}_1]$ and $(\boldsymbol{a}_2, \boldsymbol{b}_2]$ in $\mathcal{A}$, where $-\infty < \boldsymbol{a}_i \leq \boldsymbol{b}_i < \infty$ for all $i = 1, 2$. Since $(\boldsymbol{a}_1, \boldsymbol{b}_1] \backslash (\boldsymbol{a}_2, \boldsymbol{b}_2]$ is an union of finitely many hypercubes (consider the case $d = 2$ for example), it is in $\mathcal{A}$.

(3) Fix any $(\boldsymbol{a}_1, \boldsymbol{b}_1]$ and $(\boldsymbol{a}_2, \boldsymbol{b}_2]$ in $\mathcal{A}$, where $-\infty < \boldsymbol{a}_i \leq \boldsymbol{b}_i < \infty$ for all $i = 1, 2$. Then, we have

$$(\boldsymbol{a}_1, \boldsymbol{b}_1] \cap (\boldsymbol{a}_2, \boldsymbol{b}_2] = \left(\begin{bmatrix} \max\{a_{1,1}, a_{2,1}\} \\ \vdots \\ \max\{a_{1,d}, a_{2,d}\} \end{bmatrix}, \begin{bmatrix} \min\{b_{1,1}, b_{2,1}\} \\ \vdots \\ \min\{b_{1,d}, b_{2,d}\} \end{bmatrix} \right] \in \mathcal{A}.$$

Here the interval is interpreted as $\varnothing \in \mathcal{A}$ if $\max\{a_{1,j}, a_{2,j}\} \geq \min\{b_{1,j}, b_{2,j}\}$ for some $j \in \{1, \ldots, d\}$.

Now we define μ_0 on $\mathcal{A}$ by $\mu_0((\boldsymbol{a}, \boldsymbol{b}]) := \Delta_{(\boldsymbol{a}, \boldsymbol{b}]} F$ for all $\boldsymbol{a} \leq \boldsymbol{b}$.

Showing the (finite) additivity, "superadditivity", and "subadditivity" of μ_0.

- *Additivity:* Fix any pairwise disjoint $(\boldsymbol{a}_1, \boldsymbol{b}_1], \ldots, (\boldsymbol{a}_n, \boldsymbol{b}_n] \in \mathcal{A}$ with $\biguplus_{i=1}^{n} (\boldsymbol{a}_i, \boldsymbol{b}_i] = (\boldsymbol{a}, \boldsymbol{b}] \in \mathcal{A}$. Then, we have

$$\mu_0 \left(\biguplus_{i=1}^{n} (\boldsymbol{a}_i, \boldsymbol{b}_i] \right) = \mu_0((\boldsymbol{a}, \boldsymbol{b}]) = \Delta_{(\boldsymbol{a}, \boldsymbol{b}]} F \stackrel{\text{(telescoping sums)}}{=} \sum_{i=1}^{n} \Delta_{(\boldsymbol{a}_i, \boldsymbol{b}_i]} F = \sum_{i=1}^{n} \mu_0(\boldsymbol{a}_i, \boldsymbol{b}_i]).$$

- *"Superadditivity":* Fix any pairwise disjoint $(\boldsymbol{a}_1, \boldsymbol{b}_1], \ldots, (\boldsymbol{a}_n, \boldsymbol{b}_n] \in \mathcal{A}$ with $\biguplus_{i=1}^{n} (\boldsymbol{a}_i, \boldsymbol{b}_i] \subseteq (\boldsymbol{a}, \boldsymbol{b}] \in \mathcal{A}$ (not "$= (\boldsymbol{a}, \boldsymbol{b}]$" here!). Note that $(\boldsymbol{a}, \boldsymbol{b}] \backslash \biguplus_{i=1}^{n} (\boldsymbol{a}_i, \boldsymbol{b}_i] \stackrel{\text{(Lemma 2.3.a)}}{=} \biguplus_{j=1}^{m} (\boldsymbol{c}_j, \boldsymbol{d}_j]$ for some pairwise disjoint $(\boldsymbol{c}_1, \boldsymbol{d}_1], \ldots, (\boldsymbol{c}_m, \boldsymbol{d}_m] \in \mathcal{A}$. Hence, we can write $(\boldsymbol{a}, \boldsymbol{b}] = \biguplus_{i=1}^{n} (\boldsymbol{a}_i, \boldsymbol{b}_i] \uplus \biguplus_{j=1}^{m} (\boldsymbol{c}_j, \boldsymbol{d}_j]$. It then follows by the previously shown additivity that

$$\mu_0((\boldsymbol{a}, \boldsymbol{b}]) = \sum_{i=1}^{n} \mu_0((\boldsymbol{a}_i, \boldsymbol{b}_i]) + \underbrace{\sum_{j=1}^{m} \mu_0((\boldsymbol{c}_j, \boldsymbol{d}_j])}_{\geq\, 0 \text{ due to } d\text{-increasing}} \geq \sum_{i=1}^{n} \mu_0((\boldsymbol{a}_i, \boldsymbol{b}_i]),$$

establishing the "superadditivity".

- *"Subadditivity":* Fix any $(\boldsymbol{a}_1, \boldsymbol{b}_1], \ldots, (\boldsymbol{a}_n, \boldsymbol{b}_n] \in \mathcal{A}$ with $(\boldsymbol{a}, \boldsymbol{b}] \subseteq \bigcup_{i=1}^{n} (\boldsymbol{a}_i, \boldsymbol{b}_i]$ (not $= (\boldsymbol{a}, \boldsymbol{b}]$" here!). Letting $(\boldsymbol{c}_i, \boldsymbol{d}_i] = (\boldsymbol{a}_i, \boldsymbol{b}_i] \backslash \bigcup_{k=1}^{i-1} (\boldsymbol{a}_k, \boldsymbol{b}_k]$ for all $i = 1, \ldots, n$, we have $(\boldsymbol{a}, \boldsymbol{b}] \subseteq \bigcup_{i=1}^{n} (\boldsymbol{a}_i, \boldsymbol{b}_i] = \biguplus_{i=1}^{n} (\boldsymbol{c}_i, \boldsymbol{d}_i]$. Also, there exists nonempty $I \subseteq \{1, \ldots, n\}$ such that $(\boldsymbol{a}, \boldsymbol{b}] \subseteq \biguplus_{i \in I} (\boldsymbol{c}_i, \boldsymbol{d}_i] = (\boldsymbol{c}, \boldsymbol{d}]$ where $-\infty < \boldsymbol{c} \leq \boldsymbol{d} < \infty$. Thus, we have

$$\mu_0((\boldsymbol{a}, \boldsymbol{b}]) = \Delta_{(\boldsymbol{a}, \boldsymbol{b}]} F \stackrel{[2.4.3]\text{b}}{\leq} \Delta_{(\boldsymbol{c}, \boldsymbol{d}]} F \stackrel{\text{(telescopic sums)}}{=} \sum_{i \in I} \Delta_{(\boldsymbol{c}_i, \boldsymbol{d}_i]} F$$

$$\leq \sum_{i=1}^{n} \Delta_{(\boldsymbol{c}_i, \boldsymbol{d}_i]} F \stackrel{[2.4.3]\text{b}}{\leq} \sum_{i=1}^{n} \Delta_{(\boldsymbol{a}_i, \boldsymbol{b}_i]} F = \sum_{i=1}^{n} \mu_0((\boldsymbol{a}_i, \boldsymbol{b}_i]),$$

establishing the "subadditivity".

Showing that μ_0 is a σ-finite premeasure on $\mathcal{A}$. We first show that μ_0 is a premeasure on $\mathcal{A}$:

(1) It follows by the definition of d-increasing that $\Delta_{(\boldsymbol{a}, \boldsymbol{b}]} F \geq 0$ for all $\boldsymbol{a} \leq \boldsymbol{b}$. Thus, μ_0 can be considered to be a function from $\mathcal{A}$ to $[0, \infty]$.

(2) We have

$$\mu_0(\varnothing) = \mu_0((\boldsymbol{a}, \boldsymbol{a}]) = \Delta_{(\boldsymbol{a}, \boldsymbol{a}]}F = \sum_{\boldsymbol{i} \in \{0,1\}^d} (-1)^{\sum_{j=1}^d i_j} \underbrace{F(a_1^{i_1} a_1^{1-i_1}, \ldots, a_d^{i_d} a_d^{1-i_d})}_{F(\boldsymbol{a}) \text{ always}}$$

$$= F(\boldsymbol{a}) \sum_{\boldsymbol{i} \in \{0,1\}^d} (-1)^{\sum_{j=1}^d i_j} = 0.$$

(3) Fix any pairwise disjoint $(\boldsymbol{a}_1, \boldsymbol{b}_1], \ldots \in \mathcal{A}$ with $\biguplus_{i=1}^{\infty}(\boldsymbol{a}_i, \boldsymbol{b}_i] = (\boldsymbol{a}, \boldsymbol{b}] \in \mathcal{A}$. Then we show "$\leq$" and "$\geq$" separately:

- $\underline{\sum_{i=1}^{\infty} \mu_0((\boldsymbol{a}_i, \boldsymbol{b}_i]) \leq \mu_0((\boldsymbol{a}, \boldsymbol{b}])}$: For all $n \in \mathbb{N}$, we have $\biguplus_{i=1}^{n}(\boldsymbol{a}_i, \boldsymbol{b}_i] \subseteq \biguplus_{i=1}^{\infty}(\boldsymbol{a}_i, \boldsymbol{b}_i] = (\boldsymbol{a}, \boldsymbol{b}]$. By superadditivity, we then have $\sum_{i=1}^{n} \mu_0((\boldsymbol{a}_i, \boldsymbol{b}_i]) \leq \mu_0((\boldsymbol{a}, \boldsymbol{b}])$. Letting $n \to \infty$ gives $\sum_{i=1}^{\infty} \mu_0((\boldsymbol{a}_i, \boldsymbol{b}_i]) \leq \mu_0((\boldsymbol{a}, \boldsymbol{b}])$.

- $\underline{\mu_0((\boldsymbol{a}, \boldsymbol{b}]) \leq \sum_{i=1}^{\infty} \mu_0((\boldsymbol{a}_i, \boldsymbol{b}_i])}$: Fix any $\varepsilon > 0$. Since F is right-continuous, the composition $\boldsymbol{x} \mapsto \Delta_{(\boldsymbol{x}, \boldsymbol{b}]}F = \mu_0((\boldsymbol{x}, \boldsymbol{b}])$ is also right-continuous. Thus, there exists $\widetilde{\boldsymbol{a}} \in (\boldsymbol{a}, \boldsymbol{b}]$ such that $\mu_0((\boldsymbol{a}, \boldsymbol{b}]) \leq \mu_0((\widetilde{\boldsymbol{a}}, \boldsymbol{b}]) + \varepsilon/2$. Similarly, we know $\boldsymbol{x} \mapsto \mu_0((\boldsymbol{a}, \boldsymbol{x}])$ is right-continuous, thus for all $i \in \mathbb{N}$ there exists $\widetilde{\boldsymbol{b}}_i \in (\boldsymbol{b}_i, \boldsymbol{b}]$ such that $\mu_0((\boldsymbol{a}_i, \widetilde{\boldsymbol{b}}_i]) \leq \mu_0((\boldsymbol{a}_i, \boldsymbol{b}_i]) + \varepsilon/2^{i+1}$.

  Now, note that $[\widetilde{\boldsymbol{a}}, \boldsymbol{b}] \subseteq (\boldsymbol{a}, \boldsymbol{b}] = \biguplus_{i=1}^{\infty}(\boldsymbol{a}_i, \boldsymbol{b}_i] \subseteq \bigcup_{i=1}^{\infty}(\boldsymbol{a}_i, \widetilde{\boldsymbol{b}}_i)$, so $\{(\boldsymbol{a}_i, \widetilde{\boldsymbol{b}}_i)\}_{i \in \mathbb{N}}$ serves as an open cover of $[\widetilde{\boldsymbol{a}}, \boldsymbol{b}]$. Since $[\widetilde{\boldsymbol{a}}, \boldsymbol{b}]$ is compact (by Heine-Borel theorem; it is closed and bounded in $\mathbb{R}^d$), the open cover has a finite subcover. WLOG, we may assume $(\widetilde{\boldsymbol{a}}, \boldsymbol{b}] \subseteq [\widetilde{\boldsymbol{a}}, \boldsymbol{b}] \subseteq \bigcup_{i=1}^{n}(\boldsymbol{a}_i, \widetilde{\boldsymbol{b}}_i)$ for some $n \in \mathbb{N}$. Thus, by subadditivity we have $\mu_0((\widetilde{\boldsymbol{a}}, \boldsymbol{b}]) \leq \sum_{i=1}^{n} \mu_0((\boldsymbol{a}_i, \widetilde{\boldsymbol{b}}_i])$. Therefore,

$$\mu_0((\boldsymbol{a}, \boldsymbol{b}]) \leq \mu_0((\widetilde{\boldsymbol{a}}, \boldsymbol{b}]) + \frac{\varepsilon}{2} \leq \sum_{i=1}^{n} \mu_0((\boldsymbol{a}_i, \widetilde{\boldsymbol{b}}_i]) + \frac{\varepsilon}{2} \leq \sum_{i=1}^{n}\left(\mu_0((\boldsymbol{a}_i, \boldsymbol{b}_i]) + \frac{\varepsilon}{2^{i+1}}\right) + \frac{\varepsilon}{2}.$$

Letting $n \to \infty$, we have

$$\mu_0((\boldsymbol{a}, \boldsymbol{b}]) \leq \sum_{i=1}^{\infty} \mu_0((\boldsymbol{a}_i, \boldsymbol{b}_i]) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \sum_{i=1}^{\infty} \mu_0((\boldsymbol{a}_i, \boldsymbol{b}_i]) + \varepsilon,$$

which implies that $\mu_0((\boldsymbol{a}, \boldsymbol{b}]) \leq \sum_{i=1}^{\infty} \mu_0((\boldsymbol{a}_i, \boldsymbol{b}_i])$ as $\varepsilon > 0$ is arbitrary.

Now, we show the σ-finiteness of μ_0. Let $A_i = (-_d i, _d i] \in \mathcal{A}$ for all $i \in \mathbb{N}$. Then we have $\mathbb{R}^d = \biguplus_{i=1}^{\infty} A_i$ and $\mu_0(A_i) = \Delta_{(-_d i, _d i]}F < \infty$ for all $i \in \mathbb{N}$, so μ_0 is σ-finite.

Completing the proof by Carathéodory extension theorem. By Carathéodory extension theorem, there exists a unique measure λ_F on $\sigma(\mathcal{A}) \overset{\text{(Proposition 1.4.g)}}{=} \mathcal{B}(\mathbb{R}^d)$ such that $\lambda_F((\boldsymbol{a}, \boldsymbol{b}]) = \mu_0((\boldsymbol{a}, \boldsymbol{b}]) = \Delta_{(\boldsymbol{a}, \boldsymbol{b}]}F$ for all $\boldsymbol{a} \leq \boldsymbol{b}$. $\qquad\square$

2.4.5 **Lebesgue-Stieltjes measure.** As mentioned in [2.3.5], we shall assume the measure λ_F to be complete. Such complete measure λ_F is then known as the **Lebesgue-Stieltjes measure** associated to F. Particularly, if $F(\boldsymbol{x}) = \prod_{j=1}^{d} x_j$ for all $\boldsymbol{x} \in \mathbb{R}^d$, then $\lambda := \lambda_F$ is known as the (familiar) **Lebesgue measure** on $\mathbb{R}^d$, with domain being the completion $\bar{\mathcal{B}}(\mathbb{R}^d)$, called **Lebesgue σ-algebra**. Every set in $\bar{\mathcal{B}}(\mathbb{R}^d)$ is said to be **Lebesgue measurable**, or a **Lebesgue set**.

By definition, every Borel set is a Lebesgue set since $\mathcal{B}(\mathbb{R}^d) \subseteq \bar{\mathcal{B}}(\mathbb{R}^d)$. As one may expect, the converse does not hold and there is a Lebesgue set that is not a Borel set. But the construction of such set turns out to be somewhat tricky, and we will only do that in [3.1.4].

2.4.6 **Properties of the Lebesgue measure.**

(a) *(Agreeing with our usual notion of "volume")* The Lebesgue measure λ satisfies $\lambda((\boldsymbol{a}, \boldsymbol{b}]) = \prod_{j=1}^{d}(b_j - a_j)$ for all $\boldsymbol{a} \leq \boldsymbol{b}$. [Note: More generally, the Lebesgue-Stieltjes measure λ_F can be interpreted as assigning hypercubes their volumes that are *distorted by F*.]

(b) The Lebesgue measure λ is invariant with respect to translations, rotations, and reflections.

(c) *(Coinciding with the product measure of Lebesgue measure on $\mathbb{R}$)* The Lebesgue measure λ on $\mathbb{R}^d$ coincides with the product measure of d copies of the Lebesgue measure on $\mathbb{R}$.

Proof.

(a) It follows from [2.4.3]c.

(b) Omitted.

(c) Since $\mathbb{R}^d = \prod_{j=1}^d \mathbb{R}$ is a product space, it can be equipped with the product σ-algebra $\bigotimes_{j=1}^d \mathcal{B}(\mathbb{R}) \overset{\text{(Proposition 1.4.b)}}{=} \sigma(\prod_{j=1}^d B_j : B_j \in \mathcal{B}(\mathbb{R}))$. Then consider the product measure $\mu = \prod_{j=1}^d \mu_j$ on $\bigotimes_{j=1}^d \mathcal{B}(\mathbb{R})$, where each μ_j is the Lebesgue measure on $\mathbb{R}$ (which is σ-finite). By the definition of product measure, we have

$$\mu((\boldsymbol{a}, \boldsymbol{b}]) = \mu\left(\prod_{j=1}^d (a_j, b_j]\right) = \prod_{j=1}^d \mu_j((a_j, b_j]) = \prod_{j=1}^d (b_j - a_j)$$

for all $-\infty < \boldsymbol{a} \leq \boldsymbol{b} < \infty$, meaning that μ and λ coincide on the semiring $\mathcal{A}$. Since $\mathcal{A}$ is a π-system and λ is σ-finite (as hinted from the proof of Theorem 2.4.a), by Proposition 2.1.a, we must have $\lambda = \mu$.

$\square$

2.4.7 **Componentwise increasing does not imply d-increasing.** The construction of Borel measures on $\mathbb{R}^d$ in Theorem 2.4.a requires the underlying function F to be *d-increasing*, and componentwise increasing is *not enough*, as the following example illustrates.

Let $F(x_1, x_2, x_3) = \max\{x_1 + x_2 + x_3 - d + 1, 0\}$ for all $(x_1, x_2, x_3) \in \mathbb{R}^3$. Then F is componentwise increasing but not d-increasing (here $d = 3$), since

$$\Delta_{(_3 1/2, _3 1]} F = \sum_{\boldsymbol{i} \in \{0,1\}^3} (-1)^{i_1 + i_2 + i_3} F\left((1/2)^{i_1} 1^{1-i_1}, (1/2)^{i_2} 1^{1-i_2}, (1/2)^{i_3} 1^{1-i_3}\right)$$

$$= \underbrace{\max\{1+1+1-3+1, 0\}}_{i_1 = i_2 = i_3 = 0} \underbrace{-3 \max\{1+1+1/2-3+1, 0\}}_{\text{exactly one } i_j = 0} + \text{other terms that are } 0$$

$$= 1 - 3/2 < 0.$$

Therefore, we cannot apply Theorem 2.4.a on this function F to induce a Borel measure λ_F.

2.4.8 **Examples of Lebesgue null sets.** Considering the measure space $(\mathbb{R}^d, \bar{\mathcal{B}}(\mathbb{R}^d), \lambda)$, we know from Section 2.2 that every set $N \in \bar{\mathcal{B}}(\mathbb{R}^d)$ with $\lambda(N) = 0$ is a λ-null set, which is also called **Lebesgue null set**. Here we will explore some examples of Lebesgue null sets:

Lebesgue null sets in $\mathbb{R}$

- *(Singleton)* For every $x \in \mathbb{R}$, the singleton $\{x\} \subseteq \mathbb{R}$ is a Lebesgue null set since $\lambda(\{x\}) = \lambda(\bigcap_{n=1}^\infty (x - 1/n, x]) \overset{\text{(continuity from above)}}{\underset{(\lambda((x-1,x]) < \infty)}{=}} = \lim_{n \to \infty} \lambda((x - 1/n, x]) = \lim_{n \to \infty}(x - (x - 1/n)) = 0$.

- *(Countable set)* Every countable subset S of $\mathbb{R}$ is a Lebesgue null set since it can be expressed as $S = \bigcup_{i=1}^\infty \{s_i\}$ where $s_i \in S$ for every $i \in \mathbb{N}$, and so by Lemma 2.2.a we know S is a null set.

- *(Uncountable set)* The **Cantor set** $\mathcal{C}$, defined by $\mathcal{C} := \bigcap_{i=1}^\infty C_i$ with $C_0 = [0, 1]$ and $C_i = \frac{1}{3} C_{i-1} \cup \left(\frac{2}{3} + \frac{C_{i-1}}{3}\right)$[6] for every $i \in \mathbb{N}$, is an uncountable Lebesgue null set.

---

[6] Here the notation $a + bS$ refers to the set $\{a + bs : s \in S\}$.

C_0

C_1

C_2
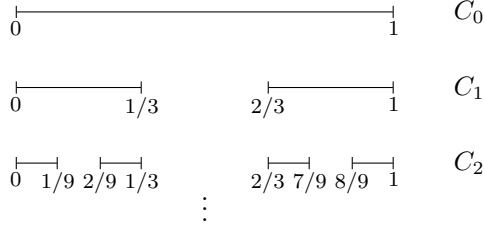
[Note: It can be shown that $\mathcal{C} = \{x \in [0,1] : x = \sum_{i=1}^{\infty} a_i 3^{-i}, a_i \in \{0,2\} \ \forall i \in \mathbb{N}\}$. The latter set contains all numbers in $[0,1]$ whose base-3 expansion only has 0 or 2 as its digits. The rough idea is that, the digits 0 or 2 correspond to the "pieces" remained in the picture above (in each step we are removing the "middle pieces", corresponding to the digit 1).]

Proof.

- *$\mathcal{C}$ is uncountable:* Assume to the contrary that $\mathcal{C}$ is countable and can be expressed as $\mathcal{C} = \{c_i : i \in \mathbb{N}\}$. List the c_i's in the following way:

  * $c_1 = 0.000\cdots$
  * $c_2 = 0.020\cdots$
  * $c_3 = 0.022\cdots$
  * $\vdots$

  (the values here are for illustration only). By changing $0 \to 2$ and $2 \to 0$ in the each of the digits in the "diagonal" above, we can obtain $c = 0.200\cdots$ (this value is for the example above), which is guaranteed to be *different* from every c_i in the list by construction, meaning that $c \notin \mathcal{C}$. However, we can certainly write $c = \sum_{i=1}^{\infty} a_i 3^{-i}$ where $a_i \in \{0,2\}$ for all $i \in \mathbb{N}$, contradiction.

- *$\mathcal{C}$ is a Lebesgue null set:* Note that for every $i \in \mathbb{N}$, the set $\mathcal{C}_i$ is obtained by removing 2^{i-1} intervals of length 3^{-i} each from C_{i-1}. Therefore, the length of all the removed intervals is

$$\lambda([0,1] \setminus \mathcal{C}) = \sum_{i=1}^{\infty} 2^{i-1} 3^{-i} = \frac{1}{2} \sum_{i=1}^{\infty} (2/3)^i = \frac{1}{2}\left(\frac{1}{1-2/3} - 1\right) = 1.$$

Then, since $1 = \lambda([0,1]) = \lambda(([0,1] \setminus \mathcal{C}) \uplus \mathcal{C}) \overset{\text{(finite additivity)}}{=} \lambda([0,1] \setminus \mathcal{C}) + \lambda(\mathcal{C}) = 1 + \lambda(\mathcal{C})$, we have $\lambda(\mathcal{C}) = 0$.

$\square$

Lebesgue null sets in $\mathbb{R}^d$

- A line in $\mathbb{R}^2$ is a Lebesgue null set.

  Proof. WLOG, consider the line $y = 0$, i.e., $\{(x,y) \in \mathbb{R}^2 : y = 0\}$, in $\mathbb{R}^2$.[7] Write $\mathbb{Q} = \{q_i : i \in \mathbb{N}\}$ and fix any $\varepsilon > 0$. Let $A_i = (q_i - \varepsilon/2, q_i + \varepsilon/2] \times (-1/2^{i+1}, 1/2^{i+1}]$ for every $i \in \mathbb{N}$. Then we have $\{(x,y) \in \mathbb{R}^2 : y = 0\} \subseteq \bigcup_{i=1}^{\infty} A_i$, and thus

$$0 \leq \lambda(\{(x,y) \in \mathbb{R}^2 : y = 0\}) \leq \lambda\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i) = \sum_{i=1}^{\infty} \varepsilon \cdot 2^{-i} = \varepsilon.$$

As this holds for every $\varepsilon > 0$, we conclude that $\lambda(\{(x,y) \in \mathbb{R}^2 : y = 0\}) = 0$. $\square$

[Note: Using a similar argument, we can show that any $\mathbb{R}^k$ with $k < d$ or its subset is a Lebesgue null set in $\mathbb{R}^d$, e.g., planes are Lebesgue null sets in $\mathbb{R}^3$.]

---

[7]Note that every line in $\mathbb{R}^2$ can be obtained from this line by translations, rotations, and reflections. As mentioned before, these actions would not change the Lebesgue measure.

2.5   Probability Measures

2.5.1   **Terminologies.** Let $(\Omega, \mathcal{F})$ be a measurable space. A **probability measure** $\mathbb{P}$ on $\mathcal{F}$ (or $(\Omega, \mathcal{F})$) is a function such that:

    (1) *(Nonnegativity)* $\mathbb{P}$ is a function from $\mathcal{F}$ to $[0, \infty]$.

    (2) *(Unitarity)* $\mathbb{P}(\Omega) = 1$.

    (3) *(σ-additivity)* $A_1, A_2, \ldots \in \mathcal{F}$ pairwise disjoint $\implies \mathbb{P}(\biguplus_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**, Ω is called a **sample space**, and each outcome $\omega \in \Omega$ is called a **sample point**. If Ω is a countable (finite resp.) set, then $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be **discrete** (**finite** resp.). Every $A \in \mathcal{F}$ is called an **event**.

2.5.2   **Properties of probability measure.**

    (a) *(Probability measure is a measure)* We have $\mathbb{P}(\varnothing) = 0$, so the probability measure $\mathbb{P}$ is indeed a measure.

        Proof. Let $A_1 = \Omega$ and $A_i = \varnothing$ for every $i = 2, 3, \ldots$. Then we have $\Omega = \biguplus_{i=1}^{\infty} A_i$, thus

$$1 = \mathbb{P}(\Omega) = \mathbb{P}\left(\biguplus_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \mathbb{P}(\Omega) + \sum_{i=2}^{\infty} \mathbb{P}(\varnothing),$$

        which implies that $\sum_{i=2}^{\infty} \mathbb{P}(\varnothing) = 0$, forcing that $\mathbb{P}(\varnothing) = 0$. $\square$

    (b) *(Numerous properties of measure)* Since $\mathbb{P}$ is a finite measure, all properties in [2.1.3] apply.

2.5.3   **Examples of probability measures.**

    (a) *(Combinatorial probability)* Let Ω be a nonempty finite sample space with $\mathcal{F} = \mathcal{P}(\Omega)$. Then we can define $\mathbb{P}(A) := |A|/|\Omega|$ for all $A \in \mathcal{F}$, which is indeed a probability measure on $\mathcal{F}$.

        Proof.

        (1) The nonnegativity follows from the fact that both $|A|$ and $|\Omega|$ are nonnegative.

        (2) We have $\mathbb{P}(\Omega) = |\Omega|/|\Omega| = 1$.

        (3) Fix any pairwise disjoint $A_1, A_2, \ldots \in \mathcal{F}$. Noting that $|\biguplus_{i=1}^{\infty} A_i| = \sum_{i=1}^{\infty} |A_i|$, we have

$$\mathbb{P}\left(\biguplus_{i=1}^{\infty} A_i \right) = \frac{|\biguplus_{i=1}^{\infty} A_i|}{|\Omega|} = \sum_{i=1}^{\infty} \frac{|A_i|}{|\Omega|} = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

                                                                             $\square$

    (b) *(Defined via mass function)* Let Ω be a countable sample space with $\mathcal{F} = \mathcal{P}(\Omega)$. Then any **mass function** $f : \Omega \to [0, 1]$ with $\sum_{\omega \in \Omega} f(\omega) = 1$ induces a **discrete probability measure** $\mathbb{P}$ on $\mathcal{F}$ by $\mathbb{P}(A) = \sum_{\omega \in A} f(\omega)$ for all $A \in \mathcal{F}$.

        Remarks:

        • Conversely, if $\mathbb{P}$ is a probability measure on $\mathcal{F}$, then $f(\omega) := \mathbb{P}(\{\omega\})$ is the corresponding pmf that induces $\mathbb{P}$.

        • The argument for showing such $\mathbb{P}$ is indeed a probability measure is analogous to that in [2.1.2]c.

    (c) *(Geometric probability space)* Let the sample space Ω be a subset of $\mathbb{R}^d$ with $\lambda(\Omega) < \infty$ and $\mathcal{F} = \bar{\mathcal{B}}(\Omega)$. Define $\mathbb{P}(A) = \lambda(A)/\lambda(\Omega)$ for all $A \in \mathcal{F}$ *(relative volume)*. Then $\mathbb{P}$ is a probability measure on $\mathcal{F}$ (which can be proved using a similar argument as in [2.5.3]a). Here, $(\Omega, \mathcal{F}, \mathbb{P})$ is called **geometric probability space**.

2.5.4   **Characterization of probability measures on $\mathbb{R}^d$.** In the case where $\Omega = \mathbb{R}^d$, there is a characterization of the probability measure through the *distribution function*. It is indeed the one you have learnt in your first course in probability, but here we will define it in a way that does not involve the notion of "random variable" (as we have not yet defined it formally!).

A **(multivariate/joint) distribution function** is a function $F : \mathbb{R}^d \to [0, 1]$ such that:

(1) $\lim_{x_j \to \infty} F(\boldsymbol{x}) = 0$ for all $j = 1, \ldots, d$ *(groundedness)* and $\lim_{\boldsymbol{x} \to \boldsymbol{\infty}} F(\boldsymbol{x}) = 1$ *(normalization)*.

(2) F is d-increasing.

(3) F is right-continuous.

The following result then tells us how distribution function characterizes probability measures on $\mathbb{R}^d$.

Theorem 2.5.a. Let $\Omega = \mathbb{R}^d$. Then:

(a) $\mathbb{P}$ *induces F:* If $\mathbb{P}$ is a probability measure on $\mathcal{B}(\mathbb{R}^d)$, then $F(\boldsymbol{x}) := \mathbb{P}((-\infty, \boldsymbol{x}])$ is a distribution function. Also, we have $\lambda_F = \mathbb{P}$ on $\mathcal{B}(\mathbb{R}^d)$ for this F.

(b) F *induces $\mathbb{P}$:* If F is a distribution function, then $\mathbb{P} := \lambda_F$ (with domain restricted to $\mathcal{B}(\mathbb{R}^d)$) is a probability measure on $\mathcal{B}(\mathbb{R}^d)$. Also, we have $\mathbb{P}((-\infty, \boldsymbol{x}]) = F(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$ for this $\mathbb{P}$.

$$\mathbb{P} \xrightarrow{\;F(\boldsymbol{x}) := \mathbb{P}((-\infty, \boldsymbol{x}])\;} \underset{(\lambda_F = \mathbb{P})}{F} \xrightarrow{\;\mathbb{P} := \lambda_F\;} \mathbb{P}$$

$$F \xrightarrow{\;\mathbb{P} := \lambda_F\;} \underset{(\mathbb{P}((-\infty, \boldsymbol{x}]) = F(\boldsymbol{x}))}{\mathbb{P}} \xrightarrow{\;F(\boldsymbol{x}) := \mathbb{P}((-\infty, \boldsymbol{x}])\;} F$$

Remarks:

- As the picture above illustrates, the two ways (mappings) of getting from $\mathbb{P}$ to F and F to $\mathbb{P}$ are inverse to each other (applying one after another would yield an identity mapping), thus forming a one-to-one correspondence between the probability measures and distribution functions.

- The distribution function $F(\boldsymbol{x}) := \mathbb{P}((-\infty, \boldsymbol{x}])$ is called the **distribution function of $\mathbb{P}$ (or λ_F)**.

- For every probability measure $\mathbb{P}$ on $\mathcal{B}(\mathbb{R}^d)$ with distribution function F, we have $\mathbb{P}((\boldsymbol{a}, \boldsymbol{b}]) = \lambda_F((\boldsymbol{a}, \boldsymbol{b}]) = \Delta_{(\boldsymbol{a}, \boldsymbol{b}]} F$. This gives rise to another interpretation of the F-volume $\Delta_{(\boldsymbol{a}, \boldsymbol{b}]} F$, namely that it is the probability of $(\boldsymbol{a}, \boldsymbol{b}]$ under the distribution function F.

- While the domain of a d-dimensional distribution function F is always $\mathbb{R}^d$ by definition, sometimes a function F that satisfies all the three properties mentioned but with domain being a subset of $\mathbb{R}^d$ is also called a distribution function; we shall adopt this convention throughout. To adapt with the definition above, we can view such function F as implicitly extended to a function $\widetilde{F}$ with domain $\mathbb{R}^d$, given by

$$\widetilde{F}(\boldsymbol{x}) = F(\min\{\max\{x_1, a_1\}, b_1\}, \ldots, \min\{\max\{x_d, a_d\}, b_d\}), \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^d,$$

where $(a_1, \ldots, a_d) := \inf \operatorname{supp}(F)$ and $(b_1, \ldots, b_d) := \sup \operatorname{supp}(F)$. In words, the extension is done by setting 1 in the "upper-right" region and 0 in all other regions outside $\operatorname{supp}(F)$.

Proof.

(a) We first show that $F(\boldsymbol{x}) := \mathbb{P}((-\infty, \boldsymbol{x}])$ is a distribution function.

(1) Note that $\lim_{x_j \to \infty} F(\boldsymbol{x}) = \lim_{n \to \infty} F(\boldsymbol{x}_{j \leftarrow -n} = \lim_{n \to \infty} \mathbb{P}((-\infty, \boldsymbol{x}_{j \leftarrow -n})$ for all $j = 1, \ldots, d$, and $F(\boldsymbol{\infty}) = \lim_{n \to \infty} \mathbb{P}((-\infty, {}_d n]) \overset{\text{(continuity from below)}}{=} \mathbb{P}(\bigcup_{n=1}^{\infty} (-\infty, {}_d n]) = \mathbb{P}(\mathbb{R}^d) = \mathbb{P}(\Omega) = 1$.

(2) Note first that

$$\begin{aligned} \Delta_{(a_1, b_1]} F &= F(b_1, x_2, \ldots, x_d) - F(a_1, x_2, \ldots, x_d) \\ &= \mathbb{P}((-\infty, (b_1, x_2, \ldots, x_d)]) - \mathbb{P}((-\infty, (a_1, x_2, \ldots, x_d)]) \\ &\overset{\text{(subtractivity)}}{=} \mathbb{P}((-\infty, (b_1, x_2, \ldots, x_d)] \setminus (-\infty, (a_1, x_2, \ldots, x_d)]) \\ &= \mathbb{P}\left(\left(\begin{bmatrix} a_1 \\ -\infty \end{bmatrix}, \begin{bmatrix} b_1 \\ \boldsymbol{x}_{\{1\}^c} \end{bmatrix}\right]\right). \end{aligned}$$

Hence, for all $\boldsymbol{a} \le \boldsymbol{b}$, applying an argument similar to above repetitively gives

$$
\begin{aligned}
\Delta_{(\boldsymbol{a},\boldsymbol{b}]}F &= \Delta_{(a_d,b_d]}\cdots\Delta_{(a_2,b_2]}\Delta_{(a_1,b_1]}F \\
&= \Delta_{(a_d,b_d]}\cdots\Delta_{(a_2,b_2]}\mathbb{P}\left(\left(\begin{bmatrix} a_1 \\ -\infty \end{bmatrix}, \begin{bmatrix} b_1 \\ \boldsymbol{x}_{\{1\}^c} \end{bmatrix}\right]\right) \\
&= \cdots = \mathbb{P}((\boldsymbol{a},\boldsymbol{b}]) \ge 0,
\end{aligned}
$$

establishing the d-increasingness.

(3) For all $\boldsymbol{x} \le \boldsymbol{y}$, we have $F(\boldsymbol{x}) = \mathbb{P}((-\infty,\boldsymbol{x}]) \overset{\text{(monotonicity)}}{\le} \mathbb{P}((-\infty,\boldsymbol{y}]) = F(\boldsymbol{y})$. Thus,

$$
\lim_{\boldsymbol{h}\to\boldsymbol{0}^+} F(\boldsymbol{x}+\boldsymbol{h}) = \lim_{n\to\infty} F(\boldsymbol{x}+\boldsymbol{1}/n) \le \lim_{\min_j\{n_j\}\to\infty} F\left(\boldsymbol{x}+{}_d\left(\frac{1}{\min_j\{n_j\}}\right)\right)
$$

$$
\overset{\text{(relabel } m \equiv \min_j\{n_j\})}{=} \lim_{m\to\infty} F(\boldsymbol{x}+{}_d(1/m)) = \lim_{m\to\infty} \mathbb{P}((-\infty,\boldsymbol{x}+{}_d(1/m)])
$$

$$
\overset{\text{(continuity from above)}}{=} \mathbb{P}\left(\bigcap_{m=1}^{\infty}(-\infty,\boldsymbol{x}+{}_d(1/m)]\right) = \mathbb{P}((-\infty,\boldsymbol{x}]) = F(\boldsymbol{x}),
$$

for all $\boldsymbol{x} \in \mathbb{R}^d$. Changing $\min_j \to \max_j$ above gives $\lim_{\boldsymbol{h}\to\boldsymbol{0}^+} F(\boldsymbol{x}+\boldsymbol{h}) \ge F(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Hence, together with the inequality above, we get $\lim_{\boldsymbol{h}\to\boldsymbol{0}^+} F(\boldsymbol{x}+\boldsymbol{h}) = F(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

Next, we show that $\lambda_F = \mathbb{P}$ on $\mathcal{B}(\mathbb{R}^d)$. By Proposition 2.1.a, it suffices to show that $\lambda_F = \mathbb{P}$ on the π-system $\mathcal{A} = \{(-\infty,\boldsymbol{b}] : \boldsymbol{b} \in \mathbb{R}^d\}$ since $\sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R}^d)$ by Proposition 1.4.g. This follows from the observation that

$$
\lambda_F((-\infty,\boldsymbol{b}]) = \Delta_{(-\infty,\boldsymbol{b}]}F \overset{\substack{[2.4.3]\text{d} \\ (F \text{ grounded})}}{=} F(\boldsymbol{b}) = \mathbb{P}((-\infty,\boldsymbol{b}])
$$

for all $\boldsymbol{b} \in \mathbb{R}^d$.

(b) By Theorem 2.4.a, we know there is a unique Borel measure λ_F such that $\lambda_F(\boldsymbol{a},\boldsymbol{b}] = \Delta_{(\boldsymbol{a},\boldsymbol{b}]}F$ for all $\boldsymbol{a} \le \boldsymbol{b}$. We now show that λ_F (with domain restricted to $\mathcal{B}(\mathbb{R}^d)$) is a probability measure on $\mathcal{B}(\mathbb{R}^d)$.

(1) The nonnegativity of λ_F follows from the fact that λ_F is a measure.

(2) Note that

$$
\lambda_F(\mathbb{R}^d) \overset{\text{(continuity from below)}}{=} \lim_{n\to\infty} \lambda_F((-{}_dn,{}_dn]) = \lim_{n\to\infty} \Delta_{(-{}_dn,{}_dn]}F
$$

$$
= \lim_{n\to\infty} \sum_{\boldsymbol{i}\in\{0,1\}^d} (-1)^{\sum_{j=1}^d i_j} F((-1)^{i_1}n,\ldots,(-1)^{i_d}n)
$$

$$
= \sum_{\boldsymbol{i}\in\{0,1\}^d} (-1)^{\sum_{j=1}^d i_j} \lim_{n\to\infty} F((-1)^{i_1}n,\ldots,(-1)^{i_d}n).
$$

Since

$$
\lim_{n\to\infty} F((-1)^{i_1}n,\ldots,(-1)^{i_d}n) = \begin{cases} 0 & \text{if } i_j = 1 \text{ for some } j \text{ by groundedness,} \\ F(\boldsymbol{\infty}) = 1 & \text{if } \boldsymbol{i} = \boldsymbol{0}, \end{cases}
$$

we conclude that $\lambda_F(\mathbb{R}^d) = 1$.

(3) The σ-additivity follows from the fact that λ_F is a measure.

Next, the result $\mathbb{P}((-\infty,\boldsymbol{x}]) = F(\boldsymbol{x})$ follows by noting that

$$
\mathbb{P}((-\infty,\boldsymbol{x}]) = \lambda_F((-\infty,\boldsymbol{x}]) \overset{\text{(continuity from below)}}{=} \lim_{n\to\infty} \lambda_F((-{}_dn,\boldsymbol{x}])
$$

$$
= \lim_{n\to\infty} \Delta_{(-{}_dn,\boldsymbol{x}]}F = \Delta_{(-\infty,\boldsymbol{x}]}F \overset{[2.4.3]\text{d}}{=} F(\boldsymbol{x})
$$

for all $\boldsymbol{x} \in \mathbb{R}^d$.

$\square$

2.5.5 Discrete, absolutely continuous, and continuous singular distribution functions. In your first probability course, you should have learnt the concepts of *discrete* and *(absolutely) continuous* random variables. Here, we will explore similar ideas for distribution functions, *without* involving the notions of random variables. Also, we will introduce the concept of *continuous singular distribution function* which, together with the discrete and absolutely continuous distribution functions, can represent *any* distribution function through a mixture.

Let F be a distribution function. The **support** of F is given by $\operatorname{supp}(F) := \{\boldsymbol{x} \in \mathbb{R}^d : \Delta_{(\boldsymbol{x}-\boldsymbol{h},\boldsymbol{x}]}F > 0 \text{ for all } \boldsymbol{h} > \boldsymbol{0}\}$, which is the set of all inputs that get positive probability assigned in the "neighbouring" region (recall that we can view $\Delta_{(\boldsymbol{x}-\boldsymbol{h},\boldsymbol{x}]}F$ as the probability $\mathbb{P}((\boldsymbol{x}-\boldsymbol{h},\boldsymbol{x}])$).

While the domain of the distribution function F is $\mathbb{R}^d$, often we are only interested in its behaviour in the support $\operatorname{supp}(F)$; the values taken outside $\operatorname{supp}(F)$ should always be 1 in the "upper-right" part (corresponding to the *normalization* condition), and 0 in the other parts (corresponding to the *groundedness* condition). Functions defined on subsets of $\mathbb{R}^d$ may then be treated as "distribution function" with the understanding that they are to be extended in this way to make their domains being $\mathbb{R}^d$.

Now, we are ready to define the three kinds of distribution functions:

- A distribution function F is **discrete** if $\operatorname{supp}(F)$ is countable, and its **mass function** is given by $f(\boldsymbol{x}) = \mathbb{P}(\{\boldsymbol{x}\}) = \Delta_{(\boldsymbol{x}-,\boldsymbol{x}]}F := \lim_{\boldsymbol{a}\to\boldsymbol{x}^-} \Delta_{(\boldsymbol{a},\boldsymbol{x}]}F$.
  [Note: We can show $\mathbb{P}(\{\boldsymbol{x}\}) = \Delta_{(\boldsymbol{x}-,\boldsymbol{x}]}F$ by:

  $$\mathbb{P}(\{\boldsymbol{x}\}) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} (\boldsymbol{x} - {}_d(1/n), \boldsymbol{x}]\right) \overset{\text{(continuity from above)}}{=} \lim_{n\to\infty} \mathbb{P}((\boldsymbol{x} - {}_d(1/n), \boldsymbol{x}])$$

  $$= \lim_{n\to\infty} \lambda_F((\boldsymbol{x} - {}_d(1/n), \boldsymbol{x}]) = \lim_{n\to\infty} \Delta_{(\boldsymbol{x} - {}_d(1/n), \boldsymbol{x}]}F = \Delta_{(\boldsymbol{x}-,\boldsymbol{x}]}F.$$

  ]

- A distribution function F is **absolutely continuous** if we can write

  $$F(\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}} f(\boldsymbol{t}) \, d\boldsymbol{t}$$

  for all $\boldsymbol{x} \in \mathbb{R}^d$, where $f : \mathbb{R}^d \to [0, \infty)$ is an *integrable* function with $\int_{-\infty}^{\infty} f(\boldsymbol{t}) \, d\boldsymbol{t} = 1$ (the concept of integrability will be discussed in Section 5). Such f is called the **density function** of F. If $\frac{\partial}{\partial \boldsymbol{x}}F(\boldsymbol{x})$ exists almost everywhere (with respect to the Lebesgue measure), then such f can be obtained by $f(\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}}F(\boldsymbol{x})$ almost everywhere (and the values taken by f outside the "almost everywhere" case would not affect the integral).

- A distribution function F is **continuous singular** if F is a continuous function with $\frac{\partial}{\partial \boldsymbol{x}}F(\boldsymbol{x}) = 0$ almost everywhere.

2.5.6 Formulas for distribution functions.

(a) *(Absolutely continuous distribution function)* If F is absolutely continuous with density f, then $\mathbb{P}((\boldsymbol{a}, \boldsymbol{b}]) = \int_{(\boldsymbol{a},\boldsymbol{b}]} f(\boldsymbol{t}) \, d\boldsymbol{t}$ (this should be somewhat familiar to you already ☺).
  Proof. We have

  $$\mathbb{P}((\boldsymbol{a}, \boldsymbol{b}]) = \Delta_{(\boldsymbol{a},\boldsymbol{b}]}F = \Delta_{(a_1,b_1]} \cdots \Delta_{(a_d,b_d]} \int_{-\infty}^{x_d} \cdots \int_{-\infty}^{x_1} f(t_1, \ldots, t_d) \, dt_1 \cdots dt_d$$

  $$= \int_{a_d}^{b_d} \cdots \int_{a_1}^{b_1} f(t_1, \ldots, t_d) \, dt_1 \cdots dt_d = \int_{(\boldsymbol{a},\boldsymbol{b}]} f(\boldsymbol{t}) \, d\boldsymbol{t}.$$

  $\square$

(b) *(Mixture representation)* Every distribution function F admits the following mixture representation

$$F(\boldsymbol{x}) = p_{\mathrm{d}} F_{\mathrm{d}}(\boldsymbol{x}) + p_{\mathrm{ac}} F_{\mathrm{ac}}(\boldsymbol{x}) + p_{\mathrm{cs}} F_{\mathrm{cs}}(\boldsymbol{x})$$

where $p_{\mathrm{d}}, p_{\mathrm{ac}}, p_{\mathrm{cs}} \geq 0$ with $p_{\mathrm{d}} + p_{\mathrm{ac}} + p_{\mathrm{cs}} = 1$, and $F_{\mathrm{d}}, F_{\mathrm{ac}}, F_{\mathrm{cs}}$ are discrete, absolutely continuous, and continuous singular distribution functions respectively. [Note: Distribution function with at least two of $p_{\mathrm{d}}, p_{\mathrm{ac}}, p_{\mathrm{cs}}$ being positive is said be of **mixed type**.]

Proof. Omitted. $\qquad\square$

(c) *(Margins of distribution functions)* For every $J \subseteq \{1, \ldots, d\}$, the *$J$-margin* of F is given by $F_J(\boldsymbol{x}_J) := \lim_{\boldsymbol{x}_{J^c} \to \boldsymbol{\infty}} F(\boldsymbol{x})$. In the special case where $J = \{j\}$ for some $j = 1, \ldots, d$, it is further called the *jth margin of F*, which is $F_j(x_j) = \lim_{\boldsymbol{x}_{j^c} \to \boldsymbol{\infty}} F(\boldsymbol{x})$. Applying a similar idea to events, for all $A = \prod_{j=1}^{d} A_j \in \bar{\mathcal{B}}(\mathbb{R}^d)$, we write $\mathbb{P}(A_J) = \mathbb{P}\left(\prod_{j \in J} A_j\right) := \mathbb{P}\left(\prod_{j=1}^{d} B_j\right)$ where $B_j = A_j$ for all $j \in J$ and $B_j = \Omega$ for all $j \notin J$.

Then, we have:

i. $F_J(\boldsymbol{x}_J) = \mathbb{P}((-\boldsymbol{\infty}, \boldsymbol{x}_J])$ for every $J \subseteq \{1, \ldots, d\}$.

ii. $F_j(x_j) = \mathbb{P}((-\infty, x_j])$ for every $j = 1, \ldots, d$.

iii. $\mathbb{P}((a_j, b_j]) = F_j(b_j) - F_j(a_j)$ for every $j = 1, \ldots, d$.

Here F is supposed to be the distribution function of $\mathbb{P}$.

Proof.

i. We have $F_J(\boldsymbol{x}_J) = \lim_{\boldsymbol{x}_{J^c} \to \boldsymbol{\infty}} F(\boldsymbol{x}) \overset{\text{(Theorem 2.5.a)}}{=} \lim_{\boldsymbol{x}_{J^c} \to \boldsymbol{\infty}} \mathbb{P}((-\boldsymbol{\infty}, \boldsymbol{x}]) \overset{\text{(continuity from below)}}{=} \mathbb{P}((-\boldsymbol{\infty}, \boldsymbol{x}_J])$.

ii. Take $J = \{j\}$ from above.

iii. We have $\mathbb{P}((a_j, b_j]) = \mathbb{P}((-\infty, b_j] \setminus (-\infty, a_j]) = \mathbb{P}((-\infty, b_j]) - \mathbb{P}((-\infty, a_j]) = F_j(b_j) - F_j(a_j)$. $\qquad\square$

2.5.7 Continuity of absolutely continuous distribution functions. As one would expect, absolutely <u>continuous</u> distribution functions are indeed continuous. We need the following lemma to prove this property.

Lemma 2.5.b (Lipschitz inequality). If F is d-increasing and grounded, then

$$|F(\boldsymbol{b}) - F(\boldsymbol{a})| \leq \sum_{j=1}^{d} |F_j(b_j) - F_j(a_j)|$$

for all $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$.

Proof. Since $F(\boldsymbol{b}) - F(\boldsymbol{a}) = \sum_{j=1}^{d} [F(b_1, \ldots, b_{j-1}, b_j, a_{j+1}, \ldots, a_d) - F(b_1, \ldots, b_{j-1}, a_j, a_{j+1}, \ldots, a_d)]$, by triangle inequality we have

$$|F(\boldsymbol{b}) - F(\boldsymbol{a})| \leq \sum_{j=1}^{d} |F(b_1, \ldots, b_{j-1}, b_j, a_{j+1}, \ldots, a_d) - F(b_1, \ldots, b_{j-1}, a_j, a_{j+1}, \ldots, a_d)|.$$

Fix any $j \in \{1, \ldots, d\}$. WLOG, suppose $a_j \leq b_j$. Then for the jth summand we have

$$|F(b_1, \ldots, b_{j-1}, b_j, a_{j+1}, \ldots, a_d) - F(b_1, \ldots, b_{j-1}, a_j, a_{j+1}, \ldots, a_d)|$$

$$\overset{[2.4.3]\mathrm{e}}{=} F(b_1, \ldots, b_{j-1}, b_j, a_{j+1}, \ldots, a_d) - F(b_1, \ldots, b_{j-1}, a_j, a_{j+1}, \ldots, a_d)$$

$$\overset{[2.4.3]\mathrm{e}}{\leq} F(\infty, b_2, \ldots, b_{j-1}, b_j, a_{j+1}, \ldots, a_d) - F(\infty, b_2, \ldots, b_{j-1}, a_j, a_{j+1}, \ldots, a_d)$$

$$\overset{[2.4.3]\mathrm{e}}{\leq} F(\infty, \infty, \ldots, b_{j-1}, b_j, a_{j+1}, \ldots, a_d) - F(\infty, \infty, \ldots, b_{j-1}, a_j, a_{j+1}, \ldots, a_d)$$

$$\overset{[2.4.3]\mathrm{e}}{\leq} \cdots \leq F_j(b_j) - F_j(a_j)$$

$$\overset{[2.4.3]\mathrm{e}}{=} |F_j(b_j) - F_j(a_j)|.$$

Hence the result follows. □

With Lemma 2.5.b, we can obtain a sufficient condition for the continuity of distribution function.

Proposition 2.5.c. If the margins $F_1, \ldots, F_d$ of F are all continuous, then so is F.

[Note: The condition is particularly satisfied by absolutely continuous distribution functions, whose margins are given by $F_j(x_j) = \mathbb{P}((-\infty, x_j]) = \int_{-\infty}^{x_j} f(t) \, dt$ for all $j = 1, \ldots, d$, which are all continuous by the fundamental theorem of calculus.]

Proof. Assume the margins $F_1, \ldots, F_d$ of F are all continuous. Then consider:

$$\lim_{\boldsymbol{h} \to \boldsymbol{0}} |F(\boldsymbol{x}) - F(\boldsymbol{x} - \boldsymbol{h})| \overset{\text{(Lemma 2.5.b)}}{\leq} \lim_{\boldsymbol{h} \to \boldsymbol{0}} \sum_{j=1}^{d} |F_j(x_j) - F_j(x_j - h_j)| \overset{(F_j\text{'s continuous})}{=} 0,$$

which implies that $\lim_{\boldsymbol{h} \to \boldsymbol{0}} F(\boldsymbol{x} - \boldsymbol{h}) = F(\boldsymbol{x})$, and hence F is continuous. □

2.5.8 **An example of continuous distribution function that is not absolutely continuous.** While we know from [2.5.7] that absolutely continuous distribution functions are always continuous, the converse does not hold. A standard counterexample is the *Cantor distribution function*, a.k.a. the *devil's staircase*. (See https://en.wikipedia.org/wiki/Cantor_function for a picture of its graph, which indeed looks like a staircase.)

Recall the Cantor set $\mathcal{C} = \{x \in [0,1] : x = \sum_{i=1}^{\infty} a_i 3^{-i}, a_i \in \{0,2\} \; \forall i \in \mathbb{N}\}$. Then the **Cantor distribution function** F is defined by

- $F(x) := \sum_{i=1}^{\infty} \frac{a_i}{2} 2^{-i}$ for every $x = \sum_{i=1}^{\infty} a_i 3^{-i} \in \mathcal{C}$, which gives the base-2 expansions of all the numbers in $[0,1]$ (hence $F(\mathcal{C}) = [0,1]$), and

- $F(x) = \sup_{y \leq x, y \in \mathcal{C}} F(y)$ for all $x \in \mathcal{C}^c$.

The Cantor distribution function F can be shown to be continuous, but F is *not* absolutely continuous, since if a density f existed, then we would have $f(t) = 0$ for all $t \in \mathcal{C}^c$ and so $\int_{[0,1]} f(t) \, dt = \int_{\mathcal{C}} f(t) \, dt + \int_{\mathcal{C}^c} f(t) \, dt = 0 + 0 = 0 \neq 1$, as the integral over a Lebesgue null set is always zero (see Section 5). [Note: The Cantor distribution function is also an example of *continuous singular* distribution function.]

2.5.9 **Quantile functions.** Apart from distribution function, the probability measure can also be characterized by a *quantile function*, whose graph is obtained by reflecting the graph of distribution function along the line $y = x$, with flat parts and jumps interchanged:
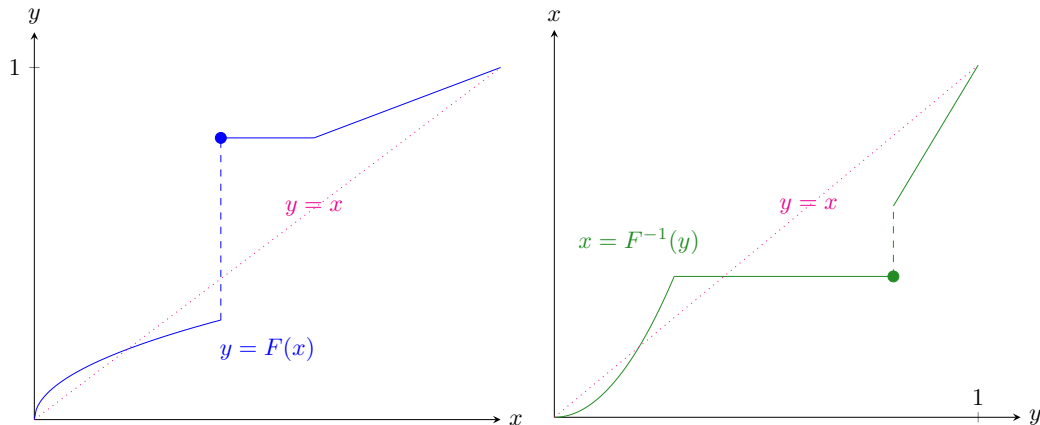
More formally, the notion of *generalized inverse* is used in the definition of quantile function. Let $F : \mathbb{R} \to \mathbb{R}$ be an increasing function. Then the **generalized inverse** of F is a function $F^{-1} : \mathbb{R} \to \bar{\mathbb{R}}$ defined by

$$F^{-1}(y) := \inf\{x \in \mathbb{R} : F(x) \geq y\} \quad \text{for all } y \in \mathbb{R},$$

where we have inf $\varnothing := \infty$ conventionally. If F is a distribution function (with codomain being $[0,1]$ instead of $\mathbb{R}$), the generalized inverse is defined in the same way except that its domain becomes $[0,1]$. In such case, the generalized inverse $F^{-1} : [0,1] \to \bar{\mathbb{R}}$ is called the **quantile function** of F.

2.5.10 **Properties of generalized inverses.** While the inverse function notation F^{-1} is used to denote generalized inverse, there are indeed some subtle differences between the generalized inverse and the ordinary inverse. Notably, the ordinary inverse only exists if the underlying function is bijective, but the generalized inverse *always* exists, even if the underlying function has jumps and/or flat parts. It is thus instructive to study some properties of generalized inverse. In the following, we will just assume that F is increasing unless otherwise specified. Also, we write $F(-\infty) := \lim_{x \to -\infty} F(x)$ and $F(\infty) := \lim_{x \to \infty} F(x)$.

(a) *(Behaviour of F^{-1})*

    i. *(Increasing)* F^{-1} is increasing.

    ii. *(Left-continuous)* If $-\infty < F^{-1}(y) < \infty$, then F^{-1} is left-continuous at y.

(b) *(Sufficient condition for coinciding with ordinary inverse)* If the underlying function F is *strictly* increasing and continuous, then the generalized inverse F^{-1} coincides with the ordinary inverse of F on ran(F).

(c) *(Behaviour of $F^{-1}(F(\cdot))$)*

    i. *("$\leq x$")* Generally, we have $F^{-1}(F(x)) \leq x$.

    ii. *("$= x$")* If F is *strictly* increasing, then we have $F^{-1}(F(x)) = x$.

(d) *(Behaviour of $F(F^{-1}(\cdot))$)* Suppose that F is right-continuous.

    i. *("$\geq y$")* If $F^{-1}(y) < \infty$, then $F(F^{-1}(y)) \geq y$.

    ii. *("$= y$")* If $y \in \text{ran}(F) \cup \{\inf \text{ran}(F), \sup \text{ran}(F)\}$, then $F(F^{-1}(y)) = y$.

    iii. *("$> y$")* If $y < \inf \text{ran}(F)$, then $F(F^{-1}(y)) > y$.

    iv. *("$< y$")* If $y > \sup \text{ran}(F)$, then $F(F^{-1}(y)) < y$.

(e) *(Inequalities about generalized inverse)*

    i. *("Applying" F^{-1} preserves "$\geq$")* If $F(x) \geq y$, then $x \geq F^{-1}(y)$.

    ii. *("Iff" for "$\geq$" under right-continuity)* Suppose that F is right-continuous. Then, $F(x) \geq y$ iff $x \geq F^{-1}(y)$.

    iii. *("Applying" F^{-1} turns "$<$" to "$\leq$")* If $F(x) < y$, then $x \leq F^{-1}(y)$.

(f) *(Relationship between continuity and strict increasingness)*

    i. F is continuous iff F^{-1} is strictly increasing on $[\inf \text{ran}(F), \sup \text{ran}(F)]$.

    ii. F is strictly increasing iff F^{-1} is continuous on ran(F).

Proof. Omitted; see Embrechts and Hofert (2013). $\qquad\qquad\square$

3   Measure Theory III: Measurable Functions

3.0.1 Apart from the *measure* itself, another important notion in measure theory is *measurable function*, which maps between different measurable spaces, providing "connections" between them. A specific kind of measurable function of interest is the *random variable* (which should be familiar to you 😊, but perhaps you have not seen its *formal* definition), which quantifies sample points in the sample spaces, or more abstractly, maps from the measurable space $(\Omega, \mathcal{F})$ (for sample points) to the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (for real numbers quantifying sample points).

3.1   Measurable Functions

3.1.1 **Definition of measurable functions and random variables/vectors/sequences.** Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be measurable spaces. Then a function $X : \Omega \to \Omega'$ is $(\mathcal{F}, \mathcal{F}')$-**measurable** (or just **measurable**) if $X^{-1}(\mathcal{F}') \subseteq \mathcal{F}$, i.e., $X^{-1}(A') \in \mathcal{F}$ for all $A' \in \mathcal{F}'$.

[Intuition 💡: The $(\mathcal{F}, \mathcal{F}')$-measurability suggests the "compatibility" of X with both $\mathcal{F}$ and $\mathcal{F}'$: Picking any set $A' \in \mathcal{F}'$, we can assign measure to the preimage $X^{-1}(A')$ as it is in $\mathcal{F}$. This property is analogous to the *continuity* of a function: Let $(\Omega, \mathcal{T})$ and $(\Omega', \mathcal{T}')$ be topological spaces. Then $X : \Omega \to \Omega'$ is **continuous** if $X^{-1}(\mathcal{T}') \subseteq \mathcal{T}$, or in words, the preimage of every open subset of Ω' under X is open in Ω.]

Special cases:

- If $(\Omega', \mathcal{F}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then every $(\mathcal{F}, \mathcal{F}')$-measurable function X is called a **random variable**.
- If $(\Omega', \mathcal{F}') = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ for some $d \in \mathbb{N}$, then every $(\mathcal{F}, \mathcal{F}')$-measurable function X is called a **random vector**.
- If $(\Omega', \mathcal{F}') = (\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}))$, then every $(\mathcal{F}, \mathcal{F}')$-measurable function X is called a **random sequence**. [Note: $\mathbb{R}^{\mathbb{N}}$ denotes the set of all functions from $\mathbb{N}$ to $\mathbb{R}$, i.e., all real-valued sequences. Sometimes we write it as $\mathbb{R}^{\infty}$, and in the following, the notation "$\mathbb{R}^d$" would also include this case ($d = \infty$), unless otherwise specified.]

Remarks:

- Every $(\mathcal{F}, \mathcal{B}(\mathbb{R}^d))$-measurable function X is also said to be $\mathcal{F}$-**measurable**, denoted by $X \in \mathcal{F}$ (this does not mean that X is an element in $\mathcal{F}$ ⚠).
- $(\mathcal{B}(\Omega), \mathcal{B}(\mathbb{R}^d))$-measurable and $(\bar{\mathcal{B}}(\Omega), \mathcal{B}(\mathbb{R}^d))$-measurable functions are said to be **Borel measurable** and **Lebesgue measurable** respectively.
- *(Completeness of $\mathcal{F}$)* To enrich the collection of measurable functions, the $\mathcal{F}$ chosen is usually the completion of a σ-algebra (i.e., $\mathcal{F} = \bar{\mathcal{F}}$; see Theorem 2.2.b), which comprises of more sets after the completion, so that it contains $X^{-1}(\mathcal{F}')$ for more functions X, thereby making more functions measurable.

  On the other hand, the $\mathcal{F}'$ chosen is often not a completion. For example, in the definition of random variable/vector/sequence, we set $\mathcal{F}'$ to be the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$ rather than the Lebesgue σ-algebra $\bar{\mathcal{B}}(\mathbb{R}^d)$ (the completion of $\mathcal{B}(\mathbb{R}^d)$). The reason is that by including fewer sets in $\mathcal{F}'$, there are less opportunities for the preimage $X^{-1}(A')$ to fall outside $\mathcal{F}$, which again allows more functions to be measurable.

3.1.2 **Examples of non-measurable and measurable functions.** To understand the concept of measurable function better, we study some examples of non-measurable and measurable functions. As you will see, the choice of $\mathcal{F}$ is quite important for the measurability.

Examples of non-measurable functions:

(1) Let Ω be a nonempty set and $\mathcal{F} = \{\varnothing, \Omega\}$ be the trivial σ-algebra on Ω. Then every non-constant function $X : \Omega \to \mathbb{R}$ is not $\mathcal{F}$-measurable.

Proof. Fix any $x \in \mathrm{ran}(X) = \{X(\omega) : \omega \in \Omega\}$. Note that $\{x\} = \bigcap_{n=1}^{\infty} (x - 1/n, x] \in \mathcal{B}(\mathbb{R})$, but $X^{-1}(\{x\})$ is neither $\varnothing$ (as $x \in \mathrm{ran}(X)$) nor Ω (as X is non-constant, so $\mathrm{ran}(X)$ contains values other than x). Thus $X^{-1}(\{x\}) \notin \mathcal{F}$. □

(2) Let $(\Omega, \mathcal{F})$ be a measurable space, and V be a non-measurable subset of Ω (i.e., $V \subseteq \Omega$ but $V \notin \mathcal{F}$)), e.g., the Vitali set for the case $(\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. [Note: By Theorem 1.2.a, it is not possible to define the Lebesgue measure on the Vitali set V, which implies that $V \notin \bar{\mathcal{B}}(\mathbb{R})$, as the domain of Lebesgue measure is $\bar{\mathcal{B}}(\mathbb{R})$. Thus we have $V \notin \mathcal{B}(\mathbb{R})$.]

Then, $X : \Omega \to \mathbb{R}$ with $X = \mathbf{1}_V$ is not $\mathcal{F}$-measurable.

Proof. Note that $\{1\} \in \mathcal{B}(\mathbb{R})$ but $X^{-1}(\{1\}) = V \notin \mathcal{F}$. $\qquad\square$

Examples of measurable functions:

(1) Let $(\Omega, \mathcal{F})$ be a measurable space. Then every constant function $X : \Omega \to \mathbb{R}$ with $X \equiv c$ is $\mathcal{F}$-measurable.

Proof. For every $B \in \mathcal{B}(\mathbb{R})$, we have

$$X^{-1}(B) = \begin{cases} \Omega & \text{if } B \text{ contains } c, \\ \varnothing & \text{if } B \text{ does not contain } c. \end{cases}$$

Hence, we have $X^{-1}(\mathcal{B}(\mathbb{R})) = \{\varnothing, \Omega\} \subseteq \mathcal{F}$ always. $\qquad\square$

(2) Let $(\Omega, \mathcal{F})$ be a measurable space, and $A \in \mathcal{F}$. Then $X : \Omega \to \mathbb{R}$ with $X = \mathbf{1}_A$ is $\mathcal{F}$-measurable.

Proof. For every $B \in \mathcal{B}(\mathbb{R})$, we have

$$X^{-1}(B) = \begin{cases} \Omega & \text{if } B \text{ contains } 0 \text{ and } 1, \\ A & \text{if } B \text{ contains } 1 \text{ but not } 0, \\ A^c & \text{if } B \text{ contains } 0 \text{ but not } 1, \\ \varnothing & \text{if } B \text{ contains none of } 0 \text{ and } 1. \end{cases}$$

Hence, $X^{-1}(\mathcal{B}(\mathbb{R})) = \{\varnothing, A, A^c, \Omega\} \subseteq \mathcal{F}$, since $A \in \mathcal{F}$. $\qquad\square$

(3) *(Simple random variables)* Let $(\Omega, \mathcal{F})$ be a measurable space, and $A_1, \dots, A_n \in \mathcal{F}$ be pairwise disjoint sets. Then, $X : \Omega \to \mathbb{R}$ with $X = \sum_{i=1}^n x_i \mathbf{1}_{A_i}$ for some $x_1, \dots, x_n \in \mathbb{R}$ is $\mathcal{F}$-measurable. [Note: Such function X is known as a **simple function**, which is essential for the construction of *Lebesgue integral* (see Section 5).]

Proof. Similar to the previous example, we have $X^{-1}(\mathcal{B}(\mathbb{R})) = \{\biguplus_{i \in I} A_i : I \subseteq \{1, \dots, n\}\} \subseteq \mathcal{F}$. $\quad\square$

3.1.3 **Properties of measurable functions.** After studying some examples of non-measurable and measurable functions, we will prove some properties of measurable functions that facilitate our checking of measurability. The following lemma will be helpful for the proof:

Lemma 3.1.a (Commutativity of $\sigma(\cdot)$ and $X^{-1}(\cdot)$)**.** Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be measurable spaces, $X : \Omega \to \Omega'$ be a function, and $\mathcal{A}' \subseteq \mathcal{P}(\Omega')$. Then we have $\sigma(X^{-1}(\mathcal{A}')) = X^{-1}(\sigma(\mathcal{A}'))$. [Note: In particular, if $\mathcal{F}' = \sigma(\mathcal{A}')$, then $\sigma(X^{-1}(\mathcal{A}')) = X^{-1}(\sigma(\mathcal{A}')) = X^{-1}(\mathcal{F}') \overset{\text{(definition)}}{=} \sigma(X)$.]

Proof. "$\subseteq$": Since $\mathcal{A}' \subseteq \sigma(\mathcal{A}')$, by [1.1.9] we have $X^{-1}(\mathcal{A}') \subseteq X^{-1}(\sigma(\mathcal{A}'))$. As $\sigma(\mathcal{A}')$ is a σ-algebra on Ω', $X^{-1}(\sigma(\mathcal{A}'))$ is a σ-algebra on Ω (namely the preimage σ-algebra; see [1.3.9]), which contains $X^{-1}(\mathcal{A}')$ as shown previously. Thus, $\sigma(X^{-1}(\mathcal{A}')) \subseteq X^{-1}(\sigma(\mathcal{A}'))$ by the minimality.

"$\supseteq$": Let $\mathcal{G}' := \{A' \subseteq \Omega' : X^{-1}(A') \in \sigma(X^{-1}(\mathcal{A}'))\}$. We first show that $\mathcal{G}'$ is a σ-algebra on Ω':

(1) $\Omega' \in \mathcal{G}'$ since $X^{-1}(\Omega') = \Omega \in \sigma(X^{-1}(\mathcal{A}'))$.

(2) Fix any $A' \in \mathcal{G}'$. Then we have $A' \subseteq \Omega'$ with $X^{-1}(A') \in \sigma(X^{-1}(\mathcal{A}'))$. So we know $A'^c \subseteq \Omega'$. Also,

$$X^{-1}(A'^c) \overset{[1.1.9]}{=} X^{-1}(A')^c \overset{\text{(closed under complementation)}}{\in} \sigma(X^{-1}(\mathcal{A}')).$$

Thus, $A'^c \in \mathcal{G}'$.

(3) Fix any $A'_1, A'_2, \ldots \in \mathcal{G}'$. Then we have $A'_i \subseteq \Omega'$ with $X^{-1}(A'_i) \in \sigma(X^{-1}(\mathcal{A}'))$ for all $i \in \mathbb{N}$. Therefore, $\bigcup_{i=1}^{\infty} A'_i \subseteq \Omega'$ and

$$X^{-1}\left(\bigcup_{i=1}^{\infty} A'_i\right) \overset{[1.1.9]}{=} \bigcup_{i=1}^{\infty} X^{-1}(A'_i) \overset{\text{(closed under countable unions)}}{\in} \sigma(X^{-1}(\mathcal{A}')),$$

meaning that $\bigcup_{i=1}^{\infty} A'_i \in \mathcal{G}'$.

Note that $X^{-1}(A') \in \sigma(X^{-1}(\mathcal{A}'))$ for all $A' \in \mathcal{A}'$, so $\mathcal{A}' \subseteq \mathcal{G}'$. As $\mathcal{G}'$ is a σ-algebra, we then have $\sigma(\mathcal{A}') \subseteq \mathcal{G}'$, which implies by [1.1.9] that $X^{-1}(\sigma(\mathcal{A}')) \subseteq X^{-1}(\mathcal{G}') \underset{\text{(definition of } \mathcal{G}')}{\subseteq} \sigma(X^{-1}(\mathcal{A}'))$. □

With the help of Lemma 3.1.a, we are now ready to prove the following properties of measurable functions.

(a) *(Checking measurability by considering preimages of generator)* Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be measurable spaces, and $\mathcal{A}' \subseteq \mathcal{P}(\Omega')$ be a *generator* of $\mathcal{F}'$, i.e., $\sigma(\mathcal{A}') = \mathcal{F}'$. Then, $X : \Omega \to \Omega'$ is measurable iff $X^{-1}(\mathcal{A}') \subseteq \mathcal{F}$.

Proof. "⇒": Since $\mathcal{A}' \subseteq \sigma(\mathcal{A}') = \mathcal{F}$, by [1.1.9] we have $X^{-1}(\mathcal{A}') \subseteq X^{-1}(\mathcal{F}') \overset{(X \text{ measurable})}{\subseteq} \mathcal{F}$.
"⇐": Assume that $X^{-1}(\mathcal{A}') \subseteq \mathcal{F}$. Since $\mathcal{F}$ is a σ-algebra, by the minimality we have $\sigma(X^{-1}(\mathcal{A}')) \subseteq \mathcal{F}$. Thus, we have $X^{-1}(\mathcal{F}') = X^{-1}(\sigma(\mathcal{A}')) \underset{\text{(Lemma 3.1.a)}}{=} \sigma(X^{-1}(\mathcal{A}')) \subseteq \mathcal{F}$. □

(b) *(Composition of measurable functions is measurable)* Let $(\Omega, \mathcal{F})$, $(\Omega', \mathcal{F}')$, and $(\Omega'', \mathcal{F}'')$ be measurable spaces. If $X : \Omega \to \Omega'$ and $Y : \Omega' \to \Omega''$ are $(\mathcal{F}, \mathcal{F}')$-measurable and $(\mathcal{F}', \mathcal{F}'')$-measurable respectively, then the composition $Y \circ X : \Omega \to \Omega''$ is $(\mathcal{F}, \mathcal{F}'')$-measurable.

Proof. For all $A'' \in \mathcal{F}''$, we have

$$(Y \circ X)^{-1}(A'') = \{\omega \in \Omega : Y(X(\omega)) \in A''\} \overset{\left(Y(\omega') \in A'' \iff \omega' \in Y^{-1}(A'')\right)}{=} \{\omega \in \Omega : X(\omega) \in Y^{-1}(A'')\}$$

$$\overset{\left(X(\omega) \in A' \iff \omega \in X^{-1}(A')\right)}{=} \{\omega \in \Omega : \omega \in X^{-1}(Y^{-1}(A''))\} = X^{-1}(\underbrace{Y^{-1}(A'')}_{\in \mathcal{F}' \text{ as } Y \text{ is measurable}}) \overset{(X \text{ measurable})}{\in} \mathcal{F}.$$

□

(c) *(Continuous functions are measurable)* Let $(\Omega, \mathcal{T})$ and $(\Omega', \mathcal{T}')$ be topological spaces. If $X : \Omega \to \Omega'$ is continuous, then X is $(\mathcal{B}(\Omega), \mathcal{B}(\Omega'))$-measurable.

Proof. By the continuity of X, we have $X^{-1}(\mathcal{T}') \subseteq \mathcal{T} \subseteq \sigma(\mathcal{T}) \overset{\text{(definition)}}{=} \mathcal{B}(\Omega)$. As $\mathcal{B}(\Omega)$ is a σ-algebra, by the minimality we have $\sigma(X^{-1}(\mathcal{T}')) \subseteq \mathcal{B}(\Omega)$. Noting that $\sigma(\mathcal{T}') = \mathcal{B}(\Omega')$ by definition, using Lemma 3.1.a gives $X^{-1}(\mathcal{B}(\Omega')) = X^{-1}(\sigma(\mathcal{T}')) = \sigma(X^{-1}(\mathcal{T}')) \subseteq \mathcal{B}(\Omega)$, as desired. □

(d) *(Monotone functions are measurable)* If $h : \mathbb{R} \to \mathbb{R}$ is a monotone (increasing or decreasing) function, then h is $(\mathcal{B}(\mathbb{R}), \mathcal{B}(\mathbb{R}))$-measurable.

Proof. WLOG, we prove only the case where h is increasing. By Proposition 1.4.g, we know $\{[t, \infty) : t \in \mathbb{R}\}$ is a generator of $\mathcal{B}(\mathbb{R})$. Thus by [3.1.3]a it suffices to show that $h^{-1}([t, \infty)) \in \mathcal{B}(\mathbb{R})$ for all $t \in \mathbb{R}$.

Fix any $t \in \mathbb{R}$ and consider the set $A_t := h^{-1}([t, \infty)) = \{x \in \mathbb{R} : h(x) \geq t\}$. Fix any $x_1 \in A_t$. Then, for every $x_2 \geq x_1$, we have $h(x_2) \overset{(h \text{ increasing})}{\geq} h(x_1) \overset{(x_1 \in A_t)}{\geq} t$, thus $x_2 \in A_t$. This forces A_t to be of the form (i) $[\inf A_t, \infty)$ if $t \in \text{ran}(h)$, or (ii) $(\inf A_t, \infty)$ if $h(\inf A_t) < t$. In either case, A_t is in $\mathcal{B}(\mathbb{R})$, as desired. □

3.1.4 **An example of Lebesgue set that is not a Borel set.** Now we have enough tools to construct a Lebesgue set that is not a Borel set, whose existence is asserted in [2.4.5].

Let $U := F^-(V)$, where F^- is the quantile function of the *Cantor distribution function* F (recall [2.5.8]), and V is the Vitali set. [Note: Here $F^-(V)$ denotes the image set $\{F^-(y) : y \in V\}$.] We claim that U is a Lebesgue set but not a Borel set.

Proof. **Showing that U is a Lebesgue set.** Since the Cantor set $\mathcal{C}$ is a countable intersection of closed sets (complements of open sets, which are in $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{T})$), we know $\mathcal{C} \in \mathcal{B}(\mathbb{R})$. Also, we have previously shown in [2.4.8] that $\lambda(\mathcal{C}) = 0$, so $\mathcal{C}$ is a λ-null set in $\mathcal{B}(\mathbb{R})$.

Noting that $V \subseteq [0,1]$, we have $U = F^{-}(V) \subseteq F^{-}([0,1]) = \mathcal{C}$. Therefore, U belongs to the Lebesgue σ-algebra $\bar{\mathcal{B}}(\mathbb{R})$ by the definition of completion, and is thus a Lebesgue set.

Showing that U is not a Borel set. By [2.5.10], we know $F^{-} : [0,1] \to \bar{\mathbb{R}}$ is increasing. By defining $F^{-}(y) := F^{-}(0)$ for all $y < 0$ and $F^{-}(y) := F^{-}(1)$ for all $y > 1$, we can extend the domain of F^{-} to $\mathbb{R}$, and we shall consider this extended F^{-} in the following. Note that F^{-} is still increasing after the extension, so F^{-} is Borel measurable by [3.1.3]d.

Now assume to the contrary that U is a Borel set. Then since F^{-} is Borel measurable, the preimage $(F^{-})^{-1}(U)$ would belong to $\mathcal{B}(\mathbb{R})$. However, we note that

$$(F^{-})^{-1}(U) = \{y \in \mathbb{R} : F^{-}(y) \in U\} = \{y \in \mathbb{R} : F^{-}(y) \in F^{-}(V)\}$$

$$\overset{\left(y \in V \iff F^{-}(y) \in F^{-}(V)\right)}{=} \{y \in \mathbb{R} : y \in V\} = V \notin \mathcal{B}(\mathbb{R}),$$

contradiction. $\qquad\square$

3.1.5 **Properties of random variables/vectors.** After studying properties of general measurable functions, we now investigate the properties of more specific measurable functions, namely random variables/vectors.

(a) *(Random vectors are vectors of random variables)* Let $(\Omega, \mathcal{F})$ be a measurable space and let $X : \Omega \to \mathbb{R}^d$ be a function. Then X is a random vector iff $X = (X_1, \ldots, X_d)$ for some random variables $X_1, \ldots, X_d$. [Note: In view of this result, often we use bold notation to denote a random vector, e.g., $\boldsymbol{X}$.]

Proof. "$\Rightarrow$": Assume X is a random vector, thus measurable. Fix any $j = 1, \ldots, d$, and consider the projection $\pi_j : \mathbb{R}^d \to \mathbb{R}$ defined by $\pi_j(x_1, \ldots, x_d) = x_j$. Note that π_j is continuous since for all $\varepsilon > 0$, we can choose $\delta = \varepsilon > 0$ such that, whenever $\|\boldsymbol{x} - \boldsymbol{y}\| < \delta$, we have $|\pi_j(\boldsymbol{x}) - \pi_j(\boldsymbol{y})| = |x_j - y_j| \le \|\boldsymbol{x} - \boldsymbol{y}\| < \delta = \varepsilon$. Thus, by [3.1.3]c, π_j is measurable, and so $\pi_j \circ X : \Omega \to \mathbb{R}$ is measurable by [3.1.3]b, hence a random variable. By letting $X_j := \pi_j \circ X$ for all $j = 1, \ldots, d$, we then have $X = (X_1, \ldots, X_d)$.

"$\Leftarrow$": Assume $X = (X_1, \ldots, X_d)$ for some random variables $X_1, \ldots, X_d$. By Proposition 1.4.g we know $\{(\boldsymbol{a}, \boldsymbol{b}] : \boldsymbol{a} < \boldsymbol{b}\}$ is a generator of $\mathcal{B}(\mathbb{R}^d)$, thus by [3.1.3]a it suffices to show that $X^{-1}((\boldsymbol{a}, \boldsymbol{b}]) \in \mathcal{F}$ for all $\boldsymbol{a} < \boldsymbol{b}$: For all $(a_1, \ldots, a_d) = \boldsymbol{a} < \boldsymbol{b} = (b_1, \ldots, b_d)$, we have

$$X^{-1}((\boldsymbol{a}, \boldsymbol{b}]) = \{\omega \in \Omega : X_j(\omega) \in (a_j, b_j] \text{ for all } j = 1, \ldots, d\} = \bigcap_{j=1}^{d} \underbrace{X_j^{-1}((a_j, b_j])}_{\in \mathcal{F} \text{ as } X_j \in \mathcal{F}} \in \mathcal{F}.$$

$\qquad\square$

(b) *(Measurable functions of random vectors are random vectors)* Let $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^k$ be a measurable function. If $\boldsymbol{X} : \Omega \to \mathbb{R}^d$ is a random vector, then $\boldsymbol{h}(\boldsymbol{X}) : \Omega \to \mathbb{R}^k$ is a random vector.

[Note: Particularly, since sums, products, minima, and maxima are continuous functions and thus measurable, sums, products, minima, and maxima of random variable are random variables.]

Proof. It follows from [3.1.3]b. $\qquad\square$

(c) *(Measurability about sequences of random variables)* Let $\{X_i\}_{i \in \mathbb{N}}$ be a sequence of random variables on a measurable space $(\Omega, \mathcal{F})$. [Note: It can be shown that X is a random sequence iff $X = \{X_i\}_{i \in \mathbb{N}}$ for some random variables $X_1, X_2, \ldots$. So the sequence here indeed constitutes a random sequence.]

i. $\inf_{k \ge n} X_k$, $\sup_{k \ge n} X_k$, $\liminf_{n \to \infty} X_n := \sup_{n \ge 1}(\inf_{k \ge n} X_k)$, and $\limsup_{n \to \infty} X_n := \inf_{n \ge 1}(\sup_{k \ge n} X_k)$ are all random variables.
[Note: "$\inf_{k \ge n} X_k$" refers to the function that takes the value $\inf_{k \ge n} X_k(\omega) = \inf\{X_k(\omega) : k \ge n\}$ with input $\omega \in \Omega$; similar for others.]

ii. If $\lim_{n\to\infty} X_n(\omega)$ exists for all $\omega \in \Omega$, then $\lim_{n\to\infty} X_n$ is a random variable.
[Note: Again, "$\lim_{n\to\infty} X_n$" refers to the function that takes the value $\lim_{n\to\infty} X_n(\omega)$ with input $\omega \in \Omega$.]

iii. The set $\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \text{ exists}\}$ is measurable (i.e., is in $\mathcal{F}$).

Proof.

i. • $\inf_{k\geq n} X_k$ *is a random variable:* For all $x \in \mathbb{R}$,

$$\left(\inf_{k\geq n} X_k\right)^{-1}((-\infty, x]) = \left\{\omega \in \Omega : \inf_{k\geq n} X_k(\omega) \leq x\right\} = \bigcup_{k\geq n} \{\omega \in \Omega : X_k(\omega) \leq x\}$$

$$= \bigcup_{k\geq n} \underbrace{X_k^{-1}((-\infty, x])}_{\in \mathcal{F} \text{ as } X_k \in \mathcal{F}} \in \mathcal{F}.$$

Thus the result follows by Proposition 1.4.g and [3.1.3]a.

• $\sup_{k\geq n} X_k$ *is a random variable:* For all $x \in \mathbb{R}$,

$$\left(\sup_{k\geq n} X_k\right)^{-1}([x, \infty)) = \left\{\omega \in \Omega : \sup_{k\geq n} X_k(\omega) \geq x\right\} = \bigcup_{k\geq n} \{\omega \in \Omega : X_k(\omega) \geq x\}$$

$$= \bigcup_{k\geq n} \underbrace{X_k^{-1}([x, \infty))}_{\in \mathcal{F} \text{ as } X_k \in \mathcal{F}} \in \mathcal{F}.$$

Thus the result follows by Proposition 1.4.g and [3.1.3]a.

• $\liminf_{n\to\infty} X_k$ *and* $\limsup_{n\to\infty} X_k$ *are random variables:* It follows from writing $\liminf_{n\to\infty} X_n = \sup_{n\geq 1}(\inf_{k\geq n} X_k)$ and $\limsup_{n\to\infty} X_n = \inf_{n\geq 1}(\sup_{k\geq n} X_k)$, and applying [3.1.3]b.

ii. Since $\lim_{n\to\infty} X_n(\omega)$ exists for all $\omega \in \Omega$, we have $\lim_{n\to\infty} X_n(\omega) = \limsup_{n\to\infty} X_n(\omega)$ for all $\omega \in \Omega$. So, the result follows from (i).

iii. Note that for all $\omega \in \Omega$, $\liminf_{n\to\infty} X_n(\omega) \leq \limsup_{n\to\infty} X_n(\omega)$ and the limit $\lim_{n\to\infty} X_n(\omega)$ exists iff the equality is achieved. Hence, we can write

$$\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \text{ exists}\}^c = \left\{\omega \in \Omega : \liminf_{n\to\infty} X_n(\omega) < \limsup_{n\to\infty} X_n(\omega)\right\}$$

$$= \bigcup_{q\in\mathbb{Q}} \left\{\omega \in \Omega : \liminf_{n\to\infty} X_n(\omega) < q < \limsup_{n\to\infty} X_n(\omega)\right\} \qquad (6)$$

$$= \bigcup_{q\in\mathbb{Q}} \left(\underbrace{\left\{\omega \in \Omega : \liminf_{n\to\infty} X_n(\omega) < q\right\}}_{\in \mathcal{F} \text{ by (i)}} \cap \underbrace{\left\{\omega \in \Omega : \limsup_{n\to\infty} X_n(\omega) > q\right\}}_{\in \mathcal{F} \text{ by (i)}}\right)$$

$$\overset{(\mathbb{Q} \text{ countable})}{\in} \mathcal{F}.$$

To show the equality in Equation (6), consider:

• "$\subseteq$": Fix any $\omega \in \Omega$ with $\liminf_{n\to\infty} X_n(\omega) < \limsup_{n\to\infty} X_n(\omega)$. By the density of $\mathbb{Q}$ on $\mathbb{R}$, there exists $q \in \mathbb{Q}$ such that $\liminf_{n\to\infty} X_n(\omega) < q < \limsup_{n\to\infty} X_n(\omega)$. Thus $\omega \in \bigcup_{q\in\mathbb{Q}}\{\omega \in \Omega : \liminf_{n\to\infty} X_n(\omega) < q < \limsup_{n\to\infty} X_n(\omega)\}$.

• "$\supseteq$": Fix any $\omega \in \Omega$ with $\liminf_{n\to\infty} X_n(\omega) < q < \limsup_{n\to\infty} X_n(\omega)$ for some $q \in \mathbb{Q}$. Then it is immediate that $\liminf_{n\to\infty} X_n(\omega) < \limsup_{n\to\infty} X_n(\omega)$.

$\square$

(d) *(Relationship between completeness and measurability)* Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and $X, Y, Z, X_1, X_2, \ldots$ be functions from Ω to $\mathbb{R}$. Then:

i. μ is complete iff (X is $\mathcal{F}$-measurable and $Y \overset{\text{a.e.}}{=} X$) $\implies$ Y is $\mathcal{F}$-measurable.

ii. μ is complete iff $(X_n$ is $\mathcal{F}$-measurable $\forall n \in \mathbb{N}$ and $Z \overset{\text{a.e.}}{=} \lim_{n\to\infty} X_n) \implies Z$ is $\mathcal{F}$-measurable.

Proof.

i. "$\Rightarrow$": Assume μ is complete. Then $\mathcal{F}$ contains all subsets of every null set. Suppose X is $\mathcal{F}$-measurable and $X \overset{\text{a.e.}}{=} Y$. Then we have $X = Y$ on N^c, where $N \in \mathcal{F}$ is a null set. Fix any set $A \in \mathcal{B}(\mathbb{R})$, and write $Y^{-1}(A) = (Y^{-1}(A) \cap N) \cup (Y^{-1}(A) \cap N^c)$. Note that:

- $Y^{-1}(A) \cap N \in \mathcal{F}$ since it is a subset of the null set N.
- $Y^{-1}(A) \cap N^c \in \mathcal{F}$ since we have $X = Y$ on N^c, thus

$$Y^{-1}(A) \cap N^c = \underbrace{X^{-1}(A)}_{\in \mathcal{F} \text{ as } X \in \mathcal{F}} \cap \underbrace{N^c}_{\in \mathcal{F} \text{ as } N \in \mathcal{F}} \in \mathcal{F}.$$

Therefore, $Y^{-1}(A) \in \mathcal{F}$, and so Y is $\mathcal{F}$-measurable.

"$\Leftarrow$": Assume we have $(X$ is $\mathcal{F}$-measurable and $Y \overset{\text{a.e.}}{=} X) \implies Y$ is $\mathcal{F}$-measurable. Fix any null set $N \in \mathcal{F}$, and consider any subset $N' \subseteq N$. Let $X = \mathbf{1}_N$ and $Y = \mathbf{1}_{N'}$. By construction, we have $X = Y = 0$ on N^c, thus $X \overset{\text{a.e.}}{=} Y$. Furthermore, X is $\mathcal{F}$-measurable (it is a simple random variable). Hence by assumption, Y is $\mathcal{F}$-measurable, meaning that $N' = Y^{-1}(\{1\}) \in \mathcal{F}$. So we conclude that μ is complete.

ii. "$\Rightarrow$": Assume μ is complete. Suppose X_n is $\mathcal{F}$-measurable for all $n \in \mathbb{N}$ and $Z \overset{\text{a.e.}}{=} \lim_{n\to\infty} X_n$. Then we have $Z(\omega) = \lim_{n\to\infty} X_n(\omega)$ for all $\omega \in N^c$, where $N \in \mathcal{F}$ is a null set. Since $\mathbf{1}_{N^c}$ is $\mathcal{F}$-measurable, by [3.1.5]b we know $X_n \mathbf{1}_{N^c}$ is $\mathcal{F}$-measurable for all $n \in \mathbb{N}$.

We claim that $\lim_{n\to\infty} X_n(\omega)\mathbf{1}_{N^c}(\omega) = Z(\omega)\mathbf{1}_{N^c}(\omega)$ for all $\omega \in \Omega$. To see this, consider:

- *Case 1:* $\omega \in N$. Then we have $\mathbf{1}_{N^c}(\omega) = 0$, so both sides of the equality are zero.
- *Case 2:* $\omega \in N^c$. The equality holds as we have $Z(\omega) = \lim_{n\to\infty} X_n(\omega)$ for all $\omega \in N^c$.

Thus, by [3.1.5]c, $Z\mathbf{1}_{N^c}$ is $\mathcal{F}$-measurable. Since $Z \overset{\text{a.e.}}{=} Z\mathbf{1}_{N^c}$ (we have $Z_n = Z_n\mathbf{1}_{N^c}$ on N^c), by the "$\Rightarrow$" direction of (i), we conclude that Z is $\mathcal{F}$-measurable.

"$\Leftarrow$": Assume we have $(X_n$ is $\mathcal{F}$-measurable $\forall n \in \mathbb{N}$ and $Z \overset{\text{a.e.}}{=} \lim_{n\to\infty} X_n) \implies Z$ is $\mathcal{F}$-measurable. Like (i), fix any null set $N \in \mathcal{F}$, and consider any subset $N' \subseteq N$. Let $X_n = \mathbf{1}_N$ for all $n \in \mathbb{N}$, and $Z = \mathbf{1}_{N'}$. Knowing that X_n is $\mathcal{F}$-measurable for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} X_n = \mathbf{1}_N \overset{\text{a.e.}}{=} \mathbf{1}_{N'} = Z$, by assumption Z is $\mathcal{F}$-measurable. Thus, $N' = Z^{-1}(\{1\}) \in \mathcal{F}$, as desired.

$\square$

3.2 Distributions

3.2.1 In your first probability course, you have always seen terms like "distributions of random variables", like normal distribution, exponential distribution, etc. Intuitively, *distribution* is a concept that describes the "probabilistic behaviour" of random variables. Here we shall investigate the *formal* definition of distribution, which indeed aligns with this intuition.

3.2.2 **Measures induced by measurable functions.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, $(\Omega', \mathcal{F}')$ be a measurable space, and $X : \Omega \to \Omega'$ be a measurable function. Then, $\mu_X := \mu \circ X^{-1}$ is a measure on $(\Omega', \mathcal{F}')$.

[Note: Here "X^{-1}" is referring to the *preimage function* rather than the inverse of X, i.e., the function $X^{-1} : \mathcal{F}' \to \mathcal{F}$ that maps the *set* $A' \in \mathcal{F}'$ to the *preimage* $X^{-1}(A') \in \mathcal{F}$.]

Proof.

(1) Since μ_X is a function from $\mathcal{F}'$ to $[0, \infty]$, nonnegativity is satisfied.

(2) We have $\mu_X(\varnothing) = \mu(X^{-1}(\varnothing)) \overset{(X^{-1}(\varnothing)=\varnothing)}{=} \mu(\varnothing) = 0$.

(3) Fix any pairwise disjoint $A'_1, A'_2, \ldots \in \mathcal{F}'$. Note that $X^{-1}(A'_1), X^{-1}(A'_2), \ldots \in \mathcal{F}$ are still pairwise disjoint since for all $i \neq j$, we have $X^{-1}(A'_i) \cap X^{-1}(A'_j) \overset{[1.1.9]}{=} X^{-1}(A'_i \cap A'_j) = X^{-1}(\varnothing) = \varnothing$. Thus,

$$\mu_X\left(\biguplus_{i=1}^{\infty} A'_i\right) = \mu\left(X^{-1}\left(\biguplus_{i=1}^{\infty} A'_i\right)\right) \overset{[1.1.9]}{=} \mu\left(\biguplus_{i=1}^{\infty} X^{-1}(A'_i)\right) = \sum_{i=1}^{\infty} \mu(X^{-1}(A'_i)) = \sum_{i=1}^{\infty} \mu_X(A_i).$$

55

$\square$

Remarks:

- The measure μ_X is also known as the **distribution** of X, or **push-forward measure**/**image measure** of μ with respect to X (μ_X is assigning measures to subsets of the *codomain* Ω' rather than the domain Ω, so the measure is "pushed forward").

- A shorthand notation that is often used for distribution is as follows: $\mu(X \in A') := \mu(\{\omega \in \Omega : X(\omega) \in A'\}) = \mu(X^{-1}(A')) = \mu_X(A')$. (Usually the "$\mu$" here is $\mathbb{P}$.)

- In case $\mu = \mathbb{P}$ is a probability measure and $(\Omega', \mathcal{F}') = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we have

$$\mathbb{P}(\boldsymbol{a} < \boldsymbol{X} \leq \boldsymbol{b}) := \mathbb{P}(\boldsymbol{X} \in (\boldsymbol{a}, \boldsymbol{b}]) = \mathbb{P}(\boldsymbol{X}^{-1}((\boldsymbol{a}, \boldsymbol{b}])) = \mathbb{P}_{\boldsymbol{X}}((\boldsymbol{a}, \boldsymbol{b}]),$$

for all $\boldsymbol{a} < \boldsymbol{b}$. Also, the distribution $\mathbb{P}_{\boldsymbol{X}}$ is a probability measure on $\mathcal{B}(\mathbb{R}^d)$ since $\mathbb{P}_{\boldsymbol{X}}(\mathbb{R}^d) = \mathbb{P}(X^{-1}(\mathbb{R}^d)) = \mathbb{P}(\Omega) = 1$.

3.2.3 Characterization of distribution by distribution function.

Due to the presence of the word "distribution" in the term *distribution function*, one would naturally anticipate that there should be a relationship between the concepts of *distribution* and *distribution function*. This is indeed the case, and distribution function actually *characterizes* distribution (i.e., they have a one-to-one correspondence), as the following result suggests.

Theorem 3.2.a (Characterization of distribution by distribution function).

(a) $\mathbb{P}_{\boldsymbol{X}}$ *induces* F: Let $\boldsymbol{X}$ be a random vector with the distribution $\mathbb{P}_{\boldsymbol{X}}$. Then, the function $F(\boldsymbol{x}) := \mathbb{P}_{\boldsymbol{X}}((-\infty, \boldsymbol{x}]) = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x})$ is the distribution function of $\mathbb{P}_{\boldsymbol{X}}$.

(b) F *induces* $\mathbb{P}_{\boldsymbol{X}}$: If F is a distribution function on $\mathbb{R}^d$, then we can define a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that there is a random vector $\boldsymbol{X} : \Omega \to \mathbb{R}^d$ with distribution $\mathbb{P}_{\boldsymbol{X}}$ that satisfies $\mathbb{P}_{\boldsymbol{X}}((-\infty, \boldsymbol{x}]) = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}) = F(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

Proof.

(a) From [3.2.2] we know that $\mathbb{P}_{\boldsymbol{X}}$ is a probability measure on $\mathcal{B}(\mathbb{R}^d)$. Therefore, by Theorem 2.5.a, the function $F(\boldsymbol{x}) := \mathbb{P}_{\boldsymbol{X}}((-\infty, \boldsymbol{x}]) = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x})$ is the distribution function of $\mathbb{P}_{\boldsymbol{X}}$.

(b) Here we only prove the case for $d = 1$. Assume F is a distribution function on $\mathbb{R}$. Letting $(\Omega, \mathcal{F}, \mathbb{P}) = ((0, 1], \mathcal{B}((0, 1]), \lambda)$ (where λ is the Lebesgue measure, with domain restricted to $\mathcal{B}((0, 1]))$ and $X := F^{-1}$, we have

$$\mathbb{P}(X \leq x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbb{P}(\{\omega \in \Omega : F^{-1}(\omega) \leq x\})$$
$$\overset{[2.5.10]}{=} \mathbb{P}(\{\omega \in \Omega : \omega \leq F(x)\}) = \mathbb{P}((0, F(x)]) = \lambda((0, F(x)]) = F(x)$$

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

$\square$

$$\mathbb{P}_{\boldsymbol{X}} \xrightarrow{F(\boldsymbol{x}) := \mathbb{P}_{\boldsymbol{X}}((-\infty, \boldsymbol{x}])} F \xrightarrow{\mathbb{P}_{\boldsymbol{X}}((-\infty, \boldsymbol{x}]) = F(\boldsymbol{x})} \mathbb{P}_{\boldsymbol{X}}$$

$$F \xrightarrow{\mathbb{P}_{\boldsymbol{X}}((-\infty, \boldsymbol{x}]) = F(\boldsymbol{x})} \mathbb{P}_{\boldsymbol{X}} \xrightarrow{F(\boldsymbol{x}) := \mathbb{P}_{\boldsymbol{X}}((-\infty, \boldsymbol{x}])} F$$

Remarks:

- In view of this result, such distribution function F is said to be the **distribution function of $\boldsymbol{X}$** (indeed this is perhaps how distribution function is *defined* in your first probability course), denoted by $\boldsymbol{X} \sim F$, and sometimes we write the distribution function as $F_{\boldsymbol{X}}$.

- By Propositions 1.4.g and 2.1.a, it can be shown that for two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{B}(\mathbb{R}^d)$, if $\mathbb{P}((-\infty, \boldsymbol{x}]) = \mathbb{Q}((-\infty, \boldsymbol{x}])$ for all $\boldsymbol{x} \in \mathbb{R}^d$, then $\mathbb{P} = \mathbb{Q}$. This explains why the picture above depicts a one-to-one correspondence between F and $\mathbb{P}_{\boldsymbol{X}}$: For the first line, applying (b) on the F induced by $\mathbb{P}_{\boldsymbol{X}}$ would indeed give you back the original $\mathbb{P}_{\boldsymbol{X}}$, as the resulting probability measure agrees with the original $\mathbb{P}_{\boldsymbol{X}}$ for every set of the form $(-\infty, \boldsymbol{x}]$, and so coincides with $\mathbb{P}_{\boldsymbol{X}}$.

- If two random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, defined on $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ respectively, have the same distribution function, then their distributions $(\mathbb{P}_1)_{\boldsymbol{X}_1}$ and $(\mathbb{P}_2)_{\boldsymbol{X}_2}$ would be the same as well. This leads to the following definition: Two random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are said to be **equal in distribution**, denoted by $\boldsymbol{X}_1 \overset{\mathrm{d}}{=} \boldsymbol{X}_2$, if we have $\boldsymbol{X}_1, \boldsymbol{X}_2 \sim F$ (which would indeed imply that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have the same distribution, as the name suggests).

- For a random vector $\boldsymbol{X} \sim F$, we have

$$\mathbb{P}(\boldsymbol{X} \in (\boldsymbol{a}, \boldsymbol{b}]) = \mathbb{P}_{\boldsymbol{X}}((\boldsymbol{a}, \boldsymbol{b}]) \overset{(\text{Theorem } 2.5.a)}{=} \lambda_F((\boldsymbol{a}, \boldsymbol{b}]) = \Delta_{(\boldsymbol{a},\boldsymbol{b}]} F,$$

offering another interpretation of the F-volume $\Delta_{(\boldsymbol{a},\boldsymbol{b}]} F$: the probability for $\boldsymbol{X} \sim F$ to take values in $(\boldsymbol{a}, \boldsymbol{b}]$.

- A random vector $\boldsymbol{X}$ is said to be **discrete**, **absolutely continuous**, **continuous singular**, or **mixed type** if its distribution function is. Since density/mass function f uniquely determines the distribution function F, sometimes we also write $\boldsymbol{X} \sim f$ to mean $\boldsymbol{X} \sim F$, whenever the density/mass function f exists. This applies similarly to other expressions that uniquely characterize F also, e.g., we write "$X \sim \mathrm{N}(\mu, \sigma^2)$" as the expression "$\mathrm{N}(\mu, \sigma^2)$" refers to a unique distribution function.

- When we have $\boldsymbol{X} \sim F$, sometimes we call $\mathrm{supp}(F) = \{\boldsymbol{x} \in \mathbb{R}^d : \Delta_{(\boldsymbol{x}-\boldsymbol{h}, \boldsymbol{x}]} F > 0 \text{ for all } \boldsymbol{h} > \boldsymbol{0}\}$ as the **support** of $\boldsymbol{X}$, denoted by $\mathrm{supp}(\boldsymbol{X})$. There are some other alternative definitions of the support of $\boldsymbol{X}$, including:

  - the *smallest closed set* $C \subseteq \mathbb{R}^d$ such that $\mathbb{P}(\boldsymbol{X} \in C) = 1$, and
  - the set $\{\boldsymbol{x} \in \mathbb{R}^d : \mathbb{P}_{\boldsymbol{X}}(B(\boldsymbol{x}, r)) > 0 \text{ for all } r > 0\}$ where $B(\boldsymbol{x}, r)$ denotes the open ball in $\mathbb{R}^d$ with center $\boldsymbol{x}$ and radius r.

  It can be shown that all these definitions are equivalent; see https://math.stackexchange.com/q/846011 for a related discussion on this matter.

- Sometimes one abuses the notation of F and writes $F(B) := \mathbb{P}_{\boldsymbol{X}}(B)$. In view of this, the notations F and $\mathbb{P}_{\boldsymbol{X}}$ are sometimes used interchangeably (see e.g., [5.2.3]a). Also, the terms "distribution" and "distribution function" are used interchangeably sometimes.

3.2.4 **Preservation of equality in distribution after applying measurable functions.** Intuitively, given two random vectors that are equal in distribution, applying an identical (measurable) function to each of them should still preserve the equality in distribution, as the function should lead to the same "change" for both distributions, so the resulting ones would still be the same as each other. The following result justifies this intuition.

Proposition 3.2.b. Let $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ be random vectors on $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ respectively. If $\boldsymbol{X}_1 \overset{\mathrm{d}}{=} \boldsymbol{X}_2$, then $\boldsymbol{h}(\boldsymbol{X}_1) \overset{\mathrm{d}}{=} \boldsymbol{h}(\boldsymbol{X}_2)$ for every measurable function $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^k$.

Proof. For every measurable function $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^k$, we have

$$F_{\boldsymbol{h}(\boldsymbol{X}_1)}(\boldsymbol{x}) = \mathbb{P}_1(\boldsymbol{h}(\boldsymbol{X}_1) \leq \boldsymbol{x}) = \mathbb{P}_1\big(\boldsymbol{X}_1 \in \boldsymbol{h}^{-1}((-\infty, \boldsymbol{x}])\big)$$

$$\overset{(\boldsymbol{X}_1 \overset{\mathrm{d}}{=} \boldsymbol{X}_2)}{=} \mathbb{P}_2\big(\boldsymbol{X}_2 \in \boldsymbol{h}^{-1}((-\infty, \boldsymbol{x}])\big) = \mathbb{P}_2(\boldsymbol{h}(\boldsymbol{X}_2) \leq \boldsymbol{x}) = F_{\boldsymbol{h}(\boldsymbol{X}_2)}(\boldsymbol{x})$$

for all $\boldsymbol{x} \in \mathbb{R}^k$. Thus, $\boldsymbol{h}(\boldsymbol{X}_1) \overset{\mathrm{d}}{=} \boldsymbol{h}(\boldsymbol{X}_2)$ by definition. $\square$

3.3 Margins

3.3.1 Margins of random vectors. Recall from [2.5.6] that the *J-margin* of F is given by $F_J(\boldsymbol{x}_j) := \lim_{\boldsymbol{x}_{J^c} \to \infty} F(\boldsymbol{x})$. Here, we consider the special case where F is the distribution function of a random vector $\boldsymbol{X}$. In such case, the J-margin of F is given by

$$F_J(\boldsymbol{x}_j) = \lim_{\boldsymbol{x}_{J^c} \to \infty} \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}) \overset{\text{(continuity from below)}}{=} \mathbb{P}(\boldsymbol{X}_J \leq \boldsymbol{x}_J)$$

for all $\boldsymbol{x}_J \in \mathbb{R}^{|J|}$. [Note: Here we view $\boldsymbol{X}$ as the vector $(X_1, \ldots, X_d)$, so the notation $\boldsymbol{X}_J$ refers to the vector $(X_j)_{j \in J}$.]

By Theorem 3.2.a, we know that this is the distribution function of $\boldsymbol{X}_J$. In view of this, the random vector $\boldsymbol{X}_J$ is called the *J*-margin of $\boldsymbol{X}$. In the special case where $J = \{j\}$ for some $j = 1, \ldots, d$, we would have the *jth margin of F*, given by $F_j(x_j) = \mathbb{P}(X_j \leq x_j)$, and hence X_j is called the *jth margin of $\boldsymbol{X}$*.

3.3.2 Margins of absolutely continuous/discrete random vectors. First consider the special case where F is absolutely continuous. The J-margin of F would then be

$$F_J(\boldsymbol{x}_J) = F(\infty_{J \leftarrow \boldsymbol{x}_J}) = \int_{(-\infty, \infty_{J \leftarrow \boldsymbol{x}_J}]} f(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z} = \int_{-\infty}^{\boldsymbol{x}_J} \int_{-\infty}^{\infty} f(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}_{J^c} \, \mathrm{d}\boldsymbol{z}_J, \quad \text{for all } \boldsymbol{x}_J \in \mathbb{R}^{|J|}.$$

Thus, F_J is also absolutely continuous with the density function being $f_J(\boldsymbol{x}_J) = \int_{-\infty}^{\infty} f(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}_{J^c}$, which is called the *J-marginal density function* of F or $\boldsymbol{X}$. This means that every lower-dimensional margin of an absolutely continuous distribution function F is absolutely continuous, and the corresponding density function can be obtained by *integrating out the joint density f* over the other variables. For the case where F is discrete, change $\int \to \sum$, "integrating" $\to$ "summing", and "density" $\to$ "mass".

However, having absolutely continuous margins does *not* imply that F is absolutely continuous. To see this, consider the following example with $\boldsymbol{X} = (U, U)$, where $U \sim \mathrm{U}(0, 1)$. Here both $X_1 = U$ and $X_2 = U$ are absolutely continuous, with the density function being $f(x) = \mathbf{1}_{(0,1)}(x)$. On the other hand, since the distribution function of $\boldsymbol{X}$ is $F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}(U \leq x_1, U \leq x_2) = \mathbb{P}(U \leq \min\{x_1, x_2\}) = \min\{x_1, x_2\}$, where $\min\{x_1, x_2\} \in (0, 1)$, we have

$$\frac{\partial^2}{\partial x_2 \, \partial x_1} F_{\boldsymbol{X}}(\boldsymbol{x}) = \begin{cases} \frac{\partial^2}{\partial x_2 \, \partial x_1} x_1 = \frac{\partial}{\partial x_2} 1 = 0 & \text{if } x_1 < x_2, \\ \frac{\partial^2}{\partial x_2 \, \partial x_1} x_2 = \frac{\partial}{\partial x_2} 0 = 0 & \text{if } x_1 > x_2, \end{cases}$$

which integrates to 0 rather than 1. This then implies that F is not absolutely continuous.

3.4 Survival Functions

3.4.1 While distribution functions are describing probabilities of the form $\mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x})$, *survival distributions* are describing probabilities of the form $\mathbb{P}(\boldsymbol{X} > \boldsymbol{x})$, which are *exceedance probabilities*. Such probabilities are of great interest in some applications, such as modelling extreme losses in actuarial science, where $\boldsymbol{X}$ could represent a collection of losses.

3.4.2 Definitions. The **survival function** of a random vector $\boldsymbol{X} \sim F$ is defined by $\bar{F}(\boldsymbol{x}) := \mathbb{P}(\boldsymbol{X} > \boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$. The *J*-margin of $\bar{F}$ is given by $\bar{F}_J(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X}_J > \boldsymbol{x}_J)$, and the *j*th marginal survival function is $\bar{F}_j(x_j) := \mathbb{P}(X_j > x_j) = 1 - F_j(x_j)$, for all $j = 1, \ldots, d$.

3.4.3 Expression of survival functions in terms of distribution functions. While we have the "nice" relationship $\bar{F}(x) = 1 - F(x)$ when $d = 1$, it breaks down when $d \geq 2$. For instance, with $d = 2$ we have $\bar{F}(x_1, x_2) = \mathbb{P}(X_1 > x_1 \cap X_2 > x_2) = 1 - \mathbb{P}(X_1 \leq x_1 \cup X_2 \leq x_2) \overset{\text{(inclusion-exclusion)}}{=} 1 - (F_1(x_1) + F_2(x_2) - F(x_1, x_2)) = 1 - F_1(x_1) - F_2(x_2) + F(x_1, x_2)$, which is clearly *not* equal to $1 - F(x_1, x_2)$ in general ⚠.

In general, the expression of $\bar{F}$ in terms of F is as follows:

$$\bar{F}(\boldsymbol{x}) = \sum_{J \subseteq \{1,\ldots,d\}} (-1)^{|J|} F(\boldsymbol{\infty}_{J \leftarrow \boldsymbol{x}_J})$$

for all $\boldsymbol{x} \in \mathbb{R}^d$.

Proof. We again utilize the inclusion-exclusion principle. First let $S_{j,d} := \sum_{J \subseteq \{1,\ldots,d\}:|J|=j} \mathbb{P}\big(\bigcap_{k \in J}\{X_k \leq x_k\}\big)$ for every $j = 0, 1, \ldots$, with the convention that $\bigcap_{k \in \varnothing}\{X_k \leq x_k\} := \Omega$. Then,

$$\bar{F}(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} > \boldsymbol{x}) = 1 - \mathbb{P}\left(\bigcup_{j=1}^{d}\{X_j \leq x_j\}\right)$$

$$\overset{\text{(inclusion-exclusion)}}{=} 1 - \sum_{j=1}^{d}(-1)^{j-1}S_{j,d} = 1 + \sum_{j=1}^{d}(-1)^j S_{j,d}$$

$$= \sum_{j=0}^{d}(-1)^j S_{j,d} = \sum_{j=0}^{d}(-1)^j \sum_{J \subseteq \{1,\ldots,d\}:|J|=j} \mathbb{P}\left(\bigcap_{k \in J}\{X_k \leq x_k\}\right)$$

$$= \sum_{j=0}^{d}(-1)^j \sum_{J \subseteq \{1,\ldots,d\}:|J|=j} F(\boldsymbol{\infty}_{J \leftarrow \boldsymbol{x}_J}) = \sum_{J \subseteq \{1,\ldots,d\}} (-1)^{|J|} F(\boldsymbol{\infty}_{J \leftarrow \boldsymbol{x}_J}).$$

$\square$

3.5 Stochastic Processes

3.5.1 Similar to [3.1.5]a, we actually have that X is a random sequence iff $X = (X_1, X_2, \ldots)$ where $X_1, X_2, \ldots$ are random variables. Such X is also known as a **discrete-time stochastic process**. [Note: Here $(X_1, X_2, \ldots)$ is actually referring to the sequence $\{X_n\}_{n \in \mathbb{N}}$.]

But apart from *discrete-time* stochastic process, you should have previously learnt also *continuous-time* stochastic processes, e.g., Brownian motion. While it appears that a continuous-time stochastic process can be obtained by just modifying the index set appropriately, the work it takes to justify the sensibility of such modification turns out to be highly non-trivial; it indeed requires a technical result known as *Kolmogorov's extension theorem*, which can be proved by the Carathéodory's extension theorem. In the following, we will state the theorem without proof.

3.5.2 **Kolmogorov's extension theorem.** The *Kolmogorov's extension theorem* allows us to <u>extend</u> from the discrete-time stochastic process $\{X_t\}_{t \in \mathbb{N}}$ to a very general stochastic process $\{\boldsymbol{X}_t\}_{t \in I}$, where $I \subseteq \mathbb{R}$ is an index set, and $\boldsymbol{X}_t : \Omega \to \mathbb{R}^d$ is a random vector for each $t \in I$.

Before stating the theorem, we first introduce a preliminary notion. Fix any $k \in \mathbb{N}$ and $t_1, \ldots, t_k \in I$ with $t_1 < \cdots < t_k$. Then the probability measure $\mathbb{P}_{t_1,\ldots,t_k}$ on $\mathcal{B}(\mathbb{R}^d)^k$ given by $\mathbb{P}_{t_1,\ldots,t_k}(\prod_{i=1}^{k} B_i) := \mathbb{P}(\boldsymbol{X}_{t_1} \in B_1, \ldots, \boldsymbol{X}_{t_k} \in B_k)$ for all $B_1, \ldots, B_k \in \mathcal{B}(\mathbb{R}^d)$ is said to be a **finite-dimensional distribution** of $\{\boldsymbol{X}_t\}_{t \in I}$.

Theorem 3.5.a (Kolmogorov's extension theorem)**.** For all $k \in \mathbb{N}$ and all $t_1, \ldots, t_k \in I$ with $t_1 < \cdots < t_k$, suppose that there is a probability measure $\mathbb{P}_{t_1,\ldots,t_k}$ on $\mathcal{B}(\mathbb{R}^d)^k$ satisfying:

- *(Permutation invariance)* $\mathbb{P}_{t_{\pi(1)},\ldots,t_{\pi(k)}}(\prod_{i=1}^{k} B_{\pi(i)}) = \mathbb{P}_{t_1,\ldots,t_k}(\prod_{i=1}^{k} B_i)$ for all $B_1, \ldots, B_k \in \mathcal{B}(\mathbb{R}^d)$ and all permutations π of $\{1, \ldots, k\}$.

- *(Compatibility)* $\mathbb{P}_{t_1,\ldots,t_{k+1}}(\prod_{i=1}^{k} B_k \times \mathbb{R}^d) = \mathbb{P}_{t_1,\ldots,t_{k+1}}(\prod_{i=1}^{k} B_i)$ for all $B_1, \ldots, B_k \in \mathcal{B}(\mathbb{R}^d)$.

Then there exists a probability space $((\mathbb{R}^d)^I, \mathcal{B}(\mathbb{R}^d)^I, \mathbb{P})$[8] on which the finite-dimensional distributions of $\{\boldsymbol{X}_t\}_{t \in I}$ are given by the $\mathbb{P}_{t_1,\ldots,t_k}$'s.

---

[8]The notation S^I denotes the set of all functions from I to S.

Proof. Omitted. □

In short, the Kolmogorov's extension theorem asserts that we can construct a general stochastic process $\{\boldsymbol{X}_t\}_{t \in I}$ by specifying its probabilistic behaviour at finitely many time points $t_1, \ldots, t_k$, through the probability measure $\mathbb{P}_{t_1,\ldots,t_k}$, as long as such specification is sufficiently "well-behaved". It provides the theoretical underpinning for studies of stochastic processes that are more general than just discrete-time stochastic processes. [Note: The famous *Brownian motion* is obtained by specifying the probability measures $\mathbb{P}_{t_1,\ldots,t_k}$ that correspond to multivariate normal distributions.]

3.6   Univariate Transforms

3.6.1   To close Section 3, we will introduce some results about *transformations* of random variables, which are helpful for performing *simulations* (see [4.2.9]).

3.6.2   **Quantile transform.** The first results suggests the distribution of a $U(0,1)$ random variable after being transformed by a <u>quantile</u> function F^{-1}.

Proposition 3.6.a (Quantile transform). Let F be a distribution function and $U \sim U(0,1)$. Then $F^{-1}(U) \sim F$.

Proof. Note that $\mathbb{P}\big(F^{-1}(U) \leq x\big) \overset{([2.5.10], \ F \text{ is right-continuous})}{=} \mathbb{P}(U \leq F(x)) = F(x)$ for all $x \in \mathbb{R}$.   □

3.6.3   **Probability transform.** Apart from the quantile transform, there is also a result that works another way round, but it requires a continuity assumption. The following lemma is used for proving the result.

Lemma 3.6.b (Strict increasingness of an univariate distribution function on its support). Let F be a distribution function on $\mathbb{R}$. Then, F is strictly increasing on its support $\text{supp}(F) = \{x \in \mathbb{R} : F(x) - F(x-h) > 0 \text{ for all } h > 0\}$, i.e., for all $x, y \in \text{supp}(F)$ with $x < y$, we have $F(x) < F(y)$.

Proof. Fix any $x, y \in \text{supp}(F)$ with $x < y$, and set $h = y - x > 0$. Then, by the definition of support we have $F(y) - F(x) = F(y) - F(y-h) > 0$, so the strict increasingness follows.   □

Proposition 3.6.c (Probability transform). If $X \sim F$ and F is continuous, then $F(X) \sim U(0,1)$.

Proof. Since F is continuous, the quantile function F^{-1} is strictly increasing on $[0,1]$ by [2.5.10]. Therefore, we have $F(X(\omega)) \leq u$ iff $F^{-1}(F(X(\omega))) \leq F^{-1}(u)$ for all $\omega \in \Omega$. Thus, we have

$$
\begin{aligned}
\mathbb{P}(F(X) \leq u) &= \mathbb{P}\big(F^{-1}(F(X)) \leq F^{-1}(u)\big) \\
&= \underbrace{\mathbb{P}\big(\{F^{-1}(F(X)) \leq F^{-1}(u)\} \cap \{X \in \text{supp}(F)\}\big)}_{\substack{[2.5.10] \\ = \\ (\text{Lemma } 3.6.\text{b})} \ \mathbb{P}(\{X \leq F^{-1}(u)\} \cap \{X \in \text{supp}(F)\})} + \underbrace{\mathbb{P}\big(\{F^{-1}(F(X)) \leq F^{-1}(u)\} \cap \{X \notin \text{supp}(F)\}\big)}_{=0 \text{ as } \mathbb{P}(X \in \text{supp}(F)) = 1} \\
&= \mathbb{P}\big(\{X \leq F^{-1}(u)\} \cap \{X \in \text{supp}(F)\}\big) + \underbrace{\mathbb{P}\big(\{X \leq F^{-1}(u)\} \cap \{X \notin \text{supp}(F)\}\big)}_{0} \\
&= \mathbb{P}\big(X \leq F^{-1}(u)\big) = F(F^{-1}(u)),
\end{aligned}
$$

which equals u for all $u \in \text{ran}(F) \cup \{0,1\} \overset{(F \text{ continuous})}{=} [0,1]$ by [2.5.10]. This implies that $F(X) \sim U(0,1)$.   □

4   Ordinary Conditional Probability, Independence, and Dependence

4.0.1   In Section 4, we will delve into notions that are *exclusive* to probability theory, namely ordinary conditional probability, independence, and dependence. The availability of these concepts is a main difference that separates the closely related *measure theory* and *probability theory.*

4.1   Ordinary Conditional Probability

4.1.1   The *ordinary conditional probability* is indeed just the conditional probability you have seen in your first probability course. We add the term "ordinary" here since we are going to introduce a much more general and abstract ☉ concept of conditioning (see Section 8).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $B \in \mathcal{F}$ be an event with $\mathbb{P}(B) > 0$. Then, the **(ordinary) conditional probability of A given B**, denoted by $\mathbb{P}(A|B)$, is given by $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ for all $A \in \mathcal{F}$. [Note: The requirement that $\mathbb{P}(B) > 0$ is indeed somewhat restrictive, e.g., it would not permit us to condition on $\{X = x\}$ where X is an absolutely continuous random variable, since $\mathbb{P}(X = x) = 0$ always. However, often one wants to do such kind of conditioning. This issue leads to the development of the more general and abstract concept of conditioning mentioned above.]

Let us verify that $\mathbb{P}(\cdot|B)$ is indeed a probability measure on $\mathcal{F}$ in the following:

Proof.

(1) The nonnegativity follows from the observation that both $\mathbb{P}(A \cap B)$ and $\mathbb{P}(B)$ are always nonnegative (the latter is always positive indeed).

(2) We have $\mathbb{P}(\Omega|B) = \mathbb{P}(\Omega \cap B)/\mathbb{P}(B) = \mathbb{P}(B)/\mathbb{P}(B) = 1$.

(3) Fix any pairwise disjoint $A_1, A_2, \ldots \in \mathcal{F}$. Then,

$$\mathbb{P}\left(\biguplus_{i=1}^{\infty} A_i \,\Big|\, B\right) = \frac{\mathbb{P}((\biguplus_{i=1}^{\infty} A_i) \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\biguplus_{i=1}^{\infty}(A_i \cap B))}{\mathbb{P}(B)} = \frac{\sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B)}{\mathbb{P}(B)}$$

$$= \sum_{i=1}^{\infty} \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B).$$

$\square$

4.1.2   **What would happen if $\mathbb{P}(B) = 0$?**   From the definition of the ordinary conditional probability above, it is clear that $\mathbb{P}(A \cap B)/\mathbb{P}(B)$ is *undefined* if $\mathbb{P}(B) = 0$. Consequently, we must not define the conditional probability $\mathbb{P}(A|B)$ using the ratio formula above. In this case, we would not define $\mathbb{P}(A|B)$ using a formulaic approach. Rather, we would just require it to satisfy $\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A \cap B)$ without actually providing an explicit definition of $\mathbb{P}(A|B)$.

With this equality, we can then obtain that $\mathbb{P}(A|B)\mathbb{P}(B) = 0$.

Proof. Note that $\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A \cap B) \overset{\text{(monotonicity)}}{\leq} \mathbb{P}(B) = 0$ and $\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A \cap B) \geq 0$. $\square$

Of course, merely having the equality $\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A \cap B)$ still leaves a lot of room for discussions on what conditioning on such zero-probability event B means. To deal with this somewhat complicated issue of "conditioning on zero-probability events", it is standard to utilize the machinery from Section 8.

4.1.3   **Two main results about ordinary conditional probability.**   You should have already learnt the following two properties about conditional probability in your first probability course, and they are stated and proved here again for reference. The first one is the *law of total probability*, which may be seen as a probability version of the *law of total measure*.

Theorem 4.1.a (Law of total probability)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $\{B_1, B_2, \dots\} \subseteq \mathcal{F}$ be a partition of Ω. Then

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

[Note: We have $\mathbb{P}(A|B_i)\mathbb{P}(B_i) = 0$ for every i where $\mathbb{P}(B_i) = 0$.]

Proof. By the law of total measure ([2.1.3]), we have

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

$\square$

The next one is the *Bayes' theorem*, which illustrates a key idea used in the *Bayesian methodology* in statistics.

Theorem 4.1.b (Bayes' theorem)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $A \in \mathcal{F}$ be an event with $\mathbb{P}(A) > 0$. Then

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

for all $B \in \mathcal{F}$. Furthermore, if $\{B_1, B_2, \dots\} \subseteq \mathcal{F}$ is a partition of Ω, then we have

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{\infty} \mathbb{P}(A|B_j)\mathbb{P}(B_j)}.$$

Proof. The second formula follows from the first by applying the law of total probability on the denominator $\mathbb{P}(A)$: $\mathbb{P}(A) = \sum_{j=1}^{\infty} \mathbb{P}(A|B_j)\mathbb{P}(B_j)$, so we will only prove the first:

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

$\square$

4.2 Independence

4.2.1 **Definitions of independence.** The conditional probability $\mathbb{P}(A|B)$ can be viewed as the probability assessment of the event A when knowing that the event B occurs. But if the event A does not in any way depend on the occurrence of event B, the probability assessment should then be the same with or without the knowledge of the occurrence of B: $\mathbb{P}(A|B) \overset{\text{(should be)}}{=} \mathbb{P}(A)$, which can be rewritten as $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. This motivates the definition of *independence* of two events, and we have more general definitions of independence in the following.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $I \subseteq \mathbb{R}$ be an arbitrary index set. Then:

(a) *(Independence of events)* The events $A_1, \dots, A_n \in \mathcal{F}$ are **independent** if $\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i)$ for <u>every</u> subset $I \subseteq \{1, \dots, n\}$. [⚠ Warning: Not just $I = \{1, \dots, n\}$!]
[Note: Here we have $\bigcap_{i \in \varnothing} A_i := \Omega$ and $\prod_{i \in \varnothing} \mathbb{P}(A_i) := 1$.]

(b) *(Independence of a collection of events)* A collection $\{A_i\}_{i \in I} \subseteq \mathcal{F}$ of events is **independent** if $A_{i_1}, \dots, A_{i_n}$ are independent for all $\{i_1, \dots, i_n\} \subseteq I$ with $n \in \mathbb{N}$ being arbitrary.
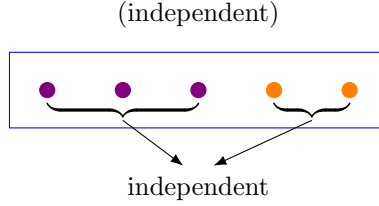
(c) *(Independence of families)* The families $\mathcal{A}_1, \dots, \mathcal{A}_n \subseteq \mathcal{F}$ of events are **independent** if $A_1, \dots, A_n$ are independent for all $A_1 \in \mathcal{A}_1, \dots, A_n \in \mathcal{A}_n$.

(d) *(Independence of a collection of families)* A collection $\{\mathcal{A}_i\}_{i \in I}$ of families of events, with $\mathcal{A}_i \subseteq \mathcal{F}$ for all $i \in I$, is **independent** if $\mathcal{A}_{i_1}, \dots, \mathcal{A}_{i_n}$ are independent for all $\{i_1, \dots, i_n\} \subseteq I$ with $n \in \mathbb{N}$ being arbitrary.

(e) *(Independence of measurable functions)* Measurable functions $X_1, \ldots, X_d$, all defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, are **independent**, if the families $\sigma(X_{i_1}), \ldots, \sigma(X_{i_d})$ are independent.

(f) *(Independence of a collection of measurable functions)* A collection $\{X_i\}_{i \in I}$ of measurable functions, all defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, is **independent** if the collection $\{\sigma(X_i)\}_{i \in I}$ is independent.

Particularly:

- If X_i is a random variable for every $i \in I$, then the collection is independent if $X_{i_1}^{-1}(B_{i_1}), \ldots, X_{i_n}^{-1}(B_{i_n})$ are independent $\forall\, B_{i_1}, \ldots, B_{i_n} \in \mathcal{B}(\mathbb{R}), \quad \forall\, \{i_1, \ldots, i_n\} \subseteq I, \quad \forall\, n \in \mathbb{N}$.

- If $X_i : \Omega \to \mathbb{R}^{d_i}$ is a random vector (written as $\boldsymbol{X}_i$ in the following) for every $i \in I$, then the collection is independent if $\boldsymbol{X}_{i_1}^{-1}(B_{i_1}), \ldots, \boldsymbol{X}_{i_n}^{-1}(B_{i_n})$ are independent $\forall\, B_{i_1} \in \mathcal{B}(\mathbb{R}^{d_1}), \ldots, B_{i_n} \in \mathcal{B}(\mathbb{R}^{d_n}), \quad \forall\, \{i_1, \ldots, i_n\} \subseteq I, \quad \forall\, n \in \mathbb{N}$.

4.2.2 Results about independence. The following results can simplify the process of verifying independence. Throughout, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

(a) *(Independence via generators)* If $\{\mathcal{A}_i\}_{i \in I}$ is a collection of independent π-systems on Ω, then the collection $\{\sigma(\mathcal{A}_i)\}_{i \in I}$ is independent.

(b) *(Independence of complements)* If $\{A_i\}_{i \in I} \subseteq \mathcal{F}$ is independent, then so is $\{A_i^c\}_{i \in I}$.

(c) *(Grouping lemma)* If $\{\mathcal{A}_i\}_{i \in I} \subseteq \mathcal{F}$ is a collection of independent π-systems, and $I = \biguplus_{j \in J} I_j$, then the collection $\{\sigma(\bigcup_{i \in I_j} \mathcal{A}_i)\}_{j \in J}$ is independent.

(independent)

independent

Proof.

(a) Fix any $n \in \mathbb{N}$, and any $\{i_1, \ldots, i_n\} \subseteq I$. Then, fix any $A_{i_k} \in \mathcal{A}_{i_k}$ for every $k = 2, \ldots, n$. Let $\mathcal{D} = \{A \in \mathcal{F} : \mathbb{P}(A \cap A_{i_2} \cap \cdots \cap A_{i_n}) = \mathbb{P}(A) \prod_{k=2}^{n} \mathbb{P}(A_{i_k})\}$.

Showing that $\mathcal{D}$ is a Dynkin system on Ω.

(1) By the definition of independence of a collection of families, we know that $\mathcal{A}_{i_2}, \ldots, \mathcal{A}_{i_n}$ are independent. Hence, we have

$$\mathbb{P}(\Omega \cap A_{i_2} \cap \cdots \cap A_{i_n}) = \mathbb{P}(A_{i_2} \cap \cdots \cap A_{i_n}) \overset{\text{(independence)}}{=} \prod_{k=2}^{n} \mathbb{P}(A_{i_k}) = \mathbb{P}(\Omega) \prod_{k=2}^{n} \mathbb{P}(A_{i_k}),$$

meaning that $\Omega \in \mathcal{D}$.

(2) Fix any $A \in \mathcal{D}$. Then we have

$$\mathbb{P}\left(A^c \cap \bigcap_{k=2}^{n} A_{i_k}\right) = \mathbb{P}\left(\bigcap_{k=2}^{n} A_{i_k} \setminus A\right) = \mathbb{P}\left(\bigcap_{k=2}^{n} A_{i_k} \setminus \left(A \cap \bigcap_{k=2}^{n} A_{i_k}\right)\right)$$

$$= \mathbb{P}\left(\bigcap_{k=2}^{n} A_{i_k}\right) - \mathbb{P}\left(A \cap \bigcap_{k=2}^{n} A_{i_k}\right)$$

$$\overset{(A \in \mathcal{D})}{=} \prod_{k=2}^{n} \mathbb{P}(A_{i_k}) - \mathbb{P}(A) \prod_{k=2}^{n} \mathbb{P}(A_{i_k}) = \mathbb{P}(A^c) \prod_{k=2}^{n} \mathbb{P}(A_{i_k}),$$

thus $A^c \in \mathcal{D}$.

(3) Fix any pairwise disjoint $A_1, A_2, \ldots \in \mathcal{D}$. Then we have

$$
\mathbb{P}\left(\left(\biguplus_{i=1}^{\infty} A_i\right) \cap \bigcap_{k=2}^{n} A_{i_k}\right) \overset{\text{(distributivity)}}{=} \mathbb{P}\left(\biguplus_{i=1}^{\infty}\left(A_i \cap \bigcap_{k=2}^{n} A_{i_k}\right)\right) = \sum_{i=1}^{\infty} \mathbb{P}\left(A_i \cap \bigcap_{k=2}^{n} A_{i_k}\right)
$$

$$
\overset{(A_1, A_2, \ldots \in \mathcal{D})}{=} \sum_{i=1}^{\infty}\left(\mathbb{P}(A_i) \prod_{k=2}^{n} \mathbb{P}(A_{i_k})\right) = \prod_{k=2}^{n} \mathbb{P}(A_{i_k}) \sum_{i=1}^{\infty} \mathbb{P}(A_i)
$$

$$
= \mathbb{P}\left(\biguplus_{i=1}^{\infty} A_i\right) \prod_{k=2}^{n} \mathbb{P}(A_{i_k}),
$$

so $\biguplus_{i=1}^{\infty} A_i \in \mathcal{D}$.

Applying Dynkin's π-λ theorem to show the independence of $\sigma(\mathcal{A}_{i_1}), \mathcal{A}_{i_2}, \ldots, \mathcal{A}_{i_n}$.
By assumption, $\mathcal{A}_{i_1}, \ldots, \mathcal{A}_{i_n}$ are independent. So, particularly, for all $A_{i_1} \in \mathcal{A}_{i_1}$, we have
$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}) = \mathbb{P}(A_{i_1}) \prod_{k=2}^{n} \mathbb{P}(A_{i_k})$, meaning that $A_{i_1} \in \mathcal{D}$. Therefore, we have
$\mathcal{A}_{i_1} \subseteq \mathcal{D}$. Since $\mathcal{A}_{i_1}$ is a π-system, by Dynkin's π-λ theorem, we have $\sigma(\mathcal{A}_{i_1}) \subseteq \mathcal{D}$. It follows that
$\sigma(\mathcal{A}_{i_1}), \mathcal{A}_{i_2}, \ldots, \mathcal{A}_{i_n}$ are independent.
Completing the proof by applying the argument repetitively. By repeating the argument
above (change "$\mathcal{A}_{i_1}$" $\to$ "$\mathcal{A}_{i_2}$" and "$\mathcal{A}_{i_2}, \ldots, \mathcal{A}_{i_n}$" $\to$ "$\sigma(\mathcal{A}_{i_1}), \mathcal{A}_{i_3}, \ldots, \mathcal{A}_{i_n}$", ...), we deduce that
$\sigma(\mathcal{A}_{i_1}), \sigma(\mathcal{A}_{i_2}), \ldots, \sigma(\mathcal{A}_{i_n})$ are independent.

(b) Assume $\{A_i\}_{i \in I} \subseteq \mathcal{F}$ is independent. Then, consider the family $\{\mathcal{A}_i\}_{i \in I}$ with $\mathcal{A}_i$ being the
singleton $\{A_i\}$ for all $i \in I$. Note that $\{\mathcal{A}_i\}_{i \in I}$ contains independent π-systems, so $\{\sigma(\mathcal{A}_i)\}_{i \in I}$ is
independent by [4.2.2]a. Since $A_i^c \in \sigma(\mathcal{A}_i) = \{\varnothing, A_i, A_i^c, \Omega\}$ for all $i \in I$, it follows that $\{A_i^c\}_{i \in I}$ is
independent.

(c) For every $j \in J$, let $\mathcal{A}_{I_j} := \{\bigcap_{i \in \widetilde{I}_j} A_i : A_i \in \mathcal{A}_i \ \forall i \in \widetilde{I}_j, \ \widetilde{I}_j \subseteq I_j, \ |\widetilde{I}_j| < \infty\}$, which is a
π-system by construction. Since $\{\mathcal{A}_i\}_{i \in I}$ is a collection of independent π-systems, by considering
the definition we deduce that $\{\mathcal{A}_{I_j}\}_{j \in J}$ is also a collection of independent π-systems. Then, by
[4.2.2]a, the collection $\{\sigma(\mathcal{A}_{I_j})\}_{j \in J}$ is independent.

Now, to complete the proof, we are going to show that $\sigma(\bigcup_{i \in I_j} \mathcal{A}_i) = \sigma(\mathcal{A}_{I_j})$:

- "$\subseteq$": Fix any $i \in I_j$. Note that

$$
\mathcal{A}_i = \{A_i : A_i \in \mathcal{A}_i\} = \left\{\bigcap_{i \in \widetilde{I}_j} A_i : A_i \in \mathcal{A}_i \ \forall i \in \widetilde{I}_j, \quad \widetilde{I}_j = \{i\}\right\} \subseteq \mathcal{A}_{I_j} \subseteq \sigma(\mathcal{A}_{I_j}),
$$

thus $\bigcup_{i \in I_j} \mathcal{A}_i \subseteq \sigma(\mathcal{A}_{I_j})$. Then, by the minimality we have $\sigma\left(\bigcup_{i \in I_j} \mathcal{A}_i\right) \subseteq \sigma(\mathcal{A}_{I_j})$.

- "$\supseteq$": Fix any $\widetilde{I}_j \subseteq I_j$ with $|\widetilde{I}_j| < \infty$, and any $A_i \in \mathcal{A}_i$ for all $i \in \widetilde{I}_j$. Then, we have
  $A_i \in \bigcup_{i \in I_j} \mathcal{A}_i \subseteq \sigma(\bigcup_{i \in I_j} \mathcal{A}_i)$ for all $i \in \widetilde{I}_j$. Hence, by the closedness under intersections for
  $\sigma(\bigcup_{i \in I_j} \mathcal{A}_i)$, we have $\bigcap_{i \in \widetilde{I}_j} A_i \in \sigma(\bigcup_{i \in I_j} \mathcal{A}_i)$.
  This implies that $\mathcal{A}_{I_j} \subseteq \sigma(\bigcup_{i \in I_j} \mathcal{A}_i)$, and hence $\sigma(\mathcal{A}_{I_j}) \subseteq \sigma(\bigcup_{i \in I_j} \mathcal{A}_i)$ by the minimality.

$\square$

4.2.3 **Zero-one laws.** Generally, a **zero-one law** refers to any result which asserts that the probability
of an event has to be <u>zero</u> or <u>one</u>, under certain conditions. Here we will go through some popular
zero-one laws, namely the *Borel-Cantelli lemmas* and the *Kolmogorov's zero-one law*.

Theorem 4.2.a (Borel-Cantelli lemmas)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F}$ be a
collection of events.

(a) *(First Borel-Cantelli lemma)* If $\sum_{i=1}^{\infty} \mathbb{P}(A_i) < \infty$, then $\mathbb{P}(\{\omega \in A_n \text{ io}\}) = 0$.

(b) *(Second Borel-Cantelli lemma)* If $\{A_i\}_{i \in \mathbb{N}}$ is independent and $\sum_{i=1}^{\infty} \mathbb{P}(A_i) = \infty$, then $\mathbb{P}(\{\omega \in A_n \text{ io}\}) = 1$.

Proof.

(a) Note that

$$\mathbb{P}(\{\omega \in A_n \text{ io}\}) \overset{[1.1.5]a}{=} \mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = \mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k\right) \overset{\text{(monotonicity)}}{\leq} \mathbb{P}\left(\bigcup_{k=n}^{\infty} A_k\right) \overset{(\sigma\text{-subadditivity})}{\leq} \sum_{k=n}^{\infty} \mathbb{P}(A_k).$$

By assumption, we have $\sum_{k=1}^{\infty} \mathbb{P}(A_k) = S < \infty$, and hence

$$\lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbb{P}(A_k) = \lim_{n\to\infty}\left(S - \sum_{k=1}^{n-1} \mathbb{P}(A_k)\right) = S - S = 0.$$

So, taking the limit $n \to \infty$ on the inequality above yields $\mathbb{P}(\{\omega \in A_n \text{ io}\}) \leq 0$. Together with the nonnegativity of $\mathbb{P}$, we must have $\mathbb{P}(\{\omega \in A_n \text{ io}\}) = 0$.

(b) With $\sum_{i=1}^{\infty} \mathbb{P}(A_i) = \infty$, we also have $\sum_{k=n}^{\infty} \mathbb{P}(A_k) = \infty$ for all $n \in \mathbb{N}$. Hence,

$$\mathbb{P}(\{\omega \in A_n \text{ io}\}) \overset{[1.1.5]a}{=} \mathbb{P}\left(\limsup_{n\to\infty} A_n\right) \overset{[1.1.5]b}{=} 1 - \mathbb{P}\left(\liminf_{n\to\infty} A_n^c\right)$$

$$\overset{[1.1.5]c}{=} 1 - \mathbb{P}\left(\lim_{n\to\infty}\inf_{k\geq n} A_k^c\right) = 1 - \mathbb{P}\left(\lim_{n\to\infty}\bigcap_{k=n}^{\infty} A_k^c\right)$$

$$\overset{\text{(continuity from below)}}{=} 1 - \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{k=n}^{\infty} A_k^c\right) \overset{[4.2.2]b}{\underset{(\{A_k\}_{k=n}^{\infty} \text{ are independent})}{=}} 1 - \lim_{n\to\infty} \prod_{k=n}^{\infty} \mathbb{P}(A_k^c)$$

$$\overset{(1-x\leq e^{-x})}{\geq} 1 - \lim_{n\to\infty} \prod_{k=n}^{\infty} e^{-\mathbb{P}(A_k)} = 1 - \lim_{n\to\infty} \underbrace{e^{-\sum_{k=n}^{\infty} \mathbb{P}(A_k)}}_{0} = 1.$$

$\square$

The next result will be the *Kolmogorov's zero-one law*. But before stating that, we first introduce a preliminary concept: *tail σ-algebra*. Let $(\Omega, \mathcal{F})$ be a measurable space and $\{\mathcal{A}_n\}_{n\in\mathbb{N}} \subseteq \mathcal{P}(\mathcal{F})$ be a collection of π-systems. Then, the **tail σ-algebra (of $\{\mathcal{A}_n\}_{n\in\mathbb{N}}$)** , denoted by $\mathcal{T}$ or $\mathcal{T}(\{\mathcal{A}_n\}_{n\in\mathbb{N}})$, is given by $\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(\bigcup_{k=n}^{\infty} \mathcal{A}_k) = \bigcap_{n=1}^{\infty} \sigma(\mathcal{A}_n, \mathcal{A}_{n+1}, \dots)$ (check that it is a σ-algebra on Ω!). Every event in $\mathcal{T}$ is said to be a **tail event**. [Intuition 💡: For the <u>tail</u> σ-algebra $\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(\mathcal{A}_n, \mathcal{A}_{n+1}, \dots)$, it is obtained by intersecting the σ-algebras generated by the "<u>tails</u>" from $\{\mathcal{A}_n\}_{n\in\mathbb{N}}$.]

Theorem 4.2.b (Kolmogorov's zero-one law)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{\mathcal{A}_n\}_{n\in\mathbb{N}} \subseteq \mathcal{P}(\mathcal{F})$ be a collection of independent π-systems, and $\mathcal{T}$ be the tail σ-algebra of $\{\mathcal{A}_n\}_{n\in\mathbb{N}}$. Then, $\mathbb{P}(A) = 0$ or 1 for all $A \in \mathcal{T}$.

Proof. Let $\mathcal{T}_n := \sigma(\bigcup_{k=1}^{n-1} \mathcal{A}_k)$ for every $n \in \mathbb{N}$, with $\mathcal{T}_\infty := \sigma(\bigcup_{k=1}^{\infty} \mathcal{A}_k)$.

Showing the independence of $\bigcup_{n=1}^{\infty} \mathcal{T}_n$ and $\mathcal{T}$. Since $\{\mathcal{A}_n\}$ is independent, by [4.2.2]c, $\mathcal{T}_n = \sigma(\bigcup_{k=1}^{n-1} \mathcal{A}_k)$ and $\sigma(\bigcup_{k=n}^{\infty} \mathcal{A}_k)$ are independent σ-algebras for all $n \in \mathbb{N}$.

Noting that $\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(\bigcup_{k=n}^{\infty} \mathcal{A}_k) \subseteq \sigma(\bigcup_{k=n}^{\infty} \mathcal{A}_k)$ for all $n \in \mathbb{N}$, we conclude that $\mathcal{T}_n$ and $\mathcal{T}$ are independent for all $n \in \mathbb{N}$. Thus, by considering the definition we have the independence of $\bigcup_{n=1}^{\infty} \mathcal{T}_n$ and $\mathcal{T}$.

Showing the independence of $\sigma(\bigcup_{n=1}^{\infty} \mathcal{T}_n)$ and $\mathcal{T}$. Next, we will show that $\bigcup_{n=1}^{\infty} \mathcal{T}_n$ is a π-system: Fix any $A_1, A_2 \in \bigcup_{n=1}^{\infty} \mathcal{T}_n$. Then we know $A_1 \in \mathcal{T}_{n_1}$ and $A_2 \in \mathcal{T}_{n_2}$ for some $n_1, n_2 \in \mathbb{N}$. With $\mathcal{T}_n \nearrow$, we then have $A_1, A_2 \in \mathcal{T}_{\max\{n_1, n_2\}}$. By the closedness under intersections for for the σ-algebra $\mathcal{T}_{\max\{n_1, n_2\}}$, $A_1 \cap A_2 \in \mathcal{T}_{\max\{n_1, n_2\}} \subseteq \bigcup_{n=1}^{\infty} \mathcal{T}_n$.

Of course, $\mathcal{T}$ is also a π-system, since it is closed under intersections. Hence, by [4.2.2]a, $\sigma(\bigcup_{n=1}^{\infty} \mathcal{T}_n)$ and $\sigma(\mathcal{T})$ are independent, which implies the independence of $\sigma(\bigcup_{n=1}^{\infty} \mathcal{T}_n)$ and $\mathcal{T} \subseteq \sigma(\mathcal{T})$.

Showing the independence of $\mathcal{T}_\infty$ and $\mathcal{T}$. We first note the following result: $\mathcal{A} \subseteq \mathcal{B} \implies \sigma(\mathcal{A}) \subseteq \sigma(\mathcal{B})$. This holds because with $\mathcal{A} \subseteq \mathcal{B} \subseteq \sigma(\mathcal{B})$, by the minimality we would have $\sigma(\mathcal{A}) \subseteq \sigma(\mathcal{B})$. Using this result, we will show that $\sigma(\bigcup_{n=1}^{\infty} \mathcal{T}_n) = \mathcal{T}_\infty$, and hence establish the independence of $\mathcal{T}_\infty$ and $\mathcal{T}$.

- "$\subseteq$": Since $\mathcal{T}_n \subseteq \mathcal{T}_\infty$ for every $n \in \mathbb{N}$, we have $\bigcup_{n=1}^\infty \mathcal{T}_n \subseteq \mathcal{T}_\infty$, thus $\sigma(\bigcup_{n=1}^\infty \mathcal{T}_n) \subseteq \mathcal{T}_\infty$.
- "$\supseteq$": For every $n \in \mathbb{N}$, we have $\mathcal{A}_n \subseteq \sigma(\bigcup_{k=1}^n \mathcal{A}_n) = \mathcal{T}_{n+1}$. Thus, $\bigcup_{n=1}^\infty \mathcal{A}_n \subseteq \bigcup_{n=1}^\infty \mathcal{T}_n$, which implies that $\mathcal{T}_\infty = \sigma(\bigcup_{n=1}^\infty \mathcal{A}_n) \subseteq \sigma(\bigcup_{n=1}^\infty \mathcal{T}_n)$ by the result above.

Showing that every $A \in \mathcal{T}$ is independent of itself. Note that $\mathcal{T} = \bigcap_{n=1}^\infty \sigma(\bigcup_{k=n}^\infty \mathcal{A}_k) \subseteq \overset{(n=1)}{\sigma(\bigcup_{k=1}^\infty \mathcal{A}_k)} = \mathcal{T}_\infty$. So, the independence of $\mathcal{T}_\infty$ and $\mathcal{T}$ implies that every $\mathcal{A} \in \mathcal{T}$ is independent of itself, which means that $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A) = \mathbb{P}(A)^2$, or $\mathbb{P}(A) = 0$ or 1, for every $A \in \mathcal{T}$. $\qquad\square$

<u>Remarks</u>:

- *(Application of Kolmogorov's zero-one law to independent events)* By taking $\mathcal{A}_n = \{A_n\}$ for every $n \in \mathbb{N}$, we know that given a collection $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$ of independent events, we have $\mathbb{P}(A) = 0$ or 1 for every $A \in \mathcal{T} = \bigcap_{n=1}^\infty \sigma(\{A_n, A_{n+1}, \dots\})$.
- *(Application of Kolmogorov's zero-one law to independent random variables)* Suppose we are given a collection $\{X_n\}_{n\in\mathbb{N}}$ of independent random variables, all defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Then, by taking $\mathcal{A}_n = \sigma(X_n)$ for every $n \in \mathbb{N}$, we have $\mathbb{P}(A) = 0$ or 1 for every $A \in \mathcal{T} = \bigcap_{n=1}^\infty \sigma(\sigma(X_n), \sigma(X_{n+1}), \dots)$. Particularly, we can indeed write $\sigma(\sigma(X_n), \sigma(X_{n+1}), \dots) = \sigma(X_n, X_{n+1}, \dots)$, where the latter is the σ-algebra generated by $\{X_k\}_{k=n}^\infty$, defined by $\sigma(\bigcup_{k=n}^\infty \sigma(X_k))$, which is the same as $\sigma(\sigma(X_n), \sigma(X_{n+1}), \dots)$.

4.2.4 Characterization of independence of random variables. By definition, to verify the independence of a collection $\{X_i\}_{i\in I}$ of random variables, we need to check that $X_{i_1}^{-1}(B_{i_1}), \dots, X_{i_n}^{-1}(B_{i_n})$ are independent $\forall\, B_{i_1}, \dots, B_{i_n} \in \mathcal{B}(\mathbb{R}), \quad \forall\, \{i_1, \dots, i_n\} \subseteq I, \quad \forall\, n \in \mathbb{N}$, which can take quite a lot of work. The following result provides us an alternative and often more convenient route for showing such independence:

Theorem 4.2.c (Characterization of independence of random variables). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $X_i : \Omega \to \mathbb{R}$ be a random variable for every $i \in I$. Then, the collection $\{X_i : i \in I\}$ is independent iff for all $n \in \mathbb{N}$ and $\{i_1, \dots, i_n\} \subseteq I$, $\mathbb{P}(X_{i_1} \leq x_{i_1}, \dots, X_{i_n} \leq x_{i_n}) = \prod_{k=1}^n \mathbb{P}(X_{i_k} \leq x_{i_k})$.

Proof. "$\Rightarrow$": Assume $\{X_i : i \in I\}$ is independent. Then by definition, $\{\sigma(X_i)\}_{i\in I}$ is independent. This implies that $\forall\, i_1, \dots, i_n \in \mathbb{R} \quad \forall\, n \in \mathbb{N}$, the events $X_{i_1}^{-1}((-\infty, x_{i_1}]), \dots, X_{i_n}^{-1}((-\infty, x_{i_n}])$ are independent, and thus

$$\mathbb{P}(X_{i_1} \leq x_{i_1}, \dots, X_{i_n} \leq x_{i_n}) = \mathbb{P}\left(\bigcap_{k=1}^n X_{i_k}^{-1}((-\infty, x_{i_k}])\right) = \prod_{k=1}^n \mathbb{P}\left(X_{i_k}^{-1}((-\infty, x_{i_k}])\right) = \prod_{k=1}^n \mathbb{P}(X_{i_k} \leq x_{i_k}).$$

"$\Leftarrow$": For every $k = 1, \dots, n$, let $\mathcal{A}_{i_k} := \{X_{i_k}^{-1}((-\infty, x])\}_{x\in\mathbb{R}}$. By the property of preservation of intersection for preimages ([1.1.9]), $\mathcal{A}_{i_k}$'s are all π-systems. By assumption, we know $\mathcal{A}_{i_1}, \dots, \mathcal{A}_{i_n}$ are independent, thus $\sigma(\mathcal{A}_{i_1}), \dots, \sigma(\mathcal{A}_{i_n})$ are independent by [4.2.2]a.

Noting that for every $k = 1, \dots, n$,

$$\sigma(\mathcal{A}_{i_k}) = \sigma(\{X_{i_k}^{-1}((-\infty, x])\}_{x\in\mathbb{R}}) \overset{\text{(notation)}}{=} \sigma(X_{i_k}^{-1}(\{(-\infty, x]\}_{x\in\mathbb{R}}))$$
$$\overset{\text{(Lemma 3.1.a)}}{=} X_{i_k}^{-1}(\sigma(\{(-\infty, x]\}_{x\in\mathbb{R}})) \overset{\text{(Proposition 1.4.g)}}{=} X_{i_k}^{-1}(\mathcal{B}(\mathbb{R})) = \sigma(X_{i_k}),$$

we conclude that $\sigma(X_{i_1}), \dots, \sigma(X_{i_n})$ are independent. Hence, by definition, $\{X_i\}_{i\in I}$ is independent. $\quad\square$

<u>Remarks</u>:

- *(Special case for finitely many random variables — factorization of distribution function)* With $I = \{1, \dots, d\}$, random variables $X_1, \dots, X_d$ with $\boldsymbol{X} = (X_1, \dots, X_d) \sim F$ and margins $F_1, \dots, F_d$ are independent iff

$$F(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}) = \prod_{j=1}^d \mathbb{P}(X_j \leq x_j) = \prod_{j=1}^d F_j(x_j), \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^d.$$

  (This condition indeed implies that $\mathbb{P}(X_{i_1} \leq x_{i_1}, \dots, X_{i_n} \leq x_{i_n}) = \prod_{k=1}^n \mathbb{P}(X_{i_k} \leq x_{i_k})$ for every $\{i_1, \dots, i_n\} \subseteq I$, by letting some components of $\boldsymbol{x}$ to ∞.)

- *(Special case for finitely many random variables — factorization of density function)* Assuming further that F has continuous partial derivatives with respect to each component in $\text{supp}(F)$, we know that F is absolutely continuous with density given by $f(\boldsymbol{x}) = \frac{\partial^d}{\partial x_d \cdots \partial x_1} F(\boldsymbol{x})$. In such case, $X_1, \ldots, X_d$ are independent iff

$$f(\boldsymbol{x}) = \frac{\partial^d}{\partial x_d \cdots \partial x_1} F(\boldsymbol{x}) = \prod_{j=1}^{d} \frac{\partial}{\partial x_j} F_j(x_j) = \prod_{j=1}^{d} f_j(x_j), \quad \text{for all } \boldsymbol{x} \in \text{supp}(F).$$

  (For discrete $\boldsymbol{X} = (X_1, \ldots, X_d)$, we have similarly that $X_1, \ldots, X_d$ are independent iff the joint mass function is $f(\boldsymbol{x}) = \prod_{j=1}^{d} f_j(x_j)$ for all $\boldsymbol{x} \in \prod_{j=1}^{d} \text{supp}(F_j)$.)

- *(Survival function version of the result)* We indeed also have a "survival function version" of the result: $\{X_i : i \in I\}$ is independent iff for all $n \in \mathbb{N}$ and $\{i_1, \ldots, i_n\} \subseteq I$, $\mathbb{P}(X_{i_1} > x_{i_1}, \ldots, X_{i_n} > x_{i_n}) = \prod_{k=1}^{n} \mathbb{P}(X_{i_k} > x_{i_k})$. This can be shown by adapting the proof slightly: change $(-\infty, x_{i_k}] \to (x_{i_k}, \infty)$ for every k and $(-\infty, x] \to (x, \infty)$.

4.2.5 **Functions of independent random variables are also independent.** An important result regarding the independence of random variable is that *functions of (disjoint sets of) independent random variables are also independent*, which can simplify the checking of independence of random variables a lot. The precise formulation of this statement is as follows.

Theorem 4.2.d (Functions of independent random variables are also independent). If $\{X_{ij}\}_{j=1,i=1}^{d_i,\infty}$ is a collection of independent random variables, and $h_i : \mathbb{R}^{d_i} \to \mathbb{R}$ is $(\mathcal{B}(\mathbb{R}^{d_i}), \mathcal{B}(\mathbb{R}))$-measurable for every $i \in \mathbb{N}$, then the collection $\{Y_i\}$ of random variables, with $Y_i := h_i(X_{i1}, \ldots, X_{id_i})$ for every $i \in \mathbb{N}$, is independent.

Proof. By assumption, the collection $\{\sigma(X_{ij})\}_{j=1,i=1}^{d_i,\infty}$ of σ-algebras is independent, so by [4.2.2]c the collection $\left\{\sigma\left(\bigcup_{j=1}^{d_i} \sigma(X_{ij})\right)\right\}_{i \in \mathbb{N}}$ is independent.

Showing that $\mathcal{F}_i$ is a σ-algebra. Fix any $i \in \mathbb{N}$. Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id_i})$. Then, we will show that

$$\mathcal{F}_i := \left\{ B \in \mathcal{B}(\mathbb{R}^{d_i}) : \boldsymbol{X}_i^{-1}(B) \in \sigma\left(\bigcup_{j=1}^{d_i} \sigma(X_{ij})\right) \right\}$$

is a σ-algebra:

(1) For $\varnothing \in \mathcal{B}(\mathbb{R}^{d_i})$, we have $\boldsymbol{X}_i^{-1}(\varnothing) = \varnothing \in \sigma\left(\bigcup_{j=1}^{d_i} \sigma(X_{ij})\right)$, thus $\varnothing \in \mathcal{F}_i$.

(2) Fix any $B \in \mathcal{F}_i$. Then, by the closedness under complements, we have $B^c \in \mathcal{B}(\mathbb{R}^{d_i})$, and also

$$\boldsymbol{X}_i^{-1}(B^c) \overset{[1.1.9]}{=} (\boldsymbol{X}_i^{-1}(B))^c \in \sigma\left(\bigcup_{j=1}^{d_i} \sigma(X_{ij})\right).$$

Hence, $B^c \in \mathcal{F}_i$.

(3) Fix any $B_1, B_2, \ldots, \in \mathcal{F}_i$. Then by the closedness under countable unions, we have $\bigcup_{n=1}^{\infty} B_n \in \mathcal{B}(\mathbb{R}^{d_i})$ and also

$$\boldsymbol{X}_i^{-1}\left(\bigcup_{n=1}^{\infty} B_n\right) \overset{[1.1.9]}{=} \bigcup_{n=1}^{\infty} \boldsymbol{X}_i^{-1}(B_n) \in \sigma\left(\bigcup_{j=1}^{d_i} \sigma(X_{ij})\right).$$

Showing that $\mathcal{F}_i = \mathcal{B}(\mathbb{R}^{d_i})$. By [1.4.15]a, $\mathcal{B}(\mathbb{R}^{d_i})$ is the product σ-algebra $\bigotimes_{j=1}^{d_i} \mathcal{B}(\mathbb{R}) = \sigma(\{\prod_{j=1}^{d_i} B_j : B_j \in \mathcal{B}(\mathbb{R}) \; \forall j = 1, \ldots, d_i\})$. Using this, we will then show that $\mathcal{F}_i = \mathcal{B}(\mathbb{R}^{d_i})$:

- "$\subseteq$": It is immediate by the definition of $\mathcal{F}_i$.

- "⊇": Fix any $B = \prod_{j=1}^{d_i} B_j$ where $B_j \in \mathcal{B}(\mathbb{R})$ for every $j = 1, \ldots, d_i$. Then, $\boldsymbol{X}_i^{-1}(B) = \bigcap_{j=1}^{d_i} X_{ij}^{-1}(B_j)$. Since $X_{ij}^{-1}(B_j) \in X_{ij}^{-1}(\mathcal{B}(\mathbb{R})) = \sigma(X_{ij}) \subseteq \sigma(\bigcup_{j=1}^{d_i} \sigma(X_{ij}))$ for every $j = 1, \ldots, d_i$, by the closedness under intersections we have $\boldsymbol{X}_i^{-1}(B) \in \sigma(\bigcup_{j=1}^{d_i} \sigma(X_{ij}))$. By definition of $\mathcal{F}_i$, this means $B \in \mathcal{F}_i$.

  Hence, we have $\{\prod_{j=1}^{d_i} B_j : B_j \in \mathcal{B}(\mathbb{R}) \ \forall j = 1, \ldots, d_i\} \subseteq \mathcal{F}_i$. By the minimality, we then have $\mathcal{B}(\mathbb{R}^{d_i}) = \bigotimes_{j=1}^{d_i} \mathcal{B}(\mathbb{R}) \subseteq \mathcal{F}_i$.

Showing that Y_i is a random variable (with $\mathcal{F} = \sigma\left(\bigcup_{j=1}^{d_i} \sigma(X_{ij})\right)$). Since $\boldsymbol{X}_i^{-1}(\mathcal{B}(\mathbb{R}^{d_i})) = \boldsymbol{X}_i^{-1}(\mathcal{F}_i) \subseteq \sigma\left(\bigcup_{j=1}^{d_i} \sigma(X_{ij})\right)$, we know $\boldsymbol{X}_i$ is $\left(\sigma\left(\bigcup_{j=1}^{d_i} \sigma(X_{ij})\right), \mathcal{B}(\mathbb{R}^{d_i})\right)$-measurable. So, by [3.1.3]b, $Y_i = h_i \circ \boldsymbol{X}_i$ is $\left(\sigma\left(\bigcup_{j=1}^{d_i} \sigma(X_{ij})\right), \mathcal{B}(\mathbb{R})\right)$-measurable, implying that Y_i is a random variable.

Showing the independence of $\{Y_i\}_{i \in \mathbb{N}}$. For every $i \in \mathbb{N}$, we have

$$\sigma(Y_i) = Y_i^{-1}(\mathcal{B}(\mathbb{R})) \overset{\text{(similar to the proof of [3.1.3]b)}}{=} \boldsymbol{X}_i^{-1}\left(h_i^{-1}(\mathcal{B}(\mathbb{R}))\right) \overset{[1.1.9]}{\subseteq} \boldsymbol{X}_i^{-1}(\mathcal{B}(\mathbb{R}^{d_i})) \subseteq \sigma\left(\bigcup_{j=1}^{d_i} \sigma(X_{ij})\right).$$

As shown previously, the collection $\left\{\sigma\left(\bigcup_{j=1}^{d_i} \sigma(X_{ij})\right)\right\}_{i \in \mathbb{N}}$ is independent. Hence, this subset inclusion implies that the collection $\{\sigma(Y_i)\}_{i \in \mathbb{N}}$ is independent also, which means that $\{Y_i\}_{i \in \mathbb{N}}$ is independent. $\quad\square$

4.2.6 **Mathematical construction of independent random variables.** In probability theory and statistics, we often work with independent random variables. Particularly, in statistics we talk about *independent and identically distributed* (iid) random variables a lot. An implicit assumption made here is that we can always construct *independent* random variables, regardless of their underlying (marginal) distributions. While this is quite reasonable and intuitive, it does take some work to mathematically show that such construction is possible:

Proposition 4.2.e (Construction of independent random variables). Let $F_1, \ldots, F_d$ be arbitrary distribution functions on $\mathbb{R}$. Then there exist a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and independent random variables $X_1 \sim F_1, \ldots, X_d \sim F_d$ defined on this probability space.

Proof. Since $F_1, \ldots, F_d$ are distribution functions, $(\Omega, \mathcal{F}, \mathbb{P}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \prod_{j=1}^d \lambda_{F_j})$ is a probability space, by Theorem 2.5.a. Now fix any $j = 1, \ldots, d$. Since the projection π_j is continuous, and thus measurable by [3.1.3]c, the function $X_j : \Omega \to \mathbb{R}$ given by $X_j(\boldsymbol{\omega}) := \pi_j(\boldsymbol{\omega}) = \omega_j \ \ \forall \ \boldsymbol{\omega} \in \mathbb{R}^d$ is a random variable. Hence, by [3.1.5]a, the function $\boldsymbol{X} : \Omega \to \mathbb{R}^d$ given by $\boldsymbol{X}(\boldsymbol{\omega}) := (X_1(\boldsymbol{\omega}), \ldots, X_d(\boldsymbol{\omega})) = \boldsymbol{\omega} \ \ \forall \ \boldsymbol{\omega} \in \mathbb{R}^d$ is a random vector.

Now, since we have

$$F(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}) = \mathbb{P}(\{\boldsymbol{\omega} \in \Omega : \boldsymbol{\omega} \leq \boldsymbol{x}\}) = \mathbb{P}\left(\bigcap_{j=1}^d \{\omega_j \in \mathbb{R} : \omega_j \leq x_j\}\right)$$

$$= \mathbb{P}\left(\prod_{j=1}^d (-\infty, x_j]\right) \overset{(\mathbb{P} = \prod_{j=1}^d \lambda_{F_j})}{=} \prod_{j=1}^d \underbrace{\lambda_{F_j}((-\infty, x_j])}_{\Delta_{(-\infty, x_j]} F_j = F_j(x_j)} = \prod_{j=1}^d F_j(x_j)$$

for all $\boldsymbol{x} \in \mathbb{R}^d$, we conclude by Theorem 4.2.c that $X_1, \ldots, X_d$ are independent.

Furthermore, letting $x_k \to \infty$ for every $k \neq j$ on the equation above gives $\mathbb{P}(X_j \leq x_j) = F_j(x_j)$. Thus, $X_j \sim F_j$ for every $j = 1, \ldots, d$. $\quad\square$

Remarks:

- *(Independent and identically distributed random variables)* If $F_1 = \cdots = F_d = F$, then $X_1, \ldots, X_d$ are said to be **independent and identically distributed (iid)** from F, or **independent copies of** $X \sim F$. This is denoted by $X_1, \ldots, X_d \overset{\text{iid}}{\sim} F$.

- *(Infinite sequence of independent random variables)* Using *Kolmogorov's extension theorem*, we can also construct infinite sequences of independent random variables.

4.2.7 **Mixture distributions.** In actuarial science, apart from working with independent random variables, often we deal with *mixture distributions*, e.g., aggregate loss $S := \sum_{i=1}^{N} X_i$, where each X_i represents a loss amount, and N is the (random) number of losses, taking values in $\mathbb{N}$. Here S possesses a mixture distribution: a distribution that is created by "mixing" two sources of randomness, one from X_i's and another from N. To be more precise, S should be interpreted as the sum of random variables $\sum_{i=1}^{n} X_i$ *given* $N = n$, for every $n \in \mathbb{N}$; from here the relationship between mixture distribution and *conditioning* becomes more transparent.

In general, let $I \sim F_I$ be a random variable with $\mathrm{supp}(F_I) = \mathbb{N}$, and $\boldsymbol{X}_i \sim F_i$ for every $i \in \mathbb{N}$. Suppose that $I, \boldsymbol{X}_1, \boldsymbol{X}_2, \dots$ are independent. Then, $\boldsymbol{X} = \boldsymbol{X}_I$ (interpreted as $\boldsymbol{X}_i$ given $I = i$, or more explicitly, $\boldsymbol{X}_I(\omega) := \boldsymbol{X}_{I(\omega)}(\omega)$ for all $\omega \in \Omega$), follows the **mixture distribution** of $F_1, F_2, \dots$ with respect to F_I.

[Note: It is indeed possible to define mixture distribution in a similar way when $\mathrm{supp}(F_I)$ is uncountable (e.g., $I \sim U(0,1)$), but some technicalities are involved so we would not discuss it in details here.]

4.2.8 **Distribution function for mixture distribution.** For $\boldsymbol{X}$ following the mixture distribution of $F_1, F_2, \dots$ with respect to F_I, its distribution function admits a "weighted average" form:

$$F(\boldsymbol{x}) = \sum_{i=1}^{\infty} p_i F_i(\boldsymbol{x}), \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^d$$

where $p_i := \mathbb{P}(I = i)$ for every $i \in \mathbb{N}$.

Proof. For all $\boldsymbol{x} \in \mathbb{R}^d$, we have

$$F(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}) = \mathbb{P}(\boldsymbol{X}_I \leq \boldsymbol{x}) = \sum_{i=1}^{\infty} p_i \mathbb{P}(\boldsymbol{X}_I \leq \boldsymbol{x} | I = i) \overset{\text{(independence)}}{=} \sum_{i=1}^{\infty} p_i \mathbb{P}(\boldsymbol{X}_i \leq \boldsymbol{x}) = \sum_{i=1}^{\infty} p_i F_i(\boldsymbol{x}).$$

$\square$

4.2.9 **Sampling.** Recall from Section 3.6 that some results concerning transformations of random variables are discussed, and there we mentioned that they are helpful for performing simulations. After learning about the concept of *independence*, we can study more details about it.

A collection of iid random vectors $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \overset{\text{iid}}{\sim} F$ is called a **random sample** from F, which is often encountered in statistical inference. Besides, for conducting simulation studies, we would need to generate *realizations* (something observable ◉) of $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. One famous technique is known as the *inversion method*, which utilizes the quantile transform in Proposition 3.6.a.

The **inversion method** for sampling from a random variable $X \sim F$ is to first obtain (independent) realizations from $U \sim U(0,1)$ (many algorithms are available for this), and then evaluate the quantile function F^{-1} on each of them. This is theoretically justified because $F^{-1}(U) \sim F$, so the resulting values $F^{-1}(u)$'s can be seen as realizations of a random vector with the distribution function F.

To sample from the iid random variables $X_1, \dots, X_n$, we can use the inversion method and consider $F^{-1}(U_1), \dots, F^{-1}(U_n) \overset{\text{iid}}{\sim} F$ with $U_1, \dots, U_n \overset{\text{iid}}{\sim} U(0,1)$. Realizations of each U_i are first obtained, and then transformed by the quantile function F^{-1}.

[Note: There are also some other more efficient methods for sampling, often equipped with a *variance reduction* component, which can increase the "precision" in the sampling (details omitted here).]

4.2.10 **Empirical distribution functions.** Given a random sample $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, the **empirical distribution function** F_n is defined by $F_n(\boldsymbol{x}) = \sum_{i=1}^{n} \frac{1}{n} \mathbf{1}_{[\boldsymbol{X}_i, \infty)}(\boldsymbol{x})$, which jumps by $1/n$ at each $\boldsymbol{X}_i$. Here $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are considered to be empirical data, and the empirical distribution function F_n can be viewed as an estimation of the true underlying distribution function that generate such $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. We also note that the empirical distribution function indeed corresponds to a mixture distribution function, with $p_i = 1/n$ and $F_i(\boldsymbol{x}) = \mathbf{1}_{[\boldsymbol{X}_i, \infty)}(\boldsymbol{x})$ for every $i = 1, \dots, n$.

Sampling can also be done on an empirical distribution function. Let $U \sim \mathrm{U}(0,1)$ and we write $I = \lceil nU \rceil$, which follows the discrete uniform distribution with support $\{1, \ldots, n\}$. Then, based on the mixture distribution we have $\boldsymbol{X}_{\lceil nU \rceil} \sim F_n$. This hints a method for sampling from the empirical distribution function F_n as follows. Given $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$:

(1) Obtain realizations of U.

(2) Based on them, obtain realizations of $I = \lceil nU \rceil$.

(3) For each realization of I, return $\boldsymbol{X}_i$ if the realization is i.

4.3 Dependence

4.3.1 After discussing *independence* in Section 4.2, it is natural to consider another related concept: *dependence*. While the concept of independence can be mathematically defined, the idea of "dependence" is much more unclear: When we speak of "dependent" random variables, often we are unclear about how they actually depend on each other. In Section 4.3, we will study a way to describe the dependence clearly and mathematically, through the usage of *copula*.

4.3.2 **Copula and its characterization.** A (*d*-**dimensional) copula** C is a distribution function on $[0,1]^d$ with $\mathrm{U}(0,1)$ margins (oftentimes, we consider the case with $d \geq 2$ so that we can meaningfully talk about "dependence"). This can be viewed as a "probabilistic" definition of copula, and the following provides an "analytic" characterization of it:

Proposition 4.3.a. A function $C : [0,1]^d \to [0,1]$ is a *d*-dimensional copula iff

(1) *(Groundedness)* $C(\boldsymbol{u}) = 0$ if $u_j = 0$ for some $j = 1, \ldots, d$.

(2) *($\mathrm{U}(0,1)$ margins)* $C(1, \ldots, 1, u_j, 1, \ldots, 1) = u_j$ for all $u_j \in [0,1]$, for all $j = 1, \ldots, d$.

(3) *(d-increasingness)* $\Delta_{(\boldsymbol{a}, \boldsymbol{b})} C \geq 0$ for all $_d 0 \leq \boldsymbol{a} \leq \boldsymbol{b} \leq {}_d 1$.

Proof. "$\Rightarrow$": It follows by the definition of distribution function.

"$\Leftarrow$": The groundedness and *d*-increasingness properties for being a distribution function are immediate; we only need to establish the normalization and right-continuous properties for C to be a distribution function (after being extended in the way mentioned in [2.5.4]):

- *Normalization:* We have $\lim_{\boldsymbol{x} \to \infty} C(\boldsymbol{x}) = C(1, \ldots, 1) \overset{(\mathrm{U}(0,1) \text{ margins})}{=} 1$.
- *Right-continuity:* Applying Lemma 2.5.b to the function C with $\mathrm{U}(0,1)$ margins here, we have $|C(\boldsymbol{b}) - C(\boldsymbol{a})| \leq \sum_{j=1}^d |b_j - a_j|$, which implies the uniform continuity of C, and hence its right-continuity.

The function C also has $\mathrm{U}(0,1)$ margins by assumption, so C is a *d*-dimensional copula. $\qquad \square$

4.3.3 **Sklar's theorem.** From the definition of copula alone, it is not too clear why it can serve for describing dependence. Mathematically, this function of copula is established by the *Sklar's theorem*, which is considered to be the foundational result that drives the study of copula:

Theorem 4.3.b (Sklar's theorem).

(a) *(Decomposition)* Given any distribution function F with margins $F_1, \ldots, F_d$, there exists a copula C such that
$$F(\boldsymbol{x}) = C(F_1(x_1), \ldots, F_d(x_d)) \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^d. \tag{7}$$
Furthermore, C is uniquely defined on $\prod_{j=1}^d \mathrm{ran}(F_j)$, and is given by $C(\boldsymbol{u}) = F\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right)$ for all $\boldsymbol{u} \in \prod_{j=1}^d \mathrm{ran}(F_j)$, where $\mathrm{ran}(F_j)$ denotes the *range* of F_j, namely $\{F_j(x) : x \in \mathbb{R}\}$ for all $j = 1, \ldots, d$.

(b) *(Combination)* Given any copula C and univariate distribution functions $F_1, \ldots, F_d$, the function F as defined in Equation (7) is a distribution function with margins $F_1, \ldots, F_d$.

Proof. We shall only prove for the case where $F_1, \ldots, F_d$ are continuous.

(a) Let $\boldsymbol{X} = (X_1, \ldots, X_d) \sim F$ and $\boldsymbol{U} = (U_1, \ldots, U_d) := (F_1(X_1), \ldots, F_d(X_d))$. By probability transform, we know $\boldsymbol{U}$ has $\mathrm{U}(0,1)$ margins, so the distribution function of U is indeed a copula, which we denote by C.

By Lemma 3.6.b, we know F_j is strictly increasing on $\mathrm{supp}(F_j)$, so by [2.5.10] we have $X_j = F_j^{-1}(F_j(X_j)) = F_j^{-1}(U_j)$ on $\mathrm{supp}(F_j)$ for all $j = 1, \ldots, d$. Therefore, using a similar argument as in the proof Proposition 3.6.c, the distribution function F of X is

$$\begin{aligned} F(\boldsymbol{x}) &= \mathbb{P}(X_j \leq x_j \text{ for all } j = 1, \ldots, d) \\ &= \mathbb{P}\big(F_j^{-1}(U_j) \leq x_j \text{ for all } j = 1, \ldots, d\big) \\ &\overset{[2.5.10]}{=} \mathbb{P}(U_j \leq F_j(x_j) \text{ for all } j = 1, \ldots, d) \\ &= C(F_1(x_1), \ldots, F_d(x_d)). \end{aligned}$$

This establishes Equation (7). Next, note that for all $\boldsymbol{u} \in \prod_{j=1}^{d} \mathrm{ran}(F_j)$ (which equals $[0,1]^d$ indeed due to the continuous, groundedness, and normalization properties of each F_j), we have

$$C(\boldsymbol{u}) \overset{[2.5.10]}{=} C(F_1(F_1^{-1}(u_1)), \ldots, F_d(F_d^{-1}(u_d)) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)).$$

(b) Let $\boldsymbol{U} = (U_1, \ldots, U_d) \sim C$ and $\boldsymbol{X} := (F_1^{-1}(U_1), \ldots, F_d^{-1}(U_d))$. Observe that

$$\mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}) \overset{[2.5.10]}{=} \mathbb{P}(U_1 \leq F_1(x_1), \ldots, U_d \leq F_d(x_d)) = C(F_1(x_1), \ldots, F_d(x_d))$$

for all $\boldsymbol{x} \in \mathbb{R}^d$, so the function F as defined in Equation (7) is a distribution function (of $\boldsymbol{X}$) with margins $F_1, \ldots, F_d$ by quantile transform.

$\square$

Remarks:

- *(Role of copula)* The decomposition property suggests that a joint distribution function can always be decomposed into two parts: (i) its margins $F_1, \ldots, F_d$ and (ii) a copula. From this, we can see that a copula indeed reflects the dependence inherit in the joint distribution function F.

- *(Construction of distribution functions via copula)* The combination property provides a method to construct joint distribution functions flexibly, according to dependence structures we want.

- We say that $\boldsymbol{X} \sim F$ **has copula** C if Equation (7) holds, where $F_1, \ldots, F_d$ are the margins of F. In case the margins are all continuous, we know by Theorem 4.3.b that such copula C would be uniquely determined (on $[0,1]^d$).

4.3.4 **Invariance principle.** Apart from Sklar's theorem, another important result about copula is the *invariance principle*, which asserts that the copula is invariant upon strictly increasing transformations; this is natural since such transformations do not change the "ordering" and hence should not influence the dependence. This is stated more precisely below.

Theorem 4.3.c (Invariance principle)**.** Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_d) \sim F$ with continuous margins $F_1, \ldots, F_d$ and copula C. If T_j is strictly increasing on $\mathrm{supp}(F_j)$ for every $j = 1, \ldots, d$, then $(T_1(X_1), \ldots, T_d(X_d))$ has the copula C as well, which is also uniquely determined.

Proof. Fix any $j = 1, \ldots, d$. On $\mathrm{supp}(F_j)$, the function T_j is strictly increasing, so it has at most countably many discontinuities, because at each discontinuity (jump in this case) there is a rational number in between. Hence, we have that T_j is right-continuous on the set $\mathrm{supp}(F_j)$ with at most countably many discontinuities removed. Since $\mathbb{P}(X_j \in \mathrm{supp}(F_j)) = 1$ and there are only at most countably discontinuities, the probability that X_j takes values in the resulting set after such removal would still be 1. So, we may assume T_j is strictly increasing and right-continuous in the following probabilistic argument.

71

By the right-continuity of T_j, we have $T_j(T_j^{-1}(x_j)) \overset{[2.5.10]}{=} x_j$ for all $x_j \in \mathrm{ran}(T_j)$. Hence, the distribution function G_j of $T_j(X_j)$ is given by

$$G_j(x_j) = \mathbb{P}(T_j(X_j) \le x_j) \overset{[2.5.10]}{=} \mathbb{P}\big(T_j(X_j) \le T_j(T_j^{-1}(x_j))\big)$$
$$\overset{(T_j \text{ strictly increasing})}{=} \mathbb{P}\big(X_j \le T_j^{-1}(x_j)\big) = F_j(T_j^{-1}(x_j))$$

for all $x_j \in \mathrm{ran}(T_j)$. Since $T_j(X_j)$ cannot possibly take values outside $\mathrm{ran}(T_j)$, the equation indeed holds for all $x_j \in \mathbb{R}$.

Since T_j is strictly increasing, by [2.5.10] T_j^{-1} is continuous on $\mathrm{ran}(T_j)$. Also, by assumption F_j is continuous, so the composition $G_j = F_j \circ T_j^{-1}$ is continuous on $\mathrm{ran}(T_j)$. Since G_j remains constant outside $\mathrm{supp}(G_j) = \mathrm{supp}(T_j(X_j)) \subseteq \mathrm{ran}(T_j)$, we conclude that G_j is continuous on $\mathbb{R}$.

Therefore, the joint distribution function of $(T_1(X_1), \ldots, T_d(X_d))$ is

$$\mathbb{P}(T_j(X_j) \le x_j \text{ for all } j = 1, \ldots, d) \overset{(G_j \text{ continuous})}{=} \mathbb{P}(T_j(X_j) < x_j \text{ for all } j = 1, \ldots, d)$$
$$\overset{(T_j \text{ right-continuous, } [2.5.10])}{=} \mathbb{P}\big(X_j < T_j^{-1}(x_j) \text{ for all } j = 1, \ldots, d\big)$$
$$\overset{(F_j \text{ continuous})}{=} \mathbb{P}\big(X_j \le T_j^{-1}(x_j) \text{ for all } j = 1, \ldots, d\big)$$
$$\overset{(\text{Sklar's theorem})}{=} C(F_1(T_1^{-1}(x_1)), \ldots, F_d(T_d^{-1}(x_d)) = C(G_1(x_1), \ldots, G_d(x_d))$$

for all $\boldsymbol{x} \in \mathbb{R}^d$.

Since $T_j(X_j) \sim G_j$ for all $j = 1, \ldots, d$, by definition we know $(T_1(X_1), \ldots, T_d(X_d))$ has copula C. Furthermore, by Sklar's theorem it is uniquely determined since $G_1, \ldots, G_d$ are all continuous. $\qquad \square$

4.3.5 **Copula from probability transforms.** By applying probability transforms and the invariance principle, we can obtain the following result, which justifies the practice of exploring the dependence between the samples after applying the respective distribution function on each of them (to "scale" the data appropriately and ensure a "fair" comparison).

Corollary 4.3.d. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a random vector with continuous margins $F_1, \ldots, F_d$. Then $\boldsymbol{X}$ has a unique copula C iff we have $(F_1(X_1), \ldots, F_d(X_d)) \sim C$.

Proof. Let $\boldsymbol{U} = (U_1, \ldots, U_d) := (F_1(X_1), \ldots, F_d(X_d))$.

"$\Rightarrow$": Knowing that F_j is strictly increasing on $\mathrm{supp}(F_j)$ for every $j = 1, \ldots, d$, the invariance principle suggests that $\boldsymbol{U}$ has the unique copula C. Also, by the probability transform, the margins of $\boldsymbol{U}$ are all $\mathrm{U}(0, 1)$. Hence, by Sklar's theorem (combination), the distribution function of $\boldsymbol{U}$ is C.

"$\Leftarrow$": By [2.5.10], we have $\boldsymbol{X} = \big(F_1^{-1}(F_1(X_1)), \ldots, F_d^{-1}(F_d(X_d))\big) = (F_1^{-1}(U_1), \ldots, F_d^{-1}(U_d))$. $\qquad \square$

4.3.6 **Three common copulas.** Three particular types of dependence structures are often of interest, namely *independence*, *extremely positive dependence (comonotonicity)*, and *extremely negative dependence (countermonotonicity)*. They are described by the following three copulas respectively:

(a) **independence copula**: $C(\boldsymbol{u}) = \Pi(\boldsymbol{u}) := \prod_{j=1}^{d} u_j$ for all $\boldsymbol{u} \in [0, 1]^d$.

(b) **comonotone copula**: $C(\boldsymbol{u}) = M(\boldsymbol{u}) := \min\{u_1, \ldots, u_d\}$ for all $\boldsymbol{u} \in [0, 1]^d$.

(c) **countermonotone copula** (when $d = 2$): $C(\boldsymbol{u}) = W(\boldsymbol{u}) := \max\{\sum_{j=1}^{d} u_j - d + 1, 0\}$ for all $\boldsymbol{u} \in [0, 1]^d$. [Note: Using a similar argument as in [2.4.7], one can show that W is never a valid copula for every $d \ge 3$.]

To prove that they are indeed valid copulas (only $d = 2$ for the countermonotone copula), consider the following:

Proof.

(a) We have $\boldsymbol{U} \sim \Pi$ for $\boldsymbol{U} = (U_1, \ldots, U_d)$ with $U_1, \ldots, U_d \overset{\text{iid}}{\sim} \mathrm{U}(0,1)$, since:

$$\mathbb{P}(\boldsymbol{U} \leq \boldsymbol{u}) = \prod_{j=1}^{d} \mathbb{P}(U_j \leq u_j) = \prod_{j=1}^{d} u_j = \Pi(\boldsymbol{u})$$

for all $\boldsymbol{u} \in [0,1]^d$.

(b) We have $\boldsymbol{U} \sim M$ for $\boldsymbol{U} = (U, , \ldots, U)$ with $U \sim \mathrm{U}(0,1)$, since:

$$\mathbb{P}(\boldsymbol{U} \leq \boldsymbol{u}) = \mathbb{P}(U \leq u_1, \ldots, U \leq u_d) = \mathbb{P}(U \leq \min\{u_1, \ldots, u_d\}) = \min\{u_1, \ldots, u_d\} = M(\boldsymbol{u})$$

for all $\boldsymbol{u} \in [0,1]^d$.

(c) When $d = 2$, we have $\boldsymbol{U} \sim W$ for $\boldsymbol{U} = (U, 1 - U)$ with $U \sim \mathrm{U}(0,1)$, since:

$$\mathbb{P}(\boldsymbol{U} \leq \boldsymbol{u}) = \mathbb{P}(U \leq u_1, 1 - U \leq u_2) = \mathbb{P}(1 - u_2 \leq U \leq u_1) \overset{\text{(by cases)}}{=} \max\{u_1 - (1 - u_2), 0\} = W(\boldsymbol{u})$$

for all $\boldsymbol{u} \in [0,1]^2$.

$\square$

4.3.7 **Interpretations of independence, comonotone, and countermonotone copulas.** For $(X_1, \ldots, X_d) \sim F$ with continuous margins $F_1, \ldots, F_d$ and copula C, we can interpret the previously introduced independence, comonotone, and countermonotone copulas as follows:

(a) *(Independence copula)* $C = \Pi$ iff $X_1, \ldots, X_d$ are <u>independent</u>.

(b) *(Comonotone copula)* $C = M$ iff $X_j \overset{\text{a.s.}}{=} T_j(X_1)$, where $T_j = F_j^{-1} \circ F_1$ is strictly increasing on $\mathrm{supp}(F_1)$, for all $j = 2, \ldots, d$.

(c) *(Countermonotone copula)* $C = W$ iff $X_2 \overset{\text{a.s.}}{=} T(X_1)$, where $T = F_2^{-1} \circ (1 - F_1)$ is strictly decreasing on $\mathrm{supp}(F_1)$.

Proof. (Sketch)

(a) Apply Theorem 4.2.c.

(b) Apply the invariance principle for "$\Leftarrow$".

(c) Apply the invariance principle for "$\Leftarrow$" also.

$\square$

4.3.8 **Fréchet-Hoeffding bounds.** For the comonotone copula M and countermonotone copula W, apart from describing the extremely positive dependence and extremely negative dependence, they also serve for bounds on *any* copula:

Theorem 4.3.e (Fréchet-Hoeffding bounds). For every d-dimensional copula C, we have $W(\boldsymbol{u}) \leq C(\boldsymbol{u}) \leq M(\boldsymbol{u})$ for all $\boldsymbol{u} \in [0,1]^d$.

Proof. Fix any $\boldsymbol{u} \in [0,1]^d$ and any d-dimensional copula C.

We first show the inequality $W \leq C$. Note that

$$1 - C(\boldsymbol{u}) = C(_d 1) - C(\boldsymbol{u}) \overset{\text{(Lemma 2.5.b)}}{\leq} \sum_{j=1}^{d} (1 - u_j) = d - \sum_{j=1}^{d} u_j,$$

and thus $C(\boldsymbol{u}) \geq (\sum_{j=1}^{d} u_j) - d + 1$. Since $C(\boldsymbol{u}) \geq 0$ (as a distribution function), this implies $C(\boldsymbol{u}) \geq \max\{(\sum_{j=1}^{d} u_j) - d + 1, 0\} = W(\boldsymbol{u})$.

Next, we show the inequality $C \leq M$. Since C is d-increasing, it is also componentwise increasing by [2.4.3]e. Thus, applying this increasingness for every component except the jth one gives

$$C(\boldsymbol{u}) \leq C(1, u_2, \ldots, u_d) \leq \cdots \leq C(1, \ldots, 1, u_j, 1, \ldots, 1) \stackrel{(\mathrm{U}(0,\,1)\ \mathrm{margins})}{=} u_j$$

for every $j = 1, \ldots, d$. This then implies that $C(\boldsymbol{u}) \leq \min\{u_1, \ldots, u_d\} = M(\boldsymbol{u})$. $\qquad \square$

Remarks:

- In view of this result, M and W are called the **Fréchet-Hoeffding upper bound** and **Fréchet-Hoeffding lower bound** respectively.

- Even if W is not a valid copula for every $d \geq 3$, it can still serve as a lower bound, as this result suggests.

5 Integration and Expectation

5.0.1 Apart from working with probability measures, often we would like to compute *expectations* in probability theory. In your first probability course, you should have learnt some formulas for computing expectations like below:

$$\mathbb{E}[X] = \begin{cases} \sum_k x_k f(x_k) & \text{if } X \text{ is discrete, taking countably many values } x_1, x_2, \ldots, \\ \int_{-\infty}^{\infty} x f(x) \, dx & \text{if } X \text{ is continuous} \end{cases}$$

where f denotes the mass function (for discrete case) or the density function (for continuous case). However, we will face difficulties when X is *neither discrete nor continuous* (this is possible; the definition for X to be continuous is not "X is not discrete" ⚠).

While these formulas are relatively easy to work with, they are not flexible enough for working with general random variables. Hence, in Section 5 we will introduce a more general (yet more abstract) definition of expectation, which is closely related to the notion of *Lebesgue integral*; let us start by studying the construction of Lebesgue integrals in Section 5.1, to better understand this concept.

5.1 Construction of Lebesgue Integrals

5.1.1 **Motivation.** The formulas of expectations mentioned above are both capturing the idea of "weighted average"; the expectation is considered to be an average weighted according to the probabilities. This idea is preserved in the general and abstract definition of expectation.

Consider the special case where the sample space Ω is countable, which forces all random variables defined on Ω to be discrete. Hence, we can express the expectation of a random variable X as follows:

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \mathbb{P}(\{\omega\}) \overset{\text{(in words)}}{=} \sum_{\text{all outcomes}} \text{outcome} \times \text{probability of having that outcome}$$

which is an alternative expression of the discrete expectation formula above. To extend this idea to the general case, we need to make sense of "weighted average" in the case where Ω is *uncountable* also. One natural idea is to replace "$\sum$" by "$\int$" in such case, and consider something like "$\int_{\Omega} X(\omega) \, d\mathbb{P}(\omega)$" or "$\int_{\Omega} X \, d\mathbb{P}$", just like the development of Riemann integral. Our goal here is to make sense of this kind of integral in a more general setting, where a general measure μ is considered.

5.1.2 **Steps for the construction.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $X : \Omega \to \mathbb{R}$ be a measurable function. Write $\bar{\mathbb{R}}_+ := [0, \infty]$ and $\bar{\mathbb{R}} := [-\infty, \infty]$. Then, the *Lebesgue integral of X with respect to μ*, denoted by $\int_{\Omega} X(\omega) \, d\mu(\omega)$ or $\int_{\Omega} X \, d\mu$, can be constructed in the following three-step process:

(1) Define $\int_{\Omega} X \, d\mu$ for every *simple function* $X : \Omega \to \mathbb{R}$.

(2) Extends the notion of $\int_{\Omega} X \, d\mu$ to every *nonnegative* $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}_+))$-measurable function $X : \Omega \to \bar{\mathbb{R}}_+$.

(3) Extends the notion of $\int_{\Omega} X \, d\mu$ to every $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}))$-measurable function $X : \Omega \to \bar{\mathbb{R}}$.

More generally, in many measure theoretic proofs, this kind of procedure is also utilized, which is known as the **standard argument**:

(1) Show that the target result holds for every *simple function* $X : \Omega \to \mathbb{R}$.

(2) Show that the target result holds for every *nonnegative* $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}_+))$-measurable function $X : \Omega \to \bar{\mathbb{R}}_+$.

(3) Show that the target result holds for every $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}))$-measurable function $X : \Omega \to \bar{\mathbb{R}}$.

Very often, step 2 is the step that takes the most work. The steps 1 and 3 are usually quite easy.

Remarks:

- Sometimes we may replace $\bar{\mathbb{R}}_+$ and $\bar{\mathbb{R}}$ by $\mathbb{R}_+$ and $\mathbb{R}$ respectively and also make some other slight adjustments, depending on the result we are trying to prove.
- For an illustration of the standard argument, see the proof of Theorem 5.2.a.

Lebesgue integrals of simple functions

5.1.3 **Definition.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Recall that a function $X : \Omega \to \mathbb{R}$ is said to be simple if it takes the form $X = \sum_{i=1}^{n} x_i \mathbf{1}_{A_i}$, where $x_1, \ldots, x_n \in \mathbb{R}$, $n \in \mathbb{N}$, and $A_1, \ldots, A_n \in \mathcal{F}$ are pairwise disjoint.

Now consider a simple function $X = \sum_{i=1}^{n} x_i \mathbf{1}_{A_i}$. The **Lebesgue integral of X with respect to μ** is

$$\int_\Omega X \, \mathrm{d}\mu := \sum_{i=1}^{n} x_i \mu(A_i),$$

with the convention that $0 \cdot \infty = 0$ (i.e., $x_i \mu(A_i) := 0$ if $x_i = 0$ and $\mu(A_i) = \infty$ for some i). Also, we write $\int_A X \, \mathrm{d}\mu := \int_\Omega X \mathbf{1}_A \, \mathrm{d}\mu$ for every $A \in \mathcal{F}$.
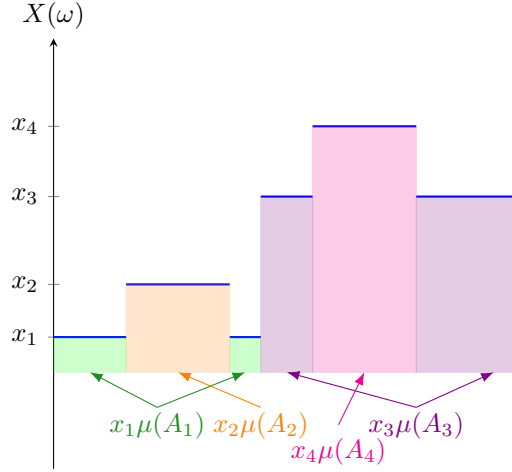
Writing $\sum_{i=1}^{n} x_i \mu(A_i)$ as $\sum_{i=1}^{n} x_i \mu(X = x_i)$, we can see that it aligns with the formula of expectation of discrete random variable X taking values $x_1, \ldots, x_n$ (when μ is a probability measure). Furthermore, from a geometric point of view, the integral $\int_\Omega X \, \mathrm{d}\mu$ also captures the notion of "area under curve", like the Riemann integral.

5.1.4 **Properties.** Let $X, Y : \Omega \to \mathbb{R}$ be simple functions, say $X = \sum_{i=1}^{n} x_i \mu(A_i)$ and $Y = \sum_{j=1}^{m} y_j \mu(B_j)$. Then:

(a) *(Nonnegativity)* $X \geq 0 \implies \int_\Omega X \, \mathrm{d}\mu \geq 0$.

(b) *(Linearity)* $\int_\Omega aX + bY \, \mathrm{d}\mu = a \int_\Omega X \, \mathrm{d}\mu + b \int_\Omega Y \, \mathrm{d}\mu$, for all $a, b \in \mathbb{R}$.

(c) *(Monotonicity)* $X \leq Y \implies \int_\Omega X \, \mathrm{d}\mu \leq \int_\Omega Y \, \mathrm{d}\mu$.

(d) If $X \geq 0$, then the function ν given by $\nu(A) := \int_A X \, \mathrm{d}\mu \ \forall A \in \mathcal{F}$ is a measure on $(\Omega, \mathcal{F})$.

(e) $\int_\Omega \mathbf{1}_A \, \mathrm{d}\mu = \mu(A)$ for every $A \in \mathcal{F}$.

Proof.

(a) Since $X \geq 0$, we have $x_i \geq 0$ for every $i = 1, \ldots, n$, and hence $\int_\Omega X \, \mathrm{d}\mu = \sum_{i=1}^{n} x_i \mu(A_i) \geq 0$.

(b) Note that $aX + bY = \sum_{i,j}(ax_i + by_j)\mathbf{1}_{A_i \cap B_j}$ is still simple, so

$$\int_\Omega aX + bY \, \mathrm{d}\mu = \sum_{i,j}(ax_i + by_j)\mu(A_i \cap B_j) = a\sum_{i=1}^{n} x_i \sum_{j=1}^{m} \mu(A_i \cap B_j) + b\sum_{j=1}^{m} y_j \sum_{i=1}^{n} \mu(A_i \cap B_j)$$

$$\overset{\text{(law of total measure)}}{=} a\sum_{i=1}^{n} x_i \mu(A_i) + b\sum_{j=1}^{m} y_j \mu(B_j) = a\int_\Omega X \, \mathrm{d}\mu + b\int_\Omega Y \, \mathrm{d}\mu.$$

76

(c) Note that $Y - X \geq 0$ is still simple, and thus

$$\int_\Omega Y \, d\mu \overset{\text{(linearity)}}{=} \int_\Omega Y - X \, d\mu + \int_\Omega X \, d\mu \overset{\text{(monotonicity)}}{\geq} \int_\Omega X \, d\mu.$$

(d) (1) For every $A \in \mathcal{F}$, we have $X \geq 0 \implies X\mathbf{1}_A \geq 0 \implies \mu(A) \geq 0$.

(2) Note that $\nu(\varnothing) = \int_\Omega X\mathbf{1}_\varnothing \, d\mu = \int_\Omega 0 \, d\mu \overset{\text{(simple integrand)}}{=} 0\mu(\Omega) = 0$ (even if $\mu(\Omega) = \infty$).

(3) Fix any pairwise disjoint $C_1, C_2, \ldots \in \mathcal{F}$. Then,

$$\mu\left(\biguplus_{k=1}^\infty C_k\right) = \int_\Omega X\mathbf{1}_{\biguplus_{k=1}^\infty C_k} \, d\mu = \int_\Omega \sum_{i=1}^n x_i \mathbf{1}_{A_i \cap \biguplus_{k=1}^\infty C_k} \, d\mu \overset{\text{(simple integrand)}}{=} \sum_{i=1}^n x_i \mu\left(A_i \cap \biguplus_{k=1}^\infty C_k\right)$$

$$\overset{\text{(σ-additivity)}}{=} \sum_{i=1}^n x_i \sum_{k=1}^\infty \mu(A_i \cap c_k) \overset{\text{(reordering)}}{\underset{\text{(Riemann series theorem)}}{=}} \sum_{k=1}^\infty \sum_{i=1}^n x_i \mu(A_i \cap C_k)$$

$$= \sum_{i=1}^\infty \int_\Omega X\mathbf{1}_{C_k} \, d\mu = \sum_{i=1}^\infty \nu(C_i).$$

(e) Note that $\int_\Omega \mathbf{1}_A \, d\mu \overset{\text{(simple integrand)}}{\underset{(\mathbf{1}_A = 1 \cdot \mathbf{1}_A + 0 \cdot \mathbf{1}_{A^c})}{=}} 1 \cdot \mu(A) + 0 \cdot \mu(A^c) = \mu(A)$.

$\square$

Lebesgue integrals of nonnegative measurable functions

5.1.5 **Definition.** The key idea to extend the notion of $\int_\Omega X \, d\mu$ to every *nonnegative* measurable function X is to approach the nonnegative measurable function X "from below", through nonnegative *simple functions* (for which the Lebesgue integral has been defined). We define the Lebesgue integral by taking the supremum over the Lebesgue integrals of all the nonnegative simple functions that "lie below" the nonnegative measurable function.

Let $X : \Omega \to \bar{\mathbb{R}}_+$ be a $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}_+))$-measurable function. Then the **Lebesgue integral of X with respect to μ** is

$$\int_\Omega X \, d\mu := \sup_{\substack{0 \leq Y \leq X, \\ Y \text{ simple}}} \int_\Omega Y \, d\mu.$$

Also, we write $\int_A X \, d\mu := \int_\Omega X\mathbf{1}_A \, d\mu$ for every $A \in \mathcal{F}$. The set of all $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}_+))$-measurable functions on Ω (i.e., all functions for which this kind of Lebesgue integral is defined) is denoted by L_+ or $L_+(\Omega, \mathcal{F}, \mu)$.

[Note: In case X is itself simple, say $X = \sum_{i=1}^n x_i \mathbf{1}_{A_i}$, we would have $\int_\Omega Y \, d\mu \leq \sum_{i=1}^n x_i \mu(A_i)$ for every simple function Y with $0 \leq Y \leq X$ by [5.1.4]. This then implies that $\sup_{\substack{0 \leq Y \leq X, \\ Y \text{ simple}}} \int_\Omega Y \, d\mu = \max_{\substack{0 \leq Y \leq X, \\ Y \text{ simple}}} \int_\Omega Y \, d\mu = \sum_{i=1}^n x_i \mu(A_i)$, so the Lebesgue integral here indeed coincides with the one defined for simple function in this case.]

5.1.6 **Preliminary results.** We first discuss some preliminary results about the Lebesgue integral of this kind.

Proposition 5.1.a (Approximating sequence). A function $X : \Omega \to \bar{\mathbb{R}}_+$ is in L_+ iff there exists a sequence $\{X_n\}$, with $X_n : \Omega \to \mathbb{R}_+ := [0, \infty)$ being nonnegative and simple for every $n \in \mathbb{N}$, such that $X_n \nearrow X$ pointwisely, with the convergence being also uniform on every set where X is bounded.

Remarks:

- Such sequence $\{X_n\}$ is said to be an **approximating sequence** to X.

77

- The notation $X_n \nearrow X$ means $X_n \to X$ with the extra emphasis that $\{X_n\}$ is increasing (denoted by $X_n \nearrow$), i.e., $X_1 \leq X_2 \leq \cdots$; we shall use similar notations for real-valued sequences also.

- *(Implications)* Based on this result, we know that the X_n's are included in the supremum "$\sup_{\substack{0 \leq Y \leq X, \\ Y \text{ simple}}}$," ($X_n \nearrow X$ implies $X_n \leq X$ for all $n \in \mathbb{N}$). Indeed, by the *monotone convergence theorem* (Theorem 5.1.c), the supremum $\sup_{\substack{0 \leq Y \leq X, \\ Y \text{ simple}}} \int_\Omega Y \, d\mu$ can be obtained as the *limit* of Lebesgue integrals of these simple X_n's as $n \to \infty$.

Proof. "$\Leftarrow$": Since X_n is simple and hence measurable for every $n \in \mathbb{N}$, we know by [3.1.5]c that X is measurable as a limit.

"$\Rightarrow$": For every $n \in \mathbb{N}$, let

$$X_n := \min\left\{\frac{\lfloor 2^n X \rfloor}{2^n}, n\right\} = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \mathbf{1}_{X^{-1}\left(\left(\frac{k-1}{2^n}, \frac{k}{2^n}\right]\right)} + n\mathbf{1}_{X^{-1}(n,\infty]}.$$

[Intuition 💡: The expression $\frac{\lfloor 2^n X \rfloor}{2^n}$ gives us a value that is "slightly" smaller than X, and taking the minimum with n avoids X_n to be "too large". Also, the final expression is related to a certain "partition" on the y-axis: each indicator $\mathbf{1}_{X^{-1}\left(\left(\frac{k-1}{2^n}, \frac{k}{2^n}\right]\right)}$ corresponds to a certain region in the y-axis.]

Since X is measurable, the sets involved in the indicators are all in $\mathcal{F}$, thus each X_n is simple. It is also clear that each X_n is nonnegative.

First, we show that $X_n \nearrow$. For every $n \in \mathbb{N}$, note that $\lfloor 2^n X \rfloor / 2^n = 2\lfloor 2^n X \rfloor / 2^{n+1} \leq \lfloor 2^{n+1} X \rfloor / 2^{n+1}$. Thus,

$$X_n = \min\left\{\frac{\lfloor 2^n X \rfloor}{2^n}, n\right\} \overset{\text{(by case)}}{\leq} \min\left\{\frac{\lfloor 2^{n+1} X \rfloor}{2^{n+1}}, n+1\right\} = X_{n+1}.$$

Next, we show the uniform convergence. On every set where X is bounded, we would have $X \leq N$ for some $N \in \mathbb{N}$. Hence, for all $n \geq N$, we have $X \in (\frac{k-1}{2^n}, \frac{k}{2^n}]$ and $X_n = (k-1)/2^n$ for some $k = 1, \ldots, n2^n$, which means that $|X_n - X| \leq 1/2^n$. This would then assert the uniform convergence on such set.

Finally, we show the pointwise convergence. This uniform convergence implies the pointwise convergence for every $\omega \in \Omega$ with $X(\omega) < \infty$ (for such ω, the singleton $\{\omega\}$ would be a set where X is bounded). So, it remains to show the pointwise convergence for every $\omega \in \Omega$ where $X(\omega) = \infty$. For every such ω, we have $X_n(\omega) = n$ for all $n \in \mathbb{N}$, hence $\{X_n(\omega)\} = \{n\} \to \infty$, establishing the pointwise convergence (which includes the possibility of "∞"). $\qquad \square$

Lemma 5.1.b (Monotonicity and scaling)**.** Let $X, Y \in L_+$. Then:

(a) *(Monotonicity)* $X \leq Y \implies \int_\Omega X \, d\mu \leq \int_\Omega Y \, d\mu$.

(b) *(Scaling)* $\int_\Omega cX \, d\mu = c\int_\Omega X \, d\mu$ for all $c \geq 0$.

Proof.

(a) With $X \leq Y$, the integral $\int_\Omega Y \, d\mu$ is a supremum over a larger set than $\int_\Omega X \, d\mu$. Hence, the former must be at least as large as the latter.

(b) If $c = 0$, then the result is immediate as we have $\int_\Omega 0 \, d\mu = 0$. So assume henceforth that $c > 0$, and consider:

$$\int_\Omega cX \, d\mu = \sup_{\substack{0 \leq Y \leq cX, \\ Y \text{ simple}}} \int_\Omega Y \, d\mu = \sup_{\substack{0 \leq cZ \leq cX, \\ cZ \text{ simple}}} \int_\Omega cZ \, d\mu$$

$$= \sup_{\substack{0 \leq Z \leq X, \\ Z \text{ simple}}} c\int_\Omega Z \, d\mu = c \cdot \sup_{\substack{0 \leq Z \leq X, \\ Z \text{ simple}}} \int_\Omega Z \, d\mu = c\int_\Omega X \, d\mu.$$

$\qquad \square$

5.1.7 **Monotone convergence theorem.** One important theorem concerning Lebesgue integrals of nonnegative measurable functions is the *monotone convergence theorem (MCT)*, which provides conditions under which the integral $\int_\Omega X \, d\mu$ can be computed as a *limit* of <u>monotone</u> sequence instead of taking the supremum. It can also be viewed as a result that provides a sufficient condition that permits the *interchange of limit and integral.*

Apart from being an important result on its own, it is also used for proving many properties of Lebesgue integral of this kind.

Theorem 5.1.c (Monotone convergence theorem (MCT))**.** Let X_n be a function on Ω in L_+ for every $n \in \mathbb{N}$. If we have $X_n \nearrow X$ pointwisely, then $X \in L_+$ and $\int_\Omega X_n \, d\mu \nearrow \int_\Omega X \, d\mu$.

[Note: For the convergence of Lebesgue integrals, we can also write $\lim_{n\to\infty} \int_\Omega X_n \, d\mu = \int_\Omega \lim_{n\to\infty} X_n \, d\mu$, from which an interchange of limit and integral can be clearly seen.]

Proof. Since X is the limit of nonnegative measurable functions X_n's, we know that $X \geq 0$ and X is measurable by [3.1.5]c, and hence $X \in L_+$. Also, with $X_n \nearrow$, by Lemma 5.1.b we have $\int_\Omega X_n \, d\mu \nearrow$, and so the limit $\lim_{n\to\infty} \int_\Omega X_n \, d\mu$ exists (but may possibly be ∞).

It then remains to establish that $\lim_{n\to\infty} \int_\Omega X_n \, d\mu = \int_\Omega X \, d\mu$:

- "$\leq$": From $X_n \nearrow X$, we have $X = \sup_{n\in\mathbb{N}} X_n \geq X_n$ for every $n \in \mathbb{N}$. So, by Lemma 5.1.b we know $\int_\Omega X_n \, d\mu \leq \int_\Omega X \, d\mu$ for every $n \in \mathbb{N}$. Thus, $\lim_{n\to\infty} \int_\Omega X_n \, d\mu \leq \int_\Omega X \, d\mu$.

- "$\geq$": Fix any simple function Y with $0 \leq Y \leq X$, and any $\alpha \in (0,1)$. Let $A_n := \{\omega \in \Omega : X_n(\omega) \geq \alpha Y(\omega)\}$ for every $n \in \mathbb{N}$. Since each $X_n - \alpha Y$ is measurable, each A_n is in $\mathcal{F}$. Also, with $X_n \nearrow$, we have $A_n \nearrow$ and

$$\lim_{n\to\infty} A_n = \bigcup_{k=1}^{\infty} A_k = \{\omega \in \Omega : X_k(\omega) \geq \alpha Y(\omega) \text{ for some } k \in \mathbb{N}\}.$$

This set equals Ω, since we have $\lim_{n\to\infty} X_n(\omega) = X(\omega) \geq Y(\omega) \geq \alpha Y(\omega)$, which implies that $X_k(\omega) \geq \alpha Y(\omega)$ must hold for sufficiently large $k \in \mathbb{N}$. So we can write $A_n \nearrow \Omega$.

By Lemma 5.1.b, we have $\int_\Omega X_n \, d\mu \geq \int_\Omega X_n \mathbf{1}_{A_n} \, d\mu \geq \alpha \int_\Omega Y \mathbf{1}_{A_n} \, d\mu = \alpha \int_{A_n} Y \, d\mu$ for all $n \in \mathbb{N}$. Taking limit then gives

$$\lim_{n\to\infty} \int_\Omega X_n \, d\mu \geq \lim_{n\to\infty} \alpha \int_{A_n} Y \, d\mu \overset{[5.1.4]}{=} \alpha \lim_{n\to\infty} \nu(A_n) \overset{\text{(continuity from below)}}{=} \alpha\nu(\Omega) \overset{\text{(definition of } \nu)}{=} \alpha \int_\Omega Y \, d\mu.$$

Letting $\alpha \to 1^-$, we have $\lim_{n\to\infty} \int_\Omega X_n \, d\mu \geq \int_\Omega Y \, d\mu$, so $\lim_{n\to\infty} \int_\Omega X_n \, d\mu$ serves as an upper bound of $\{\int_\Omega Y \, d\mu : 0 \leq Y \leq X, \quad Y \text{ simple}\}$. Thus, by the definition of supremum, we have $\lim_{n\to\infty} \int_\Omega X_n \, d\mu \geq \int_\Omega X \, d\mu$.

$\square$

5.1.8 **Properties.** Now we are ready to prove more properties of Lebesgue integral of nonnegative measurable function. Let $X, Y \in L_+$ and $X_n \in L_+$ for each $n \in \mathbb{N}$. Then:

(a) $\int_\Omega X \, d\mu = 0$ iff $X \overset{\text{a.e.}}{=} 0$.

(b) *(Linearity)* $\int_\Omega aX + bY \, d\mu = a \int_\Omega X \, d\mu + b \int_\Omega Y \, d\mu$ for all $a, b \geq 0$.

(c) *(Commutativity of integral and countable sum)* $\sum_{n=1}^{\infty} X_n \in L_+$ and $\int_\Omega \sum_{n=1}^{\infty} X_n \, d\mu = \sum_{n=1}^{\infty} \int_\Omega X_n \, d\mu$.

(d) *("A.e. version" of MCT)* $X_n \nearrow X$ a.e. $\implies \lim_{n\to\infty} \int_\Omega X_n \, d\mu = \int_\Omega X \, d\mu$. [Note: "$X_n \nearrow X$ a.e." means that we have $X_n(\omega) \nearrow X(\omega)$ for all $\omega \in \Omega \setminus N$ where N is a null set.]

(e) If $\int_\Omega X \, d\mu < \infty$, then $\mu(X = \infty) = 0$, i.e., X is finite a.e.

(f) The function ν given by $\nu(A) := \int_A X \, d\mu \; \forall A \in \mathcal{F}$ is a measure on $(\Omega, \mathcal{F})$.

Proof.

79

(a) We first consider the case where X is simple. Write $X = \sum_{i=1}^{n} x_i \mathbf{1}_{A_i}$ where $x_1, \ldots, x_n \geq 0$. Then, we have

$$\int_\Omega X \, d\mu = \sum_{i=1}^{n} x_i \mu(A_i) = 0$$

$$\iff x_i = 0 \text{ or } \mu(A_i) = 0 \text{ for all } i = 1, \ldots, n$$

$$\iff \mu(X > 0) = 0$$

$$\iff X \stackrel{\text{a.e.}}{=} 0,$$

which establishes the result for this case.

Next, consider the general case where $X \in L_+$.

"$\Leftarrow$": Assume $X \stackrel{\text{a.e.}}{=} 0$. Fix any simple Y with $0 \leq Y \leq X$. Since $X \stackrel{\text{a.e.}}{=} 0$, we have $Y \stackrel{\text{a.e.}}{=} 0$. Hence, by the result for simple function, we have $\int_\Omega Y \, d\mu = 0$. Therefore,

$$\int_\Omega X \, d\mu = \sup_{\substack{0 \leq Y \leq X, \\ Y \text{ simple}}} \underbrace{\int_\Omega Y \, d\mu}_{0} = 0.$$

"$\Rightarrow$": Assume $\int_\Omega X \, d\mu = 0$. Let $A_n := \{\omega \in \Omega : X(\omega) \geq 1/n\}$ for every $n \in \mathbb{N}$. Then, we have $X \geq \mathbf{1}_{A_n}/n$ for all $n \in \mathbb{N}$ and $\mu(\bigcup_{n=1}^{\infty} A_n) = \mu(\{\omega \in \Omega : X(\omega) > 0\}) = \mu(X > 0)$. So, it suffices to show that $\mu(\bigcup_{n=1}^{\infty} A_n) = 0$.

Assume to the contrary that $\mu(\bigcup_{n=1}^{\infty} A_n) > 0$. Then there exists $n_0 \in \mathbb{N}$ such that $\mu(A_{n_0}) > 0$; otherwise we would have $\mu(A_n) = 0$ for all $n \in \mathbb{N}$, so $\mu(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \mu(A_n) = 0$, which is impossible. Hence, we have $\int_\Omega X \, d\mu \geq \int_\Omega \frac{1}{n_0} \mathbf{1}_{A_{n_0}} \, d\mu \stackrel{\text{(simple integrand)}}{=} \frac{1}{n_0} \mu(A_{n_0}) > 0$, contradiction.

(b) Due to the scaling property, it suffices to show the result with $a = b = 1$. By Proposition 5.1.a, there exist sequences $\{X_n\}$ and $\{Y_n\}$ of nonnegative simple functions such that $X_n \nearrow X$ and $Y_n \nearrow Y$. From this, we know $\{X_n + Y_n\}$ is a sequence of nonnegative simple functions with $X_n + Y_n \nearrow X + Y$. Thus,

$$\int_\Omega X + Y \, d\mu \stackrel{\text{(MCT)}}{=} \lim_{n \to \infty} \int_\Omega X_n + Y_n \, d\mu \stackrel{\text{(linearity for simple functions)}}{=} \lim_{n \to \infty} \int_\Omega X_n \, d\mu + \lim_{n \to \infty} \int_\Omega Y_n \, d\mu$$

$$\stackrel{\text{(MCT)}}{=} \int_\Omega X \, d\mu + \int_\Omega Y \, d\mu.$$

(c) Note that we have $\sum_{n=1}^{\infty} X_n \in L^+$ by MCT, and

$$\int_\Omega \sum_{n=1}^{\infty} X_n \, d\mu \stackrel{\text{(MCT)}}{=} \lim_{N \to \infty} \int_\Omega \underbrace{\sum_{n=1}^{N} X_n}_{\nearrow} \, d\mu \stackrel{\text{(b)}}{=} \lim_{N \to \infty} \sum_{n=1}^{N} \int_\Omega X_n \, d\mu = \sum_{n=1}^{\infty} \int_\Omega X_n \, d\mu.$$

(d) By assumption, we have $X_n(\omega) \nearrow X(\omega)$ for all $\omega \in \Omega \setminus N$ where N is a null set. Since $X_n - X_n \mathbf{1}_{N^c} \stackrel{\text{a.e.}}{=} 0$ for all $n \in \mathbb{N}$ and $X - X \mathbf{1}_{N^c} \stackrel{\text{a.e.}}{=} 0$, by (a) we have $\int_\Omega X_n - X_n \mathbf{1}_{N^c} \, d\mu = 0$ for all $n \in \mathbb{N}$ and $\int_\Omega X - X \mathbf{1}_{N^c} \, d\mu = 0$, which implies by (b) that $\int_\Omega X_n \, d\mu = \int_\Omega X_n \mathbf{1}_{N^c} \, d\mu$ for all $n \in \mathbb{N}$ and $\int_\Omega X \, d\mu = \int_\Omega X \mathbf{1}_{N^c} \, d\mu$.

Therefore,

$$\lim_{n \to \infty} \int_\Omega X_n \, d\mu = \lim_{n \to \infty} \int_\Omega X_n \mathbf{1}_{N^c} \, d\mu \stackrel{\text{(MCT)}}{\underset{(X_n \mathbf{1}_{N^c} \nearrow X \mathbf{1}_{N^c} \text{ pointwisely})}{=}} \int_\Omega X \mathbf{1}_{N^c} \, d\mu = \int_\Omega X \, d\mu.$$

(e) Let $A_n := \{X \geq n\}$ for all $n \in \mathbb{N} \cup \{\infty\}$. Assume to the contrary that $\mu(X = \infty) = \mu(A_\infty) > 0$. Then, we have

$$\int_\Omega X \, d\mu = \int_\Omega X \mathbf{1}_{A_n} \, d\mu \geq \int_\Omega n \mathbf{1}_{A_n} \, d\mu \stackrel{\text{(simple integrand)}}{=} n\mu(A_n) \stackrel{(A_n \supseteq A_\infty)}{\geq} n\mu(A_\infty) \to \infty \text{ as } n \to \infty,$$

contradiction.

(f) (1) For every $A \in \mathcal{F}$, we have $\nu(A) = \int_\Omega \underbrace{X \mathbf{1}_A}_{\geq 0} \, d\mu \overset{\text{(monotonicity)}}{\geq} 0.$

    (2) We have $\nu(\varnothing) = \int_\Omega \underbrace{X \mathbf{1}_\varnothing}_{0} \, d\mu = 0.$

    (3) Fix any pairwise disjoint $A_1, A_2, \ldots \in \mathcal{F}$. Then $\nu(\biguplus_{i=1}^\infty A_i) = \int_\Omega X \mathbf{1}_{\biguplus_{i=1}^\infty A_i} \, d\mu = \int_\Omega \sum_{i=1}^\infty X \mathbf{1}_{A_i} \, d\mu \overset{\text{(c)}}{=}$
    $\sum_{i=1}^\infty \int_\Omega X \mathbf{1}_{A_i} \, d\mu = \sum_{i=1}^\infty \nu(A_i).$

$\square$

Lebesgue integrals of measurable functions

5.1.9 **Definition.** The last step in the construction of Lebesgue integral is to extend the notion of $\int_\Omega X \, d\mu$ to every $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}))$-measurable function $X : \Omega \to \bar{\mathbb{R}}$. The key idea is to write such function X as the difference $X = X^+ - X^-$ where $X^+ := \max\{X, 0\} \geq 0$ and $X^- := \max\{-X, 0\} \geq 0$ are the **positive part** and **negative part** of X respectively. Note also that we have $|X| = X^+ + X^-$, and that X is measurable iff X^+ and X^- both are (by [3.1.5]b).

Let $X : \Omega \to \bar{\mathbb{R}}$ be $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}))$-measurable. If $\int_\Omega X^+ \, d\mu < \infty$ <u>or</u> $\int_\Omega X^- \, d\mu < \infty$, then X is said to be **quasi-integrable** and the **Lebesgue integral of X with respect to μ** is

$$\int_\Omega X \, d\mu := \int_\Omega X^+ \, d\mu - \int_\Omega X^- \, d\mu.$$

Also, we write $\int_A X \, d\mu := \int_\Omega X \mathbf{1}_A \, d\mu$. Furthermore, if $\int_\Omega X^+ \, d\mu < \infty$ <u>and</u> $\int_\Omega X^- \, d\mu < \infty$, then X is said to be **integrable**; if $\int_\Omega X^+ \, d\mu = \infty$ <u>and</u> $\int_\Omega X^- \, d\mu = \infty$, then X is said to be **non-integrable**. [⚠ Warning: Here, "non-integrable" is not the same as "not integrable"! The latter covers the possibility of quasi-integrable also.]

The set of all integrable and $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable functions is denoted by L^1 or $L^1(\Omega, \mathcal{F}, \mu)$; note that here we only consider functions with codomain $\mathbb{R}$ rather than $\bar{\mathbb{R}}$ so that the values taken by such functions are always finite, for mathematical tractability. (To be compatible with the definition above, we may still view such functions as having the codomain $\bar{\mathbb{R}}$ notionally, but they are not allowed to take the values $\pm\infty$.)

Remarks:

- *(More on quasi-integrability)* If $\int_\Omega X^+ \, d\mu = \infty$ and $\int_\Omega X^+ \, d\mu < \infty$, then $\int_\Omega X \, d\mu = \infty$. If $\int_\Omega X^+ \, d\mu < \infty$ and $\int_\Omega X^+ \, d\mu = \infty$, then $\int_\Omega X \, d\mu = -\infty$.

- *(More on integrability)* Since $|X| = X^+ + X^-$, we have $\int_\Omega |X| \, d\mu < \infty \iff (\int_\Omega X^+ \, d\mu < \infty$ and $\int_\Omega X^- \, d\mu < \infty)$. Thus, we know X is integrable iff $\int_\Omega |X| \, d\mu < \infty$, which gives us a helpful criterion for determining integrability.

- *(Lebesgue integral of complex-valued function)* Lebesgue integral of a *complex-valued* function can be constructed by handling its real and imaginary parts separately. For a complex-valued function $X : \Omega \to \mathbb{C}$ that is $(\mathcal{F}, \mathcal{B}(\mathbb{C}))$-measurable (note that we have $\mathcal{B}(\mathbb{C}) = \mathcal{B}(\mathbb{R}^2)$), it is **integrable** if $\int_\Omega |\operatorname{Re}(X)| \, d\mu < \infty$ and $\int_\Omega |\operatorname{Im}(X)| \, d\mu < \infty$ where $\operatorname{Re}(X)$ and $\operatorname{Im}(X)$ denote the real and imaginary parts of X respectively. If it is integrable, then we define $\int_\Omega X \, d\mu := \int_\Omega \operatorname{Re}(X) \, d\mu + i \int_\Omega \operatorname{Im}(X) \, d\mu$.
Since $|X| \overset{\text{(triangle inequality)}}{\leq} |\operatorname{Re}(X)| + |\operatorname{Im}(X)| \overset{\text{(Pythagorean theorem)}}{\leq} |X| + |X| = 2|X|$, we have $\int_\Omega |X| \, d\mu < \infty$ iff $(\int_\Omega |\operatorname{Re}(X)| \, d\mu < \infty$ and $\int_\Omega |\operatorname{Im}(X)| \, d\mu < \infty)$, we again have X is integrable iff $\int_\Omega |X| \, d\mu < \infty$ here (where $|X|$ denotes the modulus of X).

- *(Special notation for Lebesgue measure)* In the case where $\Omega = \mathbb{R}^d$ and $\mu = \lambda$ is the Lebesgue measure, we often just write $\int_\Omega f(\boldsymbol{x}) \, d\boldsymbol{x}$ instead of $\int_\Omega f(\boldsymbol{x}) \, d\lambda(\boldsymbol{x})$.

5.1.10 **Properties of quasi-integrable functions.** Let $X, Y : \Omega \to \bar{\mathbb{R}}$ be quasi-integrable (and measurable) functions. Then:

(a) *(Additivity)* If we have $(\int_\Omega X^- \, d\mu < \infty$ and $\int_\Omega Y^- \, d\mu < \infty)$ or $(\int_\Omega X^+ \, d\mu < \infty$ and $\int_\Omega Y^+ \, d\mu < \infty)$, then $\int_\Omega X + Y \, d\mu = \int_\Omega X \, d\mu + \int_\Omega Y \, d\mu$.

(b) *(Scaling)* $\int_\Omega cX \, d\mu = c \int_\Omega X \, d\mu$.

(c) *(Nonnegativity)* $X \geq 0 \implies \int_\Omega X \, d\mu \geq 0$.

(d) *(Monotonicity)* $X \leq Y \implies \int_\Omega X \, d\mu \leq \int_\Omega Y \, d\mu$.

(e) *(Triangle inequality)* $|\int_\Omega X \, d\mu| \leq \int_\Omega |X| \, d\mu$.

(f) *(Integral over null set is always zero)* $\int_A X \, d\mu = 0$ for every $A \in \mathcal{F}$ with $\mu(A) = 0$.

[Note: If X and Y are *integrable*, then combining (a) and (b) would yield the *linearity*: $\int_\Omega aX + bY \, d\mu = a \int_\Omega X \, d\mu + b \int_\Omega Y \, d\mu$ for all $a, b \in \mathbb{R}$.]

Proof.

(a) First note that
$$\begin{cases} X + Y = (X+Y)^+ - (X+Y)^-, \\ X + Y = X^+ - X^- + Y^+ - Y^-, \end{cases}$$

thus we have $(X+Y)^+ + X^- + Y^- = (X+Y)^- + X^+ + Y^+$. As all the functions here are nonnegative, by [5.1.8] we have

$$\int_\Omega (X+Y)^+ \, d\mu + \int_\Omega X^- \, d\mu + \int_\Omega Y^- \, d\mu = \int_\Omega (X+Y)^- \, d\mu + \int_\Omega X^+ \, d\mu + \int_\Omega Y^+ \, d\mu. \quad (8)$$

WLOG, assume we are in the case where $\int_\Omega X^- \, d\mu < \infty$ and $\int_\Omega Y^- \, d\mu < \infty$. Since we have $(X+Y)^- \leq X^- + Y^-$, by monotonicity we get $\int_\Omega (X+Y)^- \, d\mu \leq \int_\Omega X^- \, d\mu + \int_\Omega Y^- \, d\mu < \infty$. Therefore, subtracting $\int_\Omega X^- \, d\mu + \int_\Omega Y^- \, d\mu + \int_\Omega (X+Y)^- < \infty$ from both sides of Equation (8) yields

$$\int_\Omega (X+Y)^+ \, d\mu - \int_\Omega (X+Y)^- \, d\mu = \int_\Omega X^+ \, d\mu - \int_\Omega X^- \, d\mu + \int_\Omega Y^+ \, d\mu - \int_\Omega Y^- \, d\mu,$$

which implies that $\int_\Omega X + Y \, d\mu = \int_\Omega X \, d\mu + \int_\Omega Y \, d\mu$.

(b) • *Case 1: $c \geq 0$.* We have

$$\int_\Omega cX \, d\mu = \int_\Omega (cX)^+ \, d\mu - \int_\Omega (cX)^- \, d\mu \overset{\text{(by case)}}{=} \int_\Omega c(X^+) \, d\mu - \int_\Omega c(X^-) \, d\mu$$
$$= c \int_\Omega X^+ \, d\mu - c \int_\Omega X^- \, d\mu = c \int_\Omega X \, d\mu.$$

• *Case 2: $c < 0$.* We have

$$\int_\Omega cX \, d\mu = \int_\Omega (cX)^+ \, d\mu - \int_\Omega (cX)^- \, d\mu \overset{\text{(by case)}}{=} \int_\Omega -c(X^-) \, d\mu - \int_\Omega -c(X^+) \, d\mu$$
$$= -c \int_\Omega X^- \, d\mu + c \int_\Omega X^+ \, d\mu = c \int_\Omega X \, d\mu.$$

(c) With $X \geq 0$, we know $X = X^+$, so by the monotonicity for Lebesgue integral of nonnegative measurable function, we have $\int_\Omega X \, d\mu = \int_\Omega X^+ \, d\mu \geq 0$.

(d) Since $Y - X \geq 0$, we have $\int_\Omega (Y-X)^- \, d\mu = \int_\Omega 0 \, d\mu = 0 < \infty$. Thus, $Y - X$ is quasi-integrable. With X and $Y - X$ being quasi-integrable, we can apply (a) and (c) to get

$$\int_\Omega Y \, d\mu \overset{\text{(a)}}{=} \int_\Omega Y - X \, d\mu + \int_\Omega X \, d\mu = \int_\Omega Y - X \, d\mu + \int_\Omega X \, d\mu \overset{\text{(c)}}{\geq} \int_\Omega X \, d\mu.$$

(e) We have

$$\left|\int_\Omega X\,\mathrm{d}\mu\right| = \left|\int_\Omega X^+\,\mathrm{d}\mu - \int_\Omega X^-\,\mathrm{d}\mu\right| \overset{\text{(triangle inequality for real numbers)}}{\leq} = \left|\int_\Omega X^+\,\mathrm{d}\mu\right| + \left|\int_\Omega X^-\,\mathrm{d}\mu\right|$$

$$\overset{(X^+,X^-\geq 0)}{=} \int_\Omega X^+\,\mathrm{d}\mu + \int_\Omega X^-\,\mathrm{d}\mu = \int_\Omega X^+ + X^-\,\mathrm{d}\mu = \int_\Omega |X|\,\mathrm{d}\mu.$$

(f) First consider the special case where $X \geq 0$, which implies $X\mathbf{1}_A \geq 0$. Fix any simple function $Y = \sum_{i=1}^n y_i\mathbf{1}_{B_i}$ with $0 \leq Y \leq X\mathbf{1}_A$, where $y_1,\ldots,y_n \geq 0$. For all $i = 1,\ldots,n$, we have $y_i\mathbf{1}_{B_i} \leq X\mathbf{1}_A$, and consider the following two cases:

- *Case 1:* $B_i \subseteq A$. By monotonicity of μ, we have $\mu(B_i) \leq \mu(A) = 0$, forcing that $\mu(B_i) = 0$.
- *Case 2:* $B_i \not\subseteq A$. In this case, there exists $\omega \in B_i$ such that $\omega \notin A$. For this ω, the inequality asserts that $y_i = y_i\mathbf{1}_{B_i}(\omega) \leq X(\omega)\mathbf{1}_A(\omega) = 0$, which forces that $y_i = 0$.

Therefore, in either case we have $y_i\mathbf{1}_{B_i} = 0$. It follows that $\int_\Omega Y\,\mathrm{d}\mu = \sum_{i=1}^n y_i\mu(B_i) = 0$. Hence, we have $\int_A X\,\mathrm{d}\mu = \int_\Omega X\mathbf{1}_A\,\mathrm{d}\mu = \sup_{\substack{0\leq Y\leq X,\\Y\text{ simple}}} \int_\Omega Y\,\mathrm{d}\mu = 0$.

Now, consider the general case. Using the result proven for the special case above, we get $\int_\Omega X\mathbf{1}_A\,\mathrm{d}\mu = \int_\Omega X^+\mathbf{1}_A\,\mathrm{d}\mu - \int_\Omega X^-\mathbf{1}_A\,\mathrm{d}\mu = 0 - 0 = 0$.

$\square$

5.1.11 **More properties of Lebesgue integrals.**

(a) *(σ-additivity of Lebesgue integral)* Let $A_1, A_2, \ldots \in \mathcal{F}$ be pairwise disjoint and let $A := \biguplus_{i=1}^\infty A_i$. Let $Z : \Omega \to \bar{\mathbb{R}}$ be a measurable function such that $Z\mathbf{1}_A \in L^1$. Then, we have

$$\int_A Z\,\mathrm{d}\mu = \sum_{i=1}^\infty \int_{A_i} Z\,\mathrm{d}\mu.$$

(b) *(A.e. finiteness of integrable functions)* If X is integrable, then $\mu(X = \pm\infty) = 0$, i.e., X is finite a.e.

(c) *(Equivalent criteria for a.e. equality of integrable functions)* Let X, Y be integrable functions. Then the following are equivalent:

   i. $X \overset{\text{a.e.}}{=} Y$.
   ii. $\int_A X\,\mathrm{d}\mu = \int_A Y\,\mathrm{d}\mu$ for all $A \in \mathcal{F}$.
   iii. $\int_\Omega |X - Y|\,\mathrm{d}\mu = 0$.

Proof.

(a) First, note that $\int_\Omega |Z|\mathbf{1}_A\,\mathrm{d}\mu = \int_\Omega |Z\mathbf{1}_A|\,\mathrm{d}\mu \overset{(Z\mathbf{1}_A\in L^1)}{<} \infty$. Thus, $\int_\Omega Z^+\mathbf{1}_A\,\mathrm{d}\mu \leq \int_\Omega |Z|\mathbf{1}_A\,\mathrm{d}\mu < \infty$ and $\int_\Omega Z^-\mathbf{1}_A\,\mathrm{d}\mu \leq \int_\Omega |Z|\mathbf{1}_A\,\mathrm{d}\mu < \infty$. So, we can write

$$\int_\Omega Z\mathbf{1}_A\,\mathrm{d}\mu = \int_\Omega (Z^+ - Z^-)\mathbf{1}_A\,\mathrm{d}\mu = \int_\Omega Z^+\mathbf{1}_A\,\mathrm{d}\mu - \int_\Omega Z^-\mathbf{1}_A\,\mathrm{d}\mu.$$

By [5.1.8], the function ν given by $\nu(B) = \int_\Omega Z^+\mathbf{1}_B\,\mathrm{d}\mu\ \forall B \in \mathcal{F}$ is a measure, so by its σ-additivity we have $\int_\Omega Z^+\mathbf{1}_A\,\mathrm{d}\mu = \nu(A) = \nu(\biguplus_{i=1}^\infty A_i) = \sum_{i=1}^\infty \nu(A_i) = \sum_{i=1}^\infty \int_\Omega Z^+\mathbf{1}_{A_i}\,\mathrm{d}\mu$, and similarly, $\int_\Omega Z^-\mathbf{1}_A\,\mathrm{d}\mu = \sum_{i=1}^\infty \int_\Omega Z^-\mathbf{1}_{A_i}\,\mathrm{d}\mu$. Hence,

$$\int_A Z\,\mathrm{d}\mu = \int_\Omega Z\mathbf{1}_A\,\mathrm{d}\mu \overset{\text{(above)}}{=} \int_\Omega Z^+\mathbf{1}_A\,\mathrm{d}\mu - \int_\Omega Z^-\mathbf{1}_A\,\mathrm{d}\mu = \sum_{i=1}^\infty \int_\Omega Z^+\mathbf{1}_{A_i}\,\mathrm{d}\mu - \sum_{i=1}^\infty \int_\Omega Z^-\mathbf{1}_{A_i}\,\mathrm{d}\mu$$

$$= \sum_{i=1}^\infty \left(\int_\Omega Z^+\mathbf{1}_{A_i}\,\mathrm{d}\mu - \int_\Omega Z^-\mathbf{1}_{A_i}\,\mathrm{d}\mu\right) = \sum_{i=1}^\infty \int_\Omega (Z^+ - Z^-)\mathbf{1}_{A_i}\,\mathrm{d}\mu$$

$$= \sum_{i=1}^\infty \int_\Omega Z\mathbf{1}_{A_i}\,\mathrm{d}\mu = \sum_{i=1}^\infty \int_{A_i} Z\,\mathrm{d}\mu.$$

(b) Since X is integrable, we have $\int_\Omega X^+ \, \mathrm{d}\mu < \infty$ and $\int_\Omega X^- \, \mathrm{d}\mu < \infty$, thus $\mu(X^+ = \infty) = \mu(X^- = \infty) = 0$. This implies that $\mu(X = \pm\infty) = \mu(\{X^+ = \infty\} \uplus \{X^- = \infty\}) = \mu(X^+ = \infty) + \mu(X^- = \infty) = 0 + 0 = 0$.

(c)
- (i) $\implies$ (iii): From $X \overset{\text{a.e.}}{=} Y$, we know $|X - Y| \overset{\text{a.e.}}{=} 0$. Hence, by [5.1.8] we have $\int_\Omega |X - Y| \, \mathrm{d}\mu = 0$.

- (iii) $\implies$ (ii): Assume $\int_\Omega |X - Y| \, \mathrm{d}\mu = 0$. Then, we have

$$0 \le \left| \int_A X \, \mathrm{d}\mu - \int_A Y \, \mathrm{d}\mu \right| = \left| \int_\Omega (X - Y)\mathbf{1}_A \, \mathrm{d}\mu \right| \overset{\text{(triangle inequality)}}{\le} \int_\Omega |X - Y|\mathbf{1}_A \, \mathrm{d}\mu$$

$$\le \int_\Omega |X - Y| \, \mathrm{d}\mu = 0,$$

  so $\int_A X \, \mathrm{d}\mu = \int_A Y \, \mathrm{d}\mu$.

- (ii) $\implies$ (i): Assume $\int_A X \, \mathrm{d}\mu = \int_A Y \, \mathrm{d}\mu$ for all $A \in \mathcal{F}$. Let $Z := X - Y$. Assume to the contrary that $\mu(Z \ne 0) = \mu(X \ne Y) > 0$. Then, we have $\mu(Z^+ > 0) > 0$ or $\mu(Z^- > 0) > 0$. WLOG, suppose that the former is the case, and let $A := \{Z^+ > 0\}$.
  On A we have $Z > 0$, so $Z\mathbf{1}_A = Z^+\mathbf{1}_A$. Hence,

$$0 = \int_A X \, \mathrm{d}\mu - \int_A Y \, \mathrm{d}\mu = \int_\Omega Z\mathbf{1}_A \, \mathrm{d}\mu = \int_\Omega Z^+\mathbf{1}_A \, \mathrm{d}\mu,$$

  which implies by [5.1.8] that $Z^+\mathbf{1}_A \overset{\text{a.e.}}{=} 0$, leading to a contradiction since we have $\mu(Z^+\mathbf{1}_A > 0) = \mu(\{Z^+ > 0\} \cap A) = \mu(A) > 0$.

$\square$

5.1.12 **Dominated convergence theorem.** Apart from *monotone convergence theorem*, another important result about Lebesgue integral is the *dominated convergence theorem (DCT)*, which provides another sufficient condition (involving a certain "domination") that allows the interchange of limit and integral. To prove DCT, we would need the *Fatou's lemma*, which suggests that interchanging limit inferior and integral would lead to an inequality in general, for nonnegative measurable functions.

Lemma 5.1.d (Fatou's lemma)**.** Let $X_n \in L_+$ for every $n \in \mathbb{N}$. Then

$$\int_\Omega \liminf_{n\to\infty} X_n \, \mathrm{d}\mu \le \liminf_{n\to\infty} \int_\Omega X_n \, \mathrm{d}\mu.$$

[Note: The limit inferior $\liminf_{n\to\infty} X_n$ can be defined as $\lim_{n\to\infty} \inf_{k\ge n} X_k$ (pointwise).]

Proof. First, note that $\inf_{k\ge n} X_k$ is in L_+ for all $n \in \mathbb{N}$. Also, notice that the sequence $\{\inf_{k\ge n} X_k\}_{n\in\mathbb{N}}$ is increasing. Hence, by monotone convergence theorem, we have

$$\int_\Omega \liminf_{n\to\infty} X_n \, \mathrm{d}\mu = \int_\Omega \lim_{n\to\infty} \inf_{k\ge n} X_k \, \mathrm{d}\mu \overset{\text{(MCT)}}{=} \lim_{n\to\infty} \int_\Omega \inf_{k\ge n} X_k \, \mathrm{d}\mu$$

$$= \liminf_{n\to\infty} \int_\Omega \inf_{k\ge n} X_k \, \mathrm{d}\mu \overset{(\inf_{k\ge n} X_k \le X_n)}{\underset{\text{(monotonicity)}}{\le}} \liminf_{n\to\infty} \int_\Omega X_n \, \mathrm{d}\mu.$$

$\square$

Theorem 5.1.e (Dominated convergence theorem (DCT))**.** Let $X_n \in L^1$ for every $n \in \mathbb{N}$, and $X : \Omega \to \mathbb{R}$ be measurable. If $X_n \to X$ a.e. and $|X_n| \le Y$ a.e. for all $n \in \mathbb{N}$, for some $Y \in L^1$ (domination), then $X \in L^1$ and $\lim_{n\to\infty} \int_\Omega X_n \, \mathrm{d}\mu = \int_\Omega X \, \mathrm{d}\mu$.

Proof. By assumption, we have $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all $\omega \in \Omega \setminus N_X$, and for each $n \in \mathbb{N}$, $|X_n(\omega)| \le Y(\omega)$ for all $\omega \in \Omega \setminus N_n$, where N_X and N_n's are null sets. Now let $N := N_X \cup \bigcup_{n=1}^\infty N_n$, which is also a null set.

By construction, we have $|X_n|\mathbf{1}_{N^c} \leq Y\mathbf{1}_{N^c}$ for every $n \in \mathbb{N}$, and thus letting $n \to \infty$ gives $|X|\mathbf{1}_{N^c} \leq Y\mathbf{1}_{N^c}$. Hence,

$$\int_\Omega |X|\,\mathrm{d}\mu = \underbrace{\int_\Omega |X|\mathbf{1}_N\,\mathrm{d}\mu}_{0 \text{ by } [5.1.10]} + \int_\Omega |X|\mathbf{1}_{N^c}\,\mathrm{d}\mu$$

$$\overset{\text{(monotonicity)}}{\leq} \int_\Omega Y\mathbf{1}_{N^c}\,\mathrm{d}\mu \overset{[5.1.11]\mathrm{c}}{\underset{(Y\mathbf{1}_{N^c}\overset{\text{a.e.}}{=}Y)}{=}} \int_\Omega Y\,\mathrm{d}\mu \overset{\text{(monotonicity)}}{\leq} \int_\Omega |Y|\,\mathrm{d}\mu < \infty,$$

which implies that $X \in L^1$. It then remains to show that $\lim_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu = \int_\Omega X\,\mathrm{d}\mu$.

Showing that $\int_\Omega X\,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu$. For every $n \in \mathbb{N}$, since $|X_n|\mathbf{1}_{N^c} \leq Y\mathbf{1}_{N^c}$, we have $(Y+X_n)\mathbf{1}_{N^c} \geq 0$ and $(Y-X_n)\mathbf{1}_{N^c} \geq 0$, which means that $(Y+X_n)\mathbf{1}_{N^c}, (Y-X_n)\mathbf{1}_{N^c} \in L_+$. Therefore, we get

$$\int_\Omega Y\,\mathrm{d}\mu + \int_\Omega X\,\mathrm{d}\mu \overset{[5.1.11]\mathrm{c}}{=} \int_\Omega (Y+X)\mathbf{1}_{N^c}\,\mathrm{d}\mu = \int_\Omega \liminf_{n\to\infty}(Y+X_n)\mathbf{1}_{N^c}\,\mathrm{d}\mu$$

$$\overset{\text{(Fatou's lemma)}}{\leq} \liminf_{n\to\infty} \int_\Omega (Y+X_n)\mathbf{1}_{N^c}\,\mathrm{d}\mu \overset{[5.1.11]\mathrm{c}}{=} \int_\Omega Y\,\mathrm{d}\mu + \liminf_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu,$$

which implies that $\int_\Omega X\,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu$.

Showing that $\int_\Omega X\,\mathrm{d}\mu \geq \limsup_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu$. Similarly, we have

$$\int_\Omega Y\,\mathrm{d}\mu - \int_\Omega X\,\mathrm{d}\mu \overset{[5.1.11]\mathrm{c}}{=} \int_\Omega (Y-X)\mathbf{1}_{N^c}\,\mathrm{d}\mu = \int_\Omega \liminf_{n\to\infty}(Y-X_n)\mathbf{1}_{N^c}\,\mathrm{d}\mu$$

$$\overset{\text{(Fatou's lemma)}}{\leq} \liminf_{n\to\infty} \int_\Omega (Y-X_n)\mathbf{1}_{N^c}\,\mathrm{d}\mu \overset{[5.1.11]\mathrm{c}}{=} \int_\Omega Y\,\mathrm{d}\mu + \liminf_{n\to\infty}\left(-\int_\Omega X_n\,\mathrm{d}\mu\right)$$

$$\overset{(\liminf_{n\to\infty} -f_n = -\limsup_{n\to\infty} f_n)}{=} \int_\Omega Y\,\mathrm{d}\mu - \limsup_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu,$$

which implies that $\int_\Omega X\,\mathrm{d}\mu \geq \limsup_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu$.

So, altogether we have

$$\limsup_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu \geq \liminf_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu \geq \int_\Omega X\,\mathrm{d}\mu \geq \limsup_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu \geq \liminf_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu,$$

which implies that $\int_\Omega X\,\mathrm{d}\mu = \liminf_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu = \limsup_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu = \lim_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu.$ $\square$

Remarks:

- *(Necessity of domination)* To illustrate the necessity of domination, consider the following example where domination fails and the DCT does not apply. Let $(\Omega, \mathcal{F}, \mu) = ([0,1], \mathcal{B}([0,1]), \lambda)$. Let $X_n := n^2 \mathbf{1}_{(0,1/n)} \geq 0$ for every $n \in \mathbb{N}$ and $X := 0$. Note that $X_n \to X$ pointwise (so a.e.), but $\lim_{n\to\infty} \int_\Omega X_n\,\mathrm{d}\mu = \lim_{n\to\infty} n^2\mu((0,1/n)) = \lim_{n\to\infty} n = \infty$, which is definitely not equal to $\int_\Omega X\,\mathrm{d}\mu = 0$.

- *(A simple corollary)* Under the conditions in DCT, we can also deduce that $\lim_{n\to\infty} \int_\Omega |X - X_n|\,\mathrm{d}\mu = 0$, which can be shown by noting that $|X - X_n| \to 0$ a.e., and $|X - X_n| \leq |X| + |X_n| \underset{\text{(a.e.)}}{\leq} |X| + |Y| \in L^1$.

We have previously shown the commutativity of integral and countable sum for Lebesgue integral of nonnegative measurable function. Using DCT, we can extend this result to Lebesgue integral of general measurable function (but with some extra condition).

Corollary 5.1.f (Commutativity of integral and countable sum). Let $X_n \in L^1$ for each $n \in \mathbb{N}$. If $\sum_{n=1}^\infty \int_\Omega |X_n|\,\mathrm{d}\mu < \infty$, then $\sum_{n=1}^\infty X_n \in L^1$ and $\int_\Omega \sum_{n=1}^\infty X_n\,\mathrm{d}\mu = \sum_{n=1}^\infty \int_\Omega X_n\,\mathrm{d}\mu$.

Proof. We first verify the conditions for applying DCT: (i) $\sum_{k=1}^{n} X_k \in L^1$ for each $n \in \mathbb{N}$, (ii) $\sum_{k=1}^{n} X_k \to \sum_{k=1}^{\infty} X_k$ a.e., and (iii) $|\sum_{k=1}^{n} X_k| \leq Y$ a.e. for every $n \in \mathbb{N}$, for some $Y \in L^1$:

(i) For each $n \in \mathbb{N}$, note that $\sum_{k=1}^{n} X_k$ is measurable and $\int_{\Omega} |\sum_{k=1}^{n} X_k| \, \mathrm{d}\mu \leq \int_{\Omega} \sum_{k=1}^{n} |X_k| \, \mathrm{d}\mu = \sum_{k=1}^{n} \int_{\Omega} |X_k| \, \mathrm{d}\mu < \infty$. Hence $\sum_{k=1}^{n} X_k \in L^1$.

(ii) It suffices to show that $\sum_{k=1}^{\infty} X_k$ converges a.e., which would then imply by definition that $\sum_{k=1}^{n} X_k \to \sum_{k=1}^{\infty} X_k$ a.e.

Note that $\int_{\Omega} \sum_{k=1}^{\infty} |X_k| \, \mathrm{d}\mu \overset{(|X_k| \in L_+ \ \forall k \in \mathbb{N})}{=} \sum_{k=1}^{\infty} \int_{\Omega} |X_k| \, \mathrm{d}\mu \overset{(\text{assumption})}{<} \infty$, and $\sum_{k=1}^{\infty} |X_k|$ is non-negative. So, we have $Y := \sum_{k=1}^{\infty} |X_k| \in L^1$. By [5.1.11]b, we have $Y(\omega) = \sum_{k=1}^{\infty} |X_k(\omega)| < \infty$ for all $\omega \in N^c$ where N is a null set. This means that $\sum_{k=1}^{\infty} X_k$ converges absolutely (and hence converges) a.e.

(iii) We consider the $Y \in L^1$ above. The domination follows by noting that for all $n \in \mathbb{N}$, we have $|\sum_{k=1}^{n} X_k| \leq \sum_{k=1}^{n} |X_k| \leq Y$.

Now, we apply the DCT to get $\sum_{k=1}^{\infty} X_k \in L^1$ and

$$\int_{\Omega} \sum_{k=1}^{\infty} X_k \, \mathrm{d}\mu = \lim_{n \to \infty} \int_{\Omega} \sum_{k=1}^{n} X_k \, \mathrm{d}\mu \overset{(\text{additivity})}{=} \lim_{n \to \infty} \sum_{k=1}^{n} \int_{\Omega} X_k \, \mathrm{d}\mu = \sum_{k=1}^{\infty} \int_{\Omega} X_k \, \mathrm{d}\mu.$$

$\square$

5.1.13 L^p **spaces.** In the definition of Lebesgue integral of measurable function, we have seen the notation L^1, which denotes the set of all integrable functions. Knowing that X is integrable iff $\int_{\Omega} |X| \, \mathrm{d}\mu < \infty$, we can express it as $L^1(\Omega, \mathcal{F}, \mu) = \{X : \Omega \to \mathbb{R} : X \text{ is } (\mathcal{F}, \mathcal{B}(\mathbb{R}))\text{-measurable and } \int_{\Omega} |X| \, \mathrm{d}\mu < \infty\}$.

More generally, we can define L^p *space* in the following. For all $p \in (0, \infty)$, let $\|X\|_p := (\int_{\Omega} |X|^p \, \mathrm{d}\mu)^{1/p}$ be the L^p **norm**, and then the L^p **space** is given by

$$L^p = L^p(\Omega, \mathcal{F}, \mu) := \{X : \Omega \to \mathbb{R} : X \text{ is } (\mathcal{F}, \mathcal{B}(\mathbb{R}))\text{-measurable and } \|X\|_p < \infty\}.$$

Remarks:

- *(Relationship between L^p and L^1 spaces)* For a $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable function X, we have $X \in L^p$ iff $\left(\int_{\Omega} |X|^p \, \mathrm{d}\mu\right)^{1/p} < \infty$ iff $\int_{\Omega} ||X|^p| \, \mathrm{d}\mu < \infty$ iff $|X|^p \in L^1$.

- *(Banach and Hilbert spaces)* For every $p \geq 1$, L^p can be shown to be a *Banach space*. Also, L^2 can be shown to be a *Hilbert space*. These spaces are important in the subject of *functional analysis* (but we shall not discuss them in details here).

- *(L^∞ space)* For $p = \infty$, we define the L^∞ **norm** by $\|X\|_\infty := \operatorname{ess\,sup} |X|$, where $\operatorname{ess\,sup} |X|$ is the **essential supremum** of $|X|$ which is given by $\operatorname{ess\,sup} |X| := \inf\{x \geq 0 : \mu(|X| > x) = 0\}$. For every $x \geq 0$ with $\mu(|X| > x) = 0$, x is an "essential" upper bound for $|X|$ in the sense that we have $|X| \leq x$ a.e. The essential supremum is then the least "essential" upper bound for $|X|$, which has the property that $|X| \leq \operatorname{ess\,sup} |X|$ a.e., and $\operatorname{ess\,sup} |X| \leq x$ for all "essential" upper bounds x for $|X|$ (meaning that the "inf" indeed coincides with "min").

  The L^∞ **space** is then defined by

  $$L^\infty = L^\infty(\Omega, \mathcal{F}, \mu) := \{X : \Omega \to \mathbb{R} : X \text{ is } (\mathcal{F}, \mathcal{B}(\mathbb{R}))\text{-measurable and } \|X\|_\infty < \infty\}.$$

  It can be shown that $X \in L^\infty$ iff X is bounded a.e.

5.1.14 **Properties for L^p spaces.**

(a) *(Hölder inequality)* Let $p, q \in [1, \infty]$ with $1/p + 1/q = 1$ (with the convention that $1/\infty := 0$) be *conjugate indices*. Then, $\|XY\|_1 \leq \|X\|_p \|Y\|_q$ for all measurable functions $X, Y : \Omega \to \mathbb{R}$.

Furthermore, if we have $p, q \in (1, \infty)$, $X \in L^p$, and $Y \in L^q$, then the equality holds iff $\alpha |X|^p \overset{\text{a.e.}}{=} \beta |Y|^q$ for some $\alpha, \beta \geq 0$.

[Note: The special case with $p = q = 2$ is known as the *Cauchy-Schwarz inequality*.]

(b) *(Minkowski's inequality)* Let $p \in [1, \infty]$. Then $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$ for all $X, Y \in L^p$.

(c) *(Jensen's inequality)* Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, $X : \Omega \to \mathbb{R}$ be a function in L^1, and φ be a convex function on $\mathbb{R}$. Then, $\varphi(\int_\Omega X \, d\mu) \leq \int_\Omega \varphi(X) \, d\mu$. [Note: In case φ is concave, we can apply the Jensen's inequality to $-\varphi$ which is convex. This would then yield the inequality $\varphi(\int_\Omega X \, d\mu) \geq \int_\Omega \varphi(X) \, d\mu$.]

(d) *(Relationship between L^p spaces and norms)* If $\mu(\Omega) < \infty$ (i.e., μ is finite) and $0 < p < q \leq \infty$, then $\|X\|_p \leq \|X\|_q \cdot \mu(\Omega)^{1/p - 1/q}$ (and hence $L^q \subseteq L^p$).

Remarks:

- If we have $\mu(\Omega) = 1$ (e.g., in a probability space), then this inequality gets simplified to $\|X\|_p \leq \|X\|_q$.
- With this inequality, we know that $\|X\|_q < \infty \implies \|X\|_p < \infty$ (as $\mu(\Omega)^{1/p - 1/q}$ must be finite), so $L^q \subseteq L^p$.

Proof.

(a) Omitted.

(b) Omitted.

(c) First, it can be shown that $\int_\Omega \varphi(X) \, d\mu \in (-\infty, \infty]$ (see, e.g., Klenke (2020, Theorem 7.9)). In case it is ∞, there is nothing to prove. So, we can assume henceforth that it is in $(-\infty, \infty)$, i.e., $\varphi(X) \in L^1$.

By the convexity of φ, for all $x_0 \in \mathbb{R}$, there exists a *supporting line* $x \mapsto \varphi(x_0) + m(x - x_0)$ of the graph of φ, satisfying that $\varphi(x) \geq \varphi(x_0) + m(x - x_0)$ for all $x \in \mathbb{R}$.
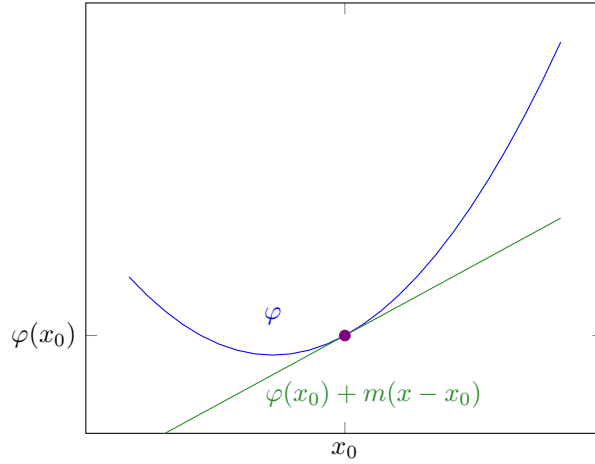
Let $\mathbb{E}[X] := \int_\Omega X \, d\mu$ (this is indeed how we define *expectation*; see Section 5.2). By setting $x_0 = \mathbb{E}[X]$, we have $\varphi(x) \geq \varphi(\mathbb{E}[X]) + m(x - \mathbb{E}[X])$ for all $x \in \mathbb{R}$. Thus, we have $\varphi(X) \geq \varphi(\mathbb{E}[X]) + m(X - \mathbb{E}[X])$. By monotonicity, we then get

$$\int_\Omega \varphi(X) \, d\mu \geq \int_\Omega \varphi(\mathbb{E}[X]) + m(X - \mathbb{E}[X]) \, d\mu = \varphi(\mathbb{E}[X]) \cdot \underbrace{\mu(\Omega)}_{1} + m \underbrace{\left(\int_\Omega X \, d\mu - \mathbb{E}[X] \right)}_{0} = \varphi \left(\int_\Omega X \, d\mu \right).$$

(d) For the case where $q = \infty$, the inequality becomes $\|X\|_p \leq \|X\|_\infty \cdot \mu(\Omega)^{1/p}$, which can be proven as follows:

$$\|X\|_p = \left(\int_\Omega |X|^p \, d\mu \right)^{1/p} \underset{\substack{(|X|^p \leq \|X\|_\infty^p \text{ a.e.})}}{\overset{\text{(monotonicity)}}{\leq}} \left(\int_\Omega \|X\|_\infty^p \, d\mu \right)^{1/p} = \|X\|_\infty \cdot \mu(\Omega)^{1/p}.$$

87

Next, for the case where $q < \infty$, we apply Hölder inequality with conjugate indices q/p and $q/(q-p)$ to get

$$\|X\|_p = \left(\int_\Omega |X|^p \cdot 1 \, d\mu \right)^{1/p} = \||X|^p \cdot 1\|_1^{1/p} \overset{\text{(Hölder)}}{\leq} \left(\||X|^p\|_{q/p} \|1\|_{q/(q-p)} \right)^{1/p}$$

$$= \left(\int_\Omega |X|^q \, d\mu \right)^{p/q \cdot 1/p} \left(\int_\Omega 1 \, d\mu \right)^{(q-p)/q \cdot 1/p} = \|X\|_q \cdot \mu(\Omega)^{1/p - 1/q}.$$

$\square$

5.2 Expectation

5.2.1 After studying Lebesgue integrals in Section 5.1, we now apply the concept in defining and computing expectations. Let $X : \Omega \to \bar{\mathbb{R}}$ be $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}))$-measurable. Then, the **expectation** or **mean** of X is given by $\mathbb{E}[X] := \int_\Omega X \, d\mu$, provided that the Lebesgue integral is defined (being finite or $\pm\infty$). [Note: If we want to emphasize the underlying measure μ, we may write $\mathbb{E}_\mu[X]$ instead.]

5.2.2 **Change of variables formula.** Although many general properties of expectations are derived based on this abstract form of expectation, in practice we rarely use such abstract formula for computing expectations. Rather than, we often utilize formulas involving the *mass function* or *density function*, which can be seen as special cases of the following *change of variables* formula.

Theorem 5.2.a (Change of variables). Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, $(\Omega', \mathcal{F}')$ be a measurable space, $X : \Omega \to \Omega'$ be measurable, and $h : \Omega' \to \mathbb{R}$ be measurable. If $\mathbb{E}[|h(X)|] < \infty$ (i.e., $h(X)$ is integrable), then we have

$$\int_\Omega h(X) \, d\mu = \int_{\Omega'} h \, d\mu_X$$

where μ_X denotes the push-forward measure of μ with respect to X.

Proof. We apply the standard argument to h here.

(1) Fix any indicator function $h(\omega') = \mathbf{1}_{A'}(\omega')$, where $A' \in \mathcal{F}$. Since $\mathbf{1}_{A'}(X(\omega)) = 1 \iff X(\omega) \in A' \iff \omega \in X^{-1}(A') \iff \mathbf{1}_{X^{-1}(A')}(\omega) = 1$, we have $\mathbf{1}_{A'}(X) = \mathbf{1}_{X^{-1}(A')}$. Hence,

$$\int_\Omega h(X) \, d\mu = \int_\Omega \mathbf{1}_{A'}(X) \, d\mu = \int_\Omega \mathbf{1}_{X^{-1}(A')} \, d\mu = \mu(X^{-1}(A')) = \mu_X(A') = \int_{\Omega'} \mathbf{1}_{A'} \, d\mu_X = \int_{\Omega'} h \, d\mu_X.$$

Then, by linearity, the equality also holds for every simple function h.

(2) Fix any $h \in L_+$. By Proposition 5.1.a, there exists a sequence $\{h_n\}$ of nonnegative simple functions on Ω' such that $h_n \nearrow h$ (pointwisely), which implies $h_n(X) \nearrow h(X)$. Thus,

$$\int_\Omega h(X) \, d\mu \overset{\text{(MCT)}}{=} \lim_{n \to \infty} \int_\Omega h_n(X) \, d\mu \overset{(1)}{=} \lim_{n \to \infty} \int_{\Omega'} h_n \, d\mu_X \overset{\text{(MCT)}}{=} \int_\Omega h \, d\mu_X.$$

(3) Fix any $h \in L^1$, and we have

$$\int_\Omega h(X) \, d\mu = \int_\Omega h^+(X) \, d\mu - \int_\Omega h^-(X) \, d\mu \overset{(2)}{=} \int_{\Omega'} h^+ \, d\mu_X - \int_{\Omega'} h^- \, d\mu_X = \int_{\Omega'} h \, d\mu_X.$$

$\square$

5.2.3 **Implications of change of variables formula.**

(a) *(Expectation as a Lebesgue-Stieltjes integral)* When applying the change of variables formula, often we are in the case where $\mu = \mathbb{P}$ is a *probability measure* and $(\Omega', \mathcal{F}') = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ (we then

change the notation $X \to \boldsymbol{X}$, as it is a random vector). Based on Theorem 3.2.a, sometimes we write $\mathrm{d}F$ in place of $\mathrm{d}\mathbb{P}_{\boldsymbol{X}}$, so the change of variables formula in this case would become

$$\mathbb{E}[h(\boldsymbol{X})] = \int_{\mathbb{R}^d} h \, \mathrm{d}F \overset{\text{(notation)}}{=} \int_{\mathbb{R}^d} h(\boldsymbol{x}) \, \mathrm{d}F(\boldsymbol{x}),$$

which takes a form that is more familiar to us. The integral $\int_{\mathbb{R}^d} h(\boldsymbol{x}) \, \mathrm{d}F(\boldsymbol{x})$ is known as the **Lebesgue-Stieltjes integral** of h with respect to F. In the special case where $F(\boldsymbol{x}) = \prod_{j=1}^d x_j$ for all $\boldsymbol{x} \in [\boldsymbol{a}, \boldsymbol{b}]$, it becomes the Lebesgue integral of h with respect to the Lebesgue measure λ.

From this expression, it is also clear that if we have $\boldsymbol{X}_1 \overset{\mathrm{d}}{=} \boldsymbol{X}_2$ (so both share the same distribution function F), then $\mathbb{E}[h(\boldsymbol{X}_1)] = \mathbb{E}[h(\boldsymbol{X}_2)]$ for every measurable function $h : \mathbb{R}^d \to \mathbb{R}$.

(b) *(Interpretations of Lebesgue-Stieltjes integral)* Based on the standard argument, one can derive the following expressions of $\mathbb{E}[h(\boldsymbol{X})]$:

- *(In terms of mass function)* If F is discrete with mass function f and countable support $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots\} \subseteq \mathbb{R}^d$, then

$$\mathbb{E}[h(\boldsymbol{X})] = \sum_{i=1}^{\infty} h(\boldsymbol{x}_i) f(\boldsymbol{x}_i).$$

- *(In terms of density function)* If F is absolutely continuous with density function f, then

$$\mathbb{E}[h(\boldsymbol{X})] = \int_{\mathbb{R}^d} h(\boldsymbol{x}) f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

These suggest the interpretations of the Lebesgue-Stieltjes integral $\int_{\mathbb{R}^d} h(\boldsymbol{x}) \, \mathrm{d}F(\boldsymbol{x})$ in the cases above.

5.2.4 **Relationship between Riemann-Stieltjes and Lebesgue-Stieltjes integrals.** The notation $\int_{[\boldsymbol{a}, \boldsymbol{b}]} h(\boldsymbol{x}) \, \mathrm{d}F(\boldsymbol{x})$ is also used for the *Riemann-Stieltjes integral* with integrand $h : [\boldsymbol{a}, \boldsymbol{b}] \to \mathbb{R}$ and integrator $F : [\boldsymbol{a}, \boldsymbol{b}] \to \mathbb{R}$. This is because such Riemann-Stieltjes integral indeed always coincides with the corresponding Lebesgue-Stieltjes integral, provided that the former exists.

The **Riemann-Stieltjes integral** $\int_{[\boldsymbol{a}, \boldsymbol{b}]} h(\boldsymbol{x}) \, \mathrm{d}F(\boldsymbol{x})$ is defined as the limit of the *Riemann-Stieltjes sums*:

$$\int_{[\boldsymbol{a}, \boldsymbol{b}]} h(\boldsymbol{x}) \, \mathrm{d}F(\boldsymbol{x}) := \lim_{n_1, \dots, n_d \to \infty} \sum_{i_d=1}^{n_d} \cdots \sum_{i_1=1}^{n_1} h(\xi_{1 i_1}, \dots, \xi_{d i_d}) \Delta_{(\boldsymbol{x}_{\boldsymbol{i}-_d 1}, \boldsymbol{x}_{\boldsymbol{i}}]} F$$

where $(\boldsymbol{x}_{\boldsymbol{i}-_d 1}, \boldsymbol{x}_{\boldsymbol{i}}]$ denotes $\left(\begin{bmatrix} x_{1, i_1-1} \\ \vdots \\ x_{d, i_d-1} \end{bmatrix}, \begin{bmatrix} x_{1 i_1} \\ \vdots \\ x_{d i_d} \end{bmatrix} \right]$, with $a_j = x_{j0} < x_{j1} < \cdots < x_{jn} = b_j$ and $\xi_{j i_j} \in [x_{j, i_j-1}, x_{j, i_j}]$ for all $j = 1, \dots, d$. Here, the measure associated to the F-volume is the Lebesgue-Stieltjes measure λ_F.

[Note: In the special case with $F(\boldsymbol{x}) = \prod_{j=1}^d x_j$ for all $\boldsymbol{x} \in [\boldsymbol{a}, \boldsymbol{b}]$, it reduces to the **Riemann integral** $\int_{[\boldsymbol{a}, \boldsymbol{b}]} h(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$. Here, the measure associated to the F-volume is the Lebesgue measure λ.]

The following relates the Riemann-Stieltjes and Lebesgue-Stieltjes integrals:

Proposition 5.2.b. Let $h : [\boldsymbol{a}, \boldsymbol{b}] \to \mathbb{R}$ be a function.

(a) If h is bounded, then h is Riemann-Stieltjes integrable with respect to F on $[\boldsymbol{a}, \boldsymbol{b}]$ iff h is continuous λ_F-a.e. on $[\boldsymbol{a}, \boldsymbol{b}]$.

(b) If h is Riemann-Stieltjes integrable with respect to F on $[\boldsymbol{a}, \boldsymbol{b}]$, then h is also Lebesgue-Stieltjes integrable with respect to λ_F, with the Lebesgue-Stieltjes integral $\int_{[\boldsymbol{a}, \boldsymbol{b}]} h \, \mathrm{d}\lambda_F$ being equal to the Riemann-Stieltjes integral $\int_{[\boldsymbol{a}, \boldsymbol{b}]} h \, \mathrm{d}F$.

Proof. See ter Horst (1984). □

[Note: In the context of computing expectation of $h(\boldsymbol{X})$ where $\boldsymbol{X}$ is a random vector under a probability measure $\mathbb{P}$, since we have $\lambda_F = \mathbb{P}_{\boldsymbol{X}}$ by Theorem 2.5.a, the Lebesgue-Stieltjes integral $\int_{[\boldsymbol{a},\boldsymbol{b}]} h \, \mathrm{d}\lambda_F$ can indeed be expressed as $\int_{[\boldsymbol{a},\boldsymbol{b}]} h(\boldsymbol{x}) \, \mathrm{d}\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{x}) \overset{\text{(notation)}}{=} \int_{[\boldsymbol{a},\boldsymbol{b}]} h(\boldsymbol{x}) \, \mathrm{d}F(\boldsymbol{x})$. It shares the same notation as the one for Riemann integral, and indeed can be computed as a Riemann-Stieltjes integral (if exists) by Proposition 5.2.b. This serves as the main tool for computing such Lebesgue-Stieltjes integral.]

A famous example concerning the difference between Riemann and Lebesgue integral is the *Dirichlet function* $h(x) = \mathbf{1}_{\mathbb{Q}}(x)$ for all $x \in [0,1]$. Since it is bounded and discontinuous on $[0,1]$ (with $\lambda([0,1]) = 1$), it follows by Proposition 5.2.b that h is not Riemann integrable. However, it is a simple function and is Lebesgue integrable: We have $\int_{[0,1]} h \, \mathrm{d}\lambda = 1 \cdot \lambda(\mathbb{Q}) + 0 \cdot \lambda([0,1] \setminus \mathbb{Q}) = 0$.

5.2.5 **Fubini-Tonelli theorem.** The main tool for computing multivariate integrals is the *Fubini-Tonelli theorem*; perhaps you have seen some special cases of this result in your previous multivariable calculus course.

Theorem 5.2.c (Fubini-Tonelli). Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be two measure spaces where μ_1 and μ_2 are σ-finite, and consider the product space $(\Omega, \mathcal{F}, \mu) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mu_1 \times \mu_2)$. Let $h : \Omega \to \bar{\mathbb{R}}$ be a $(\mathcal{F}_1 \otimes \mathcal{F}_2, \mathcal{B}(\bar{\mathbb{R}}))$-measurable function.

If $h \in L_+(\Omega, \mathcal{F}, \mu)$ *(Tonelli)* or $h \in L^1(\Omega, \mathcal{F}, \mu)$ *(Fubini)*, then $\omega_1 \mapsto \int_{\Omega_2} h(\omega_1, \omega_2) \, \mathrm{d}\mu_2(\omega_2)$ and $\omega_2 \mapsto \int_{\Omega_1} h(\omega_1, \omega_2) \, \mathrm{d}\mu_1(\omega_1)$ are $\mathcal{F}_1$-measurable and $\mathcal{F}_2$-measurable respectively, and

$$\int_\Omega h \, \mathrm{d}\mu = \int_{\Omega_1} \left(\int_{\Omega_2} h(\omega_1, \omega_2) \, \mathrm{d}\mu_2(\omega_2) \right) \mathrm{d}\mu_1(\omega_1) = \int_{\Omega_2} \left(\int_{\Omega_1} h(\omega_1, \omega_2) \, \mathrm{d}\mu_1(\omega_1) \right) \mathrm{d}\mu_2(\omega_2).$$

Proof. Omitted. □

Remarks:

- *Tonelli's theorem* refers to the result with condition $h \in L_+$, and *Fubini's theorem* refers to the result with $h \in L^1$. The result above is a combination of these two theorems.

- The Fubini-Tonelli theorem gives us conditions under which changing the order of integration is allowed. The result can be extended to the case where finitely many (not just two) integrals are involved by induction.

- A typical procedure in applying the Fubini-Tonelli theorem is as follows:
  (1) Apply *Tonelli's theorem* to $|h| \in L_+$ for calculating $\int_{\mathbb{R}^d} |h(\boldsymbol{x})| \, \mathrm{d}F(\boldsymbol{x})$.
  (2) If $\int_{\mathbb{R}^d} |h(\boldsymbol{x})| \, \mathrm{d}F(\boldsymbol{x}) < \infty$, then $h \in L^1$. After that, apply *Fubini's theorem* to $h \in L^1$ for calculating $\int_{\mathbb{R}^d} h(\boldsymbol{x}) \, \mathrm{d}F(\boldsymbol{x})$ (the actual integral of interest).

5.2.6 **Expectation of products under independence.** With the Fubini-Tonelli theorem, we can establish the following result which can often substantially simplify the calculations of expectations under independence.

Proposition 5.2.d. Let $X_1, \ldots, X_d \in L^1$ be independent. Then, we have $\mathbb{E}[X_1 \cdots X_d] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_d]$.

Proof. First note that under the independence of $X_1, \ldots, X_d$, by Theorem 4.2.c the joint distribution function F of $X_1, \ldots, X_d$ is given by $F(\boldsymbol{x}) = \prod_{j=1}^d F_j(x_j)$, where F_j is the distribution function of X_j for each $j = 1, \ldots, d$. Interpreting F and each F_j as distributions (push-forward measures), we can then see that F can be seen as the product measure $\prod_{j=1}^d F_j$.

Then, we follow the typical procedure in applying the Fubini-Tonelli theorem suggested above.

(1) By *Tonelli's theorem*, we have

$$\mathbb{E}[|X_1 \cdots X_d|] = \mathbb{E}[|X_1| \cdots |X_d|] = \int_{\mathbb{R}^d} |x_1| \cdots |x_d| \, \mathrm{d}F(\boldsymbol{x}) \stackrel{\text{(Tonelli)}}{=} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} |x_1| \cdots |x_d| \, \mathrm{d}F_1(x_1) \cdots \mathrm{d}F_d(x_d)$$

$$\stackrel{\text{(taking out constants)}}{=} \prod_{j=1}^{d} \int_{\mathbb{R}} |x_j| \, \mathrm{d}F_j(x_j) = \mathbb{E}[|X_1|] \cdots \mathbb{E}[|X_d|] \stackrel{(X_1,\ldots,X_d \in L^1)}{<} \infty.$$

Hence, $X_1 \cdots X_d \in L^1$.

(2) Applying *Fubini's theorem* gives

$$\mathbb{E}[X_1 \cdots X_d] = \int_{\mathbb{R}^d} x_1 \cdots x_d \, \mathrm{d}F(\boldsymbol{x}) \stackrel{\text{(Fubini)}}{=} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} x_1 \cdots x_d \, \mathrm{d}F_1(x_1) \cdots \mathrm{d}F_d(x_d)$$

$$= \prod_{j=1}^{d} \int_{\mathbb{R}} x_j \, \mathrm{d}F_j(x_j) = \mathbb{E}[X_1] \cdots \mathbb{E}[X_d].$$

$\square$

5.2.7 Expectation formula in terms of quantile/survival function. While the change of variables formula in Theorem 5.2.a is a common tool for computing expectations, it is not the only one. Below, we will provide two alternative methods for computing expectations: one is based on *quantile function* and another is based on *survival function*.

Proposition 5.2.e. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $X \sim F$, and $X \in L^1$. Then, $\mathbb{E}[X] = \int_0^1 F^{-1}(u) \, \mathrm{d}u = \int_0^\infty \bar{F}(x) \, \mathrm{d}x - \int_{-\infty}^0 F(x) \, \mathrm{d}x$, where F^{-1} and $\bar{F}$ denote quantile function and survival function respectively.

[Note: In the special case where X is nonnegative, we can simplify the formula in terms of survival function to $\mathbb{E}[X] = \int_0^\infty \bar{F}(x) \, \mathrm{d}x$, which is very useful in *survival analysis* (e.g., X represents lifetime).]

Proof. By quantile transform, we have $X \stackrel{\mathrm{d}}{=} F^{-1}(U)$ with $U \sim \mathrm{U}(0,1)$. Hence, $\mathbb{E}[X] = \mathbb{E}[F^{-1}(U)] = \int_0^1 F^{-1}(u) \, \mathrm{d}u$, establishing the formula in terms of quantile function. The formula in terms of survival function can be shown analytically through integration by parts, but is best understood geometrically as follows:
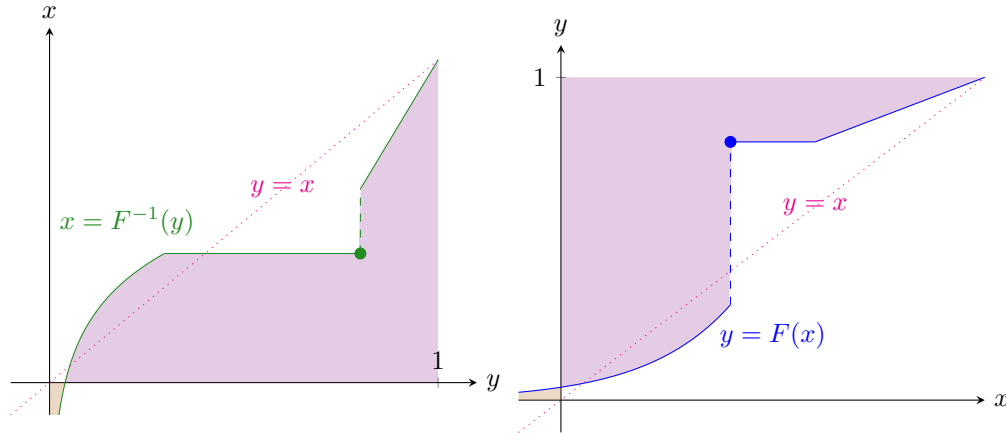
For the graph of distribution function (right), the violet area corresponds to the integral $\int_0^\infty \bar{F}(x) \, \mathrm{d}x = \int_0^\infty 1 - F(x) \, \mathrm{d}x$, while the brown area corresponds to the integral $\int_{-\infty}^0 F(x) \, \mathrm{d}x$. $\square$

5.3   Variance, Covariance, and Correlation

5.3.1 Apart from expectation, you should have also learnt other kinds of probabilistic quantities like *variance*, *covariance*, and *correlation* in your first probability course. Here, we will briefly review them and study some related results.

5.3.2 **Definitions.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X, Y \in L^2$. Then the **variance** and **standard deviation** of X are $\mathrm{Var}(X) := \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$ and $\mathrm{SD}(X) := \sqrt{\mathrm{Var}(X)}$ respectively. The **covariance** and **correlation** of X and Y are $\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ and $\mathrm{Corr}(X, Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$ respectively.

[Note: The correlation $\mathrm{Corr}(X, Y)$ serves as a measure of *linear* dependence between X and Y. Some other measures of dependence between X and Y are available, such as the **Kendall's tau** $\tau = \mathrm{Corr}\big(\mathbf{1}_{\{F_X(X) \leq F_X(X')\}}, \mathbf{1}_{\{F_X(X) \leq F_X(X')\}}\big)$ and the **Spearman's rho** $\rho_S = \mathrm{Corr}(F_X(X), F_Y(Y))$, where F_X and F_Y are distribution functions of X and Y respectively, and X', Y' are *independent copies* of X, Y respectively, i.e., $X, X' \overset{\text{iid}}{\sim} F_X$ and $Y, Y' \overset{\text{iid}}{\sim} F_Y$. These measures can also capture some nonlinear dependence between X and Y.

While the expression of Kendall's tau is more complicated, it turns out to be easier to compute than Spearman's rho.]

5.3.3 **Properties of variance, covariance, and correlation.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X, Y \in L^2$.

(a) *(Well-known formula for variance)* $\mathrm{Var}(X) = \mathbb{E}\big[X^2\big] - \mathbb{E}[X]^2$.

(b) *(Well-known formula for covariance)* $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

(c) $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.

(d) *(Symmetry)* $\mathrm{Cov}(Y, X) = \mathrm{Cov}(X, Y)$.

(e) $\mathrm{Cov}(X, c) = 0$ for all $c \in \mathbb{R}$.

(f) *(Criterion for zero variance)* $\mathrm{Var}(X) = 0$ iff $X \overset{\text{a.s.}}{=} \mathbb{E}[X]$.

(g) $\mathrm{Var}(aX + bY) = a^2 \mathrm{Var}(X) + b^2 \mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y)$ for all $a, b \in \mathbb{R}$.
[Note: If we have $Y \equiv 1$, then this formula reduces to $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$.]

(h) *(Independence implies uncorrelated)* If X and Y are independent, then $\mathrm{Cov}(X, Y) = \mathrm{Corr}(X, Y) = 0$ (in this case we say that X and Y are **uncorrelated**).
[⚠ Warning: It should be well known that the converse does <u>not</u> hold. For instance, if we have $X \sim \mathrm{U}(-1, 1)$ and $Y = X^2$, then $\mathrm{Cov}(X, Y) = \mathbb{E}\big[X^3\big] - \mathbb{E}[X]\mathbb{E}\big[X^2\big] = 0 - 0 = 0$ (and also $\mathrm{Corr}(X, Y) = 0$) but definitely X and Y are not independent.]

Proof.

(a) $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}\Big[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\Big] = \mathbb{E}\big[X^2\big] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}\big[X^2\big] = \mathbb{E}\big[X^2\big] - \mathbb{E}[X]^2$.

(b) $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY - \mathbb{E}[X]Y - X\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]] = \mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

(c) $\mathrm{Cov}(X, X) \overset{\text{(b)}}{=} \mathbb{E}\big[X^2\big] - \mathbb{E}[X]^2 \overset{\text{(a)}}{=} \mathrm{Var}(X)$.

(d) $\mathrm{Cov}(Y, X) \overset{\text{(b)}}{=} \mathbb{E}[YX] - \mathbb{E}[Y]\mathbb{E}[X] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \overset{\text{(b)}}{=} \mathrm{Cov}(X, Y)$.

(e) $\mathrm{Cov}(X, c) \overset{\text{(b)}}{=} \mathbb{E}[cX] - \mathbb{E}[c]\mathbb{E}[X] = c\mathbb{E}[X] - c\mathbb{E}[X] = 0$.

(f) $\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = 0 \overset{[5.1.8]}{\Longleftrightarrow} (X - \mathbb{E}[X])^2 \overset{\text{a.s.}}{=} 0 \iff X \overset{\text{a.s.}}{=} \mathbb{E}[X]$.

(g) $\mathrm{Var}(aX + bY) = \mathbb{E}\Big[\big((aX + bY) - \mathbb{E}[aX + bY]\big)^2\Big] = \mathbb{E}\Big[\big(a(X - \mathbb{E}[X]) + b(Y - \mathbb{E}[Y])\big)^2\Big] = \mathbb{E}\big[a^2(X - \mathbb{E}[X])^2\big] + \mathbb{E}[2ab(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] + \mathbb{E}\big[b^2(Y - \mathbb{E}[Y])^2\big] = a^2 \mathrm{Var}(X) + 2ab\,\mathrm{Cov}(X, Y) + b^2 \mathrm{Var}(Y)$.

(h) Since X and Y are independent, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ by Proposition 5.2.d. Hence, $\mathrm{Cov}(X,Y) \overset{\text{(b)}}{=} \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$, which implies by definition that $\mathrm{Corr}(X,Y) = 0$.

$\square$

5.3.4 **Cauchy-Schwarz inequality.** We have already mentioned the *Cauchy-Swartz inequality* in [5.1.14]a, as a special case of the Hölder inequality. Now, we will express it in terms of the probabilistic quantities introduced here and prove the inequality, together with an associated result about correlation.

Proposition 5.3.a (Cauchy-Swartz inequality and bounds on correlation)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X, Y \in L^2$. Then,

(a) *(Cauchy-Swartz inequality)* $|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$ where the equality holds iff $Y \overset{\text{a.s.}}{=} mX$ for some $m \in \mathbb{R}$.

(b) *(Bounds on correlation)* $-1 \leq \mathrm{Corr}(X,Y) \leq 1$, and we have $\mathrm{Corr}(X,Y) = 1$ (-1 resp.) iff $Y \overset{\text{a.s.}}{=} mX + c$ where $m > 0$ ($m < 0$ resp.).

Proof.

(a) Let $Z_t := tX + Y$ for every $t \in \mathbb{R}$. Then, we have $0 \leq \mathbb{E}[Z_t^2] = t^2\mathbb{E}[X^2] + 2t\mathbb{E}[XY] + \mathbb{E}[Y^2] =: at^2 + bt + c$. So $at^2 + bt + c$ is a polynomial in t with at most one root (as it is always nonnegative). Hence, its discriminant satisfies $\Delta = b^2 - 4ac \leq 0$, which implies that $4\mathbb{E}[XY]^2 - 4\mathbb{E}[X^2]\mathbb{E}[Y^2] \leq 0$, and hence $|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$.

Also, we have

$$
\begin{aligned}
\text{the equality holds} &\iff b^2 - 4ac = 0 \\
&\iff at^2 + bt + c = 0 \text{ for some unique } t \in \mathbb{R} \\
&\iff \mathbb{E}[Z_t^2] = 0 \text{ for such } t \in \mathbb{R} \\
&\overset{[5.1.8]}{\iff} Z_t \overset{\text{a.s.}}{=} 0 \text{ for such } t \in \mathbb{R} \\
&\iff Y \overset{\text{a.s.}}{=} -tX \text{ for such } t \in \mathbb{R}.
\end{aligned}
$$

(b) Applying the Cauchy-Swartz inequality to $\widetilde{X} := X - \mathbb{E}[X]$ and $\widetilde{Y} := Y - \mathbb{E}[Y]$ yields

$$
|\mathrm{Cov}(X,Y)| \overset{\text{(by construction)}}{=} \left|\mathbb{E}[\widetilde{X}\widetilde{Y}]\right| \overset{\text{(Cauchy-Swartz)}}{\leq} \left(\mathbb{E}[\widetilde{X}^2]\mathbb{E}[\widetilde{Y}^2]\right)^{1/2} \overset{\text{(by construction)}}{=} (\mathrm{Var}(X)\mathrm{Var}(Y))^{1/2},
$$

where the equality holds iff $\widetilde{Y} \overset{\text{a.s.}}{=} m\widetilde{X}$ for some $m \in \mathbb{R}$ iff $Y \overset{\text{a.s.}}{=} mX + c$ for such $m \in \mathbb{R}$, with $c = \mathbb{E}[Y] - m\mathbb{E}[X]$.

This implies that $|\mathrm{Corr}(X,Y)| = \frac{|\mathrm{Cov}(X,Y)|}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \leq 1$, and the equality holds (i.e., $\mathrm{Corr}(X,Y) = \pm 1$) iff $Y \overset{\text{a.s.}}{=} mX + c$; in such case, we have $\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X)mX+c}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(mX+c)}} = \frac{m\mathrm{Var}(X)}{\sqrt{m^2\mathrm{Var}(X)^2}} = \frac{m}{|m|}$, which equals 1 ($-1$ resp.) iff $m > 0$ ($m < 0$ resp.).

$\square$

5.3.5 **Hoeffding's lemma.** Another less well-known result about covariance is the *Hoeffding's lemma*, which can be used to determine the maximum/minimum possible values of correlations given the two marginals (it turns out that some values of correlation *are not attainable* for certain margins!).

Lemma 5.3.b (Hoeffding's lemma)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $(X_1, X_2) \sim F$ with margins F_1, F_2 respectively and $X_1, X_2 \in L^2$. Then,

$$
\mathrm{Cov}(X_1, X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(x_1, x_2) - F_1(x_1)F_2(x_2) \, \mathrm{d}x_1 \, \mathrm{d}x_2.
$$

Proof. Let (X_1', X_2') be an *independent copy* of (X_1, X_2), i.e., $(X_1, X_2), (X_1', X_2') \overset{\text{iid}}{\sim} F$. The result then follows from the following chain of equalities:

$$2\operatorname{Cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] + \mathbb{E}[(X_1' - \mathbb{E}[X_1'])(X_2' - \mathbb{E}[X_2'])]$$

$$\overset{\text{(independence)}}{=} \mathbb{E}\big[\big((X_1 - \mathbb{E}[X_1]) - (X_1' - \mathbb{E}[X_1'])\big) \cdot \big((X_2 - \mathbb{E}[X_2]) - (X_2' - \mathbb{E}[X_2'])\big)\big]$$

$$\overset{(\mathbb{E}[X_j] = \mathbb{E}[X_j'])}{=} \mathbb{E}[(X_1 - X_1')(X_2 - X_2')]$$

$$= \mathbb{E}\left[\left(\int_{-\infty}^{\infty} \mathbf{1}_{\{X_1' \leq x_1\}} - \mathbf{1}_{\{X_1 \leq x_1\}}\, \mathrm{d}x_1\right)\left(\int_{-\infty}^{\infty} \mathbf{1}_{\{X_2' \leq x_2\}} - \mathbf{1}_{\{X_2 \leq x_2\}}\, \mathrm{d}x_1\right)\right]$$

$$= \mathbb{E}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\mathbf{1}_{\{X_1' \leq x_1\}} - \mathbf{1}_{\{X_1 \leq x_1\}}\right)\left(\mathbf{1}_{\{X_2' \leq x_2\}} - \mathbf{1}_{\{X_2 \leq x_2\}}\right)\mathrm{d}x_1\, \mathrm{d}x_2\right]$$

$$\overset{\text{(Fubini)}}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{E}\left[\left(\mathbf{1}_{\{X_1' \leq x_1\}} - \mathbf{1}_{\{X_1 \leq x_1\}}\right)\left(\mathbf{1}_{\{X_2' \leq x_2\}} - \mathbf{1}_{\{X_2 \leq x_2\}}\right)\right]\mathrm{d}x_1\, \mathrm{d}x_2$$

$$= 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x_1, x_2) - F_1(x_1)F_2(x_2) - F_1(x_1)F_2(x_2) + F(x_1, x_2)\, \mathrm{d}x_2$$

$$= 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x_1, x_2) - F_1(x_1)F_2(x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2.$$

$\square$

[Note: To apply Hoeffding's lemma for determining the maximum/minimum possible values of correlations given two marginals, we can use the Sklar's theorem (Theorem 4.3.b) in conjunction with the Fréchet-Hoeffding bounds (Theorem 4.3.e), which provide bounds for $F(x_1, x_2)$, namely $W(F_1(x_1), F_2(x_2)) \leq F(x_1, x_2) \leq M(F_1(x_1), F_2(x_2))$, by writing $F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$ where C is a copula of F.

Once the margins are fixed, the upper and lower bounds here are fixed (so do the variances of X_1 and X_2). Hence, we can then bound the covariance (and thus also correlation) by the Hoeffding lemma as follows:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W(F_1(x_1), F_2(x_2)) - F_1(x_1)F_2(x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 \leq \operatorname{Cov}(X_1, X_2) \leq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} M(F_1(x_1), F_2(x_2)) - F_1(x_1)F_2(x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2.$$

]

5.4 Multivariate Versions of Probabilistic Quantities

5.4.1 After studying some probabilistic quantities, here we are going to study their *multivariate versions* (serving as generalizations of the concepts to higher dimensions), which may have been discussed in your previous linear regression course.

5.4.2 **Definitions.** Let $X_1, \ldots, X_d \in L^1$. Then, $\mathbb{E}[\boldsymbol{X}] := (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])$ is the **mean vector** or **expectation** of $\boldsymbol{X}$.

Let $X_1, \ldots, X_d, Y_1, \ldots, Y_p \in L^2$. Then, $\operatorname{Cov}(\boldsymbol{X}, \boldsymbol{Y}) := \big[\operatorname{Cov}(X_i, Y_j)\big]_{i,j=1}^{d,p}$ is the **cross-variance matrix** of $\boldsymbol{X} = (X_1, \ldots, X_d)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$. The **covariance matrix** of $\boldsymbol{X}$ is $\operatorname{Cov}(\boldsymbol{X}) := \operatorname{Cov}(\boldsymbol{X}, \boldsymbol{X})$ and the **correlation matrix** of $\boldsymbol{X}$ is $\operatorname{Corr}(\boldsymbol{X}) := \big[\operatorname{Corr}(X_i, Y_j)\big]_{i,j=1}^{d,d}$.

[Note: Here, the notation $\big[a_{ij}\big]_{i,j=1}^{m,n}$ denotes the $m \times n$ matrix $\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$.]

5.4.3 **Properties of multivariate probabilistic quantities.**

(a) *(Linearity)* Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ with $X_1, \ldots, X_d \in L^1$. Then, $\mathbb{E}[A\boldsymbol{X} + \boldsymbol{b}] = A\mathbb{E}[\boldsymbol{X}] + \boldsymbol{b}$, for all A and $\boldsymbol{b}$ such that the matrix-vector operations are defined.

(b) *(Effects of linear operations on cross-covariance matrix)* Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$ with $X_1, \ldots, X_d, Y_1, \ldots, Y_p \in L^2$. Then, $\mathrm{Cov}(A\boldsymbol{X} + \boldsymbol{b}, C\boldsymbol{Y} + \boldsymbol{d}) = A\,\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})C^T$ for all $A, \boldsymbol{b}, C, \boldsymbol{d}$ such that the matrix-vector operations are defined.

(c) *(Covariance matrix of a sum)* $\boldsymbol{X} = (X_1, \ldots, X_d)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ with $X_1, \ldots, X_d, Y_1, \ldots, Y_d \in L^2$. Then, $\mathrm{Cov}(\boldsymbol{X} + \boldsymbol{Y}) = \mathrm{Cov}(\boldsymbol{X}) + \mathrm{Cov}(\boldsymbol{Y}) + \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) + \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})^T$.

Proof. We shall use the notations $\boldsymbol{a}_j$ and A_{ij} to denote the jth entry of the vector $\boldsymbol{a}$ and the (i, j)th entry of the matrix A respectively.

(a) Entrywise, we have

$$\mathbb{E}[A\boldsymbol{X} + \boldsymbol{b}]_j = \mathbb{E}\left[\sum_{k=1}^{d} a_{jk}X_k + b_j\right] = \sum_{k=1}^{d} a_{jk}\mathbb{E}[X_k] + b_j = (A\mathbb{E}[\boldsymbol{X}] + \boldsymbol{b})_j.$$

(b) Entrywise, we have

$$\mathrm{Cov}(A\boldsymbol{X} + \boldsymbol{b}, C\boldsymbol{Y} + \boldsymbol{d})_{ij} = \mathrm{Cov}((A\boldsymbol{X} + \boldsymbol{b})_i, (C\boldsymbol{Y} + \boldsymbol{d})_j) = \mathrm{Cov}\left(\sum_{k=1}^{d} a_{ik}X_k + b_i, \sum_{\ell=1}^{p} c_{j\ell}Y_\ell + d_j\right)$$

$$= \mathbb{E}\left[\left(\sum_{k=1}^{d} a_{ik}(X_k - \mathbb{E}[X_k])\right)\left(\sum_{\ell=1}^{p} c_{j\ell}(Y_\ell - \mathbb{E}[Y_\ell])\right)\right]$$

$$= \mathbb{E}\left[\sum_{k=1}^{d}\sum_{\ell=1}^{p} a_{ik}(X_k - \mathbb{E}[X_k])c_{j\ell}(Y_\ell - \mathbb{E}[Y_\ell])\right]$$

$$= \sum_{k=1}^{d}\sum_{\ell=1}^{p} a_{ik}c_{j\ell}\mathbb{E}[(X_k - \mathbb{E}[X_k])(Y_\ell - \mathbb{E}[Y_\ell])] = \sum_{k=1}^{d}\sum_{\ell=1}^{p} a_{ik}c_{j\ell}\,\mathrm{Cov}(X_k, Y_\ell)$$

$$= (A\,\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})C^T)_{ij}.$$

(c) Entrywise, we have

$$\mathrm{Cov}(\boldsymbol{X} + \boldsymbol{Y})_{ij} = \mathrm{Cov}((\boldsymbol{X} + \boldsymbol{Y})_i, (\boldsymbol{X} + \boldsymbol{Y})_j) = \mathrm{Cov}(X_i + Y_i, X_j + Y_j)$$

$$= \mathbb{E}[(X_i + Y_i - \mathbb{E}[X_i + Y_i])(X_j + Y_j - \mathbb{E}[X_j + Y_j])]$$

$$= \mathbb{E}[(X_i - \mathbb{E}[X_i] + Y_i - \mathbb{E}[Y_i])(X_j - \mathbb{E}[X_j] + Y_j\mathbb{E}[Y_j])]$$

$$= \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] + \mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_j - \mathbb{E}[Y_j])]$$

$$\quad + \mathbb{E}[(Y_i - \mathbb{E}[Y_i])(X_j - \mathbb{E}[X_j])] + \mathbb{E}[(Y_i - \mathbb{E}[Y_i])(Y_j - \mathbb{E}[Y_j])]$$

$$= (\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{X}))_{ij} + (\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}))_{ij} + (\mathrm{Cov}(\boldsymbol{Y}, \boldsymbol{X}))_{ij} + (\mathrm{Cov}(\boldsymbol{Y}, \boldsymbol{Y}))_{ij}$$

$$= (\mathrm{Cov}(\boldsymbol{X}))_{ij} + (\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}))_{ij} + (\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}))_{ji} + (\mathrm{Cov}(\boldsymbol{Y}))_{ij}$$

$$= (\mathrm{Cov}(\boldsymbol{X}) + \mathrm{Cov}(\boldsymbol{Y}) + \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) + \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})^T)_{ij}.$$

$\square$

5.4.4 **Implications of the properties.** The properties in [5.4.3] are quite general and can be used to derive some more familiar properties as special cases:

(a) *([5.4.3]a with $A = \boldsymbol{a}^T$)* $\mathbb{E}[\boldsymbol{a}^T\boldsymbol{X}] = \boldsymbol{a}^T\mathbb{E}[\boldsymbol{X}]$.

(b) *([5.4.3]b with $C = A$ and $\boldsymbol{d} = \boldsymbol{b}$)* $\mathrm{Cov}(A\boldsymbol{X} + \boldsymbol{b}) = A\,\mathrm{Cov}(\boldsymbol{X})A^T$

(c) *([5.4.3]b with $C = A = \boldsymbol{a}^T$ and $\boldsymbol{d} = \boldsymbol{b} = \boldsymbol{0}$)* $\mathrm{Var}(\boldsymbol{a}^T\boldsymbol{X}) = \boldsymbol{a}^T\,\mathrm{Cov}(\boldsymbol{X})\boldsymbol{a}$ (a scalar).

    i. Setting $\boldsymbol{a} = (1, \ldots, 1)$ gives the *variance of sum* formula:

$$\mathrm{Var}\left(\sum_{i=1}^{d} X_i\right) = \sum_{i=1}^{d}\sum_{j=1}^{d} \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{d} \mathrm{Var}(X_i) + 2\sum_{1 \le i < j \le d} \mathrm{Cov}(X_i, X_j).$$

If we have further that X_i and X_j are uncorrelated for all $i \neq j$, then it reduces to $\mathrm{Var}\left(\sum_{i=1}^{d} X_i\right) = \sum_{i=1}^{d} \mathrm{Var}(X_i)$.

ii. Setting $\boldsymbol{a} = (1/d, \ldots, 1/d)$ and assuming that X_i and X_j are uncorrelated for all $i \neq j$, we have $\mathrm{Var}\left((\sum_{i=1}^{d} X_i)/d\right) = (\sum_{i=1}^{d} \mathrm{Var}(X_i))/d^2$.

If we have further that $X_1, \ldots, X_d$ are identically distributed, then it reduces to $\mathrm{Var}\left((\sum_{i=1}^{d} X_i)/d\right) = (\mathrm{Var}(X_1))/d$.

5.4.5 Characterization of covariance matrix. To close Section 5.4, we will introduce a characterization of covariance matrix. To prove such characterization, the following result from linear algebra is helpful.

Lemma 5.4.a (Cholesky decomposition). Let $\Sigma \in \mathbb{R}^{d \times d}$ be a symmetric and positive semi-definite matrix. Then, we can write $\Sigma = AA^T$ for some unique $A \in \mathbb{R}^{d \times d}$, which is a lower triangular matrix with nonnegative diagonal elements (which are further positive if Σ is *positive definite*); such A is known as the **Cholesky factor**.

[Note: A symmetric matrix $\Sigma \in \mathbb{R}^{d \times d}$ is **positive semi-definite** if $\boldsymbol{x}^T \Sigma \boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$, and is **positive definite** if $\boldsymbol{x}^T \Sigma \boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$. From these definitions, it should be clear that a positive definite matrix is always positive semi-definite also, but not vice versa.]

Proposition 5.4.b (Characterization of covariance matrix). A real symmetric matrix $\Sigma \in \mathbb{R}^{d \times d}$ is a covariance matrix iff it is positive semi-definite.

Proof. "$\Rightarrow$": Assume Σ is the covariance matrix of a random vector $\boldsymbol{X}$. Then for all $\boldsymbol{a} \in \mathbb{R}^d$ we have $\boldsymbol{a}^T \Sigma \boldsymbol{a} = \mathrm{Var}(\boldsymbol{a}^T \boldsymbol{X}) \geq 0$, so Σ is positive definite.

"$\Leftarrow$": Assume Σ is positive semi-definite with the Cholesky factor A. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ with $X_1, \ldots, X_d \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$.[9] Then, its covariance matrix is $\mathrm{Cov}(\boldsymbol{X}) = I_d$, and thus $\mathrm{Cov}(A\boldsymbol{X}) = A\,\mathrm{Cov}(\boldsymbol{X})A^T = AA^T = \Sigma$. This means that Σ is the covariance matrix of $A\boldsymbol{X}$. $\qquad \square$

Remarks:

- *(Characterization of correlation matrix)* This result also implies a characterization of *correlation matrix*, namely that a real symmetric matrix P is a correlation matrix iff it is positive semi-definite with every diagonal entry being 1. This is because the correlation matrix of $(X_1, \ldots, X_d)$ is the same as the covariance matrix of $(X_1/\mathrm{SD}(X_1), \ldots, X_d/\mathrm{SD}(X_d))$, whose diagonal entries are always 1.

- *(Invertibility of covariance matrix)* From linear algebra, $\Sigma \in \mathbb{R}^{d \times d}$ is invertible iff it is positive *definite* (not positive *semi-definite*). Hence, some covariance matrices can be non-invertible and we need positive definiteness to ensure the invertibility.

5.5 The Lebesgue-Radon-Nikodym Theorem

5.5.1 In Section 5.5, we will explore results about *changes of measures*, which are applied in *importance sampling* in statistics (see [5.5.5]) and *risk-neutral pricing* in financial economics. To start with, we shall study the three types of relationships between different measures, namely *dominating*, *equivalent*, and *singular*.

5.5.2 **Dominating, equivalent, and singular.** Let μ and ν be measures on a measurable space $(\Omega, \mathcal{F})$. Then:

- μ **dominates** ν (or ν is **absolutely continuous** with respect to μ), denoted by $\nu \ll \mu$, if $\mu(A) = 0 \implies \nu(A) = 0$ for all $A \in \mathcal{F}$.

- μ and ν are **equivalent** if $\nu \ll \mu$ and $\mu \ll \nu$, i.e., $\mu(A) = 0 \iff \nu(A) = 0$ for all $A \in \mathcal{F}$ *(sharing the same null sets)*

---

[9]Indeed, any other random variables that are pairwise uncorrelated and have unit variance each would also work.

- μ and ν are **singular**, denoted by $\mu \perp \nu$, if there exists $A \in \mathcal{F}$ such that $\mu(A) = 0$ *(μ lives on A^c)* and $\nu(A^c) = 0$ *(ν lives on A)*.

  [Note: The singularity $\mu \perp \nu$ implies that for every $B \in \mathcal{F}$ with $\mu(B) > 0$ (from here we deduce that $B \subseteq A^c$), we have $\nu(B) = 0$ (since $0 \leq \nu(B) \leq \nu(A^c) = 0$), and vice versa. Symbolically, we can write $\mu(B) > 0 \implies \nu(B) = 0$ and $\nu(B) > 0 \implies \mu(B) = 0$ for all $B \in \mathcal{F}$.]

5.5.3 Some lemmas. To prove the *Lebesgue-Radon-Nikodym theorem*, the following lemmas are needed.

Lemma 5.5.a (Preservation of domination/singularity upon summation)**.** Let ν_n be a measure on a measurable space $(\Omega, \mathcal{F})$ for each $n \in \mathbb{N}$. For the measure $\nu := \sum_{n=1}^{\infty} \nu_n$ on $(\Omega, \mathcal{F})$ (check that it is indeed a measure!), we have:

(a) *(Preserving domination)* If $\nu_n \ll \mu$ for every $n \in \mathbb{N}$, then $\nu \ll \mu$.

(b) *(Preserving singularity)* If $\nu_n \perp \mu$ for every $n \in \mathbb{N}$, then $\nu \perp \mu$.

Proof.

(a) For every $A \in \mathcal{F}$ with $\mu(A) = 0$, we have $\nu(A) = \sum_{n=1}^{\infty} \mu_n(A) = \sum_{n=1}^{\infty} 0 = 0$ since $\nu_n \ll \mu$ for every $n \in \mathbb{N}$. Thus, $\nu \ll \mu$.

(b) For each $n \in \mathbb{N}$, since $\nu_n \perp \mu$, there exists $A_n \in \mathcal{F}$ such that $\nu_n(A_n) = 0$ and $\mu(A_n^c) = 0$. Now let $A := \bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$. Then, we have

$$0 \leq \nu(A) = \sum_{n=1}^{\infty} \nu_n(A) \overset{\text{(monotonicity)}}{\leq} \sum_{n=1}^{\infty} \nu_n(A_n) = 0$$

and

$$0 \leq \mu(A^c) \underset{(A^c = \bigcup_{n=1}^{\infty} A_n^c)}{\overset{\text{(σ-subadditivity)}}{\leq}} \sum_{n=1}^{\infty} \mu(A_n^c) = 0.$$

Hence, $\nu \perp \mu$.

$\square$

Lemma 5.5.b (Relationship between finite measures)**.** Let μ and ν be finite measures on $(\Omega, \mathcal{F})$. Then we have either $\mu \perp \nu$ or $\nu|_A \geq \varepsilon \mu|_A$ for some $\varepsilon > 0$ and $A \in \mathcal{F}$ where $\mu(A) > 0$.

Proof. It follows from applying the *Hahn's decomposition theorem* (Folland, 1999, Theorem 3.3) on the *signed measure* $\nu - (1/n)\mu$. For more details, see Folland (1999, Lemma 3.7). $\square$

5.5.4 The Lebesgue-Radon-Nikodym theorem.

Theorem 5.5.c (Lebesgue-Radon-Nikodym theorem)**.** Let μ and ν be σ-finite measures on a measurable space $(\Omega, \mathcal{F})$. Then:

(a) *(Lebesgue decomposition)* There exist unique σ-finite measures ν_a and ν_s on $(\Omega, \mathcal{F})$ such that $\nu = \nu_a + \nu_s$ where $\nu_a \ll \mu$ *(absolutely continuous with respect to μ)* and $\nu_s \perp \mu$ *(singular with μ)*.

(b) *(Radon-Nikodym theorem)* There exists a μ-a.e. unique and measurable $f : \Omega \to [0, \infty)$ such that $\nu_a(A) = \int_A f \, \mathrm{d}\mu$ for all $A \in \mathcal{F}$.

  Remarks:

  - *(Meaning of a.e. unique)* Here "μ-a.e. unique" suggests that if there is another function g satisfying such conditions, then we have $f = g$ μ-a.e.
  - *(Integrability of f)* In general, f may not be (μ-)integrable. However, if ν_a is finite, i.e., $\nu_a(\Omega) = \int_{\Omega} f \, \mathrm{d}\mu < \infty$, then the nonnegative function f is integrable by definition.
  - *(Terminologies)* In such case, we say that ν_a has **density** f with respect to μ, denoted by $\mathrm{d}\nu_a = f \, \mathrm{d}\mu$ or $f = \frac{\mathrm{d}\nu_a}{\mathrm{d}\mu}$ (the notations are inspired by the formula $\nu_a(A) = \int_A \mathrm{d}\nu_a = \int_A f \, \mathrm{d}\mu$). Such density f is also known as the **Radon-Nikodym derivative** of ν_a with respect to μ (unique a.e.).

Proof. We will prove both results together.

Step 1

We start by proving the special case where μ and ν are *finite* (so $\mu(\Omega) < \infty$ and $\nu(\Omega) < \infty$). The first step is to consider the collection $\mathcal{A} := \{g : \Omega \to [0, \infty] : g \text{ is integrable wrt } \mu, \int_A g \, d\mu \leq \nu(A) \text{ for all } A \in \mathcal{F}\}$.

Showing that $\mathcal{A}$ is nonempty and closed under maxima (of finitely many elements).

- *Nonempty:* Since $\mathcal{A}$ contains the zero function, it is nonempty.
- *Closed under maxima:* Fix any $g_1, g_2 \in \mathcal{A}$. We want to show that $\max\{g_1, g_2\} \in \mathcal{A}$ (which would then imply by induction that $g_1, \ldots, g_n \in \mathcal{A} \implies \max\{g_1, \ldots, g_n\} \in \mathcal{A}$).
  Let $B := \{\omega \in \Omega : g_1(\omega) < g_2(\omega)\} \in \mathcal{F}$. Since

  $$\int_A \max\{g_1, g_2\} \, d\mu = \int_A \max\{g_1, g_2\} \mathbf{1}_B \, d\mu + \int_A \max\{g_1, g_2\} \mathbf{1}_{B^c} \, d\mu$$
  $$= \int_{A \cap B} g_2 \, d\mu + \int_{A \cap B^c} g_1 \, d\mu$$
  $$\overset{(g_1, g_2 \in \mathcal{A})}{\leq} \nu(A \cap B) + \nu(A \cap B^c) = \nu(A),$$

  we have $\max\{g_1, g_2\} \in \mathcal{A}$.

Constructing the density f. Let $s := \sup_{g \in \mathcal{A}} \int_\Omega g \, d\mu$. For every $g \in \mathcal{A}$, we have $\int_\Omega g \, d\mu \leq \nu(\Omega)$, so $s \leq \nu(\Omega) < \infty$. By definition of supremum, there exists a sequence $\{g_n\} \subseteq \mathcal{A}$ such that $\int_\Omega g_n \, d\mu \to s$ as $n \to \infty$.

Now, for each $n \in \mathbb{N}$, let $f_n := \max\{g_1, \ldots, g_n\} \overset{\text{(closed under maxima)}}{\in} \in \mathcal{A}$, and let $f := \sup_{n \in \mathbb{N}} g_n$. Then by construction we have $f_n \nearrow f$ pointwisely. Since each g_n is nonnegative, we have $f_n \in L_+$ for every $n \in \mathbb{N}$. Thus, by MCT, $\int_\Omega f \, d\mu = \lim_{n \to \infty} \int_\Omega f_n \, d\mu$ and f is integrable since $s < \infty$.

Moreover, we can establish the following properties of f:

- $\underline{f \in \mathcal{A}}$: Fix any $A \in \mathcal{F}$. Since $f_n \in \mathcal{A} \ \forall n \in \mathbb{N}$, we have $\int_A f_n \, d\mu \leq \nu(A) \ \forall n \in \mathbb{N}$. Thus,

  $$\int_A f \, d\mu = \int_\Omega f \mathbf{1}_A \, d\mu \overset{\text{(MCT)}}{=} \lim_{n \to \infty} \int_\Omega f_n \mathbf{1}_A \, d\mu = \lim_{n \to \infty} \int_A f_n \, d\mu \leq \nu(A),$$

  which means that $f \in \mathcal{A}$.
- $\underline{\int_\Omega f \, d\mu = s}$: Note that

  $$\int_\Omega f \, d\mu = \lim_{n \to \infty} \int_\Omega f_n \, d\mu \overset{(f_n \geq g_n)}{\geq} \lim_{n \to \infty} \int_\Omega g_n \, d\mu = s$$

  and

  $$\int_\Omega f \, d\mu = \lim_{n \to \infty} \int_\Omega f_n \, d\mu \overset{\underset{(\int_\Omega f_n \, d\mu \leq s \ \forall n \in \mathbb{N})}{\leq}}{\overset{(f_n \in \mathcal{A} \ \forall n \in \mathbb{N})}{}} s.$$

  Thus, $\int_\Omega f \, d\mu = s$.
- $\underline{f : \Omega \to [0, \infty)}$: Since $f_n \nearrow f$ pointwisely, we already know that f is a function from Ω to $[0, \infty]$. Also, f is integrable, so we have $f < \infty$ a.e. by [5.1.11]b. Hence by redefining f to take nonnegative real values (say 0) on a null set (where f takes the value ∞ originally) if needed, we may assume that f is a function from Ω to $[0, \infty)$. Due to this redefinition, such density f can at most be μ-a.e. unique (we will show that f is indeed a.e. unique below).

Showing the existence of such σ-finite ν_a and ν_s. We define $\nu_a(A) := \int_A f \, d\mu \ \forall A \in \mathcal{F}$ and $\nu_s(A) := \nu(A) - \nu_a(A) \overset{(f \in \mathcal{A})}{\geq} 0$. By [5.1.8], ν_a is a measure on $(\Omega, \mathcal{F})$. It is also straightforward to show that ν_s is a measure on $(\Omega, \mathcal{F})$. Regarding their σ-finiteness, consider:

- ν_a: Since f is integrable, we have $\nu_a(\Omega) = \int_\Omega f \, d\mu < \infty$, so ν_a is finite, hence σ-finite.

- ν_s: Since $\nu(\Omega) < \infty$ and $\nu_a(\Omega) < \infty$, we have $\mu_s(\Omega) < \infty$, and so ν_s is again finite, hence σ-finite.

It then remains to show that $\nu_a \ll \mu$ and $\nu_s \perp \mu$:

- $\underline{\nu_a \ll \mu}$: For every $A \in \mathcal{F}$ with $\mu(A) = 0$, by [5.1.10] we know $\nu_a(A) = 0$. Thus, $\nu_a \ll \mu$.
- $\underline{\nu_s \perp \mu}$: Assume to the contrary that $\nu_s \not\perp \mu$. Then by Lemma 5.5.b, there must exist $\varepsilon > 0$ and $A \in \mathcal{F}$ with $\mu(A) > 0$ such that $\nu_s|_A \geq \varepsilon\mu|_A$. This implies that for all $B \in \mathcal{F}$ we have

$$\int_B \varepsilon \mathbf{1}_A \, d\mu = \varepsilon\mu(\underbrace{A \cap B}_{\subseteq A}) = \varepsilon\mu|_A(A \cap B) \leq \nu_s(A \cap B) \overset{\text{(monotonicity)}}{\leq} \nu_s(B) = \nu(B) - \int_B f \, d\mu.$$

  Rearranging this inequality then gives $\int_B (f + \varepsilon\mathbf{1}_A) \, d\mu \leq \mu(B)$, which implies that $f + \varepsilon\mathbf{1}_A \in \mathcal{A}$, and therefore $\int_\Omega f + \varepsilon\mathbf{1}_A \, d\mu \leq s$.

  But on the other hand, we have $\int_\Omega f \, d\mu = s$ as shown previously, so $\int_\Omega f + \varepsilon\mathbf{1}_A \, d\mu = s + \varepsilon\mu(A) > s$, contradiction.

Showing the uniqueness of ν_a and ν_s, and the a.e. uniqueness of f. Suppose that we have $\nu = \nu_a' + \nu_s'$ with $\nu_a'(A) = \int_A f' \, d\mu \; \forall A \in \mathcal{F}$, where ν_a', ν_s', f' satisfy the conditions mentioned in the result.

Then, rearranging $\nu_a + \nu_s = \nu = \nu_a' + \nu_s'$ gives $\nu_s - \nu_s' = \nu_a' - \nu_a$, so we can write $\nu_s(B) = \nu_s'(B) = \int_B f' - f \, d\mu = \nu_a'(B) - \nu_a(B)$ for all $B \in \mathcal{F}$. Next, consider two cases:

- *Case 1:* $\mu(B) > 0$. Since $\nu_s \perp \mu$ and $\nu_s' \perp \mu$, we have $\nu_s(B) - \nu_s'(B) = 0 - 0 = 0$, so $\nu_a'(B) - \nu_a(B) = 0$ also.
- *Case 2:* $\mu(B) = 0$. By [5.1.10], we know $\int_B f' - f \, d\mu = 0$, so $\nu_s(B) - \nu_s'(B) = \nu_a'(B) - \nu_a(B) = 0$.

Therefore, we have $\nu_s(B) = \nu_s'(B)$ and $\nu_a'(B) = \nu_a(B)$ for all $B \in \mathcal{F}$, establishing the uniqueness of ν_a and ν_s. Furthermore, since $\int_B f' - f \, d\mu = 0$ for all $B \in \mathcal{F}$, by [5.1.11]c we know $f' = f$ μ-a.e., establishing the a.e. uniqueness of f.

Step 2

We have completed the proof for the case where μ and ν are finite. Next, we are going to extend it to the case where they are σ-finite.

Constructing pairwise disjoint $A_1, A_2, \ldots \in \mathcal{F}$ that "simultaneously satisfy" the σ-additivity of μ and ν. Our goal here is to construct pairwise disjoint $A_1, A_2, \ldots \in \mathcal{F}$ such that $\biguplus_{n=1}^{\infty} A_n = \Omega$ with $\mu(A_n) < \infty$ and $\nu(A_n) < \infty$ for each $n \in \mathbb{N}$. This can be done by intersecting the sets arising respectively from the σ-additivity of μ and ν.

Due to the σ-additivity of μ and ν, there exist $A_{\mu,1}, A_{\mu,2}, \ldots \in \mathcal{F}$ and $A_{\nu,1}, A_{\nu,2}, \ldots \in \mathcal{F}$ such that $\bigcup_{m=1}^{\infty} A_{\mu,m} = \Omega$ and $\bigcup_{\ell=1}^{\infty} A_{\nu,\ell} = \Omega$ with $\mu(A_{\mu,m}) < \infty$ and $\nu(A_{\nu,\ell}) < \infty$ for all $m, n \in \mathbb{N}$. WLOG (by performing *disjointification* on them if needed), we may assume that $A_{\mu,1}, A_{\mu,2}, \ldots \in \mathcal{F}$ and $A_{\nu,1}, A_{\nu,2}, \ldots \in \mathcal{F}$ are respectively pairwise disjoint.

Now, consider $\Omega = \Omega \cap \Omega = (\biguplus_{m=1}^{\infty} A_{\mu,m}) \cap (\biguplus_{\ell=1}^{\infty} A_{\nu,\ell}) = \biguplus_{m=1}^{\infty} \biguplus_{\ell=1}^{\infty} (A_{\mu,m} \cap A_{\nu,\ell})$. Note that there are countably many $(A_{\mu,m} \cap A_{\nu,\ell})$'s, so after relabelling them to $A_1, A_2, \ldots \in \mathcal{F}$, we have constructed the desired sets.

Constructing finite measures and applying the result from step 1 to them. Fix any $n \in \mathbb{N}$. Define $\mu_n(B) := \mu(B \cap A_n)$ and $\nu_n(B) := \nu(B \cap A_n)$ for all $B \in \mathcal{F}$. By construction of A_n's, μ_n and ν_n are both finite. Hence, applying the result from step 1 to them gives

$$\nu_n(B) = \nu_{n,a}(B) + \nu_{n,s}(B) = \int_B f_n \, d\mu_n + \nu_{n,s}(B) \quad \forall B \in \mathcal{F},$$

where $\nu_{n,a}, \nu_{n,s}, f_n$ satisfy the conditions mentioned in the result, and particularly we know from step 1 that $\nu_{n,a}$ and $\nu_{n,s}$ are indeed *finite*.

Showing the existence of such ν_a, ν_s, and f. By construction of μ_n and ν_n, we have $\mu_n(A_n^c) = \nu_n(A_n^c) = 0$. So, we have $\nu_{n,a}(A_n^c) = \int_{A_n^c} f_n \, \mathrm{d}\mu_n = 0$ by [5.1.10], and hence $\nu_{n,s}(A_n^c) = \nu_n(A_n^c) - \int_{A_n^c} f_n \, \mathrm{d}\mu_n = 0 - 0 = 0$. Furthermore, we may assume that $f_n|_{A_n^c} = 0$ (after redefinition if needed) without affecting the μ_n-a.e. uniqueness of f_n, since $\mu_n(A_n^c) = 0$.

Let $f := \sum_{n=1}^{\infty} f_n$, and we will establish the measurability, nonnegativity, and finiteness of f in the following:

- *f is measurable:* As a limit of measurable functions (finite sums of f_n's are measurable), by [3.1.5]c we know that f is measurable also.

- *$f : \Omega \to [0, \infty)$:* Since $\biguplus_{n=1}^{\infty} A_n = \Omega$ and $f_n|_{A_n^c} = 0$ for each $n \in \mathbb{N}$, for all $\omega \in \Omega$ we have $f(\omega) = f_n(\omega) \in [0, \infty)$, for the (unique) $n \in \mathbb{N}$ where A_n contains ω. Hence, f is a function from Ω to $[0, \infty)$.

Now, let $\nu_a(B) := \int_B f \, \mathrm{d}\mu \ \forall B \in \mathcal{F}$, and let $\nu_s := \sum_{n=1}^{\infty} \nu_{n,s}$. We then establish the rest of target properties as follows:

- *$\nu_a \ll \mu$ and $\nu_s \perp \mu$:* By Lemma 5.5.a, since $\nu_{n,s} \perp \mu \ \forall n \in \mathbb{N}$, we have $\nu_s \perp \mu$. Also, we know by [5.1.10] that $\nu_a \ll \mu$.

- *ν_a and ν_s are σ-finite:* Since $\nu_{n,a}(A_n^c) = \nu_{n,s}(A_n^c) = 0$ and $\nu_{n,a}, \nu_{n,s}$ are finite for each $n \in \mathbb{N}$, we know $\nu_{n,a}(A_n), \nu_{n,s}(A_n)$ are finite for each $n \in \mathbb{N}$. Now, as A_n's are pairwise disjoint, we know

$$\nu_a(A_n) = \int_{A_n} f \, \mathrm{d}\mu \overset{[5.1.8]}{=} \sum_{k=1}^{\infty} \int_{A_n} f_k \, \mathrm{d}\mu \overset{(\mu_k|_{A_k} = \mu|_{A_k})}{\underset{(f_k|_{A_k^c} = 0)}{=}} \sum_{k=1}^{\infty} \int_{A_n} f_k \, \mathrm{d}\mu_k = \sum_{k=1}^{\infty} \nu_{k,a}(A_n) \overset{(\text{disjoint})}{\underset{(\nu_{k,a}(A_k^c) = 0)}{=}} \nu_{n,a}(A_n)$$

and $\nu_s(A_n) = \sum_{k=1}^{\infty} \nu_{k,s}(A_n) = \nu_{n,s}(A_n)$ are finite for each $n \in \mathbb{N}$. The σ-finiteness of ν_a, ν_s then follows from noting that $\biguplus_{n=1}^{\infty} A_n = \Omega$.

- *$\nu(B) = \nu_a(B) + \nu_s(B) \ \forall B \in \mathcal{F}$:* Consider

$$\nu(B) = \nu\left(B \cup \biguplus_{n=1}^{\infty} A_n\right) = \nu\left(\biguplus_{n=1}^{\infty} B \cap A_n\right) = \sum_{n=1}^{\infty} \nu(B \cap A_n) = \sum_{n=1}^{\infty} \nu_n(B)$$

$$= \sum_{n=1}^{\infty} \left(\int_B f_n \, \mathrm{d}\mu_n + \nu_{n,s}(B)\right) \overset{(\mu_n|_{A_n} = \mu|_{A_n})}{\underset{(f_n|_{A_n^c} = 0)}{=}} \sum_{n=1}^{\infty} \int_B f_n \, \mathrm{d}\mu + \nu_s(B)$$

$$\overset{[5.1.8]}{=} \int_B \sum_{n=1}^{\infty} f_n \, \mathrm{d}\mu + \nu_s(B) = \int_B f \, \mathrm{d}\mu + \nu_s(B) = \nu_a(B) + \nu_s(B).$$

Showing the uniqueness of ν_a and ν_s, and the a.e. uniqueness of f. The argument is exactly the same as the one in step 1. $\qquad \square$

Sometimes the following result, which is a corollary of the Lebesgue-Radon-Nikodym theorem, is referred to as the *Radon-Nikodym theorem*. For studying changes of measures, quite often this special case is utilized instead of the more general Lebesgue-Radon-Nikodym theorem.

Corollary 5.5.d (Radon-Nikodym theorem). If μ and ν are σ-finite measures on a measurable space $(\Omega, \mathcal{F})$ with $\nu \ll \mu$, then there exists a μ-a.e. unique and measurable $f : \Omega \to [0, \infty)$ such that $\nu(A) = \int_A f \, \mathrm{d}\mu$ for all $A \in \mathcal{F}$ (which is the Radon-Nikodym derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$).

Proof. With $\nu \ll \mu$, we know from the Lebesgue-Radon-Nikodym theorem that the Lebesgue decomposition must be given by $\nu = \nu_a$ ($\nu_s = 0$), and so the result follows. $\qquad \square$

5.5.5 **Properties of Radon-Nikodym derivatives.** Here, we are going to show some properties of Radon-Nikodym derivatives, which are quite consistent with the notation $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.

Let μ, ν, λ be σ-finite measures on $(\Omega, \mathcal{F})$ with $\nu \ll \mu$ and $\mu \ll \lambda$. Then:

(a) *("Integration by substitution")* If $g \in L^1(\Omega, \mathcal{F}, \nu)$, then $g\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^1(\Omega, \mathcal{F}, \mu)$ and $\int_\Omega g \, \mathrm{d}\nu = \int_\Omega g\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu$.

(b) *("Chain rule")* $\nu \ll \lambda$ and $\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ λ-a.e.

Proof.

(a) We apply the standard argument on g.

(1) Fix any indicator function $g = \mathbf{1}_A$ where $A \in \mathcal{F}$. We then have

$$\int_\Omega g \, \mathrm{d}\nu \overset{\text{(simple)}}{=} \nu(A) \overset{\text{(Radon-Nikodym)}}{\underset{(\nu \ll \mu)}{=}} \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu = \int_\Omega g\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu$$

By the linearity of integral, the equation also holds for all simple functions.

(2) Fix any $g \in L_+(\Omega, \mathcal{F}, \nu)$. By Proposition 5.1.a, there exists a sequence $\{g_n\}$ of nonnegative simple functions on Ω' such that $g_n \nearrow g$, which implies that $g_n\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \nearrow g\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ with $g_n\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L_+$ for each $n \in \mathbb{N}$. Hence,

$$\int_\Omega g \, \mathrm{d}\nu \overset{\text{(MCT)}}{=} \lim_{n \to \infty} \int_\Omega g_n \, \mathrm{d}\nu \overset{(1)}{=} \lim_{n \to \infty} \int_\Omega g_n\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu \overset{\text{(MCT)}}{=} \int_\Omega g\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu.$$

(3) Fix any $g \in L^1(\Omega, \mathcal{F}, \nu)$. Then we have

$$\int_\Omega g \, \mathrm{d}\nu = \int_\Omega g^+ \, \mathrm{d}\nu - \int_\Omega g^- \, \mathrm{d}\nu \overset{(2)}{=} \int_\Omega g^+\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu - \int_\Omega g^-\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu = \int_\Omega g\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu.$$

This also implies that $g\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \in L^1(\Omega, \mathcal{F}, \mu)$ as $\int_\Omega g^+\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu = \int_\Omega g^+ \, \mathrm{d}\nu < \infty$ and $\int_\Omega g^-\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu = \int_\Omega g^- \, \mathrm{d}\nu < \infty$.

(b) For all $A \in \mathcal{F}$, we have $\lambda(A) = 0 \overset{(\mu \ll \lambda)}{\Longrightarrow} \mu(A) = 0 \overset{(\nu \ll \lambda)}{\Longrightarrow} \nu(A) = 0$, and hence $\nu \ll \lambda$. Next, we have

$$\int_A \frac{\mathrm{d}\nu}{\mathrm{d}\lambda} \, \mathrm{d}\lambda \overset{\text{(Radon-Nikodym)}}{=} \nu(A) \overset{\text{(Radon-Nikodym)}}{=} \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu = \int_\Omega \mathbf{1}_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu \overset{\text{(a)}}{=} \int_\Omega \mathbf{1}_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \, \mathrm{d}\mu = \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \, \mathrm{d}\lambda$$

for all $A \in \mathcal{F}$, so the result follows by [5.1.11]c.

$\square$

Remarks:

- *(Relationship between Radon-Nikodym derivative and density function)* Consider a probability space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P}_{\boldsymbol{X}})$. If the distribution $\mathbb{P}_{\boldsymbol{X}}$ or F is absolutely continuous with respect to the Lebesgue measure λ, then by Radon-Nikodym theorem we have $F(B) = \int_B \mathrm{d}F = \int_B f \, \mathrm{d}\lambda$ for all $B \in \mathcal{B}(\mathbb{R}^d)$, where $f = \frac{\mathrm{d}F}{\mathrm{d}\lambda}$ is the Radon-Nikodym derivative.

  Particularly, setting $B = (-\infty, \boldsymbol{x}]$ gives $F(\boldsymbol{x}) = F((-\infty, \boldsymbol{x}] = \int_{(-\infty,\boldsymbol{x}]} f \, \mathrm{d}\lambda = \int_{(-\infty,\boldsymbol{x}]} f(\widetilde{\boldsymbol{x}}) \, \mathrm{d}\lambda(\widetilde{\boldsymbol{x}}) \overset{\text{(notation)}}{=} \int_{(-\infty,\boldsymbol{x}]} f(\widetilde{\boldsymbol{x}}) \, \mathrm{d}\widetilde{\boldsymbol{x}}$, which suggests that the Radon-Nikodym derivative is indeed the density function of F. In such case, recall that F is said to be *absolutely continuous* — this relates with the terminology here. Indeed, a distribution function F is absolutely continuous iff its corresponding distribution (also denoted by F) is absolutely continuous wrt the Lebesgue measure λ.

  Applying [5.5.5]a, we also obtain $\int_{\mathbb{R}^d} g(\boldsymbol{x}) \, \mathrm{d}F(\boldsymbol{x}) = \int_{\mathbb{R}^d} g(\boldsymbol{x})f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$, if $g \in L^1(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), F)$.

- *(Importance sampling)* [5.5.5]a can be expressed more succinctly as "$\mathbb{E}_\nu[g] = \mathbb{E}_\mu\left[g\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right]$ if $\nu \ll \mu$". This formula is applied in *importance sampling* in statistics. To see this, consider $\boldsymbol{X} \sim F$ and $\widetilde{\boldsymbol{X}} \sim H$, with f and h being the densities of F and H (interpreted as distribution functions/distributions) respectively. Then the formula suggests that, for every density h satisfying that $h(\boldsymbol{x}) = 0 \implies f(\boldsymbol{x}) = 0$,

$$\mathbb{E}_F[g(\boldsymbol{X})] = \int_{\mathbb{R}^d} g(\boldsymbol{x})f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d} \underbrace{\frac{g(\boldsymbol{x})f(\boldsymbol{x})}{h(\boldsymbol{x})}}_{:=0 \text{ if } h(\boldsymbol{x}) = 0} h(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \mathbb{E}_H\left[\frac{g(\widetilde{\boldsymbol{X}})f(\widetilde{\boldsymbol{X}})}{h(\widetilde{\boldsymbol{X}})}\right],$$

which is utilized in importance sampling.

6 Modes of Convergence

6.0.1 In Section 6, we will explore the different notions of *convergence* in probability theory. These concepts are crucial for developments of asymptotic methods, and also for many important results like *laws of large numbers* and *central limit theorem*.

6.1 Almost Sure Convergence and Convergence in Probability

6.1.1 Two modes of convergence that are commonly studied and have wide applicability are *almost sure convergence* and *convergence in probability*. There is also another less common mode of convergence, known as *complete convergence*, which is stronger than both of them and is mainly of theoretical interest. Now, we shall start by defining these three types of convergence.

6.1.2 Definitions. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\boldsymbol{X}$ be a random vector, and $\{\boldsymbol{X}_n\}$ be a sequence of random vectors, where $\boldsymbol{X}$ and all $\boldsymbol{X}_n$'s are defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Then:

- $\{\boldsymbol{X}_n\}$ **converges completely** to $\boldsymbol{X}$, denoted by $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{c.c.}} \boldsymbol{X}$, if for all $\varepsilon > 0$, we have $\sum_{n=1}^{\infty} \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) < \infty$.

- $\{\boldsymbol{X}_n\}$ **converges almost surely** to $\boldsymbol{X}$, denoted by $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}$, if $\mathbb{P}(\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}) = 1$.
  [Note: More explicitly, we can write $\mathbb{P}(\{\omega \in \Omega : \lim_{n\to\infty} \boldsymbol{X}_n(\omega) = \boldsymbol{X}(\omega)\}) = 1$; here for each fixed $\omega \in \Omega$, $\{\boldsymbol{X}_n(\omega)\}_{n\in\mathbb{N}}$ is a real-valued sequence.]

- $\{\boldsymbol{X}_n\}$ **converges in probability** to $\boldsymbol{X}$, denoted by $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{P}} \boldsymbol{X}$, if for all $\varepsilon > 0$, we have $\lim_{n\to\infty} \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) = 0$.
  [Note: More explicitly, we have that for all $\varepsilon > 0$ and $\varepsilon' > 0$, there exists $m \in \mathbb{N}$ such that $\mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) < \varepsilon'$ for all $n \geq m$.]

6.1.3 Characterization of almost sure convergence by convergence in probability. One important result about almost sure convergence and convergence in probability is the characterization of the former in terms of the latter. To show such characterization, we need the following lemma.

Lemma 6.1.a (Interchanges of (limit) supremum with other operations)**.** For all $\varepsilon > 0$, we have:

(a) $\sup_{k \geq n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\} = \{\sup_{k \geq n} \|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}$ and $\limsup_{n\to\infty}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\} = \{\limsup_{n\to\infty} \|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}$.

(b) $\lim_{n\to\infty} \mathbb{P}\big(\sup_{k \geq n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}\big) = \mathbb{P}(\limsup_{n\to\infty}\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon\})$.

Proof.

(a) Note that

$$\omega \in \sup_{k\geq n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\} = \bigcup_{k\geq n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}$$

$$\iff \|\boldsymbol{X}_k(\omega) - \boldsymbol{X}(\omega)\| > \varepsilon \text{ for some } k \geq n$$

$$\iff \sup_{k\geq n} \|\boldsymbol{X}_k(\omega) - \boldsymbol{X}(\omega)\| > \varepsilon$$

and

$$\omega \in \limsup_{n\to\infty}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\} = \bigcap_{n=1}^{\infty} \bigcup_{k\geq n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}$$

$$\iff \|\boldsymbol{X}_k(\omega) - \boldsymbol{X}(\omega)\| > \varepsilon \text{ for some } k \geq n, \text{ for all } n \in \mathbb{N}$$

$$\iff \limsup_{n\to\infty} \|\boldsymbol{X}_k(\omega) - \boldsymbol{X}(\omega)\| > \varepsilon.$$

(b) Let $A_{n,\varepsilon} := \sup_{k \geq n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}$ and $A_{\infty,\varepsilon} := \lim_{n \to \infty} A_{n,\varepsilon} = \limsup_{n \to \infty}\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon\}$. The result then follows from the equality $\lim_{n \to \infty} \mathbb{P}(A_{n,\varepsilon}) = \mathbb{P}(A_{\infty,\varepsilon})$.

$\square$

Theorem 6.1.b (Characterization of almost sure convergence by convergence in probability). We have $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}$ iff $\sup_{k \geq n}\|\boldsymbol{X}_k - \boldsymbol{X}\| \xrightarrow[n\to\infty]{\text{P}} 0$.

Proof. "$\Rightarrow$": Assume $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}$, so we have $\lim_{n\to\infty}\boldsymbol{X}_n(\omega) = \boldsymbol{X}(\omega)$ for all $\omega \in N^c$ where $N \in \mathcal{F}$ with $\mathbb{P}(N) = 0$. Fix any $\varepsilon > 0$. By the definition of limit, for all $\omega \in N^c$, there exists $m \in \mathbb{N}$ such that $\|\boldsymbol{X}_n(\omega) - \boldsymbol{X}(\omega)\| < \varepsilon$ for all $n \geq m$.

Let $A_{n,\varepsilon} := \bigcup_{k \geq n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\} = \sup_{k \geq n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}$ and $A_{\infty,\varepsilon} := \bigcap_{n=1}^{\infty} A_{n,\varepsilon} = \lim_{n\to\infty} A_{n,\varepsilon} = \limsup_{n\to\infty}\{\|\bar{\boldsymbol{X}}_k - \boldsymbol{X}\| > \varepsilon\}$. By construction, we know for all $\omega \in N^c$, there exists $n \in \mathbb{N}$ such that $\omega \notin A_{n,\omega}$, and thus $\omega \notin A_{\infty,\varepsilon}$. Hence, $A_{\infty,\varepsilon} \subseteq N$.

This implies that

$$0 \leq \lim_{n\to\infty} \mathbb{P}\left(\left|\sup_{k\geq n}\|\boldsymbol{X}_k - \boldsymbol{X}\| - 0\right| > \varepsilon\right) = \lim_{n\to\infty}\mathbb{P}\left(\sup_{k\geq n}\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\right)$$

$$\overset{\text{(Lemma 6.1.a)}}{=} \lim_{n\to\infty}\mathbb{P}\left(\sup_{k\geq n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}\right) = \mathbb{P}\left(\limsup_{n\to\infty}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}\right) = \mathbb{P}(A_{\infty,\varepsilon}) \leq \mathbb{P}(N) = 0,$$

and thus $\sup_{k\geq n}\|\boldsymbol{X}_k - \boldsymbol{X}\| \xrightarrow[n\to\infty]{\text{P}} 0$.

"$\Leftarrow$": Assume $\sup_{k\geq n}\|\boldsymbol{X}_k - \boldsymbol{X}\| \xrightarrow[n\to\infty]{\text{P}} 0$. Then, we have

$$\mathbb{P}\left(\lim_{n\to\infty}\boldsymbol{X}_n \neq \boldsymbol{X}\right) \overset{\text{(definition of limit)}}{=} \mathbb{P}\left(\lim_{n\to\infty}\|\boldsymbol{X}_n - \boldsymbol{X}\| > 0\right) = \mathbb{P}\left(\limsup_{n\to\infty}\|\boldsymbol{X}_n - \boldsymbol{X}\| > 0\right)$$

$$= \mathbb{P}\left(\bigcup_{m=1}^{\infty}\left\{\limsup_{n\to\infty}\|\boldsymbol{X}_n - \boldsymbol{X}\| > 1/m\right\}\right)$$

$$\overset{(\sigma\text{-subadditivity})}{\leq} \sum_{m=1}^{\infty}\mathbb{P}\left(\left\{\limsup_{n\to\infty}\|\boldsymbol{X}_n - \boldsymbol{X}\| > 1/m\right\}\right)$$

$$\overset{\text{(Lemma 6.1.a)}}{=} \sum_{m=1}^{\infty}\mathbb{P}\left(\limsup_{n\to\infty}\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > 1/m\}\right)$$

$$\overset{\text{(Lemma 6.1.a)}}{=} \sum_{m=1}^{\infty}\lim_{n\to\infty}\mathbb{P}\left(\sup_{k\geq n}\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > 1/m\}\right)$$

$$\overset{\text{(Lemma 6.1.a)}}{=} \sum_{m=1}^{\infty}\lim_{n\to\infty}\mathbb{P}\left(\sup_{k\geq n}\|\boldsymbol{X}_n - \boldsymbol{X}\| > 1/m\right) \overset{\text{(assumption)}}{=} \sum_{m=1}^{\infty} 0 = 0,$$

which implies that $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}$. $\square$

6.1.4 **Properties of these modes of convergence.** By Theorem 6.1.b, we can derive the following properties about the modes of convergence discussed here fairly easily:

(a) $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{c.c.}} \boldsymbol{X} \implies \boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}$.

(b) $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X} \implies \boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{P}} \boldsymbol{X}$.

(c) $\boldsymbol{X}_n \overset{\text{P}}{\underset{n\to\infty}{\nearrow}} \boldsymbol{X}$ or $\boldsymbol{X}_n \overset{\text{P}}{\underset{n\to\infty}{\searrow}} \boldsymbol{X} \implies \boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}$.

   [Note: Here $\boldsymbol{X}_n \overset{\text{P}}{\underset{n\to\infty}{\nearrow}} \boldsymbol{X}$ means $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{P}} \boldsymbol{X}$ and $\boldsymbol{X}_n \nearrow$ (similar for $\boldsymbol{X}_n \overset{\text{P}}{\underset{n\to\infty}{\searrow}} \boldsymbol{X}$).]

Proof.

(a) Assume $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{c.c.}} \boldsymbol{X}$. Fix any $\varepsilon > 0$. Since $\sum_{n=1}^{\infty} \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) < \infty$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{k\geq n} \|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\right) \overset{\text{(Lemma 6.1.a)}}{=} \lim_{n\to\infty} \mathbb{P}\left(\sup_{k\geq n}\{\|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon\}\right)$$

$$\overset{\text{(Lemma 6.1.a)}}{=} \mathbb{P}\left(\limsup_{n\to\infty}\{\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon\}\right) \overset{\text{(first Borel-Cantelli lemma)}}{=} 0.$$

Thus by Theorem 6.1.b we have $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}$.

(b) Assume $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}$. By Theorem 6.1.b, we have $\sup_{k\geq n} \|\boldsymbol{X}_k - \boldsymbol{X}\| \xrightarrow[n\to\infty]{\text{P}} 0$, which means that $\lim_{n\to\infty} \mathbb{P}(\sup_{k\geq n} \|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon) = 0$ for all $\varepsilon > 0$. For every $n \in \mathbb{N}$, since $\|\boldsymbol{X}_n - \boldsymbol{X}\| \leq \sup_{k\geq n} \|\boldsymbol{X}_k - \boldsymbol{X}\|$, we have $\mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) \leq \mathbb{P}(\sup_{k\geq n} \|\boldsymbol{X}_k - \boldsymbol{X}\| > \varepsilon) \; \forall \varepsilon > 0$, forcing that $\lim_{n\to\infty} \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) = 0 \; \forall \varepsilon > 0$. Therefore, $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{P}} \boldsymbol{X}$.

(c) Note that $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{P}} \boldsymbol{X} \implies \|\boldsymbol{X}_n - \boldsymbol{X}\| \xrightarrow[n\to\infty]{\text{P}} 0$. Also, since $\boldsymbol{X}_n \overset{\text{P}}{\underset{n\to\infty}{\nearrow}} \boldsymbol{X}$, we have $\sup_{k\geq n} \|\boldsymbol{X}_k - \boldsymbol{X}\| \overset{\text{a.s.}}{=} \|\boldsymbol{X}_n - \boldsymbol{X}\| \xrightarrow[n\to\infty]{\text{P}} 0$. This implies that $\sup_{k\geq n} \|\boldsymbol{X}_k - \boldsymbol{X}\| \xrightarrow[n\to\infty]{\text{P}} 0$, and hence by Theorem 6.1.b we have $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}$.

$\square$

6.1.5 **Subsequence principle.** A remarkable result that relates the three modes of convergence is the *subsequence principle*, which involves "subsequence of subsequence".

Theorem 6.1.c (Subsequence principle)**.** The following are equivalent.

(a) $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{P}} \boldsymbol{X}$.

(b) For every subsequence $\{\boldsymbol{X}_{n_k}\}_{k\in\mathbb{N}} \subseteq \{\boldsymbol{X}_n\}_{n\in\mathbb{N}}$, there exists a *further* subsequence $\{\boldsymbol{X}_{n_{k_\ell}}\}_{\ell\in\mathbb{N}} \subseteq \{\boldsymbol{X}_{n_k}\}_{k\in\mathbb{N}}$ such that $\boldsymbol{X}_{n_{k_\ell}} \xrightarrow[\ell\to\infty]{\text{c.c.}} \boldsymbol{X}$.

(c) For every subsequence $\{\boldsymbol{X}_{n_k}\}_{k\in\mathbb{N}} \subseteq \{\boldsymbol{X}_n\}_{n\in\mathbb{N}}$, there exists a *further* subsequence $\{\boldsymbol{X}_{n_{k_\ell}}\}_{\ell\in\mathbb{N}} \subseteq \{\boldsymbol{X}_{n_k}\}_{k\in\mathbb{N}}$ such that $\boldsymbol{X}_{n_{k_\ell}} \xrightarrow[\ell\to\infty]{\text{a.s.}} \boldsymbol{X}$.

Proof.

- (a) $\implies$ (b): By (a), for every subsequence $\{\boldsymbol{X}_{n_k}\}_{k\in\mathbb{N}} \subseteq \{\boldsymbol{X}_n\}_{n\in\mathbb{N}}$, we have $\boldsymbol{X}_{n_k} \xrightarrow[k\to\infty]{\text{P}} \boldsymbol{X}$. By definition of convergence in probability, for all $\varepsilon > 0$ and $\varepsilon' > 0$, there exists $K \in \mathbb{N}$ such that $\mathbb{P}(\|\boldsymbol{X}_{n_k} - \boldsymbol{X}\| > \varepsilon) < \varepsilon'$ for all $k \geq K$. Thus, for every $\ell \in \mathbb{N}$, by setting $\varepsilon = \varepsilon' = 2^{-\ell}$, we can choose $k_\ell \in \mathbb{N}$ such that $\mathbb{P}\left(\|\boldsymbol{X}_{n_{k_\ell}} - \boldsymbol{X}\| > 2^{-\ell}\right) < 2^{-\ell}$.

  Hence, for all $\varepsilon > 0$ we have

$$\sum_{\ell=1}^{\infty} \mathbb{P}\left(\|\boldsymbol{X}_{n_{k_\ell}} - \boldsymbol{X}\| > \varepsilon\right) = \underbrace{\sum_{\ell\in\mathbb{N}: 2^{-\ell} > \varepsilon} \mathbb{P}\left(\|\boldsymbol{X}_{n_{k_\ell}} - \boldsymbol{X}\| > \varepsilon\right)}_{\text{finitely many summands}} + \underbrace{\sum_{\ell\in\mathbb{N}: 2^{-\ell} \leq \varepsilon} \underbrace{\mathbb{P}\left(\|\boldsymbol{X}_{n_{k_\ell}} - \boldsymbol{X}\| > \varepsilon\right)}_{\leq \mathbb{P}\left(\|\boldsymbol{X}_{n_{k_\ell}} - \boldsymbol{X}\| > 2^{-\ell}\right) < 2^{-\ell}}}_{< \sum_{\ell\in\mathbb{N}} 2^{-\ell} = 1} < \infty,$$

  which means by definition that $\boldsymbol{X}_{n_{k_\ell}} \xrightarrow[\ell\to\infty]{\text{c.c.}} \boldsymbol{X}$.

- (b) $\implies$ (c): It follows from [6.1.4]a.

104

- (c) $\implies$ (a): We prove by contrapositive. Assume that $\boldsymbol{X}_n \overset{\text{p}}{\underset{n\to\infty}{\not\to}} \boldsymbol{X}$. Then by negating the definition, we know that there exist $\varepsilon, \varepsilon' > 0$ such that $\mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \varepsilon) \geq \varepsilon'$ for infinitely many n's. So we can form a subsequence $\{\boldsymbol{X}_{n_k}\}_{k\in\mathbb{N}} \subseteq \{\boldsymbol{X}_n\}_{n\in\mathbb{N}}$ such that $\mathbb{P}(\|\boldsymbol{X}_{n_k} - \boldsymbol{X}\| > \varepsilon) \geq \varepsilon'$ for all $k \in \mathbb{N}$. Therefore, for every further subsequence $\{\boldsymbol{X}_{n_{k_\ell}}\}_{\ell\in\mathbb{N}} \subseteq \{\boldsymbol{X}_{n_k}\}_{k\in\mathbb{N}}$, we have $\boldsymbol{X}_{n_{k_\ell}} \overset{\text{p}}{\underset{\ell\to\infty}{\not\to}} \boldsymbol{X}$, and so $\boldsymbol{X}_{n_{k_\ell}} \overset{\text{a.s}}{\underset{\ell\to\infty}{\not\to}} \boldsymbol{X}$ by [6.1.4]b, which means (c) does not hold.

$\square$

6.1.6 **Dominated convergence theorem for convergence in probability.** Under a probability space, we can apply Theorem 6.1.c to extend the dominated convergence theorem by replacing the almost sure convergence by the convergence in probability in its conditions, thereby widening its applicability.

Corollary 6.1.d (Dominated convergence theorem (convergence in probability))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $X_n \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ for every $n \in \mathbb{N}$, and $X : \Omega \to \mathbb{R}$ be measurable. If $X_n \overset{\text{p}}{\underset{n\to\infty}{\to}} X$ and $|X_n| \leq Y$ a.s. for all $n \in \mathbb{N}$, for some $Y \in L^1$ *(domination)*, then $X \in L^1$ and $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$.

Proof. Since $X_n \overset{\text{p}}{\underset{n\to\infty}{\to}} X$, by subsequence principle we know for every subsequence $\{\boldsymbol{X}_{n_k}\}_{k\in\mathbb{N}} \subseteq \{\boldsymbol{X}_n\}_{n\in\mathbb{N}}$, there exists a further subsequence $\{\boldsymbol{X}_{n_{k_\ell}}\}_{\ell\in\mathbb{N}} \subseteq \{\boldsymbol{X}_{n_k}\}_{k\in\mathbb{N}}$ such that $\boldsymbol{X}_{n_{k_\ell}} \overset{\text{a.s.}}{\underset{\ell\to\infty}{\to}} \boldsymbol{X}$. Applying the original dominated convergence theorem (Theorem 5.1.e) on $\{\boldsymbol{X}_{n_{k_\ell}}\}_{\ell\in\mathbb{N}}$, we get $X \in L^1$ and $\lim_{\ell\to\infty} \mathbb{E}\left[X_{n_{k_\ell}}\right] = \mathbb{E}[X]$.

Claim: We also have the subsequence principle for real-valued sequence, i.e., $a_n \to a$ iff for every subsequence $\{a_{n_k}\}_{k\in\mathbb{N}} \subseteq \{a_n\}_{n\in\mathbb{N}}$, there exists a *further* subsequence $\{a_{n_{k_\ell}}\}_{\ell\in\mathbb{N}} \subseteq \{a_{n_k}\}_{k\in\mathbb{N}}$ such that $a_{n_{k_\ell}} \to a$.

Proof.

- "$\Rightarrow$": It follows by considering the definition of sequence.
- "$\Leftarrow$": We prove by contrapositive. Assume that $a_n \not\to a$. Then there exists $\varepsilon > 0$ such that for all $k \in \mathbb{N}$, $|a_{n_k} - a| > \varepsilon$ for some $n_k \geq k$. From this we can construct a subsequence $\{a_{n_k}\} \subseteq \{a_n\}$ such that $|a_{n_k} - a| > \varepsilon$ for all $k \in \mathbb{N}$. The result then follows by noting that it does not contain any further subsequence that converges to a.

$\square$

By the subsequence principle for real-valued sequence, we then have $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$. $\square$

6.1.7 **Continuous mapping theorem for almost sure convergence and convergence in probability.** When working with modes of convergence, a nice and useful result is the *continuous mapping theorem* (CMT), which suggests that some modes of convergence are preserved after applying continuous function. Here, we will state and prove the CMT for almost sure convergence and convergence in probability, but it actually also holds for *convergence in distribution*; see Theorem 6.3.b.

Theorem 6.1.e (Continuous mapping theorem for almost sure convergence and convergence in probability)**.** Let $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^k$ be continuous. Then:

(a) $\boldsymbol{X}_n \overset{\text{a.s.}}{\underset{n\to\infty}{\to}} \boldsymbol{X} \implies \boldsymbol{h}(\boldsymbol{X}_n) \overset{\text{a.s.}}{\underset{n\to\infty}{\to}} \boldsymbol{h}(\boldsymbol{X})$.

(b) $\boldsymbol{X}_n \overset{\text{P}}{\underset{n\to\infty}{\to}} \boldsymbol{X} \implies \boldsymbol{h}(\boldsymbol{X}_n) \overset{\text{P}}{\underset{n\to\infty}{\to}} \boldsymbol{h}(\boldsymbol{X})$.

Proof.

(a) Assume $\boldsymbol{X}_n \overset{\text{a.s.}}{\underset{n\to\infty}{\to}} \boldsymbol{X}$. Then $\lim_{n\to\infty} \boldsymbol{X}_n(\omega) = \boldsymbol{X}(\omega)$ for all $\omega \in N^c$ for some $N \in \mathcal{F}$ with $\mathbb{P}(N) = 0$.

Hence, we have for all $\omega \in N^c$, $\lim_{n\to\infty} \boldsymbol{h}(\boldsymbol{X}_n(\omega)) \overset{(\boldsymbol{h} \text{ continuous})}{=} \boldsymbol{h}(\lim_{n\to\infty} \boldsymbol{X}_n(\omega)) = \boldsymbol{h}(\boldsymbol{X}(\omega))$. This shows that $\{\omega \in N^c : \lim_{n\to\infty} \boldsymbol{X}_n(\omega) = \boldsymbol{X}(\omega)\} \subseteq \{\omega \in N^c : \lim_{n\to\infty} \boldsymbol{h}(\boldsymbol{X}_n(\omega)) = \boldsymbol{h}(\boldsymbol{X}(\omega))\}$.

Hence, we have

$$\mathbb{P}\Big(\lim_{n\to\infty} \boldsymbol{h}(\boldsymbol{X}_n) = \boldsymbol{h}(\boldsymbol{X})\Big) = \mathbb{P}\Big(\big\{\omega \in \Omega : \lim_{n\to\infty} \boldsymbol{h}(\boldsymbol{X}_n) = \boldsymbol{h}(\boldsymbol{X})\big\}\Big)$$

$$\overset{(\mathbb{P}(N)=0)}{=} \mathbb{P}\Big(\big\{\omega \in N^c : \lim_{n\to\infty} \boldsymbol{h}(\boldsymbol{X}_n) = \boldsymbol{h}(\boldsymbol{X})\big\}\Big)$$

$$\overset{\text{(monotonicity)}}{\geq} \mathbb{P}\Big(\big\{\omega \in N^c : \lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}\big\}\Big)$$

$$\overset{(\mathbb{P}(N)=0)}{=} \mathbb{P}\Big(\big\{\omega \in \Omega : \lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}\big\}\Big)$$

$$= \mathbb{P}\Big(\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}\Big) = 1.$$

(b) For this part we will provide two proofs:

    i. *Method 1: using subsequence principle.* Since $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{p}} \boldsymbol{X}$, by the subsequence principle we know for every subsequence $\{\boldsymbol{X}_{n_k}\}_{k\in\mathbb{N}} \subseteq \{\boldsymbol{X}_n\}_{n\in\mathbb{N}}$ (or $\{\boldsymbol{h}(\boldsymbol{X}_{n_k})\}_{k\in\mathbb{N}} \subseteq \{\boldsymbol{h}(\boldsymbol{X}_n)\}_{n\in\mathbb{N}}$), there exists a further subsequence $\{\boldsymbol{X}_{n_{k_\ell}}\}_{\ell\in\mathbb{N}} \subseteq \{\boldsymbol{X}_{n_k}\}_{k\in\mathbb{N}}$ (or $\{\boldsymbol{h}(\boldsymbol{X}_{n_{k_\ell}})\}_{\ell\in\mathbb{N}} \subseteq \{\boldsymbol{h}(\boldsymbol{X}_{n_k})\}_{k\in\mathbb{N}}$) such that $\boldsymbol{X}_{n_{k_\ell}} \xrightarrow[\ell\to\infty]{\text{a.s.}} \boldsymbol{X}$, which implies by (a) that $\boldsymbol{h}(\boldsymbol{X}_{n_{k_\ell}}) \xrightarrow[\ell\to\infty]{\text{a.s.}} \boldsymbol{h}(\boldsymbol{X})$. Hence, applying the subsequence principle on $\{\boldsymbol{h}(\boldsymbol{X}_n)\}_{n\in\mathbb{N}}$ gives $\boldsymbol{h}(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\text{p}} \boldsymbol{h}(\boldsymbol{X})$.

    ii. *Method 2: arguing by definition.* Fix any $\varepsilon > 0$. For each $m \in \mathbb{N}$, let $E_m := \{\boldsymbol{x} \in \mathbb{R}^d :$ there exists $\boldsymbol{y} \in \mathbb{R}^d$ such that $\|\boldsymbol{y} - \boldsymbol{x}\| < 1/m$ and $\|\boldsymbol{h}(\boldsymbol{y}) - \boldsymbol{h}(\boldsymbol{x})\| > \varepsilon\}$. Observe that $E_m \searrow$ and since $\boldsymbol{h}$ is continuous, $\lim_{n\to\infty} E_m = \bigcap_{m=1}^\infty E_m = \varnothing$ by considering the definition of continuity. Thus, by the continuity from above of $\mathbb{P}_{\boldsymbol{X}}$, we have $\lim_{m\to\infty} \mathbb{P}(\boldsymbol{X} \in E_m) = \mathbb{P}(\boldsymbol{X} \in \varnothing) = 0$.

Therefore, for all $m, n \in \mathbb{N}$,

$$\mathbb{P}(\|\boldsymbol{h}(\boldsymbol{X}_n) - \boldsymbol{h}(\boldsymbol{X})\| > \varepsilon) = \mathbb{P}(\|\boldsymbol{h}(\boldsymbol{X}_n) - \boldsymbol{h}(\boldsymbol{X})\| > \varepsilon, \|\boldsymbol{X}_n - \boldsymbol{X}\| < 1/m)$$
$$+ \mathbb{P}(\|\boldsymbol{h}(\boldsymbol{X}_n) - \boldsymbol{h}(\boldsymbol{X})\| > \varepsilon, \|\boldsymbol{X}_n - \boldsymbol{X}\| \geq 1/m)$$
$$\leq \mathbb{P}(\boldsymbol{X} \in E_n) + \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > 1/m).$$

By assumption, we know $\lim_{n\to\infty} \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{X}\| > 1/m) = 0$ for every $m \in \mathbb{N}$. Hence, taking $n \to \infty$ gives $0 \leq \lim_{n\to\infty} \mathbb{P}(\|\boldsymbol{h}(\boldsymbol{X}_n) - \boldsymbol{h}(\boldsymbol{X})\| > \varepsilon) = \lim_{n\to\infty} \mathbb{P}(\boldsymbol{X} \leq E_n) + 0 = 0$, implying that $\boldsymbol{h}(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\text{p}} \boldsymbol{h}(\boldsymbol{X})$.

$\square$

6.2 Convergence in L^p

6.2.1 **Definition.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X be a random variable, and $\{X_n\}$ be a sequence of random variables, where X and all X_n's are in $L^p = L^p(\Omega, \mathcal{F}, \mathbb{P})$, for some $p \in [1, \infty]$. Then $\{X_n\}$ **converges to X in L^p (or in the pth mean)**, denoted by $X_n \xrightarrow[n\to\infty]{L^p} X$, if $\lim_{n\to\infty} \|X_n - X\|_p = 0$.

[Note: It is customary to exclude the case where $p \in (0, 1)$ since for such p the L^p norm does not define a valid norm (in mathematical sense).]

6.2.2 **Properties of convergence in L^p.** Now, we will consider some properties of the convergence in L^p. In the proof, the following lemma that bounds tail probabilities is utilized:

Lemma 6.2.a (Tail probability bounds). Let $h : [0, \infty) \to [0, \infty)$ be a strictly increasing function and X be a random variable. Then, $\mathbb{P}(|X| \geq x) \leq \mathbb{E}[h(|X|)]/h(x)$ for all $x > 0$.

Proof. For all $x > 0$, we have

$$\mathbb{P}(|X| \geq x) \overset{(h \text{ strictly increasing})}{=} \mathbb{P}(h(|X|) \geq h(x))$$

$$= \mathbb{E}\left[\mathbf{1}_{\{h(|X|)\geq h(x)\}}\right] \overset{\text{(monotonicity)}}{\underset{(h(|X|)/h(x)\geq 1 \text{ if } h(|X|)\geq h(x))}{\leq}} \mathbb{E}\left[\frac{h(|X|)}{h(x)}\mathbf{1}_{\{h(|X|)\geq h(x)\}}\right]$$

$$\overset{\text{(monotonicity)}}{\leq} \mathbb{E}\left[\frac{h(|X|)}{h(x)}\right] = \frac{\mathbb{E}[h(|X|)]}{h(x)}.$$

Here, note that $h(x) > h(0) \geq 0$, so the division by $h(x)$ is well-defined. $\qquad\square$

[Note: By taking $h(x) = x$, the inequality reduces to $\mathbb{P}(|X| \geq x) \leq \mathbb{E}[|X|]/x \ \forall x > 0$, which is known as the **Markov's inequality**. Also, taking $h(x) = x^2$ gives $\mathbb{P}(|X| \geq x) \leq \mathbb{E}[X^2]/x^2 \ \forall x > 0$, which is known as the **Chebyshev's inequality**.]

Proposition 6.2.b.

(a) *(Higher order convergence implies lower order convergence)* For all $1 \leq p < q \leq \infty$, we have
$$X_n \xrightarrow[n\to\infty]{L^q} X \implies X_n \xrightarrow[n\to\infty]{L^p} X.$$

(b) *(Stronger than convergence in probability)* For all $p \in [1, \infty]$, we have $X_n \xrightarrow[n\to\infty]{L^p} X \implies X_n \xrightarrow[n\to\infty]{\text{P}} X$.

(c) *(Convergence of L^p norms)* For all $p \in [1, \infty]$, we have $X_n \xrightarrow[n\to\infty]{L^p} X \implies \lim_{n\to\infty} \|X_n\|_p = \|X\|_p$.

Proof.

(a) Assume $X_n \xrightarrow[n\to\infty]{L^q} X$. By definition we have $\lim_{n\to\infty} \|X_n - X\|_q = 0$. So, by [5.1.14]d, we have $0 \leq \lim_{n\to\infty} \|X_n - X\|_p \leq \lim_{n\to\infty} \|X_n - X\|_q = 0$ and thus the result follows.

(b) • *Case 1: $p \in [1, \infty)$.* Fix any $\varepsilon > 0$. For every $n \in \mathbb{N}$, by Lemma 6.2.a we know $\mathbb{P}(|X_n - X| > \varepsilon) \leq \mathbb{E}[|X_n - X|^p]/\varepsilon^p$. Since $X_n \xrightarrow[n\to\infty]{L^p} X \implies \lim_{n\to\infty} \|X_n - X\|_p^p = \lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$, we have $0 \leq \lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) \leq \lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$, implying that $X_n \xrightarrow[n\to\infty]{\text{P}} X$.

• *Case 2: $p = \infty$.* Since $|X_n - X| \overset{\text{a.s.}}{\leq} \text{ess sup} |X_n - X| = \|X_n - X\|_\infty$, i.e., $|X_n(\omega) - X(\omega)| \leq \|X_n - X\|_\infty$ for all $\omega \in N^c$ for some null set N. Hence, for all $\varepsilon > 0$, we have $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(\{|X_n - X| > \varepsilon\} \cap N^c) \leq \mathbb{P}(\{\|X_n - X\|_\infty > \varepsilon\} \cap N^c) = \mathbb{P}(\|X_n - X\|_\infty > \varepsilon)$ for every $n \in \mathbb{N}$.

It then follows that $0 \leq \lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) \leq \lim_{n\to\infty} \mathbb{P}(\|X_n - X\|_\infty > \varepsilon) \overset{\text{(assumption)}}{=} 0$, and so $X_n \xrightarrow[n\to\infty]{\text{P}} X$.

(c) Note first that

$$\begin{cases} \|X_n\|_p = \|X_n - X + X\|_p \overset{\text{(Minkowski)}}{\leq} \|X_n - X\|_p + \|X\|_p, \\ \|X_n\|_p = \|X - X_n + X_n\|_p \overset{\text{(Minkowski)}}{\leq} \|X_n - X\|_p + \|X_n\|_p, \end{cases}$$

which implies that

$$-\|X_n - X\|_p \leq \|X_n\|_p - \|X\|_p \leq \|X_n - X\|_p,$$

or $0 \leq \big|\|X_n\|_p - \|X\|_p\big| \leq \|X_n - X\|_p$. By assumption, $\lim_{n\to\infty} \|X_n - X\|_p = 0$, so we have $\lim_{n\to\infty}\big|\|X_n\|_p - \|X\|_p\big| = 0$, and hence $\lim_{n\to\infty} \|X_n\|_p = \|X\|_p$ by the definition of limit.

$\qquad\square$

6.3 Convergence in Distribution

6.3.1 **Definition.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\boldsymbol{X}$ be a random vector, and $\{\boldsymbol{X}_n\}$ be a sequence of random vectors, where $\boldsymbol{X}$ and all $\boldsymbol{X}_n$'s are from Ω to $\mathbb{R}^d$. Let $\boldsymbol{X}_n \sim F_n$ for every $n \in \mathbb{N}$ and $\boldsymbol{X} \sim F$. Then, $\{\boldsymbol{X}_n\}$ **converges to $\boldsymbol{X}$ in distribution (or weakly)**, denoted by $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathrm{d}} \boldsymbol{X}$, if $\lim_{n\to\infty} F_n(\boldsymbol{x}) = F(\boldsymbol{x})$ for all $\boldsymbol{x} \in C(F) := \{\boldsymbol{x} \in \mathbb{R}^d : F \text{ is continuous at } \boldsymbol{x}\}$, which is the set of all continuity points of F.

[Note: Since F is increasing, there are at most countably many discontinuities (jumps), so $C(F)^c$ is countable, and hence a Lebesgue null set. Therefore, with $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathrm{d}} \boldsymbol{X}$, we have $\lim_{n\to\infty} F_n(\boldsymbol{x}) = F(\boldsymbol{x})$ a.e. (wrt Lebesgue measure λ).]

6.3.2 **Uniqueness of limiting distribution.** The limiting distribution function F appearing in the definition of convergence in distribution is indeed unique.

Proof. Suppose $\lim_{n\to\infty} F_n(\boldsymbol{x}) = F(\boldsymbol{x})\ \forall \boldsymbol{x} \in C(F)$ and $\lim_{n\to\infty} F_n(\boldsymbol{x}) = \widetilde{F}(\boldsymbol{x})\ \forall \boldsymbol{x} \in C(\widetilde{F})$. Then, we have $F(\boldsymbol{x}) = \lim_{n\to\infty} F_n(\boldsymbol{x})j = \lim_{n\to\infty} \widetilde{F}_n(\boldsymbol{x}) = \widetilde{F}(\boldsymbol{x})\ \forall \boldsymbol{x} \in C(F) \cap C(\widetilde{F}) = (C(F)^c \cup C(\widetilde{F}^c)^c =: N^c$, with N being a Lebesgue null set. Hence, we have $F = \widetilde{F}$ on N^c.

It then remains to establish that $F = \widetilde{F}$ on N also. For all $\boldsymbol{x} \in \mathbb{N}$, by the right-continuity of F and $\widetilde{F}$ we have $F(\boldsymbol{x}) = \lim_{\substack{\boldsymbol{z}\to\boldsymbol{x}^+ \\ \boldsymbol{z}\in N^c}} F(\boldsymbol{z}) = \lim_{\substack{\boldsymbol{z}\to\boldsymbol{x}^+ \\ \boldsymbol{z}\in N^c}} \widetilde{F}(\boldsymbol{z}) = \widetilde{F}(\boldsymbol{x})$ where we can avoid choosing $\boldsymbol{z}$'s from N since N is countable. This establishes the uniqueness. $\qquad\square$

6.3.3 **Portmanteau theorem.** To work with convergence in distribution, we often rely on the *Portmanteau theorem*, which provides a useful characterization of convergence in distribution.

Theorem 6.3.a (Portmanteau). We have $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathrm{d}} \boldsymbol{X}$ iff $\lim_{n\to\infty} \mathbb{E}[h(\boldsymbol{X}_n)] = \mathbb{E}[h(\boldsymbol{X})]$ for every bounded and continuous function $h : \mathbb{R}^d \to \mathbb{R}$.

Proof. Let F_n denote the distribution function of $\boldsymbol{X}_n$ for every $n \in \mathbb{N}$, and F denote the distribution function of $\boldsymbol{X}$.

"$\Rightarrow$": We are going to utilize the definition of limit.

Fix any $\varepsilon > 0$ and any continuous and bounded function h. Then, by the boundedness we have $|h(\boldsymbol{x})| \leq M$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

Showing that $\mathbb{P}(\boldsymbol{X} \in I^c)$ is sufficiently small with $I := [\boldsymbol{a}, \boldsymbol{b}]$. By the contrapositive of Proposition 2.5.c, we know that at each discontinuity point (jump) of F, it corresponds to a discontinuity point (jump) of some margin of F (and also vice versa). Together with the fact that every margin can have at most countably many discontinuity points (jumps), we know that the set of discontinuity points of F can be expressed as $D = \prod_{j=1}^d D_j$ where D_j is the countable set of discontinuities of of the margin F_j of F, which is a Lebesgue null set.

Since $\lim_{\boldsymbol{a}\to-\infty, \boldsymbol{b}\to\infty} \Delta_{(\boldsymbol{a},\boldsymbol{b}]} F = 1$, for every $\varepsilon > 0$, there exist $\boldsymbol{a}, \boldsymbol{b} \in D^c$ such that $\Delta_{[\boldsymbol{a},\boldsymbol{b}]} F \overset{(F \text{ is continuous at } \boldsymbol{a})}{=} \Delta_{(\boldsymbol{a},\boldsymbol{b}]} F \geq 1 - \varepsilon/6M$; here again we can choose $\boldsymbol{a}, \boldsymbol{b}$ from D^c since D is countable. Therefore, $\mathbb{P}(\boldsymbol{X} \in I^c) = 1 - \Delta_{[\boldsymbol{a},\boldsymbol{b}]} F \leq \varepsilon/6M$.

Constructing an approximation h_ε that is sufficiently close to h uniformly. Since h is continuous on the compact I, it is uniformly continuous on I. Hence, there exists a partition $I = \biguplus_{k=1}^m I_k$ with $m \in \mathbb{N}$, where each I_k is a rectangle with endpoints in D^c (possible since D is countable), such that $\sup_{\boldsymbol{x},\boldsymbol{y}\in I_k} |h(\boldsymbol{x}) - h(\boldsymbol{y})| \leq \varepsilon/6$ (from the uniform continuity).

Using this partition, we can construct h_ε by choosing $\boldsymbol{x}_k \in I_k$ for each $k = 1, \ldots, m$, and then defining $h_\varepsilon(\boldsymbol{x}) := \sum_{k=1}^m h(\boldsymbol{x}_k)\mathbf{1}_{I_k}(\boldsymbol{x})$. Hence, by construction we have $|h(\boldsymbol{x}) - h_\varepsilon(\boldsymbol{x})| \leq \varepsilon/6$ for all $\boldsymbol{x} \in I$.

Showing that $\mathbb{E}[h(\boldsymbol{X})]$ and $\mathbb{E}[h_\varepsilon(\boldsymbol{X})]$ are sufficiently close. By the sufficiently tight upper bounds

on $\mathbb{P}(\boldsymbol{X} \in I^c)$ and $|h(\boldsymbol{x}) - h_\varepsilon(\boldsymbol{x})|$ established above, we have

$$\left|\mathbb{E}[h(\boldsymbol{X})] - \mathbb{E}[h_\varepsilon(\boldsymbol{X})]\right| = \left|\mathbb{E}[h(\boldsymbol{X}) - h_\varepsilon(\boldsymbol{X})]\right| \overset{\text{(triangle inequality)}}{\leq} \mathbb{E}[|h(\boldsymbol{X}) - h_\varepsilon(\boldsymbol{X})|]$$

$$= \mathbb{E}[|h(\boldsymbol{X}) - h_\varepsilon(\boldsymbol{X})|\mathbf{1}_I(\boldsymbol{X})] + \mathbb{E}[|h(\boldsymbol{X}) - h_\varepsilon(\boldsymbol{X})|\mathbf{1}_{I^c}(\boldsymbol{X})]$$

$$\overset{\text{(bound, } h_\varepsilon = 0 \text{ on } I^c)}{\leq} \mathbb{E}[(\varepsilon/6)\mathbf{1}_I(\boldsymbol{X})] + \mathbb{E}[|h(\boldsymbol{X}) - 0|\mathbf{1}_{I^c}(\boldsymbol{X})] \overset{(|h(\boldsymbol{x})|\leq M)}{\leq} (\varepsilon/6)\mathbb{P}(\boldsymbol{X} \in I) + \mathbb{E}[M\mathbf{1}_{I^c}(\boldsymbol{X})]$$

$$= (\varepsilon/6)\mathbb{P}(\boldsymbol{X} \in I) + M\mathbb{P}(\boldsymbol{X} \in I^c) \overset{\text{(bound)}}{\leq} (\varepsilon/6)\mathbb{P}(\boldsymbol{X} \in I) + M(\varepsilon/6M) = \varepsilon/3.$$

Showing that $\mathbb{E}[h(\boldsymbol{X}_n)]$ and $\mathbb{E}[h_\varepsilon(\boldsymbol{X}_n)]$ are sufficiently close with sufficiently large n.

Note that we have $\lim_{n\to\infty} \mathbb{P}(\boldsymbol{X}_n \in I^c) = 1 - \lim_{n\to\infty} \Delta_I F_n \overset{\text{(endpoints of } I \text{ are continuity points)}}{\underset{(\boldsymbol{X}_n \xrightarrow{\text{d}}_{n\to\infty} \boldsymbol{X})}{=}} 1 - \lim_{n\to\infty} \Delta_I F =$

$\lim_{n\to\infty} \mathbb{P}(\boldsymbol{X} \in I^c) \leq \varepsilon/6M$. Therefore, for all n that are sufficiently large, we have $\mathbb{P}(\boldsymbol{X}_n \in I^c) \leq \varepsilon/3M$, and thus as in above we get $|\mathbb{E}[h(\boldsymbol{X}_n)] - \mathbb{E}[h_\varepsilon(\boldsymbol{X}_n)]| \leq \varepsilon/6 + M\mathbb{P}(\boldsymbol{X}_n \in I^c) \leq \varepsilon/6 + \varepsilon/3 = \varepsilon/2$.

Showing that $\mathbb{E}[h_\varepsilon(\boldsymbol{X}_n)]$ and $\mathbb{E}[h_\varepsilon(\boldsymbol{X})]$ are sufficiently close with sufficiently large n. With sufficiently large n, we have

$$\left|\mathbb{E}[h_\varepsilon(\boldsymbol{X}_n)] - \mathbb{E}[h_\varepsilon(\boldsymbol{X})]\right| = \left|\mathbb{E}[h_\varepsilon(\boldsymbol{X}_n) - h_\varepsilon(\boldsymbol{X})]\right| = \left|\sum_{k=1}^m h(\boldsymbol{x}_k)(\mathbb{P}(\boldsymbol{X}_n \in I_k) - \mathbb{P}(\boldsymbol{X} \in I_k))\right|$$

$$\overset{\text{(triangle inequality)}}{\leq} \sum_{k=1}^m |h(\boldsymbol{x}_k)| \cdot \underbrace{|\mathbb{P}(\boldsymbol{X}_n \in I_k) - \mathbb{P}(\boldsymbol{X} \in I_k)|}_{\substack{\to 0 \text{ by assumption as} \\ \text{endpoints of } I_k \text{ are continuity points}}}$$

$$\overset{(n \text{ sufficiently large})}{<} \varepsilon/6.$$

Showing that $\mathbb{E}[h(\boldsymbol{X}_n)]$ and $\mathbb{E}[h(\boldsymbol{X})]$ are sufficiently close through $\mathbb{E}[h_\varepsilon(\boldsymbol{X}_n)]$ and $\mathbb{E}[h_\varepsilon(\boldsymbol{X})]$. Combining all the sufficiently tight bounds obtained above, we get

$$|\mathbb{E}[h(\boldsymbol{X}_n)] - \mathbb{E}[h(\boldsymbol{X})]| = |\mathbb{E}[h(\boldsymbol{X}_n)] - \mathbb{E}[h_\varepsilon(\boldsymbol{X}_n)] + \mathbb{E}[h_\varepsilon(\boldsymbol{X}_n)] - \mathbb{E}[h_\varepsilon(\boldsymbol{X})] + \mathbb{E}[h_\varepsilon(\boldsymbol{X})] - \mathbb{E}[h(\boldsymbol{X})]|$$

$$\overset{\text{(triangle inequality)}}{\leq} |\mathbb{E}[h(\boldsymbol{X}_n)] - \mathbb{E}[h_\varepsilon(\boldsymbol{X}_n)]| + |\mathbb{E}[h_\varepsilon(\boldsymbol{X}_n)] - \mathbb{E}[h_\varepsilon(\boldsymbol{X})]| + |\mathbb{E}[h_\varepsilon(\boldsymbol{X})] - \mathbb{E}[h(\boldsymbol{X})]|$$

$$< \varepsilon/2 + \varepsilon/6 + \varepsilon/3 = \varepsilon.$$

"$\Leftarrow$": We will establish that $\lim_{n\to\infty} F_n(\boldsymbol{x}) = F(\boldsymbol{x}) \ \forall \boldsymbol{x} \in C(F)$ by considering $\liminf_{n\to\infty} F_n(\boldsymbol{x})$ and $\limsup_{n\to\infty} F_n(\boldsymbol{x})$.

Showing that $F(\boldsymbol{x}) \leq \liminf_{n\to\infty} F_n(\boldsymbol{x}) \ \forall \boldsymbol{x} \in C(F)$ by constructing a multilinear h_ε. Fix any $\boldsymbol{x} \in C(F)$ and any $\boldsymbol{\varepsilon} > 0$. Construct the multilinear function $h_{\boldsymbol{\varepsilon}}(\boldsymbol{z}) = \prod_{j=1}^d \max\{\min\{(x_j - z_j)/\varepsilon_j, 1\}, 0\} \ \forall \boldsymbol{z} \in \mathbb{R}^d$, which satisfies $\mathbf{1}_{(-\infty, \boldsymbol{x}-\boldsymbol{\varepsilon}]}(\boldsymbol{z}) \leq h_{\boldsymbol{\varepsilon}}(\boldsymbol{z}) \leq \mathbf{1}_{(-\infty, \boldsymbol{x}]}(\boldsymbol{z}) \ \forall \boldsymbol{z} \in \mathbb{R}^d$. [Intuition 💡: The function $h_{\boldsymbol{\varepsilon}}$ is obtained by linear interpolating of the indicator functions $\mathbf{1}_{(-\infty, \boldsymbol{x}-\boldsymbol{\varepsilon}]}$ and $\mathbf{1}_{(-\infty, \boldsymbol{x}]}$.

]
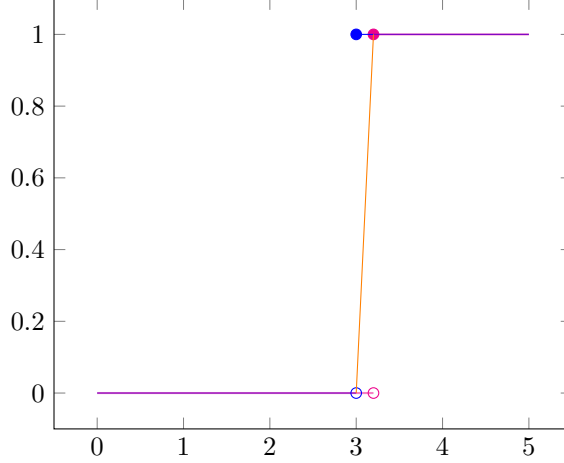
Then, we have $F(\boldsymbol{x} - \boldsymbol{\varepsilon}) = \mathbb{E}\big[\mathbf{1}_{(-\infty, \boldsymbol{x} - \boldsymbol{\varepsilon}]}(\boldsymbol{X})\big] \leq \mathbb{E}[h_\varepsilon(\boldsymbol{X})]$ and $\mathbb{E}[h_\varepsilon(\boldsymbol{X}_n)] \leq \mathbb{E}\big[\mathbf{1}_{(-\infty, \boldsymbol{x}]}(\boldsymbol{X}_n)\big] = F_n(\boldsymbol{x})$. Since $h_{\boldsymbol{\varepsilon}}$ is continuous and bounded, we have

$$\liminf_{n \to \infty} F_n(\boldsymbol{x}) \geq \liminf_{n \to \infty} \mathbb{E}[h_\varepsilon(\boldsymbol{X}_n))] \overset{\text{(assumption)}}{=} \lim_{n \to \infty} \mathbb{E}[h_\varepsilon(\boldsymbol{X}_n)] \overset{\text{(assumption)}}{=} \mathbb{E}[h_\varepsilon(\boldsymbol{X}))] \geq F(\boldsymbol{x} - \boldsymbol{\varepsilon}).$$

As this holds for all $\boldsymbol{\varepsilon} > \mathbf{0}$, we have $F(\boldsymbol{x}) \overset{(\boldsymbol{x} \in C(F))}{=} \lim_{\boldsymbol{\varepsilon} \to \mathbf{0}^+} F(\boldsymbol{x} - \boldsymbol{\varepsilon}) \leq \liminf_{n \to \infty} F_n(\boldsymbol{x})$.

Showing that $\limsup_{n \to \infty} F_n(\boldsymbol{x}) \leq F(\boldsymbol{x}) \ \forall \boldsymbol{x} \in C(F)$ **by constructing a multilinear** $h_{\boldsymbol{\varepsilon}}$. Using a similar idea, we construct a multilinear function $h_{\boldsymbol{\varepsilon}}$, given by $h_{\boldsymbol{\varepsilon}}(\boldsymbol{z}) := \prod_{j=1}^d \max\{\min\{(x_j + \varepsilon_j - z_j)/\varepsilon_j, 1\}, 0\} \ \forall \boldsymbol{z} \in \mathbb{R}^d$, which satisfies $\mathbf{1}_{(-\infty, \boldsymbol{x}]}(\boldsymbol{z}) \leq h_{\boldsymbol{\varepsilon}}(\boldsymbol{z}) \leq \mathbf{1}_{(-\infty, \boldsymbol{x} + \boldsymbol{\varepsilon}]}(\boldsymbol{z}) \ \forall \boldsymbol{z} \in \mathbb{R}^d$.

Similarly, we have $F_n(\boldsymbol{x}) = \mathbb{E}\big[\mathbf{1}_{(-\infty, \boldsymbol{x}]}(\boldsymbol{X}_n)\big] \leq \mathbb{E}[h_{\boldsymbol{\varepsilon}}(\boldsymbol{X}_n)]$ and $\mathbb{E}[h_{\boldsymbol{\varepsilon}}(\boldsymbol{X}))] \leq \mathbb{E}\big[\mathbf{1}_{(-\infty, \boldsymbol{x} + \boldsymbol{\varepsilon}]}(\boldsymbol{X})\big] = F(\boldsymbol{x} + \boldsymbol{\varepsilon})$. Likewise, $h_{\boldsymbol{\varepsilon}}$ is continuous and bounded, so

$$\limsup_{n \to \infty} F_n(\boldsymbol{x}) \leq \limsup_{n \to \infty} \mathbb{E}[h_{\boldsymbol{\varepsilon}}(\boldsymbol{X}_n)] = \lim_{n \to \infty} \mathbb{E}[h_{\boldsymbol{\varepsilon}}(\boldsymbol{X}_n)] = \mathbb{E}[h_{\boldsymbol{\varepsilon}}(\boldsymbol{X})] \leq F(\boldsymbol{x} + \boldsymbol{\varepsilon}),$$

meaning that $\limsup_{n \to \infty} F_n(\boldsymbol{x}) \leq \lim_{\boldsymbol{\varepsilon} \to \mathbf{0}^+} F(\boldsymbol{x} + \boldsymbol{\varepsilon}) = F(\boldsymbol{x})$.

Completing the proof. Combining the two inequalities obtained gives $F(\boldsymbol{x}) \leq \liminf_{n \to \infty} F_n(\boldsymbol{x}) \leq \limsup_{n \to \infty} F_n(\boldsymbol{x}) \leq F(\boldsymbol{x}) \ \forall \boldsymbol{x} \in C(F)$, which implies that $\liminf_{n \to \infty} F_n(\boldsymbol{x}) = \limsup_{n \to \infty} F_n(\boldsymbol{x}) = F(\boldsymbol{x}) \ \forall \boldsymbol{x} \in C(F)$, and thus $\lim_{n \to \infty} F_n(\boldsymbol{x}) = F(\boldsymbol{x}) \ \forall \boldsymbol{x} \in C(F)$. □

6.3.4 **Properties of convergence in distribution.** While it takes quite some work to prove the Portmanteau theorem, it can be utilized for establishing many properties about convergence in distribution conveniently, such as the following:

(a) $\boldsymbol{X}_n \xrightarrow[n \to \infty]{\text{p}} \boldsymbol{X} \implies \boldsymbol{X}_n \xrightarrow[n \to \infty]{\text{d}} \boldsymbol{X}$.

(b) $\boldsymbol{X}_n \xrightarrow[n \to \infty]{\text{d}} \boldsymbol{c} \implies \boldsymbol{X}_n \xrightarrow[n \to \infty]{\text{p}} \boldsymbol{c}$, where $\boldsymbol{c} \in \mathbb{R}^d$ is a constant.

Proof.

(a) We will apply the Portmanteau theorem and the dominated convergence theorem for convergence in probability (Corollary 6.1.d). Now, fix any bounded and continuous function $h : \mathbb{R}^d \to \mathbb{R}$, and we check the conditions for Corollary 6.1.d:

- Since h is continuous, it is measurable by [3.1.3]c. Hence, by [3.1.3]b, $h(\boldsymbol{X}_n)$ is measurable for all $n \in \mathbb{N}$. Also, $h(\boldsymbol{X}_n)$ is bounded as h is bounded. Therefore, $h(\boldsymbol{X}_n)$ is integrable, i.e., in L^1, for each $n \in \mathbb{N}$.

- We have $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{P}} \boldsymbol{X} \stackrel{\text{(CMT)}}{\implies} h(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\text{P}} h(\boldsymbol{X})$.
- By the boundedness, we have $|h| \leq M < \infty$ for some $M > 0$. Hence, we know $|h(\boldsymbol{X}_n)| \leq M$ for all $n \in \mathbb{N}$, satisfying the dominating condition as the constant function (degenerate random variable) M is in L^1 (we have $\mathbb{E}[|M|] = |M|\mathbb{P}(\Omega) = |M| < \infty$).

Hence, by Corollary 6.1.d, we have $\lim_{n\to\infty} \mathbb{E}[h(\boldsymbol{X}_n)] = \mathbb{E}[h(\boldsymbol{X})]$, and the result follows by the Portmanteau theorem.

(b) Here we will use the fact that the *Euclidean norm* (2-norm) of $\boldsymbol{x}$ is less than or equal to the *taxicab norm* (1-norm) of $\boldsymbol{x}$: $(\sum_{j=1}^{d} x_j^2)^{1/2} =: \|\boldsymbol{x}\| \leq \|\boldsymbol{x}\|_1 := \sum_{j=1}^{d} |x_j|$.

Note that the distribution function of the degenerate random vector $\boldsymbol{c}$ is $F(\boldsymbol{x}) = \mathbf{1}_{[\boldsymbol{c},\boldsymbol{\infty})}(\boldsymbol{x})$ with $C(F) = \mathbb{R}^d \setminus \{\boldsymbol{x} : x_j = c_j \text{ and } x_k \geq c_k \ \forall k \geq j\}$.

Fix any $n \in \mathbb{N}$ and $\varepsilon > 0$. Since $\{\omega \in \Omega : \max_{j=1,\dots,d} |X_{nj}(\omega) - c_j| \leq \varepsilon/d\} \subseteq \{\omega \in \Omega : \|\boldsymbol{X}_n - \boldsymbol{c}\|_1 \leq \varepsilon\} \stackrel{(\|\cdot\|\leq\|\cdot\|_1)}{\subseteq} \{\omega \in \Omega : \|\boldsymbol{X}_n - \boldsymbol{c}\| \leq \varepsilon\}$, we have

$$\mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{c}\| > \varepsilon) = 1 - \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{c}\| < \varepsilon) \leq 1 - \mathbb{P}\left(\max_{j=1,\dots,d} |X_{nj} - c_j| \leq \frac{\varepsilon}{d}\right)$$

$$= 1 - \mathbb{P}\left(|X_{nj} - c_j| \leq \frac{\varepsilon}{d} \ \forall j = 1,\dots,d\right) = 1 - \Delta_{[\boldsymbol{c}-\frac{\boldsymbol{\varepsilon}}{d},\boldsymbol{c}+\frac{\boldsymbol{\varepsilon}}{d}]} F_n,$$

where F_n is the distribution function of $\boldsymbol{X}_n = (X_{n1},\dots,X_{nd})$ and $\boldsymbol{\varepsilon} = (\varepsilon,\dots,\varepsilon)$.

Since all endpoints of $[\boldsymbol{c}-\frac{\boldsymbol{\varepsilon}}{d},\boldsymbol{c}+\frac{\boldsymbol{\varepsilon}}{d}]$ are in $C(F)$ (as the jth component cannot be c_j), we have

$$\lim_{n\to\infty} \mathbb{P}(\|\boldsymbol{X}_n - \boldsymbol{c}\| > \varepsilon) = 1 - \lim_{n\to\infty} \Delta_{[\boldsymbol{c}-\frac{\boldsymbol{\varepsilon}}{d},\boldsymbol{c}+\frac{\boldsymbol{\varepsilon}}{d}]} F_n$$

$$\stackrel{\left(\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{d}} \boldsymbol{c}\right)}{=} 1 - \Delta_{[\boldsymbol{c}-\frac{\boldsymbol{\varepsilon}}{d},\boldsymbol{c}+\frac{\boldsymbol{\varepsilon}}{d}]} F \stackrel[\text{(other terms are zero)}]{\text{(expand by definition)}}{=} 1 - F(c_1 + \varepsilon/d,\dots,c_d + \varepsilon/d)$$

$$\stackrel{(c_j \leq c_1 + \varepsilon/d \ \forall j)}{=} 1 - 1 = 0,$$

so the result follows.

$\square$

6.3.5 **Continuous mapping theorem for convergence in distribution.** As mentioned in [6.1.7], there is also a continuous mapping theorem for convergence in distribution, and here we are ready to prove it using the Portmanteau theorem (elegantly!).

Theorem 6.3.b (Continuous mapping theorem for convergence in distribution)**.** Let $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^k$ be continuous. Then, $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{d}} \boldsymbol{X} \implies \boldsymbol{h}(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\text{d}} \boldsymbol{h}(\boldsymbol{X})$.

Proof. Fix any bounded and continuous function $g : \mathbb{R}^k \to \mathbb{R}$. Then the composition $g \circ \boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}$ is also bounded and continuous. Hence, by "$\Rightarrow$" direction of Portmanteau theorem on $g \circ \boldsymbol{h}(\boldsymbol{X}_n)$, we have $\lim_{n\to\infty} \mathbb{E}[g(\boldsymbol{h}(\boldsymbol{X}_n))] = \mathbb{E}[g(\boldsymbol{h}(\boldsymbol{X}))]$.

As this holds for every bounded and continuous function $g : \mathbb{R}^k \to \mathbb{R}$, applying the "$\Leftarrow$" direction of Portmanteau theorem on $g(\boldsymbol{h}(\boldsymbol{X}_n))$ gives $\boldsymbol{h}(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\text{d}} \boldsymbol{h}(\boldsymbol{X})$. $\square$

6.4 Uniform Integrability

6.4.1 In Section 6.2, we do not find a mode of convergence that implies the convergence in L^p. As it turns out, neither almost sure convergence nor convergence in probability would imply convergence in L^p; see Section 6.6. However, if the *uniform integrability* is further assumed, then such implication can be obtained; see Theorem 6.4.c. Let us start by defining what uniform integrability is and looking at some examples.

6.4.2 **Definition.** A collection $\{X_i\}_{i\in I} \subseteq L^1$ is said to be **uniformly integrable** if $\lim_{a\to\infty} \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] = 0$.

[Intuition 💡: Uniform integrability suggests that the "tail" part of $|X_i|$ (indicated by $\mathbf{1}_{\{|X_i|>a\}}$) would not "explode" and hence we have the <u>integrability</u>. The word "<u>uniform</u>" refers to the fact that we are considering the supremum over all $i \in I$ in the definition.]

To better understand uniform integrability, consider the properties in [6.4.3].

6.4.3 Properties of uniform integrability.

(a) The singleton $\{X\}$ with $X \in L^1$ is uniformly integrable.

(b) If $|X_i| \leq Y$ for all $i \in I$ with $Y \in L^1$, then $\{X_i\}_{i\in I}$ is uniformly integrable.

(c) A finite collection $\{X_i\}_{i=1}^n \subseteq L^1$ is uniformly integrable.

Proof.

(a) Since $X \in L^1$, we know X is finite a.e. by [5.1.11]b, and thus $\lim_{a\to\infty} |X|\mathbf{1}_{\{|X|>a\}} = 0$ a.e. Also, we have $|X|\mathbf{1}_{\{|X|>a\}} \leq |X| \; \forall a \geq 0$ with $|X| \in L^1$. Hence, by DCT we have $\lim_{a\to\infty} \mathbb{E}\big[|X|\mathbf{1}_{\{|X|>a\}}\big] = 0$, and so $\{X\}$ is uniformly integrable.

(b) Since $|X_i| \leq Y$ for all $i \in I$, we have $0 \leq \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] \leq \sup_{i\in I} \mathbb{E}\big[Y\mathbf{1}_{\{Y>a\}}\big] \xrightarrow[a\to\infty]{(a)} 0$. Therefore, we have $\lim_{a\to\infty} \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] = 0$.

(c) Note that we have $|X_j| \leq \sum_{i=1}^n |X_i|$ with $\sum_{i=1}^n |X_i| \in L^1$ for every $j = 1, \ldots, n$, and hence $\{X_i\}_{i=1}^n$ is uniformly integrable by (b). $\qquad\square$

6.4.4 Characterization of uniform integrability. The following characterization of uniform integrability is helpful for establishing results about uniform integrability.

Theorem 6.4.a. A collection $\{X_i\}_{i\in I} \subseteq L^1$ is uniformly integrable iff

(a) *(uniform bounded first absolute moments)* $\sup_{i\in I} \mathbb{E}[|X_i|] < \infty$, and

(b) *(uniform absolute continuity)* for all $\varepsilon > 0$, there exists $\delta > 0$ such that $\sup_{i\in I} \mathbb{E}[|X_i|\mathbf{1}_A] < \varepsilon$ for every $A \in \mathcal{F}$ with $\mathbb{P}(A) < \delta$.

Proof.

- "$\Rightarrow$": For all $i \in I$ and $a \in (0, \infty)$, we have $\mathbb{E}[|X_i|\mathbf{1}_A] = \mathbb{E}\big[|X_i|\mathbf{1}_{A\cap\{|X_i|\leq a\}}\big] + \mathbb{E}\big[|X_i|\mathbf{1}_{A\cap\{|X_i|>a\}}\big] \leq a\mathbb{P}(A) + \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big]$. Thus, we have $\sup_{i\in I} \mathbb{E}[|X_i|\mathbf{1}_A] \leq a\mathbb{P}(A) + \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] \; \forall a \in (0, \infty)$.

  Taking $A = \Omega$ gives (a). Next, fix any $\varepsilon > 0$. By the uniform integrability we have $\lim_{a\to\infty} \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] = 0$, so there exists sufficiently large $a > 0$ such that $\sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] < \varepsilon/2$. We then choose $\delta = \varepsilon/2a$. With this δ, for every A with $\mathbb{P}(A) < \delta$, we have $\sup_{i\in I} \mathbb{E}[|X_i|\mathbf{1}_A] \leq a\mathbb{P}(A) + \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] < a \cdot \varepsilon/2a + \varepsilon/2 = \varepsilon$.

- "$\Leftarrow$": Fix any $\varepsilon > 0$. Applying (b) with $A = \{|X_i| > a\}$, we know there exists $\delta > 0$ such that $\sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] < \varepsilon/2$ for all i with $\mathbb{P}(|X_i| > a) < \delta$. Since $\sup_{i\in I} \mathbb{P}(|X_i| > a) \overset{\text{(Markov)}}{\leq} \sup_{i\in I} \mathbb{E}[|X_i|]/a =: c/a \overset{(a)}{<} \infty$, there exists sufficiently large a such that $\mathbb{P}(|X_i| > a) < \delta$ for all $i \in I$.

  With such choice of a, we then have $\mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] < \varepsilon/2$ for all $i \in I$, and hence $0 \leq \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] \leq \varepsilon/2 < \varepsilon$. By the definition of limit, we have $\lim_{a\to\infty} \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] = 0$. $\qquad\square$

6.4.5 **Convergence in probability implies convergence in L^p under uniform integrability.** To prove the implication we have suggested at the beginning of Section 6.4, the following lemma is needed.

Lemma 6.4.b. Let $X \in L^1$ and $A_n \in \mathcal{F}$ for each $n \in \mathbb{N}$. If $\lim_{n\to\infty} \mathbb{P}(A_n) = 0$, then $\lim_{n\to\infty} \mathbb{E}[X\mathbf{1}_{A_n}] = 0$.

Proof. Fix any $\varepsilon > 0$. Since $\{X\}$ is uniformly integrable by [6.4.3], there exists $a > 0$ such that $\mathbb{E}[|X|\mathbf{1}_{\{|X|>a\}}] < \varepsilon/2$. Also, by assumption there exists $n_\varepsilon \in \mathbb{N}$ such that $\mathbb{P}(A_n) < \varepsilon/2a$ for all $n \geq n_\varepsilon$. Hence,

$$
\begin{aligned}
|\mathbb{E}[X\mathbf{1}_{A_n}]| &\overset{\text{(Jensen)}}{\leq} \mathbb{E}[|X|\mathbf{1}_{A_n}] = \mathbb{E}\big[|X|\mathbf{1}_{A_n\cap\{|X|\leq a\}}\big] + \mathbb{E}\big[|X|\mathbf{1}_{A_n\cap\{|X|>a\}}\big] \\
&\leq \mathbb{E}[a\mathbf{1}_{A_n}] + \mathbb{E}\big[|X|\mathbf{1}_{\{|X|>a\}}\big] = a\mathbb{P}(A_n) + \mathbb{E}\big[|X|\mathbf{1}_{\{|X|>a\}}\big] \\
&< a \cdot \varepsilon/2a + \varepsilon/2 = \varepsilon.
\end{aligned}
$$

$\square$

Theorem 6.4.c. Let X be a random variable and $\{X_n\}$ be a sequence of random variables. If $X_n \xrightarrow[n\to\infty]{\text{P}} X$ and $\{|X_n|^p\}$ is uniformly integrable for some $p \in [1, \infty)$, then $X_n \xrightarrow[n\to\infty]{L^p} X$.

Proof. **Showing that $X \in L^p$.** Since $X_n \xrightarrow[n\to\infty]{\text{P}} X$, by subsequence principle we know for every subsequence $\{X_{n_k}\}_{k\in\mathbb{N}} \subseteq \{X_n\}_{n\in\mathbb{N}}$, there exists a further subsequence $\{X_{n_{k_\ell}}\}_{\ell\in\mathbb{N}} \subseteq \{X_{n_k}\}_{k\in\mathbb{N}}$ such that $X_{n_{k_\ell}} \xrightarrow[\ell\to\infty]{\text{a.s.}} X$. Hence, we have $\mathbb{E}[|X|^p] \overset{\left(|X_{n_{k_\ell}}|^p \overset{\text{a.s.}}{\underset{\ell\to\infty}{\longrightarrow}} |X|^p\right)}{=} \mathbb{E}\big[\liminf_{\ell\to\infty}|X_{n_{k_\ell}}|^p\big] \overset{\text{(Fatou)}}{\leq} \liminf_{\ell\to\infty} \mathbb{E}\big[|X_{n_{k_\ell}}|^p\big] \leq \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|^p] \overset{\text{(Theorem 6.4.a)}}{<} \infty$. Thus, $X \in L^p$.

Getting an upper bound for $\mathbb{E}[|X_n - X|^p]$. Note that $|X_n - X|^p \leq (|X_n|+|X|)^p \leq (2\max\{|X_n|, |X|\})^p = 2^p \max\{|X_n|^p, |X|^p\} \leq 2^p(|X_n|^p + |X|^p)$. Fix any $\varepsilon > 0$. Then we have

$$
\begin{aligned}
\mathbb{E}[|X_n - X|^p] &= \mathbb{E}\big[|X_n - X|^p\mathbf{1}_{\{|X_n-X|\leq\varepsilon\}}\big] + \mathbb{E}\big[|X_n - X|^p\mathbf{1}_{\{|X_n-X|>\varepsilon\}}\big] \\
&\leq \mathbb{E}\big[(\varepsilon)^p\mathbf{1}_{\{|X_n-X|\leq\varepsilon\}}\big] + \mathbb{E}\big[2^p(|X_n|^p + |X|^p)\mathbf{1}_{\{|X_n-X|>\varepsilon\}}\big] \\
&\leq \varepsilon \cdot 1 + 2^p\mathbb{E}\big[|X_n|^p\mathbf{1}_{\{|X_n-X|>\varepsilon\}}\big] + 2^p\mathbb{E}\big[|X|^p\mathbf{1}_{\{|X_n-X|>\varepsilon\}}\big].
\end{aligned}
$$

Completing the proof by getting bounds in terms of ε. Since $X \in L^p$, we have $|X|^p \in L^1$ and hence by Lemma 6.4.b we have $\lim_{n\to\infty} \mathbb{E}\big[|X|^p\mathbf{1}_{\{|X_n-X|>\varepsilon\}}\big] = 0$, because $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$ by the assumption that $X_n \xrightarrow[n\to\infty]{\text{P}} X$. Thus, there exists $n_\varepsilon \in \mathbb{N}$ such that $\mathbb{E}\big[|X|^p\mathbf{1}_{\{|X_n-X|>\varepsilon\}}\big] < \varepsilon$ for all $n \geq n_\varepsilon$.

Also, since $\{|X_n|^p\}$ is uniformly integrable, by Theorem 6.4.a there exists $\delta > 0$ such that $\sup_{i\in\mathbb{N}} \mathbb{E}\big[|X_i|^p\mathbf{1}_{\{|X_n-X|>\varepsilon\}}\big] < \varepsilon$ whenever $\mathbb{P}(|X_n - X| > \varepsilon) < \delta$. Now, by the assumption that $X_n \xrightarrow[n\to\infty]{\text{P}} X$, we know there exists $\widetilde{n}_\varepsilon \in \mathbb{N}$ such that $\mathbb{P}(|X_n - X| > \varepsilon) < \delta$, which implies that

$$
\mathbb{E}\big[|X_n|^p\mathbf{1}_{\{|X_n-X|>\varepsilon\}}\big] \leq \sup_{i\in\mathbb{N}} \mathbb{E}\big[|X_i|^p\mathbf{1}_{\{|X_n-X|>\varepsilon\}}\big] < \varepsilon,
$$

for all $n \geq \widetilde{n}_\varepsilon$.

Therefore, for all $n \geq \max\{n_\varepsilon, \widetilde{n}_\varepsilon\}$, we have $\mathbb{E}[|X_n - X|^p] < \varepsilon + 2^p\varepsilon + 2^p\varepsilon = (1 + 2^{p+1})\varepsilon$, which then implies that $\lim_{n\to\infty} \|X_n - X\|_p = 0$ (after changing $\varepsilon \to \varepsilon/(1 + 2^{p+1})$ above). $\square$

6.5 Slutsky's Theorem

6.5.1 Apart from the *continuous mapping theorem*, another crucial result for working with modes of convergence is the *Slutsky's theorem*, which suggests conditions under which we can "combine" individual convergences together.

Based on the CMT, we have $(\boldsymbol{X}_n, \boldsymbol{Y}_n) \xrightarrow[n\to\infty]{d} (\boldsymbol{X}, \boldsymbol{Y}) \implies h(\boldsymbol{X}_n, \boldsymbol{Y}_n) \xrightarrow[n\to\infty]{d} h(\boldsymbol{X}, \boldsymbol{Y})$ for every continuous function h. Particularly, assuming the $\boldsymbol{X}$'s and $\boldsymbol{Y}$'s both take values in $\mathbb{R}^d$, we would have $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow[n\to\infty]{d} \boldsymbol{X} + \boldsymbol{Y}$.

However, if we only have the individual convergences $\boldsymbol{X}_n \xrightarrow[n\to\infty]{d} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow[n\to\infty]{d} \boldsymbol{Y}$, then we do not have $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow[n\to\infty]{d} \boldsymbol{X} + \boldsymbol{Y}$ in general. To see this, take $X \sim \mathrm{N}(0,1)$ and $Y = -X \stackrel{d}{=} X$. Now, take a sequence $\{X_n\}$ such that $X_n \xrightarrow[n\to\infty]{d} X$. Since $Y \stackrel{d}{=} X$, by taking $\{Y_n\} = \{X_n\}$, we have $Y_n \xrightarrow[n\to\infty]{d} Y$ also. So, in such case, we have $X_n \xrightarrow[n\to\infty]{d} X$ and $Y_n \xrightarrow[n\to\infty]{d} Y$, but $X_n + Y_n = 2X_n \xrightarrow[n\to\infty]{d} 2X$. Of course, $2X$ and $X + Y = 0$ do not have the same distribution.

6.5.2 **Condition for joint convergence in distribution.** Other than imposing the independence condition, there is a relatively less restrictive condition that can ensure joint convergence in distribution from individual convergence, which requires one of the sequences to converge in distribution to a *constant*. Intuitively, this works because converging to a constant can avoid influencing the tail behaviour of another sequence.

Theorem 6.5.a. If $\boldsymbol{X}_n \xrightarrow[n\to\infty]{d} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow[n\to\infty]{d} \boldsymbol{c}$ where $\boldsymbol{c} \in \mathbb{R}^d$ is a constant, then $(\boldsymbol{X}_n, \boldsymbol{Y}_n) \xrightarrow[n\to\infty]{d} (\boldsymbol{X}, \boldsymbol{c})$.
[Note: Here, it is not necessary for the $\boldsymbol{X}$'s to take values from the same space as the $\boldsymbol{Y}$'s.]

Proof. Let F_n and F denote the distribution functions of $(\boldsymbol{X}_n, \boldsymbol{Y}_n)$ and $(\boldsymbol{X}, \boldsymbol{c})$ respectively. Then, we have $F(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{P}(\boldsymbol{X} \le \boldsymbol{x}, \boldsymbol{c} \le \boldsymbol{y}) = F_{\boldsymbol{X}}(\boldsymbol{x})\mathbf{1}_{[\boldsymbol{c},\infty)}(\boldsymbol{y})$. Now fix any $(\boldsymbol{x}, \boldsymbol{y}) \in C(F)$ and $\varepsilon > 0$ (which implies that $\boldsymbol{x} \in C(F_{\boldsymbol{X}})$).

Upper bounding $\limsup_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y})$. Consider

$$
\begin{aligned}
F_n(\boldsymbol{x}, \boldsymbol{y}) &= \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x}, \boldsymbol{Y}_n \le \boldsymbol{y}, \|\boldsymbol{Y}_n - \boldsymbol{c}\| \le \varepsilon) + \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x}, \boldsymbol{Y}_n \le \boldsymbol{y}, \|\boldsymbol{Y}_n - \boldsymbol{c}\| > \varepsilon) \\
&\le \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x}, \boldsymbol{c} \le \boldsymbol{y} + {}_d\varepsilon) + \mathbb{P}(\|\boldsymbol{Y}_n - \boldsymbol{c}\| > \varepsilon) \\
&= F_{\boldsymbol{X}_n}(\boldsymbol{x})\mathbf{1}_{[\boldsymbol{c},\infty)}(\boldsymbol{y} + {}_d\varepsilon) + \mathbb{P}(\|\boldsymbol{Y}_n - \boldsymbol{c}\| > \varepsilon).
\end{aligned}
$$

By [6.3.4], we have $\boldsymbol{Y}_n \xrightarrow[n\to\infty]{d} \boldsymbol{c} \implies \boldsymbol{Y}_n \xrightarrow[n\to\infty]{p} \boldsymbol{c}$, which gives

$$
\limsup_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y}) \le \limsup_{n\to\infty} F_{\boldsymbol{X}_n}(\boldsymbol{x})\mathbf{1}_{[\boldsymbol{c},\infty)}(\boldsymbol{y} + {}_d\varepsilon) + \underbrace{\limsup_{n\to\infty} \mathbb{P}(\|\boldsymbol{Y}_n - \boldsymbol{c}\| > \varepsilon)}_{0}
$$

$$
\stackrel{(\boldsymbol{x}\in C(F_{\boldsymbol{X}}))}{=} F_{\boldsymbol{X}}(\boldsymbol{x})\mathbf{1}_{[\boldsymbol{c},\infty)}(\boldsymbol{y} + {}_d\varepsilon) = F(\boldsymbol{x}, \boldsymbol{y} + {}_d\varepsilon).
$$

Lower bounding $\liminf_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y})$. Since $\boldsymbol{x} \in C(F_{\boldsymbol{X}})$, by definition of continuity, for all $\delta > 0$ there exists $n_0 \in \mathbb{N}$ such that $F_{\boldsymbol{X}_n}(\boldsymbol{x}) \ge F_{\boldsymbol{X}}(\boldsymbol{x}) - \delta$ for all $n \ge n_0$. Hence, for all $n \ge n_0$,

$$
\begin{aligned}
F(\boldsymbol{x}, \boldsymbol{y} - {}_d\varepsilon) &= F_{\boldsymbol{X}}(\boldsymbol{x})\mathbf{1}_{[\boldsymbol{c},\infty)}(\boldsymbol{y} - {}_d\varepsilon) \\
&\le (F_{\boldsymbol{X}_n}(\boldsymbol{x}) + \delta)\mathbf{1}_{[\boldsymbol{c},\infty)}(\boldsymbol{y} - {}_d\varepsilon) \le F_{\boldsymbol{X}_n}(\boldsymbol{x})\mathbf{1}_{[\boldsymbol{c},\infty)}(\boldsymbol{y} - {}_d\varepsilon) + \delta \\
&= \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x}, \boldsymbol{c} \le \boldsymbol{y} - {}_d\varepsilon) + \delta \\
&= \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x}, \boldsymbol{c} \le \boldsymbol{y} - {}_d\varepsilon, \|\boldsymbol{Y}_n - \boldsymbol{c}\| \le \varepsilon) + \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x}, \boldsymbol{c} \le \boldsymbol{y} - {}_d\varepsilon, \|\boldsymbol{Y}_n - \boldsymbol{c}\| > \varepsilon) + \delta \\
&\le \mathbb{P}(\boldsymbol{X}_n \le \boldsymbol{x}, \boldsymbol{Y}_n \le \boldsymbol{y}) + \mathbb{P}(\|\boldsymbol{Y}_n - \boldsymbol{c}\| > \varepsilon) + \delta \\
&= F_n(\boldsymbol{x}, \boldsymbol{y}) + \mathbb{P}(\|\boldsymbol{Y}_n - \boldsymbol{c}\| > \varepsilon) + \delta.
\end{aligned}
$$

Therefore, like before we get $F(\boldsymbol{x}, \boldsymbol{y} - {}_d\varepsilon) \le \liminf_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y}) + 0 + \delta$.

Completing the proof by letting $\delta \to 0^+$ and $\varepsilon \to 0^+$. Combining the bounds obtained previously, we get

$$
F(\boldsymbol{x}, \boldsymbol{y} - {}_d\varepsilon) - \delta \le \liminf_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y}) \le \limsup_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y}) \le F(\boldsymbol{x}, \boldsymbol{y} + {}_d\varepsilon).
$$

Letting $\delta \to 0^+$ yields

$$F(\boldsymbol{x}, \boldsymbol{y} - {}_d\varepsilon) \le \liminf_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y}) \le \limsup_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y}) \le F(\boldsymbol{x}, \boldsymbol{y} + {}_d\varepsilon).$$

Since $(\boldsymbol{x}, \boldsymbol{y}) \in C(F)$, we have $\lim_{\varepsilon\to 0^+} F(\boldsymbol{x}, \boldsymbol{y} - {}_d\varepsilon) = F(\boldsymbol{x}, \boldsymbol{y})$ and $\lim_{\varepsilon\to 0^+} F(\boldsymbol{x}, \boldsymbol{y} + {}_d\varepsilon) = F(\boldsymbol{x}, \boldsymbol{y})$. Letting $\varepsilon \to 0^+$ thus yields

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

which means $\lim_{n\to\infty} F_n(\boldsymbol{x}, \boldsymbol{y}) = F(\boldsymbol{x}, \boldsymbol{y})$. $\qquad\square$

6.5.3 **Slutsky's theorem.** Slutsky's theorem is indeed an immediate corollary of Theorem 6.5.a (but is a more well-known result).

Theorem 6.5.b (Slutsky's theorem). *If $\boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathrm{d}} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow[n\to\infty]{\mathrm{d}} \boldsymbol{c}$ where $\boldsymbol{c} \in \mathbb{R}^d$ is a constant, then $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow[n\to\infty]{\mathrm{d}} \boldsymbol{X} + \boldsymbol{c}$ and $\boldsymbol{X}_n \boldsymbol{Y}_n \xrightarrow[n\to\infty]{\mathrm{d}} \boldsymbol{c}\boldsymbol{X}$, where the $\boldsymbol{X}$'s and $\boldsymbol{Y}$'s are assumed to all take values from $\mathbb{R}^d$, and the "multiplications" of vectors are interpreted in the componentwise sense.*

Proof. The result follows by applying Theorem 6.5.a and the continuous mapping theorem with the continuous functions $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} + \boldsymbol{y}$ and $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}\boldsymbol{y} = (x_1 y_1, \dots, x_d y_d)$ respectively. $\qquad\square$

6.6 Counterexamples About Implications Between Modes of Convergence

6.6.1 **A summary of implications between modes of convergence.** The implications between the major modes of convergence can be summarized in the picture below. Naturally, we would also like to investigate whether the converse of each implication holds. In Section 6.6, we will provide counterexamples to illustrate that the converse indeed does not hold (without additional assumptions).

$$\boldsymbol{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} \boldsymbol{X}$$
$$\Downarrow$$
$$\boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathrm{p}} \boldsymbol{X} \quad\Longrightarrow\quad \boldsymbol{X}_n \xrightarrow[n\to\infty]{\mathrm{d}} \boldsymbol{X}$$
$$\Uparrow$$
$$X_n \xrightarrow[n\to\infty]{L^p} X$$

6.6.2 **An useful lemma for generating counterexamples.** The following lemma provides us a systematic way to generate counterexamples about the implications between modes of convergence.

Lemma 6.6.a. *Let $\{X_n\}_{n\in\mathbb{N}}$ be independent with $\mathbb{P}(X_n = 0) = 1 - 1/n^\alpha$ and $\mathbb{P}(X_n = n) = 1/n^\alpha$ for each $n \in \mathbb{N}$, where $\alpha > 0$. Then:*

(a) $X_n \xrightarrow[n\to\infty]{\text{a.s.}} 0$ *iff* $\alpha > 1$.

(b) $X_n \xrightarrow[n\to\infty]{\mathrm{p}} 0$ *for every* $\alpha > 0$.

(c) $X_n \xrightarrow[n\to\infty]{L^p} 0$ *iff* $\alpha > p$.

Proof.

(a) Fix any $\varepsilon > 0$. Note that $\mathbb{P}(|X_n - 0| > \varepsilon) = \mathbb{P}(X_n > \varepsilon) = \mathbb{P}(X_n = n) = 1/n^\alpha$. Hence, $\sum_{n=1}^\infty \mathbb{P}(|X_n - 0| > \varepsilon) = \sum_{n=1}^\infty 1/n^\alpha < \infty$ iff $\alpha > 1$ by the convergence criterion for p-series. Now, by the first and second Borel-Cantelli lemmas, we have $\mathbb{P}(X_n = n \text{ io}) = \mathbf{1}_{(0,1]}(\alpha)$, and thus $\mathbb{P}(\lim_{n\to\infty} X_n = 0) = \mathbb{P}(X_n = 0 \text{ abfm}) = 1 - \mathbb{P}(X_n = n \text{ io}) = \mathbf{1}_{\{\alpha>1\}} = 1$ iff $\alpha > 1$, as desired.

(b) Fix any $\varepsilon > 0$ and any $\alpha > 0$. Since we have $\mathbb{P}(|X_n - 0| > \varepsilon) \overset{\text{(same as (a))}}{=} 1/n^\alpha$ for each $n \in \mathbb{N}$, it follows that $\lim_{n\to\infty} \mathbb{P}(|X_n - 0| > \varepsilon) = \lim_{n\to\infty} 1/n^\alpha = 0$.

(c) Fix any $\alpha > 0$ and $p > 0$. The result follows by noting that $\lim_{n\to\infty} \mathbb{E}[|X_n - 0|^p] = \lim_{n\to\infty} \mathbb{E}[X_n^p] = \lim_{n\to\infty} n^p \cdot (1/n^\alpha) = \lim_{n\to\infty} n^{p-\alpha} = 0$ iff $\alpha > p$.

$\square$

6.6.3 **List of counterexamples.** With the help of Lemma 6.6.a, we can obtain many counterexamples fairly easily:

(1) $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X \Longleftarrow\!\!\!/\; X_n \xrightarrow[n\to\infty]{\text{P}} X$: The sequence in Lemma 6.6.a with $\alpha \in (0,1]$.

Proof. By Lemma 6.6.a, we know $X_n \xrightarrow[n\to\infty]{\text{P}} 0$ while $X_n \xrightarrow[n\to\infty]{\text{a.s.}}\!\!\!\!\!\!\!/\;\; 0$. $\square$

(2) $X_n \xrightarrow[n\to\infty]{L^p} X \Longleftarrow\!\!\!/\; X_n \xrightarrow[n\to\infty]{\text{P}} X$: The sequence in Lemma 6.6.a with $\alpha \in (0,p)$.

Proof. By Lemma 6.6.a, we know $X_n \xrightarrow[n\to\infty]{\text{P}} 0$ while $X_n \xrightarrow[n\to\infty]{L^p}\!\!\!\!\!\!\!/\;\; 0$. $\square$

(3) $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X \Longrightarrow\!\!\!\!/\; X_n \xrightarrow[n\to\infty]{L^p} X$: The sequence in Lemma 6.6.a with $\alpha \in (1,p]$.

Proof. By Lemma 6.6.a, we know $X_n \xrightarrow[n\to\infty]{\text{a.s.}} 0$ while $X_n \xrightarrow[n\to\infty]{L^p}\!\!\!\!\!\!\!/\;\; 0$. $\square$

(4) $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X \Longleftarrow\!\!\!/\; X_n \xrightarrow[n\to\infty]{L^p} X$: The *typewriter sequence* defined by $X_1 = \mathbf{1}_{[0,1]}, X_2 = \mathbf{1}_{[0,1/2]}, X_3 = \mathbf{1}_{[1/2,1]}, X_4 = \mathbf{1}_{[0,1/3]}, X_5 = \mathbf{1}_{[1/3,2/3]}, X_6 = \mathbf{1}_{[2/3,1]}, \ldots$, where each X_n is on $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \bar{\mathcal{B}}([0,1]), \lambda)$. The sequence can be more compactly expressed by $X_{\frac{n(n-1)}{2}+k} = \mathbf{1}_{[\frac{k-1}{n}, \frac{k}{n}]}$ for all $k = 1, \ldots, n$ and $n \in \mathbb{N}$. [Intuition 💡: The intervals in the indicator functions "move" like a typewriter.]

Proof. For all $k = 1, \ldots, n$, we have $\mathbb{E}\left[\left|X_{\frac{n(n-1)}{2}+k} - 0\right|^p\right] = \mathbb{E}\left[\mathbf{1}_{[\frac{k-1}{n}, \frac{k}{n}]}\right] = \lambda([\frac{k-1}{n}, \frac{k}{n}]) = \frac{1}{n} \to 0$ as $n \to \infty$, for every $p \in [1, \infty)$. Consequently, we have $X_m \xrightarrow[m\to\infty]{L^p} 0$.

However, for all $\omega \in [0,1]$ and all $n \in \mathbb{N}$, there exists a unique $k = 1, \ldots, n$ such that $X_{\frac{n(n-1)}{2}+k}(\omega) = 1$ (the ω falls into exactly one of those intervals), and $X_{\frac{n(n-1)}{2}+k}(\omega) = 0$ for other k's. Therefore, we have $\limsup_{m\to\infty} X_m(\omega) = 1$ and $\liminf_{m\to\infty} X_m(\omega) = 0$, which means that $\{X_m\}$ does not converge at every $\omega \in \Omega$, and thus definitely not converging almost surely. $\square$

(5) $X_n \xrightarrow[n\to\infty]{\text{P}} X \Longleftarrow\!\!\!/\; X_n \xrightarrow[n\to\infty]{\text{d}} X$: Let X follow the *Rademacher distribution*, which is given by $\mathbb{P}(X = -1) = \mathbb{P}(X = 1) = 1/2$, and $X_n := (-1)^n X$ for every $n \in \mathbb{N}$.

Proof. Since $X, X_1, X_2, \ldots$ all have the same distribution, we have $X_n \xrightarrow[n\to\infty]{\text{d}} X$. However, for every $\varepsilon \in (0,2)$, we have

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}\left(\left|((-1)^n - 1)X\right| > \varepsilon\right) = \begin{cases} \mathbb{P}(0 > \varepsilon) = 0 & \text{if } n \text{ is even,} \\ \mathbb{P}(2 > \varepsilon) = 1 & \text{if } n \text{ is odd,} \end{cases}$$

implying that $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon)$ does not even exist, and so $\{X_n\}$ does not converge in probability. $\square$

With these counterexamples, we can enrich our picture above as follows:

$$X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$$

$$X_n \xrightarrow[n\to\infty]{\text{p}} X \quad \overset{\Longleftarrow}{\Longrightarrow} \quad X_n \xrightarrow[n\to\infty]{\text{d}} X$$

$$X_n \xrightarrow[n\to\infty]{L^p} X$$

6.7 Applications of Modes of Convergence

6.7.1 After studying different modes of convergence and their relationships, we will now discuss some of their applications, including (i) convergence of quantile functions, (ii) *laws of large numbers*, and (iii) *Glivenko-Cantelli theorem*, which is an important theorem for statistics. For the famous *central limit theorem*, we will discuss it in Section 7.

6.7.2 Convergence of quantile functions.

Proposition 6.7.a. If $\lim_{n\to\infty} F_n(x) = F(x) \ \forall x \in C(F)$ *(convergence in distribution)*, then $\lim_{n\to\infty} F_n^{-1}(u) = F^{-1}(u) \ \forall u \in (0,1) \cap C(F^{-1})$.

Proof. Fix any $u \in (0,1) \cap C(F^{-1})$.

Showing that $\liminf_{n\to\infty} F_n^{-1}(u) \geq F^{-1}(u)$**.** Since F has at most countably many discontinuities (jumps), we know $C(F)^c$ is countable. Hence, for all $\varepsilon > 0$, we can choose $x \in C(F)$ such that $F^{-1}(u) - \varepsilon < x < F^{-1}(u)$. Now, as $x \in C(F)$, we have $\lim_{n\to\infty} F_n(x) = F(x) < u$, where $F(x) < u$ holds because $x < F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}$, which means that $\inf\{x \in \mathbb{R} : F(x) \geq u\}$ is not a lower bound for x, implying that $F(x) < u$.

By the definition of limit, we know that there exists $n_0 \in \mathbb{N}$ such that $F_n(x) < u$ for all $n \geq n_0$. This implies by [2.5.10] that $F_n^{-1}(u) \geq x > F^{-1}(u) - \varepsilon$ for all $n \geq n_0$, and hence $\liminf_{n\to\infty} F_n^{-1}(u) \geq x > F^{-1}(u) - \varepsilon$. Letting $\varepsilon \to 0^+$ then yields $\liminf_{n\to\infty} F_n^{-1}(u) \geq F^{-1}(u)$.

Showing that $\limsup_{n\to\infty} F_n^{-1}(u) \leq F^{-1}(u)$**.** Again, since $C(F)^c$ is countable, for all $\varepsilon > 0$ and all $u' > u$, there exists $x' \in C(F)$ such that $F^{-1}(u') < x' < F^{-1}(u') + \varepsilon$. Then, we have $\lim_{n\to\infty} F_n(x') \overset{(x' \in C(F))}{=} F(x') \overset{(F^{-1}(u') < x')}{\geq} u' > u$, and hence there exists $n_0' \in \mathbb{N}$ such that $F_n(x') > u$ for all $n \geq n_0$.

Likewise, by [2.5.10] we have $F_n^{-1}(u) \leq x' < F^{-1}(u') + \varepsilon$ for all $n \geq n_0$, and so $\limsup_{n\to\infty} F_n^{-1}(u) \leq x' < F^{-1}(u') + \varepsilon$. Letting $\varepsilon \to 0^+$ then yields $\limsup_{n\to\infty} F_n^{-1}(u) \leq F^{-1}(u')$. Further letting $u' \to u^+$ yields $\limsup_{n\to\infty} F_n^{-1}(u) \leq F^{-1}(u)$, as $u \in C(F^{-1})$.

Completing the proof. Collecting the two inequalities above gives $F^{-1}(u) \leq \liminf_{n\to\infty} F_n^{-1}(u) \leq \limsup_{n\to\infty} F_n^{-1}(u) \leq F^{-1}(u)$, which implies that $\lim_{n\to\infty} F_n^{-1}(u) = F^{-1}(u)$, as desired. $\qquad\square$

6.7.3 Laws of large numbers.

Theorem 6.7.b (Laws of large numbers)**.**

(a) *(Weak law of large numbers (WLLN))* Let $\{X_n\} \subseteq L^2$ be iid with mean $\mu = \mathbb{E}[X_1]$ and variance $\sigma^2 = \text{Var}(X_1)$ $(< \infty$ as the random variables are in $L^2)$. Then, $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow[n\to\infty]{\text{p}} \mu$.

(b) *(Strong law of large numbers (SLLN))* Let $\{X_n\} \subseteq L^1$ be iid with mean $\mu = \mathbb{E}[X_1]$. Then, $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow[n\to\infty]{\text{a.s.}} \mu$.

[Note: As their names suggest, the SLLN indeed implies the weak law of large numbers. While it may seem to be unnecessary to cover the WLLN then, the proof of the WLLN (without using the SLLN) turns out

to be much easier than the one for the SLLN. Also, in many statistical applications, the WLLN is already enough.]

Proof. We will only cover the proof of the weak law of large number here. For all $\varepsilon > 0$, we have

$$0 \leq \mathbb{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) \overset{\text{(monotonicity)}}{\leq} \mathbb{P}\big(|\bar{X}_n - \mu| \geq \varepsilon\big) \overset{\text{(Chebyshev)}}{\leq} \frac{\mathbb{E}\big[(\bar{X}_n - \mu)^2\big]}{\varepsilon^2} = \frac{\text{Var}\big(\bar{X}_n\big)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Letting $n \to \infty$ then gives $\lim_{n\to\infty} \mathbb{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) = 0$, as desired. $\qquad\square$

6.7.4 **Glivenko-Cantelli theorem.** The Glivenko-Cantelli theorem is about convergence of *empirical distribution functions*. Let $\{\boldsymbol{X}_n\}_{n\in\mathbb{N}}$ be iid. For every fixed $\boldsymbol{x} \in \mathbb{R}^d$, the collection $\{\mathbf{1}_{\{\boldsymbol{X}_n \leq \boldsymbol{x}\}}\}_{n\in\mathbb{N}}$ contains iid random variables in L^1. So, by the SLLN, for the empirical distribution function $F_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\boldsymbol{X}_i \leq \boldsymbol{x}\}}$, we have the *pointwise* almost sure convergence:

$$F_n(\boldsymbol{x}) \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}\big[\mathbf{1}_{\{\boldsymbol{X}_1 \leq \boldsymbol{x}\}}\big] = \mathbb{P}(\boldsymbol{X}_1 \leq \boldsymbol{x}) = F(\boldsymbol{x})$$

for all $\boldsymbol{x} \in \mathbb{R}^d$.

The Glivenko-Cantelli theorem asserts an even stronger result, namely that such almost sure convergence is indeed also *uniform*.

Theorem 6.7.c (Glivenko-Cantelli). If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{iid}}{\sim} F$, then $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |F_n(\boldsymbol{x}) - F(\boldsymbol{x})| \xrightarrow[n\to\infty]{\text{a.s.}} 0$.

Proof. We only prove the case with $d = 1$ here. For the case with $d \geq 2$, see, e.g., Kiefer and Wolfowitz (1958).

Upper bounding $|F_n(x) - F(x)|$. Fix any $x \in [a, b)$. Since each F_n and F are increasing, we have

$$\begin{cases} F_n(x) - F(x) \geq F_n(a) - F(b-) = \big(F_n(a) - F(a)\big) - \big(F(b-) - F(a)\big), \\ F_n(x) - F(x) \leq F_n(b-) - F(a) = \big(F_n(b-) - F(b-)\big) + \big(F(b-) - F(a)\big), \end{cases}$$

where $F(b-) := \lim_{x\to b-} F(x)$. This implies that

$$|F_n(x) - F(x)| \leq \max\{|\big(F_n(a) - F(a)\big) - \big(F(b-) - F(a)\big)|, |\big(F_n(b-) - F(b-)\big) + \big(F(b-) - F(a)\big)|\}$$

$$\overset{\text{(triangle)}}{\leq} \max\{|F_n(a) - F(a)|, |F_n(b-) - F(b-)|\} + \big(F(b-) - F(a)\big).$$

Partitioning the y-axis for F with sufficiently close end-points.

Claim: For every $\varepsilon \in (0, 3]$, there exists $n_\varepsilon \in \mathbb{N}$ and a partition $-\infty =: z_0 < z_1 \cdots < z_{n_\varepsilon} := \infty$ such that $F(z_k-) - F(z_{k-1}) \leq \varepsilon/3$ for all $k = 1, \ldots, n_\varepsilon$.

Proof. In case F is continuous, taking $z_k = F^{-1}(\varepsilon k/3)$ for all $k = 0, 1, \ldots, \lfloor 3/\varepsilon \rfloor$, with $n_\varepsilon = \lfloor 3/\varepsilon \rfloor + 1$, would work. For the case where F has some discontinuities (jumps), note that there are only finitely many x for which $F(x) - F(x-) > \varepsilon/3$; otherwise, F could take value larger than 1. By including those x's as the end-points in the partition (raising n_ε by a finite number), we can then ensure that $F(z_k-) - F(z_{k-1}) \leq \varepsilon/3$ for all k still. $\qquad\square$

Upper bounding the supremum by maximal distances from the partition and applying the SLLN. Since every $x \in \mathbb{R}$ lies in exactly one of the partition intervals $[z_{k-1}, z_k)$, applying the upper bound on $|F_n(x) - F(x)|$ with $a = z_{k-1}$ and $b = z_k$ for all $k = 1, \ldots, n_\varepsilon$ gives

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \leq \max_{k=1,\ldots,n_\varepsilon} \Big\{\max\{|F_n(z_{k-1}) - F(z_{k-1})|, |F_n(z_k-) - F(z_k-)|\} + \underbrace{\big(F(z_k-) - F(z_{k-1})\big)}_{\leq \varepsilon/3}\Big\}$$

$$\leq \max_{k=1,\ldots,n_\varepsilon} \Big\{\underbrace{|F_n(z_{k-1}) - F(z_{k-1})|}_{\xrightarrow[n\to\infty]{\text{a.s.}} 0 \text{ by SLLN}}\Big\} + \max_{k=1,\ldots,n_\varepsilon} \Big\{\underbrace{|F_n(z_k-) - F(z_k-)|}_{\xrightarrow[n\to\infty]{\text{a.s.}} 0 \text{ by SLLN}}\Big\} + \varepsilon/3$$

$$\overset{(n \text{ sufficiently large})}{<} \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon,$$

implying that $\sup_{\boldsymbol{x} \in \mathbb{R}^d} |F_n(\boldsymbol{x}) - F(\boldsymbol{x})| \xrightarrow[n \to \infty]{\text{a.s.}} 0.$ □

Remarks:

- For a fixed $\boldsymbol{x} \in \mathbb{R}^d$, $F_n(\boldsymbol{x}) \xrightarrow[n \to \infty]{\text{a.s.}} F(\boldsymbol{x})$ is equivalent to $|F_n(\boldsymbol{x}) - F(\boldsymbol{x})| \xrightarrow[n \to \infty]{\text{a.s.}} 0$. So the statement is suggesting $F_n(\boldsymbol{x})$ is converging almost surely to $F(\boldsymbol{x})$ at an "uniform rate" for all $\boldsymbol{x} \in \mathbb{R}^d$.

- *(Rate of convergence)* The rate of convergence for $\sup_{\boldsymbol{x} \in \mathbb{R}^d} |F_n(\boldsymbol{x}) - F(\boldsymbol{x})|$ is quantified by the *Dvoretzky-Kiefer-Wolfowitz (DKW) inequality*: For all $\varepsilon > 0$, we have

$$\begin{cases} \mathbb{P}(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > \varepsilon) \leq 2e^{-2n\varepsilon^2} & \text{if } d = 1, \\ \mathbb{P}(\sup_{\boldsymbol{x} \in \mathbb{R}^d} |F_n(\boldsymbol{x}) - F(\boldsymbol{x})| > \varepsilon) \leq (n+1)de^{-2n\varepsilon^2} & \text{if } d \geq 2. \end{cases}$$

7 Characteristic Functions

7.0.1 From Theorem 3.2.a, we know that distribution can be characterized via the *distribution function*. Here, we will study another kind of function that can also characterize distributions, known as the *characteristic function* (cf). It serves as an alternative to the *moment generating function* you have learnt before, with the critical advantage that it always exists, unlike the moment generating function. As we shall see in Section 7.2, the characteristic function is the basis for proving the famous *central limit theorem*. Before going through the proof, we need to lay some groundwork in Section 7.1 first.

7.1 Definition and Properties of Characteristic Function

7.1.1 **Definition.** The **characteristic function** of $\boldsymbol{X}$ is a complex-valued function $\phi_{\boldsymbol{X}} : \mathbb{R}^d \to \mathbb{C}$ given by $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\left[e^{i\boldsymbol{t}^T\boldsymbol{X}}\right]$ for all $\boldsymbol{t} \in \mathbb{R}^d$.

Comparing with the moment generating function $M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\left[e^{\boldsymbol{t}^T\boldsymbol{X}}\right]$, the cf has an additional "i" in the power. So, the cdf takes a similar "form" as the moment generating function, and as one may expect, also carries similar properties as the moment generating function.

7.1.2 **Properties of characteristic functions.**

   (a) *(Explicit form of cfs)* $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\left[\cos(\boldsymbol{t}^T\boldsymbol{X})\right] + i\mathbb{E}\left[\sin(\boldsymbol{t}^T\boldsymbol{X})\right]$ for all $\boldsymbol{t} \in \mathbb{R}^d$.

   (b) $\mathbb{E}\left[\left|e^{i\boldsymbol{t}^T\boldsymbol{X}}\right|\right] = 1$ for all $\boldsymbol{t} \in \mathbb{R}^d$.

   (c) *(Existence)* The characteristic function $\phi_{\boldsymbol{X}}(\boldsymbol{t})$ exists for all $\boldsymbol{t} \in \mathbb{R}^d$.

   (d) *(Bounded)* $|\phi_{\boldsymbol{X}}(\boldsymbol{t})| \le 1$ for all $\boldsymbol{t} \in \mathbb{R}^d$.

   (e) $\phi_{\boldsymbol{X}}(\boldsymbol{0}) = 1$.

   (f) *(Uniform continuity)* The characteristic function $\phi_{\boldsymbol{X}}$ is uniformly continuous.

   (g) *(Effects after linear operations)* We have $\phi_{A\boldsymbol{x}+\boldsymbol{b}}(\boldsymbol{t}) = e^{i\boldsymbol{t}^T\boldsymbol{b}}\phi_{\boldsymbol{X}}(A^T\boldsymbol{t})$ for all $\boldsymbol{t} \in \mathbb{R}^d$.

   (h) *(Factorization under independence)* If $X_1, \ldots, X_d$ are independent, then $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \prod_{j=1}^{d} \phi_{X_j}(t_j)$ for all $\boldsymbol{t} \in \mathbb{R}^d$, and particularly, $\phi_{\sum_{j=1}^{d} X_j}(t) = \prod_{j=1}^{d} \phi_{X_j}(t)$ for all $t \in \mathbb{R}$.

   Proof.

   (a) By the Euler's formula $e^{ix} = \cos(x) + i\sin(x) \ \forall x \in \mathbb{R}$ and the definition of expectation of complex-valued function (see [5.1.9]), we have $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\left[\cos(\boldsymbol{t}^T\boldsymbol{X})\right] + i\mathbb{E}\left[\sin(\boldsymbol{t}^T\boldsymbol{X})\right]$.

   (b) By the definition of modulus, we have $|a + ib| = \sqrt{a^2 + b^2}$ for all $a, b \in \mathbb{R}$. Hence, for all $\boldsymbol{t} \in \mathbb{R}^d$,

$$\left|e^{i\boldsymbol{t}^T\boldsymbol{X}}\right| = \cos(\boldsymbol{t}^T\boldsymbol{X}) + i\sin(\boldsymbol{t}^T\boldsymbol{X})| = \sqrt{\cos^2(\boldsymbol{t}^T\boldsymbol{X}) + \sin^2(\boldsymbol{t}^T\boldsymbol{X})} = 1, \tag{9}$$

   and so the result follows.

   (c) By (b) we have $\mathbb{E}\left[\left|e^{i\boldsymbol{t}^T\boldsymbol{X}}\right|\right] = 1 < \infty$ for all $\boldsymbol{t} \in \mathbb{R}^d$. Thus, $e^{i\boldsymbol{t}^T\boldsymbol{X}}$ is integrable and hence $\phi_{\boldsymbol{X}}(\boldsymbol{t})$ exists for all $\boldsymbol{t} \in \mathbb{R}^d$.

   (d) By triangle inequality, we have $|\phi_{\boldsymbol{X}}(\boldsymbol{t})| = \left|\mathbb{E}\left[e^{i\boldsymbol{t}^T\boldsymbol{X}}\right]\right| \overset{\text{(triangle)}}{\le} \mathbb{E}\left[\left|e^{i\boldsymbol{t}^T\boldsymbol{X}}\right|\right] \overset{\text{(b)}}{=} 1$ for all $\boldsymbol{t} \in \mathbb{R}^d$.

   (e) Note that $\phi_{\boldsymbol{X}}(\boldsymbol{0}) = \mathbb{E}\left[e^0\right] = 1$.

   (f) Note first that

$$|\phi_{\boldsymbol{X}}(\boldsymbol{t}+\boldsymbol{h}) - \phi_{\boldsymbol{X}}(\boldsymbol{t})| = \left|\mathbb{E}\left[e^{i(\boldsymbol{t}+\boldsymbol{h})^T\boldsymbol{X}}\right] - \mathbb{E}\left[e^{i\boldsymbol{t}^T\boldsymbol{X}}\right]\right| = \left|\mathbb{E}\left[e^{i\boldsymbol{t}^T\boldsymbol{X}}(e^{i\boldsymbol{h}^T\boldsymbol{X}} - 1)\right]\right|$$

$$\overset{\text{(triangle)}}{\le} \mathbb{E}\left[\left|e^{i\boldsymbol{t}^T\boldsymbol{X}}(e^{i\boldsymbol{h}^T\boldsymbol{X}} - 1)\right|\right] \overset{\text{(Equation (9))}}{=} \mathbb{E}\left[\left|e^{i\boldsymbol{h}^T\boldsymbol{X}} - 1\right|\right].$$

Now fix any sequence $\{\boldsymbol{h}_n\} \underset{n\to\infty}{\longrightarrow} \boldsymbol{0}$. Since we have $\{|e^{i\boldsymbol{h}_n^T\boldsymbol{x}} - 1|\} \underset{n\to\infty}{\longrightarrow} 0$ and $|e^{i\boldsymbol{h}_n^T\boldsymbol{x}} - 1| \overset{\text{(triangle)}}{\leq} |e^{i\boldsymbol{h}_n^T\boldsymbol{x}}| + 1 \leq 2$ for all $n \in \mathbb{N}$, by DCT we conclude that $\lim_{\boldsymbol{h}_n\to\boldsymbol{0}} \mathbb{E}\left[\left|e^{i\boldsymbol{h}_n^T\boldsymbol{X}} - 1\right|\right] = 0$. As this holds for every sequence $\{\boldsymbol{h}_n\} \underset{n\to\infty}{\longrightarrow} \boldsymbol{0}$, the characteristic function $\phi_{\boldsymbol{X}}$ is uniformly continuous.

(g) For all $\boldsymbol{t} \in \mathbb{R}^d$, we have

$$\phi_{A\boldsymbol{X}+\boldsymbol{b}}(\boldsymbol{t}) = \mathbb{E}\left[e^{i\boldsymbol{t}^T(A\boldsymbol{X}+\boldsymbol{b})}\right] = e^{i\boldsymbol{t}^T\boldsymbol{b}}\mathbb{E}\left[e^{i\boldsymbol{t}^T(A\boldsymbol{X})}\right] = e^{i\boldsymbol{t}^T\boldsymbol{b}}\mathbb{E}\left[e^{i(A^T\boldsymbol{t})^T\boldsymbol{X}}\right] = e^{i\boldsymbol{t}^T\boldsymbol{b}}\phi_{\boldsymbol{X}}(A^T\boldsymbol{t}).$$

(h) Under the independence, we have

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\left[e^{i\boldsymbol{t}^T\boldsymbol{X}}\right] = \mathbb{E}\left[e^{i\sum_{j=1}^d t_j X_j}\right] = \mathbb{E}\left[\prod_{j=1}^d e^{it_j X_j}\right] \overset{\text{(Proposition 5.2.d)}}{=} \prod_{j=1}^d \mathbb{E}\left[e^{it_j X_j}\right] = \prod_{j=1}^d \phi_{X_j}(t_j)$$

for all $\boldsymbol{t} \in \mathbb{R}^d$. Particularly, for all $t \in \mathbb{R}$ we have $\phi_{\sum_{j=1}^d X_j}(t) = \mathbb{E}\left[it\sum_{j=1}^d X_j\right] \overset{(\boldsymbol{t}=(t,\ldots,t))}{=} \phi_{\boldsymbol{X}}(t) = \prod_{j=1}^d \phi_{X_j}(t)$.

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

7.1.3 Lévy's continuity theorem.
An important result that relates the *convergence in distribution* and the limit of characteristic functions is the *Lévy's continuity theorem*. It is a very helpful tool for many proofs involving characteristic functions, notably the *central limit theorem*.

Theorem 7.1.a (Lévy's continuity theorem)**.**

(a) If $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{d}}{\longrightarrow}} \boldsymbol{X}$, then $\lim_{n\to\infty} \phi_{\boldsymbol{X}_n}(\boldsymbol{t}) = \phi_{\boldsymbol{X}}(\boldsymbol{t})$ for all $\boldsymbol{t} \in \mathbb{R}^d$.

(b) If $\lim_{n\to\infty} \phi_{\boldsymbol{X}_n}(\boldsymbol{t})$ exists for all $\boldsymbol{t} \in \mathbb{R}^d$ and is continuous at $\boldsymbol{0}$ (as a function of $\boldsymbol{t}$), then $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{d}}{\longrightarrow}} \boldsymbol{X}$ where $\boldsymbol{X}$ is a random vector having characteristic function $\phi(\boldsymbol{t}) := \lim_{n\to\infty} \phi_{\boldsymbol{X}_n}(\boldsymbol{t})$.

Proof.

(a) Fix any $\boldsymbol{t} \in \mathbb{R}^d$. Consider the function $\boldsymbol{h}(\boldsymbol{x}) = e^{i\boldsymbol{t}^T\boldsymbol{x}}$. By Equation (9), we know that $\boldsymbol{h}$ is bounded. Also, note that $\boldsymbol{h}$ is continuous, as a composition of continuous functions. Hence, by the Portmanteau theorem, we have $\lim_{n\to\infty} \phi_{\boldsymbol{X}_n}(\boldsymbol{t}) = \lim_{n\to\infty} \mathbb{E}[\boldsymbol{h}(\boldsymbol{X}_n)] = \mathbb{E}[\boldsymbol{h}(\boldsymbol{X})] = \phi_{\boldsymbol{X}}(\boldsymbol{t})$.

(b) Omitted.

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

7.1.4 Cramér-Wold device.
Based on the characteristic function, we can establish the following result about the convergence in distribution of linear combinations.

Theorem 7.1.b (Cramér-Wold device)**.** Let $\boldsymbol{X}, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be random vectors. We have $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{d}}{\longrightarrow}} \boldsymbol{X}$ iff $\boldsymbol{a}^T\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{d}}{\longrightarrow}} \boldsymbol{a}^T\boldsymbol{X}$ for all $\boldsymbol{a} \in \mathbb{R}^d$.

Proof. "$\Rightarrow$": It follows by applying the CMT with $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x}$.
"$\Leftarrow$": For all $\boldsymbol{t} \in \mathbb{R}^d$, we have

$$\phi_{\boldsymbol{X}_n}(\boldsymbol{t}) = \mathbb{E}\left[e^{i(1)\boldsymbol{t}^T\boldsymbol{X}_n}\right] = \phi_{\boldsymbol{t}^T\boldsymbol{X}_n}(1) \overset{\text{(Theorem 7.1.a, assumption)}}{\underset{n\to\infty}{\longrightarrow}} \phi_{\boldsymbol{t}^T\boldsymbol{X}}(1) = \phi_{\boldsymbol{X}}(\boldsymbol{t}).$$

Since $\phi_{\boldsymbol{X}}$ is continuous at $\boldsymbol{0}$ by [7.1.2]f, applying Theorem 7.1.a again gives $\boldsymbol{X}_n \underset{n\to\infty}{\overset{\text{d}}{\longrightarrow}} \boldsymbol{X}$. $\qquad\square$

In particular, Theorem 7.1.b establishes the following helper result for dealing with equality in distribution.

Corollary 7.1.c. Let X and Y be random vectors. Then, $X \stackrel{\mathrm{d}}{=} Y$ iff $a^T X \stackrel{\mathrm{d}}{=} a^T Y$.

Proof. "$\Rightarrow$": Assume $X \stackrel{\mathrm{d}}{=} Y$. Taking $X_n = X \stackrel{\mathrm{d}}{=} Y$ for every $n \in \mathbb{N}$, we have $X_n \xrightarrow[n\to\infty]{\mathrm{d}} X$ and $X_n \xrightarrow[n\to\infty]{\mathrm{d}} Y$. Hence, applying Theorem 7.1.b twice yields $a^T X_n \xrightarrow[n\to\infty]{\mathrm{d}} a^T X$ and $a^T X_n \xrightarrow[n\to\infty]{\mathrm{d}} a^T Y$ for all $a \in \mathbb{R}^d$. Due to the uniqueness of limiting distribution ([6.3.2]), we must then have $a^T X \stackrel{\mathrm{d}}{=} a^T Y$ for all $a \in \mathbb{R}^d$.

"$\Leftarrow$": The idea is similar. Fix any $a \in \mathbb{R}^d$ and assume $a^T X \stackrel{\mathrm{d}}{=} a^T Y$. Taking $X_n = X$ for every $n \in \mathbb{N}$, we have $a^T X_n \xrightarrow[n\to\infty]{\mathrm{d}} a^T X$ and $a^T X_n \xrightarrow[n\to\infty]{\mathrm{d}} a^T Y$. Hence, applying Theorem 7.1.b twice yields $X_n \xrightarrow[n\to\infty]{\mathrm{d}} X$ and $X_n \xrightarrow[n\to\infty]{\mathrm{d}} Y$ for all $a \in \mathbb{R}^d$. Due to the uniqueness of limiting distribution ([6.3.2]), we must then have $X \stackrel{\mathrm{d}}{=} Y$ for all $a \in \mathbb{R}^d$. $\qquad\square$

7.1.5 **Characterizing distribution by characteristic function.** As mentioned at the beginning of Section 7, distribution can be characterized by the characteristic function. This is justified by the following result.

Theorem 7.1.d. We have $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}^d$ iff $X \stackrel{\mathrm{d}}{=} Y$.

Proof. (Idea only) Quite a lot of analytical arguments (mostly dealing with integrals) are needed in the proof, so we shall only give a rough proof idea here, without covering the technical analytic details.

For "$\Leftarrow$", it just follows by definition. So the only hard part is to prove "$\Rightarrow$". The key idea is to derive an *inversion formula* for $d = 1$, i.e., a formula that expresses the *distribution function* in terms of the *characteristic function*, which is the other way round as the usual expression of characteristic function as an expectation, which can in turn be expressed in terms of distribution function. Here, we skip all the technical details in the derivation, and just directly provide the inversion formula:

$$F(b) - F(a) + \frac{\mathbb{P}(X = a)}{2} - \frac{\mathbb{P}(X = b)}{2} = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \phi_X(t)\, \mathrm{d}t. \qquad (10)$$

This establishes the "$\Rightarrow$" direction for $d = 1$. For the case with $d > 1$, the "$\Rightarrow$" direction follows by noting that

$$\phi_{a^T X}(t) = \mathbb{E}\left[e^{ita^T X}\right] = \mathbb{E}\left[e^{i(ta)^T X}\right] = \phi_X(ta) \stackrel{(X \stackrel{\mathrm{d}}{=} Y)}{=} \phi_Y(ta) = \phi_{a^T Y}(t)$$

for all $t \in \mathbb{R}$ and $a \in \mathbb{R}^d$, which implies by the established $d = 1$ case that $a^T X \stackrel{\mathrm{d}}{=} a^T Y$. Therefore, by Corollary 7.1.c we have $X \stackrel{\mathrm{d}}{=} Y$. $\qquad\square$

Using Theorem 7.1.d, we can derive the following criteria about the realness of characteristic function.

Corollary 7.1.e. Let ϕ_X be a characteristic function. The following are equivalent.

(a) ϕ_X is real (i.e., $\phi_X(t) \in \mathbb{R}$ for all $t \in \mathbb{R}^d$).
(b) $\phi_X(t) = \phi_{-X}(t)$ for all $t \in \mathbb{R}^d$.
(c) $X \stackrel{\mathrm{d}}{=} -X$.

Proof.

- (a) $\iff$ (b): Note that ϕ_X is real iff for all $t \in \mathbb{R}^d$,

$$\phi_X(t) = \underbrace{\overline{\phi_X(t)}}_{\text{complex conjugate}} = \mathbb{E}\left[\cos(t^T X)\right] - i\mathbb{E}\left[\sin(t^T X)\right] \stackrel{(\cos(-x)=x)}{\underset{(\sin(-x)=-\sin(x))}{=}} \mathbb{E}\left[\cos((-t)^T X)\right] + i\mathbb{E}\left[\sin((-t)^T X)\right]$$

$$= \phi_X(-t) = \mathbb{E}\left[e^{i(-t)^T X}\right] = \mathbb{E}\left[e^{it^T(-X)}\right] = \phi_{-X}(t).$$

122

- (b) $\iff$ (c): It follows from Theorem 7.1.d.

$\square$

7.1.6 **Properties of multivariate normal distribution.** Theorem 7.1.d is also helpful for establishing numerous properties of multivariate normal distribution. We start with the following result that provides a characterization of multivariate normal distribution.

Proposition 7.1.f. We have $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ iff $\boldsymbol{a}^T \boldsymbol{X} \sim \mathrm{N}\big(\boldsymbol{a}^T \boldsymbol{\mu}, \boldsymbol{a}^T \Sigma \boldsymbol{a}\big)$ for all $\boldsymbol{a} \in \mathbb{R}^d$.

Proof. "$\Rightarrow$": Here we will use the fact that the characteristic function of $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$ is given by $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = e^{i\boldsymbol{t}^T \boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T \Sigma \boldsymbol{t}}$ for all $\boldsymbol{t} \in \mathbb{R}^d$.

Assume that $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$, and fix any $\boldsymbol{a} \in \mathbb{R}^d$. Then, we have

$$\phi_{\boldsymbol{a}^T \boldsymbol{X}}(t) = \mathbb{E}\Big[e^{i(t\boldsymbol{a})^T \boldsymbol{X}}\Big] = \phi_{\boldsymbol{X}}(t\boldsymbol{a}) = e^{i(t\boldsymbol{a})^T \boldsymbol{\mu} - \frac{1}{2}(t\boldsymbol{a})^T \Sigma (t\boldsymbol{a})} = e^{it(\boldsymbol{a}^T \boldsymbol{\mu}) - \frac{1}{2}t^2 (\boldsymbol{a}^T \Sigma \boldsymbol{a})}$$

for all $t \in \mathbb{R}$. Therefore, by Theorem 7.1.d, we have $\boldsymbol{a}^T \boldsymbol{X} \sim \mathrm{N}\big(\boldsymbol{a}^T \boldsymbol{\mu}, \boldsymbol{a}^T \Sigma \boldsymbol{a}\big)$.

"$\Leftarrow$": Assume $\boldsymbol{a}^T \boldsymbol{X} \sim \mathrm{N}\big(\boldsymbol{a}^T \boldsymbol{\mu}, \boldsymbol{a}^T \Sigma \boldsymbol{a}\big)$ for all $\boldsymbol{a} \in \mathbb{R}^d$. Let $\boldsymbol{Y} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$, and fix any $\boldsymbol{a} \in \mathbb{R}^d$. By the "$\Rightarrow$" direction, we have $\boldsymbol{a}^T \boldsymbol{Y} \sim \mathrm{N}\big(\boldsymbol{a}^T \boldsymbol{\mu}, \boldsymbol{a}^T \Sigma \boldsymbol{a}\big)$, which means that $\boldsymbol{a}^T \boldsymbol{Y} \overset{\mathrm{d}}{=} \boldsymbol{a}^T \boldsymbol{X}$. Hence, by Corollary 7.1.c, we have $\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{Y}$, implying that $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$. $\square$

Proposition 7.1.f leads to many properties of normal distribution, suggested by the following result.

Corollary 7.1.g. Let $\boldsymbol{X} \sim \mathrm{N}_d(\boldsymbol{\mu}, \Sigma)$.

(a) *(Marginal normality of multivariate normal distribution)* We have $X_j \sim \mathrm{N}(\mu_j, \Sigma_{jj})$ for all $j = 1, \ldots, d$.

(b) *(Distribution of sum of normal random variables)* We have $X_1 + \cdots + X_d \sim \mathrm{N}\Big(\sum_{j=1}^d \mu_j, \sum_{i=1}^d \sum_{j=1}^d \Sigma_{ij}\Big)$.
Particularly, if $X_1, \ldots, X_d$ are pairwise uncorrelated, then $X_1 + \cdots + X_d \sim \mathrm{N}\Big(\sum_{j=1}^d \mu_j, \sum_{j=1}^d \Sigma_{jj}\Big)$.

(c) *(Distribution of mean of normal random variables)* We have

$$\frac{X_1 + \cdots + X_d}{d} \sim \mathrm{N}\left(\frac{1}{d}\sum_{j=1}^d \mu_j, \frac{1}{d^2}\sum_{i=1}^d \sum_{j=1}^d \Sigma_{ij}\right).$$

Particularly, if $X_1, \ldots, X_d$ are pairwise uncorrelated, then

$$\frac{X_1 + \cdots + X_d}{d} \sim \mathrm{N}\left(\frac{1}{d}\sum_{j=1}^d \mu_j, \frac{1}{d^2}\sum_{j=1}^d \Sigma_{jj}\right).$$

Proof.

(a) Use the "$\Rightarrow$" direction of Proposition 7.1.f with $\boldsymbol{a} = \boldsymbol{e}_j$ for every $j = 1, \ldots, d$.

(b) Use the "$\Rightarrow$" direction of Proposition 7.1.f with $\boldsymbol{a} = (1, \ldots, 1)$.

(c) Use the "$\Rightarrow$" direction of Proposition 7.1.f with $\boldsymbol{a} = (1/d, \ldots, 1/d)$.

$\square$

7.2 Central Limit Theorem

7.2.1 Preliminary analytical results. To prove the central limit theorem, two more preliminary analytical results are needed, as follows.

Lemma 7.2.a. If $\lim_{n \to \infty} a_n = a$, then $\lim_{n \to \infty} (1 + a_n/n)^n = e^a$.

[⚠ Warning: In general, $\{f_n\} \xrightarrow[n \to \infty]{} f$ and $\{a_n\} \xrightarrow[n \to \infty]{} a$ do *not* imply that $\{f_n(a_n)\} \xrightarrow[n \to \infty]{} f(a)$. For example, take $f_n(x) = x^n$ and $a_n = 1 - 1/n$ for each $n \in \mathbb{N}$. Then, we have $\{f_n\} \xrightarrow[n \to \infty]{} f$ with $f(x) = \mathbf{1}_{\{x=1\}}$ and $\{a_n\} \xrightarrow[n \to \infty]{} a = 1$, while $\{f_n(a_n)\} = \{(1 - 1/n)^n\} \xrightarrow[n \to \infty]{} e^{-1} \neq 1 = f(a)$.]

Proof. By Taylor series expansion, we have $\ln(1+x) = -\sum_{k=1}^{\infty} (-x)^k/k$ for all $-1 < x < 1$. Hence, for all $-1/2 < x < 1/2$ we have

$$|\ln(1+x) - x| = \left| \sum_{k=2}^{\infty} \frac{(-x)^k}{k} \right| \overset{\text{(triangle)}}{\leq} x^2 \sum_{k=2}^{\infty} \frac{|x|^{k-2}}{k} \overset{(j=k-2)}{=} x^2 \sum_{j=0}^{\infty} \frac{|x|^j}{j+2}$$

$$\leq \frac{x^2}{2} \sum_{j=0}^{\infty} |x|^j \overset{(|x| \leq 1/2)}{\leq} \frac{x^2}{2} \sum_{k=0}^{\infty} (1/2)^k = x^2.$$

Next, since $\lim_{n \to \infty} a_n = a$, we have $\lim_{n \to \infty} a_n/n = (\lim_{n \to \infty} a_n)(\lim_{n \to \infty} 1/n) = a \cdot 0 = 0$, by setting $\varepsilon = 1/2$, we know there exists $n_0 \in \mathbb{N}$ such that $|a_n/n| < 1/2$ for all $n \geq n_0$. Also, since $\{a_n\}$ is convergent, it is bounded and we can write $|a_n| \leq M$ for all $n \in \mathbb{N}$, for some $M > 0$.

Therefore, for all $n \geq n_0$, we have

$$\left| n \ln\left(1 + \frac{a_n}{n}\right) - a_n \right| = n \left| \ln\left(1 + \frac{a_n}{n}\right) - \frac{a_n}{n} \right| \overset{(|a_n/n| < 1/2)}{\leq} n \left(\frac{a_n}{n}\right)^2 \leq \frac{M^2}{n}.$$

This implies that

$$0 \leq \lim_{n \to \infty} \left| n \ln\left(1 + \frac{a_n}{n}\right) - a \right| \overset{\text{(triangle)}}{\leq} \lim_{n \to \infty} \left| n \ln\left(1 + \frac{a_n}{n}\right) - a_n \right| + \lim_{n \to \infty} |a_n - a| \leq \lim_{n \to \infty} \frac{M^2}{n} + 0 = 0,$$

and so $\lim_{n \to \infty} \ln(1 + a_n/n)^n = \lim_{n \to \infty} n \ln(1 + a_n/n) = a$. Since the exponential function is continuous, we get $\lim_{n \to \infty} (1 + a_n/n)^n = e^a$. $\qquad \square$

Lemma 7.2.b. If $\mathbb{E}[|X|^m] < \infty$ for some $m \in \mathbb{N}$, then $\phi_X(t) = \sum_{k=0}^{m} \mathbb{E}[X^k](it)^k/k! + o(|t|^m)$, where we have $h = o(g)$ if $\lim_{t \to 0} |h(t)|/g(t) = 0$.

Proof. Omitted. $\qquad \square$

7.2.2 Central limit theorem. We can now prove the central limit theorem (CLT) as follows.

Theorem 7.2.c (Central limit theorem). If $\{X_n\}$ is a sequence of iid random variables with $\mu = \mathbb{E}[X_1] \in \mathbb{R}$ and $0 < \sigma^2 = \text{Var}(X_1) < \infty$, then

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n \to \infty]{d} N(0, 1),$$

where $\bar{X}_n = (\sum_{i=1}^{n} X_i)/n$.

Proof. Let $Z_k = (X_k - \mu)/\sigma$ for every $k \in \mathbb{N}$. Then, it suffices to show that $\sqrt{n}\bar{Z}_n \xrightarrow[n \to \infty]{d} N(0, 1)$, where $\bar{Z}_n = (Z_1 + \cdots + Z_n)/n$. Fix any $t \in \mathbb{R}$, and consider

$$\phi_{\sqrt{n}\bar{Z}_n}(t) = \mathbb{E}\left[e^{i(t/\sqrt{n}) \sum_{k=1}^{n} Z_k}\right] \overset{\text{(Proposition 5.2.d)}}{=} \prod_{k=1}^{n} \mathbb{E}\left[e^{i(t/\sqrt{n})Z_k}\right] \overset{\text{(identically distributed)}}{=} \phi_{Z_1}(t/\sqrt{n})^n$$

$$\overset{\text{(Lemma 7.2.b, } m=2)}{=} \left(1 + i\frac{t}{\sqrt{n}}\mathbb{E}[Z_1] - \frac{t^2/n}{2}\mathbb{E}[Z_1]^2 + o(t^2/n)\right)^n \overset{(\mathbb{E}[Z_1]=0)}{\underset{(\mathbb{E}[Z_1^2]=1)}{=}} \left(1 - \frac{t^2/2 - o(t^2/n)/(1/n)}{n}\right)^n$$

$$= \left(1 - \frac{t^2/2 - t^2 o(1/n)/(1/n)}{n}\right)^n.$$

Since $\lim_{n\to\infty} t^2/2 - t^2 o(1/n)/(1/n) = t^2/2 - t^2(0) = t^2/2$, by Lemma 7.2.a we have $\lim_{n\to\infty} \phi_{\sqrt{n}\bar{Z}_n}(t) = e^{-t^2/2}$. Therefore, Theorem 7.1.a gives $\sqrt{n}\bar{Z}_n \xrightarrow[n\to\infty]{\text{d}} Z$ where the characteristic function of Z is $\phi_Z(t) = e^{-t^2/2}$, which is the one for $\mathrm{N}(0,1)$. Hence, by Theorem 7.1.d we know $Z \sim \mathrm{N}(0,1)$. $\qquad\square$

Remarks:

- *(Sample sum form)* The form of CLT is expressed in terms of *sample mean*. By multiplying both the numerator and denominator by n, we can get the *sample sum* form: $(\sum_{i=1}^n X_i - n\mu)/(\sqrt{n}\sigma) \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}(0,1)$.

- *(Role of the factor $\sqrt{n}$)* Intuitively, the factor $\sqrt{n}$ multiplied to the expression serves for "amplifying" the difference between sample mean and the true mean, adjusted by σ: $(\bar{X}_n - \mu)/\sigma$, so that its limiting distribution does not just "vanish"; note that we have $(\bar{X}_n - \mu)/\sigma \xrightarrow[n\to\infty]{\text{a.s.}} 0$ by the SLLN.

- *(Practical usage of CLT)* In practice, with a large n, we know that $\bar{X}_n \overset{\text{approx.}}{\sim} \mathrm{N}(\mu, \sigma^2/n)$ and $\sum_{i=1}^n X_i \overset{\text{approx.}}{\sim} \mathrm{N}(n\mu, n\sigma^2)$, as long as $X_1, \ldots, X_n$ are iid with finite mean and variance (not necessarily normally distributed themselves). However, there is not a clear mathematical meaning of "$\overset{\text{approx.}}{\sim}$" and formally we still need to interpret these based on the convergence in distribution.

  Here, we also note that "$\overset{\text{approx.}}{\sim}$" would become "$\sim$" (exactly) if $X_1, \ldots, X_n$ are also normally distributed, by Corollary 7.1.g.

7.2.3 Extensions to CLT. The result in Theorem 7.2.c is sometimes known as the *classical* CLT since there are indeed numerous extensions to the result, including the following:

(a) *(Multivariate CLT)* If $\{\boldsymbol{X}_n\}$ is a sequence of iid random vectors with mean vector $\boldsymbol{\mu} = (\mathbb{E}[X_{11}], \ldots, \mathbb{E}[X_{1d}]) \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, whose (i,j)th entry is $\Sigma_{ij} = \mathrm{Cov}(X_{1i}, X_{1j})$, then

$$\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}_d(\boldsymbol{0}, \Sigma).$$

(b) *(Lindeberg-Feller CLT)* Let $\{X_n\}$ be a sequence of independent random variables, where the mean of variance of X_n are $\mu_n = \mathbb{E}[X_n] \in \mathbb{R}$ and $\sigma_n^2 = \mathrm{Var}(X_n) < \infty$ respectively for each $n \in \mathbb{N}$. Let $s_n^2 = \mathrm{Var}(\sum_{k=1}^n X_k) = \sum_{k=1}^n \sigma_k^2$. If the **Lindeberg condition**

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^n \mathbb{E}\left[(X_k - \mu_k)^2 \mathbf{1}_{\left\{\frac{|X_k-\mu_k|}{s_n} > \varepsilon\right\}}\right] = 0 \quad \text{for all } \varepsilon > 0$$

holds, then

$$\frac{\sum_{k=1}^n (X_k - \mu_k)}{s_n} \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}(0,1).$$

Furthermore, the Lindeberg condition holds if the **Lyapunov condition**

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{k=1}^n \mathbb{E}\left[|X_k - \mu_k|^{2+\delta}\right] = 0 \quad \text{for some } \delta > 0$$

holds.

Proof.

(a) Fix any $\boldsymbol{a} \in \mathbb{R}^d$. Consider first the case where $\boldsymbol{a}^T \Sigma \boldsymbol{a} = 0$, which means that $\mathrm{Var}(\boldsymbol{a}^T \boldsymbol{X}_n) = \boldsymbol{a}^T \Sigma \boldsymbol{a} = 0$ for each $n \in \mathbb{N}$. Then, we know that $\boldsymbol{a}^T \boldsymbol{X}_n \overset{\text{a.s.}}{=} \boldsymbol{a}^T \boldsymbol{\mu}$ $\forall n \in \mathbb{N}$ by [5.3.3]. Therefore, we have $\boldsymbol{a}^T(\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu})) \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}(0, \boldsymbol{a}^T \Sigma \boldsymbol{a})$, where $\mathrm{N}(0,0)$ refers to the distribution of a random variable taking the value 0 always.

Next, consider the case where $\boldsymbol{a}^T\Sigma\boldsymbol{a} > 0$. By Proposition 7.1.f we have $\boldsymbol{a}^T\boldsymbol{X}_n \sim N(\boldsymbol{a}^T\boldsymbol{\mu}, \boldsymbol{a}^T\Sigma\boldsymbol{a})$ for all $n \in \mathbb{N}$. Since $\{\boldsymbol{a}^T\boldsymbol{X}_n\}$ forms a sequence of iid random variables with finite mean $\boldsymbol{a}^T\boldsymbol{\mu}$ and variance $\boldsymbol{a}^T\Sigma\boldsymbol{a}$, by the classical CLT (Theorem 7.2.c) we have

$$\sqrt{n}\frac{(\sum_{k=1}^n \boldsymbol{a}^T\boldsymbol{X}_k)/n - \boldsymbol{a}^T\boldsymbol{\mu}}{\sqrt{\boldsymbol{a}^T\Sigma\boldsymbol{a}}} = \sqrt{n}\frac{\boldsymbol{a}^T(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu})}{\sqrt{\boldsymbol{a}^T\Sigma\boldsymbol{a}}} \xrightarrow[n\to\infty]{\mathrm{d}} N(0,1).$$

This implies that $\boldsymbol{a}^T(\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu})) \xrightarrow[n\to\infty]{\mathrm{d}} N(0, \boldsymbol{a}^T\Sigma\boldsymbol{a})$, by the CMT. Now, let $\boldsymbol{X} \sim N_d(\boldsymbol{0}, \Sigma)$. Noting that $\boldsymbol{a}^T\boldsymbol{X} \sim N(0, \boldsymbol{a}^T\Sigma\boldsymbol{a})$, by Theorem 7.1.b we conclude that $\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow[n\to\infty]{\mathrm{d}} N_d(\boldsymbol{0}, \Sigma)$.

(b) Omitted.

$\square$

8   Conditional Expectation

8.0.1   We have finally reached the last section of this notes, which is about the concept of *conditional expectation*. This notion plays a fundamental role in the study of probability theory, which formalizes the idea of computing probabilistic quantities with some *additional information* given *(conditioning)*. While you should have already learnt basic ideas about conditional expectation before, here we will analyze it in full mathematical details and introduce a *measure-theoretic* way to define conditional expectation. Such definition is quite general and frees us from many restrictions, but is also rather abstract unfortunately 😵.

8.1   Ordinary Conditioning

8.1.1   Before providing the measure-theoretic definition of conditional expectation, we first have a glance on how "ordinary" conditioning (i.e., the one you have seen before) works, and its potential restrictions, as a motivation of the measure-theoretic study of conditional expectation.

8.1.2   **Ordinary conditional probability and implied conditional expectation.** Recall from Section 4.1 that, on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, for an event $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, the *(ordinary) conditional probability* of A given B is defined by $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ for every $A \in \mathcal{F}$. Based on this idea, we can then define *conditional distribution function* (with some restrictions) as follows. Consider a random vector $\boldsymbol{X}_2 : \Omega \to \mathbb{R}^{d_2}$. For every event $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, we define the **(ordinary) conditional distribution function of $\boldsymbol{X}_2$ given B** by

$$F_{\boldsymbol{X}_2|B}(\boldsymbol{x}_2) := \mathbb{P}(\boldsymbol{X}_2 \leq \boldsymbol{x}_2|B) = \frac{\mathbb{P}(\{\boldsymbol{X}_2 \leq \boldsymbol{x}_2\} \cap B)}{\mathbb{P}(B)}$$

for all $\boldsymbol{x}_2 \in \mathbb{R}^{d_2}$. The mean of $F_{\boldsymbol{X}_2|B}$ (i.e., mean of a random variable with distribution function being $F_{\boldsymbol{X}_2|B}$), if exists, is called the **(ordinary) conditional expectation of $\boldsymbol{X}_2$ given B**. Often, we are handling the case with $B = \{\boldsymbol{X}_1 = \boldsymbol{x}_1\}$ for some random variable $\boldsymbol{X}_1$. However, the definition above only works if $\mathbb{P}(B) = \mathbb{P}(\boldsymbol{X}_1 = \boldsymbol{x}_1) > 0$ (e.g., discrete case), and breaks down if $\mathbb{P}(\boldsymbol{X}_1 = \boldsymbol{x}_1) = 0$ (e.g., continuous case).

8.1.3   **How to deal with conditioning on zero-probability event?** If $F_{\boldsymbol{X}_1}$ is discrete, we know that $B = \{\boldsymbol{X}_1 = \boldsymbol{x}_1\}$ has positive probability for every $\boldsymbol{x}_1 \in \mathrm{supp}(F_{\boldsymbol{X}_1})$, and zero probability for every $\boldsymbol{x}_1$ outside the support. In this case, we may handle the potential conditioning on zero probability through, e.g., setting the conditional distribution function as zero for every $\boldsymbol{x}_1$ outside the support:

$$F_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_1) = \begin{cases} \mathbb{P}(\boldsymbol{X}_2 \leq \boldsymbol{x}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1) & \text{if } \boldsymbol{x}_1 \in \mathrm{supp}(F_{\boldsymbol{X}_1}), \\ 0 & \text{otherwise.} \end{cases}$$

This is still meaningful as we have plenty of $\boldsymbol{x}_1 \in \mathrm{supp}(F_{\boldsymbol{X}_1})$ to work with, and we can just "neglect" the behaviour of $F_{\boldsymbol{X}_2|\boldsymbol{X}_1}$ at $\boldsymbol{x}_1$ outside the support.

However, in case $F_{\boldsymbol{X}_1}$ is *continuous*, the event $B = \{\boldsymbol{X}_1 = \boldsymbol{x}_1\}$ would have zero probability for *every* $\boldsymbol{x}_1$, making the method suggested above obsolete, as we would need to artificially define the value taken by $F_{\boldsymbol{X}_2|\boldsymbol{X}_1}$ at *every* point, making this concept not meaningful anymore.

This leads us to the idea of considering the following limiting argument. Suppose we have $(\boldsymbol{X}_1, \boldsymbol{X}_2) \sim F$ with a joint density f. Recall from your first probability course that the *conditional density* is defined by

$$f_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2|\boldsymbol{x}_1) = \frac{f(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f_{\boldsymbol{X}_1}(\boldsymbol{x}_1)}$$

if $f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) > 0$. It is then natural to define

$$F_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2|\boldsymbol{x}_1) := \mathbb{P}(\boldsymbol{X}_2 \leq \boldsymbol{x}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1) = \int_{-\infty}^{\boldsymbol{x}_2} f_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\widetilde{\boldsymbol{x}}_2|\boldsymbol{x}_1) \, \mathrm{d}\widetilde{\boldsymbol{x}}_2$$

provided that $f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) > 0$. Nevertheless, such "definition" may actually lead to potential ill-definedness, as the *Borel-Kolmogorov paradox* illustrates.

8.1.4 **Borel-Kolmogorov paradox.** Let $(X_1, X_2) \sim U(D)$ with $D = \{(x_1, x_2) : -1 \leq x_1 \leq 1, 0 \leq x_2 \leq \sqrt{1 - x_1^2}\}$ being the upper half of unit disk centered at $(0, 0)$. Here, $U(D)$ refers to the uniform distribution on D (i.e., constant density over D). We then consider the conditional distribution using two coordinates.

(a) *(Cartesian coordinates)* Using the Cartesian coordinates, the joint density is

$$f(x_1, x_2) = \frac{1}{\pi/2} \mathbf{1}_{\{(x_1, x_2) \in D\}} = \frac{2}{\pi} \mathbf{1}_{\{(x_1, x_2) \in D\}},$$

and we have

$$f_{X_1}(x_1) = \int_0^{\sqrt{1-x_1^2}} f(x_1, x_2) \, dx_2 = \frac{2}{\pi} \mathbf{1}_{\{-1 \leq x_1 \leq 1\}} \int_0^{\sqrt{1-x_1^2}} dx_2 = \frac{2\sqrt{1-x_1^2}}{\pi} \mathbf{1}_{\{-1 \leq x_1 \leq 1\}}.$$

Hence, for every $x_1 \in (-1, 1)$, we get

$$f_{X_2|X_1}(x_2|x_1) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)} = \frac{1}{\sqrt{1-x_1^2}} \mathbf{1}_{\{(x_1, x_2) \in D\}},$$

Particularly, we have

$$\mathbb{P}(X_2 \in [0, 1/2]|X_1 = 0) = \int_0^{1/2} 1 \, dx_2 = \frac{1}{2}.$$

(b) *(Polar coordinates)* We introduce the polar coordinates with $X_1 = R\cos(\Theta)$ and $X_2 = R\sin(\Theta)$. In polar coordinates, D can be expressed as $\{(r, \theta) : 0 \leq r \leq 1, 0 \leq \theta \leq \pi\} = [0, 1] \times [0, \pi]$. Applying the change-of-variables formula for density, we get

$$f_{R,\Theta}(r, \theta) = f(x_1, x_2) \begin{vmatrix} \partial r\cos\theta/\partial r & \partial r\cos\theta/\partial\theta \\ \partial r\sin\theta/\partial r & \partial r\sin\theta/\partial\theta \end{vmatrix} = f(x_1, x_2) \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix}$$

$$= \frac{2r}{\pi} \mathbf{1}_{\{(x_1, x_2) \in D\}} = \frac{2r}{\pi} \mathbf{1}_{\{(r, \theta) \in [0,1] \times [0, \pi]\}}.$$

Therefore, we have

$$f_\Theta(\theta) = \mathbf{1}_{\{\theta \in [0, \pi]\}} \int_0^1 \frac{2r}{\pi} \, dr = \frac{1}{\pi} \mathbf{1}_{\{\theta \in [0, \pi]\}}.$$

Thus, for every $\theta \in [0, \pi]$ we get

$$f_{R|\Theta}(r|\theta) = \frac{2r/\pi}{1/\pi} \mathbf{1}_{\{0 \leq r \leq 1\}} = 2r \mathbf{1}_{\{0 \leq r \leq 1\}}.$$

Particularly, we have

$$\mathbb{P}(R \in [0, 1/2]|\Theta = \pi/2) = \int_0^{1/2} 2r \, dr = \frac{1}{4}.$$

In this case, while both $\mathbb{P}(X_2 \in [0, 1/2]|X_1 = 0)$ and $\mathbb{P}(R \in [0, 1/2]|\Theta = \pi/2)$ should just be describing the same probability expressed in different coordinates system, their values turn out to differ.
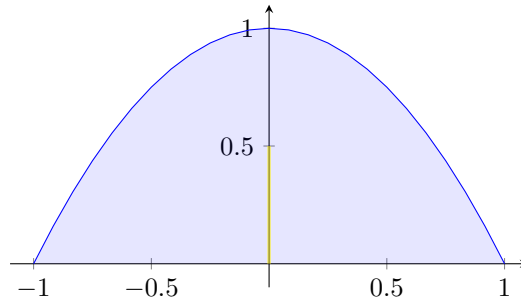
Intuitively, this phenomenon appears since the two coordinates system lead to different ways to "approach" this probability as a limit. This is best understood through the following pictures:
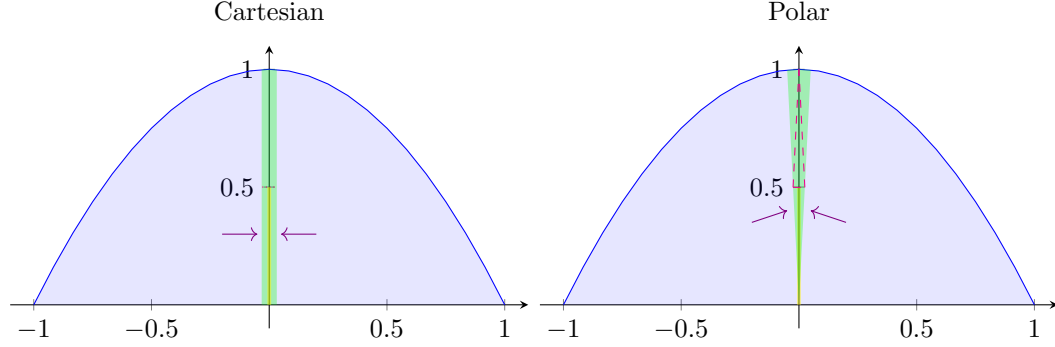
These pictures also intuitively illustrate why the probability is $1/2$ for Cartesian coordinates and is $1/4$ for polar coordinates. For Cartesian coordinates, the "region of interest" occupies a half of the "limiting" region (approached with "rectangles" having shrinking widths). On the other hand, for polar coordinates, it only occupies a quarter of the "limiting region" (approached with "circular sectors" having shrinking central angles). The dependency of the probability calculation on the way of how the limit is approached is an undesirable trait, and can lead to ambiguities. This issue can be avoided by following the *measure-theoretic* approach instead.

8.2 Conditioning in a Measure-Theoretic Framework

8.2.1 In view of the potential issues from the "ordinary" way of conditioning suggested above, we are then motivated to study the conditioning in a *measure-theoretic* framework. Let us start with the measure-theoretic definition of conditional expectation.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra on Ω, and $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. The **conditional expectation of X given $\mathcal{G}$**, denoted by $\mathbb{E}[X|\mathcal{G}]$, is any function $Y : \Omega \to \mathbb{R}$ satisfying:

(1) *(Measurability)* Y is $\mathcal{G}$-measurable.

(2) *(Partial averaging)* $\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A]$ for every $A \in \mathcal{G}$.
[Note: Due to the presence of the indicator function $\mathbf{1}_A$, the "averaging" (mean) is only taken over "part" of Ω, hence the name "partial averaging".]

Any such Y is said to be a **version** of $\mathbb{E}[X|\mathcal{G}]$. We also have the following shorthand notations about conditional expectation:

- $\mathbb{E}[X|\mathcal{A}] := \mathbb{E}[X|\sigma(\mathcal{A})]$ with $\mathcal{A} \subseteq \mathcal{F}$ being a collection of events.
- $\mathbb{E}[X|Z] := \mathbb{E}[X|\sigma(Z)]$ with $Z : \Omega \to \Omega'$ being a function (e.g., random variable, random vector, etc.).
- $\mathrm{Var}(X|\mathcal{G}) = \mathbb{E}\big[(X - \mathbb{E}[X|\mathcal{G}])^2|\mathcal{G}\big]$ is the **conditional variance of X given $\mathcal{G}$**.

With the measurability and partial averaging properties satisfied, the function Y carries the characteristics of a random variable and an expectation "with respect to $\mathcal{G}$" (i.e., we only require $\mathcal{G}$-measurability for random variable, and consider only set in $\mathcal{G}$ for the partial averaging).

8.2.2 **Integrability, existence, and uniqueness.** With this nonconstructive definition of conditional expectation, the first thing we should do is to examine whether the object defined "makes sense", by investigating some core properties about it, namely *integrability*, *existence*, and *uniqueness*.

Proposition 8.2.a (Integrability). Every version Y of $\mathbb{E}[X|\mathcal{G}]$ is integrable, i.e., in $L^1(\Omega, \mathcal{G}, \mathbb{P})$.

Proof. Let $A = \{Y \geq 0\} = Y^{-1}([0, \infty)) \in \mathcal{G}$. Then, we have $A^c \in \mathcal{G}$, and hence

$$\mathbb{E}[|Y|] = \mathbb{E}[|Y|\mathbf{1}_A] + \mathbb{E}[|Y|\mathbf{1}_{A^c}] = \mathbb{E}[Y\mathbf{1}_A] + \mathbb{E}[(-Y)\mathbf{1}_{A^c}] = \int_A Y \, d\mathbb{P} + \int_{A^c} (-Y) \, d\mathbb{P}$$

$$\overset{\text{(partial averaging)}}{=} \int_A X \, d\mathbb{P} + \int_{A^c} (-X) \, d\mathbb{P} \overset{\text{(monotonicity)}}{\leq} \int_A |X| \, d\mathbb{P} + \int_{A^c} |X| \, d\mathbb{P} = \int_\Omega |X| \, d\mathbb{P} = \mathbb{E}[|X|] \overset{(X \in L^1)}{<} \infty.$$

$\square$

Theorem 8.2.b (Existence and a.s. uniqueness). The conditional expectation $\mathbb{E}[X|\mathcal{G}]$ exists and is unique a.s.

Proof. We first prove the existence. Consider first the case where $X \geq 0$. By [5.1.8], the function ν given by $\nu(A) = \mathbb{E}[X\mathbf{1}_A] = \int_A X \, d\mathbb{P}$ for all $A \in \mathcal{G}$ is a measure on $(\Omega, \mathcal{G})$. Also, by [5.1.10] we have $\nu(A) = 0$ for every $A \in \mathcal{G}$ with $\mathbb{P}(A) = 0$, i.e., $\nu \ll \mathbb{P}$, with $\mathbb{P}$ being the original probability measure restricted on $\mathcal{G}$ (we keep this notation for convenience). Applying Theorem 5.5.c then gives, for all $A \in \mathcal{G}$,

$$\int_A X \, d\mathbb{P} = \nu(A) = \int_A \frac{d\nu}{d\mathbb{P}} \, d\mathbb{P}$$

where $d\nu/d\mathbb{P}$ is the $\mathbb{P}$-a.s. unique and $\mathcal{G}$-measurable Radon-Nikodym derivative. By definition of conditional expectation, $d\nu/d\mathbb{P}$ is a version of $\mathbb{E}[X|\mathcal{G}]$, establishing the existence in this case.

Next, consider the general case where X may not be nonnegative. We write $X = X^+ - X^-$, and let $Y^+ = \mathbb{E}[X^+|\mathcal{G}]$ and $Y^- = \mathbb{E}[X^-|\mathcal{G}]$ (which exist by the case above). By Proposition 8.2.a, both Y^+ and Y^- are integrable, which implies that $Y := Y^+ - Y^-$ is also integrable (and hence $\mathcal{G}$-measurable). Furthermore, we have

$$\int_A X \, d\mathbb{P} = \int_A X^+ \, d\mathbb{P} - \int_A X^- \, d\mathbb{P} \overset{\text{(partial averaging)}}{=} \int_A Y^+ \, d\mathbb{P} - \int_A Y^- \, d\mathbb{P} = \int_A Y \, d\mathbb{P}$$

for all $A \in \mathcal{G}$. Therefore, Y is a version of $\mathbb{E}[X|\mathcal{G}]$. This establishes the existence.

Now, we prove the a.s. uniqueness. Suppose both Y and $\widetilde{Y}$ are versions of $\mathbb{E}[X|\mathcal{G}]$. Then, by the partial averaging property we have $\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[\widetilde{Y}\mathbf{1}_A]$ for every $A \in \mathcal{G}$. Furthermore, by Proposition 8.2.a, Y and $\widetilde{Y}$ are integrable. Hence, by [5.1.11]c we have $Y \overset{\text{a.s.}}{=} \widetilde{Y}$, establishing the a.s. uniqueness. $\square$

As Theorem 8.2.b suggests, the conditional expectation is only unique *almost surely*. Consequently, equalities/inequalities involving conditional expectation like $Y = \mathbb{E}[X|\mathcal{G}]$ should all be understood to hold *almost surely* only. Nonetheless, to avoid making the notations cumbersome, we often just write "$=$" to mean "$\overset{\text{a.s.}}{=}$"; this should be clear from context. When considering conditional expectation in the measure-theoretic framework here, we are actually working with a specific version of $\mathbb{E}[X|\mathcal{G}]$ implicitly. But since it equals every other version a.s., we usually do not care much about the choice of version.

8.2.3 **Regular conditional probability measures.** Having the conditional expectation $\mathbb{E}[X|\mathcal{G}]$ defined, it is then tempting to define the *conditional probability measure* $\mathbb{P}(\cdot|\mathcal{G})$ by $\mathbb{P}(A|\mathcal{G}) = \mathbb{E}[\mathbf{1}_A|\mathcal{G}]$ for every $A \in \mathcal{F}$. However, since $\mathbb{E}[\mathbf{1}_A|\mathcal{G}]$ is only unique a.s. in general, such function may not be well-defined.

In view of this, the first thing to be done is to specify a certain version of $\mathbb{E}[\mathbf{1}_A|\mathcal{G}]$ for each $A \in \mathcal{F}$ in the definition, so that the function is well-defined. Afterwards, we need to ensure that $\mathbb{P}(\cdot|\mathcal{G})$ is a valid probability measure, by requiring the conditions there to be satisfied. This discussion leads to the definition of *regular conditional probability measure*.

Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be measurable spaces, and $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. The function $\mathbb{P}^* : \mathcal{F} \times \Omega \to [0, \infty]$ is called a **regular conditional probability measure given $\mathcal{G}$** if

(1) *(Specifying versions)* For every fixed $A \in \mathcal{F}$, the function $\omega \mapsto \mathbb{P}^*(A, \omega)$ is a specific version of $\mathbb{P}(A|\mathcal{G}) := \mathbb{E}[\mathbf{1}_A|\mathcal{G}]$.

(2) *(Qualifying as probability measure)* For every fixed $\omega \in \Omega$, the function $A \mapsto \mathbb{P}^*(A, \omega)$ is a probability measure on $(\Omega, \mathcal{F})$.

In a similar way, we call the function $F^*_{\boldsymbol{X}|\mathcal{G}} : \mathbb{R}^d \times \Omega \to [0,1]$ a **regular conditional distribution function of $\boldsymbol{X}$ given $\mathcal{G}$** if

(1) *(Specifying versions)* For every fixed $\boldsymbol{x} \in \mathbb{R}^d$, the function $\omega \mapsto F^*_{\boldsymbol{X}|\mathcal{G}}(\boldsymbol{x}, \omega)$ is a specific version of $\mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}|\mathcal{G})$.

(2) *(Qualifying as distribution function)* For every fixed $\omega \in \Omega$, the function $\boldsymbol{x} \mapsto F_{\boldsymbol{X}|\mathcal{G}}(\boldsymbol{x}, \omega)$ is a distribution function.

The definitions of regular conditional probability measure and distribution function here are again *nonconstructive*, so it is not immediately clear whether they actually exist (just like the case for conditional expectation). It turns out that as long as there is a bijective function $\varphi : \Omega' \to \mathbb{R}^d$ such that both φ and φ^{-1} are measurable *(($\Omega', \mathcal{F}'$) is nice)*, then both of them exist. In many cases of practical interest, such "niceness" is satisfied, e.g., $(\Omega', \mathcal{F}') = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is nice (by taking φ to be the identity function). So, henceforth we will assume the niceness and always work with the regular ones (implicitly) when considering the notations like $\mathbb{P}(A|\mathcal{G})$ and $\mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}|\mathcal{G})$.

8.2.4 **Formula of conditional expectation for $\mathcal{G}$ generated by a countable partition.** Generally, we do not have an explicit formula for finding out the conditional expectation. However, in the special case where $\mathcal{G}$ is generated by a countable partition, we do have an explicit formula for conditional expectation (up to a.s. equality), which involves expressions that should have appeared in your first probability course.

Proposition 8.2.c. Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{A} = \{A_n : n \in \mathbb{N}\} \subseteq \mathcal{F}$ be a partition of Ω, and $\mathcal{G} = \sigma(\mathcal{A}) \overset{\text{(Lemma 1.4.a)}}{=} \{\biguplus_{i \in I} A_i : A_i \in \mathcal{A} \; \forall i \in I \text{ and } I \subseteq \mathbb{N}\} \subseteq \mathcal{F}$. Then, we have

$$\mathbb{E}[X|\mathcal{G}] = \sum_{n=1}^{\infty} \mathbb{E}_{A_n}[X]\mathbf{1}_{A_n}, \quad \text{where } \mathbb{E}_{A_n}[X] := \begin{cases} \mathbb{E}[X\mathbf{1}_{A_n}]/\mathbb{P}(A_n) & \text{if } \mathbb{P}(A_n) > 0, \\ 0 & \text{if } \mathbb{P}(A_n) = 0. \end{cases}$$

Remarks:

- The value assigned for $\mathbb{E}_{A_n}[X]$ in the case $\mathbb{P}(A_n) = 0$ can indeed be replaced by other real numbers (not necessarily 0); this also explains why the equality only holds a.s.

- *(More explicit form of the formula)* For convenience, suppose that there are only k nonempty sets $A_1, \ldots, A_k$ in the partition, each with positive probability. Then, we can write

$$\mathbb{E}[X|\mathcal{G}](\omega) = \begin{cases} \mathbb{E}[X\mathbf{1}_{A_1}]/\mathbb{P}(A_1) & \text{if } \omega \in A_1, \\ \mathbb{E}[X\mathbf{1}_{A_2}]/\mathbb{P}(A_2) & \text{if } \omega \in A_2, \\ \vdots \\ \mathbb{E}[X\mathbf{1}_{A_k}]/\mathbb{P}(A_k) & \text{if } \omega \in A_k. \end{cases}$$

From this expression, it is transparent that the conditional expectation in this case is a "piecewise constant" function in ω (the output stays the same within a certain A_j).

Proof. Let $Y := \sum_{n=1}^{\infty} \mathbb{E}_{A_n}[X]\mathbf{1}_{A_n} : \Omega \to \mathbb{R}$. We now verify that Y satisfies both the measurability and partial averaging properties.

(1) Let $Y_m := \sum_{n=1}^{m} \mathbb{E}_{A_n}[X]\mathbf{1}_{A_n} = \sum_{n=1}^{m} \mathbb{E}_{A_n}[X]\mathbf{1}_{A_n} + 0 \cdot \mathbf{1}_{\biguplus_{k=m+1}^{\infty} A_k}$ for every $m \in \mathbb{N}$. Since each Y_m is simple with all the indicator sets being in $\mathcal{G}$, we know $Y_m \in \mathcal{G}$ for all $m \in \mathbb{N}$. Hence, by [3.1.5]c we have $Y = \lim_{m \to \infty} Y_m \in \mathcal{G}$.

(2) **Showing that $Y\mathbf{1}_A \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ for all $A \in \mathcal{G}$.** Note that $|Y| \overset{\text{(triangle)}}{\leq} \sum_{n=1}^{\infty} |\mathbb{E}_{A_n}[X]|\mathbf{1}_{A_n} \overset{\text{(triangle)}}{\leq}$ $\sum_{n=1}^{\infty} \mathbb{E}_{A_n}[|X|]\mathbf{1}_{A_n}$. Hence,

$$\mathbb{E}[|Y|] \overset{\text{(monotonicity)}}{\leq} \mathbb{E}\left[\sum_{n=1}^{\infty} \mathbb{E}_{A_n}[|X|]\mathbf{1}_{A_n}\right] \overset{\text{(MCT)}}{=} \lim_{N\to\infty} \mathbb{E}\left[\sum_{n=1}^{N} \mathbb{E}_{A_n}[|X|]\mathbf{1}_{A_n}\right]$$

$$\overset{\text{(linearity)}}{=} \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{E}_{A_n}[|X|]\mathbb{P}(A_n) = \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{E}[|X|\mathbf{1}_{A_n}]$$

$$\overset{\text{(linearity)}}{=} \lim_{N\to\infty} \mathbb{E}\left[\sum_{n=1}^{N} |X|\mathbf{1}_{A_n}\right] = \lim_{N\to\infty} \mathbb{E}\left[|X|\mathbf{1}_{\bigcup_{n=1}^{N} A_n}\right] \overset{\text{(MCT)}}{=} \mathbb{E}[|X|] < \infty.$$

Therefore, for every $A \in \mathcal{G}$, we have $\mathbb{E}[|Y\mathbf{1}_A|] \overset{\text{(monotonicity)}}{\leq} \mathbb{E}[|Y|] < \infty$, and hence $Y\mathbf{1}_A \in L^1(\Omega, \mathcal{F}, \mathbb{P})$.

Showing the partial averaging property. Fix any $A \in \mathcal{G}$. We can then write $A = \biguplus_{i \in I} A_i$ for some countable $I \subseteq \mathbb{N}$. Therefore,

$$\mathbb{E}[Y\mathbf{1}_A] \overset{(Y\mathbf{1}_A \in L^1, \; [5.1.11]\text{a})}{=} \sum_{i \in I} \mathbb{E}[Y\mathbf{1}_{A_i}] = \sum_{i \in I} \mathbb{E}[\mathbb{E}_{A_i}[X]\mathbf{1}_{A_i}]$$

$$\overset{(\mathbb{E}_{A_i}[X] \text{ deterministic})}{=} \sum_{i \in I} \mathbb{E}_{A_i}[X]\mathbb{E}[\mathbf{1}_{A_i}] = \sum_{i \in I} \mathbb{E}_{A_i}[X]\mathbb{P}(A_i).$$

Noting that for each $i \in I$,

$$\mathbb{E}_{A_i}[X]\mathbb{P}(A_i) = \begin{cases} \frac{\mathbb{E}[X\mathbf{1}_{A_i}]}{\mathbb{P}(A_i)}\mathbb{P}(A_i) = \mathbb{E}[X\mathbf{1}_{A_i}] & \text{if } \mathbb{P}(A_i) > 0, \\ 0 \cdot 0 & \text{if } \mathbb{P}(A_i) = 0, \end{cases}$$

$$\overset{[5.1.10]}{=} \mathbb{E}[X\mathbf{1}_{A_i}],$$

we have

$$\mathbb{E}[Y\mathbf{1}_A] = \sum_{i \in I} \mathbb{E}[X\mathbf{1}_{A_i}] \overset{[5.1.11]\text{a}}{=} \mathbb{E}[X\mathbf{1}_A].$$

$\square$

Remarks:

- *(Interpretation of formula)* The formula of conditional expectation in Proposition 8.2.c can be more intuitively understood as follows. From an "information" perspective, we can interpret "given $\mathcal{G}$" as suggesting that, for the unknown outcome $\omega \in \Omega$, the given information is enough for us to determine whether each set in $\mathcal{G}$ contains ω. In the case here, we have $\mathcal{G} = \{\biguplus_{i \in I} A_i : A_i \in \mathcal{A} \; \forall i \in I \text{ and } I \subseteq \mathbb{N}\}$, so based on the information we have $\omega \in A_n$ for some known $n \in \mathbb{N}$ (but we still do not know what ω is exactly). With this piece of information, the "expectation" or "best prediction" for $X = X(\omega)$ should naturally be the mean of X over A_n (containing all the possible candidates of ω), scaled by for the probability of A_n, i.e., $\mathbb{E}[X|\mathcal{G}](\omega) \overset{(\omega \in A_n)}{=} \mathbb{E}[X\mathbf{1}_{A_n}]/\mathbb{P}(A_n) = \mathbb{E}_{A_n}[X]$.

- *(Conditional probability)* Proposition 8.2.c also gives rise to formulas of conditional probability. By taking $X = \mathbf{1}_A$ with $A \in \mathcal{F}$, we have

$$\mathbb{P}(A|\mathcal{G}) = \mathbb{E}[\mathbf{1}_A|\mathcal{G}] = \sum_{n=1}^{\infty} \frac{\mathbb{E}[\mathbf{1}_A\mathbf{1}_{A_n}]}{\mathbb{P}(A_n)}\mathbf{1}_{A_n} = \sum_{n=1}^{\infty} \frac{\mathbb{E}[\mathbf{1}_{A \cap A_n}]}{\mathbb{P}(A_n)}\mathbf{1}_{A_n} = \sum_{n=1}^{\infty} \frac{\mathbb{P}(A \cap A_n)}{\mathbb{P}(A_n)}\mathbf{1}_{A_n} = \sum_{n=1}^{\infty} \mathbb{P}(A|A_n)\mathbf{1}_{A_n}$$

with $\mathbb{P}(A|A_n) := 0$ if $\mathbb{P}(A_n) = 0$. This formula suggests that $\mathbb{P}(A|\mathcal{G})(\omega) = \mathbb{P}(A|A_n)$ whenever $\omega \in A_n$, which is quite intuitive.

- *(Conditional probability given discrete $\boldsymbol{X}$)* The formula above can be reduced to a more familiar one by taking $\mathcal{G} = \sigma(\boldsymbol{X})$ where $\boldsymbol{X}$ has support $\mathrm{supp}(\boldsymbol{X}) = \{\boldsymbol{x}_n : n \in \mathbb{N}\}$. In such case, we can write $\mathcal{G} = \{\boldsymbol{X}^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^d)\} = \sigma(\mathcal{A})$ where $\mathcal{A}$ is a partition of Ω, including every set $\{\boldsymbol{X} = \boldsymbol{x}_n\}$ (and also the set $\{\boldsymbol{X} \notin \mathrm{supp}(\boldsymbol{X})\}$, if nonempty, which always has probability zero and thus can be omitted in the sum below). Hence, by the formula above we have

$$\mathbb{P}(A|\boldsymbol{X}) := \mathbb{P}(A|\sigma(\boldsymbol{X})) = \sum_{n=1}^{\infty} \mathbb{P}(A|\boldsymbol{X} = \boldsymbol{x}_n)\mathbf{1}_{\{\boldsymbol{X}=\boldsymbol{x}_n\}},$$

  meaning that if we know that $\boldsymbol{X} = \boldsymbol{x}_n$ (i.e., $\omega \in \{\boldsymbol{X} = \boldsymbol{x}_n\}$), the conditional probability $\mathbb{P}(A|\boldsymbol{X})$ is given by $\mathbb{P}(A|\boldsymbol{X} = \boldsymbol{x}_n)$ (very natural).

8.2.5 **Properties of conditional expectation.** Based on the definition of conditional expectation here, we can indeed deduce many properties that are perhaps somewhat familiar to you, and also allow us to work with conditional expectations more efficiently.

Proposition 8.2.d. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{H}, \mathcal{G}$ be σ-algebras such that $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$, and $X, Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$.

(a) *(No information)* $\mathbb{E}[X|\{\varnothing, \Omega\}] = \mathbb{E}[X]$.

(b) *(No relevant information)* If X is independent of $\mathcal{G}$ (i.e., $\sigma(X)$ and $\mathcal{G}$ are independent), then $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$.

(c) *(Full information)* If $X \in \mathcal{G}$, then $\mathbb{E}[X|\mathcal{G}] = X$. Particularly, $\mathbb{E}[c|\mathcal{G}] = c$ for every $c \in \mathbb{R}$.

(d) *(Linearity)* $\mathbb{E}[aX + bY|\mathcal{G}] = a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]$ for all $a, b \in \mathbb{R}$.

(e) *(Monotonicity)* If $X \leq Y$, then $\mathbb{E}[X|\mathcal{G}] \leq \mathbb{E}[Y|\mathcal{G}]$.

(f) *(Triangle inequality)* $|\mathbb{E}[X|\mathcal{G}]| \leq \mathbb{E}[|X||\mathcal{G}]$.

(g) *(Tower property)* $\mathbb{E}[\mathbb{E}[X|\mathcal{H}]|\mathcal{G}] = \mathbb{E}[X|\mathcal{H}] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}]$. Particularly, $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]]$ *(law of total expectation)*.

Proof.

(a) (1) Since $\mathbb{E}[X]$ is constant, it is $\{\varnothing, \Omega\}$-measurable.

   (2) We have $\mathbb{E}[\mathbb{E}[X]\mathbf{1}_\varnothing] \overset{[5.1.10]}{=} 0 \overset{[5.1.10]}{=} \mathbb{E}[X\mathbf{1}_\varnothing]$ and $\mathbb{E}[\mathbb{E}[X]\mathbf{1}_\Omega] = \mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] = \mathbb{E}[X\mathbf{1}_\Omega]$.

(b) (1) Since $\mathbb{E}[X]$ is constant, it is $\mathcal{G}$-measurable.

   (2) For all $A \in \mathcal{G}$, we have $\mathbb{E}[\mathbb{E}[X]\mathbf{1}_A] \overset{\text{(linearity)}}{=} \mathbb{E}[X]\mathbb{E}[\mathbf{1}_A] \overset{\text{(independence)}}{=} \mathbb{E}[X\mathbf{1}_A]$.

(c) Let $Y = X$. We have:

   (1) Y is $\mathcal{G}$-measurable by definition.

   (2) For all $A \in \mathcal{G}$, we have $\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A]$.

   Hence, $\mathbb{E}[X|\mathcal{G}] = Y = X$. In particular, the random variable (constant function) c is always $\mathcal{G}$-measurable, and so $\mathbb{E}[c|\mathcal{G}] = c$.

(d) (1) Since $\mathbb{E}[X|\mathcal{G}]$ and $\mathbb{E}[Y|\mathcal{G}]$ are $\mathcal{G}$-measurable, $a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]$ is also $\mathcal{G}$-measurable.

   (2) For all $A \in \mathcal{G}$, we have

$$\mathbb{E}[(a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}])\mathbf{1}_A] \overset{\text{(linearity)}}{=} a\mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A] + b\mathbb{E}[\mathbb{E}[Y|\mathcal{G}]\mathbf{1}_A]$$
$$\overset{\text{(partial averaging)}}{=} a\mathbb{E}[X\mathbf{1}_A] + b\mathbb{E}[Y\mathbf{1}_A] \overset{\text{(linearity)}}{=} \mathbb{E}[(aX + bY)\mathbf{1}_A].$$

(e) For all $A \in \mathcal{G}$, we have

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A] \overset{\text{(partial averaging)}}{=} \mathbb{E}[X\mathbf{1}_A] \overset{(X\mathbf{1}_A \leq Y\mathbf{1}_A)}{\leq} \mathbb{E}[Y\mathbf{1}_A] \overset{\text{(partial averaging)}}{=} \mathbb{E}[\mathbb{E}[Y|\mathcal{G}]\mathbf{1}_A].$$

Hence, we have

$$\mathbb{E}[(\mathbb{E}[X|\mathcal{G}] - \mathbb{E}[Y|\mathcal{G}])\mathbf{1}_A] \leq 0.$$

Now fix any $\varepsilon > 0$ and let $A_\varepsilon := \{\mathbb{E}[X|\mathcal{G}] - \mathbb{E}[Y|\mathcal{G}] \geq \varepsilon\} \in \mathcal{G}$ (as $\mathbb{E}[X|\mathcal{G}] - \mathbb{E}[Y|\mathcal{G}]$ is $\mathcal{G}$-measurable). Applying the inequality above on A_ε gives

$$0 \geq \mathbb{E}[(\mathbb{E}[X|\mathcal{G}] - \mathbb{E}[Y|\mathcal{G}])\mathbf{1}_{A_\varepsilon}] \overset{\text{(monotonicity)}}{\geq} \mathbb{E}[\varepsilon \mathbf{1}_{A_\varepsilon}] = \varepsilon \mathbb{P}(A_\varepsilon) \geq 0.$$

This implies that $\mathbb{P}(A_\varepsilon) = 0$ for all $\varepsilon > 0$, and hence

$$0 \leq \mathbb{P}(\mathbb{E}[X|\mathcal{G}] > \mathbb{E}[Y|\mathcal{G}]) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_{1/n}\right) \overset{\text{(subadditivity)}}{\leq} \sum_{n=1}^{\infty} \mathbb{P}(A_{1/n}) = 0,$$

meaning that $\mathbb{P}(\mathbb{E}[X|\mathcal{G}] > \mathbb{E}[Y|\mathcal{G}]) = 0$. Therefore, we have $\mathbb{E}[X|\mathcal{G}] \leq \mathbb{E}[Y|\mathcal{G}]$ (a.s.).

(f) Consider

$$\left|\mathbb{E}[X|\mathcal{G}]\right| = \left|\mathbb{E}[X^+ - X^-|\mathcal{G}]\right| \overset{\text{(linearity)}}{=} \left|\mathbb{E}[X^+|\mathcal{G}] - \mathbb{E}[X^-|\mathcal{G}]\right|$$

$$\overset{\text{(triangle)}}{\leq} \big|\underbrace{\mathbb{E}[X^+|\mathcal{G}]}_{\substack{\geq \ 0 \\ \text{(monotonicity)}}}\big| + \big|\underbrace{\mathbb{E}[X^-|\mathcal{G}]}_{\substack{\geq \ 0 \\ \text{(monotonicity)}}}\big| = \mathbb{E}[X^+|\mathcal{G}] + \mathbb{E}[X^-|\mathcal{G}] = \mathbb{E}[|X||\mathcal{G}].$$

(g) Since $\mathbb{E}[X|\mathcal{H}] \in \mathcal{H}$ and $\mathcal{H} \subseteq \mathcal{G}$, we have $\mathbb{E}[X|\mathcal{H}] \in \mathcal{G}$. Hence, by (c) we have $\mathbb{E}[\mathbb{E}[X|\mathcal{H}]|\mathcal{G}] = \mathbb{E}[X|\mathcal{H}]$. This proves the first equality. Now, consider the second equality:

(1) By definition, $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}]$ is $\mathcal{H}$-measurable.

(2) For all $A \in \mathcal{H} \subseteq \mathcal{G}$, we have

$$\mathbb{E}[\mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}]\mathbf{1}_A] \overset{\substack{\text{(partial averaging)} \\ (A \in \mathcal{H})}}{=} \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A] \overset{\substack{\text{(partial averaging)} \\ (A \in \mathcal{G})}}{=} \mathbb{E}[X\mathbf{1}_A] \overset{\substack{\text{(partial averaging)} \\ (A \in \mathcal{H})}}{=} \mathbb{E}[\mathbb{E}[X|\mathcal{H}]\mathbf{1}_A].$$

To get the law of total expectation, take $\mathcal{H} = \{\varnothing, \Omega\} \subseteq \mathcal{G}$. Then, we have

$$\mathbb{E}[X] \overset{\text{(a)}}{=} \mathbb{E}[X|\{\varnothing, \Omega\}] \overset{\text{(tower)}}{=} \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\{\varnothing, \Omega\}] \overset{\text{(a)}}{=} \mathbb{E}[\mathbb{E}[X|\mathcal{G}]].$$

$\square$

Remarks:

- *("Tower" in tower property)* In the tower property, we have $\mathbb{E}[\mathbb{E}[X|\mathcal{H}]|\mathcal{G}] = \mathbb{E}[X|\mathcal{H}] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}]$. There are several "layers" of conditioning involved, which looks like a "tower", hence the name "tower property".

- *(Interpretation of tower property)* To interpret the tower property more intuitively, we can consider the impact of the coarseness/fineness[10] of $\mathcal{G}$ on the conditional expectation (averaging): For a coarser (finer) $\mathcal{G}$, generally "larger" ("smaller") pieces are contained. Hence, "stronger" ("weaker") averaging takes place over the "pieces", making $\mathbb{E}[X|\mathcal{G}]$ retain less (more) information about X. Parts (a) and (c) provide two extreme examples: (i) For $\mathcal{G} = \{\varnothing, \Omega\}$, it is so coarse that no information is retained except just the mean of X: $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$, and (ii) For $\mathcal{G}$ with $X \in \mathcal{G}$, it is so fine that full information about X is retained: $\mathbb{E}[X|\mathcal{G}] = X$ (there is no "averaging" taking place).

  Armed with this idea, the tower property can then be interpreted as saying that *the coarser σ-algebra remains* ($\mathcal{H}$ in the case with $\mathcal{G} \subseteq \mathcal{H}$), which can be intuitively understood as follows:

  - $\mathbb{E}[\mathbb{E}[X|\mathcal{H}]|\mathcal{G}] = \mathbb{E}[X|\mathcal{H}]$: Conditioning on the coarser $\mathcal{H}$ already retains only a little information about X, and conditioning on the finer $\mathcal{G}$ afterwards cannot "bring back" the information lost.

---

[10]We say that $\mathcal{H}$ is coarser than $\mathcal{G}$ (or $\mathcal{G}$ is finer than $\mathcal{H}$) if $\mathcal{H} \subseteq \mathcal{G}$, i.e., $\mathcal{H}$ contains fewer "pieces" and hence they are generally "larger"/"rougher".

– $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}] = \mathbb{E}[X|\mathcal{H}]$ Conditioning on the finer $\mathcal{G}$ first allows us to retain more information about X, but conditioning on the coarser $\mathcal{H}$ afterwards just leads to more loss in information, and so at the end, still a little information about X is retained.

8.2.6 **Inequalities about conditional expectation.** In [5.1.14]a to [5.1.14]c, we have studied various inequalities about expectation. It turns out that there is an analogous versions of them for *conditional* expectation. But before stating them, we first need to define a "conditional version" of L^p norm. We define $\|X|\mathcal{G}\|_p := \mathbb{E}[|X|^p|\mathcal{G}]^{1/p}$ for every $p \in (0,\infty]$ (though we typically only consider $p \in [1,\infty]$). The inequalities are as follows. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

(a) *(Conditional Hölder's inequality)* Let $p, q \in [1,\infty]$ with $1/p + 1/q = 1$ (with the convention that $1/\infty := 0$) be conjugate indices. Then, $\|XY|\mathcal{G}\|_1 \leq \|X|\mathcal{G}\|_p \|Y|\mathcal{G}\|_q$ for all $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ and $Y \in L^q(\Omega, \mathcal{F}, \mathbb{P})$ (which can then be shown to imply $XY \in L^1(\Omega, \mathcal{F}, \mathbb{P})$).

[Note: The special case with $p = q = 2$ is known as the *conditional Cauchy-Schwarz inequality*.]

(b) *(Conditional Minkowski's inequality)* Let $p \in [1,\infty]$. Then $\|X + Y|\mathcal{G}\|_p \leq \|X|\mathcal{G}\|_p + \|Y|\mathcal{G}\|_p$ for all $X, Y \in L^p(\Omega, \mathcal{F}, \mathbb{P})$.

(c) *(Conditional Jensen's inequality)* Let $X : \Omega \to \mathbb{R}$ be a function in $L^1(\Omega, \mathcal{F}, \mathbb{P})$, and φ be a convex function on $\mathbb{R}$ such that $\varphi(X) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. Then, $\varphi(\mathbb{E}[X|\mathcal{G}]) \leq \mathbb{E}[\varphi(X)|\mathcal{G}]$.

[Note: Like [5.1.14]c, if φ is concave (with $\varphi(X) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ still), then we have $\varphi(\mathbb{E}[X|\mathcal{G}]) \geq \mathbb{E}[\varphi(X)|\mathcal{G}]$.]

Proof. Omitted. $\qquad \square$

8.2.7 **Contraction and convergence in L^p for conditional expectation.**

(a) *(Contraction in L^p)* If $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ and $p \in [1,\infty)$, then $\|\mathbb{E}[X|\mathcal{G}]\|_p \leq \|X\|_p$.

(b) *(Convergence in L^p)* Let $X, X_1, X_2, \ldots$ be in $L^p(\Omega, \mathcal{F}, \mathbb{P})$, with $p \in [1,\infty)$. If $X_n \xrightarrow[n\to\infty]{L^p} X$, then $\mathbb{E}[X_n|\mathcal{G}] \xrightarrow[n\to\infty]{L^p} \mathbb{E}[X|\mathcal{G}]$.

Proof.

(a) Note that

$$\|\mathbb{E}[X|\mathcal{G}]\|_p = \mathbb{E}[|\mathbb{E}[X|\mathcal{G}]|^p]^{1/p} \overset{\text{(conditional Jensen)}}{\leq} \mathbb{E}[\mathbb{E}[|X|^p|\mathcal{G}]]^{1/p} \overset{\text{(total expectation)}}{=} \mathbb{E}[|X|^p]^{1/p} = \|X\|_p.$$

(b) Note that

$$\|\mathbb{E}[X_n|\mathcal{G}] - \mathbb{E}[X|\mathcal{G}]\|_p \overset{\text{(linearity)}}{=} \|\mathbb{E}[X_n - X|\mathcal{G}]\|_p \overset{\text{(a)}}{\leq} \|X_n - X\|_p \xrightarrow[n\to\infty]{\text{(assumption)}} 0.$$

Hence, by definition we have $\mathbb{E}[X_n|\mathcal{G}] \xrightarrow[n\to\infty]{L^p} \mathbb{E}[X|\mathcal{G}]$.

$\qquad \square$

8.2.8 **Conditional versions of convergence results.** In Theorems 5.1.c and 5.1.e and lemma 5.1.d, we have studied three notable convergence results for expectation, namely MCT, Fatou's lemma, and DCT. Like the inequalities, there are also conditional versions for them, as follows.

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

(a) *(Conditional monotone convergence theorem)* Let X_n be an integrable function in L_+ for every $n \in \mathbb{N}$. If we have $X_n \nearrow X$ pointwisely, then X is in L_+ and is integrable, and $\mathbb{E}[X_n|\mathcal{G}] \nearrow \mathbb{E}[X|\mathcal{G}]$ a.s.

(b) *(Conditional Fatou's lemma)* Let X_n be an integrable function in L_+ for every $n \in \mathbb{N}$. Then, $\mathbb{E}[\liminf_{n\to\infty} X_n|\mathcal{G}] \leq \liminf_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}]$ a.s.

(c) *(Conditional dominated convergence theorem)* Let $X_n \in L^1$ for every $n \in \mathbb{N}$, and $X : \Omega \to \mathbb{R}$ be measurable. If $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$ and $|X_n| \le Y$ a.s. for all $n \in \mathbb{N}$, for some $Y \in L^1$ *(domination)*, then $X \in L^1$ and $\lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}] = \mathbb{E}[X|\mathcal{G}]$ a.s.

Proof.

(a) By monotonicity, $X_n \nearrow X$ pointwisely implies $\mathbb{E}[X_n|\mathcal{G}] \nearrow$ pointwisely (a.s.), and so we can let $Y := \lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}]$ (a.s.). By [3.1.5]c, Y is $\mathcal{G}$-measurable. Also, for all $A \in \mathcal{G}$ we have

$$\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}\Big[\lim_{n\to\infty}(\mathbb{E}[X_n|\mathcal{G}]\mathbf{1}_A)\Big] \overset{(\text{MCT})}{=} \lim_{n\to\infty} \mathbb{E}[\mathbb{E}[X_n|\mathcal{G}]\mathbf{1}_A] \overset{(\text{partial averaging})}{=} \lim_{n\to\infty} \mathbb{E}[X_n\mathbf{1}_A] \overset{(\text{MCT})}{=} \mathbb{E}[X\mathbf{1}_A].$$

By Proposition 8.2.a, we know that $Y \in L^1$. Hence, by taking $A = \Omega \in \mathcal{G}$ we know $X \in L_+$ is integrable (MCT implies that $X \in L_+$). Furthermore, we have $\mathbb{E}[X|\mathcal{G}] = Y = \lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}]$.

(b) Like the proof of Fatou's lemma, let $Y_n := \inf_{k\ge n} X_k$ for every $n \in \mathbb{N}$. Since $0 \le Y_n \le X_n$ for each $n \in \mathbb{N}$, each Y_n is in L_+ and is integrable. Also, $Y_n \nearrow$ pointwisely by construction, thus we can let $Y := \lim_{n\to\infty} Y_n = \liminf_{n\to\infty} X_n$. Hence,

$$\mathbb{E}\Big[\liminf_{n\to\infty} X_n|\mathcal{G}\Big] = \mathbb{E}[Y|\mathcal{G}] \overset{(\text{conditional MCT})}{=} \lim_{n\to\infty} \mathbb{E}[Y_n|\mathcal{G}] = \liminf_{n\to\infty} \mathbb{E}[Y_n|\mathcal{G}]$$

$$= \liminf_{n\to\infty} \mathbb{E}\Big[\inf_{k\ge n} X_k\Big|\mathcal{G}\Big] \overset{(\inf_{k\ge n} X_k \le X_n)}{=} \liminf_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}].$$

(c) By DCT, we immediately have $X \in L^1$. So it remains to show that $\lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}] = \mathbb{E}[X|\mathcal{G}]$ a.s. Noting that

$$0 \le |\mathbb{E}[X_n|\mathcal{G}] - \mathbb{E}[X|\mathcal{G}]| \overset{(\text{linearity})}{=} |\mathbb{E}[X_n - X|\mathcal{G}]| \overset{(\text{triangle})}{\le} \mathbb{E}[|X_n - X||\mathcal{G}] \overset{(\text{monotonicity})}{\le} \mathbb{E}\Big[\sup_{k\ge n}|X_k - X|\Big|\mathcal{G}\Big],$$

it suffices to show that $\lim_{n\to\infty} \mathbb{E}[Z_n] = 0$ a.s., with $Z_n := \sup_{k\ge n}|X_k - X|$ for every $n \in \mathbb{N}$. By Theorem 6.1.b, we have $Z_n \xrightarrow[n\to\infty]{\text{P}} 0$. Also, for each $n \in \mathbb{N}$, we have

$$0 \le Z_n \overset{(\text{triangle})}{\le} \sup_{k\ge n}(\underbrace{|X_n|}_{\substack{\le Y \\ \text{a.s.}}} + \underbrace{|X|}_{\substack{\le Y \\ \text{a.s.}}}) \overset{\text{a.s.}}{\le} 2Y \in L^1.$$

It then follows by Corollary 6.1.d that $\lim_{n\to\infty} \mathbb{E}[Z_n] = \mathbb{E}[0] = 0$.

Since $Z_n \ge 0$ and $Z_n \searrow$, by monotonicity we know $\mathbb{E}[Z_n|\mathcal{G}] \ge 0$ and $\mathbb{E}[Z_n|\mathcal{G}] \searrow$. Therefore, we can let $Z := \lim_{n\to\infty} \mathbb{E}[Z_n|\mathcal{G}]$, which is $\mathcal{G}$-measurable by [3.1.5]c. Also, since $0 \le Z \le \mathbb{E}[Z_n|\mathcal{G}] \overset{(\text{Proposition 8.2.a})}{\in} L^1$ for each $n \in \mathbb{N}$, we know Z is in L_+ and is integrable, and also

$$0 \overset{(\text{monotonicity})}{\le} \mathbb{E}[Z] \overset{(\text{monotonicity})}{\le} \mathbb{E}[\mathbb{E}[Z_n|\mathcal{G}]] \overset{(\text{total expectation})}{=} \mathbb{E}[Z_n] \xrightarrow[n\to\infty]{(\text{above})} 0,$$

which implies that $\mathbb{E}[Z] = 0$. Hence, $\lim_{n\to\infty} \mathbb{E}[Z_n|\mathcal{G}] \overset{(\text{definition})}{=} Z \overset{\text{a.s.}}{=} 0$.
[5.1.8]

$\square$

8.2.9 **"Removing" σ-algebra in conditional expectation based on independence.** By the *no relevant information* property in Proposition 8.2.d, we know that $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$ if $\mathcal{G}$ is independent of $\sigma(X)$. Here, the σ-algebra $\mathcal{G}$ can be "removed" from the conditional expectation due to the independence. This property is generalized by the following result, which suggests when such "removal" can take place with two σ-algebras involved.

Proposition 8.2.e. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ be σ-algebras, and $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. If $\mathcal{H}$ is independent of $\sigma(\sigma(X), \mathcal{G})$, then $\mathbb{E}[X|\sigma(\mathcal{G}, \mathcal{H})] = \mathbb{E}[X|\mathcal{G}]$.

Proof. Consider first the case with $X \geq 0$. Let Y be a version of $\mathbb{E}[X|\mathcal{G}]$. Then Y is $\mathcal{G}$-measurable by definition, and $Y \geq 0$ a.s. by monotonicity. By [5.1.8], the functions μ and ν defined by $\mu(A) = \mathbb{E}[Y\mathbf{1}_A]$ and $\nu(A) = \mathbb{E}[X\mathbf{1}_A]$ for all $A \in \mathcal{F}$ are measures on $(\Omega, \mathcal{F})$. Note that $\mu(\Omega) = \mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]] \overset{\text{(total expectation)}}{=} \mathbb{E}[X] < \infty$ and $\nu(\Omega) = \mathbb{E}[X] < \infty$ also. Hence both measures are finite, thus also σ-finite.

For all $G \in \mathcal{G}$ and $H \in \mathcal{H}$, we have

$$\mu(G \cap H) = \mathbb{E}[Y\mathbf{1}_{G \cap H}] = \mathbb{E}[Y\mathbf{1}_G\mathbf{1}_H] \overset{(Y \in \mathcal{G},\ \mathcal{H}\ \text{independent of } \mathcal{G})}{=} \mathbb{E}[Y\mathbf{1}_G]\mathbb{E}[\mathbf{1}_H]$$

$$\overset{\text{(partial averaging)}}{=} \mathbb{E}[X\mathbf{1}_G]\mathbb{E}[\mathbf{1}_H] \overset{(\mathcal{H}\ \text{independent of } \sigma(\sigma(X), \mathcal{G}))}{=} \mathbb{E}[X\mathbf{1}_G\mathbf{1}_H] = \mathbb{E}[X\mathbf{1}_{G \cap H}] = \nu(G \cap H).$$

This means that μ and ν are σ-finite measures that coincide on the π-system $\{G \cap H : G \in \mathcal{G}, H \in \mathcal{H}\}$. Thus, by Proposition 2.1.a, μ and ν coincide on $\sigma(\{G \cap H : G \in \mathcal{G}, H \in \mathcal{H}\} \overset{[1.4.5]}{=} \sigma(\mathcal{G}, \mathcal{H})$. In other words, for every $A \in \sigma(\mathcal{G}, \mathcal{H})$, we have $\mu(A) = \mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A] = \nu(A)$. This shows the partial averaging property.

Therefore, $Y = \mathbb{E}[X|\mathcal{G}]$ is also a version of $\mathbb{E}[X|\sigma(\mathcal{G}, \mathcal{H})]$. It then follows by the a.s. uniqueness of conditional expectation (Theorem 8.2.b) that $\mathbb{E}[X|\mathcal{G}] \overset{\text{a.s.}}{=} \mathbb{E}[X|\sigma(\mathcal{G}, \mathcal{H})]$.

Now, consider the general case where X may not be nonnegative. We write $X = X^+ - X^-$. Then, by the proven case and linearity, we get

$$\mathbb{E}[X|\mathcal{G}] \overset{\text{a.s.}}{=} \mathbb{E}[X^+|\mathcal{G}] - \mathbb{E}[X^-|\mathcal{G}] \overset{\text{a.s.}}{=} \mathbb{E}[X^+|\sigma(\mathcal{G}, \mathcal{H})] - \mathbb{E}[X^-|\sigma(\mathcal{G}, \mathcal{H})] \overset{\text{a.s.}}{=} \mathbb{E}[X|\sigma(\mathcal{G}, \mathcal{H})].$$

$\square$

To see how Proposition 8.2.e generalizes the *no relevant information* property (i.e., includes it as a special case), take $\mathcal{G} = \{\varnothing, \Omega\}$. Then, we have

$$\sigma(\sigma(X), \mathcal{G}) \overset{\text{(consider definition)}}{=} \sigma(\sigma(X)) \overset{(\sigma(X)\ \text{is the smallest } \sigma\text{-algebra containing itself})}{=} \sigma(X)$$

and

$$\sigma(\mathcal{G}, \mathcal{H}) \overset{\text{(consider definition)}}{=} \sigma(\mathcal{H}) \overset{(\mathcal{H}\ \text{is the smallest } \sigma\text{-algebra containing itself})}{=} \mathcal{H}.$$

Therefore, in this case Proposition 8.2.e is just asserting that $\mathbb{E}[X|\mathcal{H}] = \mathbb{E}[X|\{\varnothing, \Omega\}] \overset{\text{(no information)}}{=} \mathbb{E}[X]$, which is the same as the *no relevant information* property.

8.2.10 **Taking out what is known.** The next property of conditional expectation to be discussed is *taking out what is known (TOWIK)*, which is frequently used. In its proof, we will utilize the following lemma, which can simplify the verification of the *partial averaging* property for conditional expectation.

Lemma 8.2.f. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra on Ω, and $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. If $\mathcal{A} \subseteq \mathcal{F}$ is a π-system such that (i) $\sigma(\mathcal{A}) = \mathcal{G}$ *(generating $\mathcal{G}$)* and (ii) there exists $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{A}$ such that $\bigcup_{i=1}^{\infty} A_i = \Omega$ *(containing a cover of Ω)*, then the function $Y : \Omega \to \mathbb{R}$ is a version of $\mathbb{E}[X|\mathcal{G}]$ if it satisfies:

(1) *(Measurability)* $Y \in \mathcal{G}$.
(2) *(Partial averaging on $\mathcal{A}$)* $\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A]$ for all $A \in \mathcal{A}$.

Proof. First, for all $A \in \mathcal{A}$, we have $\mathbb{E}[|X\mathbf{1}_A|] \overset{\text{(monotonicity)}}{\leq} \mathbb{E}[|X|] < \infty$, and so $X\mathbf{1}_A$ is integrable. Now, by the partial averaging on $\mathcal{A}$, we have $\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A]$, which means that $Y\mathbf{1}_A$ is integrable also. Therefore, by the integrability of $X\mathbf{1}_A$ and $Y\mathbf{1}_A$ for all $A \in \mathcal{A}$, we get the following chain of equivalences:

$$\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A] \quad \forall A \in \mathcal{A}$$
$$\Longleftrightarrow \mathbb{E}[Y^+\mathbf{1}_A] - \mathbb{E}[Y^-\mathbf{1}_A] = \mathbb{E}[X^+\mathbf{1}_A] - \mathbb{E}[X^-\mathbf{1}_A] \quad \forall A \in \mathcal{A}$$
$$\Longleftrightarrow \mu(A) := \mathbb{E}[(Y^+ + X^-)\mathbf{1}_A] = \mathbb{E}[(X^+ + Y^-)\mathbf{1}_A] =: \nu(A) \quad \forall A \in \mathcal{A}.$$

Since $X \in \mathcal{F}$ and $Y \in \mathcal{G}$ (hence also $Y \in \mathcal{F}$), we know $Y^+ + X^- \in \mathcal{F}$ and $X^+ + Y^- \in \mathcal{F}$. Therefore, by [5.1.8] we know μ and ν are measures on $\mathcal{F}$ (hence also on $\mathcal{G}$).

With $\mu|_{\mathcal{A}} = \nu|_{\mathcal{A}}$ and the integrability of X, Y (thus also X^+, X^-, Y^+, Y^-) on $\mathcal{A} \subseteq \sigma(\mathcal{A}) = \mathcal{G}$, we get for each $i \in \mathbb{N}$, $\mu(A_i) \stackrel{(A_i \in \mathcal{A})}{=} \nu(A_i) \stackrel{\text{(integrability)}}{<} \infty$. Since $\bigcup_{i=1}^{\infty} A_i = \Omega$ by assumption, we have

$$\mu|_{\mathcal{G}} = \mu|_{\sigma(\mathcal{A})} \stackrel{\text{(Proposition 2.1.a)}}{=} \nu|_{\sigma(\mathcal{A})} = \nu|_{\mathcal{G}}.$$

This means that

$$\mu(A) = \mathbb{E}\big[(Y^+ + X^-)\mathbf{1}_A\big] = \mathbb{E}\big[(X^+ + Y^-)\mathbf{1}_A\big] = \nu(A) \quad \forall A \in \mathcal{G},$$

which is equivalent to

$$\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A] \quad \forall A \in \mathcal{G},$$

i.e., the partial averaging property for conditional expectation. Hence, by definition, Y is a version of $\mathbb{E}[X|\mathcal{G}]$. $\qquad\square$

Now we are ready to prove the *taking out what is known* property.

Proposition 8.2.g (Taking out what is known (TOWIK))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra on Ω. If XY and X are integrable and $Y \in \mathcal{G}$, then $\mathbb{E}[XY|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}]$ *(taking out the "known" Y).*

Proof.

(1) Since $Y \in \mathcal{G}$ and $\mathbb{E}[X|\mathcal{G}] \in \mathcal{G}$, we have $Y\mathbb{E}[X|\mathcal{G}] \in \mathcal{G}$.

(2) We apply the standard argument on Y.

    i. Fix any indicator function $Y = \mathbf{1}_B$ with $B \in \mathcal{G}$. For all $A \in \mathcal{G}$,

$$\mathbb{E}[Y\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A] = \mathbb{E}[\mathbf{1}_B\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_{A \cap B}] \stackrel{(A \cap B \in \mathcal{G})}{=} \mathbb{E}[X\mathbf{1}_{A \cap B}] = \mathbb{E}[\mathbf{1}_B X\mathbf{1}_A] = \mathbb{E}[XY\mathbf{1}_A].$$

    By linearity, the equality also holds for every simple $Y \in \mathcal{G}$ (i.e., those with the form $Y = \sum_{i=1}^n y_i \mathbf{1}_{A_i}$ where $y_1, \ldots, y_n \in \mathbb{R}$ and $A_1, \ldots, A_n \in \mathcal{G}$ are pairwise disjoint).

    ii. Fix any nonnegative $Y \in \mathcal{G}$. Consider first the case where $X \geq 0$.
    By Proposition 5.1.a, we know there exists a sequence $\{Y_n\}$ of nonnegative simple functions such that $Y_n \nearrow Y$. Hence, for all $A \in \mathcal{G}$, we have $Y_n\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A \nearrow Y\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A$ and $XY_n\mathbf{1}_A \nearrow XY\mathbf{1}_A$, both pointwisely, which imply that

$$\mathbb{E}[Y\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A] \stackrel{\text{(MCT)}}{=} \lim_{n \to \infty} \mathbb{E}[Y_n\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A] \stackrel{\text{(simple case)}}{\underset{\text{(partial averaging)}}{=}} \lim_{n \to \infty} \mathbb{E}[XY_n\mathbf{1}_A] \stackrel{\text{(MCT)}}{=} \mathbb{E}[XY\mathbf{1}_A].$$

    Next, consider the general case where X may not be nonnegative. Write $X = X^+ - X^-$, and then we get

$$\mathbb{E}[Y\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A] = \mathbb{E}\big[Y\mathbb{E}\big[X^+|\mathcal{G}\big]\mathbf{1}_A\big] - \mathbb{E}\big[Y\mathbb{E}\big[X^-|\mathcal{G}\big]\mathbf{1}_A\big] \stackrel{\text{(above case)}}{\underset{\text{(partial averaging)}}{=}} \mathbb{E}\big[X^+Y\mathbf{1}_A\big] - \mathbb{E}\big[X^-Y\mathbf{1}_A\big]$$

$$= \mathbb{E}[XY\mathbf{1}_A]$$

    for all $A \in \mathcal{G}$.

    iii. Fix any $Y \in \mathcal{G}$. Write $Y = Y^+ - Y^-$. Then, for all $A \in \mathcal{G}$ we have

$$\mathbb{E}[Y\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A] = \mathbb{E}\big[Y^+\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A\big] - \mathbb{E}\big[Y^-\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A\big] \stackrel{\text{(nonnegative case)}}{\underset{\text{(partial averaging)}}{=}} \mathbb{E}\big[XY^+\mathbf{1}_A\big] - \mathbb{E}\big[XY^-\mathbf{1}_A\big]$$

$$= \mathbb{E}[XY\mathbf{1}_A].$$

$\square$

A corollary of the TOWIK property is a "conditional version" of Proposition 5.2.d (expectation of product is product of expectations), with two random variables.

Corollary 8.2.h. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra on Ω. If X, Y, and XY are integrable, and Y is independent of $\sigma(\sigma(X), \mathcal{G})$, then $\mathbb{E}[XY|\mathcal{G}] = \mathbb{E}[Y]\mathbb{E}[X|\mathcal{G}]$.

Proof. Note that

$$\mathbb{E}[XY|\mathcal{G}] \overset{\text{(tower, } \mathcal{G} \subseteq \sigma(\sigma(X), \mathcal{G})))}{=} \mathbb{E}[\mathbb{E}[XY|\sigma(\sigma(X), \mathcal{G})]|\mathcal{G}] \overset{\text{(TOWIK)}}{=} \mathbb{E}[X\mathbb{E}[Y|\sigma(\sigma(X), \mathcal{G})]|\mathcal{G}]$$
$$\overset{\text{(Proposition 8.2.d)}}{=} \mathbb{E}[X\mathbb{E}[Y]|\mathcal{G}] = \mathbb{E}[Y]\mathbb{E}[X|\mathcal{G}].$$

$\square$

8.2.11 **Factorization/Doob-Dynkin lemma.** Here, we are going to study an important result, known as *factorization* or *Doob-Dynkin lemma*, that justifies the method of finding conditional expectation learnt in your first probability course. It turns out that, while the definition of conditional expectation given here is rather abstract, it is possible to be reduced to some more "comprehensible" forms that you have previously seen.

Theorem 8.2.i (Factorization/Doob-Dynkin lemma)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\Omega', \mathcal{F}')$ be a measurable space, and $Z : \Omega \to \Omega'$. The function $Y : \Omega \to \mathbb{R}$ is $(\sigma(Z), \mathcal{B}(\mathbb{R}))$-measurable iff there exists a $(\mathcal{F}', \mathcal{B}(\mathbb{R}))$-measurable function $h : \Omega' \to \mathbb{R}$ such that $Y = h(Z)$.

"factorize"
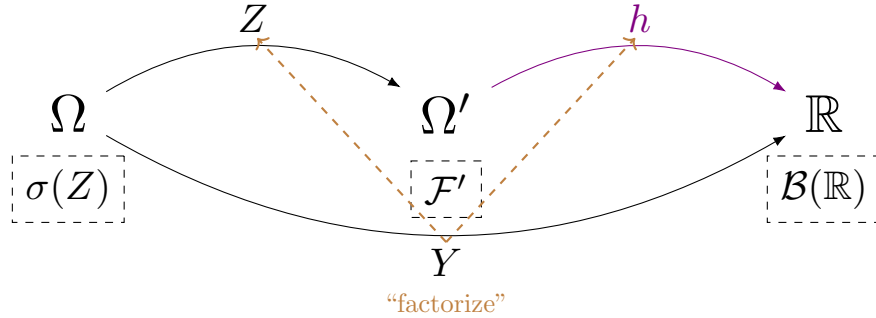
Proof. "$\Leftarrow$": Since Z is $(\sigma(Z), \mathcal{F}')$-measurable and h is $(\mathcal{F}', \mathcal{B}(\mathbb{R}))$-measurable, by [3.1.3]b we know that $Y = h(Z)$ is $(\sigma(Z), \mathcal{B}(\mathbb{R}))$-measurable.

"$\Rightarrow$": Let Y be $(\sigma(Z), \mathcal{B}(\mathbb{R}))$-measurable. We then apply the standard argument on Y.

(1) Fix any simple $Y = \sum_{i=1}^{n} y_i \mathbf{1}_{A_i}$ with $n \in \mathbb{N}$ and $A_i \in \sigma(Z)$ for every $i = 1, \ldots, n$. Noting that $A_i = Z^{-1}(A_i')$ with $A_i' \in \mathcal{F}'$ for each $i = 1, \ldots, n$, we have

$$Y(\omega) = \sum_{i=1}^{n} y_i \mathbf{1}_{A_i}(\omega) = \sum_{i=1}^{n} y_i \mathbf{1}_{Z^{-1}(A_i')}(\omega) \overset{(\omega \in Z^{-1}(A_i') \iff Z(\omega) \in A_i')}{=} \sum_{i=1}^{n} y_i \mathbf{1}_{A_i'}(Z(\omega)) = h(Z(\omega))$$

where $h(z) := \sum_{i=1}^{n} y_i \mathbf{1}_{A_i'}(z)$ is $(\mathcal{F}', \mathcal{B}(\mathbb{R}))$-measurable.

(2) Fix any $Y \geq 0$ that is $(\sigma(Z), \mathcal{B}(\mathbb{R}))$-measurable. By Proposition 5.1.a, there exists a sequence $\{Y_n\}$ of simple functions such that $Y_n \nearrow Y$. Now, for each $n \in \mathbb{N}$, we know by (1) that there exists a $(\mathcal{F}', \mathcal{B}(\mathbb{R}))$-measurable function h_n such that $Y_n = h_n(Z)$. Since $Y_n \nearrow$, we have $h_n \nearrow$ also. Thus, we have

$$Y = \lim_{n \to \infty} Y_n = \lim_{n \to \infty} h_n(Z) = \sup_{n \in \mathbb{N}} h_n(Z) =: h(Z),$$

where $h(z) = \sup_{n \in \mathbb{N}} h_n(z)$ is $(\mathcal{F}', \mathcal{B}(\mathbb{R}))$-measurable by [3.1.5]c.

(3) Fix any $(\sigma(Z), \mathcal{B}(\mathbb{R}))$-measurable Y. Write $Y = Y^+ - Y^-$. By (2), there exist $(\mathcal{F}', \mathcal{B}(\mathbb{R}))$-measurable functions h^+ and h^- such that $Y^+ = h^+(Z)$ and $Y^- = h^-(Z)$. Then, we have $Y = h(Z)$ with $h := h^+ - h^-$ being $(\mathcal{F}', \mathcal{B}(\mathbb{R}))$-measurable.

$\square$

Using Theorem 8.2.i, we can derive the following corollary which justifies the "elementary" way of computing conditional expectation.

Corollary 8.2.j (Factorization for conditional expectation). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $X \in L^1$ and $\boldsymbol{Z} : \Omega \to \mathbb{R}^d$ is a random vector, then there exists a measurable $h : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}[X|\boldsymbol{Z}] = h(\boldsymbol{Z})$.

[Note: For every $\boldsymbol{z} \in \mathbb{R}^d$, the value $\mathbb{E}[X|\boldsymbol{Z} = \boldsymbol{z}] := h(\boldsymbol{z})$ is said to be the **conditional expectation of X given $\boldsymbol{Z} = \boldsymbol{z}$**.]

Proof. Note that X is $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable, $\boldsymbol{Z}$ is $(\mathcal{F}, \mathcal{B}(\mathbb{R}^d))$-measurable, and $\mathbb{E}[X|\boldsymbol{Z}]$ is $(\sigma(Z), \mathcal{B}(\mathbb{R}))$-measurable. Therefore, the result follows by taking $\Omega' = \mathbb{R}^d$, $\mathcal{F}' = \mathcal{B}(\mathbb{R}^d)$, and $Y = \mathbb{E}[X|\boldsymbol{Z}]$ in Theorem 8.2.i. $\square$

8.2.12 **Factorization of conditional expectation under independence.** Both Theorem 8.2.i and Corollary 8.2.j suggest the *existence* of factorization, but do not give us an explicit "formula" for finding out what h is.[11] In the following, we will consider a special case where we do have an explicit formula for such h.

Proposition 8.2.k. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $\boldsymbol{X} : \Omega \to \mathbb{R}^{d_{\boldsymbol{x}}}$ and $\boldsymbol{Z} : \Omega \to \mathbb{R}^{d_{\boldsymbol{z}}}$ are independent random vectors, and $g : \mathbb{R}^{d_{\boldsymbol{x}}+d_{\boldsymbol{z}}} \to \mathbb{R}$ is a measurable function such that $g(\boldsymbol{X}, \boldsymbol{Z}) \in L^1$, then $\mathbb{E}[g(\boldsymbol{X}, \boldsymbol{Z})|\boldsymbol{Z}] = h(\boldsymbol{Z})$ (a.s.), where $h(\boldsymbol{z}) := \mathbb{E}[g(\boldsymbol{X}, \boldsymbol{z})]$.

Proof.

(1) It can be shown that h is $(\mathcal{B}(\mathbb{R}^{d_{\boldsymbol{z}}}), \mathcal{B}(\mathbb{R}))$-measurable (see Resnick (2014, Section 10.3) for more details), hence $(\mathcal{B}(\mathbb{R}^{d_{\boldsymbol{z}}}), \sigma(\boldsymbol{Z}))$-measurable. Together with the assumption that $\boldsymbol{Z}$ is $(\sigma(\boldsymbol{Z}), \mathcal{B}(\mathbb{R}^d))$-measurable, by [3.1.3]b we know $h(\boldsymbol{Z})$ is $\sigma(\boldsymbol{Z})$-measurable.

(2) Fix any $A \in \sigma(\boldsymbol{Z})$. Then, we can write $A = \boldsymbol{Z}^{-1}(B)$ for some $B \in \mathcal{B}(\mathbb{R}^{d_{\boldsymbol{z}}})$. With this expression, we note that $\mathbf{1}_A(\omega) = \mathbf{1}_{\boldsymbol{Z}^{-1}(B)}(\omega) = \mathbf{1}_B(\boldsymbol{Z}(\omega))$ for all $\omega \in \Omega$. Now, since $g(\boldsymbol{X}, \boldsymbol{Z}) \in L^1$, applying the Fubini's theorem yields

$$\int_A g(\boldsymbol{X}, \boldsymbol{Z})\,\mathrm{d}\mathbb{P} = \int_\Omega g(\boldsymbol{X}, \boldsymbol{Z})\mathbf{1}_A\,\mathrm{d}\mathbb{P} = \int_\Omega g(\boldsymbol{X}, \boldsymbol{Z})\mathbf{1}_B(\boldsymbol{Z})\,\mathrm{d}\mathbb{P}$$

$$\overset{\text{(Theorem 5.2.a)}}{=} \int_{\mathbb{R}^{d_{\boldsymbol{x}}+d_{\boldsymbol{z}}}} g(\boldsymbol{x}, \boldsymbol{z})\mathbf{1}_B(\boldsymbol{z})\,\mathrm{d}F_{\boldsymbol{X}, \boldsymbol{Z}}(\boldsymbol{x}, \boldsymbol{z})$$

$$\overset{\text{(Fubini, independence)}}{=} \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} \int_{\mathbb{R}^{d_{\boldsymbol{x}}}} g(\boldsymbol{x}, \boldsymbol{z})\mathbf{1}_B(\boldsymbol{z})\,\mathrm{d}F_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}F_{\boldsymbol{Z}}(\boldsymbol{z})$$

$$= \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} \mathbf{1}_B(\boldsymbol{z}) \underbrace{\int_{\mathbb{R}^{d_{\boldsymbol{x}}}} g(\boldsymbol{x}, \boldsymbol{z})\,\mathrm{d}F_{\boldsymbol{X}}(\boldsymbol{x})}_{\mathbb{E}[g(\boldsymbol{X}, \boldsymbol{z})] = h(\boldsymbol{z})}\,\mathrm{d}F_{\boldsymbol{Z}}(\boldsymbol{z})$$

$$= \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} \mathbf{1}_B(\boldsymbol{z})h(\boldsymbol{z})\,\mathrm{d}F_{\boldsymbol{Z}}(\boldsymbol{z}) \overset{\text{(Theorem 5.2.a)}}{=} \int_\Omega \mathbf{1}_B(\boldsymbol{Z})h(\boldsymbol{Z})\,\mathrm{d}\mathbb{P}$$

$$= \int_\Omega \mathbf{1}_A h(\boldsymbol{Z})\,\mathrm{d}\mathbb{P} = \int_A h(\boldsymbol{Z})\,\mathrm{d}\mathbb{P}.$$

$\square$

---

[11]While the proof of Theorem 8.2.i does suggest how such h can be constructed, it is not in a very explicit form and is generally hard to obtain.

[Note: Based on Proposition 8.2.k, we can compute conditional expectation of the form $\mathbb{E}[g(\boldsymbol{X}, \boldsymbol{Z})|\boldsymbol{Z}]$, where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent, by first computing $h(\boldsymbol{z}) = \mathbb{E}[g(\boldsymbol{X}, \boldsymbol{z})]$ *(replacing $\boldsymbol{Z}$ by $\boldsymbol{z}$ in the expression and removing the condition)*, and then conclude that the conditional expectation is $h(\boldsymbol{Z})$ *(replacing every $\boldsymbol{z}$ by $\boldsymbol{Z}$ in the final expression).*]

8.3 Applications of Conditional Expectation

8.3.1 After having a solid foundation on the measure-theoretic backbone of conditional expectation in Section 8.2, we now study some applications of conditional expectation, including deviations of formulas about *conditional distribution functions*, *variances*, and *random sums*, and also about the usage of conditional expectation in *regression analysis*.

8.3.2 **Conditional distribution formula.** Conditional expectation can be used for deriving the following versatile formula for computing distribution functions via conditioning.

Proposition 8.3.a. Let $\boldsymbol{X}$ and $\boldsymbol{Z}$ be random vectors. If $(\boldsymbol{X}, \boldsymbol{Z}) \sim F$ for a $(d_{\boldsymbol{x}} + d_{\boldsymbol{z}})$-dimensional distribution function F, then

$$F(\boldsymbol{x}, \boldsymbol{z}) = \int_{(-\infty, \boldsymbol{z}]} F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}|\widetilde{\boldsymbol{z}}) \, \mathrm{d}F_{\boldsymbol{Z}}(\widetilde{\boldsymbol{z}}) \quad \text{for all } (\boldsymbol{x}, \boldsymbol{z}) \in \mathbb{R}^{d_{\boldsymbol{x}} + d_{\boldsymbol{z}}}.$$

Proof. Let $h(\widetilde{\boldsymbol{z}}) = \mathbb{E}\big[\mathbf{1}_{\{\boldsymbol{X} \leq \boldsymbol{x}\}}|\boldsymbol{Z} = \widetilde{\boldsymbol{z}}\big] = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}|\boldsymbol{Z} = \widetilde{\boldsymbol{z}}) = F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}|\widetilde{\boldsymbol{z}})$. Consider

$$F(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}\big[\mathbf{1}_{\{\boldsymbol{X} \leq \boldsymbol{x}, \boldsymbol{Z} \leq \boldsymbol{z}\}}\big] = \mathbb{E}\big[\mathbf{1}_{\{\boldsymbol{X} \leq \boldsymbol{x}\}}\mathbf{1}_{\{\boldsymbol{Z} \leq \boldsymbol{z}\}}\big] \overset{\text{(total expectation)}}{=} \mathbb{E}\big[\mathbb{E}\big[\mathbf{1}_{\{\boldsymbol{X} \leq \boldsymbol{x}\}}\mathbf{1}_{\{\boldsymbol{Z} \leq \boldsymbol{z}\}}|\boldsymbol{Z}\big]\big]$$

$$\overset{\text{(TOWIK)}}{=} \mathbb{E}\big[\mathbf{1}_{\{\boldsymbol{Z} \leq \boldsymbol{z}\}}\mathbb{E}\big[\mathbf{1}_{\{\boldsymbol{X} \leq \boldsymbol{x}\}}|\boldsymbol{Z}\big]\big] = \mathbb{E}\big[\mathbf{1}_{\{\boldsymbol{Z} \leq \boldsymbol{z}\}}h(\boldsymbol{Z})\big] \overset{\text{(Theorem 5.2.a)}}{=} \int_{\mathbb{R}^{d_{\boldsymbol{z}}}} \mathbf{1}_{\{\widetilde{\boldsymbol{z}} \leq \boldsymbol{z}\}}h(\widetilde{\boldsymbol{z}}) \, \mathrm{d}F_{\boldsymbol{Z}}(\widetilde{\boldsymbol{z}})$$

$$= \int_{(-\infty, \boldsymbol{z}]} h(\widetilde{\boldsymbol{z}}) \, \mathrm{d}F_{\boldsymbol{Z}}(\widetilde{\boldsymbol{z}}) = \int_{(-\infty, \boldsymbol{z}]} F_{\boldsymbol{X}|\boldsymbol{Z}}(\boldsymbol{x}|\widetilde{\boldsymbol{z}}) \, \mathrm{d}F_{\boldsymbol{Z}}(\widetilde{\boldsymbol{z}}).$$

$\square$

8.3.3 **Law of total variance.** Apart from the *law of total expectation* (Proposition 8.2.d), there is also the *law of total variance*, but the formula is slightly more complex. The following lemma, which gives a conditional version of the "2nd moment minus 1st moment squared" formula for the variance, is helpful for establishing the law of total variance.

Lemma 8.3.b. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra, and $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Then, $\mathrm{Var}(X|\mathcal{G}) = \mathbb{E}\big[X^2|\mathcal{G}\big] - \mathbb{E}[X|\mathcal{G}]^2$.

Proof. Note that

$$\mathrm{Var}(X|\mathcal{G}) = \mathbb{E}\big[(X - \mathbb{E}[X|\mathcal{G}])^2|\mathcal{G}\big] \overset{\text{(linearity)}}{=} \mathbb{E}\big[X^2|\mathcal{G}\big] - 2\mathbb{E}[X\mathbb{E}[X|\mathcal{G}]|\mathcal{G}] + \mathbb{E}\big[(\mathbb{E}[X|\mathcal{G}])^2|\mathcal{G}\big]$$

$$\overset{\text{(TOWIK twice)}}{=} \mathbb{E}\big[X^2|\mathcal{G}\big] - 2\mathbb{E}[X|\mathcal{G}]\mathbb{E}[X|\mathcal{G}] + (\mathbb{E}[X|\mathcal{G}])^2 = \mathbb{E}\big[X^2|\mathcal{G}\big] - (\mathbb{E}[X|\mathcal{G}])^2.$$

$\square$

Proposition 8.3.c (Law of total variance)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra, and $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Then, $\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X|\mathcal{G})] + \mathrm{Var}(\mathbb{E}[X|\mathcal{G}])$.

Proof. We have

$$\mathrm{Var}(X) = \mathbb{E}\big[X^2\big] - \mathbb{E}[X]^2 \overset{\text{(total expectation twice)}}{=} \mathbb{E}\big[\mathbb{E}\big[X^2|\mathcal{G}\big]\big] - \mathbb{E}[\mathbb{E}[X|\mathcal{G}]]^2$$

$$\overset{\text{(Lemma 8.3.b)}}{=} \mathbb{E}\big[\mathrm{Var}(X|\mathcal{G}) + \mathbb{E}[X|\mathcal{G}]^2\big] - \mathbb{E}[\mathbb{E}[X|\mathcal{G}]]^2 \overset{\text{(linearity)}}{=} \mathbb{E}[\mathrm{Var}(X|\mathcal{G})] + \mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]^2\big] - \mathbb{E}[\mathbb{E}[X|\mathcal{G}]]^2$$

$$= \mathbb{E}[\mathrm{Var}(X|\mathcal{G})] + \mathrm{Var}(\mathbb{E}[X|\mathcal{G}]).$$

$\square$

8.3.4 Formulas of expectation and variance of random sum. In [4.2.7], we have seen the appearance of random variable like $S = \sum_{i=1}^{N} X_i$, which is a *random sum* with both the number of summands and the summands themselves being random variables. To compute the expectation and variance of such random sum, the following formulas are often used.

Proposition 8.3.d. Let $N, X_1, X_2, \ldots$ be independent, $N \in \mathbb{N}_0$, $X_1, X_2, \ldots$ be identically distributed, and $S = \sum_{i=1}^{N} X_i$.

(a) *(Wald's equation)* If $N, X_1 \in L^1$, then $S \in L^1$ and $\mathbb{E}[S] = \mathbb{E}[N]\mathbb{E}[X_1]$.

(b) *(Blackwell-Girshick equation)* If $N, X_1 \in L^2$, then $S \in L^2$ and $\operatorname{Var}(S) = \mathbb{E}[N]\operatorname{Var}(X_1) + \operatorname{Var}(N)\mathbb{E}[X_1]^2$.

Proof.

(a) We first show $S \in L^1$. Let $S_n = \sum_{i=1}^{n} X_i$ for each $n \in \mathbb{N}_0$. Then, we have $S = \sum_{n=1}^{\infty} S_n \mathbf{1}_{\{N=n\}}$, and thus

$$\mathbb{E}[|S|] \overset{\text{(triangle)}}{\leq} \sum_{n=1}^{\infty} \mathbb{E}\big[|S_n|\mathbf{1}_{\{N=n\}}\big] \overset{\text{(independence)}}{=} \sum_{n=1}^{\infty} \mathbb{E}[|S_n|]\mathbb{E}\big[\mathbf{1}_{\{N=n\}}\big]$$
$$\overset{\text{(triangle)}}{\leq} \sum_{n=1}^{\infty} n\mathbb{E}[|X_1|]\mathbb{P}(N=n) = \mathbb{E}[|X_1|]\sum_{n=1}^{\infty} n\mathbb{P}(N=n) = \mathbb{E}[|X_1|]\mathbb{E}[N] < \infty.$$

Repeating the same argument without the absolute value $|\cdot|$ (inequalities becomes equalities as there is no absolute value) yields $\mathbb{E}[S] = \mathbb{E}[N]\mathbb{E}[X_1]$. Alternatively, we can apply Proposition 8.2.k as follows. Noting that $S = \sum_{i=1}^{N} X_i = g(\boldsymbol{X}, N) \in L^1$, we have

$$\mathbb{E}[S|N] = \mathbb{E}[g(\boldsymbol{X}, N)|N] \overset{\text{(Proposition 8.2.k)}}{=} h(N),$$

where $h(n) = \mathbb{E}[g(\boldsymbol{X}, n)] = \mathbb{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n\mathbb{E}[X_1]$. Hence, we get

$$\mathbb{E}[S] \overset{\text{(total expectation)}}{=} \mathbb{E}[\mathbb{E}[S|N]] = \mathbb{E}[h(N)] = \mathbb{E}[N\mathbb{E}[X_1]] = \mathbb{E}[N]\mathbb{E}[X_1].$$

(b) Again, we first show $S \in L^2$. Let $S_n = \sum_{i=1}^{n} X_i$ for each $n \in \mathbb{N}_0$. Then, we have $S^2 = \sum_{n=1}^{\infty} S_n^2 \mathbf{1}_{\{N=n\}}$, and thus

$$\mathbb{E}\big[S^2\big] = \sum_{n=1}^{\infty} \mathbb{E}\big[S_n^2 \mathbf{1}_{\{N=n\}}\big] \overset{\text{(independence)}}{=} \sum_{n=1}^{\infty} \mathbb{E}\big[S_n^2\big]\mathbb{P}(N=n)$$
$$= \sum_{n=1}^{\infty} (\operatorname{Var}(S_n) + \mathbb{E}[S_n]^2)\mathbb{P}(N=n) \overset{\substack{\text{(independence)}\\\text{(linearity)}}}{=} \sum_{n=1}^{\infty} (n\operatorname{Var}(X_1) + (n\mathbb{E}[X_1])^2)\mathbb{P}(N=n)$$
$$= \mathbb{E}[N]\operatorname{Var}(X_1) + \mathbb{E}\big[N^2\big]\mathbb{E}[X_1]^2 < \infty,$$

which implies that $S \in L^2$.

Next, to show the equation, there are again two approaches:

- *Method 1: Using the relationship* $\operatorname{Var}(S) = \mathbb{E}\big[S^2\big] - \mathbb{E}[S]^2$. Note that

$$\operatorname{Var}(S) = \mathbb{E}[N]\operatorname{Var}(X_1) + \mathbb{E}\big[N^2\big]\mathbb{E}[X_1]^2 - (\mathbb{E}[N]\mathbb{E}[X_1])^2 = \mathbb{E}[N]\operatorname{Var}(X_1) + \mathbb{E}\big[N^2\big]\operatorname{Var}(X_1).$$

- *Method 2: Using Proposition 8.2.k.* First, we have

$$\operatorname{Var}(S|N) = \mathbb{E}\big[(S - \mathbb{E}[S|N])^2|N\big] \overset{\text{(see (a))}}{=} \mathbb{E}\big[(S - N\mathbb{E}[X_1])^2|N\big]$$
$$\overset{(g(\boldsymbol{X},N)=(S-N\mathbb{E}[X_1])^2)}{=} \mathbb{E}[g(\boldsymbol{X}, N)|N] \overset{\text{(Proposition 8.2.k)}}{=} h(N)$$

where $h(n) = \mathbb{E}[g(\boldsymbol{X}, n)] = \mathbb{E}\big[(\sum_{i=1}^n X_i - n\mathbb{E}[X_1])^2\big] = \text{Var}(\sum_{i=1}^n X_i) \overset{\text{(independence)}}{=} n\,\text{Var}(X_1)$.
Hence, $\text{Var}(S|N) = N\,\text{Var}(X_1)$. Therefore, by the law of total variance,

$$\text{Var}(S) = \mathbb{E}[\text{Var}(S|N)] + \text{Var}(\mathbb{E}[S|N]) \overset{\text{(see (a))}}{=} \mathbb{E}[N\,\text{Var}(X_1)] + \text{Var}(N\mathbb{E}[X_1])$$
$$= \mathbb{E}[N]\,\text{Var}(X_1) + \text{Var}(N)\mathbb{E}[X_1]^2.$$

$\square$

8.3.5 **Role of conditional expectation in regression.** In statistics (regression analysis), the function $\boldsymbol{z} \mapsto \mathbb{E}[g(\boldsymbol{X}, \boldsymbol{Z})|\boldsymbol{Z} = \boldsymbol{z}] := h(\boldsymbol{z})$ is called the **regression function** of $g(\boldsymbol{X}, \boldsymbol{Z})$ on $\boldsymbol{Z}$ at $\boldsymbol{z}$. If $g(\boldsymbol{X}, \boldsymbol{Z}) \in L^2$, then $h(\boldsymbol{z})$ is the best L^2-approximation of $g(\boldsymbol{X}, \boldsymbol{Z})$ with $\boldsymbol{Z} = \boldsymbol{z}$ observed, in terms of mean squared error, as the following result illustrates.

Proposition 8.3.e. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra, and $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Then, $\mathbb{E}[X|\mathcal{G}]$ is the $Y \in L^2(\Omega, \mathcal{G}, \mathbb{P})$ which minimizes $\mathbb{E}\big[(X - Y)^2\big]$, i.e.,

$$\mathbb{E}\big[(X - Y)^2\big] = \min_{Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})} \mathbb{E}\big[(X - Z)^2\big] \quad \text{iff} \quad Y \overset{\text{a.s.}}{=} \mathbb{E}[X|\mathcal{G}].$$

Proof. Note first that

$$\mathbb{E}\big[(X - Y)^2\big] = \mathbb{E}\big[(X - \mathbb{E}[X|\mathcal{G}] + \mathbb{E}[X|\mathcal{G}] - Y)^2\big]$$
$$= \underbrace{\mathbb{E}\big[(X - \mathbb{E}[X|\mathcal{G}])^2\big]}_{\text{does not depend on } Y} + 2\mathbb{E}[(X - \mathbb{E}[X|\mathcal{G}])(\mathbb{E}[X|\mathcal{G}] - Y)] + \underbrace{\mathbb{E}\big[(\mathbb{E}[X|\mathcal{G}] - Y)^2\big]}_{\text{smallest } (= 0) \text{ iff } Y \overset{\text{a.s.}}{=} \mathbb{E}[X|\mathcal{G}] \text{ by [5.1.8]}}.$$

Therefore, it suffices to show that $\mathbb{E}[(X - \mathbb{E}[X|\mathcal{G}])(\mathbb{E}[X|\mathcal{G}] - Y)] = 0$ for every $Y \in L^2(\Omega, \mathcal{G}, \mathbb{P})$:

$$\mathbb{E}[(X - \mathbb{E}[X|\mathcal{G}])(\mathbb{E}[X|\mathcal{G}] - Y)] \overset{\text{(total expectation)}}{=} \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X|\mathcal{G}])(\mathbb{E}[X|\mathcal{G}] - Y)|\mathcal{G}]]$$
$$\overset{\text{(TOWIK)}}{=} \mathbb{E}[(\mathbb{E}[X|\mathcal{G}] - Y)\mathbb{E}[(X - \mathbb{E}[X|\mathcal{G}])|\mathcal{G}]] \overset{\text{(linearity)}}{=} \mathbb{E}[(\mathbb{E}[X|\mathcal{G}] - Y)(\mathbb{E}[X|\mathcal{G}] - \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{G}])]$$
$$\overset{\text{(TOWIK)}}{=} \mathbb{E}[(\mathbb{E}[X|\mathcal{G}] - Y)(\mathbb{E}[X|\mathcal{G}] - \mathbb{E}[X|\mathcal{G}])] = \mathbb{E}[0] = 0.$$

$\square$

More geometrically, the conditional expectation $\mathbb{E}[X|\mathcal{G}]$ can actually be interpreted as the *orthogonal projection* of X onto the closed subspace $L^2(\Omega, \mathcal{G}, \mathbb{P})$ of the Hilbert space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ equipped with the inner product $\langle X, Y \rangle := \mathbb{E}[XY]$. Here, such orthogonal projection is given by

$$\underset{Y \in L^2(\Omega, \mathcal{G}, \mathbb{P})}{\arg\min} \mathbb{E}\big[(X - Y)^2\big] \overset{\text{(Proposition 8.3.e)}}{=} \mathbb{E}[X|\mathcal{G}].$$

143
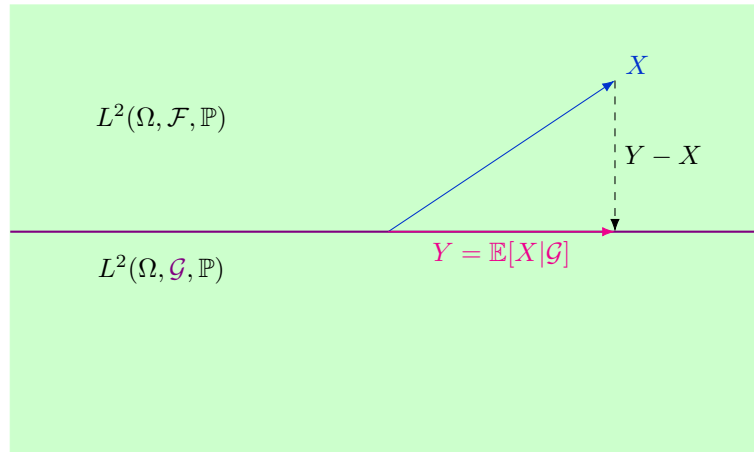
We can verify the orthogonality as follows:

$$\langle Y, Y - X \rangle = \mathbb{E}[Y(Y - X)] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}](\mathbb{E}[X|\mathcal{G}] - X)] = \mathbb{E}\left[\mathbb{E}[X|\mathcal{G}]^2 - X\mathbb{E}[X|\mathcal{G}]\right] = \mathbb{E}\left[\mathbb{E}[X|\mathcal{G}]^2\right] - \mathbb{E}[X\mathbb{E}[X|\mathcal{G}]]$$

$$\overset{\text{(total expectation)}}{=} \mathbb{E}\left[\mathbb{E}[X|\mathcal{G}]^2\right] - \mathbb{E}[\mathbb{E}[X\mathbb{E}[X|\mathcal{G}]|\mathcal{G}]] \overset{\text{(TOWIK)}}{=} \mathbb{E}\left[\mathbb{E}[X|\mathcal{G}]^2\right] - \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbb{E}[X|\mathcal{G}]]$$

$$= \mathbb{E}\left[\mathbb{E}[X|\mathcal{G}]^2\right] - \mathbb{E}\left[\mathbb{E}[X|\mathcal{G}]^2\right] = 0.$$

References

Durrett, R. (2019). *Probability: Theory and examples*. Cambridge University Press.

Embrechts, P., & Hofert, M. (2013). A note on generalized inverses. *Mathematical Methods of Operations Research*, *77*, 423–432.

Folland, G. B. (1999). *Real analysis: Modern techniques and their applications*. John Wiley & Sons.

Kiefer, J., & Wolfowitz, J. (1958). On the deviations of the empiric distribution function of vector chance variables. *Transactions of the American Mathematical Society*, *87*(1), 173–186.

Klenke, A. (2020). *Probability theory: A comprehensive course*. Springer Science & Business Media.

Resnick, S. I. (2014). *A probability path*. Springer Science & Business Media.

Schilling, R. L. (2017). *Measures, integrals and martingales*. Cambridge University Press.

ter Horst, H. J. (1984). Riemann-Stieltjes and Lebesgue-Stieltjes integrability. *The American Mathematical Monthly*, *91*(9), 551–559.

Williams, D. (1991). *Probability with martingales*. Cambridge University Press.

Links to Definitions

Results

Section 1

- [1.1.5]a: interpretation of limit infimum and limit supremum (set)
- [1.1.5]b: relationship between limit infimum and limit supremum (set)
- [1.1.5]c: limits of monotone sequences of sets
- [1.1.9]: properties of preimages
- Theorem 1.2.a: Vitali's theorem
- Theorem 1.2.b: Banach-Tarski paradox
- Proposition 1.3.a: relationships between systems of sets
- [1.4.2]: interpretation of $\sigma(\mathcal{A})$
- Lemma 1.4.a: expressions of σ-algebras generated by partitions
- [1.4.5]: alternative expression of $\sigma(\mathcal{F}_1, \mathcal{F}_2)$
- Proposition 1.4.b: interpretation of product σ-algebra with countably many σ-algebras involved
- Proposition 1.4.c: properties of Dynkin systems
- Theorem 1.4.d: Dynkin's π-λ theorem
- Proposition 1.4.f: $\mathcal{B}(\mathbb{R})$ is generated by all open intervals
- Proposition 1.4.g: generators of $\mathcal{B}(\mathbb{R}^d)$

Section 2

- [2.1.3]: properties of measures
- Proposition 2.1.a: uniqueness of measures
- Lemma 2.2.a: countable union of null sets is still null set
- Theorem 2.2.b: measures can always be completed uniquely
- Lemma 2.3.a: representation of set differences in terms of disjoint union
- Theorem 2.3.b: Carathéodory extension theorem
- [2.4.3]: properties of F-volumes
- Theorem 2.4.a: construction of Borel measures on $\mathbb{R}^d$
- [2.4.6]: properties of the Lebesgue measure
- [2.4.7]: componentwise increasing does not imply d-increasing
- [2.5.2]: properties of probability measures
- Theorem 2.5.a: characterization of probability measures on $\mathbb{R}^d$ by distribution function
- [2.5.6]: formulas for distribution functions
- Lemma 2.5.b: Lipschitz inequality
- Proposition 2.5.c: a distribution function with continuous margins is continuous
- [2.5.8]: an example of continuous distribution function that is not absolutely continuous
- [2.5.10]: properties of generalized inverses

Section 3

- Lemma 3.1.a: commutativity of $\sigma(\cdot)$ and $X^{-1}(\cdot)$

- [3.1.3]a: checking measurability by considering preimages of generator

- [3.1.3]b: composition of measurable functions is measurable

- [3.1.3]c: continuous functions are measurable

- [3.1.3]d: monotone functions are measurable

- [3.1.5]a: random vectors are vectors of random variables

- [3.1.5]b: measurable functions of random vectors are random vectors

- [3.1.5]c: measurability about sequences of random variables

- [3.1.5]d: relationship between completeness and measurability

- [3.2.2]: measurable functions induce measures

- Theorem 3.2.a: characterization of distribution by distribution function

- Proposition 3.2.b: preservation of equality in distribution after applying measurable functions

- [3.4.3]: expression of survival functions in terms of distribution functions

- Theorem 3.5.a: Kolmogorov's extension theorem

- Proposition 3.6.a: quantile transform

- Lemma 3.6.b: strict increasingness of an univariate distribution function on its support

- Proposition 3.6.c: probability transform

Section 4

- Theorem 4.1.a: law of total probability

- Theorem 4.1.b: Bayes' theorem

- [4.2.2]a: independence via generators

- [4.2.2]b: independence of complements

- [4.2.2]c: grouping lemma

- Theorem 4.2.a: Borel-Cantelli lemmas

- Theorem 4.2.b: Kolmogorov's zero-one law

- Theorem 4.2.c: characterization of independence of random variables

- Theorem 4.2.d: functions of independent random variables are independent

- Proposition 4.2.e: construction of independent random variables

- [4.2.8]: formula of distribution function for mixture distribution

- Proposition 4.3.a: analytic characterization of copula

- Theorem 4.3.b: Sklar's theorem

- Theorem 4.3.c: invariance principle for copula

- Corollary 4.3.d: copula from probability transforms

- [4.3.7]: interpretations of independence, comonotone, and countermonotone copulas

- Theorem 4.3.e: Fréchet-Hoeffding bounds

Section 5

- [5.1.4]: properties of Lebesgue integral of simple function

- Proposition 5.1.a: approximating sequence for function in L_+

- Lemma 5.1.b: monotonicity and scaling for Lebesgue integral of nonnegative measurable function

- Theorem 5.1.c: monotone convergence theorem

- [5.1.8]: properties of Lebesgue integral of nonnegative function

- [5.1.10]: properties of Lebesgue integral of quasi-integrable function

- [5.1.11]a: σ-additivity of Lebesgue integral of measurable function

- [5.1.11]b: a.e. finiteness of integrable functions

- [5.1.11]c: equivalent criteria for a.e. equality of integrable functions

- Lemma 5.1.d: Fatou's lemma

- Theorem 5.1.e: dominated convergence theorem

- Corollary 5.1.f: commutativity of integral and countable sum for Lebesgue integral of measurable function

- [5.1.14]a: Hölder inequality (or Cauchy-Swartz inequality for $p = q = 2$)

- [5.1.14]b: Minkowski's inequality

- [5.1.14]c: Jensen's inequality

- [5.1.14]d: relationship between L^p spaces and norms

- Theorem 5.2.a: change of variables formula

- Theorem 5.2.c: Fubini-Tonelli theorem

- Proposition 5.2.d: expectation of product is product of expectations under independence

- Proposition 5.2.e: expectation formula in terms of quantile/survival function

- [5.3.3]: properties of variance, covariance, and correlation

- Proposition 5.3.a: Cauchy-Swartz inequality and correlation bounds

- Lemma 5.3.b: Hoeffding's lemma

- [5.4.3]a: linearity of mean vectors

- [5.4.3]b: effects of linear operations on cross-covariance matrix

- [5.4.3]c: covariance matrix of sums

- Lemma 5.4.a: Cholesky decomposition

- Proposition 5.4.b: characterization of covariance matrix

- Lemma 5.5.a: preservation of domination/singularity upon summation

- Lemma 5.5.b: relationship between finite measures

- Theorem 5.5.c: Lebesgue-Radon-Nikodym theorem

- Corollary 5.5.d: Radon-Nikodym theorem

- [5.5.5]a: "integration by substitution" for Radon-Nikodym derivatives

- [5.5.5]b: "chain rule" for Radon-Nikodym derivatives

Section 6

- Lemma 6.1.a: interchanges of (limit) supremum with other operations

- Theorem 6.1.b: characterization of almost sure convergence by convergence in probability

- [6.1.4]a: complete convergence implies almost sure convergence

- [6.1.4]b: almost sure convergence implies convergence in probability

- [6.1.4]c: monotone convergence in probability implies almost sure convergence

- Theorem 6.1.c: subsequence principle

- Corollary 6.1.d: dominated convergence theorem for convergence in probability

- Theorem 6.1.e: continuous mapping theorem for almost sure convergence and convergence in probability

- Lemma 6.2.a: tail probability bounds

- Proposition 6.2.b: properties of convergence in L^p

- [6.3.2]: uniqueness of limiting distribution for convergence in distribution

- Theorem 6.3.a: Portmanteau theorem

- [6.3.4]: properties of convergence in distribution

- Theorem 6.3.b: continuous mapping theorem for convergence in distribution

- [6.4.3]: properties of uniform integrability

- [6.4.4]: characterization of uniform integrability

- Lemma 6.4.b: limiting behaviour of integral over sets with probabilities converging to zero

- Theorem 6.4.c: convergence in probability with uniform integrability implies convergence in L^p

- Theorem 6.5.a: condition for joint convergence in distribution

- Theorem 6.5.b: Slutsky's theorem

- Lemma 6.6.a: independent two-point distributions used for counterexamples

- Proposition 6.7.a: convergence in quantile function

- Theorem 6.7.b: laws of large numbers

- Theorem 6.7.c: Glivenko-Cantelli theorem

Section 7

- [7.1.2]: properties of characteristic function

- Theorem 7.1.a: Lévy's continuity theorem

- Theorem 7.1.b: Cramér-Wold device

- Corollary 7.1.c: criterion for equality in distribution about linear combinations

- Theorem 7.1.d: characteristic function characterizes distribution

- Equation (10): inversion formula for characteristic function

- Corollary 7.1.e: criteria for the realness of characteristic function

- Proposition 7.1.f: characterization of multivariate normal distribution

- Corollary 7.1.g: properties of (multivariate) normal distribution

- Lemma 7.2.a: convergence to exponential

- Lemma 7.2.b: expansion of characteristic function about zero

- Theorem 7.2.c: central limit theorem

- [7.2.3]a: multivariate CLT

- [7.2.3]b: Lindeberg-Feller CLT

Section 8

- Proposition 8.2.a: integrability of conditional expectation

- Theorem 8.2.b: existence and a.s. uniqueness of conditional expectation

- Proposition 8.2.c: formula of conditional expectation for $\mathcal{G}$ generated by a countable partition

- Proposition 8.2.d: properties of conditional expectation

- [8.2.6]a: conditional Hölder inequality (or conditional Cauchy-Swartz inequality for $p = q = 2$)

- [8.2.6]b: conditional Minkowski's inequality

- [8.2.6]c: conditional Jensen's inequality

- [8.2.7]a: contraction in L^p for conditional expectation

- [8.2.7]b: convergence in L^p for conditional expectation

- Proposition 8.2.e: "removing" σ-algebra in conditional expectation under certain independence

- Lemma 8.2.f: verifying partial averaging on a π-system satisfying some conditions

- Proposition 8.2.g: taking out what is known (TOWIK)

- Corollary 8.2.h: conditional version of "expectation of product is product of expectations"

- Theorem 8.2.i: factorization/Doob-Dynkin lemma

- Corollary 8.2.j: factorization for conditional expectation given $\boldsymbol{Z}$

- Proposition 8.2.k: explicit formula for factorization under independence

- Proposition 8.3.a: conditional distribution formula

- Lemma 8.3.b: conditional version of "2nd moment minus 1st moment squared"

- Proposition 8.3.c: law of total variance

- Proposition 8.3.d: expectation and variance of random sum

- Proposition 8.3.e: conditional expectation as the best L^2 approximation