

## 2. Preliminaries

In this preliminary chapter, we review some of the basic concepts and results used in the remainder of the book. The main purpose of this chapter is to collect all of these results in a single place, so that a reader can become aware at once as to what background he/she needs in order to be able to understand the material that follows. Though some attempt is made to begin from first principles and to give a few proofs and examples, it must be emphasized that the chapter is *not* intended to be a complete treatment of the topics mentioned. In particular, a reader who is encountering a topic for the first time is encouraged to consult one of the standard references cited at the end of the chapter.

### 2.1 Pseudometric Spaces, Packing and Covering Numbers

#### 2.1.1 Pseudometric Spaces

Suppose  $X$  is a set. A function  $\rho : X \times X \rightarrow \mathbb{R}_+$  is said to be a **pseudometric** if

- (i)  $\rho(x, x) = 0, \forall x \in X,$
- (ii)  $\rho(x, y) = \rho(y, x), \forall x, y \in X,$  and
- (iii)  $\rho(x, z) \leq \rho(x, y) + \rho(y, z), \forall x, y, z \in X.$

If, in addition,  $\rho$  satisfies

- (iv)  $\rho(x, y) = 0 \Rightarrow x = y,$

then  $\rho$  is said to be a **metric**.

If  $\rho$  is a pseudometric (metric) on  $X$ , we say that  $(X, \rho)$  is a **pseudometric space (metric space)**.

Suppose  $(X, \rho)$  is a pseudometric space, and suppose a binary relation  $\sim$  on  $X$  is defined as follows:

$$x \sim y \Leftrightarrow \rho(x, y) = 0.$$

Then it is easy to verify that  $\sim$  is in fact an *equivalence* relation on  $X$ . The reflexivity of  $\sim$  follows from (i) above, the symmetry from (ii), and

the transitivity from (iii). Hence  $X$  can be partitioned into its equivalence classes under  $\sim$ . Let  $\tilde{X}$  denote the resulting collection of equivalence classes; thus a typical element of  $\tilde{X}$  is of the form  $[x]$  where  $x \in X$  and  $[x]$  is the corresponding equivalence class under  $\sim$ . Now, given  $[x], [y] \in \tilde{X}$ , define

$$\tilde{\rho}([x], [y]) = \rho(x, y).$$

It can be easily verified that  $\tilde{\rho}$  is well-defined; that is,  $\tilde{\rho}([x], [y])$  is independent of the particular  $x \in [x], y \in [y]$  used in the right side of the above equation. Also,  $\tilde{\rho}$  is a *metric* on  $\tilde{X}$ . Thus the fact that a pseudometric  $\rho$  might not satisfy (iv) above need not cause much consternation, and in fact, all the familiar metric space concepts of neighbourhoods, open sets, closed sets, etc. can be readily adapted to pseudometric spaces.

Suppose  $(X, \rho)$  is a pseudometric space, and that  $x \in X, \epsilon > 0$ . Then we denote

$$\mathcal{B}(\epsilon, x, \rho) := \{y \in X : \rho(x, y) < \epsilon\}, \text{ and}$$

$$\bar{\mathcal{B}}(\epsilon, x, \rho) := \{y \in X : \rho(x, y) \leq \epsilon\}.$$

Thus  $\mathcal{B}(\epsilon, x, \rho)$  and  $\bar{\mathcal{B}}(\epsilon, x, \rho)$  are respectively the open and closed balls (with respect to the pseudometric  $\rho$ ) of radius  $\epsilon$  centered at  $x$ .

### 2.1.2 Packing and Covering Numbers

Suppose  $(X, \rho)$  is a pseudometric space, and that  $S \subseteq X$ . Given  $\epsilon > 0$ , a set  $\{a_1, \dots, a_n\} \subseteq S$  is said to be an  $\epsilon$ -**cover** of  $S$  if, for each  $x \in S$ , there exists an index  $i$  such that  $\rho(x, a_i) \leq \epsilon$ . Equivalently, a set  $\{a_1, \dots, a_n\}$  is an  $\epsilon$ -*cover* of  $S$  if  $a_i \in S$  for all  $i$ , and in addition

$$\bigcup_{i=1}^n \bar{\mathcal{B}}(\epsilon, a_i, \rho) \supseteq S.$$

The  $\epsilon$ -**covering number** of  $S$  (with respect to the pseudometric  $\rho$ ) is defined as the smallest number  $n$  such that  $S$  has an  $\epsilon$ -cover of cardinality  $n$ , and is denoted by  $N(\epsilon, S, \rho)$ . An  $\epsilon$ -cover of this cardinality is said to be a **minimal  $\epsilon$ -cover**. Note that a minimal  $\epsilon$ -cover need not be unique, but the  $\epsilon$ -covering number is well-defined (and could perhaps be infinite). Similarly, a set  $\{b_1, \dots, b_l\} \subseteq X$  is said to be an **external  $\epsilon$ -cover** of  $S$  if

$$\bigcup_{i=1}^l \bar{\mathcal{B}}(\epsilon, b_i, \rho) \supseteq S.$$

The key point to note here is that the  $b_i$ 's need not themselves belong to  $S$ . The **external  $\epsilon$ -covering number** of  $S$  is defined as the smallest number  $l$  such that  $S$  has an external  $\epsilon$ -cover of cardinality  $l$ , and is denoted by  $L(\epsilon, S, \rho)$ .

The above definitions are not quite standard, in two ways. First, some authors define the (external)  $\epsilon$ -covering number as the smallest number of *open* balls of radius  $\epsilon$  needed to cover  $S$ , as opposed to the smallest number of *closed* balls of radius  $\epsilon$ , as is done here. In the context of learning theory, the definition adopted here offers some advantages. Second, the term “covering number” is used with different meanings by different authors. For instance, Vapnik [187] uses “cover” to mean what we call here as an “external cover,” while our “cover” is his “proper cover.”

**Lemma 2.1.** *For each  $S \subseteq X$  and each  $\epsilon > 0$ , it is true that*

$$N(2\epsilon, S, \rho) \leq L(\epsilon, S, \rho) \leq N(\epsilon, S, \rho).$$

*In particular, the following statements are equivalent:*

- (i) *The  $\epsilon$ -covering number of  $S$  is finite for each  $\epsilon$ .*
- (ii) *The external  $\epsilon$ -covering number of  $S$  is finite for each  $\epsilon$ .*

*Proof.* Obviously the right inequality is valid, because every  $\epsilon$ -cover is also an external  $\epsilon$ -cover. To prove the left inequality, suppose  $\{b_1, \dots, b_m\} \subseteq X$  (not  $S$ !) is an external  $\epsilon$ -cover of  $S$  of minimal cardinality. Then each closed ball  $\bar{B}(\epsilon, b_i, \rho)$  contains an element of  $S$  – if not, then  $b_i$  can be dropped from the  $\epsilon$ -cover, thus contradicting the minimality of the cover. For each  $i = 1, \dots, m$ , choose an  $a_i \in S \cap \bar{B}(\epsilon, b_i, \rho)$ . Then, by the triangle inequality, it follows that  $\{a_1, \dots, a_m\}$  is a  $2\epsilon$ -cover of  $S$ , because every  $x \in S$  is within a distance  $\epsilon$  of some  $b_i$ , which in turn is within  $\epsilon$  of  $a_i$ . ■

A set  $\{b_1, \dots, b_m\} \subseteq S$  is said to be  $\epsilon$ -**separated** if  $\rho(b_i, b_j) > \epsilon \forall i \neq j$ . The  $\epsilon$ -**packing number** of  $S$  is defined as the *largest* number  $m$  such that  $S$  contains an  $\epsilon$ -separated set of cardinality  $m$ , and is denoted by  $M(\epsilon, S, \rho)$ . An  $\epsilon$ -separated set of this cardinality is called a **maximal  $\epsilon$ -separated set**.

Note that some authors call a set  $\{b_1, \dots, b_m\}$  “ $\epsilon$ -separated” if  $\rho(b_i, b_j) \geq \epsilon$  for all  $i \neq j$ , as opposed to the present definition which requires that  $\rho(b_i, b_j) > \epsilon \forall i \neq j$ .

**Lemma 2.2.** *For each  $S \subseteq X$  and  $\epsilon > 0$ , it is true that*

$$M(2\epsilon, S, \rho) \leq L(\epsilon, S, \rho) \leq N(\epsilon, S, \rho) \leq M(\epsilon, S, \rho).$$

*Proof.* Suppose  $\{a_1, \dots, a_k\}$  is a maximal  $2\epsilon$ -separated set in  $S$ . Then no closed ball of radius  $\epsilon$  can contain more than one  $a_i$  (by the triangle inequality). This is true irrespective of whether the center of the ball belongs to  $S$  or not. This proves the left inequality. The middle inequality is already proved in Lemma 2.1.

Suppose  $\{b_1, \dots, b_m\} \subseteq S$  is a maximal  $\epsilon$ -separated set. Then  $\{b_1, \dots, b_m\}$  must be an  $\epsilon$ -cover of  $S$  – otherwise there would exist a  $b_{m+1}$  that is more than  $\epsilon$ -far from each of  $b_1, \dots, b_m$ , thus contradicting the maximality. This establishes the right inequality. ■

See [106] for an excellent discussion of packing and covering numbers, as well as a wealth of examples.

**Lemma 2.3.** *Suppose  $(X, \rho)$  is a pseudometric space, and let  $S \subseteq X$ . Then each of the three functions  $L(\epsilon, S, \rho)$ ,  $N(\epsilon, S, \rho)$ , and  $M(\epsilon, S, \rho)$  is a nondecreasing function of  $\epsilon$  as  $\epsilon$  decreases towards zero; that is*

$$L(\epsilon_1, S, \rho) \geq L(\epsilon_2, S, \rho) \text{ if } \epsilon_1 \leq \epsilon_2,$$

and similarly for the other two functions. Each function is continuous from the right; that is

$$L(\epsilon_0, S, \rho) = \lim_{\epsilon \rightarrow \epsilon_0^+} L(\epsilon, S, \rho), \quad \forall \epsilon_0 > 0,$$

and similarly for the other two functions.

*Proof.* The fact that each function is nondecreasing is obvious. For a proof of the second part of the lemma, see [106], Theorem III. ■

### 2.1.3 Compact and Totally Bounded Sets

Suppose  $(X, \rho)$  is a pseudometric space, and that  $S \subseteq X$ . Then we say that  $S$  is **compact** if every open cover of  $S$  has a finite subcover, and that  $S$  is **totally bounded** if

$$N(\epsilon, S, \rho) < \infty \quad \forall \epsilon > 0,$$

i.e., if  $S$  has a finite  $\epsilon$ -covering number for each  $\epsilon > 0$ . Note that, from Lemmas 2.3 and 2.1, the above condition is equivalent to

$$L(\epsilon, S, \rho) < \infty \quad \forall \epsilon > 0,$$

and to

$$M(\epsilon, S, \rho) < \infty \quad \forall \epsilon > 0.$$

Note that, instead of “totally bounded,” one could also call such a set “precompact,” because of the following result:

**Lemma 2.4.** *Suppose  $(X, \rho)$  is a pseudometric space, and that  $S \subseteq X$ . Then  $S$  is compact if and only if it is totally bounded and closed.*

*Proof.* See [99], p. 198, Theorem 32. ■

Thus total boundedness is “almost” the same as compactness, the only difference being that a compact set is also closed, whereas a totally bounded set may or may not be closed.

## 2.2 Probability Measures

### 2.2.1 Definition of a Probability Space

Suppose  $X$  is a set. A (nonempty) collection  $\mathcal{S}$  of subsets of  $X$  is said to be a  $\sigma$ -algebra if it satisfies the following:

- (i)  $\mathcal{S}$  is closed under complementation; i.e.,  $A \in \mathcal{S} \Rightarrow A^c \in \mathcal{S}$ .
- (ii)  $\mathcal{S}$  is closed under countable union; i.e., if  $A_i \in \mathcal{S}$  for  $i = 1, 2, \dots$ , then  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{S}$ .

It is an easy consequence of (i) and (ii) that  $\mathcal{S}$  is also closed under countable intersection.

Suppose  $(X, \rho)$  is a pseudometric space. Then the smallest  $\sigma$ -algebra of subsets of  $X$  that contains every closed subset of  $X$  is called the **Borel  $\sigma$ -algebra** of  $(X, \rho)$ . Note that, by Condition (i) above, the Borel  $\sigma$ -algebra also contains every open subset of  $X$ .

If  $S$  is a set and  $\mathcal{S}$  is a  $\sigma$ -algebra of subsets of  $X$ , then the pair  $(X, \mathcal{S})$  is called a **measurable space**. Suppose  $(X, \mathcal{S})$ ,  $(Y, \mathcal{T})$  are measurable spaces, and that  $f : X \rightarrow Y$ . Then  $f$  is said to be a **measurable function** if  $f^{-1}(T) \in \mathcal{S}$  whenever  $T \in \mathcal{T}$ .

A function  $\mu : \mathcal{S} \rightarrow \mathbb{R}_+$  is said to be a **measure** if  $\mu(\emptyset) = 0$ , and  $\mu$  is countably additive; that is, if  $A_i \in \mathcal{S}$ ,  $i = 1, 2, \dots$  is a finite or countable collection of pairwise disjoint sets, then

$$\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i).$$

A ready consequence of the above property is the subadditivity property: Suppose  $A_i \in \mathcal{S}$ ,  $i = 1, 2, \dots$  is a countable collection of sets (not necessarily pairwise disjoint). Then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i).$$

A measure  $P : \mathcal{S} \rightarrow \mathbb{R}_+$  is said to be a **probability measure** if  $P(X) = 1$ . We refer to  $(X, \mathcal{S}, P)$  as a **probability space**.

We do not discuss at all the notion of *integrating* a measurable function, as such a discussion would take us too far afield. The reader is referred instead to [76] for a thorough treatment. It is not really necessary for the reader to master the various intricacies of of integration in the measure-theoretic sense in order to follow the contents of the book.

### 2.2.2 A Pseudometric Induced by a Probability Measure

Suppose  $(X, \mathcal{S}, P)$  is a probability space. Then  $P$  induces a pseudometric on  $\mathcal{S}$ , as follows: For each  $A, B \subseteq X$ , define their **symmetric difference**  $A\Delta B$  by

$$A\Delta B = (A^c \cap B) \cup (A \cap B^c),$$

where  $A^c$  denotes the complement of the set  $A$ . An equivalent definition is:

$$A\Delta B = (A \cup B) - (A \cap B).$$

Evidently,  $A\Delta B$  is the set of points that belong to *exactly* one of the two sets  $A$  and  $B$ . It is easy to see that  $A\Delta B = B\Delta A$ , so that  $\Delta$  is indeed symmetric. Also, it is tedious but routine to verify that, for three sets  $A, B, C$ , we have

$$A\Delta B = (A\Delta C) \Delta (B\Delta C).$$

Clearly, if  $A, B \in \mathcal{S}$ , then  $A\Delta B \in \mathcal{S}$ . Thus it is possible to define the function  $d_P : \mathcal{S} \times \mathcal{S} \rightarrow [0, 1]$  by

$$d_P(A, B) = P(A\Delta B).$$

Now it is a routine matter to verify that  $d_P$  is a pseudometric on  $\mathcal{S}$ . Axioms (i) and (ii) of the definition follow readily. To prove the triangle inequality, one uses the fact that

$$A\Delta C \subseteq (A\Delta B) \cup (B\Delta C),$$

whence

$$P(A\Delta C) \leq P(A\Delta B) + P(B\Delta C).$$

However,  $d_P$  is in general *not* a metric because  $d_P(A, B) = 0$  whenever  $A\Delta B$  is a set of zero measure, even if  $A\Delta B \neq \emptyset$ .

More generally, let  $[0, 1]^X$  denote the set of *measurable* functions mapping  $X$  into  $[0, 1]$ , when  $[0, 1]$  is equipped with the Borel  $\sigma$ -algebra.<sup>1</sup> Then one can define a pseudometric  $d_P$  on  $[0, 1]^X$  by

$$d_P(f, g) = \int_X |f(x) - g(x)| P(dx), \quad \forall f, g \in [0, 1]^X.$$

It is easy to verify that this  $d_P$  is also a pseudometric. In general it is not a metric, because  $d_P(f, g) = 0$  whenever  $f$  and  $g$  differ on a set of measure zero, even if  $f \neq g$ . Actually, this  $d_P$  is a *generalization* of the earlier  $d_P$  defined on  $\mathcal{S}$ , which justifies the use of the same symbol for both. To see this, observe that there is a one-to-one correspondence between sets in  $\mathcal{S}$  and (measurable) functions mapping  $X$  into  $\{0, 1\}$ . Specifically, if  $A \in \mathcal{S}$ , then its indicator function  $I_A(\cdot)$  defined by

<sup>1</sup> This is a slight abuse of notation because, strictly speaking,  $[0, 1]^X$  should denote the set of *all* functions mapping  $X$  into  $[0, 1]$ , measurable or otherwise.

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A, \end{cases}$$

is measurable and maps  $X$  into  $\{0, 1\}$ . Conversely, suppose  $f : X \rightarrow \{0, 1\}$  is measurable. Then its support defined by

$$\text{supp}(f) = \{x \in X : f(x) = 1\}$$

belongs to  $\mathcal{S}$ . Now, if  $A, B \in \mathcal{S}$ , then it is easy to see that

$$d_P(A, B) = d_P(I_A, I_B),$$

where the  $d_P$  on the left side is defined on  $\mathcal{S}$  while the  $d_P$  on the right side is defined on  $[0, 1]^X$ . This justifies the use of the same symbol for both quantities.

### 2.2.3 A Metric on the Set of Probability Measures

Suppose  $(X, \mathcal{S})$  is a measurable space, and let  $\mathcal{P}^*$  denote the set of all probability measures on  $(X, \mathcal{S})$ . It is possible to define a metric on  $\mathcal{P}^*$  as follows: Given  $P, Q \in \mathcal{P}^*$ , let

$$\rho(P, Q; \mathcal{S}) := \sup_{A \in \mathcal{S}} |P(A) - Q(A)|.$$

The function  $\rho$  is indeed a metric (and not merely a pseudometric) because, if  $P, Q$  are probability measures on  $(X, \mathcal{S})$  and  $P \neq Q$ , then there exists at least one set  $A \in \mathcal{S}$  such that  $P(A) \neq Q(A)$ ; hence  $\rho(P, Q) > 0$ . Note that  $\rho$  is called the **total variation metric** on  $\mathcal{P}^*$ . In the above definition, the underlying  $\sigma$ -algebra  $\mathcal{S}$  is explicitly highlighted, since there will be cases when we will compute the total variation metric between the same pair of probability measures, but with respect to different  $\sigma$ -algebras. However, if  $\mathcal{S}$  is obvious from the context, then it can be omitted, and we can simply write  $\rho(P, Q)$ .

Suppose  $X \subseteq \mathbb{R}$ , and that  $P, Q$  are probability measures with densities  $p(\cdot)$  and  $q(\cdot)$  respectively. Thus, if  $A \subseteq X$  is measurable, then

$$P(A) = \int_A p(x) dx, \text{ and } Q(A) = \int_A q(x) dx.$$

Then the total variation metric between  $P$  and  $Q$  equals

$$\rho(P, Q) = \int_X |f(x) - g(x)| dx;$$

that is,  $\rho(P, Q)$  is the  $L_1$ -distance between the densities  $p(\cdot)$  and  $q(\cdot)$ .

Now it is shown that the total variation metric has a very useful property. In order to present it, we begin by discussing the notion of product  $\sigma$ -algebras and product probability measures.

Suppose  $(X, \mathcal{S}, P)$  and  $(Y, \mathcal{T}, R)$  are probability spaces. Then a set of the form  $A \times B$ ,  $A \in \mathcal{S}, B \in \mathcal{T}$  is called a “cylinder set.” The smallest  $\sigma$ -algebra on  $X \times Y$  that contains all such cylinder sets is called the “product  $\sigma$ -algebra” and is denoted by  $\mathcal{S} \times \mathcal{T}$ . By defining

$$R(A \times B) := P(A) \cdot Q(B), \forall A \in \mathcal{S}, B \in \mathcal{T},$$

we can define the measure of cylinder sets, which can then be extended, via the Kolmogorov extension theorem, to all of  $\mathcal{S} \times \mathcal{T}$ . The resulting (probability) measure  $R$  is called the “product measure” and is denoted by  $P \times Q$ . Clearly the idea can be extended to any finite number of probability spaces, and indeed, to even more general situations; see Section 2.4.

**Lemma 2.5.** *Suppose  $(X, \mathcal{S}, P), (Y, \mathcal{T}, Q)$  and  $(Y, \mathcal{T}, R)$  are probability spaces. Then*

$$\rho(P \times Q, P \times R) = \rho(Q, R).$$

*Proof.* Consider the collection of sets of the form  $C = \cup_{i=1}^{\infty} (A_i \times B_i)$ , where  $A_i \in \mathcal{S}, B_i \in \mathcal{T}$ , and the  $A_i$  are pairwise disjoint. Such a collection of sets forms an “algebra” in that it is closed under complementation and finite union; but it might not be a  $\sigma$ -algebra since it might not be closed under *countable* union. However, it can be shown that  $\rho(P \times Q, P \times R)$  equals the supremum of the difference  $|(P \times Q)(C) - (P \times R)(C)|$  over all such sets  $C$ .

Now

$$\begin{aligned} |(P \times Q)(C) - (P \times R)(C)| &\leq \sum_{i=1}^{\infty} |(P \times Q)(A_i \times B_i) - (P \times R)(A_i \times B_i)| \\ &= \sum_{i=1}^{\infty} P(A_i) |Q(B_i) - R(B_i)| \\ &\leq \left( \sum_{i=1}^{\infty} P(A_i) \right) \cdot \rho(Q, R) \\ &\leq \rho(Q, R). \end{aligned}$$

This shows that  $\rho(P \times Q, P \times R) \leq \rho(Q, R)$ . The opposite inequality can be proven by considering sets of the form  $X \times B, B \in \mathcal{T}$ . ■

**Lemma 2.6.** *Suppose  $(X, \mathcal{S})$  is a measurable space, and that  $P, Q$  are probability measures on this space. Then*

$$\rho(P^k, Q^k) \leq k\rho(P, Q).$$

*Proof.* By the triangle inequality, we have

$$\rho(P^k, Q^k) \leq \rho(P^k, P^{k-1}Q) + \dots + \rho(PQ^{k-1}, Q^k).$$

By Lemma 2.5, each of the quantities on the right side equals  $\rho(P, Q)$ . ■



One consequence of Lemma 2.6 is that if  $\{P_i\}$  is a sequence of probability measures on  $(X, \mathcal{S})$  converging to the probability measure  $Q$ , then  $P_i^k \rightarrow Q^k$  for each *fixed* integer  $k$ . At the same time, it can also be shown that  $P, Q$  are distinct probability measures on  $(X, \mathcal{S})$ , then  $\rho(P^k, Q^k) \rightarrow 1$  as  $k \rightarrow \infty$ . That is, as  $k \rightarrow \infty$ , the measures  $P^k, Q^k$  tend to become “mutually singular” in that they are supported on disjoint sets.

#### 2.2.4 Random Variables

Suppose  $(\Omega, \mathcal{T}, Q)$  is a probability space. Thus  $\Omega$  is a set,  $\mathcal{T}$  is a  $\sigma$ -algebra of subsets of  $\Omega$ , and  $Q$  is a probability measure on  $(\Omega, \mathcal{T})$ . Suppose  $(X, \mathcal{S})$  is a measurable space. Then an  **$X$ -valued random variable** is defined as a measurable map, call it  $\mathbf{X}$ , from  $(\Omega, \mathcal{T}, Q)$  to  $(X, \mathcal{S})$ . Note that in the probability literature, it is common to restrict the term “random variable” to the situation where  $X = \mathbb{R}$  and  $\mathcal{S}$  is the Borel  $\sigma$ -algebra. At best,  $\mathbf{X}$  is taken to be  $\mathbb{R}^k$  for some integer  $k$  and  $\mathcal{S}$  is taken to be the associated Borel  $\sigma$ -algebra. However, for present purposes, it is desirable to adopt the more general usage stated above.

Suppose  $f$  is a measurable map from a probability space  $(\Omega, \mathcal{T}, Q)$  into  $\mathbb{R}$ . Thus one can also think of  $f$  as a real-valued random variable. The **expected value** of the function  $f$  is defined as

$$E(f, Q) := \int_{\Omega} f(\omega) Q(d\omega),$$

assuming of course that the integral is well-defined. In particular, we can speak of the expected value of a random variable  $\mathbf{X}$  with probability measure  $P$ , and denote it by  $E(\mathbf{X}, P)$ . If  $P$  is clear from the context (or if it does not matter what  $P$  is), then we simply write  $E(\mathbf{X})$ .

Suppose  $\mathbf{X}$  is a real-valued random variable. Then the **distribution function** of  $\mathbf{X}$  is the function  $P_{\mathbf{X}}$  mapping  $\mathbb{R}$  into  $[0, 1]$  defined by

$$P_{\mathbf{X}}(a) := Q\{\omega \in \Omega : \mathbf{X}(\omega) \leq a\}.$$

It is obvious that  $P_{\mathbf{X}}(a)$  is nondecreasing as a function of  $a$ . The distribution function has a property known as “cadlag,” which is an abbreviation for the French expression “continuité a droite, limité a gauche.” What it means is that the distribution function is continuous from the right, and has well-defined limits from the left. Thus

$$\lim_{a \rightarrow a_0^+} P_{\mathbf{X}}(a) = P_{\mathbf{X}}(a_0), \quad \forall a_0 \in \mathbb{R}, \quad \text{and} \quad \lim_{a \rightarrow a_0^-} P_{\mathbf{X}}(a) \text{ exists.}$$

If  $\mathbf{X}_1, \mathbf{X}_2$  are real-valued random variables defined on a common probability space  $(\Omega, \mathcal{T}, Q)$ , then one can define their **joint distribution** as follows:

$$P_{\mathbf{X}_1, \mathbf{X}_2}(a_1, a_2) := Q\{\omega \in \Omega : \mathbf{X}_1(\omega) \leq a_1, \mathbf{X}_2(\omega) \leq a_2\}.$$

It is obvious that the notion of a joint distribution can be readily extended to any finite number of real-valued random variables.

Suppose  $(\Omega, \mathcal{T}, Q)$  and  $(\Omega', \mathcal{T}', Q')$  are probability spaces, and that  $\mathbf{X}, \mathbf{X}'$  are random variables mapping  $(\Omega, \mathcal{T}, Q)$  into  $\mathbb{R}$  and  $(\Omega', \mathcal{T}', Q')$  into  $\mathbb{R}$  respectively. Then the random variables  $\mathbf{X}$  and  $\mathbf{X}'$  are said to have the same “law” if they have the same distribution function, that is,  $P_{\mathbf{X}}(\cdot) = P_{\mathbf{X}'}(\cdot)$ . The point is that the domain of a random variable is really not important.

Suppose  $(\Omega, \mathcal{T}, Q)$  is a probability space, and that  $A, B \in \mathcal{T}$ . Thus  $A$  and  $B$  are deemed to be “events.” The events  $A$  and  $B$  are said to be **independent** under the probability measure  $Q$  if  $Q(A \cap B) = Q(A) \cdot Q(B)$ . With a little bit of work, this notion can be extended to define the notion of independence for random variables.

Suppose  $\mathbf{X}_1, \mathbf{X}_2$  are random variables defined on a common probability space  $(\Omega, \mathcal{T}, Q)$ . Thus  $\mathbf{X}_i$  maps  $\Omega$  into a measurable space  $(X_i, \mathcal{S}_i)$  for  $i = 1, 2$ . Then the random variables  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are independent if, for every  $A \in \mathcal{S}_1, B \in \mathcal{S}_2$ , the preimages  $\mathbf{X}_1^{-1}(A)$  and  $\mathbf{X}_2^{-1}(B)$  are independent. Equivalently, the random variables  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are independent if

$$Q\{\mathbf{X}_1(\omega) \in A, \mathbf{X}_2(\omega) \in B\} = Q\{\mathbf{X}_1(\omega) \in A\} Q\{\mathbf{X}_2(\omega) \in B\}.$$

Note that in the above equation we employ the commonly used abbreviation, whereby  $Q\{S\}$  denotes  $Q\{\omega \in \Omega : S \text{ is true}\}$ . In particular, if  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are real-valued random variables, then  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are independent if and only if the joint distribution function is a product of the individual distribution functions, that is, if and only if

$$P_{\mathbf{X}_1, \mathbf{X}_2}(a_1, a_2) = P_{\mathbf{X}_1}(a_1) \cdot P_{\mathbf{X}_2}(a_2).$$

More generally, suppose as above that  $\mathbf{X}_1, \mathbf{X}_2$  are random variables defined on a common probability space. Define  $X = X_1 \times X_2$ , and let  $\mathcal{S}$  denote the *product  $\sigma$ -algebra*,  $\mathcal{S}_1 \times \mathcal{S}_2$ , which consists of the smallest  $\sigma$ -algebra that contains all “cylinder sets” of the form  $A \times B$ ,  $A \in \mathcal{S}_1, B \in \mathcal{S}_2$ . Then one can think of the map  $\omega \mapsto (\mathbf{X}_1(\omega), \mathbf{X}_2(\omega))$  as a random variable taking values in the measurable space  $(X, \mathcal{S})$ . Then the collection  $\mathbf{X}_1^{-1}(A)$ ,  $A \in \mathcal{S}_1$  is a  $\sigma$ -algebra on  $\Omega$ , called the  $\sigma$ -algebra **generated** by  $\mathbf{X}_1$ , and denoted by  $\Sigma(\mathbf{X}_1)$ . Similarly,  $\mathbf{X}_2$  also generates a  $\sigma$ -algebra on  $\Omega$ , denoted by  $\Sigma(\mathbf{X}_2)$ . Now consider the product  $\sigma$ -algebra generated by  $\Sigma(\mathbf{X}_1)$  and  $\Sigma(\mathbf{X}_2)$ . Denote it by  $\Sigma(\mathbf{X}_1, \mathbf{X}_2)$ , since it is the  $\sigma$ -algebra generated jointly by the two variables  $\mathbf{X}_1$  and  $\mathbf{X}_2$ . Since both  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are measurable maps,  $\Sigma(\mathbf{X}_1, \mathbf{X}_2) \subseteq \mathcal{T}$ . Now the probability measure  $Q$ , restricted to  $\Sigma(\mathbf{X}_i)$ , is called the **one-dimensional marginal probability measure** corresponding to  $\mathbf{X}_i$ , and denoted by  $Q_{\Sigma(\mathbf{X}_i)}$ . Note that the original  $\sigma$ -algebra  $\mathcal{T}$  could be strictly bigger than  $\Sigma(\mathbf{X}_1, \mathbf{X}_2)$ , but for the purposes of checking whether or not  $\mathbf{X}_1, \mathbf{X}_2$  are independent, we do not need to work with all of  $\mathcal{T}$ , only  $\Sigma(\mathbf{X}_1, \mathbf{X}_2)$ . With these definitions, it is easy to see that  $\mathbf{X}_1, \mathbf{X}_2$  are independent if and only if

$$Q_{\Sigma(\mathbf{X}_1, \mathbf{X}_2)} = Q_{\Sigma(\mathbf{X}_1)} \times Q_{\Sigma(\mathbf{X}_2)}.$$

### 2.2.5 Conditional Expectations

In this we give a strictly functional description of conditional expectations. The present treatment is much less general than is normally found in most texts on probability, but it is good enough for present purposes.

Suppose  $(\Omega, \mathcal{T}, P)$  is a probability space, and that  $f$  is a measurable map from  $(\Omega, \mathcal{T})$  into the real numbers. Let  $\mathcal{B}$  denote the Borel  $\sigma$ -algebra on  $\mathbb{R}$ . Then the collection of preimages under  $f$  of Borel measurable sets is called the  **$\sigma$ -algebra generated by  $f$** , and is denoted by  $\Sigma(f)$ , or sometimes by  $\mathcal{F}$ . Thus

$$\mathcal{F} = \Sigma(f) := \{f^{-1}(S), S \in \mathcal{B}\}.$$

Suppose  $\Sigma$  is a subalgebra of  $\mathcal{T}$ , and let  $\mathcal{M}(\Sigma)$  denote the set of all functions from  $\Omega$  into  $\mathbb{R}$  that are measurable with respect to  $\Sigma$ . Clearly  $\mathcal{M}(\Sigma)$  is a linear vector space. Suppose now that  $g$  is another measurable mapping from  $(\Omega, \mathcal{T}, P)$  into  $\mathbb{R}$  with finite variance. This means that

$$\int_{\Omega} g^2(\omega) P(d\omega) < \infty.$$

To put it another way,  $g(\cdot)$  belongs to  $L_2(\Omega, \mathcal{T}, P)$ . Then the **conditional expectation of  $g$  with respect to the  $\sigma$ -algebra  $\Sigma$**  is defined as the unique function  $h \in \mathcal{M}(\Sigma)$  that minimizes the mean-squared error

$$\int_X |h(\omega) - g(\omega)|^2 P(d\omega).$$

The conditional expectation of  $g$  with respect to  $\Sigma$  is denoted by  $g_{\Sigma}$ . Note that, whereas the expected value of a random variable is a real number, the conditional expectation of a random variable is itself a random variable. This is why, even though it is customary in the probability literature to denote the conditional expectation by  $E(g|\Sigma)$ , we prefer to use the notation  $g_{\Sigma}$ . Moreover, suppose  $X_1, \dots, X_n$  are random variables on  $(\Omega, \mathcal{T}, P)$ , and let  $\Sigma$  denote the  $\sigma$ -algebra generated by these random variables. Then any function  $h \in \mathcal{M}(\Sigma)$  can be written as  $h = h(X_1, \dots, X_n)$  where  $h$  is a measurable function from  $\mathbb{R}^n$  into  $\mathbb{R}$ . Therefore, if  $h = g_{\Sigma}$ , then we can write  $g_{\Sigma} = h(X_1, \dots, X_n)$ . This relationship expresses the “best estimate” of  $g$  very explicitly as a function of the  $n$  random variables, and is the motivation for the notation  $E(g|X_1, \dots, X_n)$ .

It is possible to define the notion of conditional expectation in much more general settings. However, by restricting attention to random variables of finite variance, the technicalities are kept to a minimum. In particular, it is a ready consequence of the projection theorem that there exists a unique conditional expectation, and that the error term  $g - g_{\Sigma}$  is “orthogonal” to the space  $\mathcal{M}(\Sigma)$ . Therefore

$$\int_X h(\omega)[g(\omega) - g_{\Sigma}(\omega)] P(d\omega) = 0, \forall h \in \mathcal{M}(\Sigma),$$

provided only that the integral is well-defined. In particular, the above equality holds for every  $h$  with finite variance. The above relationship can be written very simply as

$$E[h(g - g_{\mathcal{F}})] = 0, \quad E(hg) = E(hg_{\mathcal{F}}), \quad \forall h \in \mathcal{M}(\mathcal{F}). \quad (2.2.1)$$

## 2.3 Large Deviation Type Inequalities

An important, and recurring, theme in these notes is the use of so-called “large deviation” type inequalities. These inequalities give an estimate of the probability that an “average” of independent random variables differs considerably from its mean value. In this section, several inequalities used in the sequel are summarized.

### 2.3.1 Chernoff Bounds

Suppose  $\mathbf{X}$  is a random variable with only two possible values, namely 0 and 1, and suppose further that the probability that  $\mathbf{X} = 1$  is  $p$ . Then clearly  $E(\mathbf{X}) = p$ . Let  $x_1, \dots, x_m$  denote independent samples of  $\mathbf{X}$ ; these are also known as **Bernoulli trials**. Now define

$$S_m = \sum_{i=1}^m x_i, \quad A_m = \frac{1}{m} \sum_{i=1}^m x_i.$$

Then  $A_m$  can be thought of as the *empirical mean* of the random variable  $\mathbf{X}$ . In other words,  $A_m$  is an *estimate* of the probability that  $\mathbf{X} = 1$  based on  $m$  trials. Note that  $A_m$  is itself a random variable. The probability that  $A_m$  exceeds a number  $r$  can be expressed as

$$\Pr\{A_m \geq r\} = \Pr\{S_m \geq mr\} = \sum_{k \geq mr}^m \binom{m}{k} p^k (1-p)^{m-k}.$$

Similarly, the probability that  $A_m$  is less than a given number  $r$  is given by

$$\Pr\{A_m \leq r\} = \sum_{k=0}^{k \leq mr} \binom{m}{k} p^k (1-p)^{m-k}.$$

The **Chernoff bounds** give upper bounds on the right side of the above inequalities. The bounds can be given in either additive form or multiplicative form. In the additive form, the bounds state that, for all  $\epsilon \in [0, 1]$ ,

$$\Pr\{A_m \geq p + \epsilon\} \leq \exp(-2m\epsilon^2), \quad \text{and}$$

$$\Pr\{A_m \leq p - \epsilon\} \leq \exp(-2m\epsilon^2).$$

Combining these two inequalities gives

$$\Pr\{|A_m - p| \geq \epsilon\} \leq 2 \exp(-2m\epsilon^2).$$

In the multiplicative form the bounds state that, for all  $\gamma \in [0, 1]$ ,

$$\Pr\{A_m \geq (1 + \gamma)p\} \leq \exp(-\gamma^2 mp/3), \text{ and}$$

$$\Pr\{A_m \leq (1 - \gamma)p\} \leq \exp(-\gamma^2 mp/2).$$

Note that, unlike the additive bounds, the multiplicative bounds are not “symmetric.” Also, the multiplicative bounds require that  $\gamma \leq 1$ , whereas the additive bounds do not place any restrictions on the size of  $\epsilon$  relative to  $p$ .

It is worth noting that, in case  $p$  is itself “small,” the multiplicative bounds are less conservative than the additive bounds. To illustrate, suppose  $p = \epsilon$ , a “small” number, and let us bound the probability that  $A_m \geq 1.5\epsilon$ . The additive form of the Chernoff bound leads to the estimate

$$\Pr\{A_m \geq 1.5\epsilon\} \leq \exp(-m\epsilon^2/2),$$

whereas the multiplicative form leads to the estimate

$$\Pr\{A_m \geq 1.5\epsilon\} \leq \exp(-m\epsilon/12)$$

after substituting  $\gamma = 0.5, p = \epsilon$ . It is important to note that the first bound contains an  $\epsilon^2$  in the exponent, whereas the second bound contains only  $\epsilon$ . Hence, when  $\epsilon$  is small, the second bound is considerably superior to the first.

In the multiplicative form of the Chernoff bound, the restriction that  $\gamma \leq 1$  is not serious when it is desired to estimate  $\Pr\{A_m \leq (1 - \gamma)p\}$ . However, it is a bit of a nuisance when it is desired to estimate  $\Pr\{A_m \geq (1 + \gamma)p\}$ , since the quantity  $(1 + \gamma)p$  is effectively limited to the range  $[p, 2p]$ . Obviously situations can arise where one would like to estimate  $\Pr\{A_m \geq (1 + \gamma)p\}$  with  $\gamma > 1$ . In such cases, the Chernoff bound cannot be applied directly. However, with a little imagination, the range of applicability of the bound can be “stretched.” Specifically, observe that the map  $p \mapsto p^k(1 - p)^{m-k}$  is nondecreasing whenever  $p \leq k/m$ . This observation leads to the following alternate form of the multiplicative Chernoff bound. Suppose  $X$  is a Bernoulli process, and that the probability  $\Pr\{X = 1\}$  is *less than or equal to*  $\mu$ . Let  $A_m$  denote the  $m$ -fold average of independent observations of  $X$ , as above. Then

$$\Pr\{A_m \geq (1 + \gamma)\mu\} \leq \exp(-\gamma^2 m\mu/3), \text{ for } 0 \leq \gamma \leq 1, E(X) \leq \mu. \quad (2.3.1)$$

Note that it is permissible for  $(1 + \gamma)\mu$  to exceed  $2p$ , where  $p = \Pr\{X = 1\}$ . The above inequality follows from

$$\Pr\{A_m \geq (1 + \gamma)\mu\} = \sum_{k \geq (1 + \gamma)\mu m} \binom{m}{k} p^k (1 - p)^{m-k}$$

$$\begin{aligned} &\leq \sum_{k \geq (1+\gamma)\mu m} \binom{m}{k} \mu^k (1-\mu)^{m-k} \\ &\leq \exp(-\gamma^2 m \mu / 3), \end{aligned}$$

where the last inequality follows from the “standard” Chernoff bound. Now suppose it is desired to use this alternate form to estimate  $\Pr\{A_m \geq (1+\delta)p\}$  where  $\delta > 1$ . Then one can apply the above bound with  $\mu = (1+\delta)p/2$  and  $\gamma = 1$ , and derive that

$$\Pr\{A_m \geq (1+\delta)p\} \leq \exp(-(1+\delta)mp/6), \quad \forall \delta > 1.$$

In this way, the multiplicative form of the Chernoff bound can be extended to cover all values in the range  $[p, 1]$ .

### 2.3.2 Chernoff-Okamoto Bound

The **Chernoff-Okamoto bound** is less conservative than the Chernoff bounds, but applies only when  $p \leq 0.5$ . It states that, if  $p \leq 0.5$  and  $r \leq p$ , then

$$\Pr\{A_m \leq r\} \leq \exp\left[-\frac{m(p-r)^2}{2p(1-p)}\right].$$

By applying the above bound with  $r = p - \epsilon$  and  $r = (1-\gamma)p$  respectively, and observing that  $2p(1-p) \leq 0.5$  for all  $p \in [0, 1]$ , one can derive two of the four Chernoff bounds above as consequences of the Chernoff-Okamoto bound.

### 2.3.3 Hoeffding’s Inequality

Hoeffding’s inequality is a very general inequality that applies to the sum of independent random variables with bounded range.

**Lemma 2.7.** *Suppose  $Y_1, \dots, Y_m$  are independent random variables, and that  $a_i \leq Y_i \leq b_i$  for each  $i$ . Suppose  $y_1, \dots, y_m$  are realizations of these random variables. Then*

$$\Pr\left\{\sum_{i=1}^m (y_i - E(Y_i)) \geq \alpha\right\} \leq \exp\left[-2\alpha^2 / \sum_{i=1}^m (b_i - a_i)^2\right].$$

**Remark:** Note that the additive form of the Chernoff bounds can be derived readily from Hoeffding’s inequality (but not the multiplicative form).

The proof of Hoeffding’s inequality uses the following auxiliary lemma.

**Lemma 2.8.** *Suppose  $X$  is a zero-mean random variable assuming values in the interval  $[a, b]$ . Then for any  $s > 0$ , we have*

$$E[\exp(sX)] \leq \exp(s^2(b-a)^2/8).$$

*Proof. (of the auxiliary lemma):* Since the exponential is a convex function, the value of  $e^{sx}$  is bounded by the corresponding convex combination of its extreme values; that is,

$$\exp(sx) \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}, \quad \forall x \in [a, b].$$

Now take the expectation of both sides, and use the fact that  $E(\mathbf{X}) = 0$ . This gives

$$\begin{aligned} E[\exp(s\mathbf{X})] &\leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\ &= (1-p + pe^{s(b-a)})e^{-ps(b-a)} \\ &=: \exp(\phi(u)), \end{aligned}$$

where  $p := -a/(b-a)$ ,  $u := s(b-a)$ , and  $\phi(u) := -pu + \ln(1-p + pe^u)$ . Clearly  $\phi(u) = 0$ . Moreover, a routine calculation shows that

$$\phi'(u) = -p + \frac{p}{p + (1-p)e^{-u}},$$

whence  $\phi'(u) = 0$  as well. Moreover,

$$\phi''(u) = \frac{p(1-p)e^{-u}}{(p + (1-p)e^{-u})^2} \leq 0.25.$$

Hence by Taylor's theorem, there exists a  $\theta \in [0, u]$  such that

$$\phi(u) = \phi''(\theta)u^2/2 \leq u^2/8 = \frac{s^2(b-a)^2}{8}.$$

This completes the proof. ■

*Proof. (of Hoeffding's inequality):* For any nonnegative random variable, we have

$$\Pr\{\mathbf{X} \geq \epsilon\} \leq \frac{E(\mathbf{X})}{\epsilon},$$

which is known as **Markov's inequality**. Hence, for every  $s > 0$ , we have

$$\Pr\{\mathbf{X} \geq \epsilon\} = \Pr\{e^{s\mathbf{X}} \geq e^{s\epsilon}\} \leq \frac{E[\exp(s\mathbf{X})]}{\exp(s\epsilon)} = e^{-s\epsilon} E[\exp(s\mathbf{X})].$$

Now apply this inequality to the random variable

$$Z_m := \sum_{i=1}^m (Y_i - E(Y_i)),$$

which has zero mean since the  $Y_i$ 's are independent. Then

$$\begin{aligned}
\Pr\{Z_m \geq \epsilon\} &\leq e^{-s\epsilon} E \left[ \exp \left( s \sum_{i=1}^m (Y_i - E(Y_i)) \right) \right] \\
&= e^{-s\epsilon} \prod_{i=1}^m E[e^{s(Y_i - E(Y_i))}] \text{ by independence} \\
&\leq e^{-s\epsilon} \prod_{i=1}^m e^{s^2(b_i - a_i)^2/8} \text{ by Lemma 2.8} \\
&= \exp \left[ -s\epsilon + s^2 \sum_{i=1}^m \frac{(b_i - a_i)^2}{8} \right] \\
&= \exp \left[ \frac{-2\epsilon^2}{\sum_{i=1}^m (b_i - a_i)^2} \right], \tag{2.3.2}
\end{aligned}$$

where the last step follows by choosing

$$s = \frac{4\epsilon}{\sum_{i=1}^m (b_i - a_i)^2}.$$

This completes the proof. ■

Suppose  $f : X \rightarrow [0, 1]$  is measurable with respect to the  $\sigma$ -algebra  $\mathcal{S}$ , and that  $P$  is a probability measure on  $(X, \mathcal{S})$ . Then

$$E_P(f) := \int_X f(x) P(dx)$$

is the **expected value** or **mean** of the function  $f$ . Now suppose  $x_1, \dots, x_m$  are i.i.d. samples drawn from  $X$  in accordance with  $P$ , and define

$$\hat{E}(f; \mathbf{x}) := \frac{1}{m} \sum_{i=1}^m f(x_i),$$

where  $\mathbf{x} = [x_1 \dots x_m]^t \in X^m$ . Then  $\hat{E}(f; \mathbf{x})$  is called the **empirical mean** of the function  $f$  corresponding to the multisample  $x_1, \dots, x_m$ . Now  $f(x_1) - E_P(f), \dots, f(x_m) - E_P(f)$  are all zero mean random variables, to which Hoeffding's inequality can be applied. This leads to the following very useful bounds:

$$\begin{aligned}
P^m \{ \mathbf{x} \in X^m : \hat{E}(f; \mathbf{x}) - E_P(f) \geq \epsilon \} &\leq \exp(-2m\epsilon^2), \\
P^m \{ \mathbf{x} \in X^m : \hat{E}(f; \mathbf{x}) - E_P(f) \leq -\epsilon \} &\leq \exp(-2m\epsilon^2), \\
P^m \{ \mathbf{x} \in X^m : |\hat{E}(f; \mathbf{x}) - E_P(f)| \geq \epsilon \} &\leq 2 \exp(-2m\epsilon^2).
\end{aligned}$$



## 2.4 Stochastic Processes, Almost Sure Convergence

In this section, we introduce the technical tools needed to conclude that various stochastic processes converge *almost surely*, as opposed to merely converging in probability. This section may be omitted by readers who are not very interested in such nuances. In that case, they should also skip through the references to almost sure convergence in subsequent chapters.

### 2.4.1 Probability Measures on Infinite Cartesian Products

Suppose  $(X, \mathcal{S}, P)$  is a probability space. In the sequel, we shall often encounter situations where we would like to study the probability of (sets of) *infinite sequences*  $\{x_1, x_2, \dots\}$  where  $x_i \in X$  for each  $i$ . The machinery in this subsection is intended to enable us to do so.

Let  $(X, \mathcal{S})$  be a given measurable space, and let  $\mathbf{N}$  denote the set  $\{1, 2, \dots\}$  of natural numbers. We begin by defining a measurable space whose underlying set is the (countably) infinite Cartesian product  $X^\infty$ , consisting of all sequences of the form  $\{x_i\}_{i \geq 1}$  where  $x_i \in X$  for each  $i$ . A **cylinder set**  $A \subseteq X^\infty$  is a set of the form  $\prod_{i=1}^{\infty} A_i$  where  $A_i \in \mathcal{S}$  for all  $i$ , and in addition,  $A_i = X$  for all except a finite number of indices  $i$ . Let  $\mathcal{S}^\infty$  denote the smallest  $\sigma$ -algebra on  $X^\infty$  that contains all the cylinder sets. Suppose now that  $P$  is a probability measure on  $(X, \mathcal{S})$ ; it is possible to define a corresponding probability measure  $P^\infty$  on the space  $(X^\infty, \mathcal{S}^\infty)$ . Given any cylinder set  $A = \prod_{i=1}^{\infty} A_i \subseteq X^\infty$ , define

$$P^\infty(A) = \prod_{i=1}^{\infty} P(A_i).$$

Observe that  $A_i = X$  for all but a finite number of  $i$ , and as a result  $P(A_i) = 1$  for all but a finite number of indices  $i$ . By the Kolmogorov extension theorem, there exists a *unique* probability measure  $P^\infty$  on  $(X^\infty, \mathcal{S}^\infty)$  that satisfies the above relationship.

### 2.4.2 Stochastic Processes

In this section, a brief introduction is given to the notion of stochastic processes. As is the case elsewhere, the treatment here is strictly minimalist, and the reader is encouraged to consult an authoritative source for a more thorough treatment.

Suppose  $(\Omega, \mathcal{T}, Q)$  is a probability space, and that  $(X, \mathcal{S})$  is a measurable space. Then an  **$X$ -valued stochastic process** is a sequence of  $X$ -valued random variables, of the form  $\{\mathbf{X}\}_{i=-\infty}^{\infty}$ . Note that it is customary in the stochastic process literature to work with two-sided infinite sequences, as opposed to one-sided sequences of the form  $\{\mathbf{X}_i\}_{i=0}^{\infty}$ .

Suppose  $X = \mathbb{R}$  and  $\mathcal{S}$  is the Borel  $\sigma$ -algebra, so that  $\{X_i\}$  is a real-valued stochastic process. For such a process, one can define the multivariate distribution function as follows: Suppose  $k$  is an integer,  $i_1, \dots, i_k$  is an increasing set of  $k$  indices, and  $a_{i_1}, a_{i_k}$  are real numbers. Then

$$P_{X_{i_1}, \dots, X_{i_k}}(a_{i_1}, \dots, a_{i_k}) := Q\{\omega \in \Omega : X_{i_j}(\omega) \leq a_{i_j}, j = 1, \dots, k\}.$$

As in the case of real-valued random variables, one can define two real-valued stochastic processes to be **equivalent** if they have the same multivariate distribution functions, for all multi-indices. Again as in the case of a single random variable, it is possible to change the underlying domain set to something that is a little more natural and easy to work with. This is called the **canonical representation**, and is defined as follows: Given a measurable space  $(X, \mathcal{S})$ , define the corresponding infinite cartesian product space and product  $\sigma$ -algebra  $(X^\infty, \mathcal{S}^\infty)$  as above, and let  $\tilde{P}$  denote a probability measure on  $(X^\infty, \mathcal{S}^\infty)$ . Then we define the canonical representation of an  $X$ -valued stochastic process as a measurable mapping  $\mathbf{X}$  from the probability space  $(X^\infty, \mathcal{S}^\infty, \tilde{P})$  into  $(X^\infty, \mathcal{S}^\infty)$ . Thus if  $\omega$  denotes a typical element of  $X^\infty$ , the image  $\mathbf{X}(\omega)$  is a *sequence*  $X_i(\omega)$  where each element belongs to  $X$ . We define the map  $\omega \mapsto X_i(\omega)$  as the **coordinate random variable**.

The stochastic process  $\{\mathbf{X}\}$  is said to be **stationary** if the probability measure is shift-invariant. This means that, for every finite set of indices  $i_1, \dots, i_l$ , the marginal probability of the  $l$ -tuple  $(X_{i_1}, \dots, X_{i_l})$  is the same as that of  $(X_{i_1+1}, \dots, X_{i_l+1})$ .

The stochastic process is said to be **i.i.d. (independent and identically distributed)** if the coordinate random variables are pairwise independent, and all have the same one-dimensional marginal probability measure.

### 2.4.3 The Borel-Cantelli Lemma and Almost Sure Convergence

Suppose  $(\Omega, \mathcal{T}, Q)$  is a probability space, and let  $\{f_m\}_{m \geq 1}$  be (the top half of) a stochastic process on  $(\Omega, \mathcal{T}, Q)$ . Thus, for each  $m$ ,  $f_m : \Omega \rightarrow \mathbb{R}$  and is measurable. Suppose  $g : \Omega \rightarrow \mathbb{R}$  is measurable. We say that  $\{f_m\}$  **converges to  $g$  in probability** if

$$Q\{\omega \in \Omega : |f_m(\omega) - g(\omega)| > \epsilon\} \rightarrow 0 \text{ as } m \rightarrow \infty, \forall \epsilon > 0.$$

We say that  $\{f_m\}$  **converges almost surely to  $g$**  if

$$Q\{\omega \in \Omega : \lim_{m \rightarrow \infty} f_m(\omega) = g(\omega)\} = 1.$$

It is easy to see that almost sure convergence implies convergence in probability; however, the converse is not true in general. On the other hand, convergence in probability is often much easier to prove than almost sure convergence. A common method of deducing almost sure convergence from “sufficiently fast” convergence in probability is to appeal to the result below, which is known as the **Borel-Cantelli Lemma**.

**Lemma 2.9.** *Suppose  $(\Omega, \mathcal{T}, Q)$  is a probability space, and let  $\{A_m\}$  be a sequence of sets in  $\mathcal{T}$ . Define*

$$B = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n.$$

*Suppose*

$$\sum_{m=1}^{\infty} Q(A_m) < \infty.$$

*Then  $Q(B) = 0$ .*

**Remarks** Note that a point  $\omega \in \Omega$  belongs to  $B$  if and only if it belongs to *infinitely many* sets  $A_m$ , that is, for each  $m \geq 1$  there exists an  $n \geq m$  such that  $\omega \in A_n$ . The point of the lemma is that, if the measures of the sets  $A_m$  decrease sufficiently rapidly that the sequence  $\{Q(A_m)\}$  is summable, then the set of points that belong to infinitely many  $A_m$  has measure zero.

*Proof.* Define

$$B_m = \bigcup_{n=m}^{\infty} A_n, \quad m \geq 1.$$

Then, by the subadditivity property of  $Q$ , it follows that

$$Q(B_m) \leq \sum_{n=m}^{\infty} Q(A_n) \rightarrow 0 \text{ as } m \rightarrow \infty,$$

because of the assumption that  $\sum_{m=1}^{\infty} Q(A_m) < \infty$ . Now note that  $B = \bigcap_{m=1}^{\infty} B_m$ . Consequently we have

$$Q(B) \leq Q(B_m) \quad \forall m \Rightarrow Q(B) \leq \inf_{m \geq 1} Q(B_m) = 0.$$

This completes the proof. ■

The Borel-Cantelli lemma leads at once to the following sufficient condition for almost sure convergence.

**Lemma 2.10.** *Suppose  $(\Omega, \mathcal{T}, Q)$  is a probability space, that  $\{f_m\}_{m \geq 1}$  is a stochastic process on  $(\Omega, \mathcal{T}, Q)$ , and that  $g : \Omega \rightarrow \mathbb{R}$  is measurable. Finally, suppose*

$$\sum_{m=1}^{\infty} Q\{\omega \in \Omega : |f_m(\omega) - g(\omega)| > \epsilon\} < \infty, \quad \forall \epsilon > 0.$$

*Then  $\{f_m\}$  converges almost surely to  $g$ .*

**Remarks:** The hypothesis of the lemma states that the stochastic process  $\{f_m(\cdot)\}$  converges in probability to  $g(\cdot)$  *sufficiently rapidly* that the sequence  $\{q_{m,\epsilon}\}$  is summable for each  $\epsilon > 0$ , where

$$q_{m,\epsilon} := Q\{\omega \in \Omega : |f_m(\omega) - g(\omega)| > \epsilon\}.$$

*Proof.* For each positive integer  $k$ , define the set

$$B_k = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} \{\omega \in \Omega : |f_n(\omega) - g(\omega)| > 1/k\}.$$

Applying the Borel-Cantelli lemma with

$$A_{m,k} = \{\omega \in \Omega : |f_m(\omega) - g(\omega)| > 1/k\},$$

we conclude that  $Q(B_k) = 0$  for all  $k$ . Hence

$$Q\left(\bigcup_{k=1}^{\infty} B_k\right) = 0,$$

by the countable subadditivity property of  $Q$ . Now note that

$$\bigcup_{k=1}^{\infty} B_k = \{\omega \in \Omega : f_m(\omega) \not\rightarrow g(\omega)\}.$$

Therefore  $\{f_m\}$  converges almost surely to  $g$ . ■

*Example 2.1.* As an application of this lemma, let us return to the problem of empirically estimating the expected value (i.e., mean) of a function based upon i.i.d. samples. Suppose  $(X, \mathcal{S}, P)$  is a probability space, and that  $f : X \rightarrow [0, 1]$  is measurable. As before, let

$$E_P(f) := \int_X f(x) P(dx)$$

denote the expected value of  $f$ . Now suppose  $x_1, \dots, x_m$  are i.i.d. samples drawn from  $X$  in accordance with  $P$ , and define

$$\hat{E}(f; \mathbf{x}_m) := \frac{1}{m} \sum_{i=1}^m f(x_i).$$

It is now shown that  $\hat{E}(f; \mathbf{x}_m)$  converges almost surely to  $E_P(f)$  in a sense to be made precise next.

Let  $\mathbf{x}^* \in X^\infty$ ; thus  $\mathbf{x}^*$  is a sequence  $\{x_i\}_{i \geq 1}$  where each  $x_i \in X$ . Now one can define  $\hat{E}_m(f; \mathbf{x}^*)$  as the random variable mapping  $X^\infty$  into  $[0, 1]$  according to

$$\hat{E}_m(f; \mathbf{x}^*) := \frac{1}{m} \sum_{i=1}^m f(x_i) = \hat{E}(f; \mathbf{x}_m).$$

Note that  $\hat{E}_m(f; \mathbf{x}^*)$  depends only on the first  $m$  components of  $\mathbf{x}^*$ . Now Hoeffding's inequality states that

$$P^m\{\mathbf{x} \in X^m : |\hat{E}_m(f; \mathbf{x}_m) - E_P(f)| > \epsilon\} \leq 2 \exp(-2m\epsilon^2).$$

One can recast this as

$$P^\infty\{\mathbf{x}^* \in X^\infty : |\hat{E}_m(f; \mathbf{x}^*) - E_P(f)| > \epsilon\} \leq 2 \exp(-2m\epsilon^2).$$

Since the sequence  $\{2 \exp(-2m\epsilon^2)\}_{m \geq 1}$  is summable for each  $\epsilon > 0$ , it follows from Lemma 2.10 that the sequence of random variables  $\{\hat{E}_m(f; \cdot)\}$  converges almost surely to  $E_P(f)$  (or more precisely, to the “random” variable whose value equals  $E_P(f)$  for all  $\mathbf{x}^* \in X^\infty$ ). This means that

$$P^\infty\{\mathbf{x}^* \in X^\infty : \hat{E}_m(f; \mathbf{x}^*) \rightarrow E_P(f)\} = 1.$$

This property is known as the “strong law of large numbers.”

As a very useful application of the above property, suppose  $A \in \mathcal{S}$  is a measurable set, and let  $f = I_A(\cdot)$ , the indicator function of the set  $A$ . Then it is easy to see that  $E_P(f)$  is the same as  $P(A)$ . Moreover, given an infinite sequence  $\mathbf{x}^* \in X^\infty$ , one can define the random variable  $\hat{P}_m(A; \mathbf{x}^*)$  by

$$\hat{P}_m(A; \mathbf{x}^*) := \frac{1}{m} \sum_{i=1}^m I_A(x_i).$$

Note that  $\hat{P}_m(A; \mathbf{x}^*)$  is just the fraction of the first  $m$  samples that belong to the set  $A$ . One can think of  $\hat{P}_m(A; \mathbf{x}^*)$  as an empirical estimate of the probability of the set  $A$ , based on the first  $m$  elements of the sequence  $\mathbf{x}^*$ . By the preceding argument, it follows that

$$P^\infty\{\mathbf{x}^* \in X^\infty : \hat{P}_m(A; \mathbf{x}^*) \rightarrow P(A)\} = 1.$$

## 2.5 Mixing Properties of Stochastic Processes

This section is devoted to a discussion of an advanced notion called the “mixing” of stochastic processes. Up to now (as in Example 2.1 for instance), we have dealt with i.i.d. processes. Indeed, much of the “classical” form of statistical learning theory is couched in terms of i.i.d. processes. However, independence is a very restrictive concept, in several ways. First, it is often an assumption, rather than a deduction on the basis of observations. Second, it is an “all or nothing” property, in the sense that two random variables are either independent or they are not – the definition does not permit an intermediate notion of being “nearly” independent. As a result, many of the proofs based on the assumption that the underlying stochastic process is i.i.d. are rather “fragile.” The notion of mixing allows one to put the notion of “near independence” on a firm mathematical foundation, and moreover, permits one to derive a “robust” rather than a “fragile” theory, by allowing one to prove that most of the desirable properties of i.i.d. stochastic processes are preserved when the underlying process is mixing.

### 2.5.1 Definitions of Various Kinds of Mixing Coefficients

There are several diverse notions of mixing used in the literature, but we shall be concerned with only two, namely  $\alpha$ -mixing and  $\beta$ -mixing. In the interests of completeness, we also define one more notion called  $\phi$ -mixing, but we also show why this is not a very useful concept, at least in learning theory.

To define these concepts, let us begin with a stationary stochastic process  $\{\mathbf{X}_i\}_{i=-\infty}^{\infty}$  defined on a probability space  $(X^\infty, \mathcal{S}^\infty, \tilde{P})$ . It is assumed that a canonical representation is used for the stochastic process, so that each  $\mathbf{X}_i$  maps  $(X^\infty, \mathcal{S}^\infty, \tilde{P})$  into  $X$ . For each index  $k$ , let  $\Sigma_{-\infty}^k$  denote the  $\sigma$ -algebra generated by the coordinate random variables  $\mathbf{X}_i, i \leq k$ , and similarly let  $\Sigma_k^\infty$  denote the  $\sigma$ -algebra generated by the coordinate random variables  $\mathbf{X}_i, i \geq k$ . Let  $\tilde{P}_{-\infty}^k$  and  $\tilde{P}_k^\infty$  denote the corresponding marginal probability measures. Then, by the Kolmogorov extension theorem, there exists a unique probability measure on  $(X^\infty, \mathcal{S}^\infty)$ , denoted by  $\tau_0(\tilde{P})$ , such that

1. The laws of  $\{\mathbf{X}_i, i \leq 0\}$  under  $\tilde{P}$  and under  $\tau_0(\tilde{P})$  are the same.
2. The laws of  $\{\mathbf{X}_j, j \geq 1\}$  under  $\tilde{P}$  and under  $\tau_0(\tilde{P})$  are the same.
3. Under the measure  $\tau_0(\tilde{P})$ , the variables  $\{\mathbf{X}_i, i \leq 0\}$  are independent of  $\{\mathbf{X}_j, j \geq 1\}$ . This means that each  $\mathbf{X}_i, i \leq 0$  is independent of each  $\mathbf{X}_j, j \geq 1$ .

Some authors denote this new probability measure by the symbol  $\tilde{P}_{-\infty}^0 \times \tilde{P}_1^\infty$ . However, in the proofs it is more convenient to use the symbol  $\tau_0(\tilde{P})$ , where the subscript 0 serves to remind us of the place at which the two halves of the stochastic process are “split.” To make the present theorem statements resemble those found in the literature, the two symbols  $\tilde{P}_{-\infty}^0 \times \tilde{P}_1^\infty$  and  $\tau_0(\tilde{P})$  are used interchangeably. For future use, let us also introduce the symbol  $\bar{\Sigma}_1^{k-1}$  to denote the  $\sigma$ -algebra generated by the random variables  $\mathbf{X}_i, i \leq 0$  as well as  $\mathbf{X}_j, j \geq k$ . Thus the bar over the  $\Sigma$  serves to remind us that the random variables between 1 and  $k-1$  are missing from the list of variables that generate  $\Sigma$ .

With this notation, we can now define various mixing coefficients.

**Definition 2.1.** *The  $\alpha$ -mixing coefficient of the stochastic process  $\{\mathbf{X}_i\}$  is defined as*

$$\alpha(k) := \sup_{A \in \Sigma_{-\infty}^0, B \in \Sigma_k^\infty} |\tilde{P}(A \cap B) - \tilde{P}(A) \cdot \tilde{P}(B)|. \quad (2.5.1)$$

*The  $\beta$ -mixing coefficient of the stochastic process is defined as*

$$\begin{aligned} \beta(k) &:= \sup_{C \in \bar{\Sigma}_1^{k-1}} |\tilde{P}(C) - (\tilde{P}_{-\infty}^0 \times \tilde{P}_1^\infty)(C)| \\ &= \rho(\tilde{P}, \tau_0(\tilde{P}); \bar{\Sigma}_1^{k-1}). \end{aligned} \quad (2.5.2)$$

*The  $\phi$ -mixing coefficient of the stochastic process is defined as*

$$\phi(k) := \sup_{A \in \Sigma_{-\infty}^0, B \in \Sigma_k^\infty} |\tilde{P}(B|A) - \tilde{P}(B)|. \quad (2.5.3)$$

In the definition of the  $\alpha$ -mixing coefficient,  $A$  is an event that depends only on the “past” random variables  $\{X_i, i \leq 0\}$  while  $B$  is an event that depends only on the “future” random variables  $\{X_i, i \geq k\}$ . Thus if the future event  $B$  were to be truly independent of the past event  $A$ , then the probability  $\tilde{P}(A \cap B)$  would exactly equal  $\tilde{P}(A)\tilde{P}(B)$ . Thus the  $\alpha$ -mixing coefficient measures how near to independence future events are of the past events, by taking the supremum of the difference between the two quantities  $\tilde{P}(A \cap B)$  and  $\tilde{P}(A)\tilde{P}(B)$ . Similarly, if the future event  $B$  were to be truly independent of the past event  $A$ , then the conditional probability  $\tilde{P}(B|A)$  would exactly equal the unconditional probability  $\tilde{P}(B)$ . The  $\phi$ -mixing coefficient measures how near to independence future events are of the past events, by taking the supremum of the difference between the two quantities  $\tilde{P}(B|A)$  and  $\tilde{P}(B)$ . The  $\beta$ -mixing coefficient has a somewhat more involved interpretation. If the future events beyond time  $k$  were to be truly independent of the past events before time 0, then the probability measure  $\tilde{P}$  would exactly equal the “split” measure  $\tau_0(\tilde{P})$ , or  $\tilde{P}_{-\infty}^0 \times \tilde{P}_k^\infty$  as some authors write it. The  $\beta$ -mixing coefficient thus measures how nearly the product measure approximates the actual measure  $\tilde{P}$ .

Now a few properties of these coefficients are discussed.

1. Note that, in the definitions of  $\alpha(k)$  and  $\phi(k)$ , we can write  $\tilde{P}_{-\infty}^0(A)$  instead of  $\tilde{P}(A)$ ; similarly we can write  $\tilde{P}_1^\infty(B)$  instead of  $\tilde{P}(B)$ .
2. In the definition of the  $\phi$ -mixing coefficient, the conditional probability  $\tilde{P}(B|A)$  is taken as  $\tilde{P}(B)$  if  $\tilde{P}(A) = 0$ .
3. Since  $\Sigma_{k+1}^\infty \subseteq \Sigma_k^\infty$ , it is obvious that the  $\alpha$ -,  $\beta$ - and  $\phi$ -mixing coefficients are all nonincreasing. Thus

$$\alpha(k+1) \leq \alpha(k), \quad \beta(k+1) \leq \beta(k), \quad \phi(k+1) \leq \phi(k), \quad \forall k.$$

4. If we write  $C = A \cap B$  where  $A \in \Sigma_{-\infty}^0$  and  $B \in \Sigma_k^\infty$ , then we have

$$(P_{-\infty}^0 \times P_1^\infty)(A \cap B) = P_{-\infty}^0(A) \cdot P_1^\infty(B).$$

Thus, when  $C$  is restricted to intersections of the above type, the right sides of (2.5.1) and 2.5.2) coincide. However, in (2.5.2), the supremum is taken not only over sets  $C$  of the form  $A \cap B$ , but over the  $\sigma$ -algebra generated by all such intersections. Thus it follows that

$$\alpha(k) \leq \beta(k), \quad \forall k \geq 1.$$

Similarly, since  $\tilde{P}(B|A) = \tilde{P}(A \cap B)/\tilde{P}(A)$  if  $\tilde{P}(A) \neq 0$ , it is easy to see that

$$\alpha(k) \leq \phi(k), \quad \forall k \geq 1.$$

It can also be shown that

$$\beta(k) \leq \phi(k), \quad \forall k \geq 1.$$

5. It is obvious that if the stochastic process  $\{X_i\}$  consists of independent and identically distributed (i.i.d.) random variables, then  $\tilde{P}$  equals the measure  $(\tilde{P}_0)^\infty$ , which denotes the measure on  $(X^\infty, \mathcal{S}^\infty)$  under which each  $X_i$  has marginal probability  $\tilde{P}_0$ , and the  $X_i$ 's are pairwise independent. In such a case, all the three mixing coefficients are zero, for each  $k$ .
6. It is somewhat ironic that some authors refer to  $\alpha$ -mixing as “strong” mixing, even though it is the weakest of the various notions of mixing studied in the literature.

**Definition 2.2.** *The stochastic process  $\{X_i\}$  is said to be  $\alpha$ -mixing, or strongly regular if  $\alpha(k) \rightarrow 0$  as  $k \rightarrow \infty$ . The stochastic process  $\{X_i\}$  is said to be  $\beta$ -mixing or completely regular if  $\beta(k) \rightarrow 0$  as  $k \rightarrow \infty$ . The stochastic process  $\{X_i\}$  is said to be  $\phi$ -mixing or uniformly regular if  $\phi(k) \rightarrow 0$  as  $k \rightarrow \infty$ .*

### 2.5.2 Inequalities for Mixing Processes

In this subsection, a few consequences of mixing processes are derived. We begin with  $\beta$ -mixing processes.

**Lemma 2.11.** *Suppose  $\{X_i\}$  is a  $\beta$ -mixing process on a probability space  $(X^\infty, \mathcal{S}^\infty, \tilde{P})$ . Suppose  $f : X^\infty \rightarrow \mathbb{R}$  is essentially bounded and is measurable with respect to the  $\sigma$ -algebra  $\tilde{\Sigma}_1^{k-1} = \Sigma(X_i, i \leq 0 \text{ or } i \geq k)$ . Then*

$$|E(f, \tilde{P}) - E(f, P_{-\infty}^0 \times P_1^\infty)| \leq \beta(k) \cdot \|f\|_\infty. \quad (2.5.4)$$

*Proof.* Note that, by definition, the quantity  $\beta(k)$  is precisely the total variation metric between the two probability measures  $\tilde{P}$  and  $P_{-\infty}^0 \times P_1^\infty$ . The desired inequality now follows readily. ■

**Theorem 2.1.** *Suppose  $\{X_i\}$  is a  $\beta$ -mixing process on a probability space  $(X^\infty, \mathcal{S}^\infty, \tilde{P})$ . Suppose  $f : X^\infty \rightarrow \mathbb{R}$  is essentially bounded and depends only on the variables  $x_{ik}, 0 \leq i \leq l$ . Let  $\tilde{P}_0$  denote the one-dimensional marginal probability of each of the  $X_i$ . Then*

$$|E(f, \tilde{P}) - E(f, \tilde{P}_0^\infty)| \leq l\beta(k) \|f\|_\infty. \quad (2.5.5)$$

The proof of theorem depends on the following auxiliary lemma.

**Lemma 2.12.** *Suppose  $P, Q$  are probability measures on  $(U, \mathcal{S})$ , that  $X, Y$  are measurable real-valued functions on  $(U, \mathcal{S})$ . Suppose further that (i)  $X, Y$  are independent random variables under each of  $P, Q$ , and (ii) the marginal probabilities of  $X$  under  $P$  and  $Q$  are equal. Then*

$$\rho(P, Q; \Sigma(X, Y)) = \rho(P, Q; \Sigma(Y)).$$



*Proof. (of the lemma):* Consider a set  $S \in \Sigma(\mathbf{X}, \mathbf{Y})$  of the form

$$S = \bigcup_{i=1}^{\infty} (A_i \cap B_i), \quad A_i \in \Sigma(\mathbf{X}), \quad B_i \in \Sigma(\mathbf{Y}), \quad \forall i,$$

where the  $A_i$  are pairwise disjoint. As in the case of Lemma 2.5,  $\rho(P, Q; \Sigma(\mathbf{X}, \mathbf{Y}))$  equals the supremum of  $|P(S) - Q(S)|$  as  $S$  varies over all such sets. Now

$$|P(S) - Q(S)| \leq \sum_{i=1}^{\infty} |P(A_i \cap B_i) - Q(A_i \cap B_i)|.$$

Next, since  $\mathbf{X}, \mathbf{Y}$  are independent under both  $P$  and  $Q$ , we have

$$P(A_i \cap B_i) = P(A_i)P(B_i), \quad Q(A_i \cap B_i) = Q(A_i)Q(B_i).$$

Moreover, since the marginal probabilities of  $\mathbf{X}$  under  $P$  and  $Q$  are equal, we have

$$P(A_i) = Q(A_i).$$

Therefore it follows that

$$\begin{aligned} |P(S) - Q(S)| &\leq \sum_{i=1}^{\infty} P(A_i) |P(B_i) - Q(B_i)| \\ &\leq \left( \sum_{i=1}^{\infty} P(A_i) \right) \rho(P, Q; \Sigma(\mathbf{Y})) \\ &\leq \rho(P, Q; \Sigma(\mathbf{Y})). \end{aligned}$$

This shows that  $\rho(P, Q; \Sigma(\mathbf{X}, \mathbf{Y})) \leq \rho(P, Q; \Sigma(\mathbf{Y}))$ . The opposite inequality follows readily since  $\Sigma(\mathbf{X}, \mathbf{Y})$  is a superset of  $\Sigma(\mathbf{Y})$ . ■

*Proof. of the theorem:* Recall the notation  $\tau_0(\tilde{P})$ , which “splits” the original measure  $\tilde{P}$  into a kind of product measure, in such a way that the variables  $\mathbf{X}_i, i \leq 0$  are independent of  $\mathbf{X}_i, i \geq 1$ , and both sets of variables have the same marginals as under  $\tilde{P}$ . Since we shall be “splitting” measures repeatedly, to make the notation less cumbersome let us define the symbol  $\tilde{Q}$  recursively as follows:

$$\tilde{Q}_1 = \tau_0(\tilde{P}), \quad \tilde{Q}_{i+1} = \tau_{ik}(\tilde{Q}_i), \quad i = 1, \dots, l-1.$$

Thus  $\tilde{Q}_l$  consists of the original measure  $\tilde{P}$  split at the time instants  $0, k, 2k, \dots, (l-1)k$ . In the interests of brevity, let us use the symbol  $\Sigma$  to denote the  $\sigma$ -algebra  $\Sigma(\mathbf{X}_{ik}, 0 \leq i \leq l)$ . We shall display it in full form when needed. Now the claim is that

$$\rho(\tilde{P}, \tilde{Q}_l; \Sigma) \leq l\beta(k). \quad (2.5.6)$$

To establish this claim, note that by the triangle inequality we have

$$\rho(\tilde{P}, \tilde{Q}_l; \Sigma) \leq \rho(\tilde{P}, \tilde{Q}_1; \Sigma) + \sum_{i=1}^{l-1} \rho(\tilde{Q}_i, \tilde{Q}_{i+1}; \Sigma). \quad (2.5.7)$$

Then (2.5.6) will follow if it can be shown that each of the above terms on the right side is less than  $\beta(k)$ . Note that, since the stochastic process is stationary, the  $\beta$ -mixing coefficient also equals

$$\beta(k) = \rho(\tilde{P}, \tau_j(\tilde{P}); \bar{\Sigma}_{j+1}^{j+k-1}).$$

In other words, the original measure  $\tilde{P}$  can be “split” at *any* time instant  $j$ , and the total variation distance does not depend on  $j$ . With this background, note first that

$$\rho(\tilde{P}, \tilde{Q}_1; \Sigma) \leq \rho(\tilde{P}, \tau_0(\tilde{P}); \bar{\Sigma}_1^{k-1}) \leq \beta(k),$$

since  $\tilde{Q}_1 = \tau_0(\tilde{P})$  and  $\Sigma$  is a subalgebra of  $\bar{\Sigma}_1^{k-1}$ . For the remaining terms, recall that  $\tilde{Q}_{i+1}$  is obtained by splitting  $\tilde{Q}_i$  at the time instant  $ik$ . Thus, under both  $\tilde{Q}_i$  and  $\tilde{Q}_{i+1}$ , the variables  $\mathbf{X}_0, \mathbf{X}_k, \dots, \mathbf{X}_{(i-1)k}$  are independent of  $\mathbf{X}_{(i+1)k}, \dots, \mathbf{X}_{lk}$ . Moreover, the marginal probabilities of  $\{\mathbf{X}_0, \dots, \mathbf{X}_{(i-1)k}\}$  are the same under both  $\tilde{Q}_i$  and  $\tilde{Q}_{i+1}$ . Hence, by Lemma 2.12, it follows that

$$\rho(\tilde{Q}_i, \tilde{Q}_{i+1}; \Sigma) \leq \rho(\tilde{Q}_i, \tilde{Q}_{i+1}; \Sigma(\mathbf{X}_{ik}, \dots, \mathbf{X}_{lk})).$$

But on this algebra,  $\tilde{Q}_i$  equals  $\tilde{P}$ , while  $\tilde{Q}_{i+1}$  equals  $\tau_{ik}(\tilde{P})$ . Moreover, this algebra is a subalgebra of  $\bar{\Sigma}_{ik+1}^{(i+1)k-1}$ . Hence, by definition, it follows that

$$\rho(\tilde{Q}_i, \tilde{Q}_{i+1}; \Sigma(\mathbf{X}_{ik}, \dots, \mathbf{X}_{lk})) \leq \rho(\tilde{P}, \tau_{ik}(\tilde{P}); \bar{\Sigma}_{ik+1}^{(i+1)k-1}) = \beta(k).$$

Hence each of the terms on the right side of (2.5.7) is less than  $\beta(k)$ . This establishes (2.5.6). Consequently, it follows that

$$|E(f, \tilde{P}) - E(f, \tilde{Q}_l)| \leq l\beta(k) \|f\|_\infty.$$

To complete the proof, note that under the measure  $\tilde{Q}_l$ , the  $l+1$  variables  $\mathbf{X}_0, \mathbf{X}_k, \dots, \mathbf{X}_{lk}$  are pairwise independent. Hence

$$E(f, \tilde{Q}_l) = E(f, \tilde{P}_0^\infty).$$

This completes the proof. ■

**Corollary 2.1.** *Suppose  $i_0 < i_1 < \dots < i_l$  are integers, and define*

$$k := \min_{0 \leq j \leq l-1} i_{j+1} - i_j.$$

*Suppose  $f$  is essentially bounded and depends only on  $\mathbf{X}_{i_0}, \dots, \mathbf{X}_{i_l}$ . Then*

$$|E(f, \tilde{P}) - E(f, \tilde{P}_0^\infty)| \leq \beta(k) \|f\|_\infty.$$

The proof is a routine modification of that of Theorem 2.1 and is left to the reader.

Now we present an inequality pertaining to  $\alpha$ -mixing processes.

**Theorem 2.2.** *Suppose  $\{\mathbf{X}_i\}$  is an  $\alpha$ -mixing process on a probability space  $(X^\infty, \mathcal{S}^\infty, \tilde{P})$ . Suppose  $f, g : X^\infty \rightarrow \mathbb{R}$  are essentially bounded, that  $f$  is measurable with respect to  $\Sigma(\mathbf{X}_i, i \leq 0)$ , and that  $g$  is measurable with respect to  $\Sigma(\mathbf{X}_i, i \geq k)$ . Then*

$$|E(fg, \tilde{P}) - E(f, \tilde{P}) E(g, \tilde{P})| \leq 4\alpha(k) \|f\|_\infty \cdot \|g\|_\infty. \quad (2.5.8)$$

Theorem 2.2 is a ready consequence of the following more general result.

**Lemma 2.13.** *Suppose  $f, g$  are essentially bounded random variables on a probability space  $(\Omega, \mathcal{T}, \mathcal{P})$ , that  $\mathcal{F}, \mathcal{G}$  are sub  $\sigma$ -algebras of  $\mathcal{T}$ , and that  $f \in \mathcal{M}(\mathcal{F})$ ,  $g \in \mathcal{M}(\mathcal{G})$ . Define the coefficient*

$$\alpha(\mathcal{F}, \mathcal{G}) := \sup_{A \in \mathcal{F}, B \in \mathcal{G}} |P(A \cap B) - P(A)P(B)|.$$

Then

$$|E(fg) - E(f)E(g)| \leq 4 \|f\|_\infty \|g\|_\infty \alpha(\mathcal{F}, \mathcal{G}).$$

*Proof.* Define  $\eta := \text{sign}(g_{\mathcal{F}} - E(g))$ . Thus

$$\eta(\omega) := \begin{cases} 1 & \text{if } g_{\mathcal{F}}(\omega) - E(g) \geq 0, \\ 0 & \text{if } g_{\mathcal{F}}(\omega) - E(g) < 0. \end{cases}$$

Then we have  $|g_{\mathcal{F}} - E(g)| = \eta(g_{\mathcal{F}} - E(g))$ . Now by (2.2.1) we have that  $E(fg) = E(fg_{\mathcal{F}})$ . Therefore

$$\begin{aligned} |E(fg) - E(f)E(g)| &= |E(fg_{\mathcal{F}}) - E[fE(g)]| \\ &= |E[f(g_{\mathcal{F}} - E(g))]| \\ &\leq \|f\|_\infty E[|g_{\mathcal{F}} - E(g)|] \\ &= \|f\|_\infty E[\eta(g_{\mathcal{F}} - E(g))] \\ &= \|f\|_\infty E[\eta(g - E(g))], \end{aligned}$$

where in the last step we use the fact that  $\eta \in \mathcal{M}(\mathcal{F})$  so that  $E(\eta g) = E(\eta g_{\mathcal{F}})$  by (2.2.1). Now define  $\xi = \text{sign}(g - E(g))$  and note that  $\xi \in \mathcal{M}(\mathcal{G})$ . Then it is possible to mimic the above argument and show that

$$|E[\eta(g - E(g))]| \leq \|g\|_\infty |E(\eta\xi) - E(\eta)E(\xi)|.$$

Combining this with the preceding inequality shows that

$$|E(fg) - E(f)E(g)| \leq \|f\|_\infty \|g\|_\infty |E(\eta\xi) - E(\eta)E(\xi)|.$$

The proof is therefore complete if it can be shown that

$$|E(\eta\xi) - E(\eta)E(\xi)| \leq 4\alpha(\mathcal{F}, \mathcal{G}). \quad (2.5.9)$$

Towards this end, define the four sets

$$\begin{aligned} A_+ &:= \{\omega : \eta(\omega) = 1\}, \quad A_- := \{\omega : \eta(\omega) = -1\}, \\ B_+ &:= \{\omega : \xi(\omega) = 1\}, \quad B_- := \{\omega : \xi(\omega) = -1\}, \end{aligned}$$

and note that  $A_+, A_- \in \mathcal{F}, B_+, B_- \in \mathcal{G}$ . Now it is easy to see that

$$\begin{aligned} E(\eta\xi) &= P(A_+ \cap B_+) + P(A_- \cap B_-) - P(A_+ \cap B_-) - P(A_- \cap B_+), \\ E(\eta) &= P(A_+) - P(A_-), \quad E(\xi) = P(B_+) - P(B_-). \end{aligned}$$

Substituting all this shows that

$$\begin{aligned} |E(\eta\xi) - E(\eta)E(\xi)| &= P(A_+ \cap B_+) + P(A_- \cap B_-) \\ &\quad - P(A_+ \cap B_-) - P(A_- \cap B_+) \\ &\quad - P(A_+)P(B_+) - P(A_-)P(B_-) \\ &\quad + P(A_+)P(B_-) + P(A_-)P(B_+). \end{aligned}$$

Now by the definition of  $\alpha(\mathcal{F}, \mathcal{G})$  we have

$$|P(A_+ \cap B_+) - P(A_+)P(B_+)| \leq \alpha(\mathcal{F}, \mathcal{G}),$$

and similarly for the other three terms. This finally leads to the desired bound (2.5.9) and completes the proof. ■

The next result is a multiplicative analog of Theorem 2.1.

**Corollary 2.2.** *Suppose  $\{X_i\}$  is an  $\alpha$ -mixing stochastic process. Suppose  $f_0, \dots, f_l$  are essentially bounded functions, where  $f_i$  depends only on  $X_{ik}$ . Then*

$$\left| E \left[ \prod_{i=0}^l f_i \right] - \prod_{i=0}^l E(f_i) \right| \leq 4l\alpha(k) \prod_{i=0}^l \|f_i\|_\infty. \quad (2.5.10)$$

The proof by induction on  $l$  is simple and is therefore omitted.

**Notes and References** The material in Sections 2.1 through 2.4 can be found in standard texts. For concepts from topology, see [99], See [106] for a survey of covering numbers, packing numbers, and their interrelationships, as well as explicit computations of these numbers for various specific sets. For concepts from probability, see [66], [73], or [117]. Hoeffding's inequality is proven in [85] and generalizes earlier work of Chernoff [45]. The material in Section 2.5 is more advanced. For definitions of mixing coefficients of stochastic processes, see [25]. For some reason, many texts on stochastic processes discuss only  $\alpha$ -mixing and  $\phi$ -mixing, but not  $\beta$ -mixing. As shown in subsequent chapters,  $\phi$ -mixing is too strong an assumption, while  $\alpha$ -mixing

is too weak; but  $\beta$ -mixing is “just right.” Theorem 2.1 is due to Yu [209], while Theorem 2.2 is due to Ibragimov [87], with the proof reproduced in [77], Theorem A.5. The importance of mixing processes originally arose from the fact that the simple law of large numbers established in Example 2.1 for i.i.d. processes can be readily extended even to  $\alpha$ -mixing processes, which satisfy the weakest type of mixing condition. Several authors have studied conditions under which the output sequence of a Markov chain is mixing in any of the three senses discussed here. We shall return to this topic again in Chapter 3.