LECTURE 2

Review of Basic Probability Theory

2.1 PROBABILITY SPACE AND AXIOMS

Probability theory provides a set of mathematical rules to assign probabilities to outcomes of random experiments, e.g., coin flips, packet arrivals, stock prices, neural spikes, noise voltages, and so on. Given a random experiment, its *sample space* Ω is the set of all outcomes. An *event* is a subset of the sample space and we say that an event $A \subseteq \Omega$ occurs if the outcome ω of the random experiment is an element of A. Let $\mathcal{F}$ be a set of events. A *probability measure* $\mathsf{P} : \mathcal{F} \to [0, 1]$ is a function that assigns probabilities to the events in $\mathcal{F}$. We refer to the triple $(\Omega, \mathcal{F}, \mathsf{P})$ as the *probability space* of the random experiment.

The probability measure P must satisfy the following.

Axioms of probability.

1. $\mathsf{P}(A) \geq 0$ for every event A in $\mathcal{F}$.
2. $\mathsf{P}(\Omega) = 1$.
3. *Countable additivity*. If $A_1, A_2, \ldots$ are *disjoint*, i.e., $A_i \cap A_j = \emptyset$, $i \neq j$, then

$$\mathsf{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathsf{P}(A_i).$$

For the probability measure P to be well-defined over all events of interest, the set of events $\mathcal{F}$ must satisfy:

1. $\emptyset \in \mathcal{F}$.
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
3. If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Due to these defining properties, $\mathcal{F}$ is often referred to as a *σ-algebra* or *σ-field*.

2.2 DISCRETE PROBABILITY SPACES

A probability space $(\Omega, \mathcal{F}, \mathsf{P})$ is said to be *discrete* if the sample space Ω is countable, i.e., finite or countably infinite.

Example 2.1 (Flipping a coin). $\Omega = \{H, T\}$, $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$, and

$$\mathsf{P}(\emptyset) = 0, \quad \mathsf{P}(\{H\}) = p, \quad \mathsf{P}(\{T\}) = 1 - p, \quad \mathsf{P}(\Omega) = 1,$$

where $p \in [0, 1]$ is the *bias* of the coin. A *fair* coin has a bias of $1/2$.

For discrete sample spaces, $\mathcal{F}$ is often the set of all subsets of Ω, namely, the *power set* 2^Ω of Ω. (Recall that $|2^\Omega| = 2^{|\Omega|}$.) In this case, the probability measure P can be fully specified by assigning probabilities to individual outcomes (or *singletons*) $\{\omega\}$ so that

$$\mathsf{P}(\{\omega\}) \geq 0, \quad \omega \in \Omega,$$

and

$$\sum_{\omega \in \Omega} \mathsf{P}(\{\omega\}) = 1.$$

Then it follows by the third axiom of probability that for any event $A \subseteq \Omega$,

$$\mathsf{P}(A) = \sum_{\omega \in A} \mathsf{P}(\{\omega\}).$$

Example 2.2 (Rolling a fair die). $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = 2^\Omega = \{\emptyset, \{1\}, \{2\}, \ldots, \Omega\}$, and

$$\mathsf{P}(\{i\}) = \frac{1}{6}, \quad i = 1, 2, \ldots, 6.$$

The probability of the event A "the outcome is even," i.e., $A = \{2, 4, 6\}$, is

$$\mathsf{P}(A) = \mathsf{P}(\{2\}) + \mathsf{P}(\{4\}) + \mathsf{P}(\{6\}) = \frac{3}{6} = \frac{1}{2}.$$

Example 2.3 (Flipping a coin n times). A coin with bias p is flipped n times. Then

$$\begin{aligned} \Omega &= \{H, T\}^n = \{\text{sequences of heads/tails of length } n\}, \\ \mathcal{F} &= 2^\Omega, \\ \mathsf{P}(\{\omega\}) &= p^i (1 - p)^{n-i}, \end{aligned}$$

where i is the number of heads in ω. The probability of the event A_k "the outcome consists of k heads and $n - k$ tails" is

$$\mathsf{P}(A_k) = \sum_{\omega : \omega \text{ has } k \text{ heads}} \mathsf{P}(\{\omega\}) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

We can verify that

$$\mathsf{P}(\Omega) = \sum_{k=0}^{n} \mathsf{P}(A_k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} = 1.$$

Example 2.4 (Flipping a coin until the first head). $\Omega = \{H, TH, TTH, TTTH, \ldots\}$, $\mathcal{F} = 2^\Omega$, and

$$\mathsf{P}(\{\omega\}) = (1-p)^i p,$$

where i is the number of tails in ω. Again we can verify that

$$\mathsf{P}(\Omega) = \sum_{\omega \in \Omega} \mathsf{P}(\{\omega\}) = \sum_{i=0}^{\infty} (1-p)^i p = 1.$$

Example 2.5 (Counting the number of packets). Consider the number of packets arriving at a node in a communication network in time interval $(0, T]$ at *rate* $\lambda \in (0, \infty)$. Then, $\Omega = \{0, 1, 2, 3, \ldots\}$, $\mathcal{F} = 2^\Omega$, and

$$\mathsf{P}(\{k\}) = \frac{(\lambda T)^k}{k!} e^{-\lambda T}, \quad k = 0, 1, 2, \ldots,$$

provided that the number of packets are *Poisson* distributed. Note that

$$\mathsf{P}(\Omega) = \sum_{k=0}^{\infty} \frac{(\lambda T)^k}{k!} e^{-\lambda T} = 1.$$

In all examples so far, $\mathcal{F} = 2^\Omega$. This is not necessarily the case.

Example 2.6. (Rolling a colored die). Suppose that each face of a die is colored, say, 1 and 2 are red, and 3 through 6 are blue. Further suppose that the observer of a die roll can only note the color of the face, not the actual number. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$ as before, but

$$\mathcal{F} = \{\emptyset, \{1, 2\}, \{3, 4, 5, 6\}, \Omega\}.$$

This is a valid σ-algebra (check!), but it is much smaller in size than the previous case and the probability measure is fully specified by $\mathsf{P}(\{1, 2\})$ alone. As an extreme, if all six faces are of the same color, then we have the *trivial* σ-algebra $\mathcal{F} = \{\emptyset, \Omega\}$, which is still valid but hardly interesting. Thus, the choice of $\mathcal{F}$ controls the level of granularity at which one can assign probabilities.

2.3 CONTINUOUS PROBABILITY SPACES

A *continuous* probability space has an uncountable number of elements in Ω. Unlike the discrete case, the choice of $\mathcal{F} = 2^\Omega$, albeit valid, is too rich to admit an interesting probability measure under the standard axioms of probability. At the same time, specifying probabilities to singletons is not sufficient to extrapolate probabilities for other events. Hence, $\mathcal{F}$ should be chosen more carefully, which is the main reason behind the intricate definitions of probability measure and σ-algebra.

Suppose that Ω is the real line $\mathbb{R}$ or its subinterval, e.g., $[0, 1] = \{x \in \mathbb{R} : 0 \le x \le 1\}$.

Then the set of events is typically taken to contain all open subintervals of Ω, i.e., all intervals of the form (a, b), $a, b \in \Omega$. More formally, let $\mathcal{F}$ be the smallest σ-algebra that contains *all open subintervals* in Ω. This σ-algebra is commonly referred to as the *Borel σ-algebra $\mathcal{B}$* and accordingly each event in $\mathcal{B}$ is called a *Borel set*.

Since $\mathcal{B}$ is a σ-algebra, it is closed under complement, countable unions, and countable intersections (cf. Problem 2.1), and contains many subsets other than open intervals. For example, since the half-open interval $(a, b]$ can be represented by a countable intersection of open intervals (Borel sets) as

$$(a, b] = \bigcap_{c \in \mathbb{Q}:\, c > b} (a, c), \tag{2.1}$$

it is also Borel. As a matter of fact, $\mathcal{B}$ contains all open subsets and thus is the smallest σ-algebra that contains all open subsets of Ω. The probability of any Borel set can be fully specified by assigning probabilities to open intervals (or to closed intervals, half-closed intervals, half-intervals, etc.).

Example 2.7 (Picking a random number between 0 and 1). $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}$, and

$$\mathsf{P}((a, b)) = b - a, \quad 0 \le a < b \le 1.$$

This is the *uniform* distribution over Ω. By (2.1) and the axioms of probability,

$$\mathsf{P}((a, b]) = \lim_{c \to b} (c - a) = b - a, \quad 0 \le a < b \le 1,$$

It can be similarly checked that

$$\mathsf{P}([a, b]) = b - a, \quad 0 \le a < b \le 1.$$

In particular, $\mathsf{P}(\{a\}) = 0$, $a \in [0, 1]$, and the probability of picking any specific number is zero.

For any reasonable Ω (such as a finite set or the d-dimensional Euclidean space $\mathbb{R}^d$, but sometimes even a space of time series or functions), the Borel σ-algebra can be defined as the smallest σ-algebra that contains all open subsets. When Ω is countable, the Borel σ-algebra is 2^Ω. Henceforth, we assume that $\mathcal{F}$ is the Borel σ-algebra of Ω and any event of our interest is Borel unless specified otherwise. Note, however, that for an uncountable Ω, there are many subsets of Ω that are not Borel (if interested in these sets, refer to any graduate-level course on *measure theory*).

2.4 BASIC PROBABILITY LAWS

We can establish the following as simple corollaries of the axioms of probability.

1. $\mathsf{P}(A^c) = 1 - \mathsf{P}(A)$.

2. If $A \subseteq B$, then $\mathsf{P}(A) \le \mathsf{P}(B)$.
3. $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$.
4. $\mathsf{P}(A \cup B) \le \mathsf{P}(A) + \mathsf{P}(B)$.

More generally, we have the following inequality, also known as *Boole's inequality*, that can be generalized to a countably infinite number of events.

Union of events bound. For any events $A_1, A_2, \ldots, A_n$,

$$\mathsf{P}\Big(\bigcup_{i=1}^{n} A_i\Big) \le \sum_{i=1}^{n} \mathsf{P}(A_i)$$

The following identity is very useful in finding the probability of a complicated event.

Law of total probability. Let $A_1, A_2, \ldots$ be events that partition Ω, that is, $A_1, A_2, \ldots$ are disjoint ($A_i \cap A_j = \emptyset$, $i \ne j$) and $\bigcup_i A_i = \Omega$. Then for any event B,

$$\mathsf{P}(B) = \sum_i \mathsf{P}(A_i \cap B).$$

2.5 CONDITIONAL PROBABILITY AND THE BAYES RULE

So far probability measures are similar to other common *measures*, such as length, area, volume, and weight, all of which are nonnegative and countably additive. The notion of conditioning is a unique feature of probability theory that is not found in the general measure theory.

Let B be an event such that $\mathsf{P}(B) \ne 0$. The *conditional probability* of the event A given B is defined to be

$$\mathsf{P}(A \,|\, B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}.$$

The function $\mathsf{P}(\cdot \,|\, B) : \mathcal{F} \to [0, 1]$ is a probability measure in itself, that is, it satisfies the three axioms of probability:

1. $\mathsf{P}(A \,|\, B) \ge 0$ for every A in $\mathcal{F}$.
2. $\mathsf{P}(\Omega \,|\, B) = 1$.
3. If $A_1, A_2, \ldots$ are *disjoint*, then

$$\mathsf{P}\Big(\bigcup_{i=1}^{\infty} A_i \,\Big|\, B\Big) = \sum_{i=1}^{\infty} \mathsf{P}(A_i \,|\, B).$$

Assume that $\mathsf{P}(A) \neq 0$ and $\mathsf{P}(B) \neq 0$. Then the conditional probability of A given B—the *a posteriori* probability (or *posterior* in short) of A—can be related to the unconditional probability of A— the *a priori probability* (or *prior* in short) of A. Using the definition of conditional probability twice, we have

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(B \mid A)}{\mathsf{P}(B)} \mathsf{P}(A). \tag{2.2}$$

By multiplying $\mathsf{P}(B)$ on both sides, we establish the following useful identity.

Chain rule. For any pair of events A and B,

$$\mathsf{P}(A \cap B) = \mathsf{P}(A)\,\mathsf{P}(B \mid A) = \mathsf{P}(B)\,\mathsf{P}(A \mid B).$$

Note that the chain rule holds even when $\mathsf{P}(A) = 0$ or $\mathsf{P}(B) = 0$ if we interpret the product of zero and an undefined number to be zero. By induction, the chain rule can be generalized to more than two events. For example,

$$\mathsf{P}(A_1 \cap A_2 \cap A_3) = \mathsf{P}(A_1)\,\mathsf{P}(A_2 \mid A_1)\,\mathsf{P}(A_3 \mid A_1 \cap A_2).$$

Let $A_1, A_2, \ldots, A_n$ be nonzero probability events that partition Ω and let B be a nonzero probability event. By (2.2),

$$\mathsf{P}(A_j \mid B) = \frac{\mathsf{P}(B \mid A_j)}{\mathsf{P}(B)} \mathsf{P}(A_j). \tag{2.3}$$

By the law of total probability,

$$\mathsf{P}(B) = \sum_{i=1}^{n} \mathsf{P}(A_i \cap B) = \sum_{i=1}^{n} \mathsf{P}(A_i)\,\mathsf{P}(B \mid A_i) \tag{2.4}$$

Substituting (2.4) into (2.3) yields the famous relationship between the priors $\mathsf{P}(A_i)$, $i = 1, 2, \ldots, n$, and the posteriors $\mathsf{P}(A_j \mid B)$, $j = 1, 2, \ldots, n$.

Bayes rule. If $A_1, A_2, \ldots, A_n$ are nonzero probability events that partition Ω, then for any nonzero probability event B,

$$\mathsf{P}(A_j \mid B) = \frac{\mathsf{P}(B \mid A_j)}{\sum_{i=1}^{n} \mathsf{P}(A_i)\,\mathsf{P}(B \mid A_i)} \mathsf{P}(A_j), \quad j = 1, 2, \ldots, n.$$

The Bayes rule also applies to a countably infinite number of events.

Example 2.8 (Binary communication channel). Consider the *probability transition diagram* for a noisy binary channel in Figure 2.1. This is a random experiment with sample space

$$\Omega = \{(0,0),(0,1),(1,0),(1,1)\},$$

where the first entry is the bit sent (the input of the channel) and the second is the bit received (the output of the channel). Define the two events

$$\begin{aligned} A &= \{0 \text{ is sent}\} = \{(0,1),(0,0)\}, \\ B &= \{0 \text{ is received}\} = \{(0,0),(1,0)\}. \end{aligned}$$

The probability measure on Ω is determined by $\mathsf{P}(A)$, $\mathsf{P}(B \,|\, A)$, and $\mathsf{P}(B^c|A^c)$, which are given on the probability transition diagram. To find $\mathsf{P}(A \,|\, B)$, we use Bayes rule:

$$\mathsf{P}(A \,|\, B) = \frac{\mathsf{P}(B \,|\, A)}{\mathsf{P}(A)\,\mathsf{P}(B \,|\, A) + \mathsf{P}(A^c)\,\mathsf{P}(B \,|\, A^c)}\;\mathsf{P}(A)$$

to obtain

$$\mathsf{P}(A \,|\, B) = \frac{0.9}{0.2 \cdot 0.9 + 0.8 \cdot 0.025} \cdot 0.2 = \frac{0.9}{0.2} \cdot 0.2 = 0.9.$$

Note that the posterior $\mathsf{P}(A \,|\, B) = 0.9$ is much larger than the prior $\mathsf{P}(A) = 0.2$; even though the observation is noisy, it still reveals some useful information about the input.

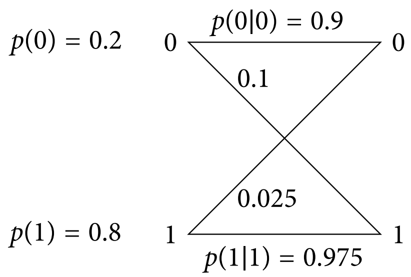


Figure 2.1. The probability transition diagram for a binary communication channel. Here $p(\cdot)$ denotes the probability of the input and $p(\cdot|\cdot)$ denotes the conditional probability of the output given the input.

2.6 INDEPENDENCE

Two events A and B are said to be *statistically independent* (or *independent* in short) if

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

When $\mathsf{P}(B) \neq 0$, this is equivalent to

$$\mathsf{P}(A \mid B) = \mathsf{P}(A).$$

In other words, knowing whether B occurs provides no information about whether A occurs.

Example 2.9. We revisit the binary channel discussed in Example 2.8. Assume that two independent bits are sent over the channel and we would like to find the probability that both bits are in error. Define the two events

$$\begin{aligned} E_1 &= \{\text{First bit is in error}\}, \\ E_2 &= \{\text{Second bit is in error}\}. \end{aligned}$$

Since the bits are sent independently, the probability that both are in error is

$$\mathsf{P}(E_1 \cap E_2) = \mathsf{P}(E_1)\,\mathsf{P}(E_2).$$

To find $\mathsf{P}(E_1)$, we express E_1 in terms of the events A_1 (0 is sent in the first transmission) and B_1 (0 is received in the first transmission) as

$$E_1 = (A_1 \cap B_1^c) \cup (A_1^c \cap B_1).$$

Since E_1 has been expressed as the union of disjoint events,

$$\begin{aligned} \mathsf{P}(E_1) &= \mathsf{P}(A_1 \cap B_1^c) + \mathsf{P}(A_1^c \cap B_1) \\ &= \mathsf{P}(A_1)\,\mathsf{P}(B_1^c \mid A_1) + \mathsf{P}(A_1^c)\,\mathsf{P}(B_1 \mid A_1^c) \\ &= 0.2 \cdot 0.1 + 0.8 \cdot 0.025 \\ &= 0.04. \end{aligned}$$

The probability that the two bits are in error is

$$\mathsf{P}(E_1 \cap E_2) = \mathsf{P}(E_1)\,\mathsf{P}(E_2) = (0.04)^2 = 1.6 \times 10^{-3}.$$

In general, the events $A_1, A_2, \ldots, A_n$ are said to be *mutually independent* (or *independent* in short) if for *every* subset $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ of the events,

$$\mathsf{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \prod_{j=1}^{k} \mathsf{P}(A_{i_j}).$$

For example, A, B, and C are independent if all of the following hold:

$$\mathsf{P}(A \cap B) = \mathsf{P}(A)\ \mathsf{P}(B), \tag{2.5}$$

$$\mathsf{P}(A \cap C) = \mathsf{P}(A)\ \mathsf{P}(C), \tag{2.6}$$

$$\mathsf{P}(B \cap C) = \mathsf{P}(B)\ \mathsf{P}(C), \tag{2.7}$$

$$\mathsf{P}(A \cap B \cap C) = \mathsf{P}(A)\ \mathsf{P}(B)\ \mathsf{P}(C). \tag{2.8}$$

Note that the last identity (2.8), or more generally,

$$\mathsf{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \prod_{j=1}^{n} \mathsf{P}(A_i)$$

is *not* sufficient for mutual independence.

Example 2.10. Roll two fair dice independently. Define the events

$$\begin{aligned}
A &= \{\text{The first die roll is 1, 2, or 3}\},\\
B &= \{\text{The first die roll is 2, 3, or 6}\},\\
C &= \{\text{The sum of the two rolls is 9}\} = \{(3,6),(4,5),(5,4),(6,3)\}.
\end{aligned}$$

Since the dice are fair and the experiments are done independently, the probability of any pair of outcomes is 1/36. Therefore

$$\mathsf{P}(A) = \frac{1}{2}, \quad \mathsf{P}(B) = \frac{1}{2}, \quad \mathsf{P}(C) = \frac{1}{9}.$$

Since $A \cap B \cap C = \{(3,6)\}$,

$$\mathsf{P}(A \cap B \cap C) = \frac{1}{36} = \mathsf{P}(A)\,\mathsf{P}(B)\,\mathsf{P}(C).$$

But A, B, and C are *not* independent because

$$\mathsf{P}(A \cap B) = \frac{1}{3} \neq \frac{1}{4} = \mathsf{P}(A)\,\mathsf{P}(B).$$

Similarly, pairwise independence, e.g., (2.5)–(2.7), does not imply independence.

Example 2.11. Flip two fair coins independently. Define the events

$$\begin{aligned}
A &= \{\text{The first coin is a head}\},\\
B &= \{\text{The second coin is a head}\},\\
C &= \{\text{Both coins are the same}\}.
\end{aligned}$$

Since the flips are independent,

$$\begin{aligned}
\mathsf{P}(A \cap B) = \mathsf{P}(B \cap C) = \mathsf{P}(C \cap A) &= \frac{1}{4}\\
&= \mathsf{P}(A)\,\mathsf{P}(B) = \mathsf{P}(B)\,\mathsf{P}(C) = \mathsf{P}(C)\,\mathsf{P}(A),
\end{aligned}$$

and A, B, and C are *pairwise* independent. However, they are *not* independent since

$$\mathsf{P}(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = \mathsf{P}(A)\,\mathsf{P}(B)\,\mathsf{P}(C).$$

Two events A and B are said to be *conditionally independent* give a third event C with $\mathsf{P}(C) > 0$ if

$$\mathsf{P}(A \cap B \mid C) = \mathsf{P}(A|C)\,\mathsf{P}(B|C).$$

Example 2.12. We continue Example 2.10. Since $A \cap C = \{(3,6)\}$, $B \cap C = \{(3,6),(6,3)\}$, and $A \cap B \cap C = \{(3,6)\}$,

$$\mathsf{P}(A \cap B \,|\, C) = \frac{1}{4} \neq \frac{1}{8} = \mathsf{P}(A|C)\,\mathsf{P}(B|C).$$

Hence, A and B are not conditionally independent given C. Now define the event

$$D = \{\text{The sum of the two rolls is 4}\} = \{(1,3),(2,2),(3,1)\}.$$

Since $A \cap D = \{(1,3),(2,2),(3,1)\}$, $B \cap D = \{(2,2),(3,1)\}$, and $A \cap B \cap D = \{(2,2),(3,1)\}$,

$$\mathsf{P}(A \cap B \,|\, D) = \frac{2}{3} = \mathsf{P}(A|D)\,\mathsf{P}(B|D).$$

Hence, A and B are conditionally independent given D.

Conditional independence of more than two events given another event C is defined similarly as independence with respect to the probability measure $\mathsf{P}(\cdot|C)$. Conditional independence neither implies nor is implied by (unconditional) independence. In Example 2.12, A and B are conditionally independent given D, but are not independent unconditionally as shown in Example 2.10. In Example 2.11, A and B are independent, but they are not conditionally independent given C as

$$\mathsf{P}(A \cup B \,|\, C) = \frac{1}{2} \neq \frac{1}{4} = \mathsf{P}(A|C)\,\mathsf{P}(B|C).$$

PROBLEMS

2.1. *σ-algebra.* Show that if $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.

2.2. *Limits of probabilities.* Show

(a) $\mathsf{P}(\bigcup_{i=1}^{\infty} A_i) = \lim_{n\to\infty} \mathsf{P}(\bigcup_{i=1}^{n} A_i)$.

(b) $\mathsf{P}(\bigcap_{i=1}^{\infty} A_i) = \lim_{n\to\infty} \mathsf{P}(\bigcap_{i=1}^{n} A_i)$.

2.3. *Extension of a probability measure.* Consider a discrete probability space $(\Omega, 2^{\Omega}, \mathsf{P})$, where Ω is a subset of $\mathbb{R}$. Show that $\mathsf{P}(\cdot \cap \Omega)$ is a valid probability measure for the sample space $\mathbb{R}$ and the set of events $2^{\mathbb{R}}$, that is, it satisfies the axioms of probability.

2.4. *Independence.* Show that the events A and B are independent if $\mathsf{P}(A|B) = \mathsf{P}(A|B^c)$.

2.5. *Conditional independence.* Let A and B be two events such that $\mathsf{P}(A \cap B) > 0$. Show that A and B are conditionally independent given $A \cap B$.

2.6. *Conditional probabilities.* Let $\mathsf{P}(A) = 0.8$, $\mathsf{P}(B^c) = 0.6$, and $\mathsf{P}(A \cup B) = 0.8$. Find

(a) $\mathsf{P}(A^c|B^c)$.

(b) $\mathsf{P}(B^c|A)$.

2.7. Let A and B be two events with $\mathsf{P}(A) \geq 0.5$ and $\mathsf{P}(B) \geq 0.75$. Show that $\mathsf{P}(A \cap B) \geq 0.25$.

2.8. *Monty Hall.* Gold is placed behind one of three curtains. A contestant chooses one of the curtains, Monty Hall (the game host) opens one of the unselected empty curtains. The contestant has a choice either to switch his selection to the third curtain or not.

(a) What is the sample space for this random experiment? (Hint: An outcome consists of the curtain with gold, the curtain chosen by the contestant, and the curtain chosen by Monty.)

(b) Assume that placement of the gold behind the three curtains is random, the contestant choice of curtains is random and independent of the gold placement, and that Monty Hall's choice of an empty curtain is random among the alternatives. Specify the probability measure for this random experiment and use it to compute the probability of winning the gold if the contestant decides to switch.

2.9. *Negative evidence.* Suppose that the evidence of an event B increases the probability of a criminal's guilt; that is, if A is the event that the criminal is guilty, then $\mathsf{P}(A|B) \geq \mathsf{P}(A)$. Does the absence of the event B decrease the criminal's probability of being guilty? In other words, is $\mathsf{P}(A|B^c) \leq \mathsf{P}(A)$? Prove or provide a counterexample.

2.10. *Random state transition.* Consider the *state diagram* in Figure 2.2. The sample space is

$$\Omega = \{(\alpha, \alpha),\ (\alpha, \beta),\ \ldots,\ (\gamma, \gamma)\},$$

where the first entry is the initial state and the second entry is the next state. Define the events

$$\begin{aligned} A_1 &= \{\text{the initial state is } \alpha\}, & A_2 &= \{\text{the next state is } \alpha\},\\ B_1 &= \{\text{the initial state is } \beta\}, & B_2 &= \{\text{the next state is } \beta\},\\ C_1 &= \{\text{the initial state is } \gamma\}, & C_2 &= \{\text{the next state is } \gamma\}. \end{aligned}$$

Assume that $\mathsf{P}(A_1) = 0.5$, $\mathsf{P}(B_1) = 0.2$, and $\mathsf{P}(C_1) = 0.3$.

(a) Find $\mathsf{P}(A_2)$, $\mathsf{P}(B_2)$, and $\mathsf{P}(C_2)$.

(b) Find $\mathsf{P}(A_1 \,|\, A_2)$, $\mathsf{P}(B_1 \,|\, B_2)$, and $\mathsf{P}(C_1 \,|\, C_2)$.

(c) Find two events among $A_1, A_2, B_1, B_2, C_1, C_2$ that are pairwise independent.

2.11. *Geometric pairs.* Consider a probability space consisting of the sample space

$$\Omega = \{1, 2, 3, \ldots\}^2 = \{(i, j) : i, j \in \mathbb{N}\},$$

i.e., all pairs of positive integers, the set of events 2^Ω, and the probability measure specified by

$$\mathsf{P}((i, j)) = p^2(1 - p)^{i+j-2}, \quad 0 < p < 1.$$

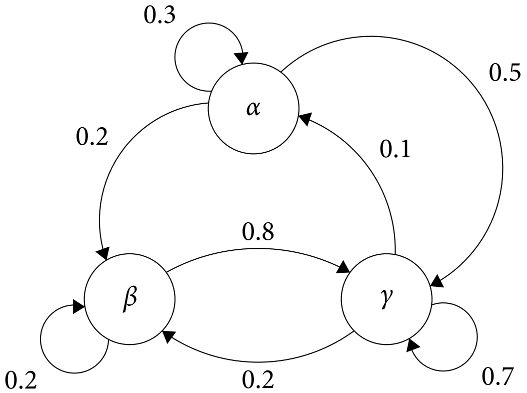


Figure 2.2. The state diagram for a three-state system. Here the label of each edge $i \to j$ denotes the transition probability from state i to state j, that is, the conditional probability that the next state is j given the initial state is i.

(a) Find $\mathsf{P}(\{(i, j) : i \geq j\})$.

(b) Find $\mathsf{P}(\{(i, j) : i + j = k\})$.

(c) Find $\mathsf{P}(\{(i, j) : i \text{ is an odd number}\})$.

(d) Describe an experiment whose outcomes (i, j), $i, j \in \mathbb{N}$, have the probabilities $\mathsf{P}((i, j))$.

2.12. *Juror's fallacy.* Suppose that $\mathsf{P}(A|B) \geq \mathsf{P}(A)$ and $\mathsf{P}(A|C) \geq \mathsf{P}(A)$. Is it always true that $\mathsf{P}(A|B \cap C) \geq \mathsf{P}(A)$? Prove or provide a counterexample.

2.13. *Polya's urn.* Suppose we have an urn containing one red ball and one blue ball. We draw a ball at random from the urn. If it is red, we put the drawn ball plus another red ball into the urn. If it is blue, we put the drawn ball plus another blue ball into the urn. We then repeat this process. At the n-th stage, we draw a ball at random from the urn with $n + 1$ balls, note its color, and put the drawn ball plus another ball of the same color into the urn.

(a) Find the probability that the first ball is red.

(b) Find the probability that the second ball is red.

(c) Find the probability that the first three balls are all red.

(d) Find the probability that two of the first three balls are red.

LECTURE 3

Random Variables

3.1 DEFINITION

It is often convenient to represent the outcome of a random experiment by a number. A *random variable* (r.v.) is such a representation. To be more precise, let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space. Then a random variable $X : \Omega \to \mathbb{R}$ is a mapping of the outcome.

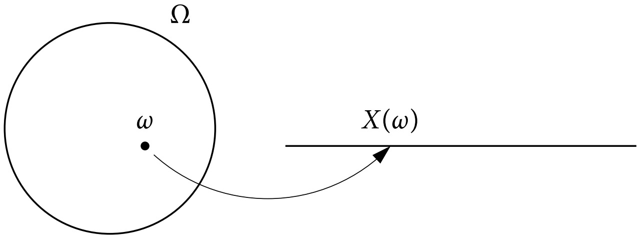


Figure 3.1. Random variable as a mapping.

Example 3.1. Let the random variable X be the number of heads in n coin flips. The sample space is $\Omega = \{H, T\}^n$, the possible outcomes of n coin flips; then

$$X \in \{0, 1, 2, \dots, n\}$$

Example 3.2. Consider packet arrival times $t_1, t_2, \dots$ in the interval $(0, T]$. The sample space Ω consists of the empty string (no packet) and all finite length strings of the form $(t_1, t_2, \dots, t_n)$ such that $0 < t_1 \le t_2 \le \cdots \le t_n \le T$. Define the random variable X to be the length of the string; then $X \in \{0, 1, 2, 3, \dots\}$.

Example 3.3. Consider the voltage across a capacitor. The sample space $\Omega = \mathbb{R}$. Define the random variables

$$X(\omega) = \omega,$$

$$Y(\omega) = \begin{cases} +1, & \omega \ge 0, \\ -1, & \text{otherwise.} \end{cases}$$

Example 3.4. Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space. For a given event $A \in \mathcal{F}$, define the *indicator random variable*

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

We use the notation $1_A(\omega)$ or $\chi_A(\omega)$ to denote the indicator random variable for A.

Throughout the course, we use uppercase letters, say, X, Y, Z, Φ, Θ, to denote random variables, and lowercase letters to denote the *values* taken by the random variables. Thus, $X(\omega) = x$ means that the random variable X takes on the value x when the outcome is ω.

As a representation of a random experiment in the probability space $(\Omega, \mathcal{F}, \mathsf{P})$, the random variable X can be viewed as an outcome of a random experiment on its own. The sample space is $\mathbb{R}$ and the set of events is the Borel σ-algebra $\mathcal{B}$. An event $A \in \mathcal{B}$ occurs if $X \in A$ and its probability is

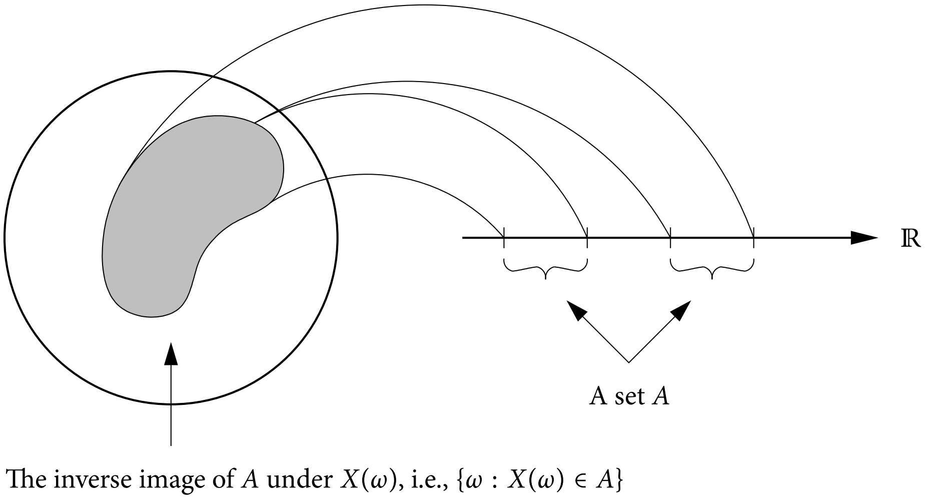
$$\mathsf{P}(\{\omega \in \Omega \colon X(\omega) \in A\}),$$

which is determined by the probability measure P of the underlying random experiment and the *inverse image* of A under the mapping $X : \Omega \to \mathbb{R}$. Thus, $(\Omega, \mathcal{F}, \mathsf{P})$ induces a probability space $(\mathbb{R}, \mathcal{B}, \mathsf{P}_X)$, where

$$\mathsf{P}_X(A) = \mathsf{P}(\{\omega \in \Omega \colon X(\omega) \in A\}), \quad A \in \mathcal{B}.$$

An implicit assumption here is that for every $A \in \mathcal{B}$, the inverse image $\{\omega \in \Omega \colon X(\omega) \in A\}$ is an event in $\mathcal{F}$. A mapping $X(\omega)$ satisfying this condition is called *measurable* (with respect to $\mathcal{F}$) and we will always assume that a given mapping is measurable.

Since we typically deal with multiple random variables on the same probability space, we will use the notation $\mathsf{P}\{X \in A\}$ instead of the more formal notation $\mathsf{P}_X(A)$ or $\mathsf{P}(\{\omega \in \Omega \colon X(\omega) \in A\})$.



The inverse image of A under $X(\omega)$, i.e., $\{\omega : X(\omega) \in A\}$

3.2 CUMULATIVE DISTRIBUTION FUNCTION

To determine $\mathsf{P}\{X \in A\}$ for any Borel set A, i.e., any set generated by open intervals via countable unions, intersections, and complements, it suffices to specify $\mathsf{P}\{X \in (a, b)\}$ or $\mathsf{P}\{X \in (a, b]\}$ for all $-\infty < a < b < \infty$. Then the probability of any other Borel set can be determined by the axioms of probability. Equivalently, it suffices to specify the *cumulative distribution function* (cdf) of the random variable X:

$$F_X(x) = \mathsf{P}\{X \leq x\} = \mathsf{P}\{X \in (-\infty, x\,]\}, \quad x \in \mathbb{R}.$$

The cdf of a random variable satisfies the following properties.

1. $F_X(x)$ is nonnegative, i.e.,
$$F_X(x) \geq 0, \quad x \in \mathbb{R}.$$

2. $F_X(x)$ is monotonically nondecreasing, i.e.,
$$F_X(a) \leq F_X(b), \quad a < b.$$

3. *Limits.*
$$\lim_{x\to-\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x\to+\infty} F_X(x) = 1.$$

4. $F_X(x)$ is right continuous, i.e.,
$$F_X(a^+) := \lim_{x\to a^+} F_X(x) = F_X(a).$$

5. *Probability of a singleton.*
$$\mathsf{P}\{X = a\} = F_X(a) - F_X(a^-),$$
where $F_X(a^-) := \lim_{x\to a^-} F_X(x)$.

Throughout, we use the notation $X \sim F(x)$ means that the random variable X has the cdf $F(x)$.

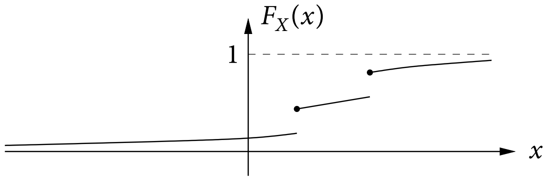


Figure 3.2. An illustration of a cumulative distribution function (cdf).

3.3 PROBABILITY MASS FUNCTION (PMF)

A random variable X is said to be *discrete* if $F_X(x)$ consists only of steps over a countable set $\mathcal{X}$ as illustrated in Figure 3.3.

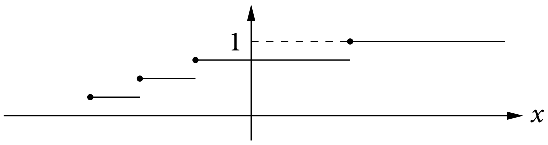


Figure 3.3. The cdf of a discrete random variable.

A discrete random variable X can be completely specified by its *probability mass function* (pmf)

$$p_X(x) = \mathsf{P}\{X = x\}, \quad x \in \mathcal{X}.$$

The set $\mathcal{X}$ is often referred to as the *alphabet* of X. Clearly, $p_X(x) \geq 0$, $\sum_{x\in\mathcal{X}} p_X(x) = 1$, and

$$\mathsf{P}(X \in A) = \sum_{x\in A\cap\mathcal{X}} p_X(x).$$

Throughout, we use the notation $X \sim p(x)$ to mean that X is a discrete random variable X with pmf $p(x)$.

We review a few famous discrete random variables.

Bernoulli. $X \sim \mathrm{Bern}(p)$, $p \in [0, 1]$, has the pmf

$$p_X(1) = p \quad \text{and} \quad p_X(0) = 1 - p.$$

This is the indicator of observing a head from flipping a coin with bias p.

Geometric. $X \sim \mathrm{Geom}(p)$, $p \in [0, 1]$, has the pmf

$$p_X(k) = p(1 - p)^{k-1}, \quad k = 1, 2, 3, \ldots.$$

This is the number of independent coin flips of bias p until the first head.

Binomial. $X \sim \mathrm{Binom}(n, p)$, $p \in [0, 1]$, $n = 1, 2, \ldots$, has the pmf

$$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

This is the number of heads in n independent coin flips of bias p.

Poisson. $X \sim \mathrm{Poisson}(\lambda)$, $\lambda > 0$, has the pmf

$$p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots.$$

[h]

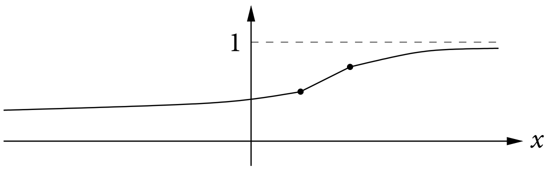


Figure 3.4. The cdf of a continuous random variable.

This is often used to characterize the number of random arrivals in a unit time interval, the number of random points in a unit area, and so on.

Let $X \sim \text{Binom}(n, \lambda/n)$. Then, its pmf for a fixed k is

$$\begin{aligned} p_X(k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}, \end{aligned}$$

which converges to the Poisson(λ) pmf $(\lambda^k/k!)e^{-\lambda}$ as $n \to \infty$. Thus, Poisson(λ) is the limit of Binom($n, \lambda/n$).

Example 3.5. In a popular lottery called "Powerball," the winning combination of numbers is selected uniformly at random among 292,201,338 possible combinations. Suppose that αn tickets are sold. What is the probability that there is no winner?

Since the number N of winners is a Binom($\alpha n, 1/n$) random variable with $n = 2.92 \times 10^8$ very large, we can use the Poission approximation to obtain

$$\mathsf{P}\{N = 0\} = \left(1 - \frac{1}{n}\right)^{\alpha n} \to e^{-\alpha}, \quad \text{as } n \to \infty.$$

If $\alpha = 1$, there is no winner with probability of 37%. If $\alpha = 2$, this probability decreases to 14%. Thus, even with 600 million tickets sold, there is a significant chance that the lottery has no winner and rolls over to the next week (with a bigger jackpot).

3.4 PROBABILITY DENSITY FUNCTION

A random variable is said to be *continuous* if its cdf is continuous as illustrated in Figure 3.4.

If $F_X(x)$ is continuous and differentiable (except possibly over a countable set), then X can be completely specified by its *probability density function* (pdf) $f_X(x)$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du.$$

If $F_X(x)$ is differentiable everywhere, then by the definition of derivative

$$\begin{aligned} f_X(x) &= \frac{dF_X(x)}{dx} \\ &= \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{\mathsf{P}\{x < X \le x + \Delta x\}}{\Delta x}. \end{aligned} \tag{3.1}$$

The pdf of a random variable satisfies the following properties.

1. $f_X(x)$ is nonnegative, i.e.,
$$f_X(x) \ge 0, \quad x \in \mathbb{R}.$$
2. *Normalization.*
$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1.$$
3. For any event $A \subset \mathbb{R}$,
$$\mathsf{P}\{X \in A\} = \int_{x \in A} f_X(x)\, dx.$$

   In particular,
$$\mathsf{P}\{a < X \le b\} = \mathsf{P}\{a < X < b\} = \mathsf{P}\{a \le X < b\} = \mathsf{P}\{a \le X \le b\} = \int_a^b f_X(x)\, dx.$$

Note that $f_X(x)$ should *not* be interpreted as the probability that $X = x$. In fact, $f_X(x)$ can be greater than 1. In light of (3.1), it is $f_X(x)\Delta x$ that can be interpreted as the approximation of the probability $\mathsf{P}\{x < X \le x + \Delta x\}$ for Δx sufficiently small.

Throughout, we use the notation $X \sim f(x)$ to mean that X is a continuous random variable with pdf $f(x)$.

We review a few famous continuous random variables.

Uniform. $X \sim \mathrm{Unif}[a, b]$, $a < b$, has the pdf

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b], \\ 0 & \text{otherwise.} \end{cases}$$

This is often used to model quantization noise.

Exponential. $X \sim \mathrm{Exp}(\lambda)$, $\lambda > 0$, has the pdf

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

This is often used to model the service time in a queue or the time between two random arrivals. An exponential random variable satisfies the *memoryless property*

$$\mathsf{P}(X > x + t \,|\, X > t) = \frac{\mathsf{P}\{X > x + t\}}{\mathsf{P}\{X > t\}} = \mathsf{P}\{X > x\}, \quad t, x > 0.$$

Example 3.6. Suppose that for every $t > 0$, the number of packet arrivals during time interval $(0, t]$ is a Poisson(λt) random variable, i.e.,

$$p_N(n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad n = 0, 1, 2, \ldots.$$

Let X be the time until the first packet arrival. Then the event $\{X > t\}$ is equivalent to the event $\{N = 0\}$ and thus

$$\begin{aligned} F_X(t) &= 1 - \mathsf{P}\{X > t) \\ &= 1 - \mathsf{P}\{N = 0\} \\ &= 1 - e^{-\lambda t}. \end{aligned}$$

Hence, $f_X(t) = \lambda e^{-\lambda t}$ and $X \sim \text{Exp}(\lambda)$.

Gaussian. $X \sim \text{N}(\mu, \sigma^2)$ has the pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

This characterizes many random phenomena such as thermal and shot noise, and is also called a *normal* random variable. The cdf of the *standard normal* random variable $\text{N}(0, 1)$ is

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \, du.$$

Its complement is

$$Q(x) = 1 - \Phi(x) = \mathsf{P}\{X > x\}.$$

The numerical values of the Q function is often used to compute probabilities of any Gaussian random variable $Y \sim \text{N}(\mu, \sigma^2)$ as

$$\mathsf{P}\{Y > y\} = \mathsf{P}\left\{X > \frac{y - \mu}{\sigma}\right\} = Q\left(\frac{y - \mu}{\sigma}\right). \tag{3.2}$$

3.5 FUNCTIONS OF A RANDOM VARIABLE

Let X be a random variable and $g : \mathbb{R} \to \mathbb{R}$ be a given function. Then $Y = g(X)$ is a random variable and its probability distribution can be expressed through that of X. For example, if X is discrete, then Y is discrete and

$$\begin{aligned} p_Y(y) &= \mathsf{P}\{Y = y\} \\ &= \mathsf{P}\{g(X) = y\} \\ &= \sum_{x:\, g(x)=y} p_X(x). \end{aligned}$$

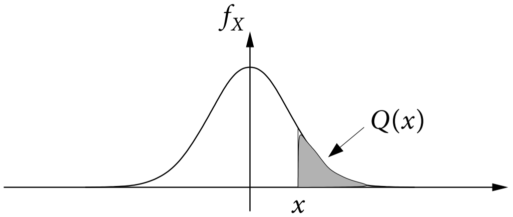


Figure 3.5. The pdf of the standard normal random variable and the Q function.

In general,

$$\begin{aligned} F_Y(y) &= \mathsf{P}\{Y \le y\} \\ &= \mathsf{P}\{g(X) \le y\}, \end{aligned}$$

which can be further simplified in many cases.

Example 3.7 (Linear function). Let $X \sim F_X(x)$ and $Y = aX + b$, $a \neq 0$. If $a > 0$, then

$$F_Y(y) = \mathsf{P}\{aX + b \le y\} = \mathsf{P}\left\{X \le \frac{y-b}{a}\right\} = F_X\left(\frac{y-b}{a}\right).$$

Taking derivative with respect to y, we have

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right)$$

We can similarly show that if $a < 0$, then

$$F_Y(y) = 1 - F_X\left(\left(\frac{y-b}{a}\right)^-\right)$$

and

$$f_Y(y) = -\frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

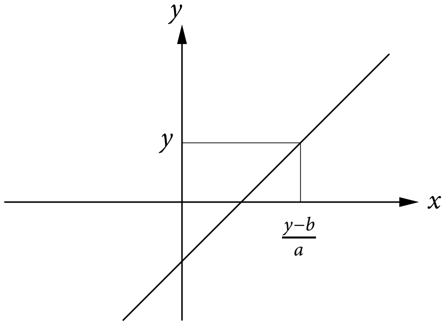


Figure 3.6. A linear function.

Combining both cases,

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

As a special case, let $X \sim \mathrm{N}(\mu, \sigma^2)$, i.e.,

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Again setting $Y = aX + b$, we have

$$\begin{aligned} f_Y(y) &= \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \\ &= \frac{1}{|a|} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(\frac{y-b}{a}-\mu\right)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi(a\sigma)^2}} e^{-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}}. \end{aligned}$$

Therefore, $Y \sim \mathrm{N}(a\mu + b,\ a^2\sigma^2)$. This result justifies the use of the Q function in (3.2) to compute probabilities for an arbitrary Gaussian random variable.

Example 3.8 (Quadratic function). Let $X \sim F_X(x)$ and $Y = X^2$. If $y < 0$, then $F_Y(y) = 0$. Otherwise,

$$F_Y(y) = \mathsf{P}\left\{-\sqrt{y} \le X \le \sqrt{y}\right\} = F_X\left(\sqrt{y}\right) - F_X\left((-\sqrt{y})^-\right)$$

If X is continuous with pdf $f_X(x)$, then

$$f_Y(y) = \frac{1}{2\sqrt{y}}\left(f_X(-\sqrt{y}) + f_X(\sqrt{y})\right).$$

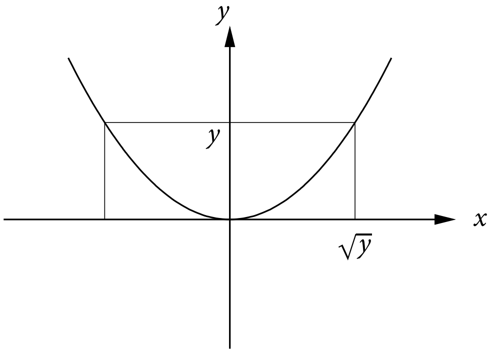


Figure 3.7. A quadratic function.

The above two examples can be generalized as follows.

Proposition 3.1. Let $X \sim f_X(x)$, $g(x)$ be differentiable, and $Y = g(X)$. Then

$$f_Y(y) = \sum_{i=1} \frac{f_X(x_i)}{|g'(x_i)|},$$

where $x_1, x_2, \ldots$ are the solutions of the equation $y = g(x)$ and $g'(x_i)$ is the derivative of g evaluated at x_i.

The distribution of Y can be written explicitly even when g is not differentiable.

Example 3.9 (Limiter). Let X be a r.v. with Laplacian pdf $f_X(x) = \frac{1}{2}e^{-|x|}$, and let Y be defined by the function of X shown in Figure 3.8. Consider the following cases.

- If $y < -a$, clearly $F_Y(y) = 0$.
- If $y = -a$,

$$\begin{aligned} F_Y(-a) &= F_X(-1) \\ &= \int_{-\infty}^{-1} \tfrac{1}{2}e^{x}\, dx = \tfrac{1}{2}e^{-1}. \end{aligned}$$

- If $-a < y < a$,

$$\begin{aligned} F_Y(y) &= \mathsf{P}\{Y \le y\} \\ &= \mathsf{P}\left\{aX \le y\right\} \\ &= \mathsf{P}\left\{X \le \frac{y}{a}\right\} = F_X\left(\frac{y}{a}\right) \\ &= \tfrac{1}{2}e^{-1} + \int_{-1}^{y/a} \tfrac{1}{2}e^{-|x|}\, dx. \end{aligned}$$

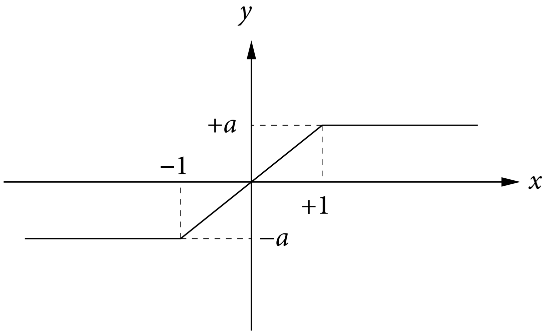


Figure 3.8. The limiter function.

- If $y \geq a$, $F_Y(y) = 1$.

Combining these cases, the cdf of Y is sketched in Figure 3.9.

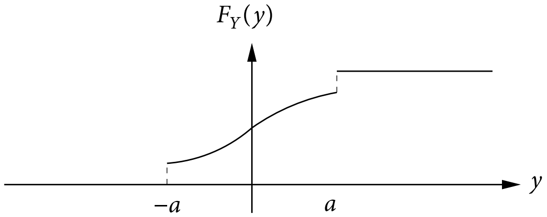


Figure 3.9. The cdf of the random variable Y.

3.6 GENERATION OF RANDOM VARIABLES

Suppose that we are given a uniform random variable $X \sim \mathrm{Unif}[0,1]$ and wish to generate a random variable Y with prescribed cdf $F(y)$. If $F(y)$ is continuous and strictly increasing, set

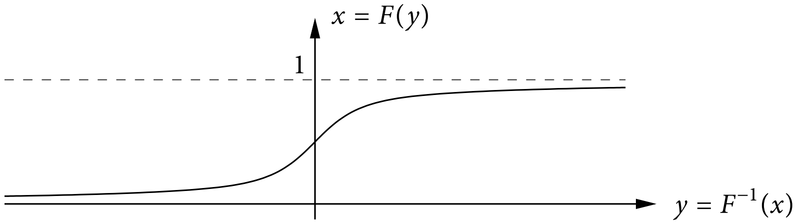


Figure 3.10. Generation of $Y \sim F(y)$ from a uniform random variable X.

$$Y = F^{-1}(X).$$

Then, since $X \sim \mathrm{Unif}[0,1]$ and $0 \leq F(y) \leq 1$,

$$\begin{aligned} F_Y(y) &= \mathsf{P}\{Y \leq y\} \\ &= \mathsf{P}\{F^{-1}(X) \leq y\} \\ &= \mathsf{P}\{X \leq F(y)\} \\ &= F(y). \end{aligned} \tag{3.3}$$

Thus, Y has the desired cdf $F(y)$. For example, to generate $Y \sim \mathrm{Exp}(\lambda)$ from $X \sim \mathrm{Unif}[0,1]$, we set

$$Y = -\frac{1}{\lambda}\ln(1 - X).$$

More generally, for an arbitrary cdf $F(y)$, we define

$$F^{-1}(x) := \min\{y\colon x \le F(y)\}, \quad x \in (0,1]. \tag{3.4}$$

Since $F(y)$ is right continuous, the above minimum is well-defined. Furthermore, since $F(y)$ is monotonically nondecreasing, $F^{-1}(x) \le y$ iff $x \le F(y)$. We now set $Y = F^{-1}(X)$ as before, but under this new definition of "inverse." It follows immediately that the equality in (3.3) continues to hold and that $Y \sim F(y)$. For example, to generate $Y \sim \text{Bern}(p)$, we set

$$Y = \begin{cases} 0 & X \le 1-p, \\ 1 & \text{otherwise.} \end{cases}$$

In conclusion, we can generate a random variable with any desired distribution from a Unif[0, 1] random variable.

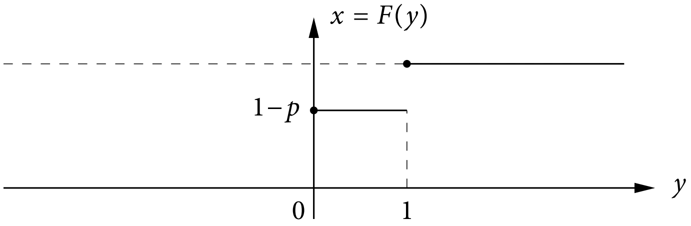


Figure 3.11. Generation of a Bern(p) random variable.

Conversely, a uniform random variable can be generated from any continuous random variable. Let X be a continuous random variable with cdf $F(x)$ and $Y = F(X)$. Since $F(x) \in [0,1]$, $F_Y(y) = \mathsf{P}\{Y \le y\} = 0$ for $y < 0$ and $F_Y(y) = 1$ for $y > 1$. For $y \in [0,1]$, let $F^{-1}(y)$ be defined as in (3.4). Then

$$\begin{aligned} F_Y(y) &= \mathsf{P}\{Y \le y\} \\ &= \mathsf{P}\{F(X) \le y\} \\ &= \mathsf{P}\{X \le F^{-1}(y)\} \\ &= F(F^{-1}(y)) \qquad (3.5) \\ &= y, \end{aligned}$$

where the equality in (3.5) follows by the definition of $F^{-1}(y)$. Hence, $Y \sim U[0,1]$. For example, let $X \sim \text{Exp}(\lambda)$ and

$$Y = \begin{cases} 1 - \exp(-\lambda X) & X \ge 0. \\ 0 & \text{otherwise.} \end{cases}$$

Then $Y \sim \text{Unif}[0,1]$.

The exact generation of a uniform random variable, which requires an infinite number of bits to describe, is not possible in any digital computer. One can instead use the following approximation. Let $X_1, X_2, \ldots X_n$ be independent and identically distributed (i.i.d.) Bern(1/2) random variables, and

$$Y = .X_1 X_2 \ldots X_n$$

be a fraction in base 2 that lies between 0 and 1. Then Y is a discrete random variable uniformly distributed over the set $\{k/2^n : k = 0, 1, \ldots, 2^n - 1\}$ and its cdf $F(y)$ converges to that of a Unif$[0,1]$ random variable for every y as $n \to \infty$. Thus, by flipping many fair coin flips, one can simulate a uniform random variable.

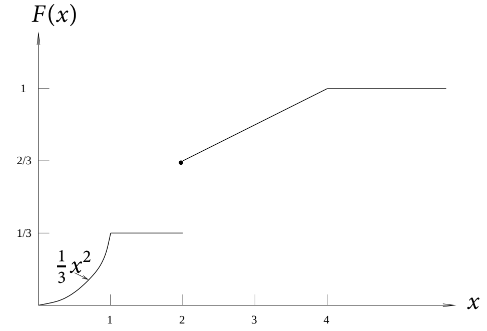
The fairness of coin flips is not essential to this procedure. Suppose that Z_1 and Z_2 are i.i.d. Bern(p) random variable. The following procedure due to von Neumann can generate a single Bern(1/2) random variable, even when the bias p is unknown. Let

$$X = \begin{cases} 0 & (Z_1, Z_2) = (0,1), \\ 1 & (Z_1, Z_2) = (1,0). \end{cases}$$

If $(Z_1, Z_2) = (0,0)$ or $(1,1)$, then the outcome is ignored. Clearly $p_X(0) = p_X(1) = 1/2$. By repeating the same procedure, one can generate a sequence of i.i.d. Bern(1/2) random variables from a sequence of i.i.d. Bern(p) random variables.

PROBLEMS

3.1. *Probabilities from a cdf.* Let X be a random variable with the cdf shown below.



Find the probabilities of the following events.

(a) $\{X = 2\}$.

(b) $\{X < 2\}$.

(c) $\{X = 2\} \cup \{0.5 \le X \le 1.5\}$.

(d) $\{X = 2\} \cup \{0.5 \le X \le 3\}$.

3.2. *Gaussian probabilities.* Let $X \sim \mathrm{N}(1000, 400)$. Express the following in terms of the Q function.

(a) $\mathsf{P}\{0 < X < 1020\}$.

(b) $\mathsf{P}\{X < 1020 | X > 960\}$.

3.3. *Laplacian.* Let $X \sim f(x) = \frac{1}{2}e^{-|x|}$.

(a) Sketch the cdf of X.

(b) Find $\mathsf{P}\{|X| \le 2 \text{ or } X \ge 0\}$.

(c) Find $\mathsf{P}\{|X| + |X-3| \le 3\}$.

(d) Find $\mathsf{P}\{X \ge 0 \mid X \le 1\}$.

3.4. *Distance to the nearest star.* Let the random variable N be the number of stars in a region of space of volume V. Assume that N is a Poisson r.v. with pmf

$$p_N(n) = \frac{e^{-\rho V}(\rho V)^n}{n!}, \qquad \text{for } n = 0, 1, 2, \ldots,$$

where ρ is the "density" of stars in space. We choose an arbitrary point in space and define the random variable X to be the distance from the chosen point to the nearest star. Find the pdf of X (in terms of ρ).

3.5. *Time until the n-th arrival.* Let the random variable N be the number of packets arriving during time $(0, t]$. Suppose that N is Poisson with pmf

$$p_N(n) = \frac{(\lambda t)^n}{n!}e^{-\lambda t} \qquad \text{for } n = 0, 1, 2, \ldots.$$

Let the random variable Y be the time to get the n-th packet. Find the pdf of Y.

3.6. *Uniform arrival.* The arrival time of a professor to his office is uniformly distributed in the interval between 8 and 9 am.

(a) Find the probability that the professor will arrive during the next minute given that he has not arrived by 8:30.

(b) Repeat for 8:50.

3.7. *Lognormal distribution.* Let $X \sim \mathrm{N}(0, \sigma^2)$. Find the pdf of $Y = e^X$ (known as the *lognormal* pdf).

3.8. *Random phase signal.* Let $Y(t) = \sin(\omega t + \Theta)$ be a sinusoidal signal with random phase $\Theta \sim U[-\pi, \pi]$. Find the pdf of the random variable $Y(t)$ (assume here that both t and the radial frequency ω are constant). Comment on the dependence of the pdf of $Y(t)$ on time t.

3.9. *Quantizer.* Let $X \sim \mathrm{Exp}(\lambda)$, i.e., an exponential random variable with parameter λ and $Y = \lfloor X \rfloor$, i.e., $Y = k$ for $k \le X < k+1$, $k = 0, 1, 2, \ldots$.

(a) Find the pmf of Y.

(b) Find the pdf of the quantization error $Z = X - Y$.

3.10. *Gambling.* Alice enters a casino with one unit of capital. She looks at her watch to generate a uniform random variable $U \sim \text{unif}\,[0,1]$, then bets the amount U on a fair coin flip. Her wealth is thus given by the r.v.

$$X = \begin{cases} 1 + U, & \text{with probability } 1/2, \\ 1 - U, & \text{with probability } 1/2. \end{cases}$$

Find the cdf of X.

3.11. *Nonlinear processing.* Let $X \sim \text{Unif}\,[-1,1]$. Define the random variable

$$Y = \begin{cases} X^2 + 1, & \text{if } |X| \geq 0.5 \\ 0, & \text{otherwise.} \end{cases}$$

Find and sketch the cdf of Y.

LECTURE 4

Pairs of Random Variables

4.1 TWO RANDOM VARIABLES

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probably space and consider two measurable mappings

$$X : \Omega \to \mathbb{R},$$
$$Y : \Omega \to \mathbb{R}.$$

In other words, X and Y are two random variables defined on the common probability space $(\Omega, \mathcal{F}, \mathsf{P})$. In order to specify the random variables, we need to determine

$$\mathsf{P}\{(X, Y) \in A\} = \mathsf{P}(\{\omega \in \Omega : (X(\omega), Y(\omega)) \in A\})$$

for every Borel set $A \subseteq \mathbb{R}^2$. By the properties of the Borel σ-field, it can be shown that it suffices to determine the probabilities of the form

$$\mathsf{P}\{a < X < b,\ c < Y < d\}, \quad a < b,\ c < d,$$

or equivalently, the probabilities of the form

$$\mathsf{P}\{X \le x,\ Y \le y\}, \quad x, y \in \mathbb{R}.$$

The latter defines their *joint cdf*

$$F_{X,Y}(x, y) = \mathsf{P}\{X \le x,\ Y \le y\}, \quad x, y \in \mathbb{R},$$

which is the shaded region in Figure 4.1.

The joint cdf satisfies the following properties:

1. $F_{X,Y}(x, y) \ge 0$.
2. If $x_1 \le x_2$ and $y_1 \le y_2$, then

$$F_{X,Y}(x_1, y_1) \le F_{X,Y}(x_2, y_2).$$

3. *Limits.*

$$\lim_{x,y\to\infty} F_{X,Y}(x, y) = 1,$$
$$\lim_{y\to-\infty} F_{X,Y}(x, y) = \lim_{x\to-\infty} F_{X,Y}(x, y) = 0.$$

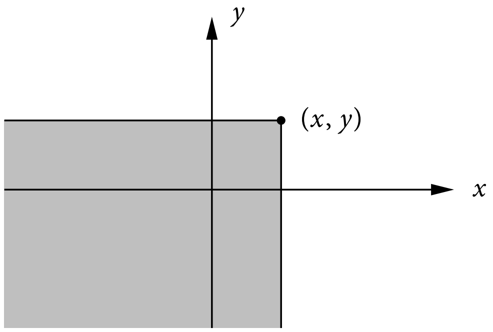


Figure 4.1. An illustration of the joint pdf of X and Y.

4. *Marginal cdfs.*

$$\lim_{y\to\infty} F_{X,Y}(x, y) = F_X(x),$$
$$\lim_{x\to\infty} F_{X,Y}(x, y) = F_Y(y).$$

The probability of any (Borel) set can be determined from the joint cdf. For example,

$$\mathsf{P}\{a < X \le b,\ c < Y \le d\} = F_{X,Y}(b, d) - F_{X,Y}(a, d) - F_{X,Y}(b, c) + F_{X,Y}(a, c)$$

as illustrated in Figure 4.2

We say that X and Y are *statistically independent* or *independent* in short if for every

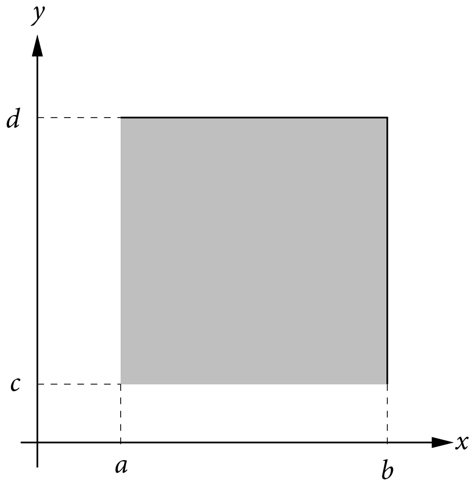


Figure 4.2. An illustration of $\mathsf{P}\{a < X \le b,\ c < Y \le d\}$.

(Borel) A and B,

$$\mathsf{P}\{X \in A,\ Y \in B\} = \mathsf{P}\{X \in A\}\,\mathsf{P}\{Y \in B\}.$$

Equivalently, X and Y are independent if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y), \quad x, y \in \mathbb{R}.$$

In the following, we focus on three special cases and discuss the joint, marginal, and conditional distributions of X and Y for each case.

- X and Y are discrete.
- X and Y are continuous.
- X is discrete and Y is continuous (mixed).

4.2 PAIRS OF DISCRETE RANDOM VARIABLES

Let X and Y be discrete random variables on the same probability space. They are completely specified by their *joint pmf*

$$p_{X,Y}(x, y) = \mathsf{P}\{X = x, Y = y\}, \quad x \in \mathcal{X},\ y \in \mathcal{Y}.$$

By the axioms of probability,

$$\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{X,Y}(x, y) = 1.$$

We use the law of total probability to find the *marginal pmf* of X:

$$p_X(x) = \sum_{y\in\mathcal{Y}} p_{X,Y}(x, y), \quad x \in \mathcal{X}.$$

The *conditional pmf* of X given $Y = y$ is defined as

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}, \quad p_Y(y) \neq 0,\ x \in \mathcal{X}.$$

Check that if $p_Y(y) \neq 0$, then $p_{X|Y}(x|y)$ is a pmf for X.

Example 4.1. Consider the pmf $p_{X,Y}(x, y)$ described by the following table

			x	
		0	1	2.5
	-3	0	$\frac{1}{4}$	$\frac{1}{8}$
y	-1	$\frac{1}{8}$	0	$\frac{1}{4}$
	2	$\frac{1}{8}$	$\frac{1}{8}$	0

Then,

$$p_X(x) = \begin{cases} 1/4 & x = 0, \\ 3/8 & x = 1, \\ 3/8 & x = 2.5, \end{cases}$$

and

$$p_{X|Y}(x|2) = \begin{cases} 1/2 & x = 0, \\ 1/2 & x = 1. \end{cases}$$

Chain rule. $p_{X,Y}(x, y) = p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y)$.

Independence. X and Y are *independent* if

$$p_{X,Y}(x, y) = p_X(x)p_Y(y), \quad x \in \mathcal{X},\ y \in \mathcal{Y},$$

which is equivalent to

$$p_{X|Y}(x\,|\,y) = p_X(x), \quad x \in \mathcal{X},\ p_Y(y) \neq 0.$$

Law of total probability. For any event A,

$$\mathsf{P}(A) = \sum_x p_X(x)\,\mathsf{P}(A \mid X = x).$$

Bayes rule. Given $p_X(x)$ and $p_{Y|X}(y|x)$ for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we can find $p_{X|Y}(x|y)$ as

$$\begin{aligned} p_{X|Y}(x\,|\,y) &= \frac{p_{X,Y}(x, y)}{p_Y(y)} \\ &= \frac{p_X(x)p_{Y|X}(y|x)}{p_Y(y)} \\ &= \frac{p_X(x)p_{Y|X}(y|x)}{\sum_{u \in \mathcal{X}} p_{X,Y}(u, y)} \\ &= \frac{p_X(x)p_{Y|X}(y|x)}{\sum_{u \in \mathcal{X}} p_X(u)p_{Y|X}(y|u)}. \end{aligned}$$

The final formula is entirely in terms of the known quantities $p_X(x)$ and $p_{Y|X}(y|x)$.

Example 4.2 (Binary symmetric channel). Consider the binary communication channel in Figure 4.3. The bit sent is $X \sim \text{Bern}(p)$, $p \in [0, 1]$, and the bit received is

$$Y = (X + Z) \bmod 2 = X \oplus Z,$$

where the noise $Z \sim \text{Bern}(\epsilon)$, $\epsilon \in [0, 1/2]$, is independent of X. We find $p_{X|Y}(x|y)$, $p_Y(y)$,

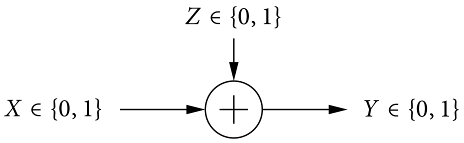


Figure 4.3. Binary symmetric channel.

and $\mathsf{P}\{X \neq Y\}$, namely, the probability of error. First, we use the Bayes rule

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)}{\sum\limits_{u\in\mathcal{X}} p_{Y|X}(y|u)p_X(u)}\, p_X(x).$$

We know $p_X(x)$, but we need to find $p_{Y|X}(y|x)$:

$$\begin{aligned} p_{Y|X}(y|x) &= \mathsf{P}\{Y = y \,|\, X = x\} = \mathsf{P}\{X \oplus Z = y \,|\, X = x\} \\ &= \mathsf{P}\{x \oplus Z = y \,|\, X = x\} = \mathsf{P}\{Z = y \oplus x \,|\, X = x\} \\ &= \mathsf{P}\{Z = y \oplus x\} \quad \text{since } Z \text{ and } X \text{ are independent} \\ &= p_Z(y \oplus x). \end{aligned}$$

Therefore

$$\begin{aligned} p_{Y|X}(0\,|\,0) &= p_Z(0 \oplus 0) = p_Z(0) = 1 - \epsilon, \\ p_{Y|X}(0\,|\,1) &= p_Z(0 \oplus 1) = p_Z(1) = \epsilon, \\ p_{Y|X}(1\,|\,0) &= p_Z(1 \oplus 0) = p_Z(1) = \epsilon, \\ p_{Y|X}(1\,|\,1) &= p_Z(1 \oplus 1) = p_Z(0) = 1 - \epsilon. \end{aligned}$$

Plugging into the Bayes rule equation, we obtain

$$\begin{aligned} p_{X|Y}(0|0) &= \frac{p_{Y|X}(0|0)}{p_{Y|X}(0|0)p_X(0) + p_{Y|X}(0|1)p_X(1)}\, p_X(0) = \frac{(1-\epsilon)(1-p)}{(1-\epsilon)(1-p)+\epsilon p}. \\ p_{X|Y}(1|0) &= 1 - p_{X|Y}(0|0) = \frac{\epsilon p}{(1-\epsilon)(1-p)+\epsilon p}. \\ p_{X|Y}(0|1) &= \frac{p_{Y|X}(1|0)}{p_{Y|X}(1|0)p_X(0) + p_{Y|X}(1|1)p_X(1)}\, p_X(0) = \frac{\epsilon(1-p)}{(1-\epsilon)p+\epsilon(1-p)}. \\ p_{X|Y}(1|1) &= 1 - p_{X|Y}(0|1) = \frac{(1-\epsilon)p}{(1-\epsilon)p+\epsilon(1-p)}. \end{aligned}$$

We already found $p_Y(y)$ as

$$\begin{aligned} p_Y(y) &= p_{Y|X}(y|0)p_X(0) + p_{Y|X}(y|1)p_X(1) \\ &= \begin{cases} (1-\epsilon)(1-p) + \epsilon p & \text{for } y = 0, \\ \epsilon(1-p) + (1-\epsilon)p & \text{for } y = 1. \end{cases} \end{aligned}$$

Now to find the probability of error $\mathsf{P}\{X \neq Y\}$, consider

$$\begin{aligned}\mathsf{P}\{X \neq Y\} &= p_{X,Y}(0,1) + p_{X,Y}(1,0)\\ &= p_{Y|X}(1|0)p_X(0) + p_{Y|X}(0|1)p_X(1)\\ &= \epsilon(1-p) + \epsilon p\\ &= \epsilon.\end{aligned}$$

Alternatively, $\mathsf{P}\{X \neq Y\} = \mathsf{P}\{Z = 1\} = \epsilon$. An interesting special case is $\epsilon = 1/2$, whence $\mathsf{P}\{X \neq Y\} = 1/2$, which is the worst possible (no information is sent), and

$$p_Y(0) = \frac{1}{2}p + \frac{1}{2}(1-p) = \frac{1}{2} = p_Y(1).$$

Therefore $Y \sim \mathrm{Bern}(1/2)$, regardless of the value of p. In this case, the bit sent X and the bit received Y are independent (check!).

4.3 PAIRS OF CONTINUOUS RANDOM VARIABLES

4.3.1 Joint and Marginal Densities

We say that X and Y are *jointly continuous* random variables if their joint cdf is continuous in both x and y. In this case, we can define their *joint pdf*, provided that it exists, as the function $f_{X,Y}(x, y)$ such that

$$F_{X,Y}(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u, v)\, du\, dv\,, \quad x, y \in \mathbb{R}.$$

If $F_{X,Y}(x, y)$ is differentiable in x and y, then

$$\begin{aligned}f_{X,Y}(x, y) &= \frac{\partial^2 F(x, y)}{\partial x \partial y}\\ &= \lim_{\Delta x, \Delta y \to 0} \frac{F_{X,Y}(x+\Delta x,\ y+\Delta y) - F_{X,Y}(x+\Delta x,\ y) - F_{X,Y}(x,\ y+\Delta y) + F_{X,Y}(x, y)}{\Delta x \Delta y}\\ &= \lim_{\Delta x, \Delta y \to 0} \frac{\mathsf{P}\{x < X \le x+\Delta x,\ y < Y \le y + \Delta y\}}{\Delta x \Delta y}.\end{aligned}$$

The joint pdf $f_{X,Y}(x, y)$ satisfies the following properties:

1. $f_{X,Y}(x, y) \ge 0$.
2. $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx\, dy = 1.$
3. The probability of any set $A \subset \mathbb{R}^2$ can be calculated by integrating the joint pdf over A:

$$\mathsf{P}\{(X, Y) \in A\} = \int\int_{(x,y)\in A} f_{X,Y}(x, y)\, dx\, dy.$$

The *marginal pdf* of X can be obtained from the joint pdf via the law of total probability:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy.$$

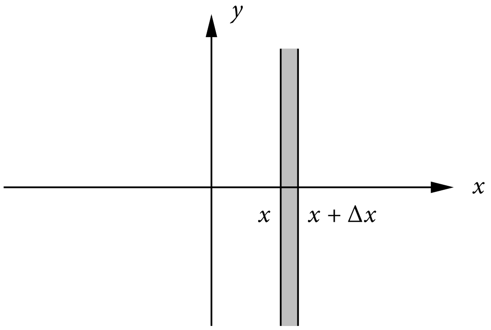
To see this, recall that

$$g(x) = \frac{d}{dx} \int_{-\infty}^{x} g(u)\, du$$

and consider

$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u, y)\, dy\, du$$

with $g(u) = \int_{-\infty}^{\infty} f_{X,Y}(u, y) dy$. We then differentiate $F_X(x)$ with respect to x. Alternatively, we can consider the following figure



and note that

$$\begin{aligned} f_X(x) &= \lim_{\Delta x \to 0} \frac{\mathsf{P}\{x < X \le x + \Delta x\}}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \lim_{\Delta y \to 0} \sum_{n=-\infty}^{\infty} \mathsf{P}\{x < X \le x + \Delta x,\ n\Delta y < Y \le (n+1)\Delta y\} \\ &= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy\, \Delta x \\ &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy. \end{aligned}$$

Example 4.3. Let $(X, Y) \sim f_{X,Y}(x, y)$, where

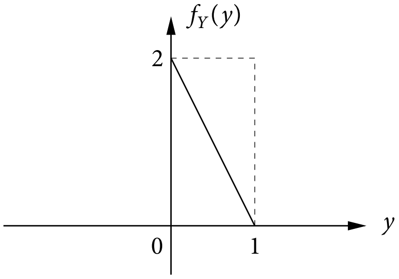
$$f X, Y(x, y) = \begin{cases} c & x \ge 0,\ y \ge 0,\ x + y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We find c, $f_Y(y)$, and $\mathsf{P}\{Y \le 2X\}$. Note first that

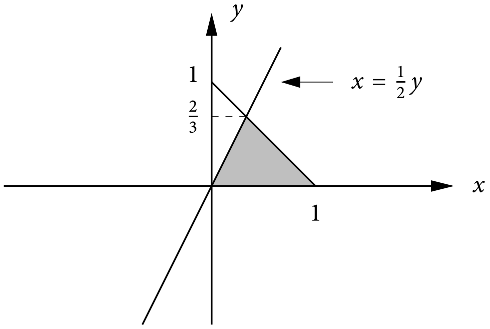
$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx\, dy = \int_0^1 \int_0^{1-y} c\, dx\, dy = c \int_0^1 (1 - y)\, dy = \tfrac{1}{2} c.$$

Hence, $c = 2$. To find $f_Y(y)$, we use the law of total probability

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \\ &= \begin{cases} \int_0^{(1-y)} 2\,dx & 0 \le y \le 1, \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 2(1-y) & 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$



To find the probability of the set $\{Y \le 2X\}$, we first sketch the set



From the figure we find that

$$\begin{aligned} \mathsf{P}\left\{X \ge \tfrac{1}{2}Y\right\} &= \int_{\left\{(x,y):x\ge\frac{1}{2}y\right\}} f_{X,Y}(x,y)\,dx\,dy \\ &= \int_0^{\frac{2}{3}} \int_{\frac{y}{2}}^{(1-y)} dx\,dy = \frac{2}{3}. \end{aligned}$$

4.3.2 Conditional Densities

Let X and Y be continuous random variables with joint pdf $f_{X,Y}(x,y)$. We wish to define

$$F_{Y|X}(y \mid X = x) = \mathsf{P}\{Y \le y \mid X = x\}.$$

We cannot define the above conditional probability as

$$\frac{\mathsf{P}\{Y \le y,\ X = x\}}{\mathsf{P}\{X = x\}}$$

since both the numerator and the denominator are equal to zero. Instead, we define the conditional probability for continuous random variables as a limit

$$\begin{aligned} F_{Y|X}(y|x) &= \lim_{\Delta x \to 0} \mathsf{P}\{Y \le y \,|\, x < X \le x + \Delta x\} \\ &= \lim_{\Delta x \to 0} \frac{\mathsf{P}\{x < X \le x + \Delta x,\ y \le y\}}{\mathsf{P}\{x < X \le x + \Delta x\}} \\ &= \lim_{\Delta x \to 0} \frac{\int_{-\infty}^{y} \int_{x}^{x+\Delta x} f_{X,Y}(u,v)\,du\,dv}{f_X(x)\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{\int_{-\infty}^{y} f_{X,Y}(x,v)\Delta x\,dv}{f_X(x)\Delta x} \\ &= \int_{-\infty}^{y} \frac{f_{X,Y}(x,v)}{f_X(x)}\,dv. \end{aligned}$$

By differentiating w.r.t. y, we thus define the conditional pdf as

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \qquad \text{if } f_X(x) \ne 0.$$

It can be readily checked that $f_{Y|X}(y|x)$ is a valid pdf and

$$F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(u|x)\,du$$

is a valid cdf for every x such that $f_X(x) > 0$. Sometimes we use the notation

$$Y \,|\, \{X = x\} \sim f_{Y|X}(y|x)$$

to denote Y has a conditional pdf $f_{Y|X}(y|x)$ given $X = x$.

Chain rule. $f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y|x) = f_Y(y) f_{X|Y}(x|y)$.

Independence. X and Y are independent iff

$$f_{X,Y}(x,y) = f_X(x) f_Y(y), \quad x, y \in \mathbb{R},$$

or equivalently,

$$f_{X|Y}(x|y) = f_X(x), \quad x \in \mathbb{R},\ f_Y(y) \ne 0.$$

Example 4.4. As in Example 4.3, let $f_{X,Y}(x, y)$ be defined as

$$f_{X,Y}(x, y) = \begin{cases} 2 & x \geq 0,\ y \geq 0,\ x + y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

We already know that

$$f_X(x) = \begin{cases} 2(1-x) & 0 \leq x \leq 1, \\ 0 & \text{otherwise} \end{cases}$$

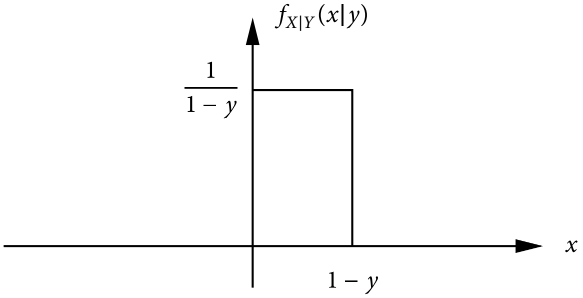
and

$$f_Y(y) = \begin{cases} 2(1-y) & 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $f_{X,Y}(1/3, 1/3) = 2$ and $f_X(1/3)f_Y(1/3) = (4/3)^2$, which implies that X and Y are *not* independent. Also note that

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \begin{cases} \dfrac{1}{1-y} & 0 \leq y < 1,\ 0 \leq x \leq 1-y, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, $X | \{Y = y\} \sim \text{Unif}[0, 1-y]$.



Law of total probability. For any event A,

$$\mathsf{P}(A) = \int_{-\infty}^{\infty} f_X(x)\, \mathsf{P}(A | X = x)\, dx.$$

Bayes rule. Given $f_X(x)$ and $f_{Y|X}(y|x)$, we can find

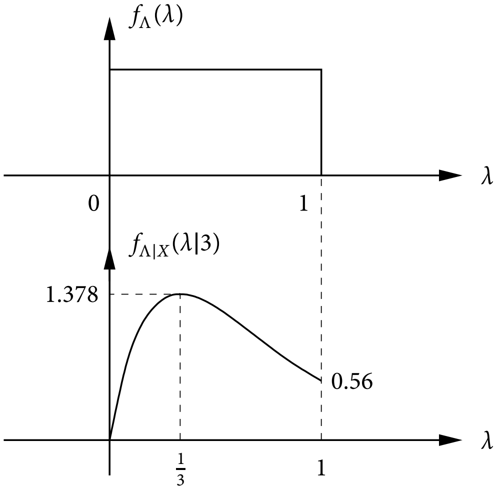
$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{Y|X}(y|x)}{f_Y(y)} f_X(x) \\ &= \frac{f_X(x) f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_{X,Y}(u, y)\, du} \\ &= \frac{f_X(x) f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(u) f_{Y|X}(y|u)\, du}. \end{aligned}$$

Example 4.5. Let $\Lambda \sim \text{Unif}[0,1]$ and the conditional pdf of X given $\{\Lambda = \lambda\}$ be

$$f_{X|\Lambda}(x|\lambda) = \lambda e^{-\lambda x}, \quad 0 < \lambda \le 1,$$

i.e., $X \,|\, \{\Lambda = \lambda\} \sim \text{Exp}(\lambda)$. Then, by the Bayes rule,

$$f_{\Lambda|X}(\lambda|3) = \frac{f_{X|\Lambda}(3|\lambda) f_\Lambda(\lambda)}{\int_0^1 f_\Lambda(u) f_{X|\Lambda}(3|u)\, du} = \begin{cases} \dfrac{\lambda e^{-3\lambda}}{\frac{1}{9}(1 - 4e^{-3})} & 0 < \lambda \le 1, \\ 0 & \text{otherwise.} \end{cases}$$



4.4 PAIRS OF MIXED RANDOM VARIABLES

Let X be a discrete random variable with pmf $p_X(x)$. For each x with $p_X(x) \ne 0$, conditioned on the event $\{X = x\}$, let Y be a continuous random variable with conditional cdf $F_{Y|X}(y|x)$ and conditional pdf

$$f_{Y|X}(y|x) = \frac{\partial}{\partial y} F_{Y|X}(y|x),$$

provided that the derivative is well-defined. Then, by the law of total probability,

$$\begin{aligned} F_Y(y) &= \mathsf{P}\{Y \le y\} \\ &= \sum_x \mathsf{P}\{Y \le y | X = x\} p_X(x) \\ &= \sum_x F_{Y|X}(y|x) p_X(x) \\ &= \sum_x \int_{-\infty}^{y} f_{Y|X}(v|x)\, dv p_X(x), \end{aligned}$$

and hence

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \sum_x f_{Y|X}(y|x)p_X(x).$$

The conditional pmf of X given $Y = y$ can be defined as a limit:

$$\begin{aligned}
p_{X|Y}(x|y) &= \lim_{\Delta y \to 0} \frac{\mathsf{P}\{X = x,\ y < Y \le y + \Delta y\}}{\mathsf{P}\{y < Y \le y + \Delta y\}} \\
&= \lim_{\Delta y \to 0} \frac{p_X(x)\,\mathsf{P}\{y < Y \le y + \Delta y | X = x\}}{\mathsf{P}\{y < Y \le y + \Delta y\}} \\
&= \lim_{\Delta y \to 0} \frac{p_X(x) f_{Y|X}(y|x)\Delta y}{f_Y(y)\Delta y} \\
&= \frac{f_{Y|X}(y|x)}{f_Y(y)}\, p_X(x).
\end{aligned}$$

Chain rule. $p_X(x)f_{Y|X}(y|x) = f_Y(y)p_{X|Y}(x|y)$.

Bayes rule. Given $p_X(x)$ and $f_{Y|X}(y|x)$,

$$p_{X|Y}(x|y) = \frac{p_X(x)f_{Y|X}(y|x)}{\sum_u p_X(u)f_{Y|X}(y|u)}.$$

Example 4.6. Let

$$X = \begin{cases} 1 & \text{with probability } p \\ 2 & \text{with probability } 1 - p, \end{cases}$$

and $Y|\{X = x\} \sim \text{Unif}[0, x]$, that is,

$$f_{Y|X}(y|x) = \begin{cases} 1/x & 0 \le y \le x \\ 0 & \text{otherwise.} \end{cases}$$

Then, by the law of total probability,

$$\begin{aligned}
f_Y(y) &= p f_{Y|X}(y|1) + (1 - p) f_{Y|X}(y|2) \\
&= \begin{cases} p + (1/2)(1 - p) & 0 \le y < 1 \\ (1/2)(1 - p) & 1 \le y \le 2 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

By the Bayes rule, for $y \in [0, 1)$,

$$\begin{aligned}
p_{X|Y}(1|y) &= \frac{p_X(1)f_{Y|X}(y|1)}{f_Y(y)} = \frac{2p}{1 + p}, \\
p_{X|Y}(2|y) &= \frac{p_X(2)f_{Y|X}(y|2)}{f_Y(y)} = \frac{1 - p}{1 + p},
\end{aligned}$$

and for $y \in [1, 2]$,

$$p_{X|Y}(1\,|\,y) = \frac{p_X(1) f_{Y|X}(y|1)}{f_Y(y)} = 0,$$
$$p_{X|Y}(2\,|\,y) = \frac{p_X(2) f_{Y|X}(y|2)}{f_Y(y)} = 1.$$

Example 4.7 (Additive Gaussian noise channel). Consider the communication channel in Figure 4.4. The signal transmitted is a binary random variable

$$X = \begin{cases} +\sqrt{P} & \text{w.p. } p, \\ -\sqrt{P} & \text{w.p. } 1 - p. \end{cases}$$

The received signal, also called the *observation*, is

$$Y = X + Z,$$

where $Z \sim \mathrm{N}(0, N)$ is additive noise and independent of X. First note that $Y \,|\, \{X = x\} \sim \mathrm{N}(x, N)$. To see this, consider

$$\begin{aligned} \mathsf{P}\{Y \le y \,|\, X = x\} &= \mathsf{P}\{X + Z \le y \,|\, X = x\} \\ &= \mathsf{P}\{Z \le y - x \,|\, X = x\} \\ &= \mathsf{P}\{Z \le y - x\}, \end{aligned}$$

which, by taking derivative w.r.t. y, implies that

$$f_{Y|X}(y|x) = f_Z(y - x) = \frac{1}{\sqrt{2\pi N}} e^{-\frac{(y-x)^2}{2N}}.$$

In other words,

$$Y \,|\, \{X = +\sqrt{P}\} \sim \mathrm{N}(+\sqrt{P}, N) \quad \text{and} \quad Y \,|\, \{X = -\sqrt{P}\} \sim \mathrm{N}(-\sqrt{P}, N).$$

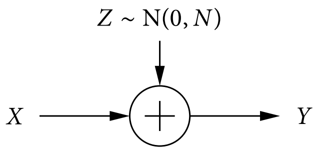


Figure 4.4. Additive Gaussian noise channel.

Next by the law of total probability,

$$\begin{aligned} f_Y(y) &= p_X(+\sqrt{P}) f_{Y|X}(y|+\sqrt{P}) + p_X(-\sqrt{P}) f_{Y|X}(y|-\sqrt{P}) \\ &= p\frac{1}{\sqrt{2\pi N}} e^{-\frac{(y-\sqrt{P})^2}{2N}} + (1-p)\frac{1}{\sqrt{2\pi N}} e^{-\frac{(y+\sqrt{P})^2}{2N}}. \end{aligned}$$

Finally, by the Bayes rule,

$$\begin{aligned} p_{X|Y}(+\sqrt{P}|y) &= \frac{\frac{p}{\sqrt{2\pi N}} e^{-\frac{(y-\sqrt{P})^2}{2N}}}{\frac{p}{\sqrt{2\pi N}} e^{-\frac{(y-\sqrt{P})^2}{2N}} + \frac{(1-p)}{\sqrt{2\pi N}} e^{-\frac{(y+\sqrt{P})^2}{2N}}} \\ &= \frac{pe^{\frac{y\sqrt{P}}{N}}}{pe^{\frac{y\sqrt{P}}{N}} + (1-p)e^{-\frac{y\sqrt{P}}{N}}} \end{aligned}$$

and

$$p_{X|Y}(-\sqrt{P}|y) = \frac{(1-p)e^{\frac{-y\sqrt{P}}{N}}}{pe^{\frac{y\sqrt{P}}{N}} + (1-p)e^{-\frac{y\sqrt{P}}{N}}}.$$

4.5 APPLICATION: SIGNAL DETECTION

Consider the general digital communication system depicted in Figure 4.5, where the signal sent is $X \sim p_X(x)$, $x \in \mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, and the observation (received signal) is

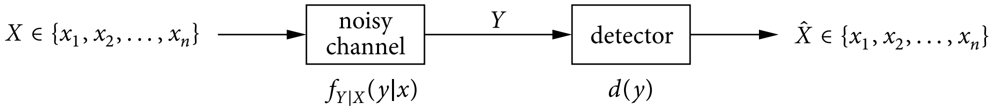

$X \in \{x_1, x_2, \ldots, x_n\}$
noisy channel
$f_{Y|X}(y|x)$
Y
detector
$d(y)$
$\hat{X} \in \{x_1, x_2, \ldots, x_n\}$

Figure 4.5. The general digital communication system.

$$Y \,|\, \{X = x\} \sim f_{Y|X}(y|x).$$

A detector is a mapping $d : \mathbb{R} \to \mathcal{X}$ that generates an estimate $\hat{X} = d(Y) \in \mathcal{X}$ of the input signal X. We wish to find an optimal detector $d^*(y)$ that minimizes the *probability of error*

$$P_e := \mathsf{P}\{\hat{X} \neq X\} = \mathsf{P}\{d(Y) \neq X\}.$$

Suppose that there is no observation and we would like to find the best guess d^* of X, that is,

$$d^* = \arg\min_d \mathsf{P}\{d \neq X\}.$$

Clearly, the best guess is the one with the largest probability, that is,

$$d^* = \arg\max_x p_X(x). \tag{4.1}$$

This simple observation continues to hold with the observation $Y = y$. Define the *maximum a posteriori probability* (MAP) detector as

$$d^*(y) = \arg\max_x p_{X|Y}(x\,|\,y). \tag{4.2}$$

Then, by (4.1)

$$\mathsf{P}\{d^*(y) \neq X \mid Y = y\} \leq \mathsf{P}\{d(y) \neq X \mid Y = y\}$$

for any detector $d(y)$. Consequently,

$$\begin{aligned}\mathsf{P}\{d^*(Y) \neq X\} &= \int \mathsf{P}\{d^*(y) \neq X \mid Y = y\} \\ &\leq \int \mathsf{P}\{d(y) \neq X \mid Y = y\} \\ &= \mathsf{P}\{d(Y) \neq X\},\end{aligned}$$

and the MAP detector minimizes the probability of error P_e.

When X is uniform on $\{x_1, x_2, \dots, x_n\}$, that is, $p_X(x_1) = p_X(x_2) = \cdots = p_X(x_n) = 1/n$,

$$p_{X|Y}(x\,|\,y) = \frac{p_X(x) f_{Y|X}(y|x)}{f_Y(y)} = \frac{f_{Y|X}(y|x)}{n f_Y(y)},$$

which depends on x only through the *likelihood* $f_{Y|X}(y|x)$ for a given y. Then, the MAP detection rule in (4.2) simplifies as the *maximum likelihood* (ML) detection rule

$$d^*(y) = \arg\max_x f_{Y|X}(y|x).$$

Example 4.8. We revisit Example 4.6. The MAP detector is

$$d^*(y) = \begin{cases} 1 & \text{if } p_{X|Y}(1|y) > p_{X|Y}(2|y), \\ 2 & \text{otherwise,} \end{cases}$$

which simplifies as

$$d^*(y) = \begin{cases} 1 & \text{if } y \in [0,1), \\ 2 & \text{otherwise.} \end{cases}$$

The *minimum* probability of error is

$$\begin{aligned}\mathsf{P}\{d^*(Y) \neq X\} &= \mathsf{P}\{X = 1\}\,\mathsf{P}\{d^*(Y) \neq 1 \mid X = 1\} + \mathsf{P}\{X = 2\}\,\mathsf{P}\{d^*(Y) \neq 2 \mid X = 2\} \\ &= \mathsf{P}\{X = 1\}\,\mathsf{P}\{1 \leq Y \leq 2 \mid X = 1\} + \mathsf{P}\{X = 2\}\,\mathsf{P}\{0 \leq Y < 1 \mid X = 2\} \\ &= p \cdot 0 + (1-p) \cdot \frac{1}{2} \\ &= \frac{1}{2}(1+p).\end{aligned}$$

Example 4.9. We revisit the additive Gaussian noise channel in Example 4.7 with signal

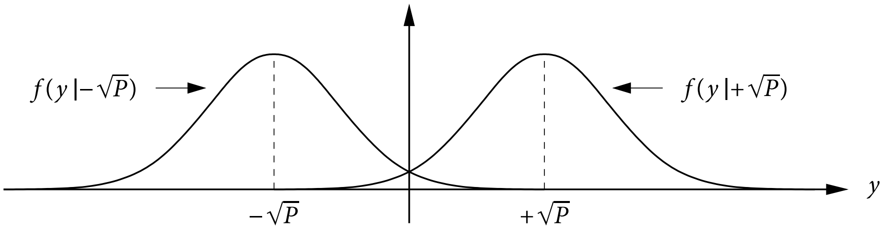
$$X = \begin{cases} +\sqrt{P} & \text{w.p. } \frac{1}{2} \\ -\sqrt{P} & \text{w.p. } \frac{1}{2} \end{cases}$$

independent noise $Z \sim \mathrm{N}(0, N)$, and output $Y = X + Z$. The MAP detector is

$$d^*(y) = \begin{cases} +\sqrt{P} & \text{if } p_{X|Y}(+\sqrt{P}|y) > p_{X|Y}(-\sqrt{P}|y), \\ -\sqrt{P} & \text{otherwise.} \end{cases}$$

Since the two signals are equally likely, the MAP detection rule reduces to the ML detection rule

$$d^*(y) = \begin{cases} +\sqrt{P} & \text{if } f_{Y|X}(y| + \sqrt{P}) > f_{Y|X}(y| - \sqrt{P}), \\ -\sqrt{P} & \text{otherwise.} \end{cases}$$



From the figure above and using the Gaussian pdf, the MAP/ML detector reduces to the *minimum distance* detector

$$d^*(y) = \begin{cases} +\sqrt{P} & (y - \sqrt{P})^2 < (y - (-\sqrt{P}))^2, \\ -\sqrt{P} & \text{otherwise,} \end{cases}$$

which further simplifies to

$$d^*(y) = \begin{cases} +\sqrt{P} & y > 0, \\ -\sqrt{P} & y \leq 0. \end{cases}$$

Now the *minimum* probability of error is

$$\begin{aligned} P_e^* &= \mathsf{P}\{d^*(Y) \neq X\} \\ &= \mathsf{P}\{X = \sqrt{P}\}\,\mathsf{P}\{d^*(Y) = -\sqrt{P} \,|\, X = \sqrt{P}\} \\ &\quad + \mathsf{P}\{X = -\sqrt{P}\}\,\mathsf{P}\{d^*(Y) = \sqrt{P} \,|\, X = -\sqrt{P}\} \\ &= \frac{1}{2}\mathsf{P}\{Y \leq 0 \,|\, X = \sqrt{P}\} + \frac{1}{2}\mathsf{P}\{Y > 0 \,|\, X = -\sqrt{P}\} \\ &= \frac{1}{2}\mathsf{P}\{Z \leq -\sqrt{P}\} + \frac{1}{2}\mathsf{P}\{Z > \sqrt{P}\} \\ &= Q\Big(\sqrt{\frac{P}{N}}\Big). \end{aligned}$$

Note that the probability of error is a decreasing function of P/N, which is often called the *signal-to-noise ratio* (SNR).

4.6 FUNCTIONS OF TWO RANDOM VARIABLES

Let $(X, Y) \sim f(x, y)$ and let $g(x, y)$ be a differentiable function. To find the pdf of $Z = g(X, Y)$, we first find the inverse image of $\{Z \leq z\}$ to compute its probability expressed as a function of z, i.e., $F_Z(z)$, and then take the derivative.

Example 4.10. Suppose that $X \sim f_X(x)$ and $Y \sim f_Y(y)$ are independent. Let $Z = X + Y$. Then

$$\begin{aligned} F_Z(z) &= \mathsf{P}\{Z \leq z\} \\ &= \mathsf{P}\{X + Y \leq z\} \\ &= \int_{-\infty}^{\infty} \mathsf{P}\{X + Y \leq z \,|\, X = x\} f_X(x)\, dx \\ &= \int_{-\infty}^{\infty} \mathsf{P}\{x + Y \leq z \,|\, X = x\} f_X(x)\, dx \\ &= \int_{-\infty}^{\infty} \mathsf{P}\{Y \leq z - x\} f_X(x)\, dx \\ &= \int_{-\infty}^{\infty} F_Y(z - x) f_X(x)\, dx. \end{aligned}$$

Taking derivative w.r.t. z, we have

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(z - x) f_X(x)\, dx,$$

which is the *convolution* of $f_X(x)$ and $f_Y(y)$. For example, if $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ are independent, then it can be readily checked (see Problem 4.20) that

$$Z = X + Y \sim \mathrm{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

A similar result also holds for the sum of two independent discrete random variables (replacing pdfs with pmfs and integrals with sums). For example, if $X \sim \mathrm{Poisson}(\lambda_1)$ and $Y \sim \mathrm{Poisson}(\lambda_2)$ are independent, then

$$Z = X + Y \sim \mathrm{Poisson}(\lambda_1) * \mathrm{Poisson}(\lambda_2) = \mathrm{Poisson}(\lambda_1 + \lambda_2).$$

The property that the sum of two independent random variables with the same distribution has the same distribution, which is obeyed by Gaussian and Poisson random variables, is referred to as *infinite divisibility*. For example, a $\mathrm{Poisson}(\lambda)$ random variable can be written as the sum of *any* number of independent $\mathrm{Poisson}(\lambda_i)$ random variables, as long as $\sum_i \lambda_i = \lambda$.

It is often easier to work with the cdf first to find the pdf of a function of X and Y (especially when $g(x, y)$ is not differentiable).

Example 4.11 (Minimum and maximum). Let $X \sim f_X(x)$ and $Y \sim f_Y(y)$ be independent. Define

$$U = \max\{X,\ Y\} \quad \text{and} \quad V = \min\{X,\ Y\}.$$

To find the pdf of U, we first find its cdf

$$F_U(u) = \mathsf{P}\{U \le u\} = \mathsf{P}\{X \le u,\ Y \le u\} = F_X(u)F_Y(u).$$

Using the product rule for derivatives,

$$f_U(u) = f_X(u)F_Y(u) + f_Y(u)F_X(u).$$

Now to find the pdf of V, consider

$$1 - F_V(v) = \mathsf{P}\{V > v\} = \mathsf{P}\{X > v,\ Y > v\} = (1 - F_X(v))(1 - F_Y(v)).$$

Thus, by taking derivatives,

$$f_V(v) = f_X(v) + f_Y(v)\ -\ f_X(v)F_Y(v)\ -\ f_Y(v)F_X(v).$$

The joint pdf of (U, V) can be found by similar arguments; see Problem 4.21.

PROBLEMS

4.1. *Geometric with conditions.* Let X be a geometric random variable with pmf

$$p_X(k) = p(1-p)^{k-1}, \quad k = 1, 2, \ldots.$$

Find and plot the conditional pmf $p_X(k|A) = \mathsf{P}\{X = k | X \in A\}$ if:

(a) $A = \{X > m\}$ where m is a positive integer.

(b) $A = \{X < m\}$.

(c) $A = \{X \text{ is an even number}\}$.

Comment on the shape of the conditional pmf of part (a).

4.2. *Conditional cdf.* Let A be a nonzero probability event A. Show that

(a) $\mathsf{P}(A) = \mathsf{P}(A|X \le x)F_X(x)\ +\ \mathsf{P}(A|X > x)(1 - F_X(x))$.

(b) $F_X(x|A) = \dfrac{\mathsf{P}(A|X \le x)}{\mathsf{P}(A)} F_X(x)$.

4.3. *Joint cdf or not.* Consider the function

$$G(x, y) = \begin{cases} 1 & \text{if } x + y \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

Can G be a joint cdf for a pair of random variables? Justify your answer.

4.4. *Time until the n-th arrival.* Let the random variable $N(t)$ be the number of packets arriving during time $(0, t]$. Suppose $N(t)$ is Poisson with pmf

$$p_N(n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \quad \text{for } n = 0, 1, 2, \ldots .$$

Let the random variable Y be the time to get the n-th packet. Find the pdf of Y.

4.5. *Diamond distribution.* Consider the random variables X and Y with the joint pdf

$$f_{X,Y}(x, y) = \begin{cases} c, & \text{if } |x| + |y| \le 1/\sqrt{2}, \\ 0, & \text{otherwise,} \end{cases}$$

where c is a constant.

(a) Find c.

(b) Find $f_X(x)$ and $f_{X|Y}(x|y)$.

(c) Are X and Y independent random variables? Justify your answer.

(d) Define the random variable $Z = (|X| + |Y|)$. Find the pdf $f_Z(z)$.

4.6. *Coin with random bias.* You are given a coin but are not told what its bias (probability of heads) is. You are told instead that the bias is the outcome of a random variable $P \sim \mathrm{U}[0, 1]$. To get more information about the coin bias, you flip it independently 10 times. Let X be the number of heads you get. Thus $X \sim \mathrm{Binom}(10, P)$. Assuming that $X = 9$, find and sketch the *a posteriori* probability of P, i.e., $f_{P|X}(p|9)$.

4.7. *First available teller.* Consider a bank with two tellers. The service times for the tellers are independent exponentially distributed random variables $X_1 \sim \mathrm{Exp}(\lambda_1)$ and $X_2 \sim \mathrm{Exp}(\lambda_2)$, respectively. You arrive at the bank and find that both tellers are busy but that nobody else is waiting to be served. You are served by the first available teller once he/she is free.

(a) What is the probability that you are served by the first teller?

(b) Let the random variable Y denote your waiting time. Find the pdf of Y.

4.8. *Optical communication channel.* Let the signal input to an optical channel be given by

$$X = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ 10 & \text{with probability } \frac{1}{2}. \end{cases}$$

The conditional pmf of the output of the channel $Y|\{X = 1\} \sim \mathrm{Poisson}(1)$, i.e., Poisson with intensity $\lambda = 1$, and $Y|\{X = 10\} \sim \mathrm{Poisson}(10)$.

(a) Show that the MAP rule reduces to

$$D(y) = \begin{cases} 1, & y < y^* \\ 10, & \text{otherwise.} \end{cases}$$

(b) Find y^* and the corresponding probability of error.

4.9. *Iocane or Sennari.* An absent-minded chemistry professor forgets to label two identically looking bottles. One bottle contains a chemical named "Iocane" and the other bottle contains a chemical named "Sennari". It is well known that the radioactivity level of "Iocane" has the U[0, 1] distribution, while the radioactivity level of "Sennari" has the Exp(1) distribution.

(a) Let X be the radioactivity level measured from one of the bottles. What is the optimal decision rule (based on the measurement X) that maximizes the chance of correctly identifying the content of the bottle?

(b) What is the associated probability of error?

4.10. *Independence.* Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be two independent discrete random variables.

(a) Show that any two events $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$ are independent.

(b) Show that any two functions of X and Y separately are independent; that is, if $U = g(X)$ and $V = h(Y)$ then U and V are independent.

4.11. *Family planning.* Alice and Bob choose a number X at random from the set $\{2, 3, 4\}$ (so the outcomes are equally probable). If the outcome is $X = x$, they decide to have children until they have a girl or x children, whichever comes first. Assume that each child is a girl with probability 1/2 (independent of the number of children and gender of other children). Let Y be the number of children they will have.

(a) Find the conditional pmf $p_{Y|X}(y|x)$ for all possible values of x and y.

(b) Find the pmf of Y.

4.12. *Radar signal detection.* The signal for a radar channel $S = 0$ if there is no target and a random variable $S \sim N(0, P)$ if there is a target. Both occur with equal probability. Thus

$$S = \begin{cases} 0 & \text{w.p. } 1/2, \\ X \sim \mathrm{N}(0, P) & \text{w.p. } 1/2. \end{cases}$$

The radar receiver observes $Y = S + Z$, where the noise $Z \sim \mathrm{N}(0, N)$ is independent of S. Find the optimal decoder for deciding whether $S = 0$ or $S = X$ and its probability of error? Provide your answer in terms of intervals of y and provide the boundary points of the intervals in terms of P and N.

4.13. *Z channel.* Suppose that the signal X is drawn as

$$X = \begin{cases} 1 & \text{with probability } 1/2, \\ 0 & \text{with probability } 1/2, \end{cases}$$

and the conditional pmf $p_{Y|X}(y|x)$ of Y given X is specified by

$$p_{Y|X}(1|1) = 1,$$
$$p_{Y|X}(1|0) = 1/2.$$

(a) Find $p_{Y|X}(0|1)$ and $p_{Y|X}(0|0)$.

(b) Find the conditional pmf $p_{X|Y}(x|y)$ of X given Y (i.e., find $p_{X|Y}(1|1)$, $p_{X|Y}(1|0)$, $p_{X|Y}(0|1)$, and $p_{X|Y}(0|0)$).

(c) Find the optimal decoder $d(y)$ that minimizes the probability of error $\mathsf{P}\{X \neq d(Y)\}$.

(d) Find the associated probability of error.

4.14. *Ternary signaling.* Let the signal S be a random variable defined as follows:

$$S = \begin{cases} -1 & \text{with probability } \frac{1}{3} \\ \ \ 0 & \text{with probability } \frac{1}{3} \\ +1 & \text{with probability } \frac{1}{3}. \end{cases}$$

The signal is sent over a channel with additive Laplacian noise Z, i.e., Z is a Laplacian random variable with pdf

$$f_Z(z) = \frac{\lambda}{2} e^{-\lambda|z|}, \quad -\infty < z < \infty\,.$$

The signal S and the noise Z are assumed to be independent and the channel output is their sum $Y = S + Z$.

(a) Find $f_{Y|S}(y|s)$ for $s = -1,\ 0,\ +1$. Sketch the conditional pdfs on the same graph.

(b) Find the optimal decoding rule $D(Y)$ for deciding whether S is -1, 0 or $+1$. Give your answer in terms of ranges of values of Y.

(c) Find the probability of decoding error for $D(y)$ in terms of λ.

4.15. *Signal or no signal.* Consider a communication system that is operated only from time to time. When the communication system is in the "normal" mode (denoted by $M = 1$), it transmits a random signal $S = X$ with

$$X = \begin{cases} +1, & \text{with probability } 1/2, \\ -1, & \text{with probability } 1/2. \end{cases}$$

When the system is in the "idle" mode (denoted by $M = 0$), it does not transmit any signal ($S = 0$). Both normal and idle modes occur with equal probability. Thus

$$S = \begin{cases} X, & \text{with probability } 1/2, \\ 0, & \text{with probability } 1/2. \end{cases}$$

The receiver observes $Y = S + Z$, where the ambient noise $Z \sim \mathrm{U}[-1, 1]$ is independent of S.

(a) Find and sketch the conditional pdf $f_{Y|M}(y|1)$ of the receiver observation Y given that the system is in the normal mode.

(b) Find and sketch the conditional pdf $f_{Y|M}(y|0)$ of the receiver observation Y given that the system is in the idle mode.

(c) Find the optimal decoder $d^*(y)$ for deciding whether the system is normal or idle. Provide the answer in terms of intervals of y.

(d) Find the associated probability of error.

4.16. *Fair coins.* We are given two coins: Coin 1 with bias (=probability of heads) 1/2 and Coin 2 with random bias $P \sim \text{Unif}[0, 1]$. We pick one at random and flip it three times independently. The value of the bias does not change during the sequence of tosses. Let X be the number of heads.

(a) Find the conditional pmf of X given that Coin 1 is selected.

(b) Find the conditional pmf of X given that Coin 2 is selected.

(c) Find the optimal decision rule $D(x) \in \{1, 2\}$ for deciding which coin is flipped such that the probability of decision error is minimized.

(d) Find the associated probability of error.

4.17. *Additive exponential noise channel.* A device has two equally likely states $S = 0$ and $S = 1$. When it is inactive ($S = 0$), it transmits $X = 0$. When it is active ($S = 1$), it transmits $X \sim \text{Exp}(1)$. Now suppose the signal is observed through the additive exponential noise channel with output

$$Y = X + Z,$$

where $Z \sim \text{Exp}(2)$ is independent of (X, S). One wishes to decide whether the device is active or not.

(a) Find $f_{Y|S}(y|0)$.

(b) Find $f_{Y|S}(y|1)$.

(c) Find $f_Y(y)$.

(d) Find $p_{S|Y}(0|y)$ and $p_{S|Y}(1|y)$.

(e) Find the decision rule $d(y)$ that minimizes the probability of error

$$\mathsf{P}(S \neq d(Y)).$$

(f) Find the corresponding probability of error.

(Hint: Recall that $Z \sim \text{Exp}(\lambda)$ means that its pdf is $f_Z(z) = \lambda e^{-\lambda z}$, $z \geq 0$.)

4.18. *Two independent uniform random variables.* Let X and Y be independently and uniformly drawn from the interval $[0, 1]$.

(a) Find the pdf of $U = \max(X, Y)$.

(b) Find the pdf of $V = \min(X, Y)$.

(c) Find the pdf of $W = U - V$.

(d) Find the pdf of $Z = (X + Y) \mod 1$.

(e) Find the probability $\mathsf{P}\{|X - Y| \geq 1/2\}$.

4.19. *Two independent Gaussian random variables.* Let X and Y be independent Gaussian random variables, both with zero mean and unit variance. Find the pdf of $|X - Y|$.

4.20. *Two independent Gaussian random variables.* Let $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ are independent. Show that

$$Z = X + Y \sim \mathrm{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

4.21. *Maximum and minimum.* Let $X \sim f_X(x)$ and $Y \sim f_Y(y)$ be independent.

$$U = \max\{X,\ Y\} \quad \text{and} \quad V = \min\{X,\ Y\}.$$

Find the joint pdf of (U, V).

4.22. *Functions of exponential random variables.* Let X and Y be independent exponentially distributed random variables with the same parameter λ. Define the following three functions of X and Y:

$$U = \max(X, Y)\,, \quad V = \min(X, Y)\,, \quad W = U - V\,.$$

(a) Find the joint pdf of U and V.

(b) Find the joint pdf of V and W. Are they independent?

Hint: You can solve part (b) either directly by finding the joint cdf or by expressing the joint pdf in terms of $f_{U,V}(u, v)$ and using the result of part (a).

4.23. *Maximal correlation.* Let $(X, Y) \sim F_{X,Y}(x, y)$ be a pair of random variables.

(a) Show that

$$F_{X,Y}(x, y) \leq \min\{F_X(x), F_Y(y)\}.$$

Now let $F(x)$ and $G(y)$ be continuous and invertible cdfs and let $X \sim F(x)$.

(b) Find the cdf of

$$Y = G^{-1}(F(X)).$$

(c) Show that

$$F_{X,Y}(x, y) = \min\{F(x),\ G(y)\}.$$

LECTURE 5

Expectation

5.1 DEFINITION AND PROPERTIES

Let $X \in \mathcal{X}$ be a discrete random variable with pmf $p_X(x)$ and let $g(x)$ be a function of x. The *expectation* (or *expected value* or *mean*) of $g(X)$ is defined as

$$\mathsf{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) p_X(x).$$

For a continuous random variable $X \sim f_X(x)$, the expected value of $g(X)$ is defined as

$$\mathsf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx.$$

The expectation operation $\mathsf{E}[\cdot]$ satisfies the following properties:

1. $\mathsf{E}[c] = c$ for every constant c.
2. If $g(X) \geq 0$ w.p. 1, then $\mathsf{E}[g(X)] \geq 0$.
3. *Linearity.* For any constant a and functions $g_1(x)$ and $g_2(x)$,

$$\mathsf{E}[a g_1(X) + g_2(X)] = a\, \mathsf{E}[g_1(X)] + \mathsf{E}[g_2(X)].$$

By considering $Y = g(X)$ as a random variable on its own, we can compute the same expectation.

Fundamental theorem of expectation. If $X \sim p_X(X)$ and $Y = g(X) \sim p_Y(y)$, then

$$\mathsf{E}[Y] = \sum_{y \in \mathcal{Y}} y p_Y(y) = \sum_{x \in \mathcal{X}} g(x) p_X(x) = \mathsf{E}[g(X)].$$

Similarly, if $X \sim f_X(x)$ and $Y = g(X) \sim f_Y(y)$, then

$$\mathsf{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y)\, dy = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx = \mathsf{E}[g(X)].$$

We prove the theorem for the discrete case. Consider

$$\begin{aligned}\mathsf{E}[Y] &= \sum_y y p_Y(y)\\ &= \sum_y y \sum_{\{x:\, g(x)=y\}} p_X(x)\\ &= \sum_y \sum_{\{x:\, g(x)=y\}} y p_X(x)\\ &= \sum_y \sum_{\{x:\, g(x)=y\}} g(x) p_X(x)\\ &= \sum_x g(x) p_X(x).\end{aligned}$$

Thus, $\mathsf{E}[Y] = \mathsf{E}[g(X)]$ can be found using either $p_X(x)$ or $p_Y(y)$. It is often much easier to use $p_X(x)$ than to first find $p_Y(y)$ and then find $\mathsf{E}[Y]$.

We already know that a random variable is completely specified, that is, any probability of a Borel set involving the random variable can be determined, by its cumulative distribution function (or its pmf and pdf in discrete and continuous cases, respectively). As a simple summary of the random variable, however, its expectation has several applications.

1. Expectation can be used to bound or estimate probabilities of interesting events, as we will see in Section 5.3.
2. Expectation provides the *optimal estimate* of a random variable under the mean square error criterion, as we will see in Section 5.5.
3. It is far easier to estimate the expectation of a random variable from data than to estimate its distribution, as we will see in Lecture #7.

5.2 MEAN AND VARIANCE

The *first moment* (or *mean*) of $X \sim f_X(x)$ is the expectation

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

The following trick is useful for computing the expectation of a nonnegative random variable. If $X \geq 0$ is continuous, then

$$\begin{aligned}\mathsf{E}[X] &= \int_0^{\infty} u f_X(u)\, du\\ &= \int_0^{\infty}\int_0^{u} dx f_X(u)\, du\\ &= \int_0^{\infty}\int_x^{\infty} f_X(u)\, du\, dx\\ &= \int_0^{\infty} 1 - F_X(x)\, dx.\end{aligned}$$

The same identity also follows by integration by parts $\int \phi\psi' = \phi\psi - \int \phi'\psi$ with $\phi = x$ and $\psi = 1 - F_X(x)$. Similarly, if $X \geq 0$ is discrete, then

$$\mathsf{E}[X] = \sum_{k=0}^{\infty}(1 - F_X(k)).$$

Let

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

Then, the expectation of the *indicator variable* is

$$\mathsf{E}[\mathbf{1}_A(X)] = \int_{-\infty}^{\infty} \mathbf{1}_A(x) f_X(x)\, dx = \int_A f_X(x)\, dx = \mathsf{P}\{X \in A\}.$$

The *second moment* (or *mean square* or *average power*) of X is

$$\mathsf{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)\, dx.$$

The *variance* of X is

$$\begin{aligned} \operatorname{Var}(X) &= \mathsf{E}[(X - \mathsf{E}(X))^2] \\ &= \mathsf{E}[X^2 - 2X\,\mathsf{E}(X) + (\mathsf{E}(X))^2] \\ &= \mathsf{E}[X^2] - 2(\mathsf{E}[X])^2 + (\mathsf{E}[X])^2 \\ &= \mathsf{E}[X^2] - (\mathsf{E}[X])^2. \end{aligned}$$

The *standard deviation* of X is defined as $\sigma_X = \sqrt{\operatorname{Var}(X)}$, i.e., $\operatorname{Var}(X) = \sigma_X^2$.

Example 5.1. We find $\mathsf{E}[X]$, $\mathsf{E}[X^2]$, and $\operatorname{Var}(X)$ for $(X, Y) \sim f_{X,Y}(x, y)$ where

$$f_{X,Y}(x, y) = \begin{cases} 2 & x \geq 0,\ y \geq 0,\ x + y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Consider

$$\begin{aligned} \mathsf{E}[X] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y)\, dx\, dy \\ &= \int_0^1 \int_0^{1-x} 2x\, dy\, dx \\ &= 2\int_0^1 (1 - x)x\, dx \\ &= 2\left(\frac{1}{2} - \frac{1}{3}\right) = \frac{1}{3} \end{aligned}$$

and

$$\mathsf{E}[X^2] = 2\int_0^1 (1 - x)x^2\, dx = 2\left(\frac{1}{3} - \frac{1}{4}\right) = \frac{1}{6}.$$

Hence,

$$\operatorname{Var}(X) = \mathsf{E}[X^2] - (\mathsf{E}[X])^2 \frac{1}{6} - (\tfrac{1}{3})^2 = \frac{1}{6} - \frac{1}{9} = \frac{1}{18}.$$

The following useful identities are direct consequences of the linearity of expectation:

$$\mathsf{E}(aX + b) = a\,\mathsf{E}(X) + b,$$
$$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X).$$

Table 5.1 summarizes the mean and variance of famous random variables.

Random Variable	Mean	Variance
$\mathrm{Bern}(p)$	p	$p(1-p)$
$\mathrm{Geom}(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$\mathrm{Binom}(n, p)$	np	$np(1-p)$
$\mathrm{Poisson}(\lambda)$	λ	λ
$\mathrm{Unif}[\,a, b\,]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathrm{Exp}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$\mathrm{N}(\mu, \sigma^2)$	μ	σ^2

Table 5.1. The mean and variance of common random variables.

Remark 5.1. Expectation can be infinite. For example, consider

$$f_X(x) = \begin{cases} 1/x^2 & 1 \le x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\mathsf{E}[X] = \int_1^\infty \frac{x}{x^2}\,dx = \infty.$$

Remark 5.2. Expectation may not exist. To find conditions under which expectation exists, consider

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx = -\int_{-\infty}^{0} |x| f_X(x)\,dx + \int_0^\infty |x| f_X(x)\,dx\,,$$

so either $\int_{-\infty}^0 |x| f_X(x)\,dx$ or $\int_0^\infty |x| f_X(x)\,dx$ must be finite.

Example 5.2. The *standard Cauchy* random variable has the pdf

$$f_X(x) = \frac{1}{\pi(1+x^2)}$$

Since both $\int_{-\infty}^{0} |x| f_X(x)\,dx$ and $\int_{0}^{\infty} |x| f_X(x)\,dx$ are infinite, its mean does not exist! (The second moment of the Cauchy is $\mathsf{E}[X^2] = \infty$, so it exists.)

5.3 INEQUALITIES

In many cases we do not know the distribution of a random variable X, but wish to find the probability of an event such as $\{X > a\}$ or $\{|X - \mathsf{E}(X)| > a\}$. The Markov and Chebyshev inequalities provide upper bounds on the probabilities of such events in terms of the mean and variance of the random variable.

Markov inequality. Let $X \geq 0$ be a random variable with finite mean. Then for any $a > 1$,

$$\mathsf{P}\{X \geq a\,\mathsf{E}[X]\} \leq \frac{1}{a}.$$

Example 5.3. If the average age in the San Diego is 35, then at most half of the population is 70 or older.

To prove the Markov inequality, let $A = \{x \geq a\,\mathsf{E}(X)\}$ and consider the indicator function $\mathbf{1}_A(x)$. As illustrated in Figure 5.1,

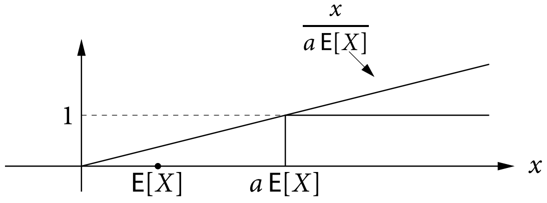


Figure 5.1. Proof of the Markov inequality.

$$\mathbf{1}_A(x) \leq \frac{x}{a\,\mathsf{E}[X]}.$$

Since $\mathsf{E}(\mathbf{1}_A(X)) = \mathsf{P}\{X \geq a\,\mathsf{E}[X]\}$, taking the expectations of both sides establishes the inequality.

The Markov inequality can be *very* loose. For example, if $X \sim \text{Exp}(1)$, then

$$\mathsf{P}\{X \geq 10\} = e^{-10} \approx 4.54 \times 10^{-5}.$$

The Markov inequality yields

$$\mathsf{P}\{X \geq 10\} \leq \tfrac{1}{10},$$

which is very pessimistic. But it is the *tightest* possible inequality on $\mathsf{P}\{X \geq a\,\mathsf{E}[X]\}$ when we are given only $\mathsf{E}[X]$. To show this, note that the inequality is tight for

$$X = \begin{cases} a\,\mathsf{E}[X] & \text{w.p. } 1/a, \\ 0 & \text{w.p. } 1 - 1/a. \end{cases}$$

In Example 5.3, if half of the population is 0 year old and half other population is 70 years old, then the average age is 35 and the Markov inequality is tight.

Chebyshev inequality. Let X be a random variable with finite mean $\mathsf{E}[X]$ and variance σ_X^2. Then for any $a > 1$,

$$\mathsf{P}\{|X - \mathsf{E}[X]| \geq a\sigma_X\} \leq \frac{1}{a^2}.$$

Example 5.4. Let X be a device parameter in an integrated circuit (IC) with known mean and variance. The IC is out-of-spec if X is more than, say, $3\sigma_X$ away from its mean. Then, by the Chebyshev inequality, the fraction of out-of-spec ICs, namely, $\mathsf{P}\{|X - \mathsf{E}(X)| \geq 3\sigma_X\}$ is no larger than $1/9$.

The proof of the Chebyshev inequality uses the Markov inequality (which is a slight twist from the teacher–student relationship between Prof. Pafnuty Chebyshev and his student Andrey Markov at Saint Petersburg University in Russia). Define the random variable $Y = (X - \mathsf{E}[X])^2 \geq 0$. Since $\mathsf{E}[Y] = \sigma_X^2$, the Markov inequality implies that

$$\mathsf{P}\{|X - \mathsf{E}(X)| \geq a\sigma_X\} = \mathsf{P}\{Y \geq a^2\sigma_X^2\} \leq \frac{1}{a^2}.$$

The Chebyshev inequality can be very loose. Let $X \sim \mathrm{N}(0, 1)$. Then, by the Chebyshev inequality,

$$\mathsf{P}\{|X| \geq 3\} \leq \tfrac{1}{9},$$

which is very pessimistic compared to the actual value $2Q(3) \approx 2 \times 10^{-3}$. But it is the tightest upper bound on $\mathsf{P}\{|X - \mathsf{E}(X)| \geq a\sigma_X\}$ given knowledge only of the mean and variance of X. Indeed, the inequality holds with equality for the random variable

$$X = \begin{cases} \mathsf{E}(X) + a\sigma_X & \text{w.p. } 1/2a^2, \\ \mathsf{E}(X) - a\sigma_X & \text{w.p. } 1/2a^2, \\ \mathsf{E}(X) & \text{w.p. } 1 - 1/a^2. \end{cases}$$

We now discuss an extremely useful inequality that is named after a Danish mathematician Johan Jensen and is centered around the notion of convexity. A function $g(x)$ is said to be *convex* if

$$g(x) \leq \frac{g(b) - g(a)}{b - a}(x - a) + g(a)$$

for all $x \in [a, b]$ and all $a < b$, that is, the function curve is below every chord across two points on the curve. If $g(x)$ is twice differentiable, then $g(x)$ is convex iff

$$g''(x) \geq 0.$$

If $-g(x)$ is convex, then $g(x)$ is called *concave*.

Example 5.5. The following functions are convex: (a) $g(x) = ax + b$. (b) $g(x) = x^2$. (c) $g(x) = |x|^p$, $p \geq 1$. (d) $g(x) = x \log x$, $x > 0$. (e) $g(x) = 1/x$, $x > 0$.

Example 5.6. The following functions are concave: (a) $g(x) = ax + b$. (b) $g(x) = \sqrt{x}$, $x > 0$. (c) $g(x) = \log x$, $x > 0$.

Jensen's inequality. Let X be a random variable with finite mean $\mathsf{E}[X]$ and $g(x)$ be a function such that $\mathsf{E}[g(X)]$ is finite. If $g(x)$ is convex, then

$$\mathsf{E}[g(X)] \geq g(\mathsf{E}[X]).$$

If $g(x)$ is concave, then

$$\mathsf{E}[g(X)] \leq g(\mathsf{E}[X]).$$

Jensen's inequality is a powerful tool to derive many inequalities.

Example 5.7. Since $g(x) = x^2$ is convex,

$$\mathsf{E}[g(X)] = \mathsf{E}[X^2] \geq (\mathsf{E}[X])^2 = g(\mathsf{E}[X]),$$

which we already know since $\mathrm{Var}(X) = \mathsf{E}[X^2] - (\mathsf{E}[X])^2 \geq 0$.

Example 5.8. Since $g(x) = 1/x$, $x > 0$, is convex,

$$\mathsf{E}[g(X^2)] = \mathsf{E}\left[\frac{1}{X^2}\right] \geq \frac{1}{\mathsf{E}[X^2]} = g(\mathsf{E}[X^2]).$$

Example 5.9 (Monotonicity of norms). For $1 \leq p \leq q$,

$$(\mathsf{E}[|X|^p])^{(1/p)} \leq (\mathsf{E}[|X|^q])^{(1/q)}. \tag{5.1}$$

To see this, consider a convex function $g(x) = |x|^{q/p}$ and use Jensen's inequality to obtain

$$\mathsf{E}[|X|^q] = \mathsf{E}[(|X|^p)^{\frac{q}{p}}] \geq (\mathsf{E}[|X|^p])^{\frac{q}{p}}.$$

Taking the q-th root of both sides establishes (5.1).

5.4 COVARIANCE AND CORRELATION

Let $(X, Y) \sim f_{X,Y}(x, y)$ and let $g(x, y)$ be a function of x and y. The expectation of $g(X, Y)$ is

$$\mathsf{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\, dx\, dy.$$

As an example, the *correlation* of X and Y is defined as

$$\mathsf{E}[XY].$$

We say that X and Y are *orthogonal* if $\mathsf{E}(XY) = 0$.

The *covariance* of X and Y is defined as

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathsf{E}[(X - \mathsf{E}[X])(Y - \mathsf{E}[Y])] \\ &= \mathsf{E}[XY - X\,\mathsf{E}[Y] - Y\,\mathsf{E}[X] + \mathsf{E}[X]\,\mathsf{E}[Y]] \\ &= \mathsf{E}[XY] - \mathsf{E}[X]\,\mathsf{E}[Y]. \end{aligned}$$

We say that X and Y are *uncorrelated* if $\mathrm{Cov}(X, Y) = 0$. Note that $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.

The *correlation coefficient* of X and Y is defined as

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

For any pair of random variables X and Y,

$$|\rho_{X,Y}| \le 1,$$

which follows by the Cauchy–Schwarz inequality

$$(\mathsf{E}[XY])^2 \le \mathsf{E}[X^2]\,\mathsf{E}[Y^2].$$

Note that $\rho_{X,Y} = \pm 1$ iff

$$\frac{X - \mathsf{E}[X]}{\sigma_X} = \pm \frac{Y - \mathsf{E}[Y]}{\sigma_Y},$$

that is, iff $X - \mathsf{E}[X]$ is a linear function of $Y - \mathsf{E}[Y]$. We shall see in Section 5.6 that $\rho_{X,Y}$ is a measure of how closely $X - \mathsf{E}[X]$ can be approximated or estimated by a *linear function* of $Y - \mathsf{E}[Y]$.

Example 5.10. We find the correlation, covariance, and correlation coefficient for $(X, Y) \sim f_{X,Y}(x, y)$ where

$$f_{X,Y}(x, y) = \begin{cases} 2 & x \ge 0,\ y \ge 0,\ x + y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Recall from Example 5.1 that $\mathsf{E}[X] = 1/3$ and $\mathrm{Var}(X) = 1/18$. By symmetry, $\mathsf{E}[Y] = 1/3$ and $\mathrm{Var}(Y) = 1/18$ as well. Consider

$$\begin{aligned}\mathsf{E}[XY] &= 2\int_0^1 \int_0^{1-x} xy\,dy\,dx \\ &= \int_0^1 x(1-x)^2\,dx \\ &= \frac{1}{12}.\end{aligned}$$

and

$$\mathrm{Cov}(X,Y) = \mathsf{E}[XY] - \mathsf{E}[X]\,\mathsf{E}[Y] = \frac{1}{12} - \frac{1}{9} = -\frac{1}{36}.$$

Finally,

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{-\frac{1}{36}}{\frac{1}{18}} = -\frac{1}{2}.$$

As noted earlier, X and Y are *uncorrelated* if $\mathrm{Cov}(X,Y) = 0$. If X and Y are independent, then they are uncorrelated, since

$$\begin{aligned}\mathsf{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\,f_{X,Y}(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\,f_X(x)f_Y(y)\,dx\,dy \\ &= \Big(\int_{-\infty}^{\infty} x f_X(x)\,dx\Big)\Big(\int_{-\infty}^{\infty} y f(y)\,dy\Big) \\ &= \mathsf{E}[X]\,\mathsf{E}[Y].\end{aligned}$$

However, that X and Y are uncorrelated does *not* necessarily imply that they are independent.

Example 5.11. Consider the pmf $p_{X,Y}(x,y)$ described by the following table

			x	
		-1	0	1
	-1	$\frac{1}{6}$	0	$\frac{1}{6}$
y	0	0	$\frac{1}{3}$	0
	1	$\frac{1}{6}$	0	$\frac{1}{6}$

Clearly X and Y are not independent. But it can be readily checked that $\mathsf{E}[X] = \mathsf{E}[Y] = \mathsf{E}[XY] = 0$. Thus $\mathrm{Cov}(X,Y) = 0$, that is, X and Y are uncorrelated.

5.5 CONDITIONAL EXPECTATION

Let $(X, Y) \sim f_{X,Y}(x, y)$. Recall that the *conditional pdf* of X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)},$$

if $f_Y(y) > 0$. Since $f_{X|Y}(x|y)$ is a pdf for X (for each y), we can define the expectation of any function $g(X, Y)$ w.r.t. $f_{X|Y}(x|y)$ as

$$\mathsf{E}[g(X, Y) \,|\, Y = y] = \int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x|y)\, dx,$$

which is a function of y.

Example 5.12. If $g(X, Y) = X$, then the conditional expectation of X given $Y = y$ is

$$\mathsf{E}[X \,|\, Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx.$$

Example 5.13. If $g(X, Y) = Y$, then $\mathsf{E}[Y \,|\, Y = y] = y$.

Example 5.14. If $g(X, Y) = XY$, then $\mathsf{E}[XY \,|\, Y = y] = y\, \mathsf{E}[X \,|\, Y = y]$.

Example 5.15. Let

$$f_{X,Y}(x, y) = \begin{cases} 2 & \text{if } x \ge 0,\ y \ge 0,\ x + y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

From Lecture #4, we already know that $X \,|\, \{Y = y\} \sim \text{Unif}[0, 1 - y]$. Thus, $\mathsf{E}[X|Y = y] = (1 - y)/2$.

Let $\phi(y) = \mathsf{E}[g(X, Y) \,|\, Y = y]$. We define the *conditional expectation* of $g(X, Y)$ given Y as

$$\mathsf{E}[g(X, Y) \,|\, Y] = \phi(Y).$$

In other words, the random variable $\mathsf{E}[g(X, Y) \,|\, Y]$ is a function of Y that takes values $\mathsf{E}[g(X, Y) \,|\, Y = y]$ when $Y = y$.

Law of iterated expectation. The following observation is very useful in computing expectation:

$$\begin{aligned} \mathsf{E}[\mathsf{E}[g(X, Y) \,|\, Y]] &= \int_{-\infty}^{\infty} \mathsf{E}[g(X, Y) \,|\, Y = y] f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x|y)\, dx \right) f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\, dx\, dy \\ &= \mathsf{E}[g(X, Y)]. \end{aligned}$$

Example 5.16. We continue Example 5.15. The conditional expectation of X given Y is the random variable

$$\mathsf{E}[X \mid Y] = \frac{1-Y}{2} =: Z.$$

The pdf of Z is

$$f_Z(z) = 8z\,, \quad 0 < z \le \tfrac{1}{2},$$

which is illustrated in Figure 5.2. Note that

$$\mathsf{E}[Z] = \int_0^{\frac{1}{2}} 8z^2\,dz = \frac{1}{3} = \mathsf{E}[X],$$

as is expected from the law of iterated expectation. Similarly,

$$\begin{aligned}\mathsf{E}[XY] &= \mathsf{E}[\mathsf{E}[XY \mid Y]]\\ &= \mathsf{E}\Big[\frac{Y(1-Y)}{2}\Big]\\ &= \int_0^1 \frac{y(1-y)}{2}\cdot 2(1-y)\,dy = \frac{1}{12},\end{aligned}$$

which agrees with the direct integration computed in Example 5.15.

Example 5.17. A coin has random bias $P \in [0,1]$ with pdf $f_P(p) = 2(1-p)$. The coin is flipped n times. Let N be the number of heads, that is, $N \mid \{P = p\} \sim \mathrm{Binom}(n, p)$. Then, by the law of iterated expectation, we can find

$$\begin{aligned}\mathsf{E}[N] &= \mathsf{E}[\mathsf{E}[N \mid P]]\\ &= \mathsf{E}[nP]\\ &= n\,\mathsf{E}[P]\\ &= n\int_0^1 2(1-p)p\,dp = \frac{1}{3}n,\end{aligned}$$

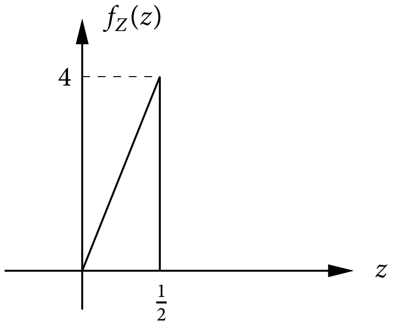


Figure 5.2. The graph of $f_Z(z)$.

which is much simpler than finding the pmf of N and computing the expectation.

Example 5.18. Let $\mathsf{E}[X \mid Y] = Y^2$ and $Y \sim \text{Unif}[0,1]$. In this case, we cannot find the pdf of X, since we do not know $f_{X|Y}(x|y)$. But using iterated expectation we can still find

$$\mathsf{E}[X] = \mathsf{E}[\mathsf{E}[X \mid Y]] = \mathsf{E}[Y^2] = \int_0^1 y^2\,dy = \frac{1}{3}.$$

We define the *conditional variance* of X given $Y = y$ as the variance of X w.r.t. $f_{X|Y}(x|y)$, i.e.,

$$\begin{aligned}\text{Var}(X \mid Y = y) &= \mathsf{E}[(X - \mathsf{E}[X|Y = y])^2 \mid Y = y] \\ &= \mathsf{E}[X^2 \mid Y = y] - 2\,\mathsf{E}[X\,\mathsf{E}[X|Y = y] \mid Y = y] + \mathsf{E}[(\mathsf{E}[X|Y = y])^2 \mid Y = y] \\ &= \mathsf{E}[X^2 \mid Y = y] - (\mathsf{E}[X \mid Y = y])^2.\end{aligned}$$

The random variable $\text{Var}(X \mid Y)$ is a function of Y that takes on the values $\text{Var}(X \mid Y = y)$. Its expected value is

$$\mathsf{E}[\text{Var}(X \mid Y)] = \mathsf{E}[\mathsf{E}[X^2 \mid Y] - (\mathsf{E}[X \mid Y])^2] = \mathsf{E}[X^2] - \mathsf{E}[(\mathsf{E}[X \mid Y])^2]. \tag{5.2}$$

Since $\mathsf{E}(X \mid Y)$ is a random variable, it has a variance

$$\begin{aligned}\text{Var}(\mathsf{E}(X \mid Y)) &= \mathsf{E}[(\mathsf{E}[X \mid Y] - \mathsf{E}[\mathsf{E}[X \mid Y]])^2] \\ &= \mathsf{E}[(\mathsf{E}[X \mid Y] - \mathsf{E}[X])^2] = \mathsf{E}[(\mathsf{E}[X \mid Y])^2] - (\mathsf{E}[X])^2. \end{aligned} \tag{5.3}$$

By adding (5.2) and (5.3), we establish the *law of conditional variances*:

$$\text{Var}(X) = \mathsf{E}[\text{Var}(X \mid Y)] + \text{Var}(\mathsf{E}[X \mid Y]).$$

5.6 MMSE ESTIMATION

Consider the signal estimation system depicted in Figure 5.3, where the original signal is $X \sim f_X(x)$ and its noisy observation is

$$Y \mid \{X = x\} \sim f_{Y|X}(y|x).$$

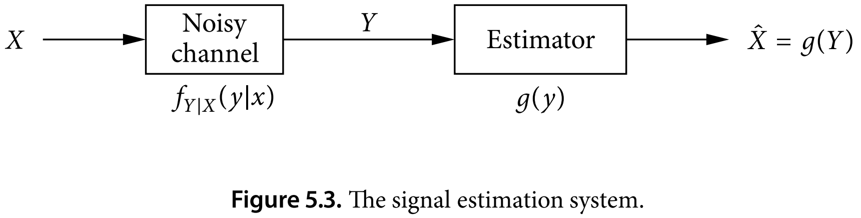


Figure 5.3. The signal estimation system.

An estimator is a mapping $g : \mathbb{R} \to \mathbb{R}$ that generates an estimate $\hat{X} = g(Y)$ of the original signal. We wish to find an optimal estimator $g^*(y)$ that minimizes the *mean square error* (MSE)

$$\mathsf{E}[(X - \hat{X})^2] = \mathsf{E}[(X - g(Y))^2]. \tag{5.4}$$

The estimator $g^*(y)$ that attains the smallest value of (5.4) is referred to as the *minimum mean square error* (MMSE) estimator of X given Y, and $\hat{X} = g^*(Y)$ is referred to as the MMSE estimate.

Suppose that there is no observation and let a^* be the MMSE estimate of X, that is,

$$a^* = \arg\min_a \mathsf{E}[(X - a)^2].$$

Then,

$$a^* = \mathsf{E}[X]. \tag{5.5}$$

In other words, the mean is the optimal summary of X under the mean square error criterion. To prove (5.5), note that for any estimate a of X,

$$\begin{aligned}
\mathsf{E}[(X - a)^2] &= \mathsf{E}[(X - \mathsf{E}[X] + \mathsf{E}[X] - a)^2] \\
&= \mathsf{E}[(X - \mathsf{E}[X])^2] + (\mathsf{E}[X] - a)^2 + \mathsf{E}[X - \mathsf{E}[X]](\mathsf{E}[X] - a) \\
&= \mathsf{E}[(X - \mathsf{E}[X])^2] + (\mathsf{E}[X] - a)^2 \\
&\geq \mathsf{E}[(X - \mathsf{E}[X])^2]
\end{aligned}$$

with equality iff $a = \mathsf{E}[X]$.

A similar observation continues to hold with the observation $Y = y$. Let $g^*(y) = \mathsf{E}[X \,|\, Y = y]$. Then, by (5.5), for any estimator $g(y)$,

$$\mathsf{E}[(X - g^*(y))^2 \,|\, Y = y] \leq \mathsf{E}[(X - g(y))^2 \,|\, Y = y].$$

Consequently,

$$\mathsf{E}[(X - g^*(Y))^2] \leq \mathsf{E}[(X - g(Y))^2]$$

and $\hat{X} = g^*(Y) = \mathsf{E}[X|Y]$ is the MMSE estimate of X given Y with the corresponding MSE

$$\mathsf{E}[\mathrm{Var}(X|Y)] = \mathsf{E}[(X - \mathsf{E}[X \,|\, Y])^2].$$

The MMSE estimate $\hat{X} = \mathsf{E}[X|Y]$ satisfies the following properties.

1. It is *unbiased*, i.e.,

$$\mathsf{E}[\hat{X}] = \mathsf{E}[X].$$

2. The estimation error $X - \hat{X}$ is unbiased for every $Y = y$, i.e.,

$$\mathsf{E}[X - \hat{X} \,|\, Y = y] = 0.$$

3. The estimation error and the estimate are *orthogonal*, i.e.,

$$\begin{aligned}\mathsf{E}[(X-\hat{X})\hat{X}] &= \mathsf{E}[\mathsf{E}[(X-\hat{X})\hat{X}\,|\,Y]] \\ &= \mathsf{E}[\hat{X}\,\mathsf{E}[X-\hat{X}\,|\,Y]] = 0.\end{aligned}$$

   In fact, the estimation error is orthogonal to *any* function $g(Y)$, i.e.,

$$\mathsf{E}[(X-\hat{X})g(Y)] = 0.$$

4. By the law of conditional variance $\operatorname{Var}(X) = \operatorname{Var}(\hat{X}) + \mathsf{E}[\operatorname{Var}(X\,|\,Y)]$, the sum of the variance of the estimate and its MSE is equal to the variance of the signal.
5. If X and Y are independent, then $\hat{X} = \mathsf{E}[X]$, that is, the observation is ignored.

Example 5.19. Again let

$$f_{X,Y}(x,y) = \begin{cases} 2 & x \ge 0,\ y \ge 0,\ x+y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We find the MMSE estimate of X given Y and its MSE. We already know that the MMSE estimate is

$$\mathsf{E}[X\,|\,Y] = \frac{1-Y}{2}$$

and that the conditional variance is

$$\operatorname{Var}[X\,|\,Y] = \frac{(1-Y)^2}{12}.$$

Hence, the MMSE is $\mathsf{E}[\operatorname{Var}(X\,|\,Y)] = 1/24$, compared to $\operatorname{Var}(X) = 1/18$. The difference is $\operatorname{Var}(\mathsf{E}[X\,|\,Y]) = 1/72$, which is the variance of the estimate.

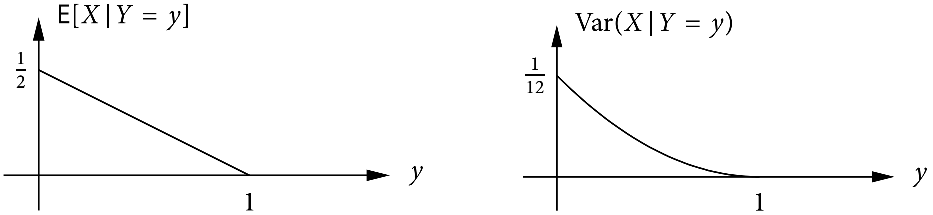


Figure 5.4. The conditional mean and variance of X given $Y = y$.

Example 5.20 (Additive Gaussian noise channel). Consider a communication channel with input $X \sim \mathrm{N}(\mu, P)$, noise $Z \sim \mathrm{N}(0, N)$, and output $Y = X + Z$. We assume that X and

Z are independent. We find the MMSE estimate of X given Y and its MSE, i.e., $\mathsf{E}[X \mid Y]$ and $\mathsf{E}[\mathrm{Var}(X \mid Y)]$. Recall that $Y \mid \{X = x\} \sim \mathrm{N}(x, N)$ and $Y \sim \mathrm{N}(\mu, P + N)$, that is,

$$f_{Y|X}(y|x) = f_Z(y - x) = \frac{1}{\sqrt{2\pi N}}\, e^{-\frac{(y-x)^2}{2N}}$$

and

$$f_Y(y) = \frac{1}{\sqrt{2\pi(P + N)}}\, e^{-\frac{(y-\mu)^2}{2(P+N)}}.$$

Hence,

$$\begin{aligned} f_{X|Y}(x \,|\, y) &= \frac{f_X(x) f_{Y|X}(y|x)}{f_Y(y)} \\ &= \frac{\frac{1}{\sqrt{2\pi P}}\, e^{-\frac{(x-\mu)^2}{2P}} \frac{1}{\sqrt{2\pi N}}\, e^{-\frac{(y-x)^2}{2N}}}{\frac{1}{\sqrt{2\pi(P+N)}}\, e^{-\frac{(y-\mu)^2}{2(P+N)}}} \\ &= \frac{1}{\sqrt{2\pi \frac{PN}{P+N}}} \exp\Bigl(-\frac{\bigl(x - (\frac{P}{P+N} y + \frac{N}{P+N}\mu)\bigr)^2}{2\frac{PN}{P+N}}\Bigr), \end{aligned}$$

or equivalently,

$$X \mid \{Y = y\} \sim \mathrm{N}\Bigl(\frac{P}{P + N} y + \frac{N}{P + N}\mu,\ \frac{PN}{P + N}\Bigr).$$

Thus,

$$\mathsf{E}[X \mid Y] = \frac{P}{P + N} Y + \frac{N}{P + N}\mu,$$

which is a convex combination of the observation Y and the mean μ (MMSE estimate without observation), and tends to Y as $N \to 0$ and to μ as $N \to \infty$. The corresponding MSE is

$$\mathsf{E}[\mathrm{Var}(X \mid Y)] = \mathsf{E}\Bigl[\frac{PN}{P + N}\Bigr] = \frac{PN}{P + N},$$

which is less than P, the MSE without the observation Y. Note that the conditional variance $\mathrm{Var}(X \mid Y)$ is independent of Y.

In the above two examples, the MMSE estimate turned out to be an affine function of Y (i.e., of the form $aY + b$). This is not always the case.

Example 5.21. Let

$$f(x|y) = \begin{cases} ye^{-yx} & x \geq 0,\ y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$\mathsf{E}[X \mid Y] = \frac{1}{Y}.$$

Remark 5.3. There can be alternative criteria for measuring goodness of estimators. For example, instead of the MSE criteria in (5.4) that was introduced in the 19th century by Legendre and Gauss, one may measure the mean absolute error (MAE)

$$\mathsf{E}[|X - g(Y)|],$$

which dates back to Boscovich and Laplace in the preceding century. It can be shown that the minimum MAE estimate is the conditional median, that is,

$$\begin{aligned}\mathsf{P}\{X \le g^*(y) \,|\, Y = y\} \ge 1/2,\\ \mathsf{P}\{X \ge g^*(y) \,|\, Y = y\} \ge 1/2.\end{aligned}$$

5.7 LINEAR MMSE ESTIMATION

To find the MMSE estimate, one needs to know the statistics of the signal and the channel, namely, $f_{X,Y}(x, y)$, or at least, $f_{X|Y}(x|y)$, which is rarely the case in practice. We typically have estimates only of the first and second moments of the signal and the observation, i.e., the means, variances, and covariance of X and Y. This is not, in general, sufficient information for computing the MMSE estimate, but as we shall see is enough to compute the *linear* MMSE (LMMSE) estimate of the signal X given the observation Y, i.e., the estimate of the form

$$\hat{X} = aY + b$$

that minimizes the mean square error

$$\mathsf{E}[(X - \hat{X})^2] = \mathsf{E}[(X - aY - b)^2].$$

We show that the LMMSE estimate of X given Y is

$$\begin{aligned}\hat{X} &= a^*Y + b^*\\ &= \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}(Y - \mathsf{E}[Y]) + \mathsf{E}[X]\\ &= \rho_{X,Y}\sigma_X\left(\frac{Y - \mathsf{E}[Y]}{\sigma_Y}\right) + \mathsf{E}[X]\end{aligned} \tag{5.6}$$

and its MSE is

$$\begin{aligned}\mathsf{E}[(X - a^*Y - b^*)^2] &= \mathrm{Var}(X) - \frac{(\mathrm{Cov}(X, Y))^2}{\mathrm{Var}(Y)}\\ &= (1 - \rho_{X,Y}^2)\sigma_X^2.\end{aligned}$$

First note that for any a,

$$\mathsf{E}[(X - aY - b)^2] = \mathsf{E}[((X - aY) - b)^2]$$

is minimized by $b^*(a) = \mathsf{E}[X - aY] = \mathsf{E}[X] - a\,\mathsf{E}[Y]$. Hence, under this choice, the MSE can be written as a quadratic function in a as

$$\begin{aligned}\mathsf{E}[(X - aY - b^*(a))^2] &= \mathsf{E}[((X - \mathsf{E}[X]) - a(Y - \mathsf{E}[Y]))^2] \\ &= \mathrm{Var}(X) - 2a\,\mathrm{Cov}(X, Y) + a^2\,\mathrm{Var}(Y),\end{aligned}$$

which is minimized at

$$a^* = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}$$

with the minimum

$$\mathrm{Var}(X) - \frac{(\mathrm{Cov}(X, Y))^2}{\mathrm{Var}(Y)}.$$

Alternatively, we can minimize

$$J(a, b) = \mathsf{E}[(X - aY - b)^2]$$

by finding (a^*, b^*) that satisfies

$$\frac{\partial}{\partial a} J(a, b) = \frac{\partial}{\partial b} J(a, b) = 0$$

and by showing that the solution attains the global minimum as illustrated in Figure 5.5.

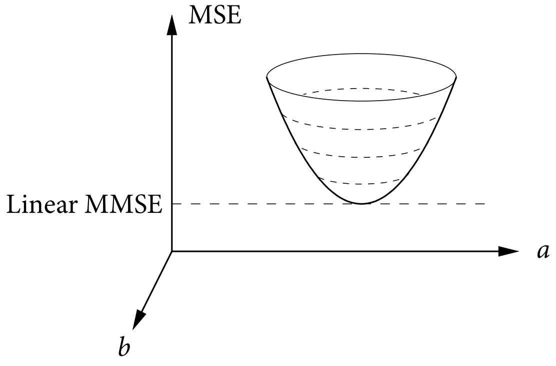


Figure 5.5. The MSE as a function of a and b.

The linear MMSE estimate $\hat{X}$ satisfies the following properties:

1. It is unbiased, i.e., $\mathsf{E}(\hat{X}) = \mathsf{E}(X)$, which was also true for the nonlinear MMSE estimate.
2. The estimation error and the estimate are *orthogonal*, i.e.,

$$\mathsf{E}[(X - \hat{X})\hat{X}] = 0.$$

In fact, the estimation error is orthogonal to *any* affine function $aY + b$, i.e.,

$$\mathsf{E}[(X - \hat{X})(aY + b)] = 0.$$

3. If $\rho_{X,Y} = 0$, i.e., X and Y are uncorrelated, then the observation is ignored and
$$\hat{X} = \mathsf{E}[X].$$
4. If $\rho_{X,Y} = \pm 1$, i.e., $(X - \mathsf{E}(X))$ and $(Y - \mathsf{E}(Y))$ are linearly dependent, then the linear estimate is perfect and $\hat{X} = X$.

The LMMSE estimate is not, in general, as good as the MMSE estimate.

Example 5.22. Let $Y \sim \text{Unif}[-1, 1]$ and $X = Y^2$. The MMSE estimate of X given Y is Y^2, which is perfect. To find the LMMSE estimate we compute
$$\begin{aligned} \mathsf{E}[Y] &= 0, \\ \mathsf{E}[X] &= \int_{-1}^{1} \frac{1}{2} y^2 \, dy = \frac{1}{3}, \end{aligned}$$
and
$$\text{Cov}(X, Y) = \mathsf{E}[XY] - 0 = \mathsf{E}[Y^3] = 0.$$
Thus, the LMMSE estimate $\hat{X} = \mathsf{E}(X) = \frac{1}{3}$, i.e., the observation Y is totally ignored, even though it completely determines X.

5.8 GEOMETRIC FORMULATION OF ESTIMATION

For both nonlinear and linear MMSE estimation problems we discussed in the previous two sections, we found that the estimation error is orthogonal to the optimal estimate. This orthogonality property is a fundamental characteristic of an optimal estimator that minimizes the MSE among a class of estimators and can be used to find the optimal estimator in a simple geometric argument.

First, we introduce some mathematical background. A *vector space* $\mathcal{V}$ consists of a set of vectors that are closed under two operations:

- *Vector addition*: if $v, w \in \mathcal{V}$ then $v + w \in \mathcal{V}$.
- *Scalar multiplication*: if $a \in \mathbb{R}$ and $v \in \mathcal{V}$, then $av \in \mathcal{V}$.

An *inner product* is a real-valued operation $v \cdot w$ satisfying these three conditions:

- *Commutativity*: $v \cdot w = w \cdot v$.
- *Linearity*: $(au + v) \cdot w = a(u \cdot w) + v \cdot w$.
- *Nonnegativity*: $v \cdot w \geq 0$ and $v \cdot v = 0$ iff $v = 0$.

A vector space with an inner product is referred to as an *inner product space*. For example, the Euclidean space
$$\mathbb{R}^n = \{\mathbf{x} = (x_1, x_2, \ldots, x_n) \colon x_1, x_2, \ldots, x_n \in \mathbb{R}\}$$

with vector addition

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n),$$

scalar multiplication

$$a\mathbf{x} = (ax_1, ax_2, \dots, ax_n),$$

and dot product

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

is an inner product space.

The inner product induces generalized notions of length, distance, and angle. The *norm* of $v \in \mathcal{V}$ is defined as $\|v\| = \sqrt{v \cdot v}$, and can be viewed as the "length" of v. A metric is a function $d(v, w)$ satisfying the following three conditions:

- *Commutativity*: $d(v, w) = d(w, v)$.
- *Nonnegativity*: $d(v, w) \geq 0$ with equality iff $v = w$.
- *Triangle inequality*: $d(u, w) \leq d(u, v) + d(v, w)$.

It is easy to verify that $\|v - w\|$ is a metric, and thus can be interpreted as the "distance" between v and w. We say that v and w are *orthogonal* (written $v \perp w$) if $v \cdot w = 0$. More generally, the "angle" θ between v and w satisfies $v \cdot w = \|v\|\|w\| \cos\theta$. For example, the norm (length) of a vector $\mathbf{x}$ in the Euclidean space is

$$\|\mathbf{x}\| = \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}},$$

the distance between the two points represented by $\mathbf{x}$ and $\mathbf{y}$ is

$$\|\mathbf{x} - \mathbf{y}\| = \left(\sum_{i=1}^{n} (x_i - y_i)^2\right)^{\frac{1}{2}},$$

and the angle between the two vectors $\mathbf{x}$ and $\mathbf{y}$ is

$$\theta = \arccos \frac{\sum_{i=1}^{n} x_i y_i}{(\sum_{i=1}^{n} x_i^2)^{1/2} (\sum_{i=1}^{n} y_i^2)^{1/2}}.$$

The Pythagorean theorem holds in an arbitrary inner product space, namely, if $v \perp w$, then

$$\|v + w\|^2 = (v + w) \cdot (v + w) = (v \cdot v) + (w \cdot w) + 2(v \cdot w) = \|v\|^2 + \|w\|^2.$$

We say that $\mathcal{W}$ is a *subspace* of a vector space $\mathcal{V}$ if $\mathcal{W} \subseteq \mathcal{V}$ is itself a vector space (i.e., closed under vector addition and scalar multiplication). A subspace of an inner product space inherits the same inner product and is also an inner product space. We now establish the following simple observation on the distance between a vector in a vector space $\mathcal{V}$ and its subspae $\mathcal{W}$.

Orthogonality principle. Let $\mathcal{V}$ be an inner product space and $\mathcal{W}$ be its subspace. Let $v \in \mathcal{V}$. Suppose that there exists $w^* \in \mathcal{W}$ such that $v - w^*$ is orthogonal to every $w \in \mathcal{W}$. Then,

$$w^* = \arg\min_{w \in \mathcal{W}} \|v - w\|.$$

As depicted in Figure 5.6, the orthogonal projection of v onto $\mathcal{W}$ (if it exists) is the closest vector of v in $\mathcal{W}$. The proof is immediate from the Pythagorean theorem. For any $w \in \mathcal{W}$, $(v - w^*) \perp (w^* - w)$ by the orthogonality condition. Hence, $\|v - w^*\|^2 + \|w^* - w\|^2 = \|v - w\|^2$ and thus $\|v - w^*\| \leq \|v - w\|$.

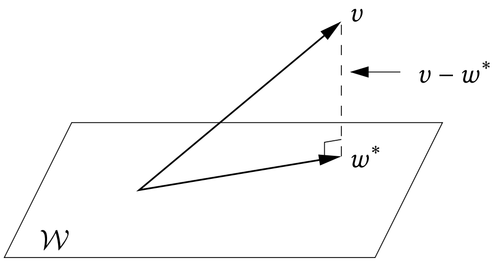


Figure 5.6. Among all vectors in $\mathcal{W}$, the orthogonal projection w^* of v is the closest.

We now consider the inner product space $\mathcal{V}$ that consists of all random variables (with finite second moment) on the same probability space, where

- the vector addition (sum) $V + W$ of random variables V and W is a random variable,
- the scalar (constant) multiplication aV is a random variable, and
- the inner product $V \cdot W = 0$ of V and W is their correlation $\mathsf{E}[VW]$ (which satisfies the three inner product axioms).

Fortuitously, two random variables V and W are *orthogonal*, i.e., $\mathsf{E}[VW] = 0$, as defined in Section 5.4 iff V and W are orthogonal as two vectors, i.e., $V \cdot W = 0$. Note that the norm of V is $\|V\| = \sqrt{\mathsf{E}[V^2]}$.

The goal of MMSE estimation can be now rephrased as follows: Given the vector space $\mathcal{V}$ of all random variables (or all random variables that are functions of X and Y) and a subspace $\mathcal{W}$ of estimators, find $\hat{X}$ that is closest to X, that is, the mean square error $\|\hat{X} - X\|^2$ is the smallest.

Example 5.23 (MMSE estimator). Let $\mathcal{W}$ be the space of all functions $g(Y)$ with finite second moment. It can be easily verified that it is an inner product space. We already know that the MMSE estimate $\hat{X} = g^*(Y) = \mathsf{E}[X|Y]$ we found in Section 5.6 has the property that the error $\hat{X} - X$ is orthogonal to every $g(Y)$. Hence, it minimizes the MSE among all functions of Y.

Example 5.24 (Mean). Let $\mathcal{W}$ be the set of all constants $a \in \mathbb{R}$. Once again it is a valid subspace. Since $X - \mathsf{E}[X]$ is orthogonal to $\mathcal{W}$, i.e., $\mathsf{E}[(X - \mathsf{E}[X])a] = 0$ for every a, $\hat{X} = a^* = \mathsf{E}[X]$ minimizes the MSE among all constants.

Example 5.25 (LMMSE estimator). Let $\mathcal{W}$ be the subspace that consists all functions of the form $aY + b$. Since $X - (a^*Y + b^*)$, where a^* and b^* are given in (5.6), is orthogonal to any $aY + b$, $\hat{X} = a^*Y + b^*$ minimizes the MSE among all affine functions of Y.

We shall later apply this orthogonality principle to find MMSE estimators in more general subspaces such as linear combinations of multiple random variables and linear filters of random processes.

5.9 JOINTLY GAUSSIAN RANDOM VARIABLES

We say that two random variables are *jointly Gaussian* if their joint pdf is of the form

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}}\, e^{-\frac{1}{2(1-\rho_{X,Y}^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho_{X,Y}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right)}.$$

Note that this pdf is a function only of μ_X, μ_Y, σ_X^2, σ_Y^2, and $\rho_{X,Y}$. Consistent with our notation, these parameters are indeed $\mathsf{E}[X]$, $\mathsf{E}[Y]$, $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, and the correlation coefficient of X and Y. In Lecture #6, we shall define jointly Gaussian random variables in a more general way.

Example 5.26. Consider the additive Gaussian noise channel in Example 5.20, where $X \sim \mathrm{N}(\mu, P)$ and $Z \sim \mathrm{N}(0, N)$ are independent and $Y = X + Z$. Then the pair X and Z, the pair X and Y, and the pair Y and Z are jointly Gaussian.

If X and Y are jointly Gaussian, contours of equal joint pdf are ellipses defined by the quadratic equation

$$\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho_{X,Y}\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} = c \geq 0.$$

The orientation of the major axis of these ellipses is

$$\theta = \frac{1}{2}\arctan\left(\frac{2\rho_{X,Y}\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2}\right).$$

Figure 5.7 shows a few examples of the joint pdf.

Jointly Gaussian random variables X and Y satisfy the following properties.

1. They are marginally Gaussian, i.e.,

$$X \sim \mathrm{N}(\mu_X, \sigma_X^2) \quad \text{and} \quad Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2).$$

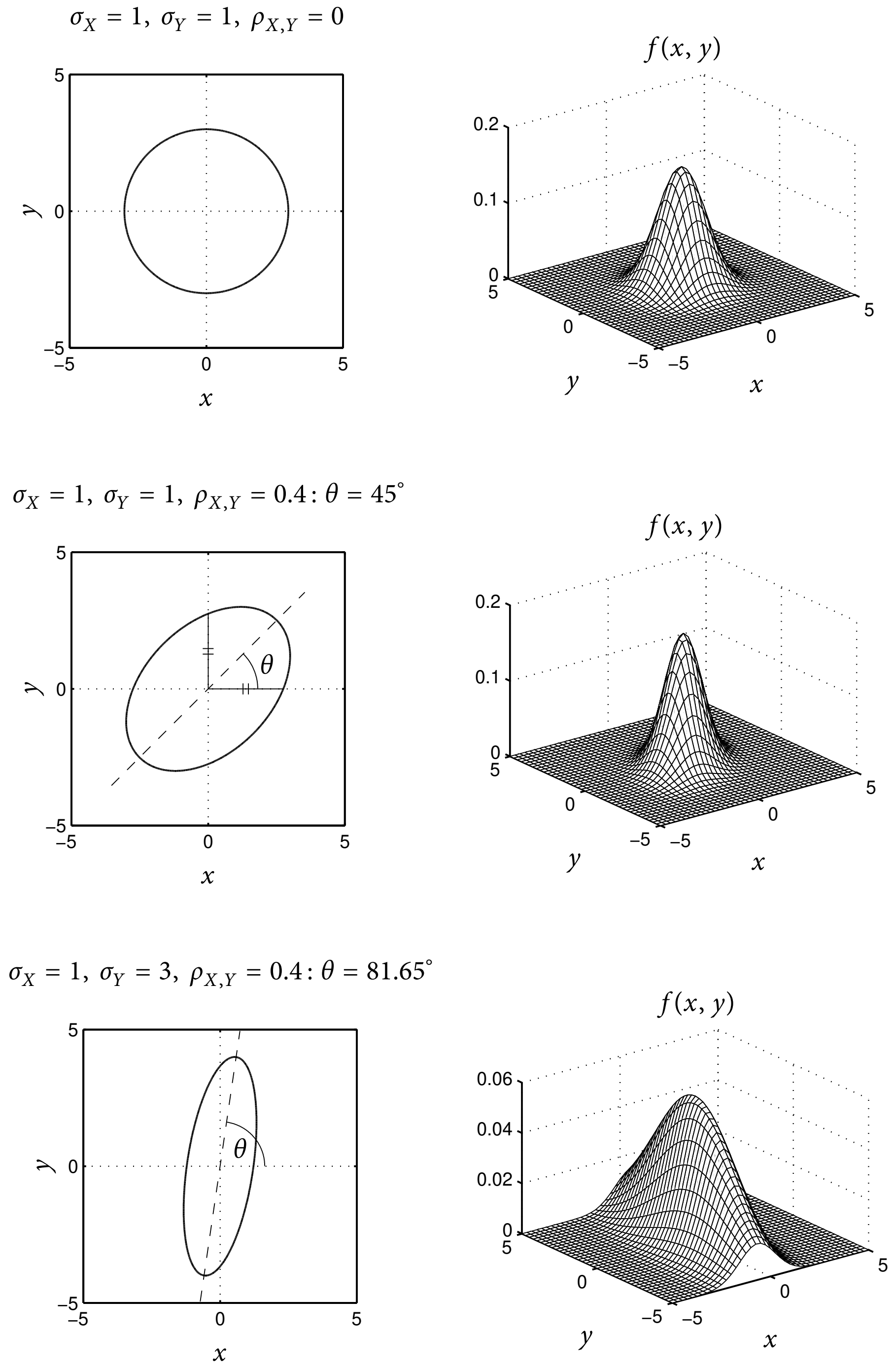


Figure 5.7. Joint pdfs of jointly Gaussian random variables.

2. The conditional pdf is Gaussian, i.e.,

$$X \mid \{Y = y\} \sim \mathrm{N}\Big(\frac{\rho_{X,Y}\sigma_X}{\sigma_Y}(y - \mu_Y) + \mu_X,\ (1 - \rho_{X,Y}^2)\sigma_X^2\Big),$$

   which shows that the MMSE estimate is linear.

3. If X and Y are jointly Gaussian and uncorrelated, i.e., $\rho_{X,Y} = 0$, then they are also independent.

The converse to the first property is not necessarily true, that is, Gaussian marginals do not necessarily mean that the random variables are *jointly* Gaussian.

Example 5.27. Let $X \sim \mathrm{N}(0,1)$ and

$$Z = \begin{cases} +1 & \text{w.p. } 1/2, \\ -1 & \text{w.p. } 1/2 \end{cases}$$

be independent and let $Y = XZ$. Clearly, $Y \sim \mathrm{N}(0,1)$. However, X and Y do not have a joint pdf. Using delta functions, "$f_{X,Y}(x,y)$" has the form shown in Figure 5.8. Note that X and Y are uncorrelated, but not independent. This does not contradict the third property since X and Y are not jointly Gaussian.

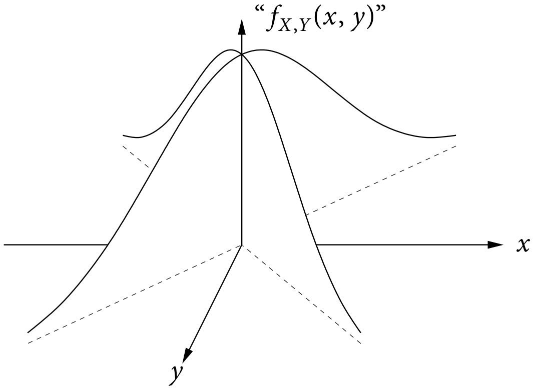


Figure 5.8. The "joint pdf" of X and Y.

PROBLEMS

5.1. *Inequalities.* Label each of the following statements with =, ≤, or ≥. Justify each answer.

(a) $1/\mathsf{E}[X^2]$ vs. $\mathsf{E}(1/X^2)$.

(b) $(\mathsf{E}[X])^2$ vs. $\mathsf{E}[X^2]$.

(c) $\mathrm{Var}(X)$ vs. $\mathrm{Var}(\mathsf{E}[X|Y])$.

(d) $\mathsf{E}[X^2]$ vs. $\mathsf{E}[(\mathsf{E}[X|Y])^2]$.

5.2. *Cauchy–Schwartz inequality.*

(a) Prove the following inequality: $(\mathsf{E}[XY])^2 \le \mathsf{E}[X^2]\,\mathsf{E}[Y^2]$. (Hint: Use the fact that for any real t, $\mathsf{E}[(X+tY)^2] \ge 0$.)

(b) Prove that equality holds if and only if $X = cY$ for some constant c. Find c in terms of the second moments of X and Y.

(c) Use the Cauchy–Schwartz inequality to show the correlation coefficient satisfies $|\rho_{X,Y}| \le 1$.

(d) Prove the *triangle inequality*: $\sqrt{\mathsf{E}[(X+Y)^2]} \le \sqrt{\mathsf{E}[X^2]} + \sqrt{\mathsf{E}[Y^2]}$.

5.3. *Two envelopes.* An amount A is placed in one envelope and the amount $2A$ is placed in another envelope. The amount A is fixed but unknown to you. The envelopes are shuffled and you are given one of the envelopes at random. Let X denote the amount you observe in this envelope. Designate by Y the amount in the other envelope. Thus

$$(X, Y) = \begin{cases} (A, 2A), & \text{w.p. } 1/2, \\ (2A, A), & \text{w.p. } 1/2. \end{cases}$$

You may keep the envelope you are given, or you can switch envelopes and receive the amount in the other envelope.

(a) Find $\mathsf{E}[X]$ and $\mathsf{E}[Y]$.

(b) Find $\mathsf{E}[X/Y]$ and $\mathsf{E}[Y/X]$.

(c) Suppose you switch. What is the expected amount you receive?

5.4. *Mean and variance.* Let X and Y be random variables with joint pdf

$$f_{X,Y}(x, y) = \begin{cases} 1 & \text{if } |x| + |y| \le 1/\sqrt{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Define the random variable $Z = |X| + |Y|$. Find the mean and variance of Z without first finding the pdf of Z.

5.5. *Tall trees.* Suppose that the average height of trees on campus is 20 feet. Argue that no more than half of the tree population is taller than 40 feet.

5.6. Let X and Y have correlation coefficient $\rho_{X,Y}$.

(a) What is the correlation coefficient between X and $3Y$?

(b) What is the correlation coefficient between $2X$ and $-5Y$?

5.7. *Random phase signal.* Let $Y(t) = \sin(\omega t + \Theta)$ be a sinusoidal signal with random phase $\Theta \sim \mathrm{Unif}[-\pi, \pi]$. Assume here that ω and t are constants. Find the mean and variance of $Y(t)$. Do they depend on t?

5.8. *Coin tosses.* A coin with bias p is tossed independently until two heads or two tails come up in a row. Find the expected value of the number of tosses X.

5.9. *Iterated expectation.* Let Λ and X be two random variables with

$$\Lambda \sim f_\Lambda(\lambda) = \begin{cases} \frac{5}{3}\lambda^{\frac{2}{3}}, & 0 \le \lambda \le 1 \\ 0, & \text{otherwise}, \end{cases}$$

and $X|\{\Lambda = \lambda\} \sim \mathrm{Exp}(\lambda)$. Find $E(X)$.

5.10. *Sum of packet arrivals.* Consider a network router with two types of incoming packets, wireline and wireless. Let the random variable $N_1(t)$ denote the number of *wireline* packets arriving during time $(0, t]$ and let the random variable $N_2(t)$ denote the number of *wireless* packets arriving during time $(0, t]$. Suppose $N_1(t)$ and $N_2(t)$ are independent Poisson with pmfs

$$\begin{aligned} \mathsf{P}\{N_1(t) = n\} &= \frac{(\lambda_1 t)^n}{n!} e^{-\lambda_1 t} \qquad \text{for } n = 0, 1, 2, \ldots \\ \mathsf{P}\{N_2(t) = k\} &= \frac{(\lambda_2 t)^k}{k!} e^{-\lambda_2 t} \qquad \text{for } k = 0, 1, 2, \ldots. \end{aligned}$$

Let $N(t) = N_1(t) + N_2(t)$ be the total number of packets arriving at the router during time $(0, t]$.

(a) Find the mean $E(N(t))$ and variance $\mathrm{Var}(N(t))$ of the total number of packet arrivals.

(b) Find the pmf of $N(t)$.

(c) Let the random variable Y be the time to receive the first packet of either type. Find the pdf of Y.

(d) What is the probability that the first received packet is wireless?

5.11. *Conditioning on an event.* Let X be a r.v. with pdf

$$f_X(x) = \begin{cases} 2(1 - x) & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

and let the event $A = \{X \ge 1/3\}$. Find $f_{X|A}(x)$, $\mathsf{E}(X|A)$, and $\mathrm{Var}(X|A)$.

5.12. *Jointly Gaussian random variables.* Let X and Y be jointly Gaussian random variables with pdf

$$f_{X,Y}(x, y) = \frac{1}{\pi\sqrt{3/4}} e^{-\frac{1}{2}(4x^2/3 + 16y^2/3 + 8xy/3 - 8x - 16y + 16)}.$$

Find $E(X)$, $E(Y)$, $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, and $\mathrm{Cov}(X, Y)$.

5.13. *Neural net.* Let $Y = X + Z$, where the signal $X \sim \mathrm{U}[-1,1]$ and noise $Z \sim \mathcal{N}(0,1)$ are independent.

(a) Find the function $g(y)$ that minimizes

$$\text{MSE} = \mathsf{E}\left[(\operatorname{sgn}(X) - g(Y))^2\right],$$

where

$$\operatorname{sgn}(x) = \begin{cases} -1 & x \leq 0 \\ +1 & x > 0. \end{cases}$$

(b) Plot $g(y)$ vs. y.

5.14. *Additive shot noise channel.* Consider an additive noise channel $Y = X + Z$, where the signal $X \sim \mathcal{N}(0,1)$, and the noise $Z|\{X = x\} \sim \mathcal{N}(0, x^2)$, i.e., the noise power of increases linearly with the signal squared.

(a) Find $E(Z^2)$.

(b) Find the best linear MSE estimate of X given Y.

5.15. *Additive uniform noise channel.* Let the signal

$$X = \begin{cases} +1, & \text{with probability } \frac{1}{2} \\ -1, & \text{with probability } \frac{1}{2}, \end{cases}$$

and the noise $Z \sim \text{Unif}[-2,2]$ be independent random variables. Their sum $Y = X + Z$ is observed. Find the minimum MSE estimate of X given Y and its MSE.

5.16. *Estimation vs. detection.* Let the signal

$$X = \begin{cases} +1, & \text{with probability } \frac{1}{2} \\ -1, & \text{with probability } \frac{1}{2}, \end{cases}$$

and the noise $Z \sim \text{Unif}[-2,2]$ be independent random variables. Their sum $Y = X + Z$ is observed.

(a) Find the best MSE estimate of X given Y and its MSE.

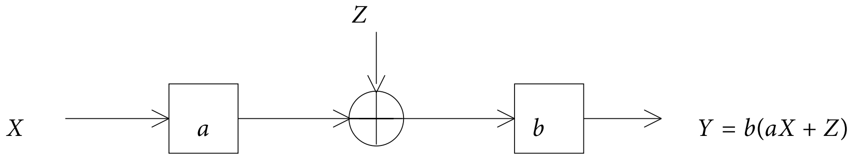
(b) Now suppose we use a decoder to decide whether $X = +1$ or $X = -1$ so that the probability of error is minimized. Find the optimal decoder and its probability of error. Compare the optimal decoder's MSE to the minimum MSE.

5.17. *Linear estimator.* Consider a channel with the observation $Y = XZ$, where the signal X and the noise Z are uncorrelated Gaussian random variables. Let $E[X] = 1$, $E[Z] = 2$, $\sigma_X^2 = 5$, and $\sigma_Z^2 = 8$.

(a) Find the best MSE linear estimate of X given Y.

(b) Suppose your friend from Caltech tells you that he was able to derive an estimator with a lower MSE. Your friend from UCLA disagrees, saying that this is not possible because the signal and the noise are Gaussian, and hence the best linear MSE estimator will also be the best MSE estimator. Could your UCLA friend be wrong?

5.18. *Additive-noise channel with path gain.* Consider the additive noise channel shown in the figure below, where X and Z are zero mean and uncorrelated, and a and b are constants.



Find the MMSE linear estimate of X given Y and its MSE in terms only of σ_X, σ_Z, a, and b.

5.19. *Worst noise distribution.* Consider an additive noise channel $Y = X + Z$, where the signal $X \sim \mathcal{N}(0, P)$ and the noise Z has zero mean and variance N. Assume X and Z are independent. Find a distribution of Z that maximizes the minimum MSE of estimating X given Y, i.e., the distribution of the worst noise Z that has the given mean and variance. You need to justify your answer.

5.20. *Image processing.* A pixel signal $X \sim \mathrm{U}[-k, k]$ is digitized to obtain

$$\tilde{X} = i + \frac{1}{2}, \text{ if } i < X \leq i + 1, \; i = -k, -k + 1, \ldots, k - 2, k - 1.$$

To improve the the visual appearance, the digitized value $\tilde{X}$ is dithered by adding an independent noise Z with mean $\mathsf{E}(Z) = 0$ and variance $\mathrm{Var}(Z) = N$ to obtain $Y = \tilde{X} + Z$.

(a) Find the correlation of X and Y.

(b) Find the best linear MSE estimate of X given Y. Your answer should be in terms only of k, N, and Y.

5.21. *Orthogonality.* Let $\hat{X}$ be the minimum MSE estimate of X given Y.

(a) Show that for any function $g(y)$, $\mathsf{E}((X - \hat{X})g(Y)) = 0$, i.e., the error $(X - \hat{X})$ and $g(Y)$ are orthogonal.

(b) Show that

$$\mathrm{Var}(X) = \mathsf{E}(\mathrm{Var}(X|Y)) + \mathrm{Var}(\hat{X}).$$

Provide a geometric interpretation for this result.

5.22. *Difference from sum.* Let X and Y be two random variables. Let $Z = X + Y$ and let $W = X - Y$. Find the best linear estimate of W given Z as a function of $\mathsf{E}(X)$, $\mathsf{E}(Y)$, σ_X, σ_Y, ρ_{XY} and Z.

5.23. *Nonlinear and linear estimation.* Let X and Y be two random variables with joint pdf

$$f(x, y) = \begin{cases} x + y, & 0 \le x \le 1,\ 0 \le y \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the MMSE estimator of X given Y.

(b) Find the corresponding MSE.

(c) Find the pdf of $Z = E(X|Y)$.

(d) Find the linear MMSE estimator of X given Y.

(e) Find the corresponding MSE.

5.24. *Additive-noise channel with signal dependent noise.* Consider the channel with correlated signal X and noise Z and observation $Y = 2X + Z$, where

$$\mu_X = 1\,, \quad \mu_Z = 0\,, \quad \sigma_X^2 = 4\,, \quad \sigma_Z^2 = 9\,, \quad \rho_{X,Z} = -\tfrac{3}{8}\,.$$

Find the best MSE linear estimate of X given Y.

LECTURE 6

Random Vectors

6.1 DEFINITION AND PROPERTIES

Let $X_1, X_2, \ldots, X_n$ be random variables on the same probability space. We define a *random vector* as

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

In other words, $\mathbf{X}$ is a tuple $(X_1, \ldots, X_n)$ of random variables written in a column vector format. A random matrix can be defined in a similar manner.

The random vector $\mathbf{X}$ is completely specified by its joint cdf

$$F_{\mathbf{X}}(\mathbf{x}) = \mathsf{P}\{X_1 \le x_1,\ X_2 \le x_2,\ \ldots,\ X_n \le x_n\}, \quad \mathbf{x} \in \mathbb{R}^n.$$

If $\mathbf{X}$ is continuous, i.e., $F_{\mathbf{X}}(\mathbf{x})$ is a continuous function of $\mathbf{x}$, then $\mathbf{X}$ can be specified by its joint pdf:

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F_{\mathbf{X}}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n.$$

If $\mathbf{X}$ is discrete, then it can be specified by its joint pmf:

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \mathsf{P}\{X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n\}, \quad \mathbf{x} \in \mathcal{X}^n.$$

A marginal cdf (pdf, pmf) is the joint cdf (pdf, pmf) for a proper subset of the random variables. For example,

$$f_{X_1}(x_1)\,,\ f_{X_2}(x_2)\,,\ f_{X_3}(x_3) f_{X_1,X_2}(x_1, x_2)\,,\ f_{X_1,X_3}(x_1, x_3)\,,\ f_{X_2,X_3}(x_2, x_3).$$

are marginal pdfs of (X_1, X_2, X_3). The marginals can be obtained from the joint in the usual way. For example,

$$F_{X_1}(x_1) = \lim_{x_2, x_3 \to \infty} F_{X_1,X_2,X_3}(x_1, x_2, x_3),$$

$$f_{X_1,X_2}(x_1, x_2) = \int_{-\infty}^{\infty} f_{X_1,X_2,X_3}(x_1, x_2, x_3)\, dx_3.$$

Conditional cdf (pdf, pmf) can also be defined in the usual way. For example,

$$f_{X_3|X_1,X_2} = \frac{f_{X_1,X_2,X_3}(x_1,x_2,x_3)}{f_{X_1,X_2}(x_1,x_2)},$$
$$f_{X_2,X_3|X_1} = \frac{f_{X_1,X_2,X_3}(x_1,x_2,x_3)}{f_{X_1}(x_1)}.$$

More generally, by writing $X^k = (X_1,\ldots,X_k)$ and $X_{k+1}^n = (X_{k+1},\ldots,X_n)$, we have

$$f_{X_{k+1}^n|X^k}(x_{k+1}^n|x^k) = \frac{f_{X^n}(x^n)}{f_{X^k}(x^k)}.$$

By telescoping numerators and denominators of conditional pdfs/pmfs, we can establish the following chain rule:

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}(x_1)f_{X_2|X_1}(x_2|x_1)f_{X_3|X_1,X_2}(x_3|x_1,x_2)\cdots f_{X_n|X^{n-1}}(x_n|x^{n-1}).$$

6.2 INDEPENDENCE AND CONDITIONAL INDEPENDENCE

The random variables $X_1,\ldots,X_n$ are said to be (mutually) independent if

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^n f_{X_i}(x_i), \quad \mathbf{x} \in \mathbb{R}^n.$$

If further $X_1,\ldots,X_n$ have the same marginal distribution, then they are said to be *independent and identically distributed* (i.i.d.).

Example 6.1. If we flip a coin n times independently, we generate i.i.d. Bern(p) random variables X^n.

Let $(X_1,X_2,X_3) \sim f_{X_1,X_2,X_3}(x_1,x_2,x_3)$. The random variables X_1 and X_3 are said to be *conditionally independent* given X_2 if

$$f_{X_1,X_3|X_2}(x_1,x_3|x_2) = f_{X_1|X_2}(x_1|x_2)f_{X_3|X_2}(x_3|x_2), \quad (x_1,x_2,x_3) \in \mathbb{R}^3.$$

Conditional independence neither implies nor is implied by independence.

Example 6.2. Let X_1 and X_2 be i.i.d. Bern(1/2), and $X_3 = X_1 \oplus X_2$. Then X_1 and X_3 are independent, but they are not conditionally independent given X_2.

Example 6.3 (Coin flips with random bias). Let $P \sim \text{Unif}[0,1]$. Given $P = p$, let X_1 and X_2 be i.i.d. Bern(p). By definition, X_1 and X_2 are conditionally independent given P, but they are not independent since

$$\mathsf{P}\{X_1 = 1\} = \mathsf{P}\{X_2 = 1\} = \int_0^1 p\,dp = \frac{1}{2}$$

while

$$\mathsf{P}\{X_1 = X_2 = 1\} = \int_0^1 p^2\,dp = \frac{1}{3} \neq \left(\frac{1}{2}\right)^2.$$

6.3 MEAN AND COVARIANCE MATRIX

The mean (vector) of the random vector $\mathbf{X}$ is

$$\mathsf{E}[\mathbf{X}] = \begin{pmatrix} \mathsf{E}[X_1] \\ \mathsf{E}[X_2] \\ \vdots \\ \mathsf{E}[X_n] \end{pmatrix}.$$

The *covariance matrix* of $\mathbf{X}$ is defined as

$$\begin{aligned} \Sigma_{\mathbf{X}} &= \begin{pmatrix} & \vdots & \\ \cdots & \operatorname{Cov}(X_i, X_j) & \cdots \\ & \vdots & \end{pmatrix} \\ &= \begin{pmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) & \cdots & \operatorname{Cov}(X_1, X_n) \\ \operatorname{Cov}(X_1, X_2) & \operatorname{Var}(X_2) & \cdots & \operatorname{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_1, X_n) & \operatorname{Cov}(X_2, X_n) & \cdots & \operatorname{Var}(X_n) \end{pmatrix}. \end{aligned}$$

Any covariance matrix $\Sigma_{\mathbf{X}}$ must satisfy the following properties.

1. $\Sigma_{\mathbf{X}}$ is *symmetric*, i.e., $\Sigma_{\mathbf{X}} = \Sigma_{\mathbf{X}}^T$.
2. $\Sigma_{\mathbf{X}}$ is *nonnegative definite (positive semidefinite)*, i.e., the *quadratic form* $\mathbf{a}^T\Sigma_{\mathbf{X}}\mathbf{a}$ is nonnegative for every $\mathbf{a} \in \mathbb{R}^n$. Equivalently, all the *eigenvalues* of $\Sigma_{\mathbf{X}}$ are nonnegative.

Conversely, any symmetric nonnegative definite matrix Σ is a covariance matrix of some random vector. To show the second property, we write

$$\Sigma_{\mathbf{X}} = \mathsf{E}\big[(\mathbf{X} - \mathsf{E}[\mathbf{X}])(\mathbf{X} - \mathsf{E}[\mathbf{X}])^T\big]$$

as the expectation of an *outer product* and note that

$$\mathbf{a}^T\Sigma_{\mathbf{X}}\mathbf{a} = \mathsf{E}\big[\mathbf{a}^T(\mathbf{X} - \mathsf{E}[\mathbf{X}])(\mathbf{X} - \mathsf{E}[\mathbf{X}])^T\mathbf{a}\big] = \mathsf{E}\big[(\mathbf{a}^T(\mathbf{X} - \mathsf{E}[\mathbf{X}]))^2\big] \ge 0.$$

Example 6.4. Consider the following matrices:

$$\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \Sigma_2 = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \qquad \Sigma_3 = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{bmatrix},$$

$$\Sigma_4 = \begin{bmatrix} -1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \qquad \Sigma_5 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 3 \end{bmatrix}, \qquad \Sigma_6 = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}.$$

Then, Σ_1, Σ_5, and Σ_6 are covariance matrices, while Σ_2, Σ_3, and Σ_4 are not.

6.4 SUMS OF RANDOM VARIABLES

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random vector and let

$$Y = X_1 + X_2 + \cdots + X_n$$

be their sum. In vector notation,

$$Y = \mathbf{1}^T \mathbf{X},$$

where $\mathbf{1}$ is the all 1 vector. By linearity of expectation, the expected value of Y is

$$\mathsf{E}[Y] = \mathsf{E}[\mathbf{1}^T \mathbf{X}] = \mathbf{1}^T \, \mathsf{E}[\mathbf{X}] = \sum_{i=1}^{n} \mathsf{E}[X_i]. \tag{6.1}$$

Example 6.5 (Mean of the binomial random variable). Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bern(p), representing whether each of n independent coin flips of bias p is a head, and $Y = \sum_{i=1}^{n} X_i$ denote the total number of heads. Then Y is a Binom(n, p) random variable and

$$\mathsf{E}[Y] = \sum_{i=1}^{n} \mathsf{E}[X_i] = np.$$

Note that we did not need independence for this result to hold, i.e., the result holds even if the coin flips are not independent.

We now compute the variance of $Y = \mathbf{1}^T \mathbf{X}$ as

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathsf{E}[(Y - \mathsf{E}(Y))^2] \\
&= \mathsf{E}[(\mathbf{1}^T(\mathbf{X} - \mathsf{E}(\mathbf{X}))^2] \\
&= \mathsf{E}[\mathbf{1}^T(\mathbf{X} - \mathsf{E}(\mathbf{X}))(\mathbf{X} - \mathsf{E}(\mathbf{X}))^T \mathbf{1}] \\
&= \mathbf{1}^T \Sigma_{\mathbf{X}} \mathbf{1} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathrm{Cov}(X_i, X_j).
\end{aligned} \tag{6.2}$$

If $X_1, \ldots, X_n$ are uncorrelated, i.e., $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$, then

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Example 6.6 (Variance of the binomial random variable). Again let $Y = \sum_{i=1}^{n} X_i$, where the $X_1, \dots, X_n$ are i.i.d. Bern(p). Since the X_is are independent, $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq$j. Hence,

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = np(1-p).$$

Example 6.7 (Hats). Suppose n people throw their hats in a box and then each picks one hat at random. Let N be the number of people that get back their own hat. We find $\mathsf{E}[N]$ and $\mathrm{Var}(N)$. We first define the *indicator random variable* X_i that takes value 1 if person i selects her own hat, and 0 otherwise. Then

$$N = \sum_{i=1}^{n} X_i.$$

Since $X_i \sim \mathrm{Bern}(1/n)$, $\mathsf{E}[X_i] = 1/n$ and $\mathrm{Var}(X_i) = (1/n)(1-1/n)$. Furthermore, since

$$p_{X_i,X_j}(1,1) = \frac{1}{n(n-1)}$$

for $i \neq j$,

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= \mathsf{E}[X_i X_j] - \mathsf{E}[X_i]\,\mathsf{E}[X_j] \\ &= \left(\frac{1}{n(n-1)} \cdot 1\right) - \left(\frac{1}{n}\right)^2 \\ &= \frac{1}{n^2(n-1)}, \quad i \neq j.\end{aligned}$$

Hence, by (6.1) and (6.2),

$$\begin{aligned}\mathsf{E}[N] &= n\,\mathsf{E}[X_1] = 1. \\ \mathrm{Var}(N) &= n\,\mathrm{Var}(X_1) + n(n-1)\,\mathrm{Cov}(X_1, X_2) \\ &= \left(1 - \frac{1}{n}\right) + n(n-1)\,\frac{1}{n^2(n-1)} = 1.\end{aligned}$$

Example 6.8 (Sample mean). Let $X_1, X_2, \dots, X_n$ be i.i.d. with finite mean $\mathsf{E}[X]$ and variance $\mathrm{Var}[X]$. The *sample mean* is defined as

$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then, $\mathsf{E}[S_n] = \mathsf{E}[X]$ and

$$\mathrm{Var}(S_n) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \cdot n\,\mathrm{Var}(X) = \frac{1}{n}\,\mathrm{Var}(X).$$

Note that $\lim_{n\to\infty} \mathrm{Var}(S_n) = 0$. This is a very important observation, which will be used in Lecture #7 to establish the weal law of large numbers.

Let N be a random variable taking positive integer values and let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with finite mean $\mathsf{E}[X]$ and variance $\mathrm{Var}(X)$, independent of N. Define the *random* sum

$$Y = \sum_{i=1}^{N} X_i.$$

Given $\mathsf{E}[N]$, $\mathrm{Var}(N)$, $\mathsf{E}[X]$, and $\mathrm{Var}(X)$, we wish to find the mean and variance of Y. By the law of iterated expectation, we have

$$\begin{aligned} \mathsf{E}[Y] &= \mathsf{E}\Big[\mathsf{E}\Big[\sum_{i=1}^{N} X_i \,\Big|\, N\Big]\Big] \\ &= \mathsf{E}\Big[\sum_{i=1}^{N} \mathsf{E}[X_i] \,\Big|\, N\Big] \\ &= \mathsf{E}[N\,\mathsf{E}[X]] \\ &= \mathsf{E}[N]\,\mathsf{E}[X]. \end{aligned}$$

Using the law of conditional variance, the variance is:

$$\begin{aligned} \mathrm{Var}(Y) &= \mathsf{E}[\mathrm{Var}(Y \mid N)] + \mathrm{Var}(\mathsf{E}[Y \mid N]) \\ &= \mathsf{E}[N\,\mathrm{Var}(X)] + \mathrm{Var}(NE[X]) \\ &= \mathsf{E}[N]\,\mathrm{Var}(X) + \mathrm{Var}(N)(\mathsf{E}[X])^2. \end{aligned}$$

Example 6.9 (Network gateway). Let $N \sim \mathrm{Geom}(p)$ be the number of data flows arriving at a gateway in a communication network in some time interval. Assume that the length of flow i is $X_i \sim \mathrm{Exp}(\lambda)$ packets, and that $X_1, X_2, \dots$ and N are mutually independent. Let $Y = \sum_{i=1}^{N} X_i$ be the total number of packets arriving at the gateway. Then,

$$\begin{aligned} \mathsf{E}(Y) &= \mathsf{E}[N]\,\mathsf{E}[X] = \frac{1}{\lambda p}, \\ \mathrm{Var}(Y) &= \mathsf{E}[N]\,\mathrm{Var}(X) + \mathrm{Var}(N)(\mathsf{E}[X])^2 \\ &= \frac{1}{\lambda^2 p} + \frac{1}{\lambda^2} \cdot \frac{1-p}{p^2} = \frac{1}{(\lambda p)^2}. \end{aligned}$$

6.5 GAUSSIAN RANDOM VECTORS

We say that $\mathbf{X} = [X_1 \ \cdots \ X_n]^T$ is a *Gaussian* random vector or $X_1, \dots, X_n$ are *jointly Gaussian* random variables if the joint pdf is of the form

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}. \tag{6.3}$$

It can be readily checked that $\boldsymbol{\mu}$ is the mean and Σ is the covariance matrix of $\mathbf{X}$. Since Σ is invertible (and nonnegative definite since it is a covariance matrix), Σ is positive definite, that is, $\mathbf{a}^T\Sigma\mathbf{a} > 0$ for every $\mathbf{a} \neq \mathbf{0}$. For $n = 2$, the joint pdf in (6.3) simplifies to what we discussed in Lecture #5. We use the notation $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$ to denote a GRV with given mean and covariance matrix.

Since Σ is positive definite, so is Σ^{-1}. Hence, if $\mathbf{x} - \boldsymbol{\mu} \neq \mathbf{0}$,

$$(\mathbf{x} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) > 0\,,$$

which implies that the contours of equal pdf are ellipsoids. The Gaussian random vector $\mathbf{X} \sim \mathrm{N}(\mathbf{0}, \sigma^2 I)$, where I is the identity matrix and $\sigma^2 > 0$, is called *white*; its contours of equal joint pdf are spheres centered at the origin.

Gaussian random vectors $\mathbf{X} = [X_1 \;\cdots\; X_n]^T$ satisfy the following properties.

1. If $X_1, \ldots, X_n$ are uncorrelated, then they are independent. For example, the components of a white Gaussian random vector $\mathbf{X} \sim \mathrm{N}(\mathbf{z}, \sigma^2 I)$ are i.i.d. $\mathrm{N}(0, \sigma^2)$.

2. A linear transformation of $\mathbf{X}$ is also Gaussian, that is, for any $m \times n$ full-rank matrix A with $m \leq n$,
$$\mathbf{Y} = A\mathbf{X} \sim \mathrm{N}(A\boldsymbol{\mu},\; A\Sigma A^T).$$
For example, if
$$\mathbf{X} \sim \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}\right)$$
and
$$\mathbf{Y} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\mathbf{X},$$
then
$$\mathbf{Y} \sim \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\right) = \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} 7 & 3 \\ 3 & 2 \end{bmatrix}\right).$$
This property can be used as an alternative definition of jointly Gaussian random variables, namely, $X_1, \ldots, X_n$ are jointly Gaussian if $\mathbf{a}^T\mathbf{X}$ is a Gaussian random variable for every $\mathbf{a}$. This definition is more general since it includes the degenerate case in which the covariance matrix is singular, i.e., $X_1, \ldots, X_n$ are linearly dependent.

3. Marginals are Gaussian. For example, if X_1, X_2, X_3 are jointly Gaussian, then so are X_1 and X_3. As discussed in Section 5.9 the converse does not hold, that is, marginally Gaussian random variables are not necessarily jointly Gaussian.

4. Conditionals are Gaussian, that is, if
$$\mathbf{X} = \left[\begin{array}{c} \mathbf{U} \\ \hline \mathbf{V} \end{array}\right] \sim \mathrm{N}\left(\left[\begin{array}{c} \boldsymbol{\mu}_{\mathbf{U}} \\ \hline \boldsymbol{\mu}_{\mathbf{V}} \end{array}\right], \left[\begin{array}{c|c} \Sigma_{\mathbf{U}} & \Sigma_{\mathbf{UV}} \\ \hline \Sigma_{\mathbf{VU}} & \Sigma_{\mathbf{V}} \end{array}\right]\right),$$
then
$$\mathbf{V} \,|\, \{\mathbf{U} = \mathbf{u}\} \sim \mathrm{N}\left(\Sigma_{\mathbf{VU}}\Sigma_{\mathbf{U}}^{-1}(\mathbf{u} - \boldsymbol{\mu}_{\mathbf{U}}) + \boldsymbol{\mu}_{\mathbf{V}}\,,\; \Sigma_{\mathbf{V}} - \Sigma_{\mathbf{VU}}\Sigma_{\mathbf{U}}^{-1}\Sigma_{\mathbf{UV}}\right).$$

For example, if

$$\begin{bmatrix} X_1 \\ \hline X_2 \\ X_3 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 1 \\ \hline 2 \\ 2 \end{bmatrix}, \left[\begin{array}{c|cc} 1 & 2 & 1 \\ \hline 2 & 5 & 2 \\ 1 & 2 & 9 \end{array}\right] \right),$$

then

$$\begin{aligned} \begin{bmatrix} X_2 \\ X_3 \end{bmatrix} \mid \{X_1 = x_1\} &\sim \mathrm{N}\left(\begin{bmatrix} 2 \\ 1 \end{bmatrix} (x_1 - 1) + \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 5 & 2 \\ 2 & 9 \end{bmatrix} - \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \end{bmatrix} \right) \\ &= \mathrm{N}\left(\begin{bmatrix} 2x_1 \\ x_1 + 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 8 \end{bmatrix} \right). \end{aligned}$$

For $n = 2$, these properties recover the properties of a pair of jointly Gaussian random variables discussed in Section 5.9.

The first property can be easily verified by noting that Σ and Σ^{-1} are diagonal for uncorrelated $X_1, \ldots, X_n$, and substituting them in the joint pdf. We prove the second property by using the *characteristic function* for $\mathbf{X}$:

$$\begin{aligned} \Phi_{\mathbf{X}}(\boldsymbol{\omega}) &= \mathsf{E}\left(e^{i\boldsymbol{\omega}^T \mathbf{X}} \right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) e^{i\boldsymbol{\omega}^T \mathbf{x}} \, d\mathbf{x}, \end{aligned}$$

where $\boldsymbol{\omega}$ is an n-dimensional real valued vector and $i = \sqrt{-1}$. Since the characteristic function is the inverse of the multi-dimensional Fourier transform of $f_{\mathbf{X}}(\mathbf{x})$, this implies that there is a one-to-one correspondence between $\Phi_{\mathbf{X}}(\boldsymbol{\omega})$ and

$$f_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(2\pi)^n} \Phi_{\mathbf{X}}(\boldsymbol{\omega}) e^{-i\boldsymbol{\omega}^T \mathbf{x}} \, d\boldsymbol{\omega}.$$

Now the characteristic function of $X \sim \mathrm{N}(\mu, \sigma^2)$ is

$$\Phi_X(\omega) = e^{-\frac{1}{2}\omega^2\sigma^2 + i\mu\omega},$$

and more generally, for $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$,

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) = e^{-\frac{1}{2}\boldsymbol{\omega}^T \Sigma \boldsymbol{\omega} + i\boldsymbol{\omega}^T \boldsymbol{\mu}}.$$

Since A is an $m \times n$ matrix, $\mathbf{Y} = A\mathbf{X}$ and $\boldsymbol{\omega}$ are m-dimensional. Therefore, the characteristic function of $\mathbf{Y}$ is

$$\begin{aligned}\Phi_{\mathbf{Y}}(\boldsymbol{\omega}) &= \mathsf{E}\left(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\right)\\ &= \mathsf{E}\left(e^{i\boldsymbol{\omega}^T A\mathbf{X}}\right)\\ &= \Phi_{\mathbf{X}}(A^T\boldsymbol{\omega})\\ &= e^{-\frac{1}{2}(A^T\boldsymbol{\omega})^T\Sigma(A^T\boldsymbol{\omega}) + i\boldsymbol{\omega}^T A\boldsymbol{\mu}}\\ &= e^{-\frac{1}{2}\boldsymbol{\omega}^T(A\Sigma A^T)\boldsymbol{\omega} + i\boldsymbol{\omega}^T A\boldsymbol{\mu}}.\end{aligned}$$

Thus $\mathbf{Y} = A\mathbf{X} \sim \mathrm{N}(A\boldsymbol{\mu}, A\Sigma A^T)$. The third property follows by the second property since a projection operation is linear; for example,

$$\mathbf{Y} = \begin{bmatrix} 1 & 0 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix} X_1\\ X_2\\ X_3\end{bmatrix} = \begin{bmatrix} X_1\\ X_3\end{bmatrix}.$$

Finally, the fourth property follows by the first and second properties, and the orthogonality principle.

6.6 MMSE ESTIMATION: THE VECTOR CASE

Let $X \sim f_X(x)$ be a random variable representing the signal and let $\mathbf{Y}$ be an n-dimensional random vector representing the noisy observations. The MMSE estimate of X given $\mathbf{Y}$ is the conditional expectation $\mathsf{E}[X\,|\,\mathbf{Y}]$.

The linear MMSE estimate is the estimate of the form

$$\hat{X} = \sum_{i=1}^{n} h_i Y_i + h_0$$

that minimizes the MSE

$$\mathsf{E}[(X - \hat{X})^2].$$

As in the scalar case, the LMMSE estimate depends only on the means, variances, and covariances of the random variables involved.

Note first that the LMMSE is attained by the estimate of the form

$$\hat{X} = \sum_{i=1}^{n} h_i(Y_i - \mathsf{E}[Y_i]) + \mathsf{E}[X] = \mathbf{h}^T(\mathbf{Y} - \mathsf{E}[\mathbf{Y}]) + \mathsf{E}[X]. \tag{6.4}$$

To characterize the optimal $\mathbf{h}$, we use the orthogonality principle discussed in Section 5.8.

We view the random variables $X, Y_1, Y_2, \ldots, Y_n$ as vectors in the inner product space, and find $\hat{X}$ such that the error vector $X - \hat{X}$ is orthogonal to any affine function of $\mathbf{Y}$, i.e.,

$$\mathsf{E}[(X - \mathsf{E}[X] - \mathbf{h}^T(\mathbf{Y} - \mathsf{E}[\mathbf{Y}]))Y_i] = 0, \quad i = 1, 2, \ldots, n,$$

or equivalently,

$$\mathsf{E}[(X - \mathsf{E}[X] - \mathbf{h}^T(\mathbf{Y} - \mathsf{E}[\mathbf{Y}]))(Y_i - \mathsf{E}[Y_i])] = 0, \quad i = 1, 2, \ldots, n.$$

Define the *cross covariance* of $\mathbf{Y}$ and X as the n-vector

$$\Sigma_{\mathbf{Y}X} = \mathsf{E}\left[(\mathbf{Y} - \mathsf{E}(\mathbf{Y}))(X - \mathsf{E}(X))\right] = \begin{bmatrix} \sigma_{Y_1X} \\ \sigma_{Y_2X} \\ \vdots \\ \sigma_{Y_nX} \end{bmatrix}.$$

Then, the orthogonality condition can be written in a vector form as

$$\Sigma_{\mathbf{Y}X} = \Sigma_{\mathbf{Y}}\mathbf{h}.$$

If $\Sigma_{\mathbf{Y}}$ is nonsingular, this equation can be solved to obtain

$$\mathbf{h} = \Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}X}.$$

Thus, by substituting in (6.4), the LMMSE estimate is

$$\hat{X} = \Sigma_{\mathbf{Y}X}^T\Sigma_{\mathbf{Y}}^{-1}(\mathbf{Y} - \mathsf{E}[\mathbf{Y}]) + \mathsf{E}[X].$$

Now to find the minimum MSE, consider

$$\begin{aligned}\mathsf{E}[(X - \hat{X})^2] &= \mathsf{E}[(X - \hat{X})(X - \mathsf{E}[X])] - \mathsf{E}[(X - \hat{X})(\hat{X} - \mathsf{E}[X])] \\ &= \mathsf{E}[(X - \hat{X})(X - \mathsf{E}[X])] \\ &= \mathsf{E}[(X - \mathsf{E}[X])^2] - \mathsf{E}[(\hat{X} - \mathsf{E}[X])(X - \mathsf{E}[X])] \\ &= \sigma_X^2 - \mathsf{E}[\Sigma_{\mathbf{Y}X}^T\Sigma_{\mathbf{Y}}^{-1}(\mathbf{Y} - \mathsf{E}[\mathbf{Y}])(X - \mathsf{E}[X])] \\ &= \sigma_X^2 - \Sigma_{\mathbf{Y}X}^T\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}X}.\end{aligned}$$

Example 6.10. Let X be the random variable representing a signal with mean μ and variance P. The observations are

$$Y_i = X + Z_i, \quad i = 1, 2, \ldots, n,$$

where $Z_1, Z_2, \ldots, Z_n$ are zero-mean uncorrelated noise random variables with variance N that are also uncorrelated with X. We find the MMSE linear estimate $\hat{X}$ of X given $\mathbf{Y}$ and its MSE. For $n = 1$, by Example 5.19, we already know that

$$\hat{X} = \frac{P}{P + N}Y_1 + \frac{N}{P + N}\mu.$$

To find the MMSE linear estimate for the general n, first note

$$\begin{aligned} \mathsf{E}[Y_i] &= \mu, \\ \operatorname{Var}(Y_i) &= P + N, \\ \operatorname{Cov}(X, Y_i) &= P, \quad i = 1, 2, \ldots, n. \end{aligned}$$

By the orthogonality principle,

$$\Sigma_{\mathbf{Y}} \mathbf{h} = \Sigma_{\mathbf{Y}X},$$

that is,

$$\begin{bmatrix} P+N & P & \cdots & P \\ P & P+N & \cdots & P \\ \vdots & \vdots & \ddots & \vdots \\ P & P & \cdots & P+N \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix} = \begin{bmatrix} P \\ P \\ \vdots \\ P \end{bmatrix}.$$

By symmetry, $h_1 = h_2 = \cdots = h_n = \dfrac{P}{nP + N}$. Thus

$$\begin{aligned} \hat{X} &= \frac{P}{nP+N} \sum_{i=1}^{n} (Y_i - \mu) + \mu \\ &= \frac{P}{nP+N} \Big(\sum_{i=1}^{n} Y_i \Big) + \frac{N}{nP+N} \mu. \end{aligned}$$

The MSE of the estimate is

$$P - \mathsf{E}[(\hat{X} - \mu)(X - \mu)] = \frac{PN}{nP + N}.$$

Thus as $n \to \infty$, the LMMSE tends to zero as $n \to \infty$, that is, the linear estimate becomes perfect, even though we do not know the complete statistics of X and Y.

PROBLEMS

6.1. *Markov chain.* Assume that the continuous random variables X_1 and X_3 are independent given X_2.

(a) Show that

$$f_{X_3|X_1,X_2}(x_3|x_1, x_2) = f_{X_3|X_2}(x_3|x_2), \quad (x_1, x_2, x_3) \in \mathbb{R}^3.$$

(b) Show that

$$f_{X_1|X_2,X_3}(x_1|x_2, x_3) = f_{X_1|X_2}(x_1|x_2), \quad (x_1, x_2, x_3) \in \mathbb{R}^3.$$

(c) Conclude that

$$\begin{aligned} f_{X_1,X_2,X_3}(x_1, x_2, x_3) &= f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_2}(x_3|x_2) \\ &= f_{X_3}(x_3) f_{X_2|X_3}(x_2|x_3) f_{X_1|X_2}(x_1|x_2), \quad (x_1, x_2, x_3) \in \mathbb{R}^3. \end{aligned}$$

6.2. *Cascade of binary symmetric channels.* Suppose that $X_1 \sim \text{Bern}(1/2)$, $Z_1 \sim \text{Bern}(p_1)$, and $Z_2 \sim \text{Bern}(p_2)$ are independent, $X_2 = X_1 \oplus Z_1$, and $X_3 = X_2 \oplus Z_2 = X_1 \oplus Z_1 \oplus Z_2$, as depicted in Figure 6.1. Assume that $0 < p_1, p_2 < 1/2$.

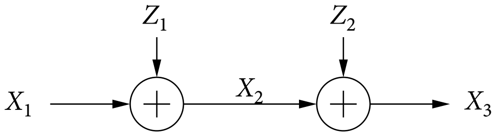


Figure 6.1. Cascade of binary symmetric channels.

(a) Are X_1 and X_2 independent?

(b) Are X_1 and X_3 independent?

(c) Are X_1 and X_2 conditionally independent given X_3?

(d) Are X_1 and X_3 conditionally independent given X_2?

6.3. *Covariance matrices.* Which of the following matrices can be a covariance matrix? Justify your answer either by constructing a random vector **X**, as a function of the i.i.d zero mean unit variance random variables Z_1, Z_2, and Z_3, with the given covariance matrix, or by establishing a contradiction.

$$\text{(a)} \begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix} \quad \text{(b)} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad \text{(c)} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \quad \text{(d)}$$

$$\begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 3 \end{bmatrix}$$

6.4. The *correlation matrix* C for a random vector **X** is the matrix whose entries are $c_{ij} = \mathsf{E}(X_i X_j)$. Show that it has the same properties as the covariance matrix, i.e., that it is real, symmetric, and positive semidefinite definite.

6.5. *Spagetti.* We have a bowl with n spaghetti strands. You randomly pick two strand ends and join them. The process is continued until there are no ends left. Let L be the number of spaghetti loops formed. Find $\mathsf{E}[L]$.

6.6. *Gaussian random vector.* Given a Gaussian random vector $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (1\ 5\ 2)^T$ and

$$\Sigma = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix}.$$

(a) Find the pdfs of

i.X_1,

ii.$X_2 + X_3$,
iii.$2X_1 + X_2 + X_3$,
iv.X_3 given (X_1, X_2), and
v.(X_2, X_3) given X_1.

(b) What is $\mathsf{P}\{2X_1 + X_2 - X_3 < 0\}$? Express your answer using the Q function.

(c) Find the joint pdf on $\mathbf{Y} = A\mathbf{X}$, where

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & -1 & 1 \end{bmatrix}.$$

6.7. *Gaussian Markov chain.* Let X, Y, and Z be jointly Gaussian random variables with zero mean and unit variance, i.e., $\mathsf{E}(X) = \mathsf{E}(Y) = \mathsf{E}(Z) = 0$ and $\mathsf{E}(X^2) = \mathsf{E}(Y^2) = \mathsf{E}(Z^2) = 1$. Let $\rho_{X,Y}$ denote the correlation coefficient between X and Y, and let $\rho_{Y,Z}$ denote the correlation coefficient between Y and Z. Suppose that X and Z are conditionally independent given Y.

(a) Find $\rho_{X,Z}$ in terms of $\rho_{X,Y}$ and $\rho_{Y,Z}$.

(b) Find the MMSE estimate of Z given (X, Y) and the corresponding MSE.

6.8. *Sufficient statistic.* The bias of a coin is a random variable $P \sim \mathrm{U}[0, 1]$. Let $Z_1, Z_2, \ldots, Z_{10}$ be the outcomes of 10 coin flips. Thus $Z_i \sim \mathrm{Bern}(P)$ and $Z_1, Z_2, \ldots, Z_{10}$ are conditionally independent given P. If X is the total number of heads, then $X|\{P = p\} \sim \mathrm{Binom}(10, p)$. Assuming that the total number of heads is 9, show that

$$f_{P|Z_1,Z_2,\ldots,Z_{10}}(p|z_1, z_2, \ldots, z_{10}) = f_{P|X}(p|9)$$

is independent of the order of the outcomes.

6.9. *Order statistics.* Let X_1, X_2, X_3 be independent and uniformly drawn from the interval $[0, 1]$. Let Y_1 be the smallest of X_1, X_2, X_3, let Y_2 be the median (second smallest) of X_1, X_2, X_3, and let Y_3 be the largest of X_1, X_2, X_3. For example, if $X_1 = .3, X_2 = .1, X_3 = .7$, then $Y_1 = .1, Y_2 = .3, Y_3 = .7$. The random variables Y_1, Y_2, Y_3 are called the *order statistics* of X_1, X_2, X_3.

(a) What is the probability $\mathsf{P}\{X_1 \le X_2 \le X_3\}$?

(b) Find the pdf of Y_1.

(c) Find the pdf of Y_3.

(d) (Difficult.) Find the pdf of Y_2.
(Hint: $Y_2 \le y$ if and only if at least two among X_1, X_2, X_3 are $\le y$.)

6.10. *Drawing balls without replacement.* Suppose that we have an urn containing one red ball and $n - 1$ white balls. Each time we draw a ball at random from the urn without replacement (so after the n-th drawing, there is no ball left in the urn). For $i = 1, 2, \ldots, n$, let

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th ball is red,} \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find $\mathsf{E}[X_i]$, $i = 1, 2, \dots, n$.

(b) Find $\mathrm{Var}(X_i)$ and $\mathrm{Cov}(X_i, X_j)$, $i, j = 1, 2, \dots, n$.

6.11. *Packet switching.* Let N be the number of packets per unit time arriving at a network switch. Each packet is routed to output port 1 with probability p and to output port 2 with probability $1 - p$, independent of N and of other packets. Let X be the number of packets per unit time routed to output port 1. Thus

$$X = \begin{cases} 0 & N = 0 \\ \sum_{i=1}^{N} Z_i & N > 0 \end{cases} \qquad \text{where} \quad Z_i = \begin{cases} 1 & \text{packet } i \text{ routed to Port 1} \\ 0 & \text{packet } i \text{ routed to Port 2,} \end{cases}$$

and $Z_1, Z_2, \dots, Z_N$ are conditionally independent given N. Suppose that $N \sim \mathrm{Poisson}(\lambda)$, i.e., has Poisson pmf with parameter λ.

(a) Find the mean and variance of X.

(b) Find the pmf of X and the pmf of $N - X$.

6.12. *Winner of a race.* Two horses are racing on a track. Let X and Y be the finish times of horse 1 and horse 2, respectively. Suppose X and Y are independent and identically distributed Exp(1) random variables, that is,

$$\mathsf{P}\{X > x, Y > y\} = e^{-x} e^{-y}$$

for all $x, y \geq 0$. Let W denote the index of the winning horse. Then $W = 1$ (i.e., horse 1 wins the race) if $X < Y$, and $W = 2$ if $X \geq Y$.

(a) Find $\mathsf{P}\{W = 2\}$.

(b) Find $\mathsf{P}\{W = 2 \,|\, Y = y\}$ for $y \geq 0$.

(c) Suppose we wish to guess which horse won the race based on the finish time of one horse only, say, Y. Find the optimal decision rule $D(y)$ that minimizes the probability of error $\mathsf{P}\{W \neq D(Y)\}$.

(d) Find the minimum probability of error in part (c).

Hint: The following facts might be useful:

$$\begin{aligned}
\int_0^t e^{-x}\, dx &= 1 - e^{-t}, \\
\int_t^\infty e^{-x}\, dx &= e^{-t}, \\
\int_0^t e^{-2x}\, dx &= \frac{1}{2}(1 - e^{-2t}), \\
\int_t^\infty e^{-2x}\, dx &= \frac{1}{2} e^{-2t}, \\
e^{-\ln 2} &= \frac{1}{2}, \\
e^{-2\ln 2} &= \frac{1}{4}.
\end{aligned}$$

6.13. *Estimation.* Let X and Y be independent and identically distributed random variables, $X, Y \sim \text{Unif}[-\frac{1}{2}, \frac{1}{2}]$ and let $Z = X + Y^2$.

(a) Find the conditional density $f_{Z|X}(z|y)$.

(b) Find the MMSE estimate of Z given Y.

(c) Find the MSE of the MMSE estimate in part (b).

6.14. *Noise cancellation.* A classical problem in statistical signal processing involves estimating a weak signal (e.g., the heart beat of a fetus) in the presence of a strong interference (the heart beat of its mother) by making two observations; one with the weak signal present and one without (by placing one microphone on the mother's belly and another close to her heart). The observations can then be combined to estimate the weak signal by "canceling out" the interference. The following is a simple version of this application.

Let the weak signal X be a random variable with mean μ and variance P, and the observations be $Y_1 = X + Z_1$ (Z_1 being the strong interference), and $Y_2 = Z_1 + Z_2$ (Z_2 is a measurement noise), where Z_1 and Z_2 are zero mean with variances N_1 and N_2, respectively. Assume that X, Z_1 and Z_2 are uncorrelated. Find the MMSE linear estimate of X given Y_1 and Y_2 and its MSE. Interpret the results.

6.15. *Additive nonwhite Gaussian noise channel.* Let $Y_i = X + Z_i$, $i = 1, 2, \ldots, n$, be n observations of a signal $X \sim \text{N}(0, P)$. The additive noise random variables $Z_1, Z_2, \ldots, Z_n$ are zero mean jointly Gaussian random variables that are independent of X and have correlation $\mathsf{E}(Z_i Z_j) = N \cdot 2^{-|i-j|}$ for $1 \le i, j \le n$.

(a) Find the best MSE estimate of X given $Y_1, Y_2, \ldots, Y_n$.

(b) Find the MSE of the estimate in part (a).

Hint: The coefficients for the best estimate are of the form $\mathbf{h}^T = [\, a \;\; b \;\; b \; \cdots \; b \;\; b \;\; a \,]$.

6.16. *Nonlinear estimator.* Consider a channel with the observation $Y = XZ$, where the signal X and the noise Z are uncorrelated Gaussian random variables. Let $\mathsf{E}[X] = 1$, $\mathsf{E}[Z] = 2$, $\sigma_X^2 = 5$, and $\sigma_Z^2 = 8$.

(a) Using the fact that $\mathsf{E}(W^3) = \mu^3 + 3\mu\sigma^2$ and $\mathsf{E}(W^4) = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$ for $W \sim \mathcal{N}(\mu, \sigma^2)$, find the mean and covariance matrix of $[X \;\; Y \;\; Y^2]^T$.

(b) Find the MMSE linear estimate of X given Y and the corresponding MSE.

(c) Find the MMSE linear estimate of X given Y^2 and the corresponding MSE.

(d) Find the MMSE linear estimate of X given Y and Y^2 and the corresponding MSE.

(e) Compare your answers in parts (b) through (d). Is the MMSE estimate of X given Y (namely, $\mathsf{E}(X|Y)$) linear?

6.17. *Prediction.* Let **X** be a random process with zero mean and covariance matrix

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n-1} \\ \alpha & 1 & \alpha & & \\ \alpha^2 & \alpha & 1 & & \\ \vdots & & & \ddots & \\ \alpha^{n-1} & & & \cdots & 1 \end{bmatrix}$$

for $|\alpha| < 1$. $X_1, X_2, \ldots, X_{n-1}$ are observed, find the best linear MSE estimate (predictor) of X_n. Compute its MSE.

LECTURE 7

Convergence

7.1 MOTIVATION

Suppose we wish to estimate the *statistics* of a random variable, e.g., its mean, variance, and distribution. To estimate such a statistic, we collect *samples* and use an *estimator* in the form of a *sample average*. But how good is the *estimator*? Does it "converge" to the true statistic? How many samples do we need to ensure with some *confidence* that we are within a certain range of the true value of the statistic? These are questions that often arise in statistics and learning, commonly referred to as parametric and nonparametric estimation.

In communications and signal processing, one faces another type of estimation, namely, estimation or detection of a signal from noisy observations. We then ask if a given estimator converges to the true signal or how many observations are needed to achieve a desired estimation accuracy.

The subject of convergence and limit theorems for random variables addresses such questions. As a motivating example, we consider the problem of estimating the mean of a random variable. Let X be a random variable with finite but unknown mean $\mathsf{E}[X]$. To estimate the mean we generate $X_1, \ldots, X_n$ i.i.d. samples drawn according to the same distribution as X and compute the *sample mean*

$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Does S_n converge to $\mathsf{E}[X]$ as we increase n? If so, how fast? More concretely, let $X_1, \ldots, X_n$ be i.i.d. N(0, 1). Figure 7.1 plots the sample mean S_n from 6 sets of outcomes of $X_1, \ldots, X_n$. Note that each s_n sequence appears to be converging to 0, the mean of the random variables, as n increases.

But what does it mean to say that a sequence of *random variables* $S_1, S_2, \ldots$ converges to $\mathsf{E}[X]$? For a sequence of *real numbers* $a_1, a_2, \ldots$, the limit

$$a = \lim_{n\to\infty} a_n$$

exists if for every $\epsilon > 0$, there exists $n(\epsilon)$ such that $|a_m - a| < \epsilon$ for every $m \geq n(\epsilon)$. For a sequence of random variables, there are several different notions of convergence.

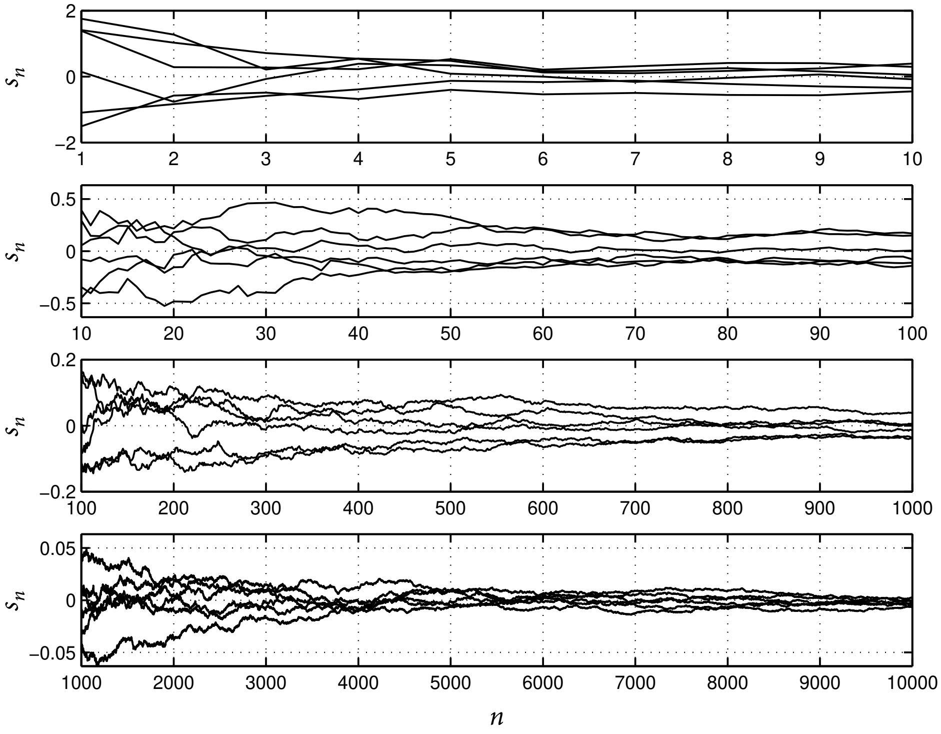


Figure 7.1. Convergence of the sample mean to the ensemble mean.

7.2 ALMOST SURE CONVERGENCE

Consider a sequence of random variables $X_1, X_2, \ldots$, all defined on the same probability space Ω. For every $\omega \in \Omega$, we obtain a *sample sequence* (or *sample path*) $X_1(\omega), X_2(\omega), \ldots$, which is a sequence of real numbers. We say that the sequence of random variables $X_1, X_2, X_3, \ldots$ converges to a random variable X *almost surely* (or *with probability 1*) if

$$\mathsf{P}\{\omega\colon \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1.$$

In other words, the set of sample paths that converge to $X(\omega)$, in the sense of a sequence converging to a limit, has probability 1. Equivalently, the sequence $X_1, X_2, \ldots$ converges to X almost surely if for every $\epsilon > 0$,

$$\begin{aligned}\lim_{m\to\infty} &\mathsf{P}\{|X_n - X| < \epsilon \text{ for every } n \geq m\} \\ &= \mathsf{P}\big(\cup_{m=1}^{\infty} \cap_{n=m}^{\infty} \{|X_n - X| < \epsilon\}\big) = 1.\end{aligned}$$

Example 7.1. Let $X_1, \ldots, X_n$ be i.i.d. Bern(1/2), and define $Y_n = 2^n \prod_{i=1}^n X_i$. Then for any $\epsilon > 0$ (such that $\epsilon < 2^m$),

$$\begin{aligned}\mathsf{P}\{|Y_n - 0| < \epsilon \text{ for all } n \geq m\} &= \mathsf{P}\{X_n = 0 \text{ for some } n \leq m\}\\ &= 1 - \mathsf{P}\{X_n = 1 \text{ for all } n \leq m\}\\ &= 1 - (\tfrac{1}{2})^m,\end{aligned}$$

which converges to 1 as $m \to \infty$. Hence, the sequence Y_n converges to 0 almost surely.

An important example of almost sure convergence is the *strong law of large numbers* (SLLN), which states that if $X_1, \ldots, X_n$ are i.i.d. with finite mean $\mathsf{E}[X]$, then the sequence of sample means S_n converges to the mean $\mathsf{E}[X]$ almost surely. The example shown in Figure 7.1 is a good demonstration of the SLLN — each of the 6 sample paths appears to be converging to 0, which is $\mathsf{E}[X]$, and the probability of such sample paths is 1. The proof of the SLLN is beyond the scope of this course and omitted.

7.3 CONVERGENCE IN MEAN SQUARE

We say that a sequence of random variables $X_1, X_2, \ldots$ converges to a random variable X *in mean square* (m.s.) if

$$\lim_{n\to\infty} \mathsf{E}[(X_n - X)^2] = 0.$$

Example 7.2. Let $X_1, X_2, \ldots$ be i.i.d. with finite mean $\mathsf{E}[X]$ and variance $\mathrm{Var}(X)$. Then, $S_n \to \mathsf{E}[X]$ in m.s. as $n \to \infty$. In other words,

$$\lim_{n\to\infty} \mathsf{E}[(S_n - \mathsf{E}(X))^2] = 0.$$

Tho show this, first note that S_n is an *unbiased* estimate of $\mathsf{E}[X]$, i.e.,

$$\mathsf{E}[S_n] = \mathsf{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathsf{E}[X_i] = \mathsf{E}[X].$$

Now to prove the convergence in m.s., consider

$$\begin{aligned}\mathsf{E}[(S_n - \mathsf{E}[X])^2] &= \mathsf{E}[(S_n - \mathsf{E}[S_n])^2]\\ &= \mathsf{E}\left[\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n \mathsf{E}[X]\right)^2\right]\\ &= \frac{1}{n^2}\mathsf{E}\left[\left(\sum_{i=1}^n (X_i - \mathsf{E}[X])\right)^2\right]\\ &= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right)\end{aligned}$$

$$\begin{aligned} &= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathrm{Var}(X_i)\right) \\ &= \frac{1}{n}\,\mathrm{Var}(X), \end{aligned}$$

which tends to zero as $n \to \infty$. Note that the proof works even if the random variables are only pairwise independent or even only uncorrelated.

Example 7.3. Consider the linear MMSE estimates Example 6.10 as a sequence of random variables $\hat{X}_1, \hat{X}_2, \ldots$, where $\hat{X}_n$ is the LMMSE estimate of X given the first n observations. This sequence converges in m.s. to X since the sequence of MSEs converges to 0 as $n \to \infty$.

Mean square convergence does not necessarily imply almost sure convergence.

Example 7.4. Consider a sequence of random variables

$$X_n = \begin{cases} 0 & \text{w.p. } 1 - 1/n, \\ 1 & \text{w.p. } 1/n. \end{cases}$$

Since $\mathsf{E}[X_n^2] = 1/n$, this sequence converges to 0 in m.s. It does not, however, converge almost surely, since for $0 < \epsilon < 1$ and any m

$$\begin{aligned} \mathsf{P}\{|X_n - 0| < \epsilon \text{ for all } n \geq m\} &= \lim_{n\to\infty} \prod_{i=m}^{n} \left(1 - \frac{1}{i}\right) \\ &= \lim_{n\to\infty} \prod_{i=m}^{n} \left(\frac{i-1}{i}\right) \\ &= \lim_{n\to\infty} \frac{(m-1)}{m}\frac{m}{(m+1)} \cdots \frac{(n-1)}{n} \\ &= \lim_{n\to\infty} \frac{m-1}{n} = 0. \end{aligned}$$

Almost sure convergence does not imply mean square convergence either.

Example 7.5. Consider the sequence Y_n in Example 7.1. Since

$$\mathsf{E}[(Y_n - 0)^2] = \left(\frac{1}{2}\right)^n 2^{2n} = 2^n\,,$$

the sequence does not converge in m.s. even though it converges almost surely.

7.4 CONVERGENCE IN PROBABILITY

We say that a sequence of random variables $X_1, X_2, \ldots$ converges to X *in probability* if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathsf{P}\{|X_n - X| < \epsilon\} = 1.$$

Clearly, almost sure convergence implies convergence in probability. The converse is not necessarily true.

Example 7.6. Let $X_1, X_2, \ldots$ be independent and

$$X_n = \begin{cases} 0 & \text{w.p. } 1 - 1/n, \\ n & \text{w.p. } 1/n. \end{cases}$$

This sequence converges in probability to 0, since

$$\mathsf{P}\{|X_n - 0| > \epsilon\} = \mathsf{P}\{X_n > \epsilon\} = \frac{1}{n} \to 0 \text{ as } n \to \infty.$$

But it does not converge almost surely, since as in Example 7.4,

$$\lim_{m\to\infty} \mathsf{P}\{|X_n - 0| < \epsilon \text{ for all } n \geq m\} = 0.$$

Convergence in mean square implies convergence in probability. Indeed, by the Markov inequality, for any $\epsilon > 0$,

$$\begin{aligned} \mathsf{P}\{|X_n - X| > \epsilon\} &= \mathsf{P}\{(X_n - X)^2 > \epsilon^2\} \\ &\leq \frac{\mathsf{E}(X_n - X)^2}{\epsilon^2}. \end{aligned}$$

Hence, if $X_n \to X$ in m.s., i.e.,

$$\lim_{n\to\infty} \mathsf{E}\left[(X_n - X)^2\right] = 0,$$

then $X_n \to X$ in probability, i.e.,

$$\lim_{n\to\infty} \mathsf{P}\{|X_n - X| > \epsilon\} = 0.$$

The converse is not necessarily true. In Example 7.6, X_n converges in probability, but

$$\mathsf{E}[(X_n - 0)^2] = n$$

and thus X_n does not converge in m.s. Hence, convergence in probability is weaker than both almost sure convergence and mean square convergence.

The most important example of convergence in probability is the weak law of large numbers (WLLN).

Weak law of large numbers. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite mean $\mathsf{E}[X]$ and variance $\mathrm{Var}(X)$. Then the sample mean

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

converges to $\mathsf{E}[X]$ in probability.

We already proved that $S_n \to \mathsf{E}[X]$ in m.s., and since convergence in m.s. implies convergence in probability, $S_n \to \mathsf{E}[X]$ in probability. Moreover, the WLLN requires only that the random variables are uncorrelated (at the cost of finite variance), while the SLLN requires independence.

Example 7.7 (Confidence interval). Given $\epsilon, \delta > 0$, how large should n, the number of samples, be so that

$$\mathsf{P}\{|S_n - \mathsf{E}[X]| \le \epsilon\} \ge 1 - \delta\,,$$

that is, S_n is within ± ϵ of $\mathsf{E}[X]$ with probability $\ge 1 - \delta$? To answer this question, we use the Chebyshev inequality:

$$\begin{aligned}\mathsf{P}\{|S_n - \mathsf{E}[X]| \le \epsilon\} &= \mathsf{P}\{|S_n - \mathsf{E}[S_n]| \le \epsilon\}\\ &\ge 1 - \frac{\mathrm{Var}(S_n)}{\epsilon^2}\\ &= 1 - \frac{\mathrm{Var}(X)}{n\epsilon^2}.\end{aligned}$$

Hence, n should satisfy

$$\frac{\mathrm{Var}(X)}{n\epsilon^2} \le \delta,$$

or equivalently,

$$n \ge \frac{\mathrm{Var}(X)}{\delta\epsilon^2}.$$

For example, if $\epsilon = 0.1\sigma_X$ and $\delta = 0.001$, then the number of samples should satisfy

$$n \ge \frac{\sigma_X^2}{0.001 \times 0.01\sigma_X^2} = 10^5.$$

In other words, 10^5 samples ensure that one can be 99.9% confident that the unknown true mean $\mathsf{E}[X]$ lies within $\pm 0.1\sigma_X$ of the sample mean S_n from the data, *regardless* of the distribution of X.

7.5 CONVERGENCE IN DISTRIBUTION

We say that a sequence of random variables $X_1, X_2, \ldots$ converges *in distribution* (or *weakly*) to X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for every x at which $F_X(x)$ is continuous. Convergence in probability implies convergence in distribution — and convergence in distribution (weak convergence) is the weakest form of convergence we discuss.

The most important example of convergence in distribution is the central limit theorem (CLT).

Central limit theorem. Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite mean $\mathsf{E}[X]$ and variance σ_X^2. Then, the *normalized* sum

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mathsf{E}(X)}{\sigma_X}$$

converges to $Z \sim \mathrm{N}(0, 1)$ in distribution, i.e., $\lim_{n\to\infty} F_{Z_n}(z) = \Phi(z)$.

Example 7.8. Let $X_1, X_2, \ldots$ be i.i.d. Unif$[-1, 1]$ random variables. Then the pdf of the normalized sum

$$Z_n = \frac{1}{\sqrt{n/3}} \sum_{i=1}^{n} X_i$$

converges to that of the standard normal random variable, as illustrated in Figure 7.2 for $n = 1, 2, 4, 16$. Note how quickly the pdf converges. It can be shown that convergence of the pdf implies convergence of the cdf (weak convergence).

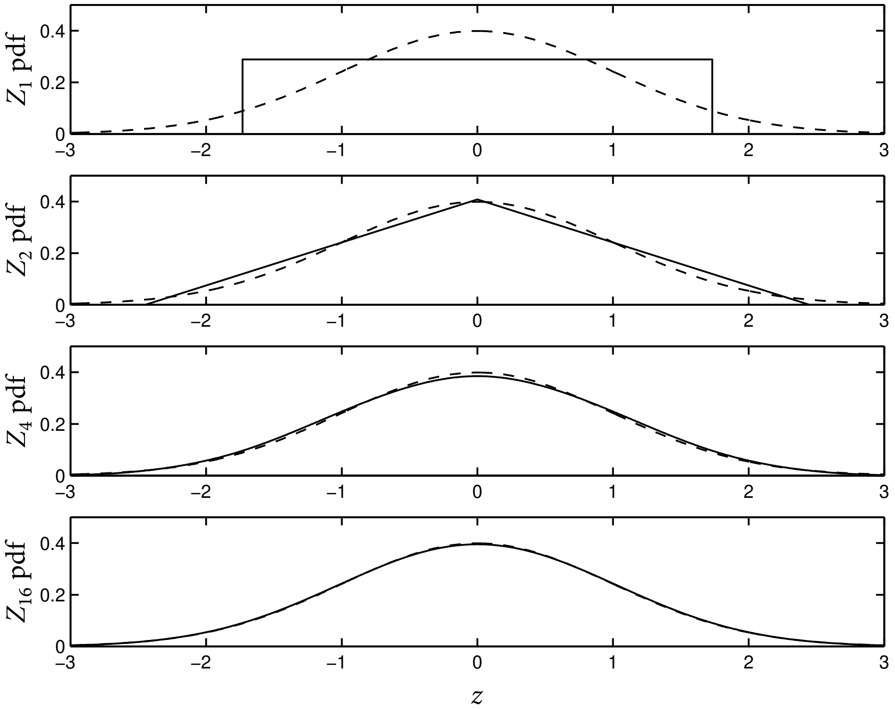


Figure 7.2. The pdfs of the normalized sums Z_n in Example 7.8.

Example 7.9. Let $X_1, X_2, \ldots$ be i.i.d. Bern(1/2). The normalized sum $Z_n = \sum_{i=1}^n (X_i - 0.5)/\sqrt{n/4}$ is discrete and thus has no pdf, but its cdf converges to the Gaussian cdf, as depicted in Figure 7.3 for $n = 10, 20, 160$.

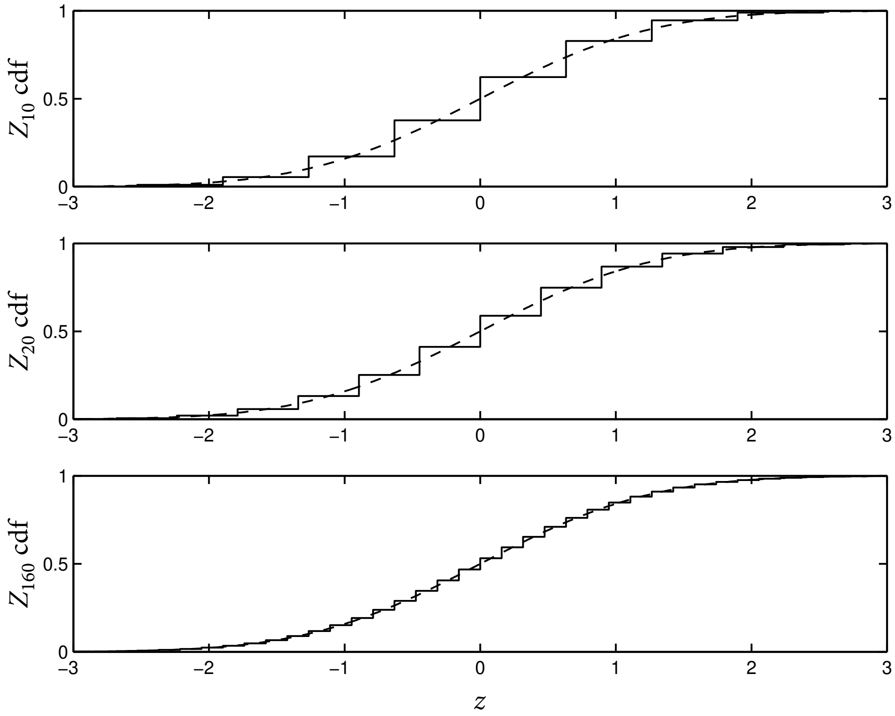


Figure 7.3. The cdfs of the normalized sums Z_n in Example 7.9.

Example 7.10 (Confidence interval). Let $X_1, X_2, \ldots$ be i.i.d. with finite mean $\mathsf{E}[X]$ and variance $\mathrm{Var}(X)$ and let S_n be the sample mean. Given ϵ, $\delta > 0$, how large should n be so that

$$\mathsf{P}\{|S_n - \mathsf{E}(X)| \le \epsilon\} \ge 1 - \delta\,?$$

We can use the CLT to find an estimate of n:

$$\begin{aligned}\mathsf{P}\{|S_n - \mathsf{E}[S_n]| \le \epsilon\} &= \mathsf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathsf{E}[X])\right| \le \epsilon\right\} \\ &= \mathsf{P}\left\{\left|\frac{1}{\sigma_X\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathsf{E}[X])\right| \le \frac{\epsilon\sqrt{n}}{\sigma_X}\right\} \\ &\approx 1 - 2Q\left(\frac{\epsilon\sqrt{n}}{\sigma_X}\right).\end{aligned}$$

For $\epsilon = 0.1\sigma_X$, $\delta = 0.001$, set $2Q(0.1\sqrt{n}) = 0.001$, so $0.1\sqrt{n} = 3.3$ or $n = 1089$ — much smaller than $n \geq 10^5$ obtained in Example 7.7 by the Chebyshev inequality.

The CLT applies to i.i.d. sequences of random vectors. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be a sequence of i.i.d. k-dimensional random vectors with finite mean $\boldsymbol{\mu}$ and nonsingular covariance matrix Σ. Define the sequence of random vectors $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$ by

$$\mathbf{Z}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{X}_i - \boldsymbol{\mu}).$$

Then the CLT for random vectors states that as $n \to \infty$

$$\mathbf{Z}_n \to \mathbf{Z} \sim \mathrm{N}(\mathbf{z}, \Sigma) \text{ in distribution.}$$

Example 7.11. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be a sequence of i.i.d. 2-dimensional random vectors with

$$f_{\mathbf{X}_1}(x_{11}, x_{12}) = \begin{cases} x_{11} + x_{12} & 0 < x_{11} < 1,\ 0 < x_{12} < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 7.2 shows the joint pdf of $\mathbf{Y}_n = \sum_{i=1}^{n} \mathbf{X}_i$ for $n = 1, 2, 3, 4$. Note how quickly it looks Gaussian.

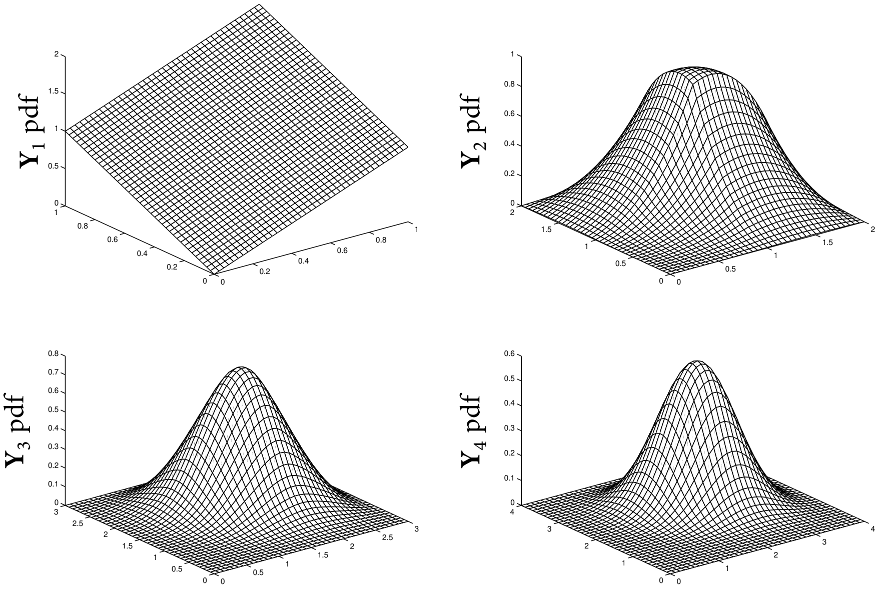


Figure 7.4. Plots of the joint pdf of $\mathbf{Y}_n$ in Example 7.11.

PROBLEMS

7.1. Consider the sequence of i.i.d. random variables $X_1, X_2, \ldots$ with mean $\mathsf{E}(X_1) = 2$ and finite variance. Define the sequence

$$Y_n = \begin{cases} X_n, & \text{for all } n \quad \text{w.p. } \frac{1}{3}, \\ \frac{1}{2}X_n, & \text{for all } n \quad \text{w.p. } \frac{1}{3}, \\ 0, & \text{for all } n \quad \text{w.p. } \frac{1}{3}. \end{cases}$$

Let

$$M_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Define the random variable (or constant) that M_n converges to (in probability) as n approaches infinity and prove the convergence.

7.2. *Roundoff errors.* The sum of a list of 100 real numbers is to be computed. Suppose that these numbers are rounded off to the nearest integer so that each number has an error that is uniformly distributed in the interval $(-0.5, 0.5)$. Use the central limit theorem to estimate the probability that the total error in the sum of the 100 numbers exceeds 6.

7.3. The signal received over a wireless communication channel can be represented by two sums

$$X_{1n} = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} Z_j \cos\Theta_j, \text{ and}$$

$$X_{2n} = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} Z_j \sin\Theta_j,$$

where $Z_1, Z_2, \ldots$ are i.i.d. with mean μ and variance σ^2 and $\Theta_1, \Theta_2, \ldots$ are i.i.d. $\mathrm{U}[0, 2\pi]$ independent of $Z_1, Z_2, \ldots$. Find the distribution of $\begin{bmatrix} X_{1n} \\ X_{2n} \end{bmatrix}$ as n approaches ∞.

7.4. *Polya's urn.* An urn initially has one red ball and one white ball. Let X_1 denote the name of the first ball drawn from the urn. Replace that ball and one like it. Let X_2 denote the name of the next ball drawn. Replace it and one like it. Continue, drawing and replacing.

(a) Argue that the probability of drawing k reds follwed by $n - k$ whites is

$$\frac{1}{2}\cdot\frac{2}{3}\cdots\frac{k}{k+1}\cdot\frac{1}{(k+2)}\cdots\frac{(n-k)}{(n+1)} = \frac{k!(n-k)!}{(n+1)!} = \frac{1}{(n+1)}\frac{1}{\binom{n}{k}}.$$

(b) Let P_n be the proportion of red balls in the urn after the n-th drawing. Argue that $\mathsf{P}\{P_n = \frac{k}{n+2}\} = \frac{1}{n+1}$, for $k = 1, 2, \ldots, n+1$. Thus all proportions are equally probable. This shows that P_n tends to a uniformly distributed random variable in distribution, i.e.,

$$\lim_{n\to\infty} \mathsf{P}\{P_n \leq t\} \longrightarrow t, \quad 0 \leq t \leq 1.$$

(c) What can you say about the behavior of the proportion P_n if you started initially with one red ball in the urn and *two* white balls? Specifically, what is the limiting distribution of P_n? Can you show $\mathsf{P}\{P_n = \frac{k}{n+3}\} = \frac{k}{n+2}$, for $k = 1, 2, \ldots, n+1$?

7.5. Let $X_1, X_2, ..., X_n$ be i.i.d. random variables with $X_i \sim \text{Exp}(\lambda)$. Show that the sequence of r.v.s $Y_n = \min\{X_1, X_2, \ldots, X_n\}$ converges w.p.1. What is the limit?

7.6. Let $X_1, X_2, \ldots$ be independent random variables with the same finite mean $\mu \neq 0$ and variance σ^2. Find the limit of $\mathsf{P}\{\frac{1}{n}\sum_{i=1}^{n} X_i < \frac{\mu}{2}\}$ as n approaches infinity.

7.7. *Convergence* Consider the following sequences of random variables defined on the probability space $(\Omega, \mathcal{F}, \mathsf{P})$, where $\Omega = \{0, 1, 2, \ldots, m-1\}$, $\mathcal{F}$ is the collection of all subsets of Ω, and P is the uniform distribution over Ω.

$$X_n(w) = \begin{cases} \frac{1}{n}, & \omega = n \bmod m \\ 0, & \text{otherwise} \end{cases}$$

$$Y_n(w) = \begin{cases} 2^n, & \omega = 1 \\ 0, & \text{otherwise} \end{cases}$$

$$Z_n(w) = \begin{cases} 1, & \omega = 1 \\ 0, & \text{otherwise} \end{cases}$$

Which of these sequences converges to zero

(a) with probability one?

(b) in mean square?

(c) in probability?

7.8. Consider a coin with random bias P. Flip it n times independently to generate $X_1, X_2, \ldots, X_n$, where $X_i = 1$ if the ith outcome is *heads*, and $X_i = 0$ otherwise. Let S_n be the sample average. Prove that S_n converges to P in probability.
(Hint: First prove convergence in mean square.)

7.9. *Convergence.*
Let $X_1, X_2, X_3, \ldots$ be i.i.d. according to the following distribution

$$f_{X_i}(x) = \begin{cases} \frac{1}{x}, & \text{if } 1 \leq x \leq e, \\ 0, & \text{otherwise.} \end{cases}$$

Let $Y_n = \sqrt[n]{(X_1 \cdot X_2 \cdots X_n)}$, for any positive integer n. Show that the sequence $Y_1, Y_2, Y_3, \ldots$ converges to $e^{\frac{1}{2}}$ with probability 1.

LECTURE 8

Random Processes

8.1 DEFINITION

A *random process* (or *stochastic process*) is an infinite indexed collection of random variables

$$\{X(t) : t \in \mathcal{T}\},$$

defined over a common probability space. The index parameter t is typically time, but can also be a spatial dimension. Random processes are used to model random experiments that evolve in space/time:

- Received sequence/waveform at the output of a communication channel
- Packet arrival times at a node in a communication network
- Thermal noise in a resistor
- Scores of an NBA team in consecutive games
- Daily price of a stock
- Winnings or losses of a gambler
- Contents in memory cells

We are interested in several questions involving random processes.

- *Dependencies of the random variables of the process.* How do future received values depend on past received values? How do future prices of a stock depend on its past values?
- *Long-term averages.* What is the proportion of time a queue is empty? What is the average noise power at the output of a circuit?
- *Extreme or boundary events.* What is the probability that a link in a communication network is congested? What is the probability that the maximum power in a power distribution line is exceeded? What is the probability that a gambler will lose all his capital?
- *Estimation/detection.* How best can one recover a signal from a noisy waveform? How can we estimate a DNA sequence from noisy reads?

A random process can be viewed as a function $X(t, \omega)$ of two variables, the time $t \in \mathcal{T}$ and the outcome $\omega \in \Omega$, where Ω is the space of the underlying random experiment. There are thus two ways to view it. First, for fixed t, $X(t, \omega)$ is a random variable over Ω. In this view, the random process is an index collection of random variables. Second, for fixed ω, $X(t, \omega)$ is a deterministic function of t, called a *sample function*, as illustrated in Figure 8.1. In this view, the random process is a randomly drawn function in t.

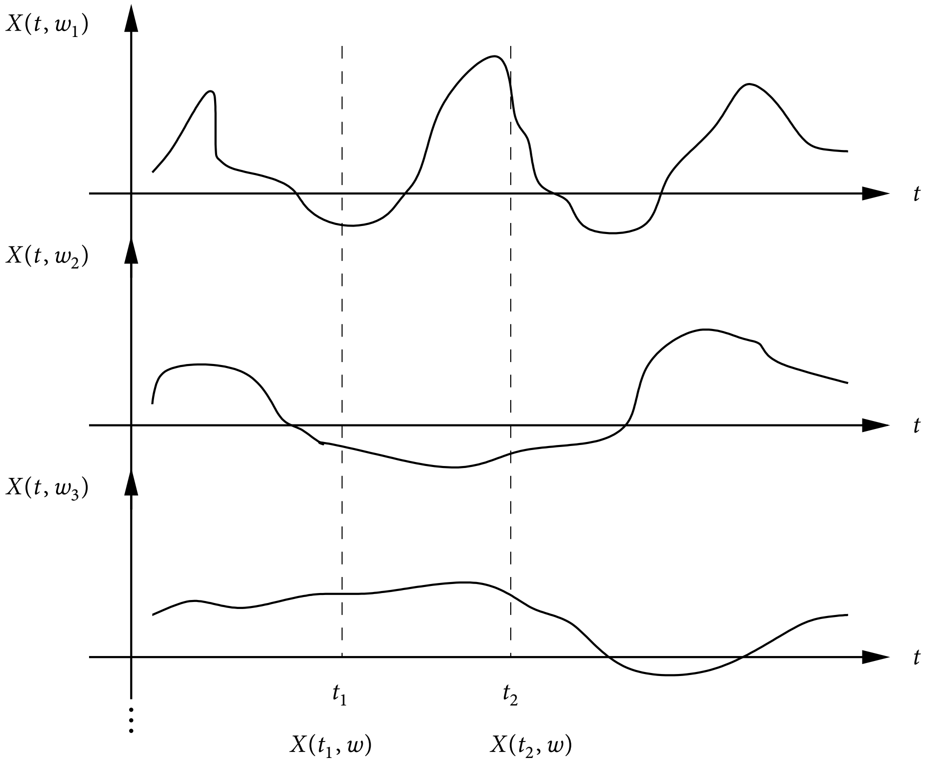


Figure 8.1. Sample functions of a random process

8.2 DISCRETE-TIME RANDOM PROCESSES

A random process is said to be *discrete-time* if $\mathcal{T}$ is a countably infinite set, e.g., $\mathbb{N} = \{0, 1, 2, \ldots\}$ and $\mathbb{Z} = \{\ldots, -2, -1, 0, +1, +2, \ldots\}$. In this case, the process is denoted by X_n, for $n \in \mathcal{N}$, a countably infinite set, and is simply an infinite sequence of random variables. A sample function for a discrete-time process is called a *sample sequence* or *sample path*. A discrete-time process can comprise discrete, continuous, or mixed random variables.

Example 8.1. Let $Z \sim \text{Unif}[0,1]$ and define the discrete time process

$$X_n = Z^n, \quad n = 1, 2, \ldots.$$

The sample paths are illustrated in Figure 8.2. The *first-order pdf* of the process, that is, the sequence of pdfs of X_n is

$$f_{X_n}(x) = \frac{1}{nx^{(n-1)/n}} = \frac{1}{n} x^{\frac{1}{n}-1}, \quad x \in [0,1],$$

which can be easily found by differentiating $\mathsf{P}\{X_n \le x\} = \mathsf{P}\{Z \le x^{1/n}\}$ w.r.t. x.

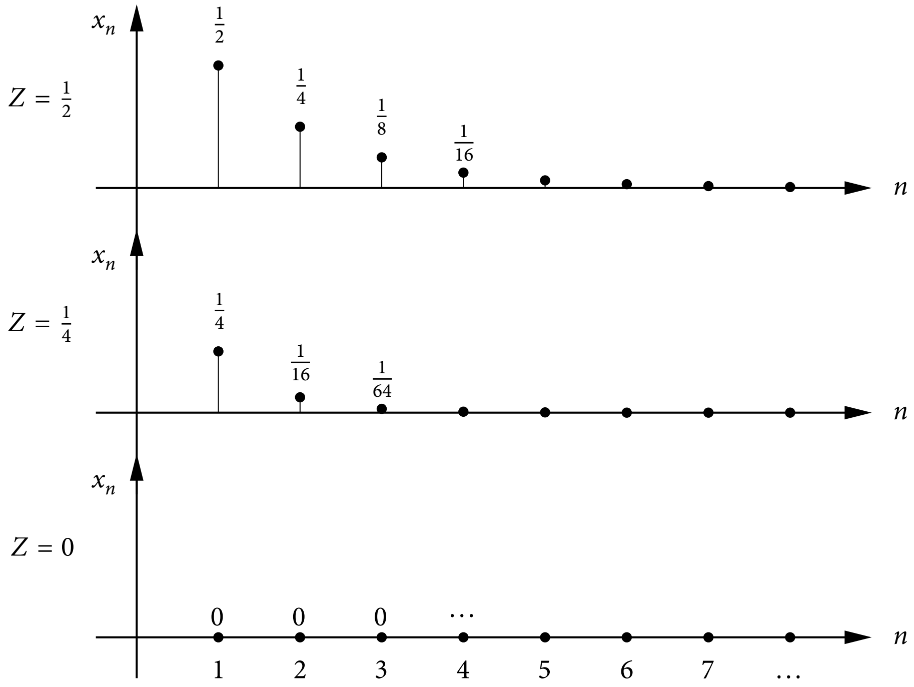


Figure 8.2. Sample paths of the random process in Example 8.1.

In the above example, we specified the random process by describing the set of sample paths and explicitly providing a probability measure over the set of events (subsets of sample paths). This way of specifying a random process has very limited applicability, and is suited only for very simple processes. A discrete-time random process is in general specified (directly or indirectly) by specifying all its k-th order cdfs (pdfs, pmfs), i.e., the joint cdf (pdf, pmf) of the samples

$$X_{n_1}, X_{n_2}, \ldots, X_{n_k}$$

for every order k and for every set of k points $n_1 < n_2 < \cdots < n_k \in \mathcal{N}$. The Kolmogorov extension theorem, the proof of which is beyond this course, guarantees that the entire process can be defined this way.

In the following, we discuss several classes of discrete-time random processes.

8.2.1 IID Processes

We say that $\{X_n : n \in \mathcal{N}\}$ is an IID process if the random variables $X_1, X_2, \ldots$ are independent and identically distributed (i.i.d.).

Example 8.2 (Bernoulli process). $X_1, X_2, \ldots$ i.i.d. Bern(p).

Example 8.3 (Discrete-time white Gaussian noise). $X_1, X_2, \ldots$ i.i.d. N$(0, N)$.

Here we specified the n-th order pmfs (pdfs) of the processes by specifying the first-order pmf (pdf) and stating that the random variables are independent. It would be quite difficult to provide the specifications for an IID process by specifying the probability measure over the subsets of the sample.

8.2.2 Random Walk

Let $Z_1, Z_2, \ldots, Z_n, \ldots$ be i.i.d., where

$$Z_n = \begin{cases} +1 & \text{w.p. } 1/2, \\ -1 & \text{w.p. } 1/2. \end{cases}$$

The *(symmetric) random walk* is defined by

$$\begin{aligned} X_0 &= 0, \\ X_n &= \sum_{i=1}^{n} Z_i = X_{n-1} + Z_n\,, \quad n = 1, 2, \ldots. \end{aligned}$$

Again this process is specified by (indirectly) specifying all n-th order pmfs. The sample path for a random walk is a sequence of integers as illustrated in Figure 8.3. The first-order pmf is $\mathsf{P}\{X_n = k\}$ as a function of n. Note that

$$\begin{aligned} k &\in \{-n, -(n-2), \ldots, -2, 0, +2, \ldots, +(n-2), +n\} \quad \text{for } n \text{ even}, \\ k &\in \{-n, -(n-2), \ldots, -1, +1, +3, \ldots, +(n-2), +n\} \quad \text{for } n \text{ odd}. \end{aligned}$$

Now if we let a be the number of $+1$'s in n steps and $n - a$ be the number of -1's, then

$$k = a - (n - a) = 2a - n,$$

or equivalently, $a = n + k/2$. Thus

$$\begin{aligned} \mathsf{P}\{X_n = k\} &= \mathsf{P}\{(n+k)/2 \text{ heads in } n \text{ independent coin tosses}\} \\ &= \binom{n}{\frac{n+k}{2}} \cdot 2^{-n} \quad \text{if } n - k \text{ is even}. \end{aligned}$$

For example, $\mathsf{P}\{X_5 = 3\} = 5/32$ and $\mathsf{P}\{X_{10} = 6\} = 45/1024$.

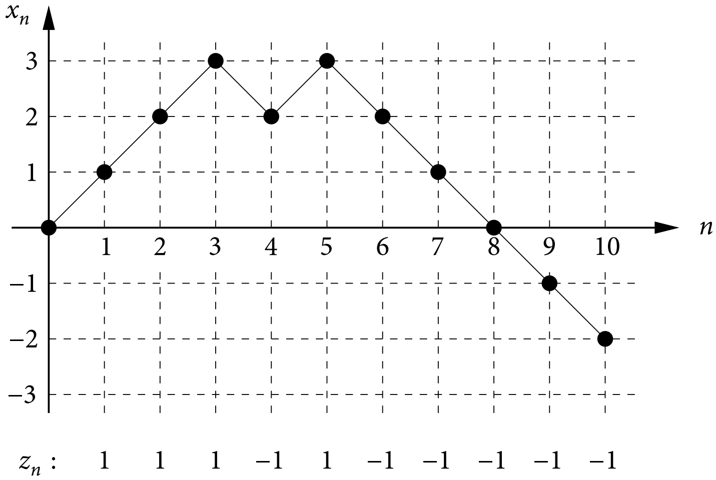


Figure 8.3. A sample path of the random walk.

8.2.3 Markov Processes

A random process $\{X_n\}$ is said to be *Markov* if the "future and the past are conditionally independent given the present." Mathematically, this can be rephrased in several ways. For example, if the random variables $X_1, X_2, \ldots$ are discrete, then the process is Markov if

$$p_{X_{n+1}|X_1,X_2,\ldots,X_n}(x_{n+1}|x_1, x_2, \ldots, x_n) = p_{X_{n+1}|X_n}(x_{n+1}|x_n)$$

for every n and every $(x_1, x_2, \ldots, x_{n+1})$. Discrete-time discrete Markov processes are often called *Markov chains*.

Example 8.4. IID processes are Markov.

Example 8.5. The random walk process is Markov. To see this, consider

$$\begin{aligned}
\mathsf{P}\{X_{n+1} = x_{n+1} \,|\, X_1 = x_1, \ldots, X_n = x_n\} &= \mathsf{P}\{X_n + Z_{n+1} = x_{n+1} \,|\, X_1 = x_1, \ldots, X_n = x_n\} \\
&= \mathsf{P}\{Z_{n+1} = x_{n+1} - x_n \,|\, X_1 = x_1, \ldots, X_n = x_n\} \\
&\overset{(a)}{=} \mathsf{P}\{Z_{n+1} = x_{n+1} - x_n\} \\
&\overset{(b)}{=} \mathsf{P}\{Z_{n+1} = x_{n+1} - x_n \,|\, X_n = x_n\} \\
&= \mathsf{P}\{X_n + Z_{n+1} = x_{n+1} \,|\, X_n = x_n\} \\
&= \mathsf{P}\{X_{n+1} = x_{n+1} \,|\, X_n = x_n\},
\end{aligned}$$

where the equalities in (a) and (b) follow since Z_{n+1} is independent of $(X_1, \ldots, X_n)$.

8.2.4 Independent Increment Processes

A random process $\{X_n\}$ is said to be *independent increment* if the increments

$$X_{n_1},\ X_{n_2} - X_{n_1},\ \ldots,\ X_{n_k} - X_{n_{k-1}}$$

are independent for all k and all $n_1 < n_2 < \cdots < n_k$.

Example 8.6. The random walk is an independent increment process since

$$\begin{aligned} X_{n_1} &= \sum_{i=1}^{n_1} Z_i, \\ X_{n_2} - X_{n_1} &= \sum_{i=n_1+1}^{n_2} Z_i, \\ &\vdots \\ X_{n_k} - X_{n_{k-1}} &= \sum_{i=n_{k-1}+1}^{n_k} Z_i \end{aligned}$$

are independent (as functions of independent random vectors). The independent increment property makes it easy to find the n-th order pmfs of the random walk process. For example,

$$\begin{aligned} \mathsf{P}\{X_5 = 3,\ X_{10} = 6,\ X_{20} = 10\} &= \mathsf{P}\{X_5 = 3,\ X_{10} - X_5 = 3,\ X_{20} - X_{10} = 4\} \\ &= \mathsf{P}\{X_5 = 3\}\,\mathsf{P}\{X_{10} - X_5 = 3\}\,\mathsf{P}\{X_{20} - X_{10} = 4\} \\ &= \mathsf{P}\{X_5 = 3\}\,\mathsf{P}\{X_5 = 3\}\,\mathsf{P}\{X_{10} = 4\} \\ &= \binom{5}{4}2^{-5}\binom{5}{4}2^{-5}\binom{10}{7}2^{-10} \\ &= 3000 \cdot 2^{-20}. \end{aligned}$$

In general if a process is independent increment, then it is also Markov. To see this, let $\{X_n\}$ be an independent increment process. Then

$$\begin{aligned} &\mathsf{P}\{X_{n+1} = x_{n+1} \mid X_1 = x_1, \ldots, X_n = x_n\} \\ &= \mathsf{P}\{X_{n+1} - X_n = x_{n+1} - x_n \mid X_1 = x_1, X_2 - X_1 = x_2 - x_1, \ldots, X_n - X_{n-1} = x_n - x_{n-1}\} \\ &\overset{(a)}{=} \mathsf{P}\{X_{n+1} - X_n = x_{n+1} - x_n\} \\ &\overset{(b)}{=} \mathsf{P}\{X_{n+1} - X_n = x_{n+1} - x_n \mid X_n = x_n\} \\ &= \mathsf{P}\{X_{n+1} = x_{n+1} \mid X_n = x_n\}, \end{aligned}$$

where the equalities in (a) and (b) follows by the independent increment property of the process. The converse is not necessarily true, e.g., IID processes are Markov but not independent increment.

8.2.5 Gauss–Markov Process

Let $Z_1, Z_2, \ldots$ be i.i.d. $\sim \mathrm{N}(0, N)$, i.e., $\{Z_n\}$ be a white Gaussian noise (WGN) process. The Gauss–Markov process is a *first-order autoregressive process* defined by

$$\begin{aligned} X_1 &= Z_1, \\ X_n &= \alpha X_{n-1} + Z_n, \quad n = 2, 3, \ldots, \end{aligned}$$

where α is a parameter such that $|\alpha| < 1$. This process can be generated by passing a WGN process through a discrete-time linear time invariant system, as we will see in Lecture #9 in more detail. The Gauss–Markov process is Markov. It is not, however, independent increment.

8.3 CONTINUOUS-TIME RANDOM PROCESSES

A random process is *continuous time* if $\mathcal{T}$ is a continuous set, e.g., $\mathbb{R} = (-\infty, \infty)$ or $\mathbb{R}^+ = [0, \infty)$.

Example 8.7 (Sinusoidal signal with random phase). Let

$$X(t) = \alpha \cos(\omega t + \Theta), \quad t \geq 0,$$

where $\Theta \sim \mathrm{Unif}[0, 2\pi]$ and α and ω are constants. The sample functions are illustrated in Figure 8.4. The first-order pdf of the process is the pdf of $X(t) = \alpha \cos(\omega t + \Theta)$. In Problem 3.7, we found it to be

$$f_{X(t)}(x) = \frac{1}{\alpha\pi\sqrt{1 - (x/\alpha)^2}}, \qquad -\alpha < x < +\alpha.$$

Note that the pdf is independent of t.

A continuous-time random process $\{X(t)\colon t \geq 0\}$ is said to be *independent increment* if $X(t+s) - X(t)$ is independent of $\{X(u)\colon 0 \leq u \leq t\}$ for every $s, t \geq 0$. In particular, $X(t_1),\ X(t_2) - X(t_1), \ldots, X(t_k) - X(t_{k-1})$ are independent for every k and $t_1 < t_2 < \cdots < t_k$. A continuous-time random process $\{X(t)\colon t \geq 0\}$ is said to be *Markov* if the future $\{X(t+s)\colon s > 0\}$ is conditionally independent of the past $\{X(u)\colon 0 \leq u \leq t\}$ given the present $X(t)$. In particular, $X(t_{k+1})$ is conditionally independent of $(X(t_1), \ldots, X(t_{k-1}))$ given $X(t_k)$ for every k and $t_1 < t_2 < \cdots < t_{k+1}$. As in the discrete-time case, an independent increment process is Markov, but the converse does not hold.

Unlike discrete-time random processes, a countinuous-time random process cannot be fully specified by k-th order distributions alone. For example, knowing k-th order distributions for every k does not answer whether the process is continuous or not. Thus, for continuous-time random processes, we often need additional properties (such as continuity).

In the following, we discuss a few famous continuous-time random processes.

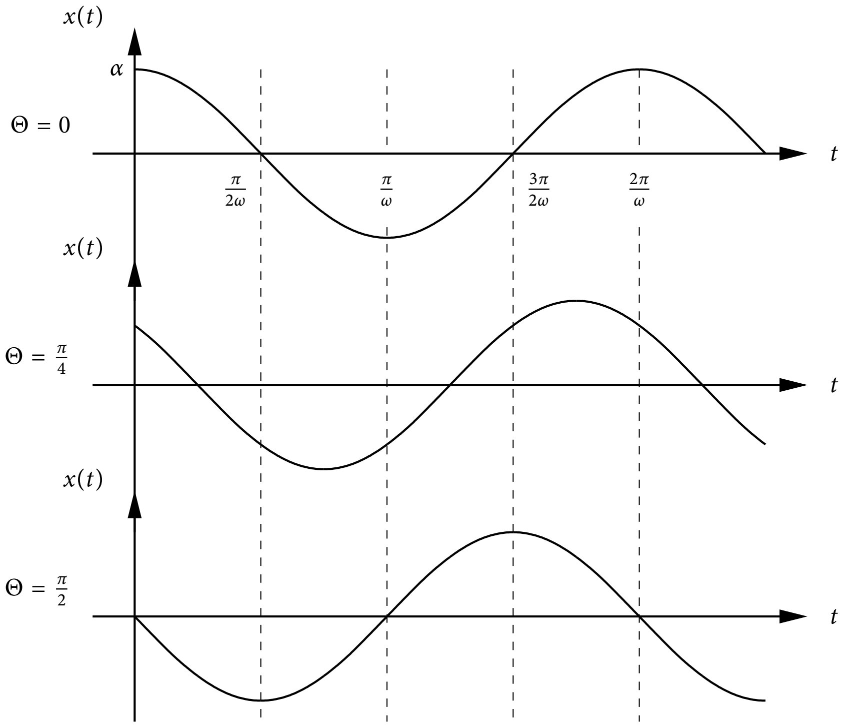


Figure 8.4. Sample functions of the sinusoidal signal with random phase.

8.3.1 Brownian Motion

A random process $\{W(t)\colon\ t \geq 0\}$ is said to be a *Brownian motion* (or *Wiener process*) if

- $W(0) = 0$,
- $\{W(t)\}$ is independent increment with $W(t) - W(s) \sim \mathrm{N}(0, t - s)$ for every $t > s$, and
- $W(t)$ is continuous for $t \geq 0$ almost surely.

The k-th order pdf can be easily computed from the independent increment property. For example,

$$\begin{aligned} f_{W(t_1),W(t_2),W(t_3)}(w_1, w_2, w_3) &= f_{W(t_1)}(w_1) f_{W(t_2)|W(t_1)}(w_2|w_1) f_{W(t_3)|W(t_2)}(w_3|w_2) \\ &= f_{W(t_1)}(w_1) f_{W(t_2-t_1)}(w_2 - w_1) f_{W(t_3-t_2)}(w_3 - w_2). \end{aligned}$$

8.3.2 Poisson Process

A random process $\{N(t)\colon\ t \geq 0\}$ is said to be *Poisson* with rate λ if

- $N(0) = 0$ and

- $\{N(t)\}$ is independent increment with $N(t) - N(s) \sim \text{Poisson}(\lambda(t-s))$ for every $t > s$.

A sample path is shown in Figure 8.5. Here $t_1, t_2, \ldots$ are the *arrival times* or the *wait times* of the events. The differences $t_1, t_2 - t_1, \ldots$ are called the *interarrival times* of the events.

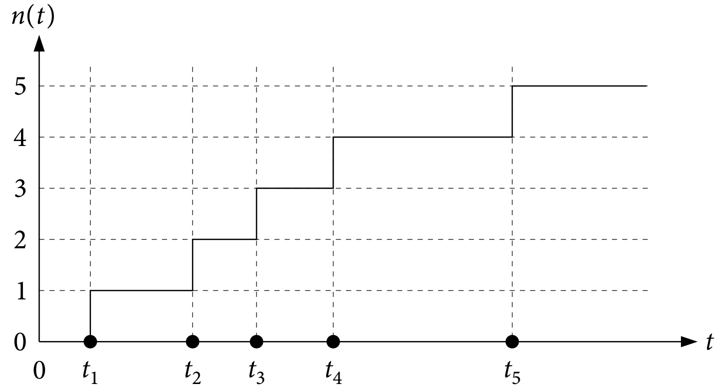


Figure 8.5. A sample path of a Poisson process.

Recalling from Example 3.6, the first arrival time T_1 can be specified by the relationship

$$\{T_1 > t\} = \{N(t) = 0\},$$

which implies that T_1 is an $\text{Exp}(\lambda)$ random variable. More generally, the interarrival times $T_1, T_2 - T_1, T_3 - T_2, \ldots$ are i.i.d. $\text{Exp}(\lambda)$.

8.4 MEAN AND AUTOCORRELATION FUNCTIONS

For a random process $X(t)$ the first and second order moments are

- *mean* function: $\mu_X(t) = \mathsf{E}[X(t)]$ for $t \in \mathcal{T}$.
- *autocorrelation* function: $R_X(t_1, t_2) = \mathsf{E}[X(t_1)X(t_2)]$ for $t_1, t_2 \in \mathcal{T}$.

The *autocovariance function* of a random process is defined as

$$\begin{aligned} C_X(t_1, t_2) &= \mathsf{E}[(X(t_1) - \mathsf{E}[X(t_1)])(X(t_2) - \mathsf{E}[X(t_2)]] \\ &= R_X(t_1, t_2) - \mu_X(t_1)\mu_X(t_2). \end{aligned}$$

Example 8.8. For an IID process X_n

$$\begin{aligned} \mu_X(n) &= \mathsf{E}[X_1], \\ R_X(n_1, n_2) &= \mathsf{E}[X_{n_1}X_{n_2}] = \begin{cases} \mathsf{E}[X_1^2] & n_1 = n_2, \\ (\mathsf{E}[X_1])^2 & n_1 \neq n_2. \end{cases} \end{aligned}$$

Example 8.9. For the random phase signal process in Example 8.7,

$$
\begin{aligned}
\mu_X(t) &= \mathsf{E}[\alpha\cos(\omega t+\Theta)] = \int_0^{2\pi} \frac{\alpha}{2\pi}\cos(\omega t+\theta)\,d\theta = 0,\\
R_X(t_1,t_2) &= \mathsf{E}[X(t_1)X(t_2)]\\
&= \int_0^{2\pi} \frac{\alpha^2}{2\pi}\cos(\omega t_1+\theta)\cos(\omega t_2+\theta)\,d\theta\\
&= \int_0^{2\pi} \frac{\alpha^2}{4\pi}\big[\cos(\omega(t_1+t_2)+2\theta)+\cos(\omega(t_1-t_2))\big]\,d\theta\\
&= \frac{\alpha^2}{2}\cos(\omega(t_1-t_2))
\end{aligned}
$$

Example 8.10. For the random walk,

$$
\mu_X(n) = \mathsf{E}\Big[\sum_{i=1}^{n} Z_i\Big] = \sum_{i=1}^{n} 0 = 0.
$$

To compute the autocorrelation function, first assume that $n_1 \le n_2$ and consider

$$
\begin{aligned}
R_X(n_1,n_2) &= \mathsf{E}[X_{n_1}X_{n_2}]\\
&= \mathsf{E}[X_{n_1}(X_{n_2}-X_{n_1}+X_{n_1})]\\
&= \mathsf{E}[X_{n_1}^2] = n_1.
\end{aligned}
$$

In general,

$$
R_X(n_1,n_2) = \min\{n_1,n_2\}.
$$

Example 8.11. For the Gauss–Markov process,

$$
\begin{aligned}
\mu_X(n) &= \mathsf{E}[X_n] = \mathsf{E}[\alpha X_{n-1}+Z_n]\\
&= \alpha\,\mathsf{E}[X_{n-1}] + \mathsf{E}[Z_n]\\
&= \alpha\,\mathsf{E}[X_{n-1}]\\
&= \alpha^{n-1}\,\mathsf{E}[Z_1] = 0.
\end{aligned}
$$

To find the autocorrelation function, assume first that $n_1 < n_2$. Then,

$$
X_{n_2} = \alpha^{n_2-n_1}X_{n_1} + \sum_{i=0}^{n_2-n_1-1} \alpha^i Z_{n_2-i}.
$$

Thus,

$$
R_X(n_1,n_2) = \mathsf{E}[X_{n_1}X_{n_2}] = \alpha^{n_2-n_1}\,\mathsf{E}[X_{n_1}^2] + 0\,,
$$

since X_{n_1} and Z_{n_2-i} are independent, zero mean for $0 \le i \le n_2-n_1-1$. Next, to find

$\mathsf{E}[X_{n_1}^2]$, consider

$$\begin{aligned}\mathsf{E}[X_1^2] &= N,\\ \mathsf{E}[X_{n_1}^2] &= \mathsf{E}[(\alpha X_{n_1-1} + Z_{n_1})^2]\\ &= \alpha^2\, \mathsf{E}[X_{n_1-1}^2] + N\\ &= \frac{1-\alpha^{2n_1}}{1-\alpha^2} N.\end{aligned}$$

Therefore, in general,

$$R_X(n_1, n_2) = \alpha^{|n_2-n_1|}\, \frac{1-\alpha^{2\min\{n_1,n_2\}}}{1-\alpha^2} N.$$

8.5 GAUSSIAN RANDOM PROCESSES

A random process is said to be *Gaussian* if

$$[X(t_1),\ X(t_2),\ \ldots,\ X(t_k)\,]^T$$

is a Gaussian random vector for every k and $t_1 < t_2 < \cdots < t_k$.

Example 8.12. The discrete time WGN process is Gaussian.

Example 8.13. The Gauss–Markov process is Gaussian. Indeed, since $X_1 = Z_1$ and $X_k = \alpha X_{k-1} + Z_k$ with $Z_1, Z_2, \ldots$ i.i.d. $\mathrm{N}(0, N)$, we have

$$\begin{bmatrix} X_1\\ X_2\\ X_3\\ \vdots\\ X_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0\\ \alpha & 1 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \alpha^{n-2} & \alpha^{n-3} & \cdots & 1 & 0\\ \alpha^{n-1} & \alpha^{n-2} & \cdots & \alpha & 1 \end{bmatrix} \begin{bmatrix} Z_1\\ Z_2\\ Z_3\\ \vdots\\ Z_n \end{bmatrix},$$

which is a linear transformation of a Gaussian random vector and is therefore Gaussian itself.

Example 8.14. The Brownian motion is Gaussian (why?).

Since the joint pdf for a Gaussian random vector is specified by its mean and covariance matrix, a discrete-time Gaussian random process is specified by its mean $\mu_X(t)$ and autocorrelation $R_X(t_1, t_2)$ functions. For the continuous-time case, mean and autocorrelation functions determine every finite-order distributions, although additional properties (such as continuity) are needed to fully specify the Gaussian random process.

PROBLEMS

8.1. *Symmetric random walk.* Let X_n be a random walk defined by

$$X_0 = 0,$$
$$X_n = \sum_{i=1}^{n} Z_i,$$

where $Z_1, Z_2, \ldots$ are i.i.d. with $\mathsf{P}\{Z_1 = -1\} = \mathsf{P}\{Z_1 = 1\} = \frac{1}{2}$.

(a) Find $\mathsf{P}\{X_{10} = 10\}$.

(b) Approximate $\mathsf{P}\{-10 \le X_{100} \le 10\}$ using the central limit theorem.

(c) Find $\mathsf{P}\{X_n = k\}$.

8.2. *Absolute-value random walk.* Consider the symmetric random walk X_n in the previous problem. Define the absolute value random process $Y_n = |X_n|$.

(a) Find $\mathsf{P}\{Y_n = k\}$.

(b) Find $\mathsf{P}\{\max_{1\le i<20} Y_i = 10 \,|\, Y_{20} = 0\}$.

8.3. *Sampled random walk.* Let $\{X_n\}$ be the (standard) symmetric random walk, i.e.,

$$X_0 = 0,$$
$$X_n = \sum_{i=1}^{n} Z_i, \quad n = 1, 2, \ldots,$$

where $Z_1, Z_2, \ldots$ are i.i.d. with $\mathsf{P}\{Z_1 = -1\} = \mathsf{P}\{Z_1 = 1\} = 1/2$. Let $\{Y_n\}$ be a sampled version of $\{X_n\}$ defined by

$$Y_n = X_{2n}, \quad n = 0, 1, 2, \ldots.$$

(a) Is $\{Y_n\}$ independent increment? Justify your answer.

(b) Is $\{Y_n\}$ Markov? Justify your answer.

(c) Find $\mathsf{E}[Y_3 \,|\, Y_2]$.

8.4. *Discrete-time Wiener process.* Let Z_n, $n \ge 0$, be a discrete time white Gaussian noise process, i.e., $Z_1, Z_2, \ldots$ are i.i.d $N(0,1)$. Define the process X_n, $n \ge 1$, such that $X_0 = 0$, and $X_n = X_{n-1} + Z_n$, for $n \ge 1$.

(a) Is X_n an independent increment process? Justify your answer.

(b) Is X_n a Gaussian process? Justify your answer.

(c) Find the mean and autocorrelation functions of X_n.

(d) Specify the first-order pdf of X_n.

(e) Specify the joint pdf of X_3, X_5, and X_8.

(f) Find $\mathsf{E}(X_{20}|X_1, X_2, \ldots, X_{10})$.

(g) Given $X_1 = 4$, $X_2 = 2$, and $0 \le X_3 \le 4$, find the minimum MSE estimate of X_4.

8.5. *Wiener process.* Recall the following definition of the (standard) Wiener process:

- $W(0) = 0$,
- $\{W(t)\}$ is independent increment with $W(t) - W(s) \sim \mathrm{N}(0, t - s)$ for all $t > s$,
- $\mathsf{P}\{\omega : W(\omega, t) \text{ is continuous in } t\} = 1$.

Let $W_1(t)$ and $W_2(t)$ be independent Wiener processes.

(a) Find the mean and the variance of

$$X(t) = \frac{1}{\sqrt{2}}(W_1(t) + W_2(t)).$$

Is $\{X(t)\}$ a Wiener process? Justify your answer.

(b) Find the mean and the variance of

$$Y(t) = \frac{1}{\sqrt{2}}(W_1(t) - W_2(t)).$$

Is $\{Y(t)\}$ a Wiener process? Justify your answer.

(c) Find $\mathsf{E}[X(t)Y(s)]$.

8.6. *Brownian bridge.* Let $\{W(t)\}_{t=0}^{\infty}$ be the standard Brownian motion (Wiener process). Recall that the process is independent-increment with $W(0) = 0$ and

$$W(t) - W(s) \sim \mathrm{N}(0, t - s), \quad 0 \le s < t.$$

In the following, we investigate several properties of the process conditioned on $\{W(1) = 0\}$.

(a) Find the conditional distribution of $W(1/2)$ given $W(1) = 0$.

(b) Find $\mathsf{E}[W(t)\,|\,W(1) = 0]$ for $t \in [0, 1]$.

(c) Find $\mathsf{E}[(W(t))^2\,|\,W(1) = 0]$ for $t \in [0, 1]$.

(d) Find $\mathsf{E}[W(t_1)W(t_2)\,|\,W(1) = 0]$ for $t_1, t_2 \in [0, 1]$.

8.7. *A random process.* Let $X_n = Z_{n-1} + Z_n$ for $n \ge 1$, where $Z_0, Z_1, Z_2, \ldots$ are i.i.d. $\sim \mathrm{N}(0, 1)$.

(a) Find the mean and autocorrelation functions of $\{X_n\}$.

(b) Is $\{X_n\}$ Gaussian? Justify your answer.

(c) Find $\mathsf{E}(X_3|X_1, X_2)$.

(d) Find $\mathsf{E}(X_3|X_2)$.

(e) Is $\{X_n\}$ Markov? Justify your answer.

(f) Is $\{X_n\}$ independent increment? Justify your answer.

8.8. *Moving average process.* Let $X_n = \frac{1}{2}Z_{n-1} + Z_n$ for $n \geq 1$, where $Z_0, Z_1, Z_2, \ldots$ are i.i.d. $\sim$ N$(0, 1)$. Find the mean and autocorrelation function of X_n.

8.9. *Autoregressive process.* Let $X_0 = 0$ and $X_n = \frac{1}{2}X_{n-1} + Z_n$ for $n \geq 1$, where $Z_1, Z_2, \ldots$ are i.i.d. $\sim$ N$(0, 1)$. Find the mean and autocorrelation function of X_n.

8.10. *Random binary waveform.* In a digital communication channel the symbol "1" is represented by the fixed duration rectangular pulse

$$g(t) = \begin{cases} 1 & \text{for } 0 \leq t < 1 \\ 0 & \text{otherwise,} \end{cases}$$

and the symbol "0" is represented by $-g(t)$. The data transmitted over the channel is represented by the random process

$$X(t) = \sum_{k=0}^{\infty} A_k g(t-k), \quad \text{for } t \geq 0,$$

where $A_0, A_1, \ldots$ are i.i.d random variables with

$$A_i = \begin{cases} +1 & \text{w.p. } \frac{1}{2} \\ -1 & \text{w.p. } \frac{1}{2}. \end{cases}$$

(a) Find its first and second order pmfs.

(b) Find the mean and the autocorrelation function of the process $X(t)$.

8.11. *Arrow of time.* Let X_0 be a Gaussian random variable with zero mean and unit variance, and $X_n = \alpha X_{n-1} + Z_n$ for $n \geq 1$, where α is a fixed constant with $|\alpha| < 1$ and $Z_1, Z_2, \ldots$ are i.i.d. $\sim N(0, 1-\alpha^2)$, independent of X_0.

(a) Is the process $\{X_n\}$ Gaussian?

(b) Is $\{X_n\}$ Markov?

(c) Find $R_X(n, m)$.

(d) Find the (nonlinear) MMSE estimate of X_{100} given $(X_1, X_2, \ldots, X_{99})$.

(e) Find the MMSE estimate of X_{100} given $(X_{101}, X_{102}, \ldots, X_{199})$.

(f) Find the MMSE estimate of X_{100} given $(X_1, \ldots, X_{99}, X_{101}, \ldots, X_{199})$.

8.12. *Convergence of random processes.* Let $\{N(t)\}_{t=0}^{\infty}$ be a Poisson process with rate λ. Recall that the process is independent increment and $N(t) - N(s)$, $0 \leq s < t$, has the pmf

$$p_{N(t)-N(s)}(n) = \frac{e^{-\lambda(t-s)}(\lambda(t-s))^n}{n!}, \quad n = 0, 1, \ldots.$$

Define

$$M(t) = \frac{N(t)}{t}, \quad t > 0.$$

(a) Find the mean and autocorrelation function of $\{M(t)\}_{t>0}$.

(b) Does $\{M(t)\}_{t>0}$ converge in mean square as $t \to \infty$, that is,

$$\lim_{t\to\infty} \mathsf{E}\big[(M(t) - M)^2\big] = 0$$

for some random variable (or constant) M? If so, what is the limit?

Now consider

$$L(t) = \frac{1}{t}\int_0^t \frac{N(s)}{s}\, ds, \qquad t > 0.$$

(c) Does $\{L(t)\}_{t>0}$ converge in mean square as $t \to \infty$? If so, what is the limit? (Hint: $\int 1/x\, dx = \ln x + C$, $\int \ln x\, dx = x\ln x - x + C$, and $\lim_{x\to 0} x\ln x = 0$.)

LECTURE 9

Stationary Processes

9.1 STRICT-SENSE STATIONARITY

Stationarity refers to *time invariance* of some, or all, of the statistics of a random process, such as mean, autocorrelation, and n-th order distribution. We define two types of stationarity: *strict-sense stationarity* (SSS) and *wide-sense stationarity* (WSS).

A random process $\{X(t)\}$ is said to be SSS (or just *stationary*) if *all* its finite-order distributions are time invariant, i.e., the joint cdfs (pdfs, pmfs) of

$$X(t_1), X(t_2), \ldots, X(t_k)$$

and

$$X(t_1 + \tau), X(t_2 + \tau), \ldots, X(t_k + \tau)$$

are the same for every k, every $t_1, t_2, \ldots, t_k$, and every time shift τ. So for a SSS process, the first-order distribution is independent of t, and the second-order distribution — the distribution of any two samples $X(t_1)$ and $X(t_2)$ — depends only on $\tau = t_2 - t_1$. To see this, note that from the definition of stationarity, for any t, the joint distribution of $X(t_1)$ and $X(t_2)$ is the same as the joint distribution of $X(t) = X(t_1 + (t - t_1))$ and $X(t_2 + (t - t_1)) = X(t + (t_2 - t_1))$.

Example 9.1. IID processes are SSS.

Example 9.2. The random walk is not SSS. In fact, no independent increment process is SSS.

Example 9.3. The Gauss–Markov process defined in Section 8.2.5 is not SSS. However, if we set X_1 to the steady-state distribution of X_n, it becomes SSS (see Problem 9.6).

9.2 WIDE-SENSE STATIONARIY

A random process $\{X(t)\}$ is said to be *wide-sense stationary* (WSS) if its mean and autocorrelation functions are time invariant, i.e.,

- $\mathsf{E}[X(t)] = \mu$ is independent of t and
- $R_X(t_1, t_2)$ is a function only of the time difference $t_2 - t_1$.

As a technical condition, we also assume that

$$\mathsf{E}[X(t)^2] < \infty.$$

Since $R_X(t_1, t_2) = R_X(t_2, t_1)$, for any wide-sense stationary process $\{X(t)\}$, the autocorrelation function $R_X(t_1, t_2)$ is a function only of $|t_2 - t_1|$. Clearly SSS implies WSS. The converse is not necessarily true.

Example 9.4. Let

$$X(t) = \begin{cases} +\sin t & \text{w.p. } 1/4, \\ -\sin t & \text{w.p. } 1/4, \\ +\cos t & \text{w.p. } 1/4, \\ -\cos t & \text{w.p. } 1/4. \end{cases}$$

Note that $\mathsf{E}[X(t)] = 0$ and $R_X(t_1, t_2) = (1/2)\cos(t_2 - t_1)$; thus $X(t)$ is WSS. But $X(0)$ and $X(\pi/4)$ have different pmfs (i.e., the first-order pmf is not time-invariant) and the process is not SSS.

For a Gaussian random process, WSS implies SSS, since every finite-order distribution of the process is completely specified by its mean and autocorrelation functions. The random walk is not WSS, since $R_X(n_1, n_2) = \min\{n_1, n_2\}$ is not time-invariant. In fact, no independent increment process can be WSS.

9.3 AUTOCORRELATION FUNCTION OF WSS PROCESSES

Let $\{X(t)\}$ be a WSS process. We relabel $R_X(t_1, t_2)$ as $R_X(\tau)$, where $\tau = t_1 - t_2$. The autocorrelation function $R_X(\tau)$ satisfies the following properties.

1. $R_X(\tau)$ is even, i.e., $R_X(\tau) = R_X(-\tau)$ for every τ.
2. $|R_X(\tau)| \le R_X(0) = \mathsf{E}[X^2(t)]$, the "average power" of $X(t)$.
3. If $R_X(T) = R_X(0)$ for some $T \ne 0$, then $R_X(\tau)$ is periodic with period T and so is $X(t)$ (with probability 1).

Example 9.5. Let $X(t) = \alpha\cos(\omega t + \Theta)$ be periodic with random phase. Then

$$R_X(\tau) = \frac{\alpha^2}{2}\cos\omega\tau,$$

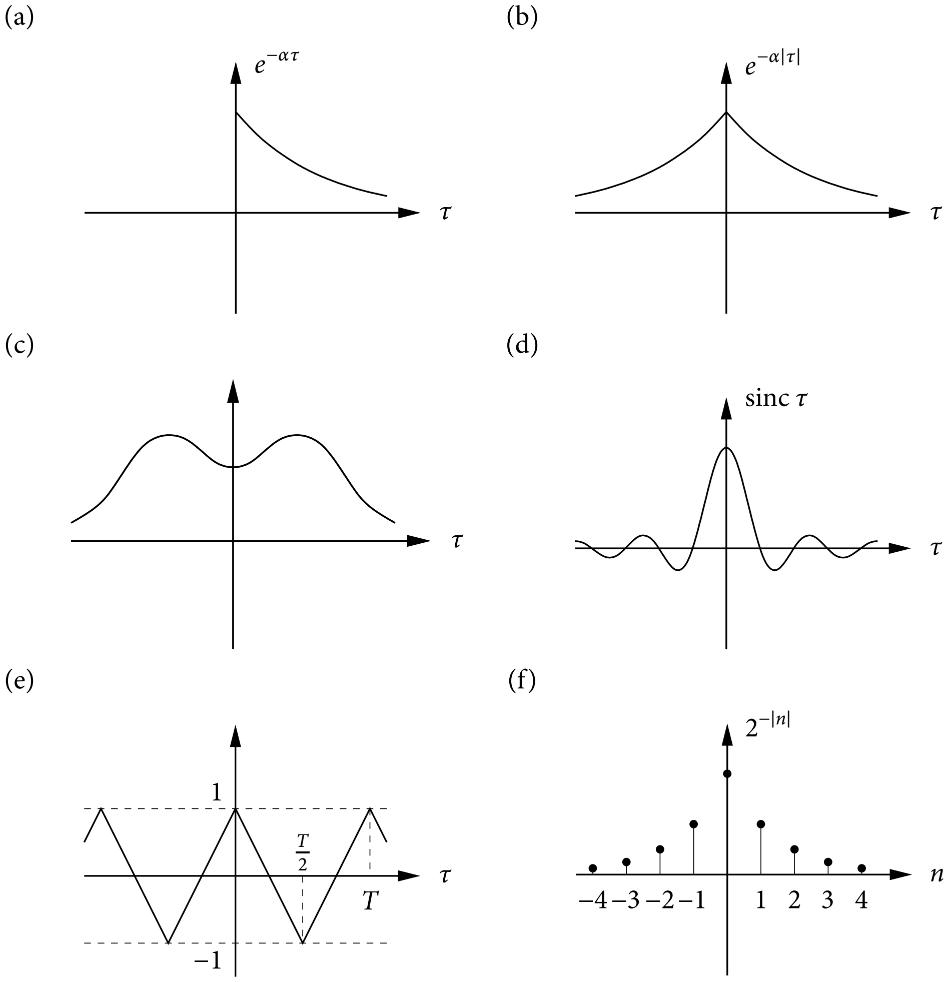
which is also periodic.

The above properties of $R_X(\tau)$ are necessary but not sufficient for a function to qualify as an autocorrelation function for a WSS process. The *necessary and sufficient* condition for a function $R(\tau)$ to be an autocorrelation function for a WSS process is that it be even and nonnegative definite, that is, for any n, any $t_1, t_2, \dots, t_n$ and any real vector $\mathbf{a} = (a_1, \dots, a_n)$,

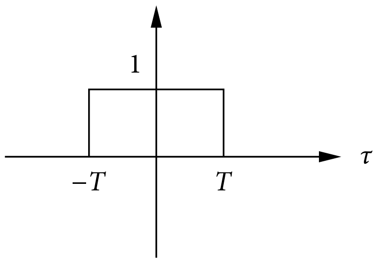
$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j R(t_i - t_j) \geq 0.$$

To see why this is necessary, recall that the correlation matrix for a random vector must be nonnegative definite, so if we take a set of n samples from the WSS random process, their correlation matrix must be nonnegative definite. The condition is sufficient since such an $R(\tau)$ can specify a zero-mean stationary Gaussian random process. The nonnegative definite condition may be difficult to verify directly. It turns out, however, to be equivalent to the condition that the Fourier transform of $R_X(\tau)$, which is called the *power spectral density* $S_X(f)$, is nonnegative for all frequencies f.

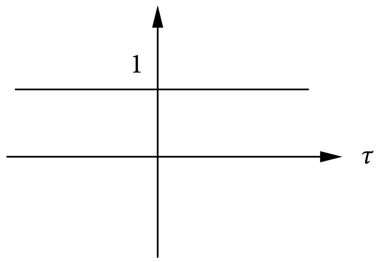
Example 9.6. Consider the following functions.



(g)

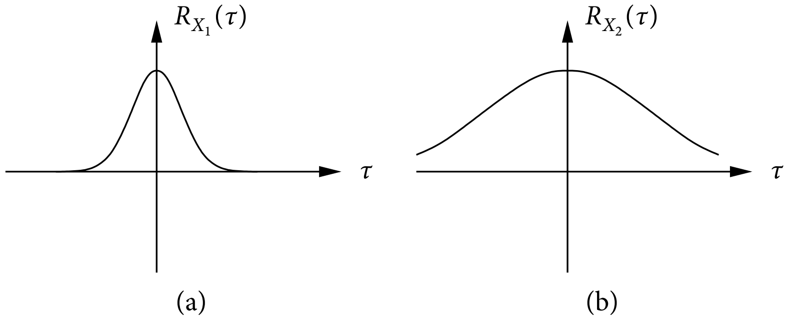


(h)



Here, the functions in (a), (c), and (g) are not autocorrelation functions, and the other functions are autocorrelation functions of some WSS processes.

If $R_X(\tau)$ drops quickly with τ, this means that samples become uncorrelated quickly as we increase τ. Conversely, if $R_X(\tau)$ drops slowly with τ, samples are highly correlated. Figure 9.1 illustrates autocorrelation functions in these two cases. Hence, $R_X(\tau)$ is a measure of the rate of change of $X(t)$ with time t. It turns out that this is not just an intuitive interpretation—as will be proved in Section 9.5, the Fourier transform of $R_X(\tau)$ (the power spectral density) is in fact the average power density of $X(t)$) over frequency.



(a) (b)

Figure 9.1. Autocorrelation functions when (a) correlation is low and (b) correlation is high.

9.4 POWER SPECTRAL DENSITY

The *power spectral density* (psd) of a WSS random process $\{X(t)\}$ is the Fourier transform of $R_X(\tau)$:

$$S_X(f) = \mathcal{F}[R_X(\tau)] = \int_{-\infty}^{\infty} R_X(\tau) e^{-i2\pi\tau f} d\tau.$$

For a discrete-time process $\{X_n\}$, the power spectral density is the discrete-time Fourier transform of the sequence $R_X(n)$:

$$S_X(f) = \sum_{n=-\infty}^{\infty} R_X(n) e^{-i2\pi n f}, \quad |f| < \tfrac{1}{2}.$$

By taking the inverse Fourier transform, $R_X(\tau)$ (or $R_X(n)$) can be recovered from $S_X(f)$ as

$$R_X(\tau) = \int_{-\infty}^{\infty} S_X(f) e^{i2\pi\tau f}\, df,$$

$$R_X(n) = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_X(f) e^{i2\pi n f}\, df.$$

The power spectral density $S_X(f)$ satisfies the following properties.

1. $S_X(f)$ is real and even.
2. $\int_{-\infty}^{\infty} S_X(f) df = R_X(0) = \mathsf{E}[X^2(t)]$, that is, the area under $S_X(f)$ is the average power.
3. $S_X(f)$ is the average power density, i.e., the average power of $X(t)$ in the frequency band $[f_1, f_2]$ is

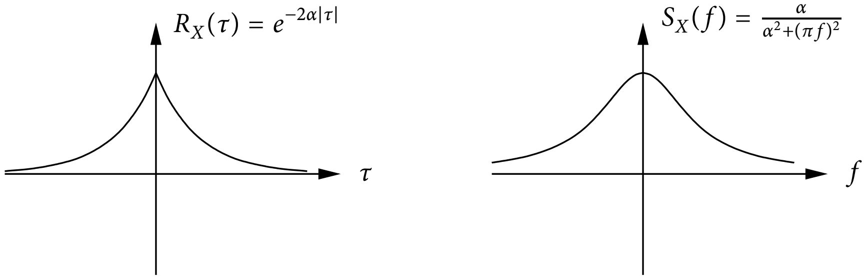
$$\int_{-f_2}^{-f_1} S_X(f)\, df + \int_{f_1}^{f_2} S_X(f)\, df = 2\int_{f_1}^{f_2} S_X(f)\, df.$$

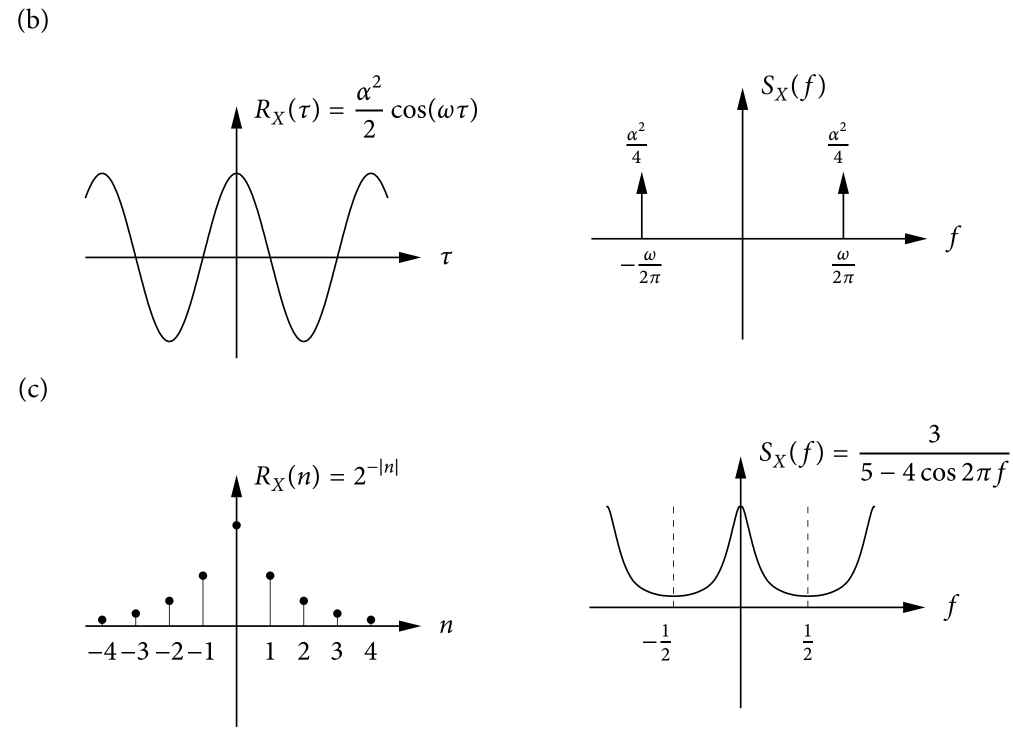
4. $S_X(f) \geq 0$.

The first two properties follow from the definition of power spectral density as the Fourier transform of a real and even function $R_X(\tau)$. The third and fourth properties will be proved later in Section 9.5. In general, a function $S(f)$ is a psd if and only if it is real, even, nonnegative, and $\int_{-\infty}^{\infty} S(f)\, df < \infty$.

Example 9.7. We consider a few pairs of autocorrelation functions and power spectral densities.

(a)





Example 9.8 (Discrete-time white noise process). Let $X_1, X_2, \ldots$ be zero mean, uncorrelated, with common variance N, for example, i.i.d. $\mathrm{N}(0, N)$. The autocorrelation function and the power spectral density are plotted in Figure 9.2.

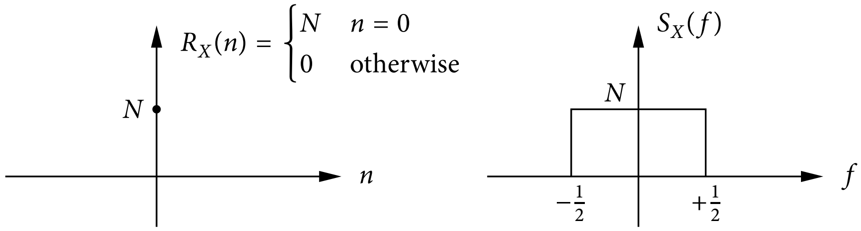


Figure 9.2. Autocorrelation function and power spectral density of the discrete-time white noise process.

Example 9.9 (Band-limited white noise process). Let $\{X(t)\}$ be a WSS zero-mean process with autocorrelation function and power spectral density plotted in Figure 9.3. Note that for any t, the samples $X\left(t \pm \frac{n}{2B}\right)$ for $n = 0, 1, 2, \ldots$ are uncorrelated.

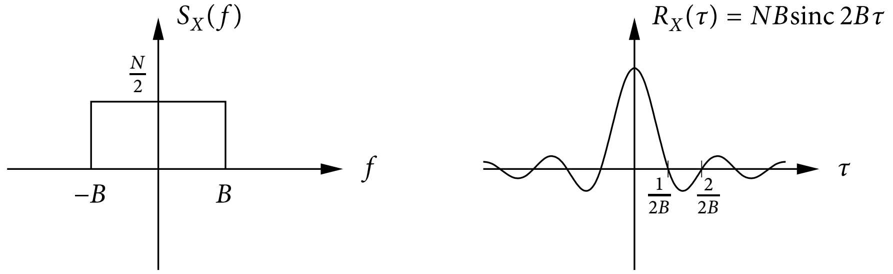


Figure 9.3. Autocorrelation function and power spectral density of the band-limited white noise process.

Example 9.10 (White noise process). If we let $B \to \infty$ in the previous example, we obtain a *white noise process*, which has

$$
\begin{aligned}
S_X(f) &= \frac{N}{2} \quad \text{for all } f, \\
R_X(\tau) &= \frac{N}{2}\delta(\tau).
\end{aligned}
$$

If, in addition, $\{X(t)\}$ is Gaussian, then we obtain the famous white Gaussian noise (WGN) process. A sample path of the white noise process is depicted in Figure 9.4. For a white noise process, all samples are uncorrelated. The process is not physically realizable, since it has infinite power. However, it plays a similar role in random processes to the role of a point mass in physics and delta function in EE. Thermal noise and shot noise are well modeled as white Gaussian noise, since they have very flat psd over very wide band (GHz).

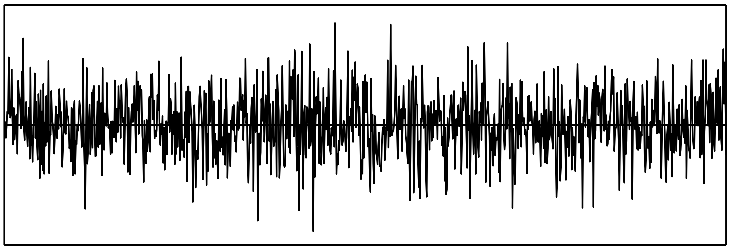


Figure 9.4. A sample function of the white noise process.

9.5 WSS PROCESSES AND LTI SYSTEMS

Consider a linear time invariant (LTI) system with impulse response $h(t)$ and transfer function $H(f) = \mathcal{F}[h(t)]$. Suppose that the input to the system is a WSS process $\{X(t) : t \in \mathbb{R}\}$, as depicted in Figure 9.5. We wish to characterize its output

$$Y(t) = X(t) * h(t) = \int_{-\infty}^{\infty} X(\tau)h(t-\tau)\,d\tau.$$

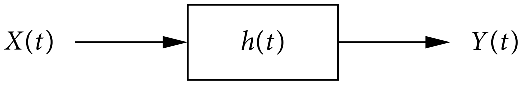


Figure 9.5. An LTI system driven by a WSS process input.

Assuming that the system is in *steady state*, it turns out (not surprisingly) that the output is also a WSS process. In fact, $X(t)$ and $Y(t)$ are *jointly WSS*, namely,

- $X(t)$ and $Y(t)$ are WSS, and
- their *crosscorrelation function*

  $$R_{XY}(t_1, t_2) = \mathsf{E}[X(t_1)Y(t_2)]$$

  is time invariant, $R_{XY}(t_1, t_2)$ is a function of $\tau = t_1 - t_2$.

Henceforth, we relabel $R_{XY}(t_1, t_2)$ as $R_{XY}(\tau)$, where $\tau = t_1 - t_2$. Note that unlike $R_X(\tau)$, $R_{XY}(\tau)$ is not necessarily even, but satisfies

$$R_{XY}(\tau) = R_{YX}(-\tau).$$

Example 9.11. Let $\Theta \sim \text{Unif}[0, 2\pi]$. Consider two processes

$$\begin{aligned} X(t) &= \alpha\cos(\omega t + \Theta), \\ Y(t) &= \alpha\sin(\omega t + \Theta). \end{aligned}$$

These processes are jointly WSS, since each is WSS (in fact SSS) and

$$\begin{aligned} R_{XY}(t_1, t_2) &= \mathsf{E}[\alpha^2\cos(\omega t_1 + \Theta)\sin(\omega t_2 + \Theta)] \\ &= \frac{\alpha^2}{4\pi}\int_0^{2\pi} \sin(\omega(t_1 + t_2) + 2\theta) - \sin(\omega(t_1 - t_2))\,d\theta \\ &= -\frac{\alpha^2}{2}\sin(\omega(t_1 - t_2)), \end{aligned}$$

which is a function only of $t_1 - t_2$.

We define the *cross power spectral density* for jointly WSS processes $\{X(t)\}$ and $\{Y(t)\}$ as

$$S_{XY}(f) = \mathcal{F}[R_{XY}(\tau)].$$

Example 9.12. Let $Y(t) = X(t) + Z(t)$, where $\{X(t)\}$ and $\{Z(t)\}$ are zero-mean uncorrelated WSS processes with power spectral densities $S_X(f)$ and $S_Z(f)$. Then, $\{X(t)\}$ and $\{Y(t)\}$ are jointly WSS. First, we show that $\{Y(t)\}$ is WSS, since it is zero mean and

$$\begin{aligned} R_Y(t_1, t_2) &= \mathsf{E}[(X(t_1) + Z(t_1))(X(t_2) + Z(t_2))] \\ &= \mathsf{E}[X(t_1)X(t_2)] + \mathsf{E}[Z(t_1)Z(t_2)] \\ &= R_X(\tau) + R_Z(\tau), \end{aligned}$$

where (a) follows since $\{X(t)\}$ and $\{Z(t)\}$ are zero mean and uncorrelated. Taking the Fourier transform of both sides,

$$S_Y(f) = S_X(f) + S_Z(f).$$

To show that $Y(t)$ and $X(t)$ are jointly WSS, consider

$$\begin{aligned} R_{XY}(t_1, t_2) &= \mathsf{E}[X(t_1)(X(t_2) + Z(t_2))] \\ &= \mathsf{E}[X(t_1)X(t_2)] + \mathsf{E}[X(t_1)Z(t_2)] \\ &= R_X(t_1, t_2), \end{aligned}$$

which is time invariant. Relabeling $R_X(t_1, t_2) = R_X(\tau)$ and taking the Fourier transform,

$$S_{XY}(f) = S_X(f).$$

Let $\{X(t)\colon t \in \mathbb{R}\}$ be a WSS process input to a LTI system with impulse response $h(t)$ and transfer function $H(f)$. If the system is *stable*, i.e.,

$$\left| \int_{-\infty}^{\infty} h(t)\, dt \right| = |H(0)| < \infty\,,$$

then the input $X(t)$ and the output $Y(t)$ are jointly WSS with the following properties:

1. $\mathsf{E}[Y(t)] = H(0)\,\mathsf{E}[X(t)]$.
2. $R_{YX}(\tau) = h(\tau) * R_X(\tau)$.
3. $R_Y(\tau) = h(\tau) * R_X(\tau) * h(-\tau)$.
4. $S_{YX}(f) = H(f)S_X(f)$.
5. $S_Y(f) = |H(f)|^2 S_X(f)$.

These properties are summarized in Figure 9.6.

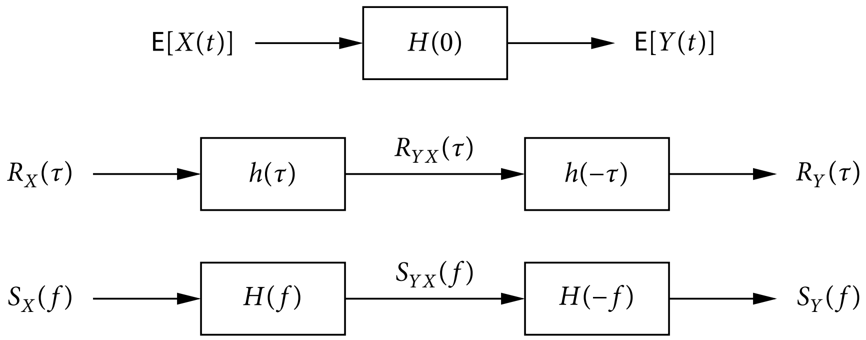


Figure 9.6. Properties of the output of an LTI system driven by a WSS process input.

To show the first property, consider

$$
\begin{aligned}
\mathsf{E}[Y(t)] &= \mathsf{E}\left[\int_{-\infty}^{\infty} X(\tau)h(t-\tau)\,d\tau\right] \\
&= \int_{-\infty}^{\infty} \mathsf{E}[X(\tau)]\,h(t-\tau)\,d\tau \\
&= \mathsf{E}[X(t)]\int_{-\infty}^{\infty} h(t-\tau)\,d\tau \\
&= \mathsf{E}[X(t)]H(0).
\end{aligned}
$$

For the second property,

$$
\begin{aligned}
R_{YX}(\tau) &= \mathsf{E}[Y(t+\tau)X(t)] \\
&= \mathsf{E}\left[\int_{-\infty}^{\infty} h(s)X(t+\tau-s)X(t)\,ds\right] \\
&= \int_{-\infty}^{\infty} h(s)R_X(\tau-s)\,ds \\
&= h(\tau) * R_X(\tau).
\end{aligned}
$$

For the third property, consider

$$
\begin{aligned}
R_Y(\tau) &= \mathsf{E}[Y(t+\tau)Y(t)] \\
&= \mathsf{E}\left[Y(t+\tau)\int_{-\infty}^{\infty} h(s)X(t-s)\,ds\right] \\
&= \int_{-\infty}^{\infty} h(s)R_{YX}(\tau+s)\,ds \\
&= R_{YX}(\tau) * h(-\tau).
\end{aligned}
$$

In the above, we have shown that $\mathsf{E}[Y(t+\tau)X(t)]$ and $\mathsf{E}[Y(t+\tau)Y(t)]$ are functions of τ (not of t), confirming that $\{X(t)\}$ and $\{Y(t)\}$ are jointly WSS. The fourth and fifth properties follow by taking the Fourier transforms of $R_{YX}(\tau)$ and $R_Y(\tau)$, respectively.

We can use the above properties to prove that $S_X(f)$ is indeed the power spectral density of $X(t)$, that is, $S_X(f)df$ is the average power contained in the frequency band $[f, f + df]$. Consider an ideal band-pass filter shown in Figure 9.7 with output

$$Y(t) = h(t) * X(t)$$

driven by a WSS process $\{X(t)\}$. Then the average power of $X(t)$ in the band $[f_1, f_2]$ is

$$\begin{aligned}
\mathsf{E}[Y^2(t)] &= \int_{-\infty}^{\infty} S_Y(f)\, df \\
&= \int_{-\infty}^{\infty} |H(f)|^2 S_X(f)\, df \\
&= \int_{-f_2}^{-f_1} S_X(f)\, df + \int_{f_1}^{f_2} S_X(f)\, df \\
&= 2\int_{f_1}^{f_2} S_X(f)\, df.
\end{aligned}$$

This argument also shows that $S_X(f) \ge 0$ for all f.

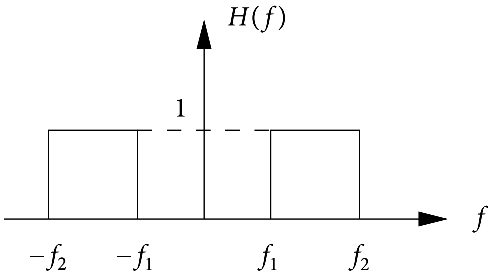


Figure 9.7. The transfer function of an ideal band-pass filter over $[f_1, f_2]$.

Example 9.13 (kT/C noise). The noise in a resistor R (in ohms) due to thermal noise is modeled as a WGN voltage source $V(t)$ in series with R; see Figure 9.8. The power spectral density of $V(t)$ is $S_V(f) = 2kTR\ \mathrm{V}^2/\mathrm{Hz}$ for all f, where k is the Boltzmann constant and T is the temperature in degrees K. We find the average output noise power for an RC circuit shown in Figure 9.9. First, note that the transfer function for the circuit is

$$H(f) = \frac{1}{1 + i2\pi fRC},$$

which implies that

$$|H(f)|^2 = \frac{1}{1 + (2\pi fRC)^2}.$$

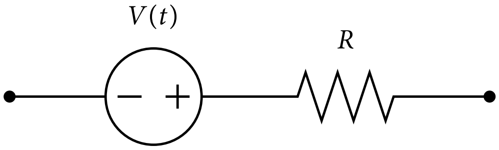


Figure 9.8. A model of the thermal noise in a resistor.

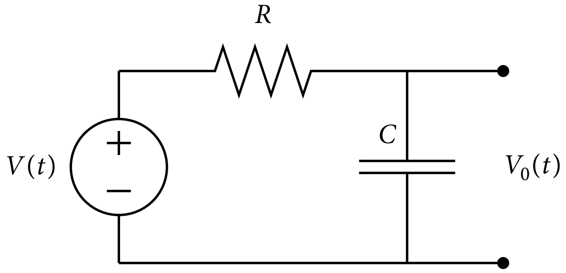


Figure 9.9. An RC circuit driven by the thermal noise.

Now we write the output power spectral density in terms of the input power spectral density as

$$S_{V_o}(f) = S_V(f)|H(f)|^2 = 2kTR\,\frac{1}{1+(2\pi fRC)^2}.$$

Thus the average output power is

$$\begin{aligned}
\mathsf{E}[V_o^2(t)] &= \int_{-\infty}^{\infty} S_{V_o}(f)df \\
&= \frac{2kTR}{2\pi RC}\int_{-\infty}^{\infty}\frac{1}{1+(2\pi fRC)^2}d(2\pi fRC) \\
&= \frac{kT}{\pi C}\int_{-\infty}^{\infty}\frac{1}{1+x^2}\,dx \\
&= \frac{kT}{\pi C}\arctan x\Big|_{-\infty}^{+\infty} \\
&= \frac{kT}{C},
\end{aligned}$$

which is independent of R.

9.6 LINEAR ESTIMATION OF RANDOM PROCESSES

Let $\{X(t)\}$ and $\{Y(t)\}$ be zero mean jointly WSS processes with known autocorrelation and crosscorrelation functions $R_X(\tau)$, $R_Y(\tau)$, and $R_{XY}(\tau)$. We observe the random process $Y(s)$ for $t - a \le s \le t + b$ $(-a \le b)$ and wish to find the linear MMSE estimate of the signal $X(t)$, i.e., the estimate of the form

$$\hat{X}(t) = \int_{-b}^{a} h(s)Y(t-s)\,ds$$

that minimizes the mean square error

$$\mathsf{E}[(X(t) - \hat{X}(t))^2].$$

By the orthogonality principle, the linear MMSE estimate must satisfy

$$(X(t) - \hat{X}(t)) \perp Y(t-s)\,, \quad -b \le s \le a,$$

or equivalently,

$$\mathsf{E}[(X(t) - \hat{X}(t))Y(t-s)] = 0\,, \quad -b \le s \le a.$$

Thus, for $-b \le s \le a$, the optimal estimation filter $h(s)$ must satisfy

$$\begin{aligned} R_{XY}(\tau) &= \mathsf{E}[X(t)Y(t-\tau)] \\ &= \mathsf{E}[\hat{X}(t)Y(t-\tau)] \\ &= \mathsf{E}\Big[\int_{-b}^{a} h(s)Y(t-s)Y(t-\tau)\,ds\Big] \\ &= \int_{-b}^{a} h(s)R_Y(\tau - s)\,ds \end{aligned}$$

for $-b \le s \le a$. To find $h(s)$, we need to solve an infinite set of integral equations. Solving these equations analytically is not possible in general. However, it can be done for two important special cases:

- *Infinite smoothing*: $a, b \to \infty$.
- *Filtering*: $a \to \infty$ and $b = 0$

We discuss only the first case. The second case leads to the *Wiener–Hopf equations* from which the famous *Wiener filter* is derived.

When $a, b \to \infty$, the integral equations for the linear MMSE estimate become

$$R_{XY}(\tau) = \int_{-\infty}^{\infty} h(s)R_Y(\tau - s)\,ds\,, \quad -\infty < \tau < +\infty.$$

In other words,

$$R_{XY}(\tau) = h(\tau) * R_Y(\tau).$$

By taking Fourier transforms, we have

$$S_{XY}(f) = H(f)S_Y(f),$$

which implies that the optimal *infinite smoothing filter* is

$$H(f) = \frac{S_{XY}(f)}{S_Y(f)}.$$

Observe the similarity between the optimal filter and the LMMSE estimate

$$\hat{X} = \Sigma_{\mathbf{Y}X}^T \Sigma_{\mathbf{Y}}^{-1}(\mathbf{Y} - \mathsf{E}[\mathbf{Y}]) + \mathsf{E}[X]$$

for the vector case discussed in Section 6.6.

The LMMSE filter achieves the MSE

$$\begin{aligned}
\mathsf{E}[(X(t) - \hat{X}(t))^2] &= \mathsf{E}[(X(t) - \hat{X}(t))X(t)] - \mathsf{E}[(X(t) - \hat{X}(t))\hat{X}(t)] \\
&\overset{(a)}{=} \mathsf{E}[(X(t) - \hat{X}(t))X(t)] \\
&= \mathsf{E}[(X(t)^2] - \mathsf{E}[\hat{X}(t)X(t)],
\end{aligned}$$

where (a) follows by orthogonality. To evaluate the second term, consider

$$\begin{aligned}
R_{X\hat{X}}(\tau) &= \mathsf{E}[X(t + \tau)\hat{X}(t)] \\
&= \mathsf{E}\Big[X(t + \tau)\int_{-\infty}^{\infty} h(s)Y(t - s)\, ds\Big] \\
&= \int_{-\infty}^{\infty} h(s)R_{XY}(\tau + s)\, ds \\
&= R_{XY}(\tau) * h(-\tau),
\end{aligned}$$

which implies that

$$S_{X\hat{X}}(f) = H(-f)S_{XY}(f).$$

Therefore,

$$\begin{aligned}
\mathsf{E}[\hat{X}(t)X(t)] &= R_{X\hat{X}}(0) \\
&= \int_{-\infty}^{\infty} S_{X\hat{X}}(f)\, df \\
&= \int_{-\infty}^{\infty} H(-f)S_{XY}(f)\, df \\
&= \int_{-\infty}^{\infty} \frac{|S_{XY}(f)|^2}{S_Y(f)}\, df
\end{aligned}$$

and the minimum MSE is

$$\begin{aligned}\mathsf{E}[(X(t)-\hat{X}(t))^2] &= \mathsf{E}[(X(t)^2] - \mathsf{E}[\hat{X}(t)X(t)]\\ &= \int_{-\infty}^{\infty} S_X(f)\,df - \int_{-\infty}^{\infty} \frac{|S_{XY}(f)|^2}{S_Y(f)}\,df\\ &= \int_{-\infty}^{\infty} \left(S_X(f) - \frac{|S_{XY}(f)|^2}{S_Y(f)}\right) df.\end{aligned}$$

This MSE can be compared to

$$\sigma_X^2 - \Sigma_{\mathbf{Y}X}^T \Sigma_{\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}X}$$

for the vector case.

Example 9.14 (Additive white noise channel). Let $\{X(t)\}$ and $\{Z(t)\}$ be zero-mean uncorrelated WSS processes with

$$S_X(f) = \begin{cases} P/2 & |f| \le B,\\ 0 & \text{otherwise,}\end{cases}$$

and

$$S_Z(f) = \frac{N}{2} \quad \text{for all } f,$$

as shown in Figure 9.10. In other words, the signal $\{X(t)\}$ is band-limited white noise and the noise $\{Z(t)\}$ is white. We find the optimal infinite smoothing filter for estimating $X(t)$ given

$$Y(\tau) = X(\tau) + Z(\tau)\,, \quad -\infty < \tau < +\infty.$$

The transfer function of the optimal filter is

$$\begin{aligned}H(f) &= \frac{S_{XY}(f)}{S_Y(f)}\\ &= \frac{S_X(f)}{S_X(f) + S_Z(f)}\\ &= \begin{cases} P/(P+N) & |f| \le B,\\ 0 & \text{otherwise,}\end{cases}\end{aligned}$$

as shown in Figure 9.11. The MSE of the optimal filter is

$$\begin{aligned}\int_{-\infty}^{\infty} S_X(f) - \frac{|S_{XY}(f)|^2}{S_Y(f)}\,df &= \int_{-B}^{B} \frac{P}{2} - \frac{(P/2)^2}{P/2 + N/2}\,df\\ &= PB - \frac{P^2/4}{(P+N)/2}\,2B\\ &= \frac{NPB}{N+P}.\end{aligned}$$

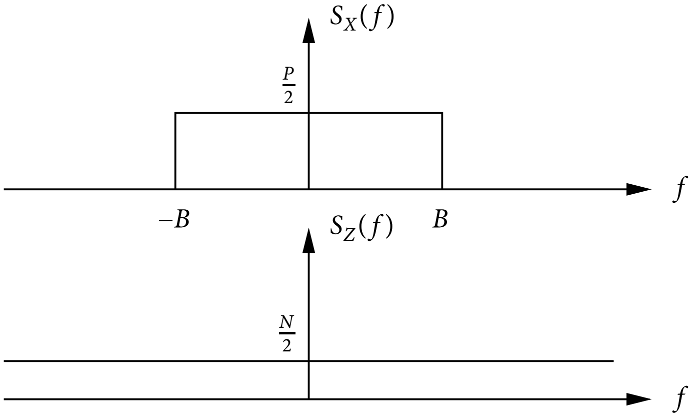


Figure 9.10. Power spectral densities $S_X(f)$ and $S_Z(f)$.

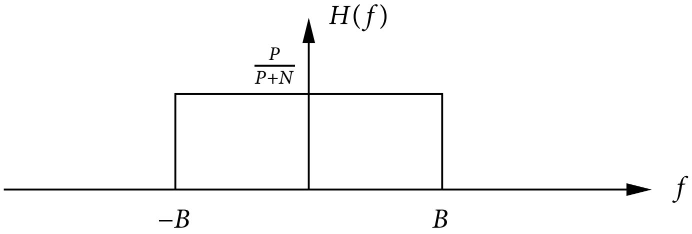


Figure 9.11. The transfer function $H(f)$ of the optimal infinite smoothing filter.

PROBLEMS

9.1. *Moving average process.* Let $Z_0, Z_1, Z_2, \ldots$ be i.i.d. $\sim \mathcal{N}(0,1)$.

(a) Let $X_n = \frac{1}{2}Z_{n-1} + Z_n$ for $n \geq 1$. Find the mean and autocorrelation function of X_n.

(b) Is $\{X_n\}$ wide-sense stationary?

(c) Is $\{X_n\}$ Gaussian?

(d) Is $\{X_n\}$ strict-sense stationary?

(e) Find $\mathsf{E}(X_3|X_1, X_2)$.

(f) Find $\mathsf{E}(X_3|X_2)$.

(g) Is $\{X_n\}$ Markov?

(h) Is $\{X_n\}$ independent increment?

(i) Let $Y_n = Z_{n-1} + Z_n$ for $n \geq 1$. Find the mean and autocorrelation functions of $\{Y_n\}$.

(j) Is $\{Y_n\}$ wide-sense stationary?

(k) Is $\{Y_n\}$ Gaussian?

(l) Is $\{Y_n\}$ strict-sense stationary?

(m)Find $\mathsf{E}(Y_3|Y_1, Y_2)$.

(n) Find $\mathsf{E}(Y_3|Y_2)$.

(o) Is $\{Y_n\}$ Markov?

(p) Is $\{Y_n\}$ independent increment?

9.2. *Random binary modulation.* Let $\{X_n\}$ be a zero-mean wide-sense stationary random process with autocorrelation function $R_X(n)$, and $Z_1, Z_2, \ldots$ be i.i.d. Bern(p) random variables, i.e.,

$$Z_i = \begin{cases} 1, & \text{with probability } p, \\ 0, & \text{with probability } 1 - p. \end{cases}$$

Let

$$Y_n = X_n \cdot Z_n, \quad n = 1, 2, \ldots.$$

(a) Find the mean and the autocorrelation function of $\{Y_n\}$ in terms of $R_X(n)$ and p.

(b) Is $\{Y_n\}$ jointly wide-sense stationary with $\{X_n\}$?

9.3. *Random binary waveform (40 pts).* Let $\{N(t)\}_{t=0}^{\infty}$ be a Poisson process with rate λ, and Z be independent of $\{N(t)\}$ with $\mathsf{P}(Z = 1) = \mathsf{P}(Z = -1) = 1/2$. Define

$$X(t) = Z \cdot (-1)^{N(t)}, \quad t \geq 0.$$

(a) Find the mean and autocorrelation function of $\{X(t)\}_{t=0}^{\infty}$.

(b) Is $\{X(t)\}_{t=0}^{\infty}$ wide-sense stationary?

(c) Find the first-order pmf $p_{X(t)}(x) = \mathsf{P}(X(t) = x)$.

(d) Find the second-order pmf $p_{X(t_1),X(t_2)}(x_1, x_2) = \mathsf{P}(X(t_1) = x_1,\ X(t_2) = x_2)$.

(Hint: $\sum_{k \text{ even}} x^k/k! = (e^x + e^{-x})/2$ and $\sum_{k \text{ odd}} x^k/k! = (e^x - e^{-x})/2$.)

9.4. *QAM random process.* Consider the random process

$$X(t) = Z_1 \cos \omega t + Z_2 \sin \omega t\,, \quad -\infty < t < \infty\,,$$

where Z_1 and Z_2 are i.i.d. discrete random variables such that $p_{Z_i}(+1) = p_{Z_i}(-1) = \frac{1}{2}$.

(a) Is $X(t)$ wide-sense stationary? Justify your answer.

(b) Is $X(t)$ strict-sense stationary? Justify your answer.

9.5. *Mixture of two WSS processes.* Let $X(t)$ and $Y(t)$ be two zero-mean WSS processes with autocorrelation functions $R_X(\tau)$ and $R_Y(\tau)$, respectively. Define the process

$$Z(t) = \begin{cases} X(t), & \text{with probability } \frac{1}{2} \\ Y(t), & \text{with probability } \frac{1}{2}. \end{cases}$$

Find the mean and autocorrelation functions for $Z(t)$. Is $Z(t)$ a WSS process?

9.6. *Stationary Gauss–Markov process.* Let

$$\begin{aligned} X_0 &\sim \mathrm{N}(0, a) \\ X_n &= \tfrac{1}{2} X_{n-1} + Z_n\,, \quad n \geq 1\,, \end{aligned}$$

where $Z_1, Z_2, Z_3, \ldots$ are i.i.d. N(0, 1) independent of X_0.

(a) Find a such that X_n is stationary. Find the mean and autocorrelation functions of X_n.

(b) (Difficult.) Consider the sample mean $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, $n \geq 1$. Show that S_n converges to the process mean in probability even though the sequence X_n is not i.i.d. (A stationary process for which the sample mean converges to the process mean is called *mean ergodic.*)

9.7. *AM modulation.* Consider the AM modulated random process

$$X(t) = A(t)\cos(2\pi t + \Theta)\,,$$

where the amplitude $A(t)$ is a zero-mean WSS process with autocorrelation function $R_A(\tau) = e^{-\frac{1}{2}|\tau|}$, the phase Θ is a Unif$[0, 2\pi)$ random variable, and $A(t)$ and Θ are independent. Is $X(t)$ a WSS process?

9.8. *Random-delay mixture.*
Let $\{X(t)\}$, $-\infty < t < \infty$, be a zero-mean wide-sense stationary process with autocorrelation function $R_X(\tau) = e^{-|\tau|}$. Let

$$Y(t) = X(t - U),$$

where U is a random delay, independent of $\{X(t)\}$. Suppose that $U \sim$ Bern(1/2), that is,

$$Y(t) = \begin{cases} X(t) & \text{with probability } 1/2, \\ X(t-1) & \text{with probabiltiy } 1/2. \end{cases}$$

(a) Find the mean and autocorrelation functions of $\{Y(t)\}$.

(b) Is $\{Y(t)\}$ wide-sense stationary? Justify your answer.

(c) Find the average power $\mathsf{E}(Y(t)^2)$ of $\{Y(t)\}$.

(d) Now suppose that $U \sim$ Exp(1), i.e., $f_U(u) = e^{-u}$, $u \geq 0$. Find the mean and autocorrelation function of $\{Y(t)\}$.

9.9. *Linear estimation.* Let $X(t)$ be a zero-mean WSS process with autocorrelation function $R_X(\tau) = e^{-|\tau|}$. Let $Y(t)$ be another zero-mean process which is jointly WSS with $X(t)$ with cross correlation function $R_{XY}(\tau) = \int_0^1 R_X(s - \tau)ds$.

(a) Find the linear MMSE estimate of $Y(t)$ given $X(t)$. Leave your answer in terms of e.

(b) Find the linear MMSE estimate of $Y(t)$ given $X(t)$ and $X(t + 1)$.

9.10. *LTI system with WSS process input.* Let $Y(t) = h(t) * X(t)$ and $Z(t) = X(t) - Y(t)$ as shown in the Figure 9.12.

(a) Find $S_Z(f)$.

(b) Find $E(Z^2(t))$.

Your answers should be in terms of $S_X(f)$ and the transfer function $H(f) = \mathcal{F}[h(t)]$.

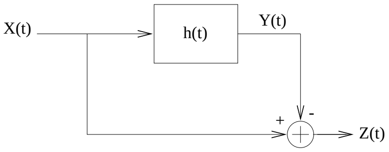


Figure 9.12. LTI system.

9.11. *Echo filtering.* A signal $X(t)$ and its echo arrive at the receiver as $Y(t) = X(t) + X(t - \Delta) + Z(t)$. Here the signal $X(t)$ is a zero-mean WSS process with power spectral density $S_X(f)$ and the noise $Z(t)$ is a zero-mean WSS with power spectral density $S_Z(f) = N_0/2$, uncorrelated with $X(t)$.

(a) Find $S_Y(f)$ in terms of $S_X(f)$, Δ, and N_0.

(b) Find the best linear filter to estimate $X(t)$ from $\{Y(s)\}_{-\infty<s<\infty}$.

9.12. *Discrete-time LTI system with white noise input.* Let $\{X_n : -\infty < n < \infty\}$ be a discrete-time white noise process, i.e., $\mathsf{E}(X_n) = 0$, $-\infty < n < \infty$, and

$$R_X(n) = \begin{cases} 1 & n = 0, \\ 0 & \text{otherwise.} \end{cases}$$

The process is filtered using a linear time invariant system with impulse response

$$h(n) = \begin{cases} \alpha & n = 0, \\ \beta & n = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find α and β such that the output process Y_n has

$$R_Y(n) = \begin{cases} 2 & n = 0, \\ 1 & |n| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

9.13. *Finding time of flight.* Finding the distance to an object is often done by sending a signal and measuring the time of flight, the time it takes for the signal to return (assuming speed of signal, e.g., light, is known). Let $X(t)$ be the signal sent and $Y(t) = X(t - \delta) + Z(t)$ be the signal received, where δ is the unknown time of flight. Assume that $X(t)$ and $Z(t)$ (the sensor noise) are uncorrelated zero mean WSS processes. The estimated crosscorrelation function of $Y(t)$ and $X(t)$, $R_{YX}(t)$ is shown in Figure 9.13. Find the time of flight δ.

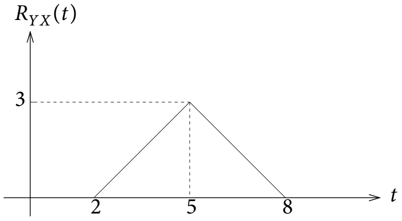


Figure 9.13. Crosscorrelation function.

9.14. *Finding impulse response of LTI system.* To find the impulse response $h(t)$ of an LTI system (e.g., a concert hall), i.e., to *identify* the system, white noise $X(t)$, $-\infty < t < \infty$, is applied to its input and the output $Y(t)$ is measured. Given the input and output sample functions, the crosscorrelation $R_{YX}(\tau)$ is estimated. Show how $R_{YX}(\tau)$ can be used to find $h(t)$.

9.15. *Generating a random process with a prescribed power spectral density.* Let $S(f) \geq 0$, for $-\infty < f < \infty$, be a real and even function such that

$$\int_{-\infty}^{\infty} S(f) df = 1.$$

Define the random process

$$X(t) = \cos(2\pi F t + \Theta),$$

where $F \sim S(f)$ and $\Theta \sim \mathrm{U}[-\pi, \pi)$ are independent. Find the power spectral density of $X(t)$. Interpret the result.

9.16. *Integration.* Let $X(t)$ be a zero-mean WSS process with autocorrelation function $R_X(\tau) = e^{-|\tau|}$. Let

$$Y(t) = \int_{t}^{t+1} X(s) ds.$$

(a) Is $Y(t)$ WSS?

(b) Is $(X(t), Y(t))$ jointly WSS?

(c) Find the linear MMSE estimate of $Y(t)$ given $X(t)$. Leave your answer in terms of e.

(d) Find the linear MMSE estimate of $Y(t)$ given $X(t)$ and $X(t+1)$.

9.17. *Integrators.* Let $Y(t)$ be a short-term integration of a WSS process $X(t)$:

$$Y(t) = \frac{1}{T}\int_{t-T}^{t} X(u)du.$$

Find $S_Y(f)$ in terms of $S_X(f)$.

9.18. *Derivatives of stochastic processes.* Let $\{X(t)\}$ be a wide-sense stationary random process with mean zero and autocorrelation function $R(\tau) = e^{-|\tau|}$. Recall that a random process $\{Y(t)\}$ is continuous in mean square if $\mathsf{E}[(Y(t+\epsilon) - Y(t))^2] \to 0$ as $\epsilon \to 0$.

(a) Find the mean and the variance of $X(t)$.

(b) Is $X(t)$ continuous in mean square? Justify your answer.

(c) Now let

$$Z_\epsilon(t) = \frac{X(t+\epsilon) - X(t)}{\epsilon}$$

be an ϵ-approximation of the derivative $\dot{X}(t)$. Find the mean and the variance of $Z_\epsilon(t)$.

(d) Find the linear MMSE estimate of $Z_\epsilon(t)$ given $(X(t), X(t+\epsilon))$ and the associated MSE.

(e) Find the linear MMSE estimate of $Z_\epsilon(t)$ given $X(t)$ and the associated MSE.

(f) Find the limiting mean and variance of $Z_\epsilon(t)$ as $\epsilon \to 0$.