So far in this course, we only considered deterministic signals. In many practical situations, this is not enough. It is much more convenient to model many processes as random signals.

For example, images acquired with a medical imaging device always have noise. The exact intensity values vary from image to image. They depend on many different factors which are beyond our control, and which therefore cannot be reliably modeled. However, what we may be able to reliably model is the average behavior of these signals. In other words, if some averaging procedure is performed over a hundred images, perhaps it will be possible to predict how the next hundred images will behave, on average. Once we have a good probabilistic model, we could pose and solve various useful estimation problems, in order to, e.g., remove noise and enhance the image quality.

Our objective for this topic\footnote{Probabilistic modeling—i.e., assigning probabilities to events—will not be studied here. We will always assume either that a probabilistic model is given, or that it can be easily constructed.} will be to develop the analysis tools for random signals. We will start by reviewing some basic facts about probability.

\section*{2.1 Introduction to Random Sequences, Detection, and Estimation}

\subsection*{2.1.1 Events and Probability}

The main concepts are as follows.

- An \textit{outcome of an experiment}, or an \textit{elementary event}.
- The \textit{sample space} $\Omega$ for an experiment is the set of all possible outcomes of the experiment.
- An \textit{event} is a set in the sample space.\footnote{Not every set in the sample space must necessarily be an event with a well-defined probability. We will not discuss this here, deferring to advanced courses on probability and measure theory.} We say that event $A$ occurs if one of the outcomes in the set $A$ occurs.
- A \textit{probability measure} is a function which assigns a probability (a number) $P(A)$ to each event $A$, and which satisfies the following properties:
1. It is non-negative: $P(A) \geq 0$.
2. The probability of the whole sample space is 1: $P(\Omega) = 1$.
3. It is countably additive:

$$P \left( \bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} P(A_i), \text{ if } A_i \cap A_j = \emptyset \text{ for any } i, j.$$ 

Recall:

$A_1 + A_2 = A_1 \cup A_2 = \text{union of } A_1, A_2 = \text{“}A_1 \text{ or } A_2\text{”}$, 
i.e., the union of $A_1$ and $A_2$ consists of every outcome which is in $A_1$ or $A_2$ (or both);

$A_1 A_2 = A_1 \cap A_2 = \text{intersection of } A_1, A_2 = \text{“}A_1 \text{ and } A_2\text{”}$, 
i.e., the intersection of $A_1$ and $A_2$ consists of every outcome which is both in $A_1$ and in $A_2$ (note that $A_1 \cap A_2 = \emptyset$ means that $A_1$ and $A_2$ are mutually exclusive events);

$A^c = \overline{A} = \text{complement of } A = \text{“}\text{not } A\text{”}$, 
i.e., the complement of $A$ consists of all the elements of $\Omega$ which are not contained in $A$.

**Example 2.1.** Consider a random experiment where a fair six-sided die is thrown once. Its sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, and $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$.

We define the following events:

$A_1 = \text{“an even number turns up”, i.e., } A_1 = \{2, 4, 6\}$.

$A_2 = \text{“a prime number turns up”, i.e., } A_2 = \{2, 3, 5\}$.

$A_3 = \text{“one turns up”, i.e., } A_3 = \{1\}$.

We can calculate

$$P(A_1) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \quad \text{by property 3 (additivity)}.$$ 

$A_1 + A_2 = A_1 \cup A_2 = \text{“the number which turns up is either even or prime (or both)”}$

$= \{2, 3, 4, 5, 6\}$

$A_1 A_2 = A_1 \cap A_2 = \text{“the number which turns up is both even and prime”}$

$= \{2\}$
\( \overline{A}_1 = \text{“an odd number turns up”} = \{1, 3, 5\} \)
\( \overline{A}_2 = \text{“a non-prime number turns up”} = \{1, 4, 6\} \)
\( \overline{A}_1 \cup \overline{A}_2 = \{1, 3, 4, 5, 6\} = \overline{A}_1 \cap \overline{A}_2 \)
\( \overline{A}_1 \cap \overline{A}_2 = \{1\} = \overline{A}_1 \cup \overline{A}_2 \)

### 2.1.2 Conditional Probability

In observing the outcomes of a random experiment, we are often interested in how the outcome of one event \( A \) is influenced by that of another event \( B \). For example, in one extreme case, \( A \) may always occur whenever \( B \) does, as in Fig. 2.1(a), while in the other extreme case, \( A \) never occurs if \( B \) does, as in Fig. 2.1(b).

![Figure 2.1](image)

(a) \( A \) always occurs if \( B \) does.  (b) \( A \) never occurs if \( B \) occurs.

**Figure 2.1.** Two illustrations of dependence between events \( A \) and \( B \).

To characterize the relationship between \( A \) and \( B \), we introduce the conditional probability of \( A \) given \( B \) which is the probability of \( A \) occurring if \( B \) is known to have occurred.

**Definition 2.1.** The conditional probability of \( A \) given \( B \) is defined as follows:

\[
P(A|B) \overset{\text{def}}{=} \frac{P(A \cap B)}{P(B)},
\]

assuming that \( P(B) > 0 \).

Conditional probabilities specify a probability law on the new universe \( B \): it can be shown that they satisfy all three axioms. Therefore, all general properties of probability laws remain valid for conditional probabilities.
Example 2.2. In the previous die-throwing experiment Ex. 2.1, what is $P(A_1|A_2)$?

$$P(A_1|A_2) = P(\text{even number turns up, given that a prime number turned up})$$

$$= \frac{P(A_1 \cap A_2)}{P(A_2)}$$

$$= \frac{P\{2\}}{P\{2,3,5\}}$$

$$= \frac{1/6}{1/2}$$

$$= \frac{1}{3}.$$

If we look at Fig. 2.2, we immediately see that $P(A_1|A_2) = \frac{1}{3}$ because, given that $A_2$ has occurred, only one outcome (i.e., 2) out of three equiprobable outcomes (2,3,5) makes $A_1$ occur.

![Figure 2.2. A diagram for the die-throwing experiment.](image)

When we are conditioning on $A_2$, we are no longer operating in the whole of $\Omega$: $A_2$ becomes the sample space and hence all probabilities conditioned on $A_2$ are normalized by $P(A_2)$. In particular,

$$P(A_2|A_2) = \frac{P(A_2)}{P(A_2)} = 1.$$

Properties of conditional probabilities

1. $0 \leq P(A|B) \leq 1$.
2. If $A \cap B = \emptyset$, then $P(A|B) = 0$.
3. If $B \subset A$ (i.e., $B \Rightarrow A$), then $P(A|B) = 1$. 
4. If $A_1, A_2, \cdots$ are mutually exclusive with

$$A = \bigcup_{k=1}^{\infty} A_k,$$

then we have

$$P(A|B) = \sum_{k=1}^{\infty} P(A_k|B).$$

5. **Total Probability Theorem.**

$$P(A|B) = \sum_{k=1}^{\infty} P(A_k|B).$$

**Figure 2.3.** When $B$ is a subset of $A$, $P(A|B) = 1$.

**Figure 2.4.** $P(A|B) = \sum_{k=1}^{\infty} P(A_k|B)$.

**Figure 2.5.** The set $B$ is contained in the union of mutually exclusive sets $A_1$, $A_2$, and $A_3$. 
If the sample space is partitioned into sets $A_1$, $A_2$, and $A_3$ as in Fig. 2.5, then the probability of any event $B$ can be computed as follows:

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3).$$

More generally,

$$P(B) = \sum_i P(B|A_i)P(A_i),$$

if $A_i$’s are mutually exclusive and $B \subset \bigcup_i A_i$.

6. **Bayes’ Rule** is used for combining evidence. It tells us how to compute probabilities $P(A_i|B)$ if $P(A_i)$ and $P(B)$ are known. Here is some commonly used terminology:

- “Prior model”: probabilities $P(A_i)$.
- “Measurement model”: $P(B|A_i) = \text{the conditional probability to observe data } B \text{ given that the truth is } A$.
- “Posterior probabilities” $P(A_i|B) = \text{the conditional probability that the truth is } A_i, \text{ given that we observed data } B$.

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B)} \overset{\text{total probability theorem}}{=} \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}.$$

7. **Multiplication Rule.**

$$P \left( \bigcap_{i=1}^n A_i \right) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P \left( A_n \bigg| \bigcap_{j=1}^{n-1} A_j \right),$$

provided that all the conditioning events have positive probability.

**Example 2.3.** You leave point $O$, choosing one the roads $OB_1$, $OB_2$, $OB_3$, $OB_4$ at random, with equal probabilities (Fig. 2.6). At every subsequent crossroads you again choose a road at random, with equal probabilities. What is the probability that you will arrive at $A$?

**Solution.**

$$P(\text{arrive at } B_k \text{ from } O) = \frac{1}{4} \text{ for } k = 1, 2, 3, 4.$$
Figure 2.6. Illustration of Example 2.3.

Applying the property 5 of conditional probabilities, we get (the notation $X \leftarrow Y$ means that we arrive at $X$ from $Y$):

\[
P(\text{arrive at } A) = P(A \leftarrow B_1|B_1 \leftarrow O) \cdot P(B_1 \leftarrow O) \\
+ P(A \leftarrow B_2|B_2 \leftarrow O) \cdot P(B_2 \leftarrow O) \\
+ P(A \leftarrow B_3|B_3 \leftarrow O) \cdot P(B_3 \leftarrow O) \\
+ P(A \leftarrow B_4|B_4 \leftarrow O) \cdot P(B_4 \leftarrow O)
\]

\[
= \frac{1}{3} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{2}{5} \cdot \frac{1}{4}
\]

\[
= \frac{1}{4} \cdot \frac{10 + 15 + 30 + 12}{30}
\]

\[
= \frac{67}{120}.
\]

This property often allows us to break complicated problems into simpler steps, and efficiently solve them. This basic principle is at the heart of Kalman filtering, and more general techniques of estimation on graphs, as well as efficient coding algorithms.

### 2.1.3 Statistical Independence

**Definition 2.2.** If $P(A \cap B) = P(A)P(B)$, then the events $A$ and $B$ are said to be statistically independent. Otherwise, they are said to be statistically dependent.

Recall that we introduced conditional probability to talk about the information that the occurrence of one event provides about another event. Independent events constitute a special case when no such information is provided. Since $P(A \cap B) =$

---

3The qualifier *statistical* is used to avoid confusion of the notion of independence of events or random variables with the notion of linear independence of vectors. When there is no danger of ambiguity, we will simply say that events $A$ and $B$ are independent.
$P(A|B)P(B) = P(B|A)P(A)$, independence is the same as

\[ P(A|B) = P(A) \quad \text{if } P(B) \neq 0 \]
\[ P(B|A) = P(B) \quad \text{if } P(A) \neq 0. \]

In other words, the occurrence of $B$ has no influence on the probability of the occurrence of $A$, and vice versa.

**Example 2.4.** Pick a card from a full deck and define the following events:

\[ A = \{ \text{the card picked at random from a full deck is a spade} \} \]
\[ B = \{ \text{the card picked at random from a full deck is a queen} \} \]

Are the events $A$ and $B$ independent?

The probabilities of events $A$ and $B$ are:

\[ P(A) = \frac{13}{52} = \frac{1}{4} \]
\[ P(B) = \frac{4}{52} = \frac{1}{13}. \]

Then the probability of their combined occurrence is:

\[ P(A \cap B) = P(\text{the card is the queen of spades}) = \frac{1}{52} = \frac{1}{4} \cdot \frac{1}{13} = P(A)P(B). \]

Hence, $A$ and $B$ are independent.

**Example 2.5.** Disjoint does not mean independent. In fact, if the events $A$ and $B$ are disjoint and if $P(A) > 0$ and $P(B) > 0$ then

\[ P(A \cap B) = P(\varnothing) = 0 \]
\[ P(A)P(B) > 0, \]

and therefore $A$ and $B$ are dependent.

**Statistical independence of a collection of events** A collection of events is called independent if information on some of the events tells us nothing about probabilities related to the remaining events.

**Definition 2.3.** $A_1, A_2, A_3, \cdots, A_n$ are called statistically independent if

\[ P\left( \bigcap_{i \in S} A_i \right) = \prod_{i \in S} P(A_i), \]
for every subset $S$ of $\{1, 2, \cdots, n\}$. In other words,

$$P(A_i \cap A_j) = P(A_i)P(A_j), \text{ for every pair } i, j.$$  

$$P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k), \text{ for every triplet } i, j, k.$$  

$$\vdots$$  

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2)\cdots P(A_n)$$

**Example 2.6.** Throw a pair dice and define the following events:

$$A_1 = \{\text{first die turns up an odd number}\}$$  

$$A_2 = \{\text{second die turns up an odd number}\}$$  

$$A_3 = \{\text{the total of both numbers is odd}\}$$

Are these events pairwise independent? Furthermore, are they independent?

**Solution.** We have: $P(A_1) = P(A_2) = 1/2$. Assuming that all 36 outcomes of the pair of throws are equally likely, $P(A_1 \cap A_2) = 9/36 = 1/4 = P(A_1)P(A_2)$, and therefore, $A_1$ and $A_2$ are independent. Moreover,  

$$P(A_3) = \frac{1}{2}.$$  

Given that $A_1$ has occurred, $A_3$ occurs if and only if the second die turns up even:

$$P(A_3|A_1) = \frac{1}{2} = P(A_3) \Rightarrow A_1 \text{ and } A_3 \text{ are independent.}$$

Similarly,  

$$P(A_3|A_2) = \frac{1}{2} = P(A_3) \Rightarrow A_2, A_3 \text{ are independent.}$$

However, $A_3$ cannot occur if $A_1$ and $A_2$ both occur, and so  

$$P(A_1 \cap A_2 \cap A_3) = 0,$$

whereas,  

$$P(A_1)P(A_2)P(A_3) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} \neq P(A_1 \cap A_2 \cap A_3).$$

Thus, $A_1$, $A_2$, and $A_3$ are not independent.

### 2.1.4 Random Variables

**Definition 2.4.** A random variable is an assignment of a value (number) to every possible outcome. We refer to this number as the numerical value or the experimental outcome of the random variable.

In other words, a random variable is a function from the sample space to the real numbers, $\Omega \rightarrow \mathbb{R}$. 
Example 2.7. Suppose our experiment consists of two independent tosses of a fair coin. Let $X =$ the number of tails. Then $X$ is a random variable.

• Notation:
  – Random variable $X$.
  – Experimental value $x$ or $k$ or any other letters.

• We will often analyze random variables without specifying the experiment which they come from. However, it is important to always remember that a random variable is a function from the sample space of some random experiment to a set of numbers.

A random variable is called discrete if its range (the set of values it can take) is finite or at most countably infinite.

Definition 2.5. Probability mass function (pmf) of a discrete random variable $X$ is:

$$p_x(x) = P(X = x).$$

$X$ is a continuous random variable if there is a function $f_x \geq 0$, called the probability density function of $X$ (or pdf of $X$) such that

$$P(X \in B) = \int_B f_x(x)dx, \text{ for any } B \subset \mathbb{R}.$$

Note that, for a discrete random variable,

$$P(X \in B) = \sum_{x \in B} p_x(x).$$

Definition 2.6. The cumulative distribution function (cdf) of a random variable $X$ is defined as

$$F_X(x) = P\{X \leq x\}.$$

This definition applies to any random variable $X$.

Example 2.8. Throw a die; $X$ is the number which turns up.
Sec. 2.1. Introduction to Random Sequences, Detection, and Estimation

Example 2.9. Toss a point “at random” onto the interval \([0, 6]\); \(X\) is the point where it lands.

Some properties:

- \(F_X(x)\) is non-decreasing.
- \(F_X(x) \to 0\) as \(x \to -\infty\).
- \(F_X(x) \to 1\) as \(x \to \infty\).
• If \( X \) is a continuous random variable,
\[
f_X(x) = \frac{dF_X(x)}{dx}.
\]

• If \( X \) is a continuous random variable,
\[
P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx
\]
\[
F_X(b) = \int_{-\infty}^b f_X(x)dx
\]
\[
\int_{-\infty}^\infty f_X(x)dx = 1.
\]

**Expectation (or expected value)**

\[
E\{g(X)\} \overset{def}{=} \int_{-\infty}^{\infty} g(x) \cdot f_X(x) \, dx
\]

is called the expectation or expected value of the random variable \( g(X) \). Note: when \( g(X) \) is a discrete, this is just
\[
E\{g(X)\} = \sum_{x=-\infty}^{\infty} g(x)P(X = x).
\]

**Expected value of \( X \):**

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

**Second moment of \( X \):**

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

**Variance of \( X \):**

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

The expectation is linear in the following sense:
\[
E\{ag(X) + bh(X) + c\} = aE\{g(X)\} + bE\{h(X)\} + c.
\]

Therefore,
\[
E\{(X - E(X))^2\} = E\{X^2 - 2XE(X) + (E(X))^2\}
\]
\[
= E\{X^2\} - 2E(X)E(X) + (E(X))^2
\]
\[
= E\{X^2\} - (E(X))^2. \quad (2.1)
\]
Example 2.10. Throw a die, and let $X$ be the number that turns up. The expected value of $X$ is

$$
E(X) = \sum_{x=1}^{6} xP(X = x)
$$

$$
= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6}
$$

$$
= \frac{21}{6}
$$

$$
= 3.5.
$$

Note that $E(X)$ is not necessarily a value that $X$ can assume.

The second moment of $X$ is

$$
E(X^2) = \sum_{x=1}^{6} x^2P(X = x)
$$

$$
= 1 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + \cdots + 36 \cdot \frac{1}{6}
$$

$$
= \frac{91}{6}
$$

$$
= 15\frac{1}{6}.
$$

The variance of $X$ can then be calculated using Eq. (2.1),

$$
\text{Var}(X) = E(X^2) - (E(X))^2
$$

$$
= \frac{91}{6} - \frac{49}{4}
$$

$$
= \frac{182 - 147}{12}
$$

$$
= \frac{35}{12}.
$$

Example 2.11. Random variable $X$ is uniformly distributed on $[0,6]$ (Fig. 2.9).

![Figure 2.9. The pdf of $X$ in Example 2.11.](image_url)
The expected value of $X$ is
\[ E(X) = \int_0^6 \frac{1}{6} x \, dx = \left. \frac{x^2}{12} \right|_0^6 = 3. \]

The second moment of $X$ is
\[ E(X^2) = \int_0^6 \frac{1}{6} x^2 \, dx = \left. \frac{x^3}{18} \right|_0^6 = 12. \]

The variance of $X$ is
\[ \text{Var}(X) = 12 - 3^2 = 3. \]

### 2.1.5 Two Random Variables

**Definition 2.7.** The joint cumulative distribution function for two random variables, $X$ and $Y$, is defined as
\[ F_{X,Y}(x, y) = P(X \leq x \text{ and } Y \leq y). \]

The joint probability density function for $X$ and $Y$ is
\[ f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}. \]

We have:
\[ F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(\alpha, \beta) \, d\beta \, d\alpha. \]

The individual pdf’s $f_X(x)$ and $f_Y(y)$ are then called marginal pdf’s to distinguish from the joint pdf $f_{X,Y}(x, y)$. How can we get $f_X(x)$ and $f_Y(y)$ from $f_{X,Y}(x, y)$? We note that
\[ P(X \leq x) = P(X \leq x, Y < \infty) = F_{X,Y}(x, \infty) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(\alpha, \beta) \, d\beta \, d\alpha. \] (2.2)

On the other hand, in the previous section, we saw that
\[ P(X \leq x) = \int_{-\infty}^{x} f_X(\alpha) \, d\alpha. \] (2.3)

Identifying the integrands in Eq. (2.2) and Eq. (2.3), we see that
\[ f_X(\alpha) = \int_{-\infty}^{\infty} f_{X,Y}(\alpha, \beta) \, d\beta \]
\[ \text{or } f_X(x) = \int_{-\infty}^{x} f_{X,Y}(x, y) \, dy. \] (2.4)
Similarly,

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

**Example 2.12.** Let

\[ f_{X,Y} = \begin{cases} A, & 0 \leq y \leq x \leq 1, \\ 0, & \text{else.} \end{cases} \]

1. Find \( A \).

2. Find \( f_X(\frac{1}{2}) \).

**Solution**

![Figure 2.10](image)

**Figure 2.10.** The support of the function \( f_{X,Y}(x, y) \) of Example 2.12 is the shaded triangle. Within the triangle, \( f_{X,Y}(x, y) = A \).

1. To find \( A \),

\[
1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx \, dy \\
= A \cdot (\text{area of the triangle}) \\
= A \cdot \frac{1}{2} \\
\Rightarrow A = 2.
\]

2. We can find \( f_X(\frac{1}{2}) \) by using Eq. (2.4),

\[
f_X\left(\frac{1}{2}\right) = \int_{-\infty}^{\infty} f_{X,Y}\left(\frac{1}{2}, y\right) \, dy \\
= \int_{0}^{\frac{1}{2}} 2 \cdot dy \\
= 1.
\]

I.e., in order to find the marginal pdf of \( X \), we integrate \( y \) out by computing the line integral of \( f_{X,Y} \) along the dashed segment in Fig 2.10. More generally,

\[
f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx = \begin{cases} \int_{0}^{x} 2 \, dy = 2x, & 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}
\]
Definition 2.8. The random variables $X$ and $Y$ are independent if

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

This definition of independence implies that any event involving $X$ is independent of any event involving $Y$.

Conditional density:

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

If $X$ and $Y$ are independent then

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

i.e., the knowledge that $Y = y$ does not provide any information about $X$.

Expected value:

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

Note that this definition is consistent with our previous definition of the expected value of a single random variable:

$$E(g(X)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x) f_{X,Y}(x, y) \, dx \, dy = \int_{-\infty}^{\infty} g(x) \left( \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy \right) \, dx = \int_{-\infty}^{\infty} g(x) f_X(x) \, dx.$$
Correlation of $X$ and $Y$ is $E(XY)$.

Covariance of $X$ and $Y$ is

$$
\lambda_{XY} = \text{Cov}(X, Y) \overset{def}{=} E \{ (X - E(X))(Y - E(Y)) \} = E(XY) - E(X)E(Y).
$$

**Definition 2.9.** $X$ and $Y$ are uncorrelated if $\lambda_{XY} = 0$.

The following are some remarks about these terms.

1. $X$ and $Y$ are uncorrelated if and only if $E(XY) = E(X)E(Y)$.

2. If $X$ and $Y$ are independent, then they are uncorrelated, because in this case

$$
E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf_{X,Y}(x,y)dx\,dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xfy_X(x)f_Y(y)dx\,dy
= \int_{-\infty}^{\infty} xf_X(x)dx \int_{-\infty}^{\infty} yf_Y(y)dy = E(X)E(Y).
$$

3. However, if $X$ and $Y$ are uncorrelated, it does not imply that they are independent. (But if $X$ and $Y$ are, e.g., Gaussian random variables, the notions of independence and uncorrelatedness are equivalent.)

4. $\text{Var}(X) = \text{Cov}(X, X)$.

5. The correlation coefficient is:

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

It is possible to show that $-1 \leq \rho_{XY} \leq 1$, and that

- $\rho_{XY} = 0 \Rightarrow X$ and $Y$ are uncorrelated.
- $\rho_{XY} = 1 \Rightarrow X = aY + b$ with $a > 0$.
- $\rho_{XY} = -1 \Rightarrow X = aY + b$ with $a < 0$.

Why do we need the notion of uncorrelatedness in addition to independence? This is because:

- it also indicates to what extent $X$ and $Y$ are related (and in some cases, is equivalent to independence);
- to determine whether $X$ and $Y$ are independent, we need to know $f_{XY}(x,y)$; to determine whether $X$ and $Y$ are uncorrelated, we only need to know the first and second moments, which are easier to estimate.
2.1.6 Random Sequences

**Definition 2.10.** A DT random (or stochastic) process (or DT random signal, or random sequence)

\[ X(n), \quad -\infty < n < \infty, \]

is a sequence of random variables

\[ \cdots, X(-1), X(0), X(1), \cdots \]

(defined on the same probability space). Alternative notation: \( X_n \),

\[ \cdots, X_{-1}, X_0, X_1, \cdots \]

**Example 2.13.** Suppose we flip a coin every second, and assign

\[ X(n) = \begin{cases} 
1 & \text{if the } n\text{-th flip is heads} \\
-1 & \text{if the } n\text{-th flip is tails.}
\end{cases} \]

This sequence of binary random variables \( X(n) \) is a random process.

The observations associated with physical processes are often most appropriately modeled as random. It is typically the case, however, that their probability distribution is unknown and has to be deduced from the data.

For example, when we flip a coin, we do not know a priori that \( P(\text{heads}) = \frac{1}{2} \): if the coin is unfair, then we may have \( P(\text{heads}) \neq \frac{1}{2} \).

2.1.7 Estimating Distributions

**Example 2.14.** Repeatedly and independently flip a coin, and let

\[ X(n) = \begin{cases} 
1 & \text{if the } n\text{-th observation is heads} \\
-1 & \text{if the } n\text{-th observation is tails.}
\end{cases} \]

\[ p \quad 1-p \]

\[ -1 \quad 1 \quad x \]

\[ P(X(n) = x) \]

\[ p \quad 1-p \]

**Figure 2.12.** Probability mass function of \( X(n) \) for the coin-tossing experiment.

Suppose we do not know \( p \). How do we learn it from observing \( X(n) \)?
Solution. We construct the empirical distribution $\hat{P}_N(x)$ of $X(n)$, by tallying the number of occurrences of heads and the number of occurrences of tails (Fig. 2.13).

![Figure 2.13. Estimated probability distribution of the coin-tossing experiment.](image)

How good is this estimate? Let

$$H(n) = \begin{cases} 1, & \text{if } X(n) = 1, \\ 0, & \text{else.} \end{cases}$$

Then

$$E(H(n)) = p \cdot 1 + (1 - p) \cdot 0 = p,$$

$$E((H(n))^2) = p, \text{ and therefore}$$

$$\text{Var}(H(n)) = p - p^2.$$

Note that

$$\hat{p}_N = \frac{\sum_{n=1}^{N} H(n)}{N},$$

and that it is a random variable. Its expected value is

$$E(\hat{p}_N) = E\left\{ \frac{1}{N} \sum_{n=1}^{N} H(n) \right\}$$

$$= \frac{1}{N} \sum_{n=1}^{N} E(H(n))$$

$$= p.$$

The expected value of our estimate $E(\hat{p}_N)$ is equal to the quantity that we are trying to estimate. Estimators that have this property are called unbiased. Let us now calculate $\text{Var}(\hat{p}_N)$. First, consider two properties of variance.
1. If we scale the random variable $W$ by a constant $\alpha$, the resulting variance will be scaled by $\alpha^2$:

$$\text{Var}(\alpha W) = E \left[ (\alpha W - E(\alpha W))^2 \right]$$

$$= E \left[ \alpha^2 (W - E(W))^2 \right]$$

$$= \alpha^2 E \left[ (W - E(W))^2 \right]$$

$$= \alpha^2 \text{Var}(W).$$

2. What is the variance of a sum of random variables? It depends. For example,

(a) Suppose $W_1 = W_2 = \cdots = W_N = W$, then

$$\text{Var} \left( \sum_{n=1}^{N} W_n \right) = \text{Var}(NW_n) = N^2 \text{Var}(W).$$

(b) Suppose $W_1, W_2, \cdots, W_N$ are independent random variables, then

$$\text{Var} \left( \sum_{n=1}^{N} W_n \right) = E \left\{ \sum_{n=1}^{N} (W_n - E(W_n))^2 \right\}$$

$$= E \left\{ \sum_{n=1}^{N} (W_n - E(W_n))^2 \right\}$$

$$+ 2 \sum_{n \neq m} E(W_n - E(W_n))(W_m - E(W_m))$$

$$= \sum_{n=1}^{N} E \left\{ (W_n - E(W_n))^2 \right\}$$

$$+ 2 \sum_{n \neq m} E(W_n - E(W_n))E(W_m - E(W_m))$$

$$= \sum_{n=1}^{N} \text{Var}(W_n).$$

Thus, if we only made one experiment but counted its result $N$ times,

$$\text{Var} \left( \hat{p}_N \right) = \text{Var} \left\{ \frac{1}{N} \sum_{n=1}^{N} H(n) \right\}$$

$$= \frac{1}{N^2} \text{Var} \left\{ \sum_{n=1}^{N} H(n) \right\}$$

$$= \text{Var}(H(n)).$$
However, if the $N$ experiments are independent, then

$$
\text{Var}(\hat{p}_N) = \frac{1}{N^2} \text{Var} \left( \sum_{n=1}^{N} H(n) \right)
$$

$$
= \frac{1}{N^2} \sum_{n=1}^{N} \text{Var}(H(n))
$$

$$
= \frac{1}{N} \text{Var}(H(n))
$$

$$
= \frac{p - p^2}{N}.
$$

Therefore, $\text{Var}(\hat{p}_N) \to 0$ as $N \to 0$. Estimators whose variance approaches zero as the sample size grows to infinity, are called consistent estimators. We have therefore just shown that $\hat{p}_N$ is a consistent estimate of $p$. In other words, as the number of experiments increases, our estimate becomes more reliable. Since the estimate itself is a random variable, the variance of the estimate gives us a measure of its reliability. Roughly speaking, a larger variance of the estimate implies a larger spread of distribution. Hence, we have a larger chance of ending up with an estimate that is far from the correct value of $p$. On the other hand, if the variance of the estimate is small, there is a higher chance of the estimate being in the vicinity of the correct value of $p$, as shown in Fig. 2.14. (Note that this figure is not meant to be a precise characterization of this particular estimator, since, for example, it is a discrete random variable and therefore its PDF is actually a sequence of impulses.)

![Distribution of $\hat{p}_N$ for small $N$.](image1)

![Distribution of $\hat{p}_N$ for large $N$.](image2)

Figure 2.14. For larger values of $N$, the variance of $\hat{p}_N$ is smaller, which means that the estimate is more likely to be close to the correct value of $p$.

**Example 2.15.** $X(1), \cdots, X(N)$ are independent, identically distributed (iid) continuous random variables with unknown pdf $f(x)$. How to estimate $f(x)$?

One possible method is:

1. Partition the $x$-axis into $L$ intervals: $[x_0, x_1], (x_1, x_2), \cdots, (x_{L-1}, x_L]$. 

2. Estimate

\[ p^{(i)} = P(x_{i-1} \leq X(n) \leq x_i) = \int_{x_{i-1}}^{x_i} f(x) \, dx \]

as follows:

\[ \hat{p}_N^{(i)} = \frac{\text{number of outcomes that fall into the } i\text{-th interval}}{N}. \]

3. Note that, if \( f(x) \) is approximately constant on the \( i\)-th interval \([x_{i-1}, x_i]\), then

\[ \int_{x_{i-1}}^{x_i} f(x) \, dx \approx f(x_i - x_{i-1}). \]

Hence, we can estimate \( f(x) \) on the \( i\)-th interval as

\[ f(x) \approx \frac{\hat{p}_N^{(i)}}{x_i - x_{i-1}}. \]

(Unless something is known a priori about how fast \( f(x) \) changes, this method will not necessarily be successful in estimating \( f(x) \).)

Let

\[ H_i(n) = \begin{cases} 1, & \text{if } x_{i-1} \leq X(n) \leq x_i, \\ 0, & \text{else.} \end{cases} \]

Then

\[ E \left( H_i(n) \right) = p^{(i)} \]
\[ E \left( \hat{p}_N^{(i)} \right) = E \left( \frac{1}{N} \sum_{n=1}^{N} H_i(n) \right) = p^{(i)}, \]

which means that our estimate of \( p^{(i)} \) is unbiased. We also have:

\[ \text{Var} \left( \hat{p}_N^{(i)} \right) = \frac{p^{(i)} - (p^{(i)})^2}{N} \]
\[ \text{Var} \left( \frac{\hat{p}_N^{(i)}}{x_i - x_{i-1}} \right) = \frac{p^{(i)} - (p^{(i)})^2}{N(x_i - x_{i-1})^2}. \]
Figure 2.16. The two upper panels illustrate the trade-off between the number of bins and the size of a bin: if the bins are too small, the estimate can be very noisy; if there are too few bins, the estimate may be too coarse. The bottom panel illustrates the influence of the number $N$ of experiments: for the same bin size, increasing $N$ leads to a less noisy estimate.

Therefore, if the bin size $(x_i - x_{i-1})$ is small, we need a large number $N$ of experiments to get a reliable estimate. However, if the bin size is too large, then $f(x)$ may not be approximately constant within each interval, resulting in an estimate which does not track the actual $f(x)$ very well. This trade-off is illustrated in Fig. 2.16.

Sometimes, it is not necessary to estimate the whole pdf. Often, there are practical situations when there are only one or several parameters of interest.

Example 2.16. Suppose we are interested in measuring some quantity $A$, but what we
actually measure is:

\[
Y(1) = A + X(1)
\]

what we actually observe

\[
Y(2) = A + X(2), \text{ second measurement, independently obtained.}
\]

\[
\vdots
\]

\[
Y(N) = A + X(N)
\]

Suppose it is known that the pdf of \(X(n)\) is the same for all \(n\), and \(E(X(n)) = 0\), but the specific form of the pdf is unknown. How do we estimate \(A\)?

**Solution.** One possibility is the following estimate:

\[
\hat{A}_N = \frac{1}{N} \sum_{n=1}^{N} Y(n).
\]

Then,

\[
E\left(\hat{A}_N\right) = \frac{1}{N} \sum_{n=1}^{N} E(Y(n)) = \frac{1}{N} \sum_{n=1}^{N} E(A + X(n)) = A - \text{unbiased.}
\]

\[
Var\left(\hat{A}_N\right) = \frac{1}{N^2} \cdot N \cdot Var(X(n)) = \frac{Var(X(n))}{N} \rightarrow 0 \text{ as } N \rightarrow \infty, \quad \text{— consistent.}
\]

The variance approaches zero as the number of measurements increases to infinity.

**Central limit theorem.** Suppose

\[
Z_n = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} X(n),
\]

where \(X(n)\) are iid with mean zero and variance \(\sigma^2\). Then, as \(N \rightarrow \infty\), the CDF of \(Z_n \rightarrow\) Gaussian CDF with mean zero and variance \(\sigma^2\). The pdf of a Gaussian (normal) random variable \(R\) with mean \(m\) and variance \(\sigma^2\) is

\[
f_R(r) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(r-m)^2}{2\sigma^2}}.
\]

- For large \(N\), the distribution of our estimate is approximately Gaussian.

- As \(N\) gets larger, it becomes more and more probable that \(\hat{A}_N\), which we compute from our observations, is near \(A\).
Sec. 2.1. Introduction to Random Sequences, Detection, and Estimation

\[f_{\hat{A}_N}(x)\]

\[\sigma_{\hat{A}_N} = \frac{\sigma_X}{\sqrt{N}}\]

\[A\]

\[x\]

Figure 2.17. The pdf of the estimate \(\hat{A}_N\) of Example 2.16.

**Example 2.17.** Estimate the variance of iid random variables \(X(1), \ldots, X(N)\).

**Solution.**

1. If the mean \(m\) is known, try

\[\hat{\lambda} = \frac{\sum_{n=1}^{N} (X(n) - m)^2}{N}.\]

Then

\[E\left(\hat{\lambda}\right) = \frac{1}{N} E\{ (X(n) - m)^2 \} = Var(X(n)) \Rightarrow \text{unbiased}.\]

2. If the mean \(m\) is unknown,

(a) estimate the mean as in Example 2.16,

\[\hat{m} = \frac{1}{N} \sum_{n=1}^{N} X(n).\]

(b) try

\[\hat{\lambda} = \frac{1}{N} \sum_{n=1}^{N} (X(n) - \hat{m})^2\]

\[= \frac{1}{N} \sum_{n=1}^{N} ((X(n) - m) - (\hat{m} - m))^2\]

\[= \frac{1}{N} \sum_{n=1}^{N} (X(n) - m)^2 - 2(\hat{m} - m) \frac{1}{N} \sum_{n=1}^{N} (X(n) - m)\]

\[\left\{ \frac{\hat{m} - m}{2(\hat{m} - m)^2} \right\}\]
\[
\begin{align*}
&+ \frac{1}{N} \sum_{n=1}^{N} (\hat{m} - m)^2 \\
&\quad \left( \hat{m} - m \right)^2 \\
&= \frac{1}{N} \sum_{n=1}^{N} (X(n) - m)^2 - (\hat{m} - m)^2.
\end{align*}
\]

\[E \left( \hat{\lambda} \right) = Var(X(n)) - Var(\hat{m})\]
\[= Var(X(n)) - \frac{Var(X(n))}{N} = \frac{N-1}{N} Var(X(n)) \quad \text{biased!}\]

\[\hat{\lambda}' = \frac{1}{N-1} \sum_{n=1}^{N} (X(n) - \hat{m})^2 \quad \text{is unbiased, because}\]

\[\hat{\lambda}' = \frac{N}{N-1} \hat{\lambda} \Rightarrow E \left( \hat{\lambda}' \right) = \frac{N}{N-1} E \left( \hat{\lambda} \right) = Var(X(n)).\]

### 2.1.8 Sampling From A Distribution

Suppose you have hypothesized or built a probabilistic model—i.e., you have assumed or estimated a probability distribution of your observation (e.g., using techniques from the previous section). It is often very important to be able to synthesize samples from this probability distribution. E.g., texture synthesis is needed in rendering, printing, and other computer vision and imaging applications. One possible approach to texture synthesis is:

1. Model texture as a 2-D random process.
2. Synthesize an image or a texture patch by sampling\(^4\) from the probabilistic model.

While thorough consideration of the texture synthesis problem is beyond the scope of this course, we now consider the problem of sampling from a distribution.

**Example 2.18.** Suppose we have a cdf \(F_X(x)\). How do we generate samples of a random variable with this cdf?

**Solution.** Suppose we know how to generate samples of a uniform variable \(Y\) whose pdf is shown in Fig. 2.18.

\(^4\)Note: the word *sampling* used here means *drawing a random sample from a probability distribution*, which is completely different from the meaning it had in the earlier sections which dealt with converting CT signals to DT and where *sampling* meant *discretization*. 
What is the cdf of $W = F_X^{-1}(Y)$?

$F_W(w) = P(W \leq w) = P(F_X^{-1}(Y) \leq w)$

because $F_X$ is monotone

$P(Y \leq F_X(w)) = F_X(w)$

So, the cdf of $W$ is $F_X$. Use $F_X^{-1}$ to transform the uniform distribution into $F_X$.

2.1.9 Filtering A Random Process

What happens to a random process when it is put through an LTI system? In general, this is a very difficult question; however, we can say what will happen to its first and second order statistics.

**Definition 2.11.** If $X(n)$, $Y(n)$ are real-valued random sequences, the autocorrelation function of $X(n)$ is:

$$r_{XX}(m, n) = E\{X(m)X(n)\}.$$  

(How strongly are the points of $X$ at $m$ and $n$ correlated?)

The cross-correlation function of $X(m)$ and $Y(n)$ is:

$$c_{XY}(m, n) = E\{X(m)Y(n)\}.$$  

(How strongly are $X$ and $Y$ correlated?)
A note on complex-valued random variables and processes:

\[ \text{Cov}(X,Y) = E\{(X - m_X)(Y - m_Y)\} \]
\[ \text{Correlation} = E(XY^*) \]
\[ r_{XX}(m,n) = E\{X(m)X^*(n)\} \]
\[ c_{XY}(m,n) = E\{X(m)Y^*(n)\} \]

**Definition 2.12.** A random sequence \( X(n) \) is “wide-sense stationary” (WSS) if

1. \( E\{X(n)\} \) is constant for all \( n \).
2. \( r_{XX}(m,n) = r_{XX}(0,n-m) \) for any \( n, m \),
   
   \( \text{abuse of notation} \Rightarrow r_{XX}(n-m) \).

If both \( X \) and \( Y \) are WSS, then

\[ c_{XY}(k,n) = c_{XY}(n-k) \].

What will happen to the mean and autocorrelation of a random process when it goes through an LTI system? Consider the system illustrated in Fig. 2.20.

**Figure 2.20.** System with a WSS input.

Given \( \mu_x, r_{XX}(n) \), and \( h(n) \), can we find \( E(Y(n)) \) and \( r_{YY} \)?
1. 

\[ E(Y(n)) = E \left\{ \sum_{m=\infty}^{\infty} h(n-m)X(m) \right\} \]

linearity of \( E \) 

\[ = \sum_{m=\infty}^{\infty} h(n-m)E(X(m)) \]

since \( X(n) \) is WSS 

\[ = \sum_{m=\infty}^{\infty} h(n-m)\mu_x \]

\[ = \mu_x \sum_{m=\infty}^{\infty} h(m) \]

2. 

\[ c_{XY}(m, n) = E \{ X(m)Y(n) \} \]

\[ = E \left\{ X(m) \sum_{k=\infty}^{\infty} h(n-k)X(k) \right\} \]

\[ = \sum_{k=\infty}^{\infty} h(n-k)E \{ X(m)X(k) \} \]

\[ = \sum_{k=\infty}^{\infty} h(n-k)r_{XX}(k-m) \]

\[ \Rightarrow \sum_{l=\infty}^{\infty} h((n-m)-l)r_{XX}(l) \]

\[ = (h \ast r_{XX})(n-m) \] — only depends on \( (n-m) \), 

\[ c_{XY}(n) = h \ast r_{XX}(n) \]
3. \[ r_{YY}(m, n) = E \{ Y(m)Y(n) \} = \left\{ Y(m) \sum_{k=-\infty}^{\infty} h(n-k)X(k) \right\} \]
\[ = \sum_{k=-\infty}^{\infty} h(n-k)E \{ X(k)Y(m) \} \]
\[ = \sum_{k=-\infty}^{\infty} h(n-k)c_{XY}(m-k) \]
\[ = \sum_{k=-\infty}^{\infty} h(n-k)c_{YX}(k-m) \]
\[ \Rightarrow \sum_{l=m-k}^{n-l} \sum_{k=-\infty}^{\infty} h((n-m)-l)c_{YX}(l) \]
\[ = (h * c_{YX})(n-m) \]
\[ = (h_- * c_{XY})(m-n), \text{ where } h_-(n) = h(-n). \]

Again, \[ r_{YY}(m, n) = r_{YY}(n, m) = r_{YY}(n-m) \]
Also, we found that \( E(Y(n)) \) is constant. \( \Rightarrow Y \) is WSS.

So, if the input to an LTI system is WSS, then the output is also WSS, with
\[ r_{YY}(n) = (h_- * c_{XY})(n) \]
\[ = (h_- * h * r_{XY})(n) \]
\[ = (h * c_{YX})(n). \]

**Figure 2.21.** Decomposition of the steps in calculating \( r_{YY}(n) \) from \( r_{XX}(n) \).

**Example 2.19.** Let \( Y(n) = X(n) + X(n-1) \) in the system illustrated in Fig. 2.22. Find \( r_{YY}(n) \). I.e., how correlated is the output with itself shifted by some lag value \( n \)?
$X(n) \begin{cases} \text{iid Gaussian,} \\ \text{zero mean, and} \\ \text{variance } \sigma_X^2. \end{cases} \rightarrow h(n) \rightarrow Y(n)$

Figure 2.22. System in Example (2.19).

**Solution:**

\[
r_{XX}(m) = E\{X(n)X(n + m)\}
= \begin{cases} \sigma_X^2, & m = 0 \\ 0, & m \neq 0 \text{ since } X \text{ is independent.} \end{cases}
= \sigma_X^2 \delta(m)
\]

A process such as this, for which $X(n)$ is uncorrelated with $X(m)$ unless $n = m$, is called a white noise process.

\[
c_{XY}(m) = h \ast r_{XX}(m)
= \begin{array}{c}
\sigma_X^2 \\
0
\end{array}
\begin{array}{c}
\sigma_X^2 \\
1
\end{array}

h(n) = \begin{array}{c}
1 \\
0
\end{array}
\begin{array}{c}
1 \\
1
\end{array}
\]

Alternatively,

\[
c_{XY}(m) = E\{X(n)Y(n + m)\}
= E\{X(n)(X(n + m) + X(n + m - 1))\}
= \begin{cases} \sigma_X^2, & \text{if } m = 0 \text{ or } m = 1. \\ 0, & \text{else, because of independence.} \end{cases}
\]

\[
r_{YY}(m) = h_- \ast c_{XY}(m) = \begin{array}{c}
\sigma_X^2 \\
2\sigma_X^2
\end{array}
\begin{array}{c}
\sigma_X^2 \\
\sigma_X^2
\end{array}
\begin{array}{c}
1 \\
-1
\end{array}
\begin{array}{c}
1 \\
0
\end{array}
\begin{array}{c}
\sigma_X^2 \\
\sigma_X^2
\end{array}
\begin{array}{c}
0 \\
1
\end{array}
\]

\[
= \begin{array}{c}
\sigma_X^2 \\
\sigma_X^2
\end{array}
\begin{array}{c}
-1 \\
0
\end{array}
\begin{array}{c}
\sigma_X^2 \\
\sigma_X^2
\end{array}
\begin{array}{c}
0 \\
1
\end{array}
\]

\[
= \begin{array}{c}
\sigma_X^2 \\
\sigma_X^2
\end{array}
\begin{array}{c}
-1 \\
0
\end{array}
\begin{array}{c}
\sigma_X^2 \\
\sigma_X^2
\end{array}
\begin{array}{c}
0 \\
1
\end{array}
\]
Figure 2.23. Scatter plots for Example 2.19.
Estimating Correlation Functions  For wide-sense stationary processes,

autocorrelation \( r_{XX}(n) = E[X(m)X(m+n)] \).

cross-correlation \( r_{XY}(n) = E[X(m)Y(m+n)] \).

How to estimate \( r_{XX}(n) \)? Again, compute sample averages:

An estimate of \( r_{XX}(2) = \frac{1}{K} \left( X(0)X(2) + X(1)X(3) + X(2)X(4) \ldots + X(K-1)X(K+1) \right) \)

More generally, if \( N \) points of \( X(n) \) are available, form

\[
r'_{XX} = \frac{1}{N - |m|} \sum_{n=0}^{N-|m|-1} X(n)X(n + |m|), \quad -(N-1) \leq m \leq N-1.
\]

What is this useful for? Well, there are a lot of different applications, one of which is radar detection. The key to this application is the following property of the autocorrelation function:

\( |r_{XX}(n)| \leq r_{XX}(0) \).

And, in fact, for many random processes which model natural phenomena, the random variables become less and less correlated as they become more separated in time:

\[
\lim_{n \to \infty} r_{XX}(n) = (E(X(n)))^2.
\]

So, a typical autocorrelation function might look like the one in Fig. 2.24.
(A random variable is very much correlated with itself, it may be somewhat correlated with its neighbors, and is basically uncorrelated from random variables which are far in the future or in the past.)

An estimate of the autocorrelation function may look like Fig. 2.25.

![Figure 2.25. An estimate of the autocorrelation function.](image)

In radar or sonar, you transmit a signal $X(n)$, and then receive its reflection from an object, $Y(n)$. By measuring the delay between $X(n)$ and $Y(n)$, you can estimate the distance to the object.

- $X(n)$, transmitted signal.
- $Y(n)$, received signal (attenuated, noisy, delayed version of $X(n)$).

It is usually impossible to tell the delay just by looking at $X(n)$ and $Y(n)$. However, by comparing an estimate of cross-correlation $c_{XY}$ such as the one shown in Fig. 2.26, to the estimate of $r_{XX}$, it becomes easy to estimate the delay.
2.1.10 Noise Removal and Detection

For the past few sections, we have been considering random sequences — that is, sequences of random variables. Exactly predicting the behavior of such a sequence is impossible, but, as you have learned in the past three sections, in some cases it is possible to estimate certain parameters — like the mean or the autocorrelation functions — from a realization of a random sequence. The main idea here was that if you have a sequence of identical random variables, preferably independent, you can estimate certain parameters by computing time averages. In this section, we continue the discussion of estimation, and consider examples of detecting or estimating a signal corrupted by additive noise.

Problem 1: Binary hypothesis testing in DT additive white Gaussian noise.

Hypothesis $H_0$: $Y = W$.

Hypothesis $H_1$: $Y = x + W$.

Where $W$ is a vector of uncorrelated, zero-mean Gaussian random variables, with variance $\sigma^2$. For example, a situation like Fig. 2.27 can be encountered in a communication system. One possible objective: having observed $Y = y$, choose $H_0$ or $H_1$ so as to maximize the likelihood of the observation:

$$P_{Y|H_1} (y|H_1) \overset{H_1}{\gtrless} P_{Y|H_0} (y|H_0)$$

Figure 2.26. An estimate of the cross-correlation function. In this figure, the delay is $n = 50$. 
Figure 2.27. Model of a communication system.

\[
\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y(n) - x(n))^2}{2\sigma^2}} \quad H_1 \geq \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y(n))^2}{2\sigma^2}} \quad H_0
\]

\[
\downarrow \log \quad \downarrow \log
\]

\[
-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y(n) - x(n))^2 \quad H_1 \geq \quad -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y(n))^2 \quad H_0
\]

\[
-\sum_{n=1}^{N} \left[ (y(n))^2 - 2y(n)x(n) + (x(n))^2 \right] \quad H_1 \geq \quad -\sum_{n=1}^{N} (y(n))^2 \quad H_0
\]

\[
2 \sum_{n=1}^{N} y(n)x(n) \quad H_1 \geq \quad -\sum_{n=1}^{N} (x(n))^2 \quad H_0
\]

\[
2\langle y, x \rangle \quad H_1 \geq \quad \| x \|^2 \quad H_0
\]

\[
\frac{\langle y, x \rangle}{\| x \|^2} \quad H_1 \geq \quad \frac{1}{2} \quad H_0
\]
I.e.,

- calculate the projection of the data $y$ onto $x$.
- if the coefficient is greater than $\frac{1}{2}$, decide that $x$ is present; if not, decide that $x$ is not present.

Figure 2.28. Vector space interpretation of hypothesis testing.

This is quite intuitive. In order to isolate $x$, we need to represent our observations $Y$ in a basis where $x$ stands out, namely, a basis where $x$ is one of the basis vectors. If indeed $Y$ was equal to $x$ plus noise, we would expect its projection onto $x$ to be large, whereas otherwise, there is no reason to expect this projection to be large.

Let us now generalize this idea to a slightly more complicated situation. Here, we do not know the underlying signal $x$ exactly, but we know, for example, that it is a sinusoidal signal or a sum of sinusoids.

If we represent a noisy sinusoid in a Fourier basis by computing its DFT, we would see two large spikes corresponding to the frequency of the sinusoid, and then there will be noise. One simple method to get rid of noise would then be to:

1. set all the coefficients which are below a certain threshold to zero, so that we are left with just these two spikes;
2. reconstruct our signal from the remaining coefficients.

Now, let us analyze this procedure a little more closely.

Problem 2: Noise removal via representations in orthogonal bases.

Let

$$ Y = x + W, $$

where

$x$ is the sum of a few non-random sinusoids of unknown frequencies, with sufficiently large amplitudes.
\( \mathbf{W} \) is the vector of uncorrelated, zero-mean random variables with variance \( \sigma^2 \):

\[ \mathbf{W} = \begin{pmatrix} W(0) \\ W(1) \\ \vdots \\ W(N-1) \end{pmatrix} \]

Take DFT:

\[ \tilde{\mathbf{Y}} = A \mathbf{Y} = A \mathbf{x} + A \mathbf{W} = \tilde{\mathbf{x}} + \tilde{\mathbf{W}}, \]

where \( \tilde{\mathbf{Y}} \) is the DFT of \( \mathbf{Y} \) and \( A \) is the DFT matrix:

\[
A = \begin{pmatrix}
1 & e^{-j\frac{2\pi}{N}1} & e^{-j\frac{2\pi}{N}2} & \ldots & e^{-j\frac{2\pi}{N}(N-1)} \\
1 & e^{-j\frac{2\pi}{N}2} & e^{-j\frac{2\pi}{N}3} & \ldots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & e^{-j\frac{2\pi}{N}(N-1)} & e^{-j\frac{2\pi}{N}(N-2)} & \ldots & e^{-j\frac{2\pi}{N}(N-1)(N-1)}
\end{pmatrix}
\]

1. Suppose \( x(n) = \alpha \cos \left( \frac{2\pi k_0}{N} n \right), \ 0 \leq n \leq N - 1 \), then

\[
\tilde{x}(k) = \begin{cases} 
\frac{\alpha N}{2}, & k = k_0 \text{ or } k = N - k_0, \\
0, & \text{otherwise.}
\end{cases}
\]

2. What is \( \tilde{\mathbf{W}} \), the DFT of white noise \( \mathbf{W} \)?

\[
E \left[ \tilde{\mathbf{W}} \right] = E[A \mathbf{W}] = AE[\mathbf{W}] = 0.
\]

Where \( E \left[ \tilde{\mathbf{W}} \right] \) denotes

\[
\begin{pmatrix}
E \left( \tilde{W}(0) \right) \\
E \left( \tilde{W}(1) \right) \\
\vdots \\
E \left( \tilde{W}(N-1) \right)
\end{pmatrix},
\]

it is the mean of \( \tilde{\mathbf{W}} \).
The covariance of \( \tilde{W} \):

\[
\Lambda_{\tilde{W}} = E \left\{ (\tilde{W} - m_{\tilde{W}})(\tilde{W} - m_{\tilde{W}})^H \right\} = E \begin{pmatrix}
\tilde{W}(0) - m_{\tilde{W}(0)} \\
\tilde{W}(1) - m_{\tilde{W}(1)} \\
\vdots \\
\tilde{W}(N-1) - m_{\tilde{W}(N-1)}
\end{pmatrix} \begin{pmatrix}
(\tilde{W}(0) - m_{\tilde{W}(0)})^* \\
\cdots
\end{pmatrix}
\]

\[
= \begin{pmatrix}
\text{Var}(\tilde{W}(0)) & \cdots & \text{Cov}(\tilde{W}(0), \tilde{W}(N-1)) \\
\vdots & \ddots & \vdots \\
\text{Cov}(\tilde{W}(N-1), \tilde{W}(0)) & \cdots & \text{Var}(\tilde{W}(N-1))
\end{pmatrix}
\]

In our case,

\[
\Lambda_{\tilde{W}} = E [\tilde{W}\tilde{W}^H] = E [AW(AW)^H] = E [AWW^HA^H] = AE [WW^H] A^H = A \begin{pmatrix}
\sigma^2 & 0 \\
0 & \sigma^2
\end{pmatrix} A^H = AA^H \sigma^2 = N\sigma^2 A A^{-1},
\]

the IDFT matrix \( B = A^{-1} = \frac{1}{N} A^H \Rightarrow A^H = N A^{-1} \),

\[
= N\sigma^2 I.
\]

So, the DFT of white noise is white noise with zero mean and standard deviation \( \sqrt{N}\sigma \). As long as

\[
\sqrt{N}\sigma \ll \frac{\alpha N}{2},
\]

we can remove noise by thresholding the DFT coefficients:

- Calculate DFT of \( \mathbf{Y} \).
- Set all coefficients whose absolute value is less than, say \( 3\sqrt{N}\sigma \), to zero.
- Reconstruct.
Figure 2.29. Thresholding noise coefficients. The dashed line represents the threshold level.

Figure 2.30. Noise removal through thresholding DFT coefficients.
Figure 2.31. Noise removal for a sum of two sinusoids.

Note that the reason why we chose the Fourier basis is that we knew that our signals of interest were sinusoids. In other scenarios, you would choose a different basis. Note that our analysis would still hold for any orthogonal basis, because in our derivation of the fact that the transform of white noise is still white noise, we never used the fact that our basis vectors were Fourier vectors.