

---

# Introduction to Analysis and Probability

## for Applied Mathematicians

---

Niushan Gao

*Webpage:* <https://math.ryerson.ca/~niushan/>

*E-mail:* [niushan@ryerson.ca](mailto:niushan@ryerson.ca)

Foivos Xanthos

*Webpage:* <https://math.ryerson.ca/~foivos/>

*E-mail:* [foivos@ryerson.ca](mailto:foivos@ryerson.ca)

Department of Mathematics

Ryerson University

Canada

July 20, 2020



## Contents

Notation and Terminology	1
Chapter 1. Measurable Sets	3
Chapter 2. Measures	15
Chapter 3. Lebesgue-Stieltjes Measures	23
Chapter 4. Random Variables	35
Chapter 5. Expectations I	45
Chapter 6. Expectations II	59
Chapter 7. Product Measures	71
Chapter 8. Distributions	83
Chapter 9. Independence	97
Chapter 10. Law of Large Numbers	107
Chapter 11. Characteristic Functions	119
Chapter 12. Central Limit Theorem	121
Chapter 13. Conditional Distribution	123
Chapter 14. Conditional Expectation	125
Appendix.	131



# Notation and Terminology

## 1. Numbers

## 2. Sets

## 3. Functions

union over empty index

sum over empty index

if  $f_n \uparrow f$  then  $\{f_n > c\} \uparrow \{f > c\}$

expand sets for measurable functions

convergence, Cauchy

0.1. Show that  $(X+Y)^- \leq X^- + Y^-$ ,  $(X+Y)^+ \leq X^+ + Y^+$ . If  $X \leq Y$ , then  $X^+ \leq Y^+$  and  $X^- \geq Y^-$ .

0.1. EXAMPLE. A monotone function has at most countable discontinuity.



## CHAPTER 1

### Measurable Sets

At any given time point, we are interested in knowing what the future will be at a later time point. But the future is full of uncertainties: tomorrow it may rain or may not rain; the lottery ticket you are buying now may win or may not win. Just like flipping a coin, we will not know which scenario would eventually become true, until the coin has been flipped. Thus, in the presence of uncertainties, instead of asking what the future will be, we shall ask what are the possible scenarios and what are their chances to become true in the future. This leads us to the realm of Probability Theory. Modern Probability Theory is built over a triple  $(\Omega, \mathcal{F}, \mathbb{P})$ , where  $\Omega$  is the collection of all possible scenarios for the future,  $\mathcal{F}$  collects all the events that are of interest, and  $\mathbb{P}$  tells the probability of each event in  $\mathcal{F}$ . We start with exploring  $\mathcal{F}$  in this chapter.

#### 1. Definition and basic properties

Suppose that we are going to conduct an experiment of flipping a coin *three times*. We use  $H$  and  $T$  to denote head and tail, respectively. Then we are facing eight possible outcomes in the future:  $HHH, HHT, HTH, HTT, THH, THT, TTH, TTT$ . We collect them together and denote it by a set

$$(1.1) \quad \Omega := \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Consider the event that the first flip is  $H$ . It means precisely that if we have conducted the experiment, then our final outcome would be one of the following four:  $HHH, HHT, HTH, HTT$ . We may thus use the set

$$F_H := \{HHH, HHT, HTH, HTT\}$$

to denote the event that the first flip is  $H$ . Other events can be similarly identified as subsets of  $\Omega$  as well. For example, we identify the event that the second flip is  $H$  with the set  $S_H := \{HHT, HHH, THH, THT\}$ .

Given that we use subsets of  $\Omega$  to denote events, how do we formulate the occurrence of an event mathematically? Say, we have conducted the experiment of flipping the coin three times, and the final outcome is  $\omega$

(which of course is an element in  $\Omega$ ). The event that the first flip is  $H$  has occurred means precisely that the realized outcome  $\omega$  is one of the four:  $HHH, HHT, HTH, HTT$ . Thus the event  $F_H$  occurs at a realization  $\omega$  iff

$$\omega \in F_H.$$

Let's now consider the collection of *all* interesting events. Say, suppose that, for some reasons, we care about only the *first two* flips. Take any event  $E$  in the collection of interesting events. At an arbitrary realization  $\omega$ ,  $E$  occurs iff  $\omega \in E$  iff  $\omega \notin E^c$  iff  $E^c$  does not occur. That is,  $E^c$  is the contrary of  $E$ . Intuitively, we shall care about the contrary of an interesting event. Thus  $E^c$  should also be included in our collection of interesting events. For example, if  $E = S_H$ , then its complement  $S_T := \{HTH, HTT, TTH, TTT\}$  is the event that the second flip is  $T$ , which is of course interesting. We may also look at multiple interesting events together. Let's take two interesting events  $E$  and  $F$ . At an arbitrary realization  $\omega$ , the intersection  $E \cap F$  occurs iff  $\omega \in E \cap F$  iff  $\omega \in E$  and  $\omega \in F$  iff both  $E$  and  $F$  occur. Intuitively, we shall care about the simultaneous occurrence of two interesting events. Thus  $E \cap F$  should lie in our collection of interesting events as well. For example, for the event  $F_H$  that the first flip is  $T$  and the event  $S_T$  that the second flip is  $T$ , the intersection  $F_H \cap S_T = \{HTH, HTT\}$  is precisely the event that the first flip is  $H$  and the second flip is  $T$ , which is again obviously interesting to us as we care about the first two flips.

The above discussions motivate us to conclude that our collection of interesting events should be closed under taking complementation and intersection. This leads us to the following notion.

**1.1. DEFINITION.** *Let  $\Omega$  be a non-empty set. Let  $\mathcal{F}$  be a collection of subsets of  $\Omega$ . We say that  $\mathcal{F}$  is a  $\sigma$ -**algebra** over  $\Omega$  if it satisfies the following conditions:*

- (a)  $\mathcal{F}$  has at least one member;
- (b) if  $E \in \mathcal{F}$ , then  $E^c \in \mathcal{F}$
- (c) if  $(E_n)_{n \in \mathbb{N}}$  is a sequence in  $\mathcal{F}$ , then  $\bigcap_{n=1}^{\infty} E_n \in \mathcal{F}$ .

Members in  $\mathcal{F}$  are also called  $\mathcal{F}$ -**measurable sets**, or simply, **measurable sets** if there is no ambiguity about the  $\sigma$ -algebra in question. In probabilistic terms,  $\Omega$  is usually called the **sample space** and members in  $\mathcal{F}$  are called **events**.



Condition (c) means that if we are interested in any given *countably infinite* many events, then we are interested in their simultaneous occurrence. A quick observation regarding this condition is the following.

1.2. PROPOSITION. *Assuming Conditions (a) and (b) in Definition 1.1, Condition (c) is equivalent to the following:*

(c') *if  $(E_n)_{n \in \mathbb{N}}$  is a sequence in  $\mathcal{F}$ , then  $\bigcup_{n=1}^{\infty} E_n \in \mathcal{F}$ .*

PROOF. Assume that Conditions (a) and (b) are satisfied by  $\mathcal{F}$ . The proof is a simple application of De Morgan's Laws.

Suppose first that (c) is satisfied by  $\mathcal{F}$ . Take any sequence  $(E_n)_{n \in \mathbb{N}}$  in  $\mathcal{F}$ . Then  $E_n^c \in \mathcal{F}$  for each  $n \in \mathbb{N}$  by Condition (b). Thus  $\bigcap_{n=1}^{\infty} E_n^c \in \mathcal{F}$  by (c). It follows from Condition (b) again that

$$\bigcup_{n=1}^{\infty} E_n = \left( \bigcap_{n=1}^{\infty} E_n^c \right)^c \in \mathcal{F}.$$

This proves that (c)  $\implies$  (c'). The reverse implication (c')  $\implies$  (c) can be proved similarly by noticing  $\bigcap_{n=1}^{\infty} E_n = (\bigcup_{n=1}^{\infty} E_n^c)^c$ .  $\square$

Some other basic properties of  $\sigma$ -algebras are listed below.

1.3. PROPOSITION. *Let  $\mathcal{F}$  be a  $\sigma$ -algebra over  $\Omega$ . The following hold.*

- (a)  $\emptyset, \Omega \in \mathcal{F}$ ;
- (b) *If  $E_1, \dots, E_n \in \mathcal{F}$ , then  $\bigcap_{k=1}^n E_k \in \mathcal{F}$ ;*
- (c) *If  $E_1, \dots, E_n \in \mathcal{F}$ , then  $\bigcup_{k=1}^n E_k \in \mathcal{F}$ ;*
- (d) *If  $E, F \in \mathcal{F}$ , then  $E \setminus F \in \mathcal{F}$ .*

PROOF. (a). By Definition 1.1 (a),  $\mathcal{F}$  has at least one member, say,  $E$ . Then by Definition 1.1 (b),  $E^c \in \mathcal{F}$ . Consider the sequence  $E^c, E, E, \dots$  in  $\mathcal{F}$ . Clearly, the intersection is  $\emptyset$  and lies in  $\mathcal{F}$  by Definition 1.1 (c); the union is  $\Omega$  and lies in  $\mathcal{F}$  by Proposition 1.2.

(b). Suppose  $E_1, \dots, E_n \in \mathcal{F}$ . Put  $E_k = \Omega$  for each  $k \geq n+1$ . Then  $\bigcap_{k=1}^n E_k = \bigcap_{k=1}^{\infty} E_k \in \mathcal{F}$  by Definition 1.1 (c). (c) can be proved similarly by setting  $E_k = \emptyset$  for each  $k \geq n+1$  and using Proposition 1.2.

(d). Suppose  $E, F \in \mathcal{F}$ . Then  $F^c \in \mathcal{F}$  by Definition 1.1 (b), and thus  $E \setminus F = E \cap F^c \in \mathcal{F}$  by (b) that we have just proved.  $\square$

1.1. EXAMPLE. Let  $\Omega$  be given in (1.1). Then  $\{\emptyset, \Omega, F_H, F_T\}$  is a  $\sigma$ -algebra over  $\Omega$ .

1.2. EXAMPLE. Let  $\Omega$  be any non-empty set. Its power set  $\mathcal{P}(\Omega)$  is a  $\sigma$ -algebra over  $\Omega$ .

## 2. Generated $\sigma$ -algebras

Sometimes we start with a collection of interesting events that is not a  $\sigma$ -algebra yet. In this case, we find the “smallest”  $\sigma$ -algebra enveloping it.

1.3. EXAMPLE. Let  $\Omega$  be given in (1.1). Consider the collection  $\mathcal{C} = \{F_H, S_T\}$ . Suppose  $\mathcal{F}$  is a  $\sigma$ -algebra and  $\mathcal{C} \subset \mathcal{F}$ .

The events in  $\mathcal{C}$  involves only the *first two* flips.

By Definition 1.1 (b), clearly  $F_T, S_H \in \mathcal{F}$ . Thus all the possible events resulting from the first flip,  $F_H$  and  $F_T$ , lie in  $\mathcal{F}$ , and also all the possible events resulting from the second flip,  $S_H$  and  $S_T$ , lie in  $\mathcal{F}$ . Using them, we conclude that all the “building-blocks” events from the first two flips all lie in  $\mathcal{F}$ :

- (a)  $F_H \cap S_H = \{HHH, HHT\}$ ; the first two flips are both  $H$ ;
- (b)  $F_H \cap S_T = \{HTH, HTT\}$ ; the first flip is  $H$  and the second flip is  $T$ ;
- (c)  $F_T \cap S_H = \{THH, THT\}$ ; the first flip is  $T$  and the second flip is  $H$ ;
- (d)  $F_T \cap S_T = \{TTH, TTT\}$ ; the first two flips are both  $T$ .

We can now produce all other events that must lie in  $\mathcal{F}$  by taking unions of these “building blocks”. Precisely,

- (a) taking union of no building blocks yields  $\emptyset$ ;
- (b) taking union of exactly one building block yields the four building blocks;
- (c) taking union of exactly two building blocks yield
  - $(F_H \cap S_H) \cup (F_H \cap S_T) = F_H$ ,
  - $(F_T \cap S_H) \cup (F_T \cap S_T) = F_T$ ,
  - $(F_H \cap S_H) \cup (F_T \cap S_H) = S_H$ ,
  - $(F_H \cap S_T) \cup (F_T \cap S_T) = S_T$ ,
  - $(F_H \cap S_H) \cup (F_T \cap S_T)$ ,
  - $(F_H \cap S_T) \cup (F_T \cap S_H)$ ;
- (d) taking union of exactly three building blocks yield
  - $(F_H \cap S_H) \cup (F_H \cap S_T) \cup (F_T \cap S_H)$ ,
  - $(F_H \cap S_H) \cup (F_H \cap S_T) \cup (F_T \cap S_T)$ ,
  - $(F_H \cap S_H) \cup (F_T \cap S_H) \cup (F_T \cap S_T)$ ,
  - $(F_H \cap S_T) \cup (F_T \cap S_H) \cup (F_T \cap S_T)$ ;
- (e) taking union of exactly four building blocks yield  $\Omega$ .

Denote by  $\sigma(\mathcal{C})$  the collection of all the sixteen events above. One sees that it is a  $\sigma$ -algebra containing  $\mathcal{C}$  as a subset. On the other hand,  $\sigma(\mathcal{C})$  is the smallest  $\sigma$ -algebra containing  $\mathcal{C}$  as a subset, in the sense that if  $\mathcal{F}$  is any other  $\sigma$ -algebra containing  $\mathcal{C}$  as a subset, then  $\sigma(\mathcal{C}) \subset \mathcal{F}$ .

In general, for any non-empty collection  $\mathcal{C}$  of subsets of  $\Omega$ , we can find the smallest  $\sigma$ -algebra enveloping it. Recall first that there is at least one  $\sigma$ -algebra containing  $\mathcal{C}$  as a subset:  $\mathcal{P}(\Omega)$ .

1.4. PROPOSITION. *Let  $\mathcal{C}$  be a non-empty collection of subsets of  $\Omega$ . Let  $\{\mathcal{F}_\lambda\}_{\lambda \in \Lambda}$  be the collection of all  $\sigma$ -algebras over  $\Omega$  containing  $\mathcal{C}$  as a subset. Put*

$$\sigma(\mathcal{C}) := \bigcap_{\lambda \in \Lambda} \mathcal{F}_\lambda.$$

*Then  $\sigma(\mathcal{C})$  is a  $\sigma$ -algebra over  $\Omega$  containing  $\mathcal{C}$  as a subset. Moreover, if  $\mathcal{F}$  is any  $\sigma$ -algebra over  $\Omega$  containing  $\mathcal{C}$  as a subset, then  $\sigma(\mathcal{C}) \subset \mathcal{F}$ .*

PROOF. By Proposition 1.3,  $\emptyset \in \mathcal{F}_\lambda$  for each  $\lambda \in \Lambda$ , so that  $\emptyset \in \bigcap_{\lambda \in \Lambda} \mathcal{F}_\lambda = \sigma(\mathcal{C})$ . Thus Condition (a) in Definition 1.1 is verified.

Take any set  $E \in \sigma(\mathcal{C})$ . Then for any  $\lambda \in \Lambda$ ,  $E \in \mathcal{F}_\lambda$ , and since  $\mathcal{F}_\lambda$  is a  $\sigma$ -algebra,  $E^c \in \mathcal{F}_\lambda$  as well. It follows that  $E^c \in \bigcap_{\lambda \in \Lambda} \mathcal{F}_\lambda$ . Thus Condition (b) in Definition 1.1 is verified.

Let  $(E_n)_{n \in \mathbb{N}}$  be a sequence in  $\sigma(\mathcal{C}) = \bigcap_{\lambda \in \Lambda} \mathcal{F}_\lambda$ . Then for each  $n \in \mathbb{N}$  and  $\lambda \in \Lambda$ ,  $E_n \in \mathcal{F}_\lambda$ . Since  $\mathcal{F}_\lambda$  is a  $\sigma$ -algebra,  $\bigcap_{n=1}^\infty E_n \in \mathcal{F}_\lambda$  for each  $\lambda \in \Lambda$ . Thus  $\bigcap_{n=1}^\infty E_n \in \bigcap_{\lambda \in \Lambda} \mathcal{F}_\lambda$ , and Condition (c) in Definition 1.1 is verified.

Finally, if  $\mathcal{F}$  is any  $\sigma$ -algebra containing  $\mathcal{C}$  as a subset, then  $\mathcal{F} = \mathcal{F}_{\lambda_0}$  for some  $\lambda_0 \in \Lambda$ . Thus  $\sigma(\mathcal{C}) = \bigcap_{\lambda \in \Lambda} \mathcal{F}_\lambda \subset \mathcal{F}_{\lambda_0} = \mathcal{F}$ .  $\square$

From now on, we call  $\sigma(\mathcal{C})$  the  **$\sigma$ -algebra generated** by  $\mathcal{C}$ . The members of  $\mathcal{C}$  are called **generators** of  $\sigma(\mathcal{C})$ .

1.5. REMARK. The last assertion in Proposition 1.4 shall be read as follows: in order for a  $\sigma$ -algebra  $\mathcal{F}$  to contain a generated  $\sigma$ -algebra  $\mathcal{G}$  as a subset, it is enough to ensure that  $\mathcal{F}$  contains all the generators of  $\mathcal{G}$ .

We are ready to introduce special  $\sigma$ -algebras over  $\mathbb{R}^d$ .

1.4. EXAMPLE. The following hold.

$$\begin{aligned} & \sigma(\{(a, b] : a, b \in \mathbb{R}, a < b\}) \\ &= \sigma(\{(-\infty, a] : a \in \mathbb{R}\}) \\ &= \sigma(\{[a, b] : a, b \in \mathbb{R}, a < b\}). \end{aligned}$$

Let's denote the  $\sigma$ -algebras by  $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ , respectively. For any  $a, b \in \mathbb{R}$  with  $a < b$ ,

$$(a, b] = (-\infty, b] \setminus (-\infty, a] \in \mathcal{F}_2.$$

Thus  $\mathcal{F}_2$  contains all the generators of  $\mathcal{F}_1$ , and by Remark 1.5,

$$\mathcal{F}_2 \supset \mathcal{F}_1.$$

Similarly, for any  $a \in \mathbb{R}$ ,

$$(-\infty, a] = \bigcup_{n \in \mathbb{N}} [a - n, a] \in \mathcal{F}_3,$$

so that  $\mathcal{F}_3$  contains all the generators of  $\mathcal{F}_2$ , and

$$\mathcal{F}_3 \supset \mathcal{F}_2.$$

For any  $a, b \in \mathbb{R}$  with  $a < b$ ,

$$[a, b] = \bigcap_{n \in \mathbb{N}} \left( a - \frac{1}{n}, b \right] \in \mathcal{F}_1,$$

so that  $\mathcal{F}_1$  contains all the generators of  $\mathcal{F}_3$ , and

$$\mathcal{F}_1 \supset \mathcal{F}_3.$$

Combining the above, we get the desired equalities.

**1.6. DEFINITION.** We denote the  $\sigma$ -algebra in the preceding example by  $\mathcal{B}$  and call it the (one-dimensional) **Borel algebra**. Every element in  $\mathcal{B}$  is called a **Borel set**.

See Exercise 1.4 for more equivalent characterizations of  $\mathcal{B}$ .

**1.5. EXAMPLE.** The following hold.

- (a)  $\{a\} \in \mathcal{B}$  for any  $a \in \mathbb{R}$ . Indeed,  $\{a\} = \bigcap_{n \in \mathbb{N}} \left( a - \frac{1}{n}, a \right] \in \mathcal{B}$ .
- (b) If  $A$  is a finite or countably infinite subset of  $\mathbb{R}$ , then  $A \in \mathcal{B}$ . Indeed,  $A$  can be expressed as a finite union or a countably infinite union of singletons; apply Propositions 1.2 and 1.3.

The higher-dimensional Borel algebras can be defined similarly.

**1.7. DEFINITION.** Let  $d \in \mathbb{N}$ . We denote by  $\mathcal{B}^d$  the  $\sigma$ -algebra generated by the collection of all bounded, left-open, right closed cubes  $\prod_{k=1}^d (a_k, b_k]$ , where  $a_k, b_k \in \mathbb{R}$  and  $a_k < b_k$ ,  $k = 1, \dots, d$ , and call it the  $d$ -dimensional Borel algebra. Elements in  $\mathcal{B}^d$  are also called Borel sets.

1.6. EXAMPLE. The following hold.

$$\mathcal{B}^d = \sigma\left(\left\{\prod_{k=1}^d (-\infty, a_k] : a_k \in \mathbb{R}, k = 1, \dots, d\right\}\right).$$

We can argue similarly as in Example 1.4; the reader may take  $d = 2$  and draw graphs to see the arguments below visually. Denote the  $\sigma$ -algebra in the right hand side by  $\mathcal{F}$ . Let  $a_k \in \mathbb{R}$ ,  $k = 1, \dots, d$ , be arbitrary. Note that

$$\prod_{k=1}^d (-\infty, a_k] = \bigcup_{n \in \mathbb{N}} \left( \prod_{k=1}^d (a_k - n, a_k] \right) \in \mathcal{B}^d.$$

Thus by Remark 1.5 again,

$$\mathcal{B}^d \supset \mathcal{F}.$$

For the reverse inclusion, we work coordinate by coordinate. Let  $a_k, b_k \in \mathbb{R}$  with  $a_k < b_k$ ,  $k = 1, \dots, d$ , be arbitrary. Then

$$(a_1, b_1] \times \prod_{k=2}^d (-\infty, b_k] = \left( \prod_{k=1}^d (-\infty, b_k] \right) \setminus \left( (-\infty, a_1] \times \prod_{k=2}^d (-\infty, b_k] \right) \in \mathcal{F}.$$

Thus

$$\begin{aligned} & \prod_{k=1}^2 (a_k, b_k] \times \prod_{k=3}^d (-\infty, b_k] \\ &= \left( (a_1, b_1] \times \prod_{k=2}^d (-\infty, b_k] \right) \setminus \left( (a_1, b_1] \times (-\infty, a_2] \times \prod_{k=3}^d (-\infty, b_k] \right) \in \mathcal{F}. \end{aligned}$$

Repeating this process, one gets that

$$\prod_{k=1}^d (a_k, b_k] \in \mathcal{F}.$$

Therefore, by Remark 1.5,

$$\mathcal{B}^d \subset \mathcal{F}.$$

Combining the above, we get the desired equality.

See Exercise 1.5 for more equivalent characterizations of  $\mathcal{B}^d$ .

### 3. Monotone class theorem

It is generally extremely difficult to figure out *all* the elements in a generated  $\sigma$ -algebra. A general approach to get around this difficulty is to study another collection of subsets of  $\Omega$  that contains the generators of the  $\sigma$ -algebra in question but satisfies some other properties and then compare this collection with the generated  $\sigma$ -algebra. We present a monotone class theorem in this spirit. It will be used later a few times.

To this end, we introduce two new notions.

1.8. DEFINITION. A non-empty collection  $\mathcal{P}$  of subsets of  $\Omega$  is called a  **$\pi$ -system** over  $\Omega$  if  $E \cap F \in \mathcal{P}$  whenever  $E, F \in \mathcal{P}$ .

$\pi$ -systems usually have relatively simpler structures.

1.7. EXAMPLE. The following collections of generators for the Borel algebra  $\mathcal{B}^d$  are both  $\pi$ -systems.

$$\mathcal{P}_1 = \{\emptyset\} \cup \left\{ \prod_{k=1}^d (a_k, b_k] : a_k, b_k \in \mathbb{R}, a_k < b_k, k = 1, \dots, d \right\},$$

$$\mathcal{P}_2 = \left\{ \prod_{k=1}^d (-\infty, a_k] : a_k \in \mathbb{R}, k = 1, \dots, d \right\}.$$

1.9. DEFINITION. A collection  $\mathcal{D}$  of subsets of  $\Omega$  is called a  **$\lambda$ -system** or **Dynkin system** over  $\Omega$  if it satisfies the following conditions:

- (a)  $\emptyset \in \mathcal{D}$ ;
- (b)  $E^c \in \mathcal{D}$  whenever  $E \in \mathcal{D}$ ;
- (c)  $\bigcup_{n \in \mathbb{N}} E_n \in \mathcal{D}$  whenever  $(E_n)_{n \in \mathbb{N}}$  is a disjoint sequence in  $\mathcal{D}$ .

The definition of  $\lambda$ -systems only “slightly” differs from that of  $\sigma$ -algebras in the third condition. A  $\sigma$ -algebra is obviously a  $\lambda$ -system but a  $\lambda$ -system need not be a  $\sigma$ -algebra (Exercise 1.12).

The monotone class theorem is stated as follows.

1.10. THEOREM. Let  $\mathcal{P}$  be a  $\pi$ -system over  $\Omega$  and  $\mathcal{D}$  be a  $\lambda$ -system over  $\Omega$  such that  $\mathcal{P} \subset \mathcal{D}$ . Then  $\sigma(\mathcal{P}) \subset \mathcal{D}$ .

For the proof, we need to exploit the notion of generated  $\lambda$ -systems.

1.1. LEMMA. Let  $\mathcal{C}$  be a non-empty collection of subsets of  $\Omega$ . Let  $\{\mathcal{D}_\lambda\}_{\lambda \in \Lambda}$  be the collection of all  $\lambda$ -systems over  $\Omega$  containing  $\mathcal{C}$  as a subset. Then

$$\mathcal{D}(\mathcal{C}) := \bigcap_{\lambda \in \Lambda} \mathcal{D}_\lambda$$

is a  $\lambda$ -system over  $\Omega$  containing  $\mathcal{C}$  as a subset. Moreover, if  $\mathcal{D}$  is any  $\lambda$ -system over  $\Omega$  containing  $\mathcal{C}$  as a subset, then  $\mathcal{D}(\mathcal{C}) \subset \mathcal{D}$ .

Its proof is straightforward verification and is similar to that of Proposition 1.4. We leave it to the reader (Exercise 1.9).

PROOF OF THEOREM 1.10. We begin with a few deductions to simpler assertions. First, by the minimality of generated  $\lambda$ -systems,  $\mathcal{D}(\mathcal{P}) \subset \mathcal{D}$ . Thus it suffices to prove that

$$\sigma(\mathcal{P}) \subset \mathcal{D}(\mathcal{P}).$$

Second, since  $\mathcal{P} \subset \mathcal{D}(\mathcal{P})$ , by the minimality of generated  $\sigma$ -algebras it is enough to show that  $\mathcal{D}(\mathcal{P})$  is a  $\sigma$ -algebra. Finally, note that a  $\lambda$ -system that is also a  $\pi$ -system is a  $\sigma$ -algebra (Exercise 1.7). Thus it only remains to be shown that the generated  $\lambda$ -system  $\mathcal{D}(\mathcal{P})$  is also a  $\pi$ -system.

Consider the collection of all sets whose intersections with every member in  $\mathcal{D}(\mathcal{P})$  remain in  $\mathcal{D}(\mathcal{P})$ :

$$\mathcal{D}_1 := \{E \subset \Omega : E \cap D \in \mathcal{D}(\mathcal{P}) \text{ for every } D \in \mathcal{D}(\mathcal{P})\}.$$

Then  $\mathcal{D}(\mathcal{P})$  being a  $\pi$ -system is clearly equivalent to  $\mathcal{D}(\mathcal{P}) \subset \mathcal{D}_1$ . Using the minimality of generated  $\lambda$ -systems again, we only need to show the following two assertions:

- (a)  $\mathcal{P} \subset \mathcal{D}_1$ ;
- (b)  $\mathcal{D}_1$  is a  $\lambda$ -system.

We start with verifying the second assertion. By Definition 1.9 (a),  $\emptyset \cap D = \emptyset \in \mathcal{D}(\mathcal{P})$  for any  $D \in \mathcal{D}(\mathcal{P})$ . Thus  $\emptyset \in \mathcal{D}_1$ . Next, take any  $E \in \mathcal{D}_1$  and any  $D \in \mathcal{D}(\mathcal{P})$ . Then  $E \cap D \in \mathcal{D}(\mathcal{P})$  by Definition of  $\mathcal{D}_1$ , and  $D^c \in \mathcal{D}(\mathcal{P})$  by Definition 1.9 (b). Consider the disjoint sequence  $E \cap D, D^c, \emptyset, \emptyset, \dots$  in  $\mathcal{D}(\mathcal{P})$ . By Definition 1.9 (c), their union, which is  $(E \cap D) \cup D^c$ , lies in  $\mathcal{D}(\mathcal{P})$ . Thus by Definition 1.9 (b) again,

$$E^c \cap D = ((E \cap D) \cup D^c)^c \in \mathcal{D}(\mathcal{P}).$$

It follows that  $E^c \in \mathcal{D}_1$ . Finally, let  $(E_n)_{n \in \mathbb{N}}$  be a disjoint sequence in  $\mathcal{D}_1$ . Then for any  $D \in \mathcal{D}(\mathcal{P})$ ,  $(E_n \cap D)_{n \in \mathbb{N}}$  is a disjoint sequence in  $\mathcal{D}(\mathcal{P})$ . By Definition 1.9 (c),

$$\left( \bigcup_{n=1}^{\infty} E_n \right) \cap D = \bigcup_{n=1}^{\infty} (E_n \cap D) \in \mathcal{D}(\mathcal{P}).$$

It follows that  $\bigcup_{n=1}^{\infty} E_n \in \mathcal{D}_1$ , completing the proof of the second assertion.

The first assertion cannot be verified directly. Instead, consider the following collection:

$$\mathcal{D}_2 := \{E \subset \Omega : E \cap P \in \mathcal{D}(\mathcal{P}) \text{ for every } P \in \mathcal{P}\}.$$

Since  $\mathcal{P}$  is a  $\pi$ -system, it is easy to see that  $\mathcal{P} \subset \mathcal{D}_2$ . Along the same lines as for  $\mathcal{D}_1$ , one also sees that  $\mathcal{D}_2$  is a  $\lambda$ -system. Therefore,  $\mathcal{D}(\mathcal{P}) \subset \mathcal{D}_2$ . That is, for any  $D \in \mathcal{D}(\mathcal{P})$ ,  $D \cap P \in \mathcal{D}(\mathcal{P})$  for every  $P \in \mathcal{P}$ . This can be restated as: for any  $P \in \mathcal{P}$ ,  $P \cap D \in \mathcal{D}(\mathcal{P})$  for every  $D \in \mathcal{D}(\mathcal{P})$ . Hence, if  $P \in \mathcal{P}$ , then  $P \in \mathcal{D}_1$ . This proves the first assertion and hence the theorem.  $\square$

This theorem is also called Dynkin's  $\pi$ - $\lambda$  Theorem.

### Exercises

1.1. Let  $\mathcal{F}$  be a  $\sigma$ -algebra over a set  $\Omega$  and  $E$  be a non-empty set in  $\mathcal{F}$ . Then

$$\mathcal{F}_{|E} := \{F : F \in \mathcal{F}, F \subset E\}$$

is a  $\sigma$ -algebra over  $E$ . Note that the universal set for  $\mathcal{F}_{|E}$  is  $E$ , not  $\Omega$ .

1.2. Let  $\Omega$  be as in (1.1). Let  $T_H$  be the event that the third flip is  $H$ . Show that  $\mathcal{P}(\Omega) = \sigma(\{F_H, S_H, T_H\})$ .

1.3. Let  $\Omega$  be a non-empty set. Let  $\{A_n\}_{n \in \mathbb{N}}$  be a given partition of  $\Omega$ , i.e.,  $\Omega = \cup_{n \in \mathbb{N}} A_n$  and  $A_j \cap A_k = \emptyset$  for any distinct  $j, k$  in  $\mathbb{N}$ . Show that

$$\sigma(\{A_n\}_{n \in \mathbb{N}}) = \left\{ \bigcup_{j \in J} A_j : J \subset \mathbb{N} \right\}.$$

1.4. Show that

$$\begin{aligned} \mathcal{B} &= \sigma(\{(a, \infty) : a \in \mathbb{R}\}) = \sigma(\{[a, \infty) : a \in \mathbb{R}\}) \\ &= \sigma(\{(-\infty, a) : a \in \mathbb{R}\}) = \sigma(\{[a, b) : a, b \in \mathbb{R}, a < b\}). \end{aligned}$$



1.5. Show that

$$\begin{aligned}
 \mathcal{B}^d &= \sigma\left(\left\{\prod_{k=1}^d (-\infty, a_k) : a_k \in \mathbb{R}, k = 1, \dots, d\right\}\right) \\
 &= \sigma\left(\left\{\prod_{k=1}^d (a_k, \infty) : a_k \in \mathbb{R}, k = 1, \dots, d\right\}\right) \\
 &= \sigma\left(\left\{\prod_{k=1}^d (a_k, b_k) : a_k, b_k \in \mathbb{R}, a_k < b_k, k = 1, \dots, d\right\}\right) \\
 &= \sigma\left(\left\{\prod_{k=1}^d (a_k, b_k) : a_k, b_k \in \mathbb{R}, a_k < b_k, k = 1, \dots, d\right\}\right).
 \end{aligned}$$

1.6. Let  $\mathcal{C}$  be a collection of subsets of  $\Omega$  such that  $\Omega \in \mathcal{C}$ . Show that  $\mathcal{C}$  is a  $\lambda$ -system over  $\Omega$  iff both of the following hold:

- (a) if  $E, F \in \mathcal{C}$  and  $E \subset F$ , then  $F \setminus E \in \mathcal{C}$ ;
- (b) if  $(E_n)_{n \in \mathbb{N}}$  is an increasing sequence in  $\mathcal{C}$ , then  $\lim_n E_n \in \mathcal{C}$ .

1.7. Show that if a collection  $\mathcal{C}$  of subsets of  $\Omega$  is both a  $\pi$ -system and a  $\lambda$ -system, then it is a  $\sigma$ -algebra.

1.8. Let  $\mathcal{D}$  be a  $\lambda$ -system. Show that if  $(E_n)_{n \in \mathbb{N}}$  is a monotone sequence in  $\mathcal{D}$  then  $\lim_n E_n \in \mathcal{D}$ .

1.9. Prove Lemma 1.1.

1.10. Show that in the proof of Theorem 1.10,  $\mathcal{D}_1 = \mathcal{D}(\mathcal{P})$ .

1.11. Let  $\mathcal{P}$  be a  $\pi$ -system over  $\Omega$ . Show that  $\sigma(\mathcal{P}) = \mathcal{D}(\mathcal{P})$ .

1.12. Construct a  $\lambda$ -system that is not a  $\sigma$ -algebra.



## CHAPTER 2

### Measures

In the previous chapter, we study the collections of interesting events. In this chapter, we study the probability of their occurrence.

#### 1. Definitions and examples

We begin with the definition of measures. In order to gather intuition for them, one may interpret it as “length”, “area”, or “weight” of objects.

2.1. DEFINITION. Let  $\mathcal{F}$  be a  $\sigma$ -algebra over  $\Omega$ . A mapping  $\mu : \mathcal{F} \rightarrow [0, \infty]$  is a **measure** on  $(\Omega, \mathcal{F})$  if it satisfies the following two conditions:

- (a)  $\mu(\emptyset) = 0$ ;
- (b) for any disjoint sequence  $(E_n)_{n \in \mathbb{N}}$  in  $\mathcal{F}$ ,

$$\mu\left(\bigcup_{n \in \mathbb{N}} E_n\right) = \sum_{n=1}^{\infty} \mu(E_n).$$

We will call the triple  $(\Omega, \mathcal{F}, \mu)$  a **measure space**. We may also simply say that  $\mu$  is a measure on  $\Omega$  if there is no doubt about  $\mathcal{F}$  in the context.

Condition (b) is referred to as **countable additivity** of  $\mu$ . Intuitively, it can be interpreted as that the total area of countably infinite non-overlapping regions equals the sum of the areas of all the sub-regions.

We first look at several illustrative but elementary examples; most important ones will be constructed in Chapter 3.

2.1. EXAMPLE. Let  $\Omega = \{HH, HT, TH, TT\}$  and  $\mathcal{F} = \mathcal{P}(\Omega)$ . Set

$$\mu(\{HH\}) = \mu(\{HT\}) = \mu(\{TH\}) = \mu(\{TT\}) = \frac{1}{4}.$$

For an arbitrary set  $E \subset \Omega$ , set

$$\mu(E) := \sum_{\omega \in E} \mu(\{\omega\}).$$

For example,  $\mu(\{HH, HT, TH\}) = \mu(\{HH\}) + \mu(\{HT\}) + \mu(\{TH\}) = \frac{3}{4}$ . One can verify that  $\mu$  is a measure over  $(\Omega, \mathcal{F})$ .

This example can be extended to much more general cases.

2.2. EXAMPLE. Let  $\Omega$  be an arbitrary non-empty set, and  $\mathcal{F} = \mathcal{P}(\Omega)$ . Suppose that for each  $\omega \in \Omega$ , there corresponds a real number  $p_\omega \geq 0$ , called the **weight** at  $\omega$ . For each set  $E \subset \Omega$ , put

$$(2.1) \quad \mu(E) := \sum_{\omega \in \Omega} p_\omega.$$

Then  $\mu(E)$  is the “total weight” of the elements in  $E$ . Again, it is easy to see that  $\mu$  is a measure on  $(\Omega, \mathcal{F})$ .

If  $p_\omega = 1$  for any  $\omega \in \Omega$ , then  $\mu$  is called the **counting measure** on  $\Omega$ , as it simply counts the number of elements in a set. In this spirit, we may term the general measure  $\mu$  in (2.1) as the **weighted counting measure** on  $\Omega$  with weights  $(p_\omega)_{\omega \in \Omega}$ .

2.3. EXAMPLE. Let  $\Omega$  be a non-empty set, and fix any  $\omega_0 \in \Omega$ . Choose the weights by  $p_{\omega_0} = 1$  and  $p_\omega = 0$  for any  $\omega \neq \omega_0$ . Then the weighted counting measure satisfies the following

$$\mu(E) = \begin{cases} 1, & \text{if } \omega_0 \in E \\ 0, & \text{if } \omega_0 \notin E \end{cases}.$$

In particular,  $\mu(\{\omega_0\}) = 1$  and  $\mu(\Omega \setminus \{\omega_0\}) = 0$ . Thus  $\mu$  concentrates all its mass, which is of size 1, at the single point  $\omega_0$ . We give it a special notation  $\delta_{\omega_0}$  and call it the **Dirac measure** at  $\omega_0$ .

We now introduce probability spaces.

2.2. DEFINITION. Let  $\mu$  be a measure on  $(\Omega, \mathcal{F})$ .

- (a) It is called a **probability measure** if  $\mu(\Omega) = 1$ . In this case, the triple  $(\Omega, \mathcal{F}, \mu)$  is called a **probability space**. From now on, probability measures will be denoted by  $\mathbb{P}$  or  $\mathbb{Q}$ , with or without subscripts.
- (b) It is said to be **finite** if  $\mu(\Omega) < \infty$ .
- (c) It is said to be  **$\sigma$ -finite** if there exists a sequence  $(E_n)_{n \in \mathbb{N}}$  in  $\mathcal{F}$  such that  $\Omega = \bigcup_{n \in \mathbb{N}} E_n$  and  $\mu(E_n) < \infty$  for each  $n \in \mathbb{N}$ .

2.4. EXAMPLE. Let  $\Omega$  be an uncountable set. Then the counting measure on  $\Omega$  is not  $\sigma$ -finite. Indeed, otherwise,  $\Omega$  can be expressed as a countable union of sets, each of which is finite. This would imply that  $\Omega$  is countable.

2.5. EXAMPLE. Let  $\Omega = \mathbb{N}$ , and let  $\mu$  be any weighted counting measure on  $\mathbb{N}$ . Then  $\mu$  is  $\sigma$ -finite. Indeed, simply note that  $\mathbb{N} = \bigcup_{n \in \mathbb{N}} \{n\}$ .

2.6. EXAMPLE. Let  $\Omega = \mathbb{N}$ , and let  $\mu$  be the weighted counting measure for given weights  $(p_k)_{k \in \mathbb{N}}$ . Then  $\mu$  is finite iff  $\sum_{k=1}^{\infty} p_k < \infty$ , and  $\mu$  is a probability measure iff  $\sum_{k=1}^{\infty} p_k = 1$ . Indeed, simply note that

$$\mu(\mathbb{N}) = \sum_{k=1}^{\infty} p_k.$$

In particular, Dirac measures are probability measures.

The following example says that taking convex combinations of probability measures still results in a probability measure.

2.7. EXAMPLE. Let  $\mathcal{F}$  be a  $\sigma$ -algebra over  $\Omega$ ,  $(\mathbb{P}_k)_{k \in \mathbb{N}}$  be a sequence of probability measures on  $(\Omega, \mathcal{F})$ , and  $(c_k)_{k \in \mathbb{N}}$  be a sequence of non-negative real numbers such that  $\sum_{k=1}^{\infty} c_k = 1$ . Put

$$\mathbb{P}(E) = \sum_{k=1}^{\infty} c_k \mathbb{P}_k(E) \quad \text{for every } E \in \mathcal{F}.$$

Clearly,  $\mathbb{P}(\emptyset) = \sum_{k=1}^{\infty} c_k \cdot 0 = 0$ , and  $\mathbb{P}(\Omega) = \sum_{k=1}^{\infty} c_k \cdot 1 = 1$ . Let  $(E_n)$  be a disjoint sequence in  $\mathcal{F}$ . Then

$$\begin{aligned} \mathbb{P}\left(\bigcup_{n=1}^{\infty} E_n\right) &= \sum_{k=1}^{\infty} c_k \mathbb{P}_k\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{k=1}^{\infty} c_k \left(\sum_{n=1}^{\infty} \mathbb{P}_k(E_n)\right) \\ &= \sum_{k=1}^{\infty} \left(\sum_{n=1}^{\infty} c_k \mathbb{P}_k(E_n)\right) = \sum_{n=1}^{\infty} \left(\sum_{k=1}^{\infty} c_k \mathbb{P}_k(E_n)\right) \\ &= \sum_{n=1}^{\infty} \mathbb{P}(E_n), \end{aligned}$$

where the second equality follows from countable additivity of  $\mathbb{P}_k$  and the fourth one is due to changing order of summation, which is always true *for double sums with non-negative terms*. It follows that  $\mathbb{P}$  is a probability measure on  $(\Omega, \mathcal{F})$ . We usually rewrite  $\mathbb{P}$  as  $\sum_{k=1}^{\infty} c_k \mathbb{P}_k$ .

A special case of the preceding example is as follows.

2.8. EXAMPLE. Let  $(x_k)_{k \in \mathbb{N}}$  be a sequence of *distinct* real numbers and  $(c_k)_{k \in \mathbb{N}}$  be a sequence of positive real numbers such that  $\sum_{k=1}^{\infty} c_k = 1$ . Then  $\mathbb{P} = \sum_{k=1}^{\infty} c_k \delta_{x_k}$  is a probability measure on  $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$ . Clearly,

$$\mathbb{P}(\{x_k\}) = c_k \quad \text{for each } k \in \mathbb{N}$$

and

$$\mathbb{P}(\mathbb{R} \setminus \{x_k : k \in \mathbb{N}\}) = 0.$$

That is,  $\mathbb{P}$  has a mass of  $c_k$  at each  $x_k$ . Such probability measures are important as they are precisely the probability distributions of so-called discrete random variables. We will revisit them in Chapters 3 and 8.

(Of course,  $\mathbb{P}$  is just the weighted counting measure on  $\mathbb{R}$  with weights  $c_k$  at each  $x_k$  and 0 elsewhere.)

*From now on, we will state results only in the framework of probability spaces. However, nearly all interesting results in Chapters 4-7 hold for  $\sigma$ -finite measure spaces. Chapters 8-14 will only deal with probability spaces.*

## 2. Basic properties

We fix an arbitrary probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  for this section.

Note first that  $\mathbb{P}$  also has finite additivity and is increasing.

2.3. PROPOSITION. (a) *Let  $E_1, \dots, E_n$  be disjoint sets in  $\mathcal{F}$ . Then*

$$\mathbb{P}(\bigcup_{k=1}^n E_k) = \sum_{k=1}^n \mathbb{P}(E_k).$$

(b) *Let  $E, F \in \mathcal{F}$  be such that  $E \subset F$ . Then  $\mathbb{P}(F \setminus E) = \mathbb{P}(F) - \mathbb{P}(E)$ .*

*In particular,  $\mathbb{P}(E) \leq \mathbb{P}(F)$ .*

PROOF. For (a), put  $E_k = \emptyset$  for  $k \geq n+1$  and apply countable additivity of  $\mathbb{P}$ . For (b), note that  $E$  and  $F \setminus E$  are disjoint. Thus by (a),

$$\mathbb{P}(E) + \mathbb{P}(F \setminus E) = \mathbb{P}(E \cup (F \setminus E)) = \mathbb{P}(F),$$

from which the desired results follow immediately.  $\square$

2.4. PROPOSITION. *Let  $(E_n)_{n \in \mathbb{N}}$  be a sequence of sets in  $\mathcal{F}$  and  $E \in \mathcal{F}$ . If  $E_n \uparrow E$ , then  $\mathbb{P}(E_n) \uparrow \mathbb{P}(E)$ .*

PROOF. The increasingness of  $\mathbb{P}(E_n)$ 's is due to Proposition 2.3 (b).

For the convergence, we cut  $E_n$ 's into disjoint sets as follows. Put  $F_1 = E_1$ . For  $n \geq 2$ , put  $F_n = E_n \setminus E_{n-1}$ . See Figure 1 below for illustration. Clearly,  $(F_n)$  is a disjoint sequence of sets in  $\mathcal{F}$ ,  $\bigcup_{k=1}^n F_k = E_n$  for any  $n \in \mathbb{N}$ , and  $\bigcup_{n=1}^{\infty} F_n = \bigcup_{n=1}^{\infty} E_n$ . Thus by the countable (and finite) additivity of  $\mathbb{P}$ , it follows that

$$\begin{aligned} \mathbb{P}(E) &= \mathbb{P}\left(\bigcup_{n=1}^{\infty} E_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} F_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(F_n) = \lim_{n \rightarrow \infty} \sum_{k=1}^n \mathbb{P}(F_k) \\ &= \lim_{n \rightarrow \infty} \mathbb{P}\left(\bigcup_{k=1}^n F_k\right) = \lim_{n \rightarrow \infty} \mathbb{P}(E_n). \end{aligned}$$

$\square$

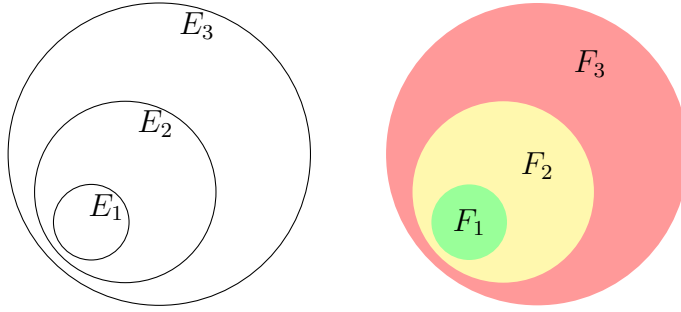


FIGURE 1. Cutting an increasing sequence of sets into a disjoint sequence

The property in this proposition is called **continuity from below**. It is an equivalent form of countable additivity (Exercise 2.5) and, as will be seen (e.g., in the proof of Theorem 5.9), is the very property of measures that guarantees the nice convergence properties of Lebesgue integrals and expectations. We now include some of its elementary corollaries below.

The following property is called **countable sub-additivity**.

2.5. COROLLARY. *Let  $(E_n)_{n \in \mathbb{N}}$  be a sequence of sets in  $\mathcal{F}$ . Then*

$$(2.2) \quad \mathbb{P}\left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(E_n).$$

PROOF. For any two sets  $F_1, F_2 \in \mathcal{F}$ , note that

$$(2.3) \quad \begin{aligned} \mathbb{P}(F_1 \cup F_2) &= \mathbb{P}(F_1 \cup (F_2 \setminus F_1)) = \mathbb{P}(F_1) + \mathbb{P}(F_2 \setminus F_1) \\ &\leq \mathbb{P}(F_1) + \mathbb{P}(F_2). \end{aligned}$$

This observation, together with induction, implies (see Exercise 2.6) that  $\mathbb{P}$  has finite sub-additivity:

$$(2.4) \quad \mathbb{P}\left(\bigcup_{k=1}^n E_k\right) \leq \sum_{k=1}^n \mathbb{P}(E_k).$$

In view of  $\bigcup_{k=1}^n E_k \uparrow_n \bigcup_{k=1}^{\infty} E_k$ , letting  $n \rightarrow \infty$  completes the proof.  $\square$

An immediate consequence of the countable sub-additivity is that union of countable negligible sets remains negligible.

2.6. DEFINITION. *A set  $E \in \mathcal{F}$  is said to be **negligible** if  $\mathbb{P}(E) = 0$ .*

2.7. COROLLARY. *Let  $(E_n)_{n \in \mathbb{N}}$  be a sequence of negligible sets. Then  $\bigcup_{n=1}^{\infty} E_n$  is also negligible.*

2.8. COROLLARY. *Let  $(E_n)_{n \in \mathbb{N}}$  be a sequence of sets in  $\mathcal{F}$ . Then*

$$\mathbb{P}(\liminf_n E_n) \leq \liminf_n \mathbb{P}(E_n).$$

PROOF. Set  $F_n = \bigcap_{k=n}^{\infty} E_k$  for  $n \in \mathbb{N}$ . Recall that  $F_n \uparrow \liminf_n E_n$ . Thus

$$\mathbb{P}(\liminf_n E_n) = \lim_n \mathbb{P}(F_n) = \sup_{n \geq 1} \mathbb{P}(F_n).$$

Furthermore, for any  $k \geq n$ , since  $F_n \subset E_k$ ,  $\mathbb{P}(F_n) \leq \mathbb{P}(E_k)$ . Thus

$$\mathbb{P}(F_n) \leq \inf_{k \geq n} \mathbb{P}(E_k).$$

Combining the above, we have

$$\mathbb{P}(\liminf_n E_n) \leq \sup_{n \geq 1} \inf_{k \geq n} \mathbb{P}(E_k) = \liminf_n \mathbb{P}(E_n).$$

□

### 3. Uniqueness

Probability measures, though defined on  $\sigma$ -algebras, are in fact determined by their values on smaller collections of sets.

2.9. THEOREM. *Let  $\mathbb{P}$  and  $\mathbb{Q}$  be probability measures over  $(\Omega, \mathcal{F})$ , where  $\mathcal{F}$  is generated by a  $\pi$ -system  $\mathcal{P}$ . If  $\mathbb{P}$  and  $\mathbb{Q}$  agree on  $\mathcal{P}$ , then they agree on  $\mathcal{F}$ .*

The proof uses a general approach of handling generated  $\sigma$ -algebras: collect all the sets satisfying the desired property and then manage to apply the monotone class theorem 1.10.

PROOF. Put

$$\mathcal{D} = \{E \in \mathcal{F} : \mathbb{P}(E) = \mathbb{Q}(E)\}.$$

Then  $\mathcal{P} \subset \mathcal{D}$ . If  $\mathcal{D}$  were a  $\lambda$ -system, then by Theorem 1.10,  $\mathcal{F} = \sigma(\mathcal{P}) \subset \mathcal{D}$ , we are done. We now verify that  $\mathcal{D}$  is a  $\lambda$ -system. First, since  $\mathbb{P}(\emptyset) = 0 = \mathbb{Q}(\emptyset)$ ,  $\emptyset \in \mathcal{D}$ . Second, take any  $E \in \mathcal{D}$ . Then by Proposition 2.3 (b),

$$\mathbb{P}(E^c) = 1 - \mathbb{P}(E) = 1 - \mathbb{Q}(E) = \mathbb{Q}(E^c),$$

so that  $E^c \in \mathcal{D}$ . Third, let  $(E_n)_{n \in \mathbb{N}}$  be any disjoint sequence in  $\mathcal{D}$ . Then by the countable additivity,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(E_n) = \sum_{n=1}^{\infty} \mathbb{Q}(E_n) = \mathbb{Q}\left(\bigcup_{n=1}^{\infty} E_n\right).$$

Thus  $\bigcup_{n=1}^{\infty} E_n \in \mathcal{D}$ . This proves that  $\mathcal{D}$  is a  $\lambda$ -system. □



2.10. COROLLARY. Let  $\mathbb{P}$  and  $\mathbb{Q}$  be probability measures over  $(\mathbb{R}^d, \mathcal{B}^d)$ . Then  $\mathbb{P} = \mathbb{Q}$  if they agree either on the collection of all cubes of the form:

$$\prod_{k=1}^d (a_k, b_k], \text{ where } a_k, b_k \in \mathbb{R}, a_k < b_k, k = 1, \dots, d,$$

or on the collection of all cubes of the form:

$$\prod_{k=1}^d (-\infty, a_k], \text{ where } a_k \in \mathbb{R}, k = 1, \dots, d.$$

PROOF. The second collection is a  $\pi$ -system generating  $\mathcal{B}^d$ . The first collection is not a  $\pi$ -system, but its union with the singleton  $\{\emptyset\}$  is a  $\pi$ -system generating  $\mathcal{B}^d$ , on which  $\mathbb{P}$  and  $\mathbb{Q}$  still agree.  $\square$

We close this chapter with the following remark.

- 2.11. REMARK. (a) All the results in Section 2 hold for a general measure space, except that  $\mathbb{P}(F \setminus E) = \mathbb{P}(F) - \mathbb{P}(E)$  in Proposition 2.3 (b), which is still valid as long as  $\infty - \infty$  does not occur. See Exercise 2.13.
- (b) Theorem 2.9 may fail for a general  $\sigma$ -finite measure space. However, under a mild additional assumption, it still holds. See Exercises 2.14 and 2.15.

### Exercises

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

- 2.1. Let  $E$  be a set in  $\mathcal{F}$  such that  $\mathbb{P}(E) > 0$ . For every  $F \in \mathcal{F}$ , put

$$\mathbb{Q}(F) = \frac{\mathbb{P}(F \cap E)}{\mathbb{P}(E)}.$$

Show that  $\mathbb{Q}$  is a probability measure on  $\mathcal{F}$ . (It is called the **conditional probability** given  $E$  and is usually written as  $\mathbb{P}(\cdot|E)$ . In contrast,  $\mathbb{P}$  is sometimes called the **unconditional probability**.)

- 2.2. Let  $(\Omega, \mathcal{F}, \mu)$  be a measure space. Let  $E$  be a set in  $\mathcal{F}$ . Put  $\nu(F) = \mu(F \cap E)$  for every  $F \in \mathcal{F}$ . Show that  $\nu$  is a measure on  $\mathcal{F}$ .

- 2.3. Let  $(\Omega, \mathcal{F}, \mu)$  be a measure space and  $c \geq 0$  is a real number. Put  $\nu(F) = c\mu(F)$  for every  $F \in \mathcal{F}$ . Show that  $\nu$  is a measure on  $\mathcal{F}$ .

- 2.4. Let  $(\mu_n)_{n \in \mathbb{N}}$  be a sequence of measures on  $(\Omega, \mathcal{F})$ . For every  $E \in \mathcal{F}$ , put  $\nu(E) = \sum_{n=1}^{\infty} \mu_n(E)$ . Show that  $\nu$  is a measure on  $\mathcal{F}$ .

The measures  $\nu$  in Exercise 2.3 and 2.4 are usually denoted as  $c\mu$  and  $\sum_{n=1}^{\infty} \mu_n$ , respectively.

2.5. Let  $\mathcal{F}$  be a  $\sigma$ -algebra over  $\Omega$ . Let  $\mu : \mathcal{F} \rightarrow [0, \infty]$  be a mapping such that  $\mu(\emptyset) = 0$  and  $\mu(\bigcup_{k=1}^n E_k) = \sum_{k=1}^n \mu(E_k)$  for any  $n \in \mathbb{N}$  and any disjoint sets  $E_1, \dots, E_n$  in  $\mathcal{F}$ . Show that  $\mu$  is a measure iff it is continuous from below, i.e., if  $E_n \uparrow E$ , then  $\mu(E_n) \uparrow \mu(E)$ .

2.6. Deduce (2.4) from (2.3).

2.7. Suppose that  $\mathbb{P}(E_n) = 1$  for any  $n \in \mathbb{N}$ . Show that  $\mathbb{P}(\bigcap_{n=1}^{\infty} E_n) = 1$ .

2.8. Let  $(E_n)_{n \in \mathbb{N}}$  be a sequence of sets in  $\mathcal{F}$  and  $E \in \mathcal{F}$  be such that  $E_n \downarrow E$ . Show that  $\mathbb{P}(E_n) \downarrow \mathbb{P}(E)$ .

2.9. Give a counterexample to show that the conclusion in Exercise 2.8 may fail for a general measure  $\mu$ . Show that it still holds if  $\mu(E_1) < \infty$ .

2.10. Let  $(E_n)_{n \in \mathbb{N}}$  be a sequence of sets in  $\mathcal{F}$ . Show that

$$\limsup_n \mathbb{P}(E_n) \leq \mathbb{P}(\limsup_n E_n).$$

2.11. Give a counterexample to show that the conclusion in Exercise 2.10 may fail for a general measure  $\mu$ . Show that it still holds if  $\mu(\bigcup_{n=1}^{\infty} E_n) < \infty$ .

2.12. Let  $E, E_n, n \in \mathbb{N}$ , be sets in  $\mathcal{F}$  such that  $E_n \rightarrow E$ . Show that  $\mathbb{P}(E_n) \rightarrow \mathbb{P}(E)$ .

2.13. Observe that the conclusions in Proposition 2.4 and Corollaries 2.5, 2.7 and 2.8 hold for a general measure. The same arguments work.

2.14. Let  $\mu$  be the counting measure on  $\mathbb{N}$  and  $\nu = 2\mu$ . Let

$$\mathcal{P} = \left\{ \{k \in \mathbb{N} : k \geq n\} : n \in \mathbb{N} \right\}.$$

Show that  $\mathcal{P}$  is a  $\pi$ -system,  $\sigma(\mathcal{P}) = \mathcal{P}(\mathbb{N})$ ,  $\mu = \nu$  on  $\mathcal{P}$ , but  $\mu \neq \nu$  on  $\mathcal{P}(\mathbb{N})$ .

2.15. Let  $\mu$  and  $\nu$  be  $\sigma$ -finite measures over  $(\Omega, \mathcal{F})$ , where  $\mathcal{F}$  is generated by a  $\pi$ -system  $\mathcal{P}$ . If  $\mu$  and  $\nu$  agree on  $\mathcal{P}$  and are finite on an increasing sequence  $(P_n)_{n \in \mathbb{N}}$  in  $\mathcal{P}$  whose union is  $\Omega$ , then they agree on  $\mathcal{F}$ .

## CHAPTER 3

### Lebesgue-Stieltjes Measures

We have not seen any non-trivial measures other than the weighted counting measures. It is because constructing non-trivial measures, even in the case of  $\mathbb{R}$ , is generally very difficult.

#### 1. An extension theorem

Say, we want to construct a measure  $m$  on  $(\mathbb{R}, \mathcal{F})$  that measures the “length” of one-dimensional objects in  $\mathbb{R}$ . Here  $\mathcal{F}$  is a  $\sigma$ -algebra over  $\mathbb{R}$ , conceptually consisting of “measurable” objects (we reasonably expect that some objects may be too complex for us to measure their length). We start with the simplest objects, intervals. It should be in common agreement that we know how to measure their length: simply specify that

$$m((a, b]) = b - a.$$

Once we agree on this, we should agree that we can also measure the length of a bit more complex objects, finite unions of intervals, e.g.,

$$\begin{aligned} & m((-4, -1] \cup (1, 2] \cup (5, 7]) \\ &= m((-4, -1]) + m((1, 2]) + m((5, 7]) \\ &= (-1 - (-4)) + (2 - 1) + (7 - 5) \\ &= 6. \end{aligned}$$

But the assumption that all the intervals are “measurable” and belong to  $\mathcal{F}$  already forces that  $\mathcal{F}$  contains all the Borel sets (Example 1.4). A general Borel set  $B$  could have a very complicated structure, and directly specifying the value  $m(B)$  could be extremely difficult.

Luckily, we have a theorem that guarantees that if it is known how to measure the length of simple objects, such as intervals or finite unions of intervals, then there is an automatic way to extend our measurement to much more complex objects, such as Borel sets.

To introduce the theorem, we need the notions of algebras and pre-measures, which are weakened versions of  $\sigma$ -algebras and measures, respectively.

3.1. DEFINITION. A collection  $\mathcal{A}$  of subsets of  $\Omega$  is called an **algebra** over  $\Omega$  if it satisfies the following conditions:

- (a)  $\emptyset \in \mathcal{A}$ ;
- (b)  $A^c \in \mathcal{A}$  whenever  $A \in \mathcal{A}$ ;
- (c)  $A \cap B \in \mathcal{A}$  whenever  $A, B \in \mathcal{A}$ .

By using induction and De Morgan's Laws, it is easy to see that  $\mathcal{A}$  is closed under taking finite intersections and unions (Proposition 1.3 holds for  $\mathcal{A}$ ). The definition of algebras differ from that of  $\sigma$ -algebras in that the latter allows to take unions of countably *infinite* objects.

3.2. DEFINITION. Let  $\mathcal{A}$  be an algebra over  $\Omega$ . A mapping  $\underline{\mu} : \mathcal{A} \rightarrow [0, \infty]$  is a **pre-measure** on  $(\Omega, \mathcal{A})$  if it satisfies the following two conditions:

- (a)  $\underline{\mu}(\emptyset) = 0$ ;
- (b) for any disjoint sequence  $(A_n)_{n \in \mathbb{N}}$  in  $\mathcal{A}$ , if  $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$ , then

$$\underline{\mu}\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n=1}^{\infty} \underline{\mu}(A_n).$$

One can easily see that  $\underline{\mu}$  has finite additivity and is increasing (cf. Proposition 2.3). The weakness of  $\underline{\mu}$  is that it is defined only on an algebra, and thus in Condition (b) of Definition 3.2, if  $(E_n)_{n \in \mathbb{N}}$  is a disjoint sequence in  $\mathcal{A}$  such that  $\bigcup_{n=1}^{\infty} E_n \notin \mathcal{A}$ , then we have no control of  $\bigcup_{n=1}^{\infty} E_n$ .

We are ready to present the Carathéodory extension theorem.

3.3. THEOREM. Let  $\underline{\mu}$  be a pre-measure on an algebra  $\mathcal{A}$  over  $\Omega$ . Then there exists a measure  $\mu$  on  $\sigma(\mathcal{A})$  such that

$$(3.1) \quad \mu(A) = \underline{\mu}(A) \quad \text{for every } A \in \mathcal{A}.$$

Furthermore, if there exists an increasing sequence  $(A_n)_{n \in \mathbb{N}}$  in  $\mathcal{A}$  such that  $\Omega = \bigcup_{n=1}^{\infty} A_n$  and  $\underline{\mu}(A_n) < \infty$  for every  $n \in \mathbb{N}$ , then there is a unique measure on  $\sigma(\mathcal{A})$  satisfying (3.1).

PROOF. For the existence part, the “automatic” way that we alluded earlier is as follows. For every  $E \in \sigma(\mathcal{A})$ , put

$$(3.2) \quad \mu(E) = \inf \left\{ \sum_{n=1}^{\infty} \underline{\mu}(A_n) : A_n \in \mathcal{A} \text{ for each } n \in \mathbb{N}, E \subset \bigcup_{n=1}^{\infty} A_n \right\}.$$

Basically, one measures  $E \in \sigma(\mathcal{A})$  by covering it with countably infinite elements from  $\mathcal{A}$ , whose objects we already how to measure: use  $\underline{\mu}$ . Though the formula is easy to state, verifying that  $\mu$  defined this way is a measure on  $\sigma(\mathcal{A})$  satisfying (3.1) is quite technical. As we will not use the techniques in the proof for the rest of the book, we put the proof to Appendix A.

For the uniqueness part, let  $\mu_1$  and  $\mu_2$  are two measures on  $\sigma(\mathcal{A})$  satisfying (3.1). Then they agree on  $\mathcal{A}$ . Observe that an algebra is a  $\pi$ -system. Thus they agree on  $\sigma(\mathcal{A})$ , by Theorem 2.9 (Exercise 2.15, to be accurate).  $\square$

## 2. Lebesgue-Stieltjes measures

We now employ the Carathéodory extension theorem to construct important, non-trivial measures on  $(\mathbb{R}, \mathcal{B})$ . Of course, the theorem reduces our work to construct pre-measures on algebras.

We start with the algebra that we will build the pre-measures on. Consider intervals that are left open and right closed, namely, intervals of one of the following forms:

$$(-\infty, a], \quad (b, c], \quad (d, \infty), \quad a, b, c, d \in \mathbb{R}, b < c.$$

Let  $\mathcal{A}$  be the collection of  $\emptyset$  and all unions of *finitely many, disjoint* such intervals. Members in  $\mathcal{A}$  have quite simple structure. For example,

$$(-\infty, -2] \cup (0, 1] \cup (2, 3] \in \mathcal{A}.$$

$$\mathbb{R} = (-\infty, 1] \cup (1, \infty) \in \mathcal{A}.$$

Note that a member in  $\mathcal{A}$  may be written in more than one form. For example,

$$(-\infty, 0] \cup (1, 7] = (-\infty, -2] \cup (-2, 0] \cup (1, 3] \cup (3, 5] \cup (5, 7].$$

3.1. LEMMA.  $\mathcal{A}$  is an algebra on  $\mathbb{R}$ .

The proof is very simple; the reader may draw a graph to illustrate it.

PROOF. Condition (a) in Definition 3.1 is clear. For Condition (b), take any  $A \in \mathcal{A}$ . If  $A = \emptyset$ , clearly  $A^c = \mathbb{R} \in \mathcal{A}$ . Otherwise, write  $A = \bigcup_{k=1}^n I_k$ , where  $I_k$ 's are disjoint intervals and each has the designated form. For each  $k$ , let  $a_k, b_k$  be the left and right endpoints of  $I_k$ , respectively. Without loss of generality, assume that  $-\infty \leq a_1 < b_1 \leq a_2 < b_2 \leq \cdots \leq a_n < b_n \leq \infty$ . Thus

$$A^c = (-\infty, a_1] \cup (b_1, a_2] \cup \cdots \cup (b_n, \infty);$$

if  $a_1 = -\infty$ , or  $b_k = a_{k+1}$  for some  $k$ , or  $b_n = \infty$ , remove the corresponding intervals from the expression. It follows clearly that  $A^c \in \mathcal{A}$ .

For Condition (c), take any  $A = \bigcup_{k=1}^n I_k$  and  $B = \bigcup_{l=1}^m J_l$  in  $\mathcal{A}$ , where  $I_k$ 's and  $J_l$ 's all are intervals of the designated form, and  $I_k$ 's as well as  $J_l$ 's are disjoint. Then

$$A \cap B = \bigcup_{k=1}^n \bigcup_{l=1}^m (I_k \cap J_l).$$

One sees that  $I_k \cap J_l$ 's are disjoint and all have the designated form, if non-empty. Remove the empty ones. It follows that  $A \cap B \in \mathcal{A}$ .

Combining the above proves that  $\mathcal{A}$  is an algebra.  $\square$

To build very general pre-measures, fix a function  $F : \mathbb{R} \rightarrow \mathbb{R}$  that is increasing and right continuous. For convenience, put

$$\begin{aligned} F(-\infty) &= \lim_{x \rightarrow -\infty} F(x) \in [-\infty, \infty), \\ F(\infty) &= \lim_{x \rightarrow \infty} F(x) \in (-\infty, \infty]. \end{aligned}$$

Put  $\underline{\mu}(\emptyset) = 0$ . For any interval  $I$  of the designated form with endpoints  $a$  and  $b$  with  $-\infty \leq a < b \leq \infty$ , put

$$\underline{\mu}(I) = F(b) - F(a).$$

(Here it explains why we want  $F$  to be increasing: to ensure that  $\underline{\mu}$  takes only non-negative values.) For a general  $A \in \mathcal{A}$ , say,  $A = \bigcup_{k=1}^n I_k$ , where  $I_k$ 's are disjoint and each  $I_k$  has the designated form, put

$$\underline{\mu}(A) = \underline{\mu}(I_1) + \underline{\mu}(I_2) + \cdots + \underline{\mu}(I_n).$$

For example,

$$\underline{\mu}((-\infty, 2] \cup (3, 4]) = F(2) - F(-\infty) + F(4) - F(3).$$

Note, however, that since  $A \in \mathcal{A}$  may have multiple expressions, we also have multiple ways to specify  $\underline{\mu}(A)$ . But all different forms give the same value. For example, let's write  $(1, 7]$  in two forms  $(1, 7]$  and  $(1, 3] \cup (3, 5] \cup (5, 7]$ . Then clearly,

$$F(7) - F(1) = F(7) - F(5) + F(5) - F(3) + F(3) - F(1).$$

Finally, note that

$$\begin{aligned} \underline{\mu}(\mathbb{R}) &= \underline{\mu}((-\infty, 1] \cup (1, \infty)) = F(\infty) - F(1) + F(1) - F(-\infty) \\ &= F(\infty) - F(-\infty). \end{aligned}$$

**3.2. LEMMA.**  $\underline{\mu}$  is a pre-measure on  $\mathcal{A}$ .

PROOF. We split the proof into a few steps.

*Step I.* The first quick observation is that if  $A, B \in \mathcal{A}$  are disjoint then

$$\underline{\mu}(A \cup B) = \underline{\mu}(A) + \underline{\mu}(B).$$

Indeed, write  $A = \bigcup_{k=1}^n I_k$  and  $B = \bigcup_{l=1}^m J_l$  in  $\mathcal{A}$ , where  $I_k$ 's and  $J_l$ 's all have the designated form, and  $I_k$ 's as well as  $J_l$ 's are disjoint. Then  $I_k$ 's and  $J_l$ 's put together are still disjoint and their union is  $A \cup B$ . Thus by definition of  $\underline{\mu}$ ,

$$\underline{\mu}(A \cup B) = \sum_{k=1}^n \underline{\mu}(I_k) + \sum_{l=1}^m \underline{\mu}(J_l) = \underline{\mu}(A) + \underline{\mu}(B).$$

By induction, it follows that if  $A_1, \dots, A_n \in \mathcal{A}$  are disjoint then

$$\underline{\mu}\left(\bigcup_{k=1}^n A_k\right) = \sum_{k=1}^n \underline{\mu}(A_k).$$

*Step II.* For any  $A, B \in \mathcal{A}$  with  $A \subset B$ , since  $A$  and  $B \setminus A$  are disjoint in  $\mathcal{A}$ , we have by Step I,

$$\underline{\mu}(B) = \underline{\mu}(A \cup (B \setminus A)) = \underline{\mu}(A) + \underline{\mu}(B \setminus A) \geq \underline{\mu}(A).$$

*Step III.* Let  $(A_n)_{n \in \mathbb{N}}$  be a disjoint sequence in  $\mathcal{A}$  such that  $A := \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$ . Then by Step II and then Step I, for any  $n \in \mathbb{N}$ ,  $\underline{\mu}(A) \geq \underline{\mu}(\bigcup_{k=1}^n A_k) = \sum_{k=1}^n \underline{\mu}(A_k)$ . Letting  $n \rightarrow \infty$ , we obtain

$$\underline{\mu}(A) \geq \sum_{n=1}^{\infty} \underline{\mu}(A_n).$$

The rest of the proof is devoted to the reverse of this inequality.

*Step IV.* Let  $A, A_1, \dots, A_n$  be in  $\mathcal{A}$  such that  $A \subset \bigcup_{k=1}^n A_k$ . Put  $B_1 = A \cap A_1$ . For  $k = 2, \dots, n$ , put  $B_k = (A_k \cap A) \setminus \bigcup_{j=1}^{k-1} (A_j \cap A)$ . Then  $B_k$ 's lie in  $\mathcal{A}$  and are disjoint such that  $\bigcup_{k=1}^n B_k = \bigcup_{k=1}^n (A_k \cap A) = A$ . By Steps I and II,

$$\underline{\mu}(A) = \sum_{k=1}^n \underline{\mu}(B_k) \leq \sum_{k=1}^n \underline{\mu}(A_k).$$

*Step V.* Take any  $I = (a, b]$ , where  $a, b \in \mathbb{R}$  and  $a < b$ . Let  $I_n = (a_n, b_n]$ ,  $n \in \mathbb{N}$ , be any disjoint sequence of intervals such that  $I = \bigcup_{n=1}^{\infty} I_n$ . Then

$$(3.3) \quad \underline{\mu}(I) = \sum_{n=1}^{\infty} \underline{\mu}(I_n).$$

Indeed, take any  $0 < \delta < b - a$  and any  $\varepsilon > 0$ . Since  $F$  is right-continuous, for each  $k$ , we can find  $\delta_k > 0$  such that

$$F(b_k + \delta_k) - F(b_k) < \frac{\varepsilon}{2^k}.$$

Note that  $[a + \delta, b] \subset \bigcup_{k=1}^{\infty} (a_k, b_k + \delta_k)$ . The Heine-Borel theorem asserts that we can find finitely many  $(a_k, b_k + \delta_k)$ 's to cover  $[a + \delta, b]$ . Thus, we can find  $N \in \mathbb{N}$  such that  $[a + \delta, b] \subset \bigcup_{k=1}^N (a_k, b_k + \delta_k)$ . Thus

$$(a + \delta, b] \subset \bigcup_{k=1}^N (a_k, b_k + \delta_k].$$

By Step IV,

$$\begin{aligned} F(b) - F(a + \delta) &= \underline{\mu}((a + \delta, b]) \\ &\leq \sum_{k=1}^N \underline{\mu}((a_k, b_k + \delta_k]) = \sum_{k=1}^N (F(b_k + \delta_k) - F(a_k)) \\ &\leq \sum_{k=1}^{\infty} (F(b_k + \delta_k) - F(a_k)) \leq \sum_{k=1}^{\infty} (F(b_k) + \frac{\varepsilon}{2^k} - F(a_k)) \\ &= \sum_{k=1}^{\infty} (F(b_k) - F(a_k)) + \sum_{k=1}^{\infty} \frac{\varepsilon}{2^k} \\ &= \sum_{k=1}^{\infty} \underline{\mu}(I_k) + \varepsilon. \end{aligned}$$

Letting  $\varepsilon \rightarrow 0$  and then  $\delta \rightarrow 0$ , using right continuity of  $F$  at  $a$ , we obtain

$$\underline{\mu}(I) \leq \sum_{n=1}^{\infty} \underline{\mu}(I_n).$$

In view of Step III, this proves the desired equality.

*Step VI.* Take any  $I = (-\infty, a]$  or  $(a, \infty)$ , where  $a \in \mathbb{R}$ . Let  $(I_n)$  be any disjoint sequence of intervals of the designated form such that  $I = \bigcup_{n=1}^{\infty} I_n$ . Then (3.3) holds. Let's prove the case  $I = (a, \infty)$ ; the other case can be proved similarly. For any  $k \in \mathbb{N}$  with  $k > a$ , we have

$$(a, k] = I \cap (a, k] = \bigcup_{n=1}^{\infty} (I_n \cap (a, k]).$$

If  $I_n \cap (a, k] = \emptyset$ , remove it from the union; the final terms in the union may be finite or countably infinite. Thus in view  $\underline{\mu}(\emptyset) = 0$ , we have, by Step I or Step V,  $F(k) - F(a) = \underline{\mu}((a, k]) = \sum_{n=1}^{\infty} \underline{\mu}(I_n \cap (a, k]) \leq \sum_{n=1}^{\infty} \underline{\mu}(I_n)$ . Letting  $k \rightarrow \infty$  and applying Step III, we obtain the desired equality.



*Step VII.* Let  $I$  be any interval of the designated form. Let  $(A_n)_{n \in \mathbb{N}}$  be any disjoint sequence in  $\mathcal{A}$  such that  $I = \bigcup_{n=1}^{\infty} A_n$ . Then

$$\underline{\mu}(I) = \sum_{n=1}^{\infty} \underline{\mu}(A_n).$$

Indeed, for each  $n \in \mathbb{N}$ , write  $A_n = \bigcup_{k=1}^{N_n} I_{n,k}$ , where  $I_{n,k}$ 's are disjoint intervals of the designated form. Putting  $I_{n,k}$ ,  $n \in \mathbb{N}$ ,  $1 \leq k \leq N_n$ , together yields a new countably infinite collection of disjoint intervals whose union is clearly  $I$ . Thus by Step V, it follows that

$$\underline{\mu}(I) = \sum_{n=1}^{\infty} \sum_{k=1}^{N_n} \underline{\mu}(I_{n,k}) = \sum_{n=1}^{\infty} \underline{\mu}(A_n).$$

*Final Step.* Let  $A \in \mathcal{A}$  and  $(A_n)_{n \in \mathbb{N}}$  be a disjoint sequence in  $\mathcal{A}$  such that  $A = \bigcup_{n=1}^{\infty} A_n$ . Write  $A = \bigcup_{k=1}^m I_k$ , where  $I_k$ 's are disjoint intervals of the designated form. For each  $k$ , we have  $I_k = A \cap I_k = \bigcup_{n=1}^{\infty} (A_n \cap I_k)$ . Thus by Step VII,  $\underline{\mu}(I_k) = \sum_{n=1}^{\infty} \underline{\mu}(A_n \cap I_k)$ . Consequently,

$$\underline{\mu}(A) = \sum_{k=1}^m \underline{\mu}(I_k) = \sum_{k=1}^m \sum_{n=1}^{\infty} \underline{\mu}(A_n \cap I_k) = \sum_{n=1}^{\infty} \sum_{k=1}^m \underline{\mu}(A_n \cap I_k) = \sum_{n=1}^{\infty} \underline{\mu}(A_n),$$

where the last equality follows from Step I and  $A_n = \bigcup_{k=1}^m (A_n \cap I_k)$ .  $\square$

To sum up, we formulate it as a theorem.

**3.4. THEOREM.** *Let  $F : \mathbb{R} \rightarrow \mathbb{R}$  be an increasing and right-continuous function. Then there exists a unique measure  $\mu$  on  $(\mathbb{R}, \mathcal{B})$  such that*

$$(3.4) \quad \mu((a, b]) = F(b) - F(a),$$

*for any  $a, b \in \mathbb{R}$  with  $a < b$ .*

**PROOF.** By Lemmas 3.1 and 3.2, we get a pre-measure  $\underline{\mu}$  on  $\mathcal{A}$  satisfying (3.4). Let  $\mu$  be any measure on  $(\mathbb{R}, \mathcal{B})$  obtained for  $\underline{\mu}$  by Theorem 3.3. It clearly satisfies (3.4) as well since it coincides with  $\underline{\mu}$  on  $\mathcal{A}$ .

Moreover, let  $\mu'$  be any measure on  $(\mathbb{R}, \mathcal{B})$  satisfying (3.4). Then it coincides with  $\underline{\mu}$  on all intervals of the form  $(a, b]$ , where  $a, b \in \mathbb{R}$  and  $a < b$ , and thus on all intervals of the form  $(-\infty, a]$  or  $(a, \infty)$ , where  $a \in \mathbb{R}$ , as well (Exercise 3.4). It follows that  $\mu'$  coincides with  $\underline{\mu}$  on  $\mathcal{A}$ . Since  $(-n, n] \uparrow \mathbb{R}$  and  $\underline{\mu}((-n, n]) = F(n) - F(-n) < \infty$  for each  $n$ ,  $\mu'$  must coincide with  $\mu$  on  $\mathcal{B}$ , by the uniqueness part in Theorem 3.3.  $\square$

We call  $\mu$  the **Lebesgue-Stieltjes measure** associated with  $F$ . We may write it as  $\mu_F$  if it is necessary to emphasize  $F$ .

3.5. REMARK. Lemma 3.2 shows the sufficiency of assuming right continuity of  $F$  to make  $\underline{\mu}$  a pre-measure. The necessity can be demonstrated easily. Consider  $(0, 1] = \bigcup_{n=1}^{\infty} (\frac{1}{n+1}, \frac{1}{n}]$ . If we want  $\underline{\mu}$  to be pre-measure, then we must have

$$\begin{aligned} F(1) - F(0) &= \underline{\mu}((0, 1]) = \sum_{n=1}^{\infty} \underline{\mu}\left(\left(\frac{1}{n+1}, \frac{1}{n}\right)\right) \\ &= \sum_{n=1}^{\infty} \left(F\left(\frac{1}{n}\right) - F\left(\frac{1}{n+1}\right)\right) = \lim_{n \rightarrow \infty} \sum_{k=1}^n \left(F\left(\frac{1}{k}\right) - F\left(\frac{1}{k+1}\right)\right) \\ &= \lim_{n \rightarrow \infty} \left(F(1) - F\left(\frac{1}{n+1}\right)\right) \\ &= F(1) - F(0+). \end{aligned}$$

Therefore,  $F(0) = F(0+)$ , i.e.,  $F$  is right continuous at 0. Right continuity at other points can be proved similarly.

### 3. Some properties and examples

3.6. PROPOSITION. *Let  $F : \mathbb{R} \rightarrow \mathbb{R}$  be increasing and right continuous. Let  $\mu$  be the Lebesgue-Stieltjes measure associated with  $F$ . Then*

$$\mu(\{a\}) = F(a) - F(a-),$$

*for any  $a \in \mathbb{R}$ . In particular,  $\mu(\{a\}) = 0$  iff  $F$  is continuous at  $a$ .*

PROOF. The second assertion is immediate by the first one. The first assertion follows from direct computation:

$$\begin{aligned} \mu(\{a\}) &= \lim_{n \rightarrow \infty} \mu\left(\left(a - \frac{1}{n}, a\right]\right) = \lim_{n \rightarrow \infty} \left(F(a) - F\left(a - \frac{1}{n}\right)\right) \\ &= F(a) - F(a-), \end{aligned}$$

where for the first equality we need to use Exercise 2.9. □

3.7. COROLLARY. *Let  $F : \mathbb{R} \rightarrow \mathbb{R}$  be increasing and continuous. Let  $\mu$  be the Lebesgue-Stieltjes measure associated with  $F$ . Then  $\mu(A) = 0$  for any finite or countably infinite set  $A$ , and*

$$\mu((a, b)) = \mu((a, b]) = \mu([a, b)) = \mu([a, b])$$

*for any  $a, b \in \mathbb{R}$  such that  $a < b$ .*

PROOF. For the first assertion, note that such a set can be expressed as a union of finitely many or countably infinitely many singletons, all of which

have measure 0 by the preceding proposition. Now apply finite or countable additivity of  $\mu$ .

For the second assertion, note that the four intervals differ from each other by a set of one or two points, which have measure 0 by the first assertion. Thus the desired equalities follow.  $\square$

We now introduce the famous Lebesgue measure.

3.1. EXAMPLE. Let  $F(x) = x$  for any  $x \in \mathbb{R}$ . We write the measure associated with it as  $m$  and call it the **Lebesgue measure** on  $\mathbb{R}$ . The Lebesgue measure of an interval equals its “natural” length: for any  $a, b \in \mathbb{R}$  with  $a < b$ ,

$$m((a, b)) = m((a, b]) = m([a, b)) = m([a, b]) = b - a.$$

In this spirit, we may say that a general Lebesgue-Stieltjes measure gives twisted length of intervals, using a twisted ruler  $F$ .

Below is an example that illustrates why we need rigorous mathematics—when things get complex, intuition just doesn’t work!

3.2. EXAMPLE. Let  $(r_n)_{n \in \mathbb{N}}$  be an enumeration of all rational numbers. Consider the set

$$E = \bigcup_{n \in \mathbb{N}} \left( r_n - \frac{1}{2^n}, r_n + \frac{1}{2^n} \right).$$

Clearly,  $E \in \mathcal{B}$ . Since the rational numbers are dense in  $\mathbb{R}$  and at every rational number, we circle an interval, one may suspect that  $E = \mathbb{R}$ . But  $E \neq \mathbb{R}$ ! In fact, far from that:

$$m(E) \leq \sum_{n=1}^{\infty} m\left(\left(r_n - \frac{1}{2^n}, r_n + \frac{1}{2^n}\right)\right) = \sum_{n=1}^{\infty} \frac{2}{2^n} = 2.$$

Finally, let’s look at the probability case.

3.8. PROPOSITION. *Let  $F : \mathbb{R} \rightarrow \mathbb{R}$  be increasing and right continuous. Let  $\mu$  be the Lebesgue-Stieltjes measure associated with  $F$ . Then  $\mu$  is a probability measure iff  $F(\infty) - F(-\infty) = 1$ .*

It is obvious because  $\mu(\mathbb{R}) = F(\infty) - F(-\infty)$ . Note that if we replace  $F$  with  $F + c$  for some constant  $c \in \mathbb{R}$ , the measures constructed will be the same. Thus in this case, by replacing  $F$  with  $F - F(-\infty)$ , we may assume

$$F(-\infty) = 0, \quad F(\infty) = 1.$$

3.9. DEFINITION. A function  $F : \mathbb{R} \rightarrow \mathbb{R}$  that is increasing, right continuous and satisfies  $F(-\infty) = 0$  and  $F(\infty) = 1$  is called a **distribution function**.

3.3. EXAMPLE. Let  $a \in \mathbb{R}$  be fixed. Suppose  $F(x) = 0$  if  $x < a$  and  $F(x) = 1$  if  $x \geq a$ . Let  $\mu$  be the Lebesgue-Stieltjes measure associated with  $F$ . Then  $\mu(\mathbb{R}) = 1$ . Moreover, by Proposition 3.6,  $\mu(\{a\}) = F(a) - F(a-) = 1 - 0 = 1$ . It follows that  $\mu(\mathbb{R} \setminus \{a\}) = \mu(\mathbb{R}) - \mu(\{a\}) = 1 - 1 = 0$ . Thus  $\mu$  is just the Dirac measure at  $a$ .

### Exercises

- 3.1. Show that Proposition 1.3 holds for an algebra  $\mathcal{A}$ .
- 3.2. Show that a pre-measure is finitely additive and increasing.
- 3.3. Show that  $\mu$  constructed in (3.2) satisfies (3.1) and the countable sub-additivity.
- 3.4. Let  $\mathcal{A}$  be as in Section 2. Show that if two measures on  $\mathcal{B}$  satisfy (3.4) then they agree on  $\mathcal{A}$ .
- 3.5. Show that every Lebesgue-Stieltjes measure is  $\sigma$ -finite.
- 3.6. Let  $\mu$  be a measure on  $(\mathbb{R}, \mathcal{B})$  such that every bounded interval has finite measure. Show that there exists a function  $F : \mathbb{R} \rightarrow \mathbb{R}$  that is increasing and right continuous such that  $\mu$  is the Lebesgue-Stieltjes measure associated with  $F$ . Any two such functions differ by a constant.

3.7. Let  $\underline{\mu}$  be the pre-measure for  $F$ . Show that for any  $E \in \mathcal{B}$ ,

$$\begin{aligned}
 & \inf \left\{ \sum_{n=1}^{\infty} \underline{\mu}(A_n) : A_n \in \mathcal{A} \text{ for each } n \in \mathbb{N}, E \subset \bigcup_{n=1}^{\infty} A_n \right\} \\
 &= \inf \left\{ \sum_{n=1}^{\infty} \underline{\mu}(A_n) : (A_n)_{n \in \mathbb{N}} \text{ is a disjoint sequence in } \mathcal{A}, E \subset \bigcup_{n=1}^{\infty} A_n \right\} \\
 &= \inf \left\{ \sum_{n=1}^{\infty} \underline{\mu}(I_n) : \text{each } I_n \text{ is an interval of the designated form,} \right. \\
 & \quad \left. E \subset \bigcup_{n=1}^{\infty} I_n \right\} \\
 &= \inf \left\{ \sum_{n=1}^{\infty} \underline{\mu}(I_n) : (I_n) \text{ is a disjoint sequence of intervals of the designated form,} \right. \\
 & \quad \left. E \subset \bigcup_{n=1}^{\infty} I_n \right\}.
 \end{aligned}$$

3.8. Let

$$F(x) = \begin{cases} 0 & \text{if } x < -4, \\ 0.2 & \text{if } -4 \leq x < -1, \\ 0.6 & \text{if } -1 \leq x < 3, \\ 1 & \text{if } x \geq 3. \end{cases}$$

Express the associated Lebesgue-Stieltjes as a convex combination of Dirac measures.



## CHAPTER 4

### Random Variables

Suppose that the stock price of a company at noon is \$50 per share, and let  $X$  be the price tomorrow noon.  $X$  should be viewed as a function defined on  $\Omega$ , where  $\Omega$  is the set of all scenarios that are possible at tomorrow noon. For example, if  $\omega_1$  is the scenario that the company announces a technological innovation by tomorrow noon and  $\omega_2$  is the scenario that an opponent company announces a technological innovation by tomorrow noon, then  $X$  clearly takes different values at them. Intuitively,  $X$  represents the numerical consequences of the uncertainties of the company's future. The sets on which  $X$  takes certain values usually have practical meaning and are of central importance. For example,  $\{\omega \in \Omega : X(\omega) \leq 40\}$  is the event that the stock price goes down at least \$10 per share, or in another word, an investor holding the stock has a loss of at least \$10 per share.

To avoid repetitions, throughout this chapter,  $\Omega$  stands for an arbitrary non-empty set  $\Omega$  and  $\mathcal{F}$  stands for an arbitrary  $\sigma$ -algebra over it.

#### 1. Definition and characterizations

If a function  $X$  defined on  $\Omega$  stands for numerical consequences of a random phenomenon that one is studying, then as is alluded earlier, sets of the form  $\{X \leq c\}$  are events that have important practical meaning. They thus should belong to  $\mathcal{F}$ , the collection of *all* the events that are under care.

4.1. DEFINITION. A function  $X : \Omega \rightarrow \mathbb{R}$  is said to be  ***$\mathcal{F}$ -measurable***, or simply ***measurable***, if  $\{X \leq c\} \in \mathcal{F}$  for every  $c \in \mathbb{R}$ . In probabilistic terms, we also call a measurable function a ***random variable***.

The simplest example of random variables is indicator functions.

4.1. EXAMPLE. Let  $E$  be a subset of  $\Omega$ . Define the ***indicator function*** of  $E$ ,  $\mathbb{1}_E : \Omega \rightarrow \mathbb{R}$ , by

$$\mathbb{1}_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E, \\ 0 & \text{if } \omega \notin E. \end{cases}$$

If  $\mathbb{1}_E$  is a measurable function, then  $E^c = \{\mathbb{1}_E \leq 0\} \in \mathcal{F}$ , so that  $E \in \mathcal{F}$ . Conversely, suppose  $E \in \mathcal{F}$ . Then

$$\{\mathbb{1}_E \leq c\} = \begin{cases} \emptyset & \text{if } c < 0, \\ E & \text{if } 0 \leq c < 1, \\ \Omega & \text{if } c \geq 1. \end{cases}$$

Thus  $\{\mathbb{1}_E \leq c\} \in \mathcal{F}$  for any  $c \in \mathbb{R}$ , and  $\mathbb{1}_E$  is measurable.

This example can be made general.

4.2. EXAMPLE. Let  $X$  be a function on  $\Omega$  that assumes only finitely many distinct values, say,  $c_1 < c_2 < \dots < c_n$ . For  $k = 1, \dots, n$ , put

$$E_k := \{X = c_k\},$$

the set where  $X$  takes the value  $c_k$ . Then  $(E_k)_{1 \leq k \leq n}$  is a partition of  $\Omega^1$ . Moreover, one easily verifies that

$$(4.1) \quad X = c_1 \mathbb{1}_{E_1} + c_2 \mathbb{1}_{E_2} + \dots + c_n \mathbb{1}_{E_n}.$$

If  $X$  is measurable, then  $E_1 = \{X \leq c_1\} \in \mathcal{F}$ , and for  $k = 2, \dots, n$ ,

$$E_k = \{X \leq c_k\} \setminus \{X \leq c_{k-1}\} \in \mathcal{F}.$$

On the other hand, for any  $c \in \mathbb{R}$ ,  $\{X \leq c\}$  collects all the  $\omega$  where  $X$  takes a value at most  $c$ . Thus

$$\{X \leq c\} = \bigcup_{k: c_k \leq c} E_k.$$

Therefore, if  $E_k$ 's are all measurable then  $X$  is measurable.

Note that in (4.1), if some  $c_k$  is zero, one may write off the term  $c_k \mathbb{1}_{E_k}$  from the expression; e.g., instead of writing  $0 \mathbb{1}_{E^c} + 1 \mathbb{1}_E$ , we simply write  $\mathbb{1}_E$ . See Exercise 4.1 for an extension of the example to the countably-infininitely-many-valued case. Such functions will be of critical importance for future developments, so we give them a name.

4.2. DEFINITION. A measurable function that takes only finitely many distinct values is called a **simple function**. In probabilistic terms, a random variable that takes only finitely many or countably infinitely many distinct values is called a **discrete random variable**.

The following proposition provides equivalent forms of measurability.

---

<sup>1</sup>That is,  $E_k$ 's are disjoint and their union is  $\Omega$ .



4.3. PROPOSITION. *Let  $X : \Omega \rightarrow \mathbb{R}$  be a function. The following statements are equivalent:*

- (a)  $X$  is measurable;
- (b)  $\{X > c\} \in \mathcal{F}$  for every  $c \in \mathbb{R}$ ;
- (c)  $\{X \geq c\} \in \mathcal{F}$  for every  $c \in \mathbb{R}$ ;
- (d)  $\{X < c\} \in \mathcal{F}$  for every  $c \in \mathbb{R}$ ;
- (e)  $\{X \in B\} \in \mathcal{F}$  for every  $B \in \mathcal{B}$ .

PROOF. Suppose (a) holds. Then for any  $c \in \mathbb{R}$ ,  $\{X \leq c\} \in \mathcal{F}$ . Thus since  $\mathcal{F}$  is a  $\sigma$ -algebra, it follows that

$$\{X > c\} = \{X \leq c\}^c \in \mathcal{F}.$$

This proves (a)  $\implies$  (b). Similarly, (b)  $\implies$  (c) follows from

$$\{X \geq c\} = \bigcap_{n=1}^{\infty} \left\{X > c - \frac{1}{n}\right\}.$$

(c)  $\implies$  (d) follows from  $\{X < c\} = \{X \geq c\}^c$ . (e)  $\implies$  (a) follows from  $(-\infty, c] \in \mathcal{B}$  and

$$\{X \leq c\} = \{X \in (-\infty, c]\} \in \mathcal{F}.$$

Finally, suppose (d) holds. Let  $\mathcal{G}$  be the collection of all subsets of  $\mathbb{R}$  whose pre-image under  $X$  belongs to  $\mathcal{F}$ . Namely,

$$\mathcal{G} := \{A \subset \mathbb{R} : \{X \in A\} \in \mathcal{F}\}.$$

For every  $c \in \mathbb{R}$ , we have

$$\{X \in (-\infty, c)\} = \{X < c\} \in \mathcal{F}.$$

Thus  $(-\infty, c) \in \mathcal{G}$  for every  $c \in \mathbb{R}$ . Recall that these intervals generate the Borel algebra  $\mathcal{B}$ . Thus if we can show that  $\mathcal{G}$  is a  $\sigma$ -algebra, then  $\mathcal{B} \subset \mathcal{G}$  (Remark 1.5), and (e) follows. Let's verify that  $\mathcal{G}$  is a  $\sigma$ -algebra. Clearly,  $\{X \in \mathbb{R}\} = \Omega \in \mathcal{F}$ , implying that  $\mathbb{R} \in \mathcal{G}$ . Take any  $A \in \mathcal{G}$ . Then

$$\{X \in A^c\} = \Omega \setminus \{X \in A\} \in \mathcal{F}.$$

Consequently,  $A^c \in \mathcal{G}$ . Finally, let  $(A_n)_{n \in \mathbb{N}}$  be any sequence in  $\mathcal{G}$ . Then

$$\left\{X \in \bigcap_{n=1}^{\infty} A_n\right\} = \bigcap_{n=1}^{\infty} \{X \in A_n\} \in \mathcal{F}.$$

It follows that  $\bigcap_{n=1}^{\infty} A_n \in \mathcal{G}$ . This proves that  $\mathcal{G}$  is a  $\sigma$ -algebra.  $\square$

4.4. REMARK. One may regard any of the other statements in Proposition 4.3 as definition of measurability.

The following example provides further intuition for measurability.

4.3. EXAMPLE. Let  $\{A, B, C\}$  be a partition of  $\Omega$ , and  $\mathcal{F} = \sigma(\{A, B, C\})$ . Suppose that  $X : \Omega \rightarrow \mathbb{R}$  is  $\mathcal{F}$ -measurable. We claim that  $X$  must be constant on each of  $A, B, C$ . Suppose otherwise that  $X$  takes at least two different values, say, on  $A$ . Then there exists  $\omega_1, \omega_2 \in A$  such that  $X(\omega_1) < X(\omega_2)$ . We can now tear up  $A$  into two parts:  $A \cap \{X \leq X(\omega_1)\}$  and  $A \cap \{X > X(\omega_1)\}$ . They are disjoint, non-empty, and both belong to  $\mathcal{F}$  by measurability of  $\mathcal{F}$ . This is impossible, since every member in  $\mathcal{F}$  is the union of some of  $A, B, C$  (cf. Exercise 1.3).

The notion of measurability can be extended to multiple dimension. Recall first the following notation:

$$\begin{aligned} & \{X_1 \leq c_1, X_2 \leq c_2, \dots, X_d \leq c_d\} \\ & \triangleq \{\omega \in \Omega : X_1(\omega) \leq c_1, X_2(\omega) \leq c_2, \dots, X_d(\omega) \leq c_d\} \\ & = \bigcap_{k=1}^d \{\omega \in \Omega : X_k(\omega) \leq c_k\} = \bigcap_{k=1}^d \{X_k \leq c_k\}. \end{aligned}$$

4.5. DEFINITION. Let  $d \in \mathbb{N}$ . Let  $(X_1, X_2, \dots, X_d) : \Omega \rightarrow \mathbb{R}^d$  be a function. It is said to be measurable if  $\{X_1 \leq c_1, X_2 \leq c_2, \dots, X_d \leq c_d\} \in \mathcal{F}$  for any  $c_1, \dots, c_d \in \mathbb{R}$ . In probabilistic terms, we may call it a **(d-dimensional) random vector**.

4.6. PROPOSITION. For a function  $(X_1, X_2, \dots, X_d) : \Omega \rightarrow \mathbb{R}^d$ , the following are equivalent:

- (a)  $(X_1, X_2, \dots, X_d)$  is measurable;
- (b) Each  $X_k$ ,  $1 \leq k \leq d$ , is measurable;
- (c)  $\{(X_1, X_2, \dots, X_d) \in B\} \in \mathcal{F}$  for every  $B \in \mathcal{B}^d$

PROOF. One may prove in the following order: (a)  $\implies$  (c)  $\implies$  (b)  $\implies$  (a). For example, suppose (c) holds. For any  $c \in \mathbb{R}$ , since  $(-\infty, c] \times \mathbb{R}^{d-1} \in \mathcal{B}^d$ ,

$$\{X_1 \leq c\} = \{(X_1, X_2, \dots, X_d) \in (-\infty, c] \times \mathbb{R}^{d-1}\} \in \mathcal{F}.$$

Thus  $X_1$  is measurable. Similar arguments work for other  $X_k$ 's. Hence, (c)  $\implies$  (b). We leave the proof of other implications to the reader.  $\square$

## 2. Elementary properties

When considering functions on  $\mathbb{R}^d$ , we usually endow  $\mathbb{R}^d$  with  $\mathcal{B}^d$ , and call a  $\mathcal{B}^d$ -measurable function  $h : \mathbb{R}^d \rightarrow \mathbb{R}$  a **Borel** measurable function, or simply measurable if no ambiguities could possibly arise.

4.7. PROPOSITION. *Let  $(X_1, X_2, \dots, X_d) : \Omega \rightarrow \mathbb{R}^d$  and  $h : \mathbb{R}^d \rightarrow \mathbb{R}$  be measurable. Then  $h(X_1, X_2, \dots, X_d) : \Omega \rightarrow \mathbb{R}$  is also measurable.*

PROOF. Take any  $B \in \mathcal{B}$ . By Proposition 4.3,  $h^{-1}(B) = \{h \in B\} \in \mathcal{B}^d$ . Thus by Proposition 4.6,

$$\{h(X_1, X_2, \dots, X_d) \in B\} = \{(X_1, X_2, \dots, X_d) \in h^{-1}(B)\} \in \mathcal{F}.$$

By Proposition 4.3 again,  $h(X_1, X_2, \dots, X_d)$  is measurable.  $\square$

In applications of this proposition, it happens often that  $h$  is continuous. We thus need the following result.

4.8. PROPOSITION. *Continuous functions are Borel-measurable.*

This result looks quite expected but its proof is very non-trivial and uses the notion of open sets. We put the proof in Appendix B.

4.4. EXAMPLE. Let  $X : \Omega \rightarrow \mathbb{R}$  be measurable. Then  $|X|, X^+, X^-, e^X$  are measurable. Indeed, take  $h(t) = |t|$  for every  $t \in \mathbb{R}$ . Then  $h$  is continuous on  $\mathbb{R}$  and is thus Borel-measurable by Propositions 4.8. Thus,  $|X| = h(X)$  is measurable by Proposition 4.7. One similarly proves measurability of the other functions. We can also establish measurability of these functions without using Proposition 4.7. For example, one verifies that

$$\{X^+ \leq c\} = \begin{cases} \emptyset & \text{if } c < 0, \\ \{X \leq c\} & \text{if } c \geq 0. \end{cases}$$

The following result demonstrates more power of Proposition 4.7.

4.9. COROLLARY. *Let  $X, Y : \Omega \rightarrow \mathbb{R}$  be measurable and  $a, b \in \mathbb{R}$ . Then  $aX + bY$  and  $XY$  are measurable.*

PROOF. Define  $h : \mathbb{R}^2 \rightarrow \mathbb{R}$  by  $h(t, s) = at + bs$  and simply note that  $aX + bY = h(X, Y)$ . For the product, define  $h(t, s) = ts$ .  $\square$

4.10. REMARK. Even if  $X, Y$  are extended-valued, measurability of  $aX + bY$  is still valid as long as  $aX + bY$  is well-defined (i.e.,  $(-\infty) + \infty$ ,  $\infty + (-\infty)$ ,  $(-\infty) - (-\infty)$  and  $\infty - \infty$  do not appear). But we need to put a bit extra

care. Let's show the case where  $a > 0$  but  $b < 0$ . Let  $X_1 = X\mathbb{1}_{\{-\infty < X < \infty\}}$  and  $Y_1 = Y\mathbb{1}_{\{-\infty < Y < \infty\}}$ . Basically,  $X_1$  knocks  $X$  down to 0 when it is  $\infty$  or  $-\infty$ . It is thus easy to see that

$$\{X_1 \leq c\} = \begin{cases} \{X \leq c\} \setminus \{X = -\infty\} & \text{if } c < 0, \\ \{X \leq c\} \cup \{X = \infty\} & \text{if } c \geq 0. \end{cases}$$

Thus  $X_1$  is measurable. Similarly, so is  $Y_1$ . Thus  $aX_1 + bY_1$  is measurable by Corollary 4.9. Take any  $c \in \mathbb{R}$ . For  $\omega \in \Omega$ , if  $X(\omega) \neq \pm\infty$  and  $Y(\omega) \neq \pm\infty$ , then  $X(\omega) = X_1(\omega)$  and  $Y(\omega) = Y_1(\omega)$ , and thus  $aX(\omega) + bY(\omega) = aX_1(\omega) + bY_1(\omega)$ . It follows that

$$\begin{aligned} & \{aX + bY \leq c\} \cap \{X \neq \pm\infty, Y \neq \pm\infty\} \\ &= \{aX + bY \leq c, X \neq \pm\infty, Y \neq \pm\infty\} \\ &= \{aX_1 + bY_1 \leq c, X \neq \pm\infty, Y \neq \pm\infty\} \\ &= \{aX_1 + bY_1 \leq c\} \cap \{X \neq \pm\infty\} \cap \{Y \neq \pm\infty\} \in \mathcal{F}. \end{aligned}$$

Moreover, if  $aX + bY \leq c$ , and if  $X = \pm\infty$  or  $Y = \pm\infty$ , then  $X = -\infty$  or  $Y = \infty$ . Thus

$$\begin{aligned} & \{aX + bY \leq c\} \setminus \{X \neq \pm\infty, Y \neq \pm\infty\} \\ &= \{aX + bY \leq c\} \cap (\{X = \pm\infty\} \cup \{Y \neq \pm\infty\}) \\ &= \{X = -\infty\} \cup \{Y = \infty\} \in \mathcal{F}. \end{aligned}$$

Since  $\{aX + bY \leq c\}$  is the union of the first terms in the two equations above, it is in  $\mathcal{F}$  as well. Thus  $aX + bY$  is measurable.

**4.5. EXAMPLE.** Corollary 4.9 can be proved directly using definition of measurability as well. Let's demonstrate it for  $XY$  when  $X, Y \geq 0$ . If  $c \leq 0$ , then  $\{XY < c\} = \emptyset \in \mathcal{F}$ . Now take any  $c > 0$ . Take any  $\omega \in \{XY < c\}$ . If  $X(\omega) = 0$ , no problem. If  $X(\omega) > 0$ , then  $Y(\omega) < \frac{c}{X(\omega)}$ . Take a rational number  $r > 0$  such that

$$Y(\omega) < r < \frac{c}{X(\omega)},$$

i.e.,  $Y(\omega) < r$  and  $X(\omega) < \frac{c}{r}$ . From these arguments, one sees that

$$\{XY < c\} = \{X = 0\} \cup \bigcup_{r > 0 \text{ rational}} \left( \{Y < r\} \cap \left\{X < \frac{c}{r}\right\} \right).$$

Each of the sets in the right hand side lies in  $\mathcal{F}$ , and since there are countable positive rational numbers, the last union is a countable union. Consequently,

$\{XY < c\} \in \mathcal{F}$ . This proves that  $XY$  is measurable as desired. We leave the proofs of other cases to the reader as exercises.

A third approach to Corollary 4.9 is to apply the following result on measurability of the limit of a sequence of measurable functions and Theorem 4.12; see Exercise 4.11.

**4.11. PROPOSITION.** *Let  $X_n : \Omega \rightarrow \mathbb{R}$  be measurable for each  $n \in \mathbb{N}$ . Then  $\sup_{n \in \mathbb{N}} X_n$ ,  $\inf_{n \in \mathbb{N}} X_n$ ,  $\limsup_{n \rightarrow \infty} X_n$ , and  $\liminf_{n \rightarrow \infty} X_n$  are all measurable.*

**PROOF.** For every  $c \in \mathbb{R}$ , since  $\sup_{n \in \mathbb{N}} X_n \leq c$  iff  $X_n \leq c$  for every  $n \in \mathbb{N}$ , it follows that

$$\left\{ \sup_{n \in \mathbb{N}} X_n \leq c \right\} = \bigcap_{n \in \mathbb{N}} \{X_n \leq c\} \in \mathcal{F}.$$

Thus  $\sup_{n \in \mathbb{N}} X_n$  is measurable. The case of  $\inf_{n \in \mathbb{N}} X_n$  is left to the reader.

Set  $Y_n = \sup_{m \geq n} X_m$  for  $n \in \mathbb{N}$ . Then every  $Y_n$  is measurable by the sup case we just proved. Thus  $\limsup_{n \rightarrow \infty} X_n = \inf_{n \in \mathbb{N}} Y_n$  is measurable. The case of  $\liminf_{n \rightarrow \infty} X_n$  is also left to the reader.  $\square$

### 3. Approximation by simple functions

The following result is of central importance in many developments in what follows. As will be seen soon, it is often used to reduce arguments from general measurable functions to simple functions.

**4.12. THEOREM.** *Let  $X$  be a non-negative measurable function on  $\Omega$ . Then there exists a sequence  $(\phi_n)_{n=1}^\infty$  of simple functions such that  $0 \leq \phi_n \uparrow X$  on  $\Omega$ .*

**PROOF.** To illustrate the idea, we first prove the theorem under the assumption that  $0 \leq X < 1$  on  $\Omega$ . Fix  $n \in \mathbb{N}$ . We cut  $[0, 1)$  into  $2^n$  small intervals:

$$\left[ \frac{k-1}{2^n}, \frac{k}{2^n} \right), \quad k = 1, 2, \dots, 2^n,$$

and thus cut  $\Omega$  into  $2^n$  subsets:

$$\left\{ \frac{k-1}{2^n} \leq X < \frac{k}{2^n} \right\}, \quad k = 1, 2, \dots, 2^n.$$

Now define

$$\phi_n = \sum_{k=1}^{2^n} \frac{k-1}{2^n} \mathbb{1}_{\left\{ \frac{k-1}{2^n} \leq X < \frac{k}{2^n} \right\}}.$$

On the set  $\{\frac{k-1}{2^n} \leq X < \frac{k}{2^n}\}$ , the value of  $X$  is floored by  $\frac{k-1}{2^n}$  and capped by  $\frac{k}{2^n}$ —total room of oscillation of  $X$  is smaller than  $\frac{k}{2^n} - \frac{k-1}{2^n} = \frac{1}{2^n}$ . Furthermore, on this set,  $\phi_n$  takes the floor value of  $\frac{k-1}{2^n}$ . Thus one sees that if  $\omega \in \{\frac{k-1}{2^n} \leq X < \frac{k}{2^n}\}$ , then

$$0 \leq X(\omega) - \phi_n(\omega) < \frac{1}{2^n}.$$

Since every  $\omega \in \Omega$  belongs to such a set,  $0 \leq X - \phi_n < \frac{1}{2^n}$  everywhere on  $\Omega$ . It follows that  $0 \leq \phi_n \leq X$  and  $\lim_n \phi_n = X$  everywhere on  $\Omega$ .

We verify that  $\phi_n \leq \phi_{n+1}$  everywhere on  $\Omega$ . Pick any  $\omega \in \Omega$ . Say,  $\omega \in \{\frac{k-1}{2^n} \leq X < \frac{k}{2^n}\}$  for some  $k = 1, \dots, 2^n$ . Then  $\phi_n(\omega) = \frac{k-1}{2^n}$ . Note that when defining  $\phi_{n+1}$ , we cut  $[0, 1)$  into intervals of the form  $[\frac{l-1}{2^{n+1}}, \frac{l}{2^{n+1}})$ . Thus, since  $\frac{k-1}{2^n} = \frac{2k-2}{2^{n+1}}$  and  $\frac{k}{2^n} = \frac{2k}{2^{n+1}}$ , the set  $\{\frac{k-1}{2^n} \leq X < \frac{k}{2^n}\}$  is split into two sets in the  $(n+1)$ -th level when defining  $\phi_{n+1}$ :

$$\left\{ \frac{2k-2}{2^{n+1}} \leq X < \frac{2k-1}{2^{n+1}} \right\} \cup \left\{ \frac{2k-1}{2^{n+1}} \leq X < \frac{2k}{2^{n+1}} \right\}.$$

If  $\omega$  lies in the first set, then

$$\phi_{n+1}(\omega) = \frac{2k-2}{2^{n+1}} = \phi_n(\omega);$$

if  $\omega$  lies in the second set, then

$$\phi_{n+1}(\omega) = \frac{2k-1}{2^{n+1}} > \phi_n(\omega).$$

Since  $\omega$  is arbitrary, this proves that  $\phi_n \leq \phi_{n+1}$  everywhere on  $\Omega$ .

Now we prove the theorem in the general case. We cut  $[0, \infty)$ , the range of  $X$ , according to the following scheme:

- $[0, 1)$ : cut it into  $2^n$  small intervals of equal length  $\frac{1}{2^n}$ ,
- $[1, 2)$ : cut it into  $2^n$  small intervals of equal length  $\frac{1}{2^n}$ ,
- $\dots$
- $[n-1, n)$ : cut it into  $2^n$  small intervals of equal length  $\frac{1}{2^n}$ ,
- $[n, \infty)$ .

Then in total, we have a big interval  $[n, \infty)$  and  $n2^n$  small intervals of length  $\frac{1}{2^n}$ , which are precisely  $[\frac{k-1}{2^n}, \frac{k}{2^n})$ ,  $k = 1, \dots, n2^n$ . Now cut  $\Omega$  accordingly

$$\Omega = \bigcup_{k=1}^{n2^n} \left\{ \frac{k-1}{2^n} \leq X < \frac{k}{2^n} \right\} \cup \{X \geq n\},$$

and set the value of  $\phi_n$  on each set as the floor value there. Namely,

$$(4.2) \quad \phi_n = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \mathbb{1}_{\{\frac{k-1}{2^n} \leq X < \frac{k}{2^n}\}} + n \mathbb{1}_{\{X \geq n\}}.$$

Take an arbitrary  $\omega \in \Omega$ . Pick any  $n \in \mathbb{N}$ . We consider two cases:

Case 1.  $X(\omega) < n$ .

In this case,  $X(\omega) \in [\frac{k-1}{2^n}, \frac{k}{2^n})$  for some  $k = 1, \dots, n2^n$ . Hence,  $\omega \in \{\frac{k-1}{2^n} \leq X < \frac{k}{2^n}\}$  and  $\phi_n(\omega) = \frac{k-1}{2^n}$ . One sees as before that  $\phi_{n+1}(\omega) = \frac{2k-2}{2^{n+1}}$  or  $\frac{2k-1}{2^{n+1}}$ , implying that  $\phi_n(\omega) \leq \phi_{n+1}(\omega)$ .

Case 2.  $X(\omega) \geq n$ .

In this case,  $X(\omega) \in [n, \infty)$ , or  $\omega \in \{X \geq n\}$ . Clearly,  $\phi_n(\omega) = n$ , the floor of  $[n, \infty)$ . When defining  $\phi_{n+1}$ , we deal with  $\{X \geq n\}$  more deliberately by splitting it as  $\{n \leq X < n+1\} \cup \{X \geq n+1\}$  and then cutting the first set further and letting  $\phi_{n+1}$  take the floor values on each set. One sees that all these floor values are at least  $n$ . Thus no matter where  $\omega$  lies,  $\phi_{n+1}(\omega) \geq n = \phi_n(\omega)$ .

Combining the above two cases, one sees that  $\phi_n \leq \phi_{n+1}$  everywhere.

Pick any  $\omega \in \Omega$ . For every  $n > X(\omega)$ , when defining  $\phi_n$ ,  $\omega$  falls into a set appearing in the summation part of (4.2). Thus as before, one sees that

$$0 \leq X(\omega) - \phi_n(\omega) < \frac{1}{2^n}.$$

Letting  $n \rightarrow \infty$ , it follows again that  $\lim_n \phi_n(\omega) = X(\omega)$ .  $\square$

### Exercises

4.1. Let  $X : \Omega \rightarrow \mathbb{R}$  be a function that assumes countably infinitely many distinct values. Show that there exist a sequence  $(c_k)_{k \in \mathbb{N}}$  of distinct real numbers and a disjoint sequence  $(E_k)_{k \in \mathbb{N}}$  of subsets of  $\Omega$  such that  $X = \sum_{k=1}^{\infty} c_k \mathbb{1}_{E_k}$ . Show that  $X$  is measurable iff each  $E_k$  is measurable.

4.2. Complete the proof of Proposition 4.6.

4.3. Let  $A_n$ 's and  $\mathcal{F}$  be as in Exercise 1.3. Show that a function  $X : \Omega \rightarrow \mathbb{R}$  is measurable iff  $X$  is constant on each  $A_n$ .

4.4. Complete the proofs in Example 4.4 using Proposition 4.7.

4.5. Complete the proofs in Example 4.4 without using Proposition 4.7.

4.6. Let  $X, Y : \Omega \rightarrow \mathbb{R}$  be measurable and  $a, b \in \mathbb{R}$ . Directly use definition of measurability to show that  $aX$  and  $X - Y$  are measurable. Conclude that  $aX + bY$  is measurable.

4.7. Let  $X, Y : \Omega \rightarrow \mathbb{R}$  be measurable. Show that  $XY$  is measurable.

4.8. Let  $f : \Omega \rightarrow \mathbb{R}$  be measurable and  $f$  is nonzero everywhere. Use definition of measurability to show that  $\frac{1}{f}$  is measurable. Conclude that  $\frac{g}{f}$  is measurable for any measurable function  $g : \Omega \rightarrow \mathbb{R}$ .

4.9. Complete the proof of Proposition 4.11.

4.10. Let  $X : \Omega \rightarrow \mathbb{R}$  be measurable. Find a sequence  $(\phi_n)_{n \in \mathbb{N}}$  of simple functions such that  $\phi_n \rightarrow X$  and  $|\phi_n| \leq |X|$  for every  $n \in \mathbb{N}$  on  $\Omega$ .

4.11. Prove Exercises 4.6 and 4.7 by showing them for simple functions first and then applying Exercise 4.10 and Proposition 4.11.

4.12. Show that an increasing function  $X : \mathbb{R} \rightarrow \mathbb{R}$  is measurable.

4.13. Let  $X, X_n, n \in \mathbb{N}$  be measurable functions on  $\Omega$ . Show that the set  $\{\omega \in \Omega : (X_n(\omega))_n \text{ converges to } X(\omega)\}$  is measurable.

4.14. Let  $(A_n)_{n \in \mathbb{N}}$  be a disjoint sequence of measurable sets. Show that  $(\mathbb{1}_{A_n})$  converges to 0 on  $\Omega$ .



## CHAPTER 5

### Expectations I

Suppose that we are in a gambling game. Let  $X$  denote the gain if one plays the game once. Suppose that with a probability of  $\frac{1}{3}$ , we win \$15, i.e.,  $X = 15$ ; with a probability of  $\frac{1}{2}$ ,  $X = 10$ , and with a probability of  $\frac{1}{6}$ ,  $X = -6$ . What we “expect” about our future if we decide to play the game once? Intuitively, our expectation should be the possible gains averaged by their chances of occurrence, namely,  $(15)\frac{1}{3} + (10)\frac{1}{2} + (-6)\frac{1}{6} = 10$ . In this chapter, we extend this naive definition of expectations from simple functions to general random variables.

Throughout this chapter,  $(\Omega, \mathcal{F}, \mathbb{P})$  stands for a fixed but arbitrary probability space. Moreover, *for the rest of the book, all sets and functions involved are assumed to be measurable, unless specified otherwise.*

#### 1. Expectations of simple functions

Imitating the example above, we make the following definition.

5.1. DEFINITION. *Let  $\phi$  be a simple function on  $\Omega$ , say,*

$$(5.1) \quad \phi = \sum_{k=1}^n c_k \mathbb{1}_{E_k},$$

*where  $c_k$ 's are all the distinct values that  $\phi$  assumes (and thus  $E_k$ 's are a partition of  $\Omega$ ). Put*

$$(5.2) \quad \mathbb{E}[\phi] := \sum_{k=1}^n c_k \mathbb{P}(E_k),$$

*and call it the **expectation** of  $\phi$ . Clearly,  $E_k = \{X = c_k\}$  since we assume that  $c_k$ 's are distinct. Thus we can rewrite  $\mathbb{E}[\phi]$  as<sup>1</sup>*

$$(5.3) \quad \mathbb{E}[\phi] = \sum_{k=1}^n c_k \mathbb{P}(X = c_k).$$

---

<sup>1</sup>We write  $\mathbb{P}(X \in B)$  instead of  $\mathbb{P}(\{X \in B\})$  for the sake of brevity.

In view of (5.3), we can interpret  $\mathbb{E}[\phi]$  as the “average” value of  $\phi$ , with values of  $\phi$  averaged by their probabilities of occurrence. In view of (5.2), we can interpret  $\mathbb{E}[\phi]$  as “area”: for each  $k$ ,  $\phi$  determines a region of height  $c_k$  and width  $\mathbb{P}(E_k)$ , and thus circling an area of  $c_k\mathbb{P}(E_k)$ .

We need to relax the conditions in the expression (5.1) for convenience of computations later. The first relaxation is as follows.

5.2. REMARK. Unlike (5.1), we may write

$$(5.4) \quad \phi = \sum_{l=1}^m d_l \mathbb{1}_{F_l},$$

where  $F_l$ 's are non-empty and still a partition of  $\Omega$  but we do not require  $d_l$ 's to be distinct. For example, the function

$$2\mathbb{1}_{(-\infty, 1]} - \mathbb{1}_{(1, \infty)}$$

can also be written as

$$2\mathbb{1}_{(-\infty, -3]} + 2\mathbb{1}_{[-3, -2]} + 2\mathbb{1}_{(-2, 1]} - \mathbb{1}_{(1, \infty)}.$$

Clearly, (5.4) is obtained from (5.1) by splitting each  $E_k$  into a few  $F_l$ 's with the heights of  $\phi$  on these  $F_l$ 's,  $d_l$ 's, all equal to  $c_k$ . Since the probability of  $E_k$  equals the sum of probabilities of the  $F_l$ 's that are split from  $E_k$ , it is easy to see that

$$(5.5) \quad \mathbb{E}[\phi] = \sum_{l=1}^m d_l \mathbb{P}(F_l).$$

For notational convenience, we may allow some  $F_l$ 's to be empty in (5.4). Note that (5.5) still holds. Indeed, if  $F_l = \emptyset$ , then  $d_l \mathbb{1}_{F_l} = 0$  on  $\Omega$ , so that the term can be removed from the sum in (5.4), and  $d_l \mathbb{P}(F_l) = 0$ , so that the term can be removed from the sum in (5.5).

The following are fundamental properties of expectations and will be extended to general random variables later.

5.1. LEMMA. *Let  $\phi$  and  $\psi$  be two simple functions. The following hold.*

- (a)  $\mathbb{E}[a\phi + b\psi] = a\mathbb{E}[\phi] + b\mathbb{E}[\psi]$  for any  $a, b \in \mathbb{R}$ ;
- (b)  $\mathbb{E}[\phi] \leq \mathbb{E}[\psi]$  if  $\phi \leq \psi$  on  $\Omega$ .

PROOF. Write  $\phi = \sum_{k=1}^n c_k \mathbb{1}_{E_k}$ , where  $c_k$ 's are all the distinct values of  $\phi$ , and  $\psi = \sum_{l=1}^m d_l \mathbb{1}_{F_l}$ , where  $d_l$ 's are all the distinct values of  $\psi$ . Note that the  $mn$  sets  $E_k \cap F_l$ 's are disjoint with union  $\Omega$  and thus constitute a partition of  $\Omega$ ; see Figure 1 for illustration.

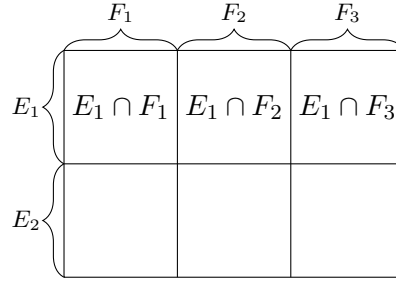


FIGURE 1. Double partition

For any  $k = 1, \dots, n$ , since  $E_k = \bigcup_{l=1}^m (E_k \cap F_l)$ , one verifies that  $\mathbb{1}_{E_k} = \sum_{l=1}^m \mathbb{1}_{E_k \cap F_l}$  (Exercise 5.1). Thus it follows that

$$\phi = \sum_{k=1}^n c_k \sum_{l=1}^m \mathbb{1}_{E_k \cap F_l} = \sum_{k=1}^n \sum_{l=1}^m c_k \mathbb{1}_{E_k \cap F_l}.$$

Similarly,

$$\begin{aligned} \psi &= \sum_{l=1}^m d_l \sum_{k=1}^n \mathbb{1}_{E_k \cap F_l} = \sum_{k=1}^n \sum_{l=1}^m d_l \mathbb{1}_{E_k \cap F_l}, \\ a\phi + b\psi &= \sum_{k=1}^n \sum_{l=1}^m (ac_k + bd_l) \mathbb{1}_{E_k \cap F_l}. \end{aligned}$$

By Remark 5.1, we have

$$\begin{aligned} \mathbb{E}[a\phi + b\psi] &= \sum_{k=1}^n \sum_{l=1}^m (ac_k + bd_l) \mathbb{P}(E_k \cap F_l) \\ &= \sum_{k=1}^n \sum_{l=1}^m ac_k \mathbb{P}(E_k \cap F_l) + \sum_{k=1}^n \sum_{l=1}^m bd_l \mathbb{P}(E_k \cap F_l) \\ &= \sum_{k=1}^n ac_k \sum_{l=1}^m \mathbb{P}(E_k \cap F_l) + \sum_{l=1}^m bd_l \sum_{k=1}^n \mathbb{P}(E_k \cap F_l) \\ &= \sum_{k=1}^n ac_k \mathbb{P}(E_k) + \sum_{l=1}^m bd_l \mathbb{P}(F_l) \\ &= a \sum_{k=1}^n c_k \mathbb{P}(E_k) + b \sum_{l=1}^m d_l \mathbb{P}(F_l) \\ &= a\mathbb{E}[\phi] + b\mathbb{E}[\psi]. \end{aligned}$$

This proves (a). Suppose now that  $\phi \leq \psi$  on  $\Omega$ . For each pair  $(k, l)$ , where  $1 \leq k \leq n$  and  $1 \leq l \leq m$ , if  $E_k \cap F_l = \emptyset$ , then  $c_k \mathbb{P}(E_k \cap F_l) = 0 =$

$d_l \mathbb{P}(E_k \cap F_l)$ . If  $E_k \cap F_l \neq \emptyset$ , then since  $\phi = c_k$  and  $\phi = d_l$  on it,  $c_k \leq d_l$ , implying that  $c_k \mathbb{P}(E_k \cap F_l) \leq d_l \mathbb{P}(E_k \cap F_l)$ . Thus

$$\mathbb{E}[\phi] = \sum_{k=1}^n \sum_{l=1}^m c_k \mathbb{P}(E_k \cap F_l) \leq \sum_{k=1}^n \sum_{l=1}^m d_l \mathbb{P}(E_k \cap F_l) = \mathbb{E}[\psi],$$

by Remark 5.2 again. This proves (b).  $\square$

The following remark continues to relax the conditions in the expression (5.1) to further ease the computation of  $\mathbb{E}[\phi]$ .

**5.3. REMARK.** By induction, it follows from Lemma 5.1 that, for any simple functions  $\phi_1, \dots, \phi_n$ ,  $\mathbb{E}[\sum_{k=1}^n \phi_k] = \sum_{k=1}^n \mathbb{E}[\phi_k]$ . Thus if we write a simple function as  $\phi = \sum_{k=1}^n c_k \mathbb{1}_{E_k}$ , where  $c_k$ 's may not be distinct and  $E_k$ 's may not be disjoint, we still have

$$\begin{aligned} \mathbb{E}[\phi] &= \sum_{k=1}^n \mathbb{E}[c_k \mathbb{1}_{E_k}] = \sum_{k=1}^n \mathbb{E}[c_k \mathbb{1}_{E_k} + 0 \mathbb{1}_{E_k^c}] = \sum_{k=1}^n (c_k \mathbb{P}(E_k) + 0 \mathbb{P}(E_k^c)) \\ &= \sum_{k=1}^n c_k \mathbb{P}(E_k). \end{aligned}$$

**5.2. LEMMA.** Let  $\phi$  and  $\phi_n$ ,  $n \in \mathbb{N}$ , be simple functions such that  $0 \leq \phi_n \uparrow \phi$  on  $\Omega$ . Then  $\mathbb{E}[\phi_n] \uparrow \mathbb{E}[\phi]$ .

**PROOF.** The increasingness of  $\mathbb{E}[\phi_n]$ 's is immediate by Lemma 5.1(b). It remains to be shown that  $\sup_n \mathbb{E}[\phi_n] = \mathbb{E}[\phi]$ .

Let's assume first that  $\phi$  is an indicator function, say,  $\phi = \mathbb{1}_E$ . For any  $\varepsilon \in (0, 1)$ , since  $\phi_n \uparrow \mathbb{1}_E$ , recall that

$$\{\phi_n > 1 - \varepsilon\} \uparrow \{\mathbb{1}_E > 1 - \varepsilon\}.$$

The last set is easily seen to be equal to  $E$ . By Proposition 2.4<sup>2</sup>,

$$(5.6) \quad \mathbb{P}(\phi_n > 1 - \varepsilon) \uparrow \mathbb{P}(E) = \mathbb{E}[\mathbb{1}_E].$$

On the set  $\{\phi_n > 1 - \varepsilon\}$ ,  $\phi_n$  is at least  $1 - \varepsilon$  and  $(1 - \varepsilon) \mathbb{1}_{\{\phi_n > 1 - \varepsilon\}}$  is exactly  $1 - \varepsilon$ ; off the set  $\{\phi_n > 1 - \varepsilon\}$ ,  $\phi_n$  is at least 0 and  $(1 - \varepsilon) \mathbb{1}_{\{\phi_n > 1 - \varepsilon\}}$  is exactly 0. Thus  $\phi_n \geq (1 - \varepsilon) \mathbb{1}_{\{\phi_n > 1 - \varepsilon\}}$  everywhere on  $\Omega$ . By Lemma 5.1(b),

$$\mathbb{E}[\phi_n] \geq \mathbb{E}[(1 - \varepsilon) \mathbb{1}_{\{\phi_n > 1 - \varepsilon\}}] = (1 - \varepsilon) \mathbb{P}(\phi_n > 1 - \varepsilon).$$

This, together with (5.6), implies that

$$\frac{\sup_n \mathbb{E}[\phi_n]}{1 - \varepsilon} \geq \sup_n \mathbb{P}(\phi_n > 1 - \varepsilon) = \mathbb{E}[\mathbb{1}_E].$$

---

<sup>2</sup>This is where countable additivity of  $\mathbb{P}$  is essentially used.

Letting  $\varepsilon \rightarrow 0$ , we obtain

$$\sup_n \mathbb{E}[\phi_n] \geq \mathbb{E}[\mathbb{1}_E].$$

Reversely, for every  $n \in \mathbb{N}$ , since  $\phi_n \leq \mathbb{1}_E$ ,  $\mathbb{E}[\phi_n] \leq \mathbb{E}[\mathbb{1}_E]$  by Lemma 5.1(b) again. Thus  $\mathbb{E}[\mathbb{1}_E] = \sup_n \mathbb{E}[\phi_n]$ , as desired.

Now we prove the general case. If  $\phi = 0$  on  $\Omega$ , then there is nothing to prove, since all the expectations are zero. Otherwise, we can write  $\phi = \sum_{l=1}^m c_l \mathbb{1}_{E_l}$ , where  $E_l$ 's are disjoint and  $c_l > 0$  for each  $l$ . Fix any  $l = 1, \dots, m$ . We have  $\phi_n \mathbb{1}_{E_l} \uparrow_n \phi \mathbb{1}_{E_l} = c_l \mathbb{1}_{E_l}$ . Thus

$$\frac{\phi_n \mathbb{1}_{E_l}}{c_l} \uparrow_n \mathbb{1}_{E_l},$$

and by the case we just proved,  $\mathbb{E}[\frac{\phi_n \mathbb{1}_{E_l}}{c_l}] \uparrow_n \mathbb{E}[\mathbb{1}_{E_l}]$ . By Lemma 5.1(a),

$$\mathbb{E}[\phi_n \mathbb{1}_{E_l}] \uparrow_n \mathbb{E}[c_l \mathbb{1}_{E_l}].$$

Summing over  $l = 1, \dots, m$ , we get by Lemma 5.1(a),

$$\mathbb{E}\left[\phi_n \sum_{l=1}^m \mathbb{1}_{E_l}\right] \uparrow_n \sum_{l=1}^m c_l \mathbb{E}[\mathbb{1}_{E_l}] = \mathbb{E}[\phi].$$

We claim that  $\phi_n \sum_{l=1}^m \mathbb{1}_{E_l} = \phi_n$ . Indeed, simply note that  $\sum_{l=1}^m \mathbb{1}_{E_l} = \mathbb{1}_{\bigcup_{l=1}^m E_l}$  and that outside the set  $\bigcup_{l=1}^m E_l$ ,  $\phi$  is zero and thus  $\phi_n$  is zero as well, since  $0 \leq \phi_n \leq \phi$ . Putting things together, we obtain  $\mathbb{E}[\phi_n] \uparrow_n \mathbb{E}[\phi]$ .  $\square$

## 2. Expectations of general functions

We define expectations of general random variables, by approximating them using simple functions, for which we already have a natural way of defining expectations, as studied in the previous section.

5.4. DEFINITION. (a) *For a non-negative random variable  $X$  on  $\Omega$ , take<sup>3</sup> a sequence  $(\phi_n)_{n \in \mathbb{N}}$  of simple functions such that  $0 \leq \phi_n \uparrow X$  on  $\Omega$  and define the expectation,  $\mathbb{E}[X]$ , of  $X$  by*

$$E[X] := \lim_n \mathbb{E}[\phi_n] = \sup_n \mathbb{E}[\phi_n]^4.$$

(b) *For a general random variable  $X$ , we define its expectation by*

$$\mathbb{E}[X] := \mathbb{E}[X^+] - \mathbb{E}[X^-],$$

*if at least one of  $\mathbb{E}[X^\pm]$  is finite. If  $\mathbb{E}[X^\pm]$  are both infinite, we say that the expectation of  $X$  is undefined.*

<sup>3</sup>By Theorem 4.12, such a sequence always exists.

<sup>4</sup>By Lemma 5.1(b),  $\mathbb{E}[\phi_n] \uparrow$ , so that  $\lim_n \mathbb{E}[\phi_n] = \sup_n \mathbb{E}[\phi_n]$ .

If  $\mathbb{E}[X]$  is defined, then  $\mathbb{E}[X] \in \mathbb{R}$  iff  $\mathbb{E}[X^\pm]$  are both finite,  $\mathbb{E}[X] = \infty$  iff  $\mathbb{E}[X^+] = \infty$  and  $\mathbb{E}[X^-] < \infty$ ,  $\mathbb{E}[X] = -\infty$  iff  $\mathbb{E}[X^+] < \infty$  and  $\mathbb{E}[X^-] = \infty$ . When  $\mathbb{E}[X] \in \mathbb{R}$ , we say that  $X$  is *integrable*.

Clearly, if  $X \geq 0$  on  $\Omega$  then  $\mathbb{E}[X] \geq 0$ .

There are two issues that need immediate dissolution.

5.5. REMARK. (a) For a non-negative random variable, we must show that our definition of  $\mathbb{E}[X]$  is independent of the choice of  $(\phi_n)$ , i.e., if we take another sequence  $(\psi_n)$  of simple functions such that  $0 \leq \psi_n \uparrow X$  on  $\Omega$ , then we must have  $\sup_n \mathbb{E}[\psi_n] = \sup_n \mathbb{E}[\phi_n]$ . Indeed, take any simple function  $\psi$  such that  $0 \leq \psi \leq X$  on  $\Omega$ . Since  $\phi_n \wedge \psi \uparrow X \wedge \psi = \psi$  and each  $\phi_n \wedge \psi$  is simple (Exercise 5.2) and non-negative, we have by Lemma 5.2,

$$\mathbb{E}[\psi] = \sup_n \mathbb{E}[\phi_n \wedge \psi] \leq \sup_n \mathbb{E}[\phi_n].$$

Thus

$$\sup \{ \mathbb{E}[\psi] : 0 \leq \psi \leq X, \psi \text{ is simple} \} \leq \sup_n \mathbb{E}[\phi_n].$$

The reverse inequality also holds, since each  $\phi_n$  is a simple function satisfying  $0 \leq \phi_n \leq X$  and thus lying in the defining set of the sup in the left hand side. It follows that

$$\sup_n \mathbb{E}[\phi_n] = \sup \{ \mathbb{E}[\psi] : 0 \leq \psi \leq X, \psi \text{ is simple} \}.$$

Clearly, the same arguments, if applied to  $(\psi_n)$ , show that  $\sup_n \mathbb{E}[\psi_n]$  is equal to the right hand as well.

(b) For a simple function  $\phi$ , we now have two methods to define its expectation: using Definition 5.1 or Definition 5.4. We must show that they coincide. Let's temporarily denote the expectation in Definition 5.1 as  $\mathbb{E}_0$ . If  $\phi \geq 0$ , let  $\phi_n = \phi$  for each  $n \in \mathbb{N}$ , then

$$\mathbb{E}[\phi] = \sup_n \mathbb{E}_0[\phi_n] = \mathbb{E}_0[\phi].$$

For an arbitrary  $\phi$ , since  $\phi = \phi^+ - \phi^-$  and  $\phi^\pm$  are both simple and non-negative, it follows from Lemma 5.1(a),

$$\mathbb{E}_0[\phi] = \mathbb{E}_0[\phi^+] - \mathbb{E}_0[\phi^-] = \mathbb{E}[\phi^+] - \mathbb{E}[\phi^-] = \mathbb{E}[\phi].$$

5.1. EXAMPLE. Let  $\Omega = \mathbb{N}$  be endowed with a probability measure  $\mathbb{P}$ . For any non-negative random variable  $X$  on  $\mathbb{N}$ , we define a sequence of

simple functions as follows. For every  $n \in \mathbb{N}$ ,

$$\phi_n(k) = \begin{cases} X(k) & \text{if } k \leq n, \\ 0 & \text{if } k > n. \end{cases}$$

That is,  $\phi_n$  knocks  $X$  to 0 on the set  $\{k \in \mathbb{N} : k > n\}$ . We can rewrite  $\phi_n$  as  $\phi_n = \sum_{k=1}^n X(k) \mathbb{1}_{\{k\}}$ . Thus

$$\mathbb{E}[\phi_n] = \sum_{k=1}^n X(k) \mathbb{P}(\{k\}).$$

One sees that for each  $k \in \mathbb{N}$ ,  $0 \leq \phi_n(k) \uparrow X(k)$ . Therefore,

$$\mathbb{E}[X] = \sup_{n \in \mathbb{N}} \mathbb{E}[\phi_n] = \sup_n \sum_{k=1}^n X(k) \mathbb{P}(\{k\}) = \sum_{k=1}^{\infty} X(k) \mathbb{P}(\{k\}).$$

For a general random variable  $X$  on  $\mathbb{N}$ ,  $\mathbb{E}[X^\pm] = \sum_{k=1}^{\infty} X(k)^\pm \mathbb{P}(\{k\})$ . If one of the sums is finite, then

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[X^+] - \mathbb{E}[X^-] = \sum_{k=1}^{\infty} X(k)^+ \mathbb{P}(\{k\}) - \sum_{k=1}^{\infty} X(k)^- \mathbb{P}(\{k\}) \\ &= \sum_{k=1}^{\infty} X(k) \mathbb{P}(\{k\}) \end{aligned}$$

Moreover,  $X$  is integrable iff  $\sum_{k=1}^{\infty} X(k)^+ \mathbb{P}(\{k\}) + \sum_{k=1}^{\infty} X(k)^- \mathbb{P}(\{k\}) < \infty$ .

Note that

$$\begin{aligned} \sum_{k=1}^{\infty} X(k)^+ \mathbb{P}(\{k\}) + \sum_{k=1}^{\infty} X(k)^- \mathbb{P}(\{k\}) &= \sum_{k=1}^{\infty} (X(k)^+ + X(k)^-) \mathbb{P}(\{k\}) \\ &= \sum_{k=1}^{\infty} |X(k)| \mathbb{P}(\{k\}) = \mathbb{E}[|X|]. \end{aligned}$$

Thus  $X$  is integrable iff  $\mathbb{E}[|X|] < \infty$ —this fact is true in general (see Exercise 5.5).

We now extend Lemma 5.1 to the general case; the extension of Lemma 5.2 is of fundamental importance and is put to the next section.

**5.6. PROPOSITION.** *Let  $X, Y$  be two random variables such that  $\mathbb{E}[X]$  and  $\mathbb{E}[Y]$  are both defined. The following statements hold.*

- (a)  $\mathbb{E}[X] \leq \mathbb{E}[Y]$  whenever  $X \leq Y$  on  $\Omega$ .
- (b)  $\mathbb{E}[aX] = a\mathbb{E}[X]$  for any  $a \in \mathbb{R}$ .
- (c) If  $\mathbb{E}[X] + \mathbb{E}[Y]$  is defined, then  $\mathbb{E}[X + Y]$  is defined, and  $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ .

PROOF. (a). Let's assume first that  $0 \leq X \leq Y$ . Take a sequence  $(\phi_n)_{n \in \mathbb{N}}$  of simple functions such that  $0 \leq \phi_n \uparrow X$  on  $\Omega$  and a sequence  $(\psi_n)_{n \in \mathbb{N}}$  of simple functions such that  $0 \leq \psi_n \uparrow Y$  on  $\Omega$ . Then each  $\phi_n \wedge \psi_n$  is simple and  $0 \leq \phi_n \wedge \psi_n \uparrow X \wedge Y = X$ . Thus

$$\mathbb{E}[X] = \sup_n \mathbb{E}[\phi_n \wedge \psi_n] \leq \sup_n \mathbb{E}[\psi_n] = \mathbb{E}[Y].$$

For general  $X, Y$ , note that since  $X \leq Y$ ,  $X^+ \leq Y^+$  and  $Y^- \leq X^-$  (Exercise 0.1). Thus by what we just proved,  $\mathbb{E}[X^+] \leq \mathbb{E}[Y^+]$  and  $\mathbb{E}[Y^-] \leq \mathbb{E}[X^-]$ , implying that  $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-] \leq \mathbb{E}[Y^+] - \mathbb{E}[Y^-] = \mathbb{E}[Y]$ .

(b). Assume that  $a \geq 0$ . We start with the special case that  $X \geq 0$ . Take any sequence  $(\phi_n)_{n \in \mathbb{N}}$  of simple functions such that  $0 \leq \phi_n \uparrow X$  on  $\Omega$ . Then each  $a\phi_n$  is simple and  $0 \leq a\phi_n \uparrow aX$  on  $\Omega$ . Thus

$$\mathbb{E}[aX] = \lim_n \mathbb{E}[a\phi_n] = \lim_n a\mathbb{E}[\phi_n] = a \lim_n \mathbb{E}[\phi_n] = a\mathbb{E}[X],$$

where the first and last equalities are definitions of expectation and the second equality is due to Lemma 5.1(a).

Now let  $X$  be general. Since  $a \geq 0$ , direct verification shows that  $(aX)^\pm = aX^\pm$ . Thus  $\mathbb{E}[(aX)^\pm] = \mathbb{E}[aX^\pm] = a\mathbb{E}[X^\pm]$ , by the special case we just proved. Since  $\mathbb{E}[X]$  is defined,  $\mathbb{E}[X^\pm]$  cannot be both  $\infty$ . Hence,  $\mathbb{E}[(aX)^\pm]$  cannot be both  $\infty$ . It follows that  $\mathbb{E}[aX]$  is defined and

$$\begin{aligned} \mathbb{E}[aX] &= \mathbb{E}[(aX)^+] - \mathbb{E}[(aX)^-] = a\mathbb{E}[X^+] - a\mathbb{E}[X^-] \\ &= a(\mathbb{E}[X^+] - \mathbb{E}[X^-]) = a\mathbb{E}[X]. \end{aligned}$$

The case where  $a < 0$  is left to the reader as exercise.

(c). Let's first consider the special case that  $X, Y \geq 0$ . Take a sequence  $(\phi_n)_{n \in \mathbb{N}}$  of simple functions such that  $0 \leq \phi_n \uparrow X$  on  $\Omega$  and a sequence  $(\psi_n)_{n \in \mathbb{N}}$  of simple functions such that  $0 \leq \psi_n \uparrow Y$  on  $\Omega$ . Then each  $\phi_n + \psi_n$  is simple and  $0 \leq \phi_n + \psi_n \uparrow X + Y$ . By Lemma 5.1(a),

$$\begin{aligned} \mathbb{E}[X + Y] &= \lim_n \mathbb{E}[\phi_n + \psi_n] = \lim_n (\mathbb{E}[\phi_n] + \mathbb{E}[\psi_n]) = \lim_n \mathbb{E}[\phi_n] + \lim_n \mathbb{E}[\psi_n] \\ &= \mathbb{E}[X] + \mathbb{E}[Y]. \end{aligned}$$

By induction, we can extend this equality as follows: for any  $X_1, \dots, X_n \geq 0$ ,

$$(5.7) \quad \mathbb{E}\left[\sum_{k=1}^n X_k\right] = \sum_{k=1}^n \mathbb{E}[X_k].$$



Consider the general case now. Note that  $X^+ - X^- + Y^+ - Y^- = X + Y = (X + Y)^+ - (X + Y)^-$ . Thus

$$X^+ + Y^+ + (X + Y)^- = (X + Y)^+ + X^- + Y^-.$$

Applying expectations to the above, we have by (5.7),

$$(5.8) \quad \mathbb{E}[X^+] + \mathbb{E}[Y^+] + \mathbb{E}[(X + Y)^-] = \mathbb{E}[(X + Y)^+] + \mathbb{E}[X^-] + \mathbb{E}[Y^-].$$

Since some terms might be infinite, we need to put a bit more attention. We discuss the following cases:

- $\mathbb{E}[X] = \infty$  or  $\mathbb{E}[Y] = \infty$ .

Without loss of generality, assume that  $\mathbb{E}[X] = \infty$ . Then  $\mathbb{E}[X^+] = \infty$  and  $\mathbb{E}[X^-] < \infty$ . Since  $\mathbb{E}[X] + \mathbb{E}[Y]$  is defined, it follows that  $\mathbb{E}[Y] \neq -\infty$ , which in turn implies that  $\mathbb{E}[Y^-] < \infty$ , and

$$\mathbb{E}[X] + \mathbb{E}[Y] = \infty.$$

Recall that  $(X + Y)^- \leq X^- + Y^-$  (Exercise 0.1). Thus by (a) and (5.7), we have

$$\mathbb{E}[(X + Y)^-] \leq \mathbb{E}[X^- + Y^-] = \mathbb{E}[X^-] + \mathbb{E}[Y^-] < \infty.$$

This proves that  $\mathbb{E}[X + Y]$  is defined. Moreover, from (5.8), it follows that  $\mathbb{E}[(X + Y)^+] = \infty$ . Consequently,

$$\mathbb{E}[X + Y] = \infty = \mathbb{E}[X] + \mathbb{E}[Y].$$

- $\mathbb{E}[X] = -\infty$  or  $\mathbb{E}[Y] = -\infty$ .

Apply (b) and consider  $-X$ ,  $-Y$  and  $-X - Y$ .

- $\mathbb{E}[X] \in \mathbb{R}$  and  $\mathbb{E}[Y] \in \mathbb{R}$ .

As in the first case, one obtains  $\mathbb{E}[(X + Y)^-] < \infty$ . (5.8) that implies all the terms in it are finite. Moving around some terms, we get  $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ .

□

To include another interesting example, we need the following proposition. We say that two random variables  $X, Y$  are ***almost surely*** equal, written as  $X = Y$  a.s., if  $\mathbb{P}(X \neq Y) = 0^5$ ; that is,  $X$  and  $Y$  coincide except on a negligible set. Intuitively, we shall not distinguish almost surely equal random variables. The following result provides support for this.

**5.7. PROPOSITION.** *If  $X = 0$  a.s., then  $\mathbb{E}[X] = 0$ .*

---

<sup>5</sup>The set  $\{X \neq Y\} = \{X - Y \neq 0\}$  is always measurable since  $X - Y$  is measurable.

Its proof uses a routine for proving many results in real analysis: first consider simple non-negative functions, then consider general non-negative functions, and finally consider general functions.

PROOF. Let  $\phi \geq 0$  be a simple function such that  $\phi = 0$  a.s. We show that  $\mathbb{E}[\phi] = 0$ . If  $\phi = 0$  on  $\Omega$ , it is clear. Otherwise, write  $\phi = \sum_{k=1}^n c_k \mathbb{1}_{E_k}$  where  $c_k$ 's are distinct and  $c_k > 0$  for each  $k = 1, \dots, n$ . Then  $E_k = \{\phi = c_k\} \subset \{\phi \neq 0\}$ , implying that  $\mathbb{P}(E_k) = 0$ , for each  $k = 1, \dots, n$ . Therefore,  $\mathbb{E}[\phi] = \sum_{k=1}^n c_k \mathbb{P}(E_k) = 0$ , as desired.

Let  $Y \geq 0$  be any function such that  $Y = 0$  a.s. Take any simple function  $\phi$  such that  $0 \leq \phi \leq Y$ . Then  $\{\phi \neq 0\} \subset \{Y \neq 0\}$  and thus  $\mathbb{P}(\phi \neq 0) = 0$ , i.e.,  $\phi = 0$  a.s. By the previous case,  $\mathbb{E}[\phi] = 0$ . Since  $\phi$  is arbitrary, we have by the definition of expectation that  $\mathbb{E}[Y] = 0$ .

For the general case, note that  $X^\pm = 0$  a.s. (Exercise 5.3). Thus by the previous case,  $\mathbb{E}[X^\pm] = 0$  and  $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-] = 0$ .  $\square$

We can similarly define that  $X \leq Y$  a.s. if  $\mathbb{P}(X > Y) = 0$ .

5.8. COROLLARY. *If  $X \leq Y$  a.s. and both  $\mathbb{E}[X]$  and  $\mathbb{E}[Y]$  are defined, then  $\mathbb{E}[X] \leq \mathbb{E}[Y]$ . If  $X = Y$  a.s. and  $\mathbb{E}[X]$  is defined, then  $\mathbb{E}[Y]$  is defined and  $\mathbb{E}[Y] = \mathbb{E}[X]$ .*

We leave the proof to the reader as an exercise.

5.2. EXAMPLE. Let  $\omega_1$  and  $\omega_2$  be two distinct points in  $\Omega$  and  $t \in [0, 1]$ . Let  $\mathbb{P} = t\delta_{\omega_1} + (1-t)\delta_{\omega_2}$ . Let's compute  $\mathbb{E}[X]$  for any random variable  $X$ . The set  $\Omega \setminus \{\omega_1, \omega_2\}$  has probability 0. This motivates us to consider  $X$  on three pieces of  $\Omega$ :  $\{\omega_1\}$ ,  $\{\omega_2\}$ , and  $\Omega \setminus \{\omega_1, \omega_2\}$ . We can write  $X$  as follows:

$$X = X(\omega_1)\mathbb{1}_{\{\omega_1\}} + X(\omega_2)\mathbb{1}_{\{\omega_2\}} + X\mathbb{1}_{\Omega \setminus \{\omega_1, \omega_2\}}.$$

Note that  $X\mathbb{1}_{\Omega \setminus \{\omega_1, \omega_2\}} \neq 0$  only possibly on  $\Omega \setminus \{\omega_1, \omega_2\}$ . Thus  $\{X\mathbb{1}_{\Omega \setminus \{\omega_1, \omega_2\}} \neq 0\} \subset \Omega \setminus \{\omega_1, \omega_2\}$  and  $\mathbb{P}(X\mathbb{1}_{\Omega \setminus \{\omega_1, \omega_2\}} \neq 0) = 0$ . It follows that  $X\mathbb{1}_{\Omega \setminus \{\omega_1, \omega_2\}} = 0$  a.s. and  $\mathbb{E}[X\mathbb{1}_{\Omega \setminus \{\omega_1, \omega_2\}}] = 0$ . Consequently, by Proposition 5.6(c),

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[X(\omega_1)\mathbb{1}_{\{\omega_1\}}] + \mathbb{E}[X(\omega_2)\mathbb{1}_{\{\omega_2\}}] + \mathbb{E}[X\mathbb{1}_{\Omega \setminus \{\omega_1, \omega_2\}}] \\ &= X(\omega_1)\mathbb{P}(\{\omega_1\}) + X(\omega_2)\mathbb{P}(\{\omega_2\}) \\ &= tX(\omega_1) + (1-t)X(\omega_2). \end{aligned}$$

See Exercise 5.8 for an extension of this example.

### 3. Convergence theorems

We extend Lemma 5.2 to general random variables. We first extend the notion of almost surety. We say a property is satisfied ***almost surely*** on  $\Omega$  if the set of points where it is not satisfied has probability 0.

5.9. THEOREM (Monotone Convergence Theorem). *Let  $X, X_n, n \in \mathbb{N}$ , be random variables such that  $X_n \geq 0$  a.s. for every  $n \in \mathbb{N}$ ,  $X_n \leq X_{n+1}$  a.s. for every  $n \in \mathbb{N}$ , and  $X = \lim_n X_n$  a.s.<sup>6</sup> Then  $\mathbb{E}[X_n] \uparrow \mathbb{E}[X]$ .*

Using  $X = \lim_n X_n$ , we can rewrite the conclusion in the theorem as  $\mathbb{E}[\lim_n X_n] = \lim_n \mathbb{E}[X_n]$ . That is, we can change the order of taking expectation and limit, under the assumptions of the theorem.

PROOF. We first prove the case where  $0 \leq X_n \uparrow X$  on  $\Omega$ . For each  $n \in \mathbb{N}$ , take a sequence  $(\phi_{n,k})_{k \in \mathbb{N}}$  of simple functions such that

$$0 \leq \phi_{n,k} \uparrow_k X_n.$$

Now for every  $n \in \mathbb{N}$ , put

$$\psi_n = \max\{\phi_{m,n} : m = 1, \dots, n\}.$$

See Figure 2 for illustration. Clearly, each  $\psi_n$  is simple and non-negative.

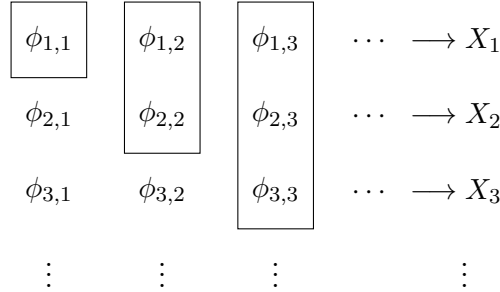


FIGURE 2. Double array of simple functions

Since each row is increasing,

$$\begin{aligned} \psi_{n+1} &= \max\{\phi_{m,n+1} : m = 1, \dots, n+1\} \geq \max\{\phi_{m,n+1} : m = 1, \dots, n\} \\ &\geq \max\{\phi_{m,n} : m = 1, \dots, n\} = \psi_n. \end{aligned}$$

Since every  $\phi_{n,k}$  is bounded by  $X_n$  and thus by  $X$ ,  $\psi_n \leq X$  for every  $n \in \mathbb{N}$ , implying that  $\lim_n \psi_n \leq X$ . For any  $k \in \mathbb{N}$ ,  $\psi_n \geq \phi_{k,n}$  whenever  $k \leq n$ . Letting  $n \rightarrow \infty$ , it follows that  $\lim_n \psi_n \geq \lim_n \phi_{k,n} = X_k$ . Letting  $k \rightarrow \infty$ ,

<sup>6</sup>The set of points of convergence is always measurable; Exercise 4.13.

it follows that  $\lim_n \psi_n \geq X$ . Therefore,  $\lim_n \psi_n = X$ . By the definition of expectation,

$$\mathbb{E}[X] = \lim_n \mathbb{E}[\psi_n].$$

Finally, note that since each row in Figure 2 is increasing, for any  $n \in \mathbb{N}$ ,

$$\psi_n = \max\{\phi_{m,n} : m = 1, \dots, n\} \leq \max\{X_m : m = 1, \dots, n\} = X_n.$$

Thus by Proposition 5.6(a),  $\mathbb{E}[\psi_n] \leq \mathbb{E}[X_n] \leq \mathbb{E}[X]$ . Using the Squeeze Law, we have  $\lim_n \mathbb{E}[X_n] = \mathbb{E}[X]$ . The increasingness of  $\mathbb{E}[X_n]$ 's is also due to Proposition 5.6(a). The special case is thus proved.

Now we prove the general case. Set

$$A = \bigcup_{n=1}^{\infty} \{X_n < 0\} \cup \bigcup_{n=1}^{\infty} \{X_n > X_{n+1}\} \cup \{(X_n)_n \text{ does not converge to } X\}.$$

Each of the sets in the right hand side has probability zero. Thus  $A$  has probability zero as well, by Corollary 2.7. We now set  $Y = X\mathbb{1}_{A^c}$  and  $Y_n = X_n\mathbb{1}_{A^c}$  for any  $n \in \mathbb{N}$ . One sees that  $Y = X$  a.s. and  $Y_n = X_n$  a.s. for every  $n \in \mathbb{N}$ . Thus by Corollary 5.8,

$$(5.9) \quad \mathbb{E}[Y] = \mathbb{E}[X], \quad \mathbb{E}[Y_n] = \mathbb{E}[X_n], \quad \text{for every } n \in \mathbb{N}.$$

Moreover, if  $\omega \in A$ , then  $Y(\omega) = 0 = Y_n(\omega)$  for each  $n \in \mathbb{N}$ ; if  $\omega \in A^c$ , then  $Y_n(\omega) = X_n(\omega)$  for every  $n \in \mathbb{N}$ ,  $0 \leq X_n(\omega) \leq X_{n+1}(\omega)$  for every  $n \in \mathbb{N}$ , and  $\lim_n X_n(\omega) = X(\omega)$ . In either case, one sees that  $0 \leq Y_n \uparrow Y$  on  $\Omega$ . Thus the proof is complete by (5.9) and the special case we just proved.  $\square$

**5.10. COROLLARY (Fatou's Lemma).** *If  $X_n \geq 0$  a.s. for every  $n \in \mathbb{N}$  and  $\liminf_n X_n \in \mathbb{R}$  on  $\Omega$ , then  $\mathbb{E}[\liminf_n X_n] \leq \liminf_n \mathbb{E}[X_n]$ .*

**PROOF.** For any  $n \in \mathbb{N}$ , set  $Y_n = \inf_{k \geq n} X_k$ . One sees that  $Y_n \geq 0$  a.s. for every  $n \in \mathbb{N}$  (why? cf. Exercise 5.17) and  $Y_n \uparrow \liminf_n X_n$  on  $\Omega$ . Thus

$$\mathbb{E}[\liminf_n X_n] = \sup_n \mathbb{E}[Y_n].$$

Now for every  $n \in \mathbb{N}$ , since  $Y_n \leq X_k$  for any  $k \geq n$ , we have by Proposition 5.6(a),

$$\mathbb{E}[Y_n] \leq \mathbb{E}[X_k] \quad \text{for any } k \geq n.$$

Taking infimum over  $k$ , we have  $\mathbb{E}[Y_n] \leq \inf_{k \geq n} \mathbb{E}[X_k]$ . Therefore,

$$\mathbb{E}[\liminf_n X_n] = \sup_n \mathbb{E}[Y_n] \leq \sup_n \inf_{k \geq n} \mathbb{E}[X_k] = \liminf_n \mathbb{E}[X_n].$$

$\square$

One may compare Theorem 5.9 and Corollary 5.10 with Proposition 2.8 and Corollary 2.4; see Exercise 5.15. See also Exercise 5.21 and 5.23 for the cases when the limit/liminf takes infinite values.

5.11. COROLLARY (Dominated Convergence Theorem). *Let  $X^* \geq 0$  be integrable. Let  $X, X_n, n \in \mathbb{N}$  be such that  $|X_n| \leq X^*$  a.s. for every  $n \in \mathbb{N}$  and  $X_n \rightarrow X$  a.s. Show that  $\mathbb{E}[X_n] \rightarrow \mathbb{E}[X]$ .*

We leave the proof of this corollary to the reader; Exercise 5.13.

### Exercises

5.1. Show that if  $E_1, \dots, E_n$  are disjoint then  $\mathbb{1}_{\bigcup_{k=1}^n E_k} = \sum_{k=1}^n \mathbb{1}_{E_k}$ .

5.2. Show that for two simple functions  $\phi$  and  $\psi$ ,  $\phi \wedge \psi$  is also simple.

5.3. Show that if  $X = 0$  a.s., then  $X^\pm = 0$  a.s.

5.4. Complete the proof of Proposition 5.6.

5.5. Show that  $X$  is integrable iff  $\mathbb{E}[|X|] < \infty$ .

5.6. Show that if  $X$  is integrable then  $X\mathbb{1}_A$  is integrable for any  $A \in \mathcal{F}$ .

5.7. Prove Corollary 5.8.

5.8. Let  $(\omega_n)_{n \in \mathbb{N}}$  be a sequence of distinct points in  $\Omega$  and let  $(t_n)_{n \in \mathbb{N}}$  be a sequence of non-negative real numbers such that  $\sum_{n=1}^{\infty} t_n = 1$ . Let  $\mathbb{P} = \sum_{n=1}^{\infty} t_n \delta_{\omega_n}$ . Show that for any non-negative function  $X$ ,  $\mathbb{E}[X] = \sum_{n=1}^{\infty} t_n X(\omega_n)$ .

5.9. Suppose that  $X_0$  is integrable and  $X_0 \leq X_n \uparrow X$  a.s. Then  $\lim_n \mathbb{E}[X_n] = \mathbb{E}[X]$ .

5.10. Suppose that  $X_0$  is integrable and  $X_0 \geq X_n \downarrow X$  a.s. Then  $\lim_n \mathbb{E}[X_n] = \mathbb{E}[X]$ .

5.11. Suppose that  $X_0$  is integrable,  $X_0 \leq X_n$  a.s. for every  $n \in \mathbb{N}$ , and  $\liminf_n X_n \in \mathbb{R}$  on  $\Omega$ . Then  $\mathbb{E}[\liminf_n X_n] \leq \liminf_n \mathbb{E}[X_n]$ .

5.12. Suppose  $X_0$  is integrable,  $X_0 \geq X_n$  a.s. for every  $n \in \mathbb{N}$ , and  $\limsup_n X_n \in \mathbb{R}$  on  $\Omega$ . Then  $\mathbb{E}[\limsup_n X_n] \geq \limsup_n \mathbb{E}[X_n]$ .

5.13. Use Exercises 5.11 and 5.12 to prove Corollary 5.11.

5.14. Let  $(A_n)$  be a sequence of subsets of a set  $\Omega$ . Show that

$$\limsup_n \mathbb{1}_{A_n} = \mathbb{1}_{\limsup_n A_n},$$

$$\liminf_n \mathbb{1}_{A_n} = \mathbb{1}_{\liminf_n A_n}.$$

Moreover,  $\mathbb{1}_{A_n} \uparrow \mathbb{1}_A$  iff  $A_n \uparrow A$ .

5.15. Use Exercise 5.14 to deduce Proposition 2.8 and Corollary 2.4 from Theorem 5.9 and Corollary 5.10, respectively.

5.16. Find a sequence  $(X_n)$  over some probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  such that  $\mathbb{E}[\liminf_n f_n] > \liminf_n \mathbb{E}[f_n]$ .

5.17. Let  $X_n$  and  $Y_n$ ,  $n \in \mathbb{N}$ , be such that  $X_n \geq Y_n$  a.s. for every  $n \in \mathbb{N}$ . Show that  $X_n \geq Y_n$  for all  $n \in \mathbb{N}$  a.s.

5.18. Show that  $X = Y$  a.s. iff  $X \geq Y$  a.s. and  $X \leq Y$  a.s.

5.19. Let  $(A_n)_{n \in \mathbb{N}}$  be a disjoint sequence of measurable sets such that  $\mu(A_k \cap A_j) = 0$  whenever  $k \neq j$ . Show that  $(\mathbb{1}_{A_n})$  converges to 0 a.s.

5.20. Let  $X, Y$  be integrable and  $a, b \in \mathbb{R}$ . Show that  $aX + bY$  is integrable.

5.21. Let  $(X_n)_{n \in \mathbb{N}}$  be a sequence of random variables such that  $X_n \geq 0$  a.s. for every  $n \in \mathbb{N}$  and  $X_n \leq X_{n+1}$  a.s. for every  $n \in \mathbb{N}$ . If  $X_n(\omega) \uparrow \infty$  for every  $\omega$  in a set of positive measure, then  $\mathbb{E}[X_n] \uparrow \infty$ .

5.22. If  $X_n \geq 0$  a.s. for every  $n \in \mathbb{N}$ , then  $\liminf_n X_n > -\infty$  a.s.

5.23. If  $X_n \geq 0$  a.s. for every  $n \in \mathbb{N}$  and  $\liminf_n X_n = \infty$  on a set of positive measure, then  $\liminf_n \mathbb{E}[X_n] = \infty$ .

## CHAPTER 6

### Expectations II

We continue to establish some further properties of expectations. Fix an arbitrary probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  in this chapter.

#### 1. Some fundamental inequalities

The first inequality controls  $\mathbb{P}(|X| \geq \varepsilon)$  in terms of expectation.

6.1. PROPOSITION (Chebyshev's inequality). *Let  $X$  be an integrable random variable on  $\Omega$ . For any  $\varepsilon > 0$ ,*

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}[|X|]}{\varepsilon}.$$

PROOF. Apply Proposition 5.6(a) to  $\varepsilon \mathbb{1}_{\{|X| \geq \varepsilon\}} \leq |X|$ .  $\square$

6.2. COROLLARY. *If  $\mathbb{E}[|X|] = 0$  then  $X = 0$  a.s.*

PROOF. By Chebyshev's inequality,  $\mathbb{P}(|X| \geq \frac{1}{k}) = 0$  for every  $k \in \mathbb{N}$ . Thus  $\mathbb{P}(X \neq 0) = \mathbb{P}(\bigcup_{k=1}^{\infty} \{|X| \geq \frac{1}{k}\}) = 0$ , by Corollary 2.7.  $\square$

The following result will be needed later.

6.3. COROLLARY. *Let  $X, Y$  be integrable. Then  $X \geq Y$  a.s. iff  $\mathbb{E}[X \mathbb{1}_A] \geq \mathbb{E}[Y \mathbb{1}_A]$  for every  $A \in \mathcal{F}$ .*

Note that  $X \mathbb{1}_A$  and  $Y \mathbb{1}_A$  are both integrable (Exercise 5.6).

PROOF. The “only if” part is immediate by Corollary 5.8. For the “if” part, take  $A = \{X < Y\}$ . Then  $0 \geq \mathbb{E}[Y \mathbb{1}_A] - \mathbb{E}[X \mathbb{1}_A] = \mathbb{E}[(Y - X) \mathbb{1}_A] \geq 0$ , where the last inequality is due to  $(Y - X) \mathbb{1}_A \geq 0$  (verify it). Therefore,  $\mathbb{E}[(Y - X) \mathbb{1}_A] = 0$ . By Corollary 6.2,  $(Y - X) \mathbb{1}_A = 0$  a.s. Since  $(Y - X) \mathbb{1}_A > 0$  at every point in  $A$ , it follows that  $\mathbb{P}(A) = 0$ , i.e.,  $X \geq Y$  a.s.  $\square$

The technique of truncating a random variable  $X$  to sets where it takes certain special values is very important. For example, it has been used in the proofs of Theorem 4.12 and Chebyshev's inequality. We demonstrate another application of it. See Exercise 6.3 for another good application.

6.1. EXAMPLE. If  $\mathbb{E}[X^2] < \infty$  then  $\mathbb{E}[|X|] < \infty$ . Indeed, note that

$$\mathbb{E}[|X|] = \mathbb{E}[|X|\mathbb{1}_{\{|X|<1\}}] + \mathbb{E}[|X|\mathbb{1}_{\{|X|\geq 1\}}].$$

Clearly,  $\mathbb{E}[|X|\mathbb{1}_{\{|X|<1\}}] \leq \mathbb{E}[\mathbb{1}_{\{|X|<1\}}] \leq 1$ . Thus the integrability of  $X$  purely depends on  $|X|\mathbb{1}_{\{|X|\geq 1\}}$ , the piece of  $X$  where it takes large values. Note that if  $\omega \in \{|X| \geq 1\}$  then  $|X(\omega)| \leq X(\omega)^2$ . Thus  $|X|\mathbb{1}_{\{|X|\geq 1\}} \leq X^2\mathbb{1}_{\{|X|\geq 1\}} \leq X^2$ . It follows that  $\mathbb{E}[|X|\mathbb{1}_{\{|X|\geq 1\}}] \leq \mathbb{E}[X^2] < \infty$ .

We now turn to the famous Hölder's Inequality and Minskowski's Inequality. For a random variable  $X$ , its  $p$ -norm is given by

$$\|X\|_p := \begin{cases} \left(\mathbb{E}[|X|^p]\right)^{\frac{1}{p}} & \text{if } 1 \leq p < \infty, \\ \inf \{M > 0 : |X| \leq M \text{ a.s.}\} & \text{if } p = \infty. \end{cases}$$

It is easy to see that

$$(6.1) \quad \||X|\|_p = \|X\|_p \text{ and } \|aX\|_p = |a|\|X\|_p$$

for any  $a \in \mathbb{R}$ . In particular, if  $0 < \|X\|_p < \infty$ , then

$$(6.2) \quad \left\| \frac{X}{\|X\|_p} \right\|_p = \frac{1}{\|X\|_p} \|X\|_p = 1.$$

If  $1 < p < \infty$ , we call  $p'$  such that  $\frac{1}{p} + \frac{1}{p'} = 1$  the conjugate index of  $p$ . Clearly,  $p' = \frac{p}{p-1} \in (1, \infty)$  and  $(p')' = p$ . The conjugate index of  $p = 1$  is  $1' = \infty$ ; the conjugate index of  $p = \infty$  is  $\infty' = 1$ .

6.4. PROPOSITION (Hölder's inequality). *For any two random variables  $X, Y$  and  $1 \leq p \leq \infty$ ,*

$$\|XY\|_1 \leq \|X\|_p \|Y\|_{p'}.$$

PROOF. We prove the inequality for  $1 < p < \infty$ ; it is left to the reader when  $p = 1$  or  $\infty$ . We begin with kicking away the trivial cases. If  $\|X\|_p = 0$ , then  $X = 0$  a.s. by Corollary 6.2. It follows that  $XY = 0$  a.s. and thus  $\|XY\|_1 = 0$  by Proposition 5.7. Therefore, the inequality holds. Similarly, if  $\|Y\|_{p'} = 0$ , then the inequality holds. Assume now that

$$\|X\|_p > 0, \quad \|Y\|_{p'} > 0.$$

If  $\|X\|_p = \infty$  or  $\|Y\|_{p'} = \infty$ , the inequality is clear since the right hand side is  $\infty$ . Thus we assume that

$$\|X\|_p < \infty, \quad \|Y\|_{p'} < \infty.$$



Dividing both sides by  $\|X\|_p \|Y\|_{p'}$ , by (6.1), the desired inequality becomes

$$\left\| \frac{|X|}{\|X\|_p} \frac{|Y|}{\|Y\|_{p'}} \right\|_1 \leq 1.$$

Thus in view of (6.2), it is enough to prove that if  $\|X\|_p = 1 = \|Y\|_{p'}$  and  $X, Y \geq 0$ , then

$$\mathbb{E}[XY] \leq 1.$$

Let's prove the last inequality. Put  $f(t) = \ln t$  on  $(0, \infty)$ . Then  $f''(t) = -t^{-2} < 0$  on  $(0, \infty)$  and  $f$  is concave, i.e.,  $t \ln x + (1-t) \ln y \leq \ln(tx + (1-t)y)$  for any  $t \in [0, 1]$  and  $x, y > 0$ . It follows that

$$x^t y^{1-t} \leq tx + (1-t)y$$

for any  $t \in [0, 1]$  and  $x, y \geq 0$ . Taking  $t = \frac{1}{p}$  (so that  $1-t = \frac{1}{p'}$ ),  $x = a^p$  and  $y = b^{p'}$ , we get the famous Young Inequality:

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

for any  $a, b \geq 0$ . In particular, at every  $\omega \in \Omega$ ,

$$X(\omega)Y(\omega) \leq \frac{X(\omega)^p}{p} + \frac{Y(\omega)^{p'}}{p'}.$$

Taking expectations we have

$$\mathbb{E}[XY] \leq \frac{\mathbb{E}[X^p]}{p} + \frac{\mathbb{E}[Y^{p'}]}{p'} = \frac{\|X\|_p^p}{p} + \frac{\|Y\|_{p'}^{p'}}{p'} = \frac{1}{p} + \frac{1}{p'} = 1.$$

This completes the proof.  $\square$

When  $p = 2$ , Hölder's Inequality is usually called Cauchy-Schwarz Inequality.

**6.5. COROLLARY** (Minkowski's Inequality). *For any two random variables  $X, Y$  and  $1 \leq p \leq \infty$ ,*

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p.$$

**PROOF.** We prove the inequality for  $1 < p < \infty$ ; it is left to the reader when  $p = 1$  or  $\infty$ . If  $\|X\|_p + \|Y\|_p = \infty$ , there is nothing to prove. Let's assume that  $\|X\|_p + \|Y\|_p < \infty$ . Note that

$$|X + Y|^p = |X + Y| \cdot |X + Y|^{p-1} \leq |X| \cdot |X + Y|^{p-1} + |Y| \cdot |X + Y|^{p-1}.$$

Applying Hölder's Inequality to the last two terms, we obtain

$$\mathbb{E}[|X + Y|^p] \leq \|X\|_p \mathbb{E}[|X + Y|^{p-1}] + \|Y\|_p \mathbb{E}[|X + Y|^{p-1}].$$

In view of  $p' = \frac{p}{p-1}$ , we have

$$\| |X + Y|^{p-1} \|_{p'} = \left( \mathbb{E} \left[ \left( |X + Y|^{p-1} \right)^{\frac{p}{p-1}} \right] \right)^{\frac{p-1}{p}} = \|X + Y\|_p^{p-1}.$$

It follows that

$$\|X + Y\|_p^p = \mathbb{E}[|X + Y|^p] \leq (\|X\|_p + \|Y\|_p) \|X + Y\|_p^{p-1}.$$

If  $\|X + Y\|_p < \infty$ , then dividing both sides by  $\|X + Y\|_p^{p-1}$  yields the desired inequality.

Let's now show that  $\|X + Y\|_p < \infty$ . Let  $a, b \geq 0$ . Then

$$(a + b)^p \leq (2 \max\{a, b\})^p = 2^p \max\{a^p, b^p\} \leq 2^p(a^p + b^p).$$

Thus

$$\begin{aligned} \mathbb{E}[|X + Y|^p] &\leq \mathbb{E}[(|X| + |Y|)^p] \leq 2^p \mathbb{E}[|X|^p + |Y|^p] \\ &= 2^p (\mathbb{E}[|X|^p] + \mathbb{E}[|Y|^p]) = 2^p (\|X\|_p^p + \|Y\|_p^p) < \infty. \end{aligned}$$

This completes the proof.  $\square$

**6.2. EXAMPLE.** Let  $1 < p < \infty$ . Consider  $\Omega = \{1, 2, \dots, n\}$  endowed with the probability  $\mathbb{P}(\{k\}) = \frac{1}{n}$  for any  $k = 1, \dots, n$ . Then  $\mathbb{E}[X] = \frac{1}{n} \sum_{k=1}^n X(k)$  for any function  $X$ . Thus Hölder's Inequality reduces to

$$\frac{1}{n} \sum_{k=1}^n |X(k)Y(k)| \leq \left( \frac{1}{n} \sum_{k=1}^n |X(k)|^p \right)^{\frac{1}{p}} \left( \frac{1}{n} \sum_{k=1}^n |Y(k)|^{p'} \right)^{\frac{1}{p'}},$$

or simply,

$$(6.3) \quad \sum_{k=1}^n |X(k)Y(k)| \leq \left( \sum_{k=1}^n |X(k)|^p \right)^{\frac{1}{p}} \left( \sum_{k=1}^n |Y(k)|^{p'} \right)^{\frac{1}{p'}}.$$

Similarly, Minkowski's Inequality reduces to

$$(6.4) \quad \left( \sum_{k=1}^n |X(k) + Y(k)|^p \right)^{\frac{1}{p}} \leq \left( \sum_{k=1}^n |X(k)|^p \right)^{\frac{1}{p}} + \left( \sum_{k=1}^n |Y(k)|^p \right)^{\frac{1}{p}}.$$

It is interesting to observe that the general inequalities can be deduced from these much simpler reduced forms. Let's illustrate to deduce Proposition 6.4 from (6.3). Take any non-negative simple functions  $\phi, \psi$  on  $\Omega$ . We

may write  $\phi = \sum_{k=1}^n a_k \mathbb{1}_{E_k}$  and  $\psi = \sum_{k=1}^n b_k \mathbb{1}_{E_k}$ . Then by (6.3),

$$\begin{aligned} \mathbb{E}[\phi\psi] &= \sum_{k=1}^n a_k b_k \mathbb{P}(E_k) = \sum_{k=1}^n a_k \mathbb{P}(E_k)^{\frac{1}{p}} \cdot b_k \mathbb{P}(E_k)^{\frac{1}{p'}} \\ &\leq \left( \sum_{k=1}^n a_k^p \mathbb{P}(E_k) \right)^{\frac{1}{p}} \left( \sum_{k=1}^n b_k^{p'} \mathbb{P}(E_k) \right)^{\frac{1}{p'}} = \|\phi\|_p \|\psi\|_{p'}. \end{aligned}$$

Now for general non-negative random variables  $X, Y$ , take two sequences of simple functions such that  $0 \leq \phi_n \uparrow X$  and  $0 \leq \psi_n \uparrow Y$ . Then  $0 \leq \phi_n \psi_n \uparrow XY$ ,  $\phi_n^p \uparrow X^p$ , and  $\psi_n^{p'} \uparrow Y^{p'}$ . Thus  $\mathbb{E}[\phi_n \psi_n] \uparrow \mathbb{E}[XY]$ ,  $\mathbb{E}[\phi_n^p] \uparrow \mathbb{E}[X^p]$ , and  $\mathbb{E}[\psi_n^{p'}] \uparrow \mathbb{E}[Y^{p'}]$ . Writing these terms in norms and letting  $n \rightarrow \infty$  in  $\mathbb{E}[\phi_n \psi_n] \leq \|\phi_n\|_p \|\psi_n\|_{p'}$ , we get  $\mathbb{E}[XY] \leq \|X\|_p \|Y\|_{p'}$ .

Put

$$L^p(\Omega, \mathcal{F}, \mathbb{P}) := \{X : \|X\|_p < \infty\}.$$

We may abbreviate it as  $L^p$ . It can be shown that  $L^p$  is a vector space for any  $p \in [1, \infty]$ ; in fact, a Banach space (Exercise 6.18). If we interpret norm as “length” of a vector, Minkowski’s inequality is then the triangle inequality.

## 2. Indefinite integrals

Let  $X$  be a non-negative integrable random variable on  $\Omega$ . We define

$$(6.5) \quad \mu(E) = \mathbb{E}[X \mathbb{1}_E] \quad \text{for every } E \in \mathcal{F}.$$

**6.6. PROPOSITION.** *For a non-negative integrable random variable  $X$ ,  $\mu$  in (6.5) is a finite measure on  $(\Omega, \mathcal{F})$  such that  $\mu(E) = 0$  whenever  $\mathbb{P}(E) = 0$ .*

**PROOF.** Clearly,  $\mu(E) \geq 0$  for any  $E \in \mathcal{F}$ . If  $\mathbb{P}(E) = 0$ , then  $\mathbb{1}_E = 0$  a.s., and thus  $X \mathbb{1}_E = 0$  a.s. It follows that  $\mu(E) = \mathbb{E}[X \mathbb{1}_E] = 0$ . In particular,  $\mu(\emptyset) = 0$ . Let  $(E_n)_{n \in \mathbb{N}}$  be a disjoint sequence in  $\mathcal{F}$ . Set  $Y_n = X \mathbb{1}_{\bigcup_{k=1}^n E_k}$  for any  $n \in \mathbb{N}$  and  $Y = X \mathbb{1}_{\bigcup_{k=1}^{\infty} E_k}$ . Then  $0 \leq Y_n \uparrow Y$ . Thus by Monotone Convergence Theorem,

$$\begin{aligned} \mu\left(\bigcup_{k=1}^{\infty} E_k\right) &= \mathbb{E}[Y] = \lim_n \mathbb{E}[Y_n] = \lim_n \mathbb{E}[X \mathbb{1}_{\bigcup_{k=1}^n E_k}] \\ &= \lim_n \mathbb{E}\left[X \sum_{k=1}^n \mathbb{1}_{E_k}\right] = \lim_n \sum_{k=1}^n \mathbb{E}[X \mathbb{1}_{E_k}] = \lim_n \sum_{k=1}^n \mu(E_k) \\ &= \sum_{k=1}^{\infty} \mu(E_k). \end{aligned}$$

This proves that  $\mu$  is a measure. It is finite since  $\mu(\Omega) = \mathbb{E}[X] < \infty$ .  $\square$

Sometimes  $\mu$  is called the *indefinite integral* of  $X$ .

Surprisingly, the converse of this proposition is also true.

**6.7. THEOREM (Radon-Nikodym).** *Let  $\mu$  be a finite measure on  $(\Omega, \mathcal{F})$  such that  $\mu(E) = 0$  whenever  $E \in \mathcal{F}$  and  $\mathbb{P}(E) = 0$ . Then there exists a non-negative integrable random variable  $X$  on  $\Omega$  satisfying (6.5).*

The proof of this theorem is very technical and beyond the scope of this book; we skip it. The random variable  $X$  is called the **Radon-Nikodym derivative** of  $\mu$  with respect to  $\mathbb{P}$  and is denoted by

$$\frac{d\mu}{d\mathbb{P}}.$$

It is unique up to a.s. equality. That is, if  $Y$  is another non-negative integrable random variable satisfying (6.5), then  $X = Y$  a.s. (Exercise 6.20).

### 3. Lebesgue and Riemann integrals

Let  $(\Omega, \mathcal{F}, \mu)$  be a general measure space. We can similarly define expectations of *non-negative* simple functions and then extend the definition to general functions as in Definition 5.4. However, in this case, we rename expectation as *integral* and rewrite it as

$$\int_{\Omega} f(\omega) d\mu(\omega),$$

or even

$$\int_{\Omega} f d\mu, \quad \int_{\Omega} f,$$

as long as there is no possible ambiguity. One can effortlessly verify that all the results in Sections 2 and 3 of Chapter 5 and in Section 1 of this chapter, except Example 6.1, still hold. Results in Section 2 of this chapter hold for  $\sigma$ -finite measures (see Exercises 6.21 and 6.22). In non-probability measure spaces, we rename almost sure to **almost everywhere**.

**6.3. EXAMPLE.** Let  $\mu$  be the counting measure over  $\mathbb{N}$ . Then for any non-negative function  $f$  on  $\mathbb{N}$ ,  $\int_{\mathbb{N}} f d\mu = \sum_{k=1}^{\infty} f(k)$ . Cf. Example 5.1.

**6.4. EXAMPLE.** We illustrate an application of Dominated Convergence Theorem. Let  $x, x_n, n \in \mathbb{N}$ , be real numbers such that  $x_n \rightarrow x$ . We want to show that

$$\lim_n \left(1 + \frac{x_n}{n}\right)^n = e^x.$$

Endow  $\Omega = \mathbb{N} \cup \{0\}$  with the counting measure. Let  $M := \sup_n |x_n| \in \mathbb{R}$ . Define the following functions on  $\mathbb{N} \cup \{0\}$ :

$$\begin{aligned} f^* : \mathbb{N} \cup \{0\} &\rightarrow \mathbb{R}; \quad k \mapsto \frac{1}{k!} M^k; \\ f : \mathbb{N} \cup \{0\} &\rightarrow \mathbb{R}; \quad k \mapsto \frac{1}{k!} x^k; \\ f_n : \mathbb{N} \cup \{0\} &\rightarrow \mathbb{R}; \quad k \mapsto \mathbb{1}_{\{k \leq n\}} \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-k+1}{n} \frac{1}{k!} x_n^k. \end{aligned}$$

One sees that  $f_n$  is simple and  $|f_n| \leq f^*$  for every  $n \in \mathbb{N}$  and that  $f_n \rightarrow f$  on  $\mathbb{N} \cup \{0\}$ . Since  $\int_{\Omega} f^* d\mu = e^M < \infty$ ,  $f^*$  is integrable. Thus by Dominated Convergence Theorem,  $\int_{\Omega} f_n d\mu \rightarrow \int_{\Omega} f d\mu = e^x$ . Finally, note that

$$\begin{aligned} \left(1 + \frac{x_n}{n}\right)^n &= \sum_{k=0}^n \binom{n}{k} \frac{x_n^k}{n^k} = \sum_{k=0}^{\infty} \mathbb{1}_{\{k \leq n\}} \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-k+1}{n} \frac{1}{k!} x_n^k \\ &= \int_{\Omega} f_n d\mu. \end{aligned}$$

We introduce one more convenient notation. For a measurable function  $f$  on  $\Omega$  and any  $E \in \mathcal{F}$ , we write

$$\int_E f d\mu := \int_{\Omega} f \mathbb{1}_E d\mu,$$

as long as the latter integral is defined.

We may extend  $\int_E f d\mu$  to functions defined only on  $E$ . Indeed, for any function  $f$  that is defined only on  $E$ , we extend it to a new function on  $\Omega$  by setting it equal to  $f$  on  $E$  and 0 off  $E$ . Abusing the notation a bit, we also write the function as  $f \mathbb{1}_E$ . One sees that  $f \mathbb{1}_E$  is measurable iff  $\{\omega \in \mathbb{E} : f(\omega) < c\} \in \mathcal{F}$  for every  $c \in \mathbb{R}$ . Now define  $\int_E f d\mu$  as above. There is an alternative way to achieve this extension; see Exercise 6.23.

So far, we've only given examples of expectations and integrals over relatively simple measure spaces, such as counting measures or Dirac's measures. Let's work on  $\mathbb{R}$ . Let  $F : \mathbb{R} \rightarrow \mathbb{R}$  be increasing and right continuous, and let  $\mu$  be its associated Lebesgue-Stieltjes measure. Instead of writing  $\int_{\mathbb{R}} f d\mu$ , we write

$$\int_{\mathbb{R}} f dF,$$

and call it the **Lebesgue-Stieltjes integral** of  $f$  with respect to  $F$ . There are two cases where Lebesgue-Stieltjes integrals are relatively computable: one is that  $\mu$  is a combination of Dirac measures (Exercise 5.8); the other is that  $F$  has a density, which we postpone to Chapter 8.

Let's look at the most important case. Recall that if  $F(x) = x$  for any  $x \in \mathbb{R}$ , then the Lebesgue-Stieltjes measure is the Lebesgue measure. The corresponding integral is called the **Lebesgue integral** and is written as

$$\int_{\mathbb{R}} f dm \quad \text{or} \quad \int_{\mathbb{R}} f dx.$$

Let  $f : [a, b] \rightarrow \mathbb{R}$  be continuous or monotone. Recall that we already have an integral of  $f$ , called the **Riemann integral** and denoted by

$$\int_a^b f dx.$$

On the other hand, as is discussed above, we can extend  $f$  to  $\mathbb{R}$  (the extended function is Borel-measurable and integrable; Exercise 6.19) and have the Lebesgue integral of  $f$ :

$$\int_{[a,b]} f dx.$$

For convenience, we use  $\int_a^b$  and  $\int_{[a,b]}$  to indicate the Riemann and Lebesgue integrals, respectively.

6.8. PROPOSITION. *Let  $f : [a, b] \rightarrow \mathbb{R}$  be continuous or monotone. Then*

$$\int_{[a,b]} f dx = \int_a^b f dx.$$

PROOF. Let's write the proof for  $a = 0$  and  $b = 1$ . For every  $n \in \mathbb{N}$ , define  $f_n : \mathbb{R} \rightarrow \mathbb{R}$  by setting it to 0 on  $(-\infty, 0)$  and  $[1, \infty)$ <sup>1</sup> and to the value of  $f$  at the left endpoint  $\frac{k-1}{n}$  over each interval  $[\frac{k-1}{n}, \frac{k}{n})$ ,  $k = 1, \dots, n$ . See Figure 1 for illustration. We can write out  $f_n$  as

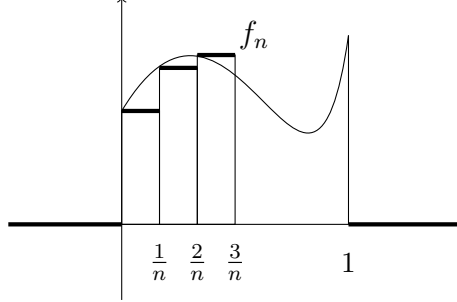
$$f_n = \sum_{k=1}^n f\left(\frac{k-1}{n}\right) \mathbb{1}_{\left[\frac{k-1}{n}, \frac{k}{n}\right)}.$$

We want to show that  $f_n \xrightarrow{a.e.} f \mathbb{1}_{[0,1]}$  on  $\mathbb{R}$ . The convergence is clear on  $(-\infty, 0) \cup (1, \infty)$  since all the functions are 0 there. We don't care about convergence at  $x = 0$  or  $1$ , since  $m(\{0, 1\}) = 0$ . We claim that if  $x \in (0, 1)$  is a continuous point of  $f$ , then  $f_n(x) \rightarrow f(x)$ . Indeed, take any  $\varepsilon > 0$ . Then continuity of  $f$  at  $x$  implies that there exists a small  $\delta > 0$  such that  $(x - \delta, x + \delta) \subset (0, 1)$  and

$$|f(x') - f(x)| < \varepsilon \quad \text{for any } x' \in (x - \delta, x + \delta).$$

---

<sup>1</sup>As one will see, inclusion or exclusion of endpoints does not matter as we only need to guarantee a.e. convergence so that Monotone Convergence Theorem is applicable; we choose this way for notational convenience.

FIGURE 1. Graph of  $f_n$ 

Let  $n_0 = \lceil \frac{1}{\delta} \rceil + 1 \in \mathbb{N}$ . For every  $n \in \mathbb{N}$ , since  $x \in (0, 1) \subset \bigcup_{k=1}^n [\frac{k-1}{n}, \frac{k}{n})$ , there exists a unique  $k$  from  $\{1, 2, \dots, n\}$  such that  $x \in [\frac{k-1}{n}, \frac{k}{n})$ . Thus if  $n \geq n_0$ ,

$$\left| x - \frac{k-1}{n} \right| < \frac{k}{n} - \frac{k-1}{n} = \frac{1}{n} \leq \frac{1}{n_0} < \delta,$$

and consequently,

$$|f_n(x) - f(x)| = \left| f\left(\frac{k-1}{n}\right) - f(x) \right| < \varepsilon.$$

This proves the claim. If  $f$  is continuous on  $[0, 1]$ , then it is immediate that  $f_n \xrightarrow{a.e.} f \mathbb{1}_{[0,1]}$  on  $\mathbb{R}$ . Let  $f$  be monotone. Recall from Example 0.1 that the set of points where  $f$  is discontinuous is finite or countably infinite and thus has Lebesgue measure 0. Therefore, we again get  $f_n \xrightarrow{a.e.} f \mathbb{1}_{[0,1]}$  on  $\mathbb{R}$ .

Next, let  $M := \sup_{x \in [0,1]} |f(x)|$ . Since  $f$  is continuous or monotone on  $[0, 1]$ ,  $M < \infty$ . Set  $f^* = M \mathbb{1}_{[0,1]}$ . Then  $\int_{\mathbb{R}} f^* dx = M$  and  $f^*$  is integrable. Moreover, it is clear that  $|f_n| \leq f^*$  on  $\mathbb{R}$  for every  $n \in \mathbb{N}$ . Thus by Dominated Convergence Theorem,

$$\int_{\mathbb{R}} f_n dx \longrightarrow \int_{\mathbb{R}} f \mathbb{1}_{[0,1]} dx = \int_{[0,1]} f dx.$$

On the other hand, each  $f_n$  is a simple function and direct computation gives  $\int_{\mathbb{R}} f_n dx = \sum_{k=1}^n f\left(\frac{k-1}{n}\right) \frac{1}{n}$ , which is a Riemann sum. Since Riemann sums converge to the Riemann integral, we get the desired equality.  $\square$

**6.9. COROLLARY.** *Let  $f : [a, \infty) \rightarrow \mathbb{R}$  be non-negative and is either continuous or monotone. Then*

$$\int_{[a, \infty)} f dx = \int_a^\infty f dx := \lim_{n \rightarrow \infty} \int_a^n f dx.$$

We leave the proof to the reader as an exercise.

### Exercises

Exercises 6.1-6.18 are set over an arbitrary probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ .

6.1. Show that if  $X \in L^1$  then  $\lim_n \mathbb{E}[|X| \mathbb{1}_{\{|X| > n\}}] = 0$ .

6.2. Show that if  $X \in L^1$  then  $\lim_{n \rightarrow \infty} n\mathbb{P}(|X| > n) = 0$ .

6.3. Show that  $X \in L^1$  iff  $\sum_{k=1}^{\infty} k\mathbb{P}(k-1 \leq |X| < k) < \infty$ .

6.4. Suppose that  $X \in L^\infty$ . Show that  $|X| \leq \|X\|_\infty$  a.s.

6.5. Deduce Proposition 6.5 from (6.4).

6.6. Prove Hölder's Inequality and Minkowski's Inequality for  $p = 1$  and  $p = \infty$ .

6.7. Let  $X \in L^2$ . The **variance** of  $X$  is defined by  $\mathbb{V}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2]$ . Show that  $\mathbb{V}[X] = 0$  iff  $X$  is a.s. equal to a constant.

6.8. Let  $X \in L^2$ . Show that  $\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ . Deduce that  $|\mathbb{E}[X]| \leq \|X\|_2$ . Deduce that  $\|X\|_1 \leq \|X\|_2$ .

6.9. For  $X, Y \in L^2$ , their **covariance** is defined by

$$\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

(If the covariance is 0, we say that the two random variables are **uncorrelated**). If  $\mathbb{V}[X], \mathbb{V}[Y] > 0$ , their **correlation** is defined by

$$\text{Cor}[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}}$$

Show that  $-1 \leq \text{Cor}[X, Y] \leq 1$ . Show that

$$\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

6.10. Show that if  $1 \leq p \leq \infty$  then  $\|X\|_1 \leq \|X\|_p \leq \|X\|_\infty$ .

6.11. Show that if  $1 \leq p \leq q \leq \infty$  then  $\|X\|_p \leq \|X\|_q$ .

6.12. Let  $1 \leq p < \infty$ . Show that if  $0 \leq X_n \uparrow X$  then  $\|X_n\|_p \uparrow \|X\|_p$  and if  $0 \leq X_n \downarrow X$  then  $\|X_n\|_p \downarrow \|X\|_p$ .

6.13. Let  $1 \leq p \leq \infty$ . Let  $X_n \geq 0$  for any  $n \in \mathbb{N}$ . Show that  $\|\sum_{n=1}^{\infty} X_n\| \leq \sum_{n=1}^{\infty} \|X_n\|_p$ .

6.14. Show that  $L^p$  is a vector space for any  $1 \leq p \leq \infty$ .



6.15. Let  $1 \leq p \leq \infty$  and  $(X_n)_{n \in \mathbb{N}}$  be a sequence of random variables in  $L^p$ . Show that if  $\sum_1^\infty \|X_n\|_p < \infty$ , then  $\sum_{n=1}^\infty |X_n| < \infty$  a.s. and there exists a random variable  $X \in L^p$  such that  $X = \sum_{n=1}^\infty X_n$  a.s. and  $\|X - \sum_{k=1}^n X_k\|_p \leq \sum_{k=n+1}^\infty \|X_k\|_p$  for every  $n \in \mathbb{N}$ .

6.16. Let  $1 \leq p \leq \infty$ . Let  $(X_n)_{n \in \mathbb{N}}$  be a sequence in  $L^p$  that is Cauchy, i.e., for any  $\varepsilon > 0$ , there exists  $n_0 \in \mathbb{N}$  such that  $\|X_n - X_m\| < \varepsilon$  whenever  $n, m \geq n_0$ . Show that there exists a strictly increasing sequence  $(n_k)_{k \in \mathbb{N}}$  in  $\mathbb{N}$  such that  $\|X_{n_{k+1}} - X_{n_k}\| \leq \frac{1}{2^k}$  for every  $k \in \mathbb{N}$ .

6.17. Let  $1 \leq p \leq \infty$ . Let  $(X_n)_{n \in \mathbb{N}}$  be a Cauchy sequence in  $L^p$ . If there exist a subsequence  $(X_{n_k})_{k \in \mathbb{N}}$  and  $X$  such that  $\|X_{n_k} - X\|_p \rightarrow 0$  then  $\|X_n - X\|_p \rightarrow 0$ .

6.18. Let  $1 \leq p \leq \infty$ . Show that  $L^p$  is a Banach space, i.e., for any Cauchy sequence  $(X_n)$  in it, there exists  $X \in L^p$  such that  $\|X_n - X\|_p \rightarrow 0$ .

6.19. Let  $f : [a, b] \rightarrow \mathbb{R}$  be continuous or monotone. Show that  $f \mathbb{1}_{[a, b]}$  is Borel-measurable and Lebesgue integrable.

6.20. Show that the Radon-Nikodym derivative is unique up to a.s. equality.

6.21. Let  $X$  be a non-negative measurable function. After replacing  $\mathbb{P}$  with a  $\sigma$ -finite measure, show that  $\mu$  defined by (6.5) is a  $\sigma$ -finite measure. Moreover,  $\mu$  is finite if  $X$  is integrable.

6.22. Let  $\mu, \nu$  be two  $\sigma$ -finite measures on  $(\Omega, \mathcal{F})$  such that  $\nu(E) = 0$  whenever  $E \in \mathcal{F}$  and  $\mu(E) = 0$ . Show that there exists a non-negative measurable function  $f : \Omega \rightarrow \mathbb{R}$  such that  $\nu(E) = \int_E f d\mu$  for any  $E \in \mathcal{F}$ .

6.23. For any non-empty set  $E \in \mathcal{F}$ , recall from Exercise 1.1 that  $\mathcal{F}|_E := \{F : F \in \mathcal{F}, F \subset E\}$  is a  $\sigma$ -algebra over  $E$ . Observe that  $f : \mathbb{E} \rightarrow \mathbb{R}$  is  $\mathcal{F}|_E$ -measurable iff  $f \mathbb{1}_E$  is measurable. Let  $\mu$  be any measure on  $\mathcal{F}$ . Define  $\mu|_E : \mathcal{F}|_E \rightarrow [0, \infty]$  by  $\mu|_E(F) = \mu(F)$  for any  $F \in \mathcal{F}|_E$ . Show that  $\int_E f d\mu|_E = \int_E f d\mu$ .

6.24. Prove Corollary 6.9.



## CHAPTER 7

### Product Measures

In this chapter, we systematically study how to build higher-dimensional measures from low-dimensional ones.

#### 1. Construction of product measures

Let  $(\Omega, \mathcal{F}, \mathbb{P})$  and  $(\Gamma, \mathcal{G}, \mathbb{Q})$  be two probability spaces. For  $A \in \mathcal{F}$  and  $B \in \mathcal{G}$ ,  $A \times B$  is called a *measurable rectangle*. Let  $\mathcal{F} \times \mathcal{G}$  be the  $\sigma$ -algebra generated by all measurable rectangles, i.e.,

$$\mathcal{F} \times \mathcal{G} := \sigma(\{A \times B : A \in \mathcal{F}, B \in \mathcal{G}\}).$$

We want to construct a probability measure  $\mu$  on  $(\Omega \times \Gamma, \mathcal{F} \times \mathcal{G})$  satisfying the following condition:

$$\mu(A \times B) = \mathbb{P}(A) \times \mathbb{Q}(B), \quad \text{for any } A \in \mathcal{F}, B \in \mathcal{G}.$$

Basically, it means that if we interpret  $\mathbb{P}(A)$  and  $\mathbb{Q}(B)$  as the “length” of  $A$  and  $B$ , respectively, then we want  $\mu$  to measure the “area” of  $A \times B$ .

We introduce two natural approaches for the construction of the desired measure. For the first approach, let’s use sets in  $\mathbb{R}^2$  for illustration. Assume that we know how to measure the length of objects in  $\mathbb{R}$ , in particular, line segments. Then we know how to measure the areas of rectangles: set the area as the product of the length of the two sides. We can then extend the measurement to more complex objects in  $\mathbb{R}^2$  by covering the object using rectangles. See Figure 1 for illustration. When we use smaller and smaller

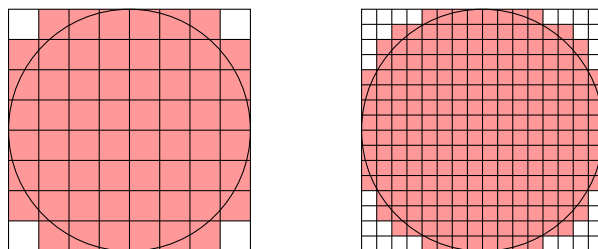


FIGURE 1. Cover a disk by rectangles.

rectangles to cover the disk, we can intuitively feel that the total area of the rectangles in the covering, which we term as the area of the covering, approximates the “area” of the disk. We put quotation marks around the word area, because we have not defined the area of the disk yet. A bit sneaky, we may define the area of the disk as the infimum of the areas of all rectangular coverings. (In Figure 1, because the disk has very nice shape, we cover it using finitely many rectangles. For a general set, we should use countably infinitely many rectangles to cover it.)

The mathematical formulation of this approach of obtaining measurement of complex objects from that of simple ones is actually (3.2). We sketch the construction explicitly. Let  $\mathcal{A}$  be the collection of all the unions of finitely many disjoint measurable rectangles, and for any  $E = \bigcup_{k=1}^n A_k \times B_k$ , where  $A_k \times B_k$ 's are disjoint measurable rectangles, set

$$\underline{\mu}(E) = \sum_{k=1}^n \mathbb{P}(A_k) \mathbb{Q}(B_k).$$

Then  $\mathcal{A}$  is an algebra and  $\underline{\mu}$  is a pre-measure over  $\mathcal{A}$ ; see Exercises 7.1-7.3. Applying Theorem 3.3, we get the desired measure  $\mu$  over  $\sigma(\mathcal{A}) = \mathcal{F} \times \mathcal{G}$ . Clearly, for any  $A \in \mathcal{F}$  and  $B \in \mathcal{G}$ ,  $\mu(A \times B) = \underline{\mu}(A \times B) = \mathbb{P}(A) \mathbb{Q}(B)$ . In particular,  $\mu(\Omega \times \Gamma) = 1$ . By (3.2), we actually know the explicit definition of  $\mu$ : For any  $E \in \mathcal{F} \times \mathcal{G}$ ,

$$(7.1) \quad \mu(E) = \inf \left\{ \sum_{n=1}^{\infty} \underline{\mu}(A_n) : A_n \in \mathcal{A} \text{ for each } n \in \mathbb{N}, E \subset \bigcup_{n=1}^{\infty} A_n \right\}.$$

One can replace sequences in  $\mathcal{A}$  by sequences of disjoint measurable rectangles in the above formula; see Exercise 7.4.

Now we focus on the second approach. Let  $E$  be a non-empty subset of  $\Omega \times \Gamma$ . The idea is to reduce the dimension of the set. We do it as follows. Pick any  $\omega \in \Omega$ . Consider the  $\omega$ -section of  $E$ :

$${}^{\omega}E := \{\gamma : (\omega, \gamma) \in E\}.$$

See the left figure in Figure 2. This is a subset of  $\Gamma$ , so we may get its length as  $\mathbb{Q}({}^{\omega}E)$ . We can then integrate the length of all sections across  $\Omega$  using  $\mathbb{P}$ :

$$(7.2) \quad \nu(E) = \int_{\Omega} \mathbb{Q}({}^{\omega}E) d\mathbb{P}(\omega).$$

See the right figure in Figure 2.

**7.1. EXAMPLE.** Let  $A \in \mathcal{F}$  and  $B \in \mathcal{G}$ . If  $\omega \in A$ , then  ${}^{\omega}A \times B = B$ ; if  $\omega \notin A$ , then  ${}^{\omega}A \times B = \emptyset$ . Thus  $\mathbb{Q}({}^{\omega}A \times B) = \mathbb{Q}(B)$  if  $\omega \in A$  and

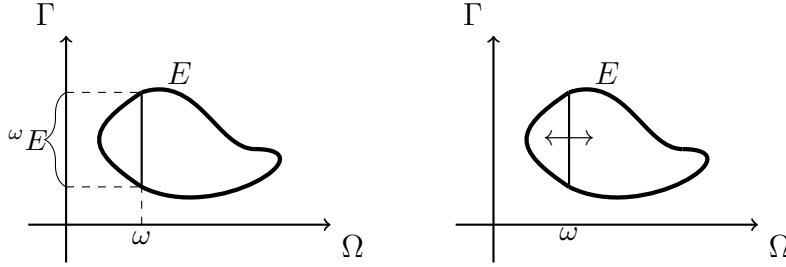


FIGURE 2. Low-dimensional sections.

$\mathbb{Q}({}^\omega A \times B) = 0$  if  $\omega \notin A$ . It follows that  $\mathbb{Q}({}^\omega A \times B) = \mathbb{Q}(B)\mathbb{1}_A(\omega)$  and hence  $\nu(A \times B) = \mathbb{P}(A)\mathbb{Q}(B)$ .

Of course, for a general set  $E \in \mathcal{F} \times \mathcal{G}$ , we shall ask whether the set  ${}^\omega E$  and the function  $\mathbb{Q}(\cdot E)$  are always measurable in their appropriate senses, so that the right hand side of (7.2) is defined. The answer is yes.

**7.1. LEMMA.** *Let  $E$  be a set in  $\mathcal{F} \times \mathcal{G}$ . Then  ${}^\omega E \in \mathcal{G}$  for any  $\omega \in \Omega$ .*

**PROOF.** Let  $\mathcal{D} = \{E \subset \Omega \times \Gamma : {}^\omega E \in \mathcal{G} \text{ for every } \omega \in \Omega\}$ . We want to show that  $\mathcal{F} \times \mathcal{G} \subset \mathcal{D}$ . Denote the collection of all measurable rectangles by  $\mathcal{P}$ . Then  $\sigma(\mathcal{P}) = \mathcal{F} \times \mathcal{G}$ ,  $\mathcal{P} \subset \mathcal{D}$  by Example 7.1, and it is immediate verification that  $\mathcal{P}$  is a  $\pi$ -system. Thus by Dynkin's  $\pi$ - $\lambda$  theorem 1.10, it suffices to show that  $\mathcal{D}$  is a  $\lambda$ -system over  $\Omega \times \Gamma$ . We verify it now. Clearly, being a measurable rectangle,  $\emptyset = \emptyset \times \Gamma \in \mathcal{D}$ <sup>1</sup>. Take any  $E \in \mathcal{D}$ . Then for any  $\omega \in \Omega$ ,  ${}^\omega E \in \mathcal{G}$ . Thus

$$\begin{aligned} {}^\omega(E^c) &= \{\gamma \in \Gamma : (\omega, \gamma) \in E^c\} = \{\gamma \in \Gamma : (\omega, \gamma) \notin E\} \\ (7.3) \quad &= \Gamma \setminus \{\gamma \in \Gamma : (\omega, \gamma) \in E\} = ({}^\omega E)^c \in \mathcal{G}. \end{aligned}$$

Consequently,  $E^c \in \mathcal{D}$ . Similarly, for any (disjoint) sequence  $(E_n)_{n \in \mathbb{N}}$  in  $\mathcal{D}$ , it follows from

$$\begin{aligned} {}^\omega\left(\bigcup_{n=1}^{\infty} E_n\right) &= \left\{\gamma \in \Gamma : (\omega, \gamma) \in \bigcup_{n=1}^{\infty} E_n\right\} \\ (7.4) \quad &= \bigcup_{n=1}^{\infty} \{\gamma \in \Gamma : (\omega, \gamma) \in E_n\} = \bigcup_{n=1}^{\infty} {}^\omega E_n \end{aligned}$$

that  $\bigcup_{n=1}^{\infty} E_n \in \mathcal{D}$ . This proves that  $\mathcal{D}$  is a  $\lambda$ -system over  $\Omega \times \Gamma$ .  $\square$

**7.2. LEMMA.**  *$\mathbb{Q}(\cdot E)$  is  $\mathcal{F}$ -measurable for any  $E \in \mathcal{F} \times \mathcal{G}$ .*

<sup>1</sup>The first empty set is a subset of  $\Omega \times \Gamma$ . The second empty set is a subset of  $\Omega$ .

PROOF. Let  $\mathcal{D} = \{E \in \mathcal{F} \times \mathcal{G} : \mathbb{Q}(\cdot E) \text{ is measurable}\}$ . By Example 7.1 again,  $\mathcal{D}$  contains all measurable rectangles. Thus as before, it suffices to show that  $\mathcal{D}$  is a  $\lambda$ -system. We apply the probability  $\mathbb{Q}$  to the first and last terms of the formulas in the proof of Lemma 7.1. Then for any  $E \in \mathcal{D}$ ,

$$\mathbb{Q}(\omega(E^c)) = \mathbb{Q}((\omega E)^c) = 1 - \mathbb{Q}(\omega E),$$

implying that  $\mathbb{Q}(\cdot(E^c))$  is measurable, so that  $E^c \in \mathcal{D}$ . For any disjoint sequence  $(E_n)_{n \in \mathbb{N}}$  in  $\mathcal{D}$ , note that  $(\omega E_n)$  is a disjoint sequence in  $\mathcal{G}$  for any  $\omega \in \Omega$ . Thus it follows from

$$(7.5) \quad \mathbb{Q}\left(\omega\left(\bigcup_{n=1}^{\infty} E_n\right)\right) = \mathbb{Q}\left(\bigcup_{n=1}^{\infty} \omega E_n\right) = \sum_{n=1}^{\infty} \mathbb{Q}(\omega E_n)$$

that  $\mathbb{Q}(\cdot(\bigcup_{n=1}^{\infty} E_n))$  is measurable, so that  $\bigcup_{n=1}^{\infty} E_n \in \mathcal{D}$ .  $\square$

7.1. THEOREM. *For any two probability spaces  $(\Omega, \mathcal{F}, \mathbb{P})$  and  $(\Gamma, \mathcal{G}, \mathbb{Q})$ , there exists a unique measure  $\nu$  on  $\mathcal{F} \times \mathcal{G}$  such that*

$$(7.6) \quad \nu(A \times B) = \mathbb{P}(A) \times \mathbb{Q}(B) \quad \text{for any } A \in \mathcal{F}, B \in \mathcal{G}.$$

PROOF. Clearly, if such a  $\nu$  exists, we have  $\nu(\Omega \times \Gamma) = \mathbb{P}(\Omega)\mathbb{Q}(\Gamma) = 1$ . Thus the uniqueness part follows from Theorem 2.9, since the set of all measurable rectangles is a  $\pi$ -system. For the existence part, we only need to show that the measure  $\nu$  in (7.2) is a measure. Indeed, it is clear that  $\nu(\emptyset) = 0$ . The countable additivity follows from taking expectations of the first and last terms in (7.5). This completes the proof.  $\square$

Of course, we can define the  $\gamma$ -sections,  $\gamma E$ , of a set  $E \in \mathcal{F} \times \mathcal{G}$  and then define a measure of  $E$  in a similar fashion as in (7.2). By the uniqueness part in Theorem 7.1, we must have, for any  $E \in \mathcal{F} \times \mathcal{G}$ ,

$$(7.7) \quad \mu(E) = \int_{\Omega} \mathbb{Q}(\omega E) d\mathbb{P}(\omega) = \int_{\Gamma} \mathbb{P}(\gamma E) d\mathbb{Q}(\gamma),$$

where  $\mu$  is as in (7.1). From now on, we rewrite the measure as  $\mathbb{P} \times \mathbb{Q}$  and call it the **product measure** of  $\mathbb{P}$  and  $\mathbb{Q}$ .

## 2. Fubini Theorem

We now study integrals with respect to the product measure. The following theorem says that integrating with respect to the product measure is the same as integrating with respect to the two measures one by one.

7.2. THEOREM (Tonelli-Fubini). *Let  $(\Omega, \mathcal{F}, \mathbb{P})$  and  $(\Gamma, \mathcal{G}, \mathbb{Q})$  be two probability spaces.*

(a) Let  $X : \Omega \times \Gamma \rightarrow [0, \infty]$  be  $\mathcal{F} \times \mathcal{G}$ -measurable. Then

$$(7.8) \quad \int_{\Omega \times \Gamma} X(\omega, \gamma) d\mathbb{P} \times \mathbb{Q}(\omega, \gamma) = \int_{\Omega} \left[ \int_{\Gamma} X(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega)$$

(b) Let  $X \in L^1(\Omega \times \Gamma)$  be such that  $\int_{\Gamma} X(\omega, \gamma) d\mathbb{Q}(\gamma)$  is defined for every  $\omega \in \Omega$ . Then  $\int_{\Gamma} X(\cdot, \gamma) d\mathbb{Q}(\gamma) \in L^1(\Omega)$  and (7.8) holds.

At a first glance, one may feel no ideas to prove the theorem. But once we connect (7.8) to (7.7), the proof of Theorem 7.2 will become transparent and almost immediate. Take any  $E \in \mathcal{F} \times \mathcal{G}$ . Note that  $\mathbb{1}_E(\omega, \gamma) = 1$  iff  $(\omega, \gamma) \in E$  iff  $\gamma \in {}^{\omega}E$  iff  $\mathbb{1}_{{}^{\omega}E}(\gamma) = 1$ . Therefore,

$$\int_{\Gamma} \mathbb{1}_E(\omega, \gamma) d\mathbb{Q}(\gamma) = \int_{\Gamma} \mathbb{1}_{{}^{\omega}E}(\gamma) d\mathbb{Q}(\gamma) = \mathbb{Q}({}^{\omega}E)$$

for any  $\omega \in \Omega$ . Consequently, with  $X = \mathbb{1}_E$ , (7.8) becomes

$$\mathbb{P} \times \mathbb{Q}(E) = \int_{\Omega} \mathbb{Q}({}^{\omega}E) d\mathbb{P}(\omega),$$

which is precisely (7.7). In other words, (7.8) holds for indicator functions. The rest of the proof will fall into our general routine: prove it for simple functions, and then for non-negative functions and for general functions.

However, before proceeding to the proof, we need to show that for each  $\omega \in \Omega$ ,  $X(\omega, \cdot)$  is  $\mathcal{G}$ -measurable, so that  $\int_{\Gamma} X(\omega, \gamma) d\mathbb{Q}(\gamma)$  is possibly defined, and also that  $\int_{\Gamma} X(\cdot, \gamma) d\mathbb{Q}(\gamma)$  is  $\mathcal{F}$ -measurable, so that the double integral  $\int_{\Omega} \left[ \int_{\Gamma} X(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega)$  is possibly defined. We include the arguments for these measurability issues in the proof of Theorem 7.2.

PROOF OF THEOREM 7.2. (a). If  $X = \mathbb{1}_E$  for some  $E \in \mathcal{F} \times \mathcal{G}$ , then as is observed above,  $\mathbb{1}_E(\omega, \cdot) = \mathbb{1}_{{}^{\omega}E}$  is  $\mathcal{G}$ -measurable for any  $\omega \in \Omega$ ,  $\int_{\Gamma} \mathbb{1}_E(\cdot, \gamma) d\mathbb{Q}(\gamma)$  is  $\mathcal{F}$ -measurable, and

$$\int_{\Omega \times \Gamma} \mathbb{1}_E(\omega, \gamma) d\mathbb{P} \times \mathbb{Q}(\omega, \gamma) = \int_{\Omega} \left[ \int_{\Gamma} \mathbb{1}_E(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega).$$

Now let  $\phi$  be any non-negative simple function on  $\Omega \times \Gamma$ , say,  $\phi = \sum_{j=1}^k c_j \mathbb{1}_{E_j}$ , where all  $c_j$ 's are non-negative and all  $E_j$ 's lie in  $\mathcal{F} \times \mathcal{G}$ . Recall that linear combinations of measurable functions are measurable (Corollary 4.9). Thus by the indicator function case,  $\phi(\omega, \cdot) = \sum_{j=1}^k c_j \mathbb{1}_{E_j}(\omega, \cdot)$  is  $\mathcal{G}$ -measurable for any  $\omega \in \Omega$ ,  $\int_{\Gamma} \phi(\cdot, \gamma) d\mathbb{Q}(\gamma) = \sum_{j=1}^k c_j \int_{\Gamma} \mathbb{1}_{E_j}(\cdot, \gamma) d\mathbb{Q}(\gamma)$  is

$\mathcal{F}$ -measurable, where the equality is due to linearity of expectations, and

$$\begin{aligned}
& \int_{\Omega \times \Gamma} \phi(\omega, \gamma) d\mathbb{P} \times \mathbb{Q}(\omega, \gamma) \\
&= \sum_{j=1}^k c_j \int_{\Omega \times \Gamma} \mathbb{1}_{E_j}(\omega, \gamma) d\mathbb{P} \times \mathbb{Q}(\omega, \gamma) = \sum_{j=1}^k c_j \int_{\Omega} \left[ \int_{\Gamma} \mathbb{1}_{E_j}(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega) \\
&= \int_{\Omega} \sum_{j=1}^k c_j \left[ \int_{\Gamma} \mathbb{1}_{E_j}(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega) = \int_{\Omega} \left[ \int_{\Gamma} \sum_{j=1}^k c_j \mathbb{1}_{E_j}(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega) \\
&= \int_{\Omega} \left[ \int_{\Gamma} \phi(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega).
\end{aligned}$$

Now let  $X \geq 0$  be general. By Theorem 4.12, we can take a sequence  $(\phi_n)_{n \in \mathbb{N}}$  of simple functions such that  $0 \leq \phi_n \uparrow X$  on  $\Omega \times \Gamma$ . Recall that the limit of a sequence of measurable functions is also measurable (Proposition 4.11). Then by the simple function case, since

$$\phi_n(\omega, \cdot) \uparrow X(\omega, \cdot),$$

$X(\omega, \cdot)$  is  $\mathcal{G}$ -measurable for any  $\omega \in \Omega$ . Taking expectation with respect to  $\mathbb{Q}$ , we have, by Monotone Convergence Theorem,

$$\int_{\Gamma} \phi_n(\omega, \gamma) d\mathbb{Q}(\gamma) \uparrow \int_{\Gamma} X(\omega, \gamma) d\mathbb{Q}(\gamma)$$

for any  $\omega \in \Omega$ , implying in particular that  $\int_{\Gamma} X(\cdot, \gamma) d\mathbb{Q}(\gamma)$  is  $\mathcal{F}$ -measurable. Taking expectation with respect to  $\mathbb{P}$  and applying Monotone Convergence Theorem again, we have

$$\int_{\Omega} \left[ \int_{\Gamma} \phi_n(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega) \uparrow \int_{\Omega} \left[ \int_{\Gamma} X(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega).$$

On the other hand, by the simple function case,

$$\begin{aligned}
& \int_{\Omega} \left[ \int_{\Gamma} \phi_n(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega) \\
&= \int_{\Omega \times \Gamma} \phi_n(\omega, \gamma) d\mathbb{P} \times \mathbb{Q}(\omega, \gamma) \\
&\uparrow \int_{\Omega \times \Gamma} X(\omega, \gamma) d\mathbb{P} \times \mathbb{Q}(\omega, \gamma),
\end{aligned}$$

where the convergence in the last step is due to Monotone Convergence Theorem applied to  $0 \leq \phi_n \uparrow X$  over the product space  $(\Omega \times \Gamma, \mathcal{F} \times \mathcal{G}, \mathbb{P} \times \mathbb{Q})$ . Combining the last two equations, we finish the proof of (a).



(b). The proof of this part has no mathematical ideas but only some technicalities. Let  $X : \Omega \times \Gamma \rightarrow [-\infty, \infty]$  be integrable. Then by the non-negative case,  $X(\omega, \cdot) = X^+(\omega, \cdot) - X^-(\omega, \cdot)$  is  $\mathcal{G}$ -measurable for any  $\omega \in \Omega$ . As functions in  $\omega$ ,  $\int_{\Gamma} X^{\pm}(\cdot, \gamma) d\mathbb{Q}(\gamma)$  may take infinite values, but their difference is defined at every point of  $\Omega$ , since we assume that  $\int_{\Gamma} X(\omega, \gamma) d\mathbb{Q}(\gamma)$  is defined for every  $\omega \in \Omega$ , which by the definition of integrals of  $X(\omega, \cdot)$  with respect to  $\mathbb{Q}$  is equal to

$$\int_{\Gamma} X^+(\omega, \gamma) d\mathbb{Q}(\gamma) - \int_{\Gamma} X^-(\omega, \gamma) d\mathbb{Q}(\gamma).$$

Thus by Remark 4.10 applied to the functions  $\int_{\Gamma} X^{\pm}(\cdot, \gamma) d\mathbb{Q}(\gamma)$  on  $\Omega$ , it follows that  $\int_{\Gamma} X(\omega, \gamma) d\mathbb{Q}(\gamma)$  is  $\mathcal{F}$ -measurable. Moreover,

$$\begin{aligned} & \int_{\Omega} \left[ \int_{\Gamma} X^{\pm}(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega) \\ & \leq \int_{\Omega} \left[ \int_{\Gamma} |X(\omega, \gamma)| d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega) = \int_{\Omega \times \Gamma} |X(\omega, \gamma)| d\mathbb{P} \times \mathbb{Q}(\omega, \gamma) < \infty, \end{aligned}$$

implying that  $\int_{\Gamma} X^{\pm}(\cdot, \gamma) d\mathbb{Q}(\gamma)$  are both integrable. Thus  $\int_{\Gamma} X(\cdot, \gamma) d\mathbb{Q}(\gamma) = \int_{\Gamma} X^+(\cdot, \gamma) d\mathbb{Q}(\gamma) - \int_{\Gamma} X^-(\cdot, \gamma) d\mathbb{Q}(\gamma)$  is integrable, and

$$\begin{aligned} & \int_{\Omega} \left[ \int_{\Gamma} X(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega) \\ & = \int_{\Omega} \left[ \int_{\Gamma} X^+(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega) - \int_{\Omega} \left[ \int_{\Gamma} X^-(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega) \\ & = \int_{\Omega \times \Gamma} X^+(\omega, \gamma) d\mathbb{P} \times \mathbb{Q}(\omega, \gamma) - \int_{\Omega \times \Gamma} X^-(\omega, \gamma) d\mathbb{P} \times \mathbb{Q}(\omega, \gamma) \\ & = \int_{\Omega \times \Gamma} X(\omega, \gamma) d\mathbb{P} \times \mathbb{Q}(\omega, \gamma), \end{aligned}$$

where the first equality is due to linearity of expectation with respect to  $\mathbb{P}$  and the second equality is due to the non-negative case.  $\square$

Of course, one may do the double integral by integrating with respect to  $\mathbb{P}$  first and then to  $\mathbb{Q}$ . Parallel results follow. Comparing the double integral in this case to that in the previous case, we obtain that for any  $\mathcal{F} \times \mathcal{G}$ -measurable function  $X : \Omega \times \Gamma \rightarrow [0, \infty]$ ,

$$(7.9) \quad \int_{\Omega} \left[ \int_{\Gamma} X(\omega, \gamma) d\mathbb{Q}(\gamma) \right] d\mathbb{P}(\omega) = \int_{\Gamma} \left[ \int_{\Omega} X(\omega, \gamma) d\mathbb{P}(\omega) \right] d\mathbb{Q}(\gamma).$$

That is, we can change the order of integration in double integrals.

The non-negative case in Theorem 7.2(a) and (7.9) is usually referred to as Tonelli Theorem, and the general case in Theorem 7.2(b) and (7.9) is referred to as Fubini Theorem.

7.3. REMARK. The results in Sections 1 and 2 hold for  $\sigma$ -finite measures spaces. In fact, Lemma 7.2 is the only place that needs additional care. We leave the verification to the reader as an exercise.

7.2. EXAMPLE. Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space and  $(X_n)_{n \in \mathbb{N}}$  be a sequence of non-negative random variables on  $\Omega$ . Consider  $\mathbb{N}$  endowed with the  $\sigma$ -algebra  $\mathcal{P}(\mathbb{N})$  and the counting measure. The product  $\sigma$ -algebra is

$$\mathcal{P}(\mathbb{N}) \times \mathcal{F} := \sigma\left(\left\{A \times E : A \subset \mathbb{N}, E \in \mathcal{F}\right\}\right).$$

Define  $F : \mathbb{N} \times \Omega \rightarrow [0, \infty]$  by  $F(n, \omega) = X_n(\omega)$ . For any  $c \in \mathbb{R}$ ,

$$\{F \leq c\} = \{(n, \omega) : X_n(\omega) \leq c\} = \bigcup_{n=1}^{\infty} \{n\} \times \{X_n \leq c\} \in \mathcal{P}(\mathbb{N}) \times \mathcal{F}.$$

Thus  $F$  is  $\mathcal{P}(\mathbb{N}) \times \mathcal{F}$ -measurable. Applying (7.9), one gets that

$$\sum_{n=1}^{\infty} \mathbb{E}[X_n] = \mathbb{E}\left[\sum_{n=1}^{\infty} X_n\right].$$

7.3. EXAMPLE. Let  $1 \leq p < \infty$ . Let  $X \geq 0$  be a random variable on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Whenever  $0 \leq t < X(\omega)^p$ , we can find a positive rational number  $r$  such that  $t < r < X(\omega)^p$ . Thus it is easy to see that

$$\begin{aligned} & \left\{(t, \omega) \in \mathbb{R} \times \Omega : 0 \leq t < X(\omega)^p\right\} \\ &= \bigcup_r \left\{(t, \omega) \in \mathbb{R} \times \Omega : 0 \leq t < r, r < X(\omega)^p\right\} \\ &= \bigcup_r [0, r) \times \left\{\omega \in \Omega : X(\omega) > r^{\frac{1}{p}}\right\} \in \mathcal{B} \times \mathcal{F}. \end{aligned}$$

Therefore,  $\mathbb{1}_{\{(t, \omega) \in \mathbb{R} \times \Omega : 0 \leq t < X(\omega)^p\}}$  is  $\mathcal{B} \times \mathcal{F}$ -measurable. Applying (7.9) with  $\mathbb{Q}$  replaced with the Lebesgue measure, one gets that

$$\begin{aligned} \mathbb{E}[X^p] &= \int_{\Omega} X(\omega)^p d\mathbb{P}(\omega) = \int_{\Omega} \left[ \int_{\mathbb{R}} \mathbb{1}_{\{(t, \omega) \in \mathbb{R} \times \Omega : 0 \leq t < X(\omega)^p\}} dt \right] d\mathbb{P}(\omega) \\ &= \int_{\mathbb{R}} \left[ \int_{\Omega} \mathbb{1}_{\{(t, \omega) \in \mathbb{R} \times \Omega : 0 \leq t < X(\omega)^p\}} d\mathbb{P}(\omega) \right] dt \\ &= \int_{\mathbb{R}} \left[ \int_{\Omega} \mathbb{1}_{\{X > t^{\frac{1}{p}}\}}(\omega) d\mathbb{P}(\omega) \right] dt \\ &= \int_{\mathbb{R}} \mathbb{1}_{[0, \infty)}(t) \mathbb{P}(X > t^{\frac{1}{p}}) dt. \end{aligned}$$

For any  $t \geq 0$ , denote  $\omega_X(t) = \mathbb{P}(X > t)$ . Then  $\omega_X$  is a decreasing function on  $[0, \infty)$ . Thus by Corollary 6.9, we have

$$\mathbb{E}[X^p] = \int_0^\infty \omega_X(t^{\frac{1}{p}}) dt = \int_0^\infty \omega_X(s) p s^{p-1} ds,$$

where the second equality follows from a change of variables  $t = s^p$  for the Riemann integrals.

### 3. Higher-dimensional constructions

We are not satisfied with constructing the product space of only two probability (or  $\sigma$ -finite) measure spaces. For example, once we know how to measure the length of objects in  $\mathbb{R}$ , in addition to knowing how to measure the area of objects in  $\mathbb{R}^2$ , we also want to know how to measure the volume of objects in  $\mathbb{R}^3$ . In another word, if we have three probability (or  $\sigma$ -finite) measure spaces  $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k)$ ,  $k = 1, 2, 3$ , how can we get their product space? Of course, we can first get the product space  $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, \mathbb{P}_1 \times \mathbb{P}_2)$  and then cross with the third space to get the following probability space

$$(\Omega_1 \times \Omega_2 \times \Omega_3, (\mathcal{F}_1 \times \mathcal{F}_2) \times \mathcal{F}_3, (\mathbb{P}_1 \times \mathbb{P}_2) \times \mathbb{P}_3).$$

Alternatively, we may cross the last two spaces first and then cross with the first one to get the following probability space

$$(\Omega_1 \times \Omega_2 \times \Omega_3, \mathcal{F}_1 \times (\mathcal{F}_2 \times \mathcal{F}_3), \mathbb{P}_1 \times (\mathbb{P}_2 \times \mathbb{P}_3)).$$

What is the relationship between these two spaces then? They are identical! Firstly, one can verify that

$$(\mathcal{F}_1 \times \mathcal{F}_2) \times \mathcal{F}_3 = \mathcal{F}_1 \times (\mathcal{F}_2 \times \mathcal{F}_3) = \sigma(\{A \times B \times C : A \in \mathcal{F}_1, B \in \mathcal{F}_2, C \in \mathcal{F}_3\})$$

(Exercise 7.8), so that the two probability spaces have the same  $\sigma$ -algebra. Secondly, observe that for any  $A \in \mathcal{F}_1, B \in \mathcal{F}_2, C \in \mathcal{F}_3$ ,

$$(\mathbb{P}_1 \times \mathbb{P}_2) \times \mathbb{P}_3(A \times B \times C) = \mathbb{P}_1 \times \mathbb{P}_2(A \times B) \mathbb{P}_3(C) = \mathbb{P}_1(A) \mathbb{P}_2(B) \mathbb{P}_3(C)$$

and

$$\mathbb{P}_1 \times (\mathbb{P}_2 \times \mathbb{P}_3)(A \times B \times C) = \mathbb{P}_1(A) \mathbb{P}_2 \times \mathbb{P}_3(B \times C) = \mathbb{P}_1(A) \mathbb{P}_2(B) \mathbb{P}_3(C),$$

and thus

$$(\mathbb{P}_1 \times \mathbb{P}_2) \times \mathbb{P}_3(A \times B \times C) = \mathbb{P}_1 \times (\mathbb{P}_2 \times \mathbb{P}_3)(A \times B \times C).$$

Since the collection  $\{A \times B \times C : A \in \mathcal{F}_1, B \in \mathcal{F}_2, C \in \mathcal{F}_3\}$  is a  $\pi$ -system generating the  $\sigma$ -algebra, we get  $(\mathbb{P}_1 \times \mathbb{P}_2) \times \mathbb{P}_3 = \mathbb{P}_1 \times (\mathbb{P}_2 \times \mathbb{P}_3)$ , by Theorem 2.9. That is, the order to get the product space of three probability

spaces does not matter. In view of this, we may simply denote the product space as  $(\Omega_1 \times \Omega_2 \times \Omega_3, \mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3, \mathbb{P}_1 \times \mathbb{P}_2 \times \mathbb{P}_3)$ . With the standard technique of passing from indicator functions to simple functions to non-negative functions, one can also easily show that, for any  $\mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3$ -measurable function  $X : \Omega_1 \times \Omega_2 \times \Omega_3 \rightarrow [0, \infty]$ ,

$$\begin{aligned} & \int_{\Omega_1 \times \Omega_2 \times \Omega_3} X(\omega_1, \omega_2, \omega_3) d\mathbb{P}_1 \times \mathbb{P}_2 \times \mathbb{P}_3 \\ &= \int_{\Omega_1} \left[ \int_{\Omega_2} \left[ \int_{\Omega_3} X(\omega_1, \omega_2, \omega_3) d\mathbb{P}_3(\omega_3) \right] d\mathbb{P}_2(\omega_2) \right] d\mathbb{P}_1(\omega_1) \\ &= \int_{\Omega_2} \left[ \int_{\Omega_3} \left[ \int_{\Omega_1} X(\omega_1, \omega_2, \omega_3) d\mathbb{P}_1(\omega_1) \right] d\mathbb{P}_3(\omega_3) \right] d\mathbb{P}_2(\omega_2), \end{aligned}$$

or in any order one may like to arrange 1, 2, 3. For a  $\mathcal{F}_1 \times \mathcal{F}_2 \times \mathcal{F}_3$ -measurable function  $X$  that may take negative values, similar results hold as long as the intermediate integrals are all defined.

In general, whenever we have  $d$  probability (or  $\sigma$ -finite) measure spaces  $(\Omega_k, \mathcal{F}_k, \mathbb{P}_k)$ ,  $k = 1, 2, \dots, d$ , we can get the product spaces by gluing them together one by one:  $\Omega_1 \times \Omega_2$ ,  $(\Omega_1 \times \Omega_2) \times \Omega_3$ ,  $((\Omega_1 \times \Omega_2) \times \Omega_3) \times \Omega_4$ , etc. We denote the final product space

$$\left( \prod_{k=1}^d \Omega_k, \prod_{k=1}^d \mathcal{F}_k, \prod_{k=1}^d \mathbb{P}_k \right).$$

Remark that

$$\prod_{k=1}^d \mathcal{F}_k = \sigma \left( \left\{ \prod_{k=1}^d E_k : E_k \in \mathcal{F}_k, k = 1, \dots, d \right\} \right)$$

(Exercise 7.7) and that  $\prod_{k=1}^d \mathbb{P}_k$  is the only measure on  $\prod_{k=1}^d \mathcal{F}_k$  such that

$$\prod_{k=1}^d \mathbb{P}_k \left( \prod_{k=1}^d E_k \right) = \prod_{k=1}^d \mathbb{P}_k(E_k)$$

for any  $E_1 \in \mathcal{F}_1, \dots, E_d \in \mathcal{F}_d$ . The order of gluing the  $d$ -spaces together and expressing the integral with respect to the product measure as a multiple integral does not matter.

**7.4. EXAMPLE.** Consider  $\prod_{k=1}^d (\mathbb{R}, \mathcal{B}, m)$ . Note that  $\prod_{k=1}^d \mathcal{B} = \mathcal{B}^d$  (Exercise 7.9). Instead of writing  $\prod_{k=1}^d m$ , we abuse the notation and still write it as  $m$ , and call it the Lebesgue measure on  $\mathbb{R}^d$ .

7.5. EXAMPLE. Let  $F_1, \dots, F_d$  be distribution functions on  $\mathbb{R}$  with associated Lebesgue-Stieltjes measures  $\mu_1, \dots, \mu_d$ . Then we can obtain the unique measure  $\mu$  on  $(\mathbb{R}^d, \mathcal{B}^d)$  such that

$$\mu\left(\prod_{k=1}^d (a_k, b_k]\right) = \prod_{k=1}^d (F_k(b_k) - F_k(a_k))$$

for any  $a_k, b_k \in \mathbb{R}$  with  $a_k < b_k$ ,  $k = 1, \dots, d$ .

### Exercises

7.1. Let  $\mathcal{A}$  be as in Section 1 and  $E \in \mathcal{A}$  be non-empty. Show that we can obtain a partition  $\{E_k\}_{1 \leq k \leq n}$  of  $\Omega$  and a partition  $\{F_j\}_{1 \leq j \leq m}$  of  $\Gamma$  such that  $E$  is the union of some of the  $E_k \times F_j$ 's. Show that  $\mathcal{A}$  is an algebra over  $\Omega \times \Gamma$ .

7.2. Let  $\underline{\mu}$  be as in Section 1. Suppose that  $(A_n \times B_n)_{n \in \mathbb{N}}$  be a disjoint sequence of measurable rectangles whose union is a measurable rectangle  $A \times B$ . Show that for any  $\omega \in \Omega$  and  $\gamma \in \Gamma$ ,

$$\mathbb{1}_A(\omega)\mathbb{1}_B(\gamma) = \sum_{n=1}^{\infty} \mathbb{1}_{A_n}(\omega)\mathbb{1}_{B_n}(\gamma).$$

Deduce  $\mathbb{1}_A(\omega)\mathbb{Q}(B) = \sum_{n=1}^{\infty} \mathbb{1}_{A_n}(\omega)\mathbb{Q}(B_n)$  and  $\mathbb{P}(A)\mathbb{Q}(B) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)\mathbb{Q}(B_n)$ .

7.3. Let  $\underline{\mu}$  and  $\mathcal{A}$  be as in Section 1. Show that  $\underline{\mu}$  is a pre-measure on  $\mathcal{A}$ .

7.4. Let  $\mathcal{A}, \underline{\mu}, \mu$  be as in Section 1. Show that for any  $E \in \mathcal{F} \times \mathcal{G}$ ,

$$\begin{aligned} & \mu(E) \\ &= \inf \left\{ \sum_{n=1}^{\infty} \underline{\mu}(A_n) : (A_n)_{n \in \mathbb{N}} \text{ is a disjoint sequence in } \mathcal{A}, E \subset \bigcup_{n=1}^{\infty} A_n \right\} \\ &= \inf \left\{ \sum_{n=1}^{\infty} \underline{\mu}(R_n) : (R_n)_{n \in \mathbb{N}} \text{ is a disjoint sequence of measurable rectangles,} \right. \\ & \quad \left. E \subset \bigcup_{n=1}^{\infty} R_n \right\}. \end{aligned}$$

7.5. Verify the first equality in (7.3) and the first equality in (7.4).

7.6. Verify that the results in Sections 1 and 2 hold for  $\sigma$ -finite measures spaces.

7.7. Suppose  $\mathcal{F}$  and  $\mathcal{G}$  are the  $\sigma$ -algebras generated by two collections  $\mathcal{C}$  and  $\mathcal{D}$ , respectively. Show that  $\mathcal{F} \times \mathcal{G}$  is generated by  $\mathcal{C} \times \mathcal{D} := \{C \times D : C \in \mathcal{C}, D \in \mathcal{D}\}$ . Use induction to extend this result to multiple  $\sigma$ -algebras.

7.8. Use Exercise 7.7 to show that

$$(\mathcal{F}_1 \times \mathcal{F}_2) \times \mathcal{F}_3 = \mathcal{F}_1 \times (\mathcal{F}_2 \times \mathcal{F}_3) = \sigma(\{A \times B \times C : A \in \mathcal{F}_1, B \in \mathcal{F}_2, C \in \mathcal{F}_3\}),$$

7.9. Use Exercise 7.7 to show that  $\mathcal{B}^{d_1} \times \mathcal{B}^{d_2} \times \cdots \times \mathcal{B}^{d_k} = \mathcal{B}^{d_1 + \cdots + d_k}$ .

7.10. Let  $X, Y$  be random variables over  $(\Omega, \mathcal{F}, \mathbb{P})$  and  $(\Gamma, \mathcal{G}, \mathbb{Q})$ , respectively. Consider the function  $XY : \Omega \times \Gamma \rightarrow [-\infty, \infty]$ . Show that  $XY$  is  $\mathcal{F} \times \mathcal{G}$ -measurable. If  $X, Y$  are real-valued, one can similarly define  $X - Y$ . Show that  $X - Y$  is  $\mathcal{F} \times \mathcal{G}$ -measurable.

7.11. Prove Example 7.5.

7.12. Let  $(\Omega, \mathcal{F}, \mu)$  and  $(\Gamma, \mathcal{G}, \nu)$  be two  $\sigma$ -finite measure spaces. Let  $X : \Omega \times \Gamma \rightarrow [0, \infty]$  be  $\mathcal{F} \times \mathcal{G}$ -measurable. For  $1 \leq p < \infty$ , show that

$$\left( \int_{\Gamma} \left( \int_{\Omega} X(\omega, \gamma) d\mu(\omega) \right)^p d\nu(\gamma) \right)^{\frac{1}{p}} \leq \int_{\Gamma} \left( \int_{\Omega} X(\omega, \gamma)^p d\nu(\gamma) \right)^{\frac{1}{p}} d\mu(\omega).$$

If  $(\Omega, \mathcal{F}, \mu)$  is  $\mathbb{N}$  with the counting measure, the formula reduces to

$$\left\| \sum_{n=1}^{\infty} X_n \right\|_p \leq \sum_{n=1}^{\infty} \|X_n\|_p,$$

which is Exercise 6.13. In view of this, the first inequality is called Minkowski Inequality in integral form.

7.13. Let  $X$  be a bounded random variable over  $(\Omega, \mathcal{F}, \mathbb{P})$ . Show that if  $\int_{\Omega \times \Omega} |X(\omega) - X(\omega')| d\mathbb{P} \times \mathbb{P}(\omega, \omega') = 0$  then  $X$  a.s. equals a constant.

## CHAPTER 8

### Distributions

Let  $\Omega$  be the set of all Canadians and let  $X : \Omega \rightarrow \mathbb{R}$  be the 2019 income of Canadians. In most cases, it will not be of economic concern what  $X(\omega)$  is for a particular Canadian  $\omega$ . But rather, it is of great importance to know, e.g., what is the probability that a randomly selected Canadian's 2019 income is below \$15k, i.e.,  $P(X < 15k)$ , or say, if the middle class 2019 income is \$80k, then what is the probability of a randomly selected Canadian's 2019 is middle-class or above, i.e.,  $P(X \geq 80k)$ ? In another word, we would like to know how  $X$  distributes its values?

We have a more intuitive example explaining the meaning of “distribution”. Say, you throw a fair die. If you get a small number 1, 2, 3, then you lose \$5; if you get a big even number 4, 6, then you win \$3; if you get 5, then you win \$1. Let  $X$  be your net gain after one toss. Then  $\mathbb{P}(X = -5) = \frac{1}{2}$ ,  $\mathbb{P}(X = 1) = \frac{1}{6}$ , and  $\mathbb{P}(X = 3) = \frac{1}{3}$ . So we may say that  $X$  distributes  $\frac{1}{2}$  of its values to  $-5$ ,  $\frac{1}{6}$  of its values to  $\frac{1}{6}$  and  $\frac{1}{3}$  of its values to 3.

In this chapter, we study distributions in details.

#### 1. Probability distributions

Let  $X$  be a random variable over a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Define  $F_X : \mathbb{R} \rightarrow [0, 1]$  by

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

One sees that  $F^X$  is increasing, right-continuous satisfying that  $F(\infty) = 1$  and  $F(-\infty) = 0$ . Thus by Proposition 3.8, it generates its Lebesgue-Stieltjes measure on  $(\mathbb{R}, \mathcal{B})$ . It can be easily verified (Exercise 8.1) that the Lebesgue-Stieltjes measure is given by follows:

$$(8.1) \quad \mathbb{P}^X(B) := \mathbb{P}(X \in B), \quad B \in \mathcal{B}.$$

We call  $\mathbb{P}^X$  the **probability distribution**, or simply **distribution**, of  $X$  because for any set  $B \in \mathcal{B}$ ,  $\mathbb{P}^X(B)$  tells the chance that  $X$  distributes its values to  $B$ , or in general,  $\mathbb{P}^X$  tells how  $X$  distributes its values. From now on,  $\mathbb{P}^X$ , instead of  $\mathbb{P}$ , is usually our focus of study.

We call  $F^X$  the **cumulative distribution function (CDF)** of  $X$ . With the new notation of  $\mathbb{P}^X$ , we have

$$(8.2) \quad F(x) = \mathbb{P}^X((-\infty, x]), \quad x \in \mathbb{R}.$$

Being the Lebesgue-Stieltjes measure of  $F^X$ ,  $\mathbb{P}^X$  is determined by  $F^X$ . Thus we may refer to  $F^X$  as the distribution of  $X$  as well.

8.1. DEFINITION. Let  $X$  be a random variable with CDF  $F$ . Its **probability mass function (PMF)**  $f : \mathbb{R} \rightarrow [0, 1]$  is defined by

$$f^X(x) = \mathbb{P}(X = x) = \mathbb{P}^X(\{x\}), \quad x \in \mathbb{R}.$$

By definition,  $f^X(x) > 0$  means precisely that the value  $x$  is taken by  $X$  with a positive probability. Recall from Proposition 3.6 that, for any  $x \in \mathbb{R}$ ,

$$f^X(x) = \mathbb{P}^X(\{x\}) = F^X(x) - F^X(x-);$$

thus  $f^X(x) > 0$  if and only if  $F^X$  has a jump at  $x$ , in which case, the size of the jump is  $f^X(x)$ .

8.1. EXAMPLE. Let  $X$  be a discrete random variable with values  $\{x_k\}_{k \in \mathbb{N}}$ , each of which has a positive probability to be taken (it does not matter if  $X$  only takes finitely many values). Then

$$f^X(x) = \begin{cases} \mathbb{P}(X = x_k) > 0 & \text{if } x = x_k \text{ for some } x_k, \\ \mathbb{P}(X = x) = 0 & x \notin \{x_k\}_{k \in \mathbb{N}}; \end{cases}$$

and

$$1 = \mathbb{P}(X \in \mathbb{R}) = \mathbb{P}(X = x_k \text{ for some } x_k) = \sum_{k=1}^{\infty} \mathbb{P}^X(X = x_k) = \sum_{k=1}^{\infty} f^X(x_k).$$

Consequently, the CDF  $F^X$  has jumps precisely at  $x_k$ ,  $k \in \mathbb{N}$ , and the total jumps are  $\sum_{k=1}^{\infty} (F(x_k) - F(x_k-)) = 1$ . Moreover, for any  $B \in \mathcal{B}$ ,

$$(8.3) \quad \mathbb{P}^X(B) = \mathbb{P}(X \in B) = \sum_{k: x_k \in B} f(x_k) = \sum_{k=1}^{\infty} f(x_k) \delta_{x_k}(B).$$

That is,

$$\mathbb{P}^X = \sum_{k=1}^{\infty} f(x_k) \delta_{x_k},$$

a form of probability measures that we have discussed in Example 2.8. In particular, the PMF determines  $\mathbb{P}^X$ —thus, we refer to the PMF as the distribution for discrete random variables as well, but of course, the PMF looks neater and is more workable as will be seen.



The converses of the above statements are also correct, namely, if the PMF of a random variable are positive at finitely many or countably infinitely many points with a sum of 1, or if the CDF has total jumps equal to 1, or if the probability distribution is of the form in Example 2.8, then it is discrete (Exercise 8.2).

8.2. EXAMPLE. Let  $X$  be a random variable that takes only two values 0, 1 both with positive probabilities. Then  $X$  is said to be **binary** or is called a **Bernoulli trial**. Its distribution is given by

$$p := f^X(1), \quad 1 - p = f^X(0),$$

or equivalently,

$$\mathbb{P}^X = p\delta_1 + (1 - p)\delta_0.$$

In practice,  $X$  may count the number of heads when flipping a coin once, with  $p$  the probability of getting a head.

8.3. EXAMPLE. Let  $n \in \mathbb{N}$  and  $p \in (0, 1)$ . A random variable  $X$  that takes values  $0, 1, \dots, n$  with the PMF

$$f^X(x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \dots, n,$$

is called a **binomial** random variable. Its distribution is called a **binomial distribution**, written as  $\text{Bi}(n, p)$ . In practice,  $X$  may count the number of heads when flipping a coin for  $n$  times, with  $p$  the probability of getting a head for each flip.

While probability mass functions work well for discrete random variables, we need to introduce probability density functions for continuous random variables.

8.2. DEFINITION. A random variable  $X$  is said to be **continuous** if  $F^X$  is continuous.  $X$  is said to be **absolutely continuous** if there exists a function  $f^X : \mathbb{R} \rightarrow [0, \infty)$  such that

$$F^X(x) = \int_{(-\infty, x]} f^X(t) dt, \quad x \in \mathbb{R};$$

or more commonly, in this case, we say that  $X$  has a **probability density function (PDF)**  $f^X$ .

8.3. REMARK. (a) If  $f_1, f_2$  are both PDFs for a random variable  $X$ , then  $f_1 = f_2$  m-a.e. (cf. XXX). In fact, in this case, the PDF is given by  $f^X = (F^X)'$  m-a.e.

- (b) For a PDF  $f^X$ ,  $\int_{-\infty}^{\infty} f^X(t)dt = 1$ .
- (c) A random variable with a PDF is always continuous (Exercise 8.3).  
The converse is not true in general, i.e., a continuous random variable may not have PDFs.

8.4. EXAMPLE. Suppose that  $X$  has PDF  $f^X$ . Then the distribution of  $X$  is given by

$$(8.4) \quad \mathbb{P}^X(B) = \int_B f^X(t)dt, \quad B \in \mathcal{B}.$$

Indeed, the indefinite integral in the right hand side is a probability measure and coincides with  $\mathbb{P}^X$  for all intervals of the form  $(-\infty, x]$ ,  $x \in \mathbb{R}$ . Thus it coincides with  $\mathbb{P}^X$  for any  $B \in \mathcal{B}$  (Corollary 2.10).

Like PMFs for discrete random variables, PDFs are referred to as the distributions of absolutely continuous random variables. These two classes of random variables are most used in reality.

8.5. EXAMPLE. A random variable  $X$  is said to follow the **normal distribution**, written as  $X \sim N(\mu, \sigma^2)$ , if it has density

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R},$$

where  $\mu \in \mathbb{R}$  and  $\sigma > 0$  are fixed constants. When  $\mu = 0$  and  $\sigma = 1$ , the distribution is called **standard normal distribution**.

- (a) If  $X \sim N(\mu, \sigma^2)$ , then  $Z := \frac{X-\mu}{\sigma} \sim N(0, 1)$ . Indeed,

$$\begin{aligned} \mathbb{P}(Z \leq z) &= \mathbb{P}\left(\frac{X-\mu}{\sigma} \leq z\right) = \mathbb{P}(X \leq \mu + \sigma z) \\ &= \int_{-\infty}^{\mu+\sigma z} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt, \end{aligned}$$

where the last step follows from a change of variable  $t = \frac{x-\mu}{\sigma}$ .

- (b) Similarly, it is easy to verify that if  $Z \sim N(0, 1)$ , then  $X := \mu + \sigma Z \sim N(\mu, \sigma^2)$  (Exercise 8.4).

## 2. The Expectation Formula

The probability distributions  $\mathbb{P}^X$  (in particular, PMFs and PDFs) bring significant convenience and simplicity for calculating expectations, as it pushes everything from  $(\Omega, \mathcal{F}, \mathbb{P})$  to  $(\mathbb{R}, \mathcal{B}, \mathbb{P}^X)$ .

8.4. THEOREM. *Let  $X$  be any random variable over  $(\Omega, \mathcal{F}, \mathbb{P})$ . Let  $\mathbb{E}^X$  be the expectation over  $(\mathbb{R}, \mathcal{B})$  with respect to  $\mathbb{P}^X$ . Let  $h : \mathbb{R} \rightarrow \mathbb{R}$  be a Borel-measurable function. Then*

$$(8.5) \quad \mathbb{E}[h(X)] = \mathbb{E}^X[h],$$

*where the expectations either both exist or both do not exist. Furthermore, if  $X$  is discrete with PMF  $f^X$ , then*

$$(8.6) \quad \mathbb{E}[h(X)] = \sum_{x_k} h(x_k) f^X(x_k),$$

*where  $x_k$ 's are the values admitted by  $X$  with positive probabilities; if  $X$  has PDF  $f^X$ , then*

$$(8.7) \quad \mathbb{E}[h(X)] = \int_{\mathbb{R}} h(t) f^X(t) dt.$$

PROOF. Without loss of generality, we assume that  $h \geq 0$ .

Step I:  $h$  is an indicator function, say,  $h = \mathbb{1}_B$  for some  $B \in \mathcal{B}$ . In this case,  $\mathbb{1}_B(X) = 1$  if  $X \in B$  and 0 otherwise, thus  $\mathbb{1}_B(X) = \mathbb{1}_{\{X \in B\}}$ . Hence,

$$\mathbb{E}[\mathbb{1}_B(X)] = \mathbb{E}[\mathbb{1}_{\{X \in B\}}] = \mathbb{P}(X \in B) = \mathbb{P}^X(B) = \mathbb{E}^X[\mathbb{1}_B].$$

Step II:  $h$  is simple, say,  $h = \sum_{k=1}^n c_k \mathbb{1}_{B_k}$ . Then by Step I and linearity of expectations with respect to both  $\mathbb{E}$  and  $\mathbb{E}^X$ ,

$$\mathbb{E}[h(X)] = \sum_{k=1}^n c_k \mathbb{E}[\mathbb{1}_{B_k}(X)] = \sum_{k=1}^n c_k \mathbb{E}^X[\mathbb{1}_{B_k}] = \mathbb{E}^X[h].$$

Step III:  $h \geq 0$  is general. Take a sequence  $(\phi_n)$  of simple functions such that  $0 \leq \phi_n \uparrow h$ . Then  $0 \leq \phi_n(X) \uparrow h(X)$ . Using Step II and applying Monotone Convergence Theorem with respect to both  $\mathbb{E}$  and  $\mathbb{E}^X$ , we have

$$\mathbb{E}[h(X)] = \lim_n \mathbb{E}[\phi_n(X)] = \lim_n \mathbb{E}^X[\phi_n] = \mathbb{E}^X[h].$$

In particular,  $\mathbb{E}[h(X)] < \infty$  if and only if  $\mathbb{E}^X[h] < \infty$ . This proves (8.5).

The proofs of (8.6) and (8.7) go along similar lines; for  $h = \mathbb{1}_B$ , they follow from (8.1) and (8.4), respectively.  $\square$

Let's do some classical examples.

8.6. EXAMPLE. Suppose  $X \sim \text{Bi}(n, p)$ . With  $h(t) = t$  for all  $t \in \mathbb{R}$ ,  $h(X) = X$ , and

$$\begin{aligned}
 \mathbb{E}[X] &= \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^n x \binom{n}{x} p^x (1-p)^{n-x} \\
 &= \sum_{x=1}^n x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
 &= np \sum_{x=1}^n \frac{(n-1)!}{(x-1)![(n-1)-(x-1)]!} p^{x-1} (1-p)^{(n-1)-(x-1)} \\
 &= np \sum_{y=0}^{n-1} \frac{(n-1)!}{y![(n-1)-y]!} p^y (1-p)^{(n-1)-y} \\
 &= np,
 \end{aligned}$$

where the fourth inequality is due to change of variable  $y = x - 1$  and the last one is due to that the new summands are the PDF of  $\text{Bi}(n-1, p)$  and thus the sum is 1. With  $h(t) = t^2$  for all  $t \in \mathbb{R}$ ,  $h(X) = X^2$ , and

$$\begin{aligned}
 \mathbb{E}[X^2] &= \sum_{x=0}^n x^2 \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^n x^2 \binom{n}{x} p^x (1-p)^{n-x} \\
 &= \sum_{x=1}^n x(x-1) \binom{n}{x} p^x (1-p)^{n-x} + \sum_{x=1}^n x \binom{n}{x} p^x (1-p)^{n-x} \\
 &= \sum_{x=2}^n x(x-1) \binom{n}{x} p^x (1-p)^{n-x} + np \\
 &= \sum_{x=2}^n x(x-1) \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} + np \\
 &= \sum_{x=2}^n \frac{n!}{(x-2)!(n-x)!} p^x (1-p)^{n-x} + np \\
 &= n(n-1)p^2 \sum_{x=2}^n \frac{(n-2)!}{(x-2)![(n-2)-(x-2)]!} p^{x-2} (1-p)^{(n-2)-(x-2)} + np \\
 &= n(n-1)p^2 \sum_{y=0}^{n-2} \frac{(n-2)!}{y![(n-2)-y]!} p^y (1-p)^{(n-2)-y} + np \\
 &= n(n-1)p^2 + np.
 \end{aligned}$$

It follows that

$$\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = np(1-p).$$

8.7. EXAMPLE. Suppose  $X \sim N(\mu, \sigma^2)$ . Then with  $h(t) = t$  for all  $t \in \mathbb{R}$ ,  $h(X) = X$ , and

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} t \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \int_{-\infty}^{\infty} (\mu + \sigma s) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{s^2}{2}} \sigma ds \\ &= \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{s^2}{2}} ds + \int_{-\infty}^{\infty} s \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{s^2}{2}} ds \\ &= \mu + 0 = \mu,\end{aligned}$$

where the second equality is due to change of variable  $t = \mu + \sigma s$  and we use the fact that  $\int_{-\infty}^{\infty} s \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{s^2}{2}} ds = 0$  (Exercise 8.5). With  $h(t) = t^2$  for all  $t \in \mathbb{R}$ ,  $h(X) = X^2$ , and

$$\begin{aligned}\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} t^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \int_{-\infty}^{\infty} (\mu + \sigma s)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{s^2}{2}} \sigma ds \\ &= \mu^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{s^2}{2}} ds + 2\mu\sigma \int_{-\infty}^{\infty} s \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{s^2}{2}} ds \\ &\quad + \sigma^2 \int_{-\infty}^{\infty} s^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{s^2}{2}} ds \\ &= \mu^2 + \sigma^2.\end{aligned}$$

Therefore,

$$\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \sigma^2.$$

### 3. Higher-dimensional analogues

What we have discussed applies to random vectors (Definition 4.5). Let  $(X_1, \dots, X_d)$  be a random vector. We define its CDF  $F^{(X_1, \dots, X_d)} : \mathbb{R}^d \rightarrow \mathbb{R}$  by

$$F^{(X_1, \dots, X_d)}(x_1, \dots, x_d) = \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d), \quad (x_1, \dots, x_d) \in \mathbb{R}^d.$$

We define its probability distribution  $\mathbb{P}^{(X_1, \dots, X_d)}$  on  $(\mathbb{R}^d, \mathcal{B}^d)$  by

$$\mathbb{P}^{(X_1, \dots, X_d)}(B) = \mathbb{P}\left((X_1, \dots, X_d) \in B\right), \quad B \in \mathcal{B}^d.$$

By Corollary 2.10, it is easy to verify that  $\mathbb{P}^{(X_1, \dots, X_d)}$  is the unique probability measure on  $(\mathbb{R}^d, \mathcal{B}^d)$  such that, for any  $(x_1, \dots, x_d) \in \mathbb{R}^d$ ,

$$F^{(X_1, \dots, X_d)}(x_1, \dots, x_d) = \mathbb{P}^{(X_1, \dots, X_d)}\left(\prod_{k=1}^d (-\infty, x_k]\right).$$

A random vector  $(X_1, \dots, X_d)$  is discrete if it admits finitely many or countably infinitely many values. We can define its PMF by

$$f^{(X_1, \dots, X_d)}(x_1, \dots, x_d) = \mathbb{P}(X_1 = x_1, \dots, X_d = x_d), \quad (x_1, \dots, x_d) \in \mathbb{R}^d.$$

Its properties are similar as outlined in Example 8.1. In particular, if  $\{\mathbf{x}_k\}_{k \in \mathbb{N}}$  are the values in  $\mathbb{R}^d$  admitted by  $(X_1, \dots, X_d)$  with positive probabilities, then we have

$$\begin{aligned} \mathbb{P}^{(X_1, \dots, X_d)}(B) &= \mathbb{P}((X_1, \dots, X_d) \in B) = \sum_{k: \mathbf{x}_k \in B} f(\mathbf{x}_k) \\ &= \sum_{k=1}^{\infty} f(\mathbf{x}_k) \delta_{\mathbf{x}_k}(B). \end{aligned}$$

That is,

$$\mathbb{P}^{(X_1, \dots, X_d)} = \sum_{k=1}^{\infty} f(\mathbf{x}_k) \delta_{\mathbf{x}_k},$$

Similarly, a random vector  $(X_1, \dots, X_d)$  is absolutely continuous or has PDF if there exists a non-negative function  $f^{(X_1, \dots, X_d)}$  such that

$$F^{(X_1, \dots, X_d)}(x_1, \dots, x_d) = \int_{\prod_{k=1}^d (-\infty, x_k]} f^{(X_1, \dots, X_d)}(t_1, \dots, t_d) d(t_1, \dots, t_d),$$

for any  $(x_1, \dots, x_d) \in \mathbb{R}^d$ . Similar properties as in Remark 8.3 hold. In particular, we have

$$(8.8) \quad \mathbb{P}^{(X_1, \dots, X_d)}(B) = \int_B f^{(X_1, \dots, X_d)}(t_1, \dots, t_d) d(t_1, \dots, t_d), \quad B \in \mathcal{B}^d.$$

As before, we may term  $F^{(X_1, \dots, X_d)}$  and  $\mathbb{P}^{(X_1, \dots, X_d)}$ , as well as the PMF and PDF whenever appropriate, as the distribution of the random vector  $(X_1, \dots, X_d)$ .

**3.1. Marginal distributions.** In Probability Theory,  $\mathbb{P}^{(X_1, \dots, X_d)}$  is often called the *joint* distribution of  $(X_1, \dots, X_d)$ ; similarly, the CDFs, PMFs and PDFs are also termed with the word “joint”. The corresponding distributions, CDFs, PMFs and PDFs for each individual  $X_k$ ’s are termed with “*marginal*”.

The joint distributions contain the marginal distributions as partial information. For example, if  $F^{(X_1, \dots, X_d)}$  is the joint CDF of  $(X_1, \dots, X_d)$ ,

$$\begin{aligned} F^{X_1}(x_1) &= \mathbb{P}(X_1 \in (-\infty, x_1]) = \mathbb{P}\left((X_1, X_2, \dots, X_d) \in (-\infty, x_1] \times \mathbb{R}^{d-1}\right) \\ &= \lim_{n \rightarrow \infty} \mathbb{P}\left((X_1, X_2, \dots, X_d) \in (-\infty, x_1] \times \prod_{k=2}^d (-\infty, n]\right) \\ &= \lim_{n \rightarrow \infty} F^{(X_1, \dots, X_d)}(x_1, n, \dots, n), \end{aligned}$$

which we can symbolically write as

$$F^{(X_1, \dots, X_d)}(x_1, \infty, \dots, \infty).$$

So the marginal CDF  $F^{X_1}$  is obtained from the joint CDF. For another example, if  $f^{(X_1, \dots, X_d)}$  is the joint density of  $(X_1, \dots, X_d)$ , then by (8.8),

$$\begin{aligned} F^{X_1}(x_1) &= \mathbb{P}(X_1 \in (-\infty, x_1]) = \mathbb{P}\left((X_1, X_2, \dots, X_d) \in (-\infty, x_1] \times \mathbb{R}^{d-1}\right) \\ &= \int_{(-\infty, x_1] \times \mathbb{R}^{d-1}} f^{(X_1, \dots, X_d)}(t_1, \dots, t_d) d(t_1, \dots, t_d) \\ &= \int_{(-\infty, x_1]} dt_1 \left( \int_{\mathbb{R}^{d-1}} f^{(X_1, \dots, X_d)}(t_1, \dots, t_d) d(t_2, \dots, t_d) \right), \end{aligned}$$

where the last equality is due to Fubini Theorem. Thus the PDF of  $X_1$  is obtained from the joint PDF:

$$\int_{\mathbb{R}^{d-1}} f^{(X_1, \dots, X_d)}(\bullet, t_2, \dots, t_d) d(t_2, \dots, t_d).$$

In particular, existence of joint PDF implies that of marginal PDFs.

This method can be extended to find the CDFs of any component of a random vector. Let's look at ***multivariate Gaussian distributions***.

8.8. EXAMPLE. A random vector  $\mathbf{X}$  is said to be normal or Gaussian if it has the following density function

$$(2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^t}, \quad \mathbf{x} \in \mathbb{R}^d,$$

where  $\Sigma$  is a  $d \times d$  positive definite matrix and  $\boldsymbol{\mu} \in \mathbb{R}^d$ . In this case, write  $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ , or simply  $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ .

Write  $\mathbf{X}_1 = (X_1, \dots, X_{d_1})$  and  $\mathbf{X}_2 = (X_{d_1+1}, \dots, X_d)$ , so  $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ . Accordingly, write  $\mathbf{x}_1 = (x_1, \dots, x_{d_1})$ ,  $\mathbf{x}_2 = (x_{d_1+1}, \dots, x_d)$ , and

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2), \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where  $\Sigma_{11}$  is the  $d_1 \times d_1$ -block in  $\Sigma$ . By symmetry,  $\Sigma_{21} = \Sigma_{12}^t$ . Then for any  $B \in \mathcal{B}^{d_1}$ , we have, by (8.8) and Fubini Theorem,

$$\begin{aligned} \mathbb{P}^{\mathbf{X}_1}(B) &= \mathbb{P}^{\mathbf{X}}(B \times \mathbb{R}^{d-d_1}) \\ &= \int_B d\mathbf{x}_1 \int_{\mathbb{R}^{d-d_1}} (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^t} d\mathbf{x}_2. \end{aligned}$$

Setting

$$\mathbf{b} = \boldsymbol{\mu}_2 + (\mathbf{x}_1 - \boldsymbol{\mu}_1)\Sigma_{11}^{-1}\Sigma_{12} \quad \text{and} \quad A = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12},$$

we have by direct simplifications (Exercise 8.6),

$$\begin{aligned} &(\mathbf{x} - \boldsymbol{\mu})\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})^t \\ (8.9) \quad &= (\mathbf{x}_1 - \boldsymbol{\mu}_1)\Sigma_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)^t + (\mathbf{x}_2 - \mathbf{b})A^{-1}(\mathbf{x}_2 - \mathbf{b})^t. \end{aligned}$$

Plugging this into the previous formula, we have

$$\begin{aligned} \mathbb{P}^{\mathbf{X}_1}(B) &= \int_B (2\pi)^{-\frac{d_1}{2}} \det(\Sigma)^{-\frac{1}{2}} \det(A)^{\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_1)\Sigma_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)^t} d\mathbf{x}_1 \\ &\quad \times \int_{\mathbb{R}^{d-d_1}} (2\pi)^{-\frac{d-d_1}{2}} \det(A)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}_2 - \mathbf{b})A^{-1}(\mathbf{x}_2 - \mathbf{b})^t} d\mathbf{x}_2. \end{aligned}$$

The second integral is equal to 1 because the integrand is precisely the PDF of  $N(\mathbf{b}, A)$ . Notice also that (Exercise 8.6)

$$(8.10) \quad \det(\Sigma)^{-1} \det(A) = \det(\Sigma_{11})^{-1}.$$

Thus

$$\mathbb{P}^{\mathbf{X}_1}(B) = \int_B (2\pi)^{-\frac{d_1}{2}} \det(\Sigma_{11})^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_1)\Sigma_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)^t} d\mathbf{x}_1.$$

Therefore,  $\mathbf{X}_1 \sim N(\boldsymbol{\mu}_1, \Sigma_{11})$ . Similarly, one shows that  $\mathbf{X}_2 \sim N(\boldsymbol{\mu}_2, \Sigma_{22})$ . Other components of  $\mathbf{X}$ , e.g.,  $(X_1, X_3, X_4)$ , can be handled in a similar fashion; they are all Gaussian whose parameters are extracted accordingly from  $\boldsymbol{\mu}$  and  $\Sigma$ . In particular, if we write  $\Sigma = (\sigma_{jk})$ , then  $X_k \sim N(\mu_k, \sigma_{kk})$  for each  $k = 1, \dots, d$ . Thus,  $\mu_k = \mathbb{E}[X_k]$  and  $\sigma_{kk} = \mathbb{V}[X_k]$ .

**3.2. The expectation formula.** The following theorem generalizes Theorem 8.4 and can be proved in the identical format.

**8.5. THEOREM.** *Let  $(X_1, \dots, X_d)$  be a random vector over  $(\Omega, \mathcal{F}, \mathbb{P})$ . Let  $\mathbb{E}^{(X_1, \dots, X_d)}$  be the expectation over  $(\mathbb{R}^d, \mathcal{B}^d)$  with respect to  $\mathbb{P}^{(X_1, \dots, X_d)}$ . Let  $h : \mathbb{R}^d \rightarrow \mathbb{R}$  be a Borel-measurable function. Then*

$$\mathbb{E}[h(X_1, \dots, X_d)] = \mathbb{E}^{(X_1, \dots, X_d)}[h],$$



where the expectations either both exist or both do not exist. Furthermore, if  $(X_1, \dots, X_d)$  is discrete with PMF  $f^{(X_1, \dots, X_d)}$ , then

$$\mathbb{E}[h(X_1, \dots, X_d)] = \sum_{\mathbf{x}_k} h(\mathbf{x}_k) f^{(X_1, \dots, X_d)}(\mathbf{x}_k),$$

where  $\mathbf{x}_k$ 's are the values admitted by  $(X_1, \dots, X_d)$  with positive probabilities; if  $(X_1, \dots, X_d)$  has PDF  $f^{(X_1, \dots, X_d)}$ , then

$$\mathbb{E}[h(X_1, \dots, X_d)] = \int_{\mathbb{R}^d} h(t_1, \dots, t_d) f^{(X_1, \dots, X_d)}(t_1, \dots, t_d) d(t_1, \dots, t_d).$$

8.9. EXAMPLE. Suppose  $\mathbf{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$ . By the expectation formula, we have

$$\begin{aligned} \text{CoV}[X_1, X_2] &= \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= \int_{\mathbb{R}^1} (x_1 - \mu_1)(x_2 - \mu_2) \frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^t} d\mathbf{x} \\ &= \int_{\mathbb{R}^1} y_1 y_2 \frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}\mathbf{y}\Sigma^{-1}\mathbf{y}} d\mathbf{y}. \end{aligned}$$

Write  $\Sigma = (\sigma_{jk})$ . Then

$$\Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{pmatrix}.$$

Consider the following change of variable

$$(t_1, t_2) = (y_1, y_2) \frac{1}{\sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2}} \begin{pmatrix} \sqrt{\sigma_{22}} & 0 \\ -\frac{\sigma_{12}}{\sqrt{\sigma_{22}}} & \sqrt{\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}} \end{pmatrix}.$$

Then (Exercise 8.7)

$$(8.11) \quad \mathbf{y}\Sigma^{-1}\mathbf{y}^t = t_1^2 + t_2^2,$$

$$\begin{aligned} (8.12) \quad \mathbf{y} &= \mathbf{t} \sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \frac{1}{\sqrt{\sigma_{22}} \sqrt{\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}}} \begin{pmatrix} \sqrt{\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}} & 0 \\ \frac{\sigma_{12}}{\sqrt{\sigma_{22}}} & \sqrt{\sigma_{22}} \end{pmatrix} \\ &= \mathbf{t} \begin{pmatrix} \frac{1}{\sqrt{\sigma_{22}}} \sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2} & 0 \\ \frac{\sigma_{12}}{\sqrt{\sigma_{22}}} & \sqrt{\sigma_{22}} \end{pmatrix}, \end{aligned}$$

and

$$d\mathbf{y} = \sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2} d\mathbf{t}.$$

Therefore,

$$\begin{aligned} \text{CoV}[X_1, X_2] &= \int_{\mathbb{R}^2} \left( \frac{1}{\sqrt{\sigma_{22}}} \sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2} t_1 + \frac{\sigma_{12}}{\sqrt{\sigma_{22}}} t_2 \right) (\sqrt{\sigma_{22}} t_2) \\ &\quad \times \frac{1}{2\pi \sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2}} e^{-\frac{1}{2}(t_1^2 + t_2^2)} \sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2} dt. \end{aligned}$$

Split the parenthesis in the integrand. Note that by symmetry, the integral of the term containing  $t_1 t_2$  is 0. Thus

$$\begin{aligned} \text{CoV}[X_1, X_2] &= \int_{\mathbb{R}^2} \sigma_{12} t_2^2 \frac{1}{2\pi \sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2}} e^{-\frac{1}{2}(t_1^2 + t_2^2)} \sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2} dt \\ &= \sigma_{12} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t_1^2}{2}} dt_1 \int_{\mathbb{R}} t_2^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{t_2^2}{2}} dt_2 \\ &= \sigma_{12}. \end{aligned}$$

That is,  $\sigma_{12}$  is precisely the covariance of  $X_1$  and  $X_2$ .

Suppose that  $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ , where  $\Sigma = (\sigma_{jk})$ . For any two distinct  $j, k$ , from Example 8.8 we know that

$$(X_j, X_k) \sim N \left( (\mu_j, \mu_k), \begin{pmatrix} \sigma_{jj} & \sigma_{jk} \\ \sigma_{kj} & \sigma_{kk} \end{pmatrix} \right).$$

Following this with an application of the above result, we obtain that  $\sigma_{jk} = \text{CoV}[X_j, X_k]$ .

For a random vector  $\mathbf{X} = (X_1, \dots, X_d)$ , we define its **mean vector** by

$$\mathbb{E}[\mathbf{X}] := (\mathbb{E}[X_1], \dots, \mathbb{E}[X_d])$$

and its **variance matrix** by

$$\mathbb{V}[\mathbf{X}] := \begin{pmatrix} \text{CoV}[X_1, X_1] & \dots & \text{CoV}[X_1, X_d] \\ \vdots & \ddots & \vdots \\ \text{CoV}[X_d, X_1] & \dots & \text{CoV}[X_d, X_d] \end{pmatrix}.$$

The preceding two examples conclude that if  $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ , then  $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$  and  $\Sigma = \mathbb{V}[\mathbf{X}]$ ; in particular,  $\Sigma$  is diagonal iff  $X_k$ 's are uncorrelated.

### Exercises

8.1. Prove that  $\mathbb{P}^X$  in (8.1) is indeed the Lebesgue-Stieltjes measure of  $F^X$ .

8.2. Prove the statements in Example 8.1.

8.3. Show that a random variable with a PDF is continuous.

8.4. Prove Example 8.5(b).

8.5. Show that

$$\int_{-\infty}^{\infty} s \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{s^2}{2}} ds = 0,$$
$$\int_{-\infty}^{\infty} s^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{s^2}{2}} ds = 1.$$

8.6. Verify (8.9) and (8.10).

8.7. Verify (8.11) and (8.12).

8.8. Uniform distribution

8.9. Poisson distribution

8.10. Geometric distribution

8.11. Exponential distribution



## CHAPTER 9

# Independence

There are two notions that draw the essentially different focus between Probability and Analysis (measure theory): independence and conditioning. In this chapter, we establish some basic facts about independence. The two well-known, fundamental results regarding independence: Law of Large Numbers and Central Limit Theorem, will be studied in the three chapters that follow. Conditioning will be studied in Chapters 13 and 14.

### 1. Characterization via distributions

We have long learned that two events  $A$  and  $B$  are independent if  $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ . Independence of random variables are defined in a similar fashion: they distribute their values in an independent way.

9.1. DEFINITION. *Let  $X_1, \dots, X_d$  be random variables over  $(\Omega, \mathcal{F}, \mathbb{P})$ . We say that they are **independent** if for any  $B_k \in \mathcal{B}$ ,  $k = 1, \dots, d$ ,*

$$(9.1) \quad \mathbb{P}(\{X_1 \in B_1\} \cap \dots \cap \{X_d \in B_d\}) = \mathbb{P}(X_1 \in B_1) \times \dots \times \mathbb{P}(X_d \in B_d).$$

*Here the left hand side is often for simplicity written as*

$$\mathbb{P}(X_1 \in B_1, \dots, X_d \in B_d).$$

Using the notion of joint distributions in Chapter 8, we can rewrite the left hand side of (9.1) as

$$\mathbb{P}^{(X_1, \dots, X_d)} \left( \prod_{k=1}^d B_k \right).$$

Using the notion of product measures in Chapter 7, we can rewrite the right hand side as

$$\mathbb{P}^{X_1}(B_1) \times \dots \times \mathbb{P}^{X_d}(B_d) = \left( \prod_{k=1}^d \mathbb{P}^{X_k} \right) \left( \prod_{k=1}^d B_k \right).$$

Therefore,  $X_1, \dots, X_d$  are independent means that, for any set  $B = \prod_{k=1}^d B_k \in \mathcal{B}^d$ ,

$$(9.2) \quad \mathbb{P}^{(X_1, \dots, X_d)}(B) = \left( \prod_{k=1}^d \mathbb{P}^{X_k} \right) (B),$$

i.e., the two measures  $\mathbb{P}^{(X_1, \dots, X_d)}$  and  $\prod_{k=1}^d \mathbb{P}^{X_k}$  coincide on all such sets.

9.2. THEOREM. *The following are equivalent:*

- (a)  $X_1, \dots, X_d$  are independent;
- (b)  $F^{(X_1, \dots, X_d)}(x_1, \dots, x_d) = F^{X_1}(x_1) \times \dots \times F^{X_d}(x_d)$  for any  $(x_1, \dots, x_d) \in \mathbb{R}^d$ ;
- (c)  $F^{(X_1, \dots, X_d)}(x_1, \dots, x_d)$  is the product of  $d$  non-negative functions each of which is a function in  $x_k$  alone,  $k = 1, \dots, d$ ;
- (d)  $\mathbb{P}^{(X_1, \dots, X_d)} = \prod_{k=1}^d \mathbb{P}^{X_k}$  as measures on  $(\mathbb{R}^d, \mathcal{B}^d)$ .

PROOF. Taking  $B_k = (-\infty, x_k]$  in (9.1), we obtain

$$\begin{aligned} & F^{(X_1, \dots, X_d)}(x_1, \dots, x_d) \\ &= \mathbb{P}(X_1 \leq x_1, \dots, x_d \leq X_d) = \mathbb{P}(X_1 \in (-\infty, x_1], \dots, X_d \in (-\infty, x_d]) \\ &= \prod_{k=1}^d \mathbb{P}(X_k \in (-\infty, x_k]) = \prod_{k=1}^d \mathbb{P}(X_k \leq x_k) \\ &= \prod_{k=1}^d F^{X_k}(x_k). \end{aligned}$$

Thus (a)  $\implies$  (b). (b) can also be translated to that  $\mathbb{P}^{(X_1, \dots, X_d)}$  and  $\prod_{k=1}^d \mathbb{P}^{X_k}$  coincide for all sets of the form  $\prod_{k=1}^d (-\infty, x_k]$ , which then, by Corollary 2.10, implies that  $\mathbb{P}^{(X_1, \dots, X_d)} = \prod_{k=1}^d \mathbb{P}^{X_k}$  as measures on  $(\mathbb{R}^d, \mathcal{B}^d)$ . Hence, (b)  $\implies$  (d). (d)  $\implies$  (a) is immediate in view of (9.2).

(b)  $\implies$  (c) is clear. The proof of (c)  $\implies$  (b) is an elementary play of functions; we include it for the sake of completeness. Assume that (c) holds, say,  $F^{(X_1, \dots, X_d)}(x_1, \dots, x_d) = \prod_{k=1}^d G_k(x_k)$ , where  $G_k \geq 0$  for each  $k$ . Each  $G_k$  cannot be identically 0 (why?). Take  $x_k^0 \in \mathbb{R}$  such that  $G_k(x_k^0) > 0$  for  $k = 2, \dots, d$ . Then

$$F^{(X_1, \dots, X_d)}(x_1, x_2^0, \dots, x_d^0) = G_1(x_1) \prod_{k=2}^d G_k(x_k^0)$$

for any  $x_1 \in \mathbb{R}$ . Since  $F^{(X_1, \dots, X_d)}$  is increasing in  $x_1$  (why?), it follows that  $G_1$  is increasing in  $x_1$ . Similarly, each  $G_k$  is increasing in  $x_k$ . Thus

$$\begin{aligned} \prod_{k=1}^d G_k(\infty) &= \lim_{n \rightarrow \infty} \prod_{k=1}^d G_k(n) = \lim_{n \rightarrow \infty} F^{(X_1, \dots, X_d)}(n, \dots, n) \\ &= \lim_{n \rightarrow \infty} \mathbb{P} \left( \bigcap_{k=1}^d \{X_k \leq n\} \right) = \mathbb{P} \left( \bigcap_{k=1}^d \{X_k \in \mathbb{R}\} \right) = 1. \end{aligned}$$

Recalling how to recover marginal CDFs from the joint CDF from Subsection 3.1 of Chapter 8, we have

$$\begin{aligned} F^{X_1}(x_1) &= \lim_{n \rightarrow \infty} F^{(X_1, \dots, X_d)}(x_1, n, \dots, n) = \lim_{n \rightarrow \infty} G_1(x_1) \prod_{k=2}^d G_k(n) \\ &= G_1(x_1) \prod_{k=2}^d G_k(\infty) = \frac{G_1(x_1)}{G_1(\infty)}. \end{aligned}$$

Similarly, one obtains the case for other  $k$ 's. This proves (c)  $\implies$  (b).  $\square$

Apparently, Condition (b) is the most convenient one to verify. The following result extends (b) to PDFs whenever existing.

**9.3. COROLLARY.** *Let  $X_1, \dots, X_d$  be random variables.*

- (a) *If they are independent and have PDFs  $f^{X_k}$ 's, then they have a joint PDF which is given by  $\prod_{k=1}^d f^{X_k}$ ;*
- (b) *If they have a joint PDF<sup>1</sup> that is a product of  $d$  non-negative functions each of which a function of  $x_k$  alone,  $k = 1, \dots, d$ , then they are independent.*

**PROOF.** (a). For any  $x_1, \dots, x_d \in \mathbb{R}$ , by independence of  $X_k$ 's and Theorem 9.2(b), we have

$$\begin{aligned} F^{(X_1, \dots, X_d)}(x_1, \dots, x_d) &= \prod_{k=1}^d F^{X_k}(x_k) = \prod_{k=1}^d \int_{(-\infty, x_k]} f^{X_k}(t_k) dt_k \\ &= \prod_{k=1}^d \int_{\mathbb{R}} \mathbb{1}_{(-\infty, x_k]}(t_k) f^{X_k}(t_k) dt_k = \int_{\mathbb{R}^d} \prod_{k=1}^d \mathbb{1}_{(-\infty, x_k]}(t_k) f^{X_k}(t_k) dt \\ &= \int_{\mathbb{R}^d} \mathbb{1}_{\prod_{k=1}^d (-\infty, x_k]}(\mathbf{t}) \prod_{k=1}^d f^{X_k}(t_k) dt = \int_{\prod_{k=1}^d (-\infty, x_k]} \prod_{k=1}^d f^{X_k}(t_k) dt, \end{aligned}$$

---

<sup>1</sup>Recall from Section 3 of Chapter 8 that if a random vector has a joint PDF then all the component random variables automatically have PDFs as well.

where the fourth equality is due to Fubini Theorem. By the definition of (joint) PDF,  $\prod_{k=1}^d f^{X_k}(t_k)$  is the PDF of  $(X_1, \dots, X_d)$ . This proves (a).

For (b), say,  $f^{(X_1, \dots, X_d)}(t_1, \dots, t_d) = \prod_{k=1}^d g_k(t_k)$  for any  $(x_1, \dots, x_d) \in \mathbb{R}^d$ . Then for any  $x_1, \dots, x_d \in \mathbb{R}$ , we have

$$\begin{aligned} F^{(X_1, \dots, X_d)}(x_1, \dots, x_d) &= \int_{\prod_{k=1}^d (-\infty, x_k]} f^{(X_1, \dots, X_d)} d\mathbf{t} = \int_{\prod_{k=1}^d (-\infty, x_k]} \prod_{k=1}^d g_k(t_k) d\mathbf{t} \\ &= \prod_{k=1}^d \int_{(-\infty, x_k]} g_k(t_k) dt_k. \end{aligned}$$

For  $k = 1, \dots, d$ , write  $G_k(x_k) = \int_{(-\infty, x_k]} g(t_k) dt_k$  for  $x_k \in \mathbb{R}$ . Then by Theorem 9.2(c),  $X_k$ 's are independent.  $\square$

9.4. REMARK. The parallel result holds for PMFs and discrete random variables.

9.1. EXAMPLE. Suppose that  $(X_1, \dots, X_d) \sim N_d(\boldsymbol{\mu}, \Sigma)$ . If  $\Sigma$  is diagonal, then their joint density is

$$\begin{aligned} &(2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^t} \\ &= (2\pi)^{-\frac{d}{2}} \left( \prod_{k=1}^d \sigma_{kk} \right)^{-\frac{1}{2}} e^{-\sum_{k=1}^d \frac{(x_k - \mu_k)^2}{2\sigma_{kk}}} = \prod_{k=1}^d \frac{1}{\sqrt{2\pi}\sigma_{kk}} e^{-\frac{(x_k - \mu_k)^2}{2\sigma_{kk}}}. \end{aligned}$$

Thus by Corollary 9.3(b),  $X_k$ 's are independent. Moreover, recall from Example 8.8 that  $X_k \sim N(\mu_k, \sigma_{kk})$  for each  $k$ . That is, if  $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$  with  $\Sigma$  diagonal, then  $X_k$ 's are independent Gaussians.

The converse is also easily seen to be true by an application of Corollary 9.3(a).

Below is another application of Corollary 9.3(a).

9.2. EXAMPLE. Let  $X_k \sim N(\mu_k, \sigma_k^2)$ ,  $k = 1, \dots, d$ , be independent. Put  $X := \sum_{k=1}^d X_k$ . Let's try to determine the distribution of  $X$ . For convenience, write  $f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$  for  $t \in \mathbb{R}$ . By Corollary 9.3(a), the PDF of  $\mathbf{X}$



is  $\prod_{k=1}^d \frac{1}{\sigma_k} f\left(\frac{x_k - \mu_k}{\sigma_k}\right)$ . Then by Corollary 8.5, for any  $x \in \mathbb{R}$ ,

$$\begin{aligned}
\mathbb{P}(X \leq x) &= \mathbb{E}[\mathbb{1}_{\{t: \sum_1^d t_i \leq x\}}(\mathbf{X})] = \int_{\mathbb{R}^d} \mathbb{1}_{\{t: \sum_1^d t_i \leq x\}}(\mathbf{x}) \prod_{k=1}^d \frac{1}{\sigma_k} f\left(\frac{x_k - \mu_k}{\sigma_k}\right) d\mathbf{x} \\
&= \int_{\mathbb{R}^{d-2}} \prod_{k=1}^{d-2} \frac{1}{\sigma_k} f\left(\frac{x_k - \mu_k}{\sigma_k}\right) d(x_1, \dots, x_{d-2}) \\
&\quad \cdot \int_{\{(t_{d-1}, t_d): t_{d-1} + t_d \leq x - \sum_1^{d-2} x_k\}} \frac{1}{\sigma_{d-1}} f\left(\frac{x_{d-1} - \mu_{d-1}}{\sigma_{d-1}}\right) \frac{1}{\sigma_d} f\left(\frac{x_d - \mu_d}{\sigma_d}\right) d(x_{d-1}, x_d) \\
&= \int_{\mathbb{R}^{d-2}} \prod_{k=1}^{d-2} \frac{1}{\sigma_k} f\left(\frac{x_k - \mu_k}{\sigma_k}\right) d(x_1, \dots, x_{d-2}) \\
&\quad \cdot \int_{\mathbb{R}} \mathbb{1}_{\{t: t \leq x - \sum_1^{d-2} x_k\}}(z) \frac{1}{\sqrt{\sigma_{d-1}^2 + \sigma_d^2}} f\left(\frac{z - (\mu_{d-1} + \mu_d)}{\sqrt{\sigma_{d-1}^2 + \sigma_d^2}}\right) dz \\
&= \int_{\mathbb{R}^{d-1}} \mathbb{1}_{\{(t_1, \dots, t_{d-1}): \sum_1^{d-1} t_i \leq x\}}(x_1, \dots, x_{d-2}, z) \prod_{k=1}^{d-2} \frac{1}{\sigma_k} f\left(\frac{x_k - \mu_k}{\sigma_k}\right) \\
&\quad \cdot \frac{1}{\sqrt{\sigma_{d-1}^2 + \sigma_d^2}} f\left(\frac{z - (\mu_{d-1} + \mu_d)}{\sqrt{\sigma_{d-1}^2 + \sigma_d^2}}\right) d(x_1, \dots, x_{d-2}, z).
\end{aligned}$$

Comparing the right hand sides of the second and last equalities, one sees that the integrands have the same pattern but the latter one has one less variable. Thus repeating the process, we obtain that

$$\mathbb{P}(X \leq x) = \int_{\mathbb{R}} \mathbb{1}_{\{t: t \leq x\}}(z) \frac{1}{\sqrt{\sum_1^d \sigma_k^2}} f\left(\frac{z - \sum_1^d \mu_k}{\sqrt{\sum_1^d \sigma_k^2}}\right) dz.$$

It follows that  $X \sim N(\sum_1^d \mu_k, \sum_1^d \sigma_k^2)$ . In particular, the sum is still a normal distribution!

We include a useful observation.

**9.5. PROPOSITION.** *Let  $X_1, \dots, X_d$  be independent random variables and  $h_k: \mathbb{R} \rightarrow \mathbb{R}$ ,  $k = 1, \dots, d$ , be Borel measurable functions. Then  $h_1(X_1), \dots, h_d(X_d)$  are also independent.*

**PROOF.** Note that  $\{h_k(X_k) \in B_k\} = \{X_k \in h_k^{-1}(B_k)\}$ . Thus

$$\begin{aligned}
\mathbb{P}(h_1(X_1) \in B_1, \dots, h_d(X_d) \in B_d) &= \mathbb{P}(X_1 \in h_1^{-1}(B_1), \dots, X_d \in h_d^{-1}(B_d)) \\
&= \prod_{k=1}^d \mathbb{P}(X_k \in h_k^{-1}(B_k)) = \prod_{k=1}^d \mathbb{P}(h_k(X_k) \in B_k),
\end{aligned}$$

which gives independence of  $h_k(X_k)$ 's.  $\square$

## 2. Characterization via expectations

So far we have not used Theorem 9.2(d). Below is a very important application.

**9.6. THEOREM.** *Let  $X_1, \dots, X_d$  be random variables. Then  $X_k$ 's are independent if and only if  $\mathbb{E}[\prod_{k=1}^d h_k(X_k)] = \prod_{k=1}^d \mathbb{E}[h_k(X_k)]$  for any Borel measurable functions  $h_k : \mathbb{R} \rightarrow [0, \infty]$ ,  $k = 1, \dots, d$ .*

**PROOF.** Taking  $h_k = \mathbb{1}_{B_k}$ , we immediately obtain the “if” part. Now suppose that  $X_1, \dots, X_d$  are independent. Then by Corollary 8.5,

$$\begin{aligned} \mathbb{E} \left[ \prod_{k=1}^d h_k(X_k) \right] &= \int_{\mathbb{R}^d} \prod_{k=1}^d h_k(x_k) d\mathbb{P}^{(X_1, \dots, X_d)}(x_1, \dots, x_d) \\ &= \int_{\mathbb{R}^d} \prod_{k=1}^d h_k(x_k) d \prod_{k=1}^d \mathbb{P}^{X_k}(x_1, \dots, x_d) \\ &= \prod_{k=1}^d \int_{\mathbb{R}} h_k(x_k) d\mathbb{P}^{X_k}. \end{aligned}$$

where the second equality is due to Theorem 9.2(d) and the last equality is due to a repeated application of Fubini Theorem. Of course, to apply Corollary 8.5, one needs to show that  $\prod_{k=1}^d h_k$  as a function on  $\mathbb{R}^d$  is Borel measurable (Exercise 9.4).  $\square$

**9.7. COROLLARY.** *Let  $X_1, \dots, X_d$  be independent random variables and  $h_k : \mathbb{R} \rightarrow [0, \infty]$ ,  $k = 1, \dots, d$ , be Borel measurable functions. If  $h_k(X_k) \in L^1$ , then  $\prod_{k=1}^d h_k(X_k) \in L^1$  and  $\mathbb{E}[\prod_{k=1}^d h_k(X_k)] = \prod_{k=1}^d \mathbb{E}[h_k(X_k)]$ . If  $\prod_{k=1}^d h_k(X_k) \in L^1$  and each  $h_k(X_k)$  is not a.s., then  $h_k(X_k) \in L^1$  for each  $k = 1, \dots, d$ .*

**PROOF.** For integrability, apply Theorem 9.6 to  $|h_k|$ ; for the second assertion, note that  $\mathbb{E}[|h_k(X_k)|] > 0$  for each  $k$ . For the asserted equality, apply  $\mathbb{E}[\prod_{k=1}^d h_k^\pm(X_k)] = \prod_{k=1}^d \mathbb{E}[h_k^\pm(X_k)]$  and reassemble the terms according to  $h_k = h_k^+ - h_k^-$  and linearity of expectations.  $\square$

**9.8. COROLLARY.** *Let  $X$  and  $Y$  be independent integrable random variables. Then  $\text{CoV}[X, Y] = 0$ .*

PROOF. By Corollary 9.7,  $(X - \mu_X)(Y - \mu_Y) = XY - \mu_X Y - X\mu_Y + \mu_X \mu_Y \in L^1$ , so that  $\text{CoV}[X, Y]$  is well-defined. Moreover,

$$\begin{aligned}\text{CoV}[X, Y] &= \mathbb{E}[XY - \mu_X Y - X\mu_Y + \mu_X \mu_Y] \\ &= \mathbb{E}[XY] - \mu_X \mathbb{E}[Y] - \mathbb{E}[X]\mu_Y + \mu_X \mu_Y \\ &= \mathbb{E}[X]\mathbb{E}[Y] - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y = 0\end{aligned}$$

□

9.9. COROLLARY. Let  $X_1, \dots, X_d$  be independent random variables in  $L^2$ . Let  $a_1, \dots, a_k$  be real numbers. Then  $\mathbb{V}[\sum_{k=1}^d a_k X_k] = \sum_{k=1}^d a_k^2 \mathbb{V}[X_k]$ .

PROOF. Write  $\mu_k = \mathbb{E}[X_k]$ . We have

$$\begin{aligned}\mathbb{V}\left[\sum_{k=1}^d a_k X_k\right] &= \mathbb{E}\left[\left(\sum_{k=1}^d a_k X_k - \sum_{k=1}^d a_k \mu_k\right)^2\right] = \mathbb{E}\left[\left(\sum_{k=1}^d a_k (X_k - \mu_k)\right)^2\right] \\ &= \mathbb{E}\left[\sum_{k=1}^d a_k^2 (X_k - \mu_k)^2 + \sum_{j \neq k} a_j a_k (X_j - \mu_j)(X_k - \mu_k)\right] \\ &= \sum_{k=1}^d a_k^2 \mathbb{V}[X_k] + \sum_{j \neq k} a_j a_k \text{Cov}[X_j, X_k] = \sum_{k=1}^d a_k^2 \mathbb{V}[X_k]\end{aligned}$$

□

### 3. Independence of random vectors

The notion of independence can be extended from random variables to random vectors.

9.10. DEFINITION. Let  $\mathbf{X}_k = (X_{k1}, X_{k2}, \dots, X_{kd_k})$ ,  $k = 1, \dots, m$ , be random vectors over  $(\Omega, \mathcal{F}, \mathbb{P})$ . We say that these  $m$  random vectors are **independent** if for any  $B_k \in \mathcal{B}^{d_k}$ ,  $k = 1, \dots, m$ ,

$$\mathbb{P}(\mathbf{X}_1 \in B_1, \dots, \mathbf{X}_m \in B_m) = \prod_{k=1}^m \mathbb{P}(\mathbf{X}_k \in B_k).$$

The following results can be proved similarly.

9.11. PROPOSITION. Let  $\mathbf{X}_k$  be a random vector of dimension  $d_k$ ,  $k = 1, \dots, m$ . Let  $h_k : \mathbb{R}^{d_k} \rightarrow \mathbb{R}^{d'_k}$ ,  $k = 1, \dots, m$ , be any Borel measurable functions. Then  $h_1(\mathbf{X}_1), \dots, h_m(\mathbf{X}_m)$  are still independent random vectors.

9.12. THEOREM. Let  $\mathbf{X}_k$  be a random vector of dimension  $d_k$ ,  $k = 1, \dots, m$ . The following are equivalent:

- (a)  $\mathbf{X}_1, \dots, \mathbf{X}_m$  are independent;
- (b)  $F^{(\mathbf{X}_1, \dots, \mathbf{X}_m)}(\mathbf{x}_1, \dots, \mathbf{x}_m) = \prod_{k=1}^m F^{\mathbf{X}_k}(\mathbf{x}_k)$  for any  $\mathbf{x}_k \in \mathbb{R}^{d_k}$ ,  $k = 1, \dots, m$ ;
- (c)  $F^{(\mathbf{X}_1, \dots, \mathbf{X}_m)}(\mathbf{x}_1, \dots, \mathbf{x}_m)$  is the product of  $m$  non-negative functions each of which is a function in  $\mathbf{x}_k \in \mathbb{R}^{d_k}$  alone,  $k = 1, \dots, m$ ;
- (d)  $\mathbb{P}^{(\mathbf{X}_1, \dots, \mathbf{X}_m)} = \prod_{k=1}^m \mathbb{P}^{\mathbf{X}_k}$  as measures on  $(\mathbb{R}^{\sum_{k=1}^m d_k}, \mathcal{B}^{\sum_{k=1}^m d_k})$ .

9.13. COROLLARY. Let  $\mathbf{X}_k$  be a random vector of dimension  $d_k$ ,  $k = 1, \dots, m$ .

- (a) If they are independent and have PDFs  $f^{\mathbf{X}_k}$ 's, then they have a joint PDF which is given by  $\prod_{k=1}^m f^{\mathbf{X}_k}$ ;
- (b) If they have a joint PDF that is a product of  $m$  non-negative functions each of which is a function of  $\mathbf{x}_k$  alone,  $k = 1, \dots, m$ , then they are independent.

9.14. THEOREM. Let  $\mathbf{X}_k$  be a random vector of dimension  $d_k$ ,  $k = 1, \dots, m$ . Then  $\mathbf{X}_k$ 's are independent if and only if  $\mathbb{E}[\prod_{k=1}^m h_k(\mathbf{X}_k)] = \prod_{k=1}^m \mathbb{E}[h_k(\mathbf{X}_k)]$  for any Borel measurable functions  $h_k : \mathbb{R}^{d_k} \rightarrow [0, \infty]$ ,  $k = 1, \dots, m$ .

Finally, we mention the following remark for clarification purposes.

- 9.15. REMARK. (a) Let  $\mathbf{X}_k$ ,  $k = 1, \dots, m$ , be independent random vectors of dimension  $d_k$ , respectively. Then the random variables in  $\mathbf{X}_k$ , i.e.,  $X_{k1}, X_{k2}, \dots, X_{kd_k}$ , of course may not be independent. But random variables such as  $X_{11}, X_{21}, \dots, X_{m1}$  are independent. Indeed, in Proposition 9.11, taking  $h_k(x_{k1}, \dots, x_{kd_k}) = x_{k1}$ , we obtain this assertion.
- (b) On the other hand, let  $X_{kj}$ ,  $k = 1, \dots, m$ ,  $j = 1, \dots, d_k$ , be given random variables. If  $X_{ij}$ 's are independent, then the random vectors  $\mathbf{X}_k$ 's are independent too, where  $\mathbf{X}_k := (X_{k1}, X_{k2}, \dots, X_{kd_k})$ ,  $k = 1, \dots, m$ . Indeed, by Theorem 9.2 (b), the CDF of  $(\mathbf{X}_1, \dots, \mathbf{X}_m)$  is the product of all  $F^{\mathbf{X}_{kj}}$ ,  $k = 1, \dots, m$ ,  $j = 1, \dots, d_k$ . Since  $X_{kj}$ ,  $j = 1, \dots, d_k$ , are also independent (Exercise 9.1), by Theorem 9.2 (b) again, the CDF of  $\mathbf{X}_k$  is the product of  $F^{\mathbf{X}_{kj}}$ ,  $j = 1, \dots, d_k$ . Thus the CDF of  $(\mathbf{X}_1, \dots, \mathbf{X}_m)$  is the product of the CDFs  $F^{\mathbf{X}_k}$ ,  $k = 1, \dots, m$ . It follows from Theorem 9.12(b) that  $\mathbf{X}_1, \dots, \mathbf{X}_m$  are independent.

**Exercises**

9.1. If  $X_1, \dots, X_d$  are independent then any few of them are also independent.

9.2. Show that

$$\begin{aligned} & \int_{\{(t_1, t_2): t_1 + t_2 \leq x\}} \frac{1}{2\pi\sigma_1\sigma_2} e^{-\left[\frac{(x_1 - \mu_1)^2}{2\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{2\sigma_2^2}\right]} d(x_1, x_2) \\ &= \int_{-\infty}^x \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} e^{-\frac{(z - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}} dz. \end{aligned}$$

Hint: apply the change of variable:  $\begin{cases} x_1 + x_2 = z \\ x_2 = \frac{\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} w + \mu_2 \end{cases}$ , repackage the terms in the exponent, and use  $\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(w-a)^2}{2\sigma_1^2}} dw = 1$  for any  $a \in \mathbb{R}$ .

9.3. Let  $Z_1, \dots, Z_d$  be standard normal distributions. Let  $a_1, \dots, a_k$  be real numbers. Determine the distribution of  $\sum_{k=1}^d a_k Z_k$ .

9.4. Let  $h_k : \mathbb{R} \rightarrow \mathbb{R}$ ,  $k = 1, \dots, d$ , be Borel measurable functions. Show that  $\prod_{k=1}^d h_k$  is Borel measurable as a function on  $\mathbb{R}^d$ .

9.5. Prove Proposition 9.11, Theorem 9.12, Corollary 9.13 and Theorem 9.14.

9.6. Find three random variables  $X, Y, Z$  such that they are not independent but any two of them are independent.



## CHAPTER 10

# Law of Large Numbers

This note briefly reviews laws of large numbers, which in a narrow sense asserts that sample means approximate the population mean as the sample size gets larger. Several applications of them to statistics will also be discussed: Monte Carlo methods, Empirical distributions, the Bootstrapping, and the Moment estimators.

### 1. Type of convergence

Laws of Large Numbers involve convergence of sequences of random variables. So far, we have encountered the most important one: almost sure convergence. Another crucial one is as follows.

10.1. DEFINITION. *Let  $X, X_n, n \in \mathbb{N}$  random variables. We say that  $(X_n)_{n \in \mathbb{N}}$  converges to  $X$  **in probability**, and write  $X_n \xrightarrow{pr} X$ , if for any  $\varepsilon > 0$ ,  $\mathbb{P}(|X_n - X| > \varepsilon) \rightarrow 0$  as  $n \rightarrow \infty$ .*

Basically, convergence in probability means that, for any given error bound  $\varepsilon > 0$ , as  $X_n$  approaches  $X$ , the probability of that the error  $|X_n - X|$  exceed  $\varepsilon$  approaches 0.

10.2. PROPOSITION. *If  $X_n \rightarrow X$  in norm or a.s., then  $X_n \rightarrow X$  in probability;*

### 2. Law of Large Numbers

**2.1. Weak law of large numbers.** The term of weak law refers to convergence in probability in the context of laws of large numbers. We decompose the proof of weak law into several short lemmas. The first one provides a typical way to yield convergence in probability.

10.1. LEMMA. *For a sequence  $(X_n)$  of rvs in  $L^2$ , if  $V[X_n] \rightarrow 0$ , then*

$$X_n - \mathbb{E}[X_n] \xrightarrow{pr} 0.$$

PROOF. For any  $\varepsilon > 0$ , by Chebyshev's inequality,

$$\mathbb{P}\left(\left|X_n - \mathbb{E}[X_n]\right| > \varepsilon\right) \leq \frac{\mathbb{E}\left[(X_n - \mathbb{E}[X_n])^2\right]}{\varepsilon^2} = \frac{V[X_n]}{\varepsilon^2} \rightarrow 0.$$

□

The second one encourages us to do truncations. Two sequences of rvs,  $(X_n)$  and  $(Y_n)$ , are said to be **equivalent** if  $\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq Y_n) < \infty$ .

- 10.2. LEMMA. (a) *If  $(X_n)$  and  $(Y_n)$  are equivalent, then  $\frac{1}{n} \sum_{k=1}^n X_k$  converges a.s. (resp., in probability) if and only if  $\frac{1}{n} \sum_{k=1}^n Y_k$  converges a.s. (resp., in probability). The limits also coincide in the case of convergence.*
- (b) *Let  $(X_n)$  a sequence of identically distributed, integrable random variables. Let  $Y_n = X_n \mathbb{1}_{\{|X_n| \leq n\}}$  for each  $n \in \mathbb{N}$ . Then  $(X_n)$  and  $(Y_n)$  are equivalent.*

PROOF. ((a)). Assume that  $(X_n)$  and  $(Y_n)$  are equivalent. By Borel-Cantelli Lemma,

$$\mathbb{P}\left(\limsup_n \{X_n \neq Y_n\}\right) = 0.$$

Take any  $\omega \in (\limsup_n \{X_n \neq Y_n\})^c = \liminf_n \{X_n = Y_n\}$ . There exists  $n_0$ , depending on  $\omega$ , such that for any  $n \geq n_0$ ,  $\omega \in \{X_n = Y_n\}$ , i.e.,  $X_n(\omega) = Y_n(\omega)$ , implying that

$$\lim_n \frac{1}{n} \sum_{k=1}^n (X_k(\omega) - Y_k(\omega)) = 0.$$

These two observations together give that  $\frac{1}{n} \sum_{k=1}^n (X_k - Y_k)$  converges to 0 a.s. and thus in probability. The assertions in ((a)) now follow immediately.

((b)) holds because

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq Y_n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_1| > n) \leq \mathbb{E}[|X_1|] < \infty.$$

□

The nice properties of truncated rvs are contained in the next lemma.

10.3. LEMMA. *Let  $(X_n)_{n \in \mathbb{N}}$  be a sequence of identically distributed integrable rvs. Let  $Y_n = X_n \mathbb{1}_{\{|X_n| \leq n\}}$  for each  $n \in \mathbb{N}$ . Then*

$$\sum_{n=1}^{\infty} \frac{V[Y_n]}{n^2} < \infty.$$



Consequently,

$$\lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{k=1}^n V[Y_k] = 0.$$

PROOF. Let  $F$  be the CDF of  $X_n$ 's. Then

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{V[Y_n]}{n^2} &\leq \sum_{n=1}^{\infty} \frac{\mathbb{E}[Y_n^2]}{n^2} = \sum_{n=1}^{\infty} \frac{1}{n^2} \int_{\{|x| \leq n\}} x^2 dF(x) = \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{k=1}^n \int_{\{k-1 < |x| \leq k\}} x^2 dF(x) \\ &= \sum_{k=1}^{\infty} \sum_{n=k}^{\infty} \frac{1}{n^2} \int_{\{k-1 < |x| \leq k\}} x^2 dF(x) \leq \sum_{k=1}^{\infty} \frac{2}{k} \int_{\{k-1 < |x| \leq k\}} x^2 dF(x) \\ &\leq \sum_{k=1}^{\infty} \frac{2}{k} \int_{\{k-1 < |x| \leq k\}} k|x| dF(x) = 2 \sum_{k=1}^{\infty} \int_{\{k-1 < |x| \leq k\}} |x| dF(x) \\ &= 2\mathbb{E}[|X|] < \infty. \end{aligned}$$

The second assertion follows from Kronecker's Lemma below on convergence of numbers, whose proof can be found in a mathematical analysis textbook and we omit.  $\square$

LEMMA (Kronecker). *Let  $(x_n)_{n \in \mathbb{N}}$  be a sequence of real numbers,  $(a_n)_{n \in \mathbb{N}}$  be a sequence of positive real numbers increasing to  $\infty$ . If  $\sum_{n=1}^{\infty} \frac{x_n}{a_n}$  converges to a real number, then*

$$\frac{1}{a_n} \sum_{k=1}^n x_k \longrightarrow 0.$$

We are now ready to present the proof of the weak law of large numbers.

10.3. THEOREM (Weak LLN). *Let  $(X_n)_{n \in \mathbb{N}}$  be a sequence of pairwise independent, identically distributed, integrable rvs. Then*

$$\frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{pr} \mu,$$

where  $\mu$  is the mean of  $X_i$ 's.

PROOF. Let  $Y_n = X_n \mathbb{1}_{\{|X_n| \leq n\}}$  for each  $n \in \mathbb{N}$ . By Lemma 10.2, it suffices to prove that

$$\frac{1}{n} \sum_{k=1}^n Y_k \xrightarrow{pr} \mu.$$

Moreover,

$$\mathbb{E}[Y_n] = \mathbb{E}[X_n \mathbb{1}_{\{|X_n| \leq n\}}] = \int_{|x| \leq n} x dF(x) = \mathbb{E}[X_1 \mathbb{1}_{\{|X_1| \leq n\}}] \longrightarrow \mathbb{E}[X_1] = \mu$$

by the Dominated Convergence Theorem, where  $F$  is the CDF of  $X_n$ 's. Thus

$$\frac{1}{n} \sum_{k=1}^n \mathbb{E}[Y_k] \longrightarrow \mu.$$

Therefore, it suffices to prove that

$$T_n := \frac{\sum_{k=1}^n (Y_k - \mathbb{E}[Y_k])}{n} \xrightarrow{pr} 0.$$

Note that  $Y_n$ 's are also pairwise independent and thus are uncorrelated. Hence,

$$\mathbb{V}[T_n] = \frac{1}{n^2} \mathbb{V}\left[\sum_{k=1}^n (Y_k - \mathbb{E}[Y_k])\right] = \frac{1}{n^2} \sum_{k=1}^n \mathbb{V}[Y_k] \longrightarrow 0,$$

by Lemma 10.3. Thus by Lemma 10.1,  $T_n = T_n - \mathbb{E}[T_n] \xrightarrow{pr} 0$ .  $\square$

**2.2. Strong law of large numbers.** The term of strong law refers to a.s. convergence. Again, we split the proof into several lemmas.

10.4. LEMMA. *Let  $(X_n)$  be a sequence of independent mean-zero rvs in  $L^2$ . Put  $S_n = \sum_{k=1}^n X_k$  for each  $n \in \mathbb{N}$ . Then for any  $\varepsilon > 0$  and  $n \in \mathbb{N}$ ,*

$$\mathbb{P}\left(\max_{1 \leq k \leq n} |S_k| > \varepsilon\right) \leq \frac{\mathbb{V}[S_n]}{\varepsilon^2}.$$

PROOF. For each  $n \in \mathbb{N}$ , set  $\mathcal{F}_n = \sigma(X_k : 1 \leq k \leq n)$ . Then

$$\mathbb{E}[S_{n+1} | \mathcal{F}_n] = \mathbb{E}[S_n + X_{n+1} | \mathcal{F}_n] = S_n + \mathbb{E}[X_{n+1} | \mathcal{F}_n] = S_n,$$

where we use the fact that since  $X_{n+1}$  is independent from  $\mathcal{F}_n$ ,  $\mathbb{E}[X_{n+1} | \mathcal{F}_n] = \mathbb{E}[X_{n+1}] = 0$ . It follows that  $\{(S_n); (\mathcal{F}_n)\}$  is a martingale, and thus  $\{(S_n^2); (\mathcal{F}_n)\}$  is a positive submartingale. Thus by Doob's maximal inequality,

$$\mathbb{P}\left(\max_{1 \leq k \leq n} |S_k| > \varepsilon\right) = \mathbb{P}\left(\max_{1 \leq k \leq n} S_k^2 > \varepsilon^2\right) \leq \frac{\mathbb{E}[S_n^2]}{\varepsilon^2} = \frac{\mathbb{V}[S_n]}{\varepsilon^2}.$$

$\square$

10.5. LEMMA. *Let  $(X_n)$  be a sequence of independent mean-zero rvs in  $L^2$ . Suppose that  $\sum_{n=1}^{\infty} \mathbb{V}[X_n] < \infty$ . Then  $\sum_{n=1}^{\infty} X_n$  converges a.s.*

PROOF. For any  $m \in \mathbb{N}$ , take  $n_m \in \mathbb{N}$  such that

$$\sum_{k=n_m+1}^{\infty} \mathbb{V}[X_k] < \frac{1}{m^4}.$$

For any  $n' > n_m$ , by Lemma 10.4, we have

$$\mathbb{P}\left(\max_{n_m+1 \leq k \leq n'} |S_k - S_{n_m}| > \frac{1}{m}\right) \leq \frac{\mathbb{V}[S_{n'} - S_{n_m}]}{\frac{1}{m^2}} = m^2 \sum_{n_m+1 \leq k \leq n'} \mathbb{V}[X_k] < \frac{1}{m^2}.$$

Putting

$$A_m = \left\{ \max_{k \geq n_m+1} |S_k - S_{n_m}| > \frac{1}{m} \right\}$$

and letting  $n' \rightarrow \infty$  above, we have

$$\mathbb{P}(A_m) \leq \frac{1}{2^m},$$

so that  $\sum_{m=1}^{\infty} \mathbb{P}(A_m) < \infty$  and thus by Borel-Catelli Lemma,

$$\mathbb{P}(\limsup_m A_m) = 0.$$

Now take any  $\omega \notin \limsup_m A_m$ , there exists some  $m \in \mathbb{N}$  such that  $\omega \notin A_m$ , which is equivalent to that  $|S_k(\omega) - S_{n_m}(\omega)| \leq \frac{1}{m}$  for any  $k \geq n_m$ . Therefore,

$$|S_k(\omega) - S_l(\omega)| \leq \frac{2}{m}, \quad \text{for any } k, l \geq n_m.$$

This proves that the partial sums of the series  $\sum_{n=1}^{\infty} S_n(\omega)$  are Cauchy, and hence the series is convergent. This completes the proof.  $\square$

**10.4. THEOREM (Strong LLN).** *Let  $(X_n)_{n \in \mathbb{N}}$  be a sequence of independent, identically distributed, integrable rvs. Then*

$$\frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{a.s.} \mu,$$

where  $\mu$  is the mean of  $X_i$ 's.

**PROOF.** Let  $Y_n = X_n \mathbb{1}_{\{|X_n| \leq n\}}$  for each  $n \in \mathbb{N}$ . As in the weak case, it suffices to prove that

$$T_n := \frac{\sum_{k=1}^n (Y_k - \mathbb{E}[Y_k])}{n} \xrightarrow{a.s.} 0.$$

By Lemmas 10.3 and 10.5,  $\sum_{n=1}^{\infty} \frac{Y_n - \mathbb{E}[Y_n]}{n}$  converges a.s. By Kronecker's Lemma,  $T_n \xrightarrow{a.s.} 0$ .  $\square$

### 3. Monte Carlo Simulations

The SLLN provides a numerical method for computing the expectation  $\mathbb{E}[X]$  of a rv via simulations. Recall that a distribution function  $F : \mathbb{R} \rightarrow \mathbb{R}$  is an increasing, right-continuous function such that  $F(-\infty) = 0$  and  $F(\infty) = 1$ . Recall also that there is a bijection between distribution functions and probability measures on  $\mathbb{R}$  via:

$$\mu((a, b]) = F(b) - F(a).$$

We will call a distribution function  $F$  of interest a **(univariate) population**. A **random sample** drawn from the population  $F$  is a sequence  $(X_n)$  of independent rvs all having  $F$  as their CDF. A **sample** drawn from the population  $F$  is a sequence  $(x_n)$  of real numbers, which is a **realization** of a random sample  $(X_n)$ , namely, there exists  $\omega$  such that

$$x_n = X_n(\omega) \quad \text{for each } n.$$

Suppose that the **population mean**  $m := \int_{\mathbb{R}} x \, dF(x)$  is finite. Let  $(X_n)$  be a random sample drawn from  $F$ . Then  $\mathbb{E}[|X_n|] = \int_{\mathbb{R}} |x| \, dF(x) < \infty$ , so that the SLLN is applicable to the sequence  $(X_n)$ . Thus, for a sample  $(x_n)$  drawn from  $F$ , the **sample means** converge to the population mean<sup>1</sup>:

$$\frac{1}{n} \sum_{k=1}^n x_k \longrightarrow \mathbb{E}[X_1] = \int_{\mathbb{R}} x \, dF(x) = m, \quad \text{as } n \rightarrow \infty.$$

In reality, the sample drawn from the population is of course a finite sequence, say,  $x_1, x_2, \dots, x_n$ , where  $n$  is called the **sample size**. When the size  $n$  is large enough, we have the following approximation:

$$\frac{1}{n} \sum_{k=1}^n x_k \approx m.$$

This algorithm of computing a population parameter using random sampling is typically referred to as **Monte Carlo methods**. For example, once we have a way to generate from  $F$  a sample, also called **random numbers** in the context of Monte Carlo methods, we can evaluate the population mean  $m$  by  $\frac{1}{n} \sum_{k=1}^n x_k$  as above.

Most computational software contain random number generators for classical distributions, such as uniform distribution and normal distributions. For example,  $x = \text{rand}(n, 1)$  returns  $n$  random numbers from the uniform distribution on  $(0, 1)$ :

```
>> x=rand(10,1)
```

```
x =
```

```
0.1622
```

```
0.7943
```

---

<sup>1</sup>Not an accurate assertion, since for any random sample, the convergence may fail on a set of probability 0. But for all practical purposes, probability-zero events are regarded as never happening.

```

0.3112
0.5285
0.1656
0.6020
0.2630
0.6541
0.6892
0.7482

```

When  $n$  is large, we can see that the sample mean  $\frac{1}{n} \sum_{k=1}^n x_k$  is indeed close to the population mean  $\mu = \int_{(0,1)} x dx = \frac{1}{2}$ . We simulate 5 samples of size one million and calculate the respective sample means; all of these five sample means are close to  $\mu = 0.5$ :

```

>> y=zeros(5,1);
for k=1:5
    x=rand(1000000,1);
    y(k)=mean(x);
end
y

y =

```

```

0.5002
0.4999
0.5000
0.5001
0.4998

```

Of course, we can use the Monte Carlo methods to compute population parameters other than the mean. Suppose that the population  $F$  has a finite second moment, i.e.,  $\int_{\mathbb{R}} x^2 dF(x) < \infty$ . Let  $(X_n)$  be a random sample drawn from  $F$ . Then  $\mathbb{E}[X_n^2] = \int_{\mathbb{R}} x^2 dF(x) < \infty$ , which further implies that  $\mathbb{E}[|X_n|] < \infty$  by the Cauchy-Schwartz inequality. Thus the SLLN applies to both  $(X_n)$  and  $(X_n^2)$ , namely,

$$\frac{1}{n} \sum_{k=1}^n X_n \xrightarrow{a.s.} \mathbb{E}[X_1] = \int_{\mathbb{R}} x dF(x), \quad \frac{1}{n} \sum_{k=1}^n X_n^2 \xrightarrow{a.s.} \mathbb{E}[X_1^2] = \int_{\mathbb{R}} x^2 dF(x).$$

It follows that for a sample  $(x_n)$ ,

$$\frac{1}{n} \sum_{k=1}^n x_k^2 - \left( \frac{1}{n} \sum_{k=1}^n x_k \right)^2 \longrightarrow \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = V[X_1],$$

where the far right term is clearly equal to

$$V[F] := \int_{\mathbb{R}} x^2 dF(x) - \left( \int_{\mathbb{R}} x dF(x) \right)^2,$$

called the **population variance**. Thus  $V[F]$  is evaluated by

$$\frac{1}{n} \sum_{k=1}^n x_k^2 - \left( \frac{1}{n} \sum_{k=1}^n x_k \right)^2 = \frac{\sum_{k=1}^n (x_k - \bar{x})^2}{n},$$

for some large enough  $n$ ; here  $\bar{x} := \frac{1}{n} \sum_{k=1}^n x_k$  is the sample mean.

We can use tricks to generate random numbers to compute more sophisticated probabilistic terms. Say, let's compute the expectation of

$$X = \frac{2^U}{e^{\sqrt{|Z|}}},$$

where  $U$  and  $Z$  are independent,  $U$  is uniform on  $(0, 1)$ , and  $Z$  is standard normal. We simulate a sample of size one million for the uniform distribution and the standard normal, respectively, aggregate them to produce random numbers for  $F$ , where  $F$  is the CDF of  $X$ , and then take the new sample mean:

```
>> u=rand(1000000,1);
z=randn(1000000,1);
x=2.^u./exp(sqrt(abs(z)));
mean(x)
```

```
ans =
```

```
0.6736
```

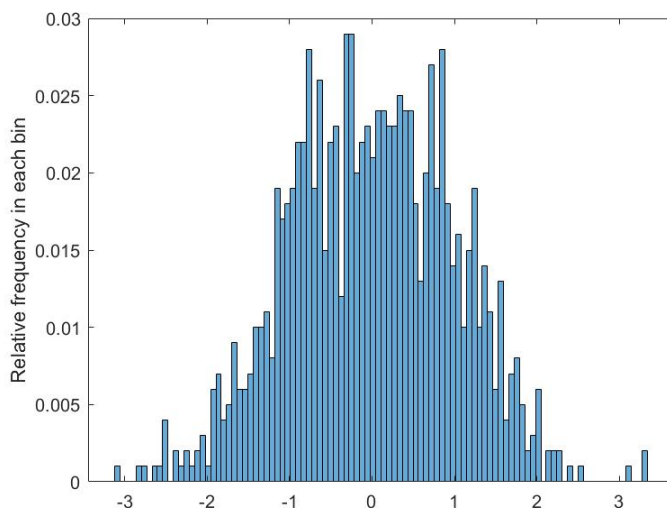
One can run these codes a few times and will see that the answer is stable around 0.673.

#### 4. Empirical Distributions

In Statistics, one often draws a histogram of data, which shows “distribution” of the data, to infer the distribution of the population. For example,

the following codes in Matlab simulate 1000 random numbers from the standard normal distribution and produces the histogram of these numbers with 50 bins, Figure 1.

```
>> x=randn(1000,1);
>> histogram(x,50)
ylabel('Relative frequency in each bin')
```



One sees that the histogram does demonstrate a shape like the graph of the density of the standard normal distribution. We now study why this happens.

In general, suppose that we collect  $n$  **observations**, i.e., a sample of size  $n$ , from the population, which we denote by  $x_1, x_2, \dots, x_n$ . In the histogram, one first cuts the  $x$ -axis into several bins. Then the histogram captures the relative frequencies of observations that belong to each bin  $(a_j, b_j]$ :

$$\frac{\#\{k : a_j < x_k \leq b_j\}}{n}.$$

We consider the following function  $F_n : \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$F_n(x) = \frac{\#\{k : x_k \leq x\}}{n}, \quad x \in \mathbb{R}.$$

Clearly,  $F_n(x)$  is the relative frequency of the observations  $x_1, \dots, x_n$  that belong to the interval  $(-\infty, x]$ . In this notation, the relative frequency in a bin  $(a_j, b_j]$  can be expressed by

$$F_n(b_j) - F_n(a_j).$$

Thus, the histogram is produced by the values of  $F_n$  at the end points of the bins.

One can easily see that  $F_n$  is an increasing, right continuous function such that  $F_n(-\infty) = 0$  and  $F_n(\infty) = 1$ . That is,  $F_n$  is also a distribution function. It is called the **empirical distribution**, because it is the distribution of the empirical evidence  $x_1, \dots, x_n$ . The assertion that histograms can be used to approximate the population distribution is mathematically equivalent to that whenever  $n$  is large enough,  $F_n(x)$  is close to  $F(x)$  at every  $x \in \mathbb{R}$ , or

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \text{ is small, whenever } n \text{ is large.}$$

10.5. THEOREM (Central Statistical Theorem). *Let  $(X_n)$  be a random sample drawn from the population  $F$ . For each  $n \in \mathbb{N}$  and  $x \in \mathbb{R}$ , put*

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{\{X_k \leq x\}}.$$

*Then*

$$\mathbb{P}\left(\limsup_n \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0\right) = 1.$$

*That is, out a set of probability 0,  $(F_n(x))$  converges to  $F(x)$ , uniformly in  $x$ .*

Clearly, at any realization  $(x_n)$  of  $(X_n)$ , the two ways of defining  $F_n(x)$  coincide.

PROOF. We only provide the proof of the following weaker version. At every  $x \in \mathbb{R}$ , outside a set of probability, we have  $F_n(x) \rightarrow F(x)$ . This is easy! Fix  $x \in \mathbb{R}$ . Since  $X_n$ 's are iid, the random variables  $\mathbb{1}_{\{X_n \leq x\}}$ 's are iid too. In fact, their common distribution is as follows:

$$\mathbb{P}(\mathbb{1}_{\{X_n \leq x\}} = 1) = \mathbb{P}(X_n \leq x) = F(x),$$

and

$$\mathbb{P}(\mathbb{1}_{\{X_n \leq x\}} = 0) = \mathbb{P}(X_n > x) = 1 - F(x).$$

Thus by the SLLN,

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{\{X_k \leq x\}} \xrightarrow{a.s.} \mathbb{E}[\mathbb{1}_{\{X_1 \leq x\}}] = \mathbb{P}(X_1 \leq x) = F(x).$$

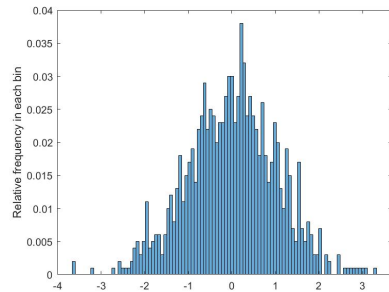
□

---

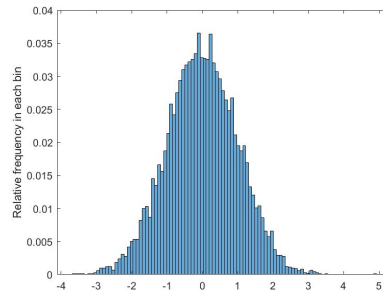
<sup>2</sup>Note that  $F_n(x)$  is in fact a rv depending on the random sample.



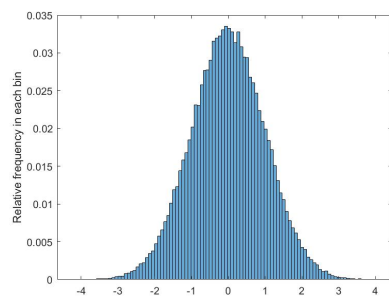
The following are histograms with 100 bins of four simulated samples from the standard normal distribution of sizes  $10^3, 10^4, 10^5, 10^6$ , respectively.



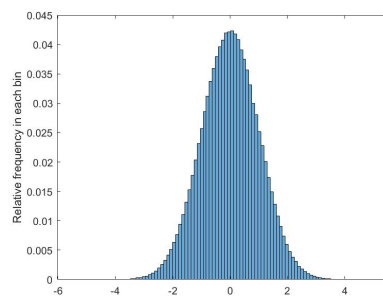
(A)  $n = 10^3$



(B)  $n = 10^4$



(C)  $n = 10^5$



(D)  $n = 10^6$

## Exercises



## CHAPTER 11

# Characteristic Functions



## CHAPTER 12

# Central Limit Theorem



## CHAPTER 13

# Conditional Distribution





## CHAPTER 14

### Conditional Expectation

**Setup.** Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a fixed probability space. Let  $\mathcal{G}$  be a given sub- $\sigma$ -algebra of  $\mathcal{F}$ . Denote by  $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  the collection of all integrable random variables over  $(\Omega, \mathcal{F}, \mathbb{P})$  (without modulo a.s. equality). Random variables are real-valued.

#### 1. Definition and Basic Properties

##### 1.1.

14.1. THEOREM. *Let  $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ . Then there exists a random variable  $Y$  such that*

- (a)  $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$
- (b)  $Y$  is  $\mathcal{G}$ -measurable,
- (c)  $\int_A Y \, d\mathbb{P} = \int_A X \, d\mathbb{P}$  for any  $A \in \mathcal{G}$ .

The reader may read the Appendix at the end for a proof.

1.2. Suppose  $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  and is  $\mathcal{G}$ -measurable. Then

$$Y \geq 0 \text{ a.s.} \iff \int_A Y \geq 0 \text{ for any } A \in \mathcal{G}.$$

Indeed, if  $Y \geq 0$  a.s., then  $Y\mathbb{1}_A \geq 0$  a.s., so that  $\int_A Y = \int \mathbb{1}_A Y \geq 0$  for any  $A \in \mathcal{G}$ .<sup>1</sup> Conversely, suppose  $\int_A Y \geq 0$  for any  $A \in \mathcal{G}$ . Since  $Y$  is  $\mathcal{G}$ -measurable,  $\{Y < 0\} \in \mathcal{G}$ , so that by assumption,  $0 \leq \int_{\{Y < 0\}} Y = \int \mathbb{1}_{\{Y < 0\}} Y = \int -Y^-$ , where the last equality follows from the identity  $\mathbb{1}_{\{Y < 0\}} Y = -Y^-$ . Therefore,  $\int Y^- \leq 0$ , and thus  $\int Y^- = 0$ . Since  $Y^- \geq 0$ , we have  $Y^- = 0$  a.s.,<sup>2</sup> implying that  $Y \geq 0$  a.s.

---

<sup>1</sup>Use the fact: over  $(\Omega, \mathcal{F}, \mathbb{P})$ , if  $X_1 \geq X_2$  a.s., then  $\int X_1 \geq \int X_2$ , as long as both integrals are well-defined.

<sup>2</sup>Use the fact: over  $(\Omega, \mathcal{F}, \mathbb{P})$ , if  $X \geq 0$  a.s., then  $X = 0$  a.s. iff  $\int X = 0$ .

**1.3.** Suppose  $Y, Z \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  and  $Y, Z$  are  $\mathcal{G}$ -measurable. Then

$$Y \geq Z \text{ a.s.} \iff \int_A Y \geq \int_A Z \text{ for any } A \in \mathcal{G}.$$

$$Y = Z \text{ a.s.} \iff \int_A Y = \int_A Z \text{ for any } A \in \mathcal{G}.$$

**1.4.** Any random variable  $Y$  satisfying the three conditions in Theorem 14.1 is called a *version* of conditional expectation of  $X$  with respect to  $\mathcal{G}$ . By 1.3, any two versions of conditional expectation of  $X$  with respect to  $\mathcal{G}$  are a.s. equal. For convenience, we now take any such a version and call it **the** conditional expectation of  $X$  with respect to  $\mathcal{G}$ , and write it as  $\mathbb{E}[X|\mathcal{G}]$ . Note that  $\mathbb{E}[X|\mathcal{G}]$  is just one version among all the versions.

We first deal with equalities regarding conditional expectations.

**1.5.** If  $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  is  $\mathcal{G}$ -measurable, then  $\mathbb{E}[X|\mathcal{G}] = X$  a.s.

For  $X$  itself satisfies the three conditions in Theorem 14.1 and is thus a version of conditional expectation of  $X$ .

**1.6.** Let  $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  and  $\mathcal{G}_1 \subset \mathcal{G}_2$  be two sub- $\sigma$ -algebras of  $\mathcal{F}$ . Then

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}_1]|\mathcal{G}_2] = \mathbb{E}[X|\mathcal{G}_1] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1] \text{ a.s.}$$

Since  $\mathbb{E}[X|\mathcal{G}_1]$  is  $\mathcal{G}_1$ -, and thus  $\mathcal{G}_2$ -, measurable, the first equality follows from 1.5. For the second one, note first that  $\mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1] \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  and is  $\mathcal{G}_1$ -measurable. Moreover, for any  $A \in \mathcal{G}_1$ ,

$$\int_A \mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1] = \int_A \mathbb{E}[X|\mathcal{G}_2] = \int_A X,$$

where the first equality follows from definition of  $\mathbb{E}[\cdot|\mathcal{G}_1]$  and the second one follows from definition of  $\mathbb{E}[\cdot|\mathcal{G}_2]$  and the fact that  $A \in \mathcal{G}_2$ . Thus  $\mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1]$  is also a version of conditional expectation of  $X$  wrt  $\mathcal{G}_1$ , so that  $\mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1] = \mathbb{E}[X|\mathcal{G}_1]$  a.s.

**1.7.** Let  $X_1, X_2 \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  and  $c, d \in \mathbb{R}$ . Then,

$$\mathbb{E}[cX_1 + dX_2|\mathcal{G}] = c\mathbb{E}[X_1|\mathcal{G}] + d\mathbb{E}[X_2|\mathcal{G}] \text{ a.s.}$$

Indeed, clearly,  $c\mathbb{E}[X_1|\mathcal{G}] + d\mathbb{E}[X_2|\mathcal{G}] \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  and is  $\mathcal{G}$ -measurable. Moreover, for any  $A \in \mathcal{G}$ ,

$$\begin{aligned} \int_A (c\mathbb{E}[X_1|\mathcal{G}] + d\mathbb{E}[X_2|\mathcal{G}]) &= c \int_A \mathbb{E}[X_1|\mathcal{G}] + d \int_A \mathbb{E}[X_2|\mathcal{G}] = c \int_A X_1 + d \int_A X_2 \\ &= \int_A (cX_1 + dX_2). \end{aligned}$$

We now deal with two simple inequalities regarding conditional expectations.

**1.8.** Let  $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  be such that  $Y$  is  $\mathcal{G}$ -measurable. By 1.3 and definition of conditional expectation,

$$\mathbb{E}[X|\mathcal{G}] \geq Y \text{ a.s.} \iff \int_A X \geq \int_A Y \text{ for any } A \in \mathcal{G}.$$

**1.9.** Let  $X_1, X_2 \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  and  $X_1 \geq X_2$  a.s. Then by 1.3 and definition of conditional expectation,

$$\mathbb{E}[X_1|\mathcal{G}] \geq \mathbb{E}[X_2|\mathcal{G}] \text{ a.s.}$$

## 2. Jensen's inequality

### 2.1. Conditional form.

**14.2. THEOREM.** Let  $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ . Let  $\Phi : \mathbb{R} \rightarrow \mathbb{R}$  be a convex function such that  $\Phi(X) \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ .<sup>3</sup> Then

$$\Phi(\mathbb{E}[X|\mathcal{G}]) \leq \mathbb{E}[\Phi(X)|\mathcal{G}] \text{ a.s.}$$

**PROOF.** We use the fact that there exist at most countably many lines  $l_n(x) = a_n x + b_n$  such that  $\Phi(x) = \sup_n l_n(x)$  for any  $x \in \mathbb{R}$ .<sup>4</sup> For any  $n \in \mathbb{N}$ , since  $\Phi(X) \geq l_n(X)$ ,

$$\mathbb{E}[\Phi(X)|\mathcal{G}] \geq \mathbb{E}[l_n(X)|\mathcal{G}] = l_n(\mathbb{E}[X|\mathcal{G}]) \text{ a.s.}$$

Let  $A_n = \left\{ \mathbb{E}[\Phi(X)|\mathcal{G}] < l_n(\mathbb{E}[X|\mathcal{G}]) \right\}$ . Then  $\mathbb{P}(\cup_n A_n) = 0$ .<sup>5</sup> For any  $\omega \in A^c$ , we have

$$\mathbb{E}[\Phi(X)|\mathcal{G}](\omega) \geq \sup_n l_n(\mathbb{E}[X|\mathcal{G}](\omega)) = \Phi(\mathbb{E}[X|\mathcal{G}](\omega)).$$

□

<sup>3</sup>Note that  $\Phi$  is continuous and thus  $\Phi(X)$  is  $\mathcal{F}$ -measurable. Similarly,  $\Phi(\mathbb{E}[X|\mathcal{G}])$  is  $\mathcal{G}$ -measurable.

<sup>4</sup>Put  $\Phi^*(y) = \sup_{x \in \mathbb{R}} (xy - \Phi(x))$  for any  $y \in \mathbb{R}$ . Recall that  $\Phi(x) = \sup_{y \in \mathbb{R}} (xy - \Phi^*(y))$  for any  $x \in \mathbb{R}$ . Note that  $\Phi^*$ , possibly taking  $\infty$ , is convex on  $\mathbb{R}$ . Thus  $I := \{y \in \mathbb{R} : \Phi^*(y) < \infty\}$  is a convex set in  $\mathbb{R}$ , and is thus an interval if not a singleton. Write  $I = (a, b) \cup E$  where  $E$  consists of possible endpoints of  $(a, b)$  that lie in  $I$ . Let  $\{y_n\}_{n \geq 1}$  be a countable set that contains  $E$  and a dense subset of  $(a, b)$ . Since  $\Phi^*$  is convex and finite on  $(a, b)$ , it is continuous there. One can now easily verify that  $\Phi(x) = \sup_{n \geq 1} (xy_n - \Phi^*(y_n))$ . The desired lines are given by  $l_n(x) = xy_n - \Phi^*(y_n)$ .

<sup>5</sup>This is why we insist on at most countably many lines.

**2.2. Unconditional form.** Putting  $\mathcal{G} = \{\emptyset, \Omega\}$  in the previous theorem, one obtains the unconditional form of Jensen's inequality:

$$\Phi(\mathbb{E}[X]) \leq \mathbb{E}[\Phi(X)].$$

**2.3.** Let  $1 \leq p < \infty$ . Let  $\Phi(t) = |t|^p$ . Then  $\Phi$  is convex<sup>6</sup>, so that  $\left|\mathbb{E}[X|\mathcal{G}]\right|^p \leq \mathbb{E}[|X|^p|\mathcal{G}]$ , implying  $|\mathbb{E}[X|\mathcal{G}]| \leq (\mathbb{E}[|X|^p|\mathcal{G}])^{\frac{1}{p}}$ . Replacing  $X$  with  $|X|$ , one has

$$(14.1) \quad \mathbb{E}[|X||\mathcal{G}] \leq \left(\mathbb{E}[|X|^p|\mathcal{G}]\right)^{\frac{1}{p}},$$

$$\mathbb{E}[|X|] \leq \left(\mathbb{E}[|X|^p]\right)^{\frac{1}{p}}.$$

Taking the expectation of the  $p$ -th power of both sides of (14.1), we have

$$\left\|\mathbb{E}[|X||\mathcal{G}]\right\|_p \leq \left(\mathbb{E}\left[\mathbb{E}[|X|^p|\mathcal{G}]\right]\right)^{1/p} = (\mathbb{E}[|X|^p])^{\frac{1}{p}} = \|X\|_p.$$

### 3. Convergence theorems

**3.1. Conditional MCT.** Suppose  $0 \leq X_n \uparrow X$  a.s. and  $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ . Then

$$\mathbb{E}[X_n|\mathcal{G}] \uparrow \mathbb{E}[X|\mathcal{G}] \text{ a.s.}$$

Indeed, let  $Y = \sup_n \mathbb{E}[X_n|\mathcal{G}]$ . Then  $Y$  is  $\mathcal{G}$ -measurable. Put  $A = \cup_n \{\mathbb{E}[X_{n+1}|\mathcal{G}] < \mathbb{E}[X_n|\mathcal{G}]\}$ . Then  $\mathbb{P}(A) = 0$ , and  $\mathbb{E}[X_n|\mathcal{G}] \uparrow Y$  on  $A^c$ , so that by the unconditional MCT,

$$\int Y = \lim_n \int \mathbb{E}[X_n|\mathcal{G}] = \lim_n \int X_n = \int X = \int \mathbb{E}[X|\mathcal{G}].$$

Put  $B = \cup_n \{E[X_n|\mathcal{G}] > \mathbb{E}[X|\mathcal{G}]\}$ . Then  $\mathbb{P}(B) = 0$ , and  $\mathbb{E}[X_n|\mathcal{G}] \leq \mathbb{E}[X|\mathcal{G}]$  on  $B$  for each  $n$ , so that  $Y \leq \mathbb{E}[X|\mathcal{G}]$  outside  $B$ . It follows that  $Y = \mathbb{E}[X|\mathcal{G}]$  a.s.<sup>7</sup>

**3.2. Conditional Fatou.** Suppose  $X_n \geq 0$  a.s. and  $X_n \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  for each  $n \in \mathbb{N}$ . Then

$$\mathbb{E}[\liminf_n X_n|\mathcal{G}] \leq \liminf_n \mathbb{E}[X_n|\mathcal{G}] \text{ a.s.}$$

**3.3. Conditional DCT.** Suppose  $X_n \xrightarrow{X}$  a.s. Suppose  $|X_n| \leq X_0$  a.s. for all  $n \in \mathbb{N}$  and some  $X_0 \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ . Then

$$\mathbb{E}[X_n|\mathcal{G}] \rightarrow \mathbb{E}[X|\mathcal{G}] \text{ a.s. and in } \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}).$$

<sup>6</sup>Use the fact: if  $\Phi'$  exists everywhere and is increasing, then  $\Phi$  is convex.

<sup>7</sup>Use the fact: if  $X_1 \leq X_2$  a.s. and  $\int X_1 = \int X_2 \in \mathbb{R}$ , then  $X_1 = X_2$  a.s.

**3.4.** Let  $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  be such that  $Y$  is  $\mathcal{G}$ -measurable. Then  $XY \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  iff  $Y\mathbb{E}[X|\mathcal{G}] \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ . In this case,

$$(14.2) \quad \mathbb{E}[XY|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}] \text{ a.s.}$$

Indeed, by splitting  $X = X^+ - X^-$  and  $Y = Y^+ - Y^-$ , one may assume that  $X, Y \geq 0$ .

Suppose first that  $XY \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ . Pick any  $A \in \mathcal{G}$ . For any  $B \in \mathcal{G}$ ,

$$\int_B \mathbb{E}[X\mathbb{1}_A|\mathcal{G}] = \int_B \mathbb{1}_A X = \int_{A \cap B} X = \int_{A \cap B} \mathbb{E}[X|\mathcal{G}] = \int_B \mathbb{1}_A \mathbb{E}[X|\mathcal{G}],$$

where the third equality is due to  $A \cap B \in \mathcal{G}$ . Thus  $\mathbb{E}[X\mathbb{1}_A|\mathcal{G}] = \mathbb{1}_A \mathbb{E}[X|\mathcal{G}]$  a.s., that is, (14.2) holds when  $Y = \mathbb{1}_A$ . Thus it also holds when  $Y$  is  $\mathcal{G}$ -simple. Now since  $Y$  is  $\mathcal{G}$ -measurable, we can take a sequence  $(Y_n)$  of  $\mathcal{G}$ -simple functions such that  $0 \leq Y_n \uparrow Y$ , so that  $0 \leq XY_n \uparrow XY$ . By Conditional MCT,

$$\mathbb{E}[XY|\mathcal{G}] = \lim_n \mathbb{E}[XY_n|\mathcal{G}] = \lim_n Y_n \mathbb{E}[X|\mathcal{G}] = Y \mathbb{E}[X|\mathcal{G}] \text{ a.s.}$$

In particular,  $Y\mathbb{E}[X|\mathcal{G}] \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ . Conversely, assume that  $Y\mathbb{E}[X|\mathcal{G}] \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ . Let  $(Y_n)$  be as before. Then by the unconditional MCT,

$$\int Y \mathbb{E}[X|\mathcal{G}] = \lim_n \int Y_n \mathbb{E}[X|\mathcal{G}] = \lim_n \int \mathbb{E}[XY_n|\mathcal{G}] = \lim_n \int XY_n = \int XY,$$

so that  $XY \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ .

**3.5.** Let  $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ . Then  $X\mathbb{E}[Y|\mathcal{G}] \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  iff  $\mathbb{E}[X|\mathcal{G}]\mathbb{E}[Y|\mathcal{G}] \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$  iff  $Y\mathbb{E}[X|\mathcal{G}] \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ . In this case,

$$\mathbb{E}[X\mathbb{E}[Y|\mathcal{G}]] = \mathbb{E}[Y\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbb{E}[Y|\mathcal{G}]].$$

Indeed, by 3.4,

$$\int X\mathbb{E}[Y|\mathcal{G}] = \int \mathbb{E}[X\mathbb{E}[Y|\mathcal{G}]|\mathcal{G}] = \int \mathbb{E}[X|\mathcal{G}]\mathbb{E}[Y|\mathcal{G}].$$

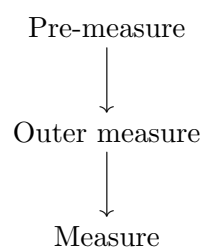
For brevity, it is conventional to suppress the explicit indication of “a.s.” when dealing with conditional expectations; for example, one may drop all the “a.s.” in the previous sections. But the reader should be aware of its existence, particularly when possible ambiguity arises and such suppression is inappropriate.

## Exercises



### **A. Outer measures**

To be specific, the process will construct an outer measure from the pre-measure and then construct a measure from the outer measure:



### **B. Open sets**