SC505

STOCHASTIC PROCESSES

Class Notes

Dept. of Electrical and Computer Engineering

Boston University
College of Engineering
8 St. Mary's Street

Boston, MA   02215

Fall 2004

Contents

List of Figures

List of Tables

Chapter 1

Introduction to Probability

What is probability theory? It is an *axiomatic* theory which *describes and predicts* the outcomes of inexact, *repeated* experiments. Note the emphases in the above definition. The basis of probabilistic analysis is to determine or estimate the probabilities that certain known events occur, and then to use the axioms of probability theory to combine this information to derive probabilities of other events of interest, and to predict the outcomes of certain experiments.

For example, consider any card game. The inexact experiment is the shuffling of a deck of cards, with the outcome being the order in which the cards appear. An estimate of the underlying probabilities would be that all orderings are equally likely; the underlying events would then be assigned a given probability.

Based on the underlying probability of the events, you may wish to compute the probability that, if you are playing alone against a dealer, you would win a hand of blackjack. Certain orderings of the cards lead to winning hands, and the probability of winning can be computed from the combined information on the orderings.

There are several interpretations of what we mean by the probability of an event occurring. The frequentist interpretation is that, if an experiment is repeated an infinite number of times, the fraction of experiments in which the event occurs is its probability. On the other hand, the subjectivist interpretation is that a probability represents an individual belief that a certain event will occur. This interpretation is most appropriate when experiments cannot be repeated, such as in economics and social situations. Independent of which interpretation is used, the same axiomatic theory is used for manipulating probabilities.

In this chapter, we review some of the key background concepts in probability theory.

1.1 Axioms of Probability

A probability space is a triple $(\Omega, \mathcal{F}, P)$ which is used to describe the outcomes of a random experiment. The set Ω is the set of all possible elementary experiment outcomes ω. The set $\mathcal{F}$ is a collection of subsets of Ω which is closed under countable unions and complementation. That is, if $A_i \in \mathcal{F}, i = 1, \ldots$, then $A_i \subset \Omega, \cup_{i=1}^{\infty} A_i \in \mathcal{F}, \bar{A}_i \in \mathcal{F}$. (The notation $\bar{A}$ is used to denote the complement of A, or $\bar{A} = \Omega - A$, where $B - A = \{x \mid x \in B x \notin A\}$.)   The set $\mathcal{F}$ is called a σ-field because of its closure under countable union, and is referred to as the set of events. An element $A \subset \mathcal{F}$ is called an event. Note that the above properties also imply that $\mathcal{F}$ is closed under countable intersections.

The measure P assigns a probability value in $[0, 1]$ to each event contained in $\mathcal{F}$; that is, it maps the set of events into the closed unit interval $[0, 1]$. Furthermore, the probability measure has some important properties, described below.

Two events A, B are said to be mutually exclusive if $A \cap B = \emptyset$, the empty set. The axioms which a probability measure must satisfy are:

1. $P(\Omega) = 1$.

2. $P(A) \geq 0$ for all $A \in \mathcal{F}$.

3. $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ if $A_i \cap A_j = \emptyset$ for all $i \neq j$. This property is called the countable additivity property of the probability measure.

Based on the above properties, probability measures can be shown to satisfy additional properties, such as:

1. $P(A) = 1 - P(\bar{A})$.

2. $P(\emptyset) = 0$.

3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

4. If $B \subset A$, then $P(B) \leq P(A)$.

Consider again the example of a shuffle of a deck of cards. The outcomes are the possible orderings. Events are combinations of outcomes; for example, an event may be the set of all orderings such that the ace of spades is the first card.

Why is the concept of event needed over and above the concept of outcome? For experiments where the set of outcomes is discrete and finite, such as a shuffle of a deck of cards, one can consider only the set of outcomes, as each outcome can be an individual event. However, there are may situations where we want to model the set of possible outcomes as continuous, rather than discrete; for instance, the experiment of picking a random number in the interval [0,1]. In such cases, it is often impossible to associate non-zero probabilities with individual outcomes; indeed, there are an uncountable number of outcomes, and none of the axioms of probability can be used to combine the probabilities of an uncountable number of outcomes. By the laws of probability, there are at most a finite number of mutually exclusive events which have probability of at least $1/n$. Thus, by defining the probability measure on events rather than on an uncountable number of outcomes, we can focus our definition on the significant outcomes of experiments, and also provide a meaningful way of combining probabilities.

Another important issue is that not every subset of Ω can be an event, because it is not possible to assign a probability to each subset in a manner which is consistent with the axioms of probability measures. For instance, consider the following construction of a subset of $[0,1]$: Let $A_s = \{x \in [0,1] | x - y \bmod 1$ is a rational number for all other $y \in A_s\}$. Construct the set $B = \{$ one element from each distinct $A_s\}$. Denote the probability measure P as the uniform measure, such that $P([a,b)) = b - a$ for $0 \leq a \leq b \leq 1$. Now, note the following properties of the constructed sets: $\cup_s A_s = [0,1], A_s \cap A_t = \emptyset$ if $A_s \neq A_t$.

Denote the translation of B as $B + r = \{y \mid y = x + r, x \in B\}$. Denote by $B_i = B + r_i$ for each rational number r_i. Note the following:

1. There are a countable number of B_i.

2. $B_i \cap B_j = \emptyset$ if $i \neq j$, because B contains one and only one element from each A_s. Note that, if the conclusion were not true, then there is are $x, y \in B, x \neq y$ such that $x + r_i = y + r_j$, which would imply that $x, y \in A_s$ for some s, contradicting the construction of B.

3. $\cup_{i=1}^{\infty} B_i = \cup_s A_s = \Omega$.

Now, consider our dilemma if B were an event, what probability would we assign to B? Clearly, by construction, $P(B_i) = P(B_j) = P(B)$. So, if $P(B) \neq 0$, then, since the B_i are mutually disjoint, $P(\Omega) = \infty$. If $P(B) = 0$, then $P(\Omega) = 0$ also! Thus, we have a set for which we cannot assign a probability which is compatible with the axioms of probability theory and the definition of the uniform probability measure.

Other useful properties of σ-fields are:

1. A σ-field $\mathcal{F}'$ is said to be a refinement of $\mathcal{F}$ (written as $\mathcal{F}' < \mathcal{F}$), if and only if, for any event $A \in \mathcal{F}'$, said event is also $A \in \mathcal{F}$.

2. Given a collections of sets $\{A_i\}, A_i \in \Omega$, there exists a smallest σ-field which contains the sets $\{A_i\}$, denoted by $\sigma(\{A_i\})$.

As a final note, in any probability space, there can be numerous events which have no probability of occurring. Thus, the difference between two events is often negligible; in such cases, we would like to define a notion of equivalence of events. This notion is stated as follows: two events $A, B \in \mathcal{F}$ are said to be equal with probability one if and only if $P(A \cup B - A \cap B) = 0$.

1.2 Conditional Probability and Independence of Events

Consider a probability space, and a pair of events $A, B \in \mathcal{F}$ such that $P(B) > 0$. We define the conditional probability of event A given that B has occurred as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.1}$$

Conditional probability functions have an interesting property: they are also probability measures, and a conditional probability space can be defined! In particular, let $P(\cdot|B)$ denote this probability measure. Then, $\{\Omega, \mathcal{F}, P(\cdot|B)\}$ is a probability space.

The total probability theorem is an important result. It can be stated as follows. Let $A_1, \ldots$ denote a countable set of pairwise mutually exclusive events with $P(A_i) > 0$ for $i = 1, \ldots$, and assume that $A_1 \cup A_2 \cup \ldots = \Omega$. Then, for any event $B \in \mathcal{F}$,

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots$$

Another important result in probability theory is Bayes' theorem, which can be combined with the total probability theorem to state:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \tag{1.2}$$

$$= \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{m} P(B|A_j)P(A_j)}. \tag{1.3}$$

Two events A, B are said to be independent if $P(A \cap B) = P(A)P(B)$. This implies that $P(B|A) = P(B), P(A|B) = P(A)$. The concept of independence can be extended to a finite sequence of sets $A_1, \ldots, A_m$, which are mutually independent if $P(A_1 \cap A_2 \cap \ldots \cap A_m) = P(A_1)P(A_2) \cdots P(A_m)$. Note that the above concept of mutual independence implies much more than pairwise independence; it is easy to construct examples of events which are pairwise independent, but not mutually independent. For example, consider the experiment of selecting an integer from 1 to 4. Consider the events $\{1, 2\}, \{1, 3\}, \{1, 4\}$; note that, if each number is equally likely, then the above events are pairwise independent but they are not mutually independent (since they all depend on the common outcome 1).

1.3 Random Variables

A random variable is similar to a function; indeed, the most common definition of a random variable is a function which assigns real values to the outcomes in Ω. In this manner, it is possible to characterize experiments with similar outcomes entirely in terms of the numerical values of their outcomes. For instance, an experiment involving tossing two unbiased coins is similar in probability to an experiment for rolling a 4-sided unbiased die, as both experiments are equally likely to create any of 4 outcomes. However, their underlying probability spaces would be different. Assigning numerical values to the outcomes of each experiment lets us recognize that the resulting probability spaces give rise to equivalent random variables.

Formally, a random variable in a probability space $(\Omega, \mathcal{F}, P)$ is a measurable function $X : \Omega \to \Re$, where $\Re$ is the space of real numbers. By a measurable function, we mean a function where the sets $\{\omega \mid X(\omega) < a\}$ are events in the original σ-field $\mathcal{F}$. Thus, not every function can generate a random variable; it must be such that inverse images of intervals are well-defined events in the original probability space. We define the Borel σ-field $\mathcal{B}$ to be generated by the sets $(-\infty, a)$, for all real numbers a. In terms of functions, a random variable is a measurable function from $(\Omega, \mathcal{F})$ into $(\Re, \mathcal{B})$.

Often, we wish to allow a random variable to take the values $+\infty$ or $-\infty$ for some specific events. We can allow this extension, provided that $P[\{\omega \mid X(\omega) = +\infty\} \cup \{\omega \mid X(\omega) = -\infty\}] = 0$.

A random variable X induces a probability measure P_X on $(\Re, \mathcal{B})$ using the function mapping. For any of the elementary events $(-\infty, a)$, this probability is given by

$$P_X((-\infty, a)) = P(\{\omega \mid X(\omega) < a\}).$$

We can extend this definition to arbitrary open intervals (b, a) as

$$P_X((b, a)) = P(\{\omega \mid b < X(\omega) < a\})$$

and, more generally, for any set $B \in \mathcal{B}$, we have

$$P_X(B) = P(\{X(\omega) \in B\}).$$

Indeed, with this induced probability, we can show that $(\Re, \mathcal{B}, P_X)$ is also a probability space. We call this space the sample space. In our study of stochastic processes, we will typically characterize random variables in terms of the properties of their sample spaces, rather than in terms of their underlying probability spaces. However, derivation of the probability measure of the sample space depends explicitly on the definition of the original experiment which gives rise to the random variable.

As an example, consider the experiment of tossing two unbiased coins. In the original space Ω, there are four outcomes: HH, HT, TH and TT, where H denotes a heads outcome and T denotes a tails outcome. We define a random variable X as follows:

$$X(\omega) = \begin{cases} -1 & \text{if } \omega \neq HH, TT \\ 1 & \text{otherwise.} \end{cases}$$

The sample space can be taken to be either $\Re$, or $\{-1, 1\}$; in cases where the number of possible sample values is discrete, the random variable is said to be a discrete random variable, and we simplify the sample space to be the range of values taken by the random variable. Note that both of the sample values, 1 and -1, are equally likely. The induced probability P_X is such that $P_X(1) = P_X(-1) = 0.5$.

Now, consider a second experiment, consisting of tossing a single unbiased coin, and define another random variable Y as

$$Y(\omega) = \begin{cases} -1 & \text{if } \omega = H \\ 1 & \text{otherwise.} \end{cases}$$

The sample space and induced probability of this random experiment and random variable Y are the same as those of the previous experiment and random variable X. Rather than treating these random variables a different, by using the sample space, we can treat them as identical random variables, as they will have identical distributions.

1.4   Characterization of Random Variables

Consider a probability space $(\Omega, \mathcal{F}, P)$, with a random variable X defined on it. We have the following definition:

Definition 1.1 (Probability Distribution Function)
The *probability distribution function* of the random variable X is defined as the function $P_X : \Re \to [0, 1]$ which satisfies:

$$P_X(a) \equiv P_X((-\infty, a]) = P(\{\omega \mid X(\omega) \leq a\}).$$

This function is also sometimes called the *cumulative distribution function*, since it is a "cumulative" measure of the probability of the random variable X falling in the interval $(-\infty, a]$. It is sometimes referred to as the "PDF" or "CDF" (all upper case) of a random variable. We will often use the notation $P(a)$ instead of $P_X(a)$ when it is clear which random variable we are referring to. In particular, for a generic argument, this is often written as $P_X(x)$ or just $P(x)$. (Note that there is a possibility for confusion between $P(\cdot)$ as used for the probability of an event vs. for the PDF. The difference should be clear based on whether the argument is a set or a point, respectively.)

Probability distribution functions have the following properties:

1. $P_X(\infty) = 1, P_X(-\infty) = 0$

2. $a \leq b$ implies that $P_X(a) \leq P_X(b)$

3. $\lim_{\epsilon \to 0^+} P_X(a + \epsilon) = P_X(a)$ (continuity from the right)

Note that the probability distribution function can be used to completely characterize the induced probability P_X on $(\Re, \mathcal{B})$, since it assigns a probability to each elementary set defining $\mathcal{B}$.

When the random variables are continuous (i.e. not discrete), it is often convenient to define a second function, the probability density function $p_X(a)$, as follows:

Definition 1.2 (Probability density function)
Assuming the function $P_X(a)$ is differentiable, the *probability density function* is defined as:

$$p_X(a) = \frac{d}{da} P_X(a),$$

with the constraints that

$$p_X(x) \geq 0 \int p_X(x) dx = 1.$$

This function is sometimes referred to as the "pdf" (lower case) of a random variable. A probability measure P_X defined on $(\Re, \mathcal{B})$ is *absolutely continuous* if, for any A such that $\int_\Re I_A(s) ds = 0$, we have also that $P_X(A) = 0$. For absolutely continuous probability measures, we have the following representation (known as the Radon-Nykodim theorem):

$$P_X(A) = \int_A p_X(s) ds,$$

where $p_X(s)$ is a non-negative measurable function corresponding to the probability density function. The probability density function can be interpreted in terms of the frequency of outcomes. If $p_X(a)$ is finite over an interval $(a, a + \epsilon]$, then, for very small ϵ, the probability that a sample value occurs in the above interval is approximately $p_X(a)\epsilon$.

Even if $P_X(a)$ is discontinuous and not differentiable, we can often define a probability density function in terms of generalized functions such as the unit impulse function (or delta function) $\delta(a)$. Recall, the impulse function is defined by the following properties:

$$\delta(a) = 0 \quad \text{if } a \neq 0$$

$$\int_b^c \delta(a) \, da = \begin{cases} 0 & \text{if } b \leq c < 0 \\ 1 & \text{if } b \leq 0 \leq c \\ 0 & \text{if } 0 < b \leq c \end{cases}$$

$$\int_{-\infty}^\infty \delta(a - s) g(s) \, ds = g(a) \quad \text{if } g \text{ is continuous at } a.$$

Probability distribution functions are discontinuous at points where a random variable can take a specific value with nonzero probability. For example,

$$p_X(x) = 0.5\delta(x + 1) + 0.5\delta(x - 1)$$

is the density of a random variable taking on the values $-1, 1$ each with equal probability.

Now let us define two types of probability measures. We say a probability measure P_X defined on $(\Re, \mathcal{B})$ is *singular* if there exists a set $A \in \Re$ such that $P_X(A) = 1$, and $\int_\Re I_A(s) ds = 0$, where $I_A(s)$ is the indicator function of the set A; that is,

$$I_A(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{otherwise.} \end{cases}$$

Probability measures on discrete-valued random variables (such as the outcomes of a coin toss) are singular. In these cases we sometimes describe the distribution at discrete points, in the same way discrete-time quantities are often represented in digital signal processing.

Definition 1.3 (Probability mass function)
For discrete-valued random variables, the probability distribution can be characterized using a *probability mass function*, where:

$$0 \leq p_X(x) \leq 1 \sum_i p_X(x_i) = 1.$$

This function is sometimes referred to as the "pmf" (lower case) of a random variable. Using the pmf, the probability of an event A is computed as

$$P_X(A) = \sum_{x_i \in A} p_X(x_i).$$

Given these definitions, every probability measure P_X defined on $(\Re, \mathcal{B})$ (or more generally on the vector space $(\Re^n, \mathcal{B}^n)$ for vectors of random variables) can be decomposed in canonical form (Lebesgue decomposition):

$$P_X = \alpha P_X^{(1)} + (1 - \alpha) P_X^{(2)},$$

where $P_X^{(1)}$ is absolutely continuous and $P_X^{(2)}$ is singular. Typically, the singular measure is represented as a sum of delta functions, and the absolutely continuous part is represented by the continuous-valued probability density function. We summarize these ideas in Table 1.1.

Example 1.1
Consider the following random variable X: With probability 1/2, it has value 0; the remaining probability 1/2 is spread uniformly in the interval $[0, 1]$. This random variable is continuous-valued, but the probability distribution function is not absolutely continuous. Indeed, if we were to write the density of this random variable, it would consist of

$$p_X(s) = 1/2\delta(s) + 1/2 I_{[0,1]}(s),$$

illustrating that the density is the sum of a singular part (corresponding to the non-zero probability at 0) and an absolutely continuous part.

For both singular and absolutely continuous cases, the relationship between probability distribution functions and probability density functions can be inverted as follows:

$$P_X(a) = \int_{-\infty}^{a} p_X(s)ds.$$

The probability density function can be used to characterize the induced probability as

$$P_X((a, b]) = P(\{X(\omega) \in (a, b]\}) = \int_{a}^{b} p_X(s)ds.$$

Using the probability density function also allows us to define certain operations and expectations of random variables. In order to avoid excess mathematics, let us define the concept of a measurable function of a random variable X to be a function for which the integral $\int_{-\infty}^{\infty} g(s)p_X(s)\,ds$ is well-defined. For any measurable function $g : \Re \to \Re$, we define its expected value as:

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(s)p_X(s)\,ds. = \int_{\Omega} g(X(\omega))P(d\omega)$$

Note that this integral may be infinite-valued if g is unbounded. For a discrete-valued random variable, it is typically more convenient to express the expected value in terms of the probability mass function

$$E\left[g(X)\right] = \sum_{k} g(x_k)p_X(x_k).$$

The expectation operation inherits all of the properties of integrals (and sums) of functions, including linearity. Thus,

$$E\left[ag(X) + bh(X)\right] = aE[g(X)] + bE[h(X)].$$

There are some standard expectations of random variables, which depend on the choice of function g. First, the **mean** or average of a random variable is defined as

$$\begin{aligned} m_X = E[X] \quad &= \quad \int_{-\infty}^{\infty} sp_X(s)ds. \quad &\text{(continuous-valued RV)} \\ &= \quad \sum_{k} x_k p_X(x_k). \quad &\text{(discrete-valued RV)} \end{aligned}$$

Case	Description	Probability Mass Function $p_X(x)$ Probability Density Function $p_X(x)$	Probability Distribution Function Cumulative Distribution Function $P_X(x)$
Discrete Distribution as pmf	$p_X(x_i) \geq 0, \quad \sum_{x_i} p_X(x_i) = 1$ $P_X(x) = \sum_{z \leq x} p_X(z)$ $P_X(A) = \sum_{x \in A} p_X(x)$	$p_X(x)$ — x_1, x_2, x	$P_X(x)$ — x_1, x_2, x
Discrete Distribution via Impulses	$p_X(x) = \sum_i p_i \delta(x - x_i), \quad \sum_i p_i = 1$ $P_X(x) = \int_{-\infty}^{x} p_X(z)\, dz$ $P_X(A) = \int_{z \in A} p_X(z)\, dz$	$p_X(x) = \dfrac{d}{dx} P_X(x)$ — x_1, x_2, x	$P_X(x)$ — x_1, x_2, x
Continuous Distribution	$p_X(x) \geq 0, \quad \int p_X(z)\, dz = 1$ $P_X(x) = \int_{-\infty}^{x} p_X(z)\, dz$ $P_X(A) = \int_{z \in A} p_X(z)\, dz$	$p_X(x) = \dfrac{d}{dx} P_X(x)$ — a, b, x	$P_X(x)$ — a, b, x
Mixed Distribution (Lebesque decomposition)	$p_X(x) = \alpha \underbrace{p(x)}_{\text{cont dist}} + (1-\alpha) \underbrace{\sum_i p_i \delta(x - x_i)}_{\text{sing dist}}$ $P_X(x) = \int_{-\infty}^{x} p_X(z)\, dz$ $P_X(A) = \int_{z \in A} p_X(z)\, dz$	$p_X(x) = \dfrac{d}{dx} P_X(x)$ — a, x_1, b, x_2, x	$P_X(x)$ — a, x_1, b, x_2, x
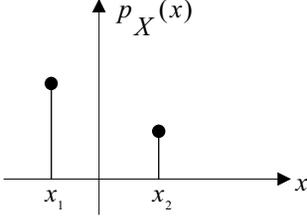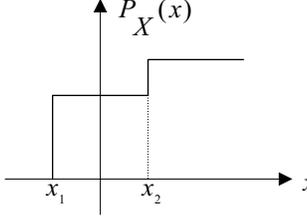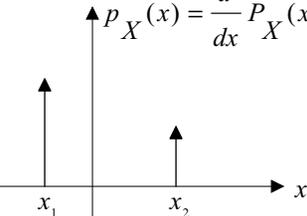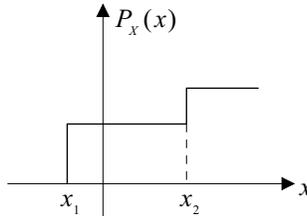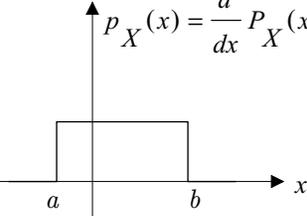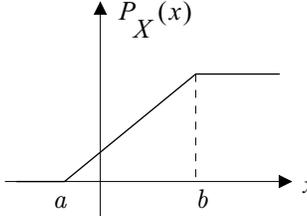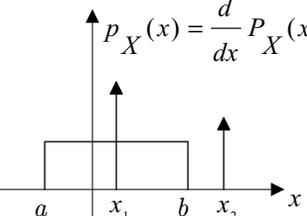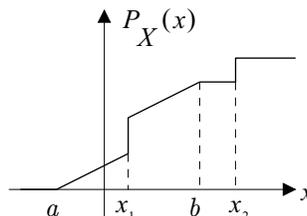
Table 1.1: Summary of probability distribution function and probability density relationships.

More generally, the n-th moment is defined as

$$
\begin{aligned}
E[X^n] &= \int_{-\infty}^{\infty} s^n p_X(s)ds \qquad \text{(continuous-valued RV)} \\
&= \sum_k x_k^n p_X(x_k). \qquad \text{(discrete-valued RV)}
\end{aligned}
$$

The **variance** of a random variable is defined in terms of its first and second moments, as

$$
\sigma_X^2 = E\left[(X - m_X)^2\right] = E[X^2] - (E[X])^2.
$$

The variance is the second-order case of the more general n-th central moment, $E[(X - m_X)^n]$.

Another important expectation is the **characteristic function**, defined as

$$
\Phi_X(w) = E[e^{jwX}] = \int_{-\infty}^{\infty} e^{jws} p_X(s)ds.
$$

The characteristic function is the Fourier transform of the probability density function, where $j = \sqrt{-1}$. It uniquely characterizes the density function, as it can be obtained as the inverse Fourier transform as

$$
p_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-jwx} \Phi_X(w)\, dw.
$$

Since the density function is integrable (it integrates to one), the characteristic function always exists. For discrete, integer-valued random variables, one often defines the **moment generating function**:

$$
G_X(z) = E[z^X] = \sum_k z^k p_X(k),
$$

which is the Z-transform of the discrete probability mass function. (Note that here we have used k instead of x_k to make the integer values explicit.) As for the continuous case, the transform can be inverted to obtain the pmf.

Both the characteristic function and the moment-generating function can be used to obtain the moments of x, assuming that the functions are differentiable, and that expectations and differentiations can be exchanged (usually, except in rare cases):

$$
\begin{aligned}
\frac{d}{dw}\Phi_X(w)|_{w=0} &= \left.\left(\frac{d}{dw}\int_{-\infty}^{\infty} e^{jws} p_X(s)\, ds\right)\right|_{w=0} \\
&= \left.\left(\int_{-\infty}^{\infty} \frac{d}{dw} e^{jws} p_X(s)\, ds\right)\right|_{w=0} \\
&= \left.\left(\int_{-\infty}^{\infty} (js) e^{jws} p_X(s)\, ds\right)\right|_{w=0} \\
&= j\int_{-\infty}^{\infty} (s) p_X(s)\, ds = jE[X]. \qquad (1.4)
\end{aligned}
$$

More generally, assuming that the characteristic function is sufficiently differentiable,

$$
\left.\frac{d^n}{dw^n}\Phi_X(w)\right|_{w=0} = (j)^n E[X^n].
$$

Similarly, for the moment generating function, we have

$$
G_X(z) = E\left[z^X\right]
$$

$$
\frac{d}{dz} G_X(z)|_{z=1} = E[X]
$$

$$
\frac{d^2}{dz^2} G_X(z)|_{z=1} = E[X^2] - E[X]
$$

and additional expressions can be developed for the higher-order moments.

1.5 Important Random Variables

There are a number of random variables that arise in many applications. These random variables model fundamental mechanisms that underlie random behavior. In this handout, we discuss several of these random variables, and their interrelations. A good reference for this material, from which this writeup is adapted, is A. Leon-Garcia's book, Probability and Random Processes, published by Addison Wesley.

1.5.1 Discrete-valued random variables

Discrete-valued random variables arise mostly in applications where counting is involved. For example, the Bernoulli random variable is a model for a single coin toss. By counting the outcomes of multiple coin tosses, other random variables such as the binomial, geometric and Poisson, are obtained.

Bernoulli random variable: Let A be an event related to the outcome of some random experiment, such as a toss of a biased coin. Define the *indicator* function of A as:

$$I_A(\omega) = \begin{cases} 0 & \text{if } \omega \text{ is not in } A \\ 1 & \text{if } \omega \text{ is in } A. \end{cases} \tag{1.5}$$

Thus, the indicator function is one if the event A occurs, and zero otherwise. Note that I_A is a random variable, with discrete values in range $\{0, 1\}$, and with probability mass function given by:

$$p_{I_A}(0) = 1 - p, \quad p_{I_A}(1) = p, \tag{1.6}$$

where $P(A) = p$. Such a random variable is called a *Bernoulli* random variable, since it identifies the outcome of a Bernoulli trial if we identify the outcome $I_A = 1$ as a success.

The important expectations of a Bernoulli random variable X are easily computed in terms of p. They are listed below:

$$E[X] = p \quad \text{Mean} \tag{1.7}$$

$$E[X^2] - E[X]^2 = p(1-p) \quad \text{Variance} \tag{1.8}$$

$$E[z^X] = 1 - p + pz \quad \text{Moment Generating Function} \tag{1.9}$$

Binomial random variable: Suppose that a random experiment is repeated n times. Let x denote the number of times that such an experiment was a success. In terms of the notation used above in the context of Bernoulli random variables, let A denote an event, and let x denote the number of times that such an event occurs out of n independent trials. Then, X is a random variable with discrete range $\{0, 1, \ldots, n\}$.

A simple representation of x is given by

$$x = I_1 + I_2 + \ldots + I_n, \tag{1.10}$$

where I_i is the indicator that event A occurs at the independent trial i. The probability mass function of X is given by

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \tag{1.11}$$

where the factorial notation $k! = \prod_{j=1}^{k} j$ is used, and p is the single-trial probability that the event A occurs.

The binomial random variable arises in various applications where there are two types of outcomes, and we are interested in the number of outcomes of one type. Such applications include repeated coin tosses, correct/erroneous bits, good/defective items, active/silent stations, etc. The important expectations of binomial random variables are given below:

$$E[X] = np \quad \text{Mean} \tag{1.12}$$

$$E[X^2] - E[X]^2 = np(1-p) \quad \text{Variance} \tag{1.13}$$

$$E[z^X] = (1 - p + pz)^n \quad \text{Moment Generating Function} \tag{1.14}$$

Geometric random variable: The binomial random variable is obtained by fixing the number of Bernoulli trials and counting the number of successes. A different random variable is obtained by counting the number of trials until the first success occurs. Denote this random variable as M; this is a *geometric* random variable, and it takes values in the discrete infinite set $\{1, 2, \ldots\}$. The probability mass function of M is given by

$$P(M = k) = (1 - p)^{k-1}p, \tag{1.15}$$

where p is the single-trial probability that the event occurs.

One of the interesting properties of the geometric random variable is that it is "memoryless"; that is,

$$P(M \geq k + j | M > j) = P(M \geq k) \quad \text{for all } j, k > 1. \tag{1.16}$$

In words, the above expression states that, if a success has not occurred in the first j trials, the probability of having to perform at least k more trials until a success is the same as the probability of initially having to perform at least k trials. Thus, the system "forgets" and begins anew as if it were performing the first trial.

The geometric random variable arises in applications where one is interested in the time between occurrence of events in a sequence of independent experiments. Such random variables have broad applications in different aspects of queuing theory. The important expectations of geometric random variables are summarized below:

$$E[M] \quad = \quad \frac{1}{p} \quad \text{Mean} \tag{1.17}$$

$$E\left[M^2\right] - E\left[M\right]^2 \quad = \quad \frac{1 - p}{p^2} \quad \text{Variance} \tag{1.18}$$

$$E\left[z^M\right] \quad = \quad \frac{pz}{1 - (1 - p)z} \quad \text{Moment Generating Function} \tag{1.19}$$

In some applications, it is useful to represent the space of outcomes as starting at zero, i.e. $\{0, 1, 2, \ldots\}$. In this case, which can be thought of as a shift by 1, the variance does not change, the mean becomes $m_x = (1 - p)/p$, and the moment generating function becomes $G_X(z) = p/[1 - (1 - p)z]$.

Poisson random variable: In many applications, we are interested in counting the number of occurrences of an event in a certain time period or in a certain region of space. The Poisson random variable arises in situations where the events occur "completely at random" in time or space; that is, where the likelihood of an event occurring at a particular time is equal to and independent of the event occurring at a different time. For example, Poisson random variables arise in counts of emissions from radioactive substances, in the number of photons emitted as a function of light intensity, in counts of demands for telephone connections, and in counts of defects in a semiconductor chip.

The probability mass function of a Poisson random variable N is given by

$$P(N = k) = \frac{\lambda^k}{k!}e^{-\lambda}, \tag{1.20}$$

where λ is the average number of event occurrences in the specified time interval or region of space.

One of the applications of the Poisson random variable is as an approximation to the binomial probabilities when the number of trials is large. If the number of trials n_t is large, and if p is small, then, letting $\lambda = n_t p$,

$$\frac{n_t!}{k!(n_t - k)!}p^k(1 - p)^{n_t - k} \approx \frac{\lambda^k}{j!}e^{-\lambda}. \tag{1.21}$$

This approximation is obtained by taking the limit $n_t \to \infty$ while keeping λ fixed.

The Poisson random variable appears naturally in many situations which can be approximated by the above limit. For example, imagine a sequence of Bernoulli trials taking place in time or space. Suppose the number of event occurrences in a T-second time interval is being counted. Divide the time interval into a very large number n_t of subintervals, where a pulse in each subinterval can be viewed as a Bernoulli trial, and assume that the probability that an event occurs in each subinterval is $p = \lambda/n_t$, where λ is the average

number of events observed in a T-second interval. Then, as $n_t \to \infty$, the limiting distribution becomes a Poisson random variable.

The important expectations of Poisson random variables are summarized below:

$$E[N] \quad = \quad \lambda \quad \text{Mean} \tag{1.22}$$

$$E\left[N^2\right] - E\left[N\right]^2 \quad = \quad \lambda \quad \text{Variance} \tag{1.23}$$

$$E\left[z^N\right] \quad = \quad e^{\lambda(z-1)} \quad \text{Moment Generating Function} \tag{1.24}$$

1.5.2 Continuous-valued random variables

Although most experimental measurements are of limited precision, it is often easier to model their outcomes in terms of continuous-valued random variables because it facilitates the resulting analysis. Furthermore, the limiting form of many discrete-valued random variables result in continuous-valued random variables. Below, we describe some of the most useful continuous-valued random variables.

Uniform random variable: The simplest continuous random variable is the uniform random variable X, where X is equally likely to achieve any value in an interval of the real line, $[a, b]$. The probability density function of X is given by:

$$p_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \tag{1.25}$$

The corresponding probability distribution function is given by

$$P_X(x) = \frac{x - a}{b - a} \tag{1.26}$$

The important expectations of uniform random variables are given by

$$E[X] \quad = \quad \frac{a + b}{2} \quad \text{Mean} \tag{1.27}$$

$$E\left[X^2\right] - E\left[X\right]^2 \quad = \quad \frac{(b - a)^2}{12} \quad \text{Variance} \tag{1.28}$$

$$\Phi_X(\omega) \quad = \quad \frac{e^{j\omega b} - e^{j\omega a}}{j\omega(b - a)} \quad \text{Characteristic Function} \tag{1.29}$$

Exponential random variable: The exponential random variable arises in the modeling of the time between occurrence of events, such as the time between customer requests for call connections in phone systems, and the modeling of lifetimes of devices and systems. The exponential random variable X with parameter α has a probability density function

$$p_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \alpha e^{-\alpha x} & \text{if } x \geq 0 \end{cases} \tag{1.30}$$

and corresponding probability distribution function

$$P_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\alpha x} & \text{if } x \geq 0 \end{cases} \tag{1.31}$$

The exponential random variable can occur as the limit of the geometric random variable, as the difference between values of a geometric random variable gets small. For example, assume that an interval of length T was subdivided into subintervals of length T/n, and assume that, for each subinterval, there is a Bernoulli trial with probability of success $p = \alpha/n$, where α is the average number of events per T seconds. Then, the number of subintervals until the occurrence of the next event is a geometric random variable M. Let X denote the time until the next successful event. Then, for any t which is a multiple of T/n,

$$P(X > t) = P\left(M > \frac{nt}{T}\right) = (1 - p)^{nt/T} = \left[\left(1 - \frac{\alpha}{n}\right)^n\right]^{t/T}$$

In the limit as $n \to \infty$, we get

$$P(X > t) \to e^{-\alpha t/T}$$

which is the complement of the probability distribution function in (1.31) of the exponential random variable.

Like the geometric random variable, the exponential random variable has the memoryless property. That is, for $h > 0$,

$$P(X > t + h | X > t) = P(X > h) \tag{1.32}$$

This can be shown analytically as:

$$P(X > t + h | X > t) = \frac{P[(X > t + h) \cap P(X > t)]}{P(X > t)} \tag{1.33}$$

$$= \frac{P(X > t + h)}{P(X > t)} = \frac{e^{-\alpha(t+h)}}{e^{-\alpha t}} \tag{1.34}$$

$$= e^{-\alpha h} = P(X > h) \tag{1.35}$$

The important expectations of exponential random variables are given by

$$E[X] = \frac{1}{\alpha} \quad \text{Mean} \tag{1.36}$$

$$E\left[X^2\right] - E\left[X\right]^2 = \frac{1}{\alpha^2} \quad \text{Variance} \tag{1.37}$$

$$\Phi_X(\omega) = \frac{\alpha}{\alpha - j\omega} \quad \text{Characteristic Function} \tag{1.38}$$

Gaussian random variable: Also known as the Normal random variable, the Gaussian random variable models many situations where the random event consists of the sum of a large number of small random variables. To develop the exact distribution of the sum of random variables is unwieldy; fortunately, the central limit theorem and the law of large numbers provide general conditions under which, as the number of components becomes large, the distribution of the sum can be approximated by that of a Gaussian random variable.

The probability density function of a Gaussian random variable is given by

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty, \tag{1.39}$$

where μ is the mean and $\sigma > 0$ is the standard deviation. The probability distribution function is given by

$$P_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} \, dt$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{\frac{-y^2}{2}} \, dy, \tag{1.40}$$

where the last expression follows from a simple substitution $y = (t - \mu)/\sigma$. The Gaussian PDF is sometimes characterized in terms of the Q-function

$$P_X(x) = 1 - Q((x - \mu)/\sigma) \quad \text{where} \quad Q(z) = \int_z^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}} \, dy, \tag{1.41}$$

and the Q function is tabulated in many texts.

Gaussian random variables occur often enough that we use the notation $N(X; \mu, \sigma^2)$ to denote the density or distribution in equations (1.39,1.40). The important expectations of Gaussian random variables are given by:

$$E[X] = \mu \quad \text{Mean} \tag{1.42}$$

$$E\left[X^2\right] - E\left[X\right]^2 = \sigma^2 \quad \text{Variance} \tag{1.43}$$

$$\Phi_X(\omega) = e^{j\omega\mu - \sigma^2\omega^2/2} \quad \text{Characteristic Function} \tag{1.44}$$

Gamma random variable: The gamma random variable appears in may applications. For example, it is often used to model the time to service customers in queuing systems, the lifetime of devices in reliability studies, and the defect clustering behavior in VLSI chips. The probability density function of the gamma random variable has two parameters $r > 0, \alpha > 0$, and is given by

$$p_X(x) = \frac{\alpha(\alpha x)^{r-1}e^{-\alpha x}}{\Gamma(r)}, \tag{1.45}$$

where $\Gamma(z)$ is the gamma function defined by the integral

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx, \quad z > 0.$$

The gamma function has the following properties:

$$
\begin{aligned}
\Gamma(0.5) &= \sqrt{\pi} & (1.46)\\
\Gamma(z+1) &= z\Gamma(z) & (1.47)\\
\Gamma(m+1) &= m! \quad \text{for } m \text{ a positive integer} & (1.48)
\end{aligned}
$$

The versatility of the gamma distribution is that, by properly choosing the two parameters, it can take a variety of shapes, which can be used to fit specific distributions. For instance, when $r = 1$, we obtain the exponential random variable. By letting $\alpha = 0.5, r = k/2$, for a positive integer k, we obtain the distribution of a chi-square random variable, which is important in statistical problems as the sum of the squares of k independent, zero-mean, unit variance Gaussian random variables. By letting $r = m$, we obtain the m-stage Erlang distribution, which is the distribution of the sum of m independent and identical exponential random variables.

The important expectations of gamma random variables are given by:

$$
\begin{aligned}
E[X] &= r/\alpha & \text{Mean} & (1.49)\\
E\left[X^2\right] - E\left[X\right]^2 &= \frac{r}{\alpha^2} & \text{Variance} & (1.50)\\
\Phi_X(\omega) &= \frac{1}{(1 - j\omega/\alpha)^r} & \text{Characteristic Function} & (1.51)
\end{aligned}
$$

Rayleigh random variable: Given a pair of independent, zero-mean, variance α^2 Gaussian random variables Y and Z, the Rayleigh random variable is the magnitude of the vector corresponding to the ordered pair (Y, Z). That is, $X = \sqrt{(Y^2 + Z^2)}$. Based upon this, we can compute the probability density function of Rayleigh random variables as

$$p_X(x) = \frac{x}{\alpha^2}e^{-x^2/2\alpha^2} \tag{1.52}$$

with corresponding expectations

$$
\begin{aligned}
E[X] &= \alpha\sqrt{\pi/2} & \text{Mean} & (1.53)\\
E\left[X^2\right] - E\left[X\right]^2 &= (2 - \pi/2)\alpha^2 & \text{Variance} & (1.54)
\end{aligned}
$$

Laplacian random variable: The Laplacian random variable models a two-sided exponential distribution. The probability density function is given by

$$p_X(x) = \frac{\alpha}{2}e^{-\alpha|x|}, \tag{1.55}$$

with expectations:

$$E[X] \quad = \quad 0 \quad \text{Mean} \tag{1.56}$$

$$E\left[X^2\right] - E\left[X\right]^2 \quad = \quad \frac{2}{\alpha^2} \quad \text{Variance} \tag{1.57}$$

$$\Phi_X(\omega) \quad = \quad \frac{\alpha^2}{\omega^2 + \alpha^2} \quad \text{Characteristic Function} \tag{1.58}$$

Cauchy random variable: The Cauchy random variable is often used as an example to illustrate distributions which do not decay fast enough as $x \to \infty$, so that no moments exist. The probability density function of Cauchy random variables is given by

$$p_X(x) = \frac{\beta/\pi}{\beta^2 + x^2}. \tag{1.59}$$

Due to its symmetry, the mean is often taken to be zero, though the formal expected value of the density does not have a unique value. It is easy to verify that the variance of this distribution does not exist, as follows. Consider the following expression for the second moment:

$$E\left[X^2\right] = \int_{-\infty}^{\infty} x^2 \frac{\beta/\pi}{\beta^2 + x^2} \, dx. \tag{1.60}$$

Note that, as $x \to \infty$, the integrand does not approach zero, so the integral will be infinite (i.e. will not exist). However, the Cauchy random variable does have a characteristic function, given by

$$\Phi_X(\omega) = e^{-\beta|\omega|} \quad \text{Characteristic Function} \tag{1.61}$$

In Table 1.2 we summarize the characteristics of important random variables, where the more general (shifted) forms of the Laplacian and Cauchy distributions are given.

1.6 Pairs of Random Variables

Now let us now consider a *pair* of random variables X, Y. In addition to the probabilistic structure and quantities associated with a single random variable, when we consider pairs of random variables we also have the additional richness of the *interrelationship between* the random variables. The joint probability density of the pair (X, Y) is denoted by $p_{X,Y}(x, y)$ or, in short, $p(x, y)$. Let $\Re^2$ denote the plane. Suppose we want to know the probability of obtaining an (x, y) pair in any subset A of the plane. For any measurable set (roughly a set with some area to it) $A \subset \Re^2$,

$$\Pr\left(\{\omega \mid (X(\omega), Y(\omega)) \in A\}\right) = \int_A p(x, y) \, dx \, dy.$$

This result just says that the probability of obtaining an outcome in this set is the integral of the probability mass over this region. In particular, suppose we want to know the probability of obtaining an x, y pair in a square located at x, y and of infinitesimal size dx, dy. Then we have:

$$\Pr\left(\{\omega \mid x < X(\omega) \leq x + dx, y < Y(\omega) \leq y + dy\}\right) = p(x, y) \, dx \, dy.$$

The joint Probability Distribution Function (PDF) (also referred to as the joint Cumulative Distribution Function (CDF)) of X, Y is defined as in the single random variable case as:

$$P_{X,Y}(x, y) = \Pr\left(\{\omega \mid X(\omega) \leq x, Y(\omega) \leq y\}\right) = \int_{-\infty}^{x} \int_{-\infty}^{y} p_{X,Y}(x, y) \, dx \, dy.$$

Thus the joint probability density and distribution functions are also related via:

$$p_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} P_{X,Y}(x, y)$$

Discrete-Valued X							
Name	Range	Parameters	pmf $p_X(x)$	PDF $P_X(x)$	Mean	Variance	$G_X(z) = E[z^X]$
Bernoulli	$\{0,1\}$	$0 \leq p \leq 1$	$p^x(1-p)^{(1-x)}$	N/A	p	$p(1-p)$	$1 - p + pz$
Binomial	$\{0,\ldots,n\}$	$0 \leq p \leq 1$	$\binom{n}{x} p^x(1-p)^{(n-x)}$	N/A	np	$np(1-p)$	$(1 - p + pz)^n$
Geometric	$\{1,\ldots\}$	$0 < p < 1$	$(1-p)^x p$	$(1-p)(1-(1-p)^x)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pz}{1-(1-p)z}$
Poisson	$\{0,1,\ldots\}$	$0 < \lambda$	$\frac{\lambda^x e^{-\lambda}}{x!}$	N/A	λ	λ	$e^{\lambda(z-1)}$

Continuous-Valued X							
Name	Range	Parameters	pdf $p_X(x)$	PDF $P_X(x)$	Mean	Variance	$\Phi_X(w) = E[e^{jwX}]$
Uniform	$[a,b]$	$a < b$	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{jwb}-e^{jwa}}{jw(b-a)}$
Gaussian	$[-\infty,\infty]$	μ, σ^2	$\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$	$1 - Q((x-\mu)/\sigma)$	μ	σ^2	$e^{(jw\mu - \frac{\sigma^2 w^2}{2})}$
Exponential	$[0,\infty]$	$\alpha > 0$	$\alpha e^{-\alpha x}$	$1 - e^{-\alpha x}$	$\frac{1}{\alpha}$	$\frac{1}{\alpha^2}$	$\frac{\alpha}{\alpha - jw}$
Erlang	$[0,\infty]$	$\alpha > 0,\, n > 0$	$\frac{\alpha^n x^{n-1} e^{-\alpha x}}{(n-1)!}$	$1 - e^{-\alpha x}\sum_{k=0}^{n-1}\frac{(\alpha x)^k}{k!}$	$\frac{n}{\alpha}$	$\frac{n}{\alpha^2}$	$\frac{\alpha^n}{(\alpha - jw)^n}$
Gamma	$[0,\infty]$	$\alpha, r > 0$	$\frac{\alpha(\alpha x)^{(r-1)} e^{-\alpha x}}{\Gamma(r)}$	N/A	$\frac{r}{\alpha}$	$\frac{r}{\alpha^2}$	$\frac{\alpha}{(\alpha - jw)^r}$
Rayleigh	$[0,\infty]$	α^2	$\frac{x}{\alpha^2}e^{-x^2/2\alpha^2}$	$1 - e^{-x^2/2\alpha^2}$	$\alpha\sqrt{\frac{\pi}{2}}$	$(2 - \frac{\pi}{2})\alpha^2$	N/A
Laplacian	$[-\infty,\infty]$	$\alpha > 0, \mu$	$\frac{\alpha}{2}e^{-\alpha|x-\mu|}$	$\begin{cases} \frac{1}{2}e^{\alpha(x-\mu)} & x < \mu \\ 1 - \frac{1}{2}e^{-\alpha(x-\mu)} & x > \mu \end{cases}$	μ	$\frac{2}{\alpha^2}$	$\frac{\alpha^2 e^{-jw\mu}}{w^2 + \alpha^2}$
Cauchy	$[-\infty,\infty]$	$\alpha \geq 0, \beta > 0$	$\frac{\beta/\pi}{\beta^2+(x-\alpha)^2}$	$\frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\frac{x-\alpha}{\beta}\right)$	Undef	Undef	$e^{j\alpha\omega - \beta|\omega|}$

Table 1.2: Important random variables. (N/A under the PDF column indicates that there is no simplified form.)

The marginal density of either X or Y can be recovered by integrating out the other variable, e.g.

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

Two random variables X, Y are said to be **independent** if $p_{X,Y}(x,y) = p_X(x)p_Y(y)$.

The expected value of a function of X, Y is given by

$$E\left[f(X,Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)p_{X,Y}(x,y)\, dx\, dy.$$

Note that this gives a consistent definition for the expected value of a function of X only:

$$
\begin{aligned}
E\left[f(X)\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)\, p_{X,Y}(x,y)\, dy\, dx \\
&= \int_{-\infty}^{\infty} f(x) \int_{-\infty}^{\infty} p_{X,Y}(x,y)\, dy\, dx \\
&= \int_{-\infty}^{\infty} f(x)\, p_X(x)\, dx
\end{aligned}
\tag{1.62}
$$

Given two random variables X and Y, there are two important expectations of interest.

Cross-Correlation: The cross-correlation (or just correlation) is given by $E[XY]$. An important property of correlations is

$$E[XY]^2 \le E[X^2]E[Y^2]. \tag{1.63}$$

This follows from a well-known inequality for integrals.

Cross-Covariance: The cross-covariance is another measure of interrelationship and is defined by

$$\sigma_{XY} = E\left[(X - m_X)(Y - m_Y)\right] = E[XY] - m_X m_Y. \tag{1.64}$$

Some very important properties of random variable pairs are defined in terms of these quantities:

Uncorrelated Random Variables: Two random variables X, Y are said to be *uncorrelated* if:

$$\sigma_{XY} = 0. \tag{1.65}$$

From the definition of the cross-covariance we can see that an equivalent statement of the uncorrelated property is: $E[XY] = E[X]E[Y]$.

Orthogonal Random Variables: The variables are said to be *orthogonal* if:

$$E[XY] = 0. \tag{1.66}$$

First note that orthogonal and uncorrelated are different concepts – be careful in your use of these terms! Also note that if two random variables are both orthogonal and uncorrelated, then the mean of at least one must be zero. Finally, for zero mean random variables, orthogonality and uncorrelated are equivalent.

Before moving on, note the extremely important (and often missed) fact that uncorrelated or orthogonal random variables are not necessarily independent, but independent random variables are always uncorrelated. Remember that independence is a *strong* property of the underlying densities, while uncorrelatedness is only a property of second order moments. Think, for example, of the difference between a random variable that is always zero and a zero mean random variable.

1.7   Conditional Probabilities, Densities, and Expectations

We defined conditional probabilities in terms of events A, B in a probability space $(\Omega, \mathcal{F}, P)$. As we have discussed, random variables can be used to define events in the original probability space; conditioning on these events can be used to define a conditional probability in the original space, which will induce a conditional probability in the sample space. In this section, we discuss the properties of such a conditional probability, and the conditional densities associated with them.

To begin with, consider a random variable X, and denote by B the event $\{\omega \mid a < x(\omega) \leq b\}$. Denote by A the event $\{\omega \mid c < X(\omega) \leq d\}$. Then, by (1.1), we have

$$
\begin{aligned}
P(A|B) &= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(\{\omega \mid X(\omega) \in (a, b] \text{ and } X(\omega) \in (c, d]\})}{P(B)} \\
&= \frac{P_X((a, b] \cap (c, d])}{P_X((a, b])} \equiv P_X((c, d]|(a, b]). \quad (1.67)
\end{aligned}
$$

Can we define the conditional probability of a single outcome, given observation of an event? Suppose we let $d = c + \epsilon$, and we let $\epsilon \to 0$. Then, if the probability distribution function P_X is differentiable at c, we know that $P_X(c) = 0$, and so will $P_X(c|(a, b])$. However, we may be able to define the conditional probability density function $p_X(c|(a, b])$ by taking derivatives, as follows:

$$
\begin{aligned}
p_X(c|(a, b]) &= \lim_{\epsilon \to 0^+} \frac{P_X((c, c + \epsilon]|(a, b])}{\epsilon} \quad (1.68) \\
&= \lim_{\epsilon \to 0^+} \frac{P_X((a, b] \cap (c, c + \epsilon])}{\epsilon P_X((a, b])} \quad (1.69) \\
&= \begin{cases} 0 & \text{if } c \notin (a, b) \\ \lim_{\epsilon \to 0^+} \frac{P_X((c, c + \epsilon])}{\epsilon P_X((a, b])} & \text{otherwise.} \end{cases} \quad (1.70)
\end{aligned}
$$

In the latter case,

$$
p_X(c|(a, b]) = \frac{p_X(c)}{P_X((a, b])}.
$$

What about the converse event, of defining the conditional probability of an event given observation of a particular outcome of a random variable? We can let $b = a + \epsilon$ in eq. (1.67), to obtain

$$
\begin{aligned}
P(A|a) &= \lim_{\epsilon \to 0^+} P(A|(a, a + \epsilon]) \quad (1.71) \\
&= \lim_{\epsilon \to 0^+} \frac{P(A \cap \{\omega \mid x(\omega) \in (a, a + \epsilon]\})}{P_X((a, a + \epsilon])} \quad (1.72) \\
&= \lim_{\epsilon \to 0^+} \frac{P(A) P_X((a, a + \epsilon]|A)}{P_X((a, a + \epsilon])} \quad (1.73) \\
&= \lim_{\epsilon \to 0^+} \frac{P_X((a, a + \epsilon]|A)}{\epsilon} \frac{\epsilon}{P_X((a, a + \epsilon])} P(A) \quad (1.74) \\
&= \frac{p_X(a|A) P(A)}{p_X(a)} \quad (1.75)
\end{aligned}
$$

assuming that $p_X(a) \neq 0$.

Using the above relationships, we have another version of the total probability theorem, expressed in terms of probability densities. In essence, since the set of possible values of a random variable automatically generates disjoint sets in the event space, we have

$$
P(A) = \int_{-\infty}^{\infty} P(A|a) p_X(a) da
$$

and the corresponding version of Bayes' rule:

$$p_X(a|A) = \frac{P(A|a)p_X(a)}{P(A)} = \frac{P(A|a)p_X(a)}{\int_{-\infty}^{\infty} P(A|a)p_X(a)\,da}.$$

Given two random variables, we can compute the conditional density of one random variable given the other as a straightforward extension of the previous conditional density relationships:

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

Similarly, Bayes' rule becomes

$$p_{X|Y}(x \mid y) = \frac{p_{Y|X}(y \mid x)p_X(x)}{p_y(y)}.$$

In particular, note that for independent random variables, $p_{X|Y}(x \mid y) = p_X(x)$. This is an equivalent condition for independence of two random variables.

Given the conditional density of a random variable based on observation of another random variable, $p_{X|Y}(x \mid y)$, we can define conditional expectation of X given $Y = y$ as

$$E[X \mid Y = y] = \int_{-\infty}^{\infty} x\, p_{X|Y}(x \mid y)\, dx. \tag{1.76}$$

Note that if a particular value y is not specified, $E[X \mid Y]$ is a function of Y, and thus can be viewed as a random variable, since it is a function of a random variable. Therefore, we can take its expectation as

$$
\begin{aligned}
E\left[E\left[X \mid Y\right]\right] &= \int_{-\infty}^{\infty} E[X \mid Y = y]p_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x\, p_{X|Y}(x \mid y)\, dx \right) p_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} p_{X,Y}(x,y)\, dy \right) dx \\
&= \int_{-\infty}^{\infty} x\, p_X(x)\, dx = E[X]. \tag{1.77}
\end{aligned}
$$

The above property is known as the smoothing property of conditional expectations. It can be very useful for finding expectations of random variables that have two sources of randomness, as in the example below.

Example 1.2
Assume that the number of people in line at the bank when you arrive is N, where N is random, having a Poisson distribution with parameter λ. The time T_i that it takes to serve each person ahead of you can be described by an exponential distribution with parameter α, and the times for different people are mutually independent. How long do you expect to wait before someone starts to serve you? Let T be the time you will wait, then

$$T = \sum_{i=1}^{N} T_i$$

and

$$E[T] = E[E[T|N]] = E[N/\alpha] = \lambda/\alpha.$$

1.8   Random Vectors

We will frequently deal with several random variables. In this case rather than extend the notation introduced for two random variables, it will prove much more convenient and insightful to use vector notation. The vector notation simply serves as a compact way to "carry around" the associated collection of random

variables. It is best if you are familiar and comfortable with such vector notation and concepts, so you should refer to the appendix or a suitable linear algebra text at this point if you need a review. Using vector notation, all of the concepts and results we developed for the cases of a single random variable and pairs of random variables can be generalized to random vectors where all of the elements are defined on a common probability space. Let:

$$\underline{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix}$$

denote a vector of N random variables. (Note that we may use the alternate notation of $\mathbf{X}$ for vectors as well.) The joint distribution function is given by

$$P_{\underline{X}}(\underline{x}) = P_{X_1,\dots,X_N}(x_1,\dots,x_N) = P(\{\omega \mid X_1(\omega) \le x_1,\dots,X_N(\omega) \le x_N\}).$$

The joint density function is

$$p_{\underline{X}}(\underline{x}) = \frac{\partial^N}{\partial x_1 \dots \partial x_N} P_{\underline{X}}(\underline{x}).$$

For any measurable set $A \in \Re^N$, we have

$$P(\underline{X}(\omega) \in A) = \int \cdots \int_A p_{\underline{X}}(\underline{x}) d\underline{x}.$$

We can have several random vectors defined on the same probability space. Given two random vectors $\underline{X}, \underline{Y}$, we have a joint density $p_{\underline{X},\underline{Y}}(\underline{x},\underline{y})$, from which we can recover marginal densities as

$$p_{\underline{X}}(\underline{x}) = \int_{-\infty}^{\infty} p_{\underline{X},\underline{Y}}(\underline{x},\underline{y})\, d\underline{y}.$$

The vectors are said to be independent if $p_{\underline{X},\underline{Y}}(\underline{x},\underline{y}) = p_{\underline{X}}(\underline{x})p_{\underline{Y}}(\underline{y})$.
   The conditional density for $\underline{X}$ given $\underline{Y}$ is given by

$$p_{\underline{X}|\underline{Y}}(\underline{x} \mid \underline{y}) = \frac{p_{\underline{X},\underline{Y}}(\underline{x},\underline{y})}{p_{\underline{Y}}(\underline{y})} = \frac{p_{\underline{Y}|\underline{X}}(\underline{y} \mid \underline{x})p_{\underline{X}}(\underline{x})}{p_{\underline{Y}}(\underline{y})}.$$

Of course, the above formulas can be extended to more than two random vectors.
   Suppose that $\underline{Y} = \underline{g}(\underline{X})$ is a function of the random vector. We can always compute the probability distribution of $\underline{Y}$, based on the distribution of $\underline{X}$, as:

$$P_{\underline{Y}}(\underline{y}) = P\left(\{\omega \mid g_1(X(\omega)) \le y_1,\dots,g_M(X(\omega)) \le y_M\}\right) = \int_{A(\underline{y})} p_{\underline{X}}(\underline{x})\, d\underline{x},$$

where

$$A(\underline{y}) = \{\underline{x} \mid g_1(x) \le y_1,\dots,g_M(x) \le y_M\}.$$

We can then obtain the density by differentiation. In the special case that $M = N$ and $\underline{g}$ is one-to-one, we can compute $p(\underline{y})$ as

$$p_{\underline{Y}}(\underline{y}) = \frac{p_{\underline{X}}\left(g^{-1}(\underline{y})\right)}{\left|\det\left(\frac{\partial \underline{g}}{\partial \underline{x}}\left[\underline{g}^{-1}(\underline{y})\right]\right)\right|},$$

where $\frac{\partial \underline{g}}{\partial \underline{x}}$ is the Jacobian matrix, det denotes the determinant of the matrix, and $|\cdot|$ denotes absolute value.
   The expectation of a function of a random vector $\underline{X}$ is given by

$$E[g(\underline{X})] = \int_{-\infty}^{\infty} g(\underline{x})p_{\underline{X}}(\underline{x})\, d\underline{x}.$$

The expectation of an M-dimensional vector-valued function $\underline{g}$ can be defined componentwise, as

$$E\left[\underline{g}(\underline{X})\right] = \left[\begin{array}{c} E\left[g_1(\underline{X})\right] \\ \vdots \\ E\left[g_M(\underline{X})\right] \end{array}\right].$$

Some important expectations are:

Mean Vector: This is just the collection of individual expected values of each element of the random vector:

$$E[\underline{X}] = m_{\underline{X}}. \tag{1.78}$$

Covariance Matrix: This is the *matrix* of variances and cross-covariances of the variables within the random vector:

$$\Sigma_{\underline{X}\underline{X}} = E\left[\left(\underline{X} - m_{\underline{X}}\right)\left(\underline{X} - m_{\underline{X}}\right)^T\right] = E\left[\underline{X}\underline{X}^T\right] - m_{\underline{X}}m_{\underline{X}}^T. \tag{1.79}$$

Thus the covariance matrix gives us information about the interrelationship between the elements within a single random vector. In particular note that the elements of $\Sigma_{\underline{X}\underline{X}}$ are just:

$$(\Sigma_{\underline{X}\underline{X}})_{ii} = \sigma_{X_i}^2; \quad (\Sigma_{\underline{X}\underline{X}})_{ij} = E\left[\left(X_i - m_{X_i}\right)\left(X_j - m_{X_j}\right)\right] = \sigma_{X_iX_j} \tag{1.80}$$

so that the covariance matrix can be seen to be a compact way of representing the collection of variance and cross-covariance information for the collection of random variables in the random vector. The covariance matrix can be seen to be the natural generalization of the concept of the variance of a random variable, extended to a random vector.

In these notes, we will often drop the double subscript of the covariance matrix for convenience and brevity. For instance, we will often write $\Sigma_{\underline{X}}$ instead of $\Sigma_{\underline{X}\underline{X}}$. Note the dimensions of the covariance matrix: if $\underline{X}$ is N-dimensional then $\Sigma_{\underline{X}}$ is a square $N \times N$ matrix. Properties of the covariance matrix are discussed in Section 1.9.

Cross-covariance Matrix: Analogous to the covariance matrix, the cross-covariance matrix is simply the matrix of cross-covariances between the elements of two different random vectors:

$$\Sigma_{\underline{X}\underline{Y}} = E\left[\left(\underline{X} - m_{\underline{X}}\right)\left(\underline{Y} - m_{\underline{Y}}\right)^T\right] = E\left[\underline{X}\underline{Y}^T\right] - m_{\underline{X}}m_{\underline{Y}}^T. \tag{1.81}$$

Thus the cross-covariance matrix gives us information about the interrelationship between the elements of two *different* random vectors. In particular note that the elements of $\Sigma_{\underline{X}\underline{Y}}$ are just:

$$\left(\Sigma_{\underline{X}\underline{Y}}\right)_{ij} = E\left[\left(X_i - m_{X_i}\right)\left(Y_j - m_{Y_j}\right)\right] = \sigma_{X_iY_j} \tag{1.82}$$

so that the cross-covariance matrix is a compact way of representing the collection of cross-covariance information between the collection of random variables in the two different random vectors. Like the covariance matrix, the cross-covariance matrix can be seen as the natural generalization of the cross-covariance between two random variables, extended to two random vectors. Note the dimensions of the cross-covariance matrix: if $\underline{X}$ is N-dimensional and $\underline{Y}$ is M-dimensional then $\Sigma_{\underline{X}\underline{Y}}$ is an $N \times M$ matrix, and so is *not* necessarily square. Properties of the cross-covariance matrix are discussed in Section 1.9.

Characteristic Function: This function is the generalization of the characteristic function we defined for a random variable to a random vector:

$$\Phi_{\underline{X}}(j\underline{w}) = E\left[e^{j\underline{w}^T\underline{X}}\right]. \tag{1.83}$$

Note that now the frequency variable $\underline{w}$ is itself a vector, since $\underline{X}$ is a vector. As a consequence the characteristic function for an N-dimensional random vector is really an N-dimensional function. High and low frequencies are defined by the *magnitude* or length of the $\underline{w}$ vector, while its orientation determines direction. It is just the multidimensional Fourier Transform of the pdf. As for the scalar case, it completely determines the pdf.

As stated above, the cross-covariance matrix gives us information about the interrelationship between two random vectors. Along these lines we have the following two definitions, which are the generalizations of our earlier definitions to pairs of random vectors:

Uncorrelated Random Vectors: Two random vectors $\underline{X}$ and $\underline{Y}$ are said to be *uncorrelated* if the cross-covariance matrix is identically zero (i.e. each element is zero):

$$(\Sigma_{\underline{XY}})_{ij} = 0 \quad \forall i, j \tag{1.84}$$

From the definition of the cross-covariance we can see that an equivalent statement of the uncorrelated property is: $E\left[\underline{X}\underline{Y}^T\right] = E\left[\underline{X}\right]E[\underline{Y}]^T$. If we are talking about a single random vector, then to say $\underline{X}$ is uncorrelated is to say that all elements are uncorrelated to each other, in which case the covariance matrix $\Sigma_{\underline{X}}$ is diagonal.

Orthogonal Random Vectors: The random vectors $\underline{X}$, $\underline{Y}$ are said to be *orthogonal* if the cross-correlation matrix is identically zero (i.e. each element is zero):

$$(E\left[\underline{X}\underline{Y}^T\right])_{ij} = 0 \qquad \forall i, j. \tag{1.85}$$

Again note that orthogonal and uncorrelated are different concepts – be careful in your use of these terms! Also note that if two random vectors are both orthogonal and uncorrelated, then the mean vector of at least one must be zero. If we are talking about a single random vector, then to say $\underline{X}$ is orthogonal is to say that all elements are orthogonal to each other, in which case the correlation matrix $E[\underline{X}\underline{X}^T]$ is diagonal. Finally, for zero mean random vectors, orthogonality and uncorrelated are equivalent (as for random variables).

We may also define conditional quantities for random vectors, in an analogous manner to our definitions for random variables.

Conditional Mean Vector:

$$E[\underline{X} \mid \underline{Y} = \underline{y}] = m_{\underline{X}|\underline{y}} = \int_{-\infty}^{\infty} \underline{x}\, p_{\underline{X}|\underline{Y}}(\underline{x} \mid \underline{y})\, d\underline{x} \tag{1.86}$$

Conditional Covariance Matrix:

$$\Sigma_{\underline{X}|\underline{y}} = \int_{-\infty}^{\infty} \left(\underline{x} - E\left[\underline{X} \mid \underline{Y} = \underline{y}\right]\right)\left(\underline{x} - E\left[\underline{X} \mid \underline{Y} = \underline{y}\right]\right)^T p_{\underline{X}|\underline{Y}}(\underline{x} \mid \underline{y})\, d\underline{x} \tag{1.87}$$

As before, these conditional quantities can have two interpretations. If we observe a particular value of $\underline{y}$ we can think of these conditional expectations as deterministic. Alternatively, $m_{\underline{X}|\underline{Y}}$ and $\Sigma_{\underline{X}|\underline{Y}}$ can be thought of as functions of $\underline{Y}$, and thus as random quantities themselves. In this latter case, these quantities have their own densities and e.g. we can find their expectations. In particular, as for the scalar case, the smoothing property of conditional expectations still holds: $E\left[E\left[\underline{X} \mid \underline{Y}\right]\right] = E[\underline{X}]$.

1.9   Properties of the Covariance Matrix

In this section we examine and summarize properties of covariance matrices and cross-covariance matrices. Since

$$\left(\Sigma_{\underline{X}}\right)_{ij} = \sigma_{X_i X_j} = \sigma_{X_j X_i} = \left(\Sigma_{\underline{X}}\right)_{ji}, \tag{1.88}$$

the first obvious property of the covariance matrix is that it is symmetric:

$$\Sigma_{\underline{X}} = \Sigma_{\underline{X}}^T. \tag{1.89}$$

To proceed, let us first understand how the covariance matrices of random vectors are transformed by linear operations such as matrix multiplication and vector addition. Let $\underline{X}$, $\underline{Y}$ be random vectors, and define a new random vector $\underline{Z}$ by linear operations as follows:

$$\underline{Z} = A\underline{X} + B\underline{Y} + \underline{c}$$

for some deterministic matrices A, B of appropriate dimensions and a deterministic vector $\underline{c}$. Since expectation is a linear operation, we can compute

$$
\begin{aligned}
E[\underline{Z}] &= E[A\underline{X} + B\underline{Y} + \underline{c}] \\
&= E[A\underline{X}] + E[B\underline{Y}] + E[\underline{c}] \\
&= AE[\underline{X}] + BE[\underline{Y}] + \underline{c} \\
&= Am_{\underline{X}} + Bm_{\underline{Y}} + \underline{c}
\end{aligned}
\tag{1.90}
$$

Similarly, we can compute the covariance Σ_z as

$$
\begin{aligned}
\Sigma_{\underline{Z}} &= E\left[\underline{Z}\underline{Z}^T\right] - E\left[m_{\underline{Z}}m_{\underline{Z}}^T\right] \\
&= E\left[(A\underline{X} + B\underline{Y} + \underline{c})(A\underline{X} + B\underline{Y} + \underline{c})^T\right] - E\left[m_{\underline{Z}}m_{\underline{Z}}^T\right] \\
&= AE\left[\underline{X}\underline{X}^T\right]A^T + AE\left[\underline{X}\underline{Y}^T\right]B^T + BE\left[\underline{Y}\underline{X}^T\right]A^T + BE\left[\underline{Y}\underline{Y}^T\right]B^T \\
&\quad + E\left[A\underline{X} + B\underline{Y}\right]\underline{c}^T + \underline{c}E\left[A\underline{X} + B\underline{Y}\right]^T + \underline{c}\,\underline{c}^T - E\left[m_{\underline{Z}}m_{\underline{Z}}^T\right] \\
&= A\Sigma_{\underline{X}}A^T + A\Sigma_{\underline{XY}}B^T + B\Sigma_{\underline{YX}}A^T + B\Sigma_{\underline{Y}}B^T
\end{aligned}
\tag{1.91}
$$

These results are general, in that they apply to any linear transformation of a pair of random vectors.

Now let's use this result to continue to characterize the covariance matrix. To this end, consider the special case arising if we define the scalar random variable Z as follows:

$$Z = \underline{a}^T\underline{X} \tag{1.92}$$

for a deterministic vector $\underline{a}$ and a random vector $\underline{X}$ of equal dimension. Then, from (1.91) we have that the variance of Z is given by:

$$\sigma_Z^2 = \underline{a}^T\Sigma_{\underline{X}}\underline{a} \tag{1.93}$$

Since we know that $\sigma_Z^2 \geq 0$ and the vector $\underline{a}$ was general we thus have that:

$$\underline{a}^T\Sigma_{\underline{X}}\underline{a} \geq 0, \quad \text{for all} \quad \underline{a} \tag{1.94}$$

A symmetric matrix that has this property is termed *positive semi-definite*. So we have just proved that *a covariance matrix must be a positive semi-definite matrix*. While the definition provided in (1.94) is correct, it is not convenient to apply (since forming the quadratic form $\underline{a}^T\Sigma_{\underline{X}}\underline{a}$ for all $\underline{a}$ is difficult). It turns out that an equivalent condition for positive semi-definiteness of a symmetric matrix is that all its eigenvalues must be nonnegative. This is not hard to see. Suppose the matrix $\Sigma_{\underline{X}}$ has an eigen-decomposition as follows:

$$\Sigma_{\underline{X}} = U\Lambda U^T = U\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix}U^T, \tag{1.95}$$

where U is the matrix whose columns are the eigenvectors of $\Sigma_{\underline{X}}$ and $\lambda_1, \cdots, \lambda_N$ are the corresponding eigenvalues. For a symmetric matrix, the eigenvector matrices can always be chosen as unitary, and so correspond to (generalized) rotation matrices. Now consider the quadratic form:

$$\underline{a}^T\Sigma_{\underline{X}}\underline{a} = \underbrace{\underline{a}^T U}_{\tilde{a}^T}\Lambda\underbrace{U^T\underline{a}}_{\tilde{a}} = \tilde{\underline{a}}^T\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix}\tilde{\underline{a}} \geq 0. \tag{1.96}$$

Now the above expression must be nonnegative for all possible choices of the vector $\underline{a}$ or, equivalently, the vector $\tilde{\underline{a}}$ (since we can always find a vector $\underline{a}$ to generate any vector $\tilde{\underline{a}} = U^T \underline{a}$). In particular, it must hold if we choose $\tilde{\underline{a}}$ as the unit coordinate vectors

$$
\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ \vdots \end{bmatrix}, \quad \dots \tag{1.97}
$$

However, this choice of vectors just picks out each eigenvalue in turn! Thus, for the quadratic form to be nonnegative each eigenvalue must be nonnegative, so nonnegativeness of the eigenvalues is necessary. Is it sufficient? Yes, since any other vector can be formed as a linear combination of these coordinate vectors. In summary, *a covariance matrix must be a symmetric, positive semi-definite matrix.*

Note that if $\Sigma_{\underline{X}}$ is positive definite, the quadratic form will always be positive or equivalently all the eigenvalues will be positive. In this case, $\Sigma_{\underline{X}}$ will be invertible and the covariance of derived random variable Z would always be strictly positive.

The singular or indefinite case, when $\Sigma_{\underline{X}}$ has a zero eigenvalue seems special. When does such a case arise? Lets take a closer look. Suppose we know that one element of the random vector $\underline{X}$ is really a linear combination of the other elements (i.e. suppose the elements are linearly dependent). In this case, there exists a vector $\underline{a}$ such that:

$$
\underline{a}^T \underline{X} = 0, \tag{1.98}
$$

which implies that:

$$
\underline{a}^T \Sigma_{\underline{X}} \underline{a} = \underline{a}^T \left(E \left[\underline{X} \underline{X}^T \right] - E \left[\underline{X} \right] E \left[\underline{X} \right]^T \right) \underline{a} = \left(E \left[\underline{a}^T \underline{X} \underline{X}^T \underline{a} \right] - E \left[\underline{a}^T \underline{X} \right] E \left[\underline{a}^T \underline{X} \right]^T \right) = 0, \tag{1.99}
$$

which implies that $\Sigma_{\underline{X}}$ is singular. Thus, if one element of a random vector is a linear combination of the other elements, then the covariance matrix of the vector will be singular!

Finally, what about the cross-covariance matrix? Again, using the definition of the cross-covariance we see that

$$
\left(\Sigma_{\underline{XY}} \right)_{ij} = \sigma_{X_i Y_j} = \sigma_{Y_j X_i} = \left(\Sigma_{\underline{YX}} \right)_{ji}, \tag{1.100}
$$

so the cross-covariance matrix satisfies:

$$
\Sigma_{\underline{XY}} = \Sigma_{\underline{YX}}^T.
$$

Unlike the entries of the covariance matrix, the entries of the cross-covariance do not satisfy any restrictions. Indeed any matrix could be the cross-covariance of some pair of random vectors. It does not even need to be square.

1.10 Gaussian Random Vectors

A special case of random vectors is the case of what are termed *Gaussian random vectors*. Recall that Gaussian random variables have at least two extremely important properties. First, their probability density functions can be completely characterized by just two quantities: the mean and the covariance. Second, linear functions of a Gaussian random variable result in another Gaussian random variable. The extensions to Gaussian random vectors will also possess generalizations of these properties. In particular, Gaussian random vectors will be completely characterized by their mean vectors and covariance matrices, and linear functions of Gaussian vectors will also result in Gaussian random vectors. This has important consequences. For example, the analysis of linear systems driven by Gaussian random variables can be restricted to analyzing the first and second-order expectations of the inputs and outputs.

Recall that, for a Gaussian random variable X, we use the notation $N(m, \sigma^2)$ to denote a Gaussian distribution of mean m and variance σ^2 (cf. eq. (1.39)). Then, the random variable $Z = aX + b$ has distribution $N(am + b, a^2\sigma^2)$, from equations (1.90) and (1.91).

Now how will we define a Gaussian random vector? The answer is in terms of quantities we already know – in particular, in terms of Gaussian random variables. An n-dimensional random vector

$$\underline{X} = \left[\begin{array}{c} X_1 \\ X_2 \\ \vdots \\ X_n \end{array} \right]$$

is defined to be a Gaussian random vector (or equivalently, $\{X_1, \ldots, X_n\}$ are defined to be a set of jointly Gaussian random variables) if, for all constants

$$\underline{a} = \left[\begin{array}{c} a_1 \\ a_2 \\ \vdots \\ a_n \end{array} \right],$$

the random variable $Y = \underline{a}^T \underline{X}$ is a Gaussian random variable. Note that it is *not enough* that each entry is marginally a Gaussian random variable for the vector to be a Gaussian random vector! *All* linear combinations of the entries must also be Gaussian. The converse, however is true: the entries of a Gaussian random vector are individually Gaussian random variables.

The probability density of Gaussian random vectors is completely described by the mean $\underline{m}_X$ and the covariance Σ_X. We use the notation $\underline{X} \sim N(\underline{m}_X, \Sigma_X)$ to denote this distribution. Then,

$$p_{\underline{X}}(\underline{x}) = N(\underline{x}; \underline{m}_X, \Sigma_X) = \frac{1}{\sqrt{(2\pi)^n |\det \Sigma_X|}} e^{-0.5(\underline{x}-\underline{m}_X)^T (\Sigma_X)^{-1}(\underline{x}-\underline{m}_X)} \tag{1.101}$$

where we have assumed that Σ_X is invertible. By extension of the properties of Gaussian random variables, we can compute explicitly other important expectations of Gaussian random vectors. In particular, the joint characteristic function of Gaussian random vectors is given by

$$\Phi_{\underline{X}}(j\underline{w}) = E\left[e^{j\underline{w}^T \underline{X}} \right] = e^{j\underline{w}^T \underline{m}_X - \underline{w}^T \Sigma_X \underline{w}/2},$$

where the above formula is valid even if Σ_X is not invertible.

Using the above formula for characteristic functions, it is easy to show that linear combinations of Gaussian random vectors are Gaussian random variables. Let $Z = \underline{a}^T \underline{X} + b$ for some constants $\underline{a}$, b. Consider now the characteristic function of Z, given by:

$$\begin{aligned} \Phi_Z(jv) &= E\left[e^{jvZ} \right] = E\left[e^{jv(\underline{a}^T \underline{X}+b)} \right] = E\left[e^{j(v\underline{a})^T \underline{X}} \right] e^{jvb} \\ &= \Phi_{\underline{X}}(jv\underline{a}) e^{jvb} = e^{jv(\underline{a}^T \underline{m}_X + b) - v^2 \underline{a}^T \Sigma_X \underline{a}/2}, \end{aligned} \tag{1.102}$$

which is the characteristic function of a Gaussian random variable with mean $\underline{a}^T \underline{m}_X + b$ and variance $\underline{a}^T \Sigma_X \underline{a}$. Recall that there is a one-to-one correspondence between characteristic functions and probability density functions.

An important property of Gaussian random vectors is that two Gaussian random vectors are independent if and only if they are uncorrelated! To see this, let $\underline{X}$ and $\underline{Y}$ be jointly Gaussian random vectors of dimensions n, m respectively. Define a Gaussian random vector $\underline{Z}$ as

$$\underline{Z} = \left[\begin{array}{c} \underline{X} \\ \hline \underline{Y} \end{array} \right].$$

Then, $\underline{Z} \sim N(\underline{m}_Z, \Sigma_Z)$, with

$$\underline{m}_Z = \left[\begin{array}{c} \underline{m}_X \\ \hline \underline{m}_Y \end{array} \right]$$

$$\Sigma_{\underline{Z}} = \left[\begin{array}{cc} \Sigma_{\underline{X}} & \Sigma_{\underline{XY}} \\ \Sigma_{\underline{YX}} & \Sigma_{\underline{Y}} \end{array} \right].$$

If $\underline{X}$ and $\underline{Y}$ are uncorrelated, it means that $\Sigma_{\underline{XY}} = \Sigma_{\underline{YX}} = 0$. Under this condition, we have:

$$\det \Sigma_{\underline{Z}} = \det \Sigma_{\underline{X}} \det \Sigma_{\underline{Y}}$$

and

$$\Sigma_{\underline{Z}}^{-1} = \left[\begin{array}{cc} \Sigma_{\underline{X}}^{-1} & 0 \\ 0 & \Sigma_{\underline{Y}}^{-1} \end{array} \right].$$

Substituting into (1.101), we get

$$
\begin{aligned}
p_{\underline{Z}}(\underline{x},\underline{y}) &= \frac{1}{\sqrt{(2\pi)^{n+m} \det \Sigma_{\underline{Z}}}} e^{-0.5(\underline{z}-\underline{m}_{\underline{Z}})^T (\Sigma_{\underline{Z}})^{-1}(\underline{z}-\underline{m}_{\underline{Z}})} \\
&= \frac{1}{\sqrt{(2\pi)^n \det \Sigma_{\underline{X}}}} \frac{1}{\sqrt{(2\pi)^m \det \Sigma_{\underline{Y}}}} e^{-0.5(\underline{z}-\underline{m}_{\underline{Z}})^T (\Sigma_{\underline{Z}})^{-1}(\underline{z}-\underline{m}_{\underline{Z}})} \\
&= \frac{1}{\sqrt{(2\pi)^n \det \Sigma_{\underline{X}}}} e^{-0.5(\underline{x}-\underline{m}_{\underline{X}})^T (\Sigma_{\underline{X}})^{-1}(\underline{x}-\underline{m}_{\underline{X}})} \frac{1}{\sqrt{(2\pi)^m \det \Sigma_{\underline{Y}}}} e^{-0.5(\underline{y}-\underline{m}_{\underline{Y}})^T (\Sigma_{\underline{Y}})^{-1}(\underline{y}-\underline{m}_{\underline{Y}})} \\
&= p_{\underline{X}}(\underline{x}) p_{\underline{Y}}(\underline{y}).
\end{aligned}
\tag{1.103}
$$

Another important property of Gaussian random vectors, which we will derive later when we deal with estimation, is that the conditional density of a Gaussian random vector, $\underline{X}$, given an observation of another Gaussian random vector $\underline{Y}$, is also Gaussian! Thus, this allows us to represent the conditional density in terms of two expectations which are readily computed: the conditional mean $E[\underline{X}|\underline{Y}]$ and the conditional covariance $\Sigma_{\underline{X}|\underline{Y}}$. Furthermore, we shall see that the conditional covariance does not depend on $\underline{Y}$, but is a constant matrix! The resulting formulas, which we will derive later in the course, are:

$$E[\underline{X} \mid \underline{Y}] = \underline{m}_{\underline{X}} + \Sigma_{\underline{XY}} \Sigma_{\underline{YY}}^{-1}(\underline{Y} - \underline{m}_{\underline{Y}})$$

$$\Sigma_{\underline{X}|\underline{Y}} = \Sigma_{\underline{X}} - \Sigma_{\underline{XY}} \Sigma_{\underline{YY}}^{-1} \Sigma_{\underline{YX}}$$

As a final note on Gaussian random vectors, since the joint density function depends only on the mean and covariance parameters, then all of the moments and other expectations must be expressible in terms of these parameters. Indeed, there are general formulas, based on using the characteristic function to obtain the moments.

1.11 Inequalities for Random Variables

In order to analyze notions of convergence of random variables, it is useful to bound the errors between the limit random variable and elements of the sequence using simple inequalities. Below, we present several of the most useful inequalities:

1.11.1 Markov inequality

Suppose that X is a non-negative random variable with known mean, and we want to obtain some bounds on the probability distribution function of X. A simple inequality is given by

$$P(X \geq a) = \int_a^\infty p_X(x)\, dx \leq E(X)/a. \tag{1.104}$$

This follows from

$$
\begin{aligned}
E(X) &= \int_a^\infty x p_X(x)\, dx + \int_0^a x p_X(x)\, dx \\
&\geq a \int_a^\infty p_X(x)\, dx = a P(X \geq a).
\end{aligned}
\tag{1.105}
$$

The above argument can be generalized as follows: Let $f(x) \geq 0$ everywhere, and let $f(x) > a > 0$ for all $x \in A$, for a subset A of the real line $\Re$. Then,

$$
\begin{aligned}
E[f(X)] &= \int_{x \in A} f(x) p_X(x)\, dx + \int_{x \notin A} f(x) p_X(x)\, dx \\
&\geq \int_{x \in A} f(x) p_X(x)\, dx \geq a \int_{x \in A} p_X(x)\, dx = a P(X \in A).
\end{aligned}
\tag{1.106}
$$

1.11.2 Chebyshev inequality

Suppose that the mean m and variance σ^2 of a random variable X are known, and we would like to bound the probability that the variable is far from its mean. The Chebyshev bound is given by

$$
P(|X - m| \geq a) \leq \frac{\sigma^2}{a^2}.
\tag{1.107}
$$

This bound is a special case of the Markov bound, since we can take the random variable $(x - m)^2$, which is nonnegative and has known variance σ^2. Then, by (1.104),

$$
P(|X - m| \geq a) = P((X - m)^2 \geq a^2) \leq \frac{\sigma^2}{a^2}.
$$

The above can be generalized for any random variable with finite higher-order moments, as

$$
P[|X - m| \geq a] = P(|X - m|^n \geq a^n) \leq \frac{E[|X - m|^n]}{a^n}
$$

or, more generally, for any real, nonnegative, even function $f(x)$ which is non-decreasing for $x > 0$, and has finite expectation. Then,

$$
P[f(X) \geq f(a)] \leq \frac{E[f(X)]}{f(a)}.
$$

1.11.3 Chernoff Inequality

Given a random variable X, we can define a new random variable Y_ϵ as:

$$
Y_\epsilon(\omega) = \begin{cases} 1 & \text{if } X(\omega) \geq \epsilon \\ 0 & \text{otherwise} \end{cases}
$$

That is, Y is the indicator random variable that $X \geq \epsilon$.

Then, for all $t \geq 0$, for all outcomes ω, the following inequality holds:

$$
e^{tX} \geq e^{t\epsilon} Y.
$$

Thus,

$$
E\left[e^{tX}\right] \geq E\left[e^{t\epsilon} Y\right] = e^{t\epsilon} P[X \geq \epsilon],
$$

which implies that

$$
P[X \geq \epsilon] \leq e^{-t\epsilon} E\left[e^{tX}\right], \quad t \geq 0.
$$

This bound can be tightened through the choice of t, as follows:

$$
P[X \geq \epsilon] \leq \min_{t \geq 0} e^{-t\epsilon} E\left[e^{tX}\right].
$$

Note that this bound requires computation of $E\left[e^{tX}\right]$, which is equivalent to computing the characteristic function of X! Thus, this bound requires extensive knowledge of the full probability density function of X, and not just its mean and variance.

1.11.4   Jensen's Inequality

A *convex* function of a continuous variable in an interval I is a function such that, for any $\alpha \in [0,1]$, any x, $y \in I$, the following is true:
$$f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha)f(y).$$

If the function f is twice differentiable, then it is convex if and only if the second derivative $\ddot{f} \ge 0$ for all $x \in I$. If it is only once differentiable, it is convex if and only if $f(x) + \dot{f}(x)(y - x) \le f(y)$ for all $x, y \in I$. Now, let X denote a random variable with probability density distributed over I, and let m denote its mean, which must be in I. Then, for any convex function f, we have
$$f(m) + \dot{f}(m)(X - m) \le f(X).$$

Taking expectation of both sides, we have
$$f(m) + \dot{f}(m)\left(E[X] - m\right) = f(m) = f\left(E\left[X\right]\right) \le E\left[f(X)\right],$$

which is known as Jensen's inequality.

A *concave* function is a function f whose negative $-f$ is convex. For concave functions, Jensen's inequality is reversed.

1.11.5   Moment Inequalities

Using Jensen's inequality, we can derive a number of inequalities involving expectations. We list some of these below.
$$E[|X + T|^r] \le c_r(E[|X|^r] + E[|Y|^r])$$

where
$$c_r = \begin{cases} 1 & \text{if } r \le 1 \\ 2^{r-1} & \text{if } r > 1 \end{cases}$$

To show this, note that the function $f(z) = z^r + (1-z)^r$ is a convex function on $(0,1)$ if $r \ge 1$, and a concave function if $r < 1$. It is symmetric about the value $z = 0.5$, and achieves its maximum or minimum at this value, taking the value 2^{1-r}. Thus, $c_r f(z) \ge 1$. Now, let $z = \frac{|x|}{|x|+|y|}$. Then,

$$c_r f(z) = c_r \frac{|x|^r + |y|^r}{(|x| + |y|)^r} \ge 1.$$

Multiplying through by the denominator gives

$$c_r(|x|^r + |y|^r) \ge (|x| + |y|)^r \ge |x + y|^r.$$

Taking expectations of both sides gives the result.

The next inequality is known as the Holder inequality; it is

$$E[|XY|] \le E[|X|^r]^{1/r} E[|Y|^p]^{1/p},$$

where $1/r + 1/p = 1$. This follows because the function $f(z) = \ln z$ is concave on the interval $(0, \infty)$, so, for $p \in [0,1]$, $x_1, x_2 > 0$, we have

$$(1-p)f(x_1) + pf(x_2) \le f((1-p)x_1 + px_2).$$

Taking exponentials of both sides yields

$$x_1^{1-p} x_2^p \le (1-p)x_1 + px_2.$$

Define

$$X_1 = \frac{|X|^r}{E[|X|^r]}, \quad X_2 = \frac{|Y|^s}{E[|Y|^s]},$$

where $1/r = 1 - p, 1/s = p$. Then, the above inequality becomes

$$\frac{|XY|}{E[|X|^r]^{1/r} E[|Y|^s]^{1/s}} \leq \frac{|X|^r}{rE[|X|^r]} + \frac{|Y|^s}{sE[|Y|^s]}.$$

Taking expectations and multiplying through, we obtain

$$
\begin{aligned}
E[|XY|] &\leq E[|X|^r]^{1/r} E[|Y|^s]^{1/s} \left(\frac{E[|X|^r]}{rE[|X|^r]} + \frac{E[|Y|^s]}{sE[|Y|^s]} \right) \\
&\leq E[|X|^r]^{1/r} E[|Y|^s]^{1/s} (1/r + 1/s) = E[|X|^r]^{1/r} E[|Y|^s]^{1/s}.
\end{aligned}
\tag{1.108}
$$

The Schwarz inequality is obtained by taking $r = s = 2$ in the Holder inequality.

The final inequality which we will show is the Lyapunov inequality. Consider the function $f(t) = \ln E[|U|^t]$, for $t \geq 0$. This is a convex function of t for any interval in which the expectation exists (the integrand is a convex function for each U, and the integral of convex functions or the sum of convex functions remains convex). Also, note that $f(0) = 0$. Thus, the slope $f(t)/t$ must increase monotonically with t. Taking exponentials maintains this monotonic property, so that $(E[|U|^t])^{1/t}$ is also a monotone function. Thus,

$$(E[|U|^t])^{1/t} \geq (E[|U|^s])^{1/s}$$

for $t \geq s \geq 0$.

Chapter 2

Sequences of Random Variables

Consider a sequence of random variables $\{x_n\}, n = 1, \ldots, \infty$, defined on a common probability space $(\Omega, \mathcal{F}, P)$. Under the analogy that random variables are functions, a natural question to ask is when would there be a random variable which would be considered a limit of the sequence? Before we can answer that question, consider some options for what we mean by a limit:

1. For every single outcome ω, the numbers $\{x_n(\omega)\}$ must approach a limit, and collectively across all outcomes, that limit is a random variable.

2. For almost every outcome ω (except for a negligible set of probability zero), the above must occur.

3. The probability distribution functions of $\{x_n\}$ must converge to a valid probability distribution.

Clearly, the above interpretations would all be useful notions of convergence, and are all important concepts of how to interpret limits of random variables. Before we begin the proper definitions, let's consider a couple of motivating examples.

Example 2.1
Let y be a random variable, selected uniformly from the interval $[0, 1]$. For $n = 1, \ldots$, define the random variable $x_n = y(1 - 1/n)$.

Example 2.2
Let y_1 be a random variable, selected uniformly from $[0, 1]$. For each $n > 1$, let y_n be a random variable, selected uniformly from the interval $[y_{n-1}, 1]$. Define

$$x_n = \begin{cases} 1 & \text{if } y_n > 1 - 1/n \\ 0 & \text{otherwise} \end{cases}$$

Consider the two sequences $\{y_n\}, \{x_n\}$. In the first example, it is clear that the sequence $\{x_n\}$ can be interpreted to converge for each outcome to whatever value y has for that outcome. However, convergence in the second example is harder to define. Consider the experiment described in this example: it is a compound experiment, where an infinite number of y_n will be generated, and each y_n will depend strongly on the values of the previous one. Intuitively, it seems that the random variables y_n are increasing towards 1, but they do so at different rates for each outcome. Furthermore, there are an uncountable sequences of outcomes which do not converge to 1, such as $y_n = 0.5(1 - 1/n)$. Does this example converge, and in what sense?

2.1 Convergence Concepts for Random Sequences

Before discussing convergence of random sequences, it is useful to recall that random variables are similar to functions; thus, let us review first the notions of convergence for sequences of functions.

Let D denote the domain of a sequence of real-valued functions $f_n(x), x \in D$, and the function f. We say that the sequence $\{f_n\}$ converges pointwise to f if and only if, for any $\epsilon > 0$, and any point $x \in D$, there is an integer $N(x, \epsilon)$ such that, for $n \geq N, |f_n(x) - f(x)| < \epsilon$. A sequence of function converges uniformly

in D to f if and only if, for any $\epsilon > 0$, there is an integer $N(\epsilon)$ such that, for any $x \in D$ and for $n \geq N$, we have $|f_n(x) - f(x)| < \epsilon$. The difference is that, in the second case, the same value of N works for all x, where in the first case, the value of N could depend on x.

The uniform convergence criterion is more concisely stated in terms of a distance norm. Let:

$$\|f\| = \sup_{x \in D} |f(x)|$$

Then, the sequence $\{f_n\}$ converges uniformly to f if and only if

$$\lim_{n \to \infty} \|f_n - f\| = 0$$

You may wonder why uniform convergence is important. The reason is essentially that it allows you to interchange certain limit operations. For example, if $\{f_n\}$ converges uniformly to f, then

$$\lim_{n \to \infty} \int_D f_n(x)dx = \int_D f(x)dx.$$

Based on the above discussion, let's define several concepts of convergence for sequences of random variables.

Definition 2.1 (Sure Convergence)
The sequence $\{x_n\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ is said to *converge surely* or everywhere to a random variable x, if, for each outcome ω, the sequence of numbers $\{x_n(\omega)\}$ converges to a limit $x(\omega)$.

Note that, in the above definition, the convergence could be at different rates for different outcomes; this is equivalent to the notion of pointwise convergence of functions. We can also define the notion of uniform convergence, by requiring that, for each $\epsilon > 0$, there would have to be an $N(\epsilon)$ such that $|x_n(\omega) - x(\omega)| < \epsilon$ for all $n > N(\epsilon)$, for all ω; the uniformity arises because $N(\epsilon)$ is the same for all outcomes ω.

Up to now, we have considered sequences of random variables as nothing more than sequences of functions, without exploiting the probability structure of the random variables. We now define some notions of convergence which take into account the probabilistic structure.

Definition 2.2 (Almost Sure Convergence)
The sequence $\{x_n\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ is said to *converge almost surely* or *almost everywhere* if, for each outcome ω except those in a set $A \in \mathcal{F}$ such that $P(A) = 0$, the sequence of numbers $\{x_n(\omega)\}$ converges to a limit $x(\omega)$. We write

$$\lim_{n \to \infty} x_n \overset{\text{a.e.}}{=} x$$

Mathematically, almost sure convergence requires that, for any given $\delta, \epsilon > 0$, there exists an $N(\epsilon, \delta)$ such that:

$$P[\cup_{n>N} \{\omega : |x_n(\omega) - x(\omega)| > \epsilon\}] < \delta$$

or equivalently,

$$P[\sup_{n>N} |x_n(\omega) - x(\omega)| < \epsilon] > 1 - \delta$$

Again, one can define the concept of uniformly almost everywhere convergence as a variation. Note that the concept of almost everywhere convergence implies that, the set of "bad" outcomes (for which some $|x_n(\omega) - x(\omega)| > \epsilon, n > N$) shrinks as N increases to a set with zero probability.

Example 2.3
Consider a probability space on the unit the interval $[0, 1]$, with a uniform probability measure. Define

$$x_n(\omega) = \begin{cases} 1 & \text{if } 2^{-n} \leq \omega < 2^{-n+1} \\ 0 & \text{otherwise} \end{cases}$$

Denote the limit as $x(\omega) = 0$ everywhere. Note that the bad sets B_n, for each n, have probability 2^{-n}. Furthermore,

$$\sum_{n=1}^{\infty} P(B_n)$$

converges. Then, for every $\delta > 0$, there exists an integer N such that

$$P[\cup_{n>N}\{\omega : |x_n(\omega) - x(\omega)| > \epsilon\}] < \delta$$

which guarantees almost sure convergence.

There are many examples of random variables which converge almost everywhere. However, in order to determine whether a particular sequence converges almost everywhere, we need to know in detail the probability law that governs the selection of ω, and the relationship between the outcome ω and the sequence. There are weaker notions of convergence, which may not require knowledge of the behavior of entire sample sequences. We list some of these below.

Definition 2.3 (Mean Square Convergence)
The sequence of random variables $\{x_n\}$ is said to *converge in the mean-square sense* to the random variable x if

$$\lim_{n\to\infty} E\left[(x_n - x)^2\right] = 0$$

We denote mean-square convergence as

$$\lim_{n\to\infty} x_n \stackrel{\mathsf{mss}}{=} x$$

Mean square convergence is also called convergence in quadratic mean. It is of great practical interest in engineering applicaions because of the interpretation of $E[(x_n - x^2)]$ as the power in the error signal, and because convergence can usually be established in terms of second-order statistics of the random variables involved. We also have a useful approach to verifying when a sequence converges in mean-square sense, without having to determine the limiting random variable first. This known as the *Cauchy Criterion*:

Definition 2.4 (Cauchy Criterion)
The Cauchy criterion for establishing mean-square convergence of a sequence of random variables $\{x_n\}$ states that $\{x_n\}$ converges in mean-square sense if and only if

$$\lim_{m,n\to\infty} E\left[(x_n - x_m)^2\right] = 0$$

Mean-square convergence does not imply that, as n increases, almost all sequences approach the limit and remain close to the limit; instead, it implies that, for each n, an increasing proportion of trajectories get close to the limit, but allows some trajectories to be far from the limit. Thus, mean-square convergence is more uniform across trajectories, but does not require that almost every trajectory converge. We illustrate the difference with two examples.

Example 2.4
Let y be selected uniformly in the interval $[0, 1]$. Define the sequence of random variables

$$x_n = e^{-n(ny-1)}$$

Note that the probability that $ny > 1$ increases to 1 as $n \to \infty$, which suggests that the x_n approach a limit of 0. As a matter of fact, for $\{\omega : y(\omega) > 0\}$ (which is an event of probability one), x_n will approach its limit of 0, and so

$$\lim_{n\to\infty} x_n \stackrel{\mathsf{a.e.}}{=} 0$$

Now, consider whether the same sequence converges in the mean-square sense. Letus use the definition:

$$
\begin{aligned}
E\left[(x_n - 0)^2\right] &= E[e^{-2n(ny-1)}] \\
&= e^{2n} \int_0^1 e^{-2n^2 y} dy = \frac{e^{2n}}{2n^2}(1 - e^{-2n^2})
\end{aligned}
\tag{2.1}
$$

This blows up as $n \to \infty$! Thus, almost sure convergence does not imply mean-square convergence.

Example 2.5
Consider a communication channel transmitting bits, and let x_n denote the random variable that the n-th bit was transmitted in error. Suppose the error mechanism in the channel is described as follows: the first bit is always in error; the next two bits have one error total, with probability distributed equally among them. The next 3 bits have one error total, distributed equally among them. The construction continues recursively as above, so that there is one error between the $m(m-1)/2$ bit and the $m(m+1)/2$ bit (right side inclusive), distributed uniformly, for each positive integer m. Note that, as $n \to \infty$, the probability that a bit is in error decreases to zero. Indeed, let's verify that this sequence converges in mean-square sense. Let $n \in (m(m-1)/2, m(m+1)/2]$. Then,

$$\lim_{n \to \infty} E[(x_n)^2] = \lim_{n \to \infty} 1/m = 0$$

so we have mean-square convergence. However, we can't have almost everywhere convergence, because, no matter how large we have n (or equivalently m), every sequence is guaranteed to have an errored bit in $[m(m-1)/2, m(m+1)/2]$ and thus, has elements that are far from zero.

Mean-square convergence has several strong implications. First, we can show that, if x is the limit in mean-square sense of $\{x_n\}$, then

$$\lim_{n \to \infty} E[x_n] = E[x]$$

This is because

$$0 \le E[x - x_n]^2 \le E[(x - x_n)^2]$$

Taking limits establishes the result. Furthermore, the limit is unique in that, if x, y are both limits in mean-square sense of $\{x_n\}$, then $P[x \ne y] = 0$.

With some additional conditions, we can obtain that mean-square convergence also implies almost everywhere convergence. Indeed, if for some $p > 0$, we have $\sum_{n=1}^{\infty} E[|x - x_n|^p] < \infty$, then the sequence is guaranteed to converge in mean-square sense and almost everywhere.

We finish this section with two weaker definitions of convergence:

Definition 2.5 (Convergence in Probability)
The sequence of random variables $\{x_n\}$ is said to *converge in probability* to the random variable x if, for any $\epsilon > 0$,

$$\lim_{n \to \infty} P\left[\{\omega \mid |x_n(\omega) - x(\omega)| > \epsilon\}\right] = 0$$

We use the following notation to denote convergence in probability:

$$\lim_{n \to \infty} x_n \overset{\mathsf{p.}}{=} x$$

Note that mean-square convergence implies convergence in probability. By the Chebyshev inequality,

$$P\left[\{\omega \mid |x_n(\omega) - x(\omega)| > \epsilon\}\right] \le \frac{E\left[(x_n - x)^2\right]}{\epsilon^2} \to 0$$

if the sequence is mean-square convergent. As in mean-square convergence, the trajectories are not required to stay close to the limit, although, for n large enough, most of them will be close to the limit.

As a final concept, we define the notion of convergence in distribution. This type of convergence does not require trajectories to remain close at all, but the resulting probability distributions must converge. Note also that convergence in probability is unique, in the sense that if x, y are both limits of the same sequence in probability, then $P[x \ne y] = 0$.

Definition 2.6 (Convergence in Distribution)
The sequence of random variables $\{x_n\}$ with probability distribution functions $P_n(x)$ is said to *converge in distribution* to the random variable x with probability distribution $P(x)$ if,

$$\lim_{n \to \infty} P_n(x) = P(x)$$

for all x at which $P(x)$ is continuous. We use the following notation to denote convergence in distribution:

$$\lim_{n \to \infty} x_n \overset{\mathsf{d.}}{=} x$$

Convergence in probability implies convergence in distribution, since the probability distribution functions are defined in terms of inequalities on the values $P([\{\omega : x_n(\omega) \le \epsilon\}])$. Thus convergence in distribution is a weaker concept. Consider the following example:

Example 2.6
Define the sequence of random variables x_n consisting of independent, identically distributed uniformly distributed random variables on the interval $[0, 1]$. Clearly, the sample sequences will not converge almost everywhere, or in mean-square sense or in probability, since each subsequent value is chosen independent of its previous ones. However, every x_n has an identical probability distribution function, so it converges trivially in distribution.

Note that a distribution function is uniquely determined by its values at points of continuity. This is because distributions are nonnegative, monotone, right-continuous and bounded by 1; thus, the number of jumps of size $1/n$ or greater is less than n. Hence, the points of discontinuity are at most countable, so that the points of continuity are dense in $(-\infty, \infty)$, and so the full distribution function can be determined from the right-continuity property.

An important property of convergence in distribution is that the moments and other statistics of the random variables also converge. That is because these statistics are defined in terms of integrals with respect to the distribution functions, which are converging.

2.2 The Central Limit Theorem and the Law of Large Numbers

The two most famous examples of convergence are the law of large numbers and the central limit theorem. We discuss these below.

Let $\{x_n\}$ denote a sequence of independent, identically distributed random variables, and define the sample mean

$$M_N = \frac{1}{N} \sum_{i=1}^{N} x_i$$

The claim is that new random sequence $\{M_n\}$ converges to the mean of the distribution of x_n. This establishes an empirical relationship for computing the mean of any random variable, by repeating independent experiments and averaging the observed values of the random variable.

Let m_x denote the mean of the random variables x_n. Then, we can compute

$$
\begin{aligned}
E[(M_n - m_x)^2] &= E[(\frac{1}{n} \sum_{i=1}^{n} x_i - m_x)^2] \\
&= E[(\frac{1}{n} \sum_{i=1}^{n} (x_i - m_x))^2] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} E[(x_i - m_x)^2] = \frac{n\sigma_x^2}{n^2}
\end{aligned}
\tag{2.2}
$$

where the last line follows from the independent, identically distributed property of the x_n, and σ_x^2 is the variance of each x_n. It is easy to see that $\lim_{n \to \infty} E[(M_n - m_x)^2] = 0$, which shows that the sequence $\{M_n\}$ converges in mean-square sense to m_x ($\lim_{n \to \infty} M_n \overset{\text{mss}}{=} m_x$). This proves the weak law of large numbers in the case that x_n has a finite variance. The more general version of the weak law of large numbers (even if there is no finite variance) is stated as:

Theorem 2.1 (Weak Law of Large Numbers)
Let $\{x_n\}$ be a sequence of independent, identically distributed random variables with finite means, and define the sequence of sample means $\{M_n\}$ as

$$M_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Then, $\{M_n\}$ converges in distribution to m_x, the mean of the random variables x_n.

We have proven a stronger statement if the random variables x_n have finite variance, that convergence is at least as strong as in the mean-square sense. The Strong Law of Large Numbers states a third result, which is:

Theorem 2.2 (Strong Law of Large Numbers)
Let $\{x_n\}$ be a sequence of independent, identically distributed random variables with finite mean m_x and finite variance. Then the sequence of sample means $\{M_n\}$:

$$M_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

converges almost everywhere to m_x.

The law of large numbers characterizes that the sample means converge to a deterministic quantity. However, it is often of interest to characterize the error. This is the purpose of the central limit theorem.

Theorem 2.3 (Central Limit Theorem)
Consider a sequence of independent, identically distributed random variables $\{x_n\}$ with finite mean m_x and finite variance σ_x^2. Denote the partial sum S_n as

$$S_n = \sum_{i=1}^{n} x_i$$

Define the new random sequence $\{y_n\}$ as

$$y_n = \frac{S_n - nm_x}{\sigma_x \sqrt{n}}$$

Then, the sequence $\{y_n\}$ converges in distribution to a Gaussian random variable with mean zero and variance 1.

The surprising part of the Central Limit Theorem is that the distribution of the individual random variables can be arbitrary. This is why Gaussian random variables are used so often in probabilistic analysis, since they approximately model sums of many independent effects.

We sketch a brief proof the the Central Limit Theorem using characteristic functions. We note that

$$y_n = \frac{1}{\sigma_x \sqrt{n}} \sum_{i=1}^{n} (x_i - m_x)$$

is also a sum of independent, zero-mean random variables. Thus, its characteristic function is given by:

$$
\begin{aligned}
\Phi_{y_n}(w) &= E[e^{jwy_n}] = E[e^{jw\frac{1}{\sigma_x \sqrt{n}} \sum_{i=1}^{n}(x_i - m_x)}] \\
&= E[\prod_{i=1}^{n} e^{jw\frac{x_i - m_x}{\sigma_x \sqrt{n}}}] \\
&= \prod_{i=1}^{n} E[e^{jw\frac{x_i - m_x}{\sigma_x \sqrt{n}}}] \\
&= (E[e^{jw\frac{x - m_x}{\sigma_x \sqrt{n}}}])^n
\end{aligned}
\tag{2.3}
$$

where the last equalities follows from the independent, identically distributed assumption. Now, we need to expand the exponential in the expresion, since, for large n, the exponent is small, and thus the exponential can be approximated by its first few terms.

$$e^{jw\frac{x - m_x}{\sigma_x \sqrt{n}}}] \approx 1 + \frac{jw(x - m_x)}{\sigma_x \sqrt{n}} - \frac{w^2(x - m_x)^2}{2\sigma_x^2 n} + \ldots \tag{2.4}$$

Keeping only the first three terms, we have

$$
\begin{aligned}
E[e^{jw\frac{x - m_x}{\sigma_x \sqrt{n}}}] &\approx 1 + \frac{jwE[x - m_x]}{\sigma_x \sqrt{n}} - \frac{w^2 E[(x - m_x)^2]}{2\sigma_x^2 n} \\
&\approx 1 - \frac{w^2}{2n}
\end{aligned}
\tag{2.5}
$$

because $E[x - m_x] = 0, E[(x - m_x)^2] = \sigma_x^2$. Thus,

$$\Phi_{y_n}(w) \approx (1 - \frac{w^2}{2n})^n$$

and, taking limits as $n \to \infty$, we get

$$\lim_{n \to \infty} \Phi_{y_n}(w) = e^{-w^2/2}$$

which is the characteristic function of a zero-mean, unit variance Gaussian random variable.

2.3   Advanced Topics in Convergence

This subsection provides additional mathematical background on convergence of sequences of random variables. In particular, concepts of sequences of events and Cauchy sequences are introduced.

A sequence of real numbers $\{x_n\}$ is a *Cauchy sequence* if, for any $\epsilon > 0$, there is an $N(\epsilon)$ such that $|x_n - x_m| < \epsilon$ for all $n, m > N$. It is a known property of the real line that all Cauchy sequences converge to a finite limit. The concept of Cauchy sequence can be generalized to any metric space, which consists of a set M and a distance metric $d(x, y), x, y \in M$, satisfying the following properties:

1. $d(x, y) > 0$ if $y \neq x$.

2. $d(x, x) = 0$

3. $d(x, y) + d(y, z) \geq d(x, z)$ for all x, y, z (Triangle inequality).

Then, a Cauchy sequence is such that for any $\epsilon > 0$, there is an $N(\epsilon)$ such that $d(x_n, x_m) < \epsilon$ for all $n, m > N$. However, for general metric spaces, Cauchy sequences are not guaranteed to converge to a limit in the space. The real numbers (and in general other Euclidean spaces) have additional properties which ensure convergence.

Cauchy sequences can be used to develop sufficient conditions for sure and almost-sure convergence. In particular, if almost everywhere, the sequence of real values $\{x_n(\omega)\}$ is a Cauchy sequence, then it converges almost everywhere to a value. Denote that value by $\{x(\omega)\}$; the question is whether the limit defined pointwise in this manner will be a random variable. The answer is affirmative, since we can write sets such as

$$\{\omega : x(\omega) < a\} = \cup_n \cap_{k \geq n} \{\omega : x_k(\omega) < a\}$$

which are countable union and intersections of events, and thus become events themselves.

We can also consider sequences of events in a probability space, as follows. let $\{A_n\}$ denote a sequence of events in $(\Omega, \mathcal{F}, P)$. The sequence is said to be increasing if $A_n \subset A_{n+1}$, and decreasing if $A_{n+1} \subset A_n$. It is monotone if it is either decreasing or increasing. For any monotone sequence, the limit is defined as

$$\lim_{n \to \infty} A_n = \begin{cases} \cup_{n=1}^{\infty} A_n & \text{if increasing} \\ \cap_{n=1}^{\infty} A_n & \text{if decreasing} \end{cases}$$

which is guaranteed to be an event. For general, non-monotone sequences, we define sup and inf limits as

$$\limsup_{n \to \infty} A_n = \cap_{n=1}^{\infty} \cup_{k \geq n} A_k$$

$$\liminf_{n \to \infty} A_n = \cup_{n=1}^{\infty} \cap_{k \geq n} A_k$$

The sup limit is the set of all outcomes which occur infinitely often, while the inf limit is the set of all outcomes which occur in all A_n except for a finite number. If the two coincide, we say the sequence has a limit.

One of the important properties of probability measures is that they are sequentially continuous with respect to limits of events. That is, if $\{A_n\}$ converges, then

$$\lim_{n \to \infty} P(A_n) = P(\lim_{n \to \infty} A_n)$$

This has the following famous lemma as a consequence:

Theorem 2.4 (Borel-Cantelli Lemma)

For an arbitrary sequence of events $\{A_n\}$, if $\sum_{n=1}^{\infty} P(A_n) < \infty$, then

$$P(\limsup_{n \to \infty} A_n) = 0.$$

The Borel-Cantelli Lemma is used primarily in proving that certain properties occur with probability one. The proof is straightforward, as

$$
\begin{aligned}
P(\limsup_{n \to \infty} A_n) &= P(\lim_{n \to \infty} \cup_{k \geq n} A_k) \\
&= \lim_{n \to \infty} P(\cup_{k \geq n} A_k) \leq \lim_{n \to \infty} \sum_{k=n}^{\infty} P(A_k) = 0
\end{aligned}
\tag{2.6}
$$

which is zero due to the summability assumption.

One of the key questions in convergence is determining conditions for when convergence in probability is equivalent to mean-square convergence or almost sure convergence. The following theorem, due to Loeve, characterizes convergence in probability:

Theorem 2.5

Let $\{x_n\}$ be a sequence of random variables. Then

1. If the sequence converges almost surely, then it converges in probability to the same limit.

2. If the sequence converges in probability, then there is a subsequence $\{x_{n_k}\}$ which converges almost surely to the same limit.

The first part of the theorem is straightforward. The second part is cumbersome to prove; if interested, see Loeve's book on Probability Theory.

The Borel-Cantelli lemma is useful in establishing the following theorem due to Gnedenko, which gives sufficient conditions for a sequence of random variables to converge almost surely:

Theorem 2.6

Suppose that, for a sequence of random variables $\{x_n\}$, for every positive integer r, we have

$$\sum_{n=1}^{\infty} P(\{\omega : |x_n - x_m| \geq 1/r\}) < \infty$$

Then, the sequence converges almost surely.

Clearly, the sequence converges in probability, since the conditions imply

$$\lim_{n \to \infty} \sup_{m \geq n} P(\{\omega : |x_n - x_m| \geq 1/r\}) = 0$$

Thus, there is a limiting random variable x. Define

$$A_n^r = \{\omega : |x_n - x| \geq 2/r\}$$

Since $A_n^r \subset \{\omega : \max(|x_n - x_m|, |x - x_m|) \geq 1/r\}$, we have

$$P(A_n^r) \leq P(|x_n - x_m| \geq 1/r) + P(|x - x_m| \geq 1/r)$$

Letting $m \to \infty$, we have

$$P(A_n^r) \leq \sup_{m \geq n} P(|x_n - x_m| \geq 1/r)$$

The conditions imply that the right-hand side is summable, so the left hand side must be summable too. The Borel-Cantelli lemma thus implies that $P(\limsup_{n \to \infty} A_n^r) = 0$, so that, for every r, $|x_n - x| > 2/r$ for at most a finite number of n, almost surely. If we define $A = \cup_{r=1}^{\infty} \limsup_{n \to \infty} A_n^r$, we see that $P(A) = 0$, and outside of A, we have $\lim_{n \to \infty} |x_n(\omega) - x(\omega)| = 0$. Thus, we have almost sure convergence.

We can also provide conditions whereby convergence in probability also implies convergence in mean square sense. The following theorem is due to Loeve:

Theorem 2.7

If the sequence $\{x_n\}$ converges to x in probability, then it converges in mean-square sense if one of the following conditions holds:

1. $\lim_{n \to \infty} E[|x_n|^2] = E[x^2] < \infty$

2. The $|x_n|^2$ variables have a uniform expectation; that is, there exists, for every $\epsilon > 0$, a value $K(\epsilon)$ such that

$$\int_B x_n(\omega) P(d\omega) < \epsilon$$

for any set $B = \{\omega : |x_n(\omega)| \geq K(\epsilon)\}$.

To tie these concepts together, convergence in mean-square sense implies condition 2 in the theorem above. In particular, there are some simple conditions which can guarantee convergence in mean-square sense when the sequence already converges in probability:

1. $\sup_n E[|x_n|^2] = c < \infty$.

2. $|x_n| < y$ for $n > N$, and $E[y^2] < \infty$.

As a final topic in convergence, let's focus on convergence in distribution. In particular, consider a sequence $\{x_n\}$ of random variables with probability distribution functions $P_n(x)$, and characteristic functions $\Phi_n(w)$. The following results are standard, and can be found in most advanced probability books:

1. If, for every bounded, continuous function $g : \mathcal{R} \to \mathcal{R}$, we have $\lim_{n \to \infty} E[g(x_n)] = E[g(x)]$, then the sequence converges in distribution to x.

2. If, for every real number w, the characteristic function $\Phi_n(w)$ converges pointwise to $\Phi(w)$, the sequence converges in distribution to x.

3. If, for any two values x_1, x_2, we have $P_n(x_1) - P_n(x_2)$ converges to $P(x_1) - P(x_2)$, the sequence converges in distribution to x.

In particular, the equivalence between the convergence of distribution functions and the convergence of characteristic functions pointwise is known as the continuity theorem of probability. This name is derived from the fact that the 1-1 correspondence between distributions and characteristic functions is preserved by the limit operation. That is, the limit of the characteristic functions is the characteristic function of the limit.

Using the additional fact that, if the limit of characteristic functions is continuous at $w = 0$, then a distribution function corresponding to this limit exists, we have a new result: A necessary and sufficient condition for convergence in distribution, such that

$$\lim_{n \to \infty} P_n(x) = P(x)$$

at all points of continuity of $P(x)$, is that the corresponding sequence of characteristic functions converges to a characteristic function which is continuous at $w = 0$.

There is a special case where convergence in probability and distribution are equivalent: when the limit random variable is a constant, such as 0. In such cases, the limiting distribution is a step function, switching from zero to 1 instantly. Then,

$$\lim_{n \to \infty} P_n(x) = u(x - c)$$

where u is the unit step function, and c is the limiting constant. Thus,

$$\lim_{n \to \infty} P[|x_n - c| > \epsilon] = 0$$

for any $\epsilon > 0$. This implies convergence in probability.

As a final result, consider when convergence in distribution implies convergence of the probability densities to the probability density of the limit. The following condition is sufficient: if $\Phi_n(w)$ and $\Phi(w)$ are absolutely integrable (i.e. $\int_{-\infty}^{\infty} |\Phi(w)| dw < \infty$), and

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} |\Phi_n(w) - \Phi(w)| dw < \infty$$

then the density functions $p_n(x)$ converge uniformly to the density $p(x)$. The integrability assumption guarantees that the densities are bounded and continuous, and defined by the inverse Fourier transform. pointwise

2.4 Martingale Sequences

In the previous subsection, we discussed conditions whereby we could establish almost sure convergence for sequences which converge only in probability. In this subsection, we present a special class of sequences of random variables, called a *martingale sequence*, for which stronger results can be established.

Definition 2.7
A sequence of random variables $\{x_n\}$ s called a *martingale* if

$$E[x_n|x_0, \ldots, x_{n-1}] = x_{n-1}$$

almost everywhere for all $n > 1$.

Thus, a martingale has the property that increments $x_{n+1} - x_n$ are zero-mean, conditioned on x_n. Martingales arise naturally in the study of sequences which are partial sums of independent random variables. For instance, in the law of large numbers, it is clear that the partial sums $s(n)$ form a martingale, since

$$
\begin{aligned}
E[s(n)|s(0), \ldots, s(n-1)] &= E[x_n + s(n-1)|x_0, \ldots, x_{n-1}] \\
&= E[s(n-1)|x_0, \ldots, x_{n-1}] = s(n-1) \quad (2.7)
\end{aligned}
$$

By the above argument, any sequence with independent, zero-mean increments will form a martingale. Martingales have the following properties

1. $E[x_n] = E[x_0]$.

2. $E[x_{n+m}x_n] = E[x_n^2]$ if $m \geq 0$.

3. $E[x_n(x_{n+m} - x_n)] = 0$ for $m > 0$.

4. $E[x_{n+m}^2] \geq E[x_n^2]$

5. $E[(x_{n+m} - x_n)^2] \geq 0$

6. For any $m \geq 0$, the sequence $y_n = x_{n+m} - x_m$ is a zero-mean martingale.

All of the above are easily established from the martingale definition, by using the smoothing property of conditional expectations. To illustrate this, we demonstrate (3) above:

$$
\begin{aligned}
E[x_{n+m}^2] &= E[(x_{n+m} - x_n + x_n)^2] \\
&= E[x_n^2] + E[(x_{n+m} - x_n)^2] - 2E[x_n(x_{n+m} - x_n)] \\
&\geq E[x_n^2] - 2E[x_n(x_{n+m} - x_n)] \quad (2.8)
\end{aligned}
$$

To complete the demonstration, use the smoothing property to show

$$
\begin{aligned}
E[x_n(x_{n+m} - x_n)] &= E[E[x_n(x_{n+m} - x_n)|x_n]] \\
&= E[x_n(E[x_{n+m}|x_n] - x_n)] = 0 \quad (2.9)
\end{aligned}
$$

since $E[x_{n+m}|x_n] = x_n$ by the definition of martingales.

The importance of martingales is that we can establish a useful bound on the convergence of a martingale to its limit. The following theorem provides such a result, similar to the Chebyshev inequality for random variables:

Theorem 2.8

For a martingale $\{x_n\}$, given any $\epsilon > 0$, any n, we have

$$P\left[\max_{0 \leq k \leq n} |x_n| \geq \epsilon\right] \leq \frac{E[x_n^2]}{\epsilon^2}$$

A proof of this result goes as follows: Construct the sets A_j as

$$A_j = \{\omega : |x_j| \geq \epsilon, |x(k)| < \epsilon \text{ for } k < j.\}$$

Then,

$$\{\max_{0 \leq k \leq n} |x_n| \geq \epsilon\} = \cup_{i=0}^n A_j$$

Let I_j be the indicator function of A_j; then

$$
\begin{aligned}
E[x_n^2] &\geq E[x_n^2 \sum_{j=1}^n I_j] \\
&\geq E[\sum_{j=1}^n x_n^2 I_j] \\
&\geq E[\sum_{j=1}^n (x_n - x_j + x_j)^2 I_j] \\
&\geq E[\sum_{j=1}^n x_j^2 I_j] + 2E[\sum_{j=1}^n x_j(x_n - x_j) I_j] \quad (2.10)
\end{aligned}
$$

Now, using the smoothing property of conditional expectations, we have

$$
\begin{aligned}
E[\sum_{j=1}^n x_j(x_n - x_j) I_j] &= \sum_{j=1}^n E[E[x_j(x_n - x_j) I_j | x(0), \ldots, x_j]] \\
&= 0 \quad (2.11)
\end{aligned}
$$

because the only random quantity in the inner expectation, conditioned on knowing $x(0), \ldots, x_j$, is $x_n - x_j$, and the martingale property guarantees that this has zero conditional mean. Thus,

$$
\begin{aligned}
E[x_n^2] &\geq \sum_{j=1}^n E[I_j x_j^2] \\
&\geq \sum_{j=1}^n \epsilon^2 E[I_j] = \sum_{j=1}^n \epsilon^2 P(A_j) \\
&= \epsilon^2 P(\cup_{0 \leq j \leq n} A_j) = \epsilon^2 P(\max_{0 \leq k \leq n} |x_n| \geq \epsilon) \quad (2.12)
\end{aligned}
$$

Based on the above bound, we can show the following theorem:

Theorem 2.9 (Martingale Convergence Theorem)

Let $\{x_n\}$ be a martingale such that

$$E[x_n^2] \leq C < \infty$$

for all n. Then, $\{x_n\}$ converges almost surely to a random variable x with finite variance.

Proof: Based on the martingale property, we know that $E[x_n^2]$ is a monotone nondecreasing sequence of n, and is bounded above by the assumption, so that it has a limit. Since this has a limit, then

$$\lim_{m,n \to \infty} E[(x_{n+m} - x_m)^2] = 0$$

which shows immediately mean-square convergence. To show almost sure convergence, we use the martingale inequality to obtain

$$P[\max_{n \geq k \geq 0} |x_{n+m} - x_m| \geq \epsilon] \leq E[(x_{n+m} - x_m)^2]$$

so that

$$\lim_{m \to \infty} P[\max_{k \geq 0} |x_{n+m} - x_m| \geq \epsilon] = 0$$

Since probabilities are continuous as a function of events, this also implies

$$P[\lim_{m \to \infty} \max_{k \geq 0} |x_{n+m} - x_m| \geq \epsilon] = 0$$

which makes this a Cauchy sequence of numbers, almost surely, so that a limit random variable x exists, and convergence is almost sure.

A simple application of the above result is the Strong Law of Large Numbers. Define

$$y_n = \sum_{i=1}^{n} \frac{x_i - m_x}{i} \tag{2.13}$$

This is a martingale, since it has independent increments. Furthermore,

$$E[y_n^2] = \sum_{i=1}^{n} \frac{\sigma_x^2}{i^2} = \sigma_x^2 \sum_{i=1}^{n} \frac{1}{i^2} \tag{2.14}$$

which is bounded above, and thus satisfies the conditions of the martingale convergence theorem. Thus, $\lim_{n \to \infty} y_n \overset{\text{a.e.}}{=} y$ for some random variable y with finite variance. Next, note that we can write the sequence of partial sums in terms of the y_n, since $x_i - m_x = i(y_i - y_{i-1})$, where $y(0) = 0$:

$$
\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} (x_i - m_x) &= \frac{1}{n} \sum_{i=1}^{n} i y_i - \frac{1}{n} \sum_{i=1}^{n} i y_{i-1} \\
&= y_n + \frac{1}{n} \sum_{i=1}^{n-1} (i - 1 - i) y_i \\
&= y_n - \frac{1}{n} \sum_{i=1}^{n-1} y_i \tag{2.15}
\end{aligned}
$$

Taking limits,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (x_i - m_x) = \lim_{n \to \infty} y_n - \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n-1} y_i = 0$$

almost surely, which proves the strong law of large numbers.

2.5 Extensions of the Law of Large Numbers and the Central Limit Theorem

Now that we have the mechanisms for analyzing convergence of random sequences, we can state the stronger versions of the Law of Large Numbers and the Central Limit Theorem. In particular, we want to relax the independent, identically distributed assumption which seemed to play such a crucial role. We begin by recalling the definition of the sequence of partial sums

$$s(n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - E[x_i])$$

A sufficient condition for convergence in probability to a limit is that $\lim_{n \to \infty} E[s(n)^2] = 0$. Note that this does not require independence, or identical distributions. If the random sequence $\{x_n\}$ has the property that each x_n has finite variance $\sigma^2_{x_n}$, and the sequence is pairwise uncorrelated, then

$$E[s(n)^2] = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2_{x_n}$$

A sufficient condition for the above is Kolmogorov's condition that

$$\lim_{n \to \infty} \sum_{i=1}^{n} \sigma^2_{x_n}/n < \infty$$

To extend this to almost sure convergence, we use the martingale convergence theorem argument as in the previous subsection. Assume that we have a sequence of independent random variables, satisfying the Kolmogorov condition. Using the definition of y_n in (2.13), equation (2.14) becomes

$$
\begin{aligned}
E[y_n^2] &= \sum_{i=1}^{n} \frac{\sigma^2_{x_i}}{i^2} \\
&\leq \left(\sum_{i=1}^{n} \sigma^2_{x_i} \right) \left(\sum_{i=1}^{n} \frac{1}{i^2} \right) \\
&\leq \frac{K}{n} \sum_{i=1}^{n} \sigma^2_{x_i}
\end{aligned}
\tag{2.16}
$$

for some constant K, which is bounded as $n \to \infty$ by assumption. Thus, y_n satisfies the conditions used to prove the strong law of large numbers above.

Similar relaxations of the central limit theorem are possible. Consider a sequence of random variables $\{x_n\}$ with finite mean m_{x_n} and finite variance $\sigma^2_{x_n}$. Denote the partial sum S_n as

$$S_n = \sum_{i=1}^{n} (x_i - m_{x_i})$$

Let σ_n^2 denote the covariance of S_n. The extensions of the central limit theorem state that the new random sequence $\{y_n\}$ as

$$y_n = \frac{S_n}{\sigma_n}$$

Then, the sequence $\{y_n\}$ converges in distribution to a Gaussian random variable with mean zero and variance 1.

One of the strongest extensions is due to Lindeberg, and is summarized below:

Theorem 2.10

If the elements of the sequence $\{x_n\}$ are independent, then the central limit theorem holds if

$$\lim_{n \to \infty} \frac{\sum_{i=1}^{n} \int_{|x - m_{x_i}| > \epsilon \sigma_n} (x - m_{x_i})^2 p_n(x) \, dx}{\sigma_n^2} = 0$$

for any $\epsilon > 0$.

The Lindeberg conditon is also necessary as long as $\lim_{n \to \infty} \sigma_n = \infty$. Assuming independence,

$$\sigma_n^2 = \sum_{i=1}^{n} (x_i - m_{x_i})^2$$

and, as long as the contributions of the individual terms $(x_i - m_{x_i})^2$ vanish sufficiently slow, then $\lim_{n \to \infty} \sigma_n = \infty$.

Proving the Lindeberg theorem is complex; Loeve has a proof in his book. Instead, we focus on a simpler result:

Theorem 2.11
If the elements of the sequence $\{x_n\}$ are independent, then the central limit theorem holds if

$$C_1 \leq E[(x_n - m_{x_n})^2] \leq C_2$$

for all n.

Note that this implies $C_2 n \geq \sigma_n^2 \geq C_1 n$. The upper bound also implies that, as $n \to \infty$, the terms

$$\int_{|x-m_{x_i}|>\epsilon\sigma_n} (x - m_{x_i})^2 p_n(x) dx$$

in the numerator in the Lindeberg condition are decaying to zero, while the denominator grows at least as nC_2. Since there are only n terms in the numerator, their sum, divided by the denominator, will decay to 0. This establishes that the Lindeberg condition is satisfied. In practice, what the Lindeberg condition requires is that each term $\frac{x_n - m_{x_n}}{\sigma_n}$ be uniformly small, so that the sum of the terms is not dominated by a finite subset of the terms, but instead is the sum of many individually negligible components.

2.6 Spaces of Random Variables

In order to get geometric insight into several of the basic operations used in this course, it is useful to understand how collections of random variables resemble the normal n-dimensional Euclidean spaces which we use in normal vector operations.

To that end, consider the collection of zero-mean, finite-variance random variables. It is clear that, if x, y belong to this collection, then, for any real numbers a, b, then $ax + by$ is also in the collection, since

$$E[ax + b] = aE[x] + bE[y] = 0$$

$$E[(ax + b)^2] = a^2 E[x^2] + b^2 E[y^2] + 2abE[xy] \leq a^2 E[x^2] + b^2 E[y^2] + 2|ab|(E[x^2]E[y^2])^{1/2} < \infty$$

Thus, like vectors, linear combinations of vectors are also in the same space. Also, we have addition and scalar multiplication defined in the collection of zero-mean, finite-variance random variables. Other properties of this collection, which are similar to vectors, include:

1. There is a zero random variable, which, when added to every other random variable, is an additive identity. Define the random variable $x(\omega) = 0$ for all $\omega \in \Omega$. Then, for any other random variable y, $x + y = y$.

2. For every random variable x, there is a second random variable in the collection, y, such that $x + y = 0$. In esssence, define $y(\omega) = -x(\omega)$; it is clear that y is also zero-mean, and has finite variance.

The above properties guarantee that the collection of zero-mean, finite variance random variables is a linear vector space, or a vector space for short. We now carry the analogy a little further: we define the concept of an inner product among vectors, or among random variables, as follows: Given any two zero-mean, finite variance random variables x, y, the inner product of x, y, denoted by $< x, y >$, is given by

$$< x, y >= E[xy]$$

Note that the above expectation is guaranteed to exist, and thus always assigns a real number to the inner product of two variables.

The inner product operation satisfies the following conditions:

1. $< x, x > \geq 0$

2. $< ax + by, z >= a < x, z > +b < y, z >$ for any real numbers a, b and random variables x, y, z.

3. $< z, ax + by >= a < z, x > +b < z, y >$

Indeed, this almost statisfies the concepts of inner product operations in Euclidean space. The one difference is that, for Euclidean vectors, $< x, x >> 0$ if $x \neq 0$, whereas we cannot say that unequivocably for zero-mean random variables. In essence, we can have some random vectors which are non-zero on sets of probability measure zero, for which $< x, x >= 0$.

The trouble is that our random variables are too many, and that two random variables can be essentially equivalent (equivalent almost everywhere), but treated as different random variables in our collection. Formally, what we have to do is to define equivalence classes of random variables, where two random variables are said to be equivalent (written as $x \equiv y$, or $x = y$ $a.e.$), if $x = y$ almost everywhere (i.e. except for a set of zero probability).

Thus, for the space of zero-mean, finite variance equivalence classes of random variables, we have the property $< x, x >> 0$ if $x \neq 0$ a.e., which is the same property satisfied by the inner product in Euclidean spaces. Thus, the linear vector space of random variables also has an inner product structure defined on it; furthermore, the inner product defines a norm $\|x\| = < x, x >^{1/2}$, and satisfies the triangle inequality:

$$< x + y, x + y >^{1/2} \leq < x, x >^{1/2} + < y, y >^{1/2}$$

Another important property of this space was established above: given any Cauchy sequence of elements in this space $\{x_n\}$ (that is, a sequence satisfying the Cauchy criterion), there is a random variable x which has zero-mean, finite variance, to which the sequence converges in mean-square sense. Note that convergence in mean-square sense is equivalent to convergence in the norm defined above, as

$$\|x - x_n\|^2 = < x - x_n, x - x_n >= E[(x - x_n)^2]$$

which is the same metric used to define mean-square sense convergence. This important property implies that all of the limits of Cauchy sequences of zero-mean, finite variance random are also zero-mean, finite variance random variables, so that this is a *complete* space. Such a vector space (a complete vector space with an inner product structure) is called a *Hilbert* space, and has mathematical properties very similar to standard n-dimensional Euclidean spaces.

An important subspace of this space is the space of zero-mean Gaussian random variables. As before, this space is closed under linear operations of addition and scalar multiplication, since the sum of Gaussian random variables is also Gaussian. Furthermore, the space is also closed under limits of Cauchy sequences with the above metric, since these Cauchy sequences converge in mean-square sense, and thus in distribution, so that the limit will also be a Gaussian random variable. Thus, the space of Gaussian random variables with zero mean and finite variance forms a closed subspace of the space of all random variables with zero mean and finite variance. The above spaces will be used extensively in the solution of estimation problems.

Chapter 3

Stochastic Processes and their Characterization

3.1 Introduction

Consider a random experiment in a probability space $(\Omega, \mathcal{F}, P)$ where, for every outcome $\omega \in \Omega$, we assign a real-valued function of time $X(t, \omega), t \in I$ according to some rule, for t in some totally ordered index set I. For most of our applications, the set I will either be the set of integers, or the set of real numbers. This collection of functions, indexed by outcomes of a probability space, is called a *stochastic process* or a random process. The index set can be continuous (e.g. the real numbers), in which case we say it is a continuous-time process, or discrete (e.g. the integers), in which case it is a discrete-time process. For a particular outcome ω, the function $X(t, \omega), t \in I$ can be thought of as a deterministic signal, and it is called a *realization* of the process.

We get a different view of stochastic processes if we fix a particular time index t, and look at the collection $X(t, \omega), \omega \in \Omega$. For each $t \in I, X(t, \omega)$ is a random variable. Thus, a stochastic process can also be viewed as an indexed collection of random variables, where the index corresponding to time is either discrete or continuous.

Example 3.1
Let $(\Omega, \mathcal{F}, P)$ be a probability space where $\Omega = [0, 1)$, $\mathcal{F}$ is the Borel Field over $[0, 1)$, and P is given by a uniform density function. For each $\omega \in \Omega$, let $X(n, \omega)$ be the n-th bit in the binary expansion of ω. For example, $X(1)$ is a random variable that takes on value 0 when $\omega \in [0, 0.5)$ and value 1 otherwise; $X(2)$ is a random variable that takes on value 0 when $\omega \in [0, 0.25) \cup [0.5, 0.75)$ and value 1 otherwise. $X(\cdot)$ is a discrete-time random process.

The random process can also be defined in terms of another random variable, as in the example below. Note that we will frequently drop the ω (or other variable) dependence in describing the random process, just as we did for random variables.

Example 3.2
Let Y be selected at random according to an exponential distribution with parameter α. Define the continuous-time random process

$$X(t) = Y \cos(t), \quad -\infty < t < \infty.$$

Note that $X(0)$ is a random variable described by an exponential distribution with parameter α, and $X(\pi/3)$ is a random variable described by an exponential distribution with parameter 2α.

Although the majority of this course is concerned with the above (scalar) definition of a stochastic process, it is easy to generalize the definition to the vector case: a *vector stochastic process* is a collection of random vectors, with values in $\mathcal{R}^n$, indexed by $t \in I$.

3.2   Complete Characterization of Stochastic Processes

We have already considered the characterization and properties of a *finite* collection of random variables, which are essentially random vectors. In particular, we characterized random vectors in terms of their joint probability distribution function. Our approach to obtaining a complete characterization of stochastic processes will build on this approach.

Consider a set of sampling times $t_1, t_2, \ldots, t_k \in I$, and let $X_i = X(t_i)$ denote the random variable obtained by fixing the value of the process at each time $t_i, i = 1, \ldots, k$. For any finite value k, we have a vector of random variables $\underline{X} = [X_1 \cdots X_k]^T$, and we can completely specify this vector through specification of its joint probability distribution function:

$$
\begin{aligned}
P_{\underline{X}}(x_1, \ldots, x_k) &= P(\{\omega : X_1(\omega) \le x_1, \ldots, X_k(\omega) \le x_k\}) \\
&= P(\{\omega : X(t_1, \omega) \le x_1, \ldots, X(t_k, \omega) \le x_k\} \\
&= P_{X(t^k)}(x_1, x_2, \ldots, x_k),
\end{aligned} \tag{3.1}
$$

where we use the notational abbreviation $X(t^k) = [X(t_1) \cdots X(t_k)]^T$. Note that this joint probability distribution function is defined for any valid finite collection of time indices $t_1, t_2, \ldots, t_k \in I$. If the stochastic process is continuous-valued, it is often easier to speak of the joint probability density function $p_{X(t^k)}(x_1, x_2, \ldots, x_k)$. We also use the notation $p_X(x_1, x_2, \ldots, x_n; t_1, \ldots, t_n)$ to denote a joint density of random variables sampled at specific times. Similarly, for discrete-valued stochastic processes, it may be more convenient to work with the joint probability mass function. Now a complete characterization of a random process can be obtained through the specification of the complete set of k-th order finite-dimensional densities in (3.1). That is, specification of (3.1) for all orders k and all possible sets of sampling points t_j is a complete characterization of a stochastic process.

At first glance, the specification of the complete set of joint probability distribution functions seems like an enormous task, as we must specify the properties of all possible subsets of time indices. However, the mechanism for generating different random processes often makes it simple to specify the set of probability distribution functions of interest. Furthermore, we often care only about first- and second-order moments; we define these in the next subsection.

3.3   First and Second-Order Moments of Stochastic Processes

The moments of time samples of a random process can be used to summarize the information in the joint probability distribution function. We define these moments as follows.

The *mean* of a random process $X(\cdot)$, denoted by $m_X(t)$ is the time function defined by

$$
m_X(t) = E[X(t)] = \int_{-\infty}^{\infty} x \, p_X(x; t) \, dx
$$

for a continuous-valued process where $p_X(x; t)$ is the density of the random variable $X(t)$, or

$$
m_X(t) = E[X(t)] = \sum_{x=-\infty}^{\infty} x \, p_X(x; t)
$$

for a discrete-valued process. (Note that the use of summation vs. integral in computing an expectation for a fixed time (or set of times) in a random process depends on the values that the associated variables take on, not whether the time index is discrete vs. continuous.)

The *autocorrelation* function of a random process $X(\cdot)$, denoted by $R_X(s, t)$ is defined as the joint moment of $X(s)$ and $X(t)$:

$$
R_X(s, t) = E[X(s)X(t)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 \, p_X(x_1, x_2; s, t) \, dx_1 \, dx_2
$$

For the discrete-valued random process, the double integral would be replaced by a double sum.

The *autocovariance* function of a random process $X(\cdot)$, denoted by $K_X(s,t)$, is defined as the covariance of $X(s)$ and $X(t)$ as

$$
\begin{aligned}
K_X(s,t) &= E[(X(s) - m_X(s))(X(t) - m_X(t))] \\
&= R_X(s,t) - m_X(s)m_X(t)
\end{aligned}
\tag{3.2}
$$

Note that the variance of $X(t)$ is given by $K_X(t,t)$.

The mean, autocovariance and autocorrelation functions of a random process are only partial descriptions of the process. There can be many different random processes with the same mean and autocorrelation functions.

It is often useful to characterize the relationship between two random processes. The *cross-correlation* function of random processes $X(\cdot)$ and $Y(\cdot)$, denoted by $R_{XY}(s,t)$ is defined as the joint moment of $X(s)$ and $Y(t)$:

$$
R_{XY}(s,t) = E[X(s)Y(t)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy p_{XY}(x,y;s,t) dx dy
$$

Again, for the discrete-valued random process, the double integral would be replaced by a double sum. Similarly, the *cross-covariance* function of random processes $X(\cdot)$ and $Y(\cdot)$, denoted by $K_{XY}(s,t)$, is defined as the covariance of $X(s)$ and $Y(t)$ as

$$
\begin{aligned}
K_{XY}(s,t) &= E[(X(s) - m_X(s))(Y(t) - m_Y(t))] \\
&= R_{XY}(s,t) - m_X(s)m_Y(t)
\end{aligned}
\tag{3.3}
$$

3.4 Special Classes of Stochastic Processes

In this subsection, we discuss special classes of random processes for which it is relatively simple to specify the joint probability distribution function for any set of times.

Definition 3.1 (Independent and Identically Distributed Process)
A discrete-time stochastic process is said to be independent and identically distributed (i.i.d.) if the joint distribution for any sampling times $n_1, \ldots, n_k$ can be expressed as the product of the first order marginal distribution:

$$
p_X(x_1, \ldots, x_k; n_1, \ldots, n_k) = \prod_{i=n_1}^{n_k} p_X(x_i),
$$

where the first order marginal $p_X(x;n) = p_X(x)$ is independent of time.

The i.i.d. process is perhaps the simplest possible class, since it can be specified completely in terms of a scalar density or mass function.

Definition 3.2 (Gaussian Stochastic Process)
A stochastic process is said to be Gaussian if the samples $X(t_1), \ldots, X(t_k)$ are jointly Gaussian random vectors for any sampling times $t_1, \ldots, t_k$.

Recall that the probability density function of jointly Gaussian random variables is determined by the vector of means and by the covariance matrix, as

$$
p_X(x_1, \ldots, x_k; t_1, \ldots, t_k) = \frac{e^{-1/2(\underline{x} - \underline{m}_X)^T \Sigma_X^{-1}(\underline{x} - \underline{m}_X)}}{(2\pi)^{k/2}(|\det \Sigma_X|)^{1/2}},
$$

where

$$
\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} \quad
\underline{m}_X = \begin{bmatrix} m_X(t_1) \\ \vdots \\ m_X(t_k) \end{bmatrix} \quad
\Sigma_X = \begin{bmatrix} K_X(t_1,t_1) & K_X(t_1,t_2) & \cdots & K_X(t_1,t_k) \\ \vdots & \vdots & & \vdots \\ K_X(t_k,t_1) & K_X(t_k,t_2) & \cdots & K_X(t_k,t_k) \end{bmatrix}.
$$

Thus, Gaussian random processes are completely specified by the process mean $m_X(t)$ and autocovariance function $K_X(t_1, t_2)$. Furthermore, Gaussian processes have additional properties which make them particularly useful in the analysis of linear systems driven by stochastic processes, as we will see in later sections.

Definition 3.3 (Independent Increments)
A stochastic process $X(t)$ is an *independent increments process* if for all $s < t$ the random variables $X(t) - X(s)$ and $X(\tau)$ are independent for any $\tau \leq s$.

Note, in particular, that this implies that if $t_1 \leq t_2 \leq \cdots$ then the increments of the process $X(t_2) - X(t_1)$, $X(t_3) - X(t_2)$, $\cdots$ are independent. This property makes it easier to compute the joint probability mass function, as follows:

$$
\begin{aligned}
p_X(x_1, \ldots, x_k; t_1, \ldots, t_k) &= \Pr[X(t_1) = x_1, \ldots, X(t_k) = x_k] \\
&= \Pr[X(t_1) = x_1, X(t_2) - X(t_1) = x_2 - x_1, \ldots, X(t_k) - X(t_{k-1}) = x_k - x_{k-1}] \\
&= \Pr[X(t_1) = x_1]\Pr[X(t_2) - X(t_1) = x_2 - x_1] \cdots \Pr[X(t_k) - X(t_{k-1}) = x_k - x_{k-1}]
\end{aligned}
$$

due to the independent increments property. Similarly, joint density functions or distribution functions can be computed in terms of a first-order marginal density or distribution function and the distributions of the independent increments.

Definition 3.4 (Markov Process)
A stochastic process $X(\cdot)$ is said to be *Markov* if the future of the process is independent of its past, conditioned on the present value of the process. That is, for any choice of sampling instances $t_1 < t_2 < \ldots < t_k$,

$$
\Pr(X(t_k) = x_k | X(t_1) = x_1, \ldots, X(t_{k-1}) = x_{k-1}) = \Pr(X(t_k) = x_k \mid X(t_{k-1}) = x_{k-1})
$$

The above equation states that $X(t_k)$ is independent of $X(t_1), \ldots, X(t_{k-2})$, conditioned on knowing the value of $X(t_{k-1})$. Note that an independent increments process is necessarily Markov; however, Markov processes are not necessarily independent increments processes.

The Markov property makes it simpler to compute joint probability density (or mass) functions, as follows:

$$
\begin{aligned}
p_X(x_1, \ldots, x_k; t_1, \ldots, t_k) &= p_X(x_1; t_1)p_X(x_2; t_2|x_1; t_1) \cdots p_X(x_k; t_k|x_1, \ldots, x_{k-1}; t_1, \ldots, t_{k-1}) \\
&= p_X(x_1; t_1)p_X(x_2; t_2|x_1; t_1) \cdots p_X(x_k; t_k|x_{k-1}; t_{k-1}) \\
&= p_X(x_1; t_1) \prod_{i=2}^{k} p_X(x_i; t_i|x_{i-1}; t_{i-1})
\end{aligned}
$$

Thus, Markov processes can be characterized by a marginal probability density $p_X(x_1; t_1)$ and transition probability densities $p_X(x_k; t_k|x_{k-1}, t_{k-1})$ (or the equivalent pmfs for discrete-valued processes). We call $p_X(x_k; t_k|x_{k-1}, t_{k-1})$ the transition probability density function (or probability mass function) of the Markov process. It is easy to establish that transition probability densities satisfy the *Chapman-Kolmogorov* equation. Let $t_1 < t_2 < t_3$; then we have:

$$
\begin{aligned}
p_X(x_3; t_3|x_1; t_1) &= \int_{-\infty}^{\infty} p_X(x_3, x_2; t_3, t_2|x_1; t_1) \, dx_2 \\
&= \int_{-\infty}^{\infty} p_X(x_3; t_3|x_2, x_1; t_2, t_1)p_X(x_2; t_2|x_1; t_1) \, dx_2 \\
&= \int_{-\infty}^{\infty} p_X(x_3; t_3|x_2; t_2)p_X(x_2; t_2|x_1; t_1) \, dx_2,
\end{aligned} \tag{3.4}
$$

where the second equality follows from the definition of conditional distributions, and the final equality follows from the Markov property.

In order to exploit the properties of Markov processes, it is useful to generalize the definition to vector-valued processes. A vector-valued stochastic process $\underline{X}(\cdot)$ is said to be Markov if, for any choice of sampling instances $t_1 < t_2 < \ldots < t_k$,

$$
p_{\underline{X}}(\underline{x}_k; t_k|\underline{x}_1, \ldots, \underline{x}_{k-1}; t_1, \ldots, t_{k-1}) = p_{\underline{X}}(\underline{x}_k; t_k|\underline{x}_{k-1}; t_{k-1}).
$$

In many cases, one can transform a non-Markov scalar process into a vector Markov process. We illustrate this below for both discrete and continuous time.

Example 3.3
Consider a discrete-time scalar-valued random process $X(n)$, where

$$p_X(x_k; t_k | x_1, \ldots, x_{k-1}; t_1, \ldots, t_{k-1}) = p_X(x_k; t_k | x_{k-1}, x_{k-2}; t_{k-1}, t_{k-2}).$$

Define a new process

$$\underline{Z}(n) = \begin{bmatrix} X(n) \\ X(n-1) \end{bmatrix}.$$

It is easy to see that $\underline{Z}(n)$ is Markov, as follows:

$$
\begin{aligned}
p_{\underline{Z}}(\underline{z}_k; t_k | \underline{z}_1, \ldots, \underline{z}_{k-1}; t_1, \ldots, t_{k-1}) &= p_X(x_k, x_{k-1}; t_k, t_{k-1} | x_0, \ldots, x_{k-1}; t_0, \ldots, t_{k-1}) \\
&= p_X(x_k, x_{k-1}; t_k, t_{k-1} | x_{k-1}, x_{k-2}; t_{k-1}, t_{k-2}) \\
&= p_{\underline{Z}}(\underline{z}_k; t_k | \underline{z}_{k-1}; t_{k-1}).
\end{aligned}
\tag{3.5}
$$

Example 3.4
Consider an independent increments process $U(\cdot)$, with enough assumptions to be integrable (a topic which we will discuss later in the notes). Define a new process

$$Y(t) = \int_0^t \int_0^s U(\tau) d\tau \, ds$$

The process is not Markov, as we can see by considering $P(y(3)|y(2), y(1))$. Note that Y satisfies the following differential equation:

$$\frac{d^2}{dt^2} Y = U$$

Thus, we can write

$$
\begin{aligned}
Y(3) &= Y(2) + \frac{d}{dt} Y(t) \Big|_{t=2} + \int_2^3 \int_2^s U(\tau) \, d\tau \, ds \\
&= Y(2) + \int_0^2 U(s) \, ds + \int_2^3 \int_2^s U(\tau) \, d\tau \, ds
\end{aligned}
\tag{3.6}
$$

It is clear from the above equation that the conditional density of $Y(3)$ depends on $Y(2)$ and $Y(1)$, since the value of $Y(1)$ will be highly correlated with the term $\int_0^2 U(s) ds$. Now, define the augmented state

$$\underline{Z} = \begin{bmatrix} Y \\ \frac{d}{dt} Y \end{bmatrix}.$$

The vector process $\underline{Z}$ is now Markov. To illustrate this, consider how $\underline{Z}(t)$ depends on previous values of $\underline{Z}(\tau), \tau < t$:

$$\underline{Z}(t) = \underline{Z}(\tau) + \int_\tau^t \begin{bmatrix} (t-s)U(s) \\ U(s) \end{bmatrix} ds.$$

Due to the independent increments property of $U(\cdot)$, it is clear that the integral on the right hand side is independent of $\underline{Z}(s)$ for any value $s < \tau$, which establishes the vector Markov property.

3.5 Properties of Stochastic Processes

There are several properties of random processes that make it easier to specify the joint distribution for arbitrary times $\{t_1, \ldots, t_k\}$ and/or simplify the first and second-order moments. Of particular importance are properties describing conditions where the nature of the process randomness does not change with time. In other words, an observation of the process on some time interval (s, t) displays the same random behavior as over the time interval $(s + \tau, t + \tau)$. This and other properties can be defined in terms of distributions or moments, corresponding to strong vs. weak conditions, respectively. Both types are included among the definitions below.

Definition 3.5 (Strict-Sense Stationary)
The stochastic process $X(\cdot)$ is called *stationary* (or *strict-sense stationary (SSS)*, or *strictly stationary*) if the joint distribution of any collection of samples depends only on their relative time. That is, for any k and any $t_1, t_2, \ldots, t_k$ and any τ, we have

$$p_X(x_1, \ldots, x_k; t_1, \ldots, t_k) = p_X(x_1, \ldots, x_k; t_1 - \tau, \ldots, t_k - \tau).$$

Clearly, the concept of stationary processes is easily extended to vector-valued processes, whereby the components are said to be jointly stationary. There are several consequences of stationarity which are useful to exploit. First, note that the mean of the process must be independent of time, since $p_X(x;t) = p_X(x;\tau)$ for all t, τ. Thus, $m_X(t) = m_X$ for all t. Similarly, the variance $\sigma_X^2(t) = \sigma_X^2$. Second, the second-order joint density functions depend only on the difference of the time indices; that is,

$$p_X(x_1, x_2; t_1, t_2) = p_X(x_1, x_2; 0, t_2 - t_1).$$

Thus, second-order moment functions such as autocorrelations and autocovariances depend only on the differences in the times! That is,

$$R_X(t_1, t_2) = R_X(0, t_2 - t_1) \equiv R_X(t_2 - t_1); \quad K_X(t_1, t_2) = K_X(0, t_2 - t_1) \equiv K_X(t_2 - t_1),$$

where we have indulged in a standard abuse of notation with the use of a single argument for time difference.

There are many processes for which we can only establish the weaker condition of stationarity of the first and second-order moment functions. We define this class of processes below:

Definition 3.6 (Wide-Sense Stationary)
The process $X(\cdot)$ is said to be *wide-sense stationary* (WSS) (or *weakly stationary*) if the mean of the process does not depend on time, and autocorrelation function depends only on the time difference of the two samples. That is,

$$m_X(t) = m_X; \qquad R_X(t_1, t_2) \equiv R_X(t_2 - t_1)$$

Another class of random processes of interest are processes whose description exhibits periodic behavior. These processes arise in many communications applications, where operations must be repeated periodically. We define two special classes of processes:

Definition 3.7 (Periodic)
A stochastic process $X(\cdot)$ is said to be *periodic* if the joint probability density (or mass) function is invariant when the time of any of the variables is shifted by integer multiples of some period T. That is, for any k, for any integers m_i and sampling times $t_1, \ldots, t_k$, we have

$$p_X(x_1, \ldots, x_k; t_1, \ldots, t_k) = p_X(x_1, \ldots, x_k; t_1 - m_1 T, \ldots, t_k - m_k T)$$

Definition 3.8 (Cyclostationary)
A stochastic process $X(\cdot)$ is said to be *cyclostationary* if the joint probability density (or mass) function is invariant when the time origin is shifted by integer multiples of some period T. That is, for any k, for any integer m and sampling times $t_1, \ldots, t_k$, we have

$$p_X(x_1, \ldots, x_k; t_1, \ldots, t_k) = p_X(x_1, \ldots, x_k; t_1 - mT, \ldots, t_k - mT)$$

The difference between periodic and cyclostationary is that, in periodic processes, each time index can be shifted by a different multiple of the period, while in cyclostationary processes, all time indices must receive the same shift. Note that periodic implies cyclostationary, but not vice versa. Again, when we care only about second-order moment functions, we have weaker definitions.

Definition 3.9 (Wide-Sense Periodic)
A stochastic process $X(\cdot)$ is said to be *wide-sense periodic* if the mean and autocorrelation of the process are invariant when the time of any of the variables is shifted by integer multiples of some period T. That is, for any k, for any integers m_1, m_2 and sampling times t_1, t_2, we have:

$$m_X(t) = m_X(t - mT); \quad R_X(t_1, t_2) = R_X(t_1 - m_1 T, t_2 - m_2 T)$$

Definition 3.10 (Wide-Sense Cyclostationary)
A stochastic process $X(\cdot)$ is said to be *wide-sense cyclostationary* if the mean and autocorrelation of the process are invariant when the time origin is shifted by integer multiples of some period T. That is, for any k, for any integer m and sampling times t_1, t_2, we have

$$m_X(t) = m_X(t - mT); \quad R_X(t_1, t_2) = R_X(t_1 - mT, t_2 - mT)$$

A final class of random processes of interest, which arises in the areas of detection and estimation are Martingale processes:

Definition 3.11 (Martingale)
A stochastic process $X(\cdot)$ is a *Martingale* if

1. $E[X(t)] < \infty$ for all t.

2. Given two times $s < t$, then $E[X(t)|\{X(s_1), s_1 \leq s\}] = X(s)$.

Thus, Martingales have properties which are similar to those of Markov processes, with respect to the operation of conditional expectation. Furthermore, their increments have zero-mean, because

$$E[X(t) - X(s)] = E[E[X(t) - X(s)|X(s)]] = E[X(s) - X(s)] = 0.$$

The relationship between the Martingale and Markov properties is similar to that between weakly and strictly stationary in the sense that the property is specified in terms of moments vs. distributions. There is an important difference, however, in that having the Markov property does not imply that a process is Martingale.

3.6 Examples of Random Processes

3.6.1 The Random Walk

The random walk model is a discrete-time process (a random sequence) which was first proposed as a way of modeling the motion of a particle. In essence, we assume the following model: Consider an infinite number of identical Bernoulli trials, with probability $1/2$ of succeeding (e.g. an unbiased coin flip). For each trial, define the random variable

$$X(n) = \begin{cases} 1 & \text{if the experiment succeeds} \\ -1 & \text{otherwise} \end{cases}$$

Note that $\{X(n)\}$ is a sequence of independent, identically distributed random variables. Define the running sum as

$$Y(n) = \sum_{i=1}^{n} X(i). \tag{3.7}$$

Then, note that $Y(n)$ is the number of successful outcomes minus the number of failures (e.g. heads - tails). The random walk process has several important properties. First, it is clearly an independent increments process, since the sequence $\{X(n)\}$ is independent, identically distributed. Second, it is a Markov process (a consequence of the independent increments property). Third, it is a Martingale, since the increments have zero-mean, so that, if $n > m$

$$E\left[Y(n) \mid Y(m)\right] = E\left[Y(m) + \sum_{i=m+1}^{n} X(i) \mid Y(m)\right] = Y(m),$$

because of the independence between $Y(n) - Y(m)$ and $Y(m)$. We can also compute the relevant second-order functions; the process is zero-mean, with autocorrelation function given by:

$$R_Y(n, m) = E[Y(n)Y(m)] = \begin{cases} E[(Y(m) + Y(n) - Y(m))Y(m)] = E[Y(m)^2] = \sum_{i=1}^{m} \sigma_X^2 & \text{if } n > m \\ E[(Y(n) + Y(m) - Y(n))Y(n)] = \sum_{i=1}^{n} \sigma_X^2 & \text{otherwise} \end{cases} \tag{3.8}$$

More concisely, we write

$$K_Y(n, m) = R_Y(n, m) = \min(m, n)\sigma_X^2, \tag{3.9}$$

using the fact that $K_Y = R_Y$ for zero-mean processes. It is important to note that the random walk process is not wide-sense stationary, since the autocovariance cannot be written as a function of $n - m$.

3.6.2   The Poisson Process

Also known as the Poisson counting process (PCP), the Poisson process is a popular model used in communications, manufacturing and other network applications to model the arrival stream of customers, calls or jobs. The Poisson process value $N(t)$ is the total number of arrivals in the interval $[0, t]$. Thus the Poisson process counts up from 0 and takes on non-negative integer values. As we will see, the PCP is equivalently defined as an independent increments process whose increments are Poisson distributed. Construction of a Poisson process can be accomplished through a series of steps, which illuminate its connections to applications. We discuss these steps next.

Step 1) First we model the times τ_k between "arrivals" of the process as a sequence of independent, identically distributed exponential random variables with parameter λ. These times τ_k are called the "interarrival times" and are illustrated in Figure 3.1. The assumption of independent, identically distributed exponential interarrival times turns out to be a good model of many physical processes, such as subway arrivals, customers entering a line, etc. The exponential density of each interarrival time is given by:

$$p_{\tau_k}(t) = \lambda e^{-\lambda t} u(t) \quad k = 1, 2, \cdots \tag{3.10}$$

where $u(t)$ is the unit step function, defined by

$$u(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ 1 & \text{otherwise} \end{cases} \tag{3.11}$$

Figure 3.1: Interarrival Times τ_k.

Step 2) Now we can define the sequence of "event times" $T(n)$, which are the times at which the arrivals or events happen. In terms of the interarrival times we have:

$$T(n) = \sum_{k=1}^{n} \tau_k, \tag{3.12}$$

where the interarrival times τ_k are independent, identically distributed exponential random variables, as described above. Figure 3.2 shows the relationship between the τ_k and $T(n)$. Now, note that $T(n)$ is defined as the sum of a series of independent, identically distributed random variables. Thus its pdf can be obtained either by convolving the individual exponential pdfs or by finding the product of the corresponding characteristic functions. It is easiest to use the characteristic function approach:

$$E\left[e^{jwT(n)}\right] = E\left[e^{jw\sum_{k=1}^{n}\tau_k}\right] = \left(E\left[e^{jw\tau_k}\right]\right)^n \tag{3.13}$$

Using the definition of the characteristic function of the exponential distribution from Table 1.2 we obtain:

$$E\left[e^{jwT(n)}\right] = \frac{\lambda^n}{(\lambda - j\omega)^n}. \tag{3.14}$$

Taking inverse Fourier transforms, we get

$$p_{T(n)}(t) = \lambda^n \frac{t^{n-1}}{(n-1)!} e^{-\lambda t} u(t), \tag{3.15}$$

Examining the form of $p_{T(n)}(t)$ and comparing to the table of common densities in Table 1.2, we can see it is an Erlang distribution. Note that this distribution has mean $m_T(n) = n/\lambda$ and variance $\sigma_T^2 = n/\lambda^2$.
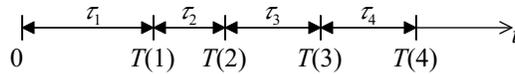
Figure 3.2: Arrival times $T(n)$ and interarrival times τ_k.

Step 3) Finally, suppose we let $T(n)$ be the times where the Poisson process takes a unit step jump. Note that these are random times, and that by construction the time between jumps has an exponential distribution. Mathematically then, the Poisson counting process (sometimes just called the counting process) is finally described as the sum of these shifted step functions:

$$N(t) = \sum_{i=1}^{\infty} u(t - T(n)). \tag{3.16}$$

so that, indeed, $N(t)$ is the number of arrivals in the interval $[0, t]$.

Now we need to construct the probability mass function $p_{N(t)}(m)$. Note that

$$p_{N(t)}(m) = \Pr(N(t) = m) = \Pr\left(T(m) \leq t, T(m+1) > t\right) = \Pr\left(T(m) \leq t, \tau_{m+1} > t - T(m)\right) \tag{3.17}$$

which is the probability of all possible ways we can have $T(m) \leq t$ and $\tau_{m+1} > t - T(m)$ for that $T(m)$. Now, by construction, $T(m)$ and τ_{m+1} are independent, so that:

$$
\begin{aligned}
p_{N(t)}(m) &= \int_0^t p_{T(m)}(u) \left[\int_{t-u}^{\infty} p_{\tau_{m+1}}(v)\, dv\right] du \\
&= \int_0^t \lambda^m \frac{u^{m-1}}{(m-1)!} e^{-\lambda u} \left[\int_{t-u}^{\infty} \lambda e^{-\lambda v}\, dv\right] du \\
&= \int_0^t \frac{\lambda^m u^{m-1}}{(m-1)!} e^{-\lambda u} e^{-\lambda(t-u)}\, du \\
&= \frac{(\lambda t)^m}{(m)!} e^{-\lambda t}, \quad t, n \geq 0
\end{aligned} \tag{3.18}
$$

which basically states that, at time t, the distribution of the Poisson process is a Poisson random variable with parameter λt. The relationship between τ_k, $T(n)$, and $N(t)$ is shown in Figure 3.3.
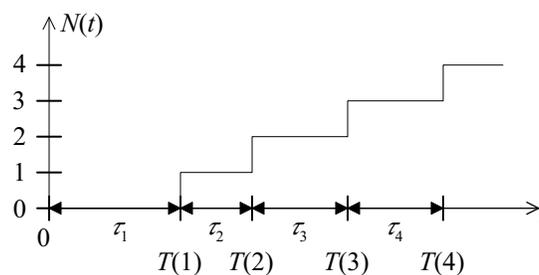
Figure 3.3: The Poisson Counting Process (PCP) $N(t)$ and the relationship between arrival times $T(n)$ and interarrival times τ_k.

By construction, it appears that the Poisson process has independent increments. However, showing this is hard, since the intervals between events are not defined as set times. In particular, we want to establish that, if $t_2 > t_1 > t_0$, then $N(t_2) - N(t_1)$ is independent of $N(t_1) - N(t_0)$. To do this requires exploiting the special structure of the exponential distribution, which was shown to be *memoryless* in Chapter 1. This structure enables us to show that when $p(\tau) = \lambda e^{-\lambda \tau} u(\tau)$ then

$$p(\tau | \tau > T) = p(\tau - T) = \lambda e^{-\lambda(t-\tau)} u(t - \tau).$$

We will show that $N(t_2) - N(t_1)$ is independent of $N(t_1)$, as follows. Note that

$$\Pr[N(t_2) - N(t_1) = n, N(t_1) = m] = \Pr[T(m) \leq t_1, T(m+1) > t_1, t_{m+n} \leq t_2, t_{m+n+1} > t_2].$$

Now, define a new random variable $S = t_1 - T(m)$. Note that $S < \tau_{m+1} = T(m+1) - T(m)$, by construction; thus, define a second random variable $V = \tau_{m+1} - S$. Due to the memoryless property of the exponential distribution, we have

$$p_{SV}(s, v) = p_{V|S}(v|s)p_S(s) = p_\tau(v + s|\tau > s)p_S(s) = p_\tau(v)p_S(s),$$

which shows that S and V are independent! Now, we can easily see that

$$\begin{aligned}
\Pr[N(t_2) - N(t_1) = n, N(t_1) = m] &= \Pr[T(m) \leq t_1, T(m+1) > t_1, T(m+n) \leq t_2, T(m+n+1) > t_2] \\
&= \Pr[T(m) + S \leq t_1, t_1 + V + \tau_2 + \ldots + \tau_{m+n} \leq t_2, t_1 + V + \tau_2 + \ldots + \tau_{m+n+1} > t_2] \\
&= \Pr[T(m) + S \leq t_1]\Pr[t_1 + V + \tau_2 + \ldots + \tau_{m+n} \leq t_2, t_1 + V + \tau_2 + \ldots + \tau_{m+n+1} > t_2] \\
&= \Pr[N(t_1) = m]\Pr[N(t_2) - N(t_1) = n].
\end{aligned} \tag{3.19}$$

The last equality follows from the independence of the interarrival times τ_i, and the decomposition of the interarrival time τ_{m+1} into two independent components S and V, thanks to the memoryless property of the exponential distribution. Thus, we have shown that the Poisson counting process is also an independent increments process!

One of the properties of Poisson random variables is that the sum of two independent Poisson random variables is also a Poisson random variable! To see this, consider the moment generating function of the sum of two Poisson random variables N, M with rates λ_N, λ_M respectively, as

$$E[z^{N+M}] = E[z^N]E[z^M] = e^{\lambda_N(z-1)}e^{\lambda_M(z-1)} = e^{(\lambda_N + \lambda_M)(z-1)},$$

which is the moment-generating function for a Poisson random variable with rate $\lambda_N + \lambda_M$! Coupled with the independent increments property of Poisson processes, this allows us to make several statements:

1. For any interval $[t_1, t_2]$, the probability that $N(t_2) - N(t_1) = n \geq 0$ occurs in that interval is Poisson distributed, with intensity $\lambda(t_2 - t_1)$. That is,

$$\Pr(N(t_2) - N(t_1) = n) = \frac{(\lambda(t_2 - t_1))^n}{(n)!}e^{-\lambda(t_2 - t_1)}.$$

   In other words, the increments of a PCP are themselves Poisson distributed random variables!

2. For any pair of disjoint intervals, the number of events which occur in those intervals is independent, and the average number of events which occur on equal-length intervals is the same.

3. The joint probability $p_N(n_1, n_2; t_1, t_2)$ for $t_2 > t_1$, in which case $n_2 \geq n_1 \geq 0$, is computed as

$$\begin{aligned}
p_N(n_1, n_2; t_1, t_2) &= p_N(n_2; t_2|n_1, t_1)p_N(n_1; t_1) \\
&= p_N(n_2 - n_1; t_2 - t_1)p_N(n_1; t_1) \\
&= \frac{(\lambda(t_2 - t_1))^{n_2 - n_1}}{(n_2 - n_1)!}e^{-\lambda(t_2 - t_1)}\frac{(\lambda(t_1))^{n_1}}{(n_1)!}e^{-\lambda t_1} \\
&= \frac{\lambda^{n_2}t_1^{n_1}(t_2 - t_1)^{n_2 - n_1}}{(n_1)!(n_2 - n_1)!}e^{-\lambda t_2}
\end{aligned} \tag{3.20}$$

4. A Poisson process is an independent-increments process, where increments over an interval $[t_1, t_2]$ are Poisson distributed with rate $\lambda(t_2 - t_1)$.

5. The sum of two independent Poisson processes $N_1(t), N_2(t)$ with intensities λ_1, λ_2 is a Poisson process with intensity $\lambda_1 + \lambda_2$.

Using the above properties, the first and second-order moment functions of Poisson processes are easy to establish. These are:

$$E[N(t)] = \lambda t; \qquad E[N(t)^2] = \lambda t + (\lambda t)^2; \qquad \sigma^2_{N(t)} = \lambda t \tag{3.21}$$

and the autocorrelation function:

$$
\begin{aligned}
R_N(t,s) &= E[N(t)N(s)] \\
&= E[(N(t) - N(s) + N(s))N(s)] \quad \text{assuming } t \geq s \text{ without loss of generality} \\
&= E[(N(t) - N(s))N(s)] + E[N(s)^2] = E[(N(t) - N(s))]E[N(s)] + E[N(s)^2] \\
&= \lambda^2(t-s)(s) + \lambda s + \lambda^2 s^2 = \lambda^2 ts + \lambda s
\end{aligned}
\tag{3.22}
$$

Using symmetry, we can write the autocorrelation and autocovariance functions more generally as

$$R_N(t,s) = \lambda^2 ts + \lambda \min(t,s); \qquad K_N(t,s) = \lambda \min(t,s)$$

Note the minimization which occurs due to the independent increments property. Also note that the Poisson process is not a stationary process; however, it will be a Markov process.

Example 3.5
When monitoring radioactivity, or also photoelectric intensity, most measuring processes are based on counting the number of particles or photons which are emitted. Due to the quantum nature of electromagnetic waves, the number of photons which are emitted in a given interval is random, and assumed to be independent over any pair of disjoint intervals; furthermore, the average number of photons emitted over any interval is constant, depending on the intensity of the source. This satisfies the assumptions of a Poisson process, which consists of independent increments and a constant rate. Indeed, one can show that these two properties, plus the fact that it is a counting process, implies that the actual number of photons generated is a Poisson process.

3.6.3 Digital Modulation: Phase-Shift Keying

A basic method for modulation of digital data is phase-shift keying (PSK). In this method, binary data, modeled by a stream of 0's and 1's, is coded onto a carrier frequency by a phase signal. Define the random phase $\theta(n)$ as follows:

$$\theta(n) = \begin{cases} \pi/2 & \text{if the } n\text{-th bit is 1,} \\ -\pi/2 & \text{otherwise} \end{cases} \tag{3.23}$$

Let T denote the duration of the signal used for each bit. Typically, T is a multiple of bit rate (the period of the carrier frequency f_c); that is, $T = m/f_c$ for some integer $m \geq 1$, so that one or more cycles are used per bit. Define the phase signal for the n-th bit as

$$\Theta(t) = \theta(n) \text{ for } nT \leq t < (n+1)T \tag{3.24}$$

The corresponding transmitted signal is given by

$$X(t) = \cos\left(\omega_c t + \Theta(t)\right) \tag{3.25}$$

where $\omega_c = 2\pi f_c$ is the carrier frequency in radians/sec.

Now, suppose that the phase process in (3.24) was an independent, identically distributed random sequence, where each $\theta(n)$ is a binary-valued random variable with probability parameter p; (i.e., $p_{\theta(n)}(\pi/2) = p$.) Then, the resulting collection of transmitted signals obtained in (3.25) form a continuous-time, continuous-valued random process. When the parameter $p = 1/2$, it is referred to as the PSK process.

What are the second-order moments of the PSK process? Remember the following trigonometric identity:

$$\cos\left(\omega_c t + \Theta(t)\right) = \cos\left(\omega_c t\right)\cos\left(\Theta(t)\right) - \sin\left(\omega_c t\right)\sin\left(\Theta(t)\right) \tag{3.26}$$

Using the above, we compute the mean of the process as

$$
\begin{aligned}
m_X(t) &= E\left[\cos\left(\omega_c t\right)\cos\left(\Theta(t)\right) - \sin\left(\omega_c t\right)\sin\left(\Theta(t)\right)\right] \\
&= \cos\left(\omega_c t\right)E\left[\cos\left(\Theta(t)\right)\right] - \sin\left(\omega_c t\right)E\left[\sin\left(\Theta(t)\right)\right] \\
&= 0 - \sin\left(\omega_c t\right)\left(p - (1-p)\right) = \sin\left(\omega_c t\right)\left(1 - 2p\right)
\end{aligned}
\tag{3.27}
$$

Note that the third equality follows because $\cos(\Theta(t)) = 0$ and $\sin(\Theta(t_1)) = \pm 1$ by definition. When $p = 0.5$, the mean is zero.

The autocorrelation (also autocovariance) is given by:

$$
\begin{aligned}
R_X(t_1, t_2) &= E[(\cos\omega_c t_1 \cos\Theta(t_1) - \sin\omega_c t_1 \sin\Theta(t_1))(\cos\omega_c t_2 \cos\Theta(t_2) - \sin\omega_c t_2 \sin\Theta(t_2))] \\
&= E[\sin(\omega_c t_1)\sin(\Theta(t_1))\sin(\omega_c t_2)\sin(\Theta(t_2))] \\
&= \sin\omega_c t_1 \sin\omega_c t_2 E[\sin\Theta(t_1)\sin\Theta(t_2)]
\end{aligned}
$$

To complete the computation, we must compute the correlation $E[\sin\Theta(t_1)\sin\Theta(t_2)]$. Note that, from (3.24), $\Theta(t_1), \Theta(t_2)$ are independent unless $nT \le t_1, t_2 < (n+1)T$ for some n. Thus,

$$
E[\sin\Theta(t_1)\sin\Theta(t_2)] = \begin{cases} 1 & \text{if } nT \le t_1, t_2 < (n+1)T \text{ for some } n \\ E[\sin\Theta(t_1)]E[\sin\Theta(t_2)] & \text{otherwise} \end{cases} \tag{3.28}
$$

and thus the autocorrelation function is given by

$$
R_X(t_1, t_2) = \begin{cases} \sin\omega_c t_1 \sin\omega_c t_2 & \text{if } nT \le t_1, t_2 < (n+1)T \text{ for some } n \\ \sin(\omega_c t_1)\sin(\omega_c t_2)(1-2p)^2 & \text{otherwise} \end{cases} \tag{3.29}
$$

and the autocovariance function is

$$
K_X(t_1, t_2) = \begin{cases} 4p(1-p)\sin\omega_c t_1 \sin\omega_c t_2 & \text{if } nT \le t_1, t_2 < (n+1)T \text{ for some } n \\ 0 & \text{otherwise} \end{cases} \tag{3.30}
$$

When $p = 0.5$ then the scaling factor is unity: $4p(1-p) = 1$. The PSK process is an example of a cyclostationary process. Note, however, that it is not periodic, since $R_X(t_1 + T, t_2) \ne R_X(t_1, t_2)$.

3.6.4   The Random Telegraph Process

The random telegraph process (sometimes also known as the random binary sequence) is a discrete-valued process which is used often as input in model identification problems because it is easy to generate sample paths of this function. The process is generated in a manner which is very similar to a Poisson process; indeed, one way to define the random binary sequence is to switch values at the jump times of a Poisson process.

Let $\{T_n\}$ denote the sequence of event times associated with Poisson process, as in eq.(3.12). The random telegraph process is generated as follows: Let $X(0)$ be a binary-valued random variable, with equal probability of achieving the values $\{-1, 1\}$. Define $T_0 = 0$ for notation; then,

$$
X(t) = \begin{cases} X(T_n) & \text{if } T_n < t < T_{n+1} \\ -X(T_n) & \text{if } t = T_{n+1} \end{cases} \tag{3.31}
$$

Due to its construction, the random telegraph process has properties which are similar to the Poisson process. In particular, the inter-event times $\tau_n = T_n - T_{n-1}$ are independent, identically distributed exponential random variables with rate λ, and the numbers of events in disjoint intervals are independent random variables.

In order to understand better the relationship between the random telegraph process and the Poisson process, let $N(t)$ denote the Poisson counting process with the same event times. If we assume that $X(T_0)$ is equally likely to be either +1 or -1, then the random telegraph process is clearly zero-mean. Assuming $t_2 \ge t_1$, the autocorrelation (and autocovariance) are given by:

$$
\begin{aligned}
R_X(t_1, t_2) &= E[X(t_1)X(t_2)] = (+1)\Pr[X(t_1) = X(t_2)] + (-1)\Pr[X(t_1) \ne X(t_2)] \\
&= \Pr[N(t_2) - N(t_1) = 2n \text{ for some } n \ge 0] - \Pr[N(t_2) - N(t_1) = 2n+1 \text{ for some } n \ge 0] \\
&= \sum_{n=0}^{\infty} \frac{[\lambda(t_2 - t_1)]^{2n}}{(2n)!} e^{-\lambda(t_2 - t_1)} - \sum_{n=0}^{\infty} \frac{[\lambda(t_2 - t_1)]^{2n+1}}{(2n+1)!} e^{-\lambda(t_2 - t_1)} \\
&= \frac{1}{2}\left(1 + e^{-2\lambda(t_2 - t_1)}\right) - \frac{1}{2}\left(1 - e^{-2\lambda(t_2 - t_1)}\right) \\
&= e^{-2\lambda(t_2 - t_1)}
\end{aligned} \tag{3.32}
$$

More generally, using symmetry of R_X, we have

$$R_X(t_1, t_2) = e^{-2\lambda|t_2 - t_1|} \tag{3.33}$$

Thus, the random telegraph process is wide-sense stationary. Indeed, we can extend the definition to $(-\infty, \infty)$ by defining event times T_n for negative integers n in an obvious manner using exponential independent, identically distributed random variables τ_n for negative integers also. Then, we can show that the random telegraph process is stationary in the strict sense, due to the stationary property of the increments of Poisson processes. That is,

$$
\begin{aligned}
p_X(x_1, x_2; t_1, t_2) &= p_X(x_2; t_2 | x_1; t_1) p_X(x_1; t_1) \\
&= \{\delta(x_1 - x_2)\Pr[N(t_2) - N(t_1) \text{ even}|X(t_1) = x_1] + \\
&\quad \delta(x_1 + x_2)\Pr[N(t_2) - N(t_1) \text{ odd}|X(t_1) = x_1]\} \, p_X(x_1; t_1)
\end{aligned}
\tag{3.34}
$$

where $\delta(\cdot)$ denotes the dirac or impulse function. Now, the increments $N(t_2) - N(t_1)$ are independent of $X(t_1)$, due to the independent increments property of the Poisson process and the fact that $X(t_1)$ depends only on events used to define $N(t_1)$. Thus,

$$
\begin{aligned}
p_X(x_1, x_2; t_1, t_2) &= \{\delta(x_1 - x_2)\Pr[N(t_2) - N(t_1) \text{ even}] + \delta(x_1 + x_2)\Pr[N(t_2) - N(t_1) \text{ odd}]\} \, p_X(x_1; t_1) \\
&= \{\delta(x_1 - x_2)\Pr[N(t_2 - t) - N(t_1 - t) \text{ even}] + \\
&\quad (1 - \delta(x_1 - x_2))\Pr[N(t_2 - t) - N(t_1 - t) \text{ odd}]\} \, p_X(x_1; t_1 - t) \\
&= p_X(x_1, x_2; t_1 - t, t_2 - t)
\end{aligned}
\tag{3.35}
\tag{3.36}
$$

since $p_X(x_1; t_1)$ is stationary, and the distribution of a Poisson increment $\Pr[N(t_2) - N(t_1)]$ is also stationary. The above argument can be generalized to an arbitrary finite number of process values $X(t_1), \ldots, X(t_n)$.

Does the random telegraph process have independent increments? No, it does not. To see this consider three ordered times $t_1 < t_2 < t_3$ and the corresponding increments $X(t_2) - X(t_1)$ and $X(t_3) - X(t_2)$. If we know nothing about the first increment $X(t_2) - X(t_1)$, then the second increment could either be $+2$, -2 or 0 (i.e. there are three possibilities, each with some probability). But if we know that $X(t_2) - X(t_1) = +2$ then $X(t_3) - X(t_2)$ can only be -2 or 0, so knowledge of $X(t_2) - X(t_1)$ is clearly affecting the uncertainty in $X(t_3) - X(t_2)$ and they cannot be independent.

The above discussion illustrates why Poisson processes are so important for the understanding of stochastic processes. We can define other processes based on the transition times of a Poisson process, and these processes inherit many of the fundamental properties of the Poisson process, such as independent increments.

3.6.5   The Wiener Process and Brownian Motion

Suppose that we have defined a discrete-time random walk process $Y(\cdot)$, as in section 3.6.1, indexed by a discrete time n. We can embed this process into continuous time by interpolating between times, as:

$$
Y_T(t) = \begin{cases} 0 & \text{if } t = 0 \\ Y(n)d & \text{if } (n-1)T < t \leq nT, n = 1, \ldots \end{cases}
\tag{3.37}
$$

for some sampling time T and some jump size d. Based on the properties of the discrete-time random walk process $Y(\cdot)$, we have the following properties for the continuous-time process $Y_T(\cdot)$:

$$E[Y_T(t)] = 0; \qquad E[Y_T(t)^2] = E[d^2 Y^2(n(t))] = d^2 n(t) = d^2 \min[n : nT \geq t]$$

Note that $n(t) \approx t/T$ in the above equation. Our goal is to define a process which is the limit of the $Y_T(t)$ process, as we let the sampling time $T \to 0$ in a manner which makes the limit a random variable. In particular, define a monotone decreasing sequence of times $T_n = 1/n$, so that $\lim_{n \to \infty} T_n = 0$, and define the corresponding sequence of processes as $Y_{T_n}(t) \equiv Y_n(t)$. As n increases, the value of the process at each time, $Y_n(t)$, is the sum of nt independent, identically-distributed, zero-mean random variables, and the variance of $Y_n(t)$ is $d^2 nt$. By the strong law of large numbers, we have that

$$\lim_{n \to \infty} \frac{Y_n(t)}{nt} \stackrel{\text{a.e.}}{=} 0 \tag{3.38}$$

The idea for the construction of the Brownian motion process is to decrease the step size d as n increases, so that the overall variance of $Y_n(t)$ remains constant. Thus, define the sequence d_n such that $d_n^2 n = 1$. Then, by the Central Limit Theorem, since $Y_n(t)$ is the sum of an increasing number of independent, identically distributed random variables,

$$\lim_{n \to \infty} \frac{Y_n(t)}{\sqrt{t}} \stackrel{\text{d.}}{=} N(0, 1) \tag{3.39}$$

That is, the limit of the normalized sequence converges in distribution to a unit variance, zero-mean Gaussian random variable. Alternatively, since the normalizing factor does not depend on n, then the sequence of random variables $Y_n(t)$ converges in distribution to a Gaussian random variable $B(t)$, with zero-mean, and variance t. We define the Brownian motion process $B(t)$ to be the limit, for each t, of the sequence of random variables $Y_n(t)$.

What properties does the process $B(t)$ have? We know that, for each n, the process $Y_n(t)$ has almost independent increments, in the sense that, for $t_1 < t_2 < t_3 < t_4$, the increments $Y_n(t_2) - Y_n(t_1)$ and $Y_n(t_4) - Y_n(t_3)$ are independent provided they are constructed from independent increments in the underlying random walk. That is, they are independent provided $t_3 - t_2 \geq 1/n$. As $n \to \infty$, the limit process $B(t)$ will have independent increments. As established before, the limit process $B(t)$ is also a Gaussian process.

Note also that, as $n \to \infty$, the size of the jumps in the process derived from the random walk, d_n, are getting smaller. Indeed, we have the following property:

$$
\begin{aligned}
\lim_{\epsilon \to 0} B(t + \epsilon) &= \lim_{\epsilon \to 0} \lim_{n \to \infty} Y_n(t + \epsilon) = \lim_{n \to \infty} \lim_{\epsilon \to 0} Y_n(t + \epsilon) \\
&= \lim_{\epsilon \to 0} \lim_{n \to \infty} Y_n(t) + (Y_n(t + \epsilon) - Y_n(t)) \\
&= \lim_{n \to \infty} [Y_n(t) + \lim_{\epsilon \to 0}(Y_n(t + \epsilon) - Y_n(t))] \approx \lim_{n \to \infty} [Y_n(t) + d_n] = B(t) \tag{3.40}
\end{aligned}
$$

where the argument is somewhat imprecise because we are dealing with random variables, and we have not defined in what sense is the limit valid. The limit indicates that $B(t)$ should be a continuous function of time. Later, we shall define more formally what we mean by continuous random processes, and show that, almost surely, the sample functions $B(t)$ are continuous.

In sum, the Brownian motion $B(t)$ (also known as the standard Wiener process), is a Gaussian, zero-mean, continuous-time random process with independent increments, and with variance $E[B^2(t)] = t$. Based on this definition, we summarize the properties of Brownian motion:

1. The sample functions $B(t)$ are almost sure continuous.

2. $E[B(t)] = 0, E[B^2(t)] = t$

3. $B(t)$ is Gaussian.

4. $B(t)$ is an independent increments process; the increments $B(t) - B(s)$ are Gaussian, zero-mean random variables with variance $t - s$ for $t > s$.

5. The autocovariance and autocorrelation functions are given by

$$R_B(t, s) = K_B(t, s) = \min(t, s) \tag{3.41}$$

In addition to the standard Wiener process or Brownian motion, it is possible to define other Gaussian processes which are very similar. In particular, a generalized Wiener process is a Gaussian, zero-mean, independent-increments process with covariance $f(t)$ for some nondecreasing function $f(t)$. In particular, we can define a Wiener process with covariance αt for some positive constant α, and interpret it simply as the limit of a random walk where the step-size $d_n^2 n = \alpha$.

3.7   Moment Functions of Vector Processes

Suppose that we have a vector-valued stochastic process $\underline{X}(t)$. We define the mean and autocorrelation functions of the vector process as

$$\underline{m}_{\underline{X}}(t) = E[\underline{X}(t)]; \qquad R_{\underline{X}}(s, t) = E[\underline{X}(s)\underline{X}(t)^T] \tag{3.42}$$

Similarly, the autocovariance function is defined as

$$K_{\underline{X}}(s,t) = R_{\underline{X}}(s,t) - \underline{m}_{\underline{X}}(s)\underline{m}_{\underline{X}}(t)^T \tag{3.43}$$

If we have two vector-valued processes $\underline{X}(t), \underline{Y}(t)$ defined on the same probability space, we define the cross-correlation function $R_{\underline{XY}}$ as

$$R_{\underline{XY}}(s,t) = E[\underline{X}(s)\underline{Y}(t)^T] \tag{3.44}$$

For complex-valued vectors (or matrices) M, we define the Hermitian adjoint of M as

$$M^H = [M^T]^*$$

where * denotes complex conjugation, element by element. For complex-valued vector processes, the above definitions can be extended as:

$$R_{\underline{X}}(s,t) = E[\underline{X}(s)\underline{X}(t)^H]; \quad K_{\underline{X}}(s,t) = R_{\underline{X}}(s,t) - \underline{m}_{\underline{X}}(s)\underline{m}_{\underline{X}}(t)^H \tag{3.45}$$

Based on the above definitions, autocorrelation (also autocovariance) functions have the following properties:

1. $R_{\underline{X}}(s,t) = R_{\underline{X}}(t,s)^H$

2. Using the Cauchy-Schwarz inequality for random variables, we get, for any appropriately-dimensioned vectors $\underline{a}, \underline{b}$,

$$\begin{aligned} |\underline{a}^H R_{\underline{X}}(s,t)\underline{b}|^2 &= E[\underline{a}^H \underline{X}(s)\underline{X}(t)^H \underline{b}]^2 \le E[(\underline{a}^H \underline{X}(s))^2]E[(\underline{b}^H \underline{X}(t))^2] \\ &= [\underline{a}^H R_{\underline{X}}(s,s)\underline{a}][\underline{b}^H R_{\underline{X}}(t,t)\underline{b}] \end{aligned} \tag{3.46}$$

In particular, when $X(\cdot)$ is a scalar-valued random process, we have

$$|R_X(s,t)|^2 \le R_X(s,s)R_X(t,t)$$

3. $R_{\underline{XY}}(s,t) = R_{\underline{YX}}(t,s)^H$

3.8 Moments of Wide-sense Stationary Processes

In this section, we restrict our attention initially to scalar, real-valued wide-sense stationary random processes When the processes are wide-sense stationary, the autocorrelation function and autocovariance functions can be expressed simply as $R_X(s,t) = E[X(s)X(t)] = E[X(0)X(t-s)] \equiv R_X(t-s)$. Letting $\tau = t - s$, we use the notation $R_X(\tau) \equiv E[X(t)X(t+\tau)]$. Assuming that $X(\cdot)$ is real-valued, we have the following properties:

1. $R_X(0) = E[X(t)^2] \ge 0$.

2. $R_X(\tau) = E[X(t)X(t+\tau)] = E[X(t-\tau)X(t)] = R_X(-\tau)$, because $X(\cdot)$ is wide-sense stationary. Thus, the autocorrelation is an even function of τ.

3. $|R_X(\tau)|^2 \le R_X(0)^2$, based on the Cauchy-Schwarz inequality.

4. If $R_X(T) = R_X(0)$ for some T, then R_X is periodic, with period T. That is, $R_X(\tau + T) = R_X(\tau)$ for all τ. This also follows from the Cauchy-Schwarz inequality, as

$$\begin{aligned} |E[(X(\tau+T) - X(\tau))X(0)]|^2 &= (R_X(\tau+T) - R_X(\tau))^2 \\ &\le E[(X(\tau+T) - X(\tau))^2]E[X(0)^2] \\ &= (E[X(\tau+T)^2] + E[X(\tau)^2] - 2E[X(\tau+T)X(\tau)])R_X(0) \\ &= (2R_X(0) - 2R_X(T))R_X(0) = 0 \end{aligned} \tag{3.47}$$

5. If $R_X(\tau)$ is continuous at $\tau = 0$, it is continuous everywhere. This follows again because of the Cauchy-Schwarz inequality

$$\begin{aligned}
\lim_{\epsilon \to 0} |R_X(\tau + \epsilon) - R_X(\tau)|^2 &= \lim_{\epsilon \to 0} |E[(X(\tau + \epsilon) - X(\tau))X(0)]|^2 \\
&\leq \lim_{\epsilon \to 0} E[(X(\tau + \epsilon) - X(\tau))]^2 E[X(0)]^2 \\
&= \lim_{\epsilon \to 0} 2(R_X(\epsilon) - R_X(0))R_X(0) = 0 \qquad (3.48)
\end{aligned}$$

because $R_X(\tau)$ is continuous at 0.

Note that all these properties also hold for $K_X(\tau)$, which can be though of as a special case of an autocorrelation function for the random process $\tilde{X}(t) = X(t) - \mu_X$.

The concept of wide-sense stationarity can also be extended to vector processes. A vector process is wide-sense stationary if its autocorrelation satisfies $R_{\underline{X}}(s,t) = R_{\underline{X}}(0, t-s) = R_{\underline{X}}(v, v+t-s)$ for any s, t and v. We write $R_{\underline{X}}(s,t) \equiv R_{\underline{X}}(\tau)$. For vector real-valued processes, we have the following natural extensions of the above results:

1. The autocorrelation matrix $R_{\underline{X}}(0)$ is positive semidefinite; that is, $\underline{a}^T E[\underline{X}(t)\underline{X}(t)^T]\underline{a} \geq 0$.

2. $R_{\underline{X}}(\tau) = E[\underline{X}(t)\underline{X}(t+\tau)^T] = E[\underline{X}(t-\tau)\underline{X}(t)^T] = R_{\underline{X}}(-\tau)^T$, because $X(\cdot)$ is wide-sense stationary.

3. If $R_{\underline{X}}(\tau)$ is continuous at $\tau = 0$, it is continuous everywhere.

   This follows again because of the Cauchy-Schwarz inequality for any vectors $\underline{a}, \underline{b}$

$$\begin{aligned}
\lim_{\epsilon \to 0} |\underline{a}^T(R_{\underline{X}}(\tau + \epsilon) - R_{\underline{X}}(\tau))\underline{b}|^2 &= \lim_{\epsilon \to 0} |E[\underline{a}(\underline{X}(\tau + \epsilon) - \underline{X}(t))\underline{X}(0)^T\underline{b}]|^2 \\
&\leq \lim_{\epsilon \to 0} E[|\underline{a}^T(\underline{X}(\tau + \epsilon) - \underline{X}(t))|^2]E[|\underline{X}(0)^T\underline{b}|^2] \\
&= \lim_{\epsilon \to 0} 2\underline{a}^T(R_{\underline{X}}(\epsilon) - R_{\underline{X}}(0))\underline{a}\underline{b}^T R_{\underline{X}}(0)\underline{b} = 0 \qquad (3.49)
\end{aligned}$$

   because $R_{\underline{X}}(\tau)$ is continuous at 0. By properly selecting the vectors, we can show that each entry in $R_{\underline{X}}$ must be continuous.

Let $X(\cdot)$ and $Y(\cdot)$ denote two scalar processes defined on the same probability space. Then, the cross-correlation function R_{XY} satisfies the following:

1. $R_{XY}(\tau) = R_{YX}(-\tau)$.

2. By the Cauchy-Schwarz inequality, $R_{XY}(\tau)^2 \leq E[X(0)^2]E[Y(\tau)^2] = R_X(0)R_Y(0)$.

3. $|R_{XY}(\tau)| \leq 1/2(R_X(0) + R_Y(0))$.
   To show this, remember the following moment inequality, which was derived from Jensen's inequality

$$E[|X + Y|^r] \leq c_r(E[|X|^r] + E[|Y|^r])$$

where

$$c_r = \begin{cases} 1 & \text{if } r \leq 1 \\ 2^{r-1} & \text{if } r > 1 \end{cases}$$

In particular, if $r = 2$, we have

$$\begin{aligned}
E[(|X(0)| + |Y(\tau)|)^2] &= R_X(0) + R_Y(0) + 2|R_{XY}(\tau)| \\
&\leq 2(R_X(0) + R_Y(0)) \qquad (3.50)
\end{aligned}$$

which establishes the inequality.

3.9   Power Spectral Density of Wide-Sense Stationary Processes

In this section, we concentrate on describing the properties of scalar wide-sense stationary processes in the frequency domain. As one might imagine, such "frequency domain" characterization will turn out to be particularly convenient when we consider the interaction of wide-sense stationary processes and linear time invariant systems, which we discuss in Section 5.

Let us assume that we have a wide-sense stationary process $x(t)$ with mean m_x and autocovariance function $K_x(\tau) \equiv K_x(t, t + \tau)$ for all t, τ. We assume also that the random process $y(t)$ is also wide-sense stationary, and that x, y are jointly wide-sense stationary. As we have discussed previously in the properties of autocovariance and autocorrelation functions, we know:

1. $|R_x(\tau)| \leq R_x(0)$

2. $|R_{xy}(\tau)| \leq \sqrt{R_x(0)R_y(0)}$

3. $R_x(\tau) = R_x^*(-\tau)$

4. For all $N > 0$, all $t_1 < t_2 < \ldots < t_N$, all complex $a_1, \ldots, a_n$, we have

$$\begin{bmatrix} a_1^* & \cdots & a_N^* \end{bmatrix} \begin{bmatrix} R_x(t_1 - t_1) & R_x(t_1 - t_2) & \cdots & R_x(t_1 - t_N) \\ R_x(t_2 - t_1) & R_x(t_2 - t_2) & \cdots & R_x(t_2 - t_N) \\ \vdots & \vdots & & \vdots \\ R_x(t_N - t_1) & R_x(t_N - t_2) & \cdots & R_x(t_N - t_N) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} \geq 0$$

5. If $x(t)$ is real, then $R_x(\tau)$ is an even function.

Notice that the autocorrelation function $R_x(\tau)$ can be viewed as a deterministic function of time, which is bounded by $R_x(0)$ in magnitude. Thus, we can define its Fourier transform (if it exists), as follows:

Definition 3.12 (Power Spectral Density)
Let $R_x(\tau)$ be the autocorrelation function of the wide-sense stationary process $x(t)$. Then, the *power spectral density* $S_x(\omega)$ of $x(t)$ is defined as the Fourier transform of $R_x(\tau)$. That is,

$$S_x(\omega) = \int_{-\infty}^{\infty} R_x(t)e^{-j\omega t}dt$$

If the Fourier transform exists, we can define the inverse Fourier transform as

$$R_x(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_x(w)e^{j\omega\tau}d\omega$$

Definition 3.13 (Cross-Power Spectral Density)
For two jointly wide-sense stationary processes $x(t), y(t)$, we define the cross power spectral density to be

$$S_{xy}(\omega) = \int_{-\infty}^{\infty} R_{xy}(t)e^{-j\omega t}dt$$

Given the properties of the autocorrelation function $R_x(\tau)$, we can establish the following properties of the power spectral density function:

1. $S_x(\omega)$ is real-valued, even if the process $x(t)$ is complex-valued.

2. If $x(t)$ is real-valued, then $S_x(\omega)$ is an even function, since $R_x(\tau)$ is an even function.

3. $S_x(\omega)$ is nonnegative.

Proving the first two properites is simple; the third property will be shown later when discussing linear systems. Consider the first property: note that $R_x(\tau) = R_x^*(-\tau)$. Thus,

$$
\begin{aligned}
\int_{-\infty}^{\infty} R_x(t)e^{-j\omega t}dt &= \int_0^{\infty} R_x(t)e^{-j\omega t}dt + \int_{-\infty}^0 R_x(t)e^{-j\omega t}dt \\
&= \int_0^{\infty} R_x(t)e^{-j\omega t}dt + \int_0^{\infty} R_x(-t)e^{j\omega t}dt \\
&= \int_0^{\infty} (R_x(t)e^{-j\omega t} + (R_x(t)e^{-j\omega t})^*)dt
\end{aligned}
$$

which is clearly real-valued, since the integrand is real-valued. If $x(t)$ is real-valued, then the integral is symmetric about $\omega = 0$, and so $S_x(\omega)$ is an even function.

The power spectral density has the interpretation of a density function for average power in the random process $x(t)$ per unit frequency. For instance, consider a wide-sense stationary process $x(t)$, and consider the random variable at frequency ω, defined by the following integral:

$$
F_T(\omega) = \int_{-T}^T x(t)e^{-j\omega t}dt
$$

This is a complex-valued random variable, and, as $T \to \infty$, looks like the Fourier transform of the sample path of $x(t)$. Note that the above integral exists as long as $x(t)$ has reasonable sample paths (e.g. bounded). The square of the magnitude of this random variable is given by

$$
F_T(\omega)F_T^*(\omega) = \int_{-T}^T \int_{-T}^T x(t)x^*(s)e^{-j\omega(t-s)}\, ds\, dt
$$

Thus,

$$
\begin{aligned}
E\left[|F_T(\omega)|^2\right] &= \int_{-T}^T \int_{-T}^T R_x(t-s)e^{-j\omega(t-s)}\, dt\, ds \\
&= 2T \int_{-2T}^{2T} \left(1 - \frac{|\tau|}{2T}\right) R_x(\tau)e^{-j\omega\tau}\, d\tau
\end{aligned}
\tag{3.51}
$$

Thus, the average power in the random variable $F_T(\omega)$ is given by

$$
\frac{1}{2T}E\left[|F_T(\omega)|^2\right] = \int_{-2T}^{2T} \left(1 - \frac{|\tau|}{2T}\right) R_x(\tau)e^{-j\omega\tau}d\tau
$$

As $T \to \infty$, we see that this integral converges to $S_x(\omega)$! Thus,

$$
S_x(\omega) = \lim_{T\to\infty} \frac{1}{2T}E\left[|F_T(\omega)|^2\right]
$$

This provides the interpretation that $S_x(w)$ is the average power in a sample of $x(t)$ at frequency ω. This also establishes that $S_x(\omega)$ must be nonnegative.

Let's consider several examples of autocorrelation functions and compute their corresponding power spectral density functions.

Example 3.6
Consider the white noise process $w(t)$, with autocorrelation function $R_w(\tau) = \delta(\tau)$. As we discussed previously, we can always compute integrals of these functions. Thus, we can take Fourier transforms to obtain $S_w(\omega) = 1$. This is why the name "white noise" is used: the white noise process contains every frequency with uniform intensity.

Example 3.7

Consider a wide-sense stationary process $x(t)$ with autocorrelation function $R_x(\tau) = e^{-a|\tau|}$, where $a > 0$. Then,

$$
\begin{aligned}
S_x(\omega) &= \int_{-\infty}^{\infty} e^{-a|t|} e^{-j\omega t} dt \\
&= \int_{-\infty}^{0} e^{(-j\omega + a)t} \, dt + \int_{0}^{\infty} e^{(-j\omega - a)t} \, dt \\
&= \frac{1}{-j\omega + a} - \frac{1}{-j\omega - a} = \frac{2a}{a^2 + \omega^2}
\end{aligned}
\tag{3.52}
$$

Example 3.8

Consider a wide-sense stationary process $x(t)$ with autocorrelation function $R_x(\tau) = \max(1 - \frac{|\tau|}{T}, 0)$. Since the triangular shape is the convolution of two rectangular shapes of width T and height $\sqrt{1/T}$, the power spectral density is the square of the transform of the power spectral density of a rectangular pulse. This is given by:

$$
\begin{aligned}
S_p(\omega) &= \int_{-T/2}^{T/2} \frac{1}{\sqrt{T}} e^{-j\omega t} \, dt \\
&= \frac{1}{j\omega\sqrt{T}}(-e^{-j\omega T/2} + e^{j\omega T/2}) = \sqrt{T}\frac{\sin \omega T/2}{\omega T/2}
\end{aligned}
\tag{3.53}
$$

so that the power spectral density of $x(t)$ is given by

$$
S_x(\omega) = T(\frac{\sin \omega T/2}{\omega T/2})^2
$$

Below is a list of properties of autocorrelation functions of wide-sense stationary processes and their corresponding power spectral density functions. Let $x(t)$ be wide-sense stationary with autocorrelation function $R_x(\tau)$ and power spectral density $S_x(\omega)$. Then,

1. For any constant a, the process $ax(t)$ has autocorrelation function $|a|^2 R_x(\tau)$ and power spectral density $|a|^2 S_x(\omega)$.

2. Let $x(t)$, $y(t)$ be orthogonal, jointly wide-sense stationary processes (that is, $R_{xy}(\tau) = 0$. Define the process $z(t) = x(t) + y(t)$. Then $R_z(\tau) = R_x(\tau) + R_y(\tau)$, and $S_z(\omega) = S_x(\omega) + S_y(\omega)$.

3. The autocorrelation function of $\frac{d}{dt}x(t)$ is given by $-\frac{d^2}{d\tau^2}R_x(\tau)$, and the power spectral density by $\omega^2 S_x(\omega)$.

4. The autocorrelation of $x(t)e^{j\omega_0 t}$ is $R_x(\tau)e^{j\omega_0 \tau}$ and its power spectral density is $S_x(\omega - \omega_0)$.

5. If $\mu_x = 0$, the autocorrelation function of $x(t) + b$ is $R_x(\tau) + |b|^2$, and its power spectral density function is $S_x(\omega) + 2\pi|b|^2\delta(\omega)$.

6. The autocorrelation of $x(t)\cos(\omega_0 t + \theta)$, where θ is uniformly distributed in $[-\pi, \pi]$ and independent of $x(t)$, is given by $\frac{1}{2}R_x(\tau)\cos \omega_0\tau$, and its power spectral density by $\frac{1}{4}S_x(\omega - \omega_0) + \frac{1}{4}S_x(\omega + \omega_0)$.

To conclude this section, consider now vector-valued wide-sense stationary processes. The concept of autocorrelation functions extends naturally to this setting, leading to matrix-valued autocorrelation functions. Clearly, a similar extension is easy to establish for power spectral density functions; for vector-valued processes, these will be matrix-valued functions of frequency ω, defined as the matrix of Fourier transforms of the elements of the autocorrelation matrix function. Properties of this matrix-valued power spectral density function follow naturally from the definition of the Fourier transform and the properties of the autocorrelation functions for vector-valued processes.

Chapter 4

Mean-Square Calculus for Stochastic Processes

The purpose of this section is to allow us to define derivatives and integrals of stochastic processes. Consider, for example, a simple electrical circuit, composed of a voltage source V, a resistor R and a capacitor C in series. The equation for the capacitor voltage $V_c(t)$ is given by

$$V(t) = RC\frac{d}{dt}V_c(t) + V_c(t)$$

and we can solve this in terms of an impulse response $h(t) = e^{\frac{-t}{RC}}u(t)$ where $u(t)$ is the unit step function. The resulting solution is

$$V_c(t) = \int_0^t h(t-s)V(s)ds$$

for $t \geq 0$.

The question before us is what happens when the input voltage $V(t)$ is a stochastic process, rather than a deterministic time function? Clearly, we should expect that, in some well-defined sense, the output $V_c(t)$ is also a stochastic process. This forces us to define what we mean by an integral of a stochastic process, particularly when the random sample functions $V(t, \omega)$ may not be continuous as a function of time. Thus, it is difficult to define integrals or derivatives of random processes strictly in terms of the individual sample functions.

4.1   Continuity of Stochastic Processes

Since definition of integration and differentiation is based on the use of limits, we want to define concepts of continuity of stochastic processes as a function of time, so that we can also define what we mean by a limit. Again, one concept would be that every sample function of the process would have to be continuous. However, this would severely restrict the class of stochastic process for which the calculus would be defined.

Given the different concepts of convergence for sequences of random variables discussed previously, the most appropriate concepts are those of almost sure convergence and mean-square convergence. Thus, we have the following definition.

Definition 4.1 (Almost Sure Continuity)
The stochastic process $x(t)$ has *almost sure continuous sample paths* at time t if

$$\lim_{\epsilon \to 0} x(t + \epsilon) \overset{\text{a.e.}}{=} x(t)$$

where the limit is interpreted in almost sure sense. If the sample paths are almost sure continuous at every t, the process is said to have continuous sample paths almost surely.

Definition 4.2 (Mean Square Continuity)
The stochastic process $x(t)$ is *continuous* in mean-square sense at time t if

$$\lim_{\epsilon \to 0} x(t + \epsilon) \overset{\mathsf{mss}}{=} x(t)$$

where the limit is interpreted in the mean-square sense. That is,

$$\lim_{\epsilon \to 0} E[(x(t + \epsilon) - x(t))^2] = 0$$

If it is mean-square continuous at every t, it is said to be mean-square continuous everywhere or simply mean-square continuous.

The advantage of mean-square continuity versus almost sure continuity of sample paths is that the mean-square continuity can be verified in terms of the second-order properties of the process, using autocorrelation functions. Note that

$$
\begin{aligned}
E[(x(t + \epsilon) - x(t))^2] &= E[x(t + \epsilon)^2] + E[x(t)^2] - 2E[x(t)x(t + \epsilon)] \\
&= R_x(t + \epsilon, t + \epsilon) - R_x(t, t + \epsilon) + R_x(t, t) - R_x(t, t + \epsilon) \quad (4.1)
\end{aligned}
$$

Thus, if $R_x(t, s)$ were continuous at $s = t = t_0$, then $x(t)$ would be mean-square continuous at t_0. If the process were stationary, then continuity of $R_x(\tau)$ at $\tau = 0$ would be sufficient. Thus, all questions of stochastic convergence can be posed in terms of questions of deterministic convergence of the corresponding autocorrelation functions!

Let's consider two examples of stochastic processes: the Brownian motion process $x(t)$, and the Poisson process $n(t)$. For the Brownian motion process, the autocorrelation function is given by $R_x(t, s) = \min(t, s)$. This function is continuous at (t, t) for any t, and thus Brownian motion is continuous everywhere in mean-square sense. As it turns out, the sample paths $x(t, \omega)$ can be shown to be continous everywhere with probability one. Thus, Brownian motion is an example of a process with continous sample functions.

What about Poisson processes? Their sample paths are discontinuous everwhere there is a discrete change in value. The autocorrelation function of Poisson processes with rate λ is give by $R_n(t, s) = \lambda^2 st + \lambda \min(s, t)$. Again, this is continuous at (t, t) for any t, and thus Poisson processes are mean-square continuous everywhere!

There are some strong implications of mean-square continuity everywhere, which are summarized below:

1. If a stochastic process x is mean-square continuous, then $\lim_{\epsilon \to 0} E[g(x(t + \epsilon))] = E[g(x(t)]$ for every continous function g.

2. x is stationary, and $R_x(t)$ is continuous at $t = 0$ if and only if x is mean-square continuous everywhere.

The first property follows because mean-square convergence of random variables implies convergence in probability (or in distribution), so that expectations being integrals of probabilities will also converge. In particular, the mean and all of the moments of the process will be continous functions of time.

As a final note on continuity, we have said little on how to show almost sure sample continuity of a stochastic process, because it is difficult to establish sufficiency theorems. However, we want to mention some results which can be used to establish this property. We state these without proof; the interested reader is referred to Loeve's book on probability theory, or other advanced probability texts.

Theorem 4.1
Let $g(h)$ and $q(h)$ be functions that are even, nondecreasing for $h > 0$, such that $g(h) \to 0, q(h) \to 0$ as $h \to 0$, and such that

$$\sum_{n=1}^{\infty} g(2^{-n}) < \infty, \sum_{n=1}^{\infty} 2^n q(2^{-n}) < \infty$$

Then, $x(t)$ has almost surely continuous sample paths if

$$P[\{\omega : |x(t + h) - x(t)| \ge g(h)\}] \le q(h)$$

for all t, h.

Theorem 4.2
Assume that one can find positive constants $p < r, k$ such that, for all t, h,

$$E[|x(t+h) - x(t)|^p] \leq \frac{k|h|}{|\log|h||^{1+r}}$$

Then, $x(t)$ has almost surely continuous sample paths.

A sufficient condition for the above result can be obtained using $p = 2$, in which case autocorrelation functions can be used! This states:

Theorem 4.3
If, for all t, h, we have

$$R_x(t+h, t+h) + R_x(t,t) - R_x(t+h,t) - R_x(t,t+h) < \frac{k|h|}{|\log|h||^s}$$

where $k > 0, s > 3$, then $x(t)$ has almost surely continuous sample paths.

In particular, the condition of the above result is satisfied if $\frac{R_x(t+h,t+h)+R_x(t,t)-R_x(t+h,t)-R_x(t,t+h)}{h^2}$ is bounded for all sufficiently small h. A sufficient condition for this is stated in terms of the autocorrelation function, as

$$\frac{\partial^2}{\partial u \partial v} R_x(u,v)|_{u=v=t} < \infty$$

As an example, consider the Brownian motion process, with autocorrelation function $R_x(t,s) = \min(t,s)$. Note that the second derivatives in the last result do not exist. However, we know that a Brownian increment is Gaussian, with variance proportional to the length of the interval. Let $0 < a < 1/2$; then,

$$P[\{\omega : |x(t+h) - x(t)| \geq |h|^a\}] = \frac{2}{\sqrt{2\pi}} \int_{|h|^{a-1/2}}^{\infty} e^{-x^2/2} ds$$

$$\leq \frac{2}{\sqrt{2\pi}} |h|^{1/2-a} e^{-0.5|h|^{2a-1}} \tag{4.2}$$

where the last inequality follows from integration by parts. Now, use the first result quoted above, by letting $g(h) = |h|^a$, and letting

$$q(h) = \frac{2}{\sqrt{2\pi}} |h|^{1/2-a} e^{-0.5|h|^{2a-1}}$$

Then,

$$\sum n = 1^\infty g(2^{-n}) = \sum n = 1^\infty (2^{-a})^n = \frac{1}{1 - 2^{-a}} < \infty$$

$$\sum n = 1^\infty 2^n q(2^{-n}) = \frac{2}{\sqrt{2\pi}} \sum n = 1^\infty 2^n 2^{-n/2+na/2} e^{-0.5 \cdot 2^{n(1-2a)}}$$

which converges because $a < 0.5$. Thus, this establishes that Brownian motion has almost surely continuous sample paths.

4.2 Mean-Square Differentiation

The normal definition for a derivative of a deterministic function of time is

$$\frac{d}{dt}x(t) = \lim_{\epsilon \to 0} \frac{x(t+\epsilon) - x(t)}{\epsilon} \tag{4.3}$$

For stochastic processes, the issue is in what sense will we define the limit. Again, we want to use the concept of mean-square convergence. The limit will also be a stochastic process.

For a stochastic process $x(t)$, if the limit in (4.3) exists in mean-square sense, then we say that the stochastic process limit $\frac{d}{dt}x(t)$ is the mean-square derivative of $x(t)$. More formally,

Definition 4.3

The stochastic process $x(t)$ has mean-square derivative $\frac{d}{dt}x(t)$ at t if

$$\lim_{\epsilon \to 0} \frac{x(t+\epsilon) - x(t)}{\epsilon} \stackrel{\text{mss}}{=} \frac{d}{dt}x(t)$$

That is,

$$\lim_{\epsilon \to 0} E\left[\left(\frac{x(t+\epsilon) - x(t)}{\epsilon} - \frac{d}{dt}x(t) \right)^2 \right] = 0$$

Given a stochastic process $x(t)$, how can we determine easily if it is mean-square differentiable at a particular time t? First, we use the Cauchy criterion for convergence of sequences, which states that, if a sequence converges, then the distance between elements of that sequence also converges. In our case, this means

$$
\begin{aligned}
0 &= \lim_{\substack{\epsilon \to 0 \\ \delta \to 0}} E\left[\left(\frac{x(t+\epsilon) - x(t)}{\epsilon} - \frac{x(t+\delta) - x(t)}{\delta} \right)^2 \right] \\
&= \lim_{\substack{\epsilon \to 0 \\ \delta \to 0}} E\left[\left(\frac{x(t+\epsilon) - x(t)}{\epsilon} \right)^2 + \left(\frac{x(t+\delta) - x(t)}{\delta} \right)^2 - 2\frac{x(t+\delta) - x(t)}{\delta}\frac{x(t+\epsilon) - x(t)}{\epsilon} \right] \\
&= \lim_{\substack{\epsilon \to 0 \\ \delta \to 0}} \left[\frac{R_x(t+\epsilon, t+\epsilon) + R_x(t,t) - 2R_x(t+\epsilon, t)}{\epsilon^2} + \frac{R_x(t+\delta, t+\delta) + R_x(t,t) - 2R_x(t+\delta, t)}{\delta^2} \right. \\
&\qquad \left. - 2\frac{R_x(t+\epsilon, t+\delta) + R_x(t,t) - R_x(t+\epsilon, t) - R_x(t+\delta, t)}{\epsilon\delta} \right]
\end{aligned}
\tag{4.4}
$$

Now, note that, if the autocorrelation function R_x was twice differentiable at (t,t), so that

$$
\begin{aligned}
\frac{\partial^2}{\partial u \partial v} R_x(u,v) \bigg|_{u=v=t} &= \lim_{\epsilon \to 0} \frac{R_x(t+\epsilon, t+\epsilon) + R_x(t,t) - 2R_x(t+\epsilon, t)}{\epsilon^2} \\
&= \lim_{\delta \to 0} \frac{R_x(t+\delta, t+\delta) + R_x(t,t) - 2R_x(t+\delta, t)}{\delta^2} \\
&= \lim_{\substack{\epsilon \to 0 \\ \delta \to 0}} \frac{R_x(t+\epsilon, t+\delta) + R_x(t,t) - R_x(t+\epsilon, t) - R_x(t+\delta, t)}{\epsilon\delta}
\end{aligned}
\tag{4.5}
$$

exists, then (4.4) is true!

We summarize the above existence conditions in the following theorems.

Theorem 4.4

A stochastic process $x(t)$ is mean-square differentiable if $\frac{\partial^2}{\partial u \partial v} R_x(u,v)|_{u=v=t}$ exists for all t.

Theorem 4.5

A *stationary* stochastic process $x(t)$ is mean-square differentiable *if and only if* $\frac{d^2}{ds^2} R_x(s)|_{s=0}$ exists.

Note the stronger conditions that we can provide for stationary processes. Mean-square differentiability provides several useful conditions for stationary processes. For any stationary process which is mean-square differentiable, we have

$$E\left[\frac{d}{dt}x(t) \right] = \lim_{\epsilon \to 0} E\left[\frac{x(t+\epsilon) - x(t)}{\epsilon} \right] = \lim_{\epsilon \to 0} \frac{m_x - m_x}{\epsilon} = 0$$

where the interchange of differentiation and expectation is possible due to the existence of the limit.

For general stochastic processes, which are mean-square differentiable, we can compute the autocorrelation statistics of their derivatives based on the statistics of the original process, as follows. Define

$y(t) = \frac{d}{dt}x(t)$ for a stochastic process $x(t)$. Then,

$$
\begin{aligned}
R_{xy}(s,t) &= E\left[x(s)y(t)\right] = E\left[x(s)\lim_{\epsilon \to 0}\left\{\frac{x(t+\epsilon)-x(t)}{\epsilon}\right\}\right] \\
&= \lim_{\epsilon \to 0}E\left[x(s)\left\{\frac{x(t+\epsilon)-x(t)}{\epsilon}\right\}\right] = \lim_{\epsilon \to 0}\frac{R_x(s,t+\epsilon)-R_x(s,t)}{\epsilon} \\
&= \frac{\partial}{\partial t}R_x(s,t)
\end{aligned}
\tag{4.6}
$$

Using a similar argument, we establish that

$$
R_y(s,t) = \frac{\partial^2}{\partial s \partial t}R_x(s,t)
\tag{4.7}
$$

As before, the above equations simplify when x is stationary, to give

$$
R_y(s,t) = R_y(t-s) = \frac{\partial}{\partial s}\frac{\partial}{\partial t}R_x(t-s) = -\frac{d^2}{d\tau^2}R_x(\tau)
\tag{4.8}
$$

where $\tau = t - s$, and the negative sign arises due to the negative sign of s in the argument of $R_x(t-s)$.

To conclude this section, let's consider our canonical examples of Brownian motion and Poisson processes. We determined that they were mean-square continuous. Are they differentiable? Note that neither process is stationary. For Brownian motion, $R_x(t,s) = \min(t,s)$. As a function of s, we can compute its derivative as

$$
\frac{\partial}{\partial s}\min(t,s) = u(s-t)
$$

where u is the unit step function. Differentiating with respect to t, we get

$$
\frac{\partial}{\partial t}u(s-t) = \delta(s-t)
$$

where δ is the generalized impulse function we use formally; in fact, the derivative is not defined in a regular form, and the limit does not exist, since the value of the impulse function is infinite when its argument is zero. Thus, Brownian motion is not a mean-square differentiable process; however, it is often useful to keep track of formal derivatives using generalized functions. Indeed, we will often use the concept of *white noise* as a process with autocorrelation function $R_x(t,s) = \delta(s-t)$ as an engineering concept; formally, our analysis above suggests that white noise can be interpreted as the "derivative" of Brownian motion. White noise is a very useful concept in the modeling of broadband noise in communicatons and radar systems, and will be studied in greater detail later in the course. However, there is a rich mathematical theory which focuses on the study of Brownian motion and white noise, which is beyond the scope of our course.

Similar computations for Poisson processes establish that $R_y(s,t) = \lambda^2 + \lambda\delta(s-t)$. Thus, Poisson processes are also not mean-square differentiable. If one thought somewhat as to what can be generalized from these examples, most continuous-time processes with independent increments will not be mean-square differentiable, because the autocorrelation function will depend on $\min(t,s)$.

4.3 Mean-Square Integration

We can define mean-square integrals using a limiting process, as before. To begin with, consider an integral of a stochastic process $x(t,\omega)$ over the interval $[s,t]$. For each sample path ω, we can construct an integral exactly the way we would construct it for deterministic functions: sampling the sample path over a discrete grid, computing Riemann sums, and taking the limit as the grid gets finer. The key question is, in what sense is the limit interpreted? As before, we will use the mean-square sense.

Mathematically, consider a sample path $x(t,\omega)$ over the interval $[s,t]$. Let $\Delta = (t-s)/N$ be the increment in a regular discretization of the interval. We define the integral $y(t,\omega)$ as

$$
y(t,\omega) = \int_s^t x(\tau,\omega)d\tau \stackrel{\text{mss}}{=} \lim_{N \to \infty}\sum_{i=1}^{N}x(s+i\Delta,\omega)\Delta
\tag{4.9}
$$

Interpreting the limit in the appropriate sense, this means

$$\lim_{N\to\infty} E\left[\left(y(t,\omega) - \sum_{i=1}^{N} x(s+i\Delta,\omega)\Delta\right)^2\right] = 0$$

As in the case of differentiation, we are interested in conditions which guarantee that a process is integrable. Expanding the above expression, we get

$$
\begin{aligned}
E[y(t)] &= E\left[\lim_{N\to\infty}\sum_{i=1}^{N} x(s+i\Delta)\Delta\right] \\
&= \lim_{N\to\infty}\sum_{i=1}^{N} E\left[x(s+i\Delta)\right]\Delta = \int_s^t m_x(\tau)d\tau
\end{aligned}
\tag{4.10}
$$

When does the integral exist? Applying the Cauchy criterion, we get

$$0 = \lim_{N,M\to\infty} E\left[\left(\sum_{i=1}^{N} x(s+i\Delta_N)\Delta_N - \sum_{j=1}^{M} x(s+j\Delta_M)\Delta_M\right)^2\right] \tag{4.11}$$

The above convergence is guaranteed if we can show that

$$E\left[\lim_{N\to\infty}\left(\sum_{i=1}^{N} x(s+i\Delta)\Delta\right)^2\right] < \infty$$

Expanding and taking expectations,

$$
\begin{aligned}
E\left[\lim_{N\to\infty}\left(\sum_{i=1}^{N} x(s+i\Delta)\Delta\right)^2\right] &= \lim_{N\to\infty}\sum_{i=1}^{N}\sum_{j=1}^{N} R_x(s+i\Delta, s+j\Delta)\Delta^2 \\
&= \int_s^t\int_s^t R_x(\sigma,\tau)d\sigma\, d\tau
\end{aligned}
\tag{4.12}
$$

This is formalized in the following result.

Theorem 4.6
The mean-square integral of a stochastic process $x(t)$ exists over an interval $[s,t]$ if

$$\int_s^t\int_s^t R_x(\sigma,\tau)\,d\sigma\,d\tau < \infty$$

Note when the process is wide-sense stationary we can simplify the condition somewhat. In particular, we can perform the following change of variables:

$$
\begin{aligned}
u &= \tau - \sigma \\
v &= \tau + \sigma
\end{aligned}
$$

The Jacobian of this transformation is 2, so that $d\sigma\,d\tau = (1/2)du\,dv$. We also need to transform the limits of integration. Note that the original limits correspond to a square in the (σ,τ) plane and in the new coordinates (which are a 45 degree rotation), this region will be a diamond. Thus the transformed integral is given by:

$$
\begin{aligned}
\int_s^t\int_s^t R_x(\tau-\sigma)\,d\sigma\,d\tau &= \int_0^{t-s}\left(\int_{2s+u}^{2t-u} R_x(u)\left(\frac{1}{2}\right)dv\right)du + \int_{-(t-s)}^0\left(\int_{2s-u}^{2t+u} R_x(u)\left(\frac{1}{2}\right)dv\right)du \\
&= \int_{-(t-s)}^{t-s} R_x(u)\left(\int_{2s+|u|}^{2t-|u|}\left(\frac{1}{2}\right)dv\right)du \\
&= \int_{-(t-s)}^{t-s} ((t-s)-|u|)\,R_x(u)\,du = 2\int_0^{t-s} ((t-s)-\tau)\,R_x(\tau)\,d\tau
\end{aligned}
$$

For the mean-square integral of a wide-sense stationary process to exist we want the integral above to exist (i.e. be finite). Note, for example, that this will be the case if $R_x(\tau)$ is absolutely integrable. This observation yields the following result:

Theorem 4.7

The mean-square integral of a wide-sense stationary stochastic process $x(t)$ exists over an interval $[s, t]$ if

$$\int_0^{t-s} |R_x(\tau)| \, d\tau < \infty$$

As in the case of differentiation, we are interested in the relationship between the autocorrelation of the integral process and the original process. If the integral exists, this is easily computed through exchange of expectation and integration, as follows. Let $y(t, \omega) = \int_s^t x(\tau, \omega) d\tau$. Then,

$$
\begin{aligned}
R_y(a, b) &= E\left[\int_s^a x(\tau, \omega) d\tau \int_s^b x(\sigma, \omega) d\sigma\right] \\
&= \int_s^a \int_s^b E[x(\tau)x(\sigma)] d\tau d\sigma \\
&= \int_s^a \int_s^b R_x(\tau, \sigma) d\tau d\sigma \quad (4.13)
\end{aligned}
$$

For wide-sense stationary processes, the above integral can be simplified, as follows: Assume $a \le b$; then

$$R_y(a, b) = \int_s^a \int_s^b R_x(\tau - \sigma) d\tau d\sigma$$

Make the variable substitution $u = \tau - \sigma, v = 0.5(\tau + \sigma)$. This results in

$$
\begin{aligned}
R_y(a, b) &= \int_{s-a}^0 \int_{s-u/2}^{a+u/2} R_x(u) dv du + \int_0^{b-a} \int_{s+u/2}^{a+u/2} R_x(u) dv du + \int_{b-a}^{b-s} \int_{s+u/2}^{b-u/2} R_x(u) dv du \\
&= \int_{s-a}^0 (a - s + u) R_x(u) du + \int_0^{b-a} (a - s) R_x(u) du + \int_{b-a}^{b-s} (b - s - u) R_x(u) du \quad (4.15)
\end{aligned}
$$

In particular, if $a = b$, this simplifies to

$$R_y(a, a) = \int_{s-a}^{a-s} (a - s - |u|) R_x(u) du \quad (4.16)$$

Example 4.1

Let $x(t)$ be a wide-sense stationary process. Consider the moving average process $y(t)$ defined as

$$y(t) = \frac{1}{2T} \int_{t-T}^{t+T} x(s) ds$$

What are the mean and covariance statistics of y?

We answer this question using the above properties of integration. First, we assume that the autocorrelation function $R_x(\tau)$ satisfies appropriate integrability conditions. Then, by (4.16), we have:

$$
\begin{aligned}
R_y(t, t) &= \frac{1}{4T^2} E\left[\int_{t-T}^{t+T} \int_{t-T}^{t+T} x(\sigma)x(\tau) \, d\sigma \, d\tau\right] \\
&= \frac{1}{4T^2} \int_{t-T}^{t+T} \int_{t-T}^{t+T} R_x(\sigma - \tau) \, d\tau \, d\sigma \\
&= \frac{1}{4T^2} \int_{-2T}^{2T} R_x(u)(2T - |u|) \, du \quad (4.17)
\end{aligned}
$$

By a similar computation, we compute m_y as

$$m_y = \frac{1}{2T} \int_{t-T}^{t+T} E[x(s)]ds = m_x \tag{4.18}$$

Thus, the covariance of $y(t)$ is given by $\sigma_y^2 = R_y(t,t) - m_x^2$. If we let the autocorrelation function of x have a special form, such as $R_x(t) = e^{-2a|t|}$, then we can evaluate this as

$$\sigma_y^2 = \frac{1}{2Ta} - \frac{1}{8a^2T^2}(1 - e^{-4aT})$$

The above example raises some interesting questions concerning the relationship of time averages of a process and statistics of a process. In particular, note that, as $T \to \infty$, as long as $R_x(u)$ decays to 0 fast enough (as it has to, in order to be integrable), then the covariance of the process must approach 0! Thus, the process must be approaching a deterministic constant! Indeed, we can determine from the above example that the constant must be the process mean, m_x. Thus, we have an example where, for every sample trajectory, if we take a long enough temporal average, this average converges to the true mean of the process, as averaged over all possible sample paths. This type of property is called an *ergodic* property. We discuss ergodicity in greater detail later.

There are other sample statistics of interest, which can be defined for each sample trajectory of wide-sense stationary processes. For instance, instead of the sampled mean, we can define the sample autocorrelation of a trajectory with itself as:

$$\langle x(t+\tau)x(t)\rangle_T = \frac{1}{2T}\int_{-T}^{T} x(t+\tau,\omega)x(t,\omega)dt \tag{4.19}$$

Since this is defined for each trajectory, the resulting time average is a random variable. The question of interest is determining conditions which guarantee that

$$\lim_{T\to\infty} \langle x(t+\tau)x(t)\rangle_T = R_x(\tau)$$

Note that, in general, the limit will exist, as long as $x(t)$ is stationary; however, it may be a random process also! In order to show that it is a constant, we must analyze its covariance, and show that the covariance goes to zero.

Example 4.2 (Strong Law of Large Numbers)
Let $y(n)$ be a sequence of independent, identically distributed, zero-mean random variables, and let s be a constant. Define $x(n) = s + y(n)$. Define the moving average of x as

$$S_N = \frac{1}{N}\sum_{i=1}^{N} x(i)$$

Note that the mean

$$m_{S_N} = \frac{1}{N}\sum_{i=1}^{N} E[x(i)] = s$$

and the variance is

$$\sigma_{S_N}^2 = \frac{1}{N^2}\sum_{i=1}^{N} E[y^2(i)] = \frac{1}{N}\sigma_y^2$$

Thus, as $N \to \infty$, the variance goes to zero, and we have

$$\lim_{N\to\infty} S_N \overset{a.e.}{=} s$$

This is the strong law of large numbers; in essence, it says that the time average of $x(n)$ converges to its expected value almost everywhere.

4.4 Integration and Differentiation of Gaussian Stochastic Processes

When the original stochastic process $x(\cdot)$ is a Gaussian random process, will its integral and derivative also be Gaussian random processes? The answer is in the affirmative, as we will show below.

The key observation is that convergence in mean-square sense implies convergence in distribution. Thus, consider a sequence $\underline{x}_n$ of jointly Gausssian random vectors, which satisfy the Cauchy criterion (i.e. a Cauchy sequence): For any $\epsilon > 0$, there exists an $N(\epsilon)$ such that, for $n, m > N(\epsilon)$

$$E\left[(\underline{x}_n - \underline{x}_m)^T (\underline{x}_n - \underline{x}_m)\right] < \epsilon$$

As we discussed in the convergence section, such a sequence is guaranteed to converge in the mean-square sense to a random vector $\underline{x}$; furthermore, each member of the sequence has a Gaussian distribution. Since mean-square sense convergence implies convergence in distribution, this requires that $\underline{x}$ also have a vector Gaussian distribution.

Now, consider a Gaussian stochastic process $x(\cdot)$. For any sampling times $t_1, \ldots, t_k$, the derivative of x will be the limit in mean-square sense of the Gaussian vector

$$\begin{bmatrix} \frac{x(t_1+\epsilon)-x(t_1)}{\epsilon} \\ \frac{x(t_2+\epsilon)-x(t_2)}{\epsilon} \\ \vdots \\ \frac{x(t_k+\epsilon)-x(t_k)}{\epsilon} \end{bmatrix}$$

By the above argument, this limit will also have a Gaussian probability density function for any sampling times $t_1, \ldots, t_k$, and thus the derivative process will also be Gaussian.

Similarly, the integral of a Gaussian stochastic process will also be Gaussian. That follows because the integral is defined as a limit in mean-square sense of a sum of jointly Gaussian random variables, and sums of jointly Gaussian random variables are Gaussian.

4.5 Generalized Mean-Square Calculus

Since white noise is a very useful abstraction in the analysis of engineering systems, we want to understand in what sense is white-noise defined. As we discussed before, if we define white noise $w(t)$ as the formal derivative of Brownian motion $b(t)$, then we should have as an autocorrelation function

$$R_w(t, s) = \delta(t - s)$$

The purpose of this subsection is to provide additional intuition into the construction of the white noise process.

Consider the increment process of Brownian motion, defined as

$$w_\Delta(t) = \frac{b(t + \Delta) - b(t)}{\Delta}$$

By the properties of Brownian motion, this is a zero-mean, Gaussian process with autocorrelation

$$\begin{aligned} R_{w_\Delta}(t, s) &= E\left[\frac{b(t + \Delta) - b(t)}{\Delta} \frac{b(s + \Delta) - b(s)}{\Delta}\right] \\ &= \begin{cases} 0 & \text{if } |t - s| > \Delta \\ \frac{\Delta - |t-s|}{\Delta^2} & \text{otherwise} \end{cases} \end{aligned} \tag{4.20}$$

Note that $R_{w_\Delta}(t, s)$ is a function only of the difference $\tau = s - t$, and thus the process w_Δ is wide-sense stationary. Since the process is Gaussian, in addition, then it is also strict-sense stationary.

A graph of the autocorrelation function $R_{w_\Delta}(\tau)$ would show a triangle of height $1/\Delta$ and base of length 2Δ. Thus, the area under the graph equals 1. As we shrink the size of Δ, the autocorrelation function $R_{w_\Delta}(\tau)$ converges to an impulse function $\delta(\tau)$, which corresponds to the white noise limit.

Note that, even though the white-noise limit is a process which is difficult to construct, or even imagine what a sample path would look like as a function, it is easy to derive its properties. In particular, in a manner completely analogous to the definition of a delta function as a generalized function, we can derive the properties of white noise from the properties of its integral. Formally, since white noise is the limit of Gaussian, stationary processes, it should also be Gaussian and stationary.

The more rigorous way of defining white noise is as follows: Consider any bounded function $h(t)$ (that is, $|h(s)| < C$ for $s \in [0, t]$.) Then, define integrals with respect to the white noise process $w(s)$ as the mean-square limits of integrals with respect to the processes $w_\Delta(s)$, as

$$y(t) \equiv \int_0^t h(s)w(s)ds \overset{\text{mss}}{=} \lim_{\Delta \to 0} \int_0^t h(s)w_\Delta(s)ds$$

Note that, in the limit, the autocorrelation of the right-hand side becomes

$$
\begin{aligned}
R_y(t, s) &= \lim_{\Delta \to 0} \int_0^t \int_0^s h(\sigma)h(\tau)R_{w_\Delta}(\tau, \sigma)\, dt\, d\sigma \\
&= \lim_{\Delta \to 0} \int_0^{\min(t,s)} \int_0^{\max(t,s)} h(\tau)h(\sigma)\frac{\Delta - |\tau - \sigma|}{\Delta^2}I(|\tau - \sigma| < \Delta)\, d\tau\, d\sigma \\
&= \int_0^{\min(t,s)} \int_0^{\max(t,s)} h(\tau)h(\sigma)\delta(\tau - \sigma)\, d\tau\, d\sigma \tag{4.21}
\end{aligned}
$$

where the last equality comes from the definition of a deterministic delta function in terms of limits of integrals. As long as the function h is bounded, the Cauchy criterion can be used to establish that the mean-square limit y will exist, and will have the above autocorrelation. Furthermore, the above expression for autocorrelation is identical to that which would arise from formally assuming that the white noise process existed as an input, with autocorrelation $R_w(t, s) = \delta(t - s)$.

As an example, suppose that we wanted to define the random process $y(t) = \int_0^t w(s)ds$, where w is white noise. Clearly, if white noise is to be viewed as the derivative of Brownian motion, then y(t) = b(t) - b(0). However, let's demonstrate this fact using the above construction. For any $\Delta > 0$, the mean-square integral

$$y_\Delta(t) = \int_0^t w_\Delta(s)ds$$

exists. Define the process $y(t)$ as the mean-square limit of $y_\Delta(t)$, to obtain

$$
\begin{aligned}
y(t) &\overset{\text{mss}}{=} \lim_{\Delta \to 0} \int_0^t w_\Delta(s)ds \overset{\text{mss}}{=} \lim_{\Delta \to 0} \frac{\int_0^t (b(s + \Delta) - b(s))ds}{\Delta} \\
&\overset{\text{mss}}{=} \lim_{\Delta \to 0} \frac{\int_0^t (b(s + \Delta))ds - \int_0^t b(s)ds}{\Delta} \\
&\overset{\text{mss}}{=} \lim_{\Delta \to 0} \frac{\int_t^{t+\Delta} b(s)ds - \int_0^\Delta b(s)ds}{\Delta} \\
&\overset{\text{mss}}{=} b(t) - b(0) \tag{4.22}
\end{aligned}
$$

because the Brownian process is integrable, and the limit converges in mean-square to the derivative of the integral! This states that, from a mean-square sense, the processes $y(t)$ and $b(t) - b(0)$ are indistinguishable; in particular, they have the same autocorrelation and mean.

The important aspect of the above definition is that the statistical properties of integrals of generalized processes obey the rules of the mean square calculus. Specifically, the autocorrelation of $y(t)$ is the limit of the autocorrelation of $y_\Delta(t)$, which is given by:

$$R_{y_\Delta}(t, s) = \int_0^t \int_0^s R_{w_\Delta}(\tau, \sigma)\, dt\, d\sigma \tag{4.23}$$

so that

$$
\begin{aligned}
R_y(t,s) &= \lim_{\Delta \to 0} \int_0^t \int_0^s R_{w_\Delta}(\tau, \sigma) \, d\tau \, d\sigma \\
&= \lim_{\Delta \to 0} \int_0^{\min(t,s)} \int_0^{\max(t,s)} \frac{\Delta - |\tau - \sigma|}{\Delta^2} I(|\tau - \sigma| < \Delta) \, d\tau \, d\sigma
\end{aligned}
$$

Note that, for Δ very small, the integrand is a deterministic integral which is approaching a delta function! Indeed, one can show that the above limit becomes $R_y(t,s) = \min(s,t)$. This is the same expression which would have been obtained from formally substituting $R_w(a,b) = \delta(a-b)$, and using the mean-square calculus to obtain

$$
\begin{aligned}
R_y(t,s) &= \int_0^t \int_0^s R_w(a,b) \, da \, db \\
&= \int_0^t \int_0^s \delta(a-b) \, da \, db \\
&= \int_0^t u(s-b) \, db = \min(s,t) \tag{4.24}
\end{aligned}
$$

which is indeed the autocorrelation function for the Brownian increment $b(t) - b(0)$. In the above equation, $u(t)$ is the unit step function, which is the integral of the delta function. In a similar manner, the mean of $y(t)$ will be the limits of the means of $y_\Delta(t)$, which are computed as

$$
m_y(t) = \lim_{\Delta \to 0} \int_0^t m_{w_\Delta}(s) ds = \lim_{\Delta \to 0} 0 = 0
$$

Thus, formally, we can define white noise as the generalized derivative of Brownian motion, and define the statistics of this derivative process as

$$
m_w(t) = \frac{d}{dt} m_b(t)
$$

$$
R_w(t,s) = \frac{d^2}{dt ds} R_b(t,s)
$$

where the derivatives are taken in a generalized sense, using delta functions. The above discussion shows that we can compute the statistics of the output process $y(t) = \int_0^t h(s) w(s) ds$ as:

$$
m_y(t) = \int_0^t h(s) m_w(s) ds
$$

$$
R_y(t,s) = \int_0^t \int_0^s h(a) h(b) R_w(a,b) da db
$$

and, even though the process $w(t)$ does not exist as a mean-square derivative, integrals of the process can be defined, and the mean-square calculus can be extended in a natural manner using generalized functions to obtain the properties of integrals of white noise.

Note that Brownian motion is not the only process which will have a generalized mean square derivative. Indeed, mean-square continuous independent increment processes such as Poisson processes will also have generalized derivatives which include delta functions. The important item to remember is that the use of delta functions is justified as the limit of ordinary functions, and that integrals of delta functions are well-defined. Thus, for stochastic processes, integrals of generalized mean-square derivatives such as white noise will be well-defined also!

In sum, the standard mean-square calculus can be extended to derivatives of processes which are mean-square continuous, but not mean-square differentiable, by defining generalized processes such as "white noise", with autocorrelation functions which use generalized functions such as delta functions and which can be obtained as the generalized derivatives of the autocorrelation functions of the original process.

To conclude, consider the process $z(t) = \int_0^t f(s)w(s)ds$. It is clear that the mean of $z(t)$ will be the integral of the mean of $f(s)w(s)$, which is zero! Also, since $w(s)$ is Gaussian, the resulting integral will also be Gaussian. Furthermore, the autocorrelation function will be given by

$$
\begin{aligned}
R_z(t,s) &= \int_0^t \int_0^s f(a)R_w(a,b)f(b) \, da \, db \\
&= \int_0^t \int_0^s f(a)\delta(a-b)f(b) \, da \, db \\
&= \int_0^t f(b)u(s-b)f(b) \, db \\
&= \int_0^{\min(t,s)} f^2(b) \, db
\end{aligned}
\tag{4.25}
$$

Note that by defining $f(s) = h(t-s)$ for some causal impulse response function h, we have an integral which looks like the response of a causal linear system to a "white noise" input!

Finally, note that $z(t)$ will be an independent increments process! In particular, since the process is Gaussian, we only have to show that nonoverlapping increments are uncorrelated. Thus, for $t > s > u$, we have

$$
\begin{aligned}
E[(z(t) - z(s))(z(s) - z(u)] &= E\left[\int_s^t f(a)w(a) \, da \int_u^s f(b)w(b) \, db \right] \\
&= \int_s^t \int_u^s f(a)f(b)E[w(a)w(b)] \, da \, db \\
&= \int_s^t \int_u^s f(a)f(b)\delta(a-b) \, da \, db = 0
\end{aligned}
\tag{4.26}
$$

because the intervals do not overlap!

4.6   Ergodicity of Stationary Random Processes

One of the important questions in the analysis of stochastic processes is the determination of the statistical properties of the process. Ideally, in order to compute statistics such as the mean and autocorrelation, we would have to repeat the same experiment many times, obtain many realizations of the required random variables, and average them, in the limit, to obtain an accurate estimate. The strong law of large numbers provides us with the necessary theory to establish that such a procedure works.

In practice, there are many situations where we do not have the flexibility to repeat an experiment! In particular, in many situations we can only gather a single sample path of the stochastic process. For instance, in stock market modeling, we only observe a single time history of the prices; we do not have the luxury of "repeating" the experiment by moving time backwards and reliving the experience. In process control or communications, we observe the noise which is present at a particular time, but again we cannot repeat that experiment at the exact same time.

For most applications, what is needed is some way of learning the needed statistical quantities from a single observed sample function of the process. This is possible primarily for stationary random processes. In particular, for stationary processes, the random variables $x(t)$ and $x(s)$ have identical distributions. Thus, if we were to observe a sample path of the process over a time interval $[-T, T]$, we can generate an estimate of the mean of the process

$$
\frac{1}{2T} \int_{-T}^{T} x(s)ds
$$

The above integral is to be interpreted in the mean-square sense, as discussed previously. Intuitively, it seems that this would be a good estimate, since we are "averaging" many identically distributed random variables. The problem is that they are *not independent*, so that the convergence properties of the law of

large numbers do not apply. Nevertheless, in many cases, the above estimate will converge to the true mean m_x as $T \to \infty$. This property is called *ergodicity in the mean*.

In essence, ergodicity is a property which establishes that certain time averages of sample functions of stochastic processes converge to their corresponding ensemble averages. Although there is a general definition of what we mean by a general ergodic process, we will focus our attention on ergodicity of certain statistics, such as ergodicity of the mean and autocorrelation. In this section we define these concepts, and discuss conditions where we can establish that processes have ergodic properties.

Before discussing the theory, let's discuss why we should expect convergence, in spite of the fact that the samples $x(t), x(s)$ are correlated across time. In particular, in our earlier handout on convergence, we discussed that the law of large numbers can be extended to correlated random variables! What was really needed was the condition that the variance of the weighted sum $s(n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - E[x_i])$ decrease to zero, since, by the Chebyshev inequality, this would imply convergence in probability to zero, and this convergence could be extended to almost-sure convergence. Thus, showing that a process is ergodic in a given statistic will correspond to showing that the variance of that statistic is converging to zero.

What is the mechanism that leads to ergodic processes? In essence, although the process is correlated with itself, what is needed is that the degree of correlation decreases sufficiently rapidly with the time increment, so that the time average in question looks like the average of many uncorrelated random variables, which by the many forms of the weak law of large numbers will converge to the appropriate expectation. Consider the 3 examples below.

Example 4.3
The most trivial example of a stationary process which is not ergodic is the constant process $x(t) = A$, where A is a random variable. Clearly, any average over time will not be a true reflection of the statistics of A, but would merely be a sample value for A.

Example 4.4
Define the stochastic process $x(t) = A \sin t + B \cos t$, where A, B are Gaussian, zero-mean, unit variance, independent random variables. To verify that this process is wide-sense stationary, the autocorrelation function is given by

$$
\begin{aligned}
E[x(t)x(s)] &= E[(A \sin t + B \cos t)(A \sin s + B \cos s)] \\
&= \sin t \sin s + \cos t \cos s = \cos(t - s) \tag{4.27}
\end{aligned}
$$

In this case, it is not clear that the process is ergodic in any statistic, since the random process at time t is strongly correlated with the value at any other time s. However, this correlation fluctuates in sign, so perhaps this can average out.

Example 4.5
Consider the stationary stochastic process $x(t)$ with autocorrelation function $R_x(\tau) = e^{-|\tau|}$. In this case, we should expect that certain statistics would be ergodic, since the autocorrelation function indicates that the strength of the correlation decreases exponentially with the time difference in the two samples.

With those examples in mind, let's proceed to define ergodicity of the different statistics of interest.

Definition 4.4 (Ergodic in the Mean)
A wide-sense stationary process $x(t)$ is *ergodic in the mean* if the time average of $x(t)$ converges in mean-square sense to the ensemble average $E[x(t)] = m_x$. That is,

$$
\lim_{T \to \infty} \langle m_x \rangle_T \equiv \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(s) ds \stackrel{\text{mss}}{=} m_x
$$

Note that, in the above equation, the sample average $\langle x \rangle_T$ is a random variable which is defined in terms of the values of the random process $x(t)$ over the interval $[-T, T]$. Thus, the limit, if it exists, is also a random variable defined in terms of the process samples $x(t), t \in [-\infty, \infty]$. Ergodicity is a statement that says that these special random variables are equal to constants!

Since $\langle m_x \rangle_T$ is a random variable defined in terms of an integral of a stochastic process, we can compute its mean and variance using the theory of mean-square integration, as:

$$
E\left[\langle m_x \rangle_T\right] = \frac{1}{2T} \int_{-T}^{T} E[x(s)] \, ds = m_x
$$

$$
\begin{aligned}
E\left[\left(\langle m_x\rangle_T - m_x\right)^2\right] &= E\left[\left(\frac{1}{2T}\int_{-T}^{T}[x(s)-m_x]\,ds\right)^2\right] \\
&= \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} E[(x(s)-m_x)(x(t)-m_x)]dtds \\
&= \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} K_x(s,t)dtds = \frac{1}{4T^2}\int_{-T}^{T}\int_{-T}^{T} K_x(s-t)dtds \qquad (4.28)
\end{aligned}
$$

Clearly, if this is to converge to a constant as $T \to \infty$, then the variance must decrease to zero. Indeed, this is a necessary and sufficient condition for convergence in the mean-square sense. However, the clumsy part about the condition in eq. (4.28) is that, although the autocovariance function is only a function of the time difference $\tau = t - s$, the integral is stated in terms of integration with respect to both t and s. We can remedy this by switching the variables of integration, using the following coordinate transformation:

$$
\tau = t - s; \sigma = t + s
$$

The Jacobian of this transformation is 2, as verified by simple computation. Thus, $dsdt = 0.5d\sigma d\tau$. We also need to transform the limits of integration. Note that the original limits correspond to a square in the (s, t) plane; in the new coordinates (which are a 45 degree rotation), it will be a diamond. Thus, the transformed integral is given by:

$$
\begin{aligned}
\int_{-T}^{T}\int_{-T}^{T} K_x(t-s)\,dt\,ds &= \int_{0}^{2T}\left(\int_{-2T+\tau}^{2T-\tau} K_x(\tau)\,d\sigma\right)0.5\,d\tau + \int_{-2T}^{0}\left(\int_{-2T-\tau}^{2T+\tau} K_x(\tau)\,d\sigma\right)0.5\,d\tau \\
&= \int_{-2T}^{2T}\left(\int_{-2T+|\tau|}^{2T-|\tau|} K_x(\tau)\,d\sigma\right)0.5\,d\tau = \int_{-2T}^{2T}(2T-|\tau|)\,K_x(\tau)\,d\tau \qquad (4.29)
\end{aligned}
$$

Thus, the condition for ergodicity in the mean becomes

$$
\lim_{T\to\infty}\frac{1}{4T^2}\int_{-2T}^{2T}(2T-|\tau|)K_x(\tau)\,d\tau = 0
$$

We formalize the above in the following result:

Theorem 4.8
A wide-sense stationary process $x(t)$ is ergodic in the mean if and only if the autocovariance function satisfies

$$
\lim_{T\to\infty}\frac{1}{2T}\int_{-2T}^{2T}\left(1-\frac{|\tau|}{2T}\right)K_x(\tau)d\tau = 0
$$

We can also obtain a sufficient condition for ergodicity in the mean. Note that $\left(1-\frac{|\tau|}{2T}\right) \le 1$ in the range of integration. Thus, $\left|\left(1-\frac{|\tau|}{2T}\right)K_x(\tau)\right| \le |K_x(\tau)|$ in the range of integration. This gives a sufficient condition which is easier to verify:

Theorem 4.9
A wide-sense stationary process $x(t)$ is ergodic in the mean if the autocovariance function satisfies

$$
\lim_{T\to\infty}\frac{1}{2T}\int_{-2T}^{2T}|K_x(\tau)|d\tau = 0
$$

In addition to the mean, there are other statistics which we use in characterizing wide-sense stationary processes. We provide definitions for ergodicity of these statistics below:

Definition 4.5 (Ergodic in Mean Square)
A wide-sense stationary stochastic process $x(t)$ is *ergodic in mean square* if

$$
\lim_{T\to\infty}\langle R_x(0)\rangle_T \equiv \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} x^2(s)\,ds \stackrel{\mathsf{mss}}{=} R_x(0)
$$

Definition 4.6 (Ergodic in Autocorrelation)
A wide-sense stationary stochastic process $x(t)$ is *ergodic in autocorrelation* if, for any shift τ,

$$\lim_{T \to \infty} \langle R_x(\tau) \rangle_T \equiv \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(s + \tau)x(s)\, ds \overset{\text{mss}}{=} R_x(\tau)$$

Like ergodicity in the mean, we can develop conditions for verifying that a process is ergodic in mean square or in autocorrelation. We have to establish that the covariance of the random variables $\langle R_x(0) \rangle_T, \langle R_x(\tau) \rangle_T$ decreases to zero in the limit. Unfortunately, this will usually require the computation of higher order (fourth-order) moments of $x(t)$, and one must show that the process has stationarity of fourth-order moments, which is stronger than wide-sense stationarity. For instance, a necessary and sufficient condition for ergodicity in autocorrelation is given in the following result:

Theorem 4.10
A wide-sense stationary stochastic process $x(t)$ is *ergodic in autocorrelation* if and only if

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-2T}^{2T} \left(1 - \frac{|\tau|}{2T} \right) K_{\Phi_s}(\tau)\, d\tau = 0$$

for all s, where

$$K_{\Phi_s}(\tau) = E[x(t + s + \tau)x(t + \tau)x(t + s)x(t)] - R_x(s)^2$$

is the autocovariance of the correlation process $\Phi_s(t) = x(t + s)x(t)$.

The key to the above result is that a new stationary process, $\Phi_s(t)$, is defined in terms of the original process. Then, ergodicity in the mean of this new process is equivalent to ergodicity in autocorrelation for the estimate $\langle R_x(s) \rangle_T$ for shift s. When this is true for all s, then we have ergodicity in autocorrelation of the original process.

Example 4.6
Consider the process $x(t) = A \cos(t + \theta)$, where $A \sim N(0, 1)$, and θ is uniform in $[0, 2\pi]$, and A, θ are independent. Note that this process has mean $m_x(t) = E[A]E[\cos(t + \theta)] = 0$. The autocorrelation of this process is given by:

$$\begin{aligned}
R_x(t, s) &= E[x(t)x(s)] = \frac{1}{2\pi} \int_0^{2\pi} \cos(t + \theta)\cos(s + \theta)d\theta \\
&= \frac{1}{4\pi} \int_0^{2\pi} [\cos(t + s + 2\theta) + \cos(t - s)]d\theta \\
&= \frac{1}{4\pi} \int_0^{2\pi} \cos(t - s)d\theta = \frac{1}{2} \cos(t - s)
\end{aligned} \tag{4.30}$$

which establishes that the process is wide-sense stationary. To show that the process is ergodic in the mean, we have to compute the variance of the estimate as

$$\frac{1}{2T} \int_{-2T}^{2T} \left(1 - \frac{|\tau|}{2T} \right) \cos(\tau)d\tau$$

and show that, in the limit, it goes to zero. Indeed, with a lot of algebra and Fourier analysis, we can show

$$\frac{1}{2T} \int_{-2T}^{2T} \left(1 - \frac{|\tau|}{2T} \right) \cos \tau d\tau = \frac{\sin T^2}{T^2} \to 0 \text{ as } T \to \infty.$$

Thus, the process is ergodic in the mean. To check whether the process is ergodic in autocorrelation, consider if it is ergodic in mean square. That is, define

$$\begin{aligned}
\langle R_x(0) \rangle_T &= \frac{1}{2T} \int_{-T}^{T} x^2(s)ds \\
&= \frac{1}{2T} \int_{-T}^{T} A^2 \cos(t + \theta)^2\, dt \\
&= A^2 \frac{1}{2T} \int_{-T}^{T} \cos(t + \theta)^2\, dt = A^2/2 + \frac{1}{2T}\epsilon
\end{aligned} \tag{4.31}$$

where ϵ is a small number, depending on how much of a period is integrated in the interval $[-T, T]$. Thus, note that the limit will always be a random variable, depending on the value of A. Thus, the process will not be ergodic in mean square.

Example 4.7

Recall the example of the stochastic process $x(t) = A \sin t + B \cos t$, where A, B are Gaussian, zero-mean, unit variance, independent random variables. The autocorrelation was given by $E[x(t)x(s)] = \cos(t - s)$. By the discussion in the previos example, this process is clearly ergodic in the mean. In the mean-square, direct computation establishes that $\langle R_x(0) \rangle_T \approx A^2 + B^2$, which has a variance which will not converge to zero.

Sometimes it is useful to describe the ergodic properties of the distribution of the process. Define the random variable

$$I_x(t) = \begin{cases} 1 & \text{if } x(t) < x \\ 0 & \text{otherwise} \end{cases}$$

Then, we can define the sample distribution function

$$\langle P_x(x) \rangle_T = \frac{1}{2T} \int_{-T}^{T} I_x(t) dt$$

and under certain conditions, this random variable converges to the true distribution function of $x(t)$; that is,

$$\lim_{T \to \infty} \langle P_x(x) \rangle_T = P_x(x)$$

We have the following result:

Theorem 4.11

The wide-sense stationary random process $x(t)$ is ergodic in distribution if and only if

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-2T}^{2T} \left[1 - \frac{|\tau|}{2T} \right] K_{I_x}(\tau) d\tau = 0$$

where

$$\begin{aligned} K_{I_x}(\tau) &= E[I_x(t + \tau) I_x(t)] - E[I_x(t)]^2 \\ &= P_x(x, x; t, t + \tau) - P_x(x; t)^2 \end{aligned} \tag{4.32}$$

Typically, the above condition would be met if, as $\tau \to \infty$, the random values $x(t), x(t+\tau)$ were asymptotically independent.

Note that verifying ergodicity is often reduced to verifying a condition of the form

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-2T}^{2T} \left[1 - \frac{|\tau|}{2T} \right] K_a(\tau) \, d\tau = 0 \tag{4.33}$$

for some wide-sense stationary random process $a(t)$, defined in terms of the original random process $x(t)$. We now provide a sufficient condition so that the above limit is zero:

Theorem 4.12

Suppose that $K_a(\tau)$ has a limit as $\tau \to \infty$. Then, eq. (4.33) is true if and only if $\lim_{\tau \to \infty} K_a(\tau) = 0$.

The proof of the above result goes as follows. Clearly, if the limit is not zero, then the ergodic condition should not hold. So the only part that needs proof is to show that, if the limit is zero, then the condition of eq. (4.33) is satisfied. Assume that the limit is zero. Then, for any $\epsilon > 0$, there exists a T_ϵ such that $|K_a(\tau)| < \epsilon$ for $t > T_\epsilon$. Let $T > T_\epsilon$. Then,

$$\begin{aligned} \frac{1}{2T} \int_{-2T}^{2T} \left[1 - \frac{|\tau|}{2T} \right] K_a(\tau) \, d\tau &\leq \frac{1}{2T} \left[\int_{-2T_\epsilon}^{2T_\epsilon} \left[1 - \frac{|\tau|}{2T} \right] K_a(\tau) d\tau + 4T\epsilon \right] \\ &= \frac{1}{2T}[M(\epsilon) + 4T\epsilon] \end{aligned} \tag{4.34}$$

Thus,

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-2T}^{2T} \left[1 - \frac{|\tau|}{2T} \right] K_a(\tau) \, d\tau \leq \lim_{T \to \infty} \frac{1}{2T}[M(\epsilon) + 4T\epsilon] = 2\epsilon$$

which establishes that the limit must be zero, since it is less than any arbitrary positive 2ϵ.

As a final concept in ergodicity, for stationary processes in the strict sense we can define the concept of ergodicity in terms of random variables defined on the samples of the process $x(t)$. Let X denote the space of all random variables which can be defined based on samples of $x(t)$; that is, $y \in X \Rightarrow y = f(\{x(t) : t \in (-\infty, \infty)\})$ for some function f. Define the shift operation $T_s y = f(\{x(t+s) : t \in (-\infty, \infty)\})$. Then, a process is said to be completely ergodic if every random variable in X with the property that $T_s y = y$ for all shifts s is almost surely a constant.

How does the above definition relate to the previous concepts in ergodicity? Note that the random variables

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} f(x(t)) dt$$

are in X, and are also invariant with respect to shifts T_s! Thus, all of the statistics which we describe above can be computed as time averages of a single sample path for a completely ergodic process. In essence, for any function f such that $E[|f(x(0))|] < \infty$, we have

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} f(x(t)) dt = E[f(x(t))]$$

In general, it is very difficult to obtain conditions for complete ergodicity. However, in the special case of Gaussian random variables, a simple sufficient condition is possible, since the Gaussian distributions are specified completely by the mean and autocorrelation function:

Theorem 4.13
A Gaussian process $x(t)$ is completely ergodic (also referred to as ergodic) if

$$\int_{-\infty}^{\infty} |K_x(\tau)| d\tau < \infty$$

Chapter 5

Linear Systems and Stochastic Processes

5.1 Introduction

In this section, we discuss the analysis of linear systems with random processes as inputs. Although most of the analysis is focused on continuous-time linear systems, the notes include some material on the analysis of discrete-time linear systems driven by random sequences as inputs. To begin with, we review some concepts from linear system theory for deterministic inputs. Then, we extend these concepts to systems with stochastic processes as inputs.

5.2 Review of Continuous-time Linear Systems

A general linear system with input $u(t)$ and output $y(t)$ has the form

$$y(t) = \int_{-\infty}^{\infty} h(t,s)u(s)ds \tag{5.1}$$

where $h(t,s)$ is referred to as the *impulse response* or *weighting function* of the system. That is, if $u(t) = \delta(t - t_0)$ where δ is the unit impulse, then $y(t) = h(t,t_0)$. The system is said to be *causal* if

$$h(t,s) = 0 \text{ for } s > t \tag{5.2}$$

or equivalently

$$y(t) = \int_{-\infty}^{t} h(t,s)u(s)ds \tag{5.3}$$

The system is said to be *time-invariant* if $h(t,s) = h(t-s,0) \equiv h(t-s)$, using the short-hand notation similar to that of autocorrelation for wide-sense stationary processes. If the system is time-invariant, then $y(t)$ is the convolution of $h(t)$ and $u(t)$. That is,

$$y(t) = \int_{-\infty}^{\infty} h(t-s)u(s)ds = \int_{-\infty}^{\infty} h(s)u(t-s)ds \tag{5.4}$$

An linear, time-invariant system is causal if and only if $h(t) = 0$ for $t < 0$.

The analysis of linear, time-invariant systems is often conducted in the frequency domain, using Laplace transforms. Denote by $X(s) \equiv \mathcal{L}[x(t)]$ the two-sided Laplace transform of the time signal $x(t)$, defined as

$$\mathcal{L}[x(t)] = \int_{-\infty}^{\infty} x(t)e^{-st}dt \equiv X(s) \tag{5.5}$$

Note that this is different from the standard one-sided Laplace transform which may be used for causal systems with initial conditions. Then, for a linear, time-invariant system, we have

$$Y(s) = H(s)U(s) \tag{5.6}$$

where $Y(s) = \mathcal{L}[y(t)], H(s) = \mathcal{L}[h(t)], U(s) = \mathcal{L}[u(t)]$.

The Fourier transform of $x(t)$ is simply $X(j\omega)$, so that, for an linear, time-invariant system,

$$Y(j\omega) = H(j\omega)U(j\omega) \tag{5.7}$$

The inverse Fourier transform is given by

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(j\omega)e^{j\omega t} d\omega \tag{5.8}$$

There are several important properties of Fourier and two-sided Laplace transforms which are summarized below:

1. Assume that $x(t)$ is real-valued. Then, $X(-j\omega) = X^*(j\omega)$, where x^* denotes the complex conjugate of x.

2. If $x(t)$ is an even function (e.g. $x(-t) = x(t)$), then $X(-s) = X(s)$.

3. If $x(t)$ is real-valued and even, so is $X(j\omega)$.

4. If $x(t) = e^{j\omega_0 t}$, then $X(j\omega) = 2\pi\delta(\omega - \omega_0)$.

5. If $x(t) = \cos(\omega_0 t)$, then $X(j\omega) = \pi(\delta(\omega - \omega_0) + \delta(\omega + \omega_0))$.

6. The Laplace transform of $\frac{d}{dt}x(t)$ is $sX(s)$.

See Appendix A for a summary of the definition and properties of continuous-time Fourier transforms.

Example 5.1
Using the above results, we see that, if the input $u(t) = A\cos(\omega_0 t)$, then the Fourier transform of the output, $Y(j\omega)$, is given by

$$Y(j\omega) = A\pi H(j\omega)(\delta(\omega - \omega_0) + \delta(\omega + \omega_0))$$

Transforming back to the time domain, we get

$$y(t) = \frac{A}{2}(H(j\omega_0)e^{j\omega_0 t} + H(-j\omega_0)e^{-j\omega_0 t})$$

Letting $H(j\omega) = |H(j\omega)|e^{j\theta(\omega)}$, we get

$$y(t) = A|H(j\omega_0)|\cos(j\omega_0 t + \theta(\omega_0))$$

Example 5.2
Consider an ideal low-pass filter, so that the transfer function

$$H(j\omega) = \begin{cases} 1 & \text{if } |\omega| \leq W, \\ 0 & \text{otherwise} \end{cases}$$

In the time domain, the impulse response of such a system is given by

$$h(t) = \frac{\sin(Wt)}{\pi t}$$

In the analysis of continuous-time linear systems, it is often useful to define two standard input signals: the unit step $u_{-1}(t)$ and its generalized derivative the unit delta function $\delta(t)$. The unit step function is defined as

$$u_{-1}(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The unit delta function has the formal property that $\frac{d}{dt}u_{-1}(t) = \delta(t)$. Furthermore, we have, for any continuous function $g(t)$,

$$\int_a^b g(t)\delta(t)dt = \begin{cases} 0 & \text{if } 0 \notin (a, b] \\ g(0) & \text{otherwise} \end{cases}$$

Example 5.3

Consider the signal $x_1(t) = e^{at}u_{-1}(t)$. Its Laplace transform is given by

$$X_1(s) = \frac{1}{s - a}$$

$X_1(s)$ is said to have a *pole* at $s = a$; this means that the denominator is 0 at that value, so that the magnitude of $X(s)$ is unbounded.

Example 5.4

Consider the function $x_2(t) = te^{at}u_{-1}(t)$. Its Laplace transform is given by

$$X_2(s) = \frac{1}{(s - a)^2}$$

$X_2(s)$ is said to have a pole of order 2. More generally, if $x_n(t) = \frac{t^n}{n!}e^{at}u_{-1}(t)$, then $X_n(s) = \frac{1}{(s-a)^{n+1}}$ has a pole of order $n + 1$.

The above examples provide the basis for inverting Laplace transforms which can be written as ratios of polynomials (also known as rational transforms); these transforms have the form

$$X(s) = \frac{b_{n-1}s^{n-1} + b_{n-2}s^{n-2} + \ldots + b_1 s + b_0}{s^n + a_{n-1}s^{n-1} + a_{n-2}s^{n-2} + \ldots + a_1 s + a_0} \equiv \frac{n(s)}{d(s)} \tag{5.10}$$

The denominator polynomial $d(s)$ can be factored as

$$d(s) = (s - \lambda_1)^{k_1}(s - \lambda_2)^{k_2} \cdots (s - \lambda_m)^{k_m}$$

for some distinct roots $\lambda_1, \ldots, \lambda_m$. Then, $X(s)$ can be written as a sum of simple factors using a *partial-fraction expansion*, as

$$\begin{aligned} X(s) &= \frac{A_{11}}{(s - \lambda_1)} + \frac{A_{12}}{(s - \lambda_1)^2} + \ldots + \frac{A_{1k_1}}{(s - \lambda_1)^{k_1}} + \frac{A_{21}}{(s - \lambda_2)} + \frac{A_{22}}{(s - \lambda_2)^2} \\ &+ \ldots + \frac{A_{2k_2}}{(s - \lambda_2)^{k_2}} + \ldots + \frac{A_{mk_m}}{(s - \lambda_m)^{k_m}} \end{aligned} \tag{5.11}$$

See Appendix B for more detail on partial-fraction expansions. The constant coefficients A_{ij} can be obtained by comparing the two equations (5.10) and (5.11) and matching coefficients of equal powers of s in the numerator. There is an alternative closed-form expression, given by

$$A_{ij} = \frac{1}{(k_i - j)!} \left\{ \frac{d^{(k_i - j)}}{ds^{(k_i - j)}} \left[(s - \lambda_i)^{k_i} X(s) \right] \right\}_{s = \lambda_i} \tag{5.12}$$

Once the partial-fraction expansion is known, the time signal $x(t)$ is easily determined from the previous examples, as

$$\begin{aligned} x(t) &= (A_{11}e^{\lambda_1 t} + A_{12}te^{\lambda_1 t} + \ldots + A_{1k_1}\frac{t^{k_1}}{k_1!}e^{\lambda_1 t} + A_{21}e_2^{\lambda}t + A_{22}te^{\lambda_2 t} + \ldots \\ &+ A_{2k_2}\frac{t^{k_2}}{k_2!}e^{\lambda_2 t} + \ldots + A_{mk_m}\frac{t^{k_m}}{k_m!}e^{\lambda_m t})u_{-1}(t) \end{aligned} \tag{5.13}$$

Example 5.5

One of the applications of rational transforms is in the solution of linear differential equations with constant coefficients. Consider a causal, linear, time-invariant system with input $u(t)$, and output $y(t)$ defined as the solution of the linear differential equation with constant coefficients

$$\frac{d^n}{dt^n}y(t) + a_{n-1}\frac{d^{(n-1)}}{dt^{(n-1)}}y(t) + \ldots + a_0 y(t) = b_m\frac{d^{(m)}}{dt^{(m)}}u(t) + b_{m-1}\frac{d^{(m-1)}}{dt^{(m-1)}}u(t) + \ldots + b_0 u(t) \qquad (5.14)$$

where $m < n$. Then, the transfer function for this system is given by

$$H(s) = \frac{Y(s)}{U(s)} = \frac{b_m s^m + b_{m-1}s^{m-1} + \ldots + b_1 s + b_0}{s^n + a_{n-1}s^{n-1} + a_{n-2}s^{n-2} + \ldots + a_1 s + a_0}$$

so that $h(t)$ can now be obtained using a partial fraction expansion. Note that $h(t)$ will be the sum of terms $e^{a_i t}u_{-1}(t), te^{a_i t}u_{-1}(t)$, where a_i is a pole of $H(s)$. The poles of $H(s)$ will be the roots of the denominator polynomial $d(s)$.

An important concept in the analysis of linear systems is the concept of stability. In particular, a linear, time-invariant system is called *bounded-input, bounded-output stable* if, whenever $|u(t)| \leq K < \infty$ for all t, then there exists a finite value M such that $|y(t)| \leq M$ for all t. Since $y(t)$ can be written in terms of $u(t)$ and the impulse response function $h(t)$, a necessary and sufficient condition for this stability is

$$\int_{-\infty}^{\infty} |h(t)|dt < \infty$$

Note that, for a causal system such as that defined in (5.14), this will be true if and only if all of the poles have negative real parts, so that they lie inside the complex left-half plane.

A system is said to be anti-causal if $h(t) = 0$ for $t > 0$. Anti-causal systems can also have rational transfer functions $H(s)$; however, for anti-causal systems, stability corresponds to all poles lying in the right-half plane! For instance, the transfer function $h(t) = e^{at}u_{-1}(-t)$ is anti-causal, and has Laplace transform $H(s) = -\frac{1}{s-a}$. Plotting $h(t)$ for a positive and negative indicates that the system is stable only if the real part of a is positive.

5.3   Review of Discrete-time Linear Systems

As in the continuous-time case, discrete-time linear systems with input $u(t)$ and output $y(t)$ can be written as

$$y(t) = \sum_{-\infty}^{\infty} h(t,s)u(s)$$

where $h(t,s)$ is the impulse response. In discrete time, the delta function $\delta(s)$ is defined as

$$\delta(t) = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{otherwise} \end{cases} \qquad (5.15)$$

The system is causal if $h(t,s) = 0$ whenever $s > t$, and anticausal if $h(t,s) = 0$ whenever $t > s$. The system is time-invariant if $h(t,s) = h(t-s,0) \equiv h(t-s)$. For linear, time-invariant systems, $y(t)$ is the convolution of $h(t)$ and $u(t)$, as

$$y(t) = \sum_{-\infty}^{\infty} h(t-s)u(s) = \sum_{-\infty}^{\infty} h(s)u(t-s) \qquad (5.16)$$

As in the continuous-time case, it is easiest to analyze linear, time-invariant systems in the transform domain. For discrete-time systems, we define the bilateral z-transform of $x(t)$ as

$$X(z) = \sum_{-\infty}^{\infty} x(t)z^{-t} \qquad (5.17)$$

For a linear, time-invariant system, we have

$$Y(z) = H(z)U(z) \qquad (5.18)$$

The Fourier transform of $x(t)$ is defined as $X(e^{j\omega})$, so that

$$X(e^{j\omega}) = \sum_{-\infty}^{\infty} x(t)e^{-j\omega t}$$

Note that $X(e^{j\omega})$ is periodic in ω, with period 2π. The inverse Fourier transform is given by

$$x(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega t} d\omega \tag{5.19}$$

Useful properties of z-transforms and Fourier transforms are:

1. If $x(t)$ is real, then $X(e^{-j\omega}) = X^*(e^{j\omega})$, and $X(z^{-1}) = X^*(z)$.

2. If $x(t)$ is even, then $X(z^{-1}) = X(z)$.

3. If $x(t)$ is real and even, so is $X(e^{j\omega})$ as a function of ω.

4. The transform of $x(t+1)$ is $zX(z)$.

5. If $x(t) = e^{j\omega_0 t}$, then $X(e^{j\omega}) = 2\pi\delta(\omega - \omega_0)$.

6. If $x(t) = A\cos(\omega_0 t)$, then, letting $H(e^{j\omega}) = |H(e^{j\omega})|e^{j\Theta(\omega)}$,

$$y(t) = A|H(e^{j\omega_0})|\cos(\omega_0 t + \Theta(\omega_0))$$

See Appendix A for a summary of the definition and properties of discrete-time Fourier transforms.

Consider the signal $x_1(t) = a^t u_{-1}(t)$; its z-transform is

$$X_1(z) = \frac{1}{1 - az^{-1}} = \frac{z}{z - a}$$

Note the presence of a pole at a. As in the continuous-time case, let $x_2(t) = ta^t u_{-1}(t)$. Then,

$$X_2(z) = \frac{az^{-1}}{(1 - az^{-1})^2} = \frac{az}{(z - a)^2}$$

The way most of these identities are derived are by noting that

$$\frac{d}{dz}X(z) = \sum_{-\infty}^{\infty} -tx(t)z^{-t-1}$$

Thus,

$$X_2(z) = \sum_{-\infty}^{\infty} tx_1(t)z^{-t} = -z\frac{d}{dz}X_1(z) = \frac{az}{(z - a)^2}$$

If $x_n(t) = t^n a^t u_{-1}(t)$, then $X_n(z) = -z\frac{d}{dz}X_{n-1}(z)$.

Using the above properties, it is useful to know the relationship between some standard z-transforms and their time functions. We summarize these in Table 1

The above discussion allows us to invert funtions with rational z-transforms using partial fraction expansions in a manner identical to the continuous-time case. As in the continuous-time case, we can also define the define the concept of stability for discrete-time linear, time-invariant systems. Such systems will be stable if and only if

$$\sum_{-\infty}^{\infty} |h(t)| < \infty$$

A necessary and sufficient condition for a causal system described by a rational transfer function to be stable is that all of the poles of the transfer function have magnitude less than 1, so that they are strictly inside the unit circle. For anti-causal systems, the stability condition is reversed, so that the poles must have magnitude strictly greater than 1.

Function $x(t)$	z-Transform $X(z)$
$\delta(t)$	1
$a^{t-1}u_{-1}(t-1)$	$\frac{1}{z-a}$
$(t-1)a^{t-2}u_{-1}(t-1)$	$\frac{1}{(z-a)^2}$
$tx(t)$	$-z\frac{d}{dz}X(z)$
$x(t+1)$	$zX(z)$
$x(t-1)$	$\frac{X(z)}{z}$

Table 5.1: Common Functions and their z-transforms.

5.4 Extensions to Multivariable Systems

All of the ideas in the previous subsections are easily extended to address linear systems with vector-valued outputs $\underline{y}(t)$ and vector-valued inputs $\underline{u}(t)$. The general form of such a multi-input, multi-output (MIMO) linear system is

$$\underline{y}(t) = \int_{-\infty}^{\infty} H(t,s)\underline{u}(s)ds \tag{5.20}$$

where $H(t,s)$ is the impulse response matrix. In the notation, we use capitals to denote matrices, and underlining to denote column vectors. Causality can be defined again in terms of $H(t,s) = 0$ if $s > t$. Time invariance corresponds to $H(t,s) = H(t-s,0) \equiv H(t-s)$. For linear, time-invariant systems, we can represent the system as a convolution, with

$$\underline{y}(t) = \int_{-\infty}^{\infty} H(t-s)\underline{u}(s)ds = \int_{-\infty}^{\infty} H(s)\underline{u}(t-s)ds \tag{5.21}$$

In order to avoid some confusion in notation, we denote the Laplace transform of the matrix $H(t)$ as the matrix $\mathrm{H}(s)$. In the transform domain, this implies

$$\underline{Y}(s) = \mathrm{H}(s)\underline{U}(s) \tag{5.22}$$

Similar extensions exist for discrete-time systems, where

$$\underline{y}(t) = \sum_{-\infty}^{\infty} H(t,s)\underline{u}(s)$$

5.5 Second-order Statistics for Vector-Valued Wide-Sense Stationary Processes

Before proceeding with the theory of linear systems driven by random processes, it will be useful to review a few definitions. To be consistent throughout the notes, we use the convention that the argument of an autocorrelation of a wide-sense stationary process is added to the time argument of the second variable. However, you must be aware that this notation is not standard, and that the answer may depend on the precise convention used above. Thus, for complex-valued wide-sense stationary processes, we have

$$R_x(t-s) = E[\underline{x}(s)\underline{x}(t)^H] = E[\underline{x}(t)\underline{x}(s)^H]^H = R_x(s-t)^H$$

Similarly, the cross-correlation between two jointly wide-sense stationary processes $\underline{x}(t), \underline{y}(t)$ is

$$R_{xy}(t-s) = E[\underline{x}(s)\underline{y}(t)^H] = E[\underline{y}(t)\underline{x}(s)^H]^H = R_{yx}(s-t)^H \tag{5.23}$$

For vector wide-sense stationary processes, we can define the power spectral density $S_x(\omega)$ and the cross-power spectral density $S_{xy}(\omega)$ as

$$S_x(\omega) = \int_{-\infty}^{\infty} R_x(\tau)e^{-j\omega\tau}d\tau; \ S_{xy}(\omega) = \int_{-\infty}^{\infty} R_{xy}(\tau)e^{-j\omega\tau}d\tau$$

Thus, we can derive the following relationship for the special case of real-valued processes:

$$S_x(-\omega) = \int_{-\infty}^{\infty} R_x(\tau)e^{j\omega\tau}d\tau = \left(\int_{-\infty}^{\infty} e^{-j\omega(-\tau)}R_x(-\tau)d\tau\right)^T = S_x(\omega)^T \tag{5.24}$$

We also know that, for any vector $\underline{a}$, the scalar process $\underline{a}^H\underline{x}(t)$ must have a nonnegative power spectral density. This means

$$\underline{a}^H S_x(\omega)\underline{a} \geq 0 \ \text{ for any vector } \underline{a}$$

which implies $S_x(\omega)$ is a positive semidefinite matrix for any ω.

Consider now modulation of a wide-sense stationary process $x(t)$ by a cosine with uniformly distributed phase. In particular, define the process $\underline{y}(t) = 2\underline{x}(t)\cos(\omega_0 t + \theta)$, where θ is uniformly distributed in $[0, 2\pi]$, independent of $x(t)$ for all t. Then, the autocorrelation of this process is given by:

$$\begin{aligned}
R_y(t, s) &= 4E\left[\underline{x}(t)\underline{x}(s)^H \cos(\omega_0 t + \theta)\cos(\omega_0 s + \theta)\right] \\
&= 4E\left[\underline{x}(t)\underline{x}(s)^H\right] E\left[\cos(\omega_0 t + \theta)\cos(\omega_0 s + \theta)\right] \\
&= 4R_x(t-s)\left(\frac{1}{2}\cos(\omega_0(t-s)) + E\left[\frac{1}{2}\cos(\omega_0(t+s) + 2\theta)\right]\right) \\
&= 2R_x(t-s)\cos(\omega_0(t-s)) \tag{5.25}
\end{aligned}$$

and the mean is

$$\underline{m}_y(t) = E[\underline{x}(t)]E[\cos(\omega_0 t + \theta)] = 0 \tag{5.26}$$

which shows that $\underline{y}(t)$ is also wide-sense stationary. The power spectral density of $\underline{y}(t)$ can be computed as

$$S_y(\omega) = \int_{-\infty}^{\infty} R_y(\tau)e^{-j\omega\tau}d\tau = \int_{-\infty}^{\infty} R_x(\tau)\left(e^{-j\omega_0\tau} + e^{j\omega_0\tau}\right)e^{-j\omega\tau}d\tau = S_x(\omega - \omega_0) + S_x(\omega + \omega_0) \tag{5.27}$$

We have already shown that $S_y(\omega)$ must be positive semidefinite for any ω. In particular, for $\omega = 0$, this states that

$$S_x(-\omega_0) + S_x(\omega_0) \geq 0$$

for any arbitrary ω_0. If $\underline{x}(t)$ is real-valued, then this can be further simplified to obtain

$$S_x(\omega) + S_x(\omega)^T \geq 0$$

a condition which is referred to as *positive real*.

As a final note, the covariance of the random vector $\underline{x}(t)$ from a wide-sense stationary process can be obtained readily from the power spectral density of the process as follows:

$$E[\underline{x}(t)\underline{x}(t)^H] = R_x(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(\omega)e^{j\omega 0}d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(\omega)d\omega \tag{5.28}$$

5.6   Continuous-time Linear Systems with Random Inputs

For the purposes of this section, it makes little difference whether the processes are scalar-valued or vector-valued, real-valued or complex-valued. In order to introduce the most general form of the results, we will make no assumptions, and describe the results in the general case. Thus, assume that $\underline{u}(t)$ is a complex-valued, vector-valued random process with mean $\underline{m}_u(t)$ and autocorrelation $R_u(t, s)$, defined as

$$\underline{m}_u(t) = E[\underline{u}(t)]; \ R_u(t, s) = E[\underline{u}(t)\underline{u}(s)^H]$$

where $\underline{a}^H = (\underline{a}^T)^*$ is the transpose of the complex conjugate of $\underline{a}$. Consider the linear system with input $\underline{u}(t)$, described by

$$\underline{y}(t) = \int_{-\infty}^{\infty} H(t, s)\underline{u}(s)ds \tag{5.29}$$

The issue is to relate the statistics of the output process $\underline{y}(t)$ to those of the input process $\underline{u}(t)$. Fortunately, we have already developed a theory of mean-square integration, which allows us to determine the properties of (5.29) for each t. In particular, we know

$$\underline{m}_y(t) = E[\int_{-\infty}^{\infty} H(t,s)\underline{u}(s)ds] = \int_{-\infty}^{\infty} H(t,s)E[\underline{u}(s)]ds = \int_{-\infty}^{\infty} H(t,s)\underline{m}_u(s)ds \qquad (5.30)$$

Furthermore, if $\underline{z}(t)$ is any other process, we know

$$
\begin{aligned}
R_{yz}(t,s) &\equiv E[\underline{y}(t)\underline{z}(s)^H] \\
&= E[\int_{-\infty}^{\infty} H(t,\sigma)\underline{u}(\sigma)\underline{z}(s)^H d\sigma] \\
&= \int_{-\infty}^{\infty} H(t,\sigma)E[\underline{u}(\sigma)\underline{z}(s)^H]d\sigma \\
&= \int_{-\infty}^{\infty} H(t,\sigma)R_{uz}(\sigma,s)d\sigma \qquad (5.31)
\end{aligned}
$$

In particular, this leads to

$$R_{yu}(t,s) = \int_{-\infty}^{\infty} H(t,\sigma)R_{uu}(\sigma,s)d\sigma \qquad (5.32)$$

and

$$
\begin{aligned}
R_{yy}(t,s) &= \int_{-\infty}^{\infty} H(t,\sigma)R_{uy}(\sigma,s)d\sigma \\
&= \int_{-\infty}^{\infty} H(t,\sigma)R_{yu}(s,\sigma)^H d\sigma \\
&= \int_{-\infty}^{\infty} H(t,\sigma)\left[\int_{-\infty}^{\infty} H(s,\tau)R_{uu}(\tau,\sigma)d\tau\right]^H d\sigma \\
&= \int_{-\infty}^{\infty} H(t,\sigma)\left[\int_{-\infty}^{\infty} R_{uu}(\tau,\sigma)^H H(s,\tau)^H d\tau\right] d\sigma \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(t,\sigma)R_{uu}(\tau,\sigma)^H H(s,\tau)^H d\tau d\sigma \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(t,\sigma)R_{uu}(\sigma,\tau)H(s,\tau)^H d\tau d\sigma \qquad (5.33)
\end{aligned}
$$

There are similar expressions for the autocovariances and cross-covariances, based on the autocovariance of the input $\underline{u}(t)$. For instance,

$$K_{yy}(t,s) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(t,\sigma)K_{uu}(\sigma,\tau)H(s,\tau)^H d\tau d\sigma \qquad (5.34)$$

and

$$K_{yu}(t,s) = \int_{-\infty}^{\infty} H(t,\sigma)K_{uu}(\sigma,s)d\sigma$$

One of the important properties of linear systems is the property of superposition. In particular, if the input $\underline{u}(t) = \underline{x}(t) + \underline{z}(t)$, where $\underline{x}(t), \underline{z}(t)$ were random processes, we would like to make some statements about the statistics of the output $\underline{y}(t)$. In particular, let

$$\underline{y}_1(t) = \int_{-\infty}^{\infty} H(t,s)\underline{x}(s)ds$$

$$\underline{y}_2(t) = \int_{-\infty}^{\infty} H(t,s)\underline{z}(s)ds$$

The question is how are the statistics of $\underline{y}(t)$ related to the statistics of $\underline{y}_1(t), \underline{y}_2(t)$?

To answer this question, let's analyze the mean of $\underline{y}(t)$, which is defined as

$$m_y(t) = \int_{-\infty}^{\infty} H(t,s)\underline{m}_u(s)ds = \int_{-\infty}^{\infty} H(t,s)(\underline{m}_x(s) + \underline{m}_z(s))ds = \underline{m}_{y_1}(t) + \underline{m}_{y_2}(t) \tag{5.35}$$

Thus, the means of the processes satisfy the deterministic superposition law! That is, the mean of the output in response to two inputs is the sum of the means of the individual outputs for each input.

Will the autocorrelation statistic satisfy a similar relationship? Consider the relationship provided by (5.33):

$$
\begin{aligned}
R_{yy}(t,s) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(t,\sigma)R_{uu}(\sigma,\tau)H(s,\tau)^H d\tau d\sigma \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(t,\sigma)[R_{xx}(\sigma,\tau) + R_{xz}(\sigma,\tau) + R_{zx}(\sigma,\tau) + R_{zz}(\sigma,\tau)]H(s,\tau)^H d\tau d\sigma \\
&= R_{y_1y_1}(t,s) + R_{y_2y_2}(t,s) + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(t,\sigma)[R_{xz}(\sigma,\tau) + R_{zx}(\sigma,\tau)]H(s,\tau)^H d\tau d\sigma \tag{5.36}
\end{aligned}
$$

Thus, if $R_{xz}(\sigma,\tau) = 0$, (which means the processes $\underline{x}(t).\underline{z}(s)$ are orthogonal), then we have

$$R_{yy}(t,s) = R_{y_1y_1}(t,s) + R_{y_2y_2}(t,s) \tag{5.37}$$

In general, if the processes $\underline{x}(t).\underline{z}(s)$ are uncorrelated, then we have superposition of the autocovariances:

$$K_{yy}(t,s) = K_{y_1y_1}(t,s) + K_{y_2y_2}(t,s) \tag{5.38}$$

A useful method for analysis of stochastic systems driven by random processes is to decompose the random process into two processes: the mean process and a zero-mean process. That is, for an input random process $\underline{u}(t)$, define the decomposed process as $\underline{u}(t) = \underline{m}_u(t) + \underline{u}'(t)$, where $\underline{u}'(t)$ is zero-mean. Then, by superposition, the mean of the output $\underline{y}(t)$ is given as the sum of the means of the individual outputs:

$$m_y(t) = \int_{-\infty}^{\infty} H(t,s)(\underline{m}_u(s) + \underline{m}_{u'}(s))ds = \int_{-\infty}^{\infty} H(t,s)\underline{m}_u(s)ds \tag{5.39}$$

so that analysis of the mean of the output depends only on the first input. Note also that the processes $\underline{m}_u(t)$ and $\underline{u}'(t)$ are uncorrelated, so that the autocovariance is given by

$$K_{yy}(t,s) = K_{y_1y_1}(t,s) + K_{y_2y_2}(t,s) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(t,\sigma)K_{u'u'}(\sigma,\tau)H(s,\tau)^H d\tau d\sigma$$

since the process $\underline{m}_u(t)$ is deterministic, and therefore has zero autocovariance. Thus, the covariance of the output depends only on the process $\underline{u}'(t)$. It will be useful in the analysis of linear systems driven by stochastic processes, to remember that we can separate the analysis of the mean and the analysis of the autocorrelation.

Suppose now that our system is linear, time-invariant and that $\underline{u}(t)$ is wide-sense stationary. Assume in addition that the system is bounded-input, bounded-output stable. For vector input systems, this means that

$$\int_{-\infty}^{\infty} \sum_{i,j} |H_{ij}(t)|dt < \infty$$

Under this conditions, $\underline{u}(t)$ and $\underline{y}(t)$ are jointly wide-sense stationary, which can be verified by direct calculation:

$$m_y(t) = \int_{-\infty}^{\infty} H(s)\underline{m}_u(t-s)ds = \int_{-\infty}^{\infty} H(s)ds\ \underline{m}_u = \underline{m}_y \tag{5.40}$$

$$R_{yu}(t,s) = \int_{-\infty}^{\infty} H(t-\sigma)E[\underline{u}(\sigma)\underline{u}(s)^H]d\sigma = \int_{-\infty}^{\infty} H(t-\sigma)R_u(s-\sigma)d\sigma = \int_{-\infty}^{\infty} H(-\tau)R_u(s-t-\tau)d\tau \equiv R_{yu}(s-t) \tag{5.41}$$

A simpler notation for writing the above equation is using the convolution operator $*$, as

$$R_{yu}(t) = H(-t) * R_u(t)$$

Furthermore, if $\underline{u}(t)$ and $\underline{z}(t)$ are jointly wide-sense stationary, then so are $\underline{y}(t)$ and $\underline{z}(t)$, and

$$R_{yz}(t,s) = \int_{-\infty}^{\infty} H(-\tau)R_{uz}(s-t-\tau)d\tau \equiv R_{yz}(s-t) \tag{5.42}$$

or, in convolution terms,

$$R_{yz}(t) = H(-t) * R_{uz}(t)$$

Extending the above discussion to $\underline{z}(t) = \underline{y}(t)$, we get

$$\begin{aligned}
R_y(t,s) &= \int_{-\infty}^{\infty} H(-\tau)R_{uy}(s-t-\tau)d\tau \\
&= \int_{-\infty}^{\infty} H(-\tau)R_{yu}(-\tau+s-t)^H d\tau \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(-\tau)R_u(\tau+t-s-\sigma)^H H(-\sigma)^H d\sigma d\tau \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(-\tau)R_u(s-t-\tau-\sigma)H(\sigma)^H d\sigma d\tau = R_y(t-s) \tag{5.43}
\end{aligned}$$

It will be useful to recognize the convolution form of the above operation. Let

$$G(t) = H(-t) * R_u(t)$$

Then,

$$G(s-t-\sigma) = \int_{-\infty}^{\infty} H(-\tau)R_u(s-t-\sigma-\tau)d\tau$$

and we can rewrite (5.43) as

$$R_y(t) = \int_{-\infty}^{\infty} G(t-\sigma)H(\sigma)^H d\sigma = G(t) * H(t)^H$$

Note that similar equations can be obtained for the auctocovariance $K_y(t-s)$ and the cross-covariance $K_{yz}(t-s)$, simply by replacing all R with K.

Since both the inputs and outputs are wide-sense stationary, and the system is time-invariant, one can take Fourier transforms of (5.41), (5.42) and (5.43) to obtain:

$$S_{yu}(\omega) = H(-j\omega)S_u(\omega) \tag{5.44}$$

$$S_{yz}(\omega) = H(-j\omega)S_{uz}(\omega) \tag{5.45}$$

Note also that the Fourier transform of $H(t)^H$ is given by

$$\int_{-\infty}^{\infty} H(t)^H e^{-j\omega t}dt = (\int_{-\infty}^{\infty} H(t)e^{j\omega t}dt)^H = H(-j\omega)^H = H(j\omega)^T$$

Hence,

$$S_y(\omega) = H(-j\omega)S_u(\omega)H(j\omega)^T \tag{5.46}$$

Using a similar analysis technique, we obtain

$$S_{uy}(\omega) = S_u(\omega)H(j\omega)^T \tag{5.47}$$

Example 5.6

One way of interpreting the power spectral density of a process is to consider what happens to that density when it is filtered under an ideal band-pass filter. Consider a wide-sense stationary scalar process $u(t)$, which is used as an input into a linear, time-invariant system with transfer function described in the frequency domain as

$$H(j\omega) = \begin{cases} 1 & \text{if } \omega \in (\omega_1, \omega_2) \\ 0 & \text{otherwise} \end{cases}$$

Denote the output as $y(t)$. Using the above relationships, we have

$$S_y(\omega) = |H(j\omega)|^2 S_u(\omega)$$

Suppose we wanted to compute the second moment of the process $y(t)$; as we have seen before, this is equal to $R_y(0)$. Then, we can use the formula relating the autocorrelation to the power spectral density, as follows:

$$R_y(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_y(\omega) e^{j\omega t} d\omega$$

so that, in particular,

$$R_y(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_y(\omega) d\omega = \frac{1}{2\pi} \int_{w_1}^{w_2} S_u(\omega) d\omega$$

The result is that the average power in the output is given as an integral of the power spectral density of the input. This is consistent with the interpretation of the power spectral density as a density, since, if one integrates the density across a frequency band, one gets the average power in that frequency band.

Note that, for scalar-valued processes, many of the matrix relationships described above simplify. We summarize these below:

$$
\begin{aligned}
S_{yu}(\omega) &= \text{H}(-j\omega) S_u(\omega) = \text{H}(j\omega)^* S_u(\omega) \\
S_{uy}(\omega) &= S_u(\omega) \text{H}(j\omega) = \text{H}(j\omega) S_u(\omega) \\
S_y(\omega) &= \text{H}(j\omega) S_u(\omega) \text{H}(j\omega)^* = \text{H}(j\omega) \text{H}(j\omega)^* S_u(\omega) = |\text{H}(j\omega)|^2 S_u(\omega)
\end{aligned}
\tag{5.48}
$$

Example 5.7

Consider the causal linear, time-invariant system described by the differential equation

$$\frac{d}{dt} y(t) = -ay(t) + u(t)$$

where $a > 0$. The transfer function of this linear system is given by

$$H(s) = \frac{1}{s + a}$$

Assume that the input is standard white noise, with power spectral density $S_u(\omega) = 1$. Then, the power spectral density of the output is given by

$$S_y(\omega) = |H(j\omega)|^2 S_u(\omega) = \frac{1}{a + j\omega} \frac{1}{a - j\omega} = \frac{1}{a^2 + \omega^2}$$

Taking inverse Fourier transforms, the autocorrelation is given by

$$R_y(\tau) = \frac{1}{2a} e^{-a|\tau|}$$

Example 5.8

Consider the causal, linear, time-invariant system described by the differential equation

$$\frac{d^2}{dt^2} y(t) = -4\frac{d}{dt} y(t) - 4y(t) + \frac{d}{dt} u(t) + u(t)$$

The transfer function of this system is

$$H(s) = \frac{s + 1}{s^2 + 4s + 4} = \frac{s + 1}{(s + 2)^2}$$

Assume that the input $u(t)$ is the sum of a standard white noise and a second wide-sense stationary process $u'(t)$, uncorrelated with the white noise, with zero mean and autocovariance $K_{u'}(\tau) = e^{-|\tau|}$. The problem is to determine the

autocovariance of the output process $y(t)$. First, note that the output process will be zero-mean, because the mean of the two input processes is zero. Second, obtain the power spectral density of the input u as

$$S_u(\omega) = 1 + S_{u'}(\omega) = 1 + \frac{2}{1 + \omega^2}$$

because the white noise process and $u'(t)$ are uncorrelated. Third, obtain the power spectral density of the output as:

$$S_y(\omega) = H(j\omega)H(-j\omega)S_u(\omega) = \frac{(1 + \omega^2)}{(\omega^2 + 4)^2}(1 + \frac{2}{1 + \omega^2})$$

Fourth, compute the power spectral density due to the white noise input only as

$$S_{y_1}(\omega) = \frac{(1 + \omega^2)}{(\omega^2 + 4)^2} = \frac{1}{(\omega^2 + 4)} + \frac{-3}{(\omega^2 + 4)^2}$$

and compute the corresponding autocorrelation as

$$R_{y_1}(\tau) = \frac{1}{4}e^{-2|\tau|} + \frac{-3}{16}(\frac{1}{2} + |t|)e^{-2|\tau|}$$

Similarly, the power spectral density due to the $u'(t)$ input is

$$S_{y_2}(\omega) = \frac{2(1 + \omega^2)}{(1 + \omega^2)(\omega^2 + 4)^2} = \frac{2}{(\omega^2 + 4)^2}$$

and the autocorrelation is given by

$$R_{y_2}(\tau) = \frac{1}{8}(\frac{1}{2} + |t|)e^{-2|\tau|}$$

Combining these, we get

$$R_y(\tau) = \frac{1}{4}e^{-2|\tau|} - \frac{1}{16}(\frac{1}{2} + |t|)e^{-2|\tau|}$$

Chapter 6

Sampling of Stochastic Processes

6.1 The Sampling Theorem

One of the most powerful results in deterministic, continuous-time signals is the sampling theorem. Briefly stated, a deterministic signal $x(t)$ is band-limited if its Fourier transform $X(j\omega)$ is zero for $|\omega| > W$, for some frequency W measured in radians per second. For band-limited deterministic signals, we can sample them at a fast enough period and recover the entire signal exactly! In particular, let the sampling period T_s be small enough, so that

$$T_s < \frac{\pi}{W}$$

Then,

$$x(t) = 2T_s \frac{W}{2\pi} \sum_{n=-\infty}^{\infty} x(nT_s) \frac{\sin W(t - nT_s)}{W(t - nT_s)} \tag{6.1}$$

This means that obtaining the samples $x(nT_s), n = \ldots, -1, 0, 1, \ldots$ is enough to determine the complete function.

Is there a corresponding sampling theorem for stochastic processes? In particular, if we define the concept of a band-limited stochastic process as a wide-sense stationary process $x(t)$ with the property that $S_x(\omega) = 0$ if $|\omega| > W$, it seems that a similar sampling theorem should hold. The purpose of this section is to establish such a sampling theorem. In order to simplify notation, we assume that the stochastic process $x(t)$ is a scalar, real-valued, continuous-time process. Extensions to vector-valued or complex-valued processes are straightforward.

In particular, it seems that (6.1) should hold independent of whether $x(t)$ is deterministic or stochastic, as long as its fluctuations are limited to a maximum frequency. Thus, define the approximate process

$$x'_N(t) = 2T_s \frac{W}{2\pi} \sum_{n=-N}^{N} x(nT_s) \frac{\sin W(t - nT_s)}{W(t - nT_s)} \tag{6.2}$$

Note that this is a random process. Thus, as $N \to \infty$, the process $x'_N(t)$ should approach a limit, and the appropriate sense is in the mean-square sense. The sampling theorem for stochastic processes states the following:

Theorem 6.1
If $x(t)$ is a band-limited process with maximum frequency less than W (i.e. $S_x(\omega) \equiv 0$ for $\omega < W$), and the sampling time T_s satisfies $T_s < \pi/W$, then

$$\lim_{N \to \infty} x'_N(t) \stackrel{\mathsf{mss}}{=} x(t)$$

The proof is as follows. Note first that, since $S_x(\omega) = 0$ outside of $[-W, W]$, we can expand $S_x(\omega)$ in a Fourier series over the interval $[-W', W']$, for $W' > W$, as

$$S_x(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{j\omega \frac{2\pi n}{2W'}}$$

where the coefficients are obtained by

$$c_n = \frac{1}{2W'} \int_{-W'}^{W'} S_x(\omega) e^{-j\omega \frac{2\pi n}{2W'}} d\omega \tag{6.3}$$

The autocorrelation function $R_x(\tau)$ is obtained as the inverse Fourier transform of the above, so that

$$
\begin{aligned}
R_x(\tau) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} S_x(\omega) e^{j\omega\tau} d\omega = \frac{1}{2\pi} \int_{-W}^{W} S_x(\omega) e^{j\omega\tau} d\omega \\
&= \frac{1}{2\pi} \int_{-W}^{W} \sum_{n=-\infty}^{\infty} c_n e^{j\omega \frac{2\pi n}{2W'}} e^{j\omega\tau} d\omega \\
&= \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} c_n \int_{-W}^{W} e^{j\omega \frac{2\pi n}{2W'}} e^{j\omega\tau} d\omega \\
&= \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} c_n \frac{2\sin W(\frac{2\pi n}{2W'} + \tau)}{(\frac{2\pi n}{2W'} + \tau)} \\
&= \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} c_n \frac{2\sin W(nT_s + \tau)}{(nT_s + \tau)} \\
&= \frac{1}{\pi} \sum_{n=-\infty}^{\infty} c_n \frac{\sin W(nT_s + \tau)}{(nT_s + \tau)} \tag{6.4}
\end{aligned}
$$

where we have defined the sampling time $T_s = \frac{\pi}{W'}$. In addition, we can recognize the following formula for the Fourier coefficients c_n:

$$c_n = \frac{2\pi}{2W'} R_x(-\frac{\pi n}{W'}) = T_s R_x(-nT_s)$$

Combining the above equations, we get:

$$
\begin{aligned}
R_x(\tau) &= \frac{T_s}{\pi} \sum_{n=-\infty}^{\infty} R_x(-nT_s) \frac{\sin W(nT_s + \tau)}{(nT_s + \tau)} \\
&= \frac{WT_s}{\pi} \sum_{n=-\infty}^{\infty} R_x(-nT_s) \frac{\sin W(nT_s + \tau)}{W(nT_s + \tau)} \\
&= \frac{WT_s}{\pi} \sum_{n=-\infty}^{\infty} R_x(nT_s) \frac{\sin W(\tau - nT_s)}{W(\tau - nT_s)} \tag{6.5}
\end{aligned}
$$

Note that, expressed in terms of frequencies measured in cycles per second rather than radians per second, the above equation becomes:

$$
\begin{aligned}
R_x(\tau) &= \frac{2\pi F T_s}{\pi} \sum_{n=-\infty}^{\infty} R_x(-nT_s) \frac{\sin 2\pi F(nT_s + \tau)}{2\pi F(nT_s + \tau)} \\
&= 2F T_s \sum_{n=-\infty}^{\infty} R_x(-nT_s) \frac{\sin 2\pi F(nT_s + \tau)}{2\pi F(nT_s + \tau)} \tag{6.6}
\end{aligned}
$$

Another useful identity comes from noting that $S_x(\omega)e^{-j\omega a}$ is also a band-limited signal, and is the power spectral density corresponding to $R_x(\tau - a)$. The Fourier coefficients of $S_x(\omega)e^{-j\omega a}$ are given by

$$c'_n = \frac{1}{2W'} \int_{-W'}^{W'} S_x(\omega)e^{-j\omega a}e^{-j\omega \frac{2\pi n}{2W'}} d\omega = R_x(-a - nT_s)$$

so that we get

$$
\begin{aligned}
R_x(\tau - a) &= \frac{WT_s}{\pi} \sum_{n=-\infty}^{\infty} R_x(-a - nT_s)\frac{\sin W(nT_s + \tau)}{W(nT_s + \tau)} \\
&= \frac{WT_s}{\pi} \sum_{n=-\infty}^{\infty} R_x(nT_s - a)\frac{\sin W(\tau - nT_s)}{W(\tau - nT_s)} \qquad (6.7)
\end{aligned}
$$

The above identities can be used to show the mean-square convergence of the approximate process $x'_N(t)$ to $x(t)$ as $N \to \infty$, as follows:

$$
\begin{aligned}
E[(x(t) - x'_N(t))^2] &= E[(x(t) - x'_N(t))x(t)] - E[(x(t) - x'_N(t))x'_N(t)] \\
&= E[x(t)^2] - 2T_s\frac{W}{2\pi} \sum_{n=-N}^{N} E[x(t)x(nT_s)]\frac{\sin W(t - nT_s)}{W(t - nT_s)} - E[(x(t) - x'_N(t))x'_N(t)] \quad (6.8)
\end{aligned}
$$

Now, consider the first two terms. Replacing expectations by autocorrelations gives

$$E[(x(t) - x'_N(t))x(t)] = R_x(0) - 2T_s\frac{W}{2\pi} \sum_{n=-N}^{N} R_x(nT_s - t)\frac{\sin W(t - nT_s)}{W(t - nT_s)}$$

which vanishes as $N \to \infty$ because of (6.7) (letting $a = t = \tau$).

Consider the next term. Then,

$$E[(x(t) - x'_N(t))x'_N(t)] = 2T_s\frac{W}{2\pi} \sum_{n=-N}^{N} E[(x(t) - x'_N(t))x(nT_s)]\frac{\sin W(t - nT_s)}{W(t - nT_s)}$$

Each of the terms involves the element

$$
\begin{aligned}
E[(x(t) - x_N(t))x(nT_s)] &= E[x(t)x(nT_s)] - T_s\frac{W}{\pi} \sum_{m=-N}^{N} E[x(mT_s)x(nT_s)]\frac{\sin W(t - mT_s)}{W(t - mT_s)} \\
&= R_x(t - nT_s) - T_s\frac{W}{\pi} \sum_{m=-N}^{N} R_x((m - n)T_s)\frac{\sin W(t - mT_s)}{W(t - mT_s)} \qquad (6.9)
\end{aligned}
$$

which again goes to zero as $N \to \infty$. This establishes the sampling theorem for stochastic processes.

A major restriction of the sampling theorem is that the sampling period T_s must be small enough; in practice, the sampling rate (in samples per second) must be faster than $\frac{W}{\pi}$, where W is the largest frequency component (in radians per second) of the power spectrum of the signal. The frequency $\frac{W}{\pi}$ is also called the Nyquist frequency.

What happens when one samples at frequencies slower than the Nyquist frequency? What is the power spectral density of the discrete-time process denoted by $y(n) = x(nT_s)$? Consider the autocorrelation (in discrete time) of the process $y(n)$, as

$$R_y(n, m) = E[x(nT_s)x(mT_s)] = R_x(nT_s - mT_s) = R_x((n - m)T_s) = R_y(n - m)$$

so that the resulting sampled process is also wide-sense stationary. The power spectral density of the process is thus obtained by:

$$S_y(\omega) = \sum_{-\infty}^{\infty} R_y(n)e^{-nj\omega} = \int_{-\infty}^{\infty} \sum_{-\infty}^{\infty} R_y(n)\delta(t-n)e^{-j\omega t}dt = \int_{-\infty}^{\infty} \sum_{-\infty}^{\infty} R_x(tT_s)\delta(t-n)e^{-j\omega t}dt$$

where we have embedded the discrete-time function into a continuous-time function! Now, the above can be simplified as follows:

$$S_y(\omega) = \int_{-\infty}^{\infty} (\sum_{-\infty}^{\infty} \delta(t-n))(R_x(tT_s))e^{-j\omega t}dt$$

and recognize that this is the Fourier transform of the product of two functions: $(\sum_{-\infty}^{\infty} \delta(t-n))$ and $(R_x(tT_s))$ Note the following equalities:

$$\int_{-\infty}^{\infty} (\sum_{-\infty}^{\infty} \delta(t-n))e^{-j\omega t}dt = \sum_{-\infty}^{\infty} e^{-j\omega n} = 2\pi \sum_{-\infty}^{\infty} \delta(\omega - 2\pi n)$$

where the last identity follows because the sum is clearly periodic in ω with period 2π, and from the observation that the Fourier series for $\delta(t)$, when restricted to a finite interval $[-\pi, \pi]$ is

$$\delta(t) = \frac{1}{2\pi} \sum_{-\infty}^{\infty} e^{jnt}$$

Similarly,

$$\int_{-\infty}^{\infty} R_x(tT_s)e^{-j\omega t}dt = \frac{1}{T_s} \int_{-\infty}^{\infty} R_x(tT_s)e^{-j\frac{\omega}{T_s}T_s t}d(tT_s) = \frac{1}{T_s}S_x(\frac{\omega}{T_s})$$

Finally, since $S_y(\omega)$ is the Fourier transform of the product of two functions, it is proportional to the convolution of the Fourier transform of the individual functions; that is,

$$\int_{-\infty}^{\infty} f(t)g(t)e^{-j\omega t}dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j(\omega - \omega_1))G(j\omega_1)d\omega_1 = \frac{1}{2\pi}F(j\omega) * G(j\omega) \qquad (6.10)$$

where $F(j\omega)$ is the Fourier transform of $f(t)$ and $G(j\omega)$ is the Fourier transform of $g(t)$. Thus, one obtains

$$\begin{aligned} S_y(\omega) &= \sum_{-\infty}^{\infty}[\delta(\omega - 2\pi n) * \frac{1}{T_s}S_x(\frac{\omega}{T_s})] \\ &= \frac{1}{T_s} \sum_{-\infty}^{\infty} S_x(\frac{\omega - 2\pi n}{T_s}) \end{aligned} \qquad (6.11)$$

Recall that, for discrete sequences such as $y(n)$, the power spectral density function is only defined in the interval $\omega \in [-\pi, \pi]$. Thus, for perfect reproduction, we want

$$S_y(\omega) = \frac{1}{T_s}S_x(\frac{\omega}{T_s}) \;\; \text{if } \omega \in [-\pi, \pi]$$

This will be true provided that the interference from other factors in (6.11) vanishes in that interval; that is,

$$S_x(\frac{\omega - 2\pi n}{T_s}) = 0 \;\; \text{if } \omega \in [-\pi, \pi] \text{ and } n \neq 0$$

An equivalent condition is $\frac{\omega - 2\pi n}{T_s} \notin [-W, W]$ for $\omega \in [-\pi, \pi]$. Equivalently, this requires $\frac{\pi}{T_s} > W$, which was the same condition for the sampling theorem to hold exactly.

What happens if T_s is too large? Then, the spectrum of the discrete-time process $y(n)$ includes some contributions beyond the original one for $n = 0$. This phenomenom is called "aliasing". The relative power

in the contribution of the aliased signal is obtained approximately as follows. The power in the original signal is given as

$$
\begin{aligned}
P_0 &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{T_s} S_x(\frac{\omega}{T_s}) d\omega \\
&= \frac{1}{2\pi T_s} \int_{-\pi}^{\pi} S_x(\frac{\omega}{T_s}) d\omega
\end{aligned}
\tag{6.12}
$$

The power contributed from the aliasing is defined as

$$
P_a = \frac{1}{2\pi T_s} \int_{-\pi}^{\pi} \sum_{n=1}^{\infty} S_x(\frac{\omega - 2\pi n}{T_s}) d\omega + \frac{1}{T_s} \int_{-\pi}^{\pi} \sum_{n=-\infty}^{-1} S_x(\frac{\omega - 2\pi n}{T_s}) d\omega
\tag{6.13}
$$

For most applications of interest, the primary contribution to P_a will come from the $n = 1$ and $n = -1$ terms, so that

$$
P_a \approx \frac{1}{2\pi T_s} \int_{-\pi}^{\pi} (S_x(\frac{\omega - 2\pi}{T_s}) + S_x(\frac{\omega + 2\pi}{T_s})) d\omega
$$

Note that, in the analysis of aliasing, we worked with the discrete time signal $y(n)$. A similar analysis is possible by defining a continuous-time, sampled-data signal, as follows. Define the sampling signal $s(t) = \sum_{-\infty}^{\infty} \delta(t - nT_s - D)$, where D is a uniform random variable, independent of $x(t)$, distributed on the interval $[0, T_s]$. Then, the sampled signal is defined as $z(t) = s(t)x(t)$. The role of D is to guarantee that the sampled signal is wide-sense stationary in continuous-time; if D were identically zero, then the sampled signal is guaranteed to be zero except at multiples of T_s, and thus would not be wide-sense stationary, but only wide-sense periodic. Indeed, the autocorrelation of $z(t)$ is given by

$$
R_z(t, s) = E[s(t)s(s)x(t)x(s)] = R_x(t - s)E[s(t)s(s)]
$$

and we can compute

$$
\begin{aligned}
E[s(t)s(s)] &= \frac{1}{T_s} \int_0^{T_s} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \delta(t - nT_s - D)\delta(s - mT_s - D) dD \\
&= \frac{1}{T_s} \sum_{n=-\infty}^{\infty} \int_0^{T_s} \delta(t - nT_s - D) \sum_{m=-\infty}^{\infty} \delta(s - mT_s - D) dD \\
&= \frac{1}{T_s} \sum_{n=-\infty}^{\infty} \delta(t - s - nT_s)
\end{aligned}
\tag{6.14}
$$

since the product of the two delta functions is zero unless $t - s$ is an integer multiple of nT_s, in which case the integral of the product of the two delta functions is a single delta function! Thus,

$$
R_z(t, s) = R_z(t - s) = R_x(t - s)\frac{1}{T_s} \sum_{n=-\infty}^{\infty} \delta(t - s - nT_s)
$$

Repeating many of the steps of the previous derivations, one can show that

$$
\int_{-\infty}^{\infty} (\sum_{-\infty}^{\infty} \delta(t - nT_s)) e^{-j\omega t} dt = \sum_{-\infty}^{\infty} e^{-j\omega nT_s} = 2\pi \sum_{-\infty}^{\infty} \delta(\omega T_s - 2\pi n) = \frac{2\pi}{T_s} \sum_{-\infty}^{\infty} \delta(\omega - \frac{2\pi n}{T_s})
$$

$$
\begin{aligned}
S_z(\omega) &= \frac{1}{2\pi} \frac{1}{T_s} [S_x(\omega) * \frac{2\pi}{T_s} \sum_{-\infty}^{\infty} \delta(\omega - \frac{2\pi n}{T_s})] \\
&= \frac{1}{T_s^2} \sum_{-\infty}^{\infty} S_x(\omega) * \delta(\omega - \frac{2\pi n}{T_s}) \\
&= \frac{1}{T_s^2} \sum_{-\infty}^{\infty} S_x(\omega - \frac{2\pi n}{T_s})
\end{aligned}
\tag{6.15}
$$

and a similar analysis of aliasing is possible. The only difference is that frequency is no longer normalized to be between $[-\pi, \pi]$ radians/sec.;the original time-scale has been preserved, so that ω can take any real value in $[-\infty, \infty]$.

Chapter 7

Model Identification for Discrete-Time Processes

In this section, we address the issue of developing a model for a discrete-time wide-sense stationary scalar-valued process $x(t)$ based on observations of a particular sample path. The approach is to represent the process as the output of a linear, time-invariant discrete-time system driven by a wide-sense stationary input, and to estimate the parameters which define the impulse response of the linear, time-invariant system. First, we discuss a few classes of models which are commonly used. Without loss of generality, we assume that the processes are real-valued, to simplify the notation. Extensions to complex-valued processes are straightforward.

We assume initially that the process $x(t)$ is zero-mean and wide-sense stationary; later in the exposition, we discuss the implications of having a non-zero mean process $x(t)$. In order to emphasize this, we will use autocovariances instead of autocorrelation functions.

7.1   Autoregressive Models

An autoregressive model is also known as an "all-pole" model, and has the general form

$$x(n) = \sum_{i=1}^{p} a_i x(n-i) + u(n) \tag{7.1}$$

where $u(n)$ is an input process, modeled as a white-noise sequence; that is, an independent, identically distributed sequence of zero-mean Gaussian random variables with covariance σ^2. The above model is referred to as an *autoregressive model of order p*, referring to the order of the transfer function. The transfer function of the discrete-time system is given by:

$$H(j\omega) = \frac{1}{1 - \sum_{i=1}^{p} a_i e^{-j\omega i}}, \text{for } \omega \in (-\pi, \pi)$$

In order for the output to be wide-sense stationary, it is necessary that the linear system in (7.1) be stable.

Given that the input is white noise, it is straightforward to obtain the statistics of the output process $x(n)$. In particular, the power spectral density is given by:

$$S_x(\omega) = \sigma^2 \Big| \frac{1}{1 - \sum_{i=1}^{p} a_i e^{-j\omega i}} \Big|^2$$

Example 7.1
Consider a first-order autoregressive model, of the form

$$x(n) = a_1 x(n-1) + u(n)$$

where $|a_1| < 1$, so that the system is stable. Under this condition, the process $x(n)$ is wide-sense stationary, and has mean given by

$$m_x = a_1 m_x + m_u = a_1 m_x + 0 = 0$$

Similarly, the variance of the process $x(n)$ can be obtained as follows: Note that $x(n)$ can be written as the sum of two independent components, $a_1 x(n-1)$ and $u(n)$. Then, $\sigma_x^2 = |a_1|^2 \sigma_x^2 + \sigma^2$. Solving this equation yields

$$\sigma_x^2 = \frac{1}{1 - |a_1|^2} \sigma^2$$

Note that we exploited the stationarity of $x(n)$ to write the above equation. Finally, we can obtain a general expression for the autocovariance $K_x(m) = E[x(n)x(n+m)]$, as follows: let $m > 0$; then

$$E[x(n)x(n+m)] = E[x(n)(a_1 x(n+m-1) + u(n+m)] = a_1 K_x(m-1) = K_x(m)$$

This equation yields a recursive equation for $K_x(m)$, with initial condition $K_x(0) = \frac{1}{1-|a_1|^2}\sigma^2$. We can solve the recursive equation to obtain

$$K_x(m) = a_1^m K_x(0)$$

The power spectral density is given by

$$S_x(\omega) = \sigma_x^2 |\frac{1}{1 - a_1 e^{-j\omega}}|^2$$

Note that, in the above example, if one were to know or estimate the value of $K_x(0)$ and $K_x(1)$, one would have a complete set of equations for defining the parameters a_1 and σ^2 of the autoregressive model! In particular, if we assume that the process $x(t)$ is ergodic in autocorrelation, then the autocorrelations would be estimated from a single sample path, thereby providing the needed information for model identification.

Let's consider now the general autoregressive model. Given that the process is wide-sense stationary, we can use equation (7.1) to obtain:

$$K_x(m) = E[x(n+m)x(n)] = E[x(n)x(n-m)] = E[(\sum_{i=1}^{p} a_i x(n-i) + u(n))x(n-m)] \tag{7.2}$$

Furthermore, since $u(n)$ is independent of $x(n-m)$ for all $m > 0$, this equation is further simplified to:

$$K_x(m) = \sum_{i=1}^{p} a_i E[x(n-i)x(n-m)] = \sum_{i=1}^{p} a_i E[x(0)x(i-m)] = \sum_{i=1}^{p} a_i K_x(i-m) \tag{7.3}$$

which is valid for all $m \geq 1$. For $m = 0$, one obtains

$$K_x(0) = E[x(n)x(n)] = \sum_{i=1}^{p} a_i K_x(i) + \sigma^2 \tag{7.4}$$

due to the correlation between $u(n)$ and $x(n)$.

Now, the idea in system identification is the following: Suppose you are given $K_x(m)$ for all m. What is the value of $a_1, \ldots, a_p$? To solve this question, consider using (7.3) for $m = 1, \ldots, p$:

$$\begin{bmatrix} K_x(1) \\ K_x(2) \\ \vdots \\ K_x(p) \end{bmatrix} = \begin{bmatrix} K_x(0) & K_x(1) & \cdots & K_x(p-1) \\ K_x(1) & K_x(0) & \cdots & K_x(p-2) \\ \vdots & & & \vdots \\ K_x(p-1) & K_x(p-2) & \cdots & K_x(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ \alpha_p \end{bmatrix} \tag{7.5}$$

This is a set of linear equations of the form $\underline{r} = R\underline{a}$, which can be solved uniquely provided R is invertible! To convince yourself that R should be invertible, note that R can be written as follows: Let

$$\underline{x}_p = \begin{bmatrix} x(p) \\ x(p-1) \\ \vdots \\ x(1) \end{bmatrix}$$

Then,

$$R = E[\underline{x}_p \underline{x}_p^T]$$

so that R is a covariance matrix which is positive semidefinite at least. To show that it is strictly positive definite (and thus invertible), one can use a contradiction argument. Let

$$\underline{c} = \begin{bmatrix} 1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix}$$

be such that $\underline{c}^T R \underline{c} = 0$. Then, $\underline{c}^T \underline{x}_p$ is a random variable with zero mean and covariance $\underline{c}^T R \underline{c} = 0$, and thus would be identically equal to zero. This means that $x(p)$ would be expressible exactly as a linear combination of $x(1), \ldots, x(p-1)$, which would imply that, from (7.1), the variance of $u(n)$ is zero! This would imply that the process $x(n)$ is purely deterministic, and so $K_x(0) = 0$. Thus, as long as $K_x(0) \neq 0$, we know that no such $\underline{c}$ can exist.

The only flaw in the above argument is if there is a $\underline{c}$ with $\underline{c}^T R \underline{c} = 0$, and its first coordinate $c_1 = 0$. Then, one can always find the first non-zero coordinate, normalize it to one, and repeat the above argument to establish that, as long as $K_x(0) > 0$, then R is invertible.

For numerical reasons, it is useful to normalize the linear equations in (7.5). These equations can be rewritten as

$$\frac{1}{K_x(0)} \begin{bmatrix} K_x(1) \\ K_x(2) \\ \vdots \\ K_x(p) \end{bmatrix} = \frac{1}{K_x(0)} \begin{bmatrix} K_x(0) & K_x(1) & \cdots & K_x(p-1) \\ K_x(1) & K_x(0) & \cdots & K_x(p-2) \\ \vdots & & & \vdots \\ K_x(p-1) & K_x(p-2) & \cdots & K_x(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ \alpha_p \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \frac{K_x(1)}{K_x(0)} & \cdots & \frac{K_x(p-1)}{K_x(0)} \\ \frac{K_x(1)}{K_x(0)} & 1 & \cdots & \frac{K_x(p-2)}{K_x(0)} \\ \vdots & & & \vdots \\ \frac{K_x(p-1)}{K_x(0)} & \frac{K_x(p-2)}{K_x(0)} & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ \alpha_p \end{bmatrix} \quad (7.6)$$

Defining the normalized correlation coefficients $r_x(n) = \frac{K_x(n)}{K_x(0)}$, we have

$$\begin{bmatrix} r_x(1) \\ r_x(2) \\ \vdots \\ r_x(p) \end{bmatrix} = \begin{bmatrix} r_x(0) & r_x(1) & \cdots & r_x(p-1) \\ r_x(1) & r_x(0) & \cdots & r_x(p-2) \\ \vdots & & & \vdots \\ r_x(p-1) & r_x(p-2) & \cdots & r_x(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ \alpha_p \end{bmatrix} \quad (7.7)$$

The above equations are called the Yule-Walker equations, and provide the relationships for optimal parameter estimation in autoregressive models.

The matrix R which occurs in the Yule-Walker equations has a special structure, called a *Toeplitz* structure. Its elements R_{ij} are a function only of the difference $i - j$! Thus, they are constant along diagonals of the matrix. This structure can be exploited to solve efficiently for the inverse of R.

7.2 Moving Average Models

The autoregressive models discussed above represent the current value of the process in terms of previous values of the process, plus the current value of the input. A different class of models, called *moving average models*, represent the current value of the process in terms of current and previous values of the inputs only. The general form of the moving average models is:

$$x(n) = \sum_{i=1}^{p} b_i u(n - i) + u(n) \quad (7.8)$$

where $u(n)$ is assumed to be a wide-sense stationary white-noise process with unknown covariance σ^2.

The transfer function of the moving average model is given by:

$$H(z) = 1 + \sum_{i=1}^{p} b_i z^{-i}$$

Moving average models have several special properties. Let $m > p$; then,

$$K_x(m) = E[x(n+m)x(n)] = E[(\sum_{i=1}^{p} b_i u(n+m-i) + u(n+m))u(n)] = 0$$

due to the independence of the $u(n)$ sequence. Furthermore, we can use the defining relationship (7.8) to obtain a sequence of expressions relating the coefficients b_i to the autocorrelation function $K_x(m)$, as follows:

$$K_x(0) = E[x(n)^2] = \sigma^2(1 + \sum_{i=1}^{p} b_i^2) \tag{7.9}$$

due to the mutual independence of the terms in (7.8). Similarly,

$$
\begin{aligned}
K_x(1) &= E[x(n)x(n-1)] = E[(\sum_{i=1}^{p} b_i u(n-i) + u(n))(\sum_{i=1}^{p} b_i u(n-1-i) + u(n-1))] \\
&= E[\sum_{i=1}^{p} b_i b_{i-1} u(n-i)^2] = \sigma^2(\sum_{i=1}^{p} b_i b_{i-1}) \tag{7.10}
\end{aligned}
$$

where the convention $b_0 = 1$ is used. Continuing the recursion, one obtains

$$K_x(2) = E[x(n)x(n-2)] = \sigma^2(\sum_{i=2}^{p} b_i b_{i-2}) \tag{7.11}$$

$$K_x(k) = \sigma^2(\sum_{i=k}^{p} b_i b_{i-k}), \quad k \le p \tag{7.12}$$

The above equations form a set of nonlinear equations for the simultaneous solution of the coefficients $b_1, \ldots, b_p$ and σ^2 in terms of the autocorrelation values $K_x(m), m \le p$. Unlike the autoregressive models, there is no simple closed-form expression for this solution due to the nonlinearity of the equations. However, for simple cases (such as $p = 1$), a closed-form solution is possible, as illustrated below:

Example 7.2

Consider a first-order moving average model. The two unknown parameters, b_1 and σ^2, are evaluated solving the following simultaneous equations:

$$K_x(0) = \sigma^2(1 + b_1^2)$$
$$K_x(1) = \sigma^2 b_1$$

Solving together, we have

$$\frac{b_1}{1 + b_1^2} = \frac{K_x(1)}{K_x(0)} = r_x(1)$$

Thus,

$$(1 + b_1^2)r_x(1) - b_1 = 0$$

Solving, we get

$$b_1 = \frac{1 \pm \sqrt{1 - 4r_x(1)^2}}{2r_x(1)}$$

Clearly, for a real-valued b_1 to exist, it is necessary that $2r_x(1) \le 1$. Having solved for b_1, one obtains σ^2 as

$$\sigma^2 = \frac{K_x(1)}{b_1}$$

7.3 Autoregressive Moving Average (ARMA) Models

One can combine the autoregressive and moving average models into an ARMA model of the form

$$x(n) = \sum_{i=1}^{p} a_i x(n-i) + \sum_{i=1}^{q} b_i u(n-i) + u(n) \tag{7.13}$$

The transfer function of this system is given by

$$H(z) = \frac{1 + \sum_{i=1}^{q} b_i z^{-i}}{1 - \sum_{i=1}^{p} a_i z^{-i}}$$

As before, this transfer function must correspond to a stable system, so that the process $x(n)$ is wide-sense stationary.

   The properties of this system are similar to the combined properties of autoregressive and moving average models. In particular, it is difficult to find a closed-form solution for the coefficients b_i using only the values of the autocorrelation coefficients $K_x(m)$. Using a similar argument as before, we have:

$$
\begin{aligned}
K_x(0) &= E\left[x(n) \left(\sum_{i=1}^{p} a_i x(n-i) + \sum_{i=1}^{q} b_i u(n-i) + u(n) \right) \right] \\
&= \sum_{i=1}^{p} a_i K_x(i) + \sigma^2 + \sum_{i=1}^{q} b_i K_{xu}(-i) \tag{7.14}
\end{aligned}
$$

Unlike the previous cases, computation of $K_{xu}(-i)$ is more involved for ARMA models. Instead of worrying about the details of such computations, we focus only on the first-order ARMA model in an example:

Example 7.3
Consider a first-order ARMA model, of the form

$$x(n) = a_1 x(n-1) + b_1 u(n-1) + u(n) \tag{7.15}$$

The autocorrelation function is given by:

$$
\begin{aligned}
K_x(0) &= a_1 K_x(1) + \sigma^2 + b_1 E[x(n)u(n-1)] \\
&= a_1 K_x(1) + \sigma^2 + b_1 E[(a_1 x(n-1) + b_1 u(n-1) + u(n))u(n-1)] \\
&= a_1 K_x(1) + \sigma^2(1 + b_1^2) + a_1 b_1 E[x(n-1)u(n-1)] = a_1 K_x(1) + \sigma^2(1 + b_1^2 + a_1 b_1) \tag{7.16}
\end{aligned}
$$

$$
\begin{aligned}
K_x(1) &= E[x(n-1)(a_1 x(n-1) + b_1 u(n-1) + u(n))] \\
&= a_1 K_x(0) + b_1 \sigma^2 \tag{7.17}
\end{aligned}
$$

$$K_x(2) = E[x(n-2)(a_1 x(n-1) + b_1 u(n-1) + u(n))] = a_1 K_x(1) \tag{7.18}$$

Note that obtaining the coefficient a_1 is easy from $K_x(2)$ and $K_x(1)$. However, solving for b_1 involves again the solution of a quadratic equation.

   In general, for ARMA models, let $m > q$. Then,

$$
\begin{aligned}
K_x(m) &= E[x(n-m)(\sum_{i=1}^{p} a_i x(n-i) + \sum_{i=1}^{q} b_i u(n-i) + u(n))] \\
&= \sum_{i=1}^{p} a_i K_x(m-i) + \sum_{i=1}^{q} b_i E[x(n-m)u(n-i)] \\
&= \sum_{i=1}^{p} a_i K_x(m-i) \tag{7.19}
\end{aligned}
$$

Thus, by obtaining sufficient values of $K_x(m), m = 1, \ldots, q+p$, one can obtain a set of linear equations relating the coefficients $a_i, i = 1, \ldots, p$ to the values of the autocorrelation functions. Thus, the autoregressive part of ARMA models is easy to determine from the autocorrelation function. However, determining the moving average coefficients b_i require the solution of nonlinear equations, as before.

7.4   Dealing with non-zero mean processes

Assume that the process $x(n)$ is wide-sense stationary, with mean m_x, and we want to obtain a model of the process in the manner of the previous subsections. The point is that, in order for $x(n)$ to be non-zero mean, the input $u(n)$ must also be non-zero mean! Indeed, by superposition, we have

$$u(n) = m_u + \tilde{u}(n); \quad x(n) = m_x + \tilde{x}(n)$$

where $\tilde{u}(n), \tilde{x}(n)$ are zero mean processes. The relationship between these processes can be summarized by the transfer function $H(j\omega)$, as follows:

$$m_x = H(0)m_u$$

$$S_{\tilde{x}}(\omega) = |H(j\omega)|^2 S_{\tilde{u}}(\omega)$$

The above equations indicate that we can identify the model parameters which define the transfer function strictly by considering the autocovariance of $x(n)$, as indicated in the previous section. To complete the model, we need to specify the input mean m_u; given knowledge of the transfer function $H(j\omega)$ as identified previously, one obtains:

$$m_u = m_x \frac{1 + \sum_{i=1}^{p} a_i}{1 - \sum_{i=1}^{q} b_i}$$

Chapter 8

Detection Theory

In this chapter we start our investigation of detection theory, also referred to as hypothesis testing or decision theory. Our goal in these problems is to estimate or infer the value of an unknown "state of nature" based on noisy observations. A general model of this process is shown in Figure 8.1. Nature generates an unknown output H. By convention, we call this output a *hypothesis*. This outcome generated by nature then probabilistically affects the quantities Y that we are allowed to observe. Based on the uncertain observation Y, we must design a rule to decide what the unknown hypothesis was. In the theory of detection the set of possible hypotheses is taken to be discrete. When the set of possibilities is continuous we are in the realm of estimation, which is discussed in Chapter 10. From Figure 8.1 we see that we will need three components in our model:

1. A model of generation processes that creates H – i.e. a model of nature.

2. A model of the observation process.

3. A decision rule $D(y)$ that maps each possible observation y to an associated decision.

In general, the first two elements are set by "nature" or the restrictions of the physical data gathering situation. For example, if we are trying to decide whether a tumor is cancerous or not, the true state of the tumor is decreed by processes outside of our control and the uncertainty or noise in the observations may arise from the physical processes in generating an X-Ray image. It is generally the last element, decision rule design, where the engineer plays the strongest role. Such decision rules can be of two types: deterministic and random. A deterministic decision rule always assigns the same decision or estimate to the same observation – i.e. when a given observation is seen the same decision is always made. In particular, deterministic decision rules can be viewed as a simple partitioning or labeling of the space of observations into disjoint regions marked with the decision corresponding to each observation, as shown in Figure 8.2. In the case of random decision rules, different decisions may arise from the same observation – i.e. when the same observation is made twice, two different decision outcomes are possible. Such random decision rules play an important role when the observed quantity y is discrete in nature. In general, however, our emphasis will be on the design of deterministic decision rules.
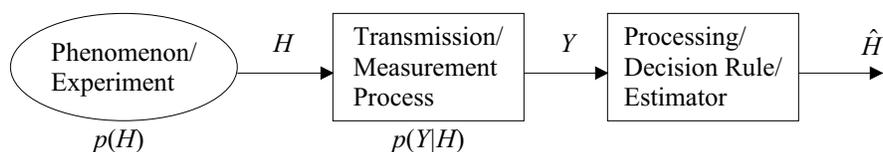

Figure 8.1: Detection problem components.

In Section 8.1 we discuss in detail the case that arises when there are only two possible hypotheses, termed binary hypothesis testing. In Section 8.4 we discuss the more general case of M hypotheses. Throughout
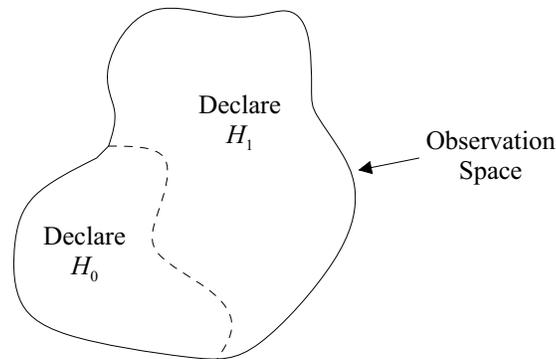
Figure 8.2: Illustration of a deterministic decision rule as a division of the observation space into disjoint regions, illustrated here for the case of two possibilities.

this chapter we focus on the case of detection based on observations of random variables. In Chapter 9 we examine the more complicated case of detection based on observations of random processes.

8.1 Bayesian Binary Hypothesis Testing

In this section we consider the simplest case when there are only two possible states of nature or hypotheses, which by convention we label as H_0 and H_1. This situation is termed "binary hypothesis testing" and the H_0 hypothesis is usually termed the "null hypothesis," due to its typical association with the absence of some quantity of interest. The binary case is of considerable practical importance, as well as having a long and rich history. To give a flavor of the possibilities, let us examine a few examples before proceeding to more detailed developments.

Example 8.1 (Communications)
Consider the following simplified version of a communication system, where a source broadcasts one bit, (either 0 or 1) The transmitter encodes this bit by a voltage, which is either 0 or E, depending on the bit. The receiver observes a noisy version of the transmitted signal, where the noise is additive, and is represented by a random variable w with zero-mean, variance σ^2, and Gaussian distribution. The receiver knows the nature of the signal E, the statistics of the noise σ^2, and the apriori probability $p(k)$ that the bit sent was k, where $k = 0, 1$. The receiver must take the received signal, y, and map this using a rule $D(y)$ into either 0 or 1, depending on the value of r. The problem is to determine the decision rule for which the probability of receiver error is minimized.

In the above example there are two possible hypotheses, H_0 and H_1, only one of which can be true. These hypotheses correspond to whether the transmitted bit was 0 or 1. There is a probabilistic relationship between the observed variable y and the hypotheses H_i. In particular, the observed variable is $y = w$ for hypothesis H_0, and $y = E + w$ for hypothesis H_1. The decision rule divides the space of possible observations into two disjoint decision regions, Z_0 and Z_1, such that, whenever an observation falls into Z_i, the decision that H_i is the correct hypothesis is made. In the example, these regions correspond to the values of y for which $D(y) = 0$ and the values of y for which $D(y) = 1$. These decision regions are established to maximize an appropriate criterion of performance, corresponding to the probability of a correct decision.

Consider other examples:

Example 8.2 (Radar)
A simple radar system makes a scalar observation y to determine the absence or presence of a target at a given range and heading. If a target is present (hypothesis H_1), the observed signal is $y = E + w$, where E is a known signal level, and $w \sim N(0, \sigma^2)$. If no target is present (hypothesis H_0), then only noise is received $y = w$. Find the decision rule for maximizing the probability of detecting the target, given a bound on the probability of false alarm.

Example 8.3 (Quality Control)
At a factory, an automatic quality control device is used to determine whether a manufactured unit is satisfactory (hypothesis H_0) or defective (hypothesis H_1), by measuring a simple quality factor q. Past statistics indicate that one out of every 10 units is defective. For satisfactory units, $q \sim N(2, \sigma^2)$, whereas for defective units, $q \sim N(1, \sigma^2)$. The quality control device is set to remove all units for which $q < t$, where t is a threshold to be designed. The problem is to determine the optimal threshold setting in order to maximize the probability of detecting a defect, subject to the constraint that the probability of removing a satisfactory unit is at most 0.005.

All of the above examples illustrate the problem of binary hypothesis testing. We will develop the relevant theory next.

8.1.1 Bayes Risk Approach and the Likelihood Ratio Test

We are now interested in obtaining "good" decision rules for the binary hypothesis testing case. A rational and common approach is to minimize a cost function given our models of the situation. Building on the development of the introduction, the elements of this approach in the binary case are:

1. **Model of Nature:** In the binary case there are only two possibilities, denoted as H_0 and H_1. Our knowledge of these possibilities is captured by the *prior probabilities* $P_i = \Pr(H = H_i)$. Note that $P_1 = 1 - P_0$.

2. **Observation Model:** As figure 8.1 indicates, the observation model captures the relationship between the observed quantity y and the unknown hypothesis H. This relationship is given by the *conditional densities* $p_{Y|H}(y \mid H_i)$.

3. **Decision Rule:** Our decision rule $D(y)$ is obtained by minimizing the average cost, called the "Bayes risk." Let C_{ij} denote the cost of deciding hypothesis $D(y) = H_i$ when hypothesis H_j is true, then the Bayes risk of the decision rule is given by:

$$E\left[C_{D(y),H}\right] = \sum_{i=0}^{1} \sum_{j=0}^{1} C_{ij} \Pr\left(D(y) = H_i, H_j \text{ true}\right) \qquad (8.1)$$

Note that the outcome of deciding H_i in (8.1) is *random*, even if the decision rule is deterministic, because y itself is random. Thus the expectation in (8.1) averages over both the randomness in the true hypothesis H_j (i.e. the randomness in the state of nature) as well as the randomness in the observation, and thus decision outcome (i.e. the randomness in the data).

There are two key assumptions in the Bayes risk approach to the hypothesis testing problem which is formulated above. First, apriori probabilities of each hypothesis occurring P_i can be determined. Second, decision costs C_{ij} can be meaningfully assigned. Under these two assumptions, the Bayes risk hypothesis testing problem above is well posed. Clearly, the key is the minimization of the Bayes risk $E\left[C_{D(y)}\right]$.

Let us now focus on finding the decision rule that minimizes the Bayes risk. Recall (Figure 8.2) that a deterministic decision rule $D(y)$ is nothing more than a division of the observation space $\mathcal{R}^n$ into disjoint decision regions Z_0 and Z_1 such that when $y \in Z_i$ our decision is H_i. Thus finding a deterministic decision rule in the binary case is simply a matter of figuring out which region to assign each observation to. Combining this insight with Bayes rule we proceed by rewriting the Bayes risk as follows:

$$E\left[C_{D(y)}\right] = E\left[E\left[C_{D(y)} \mid y\right]\right] = \int E\left[C_{D(y)} \mid y\right] p_Y(y)\, dy \qquad (8.2)$$

Now $p_Y(y)$ is always non-negative and the value of $E\left[C_{D(y)} \mid y\right]$ only depends on the decision region to which we assign the particular value y, so we can minimize (8.2) by minimizing $E\left[C_{D(y)} \mid y\right]$ for each value of y. Thus, the optimal decision is to choose the hypothesis that gives the smallest value of the conditional expected cost $E\left[C_{D(y)} \mid y\right]$ for the given value of y.

Now the conditional expected cost is given by:

$$E\left[C_{D(y)} \mid y\right] = \sum_{i=0}^{1} \sum_{j=0}^{1} C_{ij} \Pr\left(D(y) = H_i, H_j \text{ true} \mid y\right) \qquad (8.3)$$

But $\Pr\left(D(y) = H_i, H_j \text{ true} \mid y\right)$ will either equal 0 or $\Pr\left(H_j \text{ true} \mid y\right)$ for a deterministic decision rule, since the decision outcome *given* y is non-random! In particular, for a given observation value y, the expected value of the conditional cost if we choose to assign the observation to H_0 is given by:

$$\text{If } D(y) = H_0: \qquad E\left[C_{D(y)=H_0} \mid y\right] = C_{00}\, p_{H|Y}\left(H_0 \mid y\right) + C_{01}\, p_{H|Y}\left(H_1 \mid y\right) \tag{8.4}$$

where $p_{H|Y}\left(H_i \mid y\right)$ denotes $\Pr\left(H_i \text{ true} \mid Y = y\right)$. Similarly, the expected value of the conditional cost if we assign this value of y to H_1 is given by:

$$\text{If } D(y) = H_1: \qquad E\left[C_{D(y)=H_1} \mid y\right] = C_{10}\, p_{H|Y}\left(H_0 \mid y\right) + C_{11}\, p_{H|Y}\left(H_1 \mid y\right) \tag{8.5}$$

Given the discussion above, the optimal thing to do is to make the decision that results in the smaller of the two conditional costs. We can compactly represent this comparison and its associated decision rule as follows:

$$C_{00}p_{H|Y}\left(H_0 \mid y\right) + C_{01}p_{H|Y}\left(H_1 \mid y\right) \underset{H_0}{\overset{H_1}{\gtrless}} C_{10}p_{H|Y}\left(H_0 \mid y\right) + C_{11}p_{H|Y}\left(H_1 \mid y\right) \tag{8.6}$$

where $\underset{H_0}{\overset{H_1}{\gtrless}}$ denotes choosing H_1 is the inequality is $>$ and choosing H_0 if the inequality is $<$. The decision rule given in (8.6) represents the optimal Bayes risk decision rule in its most fundamental form.

Now from Bayes rule we have that:

$$p_{H|Y}(H_i \mid y) = \frac{p_{Y|H}(y \mid H_i)p_H(H_i)}{p_Y(y)} \tag{8.7}$$

Substituting (8.7) into (8.6) and dividing through by $p_Y(y)$ we obtain:

$$\left(C_{01} - C_{11}\right) P_1 p_{Y|H}(y \mid H_1) \underset{H_0}{\overset{H_1}{\gtrless}} \left(C_{10} - C_{00}\right) P_0 p_{Y|H}(y \mid H_0) \tag{8.8}$$

which expresses the optimum Bayes risk decision rule in terms of the prior probabilities P_i and the data "likelihoods" $p_{Y|H}(y \mid H_i)$. Note that the expressions (8.7) and (8.8) are valid for any assignments of the costs C_{ij}.

If we further make the reasonable assumption that errors are more costly than correct decisions, so that

$$\left(C_{01} - C_{11}\right) \;>\; 0 \tag{8.9}$$
$$\left(C_{10} - C_{00}\right) \;>\; 0 \tag{8.10}$$

we can rewrite the optimum Bayes risk decision rule $D(y)$ in (8.8) as follows:

$$\mathcal{L}(y) = \left[\frac{p_{Y|H}(y \mid H_1)}{p_{Y|H}(y \mid H_0)}\right] \quad \underset{H_0}{\overset{H_1}{\gtrless}} \quad \frac{\left(C_{10} - C_{00}\right) P_0}{\left(C_{01} - C_{11}\right) P_1} \equiv \eta \tag{8.11}$$

The consequences of (8.11) are considerable and we will take some time to discuss them. First, examining (8.11) we see that the form of the optimal Bayes risk decision rule is to compare the ratio $\mathcal{L}(y)$, which is termed the *likelihood ratio*, to a threshold, which is given by η. The value of this threshold is determined, in general, by both the prior probabilities and the assigned cost structure, both of which are known at the outset of the problem (i.e. involve prior knowledge). The test (8.11) is called a *likelihood ratio test* or LRT, for obvious reasons, and thus *all* optimal decision rules (in the Bayes risk sense) are LRTs (with perhaps different thresholds). Thus, while as engineers we may disagree on such details as the assignment of costs and prior probabilities, the *form* of the optimal test (i.e. the data processing) is always the same and given by the LRT. Indeed, while the threshold η can be set by choosing costs and prior probability assignments, it is also possible to view it simply as a tunable parameter.

Second, examining (8.11) we see that the data or observations enter the decision *only* through the likelihood ratio $\mathcal{L}(y)$. Because it is a function of the uncertain observation y, it is itself a random variable.

Since this scalar function of the data is all that is needed to perform the optimal test, it is a *sufficient statistic* for the detection problem. That is, instead of making a decision based on the original observations y, it is sufficient to make the decision based only on the likelihood ratio, which is a function of y.

Finally, note that the sufficient statistic $\mathcal{L}(y)$ is a scalar. Thus the LRT is a scalar test, independent of the dimension of the observation space. This means we can make a decision in the binary case by making a single comparison, independent of whether we have 1 observation or 1 million.

Before moving on to look at special cases we note that there is another form of (8.11) that is sometimes used. In particular, taking logarithms of both sides of (8.11) does not change the inequality and results in the following equivalent test:

$$\ln\left[\mathcal{L}(y)\right] \underset{H_0}{\overset{H_1}{\gtrless}} \ln\left[\frac{(C_{10} - C_{00})\, P_0}{(C_{01} - C_{11})\, P_1}\right] \tag{8.12}$$

The quantity on the left hand side of (8.12) is called the *log-likelihood ratio*, and as we will see, is conveniently used in Gaussian problems.

8.1.2 Special Cases

Let us now consider some common special cases of the Bayes risk and the associated decision rules corresponding to them.

MPE cost assignment and the MAP rule

Suppose we use the following cost assignment:

$$C_{ij} = 1 - \delta_{ij} \tag{8.13}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. Then the cost of all errors ($C_{10} = C_{01} = 1$) are the same and there is no cost for correct decisions ($C_{00} = C_{11} = 0$). In this case, the Bayes risk is given by:

$$
\begin{aligned}
E\left[C_{D(y)}\right] &= C_{00}\Pr\left[\text{Decide } H_0,\, H_0 \text{ true}\right] \\
&\quad + C_{01}\Pr\left[\text{Decide } H_0,\, H_1 \text{ true}\right] \\
&\quad + C_{10}\Pr\left[\text{Decide } H_1,\, H_0 \text{ true}\right] \\
&\quad + C_{11}\Pr\left[\text{Decide } H_1,\, H_1 \text{ true}\right] \\
&= \Pr\left[\text{Decide } H_0,\, H_1 \text{ true}\right] + \Pr\left[\text{Decide } H_1,\, H_0 \text{ true}\right] \\
&= \Pr\left[\text{Error}\right]
\end{aligned}
\tag{8.14, 8.15}
$$

Thus the optimal detector for this cost assignment minimizes the probability of error. The corresponding decision rule is termed the minimum probability of error (MPE) decision rule and is given by:

$$\frac{p_{Y|H}(y \mid H_1)}{p_{Y|H}(y \mid H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P_0}{P_1} \tag{8.16}$$

Since $p_{Y|H}(y \mid H_i)P_i = p_{H|Y}(H_i \mid y)p_Y(y)$ we can rewrite the MPE decision rule (8.16) in the following form:

$$p_{H|Y}(H_1 \mid y) \underset{H_0}{\overset{H_1}{\gtrless}} p_{H|Y}(H_0 \mid y) \tag{8.17}$$

This decision rule says that for minimum probability of error choose the hypothesis whose posterior probability is higher. This is termed the *Maximum aposteriori probability or MAP rule*. Thus we see that the MAP rule is also the MPE rule independent of prior probabilities.

The ML rule

Now suppose we again use the MPE cost criterion with $C_{ij} = 1 - \delta_{ij}$, but also have both hypotheses equally likely apriori so that $P_0 = P_1 = 1/2$. In this case we essentially have no prior preference for one hypothesis over the other. With these assignments we can see that the threshold in (8.11) is given by $\eta = 1$ so that the decision rule becomes:

$$p_{Y|H}(y \mid H_1) \underset{H_0}{\overset{H_1}{\gtrless}} p_{Y|H}(y \mid H_0) \tag{8.18}$$

In this case the decision rule is to choose the hypothesis that gives the higher likelihood of the observation. For this reason this rule is called the *maximum likelihood or ML rule*

Scalar Gaussian Detection

Here we consider the problem of deciding which of two possible Gaussian distributions a single scalar observation comes from. In particular, under hypothesis H_i the observation is distributed according to:

$$p_{Y|H}(y \mid H_i) \;=\; N\left(y; m_i, \sigma_i\right) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}\frac{(y-m_i)^2}{\sigma_i^2}} \tag{8.19}$$

These two possibilities are depicted in Figure 8.3. The likelihood ratio for this case is given by:
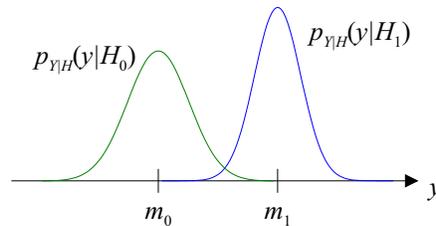

Figure 8.3: General scalar Gaussian case

$$\mathcal{L}(y) \;=\; \left[\frac{\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right) e^{-\frac{(y-m_1)^2}{2\sigma_1^2}}}{\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right) e^{-\frac{(y-m_0)^2}{2\sigma_0^2}}} \right] \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{8.20}$$

Now taking natural logs of both sides as in (8.12) and rearranging terms results in the following form of the optimal decision rule:

$$-\frac{(y-m_1)^2}{2\sigma_1^2} + \frac{(y-m_0)^2}{2\sigma_0^2} \underset{H_0}{\overset{H_1}{\gtrless}} \ln\left(\frac{\sigma_1}{\sigma_0}\eta\right) \tag{8.21}$$

Same Variances, Different Means: Let us consider some special sub-cases. First, suppose $\sigma_0 = \sigma_1 = \sigma$ and $m_1 > m_0$. In this case the Gaussian distributions have the same variance but different means and the task is to decide whether the observation came from the Gaussian with the greater or lesser mean. After simplification, (8.21) can be reduced to the following form:

$$y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{m_0 + m_1}{2} + \frac{\sigma^2 \ln(\eta)}{(m_1 - m_0)} \equiv \Gamma \tag{8.22}$$

This situation is depicted in Figure 8.4. There are some interesting things to note about this result. First there are two decision regions separated by Γ. In general, the boundary between the decision regions is an

adjusted threshold, which takes into account both the costs and the prior probabilities. For example, if we consider the ML rule (i.e. the MPE cost structure with equally likely hypotheses), then $\Gamma = (m_0 + m_1)/2$ and the boundary between decision regions is halfway between the means. In particular, in this case $\eta = 1$ and we can write the decision rule in the form:

$$\|y - m_0\|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \|y - m_1\|^2 \tag{8.23}$$

which says to choose the hypothesis "closest" to the corresponding mean. If, however, instead, we use the MPE cost structure, but $P_1 > P_0$ the decision boundary will move closer to m_0, since we expect to see the H_1 case more frequently. In any case, the data processing is linear. This will not always be the case.
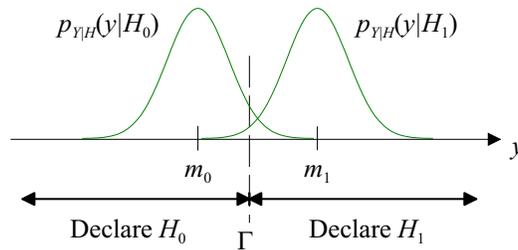

Figure 8.4: Scalar Gaussian case with equal variances

Different Variances, Same Means: Now consider what happens if we instead suppose $\sigma_0 < \sigma_1$ and $m_1 = m_0 = 0$. In this case the Gaussian distributions have the same mean, but different variances and the task is to decide whether the observation came from the Gaussian with the greater or lesser variance. After simplification, (8.21) can be reduced to the following form:

$$y^2 \underset{H_0}{\overset{H_1}{\gtrless}} 2 \left(\frac{\sigma_1^2 \sigma_0^2}{\sigma_1^2 - \sigma_0^2} \right) \ln \left(\frac{\sigma_1}{\sigma_0} \eta \right) \equiv \Gamma' \tag{8.24}$$

This situation is depicted in Figure 8.5. Note that the decision regions are no longer simple connected segments of the real line. Further, the decision rule is a nonlinear function of the observation y.
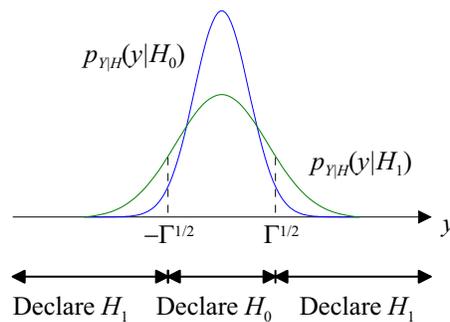

Figure 8.5: Scalar Gaussian case with equal means

8.1.3 Examples

Let us consider some examples.

Example 8.4 (Radar)
Consider the radar example, Example 8.2, discussed earlier. This is really just a scalar Gaussian detection problem. The likelihood ratio for this example is given by:

$$\mathcal{L}(y) = \frac{e^{\frac{-(y-E)^2}{2\sigma^2}}}{e^{\frac{-(y)^2}{2\sigma^2}}} = e^{\frac{2Ey-E^2}{2\sigma^2}} \tag{8.25}$$

Thus, the optimal decision rule is given by:

$$e^{\frac{2Ey-E^2}{2\sigma^2}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{8.26}$$

Taking logarithms of both sides means that the new decision rule can be restated as:

$$y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{E}{2} + \frac{\sigma^2 \ln(\eta)}{E} \tag{8.27}$$

In the case that the cost criterion is minimum probability of error (MPE) so that $C_{00} = C_{11} = 0, C_{01} = C_{10} = 1$, and the probability of each hypothesis is apriori equal ($P_0 = P_1 = 1/2$), we have that $\eta = 1$. Note that the optimal detection test in this case is to compute which mean the measurement is closer to! This is just an example of the scalar Gaussian detection problem treated above.

Example 8.5 (Multiple Observations)
Consider the radar detection example, except that N independent pulses are sent out, so that a vector of measurements is collected. This is the typical situation in radar systems, where multiple pulses are processed to improve the signal-to-noise ratio and thus obtain better detection performance. We assume that each pulse provides a measurement y_i, where

$$y_i = \begin{cases} n_i & \text{if hypothesis } H_0 \text{ is true (no target present)} \\ E + n_i & \text{if hypothesis } H_1 \text{ is true (target present)} \end{cases}$$

and n_i is a set of independent, identically distributed $N(0, \sigma^2)$ random variables. In this case, the likelihood ratio is given by:

$$\mathcal{L}(y) = \frac{p_{Y_1,\cdots Y_N | H}(y_1, \cdots, y_N \mid H_1)}{p_{Y_1,\cdots Y_N | H}(y_1, \cdots, y_N \mid H_0)} = \prod_{i=1}^{N} \frac{e^{\frac{-(y_i-E)^2}{2\sigma^2}}}{e^{\frac{-(y_i)^2}{2\sigma^2}}} = \prod_{i=1}^{N} e^{\frac{2Ey_i-E^2}{2\sigma^2}} = e^{\frac{2E\left(\sum_{i=1}^{N} y_i\right)-NE^2}{2\sigma^2}} \tag{8.28}$$

By again taking logs of both sides the decision rule can be reduced to:

$$\frac{1}{N}\sum_{i=1}^{N} y_i \underset{H_0}{\overset{H_1}{\gtrless}} \frac{E}{2} + \frac{\sigma^2 \ln(\eta)}{NE} \tag{8.29}$$

Comparing with (8.27), the effect of using the extra measurements is to reduce the measurement covariance by a factor of $N^{1/2}$.

Before, we said that the likelihood ratio was a sufficient statistic. It may not be the simplest sufficient statistic however. Whenever there is a function of the data, $g(y)$ such that the likelihood ratio can be computed strictly from $g(y)$, this value is *also* a sufficient statistic. Thus sufficient statistics are not unique. In the above example, it is clear that the sample mean, $\frac{1}{N}\sum_{i=1}^{N} y_i$, is a sufficient statistic for the detection problem; note that this is a linear function of the measurement vector y and much simpler than the likelihood ratio $\mathcal{L}(y_i)$ in (8.28).

Example 8.6
Assume that, under hypothesis H_0, we have a vector of N observations y, with independent, identically distributed $N(0, \sigma_0^2)$ components y_i. Under hypothesis H_1, we have a vector of N observations y, with independent, identically distributed $N(0, \sigma_1^2)$ components y_i. Thus, the two hypothesis correspond to multiple observations of independent identically

distributed random variables with the same mean but different covariances. The likelihood ratio is given by:

$$
\begin{aligned}
\mathcal{L}(y) &= \frac{\dfrac{e^{-\frac{\sum_{i=1}^{N} y_i^2}{2\sigma_1^2}}}{(2\pi\sigma_1^2)^{N/2}}}{\dfrac{e^{-\frac{\sum_{i=1}^{N} y_i^2}{2\sigma_0^2}}}{(2\pi\sigma_0^2)^{N/2}}} \\[2mm]
&= \frac{\sigma_1^N}{\sigma_2^N} e^{-\frac{\sum_{i=1}^{N} y_i^2}{2\sigma_1^2} + \frac{\sum_{i=1}^{N} y_i^2}{2\sigma_0^2}}
\end{aligned}
\tag{8.30}
$$

Again, after taking logs the optimal decision rule can be rewritten in terms of a simpler test, as:

$$
\frac{1}{N} \sum_{i=1}^{N} y_i^2 \underset{H_0}{\overset{H_1}{\gtrless}} 2\frac{\sigma_0^2 \sigma_1^2}{\sigma_1^2 - \sigma_0^2} \ln\left(\eta^{1/N} \frac{\sigma_1}{\sigma_0}\right)
\tag{8.31}
$$

Clearly, a sufficient statistic for this problem is the quadratic function of the measurements: $\frac{1}{N}\sum_{i=1}^{N} y_i^2$.

Before proceeding to another section, consider a problem which does not involve Gaussian random variables.

Example 8.7
Assume that we observe a random variable y which is Poisson distributed with mean m_0 when H_0 is true, and with mean m_1 when H_1 is true. Thus the likelihoods are given by:

$$
p_{Y|H}(y \mid H_i) = \frac{m_i^y e^{-m_i}}{k!}
\tag{8.32}
$$

Note that the measurements are discrete-valued; thus, the likelihood ratios will involve probability distributions rather than densities. The likelihood ratio is given by:

$$
\mathcal{L}(y) = \frac{p_{Y|H}(y \mid H_1)}{p_{Y|H}(y \mid H_0)} = \frac{m_1^y e^{-m_1}}{m_2^y e^{-m_2}}
\tag{8.33}
$$

Thus, the optimal decision rule can be written as:

$$
y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{(m_1 - m_0) + \ln(\eta)}{\ln\left(\frac{m_1}{m_0}\right)}
\tag{8.34}
$$

8.2   Performance and the Receiver Operating Characteristic

In the discussion so far we have focused on the form of the optimal test and on the nature of the data processing involved. We have found that the optimum Bayes risk test is the likelihood ratio test, where a function of the data (the likelihood ratio) is compared to a threshold. Let us now turn our attention to characterizing the performance of decision rules in general and LRT-based decision rules in particular. To aid in this discussion let us define the following standard terminology, arising from classical radar detection theory:

$$
\begin{aligned}
P_F &\equiv \Pr(\text{Choose } H_1 \mid H_0 \text{ True}) = \text{Probability of False Alarm} \quad (\text{called a ``Type I'' Error}) \\
P_D &\equiv \Pr(\text{Choose } H_1 \mid H_1 \text{ True}) = \text{Probability of Detection} \\
P_M &\equiv \Pr(\text{Choose } H_0 \mid H_1 \text{ True}) = \text{Probability of Miss} \quad (\text{called a ``Type II'' Error})
\end{aligned}
$$

The quantity P_F is the probability that the decision rule will declare H_1 when H_0 is true, while P_D is the probability that the decision rule will declare H_1 when H_1 is true and P_M is the probability that the decision rule will declare H_0 when H_1 is true. Note carefully that these are *conditional* probabilities!

Now there are two natural metrics to evaluate the performance of a decision rule. The first metric is the expected value of the cost $E[C_{D(y)}]$, i.e. the value of the Bayes risk. Let us examine this cost in more

detail. Following (8.14), and using Bayes rule and the definitions of the conditional densities P_F, P_M, and P_D above, the Bayes risk can is given by:

$$
\begin{aligned}
E\left[C_{D(y)}\right] &= \quad C_{00}\mathrm{Pr}\left[\text{Decide } H_0 \mid H_0\right] P_0 + C_{01}\mathrm{Pr}\left[\text{Decide } H_0 \mid H_1\right] P_1 \qquad\qquad (8.35)\\
&\quad + C_{10}\mathrm{Pr}\left[\text{Decide } H_1 \mid H_0\right] P_0 + C_{11}\mathrm{Pr}\left[\text{Decide } H_1 \mid H_1\right] P_1 \\
&= \quad C_{00}(1-P_F)P_0 + C_{01}(1-P_D)P_1 + C_{10}P_F P_0 + C_{11}P_D P_1 \\
&= \quad \underbrace{C_{00}P_0 + C_{01}P_1}_{\text{Fixed Cost}} + \underbrace{(C_{10}-C_{00})\,P_0 P_F - (C_{01}-C_{11})\,P_1 P_D}_{\text{Varies as function of decision rule}}
\end{aligned}
$$

Note that this cost has two components. The first component is independent of the decision rule used, is based only on the "prior" components of the problem, and represents a fixed cost. The second component varies as a function of the decision rule (e.g. as the threshold η of the LRT is varied). In particular, of the elements in this second component it is P_F and P_D that will vary as the decision rule is changed. Thus, from a performance standpoint, we can say that $E[C_{D(y)}]$ can be expressed purely as a function of P_F and P_D (where we assume C_{ij} and P_i are fixed).

A second natural performance metric of decision rules is the probability of error Pr[error]. Starting from (8.14) and again using Bayes rule and the definitions of P_F, P_M, and P_D we find:

$$
\begin{aligned}
\mathrm{Pr}\left[\text{Error}\right] &= \quad \mathrm{Pr}\left[\text{Decide } H_0,\ H_1 \text{ true}\right] + \mathrm{Pr}\left[\text{Decide } H_1,\ H_0 \text{ true}\right] \qquad\qquad (8.36)\\
&= \quad P_M P_1 + P_F P_0 \\
&= \quad (1-P_D)P_1 + P_F P_0
\end{aligned}
$$

Again, the parts of this expression that will vary as the decision rule is changed are P_D and P_F. Thus, we can also express Pr[Error] as a function of just P_D and P_F (again, assuming C_{ij} and P_i are fixed).

Let us summarize the development thus far. Given *any* decision rule we can determine its performance (i.e. either its corresponding Bayes risk $E\left[C_{D(y)}\right]$ or its Pr[Error]) by calculating P_D and P_F for the decision rule. Further, we know that "good" decision rules (i.e. those optimal in the Bayes risk sense) are likelihood ratio test – i.e. they compare the likelihood ratio to a fixed threshold to make their decision. The only undetermined quantity in a LRT is its threshold. Given this discussion it seems reasonable to limit ourselves to consideration of LRT decision rules and to calculate P_D and P_F for every possible value of the threshold η. Given this information, we have essentially characterized every possible "reasonable" decision rule. This information may be conveniently and compactly represented as graph of $P_D(\eta)$ versus $P_F(\eta)$ – that is, a plot of the points (P_F, P_D) as the parameter η is varied. Such an important plot for a decision rule has a special name – it is called the *Receiver Operating Characteristic* or ROC for the detection problem. An illustration of a ROC is given in Figure 8.6.
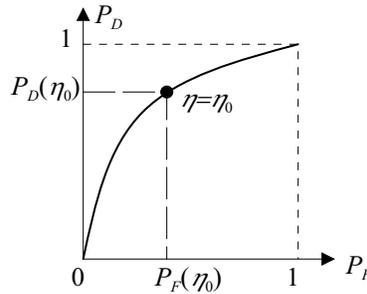

Figure 8.6: Illustration of ROC.

Let us emphasize some features of the ROC. First, note that the threshold η is a parameter along the curve. Thus any one point on the ROC corresponds to a particular choice of threshold (and vice versa). The ROC itself does not depend on the costs C_{ij} or the apriori probabilities P_i. These terms can be used, however, to determine a particular threshold, and thus a particular operating point corresponding to the

optimal Bayes risk detector. Finding appropriate values of these costs and densities can be challenging, and the ROC allows us to characterize the performance of all possible optimal detectors.

The key challenge in generating the ROC for a particular problem is finding the quantities P_D and P_F as a function of a threshold parameter. To this end, note that a general LRT decision rule can always be expressed in the following form:

$$\ell(y) \underset{H_0}{\overset{H_1}{\gtrless}} \Gamma \tag{8.37}$$

where $\ell(y)$ is a sufficient statistic for the detection problem and Γ is a corresponding threshold. The sufficient statistic might be the original likelihood ratio $\mathcal{L}(y) = p_{Y|H}(y \mid H_1)/p_{Y|H}(y \mid H_0)$ or it might be a simpler function of the observations, as we saw in the radar example. The important thing is that it completely captures the influence of the observations. Note that $\ell(y)$ is itself a random variable, since it is a function of y.

Now we can express P_D and P_F as follows:

$$
\begin{aligned}
P_D &= \Pr(\text{Choose } H_1 \mid H_1 \text{ True}) && (8.38)\\
&= \int_{\{y|\text{Choose } H_1\}} p_{Y|H}(y \mid H_1)\, dy && (8.39)\\
&= \int_{\ell > \Gamma} p_{L|H}(\ell \mid H_1)\, d\ell && (8.40)
\end{aligned}
$$

$$
\begin{aligned}
P_F &= \Pr(\text{Choose } H_1 \mid H_0 \text{ True}) && (8.41)\\
&= \int_{\{y|\text{Choose } H_1\}} p_{Y|H}(y \mid H_0)\, dy && (8.42)\\
&= \int_{\ell > \Gamma} p_{L|H}(\ell \mid H_0)\, d\ell && (8.43)
\end{aligned}
$$

The expressions (8.39) and (8.42) express the probabilities in terms of quantities in the space of the observations, i.e. in terms of the likelihoods. The expressions (8.40) and (8.43) express the probabilities in terms of quantities in the space of the test statistic and its densities. Both expressions are correct, and the choice of which to use is usually based on convenience, as we will see. Note that the region of integration (i.e. the set of values of y or ℓ used in calculation) is the *same* for both P_D and P_F, it is just the densities used that are different. We illustrate these ideas with an example.

Example 8.8 (Scalar Gaussian Detection)
Consider again the problem of determining which of two Gaussian densities of scalar observation comes from. In particular, suppose y is scalar and distributed $N(0, \sigma^2)$ under H_0 and distributed $N(E, \sigma^2)$ under H_1. We have seen in (8.27) that the optimal decision rule was:

$$\ell(y) = y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{E}{2} + \frac{\sigma^2 \ln(\eta)}{E} = \Gamma$$

In this case $\ell(y) = y$ so the observation space is the same as the space of the test statistic and it is easy to see that $\ell(y)$ will be a Gaussian random variable under either hypothesis. In particular, we have:

$$
\begin{aligned}
p_{L|H_1}(\ell \mid H_1) &= N\left(\ell; E, \sigma^2\right) && (8.44)\\
p_{L|H_0}(\ell \mid H_0) &= N\left(\ell; 0, \sigma^2\right) && (8.45)
\end{aligned}
$$

Now we can combine these densities with (8.40) and (8.43) to find P_D and P_F as we vary Γ from $(-\infty, \infty)$, which is the range of Γ which results from variations in η. Explicitly, we have

$$
\begin{aligned}
P_D &= \int_{\Gamma}^{\infty} p_{L|H_1}(\ell \mid H_1)\, d\ell && (8.46)\\
&= \int_{\Gamma}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\ell-E)^2}{2\sigma^2}}\, d\ell
\end{aligned}
$$

$$P_F = \int_\Gamma^\infty p_{L|H_0}(\ell \mid H_0)\, d\ell \tag{8.47}$$

$$= \int_\Gamma^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\ell^2}{2\sigma^2}}\, d\ell$$

These calculations of P_D and P_F are illustrated in Figure 8.7. Since these probabilities depend on the integral of Gaussian densities, we can express them in terms of the standard Q function $Q(x) = \frac{1}{2\pi}\int_x^\infty e^{-z^2/2} dz$ as follows:

$$P_D = Q\left(\frac{\Gamma - E}{\sigma}\right) \tag{8.48}$$

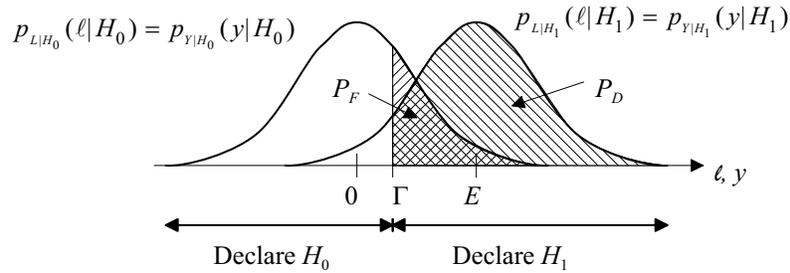$$P_F = Q\left(\frac{\Gamma}{\sigma}\right) \tag{8.49}$$

Figure 8.7: Illustration of P_D and P_F calculation.

Note that for this Gaussian detection example the performance of the detection rule really only depends on the separation of the means of the test statistic $\ell(y)$ under each hypothesis relative to the variance of the test statistic under each hypothesis – i.e. the normalized "distance" between the conditional densities. This relative or normalized distance is often an important indicator of the difficulty of a detection problem. As a result, this idea has been formalized in the definition of the so called "d^2 statistic":

$$d^2 \equiv \frac{\left(E\left[\ell \mid H_1\right] - E\left[\ell \mid H_0\right]\right)^2}{\sqrt{\mathrm{Var}\left(\ell \mid H_1\right)\mathrm{Var}\left(\ell \mid H_0\right)}} \tag{8.50}$$

The quantity d^2 can be seen to be a measure of the normalized distance between two hypotheses. In general, larger values of d^2 correspond to easier detection problems.

Example 8.9 (Scalar Gaussian Detection)
Let us continue Example 8.8. Note that:

$$d^2 = \frac{E^2}{\sigma^2} \tag{8.51}$$

which is a measure of the relative separation of the means under each hypothesis. Further we can express P_D and P_F in terms of d as follows:

$$P_F = Q\left(\frac{\Gamma}{\sigma}\right) \qquad P_D = Q\left(\frac{\Gamma}{\sigma} - d\right) \tag{8.52}$$

Larger values of d result in higher values of P_D for a given value of P_F.

8.2.1   Properties of the ROC

If we examine the expressions for P_D and P_F for Example 8.8 in more detail we can see that the corresponding ROC will possess a number of properties. First, $P_D \geq P_F$ for all thresholds Γ or η. In addition,

$\lim_{\Gamma \to -\infty} P_D = \lim_{\Gamma \to -\infty} = 1$. At the other extreme, $\lim_{\Gamma \to +\infty} P_D = \lim_{\Gamma \to +\infty} P_F = 0$. Finally, $P_D \leq 1$ and $P_F \leq 1$. Thus, the sketch in Figure 8.6 reasonably reflects this ROC. More interestingly, these properties (and others) are true for general ROC curves, and not just for the present example. We discuss these properties of the ROC next, starting with those we have just seen for our Gaussian example. We consider general likelihood ratio tests with threshold η as given in (8.11).

Property 1. The points $(P_F, P_D) = (0, 0)$ and $(P_F, P_D) = (1, 1)$ are always on the ROC. To see this, suppose we set the threshold $\eta = 0$. In this case since the densities are non-negative, the decision rule will always select H_1. In this case, $P_D = P_F = 1$. At the other extreme, assume the threshold $\eta = +\infty$. In this case the hypothesis H_0 is always selected[1]. Since H_0 is always selected $P_F = 0$ and $P_D = 0$.

Property 2. The ROC is the boundary between what is achievable by *any* decision rule and what is not. In particular, the (P_F, P_D) curve of any detection rule (including detection rules that are not LRTs) cannot lie in the shaded region shown in Figure 8.8.

Now, it is straightforward to see that we cannot get better P_D for a given P_F than that achieved by the LRT for the problem, since that would imply a detection rule resulting in lower Bayes risk (which would contradict our finding that the optimal Bayes risk decision rule is a LRT). What is perhaps less immediately obvious is that no decision rule can perform *worse* than the performance corresponding to the "reflection" of the ROC below the 45 degree line. The detector with this maximally bad performance is obtained by simply switching the decision regions for each value of η (and thus is doing the worst thing to do for every threshold). The reason is simple – if it were possible to design a decision rule with arbitrarily bad performance, than by just exchanging the decision regions we could obtain a decision rule with arbitrarily good performance. Note that the result of swapping the decision regions is that $P_D \Rightarrow 1 - P_D$ and $P_F \Rightarrow 1 - P_F$.
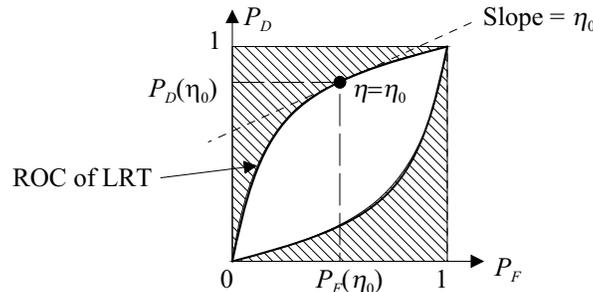

Figure 8.8: Illustration ROC properties.

Property 3. For a LRT with threshold η, the slope of the (continuous) ROC at the corresponding $(P_F(\eta), P_D(\eta))$ point is η.

To show this, first note that we may express P_D as follows:

$$P_D = \int_{\{y | \mathcal{L}(y) > \eta\}} p_{Y|H}(y \mid H_1) \, dy = \int_{\{y | \mathcal{L}(y) > \eta\}} \mathcal{L}(y) \, p_{Y|H}(y \mid H_0) \, dy \qquad (8.53)$$

$$= \int_{\eta}^{\infty} Z \, p_{\mathcal{L}|H_0}(Z \mid H_0) \, dZ \qquad (8.54)$$

Now, differentiating (8.54) with respect to η we obtain:

$$\frac{dP_D(\eta)}{d\eta} = -\eta \, p_{\mathcal{L}|H_0}(\eta \mid H_0) \qquad (8.55)$$

---

[1]Note that the only way that H_1 would be selected is if we had an observation such that $p_{Y|H}(y \mid H_0) = 0$. However, for such observations there is no possibility of a false alarm, since those value cannot be generated under H_0!

Now we also know that

$$P_D = \int_\eta^\infty p_{\mathcal{L}|H_1}(\mathcal{L} \mid H_1) \, d\mathcal{L} \tag{8.56}$$

$$P_F = \int_\eta^\infty p_{\mathcal{L}|H_0}(\mathcal{L} \mid H_0) \, d\mathcal{L} \tag{8.57}$$

Differentiating these expressions with respect to η we also obtain:

$$\frac{dP_D}{d\eta} = -p_{\mathcal{L}|H_1}(\eta \mid H_1) \tag{8.58}$$

$$\frac{dP_F}{d\eta} = -p_{\mathcal{L}|H_0}(\eta \mid H_0) \tag{8.59}$$

Now equating (8.55) to (8.58) we obtain the result that:

$$\frac{p_{\mathcal{L}|H_1}(\eta \mid H_1)}{p_{\mathcal{L}|H_0}(\eta \mid H_0)} = \eta \tag{8.60}$$

Finally, the slope of the ROC is given by the derivative of P_D with respect to P_F:

$$\frac{dP_D}{dP_F} = \frac{\frac{dP_D}{d\eta}}{\frac{dP_F}{d\eta}} = \frac{-p_{\mathcal{L}|H_1}(\eta \mid H_1)}{-p_{\mathcal{L}|H_0}(\eta \mid H_0)} = \eta \tag{8.61}$$

which shows the result.

This property is illustrated in Figure 8.8. Note that a consequence of this property is that the ROC has zero slope at the point $(P_F, P_D) = (1,1)$ ($\eta = 0$) and infinite slope at the point $(P_F, P_D) = (0,0)$ ($\eta = \infty$).

Property 4. The ROC of the LRT is convex downward. In particular, $P_D \geq P_F$.

To show this property we use the concept of randomized decision rules, discussed in the following section on detection from discrete-valued observations. Suppose we select the endpoints of a randomized decision rule to be on the optimal ROC itself, as illustrated in Figure 8.9. Note that such a randomized decision rule is not necessary optimal. As a result, the optimal test must have performance (i.e. P_D for a given P_F) that is better than any randomized test. In particular, if (P_F^*, P_D^*) are the points on the ROC for the optimal Bayes decision rule, then we must have:

$$P_D^* \geq P_D(p) \quad \text{when} \quad P_F^* = P_F(p) \tag{8.62}$$

This argument shows that points on the optimal ROC between our chosen endpoints must lie above the line connecting the endpoints, and thus that the optimal ROC is convex, as shown in Figure 8.9

To see how the ROC can be used to compare the performance of different problems and detection rules, consider the following example, where we examine how the ROC changes as a function of the amount of data.

Example 8.10

Suppose we observe N independent samples of a random variable: y_i, $i = 1, \cdots, N$. Under hypothesis H_0, $p_{Y_i|H_0}(y_i \mid H_0) \sim N(0, \sigma^2)$, and under H_1, $p_{Y_i|H_0}(y_i \mid H_1) \sim N(1, \sigma^2)$. Define the vector $\underline{y}$ to be the collection of samples. Our problem is to decide whether our vector of observations came from the H_0 distribution or the H_1 distribution. This problem is similar to the N-pulse radar detection problem of Example 8.5. Using our analysis there we find that the optimal test can be written as:

$$\ell(\underline{y}) = \frac{1}{N} \sum_{i=1}^N y_i \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{2} + \frac{\sigma^2 \ln(\eta)}{N} \tag{8.63}$$
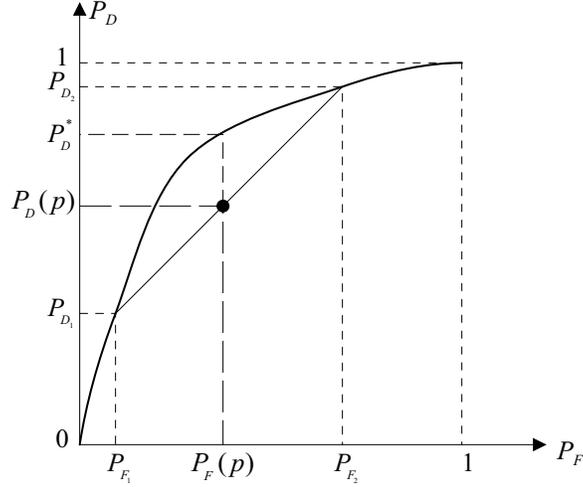
Figure 8.9: Illustration ROC convexity using randomized decision rules.

Now note that since the observations y_i are independent, the sufficient statistic for the test $\ell(\underline{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i$ has the following probability density functions under each hypothesis:

$$H_0 \quad : \quad p_{L|H_0}(\ell \mid H_0) \sim N(0, \sigma^2/N) \tag{8.64}$$

$$H_1 \quad : \quad p_{L|H_1}(\ell \mid H_1) \sim N(1, \sigma^2/N) \tag{8.65}$$

Thus, the probability of false alarm for a given threshold $\Gamma = 1/2 + \frac{\sigma^2 \ln(\eta)}{N}$ is given by

$$P_F = \frac{1}{\sqrt{2\pi\sigma^2/n}} \int_{\Gamma}^{\infty} e^{-\frac{Nx^2}{2\sigma^2}} \, dx = Q\left(\frac{N^{1/2}\Gamma}{\sigma}\right) \tag{8.66}$$

where $Q(\Gamma) = \frac{1}{\sqrt{2\pi}} \int_{\Gamma}^{\infty} e^{-x^2/2} dx$. Similarly,

$$P_D = \frac{1}{\sqrt{2\pi\sigma^2/N}} \int_{\Gamma}^{\infty} e^{-\frac{N(x-1)^2}{2\sigma^2}} \, dx = Q\left(\frac{N^{1/2}(\Gamma-1)}{\sigma}\right) \tag{8.67}$$

Note that, as N increases the ROC curves are monotonically increasing in P_D for the same P_F, and thus nest. In particular, as we make more independent observations the curves move to the northwest and closer to their bounding box. In the limit, we have $\lim_{N\to\infty} P_D = 1, \lim_{N\to\infty} P_F = 0$, which indicates that, as $N \to \infty$. This effect is shown in Figure 8.10. Simply looking at the ROC curves for the different cases we can see the positive effect of using more observations.

Finally, note that the idea of using the ROC to evaluate the performance of decision rules is so powerful and pervasive that it is used to evaluate decision rules even when they are not, strictly speaking LRT rules for binary hypothesis testing problems.

8.2.2   Detection Based on Discrete-Valued Random Variables

The theory behind detection based on observations y that are discrete valued is essentially the same as when y is continuous valued. In particular, the LRT (8.11) is still the optimal decision rule, as considered in Example 8.7. There are some important unique characteristics of the discrete valued case that are worth discussing, however. When the observations y are discrete-valued the likelihood ratio $\mathcal{L}(y)$ will also be discrete-valued. In this case, varying the threshold η will have no effect on the values of P_F, P_D until the threshold crosses one of the discrete values of $\mathcal{L}(y)$. After crossing this discrete-value, the values of P_F, P_D will then change by a finite amount. As a result, the ROC "curve" in such a discrete observation case, obtained by varying the value of the threshold, will be a series of disconnected and isolated points. This is illustrated in the following examples.
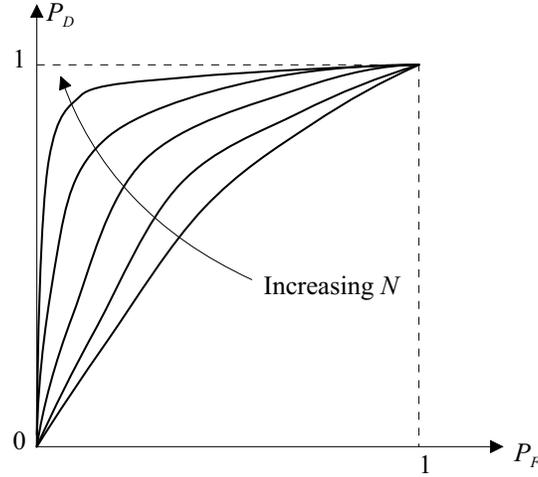
Figure 8.10: Illustration ROC behavior as we obtain more independent observations.

Example 8.11

Assume that y is a binomial random variable, resulting from the sum of two independent, identically distributed Bernoulli random variables:

$$y = x_1 + x_2 \tag{8.68}$$

The probabilities of the x_i under each hypothesis are given by:

$$\text{Under } H_0: \qquad \Pr(x_i = 1) = \frac{1}{4}; \Pr(x_i = 0) = \frac{3}{4}; \tag{8.69}$$

$$\text{Under } H_1: \qquad \Pr(x_i = 1) = \frac{1}{2}; \Pr(x_i = 0) = \frac{1}{2}; \tag{8.70}$$

Note that y can only take 3 values: 0, 1, or 2. Under these conditions, the likelihood ratio for the problem is given by:

$$\mathcal{L}(y) \quad = \quad \frac{p_{Y|H}(y \mid H_1)}{p_{Y|H}(y \mid H_0)} = \frac{\frac{2!}{y!(2-y)!}\left(\frac{1}{2}\right)^y\left(\frac{1}{2}\right)^{2-y}}{\frac{2!}{y!(2-y)!}\left(\frac{1}{4}\right)^y\left(\frac{3}{4}\right)^{2-y}} = \frac{1/4}{(1/4)^y(3/4)^{2-y}} \tag{8.71}$$

$$= \quad \frac{4}{3^{2-y}} \tag{8.72}$$

The LRT for this problem is then given by:

$$\frac{4}{3^{2-y}} \quad \underset{H_0}{\overset{H_1}{\gtrless}} \quad \eta \tag{8.73}$$

Now note that the likelihood ratio can only take the values:

$$\mathcal{L}(y) \quad = \quad \begin{cases} 4/9 & \text{if } y = 0 \\ 4/3 & \text{if } y = 1 \\ 4 & \text{if } y = 2 \end{cases} \tag{8.74}$$

Now let us examine how P_D and P_F vary as we change η. If $\eta > 4$, hypothesis H_0 is always selected so that $P_D = 0$ and $P_F = 0$. Thus, these values of η correspond to the point $(P_F, P_D) = (0,0)$ on the ROC. As η is reduced so that $4/3 < \eta < 4$, hypothesis H_1 is selected only when $y = 2$. The probability of detection is $P_D = P(y = 2 \mid H_1) = 1/4$, whereas the probability of false alarm is $P_F = P(y = 2 \mid H_0) = 1/16$. Note that P_D and P_F will have these values for *any* value of η in the range $4/3 < \eta < 4$. Thus, this entire range of η corresponds to the (isolated) point $(P_F, P_D) = (1/16, 1/4)$ on the ROC. Further reducing the threshold η so that $4/9 < \eta < 4/3$ implies that H_0 is selected only when $y = 0$. In this case, the probability of detection is $P_D = 1 - P(y = 0 \mid H_1) = 3/4$, and the probability of false alarm is $1 - P(y = 0 \mid H_0) = 7/16$. Again, note that P_D and P_F will have these values for *any* value of η in the range $4/9 < \eta < 4/3$. Again, this entire range of η thus corresponds to the (isolated) point $(P_F, P_D) = (7/16, 3/4)$ on the
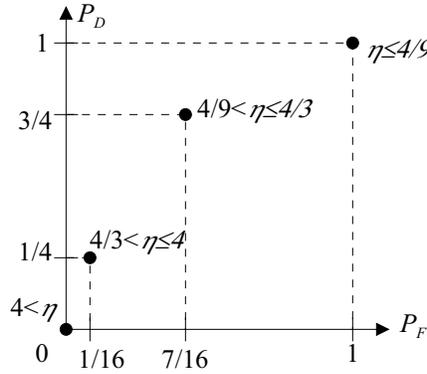
Figure 8.11: Illustration ROC for a discrete valued problem of Example 8.11.

ROC. Finally, as the threshold is lowered so that $\eta < 4/9$, hypothesis H_0 is never selected, so that $P_D = 1$ and $P_F = 1$. These values of η therefore correspond to the point $(P_F, P_D) = (1, 1)$ on the ROC. In summary, varying the threshold η produces 4 isolated points for the ROC curve for this problem, as illustrated in Figure 8.11

Let us consider another discrete valued example, this time involving Poisson random variables.

Example 8.12
Consider observing a scalar value y, which is Poisson distributed under H_0 with mean m_0, and Poisson distributed under H_1 with mean m_1. This situation was considered in Example 8.7. The optimal decision rule for this problem was found in (8.34 to be given by:

$$y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{(m_1 - m_0) + \ln(\eta)}{\ln\left(\frac{m_1}{m_0}\right)} = \Gamma \tag{8.75}$$

Since y is discrete-valued, fractional parts of the effective threshold Γ on the right hand side of (8.75) will have no effect, and the ROC will again have a countable number of points.

The probability of false alarm is thus a function of the integer part of the threshold Γ, and is given by:

$$P_F(\Gamma) = \sum_{y=\lceil \Gamma \rceil}^{\infty} \frac{m_0^y}{y!} e^{-m_0} \tag{8.76}$$

where $\lceil \Gamma \rceil$ denotes the smallest integer greater than Γ. Similarly, the probability of detection is given by:

$$P_D(\Gamma) = \sum_{y=\lceil \Gamma \rceil}^{\infty} \frac{m_1^y}{y!} e^{-m_1} \tag{8.77}$$

The ROC for this problem is illustrated in Figure 8.12

The discrete nature of the ROC when the observation is discrete-valued seems to suggest that we can only obtain detection performance at a finite number of (P_F, P_D) pairs. While this observation is true if we limit ourselves to deterministic decision rules, by introducing the concept of a *randomized decision rule* we can get a much wider set of detection performance points (i.e. (P_F, P_D) points).

To introduce the idea of a randomized decision rule, suppose we have a likelihood ratio $\mathcal{L}(y)$ for an arbitrary problem (i.e. not necessarily with discrete-valued observations) and two thresholds η_0 and η_1. We then essentially have two likelihood ratio decision rules. Assume the decision rule corresponding to η_0 has performance (P_{F_0}, P_{D_0}) and the decision rule corresponding to η_1 has performance (P_{F_1}, P_{D_1}). Suppose we now define a new (random) decision rule by deciding between H_0 and H_1 according to the following probabilistic scheme:

1. Select a Bernoulli random variable Z with $\Pr(Z = 1) = p$. This is equivalent to flipping a biased coin with $\Pr(\text{heads}) = p$.
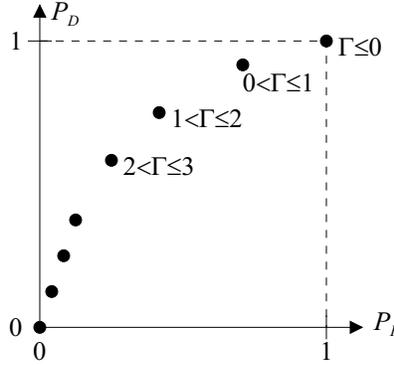
Figure 8.12: Illustration ROC for a discrete valued problem of Example 8.12.

2. If $Z = 1$ use a LRT with the threshold $\eta = \eta_1$ to make the decision.

$$\mathcal{L}(y) \underset{H_0}{\overset{H_1}{\gtrless}} \eta_1 \tag{8.78}$$

If $Z = 0$ use a LRT with the threshold $\eta = \eta_0$ to make the decision.

$$\mathcal{L}(y) \underset{H_0}{\overset{H_1}{\gtrless}} \eta_0 \tag{8.79}$$

Note that the resulting overall rule will result in a random decision. The $P_D(p)$, $P_F(p)$ performance of the overall new detection rule as a function of p can be found as:

$$
\begin{aligned}
P_D(p) &= \Pr(\text{Decide } H_1 \mid H_1) \\
&= \Pr(\text{Decide } H_1 \mid H_1, Z = 1)\Pr(Z = 1) + \Pr(\text{Decide } H_1 \mid H_1, Z = 0)\Pr(Z = 0) \\
&= pP_{D_1} + (1 - p)P_{D_0}
\end{aligned}
\tag{8.80}
$$

$$
\begin{aligned}
P_F(p) &= \Pr(\text{Decide } H_1 \mid H_0) \\
&= \Pr(\text{Decide } H_1 \mid H_0, Z = 1)\Pr(Z = 1) + \Pr(\text{Decide } H_1 \mid H_0, Z = 0)\Pr(Z = 0) \\
&= pP_{F_1} + (1 - p)P_{F_0}
\end{aligned}
\tag{8.81}
$$

Thus, the performance of the randomized decision rule is on the line connecting the points (P_{F_1}, P_{D_1}) and (P_{F_0}, P_{D_0}). These ideas are illustrated for a generic decision problem in Figure 8.13. By varying p we can obtain a decision rule with performance given by any (P_F, P_D) pair on the line connecting the points (P_{F_1}, P_{D_1}) and (P_{F_0}, P_{D_0}).

Now, let us return to the discrete-valued observation case. By using such randomized decision rules with the isolated points of the ROC of the deterministic decision rule as endpoints, we can obtain any (P_F, P_D) performance on the lines connecting these points. For example, the resulting ROC for Example 8.11 would be as shown in Figure 8.14. In general, we can obtain an ROC curve which is a piecewise-linear concave curve connecting the isolated points of the deterministic decision rule ROC. Further, c.f. ROC Property 2, it is impossible to get performance that is above this piecewise-linear curve (or below its mirror image).

Finally, note that ROC Property 3 can also be extended to discrete-valued random variables. Note that in this case the ROC curve is not differentiable at the discrete-valued points so the slope of the ROC curve is not defined at these points. At such points of non-differentiability, there is a range of possible slopes, defined by the slopes of the straight lines to the right and to the left of the isolated points. At these points, the value of η must be included in this range of possible slopes.
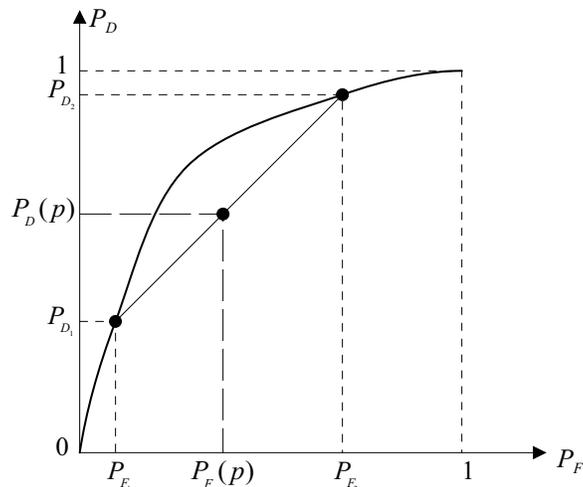
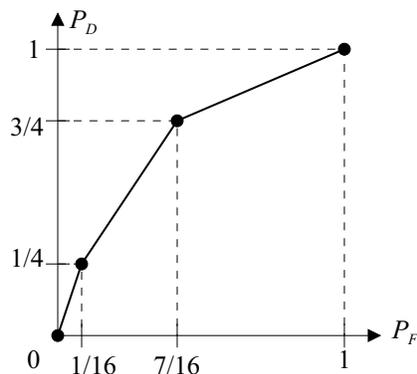Figure 8.13: Illustration of the performance of a randomized decision rule.

Figure 8.14: Illustration of the overall ROC obtained for a discrete valued observation problem using randomized rules.

8.3 Other Threshold Strategies

We have now determined that the form of the optimal Bayes risk test is the likelihood ratio test and have studied the performance of decision rules through use of the ROC. We have seen that the ROC compactly represents the performance of the LRT for all choices of the threshold η. In the general Bayes formulation the specific threshold η used for a given detection problem (and thus the specific operating point chosen on the ROC) is a function of the prior probabilities $P_i = \Pr(H_i)$ and the cost assignment C_{ij}:

$$\eta \equiv \frac{(C_{10} - C_{00}) \, P_0}{(C_{01} - C_{11}) \, P_1} \tag{8.82}$$

If we have knowledge of all these elements, then this is obviously the right (and easy) thing to do. Often, however, determining either the P_i or the C_{ij} is fraught with difficulties and an alternative strategy for picking the operating point is used. We discuss two such alternatives next.

8.3.1   Minimax Hypothesis Testing

For a given detection problem suppose that we have a cost assignment C_{ij} we believe in, but are unsure of the true prior probabilities used by nature, which are P_1^* and P_0^*. Now suppose we design a decision rule (i.e., choose a threshold) based on the costs C_{ij} and a set of *assumed* (but possibility incorrect) prior probabilities P_1 and $P_0 = 1 - P_1$. Let the performance of the resulting decision rule be given by the operating point $(P_F(P_1), P_D(P_1))$, which, as we have indicated, will be a function of our choice of P_1. Since, in general, the P_i we use to design our decision rule will be different from the true underlying P_i^*, our test will not have the minimum cost or Bayes risk for this problem. One reasonable approach in such a situation is to assume that nature will do the worst thing possible and to choose our design values of P_i (i.e. choose our threshold η) to minimize the maximum value of the cost or Bayes risk as a function of the true values P_i^*. Such a strategy leads to the *minimax decision rule.*

Now, from (8.36), the resulting cost (i.e. the Bayes risk) of a decision rule using assumed values P_i when truth is P_i^* is given by:

$$
\begin{aligned}
E\left(C, P_1, P_1^*\right) &= C_{00}P_0^* + C_{01}P_1^* + (C_{10} - C_{00})P_0^*P_F(P_1) - (C_{01} - C_{11})P_1^*P_D(P_1) &\text{(8.83)}\\
&= \left[(C_{01} - C_{00}) - (C_{10} - C_{00})P_F - (C_{01} - C_{11})P_D\right]P_1^* + C_{00} + (C_{10} - C_{00})P_F
\end{aligned}
$$

where we have used the fact that $P_0^* = (1 - P_1^*)$.

On the left in Figure 8.15 we illustrate how the expected cost changes as the true prior probability P_1^* is varied. When an arbitrary fixed value of P_1 is used, the threshold is fixed, so the corresponding values of P_F and P_D are fixed. In this case we see from (8.83) that $E(C)$ will be a linear function of the true prior probability P_1^*. This is plotted as the upper curve in Figure 8.15 (left). Now if we knew P_1^* we could design an optimal LRT using an optimal threshold. In this case the threshold would change as P_1^* varied and thus so would P_F and P_D and the resulting cost. The cost of this optimal decision rule is the lower curve in Figure 8.15 (left). The two curves touch when the design value of P_1 matches the true value of P_1^*. Thus, they will always be tangent at this matched point. For the example in the figure, the maximum value of the expected cost for the non-optimal rule is obtained at the left endpoint of the curve.
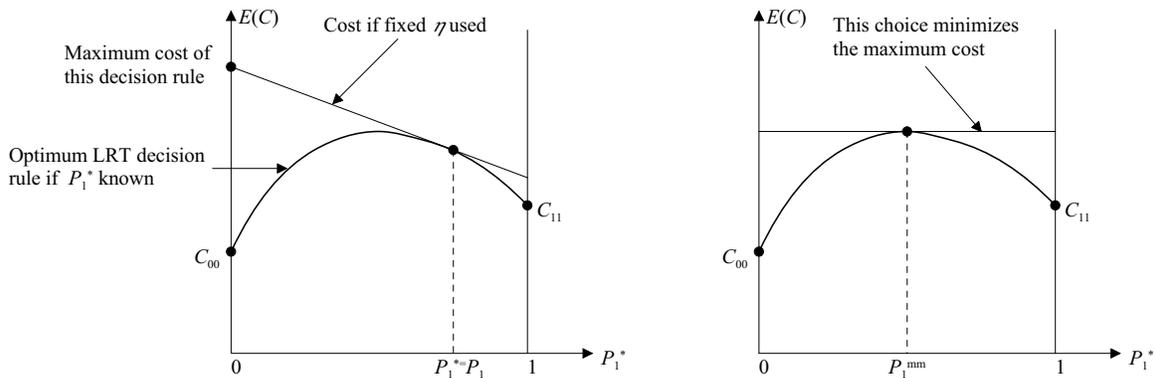

Figure 8.15: Left: Illustration of the expected cost of a decision rule using an arbitrary fixed threshold as a function of the true prior probability P_1^*. The maximum cost of this decision rule is at the left endpoint. The lower curve is the corresponding expected cost of the optimal LRT. Right: The expected cost of the minimax decision rule as a function of the true prior probability P_1^*.

In general, we would like to minimize the maximum value of (8.83). Examining Figure 8.15, we can see that this goal is accomplished if we choose our value of P_1 (or equivalently, our operating point on the ROC) so that the line (8.83) is tangent to the optimal Bayes risk curve at its maximum, as shown on the right in the figure. This happens when the slope of the curve is zero, i.e. when:

$$
\left[(C_{01} - C_{00}) - (C_{10} - C_{00})P_F - (C_{01} - C_{11})P_D\right] = 0 \tag{8.84}
$$

This result is valid as long as the maximum of the optimal Bayes cost curve is interior to the interval. When the maximum is at the boundary of the interval, then that is value of P_1 to choose.

Equation (8.84) is sometimes termed the *minimax equation* and defines the general minimax operating point. We can rewrite (8.84) in the following form:

$$P_D = \left(\frac{C_{01} - C_{00}}{C_{01} - C_{11}}\right) - \left(\frac{C_{10} - C_{00}}{C_{01} - C_{11}}\right) P_F \qquad (8.85)$$

which is just a line in (P_F, P_D) space. Thus the minimax choice of operating point can be found as the intersection of the straight line (8.85) with the ROC for the optimal LRT, as shown in Figure 8.16. For example, if we use the MPE cost assignment, $C_{01} = C_{10} = 1$, $C_{00} = C_{11} = 0$, then (8.85) reduces to $P_D = 1 - P_F$ and the minimax line is just the -45 degree line.
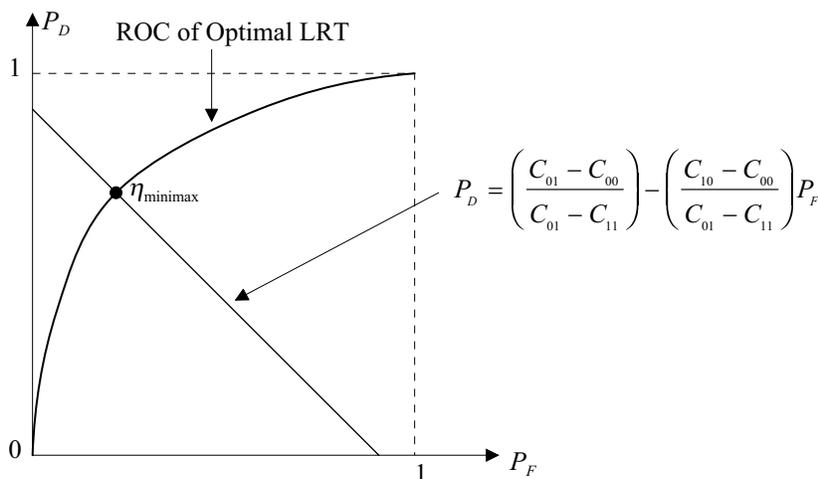

Figure 8.16: Finding the minimax operating point by intersecting (8.85) with the ROC for the optimal LRT.

8.3.2 Neyman-Pearson Hypothesis Testing

In the minimax case we assume that the costs C_{ij} can be meaningfully assigned, but that we do not know the prior probabilities. In many cases, finding such meaningful costs assignments can be difficult. This raises the question of how to choose an operating point when *neither* the prior probabilities P_i or the costs C_{ij} can be found. In general, we would like to make P_F as small as possible and P_D as large as possible. As the ROC shows, these two desires work in opposition to each other. What is often done is practice is to constrain P_F and then to maximize P_D subject to this constraint. Mathematically, one wants to solve:

$$\max P_D \quad \text{subject to } P_F \leq \alpha \qquad (8.86)$$

The solution of this problem is called a *Neyman-Pearson detection rule* or "NP rule".

Note that the optimal Bayes LRT has the highest P_D for any P_F, and thus the solution of the Neyman-Pearson problem must be an optimal LRT for some choice of threshold η. So we are again in the position of needing to find an appropriate operating point on the optimal ROC. Since the ROC of the optimal LRT has P_D as a monotonically non-decreasing function of P_F, the solution of the NP problem must correspond to the point $(\alpha, P_D(\alpha))$. In the continuous-observation case, the corresponding optimal threshold η is then the slope of the ROC at this point. When the observations are discrete, we can use randomized decision rules to obtain the best P_D for any $P_F = \alpha$ and the corresponding threshold η can be found from the thresholds of the endpoint. Indeed, the desire to perform NP decision rules is one motivation for randomized decision rules in the discrete case!

Example 8.13 (Neyman-Pearson)
Suppose that the likelihoods under each hypothesis for a binary detection problem are as given in Figure 8.17. We want to find the decision rule that maximizes P_D subject to $P_F \leq 1/2$.
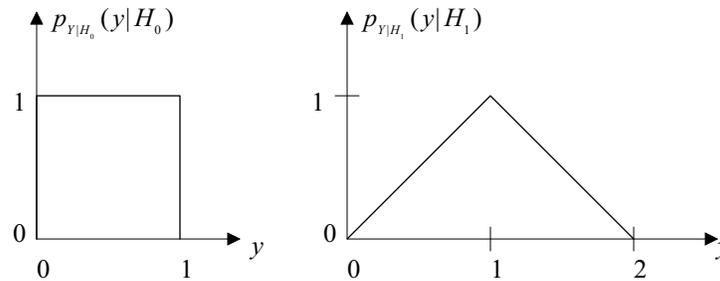

Figure 8.17: Likelihoods for a Neyman-Pearson problem.

This decision rule will be a Neyman-Pearson rule. The observation is continuous valued, so the ROC will be as well. Thus the optimal NP rule will be a LRT with threshold η chosen so that $P_F = 1/2$. We can write this rule as follows:

$$p_{Y|H}(y \mid H_1) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \, p_{Y|H}(y \mid H_0) \tag{8.87}$$

Figure 8.18 shows $p_{Y|H}(y \mid H_1)$ and $\eta \, p_{Y|H}(y \mid H_0)$ on the same axes when $\eta < 1$. The corresponding decision regions are also shown. On the right of Figure 8.18 the corresponding value of $P_F = (1 - \eta)$ is shown. Now we want $P_F = 1/2$, thus we have:

$$\eta = 1 - 1/2 = 1/2 \tag{8.88}$$

The resulting decision rule is given by:

$$\frac{p_{Y|H}(y \mid H_1)}{p_{Y|H}(y \mid H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{2} \tag{8.89}$$
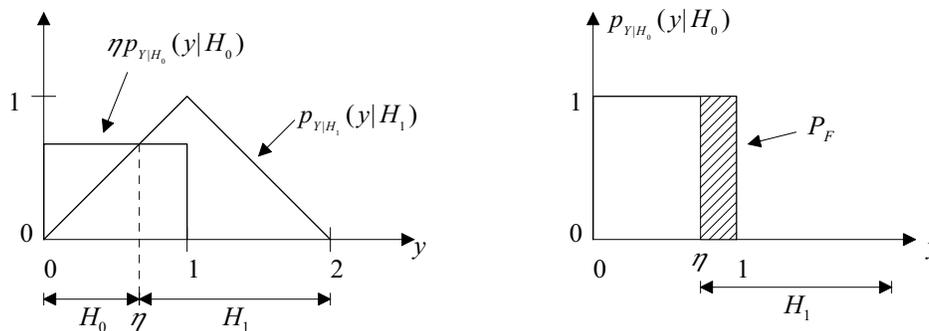
Figure 8.18: Scaled densities, decision regions and P_F for the problem of Example 8.13.

In practical problems, the bound α on P_F is determined by engineering considerations, and includes such constraints as the amount of computing power or other resources available to process false alarms. For example, a common situation we have all experienced relating to false alarm rate is in connection with car alarms. If the threshold of the car alarm is set too high, it will not trigger when the car is assaulted by thieves. On the other hand, if the threshold is set too low, the alarm will often go off even when no thief is present – creating a false alarm. If too many false alarms are generated people become exhausted and cease to check them out.

8.4 M-ary Hypothesis Testing

The exposition so far has focused on binary hypothesis testing problems. When there are M possibilities or hypotheses, we term the problem an *M-ary detection or hypothesis testing problem*. We can again take a minimum Bayes risk approach, with the same 3 three problem elements we had in the binary case:

1. **Model of Nature:** In the M-ary case there are M possibilities, denoted as H_i, $i = 0, \cdots, M-1$. Our knowledge of these possibilities is captured by the prior probabilities $P_i = \Pr(H = H_i)$, $i = 0, \cdots, M-1$. Note that $\sum_i P_i = 1$.

2. **Observation Model:** This relationship is given in the M-ary case by the M conditional densities $p_{Y|H}(y \mid h_i)$.

3. **Decision Rule:** Our decision rule $D(y)$ will again be obtained by minimizing the average cost or Bayes risk. Again, C_{ij} denotes the cost of deciding hypothesis $D(y) = H_i$ when hypothesis H_j is true and the Bayes risk is given by $E\left[C_{D(y),H}\right]$.

Note that in the M-ary case, the decision rule $D(y)$ is nothing more than a labeling of each point in the observation space with one of the corresponding possible decision outcomes H_i.

In an identical argument to the binary case, we have that the expected value of the cost is given by:

$$E\left[C_{D(y)}\right] \quad = \quad \int E\left[C_{D(y)} \mid y\right] p_Y(y)\,dy \tag{8.90}$$

and as before the expression is minimized by minimizing $E\left[C_{D(y)} \mid y\right]$. In particular, we should choose the decision resulting in the smallest value of this quantity. Now the expected cost of deciding H_k given y is:

$$E\left[C_{D(y)=H_k} \mid y\right] \quad = \quad \sum_{j=0}^{M-1} C_{kj} p_{H|Y}\left(H_j \mid y\right) \tag{8.91}$$

Thus the optimal decision rule is to choose hypothesis H_k given the observation y if:

$$\sum_{j=0}^{M-1} C_{kj} p_{H|Y}\left(H_j \mid y\right) \leq \sum_{j=0}^{M-1} C_{ij} p_{H|Y}\left(H_j \mid y\right) \quad \forall i \tag{8.92}$$

The left hand side of (8.92) is the conditional cost of assigning y to the H_k decision region and the right hand side of (8.92) is the conditional cost of assigning y to the H_i decision region. Note that if the left hand side is the smallest, than assigning the given observation y to H_k is the best thing to do. Unlike the binary case, however, if the left hand side is not the smallest, we do not immediately know what the optimal hypothesis assignment is. All we know is that it is not H_k. Using this insight we can recast (8.92) in the following form, which is similar in spirit to (8.6):

$$\sum_{j=0}^{M-1} C_{kj} p_{H|Y}\left(H_j \mid y\right) \mathop{\gtrless}_{\text{Not } H_i}^{\text{Not } H_k} \sum_{j=0}^{M-1} C_{ij} p_{H|Y}\left(H_j \mid y\right) \quad \forall\,\text{unique } i, k \text{ pairs} \tag{8.93}$$

where $\mathop{\gtrless}\limits_{\text{Not } H_i}^{\text{Not } H_k}$ denotes eliminating hypothesis H_k if the inequality is $>$ and eliminating hypothesis H_i if the inequality is $<$. In the binary case there is only one comparison needed to define the optimal decision rule. In contrast, in the M-ary case, we need $\frac{M(M-1)}{2}$ comparisons to define the optimal decision rule. Each such comparison eliminates one of the hypotheses.

We can make (8.93) more similar to the binary case through some manipulations. In analogy with (8.11), let us define the following set of likelihood ratios:

$$\mathcal{L}_j(y) = \frac{p_{Y|H}(y \mid H_j)}{p_{Y|H}(y \mid H_0)} \quad j = 0, \cdots, M-1 \tag{8.94}$$

where we take $\mathcal{L}_0(y) = 1$. Then, combining these likelihood ratios with Bayes rule (8.7) we have the following form for the optimal Bayes M-ary decision rule:

$$\sum_{j=0}^{M-1} C_{kj} P_j \mathcal{L}_j(y) \underset{\text{Not } H_i}{\overset{\text{Not } H_k}{\gtrless}} \sum_{j=0}^{M-1} C_{ij} P_j \mathcal{L}_j(y) \quad \forall \text{ unique } i, k \text{ pairs} \tag{8.95}$$

Note that quantities $\mathcal{L}_j(y)$ form a set of sufficient statistics for the M-ary detection problem. Further, this set of inequalities defines $M(M-1)/2$ linear decision boundaries in the space of the sufficient statistics $\mathcal{L}_i(y)$.

For example, consider the three-hypothesis case where $M = 3$. In this case, there are three comparisons that need to be performed:

$$k=0, i=1: \quad P_1 \left(C_{01} - C_{11} \right) \mathcal{L}_1(y) \underset{\text{Not } H_1}{\overset{\text{Not } H_0}{\gtrless}} P_0 \left(C_{10} - C_{00} \right) + P_2 \left(C_{12} - C_{02} \right) \mathcal{L}_2(y) \tag{8.96}$$

$$k=1, i=2: \quad P_1 \left(C_{11} - C_{21} \right) \mathcal{L}_1(y) \underset{\text{Not } H_2}{\overset{\text{Not } H_1}{\gtrless}} P_0 \left(C_{20} - C_{10} \right) + P_2 \left(C_{22} - C_{12} \right) \mathcal{L}_2(y) \tag{8.97}$$

$$k=2, i=0: \quad P_1 \left(C_{21} - C_{01} \right) \mathcal{L}_1(y) \underset{\text{Not } H_0}{\overset{\text{Not } H_2}{\gtrless}} P_0 \left(C_{00} - C_{20} \right) + P_2 \left(C_{02} - C_{22} \right) \mathcal{L}_2(y) \tag{8.98}$$

These comparisons are shown for a generic case in Figure 8.19. Each comparison eliminates one hypothesis. Taken together the set of comparisons labels the space of the test statistics. Note that in the space of the likelihood ratio test statistics the decision regions are always linear. In the space of the observations y this will not be true, in general. Further, the dimension of the "likelihood space" is dependent on number of hypotheses, not the dimension of the observation, which may be greater than or less than the likelihood dimension.

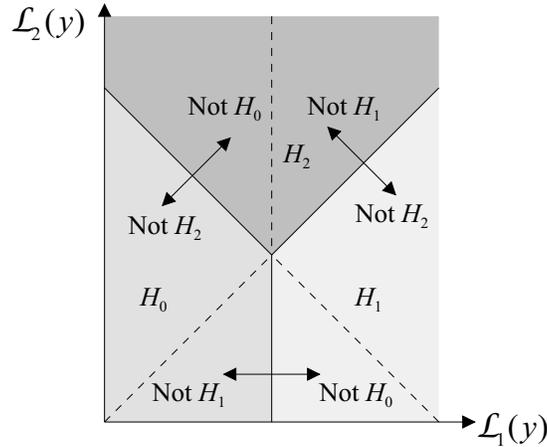Figure 8.19: Decision boundaries in the space of the likelihoods for an M-ary problem.

8.4.1   Special Cases

Let us now consider some common special cases of the Bayes risk and the associated decision rules corresponding to them for the M-ary case.

MPE cost assignment and the MAP rule

Suppose we use the following "zero-one" cost assignment for an M-ary problem:

$$C_{ij} = 1 - \delta_{ij} \tag{8.99}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. Then the cost of all errors ($C_{10} = C_{01} = 1$) are the same and there is no cost for correct decisions ($C_{00} = C_{11} = 0$). As in the binary case, this cost assignment results in the Bayes risk also equaling the probability or error:

$$E\left[C_{D(y)}\right] \;=\; \sum_{j=0}^{M-1} \sum_{\substack{i=0 \\ i \neq j}}^{M-1} \Pr\left[\text{Decide } H_i,\, H_j \text{ true}\right] = \Pr\left[\text{Error}\right] \tag{8.100}$$

Thus the optimal decision rule for this cost assignment in the M-ary case also minimizes the probability of error. The corresponding decision rule (again termed the minimum probability of error (MPE) decision rule) is to choose hypothesis H_k given the observation y if:

$$p_{H|Y}\left(H_k \mid y\right) \geq p_{H|Y}\left(H_i \mid y\right) \quad \forall\, i \tag{8.101}$$

This decision rule says that for minimum probability of error choose the hypothesis with the highest posterior probability. As in the binary case, this is termed the Maximum aposteriori probability or MAP rule. So again, the MPE cost assignment results in the MAP rule (independent of prior probabilities).

The MAP decision rule can also be expressed in terms of a series of comparisons of likelihood ratios, as in (8.95). By substituting the MPE cost structure into (8.95) and simplifying we obtain the following equivalent expression of the Bayes optimal M-ary MAP rule:

$$P_i \mathcal{L}_i(y) \;\underset{\text{Not } H_i}{\overset{\text{Not } H_k}{\gtrless}}\; P_k \mathcal{L}_k(y) \quad \forall\, \text{unique } i, k \text{ pairs} \tag{8.102}$$

Note that the details of the densities are hidden in the expressions for the likelihood ratios $\mathcal{L}_i(y)$.

The ML rule

Now suppose we again use the MPE cost criterion with $C_{ij} = 1 - \delta_{ij}$, but also have both hypotheses equally likely apriori so that $P_i = 1/M$. In this case we essentially have no prior preference for one hypothesis over the other. Applying these conditions together with Bayes rule to (8.101), this decision rule is to choose hypothesis H_k given the observation y if:

$$p_{Y|H}\left(y \mid H_k\right) \geq p_{Y|H}\left(y \mid H_i\right) \quad \forall\, i \tag{8.103}$$

In this case the decision rule is to choose the hypothesis that gives the highest likelihood of the observation, which is again the maximum likelihood or ML rule.

As for the MAP rule, the ML decision rule can also be expressed in terms of a series of comparisons of likelihood ratios, as in (8.102). Note that the expression (8.102) already reflects the impact of the MPE cost structure. If we further incorporate the fact that $P_i = P_j$ into (8.102), we obtain the following equivalent expression of the Bayes optimal M-ary ML rule:

$$\mathcal{L}_i(y) \;\underset{\text{Not } H_i}{\overset{\text{Not } H_k}{\gtrless}}\; \mathcal{L}_k(y) \quad \forall\, \text{unique } i, k \text{ pairs} \tag{8.104}$$

8.4.2 Examples

Let us now consider some examples.

Example 8.14 (Known means in White Gaussian Noise)
Suppose we want to detect which of three possible N-dimensional signals is being received in the presence of noise. In particular, suppose that under hypothesis H_k the observation is given by:

$$\text{Under } H_k: \quad \underline{y} = \underline{m}_k + \underline{w} \quad k = 0, 1, 2 \tag{8.105}$$

where $\underline{w} \sim N(\underline{0}, I)$. Note that this implies that the observation densities under the different hypotheses are Gaussian, given by:

$$p_{\underline{Y}|H}\left(\underline{y} \mid H_k\right) = N\left(\underline{y}; \underline{m}_k, I\right) \qquad k = 0, 1, 2 \tag{8.106}$$

Assume that we want a minimum probability of error decision rule, which means we want the cost assignment $C_{ij} = 1 - \delta_{ij}$ and results in the MAP rule (8.101). We can also express this rule in the form (8.95). Substituting the densities given in (8.106) and simplifying, we obtain for the optimal decision rule for this example:

$$\ell_{ik}(\underline{y}) = \underline{y}^T \left(\frac{\underline{m}_k - \underline{m}_i}{\|\underline{m}_k - \underline{m}_i\|} \right) \mathop{\gtrless}_{\text{Not } H_k}^{\text{Not } H_i} \frac{1}{\|\underline{m}_k - \underline{m}_i\|} \left[\frac{\underline{m}_k^T \underline{m}_k - \underline{m}_i^T \underline{m}_i}{2} + \ln\left(\frac{P_i}{P_k} \right) \right] = \Gamma_{ik} \tag{8.107}$$

where we perform the comparisons over all unique i, k pairs.

Note a number of things. First, the set of $\ell_{ik}(\underline{y})$ are a set of sufficient statistics for the problem (as are the set of likelihood ratios $\mathcal{L}_i(\underline{y})$). In addition, the computation of these sufficient statistics (i.e. the processing of the data) consists of projecting the data vector onto the line between the different means and then comparing the result to a threshold. These ideas are illustrated in Figure 8.20 for a two-dimensional case. Note that the dimension of the space of the observation is independent of the number of hypotheses. Further, in this example of Gaussian densities with identical covariance matrices but different means, the decision boundaries of each comparison in (8.107) are lines (or hyperplanes, when the observations are higher dimensional). In general (i.e. when the likelihood densities are not Gaussian), these decision boundaries will not be simple linear/planar shapes.
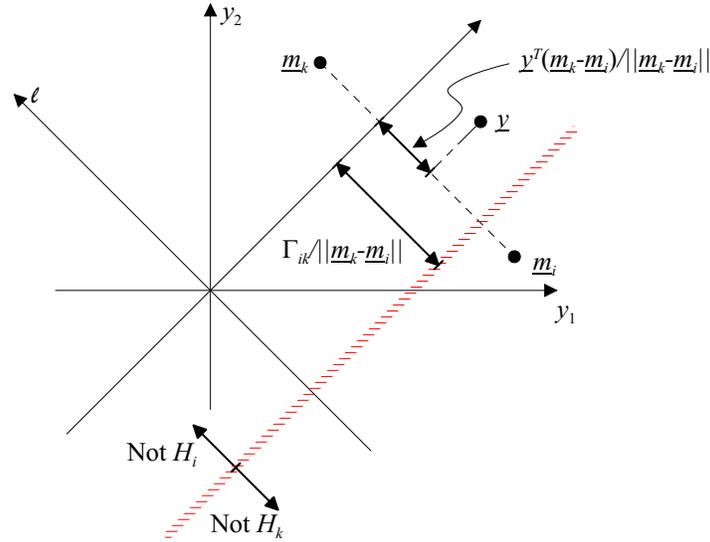

Figure 8.20: Illustration of the decision rule in the original data space.

Of course, we can also depict the decision rule for the MAP decision problem in the space of the likelihood ratios $\mathcal{L}_i$, as was done in Figure 8.19. In particular, if we express the MAP rule in the space of the original likelihood ratios for this 3 hypothesis case (i.e. by specializing (8.102) to the three hypotheses) we can express this rule as:

$$k = 0, i = 1: \qquad \mathcal{L}_1(y) \mathop{\gtrless}_{\text{Not } H_1}^{\text{Not } H_0} \frac{P_0}{P_1} \tag{8.108}$$

$$k = 1, i = 2: \qquad \mathcal{L}_2(y) \mathop{\gtrless}_{\text{Not } H_2}^{\text{Not } H_1} \left(\frac{P_1}{P_2} \right) \mathcal{L}_1(y) \tag{8.109}$$

$$k = 2, i = 0: \qquad \mathcal{L}_2(y) \mathop{\gtrless}_{\text{Not } H_2}^{\text{Not } H_0} \frac{P_0}{P_2} \tag{8.110}$$

In Figure 8.21 we show this decision rule in the likelihood space. When expressed in this way, the decision boundaries are independent of the specific likelihoods of the problem! That is, the decision regions for MAP rule for *any* 3-ary decision
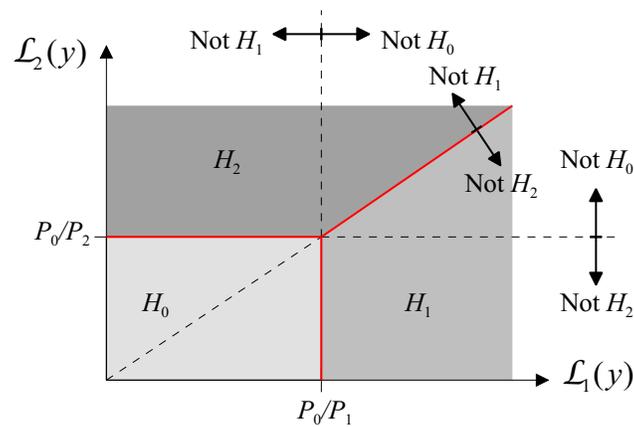
Figure 8.21: Illustration of the decision rule in the likelihood space.

problems is as given Figure 8.21. What has happened is that these likelihood details have been hidden in the likelihood ratios $\mathcal{L}(y)_i$.

Continuing with this example, suppose we additionally believe that each hypothesis is equally likely, so that $P_i = P_j = 1/3$. In this case, the decision rule will be the ML rule (8.103). Examining (8.107) and Figure 8.20, we can see that for our Gaussian example the ML rule but decision boundaries in the observation space halfway between each pair of means. Overall, the ML decision rule for this example becomes: Choose H_k if, for all i:

$$\|\underline{y} - \underline{m}_k\| \leq \|\underline{y} - \underline{m}_i\| \tag{8.111}$$

In particular, the decision rule chooses the hypothesis whose mean is closest to the given observation, resulting in the decision regions in the observation space shown in Figure 8.22 for a two-dimensional case. The decision boundaries are the bisectors of the lines connecting the means under the different hypotheses. In general, this type of decision strategy is called a *nearest neighbor classifier* or a *minimum distance receiver* in the literature. It is a strategy that is used rather widely in practice, even when it is not the optimum detector, due to its ease of implementation and understanding.
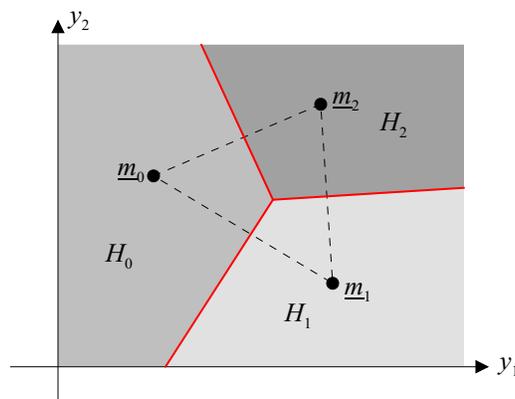
Figure 8.22: Illustration of the ML decision rule in the observation space.

Example 8.15 (Gaussians with different variances)
In this example, suppose we observe a one-dimensional random variable y and wish to determine which one of three-possible densities it could have come from. Under each of the three hypotheses the likelihoods are given by:

$$p_{Y|H}(y \mid H_i) = N(y; 0, \sigma_i^2) \quad i = 0, 1, 2 \tag{8.112}$$

where $\sigma_0 < \sigma_1 < \sigma_2$. Further, suppose the hypotheses are equally likely and we wish to minimize the probability of error. In this case the decision rule will be the ML rule. Applying (8.104) and simplifying we obtain the following decision rule for this case:

$$y^2 \mathop{\gtreqless}\limits_{\substack{\text{Not } H_k \\ \text{Not } H_i}}^{\text{Not } H_k} 2\left(\frac{\sigma_i^2 \sigma_k^2}{\sigma_i^2 - \sigma_k^2}\right) \ln\left(\frac{\sigma_i}{\sigma_k}\right) = \Gamma_{ik} \qquad \forall\, \text{unique } i, k \text{ pairs} \tag{8.113}$$

This decision rule is shown in Figure 8.23. The decision rule in the space of the likelihoods is essentially the same as that in Figure 8.21 with $P_i/P_j = 1$.
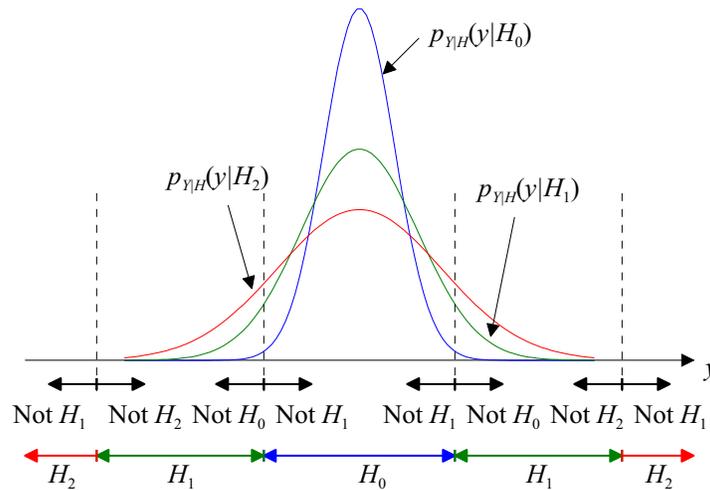

Figure 8.23: Illustration of decision rule in the observation space.

8.4.3 M-Ary Performance Calculations

The two performance metrics of the binary hypothesis testing problem were the expected value of the cost $E(C_{D(y)})$ and the probability of error $\Pr(\text{Error})$. Both these criteria still make sense in the M-ary case, though the expressions are a bit different. In particular, whereas in the binary case we could express both the metrics in terms of only two conditional densities (P_D and P_F), in the M-ary case we need $M(M-1)$ conditional densities to express them.

First let us consider the expected value of the cost:

$$E\left[C_{D(y)}\right] = \sum_{i=1}^{M-1} \sum_{j=1}^{M-1} C_{ij} \Pr\left(\text{Decide } H_i \mid H_j\right) P_j \tag{8.114}$$

Thus, we now need $M(M-1)$ conditional densities to express the expected cost or Bayes risk versus the two needed in the binary case (i.e. P_D and P_F). So the situation is more complicated, but the idea is the same. To find the expected value of the cost (that is, the Bayes risk), we have to find a set of conditional probabilities, as before.

Consider the problem of Example 8.14 with the ML decision rule, shown in Figure 8.22. To find $\Pr\left(\text{Decide } H_0 \mid H_1\right)$ in the observation space we need to integrate the conditional density $p_{Y|H}(y \mid H_1)$ over the region of the space where we would choose hypothesis H_0. The density $p_{Y|H}(y \mid H_1)$ is a circularly symmetric Gaussian centered at the mean $\underline{m}_1$. Referring to Figure 8.22, the H_0 region of the space is the shaded region on the left. The term $\Pr\left(\text{Decide } H_0 \mid H_1\right)$ is thus the area of the Gaussian in the H_0 part of the space, as shown in Figure 8.24. The calculation of the other conditional densities is similar, where, in general, both the region of integration changes and the density being integrated changes.

Of course, if it is more convenient, we can also find these conditional densities in the space of a sufficient statistic. The basic idea is the same. Consider again the Example 8.14 with the ML decision rule, shown
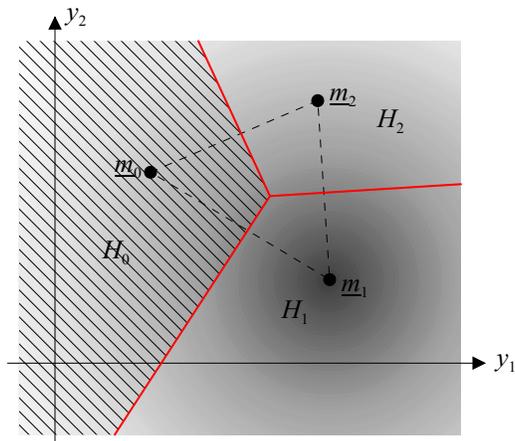
Figure 8.24: Illustration of the calculation of $\Pr\left(\text{Decide } H_0 \mid H_1\right)$ in the observation space.

this time in the space of the sufficient statistic provided by the likelihood ratios $\mathcal{L}_i(\underline{y})$ in Figure 8.21. To find $\Pr\left(\text{Decide } H_0 \mid H_1\right)$ we need to integrate the joint conditional density for the likelihood ratio sufficient statistics $p_{\mathcal{L}_1(y),\mathcal{L}_2(y)\mid H}(\mathcal{L}_1(y),\mathcal{L}_1(y) \mid H_0)$ over that part of the space of the likelihood ratios where we decide H_1. While the region of the likelihood space space is simply determined in this case, the required density may not be. In Example 8.14, even though the observations are Gaussian under any hypothesis, the likelihood ratios, being of the form $e^{\underline{y}^T \Sigma \underline{y}}$, will not be Gaussian random variables! All sufficient statistics are not equal, however, and a different choice of sufficient statistic may make the problem easier. Note for this example that the sufficient statistics $\ell_{ik}(y)$ defined in (8.107) are simply linear functions of the observations, and thus are themselves Gaussian random variables under any hypothesis. The decision regions are also relatively simple for these particular sufficient statistics. This discussion illustrates the issues we face in general when performing such calculations. The challenge is to find a sufficient statistic whose combination of decision regions and densities lead to a tractable set of calculations.

Our other performance metric was the probability of error $\Pr\left[\text{Error}\right]$. In the M-ary case this is given as:

$$\Pr\left[\text{Error}\right] \quad = \quad \sum_{j=0}^{M-1} \sum_{\substack{i=0 \\ i \neq j}}^{M-1} \Pr\left[\text{Decide } H_i \mid H_j \text{ true} P_j\right] \tag{8.115}$$

As in the calculation of the expected cost, the key is again the calculation of the conditional densities $\Pr\left[\text{Decide } H_i \mid H_j \text{ true} P_j\right]$. These probabilities can be calculated as illustrated in Figure 8.24 for Example 8.14. In the case of the $\Pr\left[\text{Error}\right]$ calculation there is an alternative form to the expression that is sometimes useful. It is based on the fact that the sum in (8.115) includes all the conditional densities except the "self term" $\Pr\left[\text{Decide } H_i \mid H_i \text{ true} P_j\right]$. As a result we may rewrite (8.115) as follows:

$$\Pr\left[\text{Error}\right] \quad = \quad \sum_{j=0}^{M-1} \left(1 - \Pr\left[\text{Decide } H_j \mid H_j \text{ true} P_j\right]\right) \tag{8.116}$$

Consider again the problem of Example 8.14 with the ML decision rule, shown in Figure 8.22. To find a self term, for example $\Pr\left(\text{Decide } H_1 \mid H_1\right)$, in the observation space we need to integrate the conditional density $p_{Y\mid H}(y \mid H_1)$ over the region of the space where we would choose hypothesis H_1. The term $\Pr\left(\text{Decide } H_1 \mid H_1\right)$ is thus the area of the Gaussian in the H_1 part of the space, as shown in Figure 8.25. The calculation of the other terms is similar. As in the case of the expected cost calculation, we may also perform such calculations in the space of a sufficient statistic if that is more convenient.
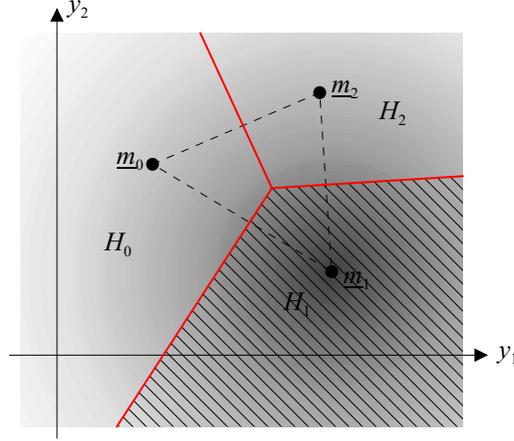
Figure 8.25: Illustration of the calculation of $\Pr\left(\text{Decide } H_1 \mid H_1\right)$ in the observation space.

8.5 Gaussian Examples

Gaussian detection problems are of general interest in many applications. In this section, several additional examples are discussed.

The general Gaussian likelihood ratio test is straightforward to compute. Let $\underline{y}$ be the n-dimensional observation vector, with hypothesized density $p_{\underline{Y}|H}(\underline{y} \mid H_0) \sim N(\underline{m}_0, \Sigma_0)$ under H_0 and density $p_{\underline{Y}|H}(\underline{y} \mid H_1) \sim N(\underline{m}_1, \Sigma_1)$ under H_1. Then the likelihood ratio test for the general Gaussian case is given by:

$$\mathcal{L}(\underline{y}) = \frac{p_{\underline{Y}|H}(\underline{y} \mid H_1)}{p_{\underline{Y}|H}(\underline{y} \mid H_0)} = \frac{\frac{1}{\sqrt{(2\pi)^N |\Sigma_1|}} e^{-\frac{1}{2}(y-\underline{m}_1)^T \Sigma_1^{-1}(y-\underline{m}_1)}}{\frac{1}{\sqrt{(2\pi)^N |\Sigma_0|}} e^{-\frac{1}{2}(y-\underline{m}_0)^T \Sigma_0^{-1}(y-\underline{m}_0)}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \qquad (8.117)$$

where $|\Sigma_i|$ is the determinant of Σ_i. Taking logarithms of both sides and clearing out factors of $1/2$, one obtains the following form of the LRT:

$$\ell(\underline{y}) = -(\underline{y} - \underline{m}_1)^T \Sigma_1^{-1}(\underline{y} - \underline{m}_1) + (y - \underline{m}_0)^T \Sigma_0^{-1}(\underline{y} - \underline{m}_0) \underset{H_0}{\overset{H_1}{\gtrless}} 2\ln(\eta) + \ln(|\Sigma_1|) - \ln(|\Sigma_0|) \qquad (8.118)$$

The above expression indicates that a sufficient statistic is $\ell(\underline{y}) = (\underline{y} - \underline{m}_0)^T \Sigma_0^{-1}(\underline{y} - \underline{m}_0) - (\underline{y} - \underline{m}_1)^T \Sigma_1^{-1}(\underline{y} - \underline{m}_1)$.

Example 8.16

Consider now the detection of known signals in additive Gaussian noise, where the elements y_j of the observation vector $\underline{y}$ are given by the following expression under each hypothesis:

$$H_i: \qquad y_j = m_{ij} + w_j \qquad (8.119)$$

where the values of m_{ij} are known, and w_j is an independent, identically distributed sequence of Gaussian random variables with distribution $N(0, \sigma^2)$. Note that, in this case, $\Sigma_1 = \Sigma_0 = \sigma^2 I$, so that the sufficient statistic becomes:

$$\ell(\underline{y}) = \frac{2}{\sigma^2}(\underline{m}_1 - \underline{m}_0)^T \underline{y} + \frac{\underline{m}_0^T \underline{m}_0 - \underline{m}_1^T \underline{m}_1}{\sigma^2} \qquad (8.120)$$

The optimal detector can be written as

$$\underline{y}^T (\underline{m}_1 - \underline{m}_0) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\underline{m}_1^T \underline{m}_1 - \underline{m}_0^T \underline{m}_0}{2} + \sigma^2 \ln(\eta) \qquad (8.121)$$

Example 8.17 (Uniformly Most Powerful Test)
One can also detect unknown signals in Gaussian noise, as follows: Assume that the observations are distributed as

$$y_j = \begin{cases} w_j & \text{if } H_0 \text{ is true} \\ x_j + w_j & \text{otherwise} \end{cases} \tag{8.122}$$

where x_j is the j-th coefficient of a Gaussian vector $\underline{x}$ which is independent of $\underline{w}$, with distribution $N(\underline{m}_x, \Sigma_x)$. Again, this is a Gaussian detection problem, with $\underline{m_1} = \underline{m}_x$, $\Sigma_1 = \Sigma_x + \sigma^2 I$, $\Sigma_0 = \sigma^2 I$, $\underline{m_0} = 0$. In this case, the sufficient statistic becomes

$$\ell(\underline{y}) = \sigma^{-2}(\underline{y}^T \underline{y}) - (\underline{y} - \underline{m}_x)^T \Sigma_1^{-1} (\underline{y} - \underline{m}_x) = \underline{y}^T[\sigma^{-2}\underline{y} - \Sigma_1^{-1}(\underline{y} - \underline{m}_x) + \Sigma_1^{-1}\underline{m}_x] - \underline{m}_x^T \Sigma_1^{-1} \underline{m}_x \tag{8.123}$$

Thus, the optimal detector is to declare H_1 whenever

$$\underline{y}^T \left[\sigma^{-2}\underline{y} - \Sigma_1^{-1} \left(\underline{y} - \underline{m}_x \right) + \Sigma_1^{-1}\underline{m}_x \right] \underset{H_0}{\overset{H_1}{\gtrless}} 2\ln(\eta) + \underline{m}_x^T \Sigma_1^{-1} \underline{m}_x + \ln\left(|\det(\Sigma_1)|\right) - \ln\left(|\det(\Sigma_0)|\right) \tag{8.124}$$

It is interesting to examine the term on the right-hand side. In particular, note the following relationships which hold true under H_1:

$$\Sigma_{yy} = E\left[\left(\underline{y} - \underline{m}_x \right) \left(\underline{y} - \underline{m}_x \right)^T \middle| H_1 \right] = \Sigma_1 = \Sigma_x + \sigma^2 I \tag{8.125}$$

$$\Sigma_{xy} = E[(\underline{x} - \underline{m}_x)(\underline{y} - \underline{m}_x)^T \mid H_1] = \Sigma_x \tag{8.126}$$

Thus,

$$\sigma^{-2}\underline{y} - \Sigma_1^{-1}\underline{y} = \Sigma_1^{-1}(\sigma^{-2}(\Sigma_x + \sigma^2 I) - I) = \sigma^{-2}\Sigma_1^{-1}\Sigma_x$$

The above expression can be given an interesting interpretation. Consider the case where $\underline{m}_x = 0$. Then, using the expression for Gaussian estimation,

$$E[\underline{x} \mid \underline{y}, H_1] = \Sigma_1^{-1}\Sigma_x \underline{y} \tag{8.127}$$

and the optimal detection rule selects H_1 whenever

$$\underline{y}^T E[\underline{x} \mid \underline{y}, H_1] > 2\sigma^2 [\ln(T) + \ln(|\det(\Sigma_1)|) - \ln(|\det(\sigma^2 I)|)] \tag{8.128}$$

In particular, this decision rule is similar to the known signal case, except that the known difference in the means is replaced by $E[\underline{x} \mid y, H_1]$.

Chapter 9

Series Expansions and Detection of Stochastic Processes

In many cases, it is easier to view a continuous-time stochastic process $x(t)$ defined over a finite interval $[0, T]$ in terms of an infinite set of random coefficients. When the sample paths of $x(t)$ are sufficiently regular (e.g. continuous), one can expand $x(t)$ in a Fourier series, for example. In this section, the properties of such expansions are discussed. To begin the discussion, it is important to review the properties of series expansions of deterministic functions.

9.1 Deterministic Functions

Let $x(t)$ be a determinsitic function defined in a time interval $[S, T]$. Assume that one is interested in representing $x(t)$ as

$$x(t) = \sum_i x_i f_i(t) \tag{9.1}$$

where $f_i(t)$ are a set of basis functions, and x_i is a set of associated coefficients.

Ideally, one would like to have a set of basis functions which are complete, in that every function $x(t)$ can be defined as in (9.1), and which are orthonormal in some sense. For instance, consider the set of functions $f_i(t) = \frac{1}{\sqrt{T-S}} e^{j2\pi it/(T-S)}, |i| = 0, 1, \dots.$ This is the standard Fourier series basis over a finite interval. This basis has the property that

$$\int_S^T f_i(t) f_k(t)^* dt = \frac{1}{T-S} \int_S^T e^{j2\pi(i-k)t/(T-S)} = \begin{cases} 1 & \text{if } i = k \\ 0 & otherwise \end{cases} \tag{9.2}$$

Indeed, one can show that, for any function $x(t)$ such that $\int_S^T |x(t)|^2 dt < \infty$, there is a set of coefficients x_i such that

$$\lim_{n \to \infty} \int_S^T |x(t) - \sum_{i=-n}^n x_i f_i(t)|^2 dt = 0 \tag{9.3}$$

The coefficients x_i can be obtained from the orthonormal property of the $f_i(t)$, by

$$x_i = \int_S^T x(t) f_i(t)^* dt \tag{9.4}$$

With an analogy to linear algebra, consider the space of all square-integrable, complex-valued functions such that $\int_S^T |x(t)|^2 dt < \infty$. This is a linear space, in that scaled versions of these functions and sums of these functions are also square-integrable. On this space, define the inner product

$$< x(t), y(t) >= \int_S^T x(t) y(t)^* dt \tag{9.5}$$

An orthonormal basis on this space is a set of functions $f_i(t)$ satisfying the property that

$$< f_i, f_k > = \int_S^T f_i(t) f_k(t)^* dt = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{otherwise} \end{cases} \tag{9.6}$$

A complete orthonormal basis is such that every element $x(t)$ of the space can be expressed as a sum (9.1), where the coefficients are computed as (9.4).

9.2 Series Expansion of Stochastic Processes

Consider now a zero-mean, complex-valued stochastic process $x(t)$, defined on the interval $[S, T]$. For stochastic processes, one can expand the process as the mean-square sense limit of an infinite series, as

$$x(t) \overset{\text{mss}}{=} \sum_i x_i f_i(t) \tag{9.7}$$

where the random coefficients x_i are given by stochastic integrals, as

$$x_i = \int_S^T x(t) f_i(t)^* dt \tag{9.8}$$

Note the following properties:

$$E[x_i] = \int_S^T E[x(t)] f_i(t)^* dt = 0 \tag{9.9}$$

$$E[x_i x_j] = \int_S^T \int_S^T E[x(t)x(s)^*] f_i(t) f_j(s)^* ds dt = \int_S^T \int_S^T K_x(t, s) f_i(t) f_j(s)^* ds dt \tag{9.10}$$

A particular case of interest is when $x(t)$ is white noise, so that $K_x(t, s) = \delta(t - s)$. In this case, one sees that

$$E[x_i x_j] = \int_S^T \int_S^T \delta(t - s) f_i(t) f_j(s)^* ds dt = \int_S^T f_i(s) f_j(s)^* ds = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{otherwise} \end{cases} \tag{9.11}$$

Thus, series expansions of white noise using orthonormal functions are such that the coefficients are orthogonal. Furthermore, since white noise is a Gaussian process, the coefficients are also Gaussian! The result is that expansion of white noise in orthogonal series results in an independent, identically distributed sequence of coefficients x_i, Gaussian with zero mean and unit variance.

Consider now the case where the process $x(t)$ is not white, but has a general autocovariance function $K_x(t, s)$. In this case, the coefficients may not be orthogonal, as

$$E[x_i x_j] = \int_S^T \int_S^T K_x(t, s) f_i(t) f_j(s)^* \, ds \, dt \neq 0 \ \text{ if } i \neq j \tag{9.12}$$

Thus, for any arbitrary complete, orthonormal basis, the process $x(t)$ can be expanded in a series, but the coefficients will not necessarily be orthogonal. However, there is a special orthonormal basis for which the coefficients would be orthogonal! Consider a basis where the basis functions satisfy the following integral equation:

$$\int_S^T K_x(t, s) f_i(s) = \lambda_i f_i(t) \tag{9.13}$$

Then, (9.12) becomes

$$E[x_i x_j] = \int_S^T \int_S^T K_x(t, s) f_i(t) f_j(s)^* ds dt = \int_S^T \lambda_j^* f_i(t) f_j(t)^* dt = \begin{cases} \lambda_j^* & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{9.14}$$

In this special basis, the coefficients x_i are orthogonal; for general stochastic processes, an expansion in an orthonormal basis of the form

$$x(t) \stackrel{\text{mss}}{=} \sum_i x_i f_i(t) \tag{9.15}$$

where the basis functions $f_i(t)$ satisfy (9.13), and the coefficients x_i are orthogonal, is known as a *Karhunen-Loeve expansion*.

The fact that a basis exists for Karhunen-Loeve expansions is a result from integral equations. The exposition below is a brief introduction to the subject. In order to best understand Karhunen-Loeve expansions, it is useful to first consider the case of a vector-valued random variable rather than a stochastic process. One can think of a vector-valued random variable as a discrete-time stochastic process where time has a finite range. Let $\underline{x}$ be an n-dimensional vector-valued random variable, with covariance matrix Σ_x. The covariance Σ_x is a positive semidefinite, symmetric matrix. The idea of a Karhunen-Loeve expansion in this problem is to seek a set of n basis functions $\underline{u}_i$ which have the following properties:

$$\Sigma_x \underline{u}_i \;\; = \;\; \lambda_i \underline{u}_i \tag{9.16}$$

$$\underline{u}_i^H \underline{u}_j \;\; = \;\; \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{9.17}$$

Note that (9.16) implies that the basis vectors $\underline{u}_i$ are eigenvectors of the matrix Σ_x with eigenvalues λ_i. Since Σ_x is positive semidefinite, its eigenvalues must be nonnegative. From the theory of symmetric matrices, one knows that a symmetric matrix has a complete set of orthonormal eigenvectors (i.e. there exists an orthogonal basis where the matrix is diagonal). In particular, one easy way of finding such a basis is by solving the following problem:

$$\lambda_1 = \max_{\|\underline{u}\|=1} \|\Sigma_x \underline{u}\| > 0 \tag{9.18}$$

and $\underline{u}_1$ is the vector which achieves the maximum. Note that such a maximum must exist, since K_x is a bounded function, and the admissible vectors are a closed and bounded set (i.e. compact), namely the unit ball. Because of the symmetry of K_x, and $\underline{u}_1$ is also an eigenvector; that is,

$$\Sigma_x \underline{u}_1 = \lambda_1 \underline{u}_1, \underline{u}_1^H \Sigma_x = \lambda_1 \underline{u}_1^H \tag{9.19}$$

Once the first vector is found, one can form the reduced matrix

$$\Sigma_1 = \Sigma_x - \lambda_1 \underline{u}_1 \underline{u}_1^H$$

Theorem 9.1
The matrix Σ_1 is also positive semidefinite.

The proof of this result lies in defining the auxiliary random vector $\underline{x}_1 = \underline{x} - (\underline{u}_1^H \underline{x})\underline{u}_1$. The covariance of x_1 is

$$\begin{aligned} \Sigma \;\; = \;\; & E[(\underline{x} - (\underline{u}_1^H \underline{x})\underline{u}_1)(\underline{x} - (\underline{u}_1^H \underline{x})\underline{u}_1)^H] \\ = \;\; & \Sigma_x - E[(\underline{u}_1^H \underline{x})\underline{u}_1 \underline{x}^H] - E[\underline{x}((\underline{u}_1^H \underline{x})\underline{u}_1)^H] + E[(\underline{u}_1^H \underline{x})\underline{u}_1 \underline{u}_1^H (\underline{u}_1^H \underline{x})^*] \end{aligned} \tag{9.20}$$

Note that $(\underline{u}_1^H \underline{x}) = (\underline{x}^H \underline{u}_1)^*$ is a scalar, and thus commutes with the matrices it is multiplied against. The above equation can be rearranged to obtain

$$\begin{aligned} \Sigma \;\; = \;\; & \Sigma_x - \underline{u}_1 \underline{u}_1^H E[\underline{x}\underline{x}^H] - E[\underline{x}\underline{x}^H]\underline{u}_1 \underline{u}_1^H + \underline{u}_1 \underline{u}_1^H (\underline{u}_1^H E[\underline{x}\underline{x}^H]\underline{u}_1) \\ = \;\; & \Sigma_x - 2\lambda_1 \underline{u}_1 \underline{u}_1^H + \lambda_1 \underline{u}_1 \underline{u}_1^H = \Sigma_x - \lambda_1 \underline{u}_1 \underline{u}_1^H = \Sigma_1 \end{aligned} \tag{9.21}$$

Hence, since Σ_1 is also a covariance matrix, it is also positive semidefinite.

Assume now that $\Sigma_1 \neq 0$. In this case, the process can be repeated again, to obtain $0 < \lambda_2 \leq \lambda_1$, and $\underline{u}_2$, based on the matrix Σ_1. The following result holds:

Theorem 9.2
$\underline{u}_2$ is an eigenvector of Σ_x, with eigenvalue λ_2. Furthermore, $\underline{u}_2^H \underline{u}_1 = 0$.

To show this, consider the definition of Σ_1. Since $\underline{u}_2$ is an eigenvector of Σ_1, the following equation holds:

$$(\Sigma_x - \lambda_1 \underline{u}_1 \underline{u}_1^H)\underline{u}_2 = \lambda_2 \underline{u}_2 \tag{9.22}$$

Multiply on the left by $\underline{u}_1^H$ to obtain

$$\underline{u}_1^H \Sigma_x \underline{u}_2 - \lambda_1 (\underline{u}_1^H \underline{u}_1)\underline{u}_1^H \underline{u}_2 = \lambda_2 \underline{u}_1^H \underline{u}_2 \tag{9.23}$$

Using the fact that $\underline{u}_1^H \underline{u}_1 = 1$, and that $\underline{u}_1$ is an eigenvector of Σ_x, this simplifies to

$$\lambda_1 (\underline{u}_1^H \underline{u}_2 - \underline{u}_1^H \underline{u}_2) = 0 = \lambda_2 \underline{u}_1^H \underline{u}_2 \tag{9.24}$$

which implies that $\underline{u}_1^H \underline{u}_2 = 0$. Substituting into (9.22) establishes that $\underline{u}_2$ is indeed an eigenvector.

The procedure can be continued recursively, until the residual covariance $\Sigma_i = 0$. Once $\Sigma_i = 0$, the procedure can be stopped, and one has the expansion:

$$\Sigma_x = \sum_{j=1}^{i} \lambda_j \underline{u}_j \underline{u}_j^H \tag{9.25}$$

Note by necessity that $i \leq n$, since, after n expansions, it is impossible to find a non-zero vector which is orthogonal to the other n vectors. Furthermore, the above construction establishes that the vector $\underline{x} - \sum_{j=1}^{i} \underline{u}_j (\underline{u}_j^H \underline{x})$ has zero covariance! This establishes the Karhunen-Loeve expansion

$$\underline{x} \stackrel{\text{mss}}{=} \sum_{j=1}^{i} x_j \underline{u}_j \tag{9.26}$$

with random, orthogonal coefficients $x_j = (\underline{u}_j^H \underline{x})$.

Consider now extending the above development to a stochastic process $x(t)$. The principal difference is that, instead of having a vector-valued random variable, the random variable takes the value of an entire function (an infinite-dimensional space). However, the autocovariance function $K_x(t, s)$ is still a positive semidefinite operator, and there are equivalent results in functional analysis which enable the development to carry through. In particular, the following result is proven in many functional analysis texts such as Hille and Yosida, or Riesz and Sz-Nagy:

Theorem 9.3
Let $K_x(t, s)$ be continuous, Hermitian, nonzero and positive semidefinite on the interval $[S, T]$. Then, there exists at least one eigenfunction $f(t)$ and one positive eigenvalue λ satisfying

$$\int_S^T K_x(t, s)f(s) = \lambda f(t) \tag{9.27}$$

$$\int_S^T f(s)f(s)^* ds < \infty \tag{9.28}$$

Using this result, a sequence of eigenvalues and eigenfunctions λ_i, $f_i(t)$ can be constructed as above, with the property that the eigenvalues λ_i are positive, and the eigenfunctions $f_i(t)$ form an orthonormal basis. After n eigenfunctions are found, one has the following expression for the reduced autocovariance $K_n(t, s)$:

$$K_n(t, s) = K_x(t, s) - \sum_{i=1}^{n} \lambda_i f_i(t) f_i(s)^* \tag{9.29}$$

which is the autocovariance of the residual process $x(t) - \sum_{i=1}^{n} f_i(t) \int_S^T x(s) f_i(s)^* ds$. One needs to show that, as $n \to \infty$, the error $K_n(t, s) \to 0$. A simple way to show this is to note that

$$K_n(t, t) = K_x(t, t) - \sum_{i=1}^{n} |f_i(t)|^2 \tag{9.30}$$

so that $\sum_{i=1}^{n} |f_i(t)|^2$ is bounded above and monotone, so it converges to a limit $\sum_{i=1}^{\infty} |f_i(t)|^2$. This also establishes that $\sum_{i=1}^{n} \lambda_i f_i(t) f_i(s)^*$ converges to a limit, since

$$
\begin{aligned}
|\sum_{i=1}^{n} \lambda_i f_i(t) f_i(s)^* - \sum_{i=1}^{m} \lambda_i f_i(t) f_i(s)^*| &= |\sum_{i=m}^{n} \lambda_i f_i(t) f_i(s)^*| \\
&\leq (\sum_{i=m}^{n} \lambda_i |f_i(t)|^2)^{1/2} (\sum_{i=m}^{n} \lambda_i |f_i(s)|^2)^{1/2} \\
&\to 0 \text{ as } m, n \to \infty.
\end{aligned}
\tag{9.31}
$$

Thus, the residual autocovariance approaches a limit $K_\infty(t, s) = K_x(t, s) - \sum_{i=1}^{\infty} \lambda_i f_i(t) f_i(s)^*$. If this limit is not identically zero, then there is a positive eigenvalue and eigenvector which can be added to the sum, to change the value of K_∞, which contradicts the assumed convergence.

The above argument establishes that the autocovariance $K_x(t, s)$ can be expanded as

$$
K_x(t, s) = \sum_{i=1}^{\infty} \lambda_i f_i(t) f_i(s)^*
\tag{9.32}
$$

in terms of the orthonormal eigenfunctions $f_i(t)$. The convergence of the sum is uniform, in $t, s \in [S, T]$, as implied by the bound in (9.31). This result is known as *Mercer's Theorem*. Mercer's Theorem implies that

$$
E[\int_S^T x^2(t)dt] = \int_S^T K_x(t, t)dt = \int_S^T \sum_{i=1}^{\infty} \lambda_i f_i(t) f_i(t)^* dt = \sum_{i=1}^{\infty} \lambda_i
\tag{9.33}
$$

Furthermore, the process $x(t)$ can be written as

$$
x(t) \overset{\text{mss}}{=} \sum_{i=1}^{\infty} x_i f_i(t)
\tag{9.34}
$$

where the coefficients of the expansion x_i are orthonormal random variables, given as

$$
x_i = \int_S^T x(t) f_i(t)^* dt
\tag{9.35}
$$

This expansion is known as the *Karhunen-Loeve expansion*.

The sequence of eigenfunctions $f_i(t)$ constructed above will form a *complete* orthonormal set if and only if $K_x(t, s)$ is positive definite. In this case, an arbitrary deterministic function $f(t)$ can be expanded using the series expansion in terms of $f_i(t)$. If $K_x(t, u)$ is only postive semidefinite, the construction above must be augmented with enough additional orthogonal functions, corresponding to the zero eigenvalues of $K_x(t, u)$, to form a complete orthonormal set. A simple way of constructing such a set is to construct the eigenfunctions for the autocovariance function $K_x(t, s) + I\delta(t - s)$, where I is the identity matrix; this modified autocovariance function has the same eigenfunctions as $K_x(t, s)$, except that it is positive definite.

Example 9.1 (Karhunen-Loeve Expansion of Wiener Process)
Let $b(t)$ be a Wiener process with rate σ^2, with autocovariance $K_b(t, s) = \sigma^2 \min(t, s)$ defined on $[0, T]$. The integral equation defining the eigenfunctions is:

$$
\lambda f(t) = \sigma^2 \int_0^T \min(t, s) f(s) ds = \sigma^2 (\int_0^t s f(s) ds + t \int_t^T f(s) ds)
\tag{9.36}
$$

To obtain a differential equation for $f(t)$, differentiate with respect to t:

$$
\lambda \frac{d}{dt} f(t) = \sigma^2 (t f(t) + \int_t^T f(s) ds - t f(t)) = \sigma^2 \int_t^T f(s) ds
\tag{9.37}
$$

Differentiating again yields

$$
\frac{d^2}{dt^2} f(t) + \frac{\sigma^2}{\lambda} f(t) = 0 \equiv \frac{d^2}{dt^2} f(t) + a^2 f(t)
\tag{9.38}
$$

where $a = \sqrt{\sigma^2/\lambda}$. The solutions to this equation are sinusoids. From the above integral equations, the following boundary conditions are required:

$$f(0) = 0; \frac{d}{dt}f(T) = 0$$

These bondary conditions are sufficient to uniquely specify the unique values of a for which a solution satisfying the boundary conditions exist. The condition $f(0) = 0$ implies $f(t) = K \sin at$; the other boundary condition implies that $aT = (n - 1/2)\pi$ for some integer n. Since the eigenvalue must be nonnegative, n is also positive, and so the eigenvalue λ can take the values

$$\lambda_n = \frac{\sigma^2 T^2}{(n - 1/2)^2 \pi^2}, n = 1, 2, \ldots \tag{9.39}$$

with corresponding eigenfunctions

$$f_n(t) = K_n \sin(n - 1/2)\pi t \tag{9.40}$$

The constants K_n are chosen to normalize the eigenfunctions, so that

$$\int_0^T K_n^2 \sin(n - 1/2)\pi t dt = 1/2 K_n^2 T = 1$$

Thus, $K_n = \sqrt{2/T}$.

The main limitation in using Karhunen-Loeve expansions is in determining the eigenvalues and eigenfunctions. This limits the applicability of the method for practical problems. However, when the process of interest is white noise, any orthonormal set of eigenfunctions can be used. Thus, solving detection problems involving white noise processes is straightforward, as discussed in the next section.

9.3   Detection of Known Signals in Additive White Noise

As an application of series expansions of random processes, consider the problem of observing a random process $y(t)$ for detecting among the following two hypotheses: Under hypothesis H_1, the random process $y(t)$ is described as:

$$y(t) = s_1(t) + w(t) \tag{9.41}$$

where $w(t)$ is a white noise process, with autocorrelation $q\delta(t - s)$, and $s_1(t)$ is a known signal. Under hypothesis H_0, the random process $y(t)$ is given by

$$y(t) = s_0(t) + w(t) \tag{9.42}$$

where $s_0(t)$ is a known signal and $w(t)$ is again white noise with autocorrelation $q\delta(t - s)$.

Assume that the process $y(t)$ is observed over the interval $[0, T]$. The problem is to design an optimal detector, in terms of either a Bayes' cost or a Maximum Aposteriori detector. The complicating factor in this problem is that it is difficult to write the likelihood ratio in terms of continuous functions (although it is possible to develop such an expression as a limit of the vector observation likelihood functions discussed previously).

Instead of dealing with the continuous-time observation process, one can use a series expansion to convert $y(t)$ to an infinite vector of coefficients $\underline{y}$, where y_i is the $i - th$ coefficient in the series expansion. It is important to select the basis functions for the expansion carefully. In particular, consider the following basis function:

$$f_1(t) = \frac{1}{(\int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds)^{1/2}} (s_1(t) - s_0(t)) \tag{9.43}$$

and select the rest of the basis functions to form a complete orthonormal set over $[0, T]$, orthogonal to

$s_1(t) - s_0(t)$ Observe the following identities:

$$\int_0^T s_1(t)f_1(t)^* dt = \frac{\int_0^T s_1(s)(s_1(s) - s_0(s))^* ds}{(\int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds)^{1/2}}$$

$$= \frac{\int_0^T s_0(s)(s_1(s) - s_0(s))^* ds + \int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds}{(\int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds)^{1/2}}$$

$$\int_0^T s_1(t)f_i(t) dt = \int_0^T s_0(t)f_i(t) dt \text{ if } i > 1$$

$$\int_0^T s_0(t)f_1(t)^* dt = \frac{\int_0^T s_0(s)(s_1(s) - s_0(s))^* ds}{(\int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds)^{1/2}} \tag{9.44}$$

One can now convert the waveform $y(t)$ into the corresponding coefficients and, under hypothesis H_1, the coefficients are independent of each other and become:

$$y_i = \begin{cases} \frac{\int_0^T s_0(s)(s_1(s) - s_0(s))^* ds + \int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds}{\left(\int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds\right)^{1/2}} + w_1 & \text{if } i = 1 \\ \int_0^T s_0(t)f_i(t) dt + w_i & \text{if } i \neq 1 \end{cases} \tag{9.45}$$

due to the orthogonal construction of the basis functions $f_i(t)$. Under hypothesis H_0, the coefficients are again independent of each other and are given by:

$$y_i = \begin{cases} \frac{\int_0^T s_0(s)(s_1(s) - s_0(s))^* ds}{\left(\int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds\right)^{1/2}} & \text{if } i = 1 \\ \int_0^T s_0(t)f_i(t) dt + w_i & \text{otherwise} \end{cases} \tag{9.46}$$

where the w_i are independent, zero-mean Gaussian random variables with covariance q, because they are the coefficients of a white noise expansion using orthonormal eigenfunctions. What the above expansion shows is that the only coefficient which differs in value between the two hypotheses is y_1. Furthermore, since the coefficients y_i are independent under each hypothesis, observation of any other coefficient $y_i, i > 1$ provides no information concerning the value of coefficient y_1, and thus is not useful for discriminating among the hypotheses. Thus, the rest of the coefficients contain no information which is useful for discriminating among the two hypotheses and can be ignored. The coefficient y_1 is a sufficient statistic for the detection problem; that is, for the purposes of solving the detection problem, it is sufficient to consider the one-dimensional Gaussian vector consisting of y_1! The reduced detection problem becomes:

$$p(y_1|H_i) = \begin{cases} N(a_1, q) & \text{if } H_1 \text{ is true} \\ N(a_0, q) & \text{if } H_0 \text{ is true} \end{cases} \tag{9.47}$$

where

$$a_1 = \frac{\int_0^T s_0(s)(s_1(s) - s_0(s))^* ds + \int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds}{(\int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds)^{1/2}}$$

$$a_2 = \frac{\int_0^T s_0(s)(s_1(s) - s_0(s))^* ds}{(\int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds)^{1/2}}$$

From the previous analysis of Gaussian detection problems, the optimal detector is given by:

$$m(y_1) = \begin{cases} H_1 & \text{if } (a_1 - a_0)y_1 > q\ln(T) + a_1^2/2 - a_0^2/2 \\ H_0 & \text{otherwise} \end{cases} \tag{9.48}$$

Simplifying, note that $a_1 - a_0$ is given by

$$a_1 - a_0 = \frac{\int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds}{(\int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds)^{1/2}} = (\int_0^T (s_1(s) - s_0(s))(s_1(s) - s_0(s))^* ds)^{1/2} \tag{9.49}$$

and

$$(a_1 - a_0)y_1 = \left((s_1(s) - s_0(s))\,(s_1(s) - s_0(s))^*\right)^{1/2} \int_0^T y(t) f_1(t) dt = \int_0^T y(t)(s_1(t) - s_0(t))^* dt \qquad (9.50)$$

Thus, an equivalent sufficient statistic is $S(y) = \int_0^T y(t)(s_1(t) - s_0(t))^* dt$. This statitstic is formed by correlating the input $y(t)$ with the known difference between the means of the two hypotheses. This is known as a *matched filter*.

Example 9.2 (Detection of Bits in Additive White Noise)
Assume that one of two signals $s_0(t) = \sin(10t)$ or $s_1(t) = \sin(10t + \pi/2)$ is transmitted over an interval $t \in [0, 2\pi]$ through a noisy channel with output modeled as $y(t) = s(t) + n(t)$, where $n(t)$ is Gaussian white noise with autocovariance $\delta(t - s)$. The problem is to design the optimal decoder at the receiving end. This is a problem of binary detection. Let H_1 denote the hypothesis that $s_1(t)$ was the transmitted signal. The sufficient statistic can be found as:

$$S(y) = \int_0^{2\pi} (\sin(10t + \pi/2) - \sin(10t))y(t)dt$$

Since $y(t)$ is a Gaussian process under either hypotheses, the statistic $S(y)$ will also be Gaussian. The mean of $S(y)$ under both hypotheses is given by

$$E[S(y)|H_i] = \int_0^{2\pi} (\sin(10t + \pi/2) - \sin(10t))s_i(t)dt = \begin{cases} \pi & \text{if } H_1 \\ -\pi & \text{if } H_0 \end{cases}$$

and the variance under both cases is given by

$$\int_0^{2\pi} (\sin(10t + \pi/2) - \sin(10t))^2 dt = 2\pi$$

Using these statistics, the optimal detector can be constructed as in the scalar Gaussian case.

9.4 Detection of Unknown Signals in White Noise

Consider the following detection problem: The observations $y(t), t \in [0, T]$ are given by:

$$y(t) = \begin{cases} x(t) + w(t) & \text{if } H_1 \text{ is true} \\ w(t) & \text{otherwise} \end{cases} \qquad (9.51)$$

where $x(t)$ is a zero-mean, finite-variance Gaussian random process with autocovariance $K_x(t, s)$, which is independent of the additive white noise $w(t)$ with autocovariance $\delta(t - s)$. The nature of the optimal detector can again be determined using a series expansion. However, the nature of the sufficient statistic is harder to determine, as shown below.

The key to solving the above detection problem is to transform the observations using the Karhunen-Loeve expansion of $x(t)$, as follows: Let $f_i(t)$ and λ_i denote the eigenfunctions and eigenvalues of the Karhunen-Loeve expansion of $x(t)$, with corresponding coefficients x_i in the expansion. Then,

$$y_i = \begin{cases} x_i + w_i \sim N(0, 1 + \lambda_i) & \text{if } H_1 \text{ is true} \\ w_i \sim N(0, 1) & \text{if } H_0 \text{ is true} \end{cases} \qquad (9.52)$$

where the coefficients y_i are again independent! Unlike the previous section, every coefficient has some information concerning the difference between the hypotheses. Thus, one cannot find a single coefficient as a sufficient statistic. However, consider the finite vector $\underline{y}_N = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}^T$. A suboptimal hypothesis test can be designed using this finite vector, which only uses the first N coefficients. Note that, as $i \to \infty$, the values of $\lambda_i \to 0$, so that the coefficients y_i have less useful information about the difference in the hypotheses.

The optimal detector using $\underline{y}_N$ can be computed using the likelihood ratio, which, since the y_i are independent, takes the simple form:

$$\mathcal{L}(\underline{y}_N) = \prod_{i=1}^{N} \mathcal{L}(y_i) = \prod_{i=1}^{N} \frac{e^{-y_i^2/2(1+\lambda_i)}}{\sqrt{1+\lambda_i}e^{-y_i^2/2}} \tag{9.53}$$

Taking logarithms, one obtains the following sufficient statistic:

$$S(\underline{y}_N) = \sum_{i=1}^{N} y_i^2 \frac{\lambda_i}{1+\lambda_i} \tag{9.54}$$

Indeed, taking limits as $N \to \infty$, one obtains the sufficient statistic for the original problem

$$s(y) = \sum_{i=1}^{\infty} y_i^2 \frac{\lambda_i}{1+\lambda_i} \tag{9.55}$$

Using this sufficient statistic, an optimal detection threshold can be designed as an extension of the vector case. In practice, it is not useful to evaluate all of the coefficients y_i, as their discrimination value decreases, so that a suboptimal detector using only N coefficients is used.

Note that, since the statistic involves the square of the coefficients y_i, the statistic will not be Gaussian. Thus, for the case of unknown signal in noise, it is useful to derive the appropriate threshold from the original likelihood ratio expression, rather than to compute it using the likelihood ratio of the sufficient statistic.

9.5 Detection of Known Signals in Colored Noise

Consider the problem of detecting the presence of a known signal $s(t)$ in the interval $[0, T]$, with observations

$$y(t) = \begin{cases} s(t) + n(t) + w(t) & \text{if } H_1 \text{ is true} \\ n(t) + w(t) & \text{otherwise} \end{cases} \tag{9.56}$$

where $n(t)$ is a zero-mean Gaussian random process with autocovariance $K_n(t, s)$, and $w(t)$ is white noise with autocovariance $\delta(t - s)$, independent of n.

Again, a solution is possible using series expansions. Let λ_i, $f_i(t)$, and n_i denote the eigenvalues, eigenfunctions, and coefficients of the Karhunen-Loeve expansion of $n(t)$. Then, the observation function $y(t)$ can be transformed using this orthonormal basis into independent coefficients as

$$y_i = \begin{cases} s_i + w_i \sim N(s_i, 1 + \lambda_i) & \text{if } H_1 \text{ is true} \\ w_i \sim N(0, 1 + \lambda_i) & \text{if } H_0 \text{ is true} \end{cases} \tag{9.57}$$

where the coefficients s_i are given by

$$s_i = \int_0^T s(t) f_i(t) dt$$

All of the coefficients y_i contain information which is useful for discrimination. Indeed, as $i \to \infty$, the information in the coefficients may improve, since the variance of the noise in each coefficient becomes smaller!

The likelihood ratio for this case takes the simple form

$$\mathcal{L}(y) = \prod_{i=1}^{\infty} \frac{e^{-(y_i-s_i)^2/2(1+\lambda_i)}}{e^{-y_i^2/2(1+\lambda_i)}}$$

Taking logarithms, one obtains the sufficient statistic

$$S(y) = \sum_{i=1}^{\infty} \frac{s_i y_i}{1 + \lambda_i} \sim \begin{cases} N(\frac{s_i^2}{1+\lambda_i}, \frac{s_i^2}{1+\lambda_i}) & \text{if } H_1 \\ N(0, \frac{s_i^2}{1+\lambda_i}) & \text{otherwise} \end{cases} \tag{9.58}$$

In practice, only a finite set of coefficients is used. However, care must be taken to include enough coefficients so that the approximation of $s(t) \approx \sum_{i=1}^{N} s_i f_i(t)$ is accurate.

Note the effect of the additive white noise is to guarantee that the covariance of the coefficients y_1 is always greater than 1. If the additive white noise were not present, the covariance would approach zero! In this case, it is possible to obtain a perfect detector by observing enough coefficients. This singularity can be avoided by including a small additive white noise in the measurement.

Chapter 10

Estimation of Parameters

10.1 Introduction

In this chapter we consider the problem of estimating or inferring the values of unknown quantities based on observation of a related set of random variables. In Chapter 11 we examine the problem of estimating a stochastic process, while in Chapter 12 we explore recursive estimation. The model of the general parameter estimation situation we are considering is depicted in Figure 10.1. The basic idea is that based on an observation y we want to estimate an unknown quantity x by using an estimation rule $\widehat{x}(y)$. In particular, this model has three components:

1. A model of nature or the parameter space that generates x

2. A model of the observation process as represented by the density $p_{Y|X}(y \mid x)$.

3. An estimation rule mapping each actual observation to a correponding estimate $\widehat{x}(y)$.

Note that this model captures the essential elements of many problems in engineering and science, including: finding the location of a target based on radar observations, estimating the heart rate of a patient from electrical measurements, discerning O^+ density in the atmosphere from brightness measurements, estimating depth in a scene from appearent motion. In all cases let us emphasize that the "estimation rule" is really nothing more than a function that maps each point in the observation space to a corresponding estimate. Thus it can be seen that estimation is closely related to detection. Indeed, the only difference is really in the nature of the variable being estimated. In particular, if the unknown x is discrete valued we generally call it a detection problem, while if x is continuously valued we say we are doing estimation.
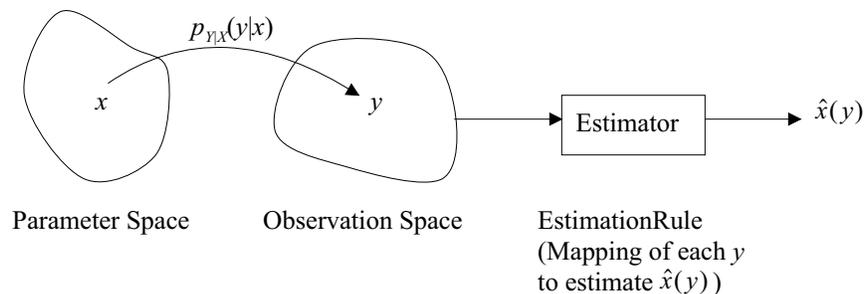
Figure 10.1: Parameter Estimation Problem Components

The first element of our general model in Figure 10.1 is the unknown quantity (or quantities) whose estimate we desire. We denote this element by X (or by the vector $\underline{X}$ if there are a number of such quantities, though we will be lax in this regard). There are two common models for this unknown X which

lead to two distinct, though related, approaches to estimation. In the first model, termed "Bayesian," X is viewed as a random quantity, which leads to what are termed "Baysian" approaches to estimation. The other model views X as unknown, but *nonrandom*, and corresponds to what is variously known as non-random parameter or "Fisher" estimation. Nonrandom parameter estimation is usually accomplished via maximum likelihood estimation. Our primary focus will be on Baysian approaches to estimation, which we discuss in Sections 10.2–10.4. We will discuss nonrandom parameter estimation, and in particular maximum likelihood estimation, in Section 10.6.

10.2   General Bayesian Estimation

In this section we focus on random parameter estimation wherein we model our unknown x as itself being random. Let us examine the three elements of an estimation problem in this light:

1. **Parameter Model:** As we discussed x is modeled as a random variable (or vector). Thus our model of nature is captured by the *prior density* $p_X(x)$ of the random variable X.

2. **Observation Model:** As figure 10.1 indicates, the observation model captures the relationship between the observed quantity y and the unknown x. In a Bayesian problem, this relationship is given by the *conditional density* $p_{Y|X}(y \mid x)$.

3. **Estimation Rule:** In a Bayesian setting the estimation rule is usually obtained by minimizing the expected value of a nonnegative cost function:

$$\widehat{x}^*(y) = \arg\min_{\widehat{x}} E\left[J(\widehat{x}, x)\right] \tag{10.1}$$

where $J(\widehat{x}, x)$ is the cost of estimating $\widehat{x}$ when x is the true value of the unknown and $\arg\min$ denotes the quantity that achieves the minimum (vs the value of the minimum). From a modeling standpoint then this component reduces to choice of an appropriate cost function $J(\widehat{x}, x)$. The quantity $E\left[J(\widehat{x}, x)\right]$ is referred to as the *Bayes risk* for the problem.

Example 10.1
Suppose we wish to estimate the random variable X from noisy observations of the form:

$$Y = x^2 + V, \quad V \sim N(0, 9) \tag{10.2}$$

We wish the estimate to minimize the mean square error: $E\left[(\widehat{X} - X)^2\right]$. In the absence of any other information we believe that X is Gaussian distributed with mean 2 and variance 4.

Given this problem statement, our first problem element (the parameter model) is given by $p_X(x) = N(x; 2, 4)$. We can derive our second element (the observation model) by using the description in (10.2). In particular, it is straightforward to show that (10.2) implies that $p_{Y|X}(y \mid x) = N(y; x^2, 9)$. Finally, since we want to minimize mean square error we would choose our cost function as the square error cost $J(\widehat{x}, x) = (\widehat{x} - x)^2$

10.2.1   General Bayes Decision Rule

Before proceeding to a discussion of the specific estimators we obtain through various choices of the cost function $J(\widehat{x}, x)$, we will derive a general expression for the Bayes decision rule which minimizes the the expected cost and the associated performance of this estimator. Recall that the Bayes estimator is derived

as the minimizer of the expected cost. Thus we have:

$$
\begin{align}
\widehat{x}^*(y) &= \arg\min_{\widehat{x}(y)} E\left[J(\widehat{x}, x)\right] \tag{10.3}\\[6pt]
&= \arg\min_{\widehat{x}(y)} E\left[E\left[J(\widehat{x}, x) \mid y\right]\right] \tag{10.4}\\[6pt]
&= \arg\min_{\widehat{x}(y)} \int E\left[J(\widehat{x}, x) \mid y\right] p_Y(y)\, dy \tag{10.5}\\[6pt]
&= \arg\min_{\widehat{x}(y)} E\left[J(\widehat{x}, x) \mid y\right] \tag{10.6}\\[6pt]
&= \arg\min_{\widehat{x}(y)} \int J(\widehat{x}, x) p_{X|Y}(x \mid y)\, dx \tag{10.7}
\end{align}
$$

In going from (10.3) to (10.4) we have used the properties of iterated expectation. In going from (10.5) to (10.6) we have used the fact that $p_Y(y)$ is always positive and independent of $\widehat{x}$, so that minimization of the conditional expected value in (10.5) for each value of y is the same as minimization of the entire function for all values of y. Recall that the estimator is nothing more than a mapping of each observation to a corresponding estimate, thus what we are saying is that we can do this minimization independently for each y. This situation is similar to what we saw in the case of detection! We can go no further than the general expression (10.7) without knowing more about the cost function $J(\widehat{x}, x)$. We will examine special cases shortly.

10.2.2 General Bayes Decision Rule Performance

Before examining specific examples of Bayes estimators let us examine the performance measures we can use for evaluating these estimators. To this end, let us define the *estimation error* of the estimate $\widehat{x}$ as: $e \equiv x - \widehat{x}$. Now one reasonable way to quantify performance is to focus on the behavior of this error. Note that e itself is a random variable, one which we want "close to zero." We present four performance measures next.

Bias: The *bias* is defined as:

$$
b \equiv E[e] = \iint [x - \widehat{x}(y)]\, p_{X,Y}(x, y)\, dx\, dy \tag{10.8}
$$

It is a measure of the average value of the error, which we want to be small. Note that since b is defined as an expected value over all the random quantities in the problem, it is just a deterministic number for an estimator. Thus it is easy to correct for if it is known. In particular, if b is known, we can create an "unbiased estimator" by just correcting our estimate: $\widehat{x}(y) + b$.

Error Covariance: The *error covariance* is defined by:

$$
\Lambda_e \equiv E\left[(e - b)(e - b)^T\right] = E\left[ee^T\right] - bb^T \tag{10.9}
$$

It is a measure of the variability of the error. Certainly we would like this small. For example, for a Gaussian problem, if the bias is zero and the error covariance zero, then the error would be nonrandom and identically zero!

Mean Square Error: The *mean square error* or "MSE" is defined as:

$$
\text{MSE} \equiv E\left[e^T e\right] = \text{tr}\left(E\left[ee^T\right]\right) = \text{tr}\left(\Lambda_e + bb^T\right) \tag{10.10}
$$

where tr denotes the trace of the matrix (see Appendix C). The MSE is the average of the squared error, and so having it small is certainly good. From the last form of the MSE we can see that it depends on *both* the bias and the variance of the error. When the bias and the variance are both zero the MSE is also zero. In general, we may have to trade off between estimators with different bias and variance to get the smallest MSE. Finally, note that when the bias is zero the MSE is equal to the trace of the error covariance.

Expected Cost: One final and obvious measure of performance is the actual value of the expected cost itself: $E[J(\widehat{x}, x)]$. If you have chosen the cost $J(\widehat{x}, x)$ to have meaning for you, then its average or expected value is certainly one reasonable measure of how well your estimator is doing. Note that *in general* the value of the expected cost is not obviously related to the MSE, the bias, and the variance (though we will see that for certain choices of the cost $J(\widehat{x}, x)$ they are related).

10.3 Bayes Least Square Estimation

We will now start our examination of different choices for the cost function $J(\widehat{x}, x)$ in the Bayesian approach to estimation and see with estimators these choices produce. Our first cost function is the square error cost:

$$J_{\text{BLSE}}(\widehat{x}, x) = (x - \widehat{x})^T(x - \widehat{x}) = \|x - \widehat{x}\|^2 = \sum_{i=1}^{M}(x_i - \widehat{x}_i)^2 \tag{10.11}$$

This cost function is depicted in Figure 10.2 as a function of the error $e = x - \widehat{x}$. Note that with this choice of cost function that $E[J_{\text{BLSE}}(\widehat{x}, x)] = E[e^T e]$ and thus this estimator is the minimimum mean square error estimator (MMSE) by design! Now let us find the estimator. Using the expression (10.7) we obtain:
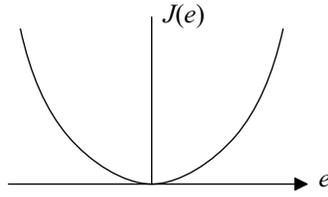

Figure 10.2: Square error cost function

$$\begin{aligned}\widehat{x}_{\text{BLSE}}(y) &= \arg\min_{\widehat{x}(y)} \int J(\widehat{x}, x) p_{X|Y}(x \mid y)\, dx = \arg\min_{\widehat{x}(y)} \int (x - \widehat{x})^T(x - \widehat{x}) p_{X|Y}(x \mid y)\, dx \quad (10.12)\\ &= \arg\min_{\widehat{x}(y)} \int \left[x^T x - \widehat{x}^T x - x^T \widehat{x} + \widehat{x}^T \widehat{x}\right] p_{X|Y}(x \mid y)\, dx = f(\widehat{x})\end{aligned}$$

Note that $\widehat{x}$ is a constant with respect to the integral. Now we can take the derivative of (10.12) with respect to $\widehat{x}$, set it equal to zero, and solve for $\widehat{x}$. We also need to check that the second derivative is positive at the solution to verify it is really a minimum.

$$\begin{aligned}\frac{\partial}{\partial \widehat{x}}\left\{\int (x - \widehat{x})^T(x - \widehat{x}) p_{X|Y}(x \mid y)\, dx\right\} &= \int \left[0 - x - x + 2\widehat{x}\right] p_{X|Y}(x \mid y)\, dx\\ &= -2\int x p_{X|Y}(x \mid y)\, dx + 2\widehat{x}\\ &= 0\end{aligned}$$

$$\frac{\partial^2}{\partial \widehat{x}^2}\left\{\int (x - \widehat{x})^T(x - \widehat{x}) p_{X|Y}(x \mid y)\, dx\right\} = 2 > 0$$

The resulting estimate is given by:

$$\boxed{\widehat{x}_{\text{BLSE}}(y) = \int x p_{X|Y}(x \mid y)\, dx = E\left[x \mid y\right]} \tag{10.13}$$

Thus the Bayes least square error estimate is just the conditional mean, i.e. the mean of the conditional density $p_{X|Y}(x \mid y)$. Note that this density is a function of y, so the estimate is as well. The Bayes least

square error estimate (BLSE) is sometimes also referred to as the Bayes minimum mean square estimate (MMSE).

Let us now examine performance of the BLSE estimator. First let us find the bias:

$$b = E\left[x - \widehat{x}_{\text{BLSE}}(y)\right] = E[x] - E\left[E\left[x \mid y\right]\right] = 0 \tag{10.14}$$

Thus the BLSE estimator is always unbiased. Independent of the prior density or the observation density the bias will always be zero. This is a very good thing.

Now let us examine the error covariance. This quantity is given by:

$$
\begin{aligned}
\Lambda_{\text{BLSE}} &= E\left[(e - b)(e - b)^T\right] = E\left[ee^T\right] = E\left[\left(x - \widehat{x}_{\text{BLSE}}(y)\right)\left(x - \widehat{x}_{\text{BLSE}}(y)\right)^T\right] \tag{10.15} \\
&= E\left[E\left\{(x - E[x \mid y])\left(x - E[x \mid y]\right)^T \mid y\right\}\right] \tag{10.16} \\
&= E\left[\Lambda_{x|y}(y)\right] \tag{10.17}
\end{aligned}
$$

Thus, the BLSE error covariance is the expected value of the conditional covariance. Note that in general the conditional covariance is a function of y, the observation. Thus the expectation in (10.17) is over the random variable y.

Now let us find the mean square error (MSE). Note that in this particular case (though not in general!) the MSE is the *same* as the value of the expected cost $E[J_{\text{BLSE}}(\widehat{x}, x)]$. Since the bias is zero, the MSE is the same as the trace of the error covariance:

$$\text{MSE} = E[J_{\text{BLSE}}(\widehat{x}, x)] = \text{tr}\left(\Lambda_{\text{BLSE}}\right) = \text{tr}\left(E\left[\Lambda_{x|y}(y)\right]\right) \tag{10.18}$$

Note that the MSE does not depend on y (due to the expectation operation) while the conditional covariance does in general.

In general, finding the BLSE and its associated performance is difficult, as the conditional density must first be found. In certain special cases it can be done however. Let us consider some examples.

Example 10.2
In this example we wish to estimate x by observing a related random variable y, where the random variables X and Y are jointly distributed with the density shown in Figure 10.3. This density is uniform over the depicted diamond shaped region. Note that this characterization provides all the information to find both a prior model for X (i.e. the marginal distribution $p_x(X)$) as well as the relationship between Y and X as given by $p_{Y|X}(y \mid x)$.
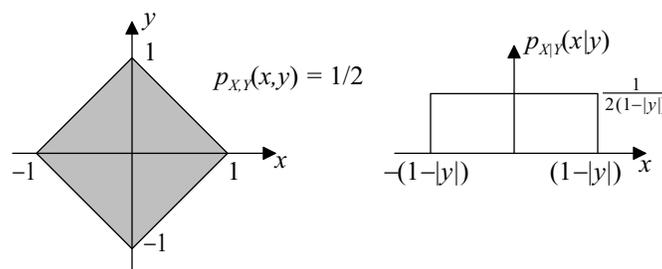

Figure 10.3: BLSE Example

To find the BLSE estimate for this problem the quantity we need to find is the conditional density $p_{X|Y}(x \mid y)$, from which we can find $E[x \mid y]$ and $\lambda_{x|y}$. To find $p_{X|Y}(x \mid y)$ we could use Bayes rule, but the geometry of the problem allows us to do this almost by inspection. Recall that $p_{X|Y}(x \mid y)$ will be a slice of the joint density $p_{X,Y}(x, y)$ parallel to the x-axis and scaled to have unit area. This conditional density is shown on the right in Figure 10.3 for any nontrivial value of y. Since the original density is "flat," each slice will be flat, so all we really need to determine are the edges. The height follows from the constraint that the density has unit area.

Now, given this density it is easy to see that the BLSE, which is the mean of the density, is given by $\widehat{x}_{\text{BLSE}}(y) = E[x \mid y] = 0$. Thus for this example the BLSE of x does not depend on y. Note that this is true in spite of the fact that X and

Y are clearly dependent random variables! This may seem strange, but remember that the BLSE is based on the *mean* of the conditional density. Next let us find the conditional variance $\lambda_{x|y}$. From the density $p_{X|Y}(x \mid y)$ this variance is easily found from the box like form of the conditional density to be $\lambda_{x|y} = (1-y)^2/3$. Note that the conditional variance depends on y. Finally let us find the MSE for the BLSE estimator. This can be shown to be MSE $= E\left[\lambda_{x|y}\right] = 1/6$. Note the MSE is independent of y, as it must be.

Example 10.3

In this example we wish to estimate x by observing a related random variable y, where the random variables X and Y are jointly distributed with density given by:

$$p_{X,Y}(x,y) = \begin{cases} 6x & 0 \le x \le y,\ 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{10.19}$$

Again, this characterization provides all the information to find both a prior model for X (i.e. the marginal distribution $p_x(X)$) as well as the relationship between Y and X as given by $p_{Y|X}(y \mid x)$.

To find the BLSE estimate for this problem we need to find the conditional density $p_{X|Y}(x \mid y)$. By integrating $p_{X,Y}(x,y)$ with respect to y we can find the marginal density for y:

$$p_Y(y) = \int_0^y 6x\,dx = 3y^2, \qquad 0 \le y \le 1. \tag{10.20}$$

Now we can use Bayes' rule to obtain the conditional density:

$$p_{X|Y}(x \mid y) \quad = \quad \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{2x}{y^2}, \qquad 0 \le x \le y \tag{10.21}$$

The mean of the conditional density is now found as:

$$E[x \mid y] \quad = \quad \int_{-\infty}^{\infty} x\,p_{X|Y}(x \mid y)\,dx = \int_0^y \frac{2x^2}{y^2}\,dx = \frac{2}{3}\frac{x^3}{y^2}\Big|_0^y = \frac{2}{3}y \tag{10.22}$$

Thus $\widehat{x}_{\mathsf{BLSE}}(y) = \frac{2}{3}y$.

Next let us find the conditional variance $\lambda_{x|y}$:

$$\lambda_{x|y} \quad = \quad E[x^2 \mid y] - E[x \mid y]^2 = \int_{-\infty}^{\infty} x^2\,p_{X|Y}(x \mid y)\,dx - E[x \mid y]^2 \tag{10.23}$$

$$= \quad \int_0^y \frac{2x^3}{y^2}\,dx - \frac{4}{9}y^2 = \frac{y^2}{2} - \frac{4y^2}{9} = \frac{y^2}{18} \tag{10.24}$$

Note that $\lambda_{x|y}$ is a function of y. Finally, the mean square error is obtained as:

$$\mathsf{MSE} \quad = \quad E\left[\lambda_{x|y}\right] = \int_{-\infty}^{\infty} \lambda_{x|y} p_Y(y)\,dy = \int_0^1 \frac{3y^4}{18}\,dy = \frac{1}{30} \tag{10.25}$$

Example 10.4

Suppose x and y are related by the following joint density function:

$$p_{X,Y}(x,y) = \begin{cases} 10x & 0 \le x \le y^2,\ 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{10.26}$$

To find the BLSE estimate for this problem we need to find the conditional density $p_{X|Y}(x \mid y)$. By integrating $p_{X,Y}(x,y)$ with respect to y we can find the marginal density for y:

$$p_Y(y) = \int_0^{y^2} 10x\,dx = 5y^4, \qquad 0 \le y \le 1. \tag{10.27}$$

Now we can use Bayes' rule to obtain the conditional density:

$$p_{X|Y}(x \mid y) \quad = \quad \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{10x}{5y^4} = \frac{2x}{y^4}, \qquad 0 \le x \le y^2 \tag{10.28}$$

The mean of the conditional density is now found as:

$$E[x \mid y] \quad = \quad \int_{-\infty}^{\infty} x\,p_{X|Y}(x \mid y)\,dx = \int_0^{y^2} \frac{2x^2}{y^4}\,dx = \frac{2}{3}\frac{x^3}{y^4}\Big|_0^{y^2} = \frac{2}{3}y^2 \tag{10.29}$$

Thus $\widehat{x}_{\mathsf{BLSE}}(y) = \frac{2}{3}y^2$. Note that this estimate is a *nonlinear* function of y in this case.

Next let us find the conditional variance $\lambda_{x|y}$:

$$\lambda_{x|y} = E[x^2 \mid y] - E[x \mid y]^2 = \int_{-\infty}^{\infty} x^2\, p_{X|Y}(x \mid y)\, dx - E[x \mid y]^2 \tag{10.30}$$

$$= \int_0^{y^2} \frac{2x^3}{y^4}\, dx - \frac{4}{9}y^4 = \frac{y^4}{2} - \frac{4y^4}{9} = \frac{y^4}{18} \tag{10.31}$$

Note that $\lambda_{x|y}$ is a function of y. Finally, the mean square error is obtained as:

$$\mathsf{MSE} = E\left[\lambda_{x|y}\right] = \int_{-\infty}^{\infty} \lambda_{x|y} p_Y(y)\, dy = \int_0^1 \frac{y^4}{18}(5y^4)\, dy = \frac{5}{162} = 0.0309 \tag{10.32}$$

Example 10.5 (Scalar Gaussian Case)

Suppose we wish to estimate x by observing a related random variable y, where the random variables X and Y are jointly Gaussian scalar random variables. Thus the random variables X and Y are completely characterized by their joint Gaussian distribution:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} m_x \\ m_y \end{bmatrix}, \begin{bmatrix} \lambda_x & \lambda_{xy} \\ \lambda_{yx} & \lambda_y \end{bmatrix} \right) \tag{10.33}$$

Again, note that from this characterization a prior model for X could be found as well as the relationship between Y and X provided by the density $p_{Y|X}(y \mid x)$.

To find the BLSE estimate for this problem we again need to find the conditional density $p_{X|Y}(x \mid y)$, from which we can find $E[x \mid y]$ and $\lambda_{x|y}$. We proceed by manipulating the expression for the conditional density until we have it in a Gaussian form, then simply read off the mean and variance of this Gaussian.

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} \tag{10.34}$$

$$\propto p_{X,Y}(x, y) \tag{10.35}$$

$$= \exp\left\{ -\frac{1}{2} \begin{bmatrix} x - m_x \\ y - m_y \end{bmatrix}^T \begin{bmatrix} \lambda_x & \lambda_{xy} \\ \lambda_{yx} & \lambda_y \end{bmatrix}^{-1} \begin{bmatrix} x - m_x \\ y - m_y \end{bmatrix} \right\} \tag{10.36}$$

$$= \exp\left\{ -\frac{1}{2} \frac{\left[x - \left(m_x + \frac{\lambda_{xy}}{\lambda_y}(y - m_y)\right)\right]^2}{\left(\lambda_x - \frac{\lambda_{xy}^2}{\lambda_y}\right)} \right\} \tag{10.37}$$

First we use Bayes rule to express the conditional density in terms of the joint density, scaled by $p_Y(y)$. For a given y this is just a normalization. Thus we concentrate on just the exponential component of the density. The last line shows that the form of this density is again a Gaussian, with mean given by the numerator term in parentheses and variance given by the denominator term. We directly obtain the BLSE and the conditional variance as:

$$\widehat{x}_{\mathsf{BLSE}}(y) = E[x \mid y] = m_x + \frac{\lambda_{xy}}{\lambda_y}(y - m_y) \tag{10.38}$$

$$\lambda_{x|y} = \lambda_x - \frac{\lambda_{xy}^2}{\lambda_y} \tag{10.39}$$

Let us make some interesting observations. First, note that the estimate is a *linear function* of the observation in the Gaussian case! This is not true in general. In addition, the conditional density $\lambda_{x|y}$ is independent of y for the Gaussian case. Thus, for the Gaussian case the MSE is the same as the conditional density (and the error variance): $\mathsf{MSE} = E[\lambda_{x|y}] = \lambda_{x|y}$. These are yet other ways in which Gaussians are special.

The structure of the BLSE for the Gaussian case has an intuitively pleasing form. For example, if the cross-correlation λ_{xy} between X and Y is zero, then X and Y are independent since they are jointly Gaussian. In this case, note that the BLSE reduces to the prior mean: $\widehat{x}_{\mathsf{BLSE}}(y) = m_x$, which is independ of y. In this case, observations do not help us. Similarly, as the observations become more variable, i.e. as $\lambda_y \to \infty$ we have that $\widehat{x}_{\mathsf{BLSE}}(y) \to m_x$. In other words, we ignore the data. Conversely, as the statistical tie between X and Y becomes larger (i.e. as λ_{xy} increases relative to λ_y) more weight is placed on the observation relative to the prior mean. Finally, note that the conditional variance (which is also the MSE and the error variance for this case) is *reduced* relative to the prior variance λ_x. In particular, we can write $\lambda_{\mathsf{BLSE}} = \lambda_{x|y} \le \lambda_x$.

Example 10.6 (Vector Gaussian Case)

Now let us examine the vector Gaussian case. In particular, suppose that $\underline{x}$ and $\underline{y}$ are joinly Gaussian random vectors with mean vectors $\underline{m}_x$, $\underline{m}_y$ and covariance matrices Λ_x, Λ_y respectively and cross-covariance matrix Λ_{xy}. Again, we can manipulate the joint density to obtain an expression for the conditional density, and from this density we can find the conditional mean and variance needed to find the BLSE. Using Bayes' rule the conditional density is given by:

$$p_{\underline{X}|\underline{Y}}(\underline{x} \mid \underline{y}) = \frac{p_{\underline{X},\underline{Y}}(\underline{x}, \underline{y})}{p_{\underline{Y}}(\underline{y})} \tag{10.40}$$

Assume that $\underline{x}$ is n-dimensional, and $\underline{y}$ is m-dimensional. In this case, the conditional probability density becomes:

$$p_{\underline{X}|\underline{Y}}(\underline{x} \mid \underline{y}) = (\sqrt{2\pi})^{m-n} \frac{\det \Lambda_y}{\det \begin{bmatrix} \Lambda_x & \Lambda_{xy} \\ \Lambda_{xy}^T & \Lambda_y \end{bmatrix}} \frac{e^{-1/2 \begin{bmatrix} \underline{x}-\underline{m}_x \\ \underline{y}-\underline{m}_y \end{bmatrix}^T \begin{bmatrix} \Lambda_x & \Lambda_{xy} \\ \Lambda_{xy}^T & \Lambda_y \end{bmatrix}^{-1} \begin{bmatrix} \underline{x}-\underline{m}_x \\ \underline{y}-\underline{m}_y \end{bmatrix}}}{e^{-1/2(\underline{y}-\underline{m}_y)^T \Lambda_y^{-1}(\underline{y}-\underline{m}_y)}} \tag{10.41}$$

The constant in front of the exponential ratio can be ignored, since it is merely a normalization factor to insure that the resulting density integrates to 1. The important term to focus on is the ratio of the exponentials. In order to understand this ratio, one needs a formula for the inverse of the joint covariance of $[\underline{x}, \underline{y}]^T$, which we do next. The following matrix identity will prove useful:

$$\begin{bmatrix} I & -\Lambda_{xy}\Lambda_y^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Lambda_x & \Lambda_{xy} \\ \Lambda_{xy}^T & \Lambda_y \end{bmatrix} = \begin{bmatrix} \Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T & 0 \\ \Lambda_{xy}^T & \Lambda_y \end{bmatrix} \tag{10.42}$$

By inverting the block triangular matrices we obtain the formula:

$$\begin{bmatrix} \Lambda_x & \Lambda_{xy} \\ \Lambda_{xy}^T & \Lambda_y \end{bmatrix}^{-1} = \begin{bmatrix} \Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T & 0 \\ \Lambda_{xy}^T & \Lambda_y \end{bmatrix}^{-1} \begin{bmatrix} I & -\Lambda_{xy}\Lambda_y^{-1} \\ 0 & I \end{bmatrix} \tag{10.43}$$

$$= \begin{bmatrix} \left(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T\right)^{-1} & 0 \\ -\Lambda_y^{-1}\Lambda_{xy}^T \left(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T\right)^{-1} & \Lambda_y^{-1} \end{bmatrix} \begin{bmatrix} I & -\Lambda_{xy}\Lambda_y^{-1} \\ 0 & I \end{bmatrix} \tag{10.44}$$

Thus, the desired inverse is given by:

$$\begin{bmatrix} \Lambda_x & \Lambda_{xy} \\ \Lambda_{xy}^T & \Lambda_y \end{bmatrix}^{-1} = \begin{bmatrix} \left(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T\right)^{-1} & -\left(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T\right)^{-1}\Lambda_{xy}\Lambda_y^{-1} \\ -\Lambda_y^{-1}\Lambda_{xy}^T \left(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T\right)^{-1} & \Lambda_y^{-1} + \Lambda_y^{-1}\Lambda_{xy}^T \left(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T\right)^{-1}\Lambda_{xy}\Lambda_y^{-1} \end{bmatrix} \tag{10.45}$$

With the above inverse, one can now compute the exponent of the exponential fraction in (10.41), as

$$\frac{e^{-1/2 \begin{bmatrix} \underline{x}-\underline{m}_x \\ \underline{y}-\underline{m}_y \end{bmatrix}^T \begin{bmatrix} \Lambda_x & \Lambda_{xy} \\ \Lambda_{xy}^T & \Lambda_y \end{bmatrix}^{-1} \begin{bmatrix} \underline{x}-\underline{m}_x \\ \underline{y}-\underline{m}_y \end{bmatrix}}}{e^{-1/2(\underline{y}-\underline{m}_y)^T \Lambda_y^{-1}(\underline{y}-\underline{m}_y)}} \tag{10.46}$$

$$= \exp\left\{ 1/2(\underline{y}-\underline{m}_y)^T\Lambda_y^{-1}(\underline{y}-\underline{m}_y) - 1/2(\underline{x}-\underline{m}_x)^T(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T)^{-1}(\underline{x}-\underline{m}_x) \right.$$
$$+ (\underline{x}-\underline{m}_x)^T(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T)^{-1}\Lambda_{xy}\Lambda_y^{-1}(\underline{y}-\underline{m}_y)$$
$$\left. - 1/2(\underline{y}-\underline{m}_y)^T \left[\Lambda_y^{-1} + \Lambda_y^{-1}\Lambda_{xy}^T(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T)^{-1}\Lambda_{xy}\Lambda_y^{-1}\right](\underline{y}-\underline{m}_y) \right\} \tag{10.47}$$

$$= \exp\left\{ -1/2(\underline{x}-\underline{m}_x)^T(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T)^{-1}(\underline{x}-\underline{m}_x) + (\underline{x}-\underline{m}_x)^T(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T)^{-1}\Lambda_{xy}\Lambda_y^{-1}(\underline{y}-\underline{m}_y) \right.$$
$$\left. - 1/2(\underline{y}-\underline{m}_y)^T\Lambda_y^{-1}\Lambda_{xy}^T(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T)^{-1}\Lambda_{xy}\Lambda_y^{-1}(\underline{y}-\underline{m}_y) \right\} \tag{10.48}$$

$$= e^{-1/2\left(\underline{x}-\underline{m}_x - \Lambda_{xy}\Lambda_y^{-1}(\underline{y}-\underline{m}_y)\right)^T \left(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T\right)^{-1}\left(\underline{x}-\underline{m}_x - \Lambda_{xy}\Lambda_y^{-1}(\underline{y}-\underline{m}_y)\right)} \tag{10.49}$$

Thus the conditional density of $\underline{x}$ given $\underline{y}$ is again Gaussian (which it must be, since they are jointly Gaussian), with an exponent given by (10.49). Now we know this conditional Gaussian distribution must be of the following general form: form of

$$p_{\underline{X}|\underline{Y}}(\underline{x} \mid \underline{y}) \propto e^{-1/2\left(\underline{x}-E[\underline{x}|\underline{y}]\right)^T \Lambda_{x|y}^{-1}\left(\underline{x}-E[\underline{x}|\underline{y}]\right)} \tag{10.50}$$

By identifying similar terms between (10.49) and (10.50) we immediately find that:

$$E\left[\underline{x} \mid \underline{y}\right] = \underline{m}_x + \Lambda_{xy}\Lambda_y^{-1}(\underline{y} - \underline{m}_y) \tag{10.51}$$

$$\Lambda_{x|y} = \Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T \tag{10.52}$$

The first of these provides the BLSE for the vector Gaussian problem:

$$\widehat{\underline{x}}_{\text{BLSE}}(\underline{y}) = E\left[\underline{x} \mid \underline{y}\right] = \underline{m}_x + \Lambda_{xy}\Lambda_y^{-1}(\underline{y} - \underline{m}_y) \tag{10.53}$$

Note the similarity between this expression and the BLSE for the scalar Gaussian case given in (10.39). Similarly, the conditional variance is given in (10.52), which does not depend on $\underline{y}$! Again, for general distributions the conditional covariance will depend on the value observed, but for jointly Gaussian random vectors, this covariance is constant! As a result, for the vector Gaussian case the MSE is simply the trace of the conditional covariance:

$$\text{MSE} = \text{tr}\left[\Lambda_{x|y}\right] \tag{10.54}$$

$$= \text{tr}\left[\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T\right] \tag{10.55}$$

The above formulas provide explicit equations for the BLSE estimator (which, recall is the MMSE estimator) and associated error for the case of jointly Gaussian random vectors. Note again that the estimator is a *linear function* of $\underline{y}$.

Let us close by summarizing the properties of BLSE estimates:

- The BLSE estimate is the conditional mean $E[x \mid y]$.

- The BLSE estimate is always unbiased.

- The BLSE estimate is always the MMSE estimate.

- In general, the BLSE is a nonlinear function of the data.

- For jointly Gaussian problems only, the BLSE estimate is linear and the conditional variance is independent of the data.

- In general, finding the BLSE estimate requires finding the conditional density, and thus is challenging.

10.4 Bayes Maximum A Posteriori (MAP) Estimation

In this section we will examine Bayes' estimation with a different choice of cost function. In particular, we now focus on the "uniform cost" function given by:

$$J_{\text{MAP}}(\widehat{x}, x) = \begin{cases} 1 & |x - \widehat{x}| \geq \epsilon \\ 0 & |x - \widehat{x}| < \epsilon \end{cases} \tag{10.56}$$

This cost function is depicted in Figure 10.4 as a function of the error $e = x - \widehat{x}$. Note that this cost function treats all errors as equally bad, no matter how large. It is reminiscent of the "0-1" cost structure we saw in our study of detection. You might expect that we will obtain similar estimates, and you will not be disappointed.
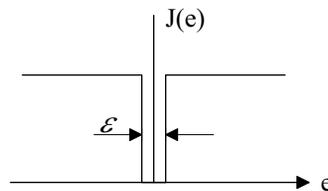

Figure 10.4: Uniform or MAP cost function.

Now let us find the Bayesian estimator corresponding to the uniform cost function. As before we start with the general Bayes' estimator definition (10.7) and go from there:

$$\widehat{x}_{\text{MAP}}(y) \quad = \quad \arg\min_{\widehat{x}(y)} \int J_{\text{MAP}}(\widehat{x}, x)\, p_{X|Y}(x \mid y)\, dx \tag{10.57}$$

$$= \quad \arg\min_{\widehat{x}(y)} \int_{\{x \,|\, |x-\widehat{x}| \geq \epsilon\}} p_{X|Y}(x \mid y)\, dx \tag{10.58}$$

$$= \quad \arg\min_{\widehat{x}(y)} \left[1 - \int_{\{x \,|\, |x-\widehat{x}| < \epsilon\}} p_{X|Y}(x \mid y)\, dx \right] \tag{10.59}$$

Now the the integral in (10.59) is over an infinitely small "gap" around $\widehat{x}$, so that the overall expression is minimized by placing the gap centered at $\widehat{x}$ at the *maximum* of the conditional density. The geometric situation is depicted in Figure 10.5. The shaded area depicts the right hand side in (10.57). Note that this
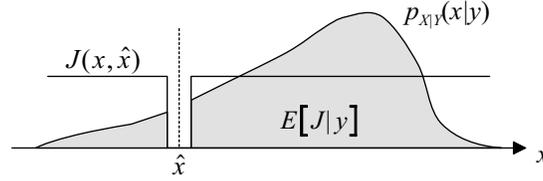
Figure 10.5: Illustration of geometry behind MAP estimate derivation.

is given by the area of the conditional density *minus* the area centered around $\widehat{x}$. We minimize the shaded area by placing the "gap" at the peak of the conditional density. This observation yields:

$$\boxed{\widehat{x}_{\text{MAP}}(y) = \arg\max_{x} p_{X|Y}(x \mid y)} \tag{10.60}$$

Thus we have the result that the optimal Bayes' estimate corresponding to the uniform cost structure in (10.56) is the value of x that maximizes the conditional density $p_{X|Y}(x \mid y)$. Since this density can be thought of as the density obtained for x *after* having observed y, this conditional density is often referred to as the "posterior density," and for this reason the corresponding estimate is referred to as the "Maximum A Posteriori" or MAP estimate. Note that whereas the BLSE estimate was the *mean* of $p_{X|Y}(x \mid y)$, the MAP estimate is the peak or "mode" of this density. Evidentally the conditional density $p_{X|Y}(x \mid y)$ plays a key role in both estimators. Note that, in general the mean and mode of a density can be different, so the BLSE and MAP estimates will be different in general.

The definition of the MAP estimate given in (10.60) is the fundamental one. For differentiable densities we can characterize the potential maxima of the density as the locations where the derivative is zero and the second derivative is negative. This approach leads to a common characterization of the MAP estimate we will derive next. First, using Bayes' rule and the monotonic properties of the natural logarithm we can rewrite the MAP estimate as follows:

$$\widehat{x}_{\text{MAP}}(y) \quad = \quad \arg\max_{x} p_{X|Y}(x \mid y) = \arg\max_{x} \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_{Y|X}(y \mid x) p_X(x)}{p_Y(y)} \tag{10.61}$$

$$= \quad \arg\max_{x} p_{Y|X}(y \mid x) p_X(x) = \arg\max_{x} \ln\left[p_{Y|X}(y \mid x) p_X(x) \right] \tag{10.62}$$

$$= \quad \arg\max_{x} \left(\ln\left[p_{Y|X}(y \mid x) \right] + \ln\left[p_X(x) \right] \right) \tag{10.63}$$

Now the MAP estimate is obtained as the maximum of the expression in parentheses in (10.63). A necessary condition for the maximum is that the derivative of this expression with respect to x be zero. Thus the MAP estimate must satisfy the following equation (if it exists!):

$$\boxed{\left. \frac{\partial \ln\left[p_{Y|X}(y \mid x) \right]}{\partial x} + \frac{\partial \ln\left[p_X(x) \right]}{\partial x} \right|_{x=\widehat{x}_{MAP}(y)} = 0} \tag{10.64}$$

This equation is sometimes referred to as the "MAP equation."

Before proceeding to some examples, note that, unlike the BLSE estimator, there are no particularly nice general expressions for either the bias or the variance of the MAP estimator. In particular, the general MAP estimator can be biased and will not be the MMSE estimator. So why do MAP estimation? One reason is that the structure of the estimator given as expressed in (10.62) rationally and, some would argue, naturally combines both an observation model (the term $p_{Y|X}(y \mid x)$) and a prior model $p_X(x)$. Another reason is that often maximizing a function is easier than averaging, which requires weighted integration of some sort. This idea of finding a solution by maximizing a function appears throughout engineering and outside of stochastic concerns. Let us look at some examples of MAP estimation.

Example 10.7

In this example let us revisit the problem of Example 10.2, but this time seek the MAP estimate. We again need the conditional density $p_{X|Y}(x \mid y)$, which is already given in Figure 10.3. Now for the MAP estimate we seek the value of x at which this density is maximum. Inspection of Figure 10.3 will show that the maximum is found for a whole range of values of x! Thus for this case the MAP estimate is *not unique* and $\widehat{x}_{MAP}(y)$ is any x in the interval $[-(1-|y|),\ (1-|y|)]$. Note that any x in this range will produce the same value for the expected risk or cost. While this is true, also note that different choices of this value will have different MSEs in general. For example, the BLSE is one consistent choice of x, which will then have minimum MSE. Suppose instead for $\widehat{x}_{MAP}(y)$ that we always choose right end of the interval so that $\widehat{x}_{MAP}(y) = (1-|y|)$. The MSE for this latter choice is:

$$\text{MSE} \;=\; E\left[(x - \widehat{x}_{MAP}(y))^2\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - (1-|y|))^2\, p_{X,Y}(x,y)\, dx\, dy \tag{10.65}$$

$$=\; \int_0^1 \int_{-1+y}^{1-y} (x - (1-y))^2\, \frac{1}{2}\, dx\, dy + \int_{-1}^0 \int_{-1-y}^{1+y} (x - (1+y))^2\, \frac{1}{2}\, dx\, dy \tag{10.66}$$

$$=\; \frac{2}{3} \tag{10.67}$$

Note that this MSE is larger than the MSE of 1/6 we found for the BLSE estimator.

Example 10.8

For this example let us revisit the problem of Example 10.3, but again seek the MAP estimate. We have already calculated the conditional density in (10.21), which is given by:

$$p_{X|Y}(x \mid y) \;=\; \frac{2x}{y^2}, \qquad 0 \le x \le y \tag{10.68}$$

The maximum of this conditional density occurs at $x = y$. Therefore:

$$\widehat{x}_{\text{MAP}}(y) = y \tag{10.69}$$

Note that in this case the MAP estimate is both unique, and different from the BLSE estimate.

Now we know the BLSE is an unbiased estimator. What about the MAP estimate? The bias for this example can be found as:

$$b \;=\; E\left[x - \widehat{x}_{\text{MAP}}(y)\right] = E[x] - E[y] \tag{10.70}$$

$$=\; \int_{-\infty}^{\infty} x\, p_X(x)\, dx - \int_{-\infty}^{\infty} y\, p_Y(y)\, dy = \int_0^1 x\, 6x(1-x)\, dx - \int_0^1 y\, 3y^2\, dy = \frac{1}{2} - \frac{3}{4} = -\frac{1}{4} \tag{10.71}$$

Thus the bias is not necessarily 0 for the MAP estimate.

We can again show that the mean square error is higher than that for the BLSE estimator by direct calculation:

$$\text{MSE} \;=\; E\left[(x - \widehat{x}_{MAP}(y))^2\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \widehat{x}_{MAP}(y))^2\, p_{X,Y}(x,y)\, dx\, dy \tag{10.72}$$

$$=\; \int_0^1 \int_0^y (x-y)^2\, 6x\, dx\, dy \tag{10.73}$$

$$=\; \frac{1}{2} \tag{10.74}$$

which is greater than the MSE of $1/30$ associated with the BLSE estimate which we found in (10.25).

Example 10.9

For this example let us revisit the problem of Example 10.4, but again seek the MAP estimate. We have already calculated the conditional density in (10.28), which is given by:

$$p_{X|Y}(x \mid y) \quad = \quad \frac{2x}{y^4}, \qquad 0 \le x \le y^2 \tag{10.75}$$

The maximum of this conditional density occurs at the endpoint of the interval $x = y^2$. Therefore:

$$\widehat{x}_{\mathsf{MAP}}(y) = y^2 \tag{10.76}$$

Note that in this case the MAP estimate is both unique, and different from the BLSE estimate. Also note that since the maximum is at the endpoint of the interval, equation (10.64) cannot be used to find the density in this case.

We can again show that the mean square error is higher than that for the BLSE estimator by direct calculation:

$$\mathsf{MSE} \quad = \quad E\left[(x - \widehat{x}_{MAP}(y))^2\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \widehat{x}_{MAP}(y))^2 \, p_{X,Y}(x,y) \, dx \, dy \tag{10.77}$$

$$= \quad \int_0^1 \int_0^{y^2} \left(x - y^2\right)^2 10x \, dx \, dy \tag{10.78}$$

$$= \quad \frac{5}{54} = 0.0926 \tag{10.79}$$

which is greater than the MSE associated with the BLSE estimate which we found in (10.32).

Example 10.10 (Gaussian Case)

Let us now examine the problems of Examples 10.5 and 10.6 with regard to the MAP estimate. Recall that the conditional density for a jointly Gaussian problem is again a Gaussian density, and e.g. is proportional to the expression in given in (10.37). Being a Gaussian, the conditional density has a single maximum which occurs at the same place as its mean. Therefore in the case of jointly Gaussian densities the MAP and BLSE estimates are *identical* with identical MSE!

$$\widehat{x}_{\mathsf{MAP}}(y) = \widehat{x}_{\mathsf{BLSE}}(y) = m_x + \frac{\lambda_{xy}}{\lambda_y}(y - m_y) \tag{10.80}$$

More generally, with a little thought we can see that the BLSE and MAP estimates will be the same whenever $p_{X|Y}(x \mid y)$ is symmetric and unimodal (i.e. has a single maximum), since in these cases the mean and maximum of the density are one and the same.

Example 10.11 (Gaussian Problems with Linear Observations)

Let us now focus on a particularly important problem: MAP estimation for problems with linear observations, Gaussian densities, and vector state. In particular, assume we have the following general problem, wherein our noisy vector observation $\underline{y}$ is linearly related to our quantity of interest $\underline{x}$, which itself is Gaussian:

$$\underline{y} \quad = \quad C\underline{x} + \underline{w}, \qquad \underline{w} \sim N(0, R) \tag{10.81}$$

$$\underline{x} \quad \sim \quad N(0, Q) \tag{10.82}$$

where $\underline{y} = [y_1, \cdots, y_N]^T$, $\underline{x} = [x_1, \cdots, x_N]^T$, $\underline{w} = [w_1, \cdots, w_N]^T$, R is the covariance matrix of the observation noise, $\underline{w}$ is independent of $\underline{x}$, and Q is the covariance matrix of $\underline{x}$.

In this problem we can see with a bit of thought that $\underline{x}$ and $\underline{y}$ will be jointly Gaussian random vectors, so the results of Example 10.10 apply and we know the MAP estimate will be given by the conditional mean. Using the fomulas (10.51) and (10.52) we obtain:

$$\widehat{\underline{x}}_{MAP} \quad = \quad QC^T \left(CQC^T + R\right)^{-1} \underline{y}$$

$$\Lambda_{MAP} \quad = \quad \Lambda_{x|y} = Q - QC^T \left(CQC^T + R\right)^{-1} CQ$$

While this result is certainly correct, we may also derive an alternative expression for the MAP estimate based directly on the MAP equation. This alternative result is widely used, so worth deriving. To this end note that $p_{Y|X}(\underline{y}|\underline{x}) = N(\underline{y}; C\underline{x}, R)$, since $\underline{w}$ is independent of $\underline{x}$, thus:

$$\widehat{\underline{x}}_{\mathsf{MAP}} \quad = \quad \arg\max_{\underline{x}} \; p_{Y|X}(\underline{y}|\underline{x}) \, p_X(\underline{x}) = \arg\max_{\underline{x}} \; \ln[p_{Y|X}(\underline{y}|\underline{x})] + \ln[p_X(\underline{x})] \tag{10.83}$$

$$= \quad \arg\max_{\underline{x}} \; -(\underline{y} - C\underline{x})^T R^{-1}(\underline{y} - C\underline{x}) - \underline{x}^T Q^{-1}\underline{x} \tag{10.84}$$

$$= \quad \arg\min_{\underline{x}} \; \left\|\underline{y} - C\underline{x}\right\|_{R^{-1}}^2 + \left\|\underline{x}\right\|_{Q^{-1}}^2 \tag{10.85}$$

$$= \quad \arg\min_{\underline{x}} \; \underline{y}^T R^{-1}\underline{y} - 2\underline{x}^T C^T R^{-1}\underline{y} + \underline{x}^T C^T R^{-1}C\underline{x} + \underline{x}^T Q^{-1}\underline{x} \tag{10.86}$$

In going from (10.83) to (10.84) we have simply inserted the densities in question and eliminated any constants not affecting the optimizations. In going from (10.84) to (10.85) we have simply switched from a maximization to a minimization by eliminating the leading minus sign and we have written the quadratic forms in (10.84) as weighted norms. Note that when the MAP problem is written in the form (10.85) we can easily see its relationship to least square minimization. The expression (10.86) is obtained by multiplying out the quadratic forms in (10.84) or (10.85).

Now, a necessary condition for our solution is that the derivative of (10.86) be zero at the MAP estimate:

$$\frac{\partial}{\partial \underline{x}} \left[\underline{y}^T R^{-1} \underline{y} - 2\underline{x}^T C^T R^{-1} \underline{y} + \underline{x}^T C^T R^{-1} C \underline{x} + \underline{x}^T Q^{-1} \underline{x} \right]\Big|_{\widehat{x}_{\text{MAP}}} = 0 \qquad (10.87)$$

$$\implies \quad -2C^T R^{-1} \underline{y} + 2C^T R^{-1} C \widehat{x}_{\text{MAP}} + 2Q^{-1} \widehat{x}_{\text{MAP}} = 0 \qquad (10.88)$$

where in going from (10.87) to (10.88) we have made use of the rules of vector calculus. Finally, we obtain that the MAP estimate must satisfy the following set of so called *normal equations*:

$$\left(C^T R^{-1} C + Q^{-1} \right) \widehat{x}_{\text{MAP}} = C^T R^{-1} \underline{y} \qquad (10.89)$$

Note that since the problem is Gaussian, this is also the Bayes and the LLSE estimate.

Note that we have specified the MAP estimate implicitly as the solution of a set of linear equations. For very large problems (as arise, for example, in image processing), the computational cost of explicitly inverting the left hand side of (10.89) is prohibitive – with a cost of $O(N^3)$. As a result, in these cases this set of equations are usually solved *iteratively* using a method such as conjugate gradient. For many such problems the left had side of (10.89) is very sparse, which is well suited to such iterative techniques.

We can also obtain an alternate expression for the estimation error covariance matrix Λ_{MAP} associated with the MAP estimate.

$$\begin{aligned}
E\left[(x - \widehat{x}_{\text{MAP}})(x - \widehat{x}_{\text{MAP}})^T \right] &= Q - E\left[\left(C^T R^{-1} C + Q^{-1} \right)^{-1} C^T R^{-1} \left(C \underline{x} + \underline{w} \right) x^T \right] \\
&\quad - E\left[\underline{x} \left(\underline{x}^T C^T + \underline{w}^T \right) R^{-1} C \left(C^T R^{-1} C + Q^{-1} \right)^{-1} \right] \\
&\quad + E\left[\left(C^T R^{-1} C + Q^{-1} \right)^{-1} C^T R^{-1} \left(C \underline{x} + \underline{w} \right) \left(C \underline{x} + \underline{w} \right)^T R^{-1} C \left(C^T R^{-1} C + Q^{-1} \right)^{-1} \right] \\
&= Q - \left(C^T R^{-1} C + Q^{-1} \right)^{-1} C^T R^{-1} C Q - Q C^T R^{-1} C \left(C^T R^{-1} C + Q^{-1} \right)^{-1} \\
&\quad + \left(C^T R^{-1} C + Q^{-1} \right)^{-1} C^T R^{-1} \left[C Q C^T + R \right] R^{-1} C \left(C^T R^{-1} C + Q^{-1} \right)^{-1} \\
&= Q - \left(C^T R^{-1} C + Q^{-1} \right)^{-1} C^T R^{-1} C Q - Q C^T R^{-1} C \left(C^T R^{-1} C + Q^{-1} \right)^{-1} \\
&\quad + \left(C^T R^{-1} C + Q^{-1} \right)^{-1} C^T R^{-1} C Q \left[C^T R^{-1} C + Q^{-1} \right] \left(C^T R^{-1} C + Q^{-1} \right)^{-1} \\
&= Q - \left(C^T R^{-1} C + Q^{-1} \right)^{-1} C^T R^{-1} C Q - Q C^T R^{-1} C \left(C^T R^{-1} C + Q^{-1} \right)^{-1} \\
&\quad + \left(C^T R^{-1} C + Q^{-1} \right)^{-1} C^T R^{-1} C Q \\
&= Q - Q C^T R^{-1} C \left(C^T R^{-1} C + Q^{-1} \right)^{-1} \\
&= \left[Q \left(C^T R^{-1} C + Q^{-1} \right) - Q C^T R^{-1} C \right] \left(C^T R^{-1} C + Q^{-1} \right)^{-1} \\
&= \left[Q C^T R^{-1} C + I - Q C^T R^{-1} C \right] \left(C^T R^{-1} C + Q^{-1} \right)^{-1} \\
&= \left(C^T R^{-1} C + Q^{-1} \right)^{-1}
\end{aligned}$$

Thus we have that

$$\Lambda_{\text{MAP}} = \left(C^T R^{-1} C + Q^{-1} \right)^{-1} \qquad (10.90)$$

The *inverse* of the error covariance for the MAP estimate is thus given by:

$$\Lambda_{\text{MAP}}^{-1} = \underbrace{C^T R^{-1} C}_{\text{Info in Obs}} + \underbrace{Q^{-1}}_{\text{Prior Info}} \qquad (10.91)$$

The inverse of a covariance (i.e. the inverse of the variability) can reasonably be taken as a measure of information. Indeed, such inverse covariances are referred to as "information matrices." We can thus see that the information in the estimate after observing data is composed of two parts, as indicated in (10.91). The prior information and the information in the observation. The interesting thing is that these two pieces of information simply add!

Example 10.12 (Gaussian Problems with Nonlinear Observations)
Here we consider a case often arising in practice, that of a nonlinear observation model with additive Gaussian noise and a Gaussian prior model. In particular, suppose we have the following model for our observation $\underline{y}$ and unknown $\underline{x}$:

$$\underline{y} = H(\underline{x}) + \underline{v}, \qquad \underline{v} \sim N(\underline{0}, R) \tag{10.92}$$

$$\underline{x} \sim N(\underline{0}, Q) \tag{10.93}$$

where we assume that the noise $\underline{v}$ is independent of $\underline{x}$. Now $p_{Y|X}(\underline{y} \mid \underline{x})$ is Gaussian. In particular, we have that:

$$\ln\left[p_{Y|X}(\underline{y} \mid \underline{x})\right] = -\frac{1}{2}\left(\underline{y} - H(\underline{x})\right)^T R^{-1}\left(\underline{y} - H(\underline{x})\right) + \text{constant} \tag{10.94}$$

Thus:

$$\frac{\partial}{\partial x}\ln\left[p_{Y|X}(\underline{y} \mid \underline{x})\right] = \underline{y}^T R^{-1} H_x(\underline{x}) - H^T(\underline{x}) R^{-1} H_x(\underline{x}) \tag{10.95}$$

where $[H_x(\underline{x})]_{ij} = \frac{\partial H_i(\underline{x})}{\partial \underline{x}_j}$ is the matrix of partial derivatives of the vector function $H(\underline{x})$ with respect to its arguments.
    Continuing, $p_X(\underline{x})$ is also Gaussian, so that we have:

$$\frac{\partial}{\partial x}\ln\left[p_X(\underline{x})\right] = -\underline{x}^T Q^{-1} \tag{10.96}$$

Combining (10.95) and (10.96) using the MAP equation (10.64) we obtain an equation for the MAP estimate for this case:

$$\underline{y}^T R^{-1} H_x(\widehat{\underline{x}}_{MAP}) - H^T(\widehat{\underline{x}}_{MAP}) R^{-1} H_x(\widehat{\underline{x}}_{MAP}) - \widehat{\underline{x}}_{MAP}^T Q^{-1} = 0 \tag{10.97}$$

$$\implies \quad H_x^T(\widehat{\underline{x}}_{MAP}) R^{-1} H(\widehat{\underline{x}}_{MAP}) + Q^{-1}\widehat{\underline{x}}_{MAP} = H_x^T(\widehat{\underline{x}}_{MAP}) R^{-1}\underline{y} \tag{10.98}$$

In general this represents a set of *nonlinear* equations, since H depends on $\underline{x}$.
    If the mapping happens to be linear, so that $H(\underline{x}) = H\underline{x}$, this yields:

$$\left(H^T R^{-1} H + Q^{-1}\right)\widehat{\underline{x}}_{MAP} = H^T R^{-1}\underline{y} \tag{10.99}$$

which is a set of linear equations for the MAP estimate. Compare this solution with that obtained for the LLSE estimate with a linear observeration model in Section 10.5 (recall the LLSE estimate is the same as the MAP estimate under a Gaussian assumption).

Implicit Gaussian Prior Models

It is sometimes convenient in MAP estimation with Gaussian models to specify our prior model for $\underline{x}$ *implicitly* rather than explicitly as in (10.82). Such a situation arises, for example, when the elements of $\underline{x}$ are related by a dynamic equation. Our problem statement in such cases can be taken to be:

$$\underline{y} = C\underline{x} + \underline{w}, \qquad \underline{w} \sim N(0, R) \tag{10.100}$$

$$L\underline{x} = \underline{v}, \qquad \underline{v} \sim N(0, Q_v) \tag{10.101}$$

where L is a matrix and v a zero-mean Gaussian process with covariance Q_v. Note that our "standard" prior model for $\underline{x}$, as given by its covariance matrix, is *implicitly* specified through (10.101). Assuming that L is invertible, it is a simple matter to obtain the explicit model of $\underline{x}$ given such an implicit model as:

$$\underline{x} \sim N\left(0, L^{-1}Q_v L^{-T}\right) = N\left(0, \left[L^T Q_v^{-1} L\right]^{-1}\right) \tag{10.102}$$

Given this prior model and our solution in (10.89), we see that the MAP estimate for an implicit prior model must satisfy:

$$\left(C^T R^{-1} C + L^T Q_v^{-1} L\right)\widehat{\underline{x}}_{\text{MAP}} = C^T R^{-1}\underline{y} \tag{10.103}$$

Note, in particular, that the normal equations can be formed without the need of inverting L. The corresponding estimation error covariance is given by:

$$\Lambda_{\mathrm{MAP}} = \left(C^T R^{-1} C + L^T Q_v^{-1} L\right)^{-1} \tag{10.104}$$

You may wonder why we would care about implicit specification of prior models as in (10.101). One case of interest arises when $\underline{x}$ is specified through an autoregressive model driven by white noise. For example, suppose the elements x_i of $\underline{x}$ are specified as the output of the following AR model:

$$\begin{aligned}
x_n &= a x_{n-1} + v_n, & v_n &\sim N(0,1) & (10.105) \\
x_0 &= v_0, & v_0 &\sim N(0,1) & (10.106)
\end{aligned}$$

Then this structure implies the following model for $\underline{x}$:

$$
\underbrace{\begin{bmatrix}
1 & 0 & \cdots & & 0 \\
-a & 1 & & & \\
 & -a & 1 & & \\
 & & \ddots & \ddots & \\
 & & & -a & 1
\end{bmatrix}}_{L}
\underbrace{\begin{bmatrix}
x_0 \\ x_1 \\ \vdots \\ x_{N-1} \\ x_N
\end{bmatrix}}_{\underline{x}}
=
\underbrace{\begin{bmatrix}
v_0 \\ v_1 \\ \vdots \\ v_{N-1} \\ v_N
\end{bmatrix}}_{\underline{v}}
\tag{10.107}
$$

where $\underline{v} \sim N(0, I)$. As can be seen, the matrix L captures the elements of the implicit AR model of the process. Indeed, it is a relatively simple matter given an arbitrary AR model to specify the elements of L. In particular, for an p-th order AR model L will have exactly p nonzero bands, consisting ones on the diagonal and the AR coefficients along the sub-diagonals.

Given the results in (10.102), and the fact that the vector $\underline{v}$ is white, it is now a simple matter to specify associated covariance matrix for the entire 1st order AR process as:

$$
Q = \left(L^T L\right)^{-1} = \begin{bmatrix}
1+a^2 & -a & & & & \\
-a & 1+a^2 & -a & & \bigcirc & \\
 & -a & 1+a^2 & -a & & \\
 & & \ddots & \ddots & \ddots & \\
 & \bigcirc & & -a & 1+a^2 & -a \\
 & & & & -a & 1
\end{bmatrix}^{-1}
\tag{10.108}
$$

Finally, as noted above, in forming the normal equations (10.103), since all we use is Q^{-1}, we need never explicitly invert L. In addition, the matrix $L^T L$ has a highly sparse and banded structure, which will be reflected in the structure of the normal equations. Such structure, which will be associated with any AR model, is typical of a host of problems appearing in science and engineering.

Let us close this section by summarizing what we have learned about MAP estimates:

- The MAP estimate is the conditional mode: $\arg\max_x p_{X|Y}(x \mid y)$.

- The MAP estimate may be biased.

- The MAP estimate is not necessarily the MMSE estimate.

- The MAP estimate may not be unique.

- In general, the MAP estimate is a nonlinear function of the data.

- For jointly Gaussian problems, the MAP estimate is the same as the BLSE estimate, and in this case is a linear estimate and MMSE.

- In general, finding the MAP estimate requires finding the conditional density, and thus is challenging.

- More generally, when $p_{X|Y}(x \mid y)$ is symmetric and unimodal, the MAP and BLSE estimates coincide.

10.5   Bayes Linear Least Square (LLSE) Estimation

In general (for non-Gaussian problems), the BLSE or MAP estimates are nonlinear functions of the data. Further, finding these estimates requires calculation of the posterior density. For these reasons, finding and implementing these estimates can be difficult in practice. As a result, we now modify our approach a bit. In particular, we will restrict our attention to estimators that are a *linear function* of the data (strictly speaking, an affine form). If we seek the estimator of this class that minimizes the square error cost function used in the BLSE estimate (i.e. minimize $E\left[J_{\text{BLSE}}(\widehat{x}, x)\right]$ over the class of linear estimators), we will see that we obtain an estimate that only requires knowledge of second-order statistics, rather than the conditional density. This resulting estimate is termed the *Linear Least Square Estimate* (LLSE). In summary, we will focus on the BLSE cost function and its mean value, but will restrict the form of the our estimator to linear functions of the data.

Let us start with the scalar development for simplicity. We restrict our estimator to have the following linear (actually affine) form:

$$\widehat{x}_{\text{LLSE}}(y) \quad = \quad c_1 y + c_2 \tag{10.109}$$

for some constants c_1 and c_2. Now we need to choose the constants c_1 and c_2 to minimize the mean square error cost criterion, i.e. so that the following mean square cost is minimized:

$$E\left[J_{\text{BLSE}}(x, c_1 y + c_2)\right] \quad = \quad E\left[(x - c_1 y - c_2)^2\right] \tag{10.110}$$
$$= \quad E\left[x^2\right] - 2c_1 E\left[xy\right] - 2c_2 E\left[x\right] + c_1^2 E\left[y^2\right] + 2c_1 c_2 E\left[y\right] + c_2^2 \tag{10.111}$$

To find the minimum of this expression we take partial derivatives with respect to c_1 and c_2 and set them to zero. Doing this yields the following pair of equations:

$$\frac{\partial}{\partial c_2} E\left[J_{\text{BLSE}}(x, c_1 y + c_2)\right] \quad = \quad -2m_x + 2c_1 m_y + 2c_2 = 0 \tag{10.112}$$
$$\frac{\partial}{\partial c_1} E\left[J_{\text{BLSE}}(x, c_1 y + c_2)\right] \quad = \quad -2\left(\lambda_{xy} + m_x m_y\right) + 2c_1 \left(\lambda_y + m_y^2\right) + 2c_2 m_y = 0 \tag{10.113}$$

Solving this simultaneous set of equations for c_1 and c_2 yields:

$$c_1 \quad = \quad \frac{\lambda_{xy}}{\lambda_y} \tag{10.114}$$

$$c_2 \quad = \quad m_x - c_1 m_y = m_x - \frac{\lambda_{xy}}{\lambda_y} m_y \tag{10.115}$$

Note that we really need to check that the second derivatives is positive at these point to assure the point is a minimum. This is left to the reader. Substituting these expressions into the LLSE form in (10.109) we obtain for the scalar LLSE estimate:

$$\boxed{\widehat{x}_{\text{LLSE}}(y) = m_x + \frac{\lambda_{xy}}{\lambda_y}\left(y - m_y\right)} \tag{10.116}$$

An alternative, perhaps more intuitive, way of deriving the LLSE estimate is as follows. As usual, let $e = x - \widehat{x}$ be the error in the estimate. Then the MSE, which we are trying to minimize, is given by:

$$E\left[J_{\text{BLSE}}(x, c_1 y + c_2)\right] \quad = \quad E\left[e^2\right] = m_e^2 + \lambda_e^2 \tag{10.117}$$

where m_e is the mean of the error and λ_e is its variance. Since both quantities are non-negative, to minimize the cost or Bayes risk for this case we want to set m_e equal to zero (make the estimate unbiased) then minimize the error variance λ_e. Doing the the first of these yields:

$$m_e = E\left[x - \widehat{x}\right] = E\left[x - c_1 y - c_2\right] = m_x - c_1 m_y - c_2 = 0 \tag{10.118}$$

or

$$c_2 = m_x - c_1 m_y \tag{10.119}$$

Thus by proper choice of c_2 we can indeed make the estimate unbiased. Note that since $m_e = 0$ for the LLSE (and incidentally for the BLSE), the MSE and the error variance are the *same* for this case. For this reason they are often referred to interchangeably (causing some confusion!).

Substituting (10.119) into the definition of the LLSE estimate (10.109) we then have that the optimal estimate is of the form:

$$\widehat{x}_{\text{LLSE}}(y) \ = \ c_1 (y - m_y) + m_x \tag{10.120}$$

Now we just need to find c_1 to minimize the error variance or the remain cost. Substituting (10.120) into the cost expression yields:

$$E\left[(x - c_1 y - c_2)^2\right] \ = \ E\left[((x - m_x) - c_1 (y - m_y))^2\right] \tag{10.121}$$

$$= \ E\left[(x - m_x)^2 - 2c_1 (x - m_x)(y - y_x) + c_1^2 (y - m_y)^2\right] \tag{10.122}$$

$$= \ \lambda_x - 2c_1 \lambda_{xy} + c_1^2 \lambda_y \tag{10.123}$$

Minimizing this expression with respect to c_1 (e.g. by taking a derivative and setting it equal to zero) yields:

$$c_1 \ = \ \frac{\lambda_{xy}}{\lambda_y} \tag{10.124}$$

Combining (10.120) and (10.124) we again obtain (10.116) for the LLSE estimate.

The corresponding MSE of the LLSE estimate is given by the value of the expected cost for this case, and can be shown by direct substitution of the estimate (10.116) into the cost function to be:

$$\text{MSE} = E\left[\left(x - \widehat{x}_{\text{LLSE}}(y)\right)^2\right] = \lambda_x - \frac{\lambda_{xy}^2}{\lambda_y} \tag{10.125}$$

Again, we note that since the bias is zero, this is also the expression for the error variance for this case λ_{LLSE}.

Note that the estimate (10.116) and the MSE (10.125) only depend on the means, covariances, and cross-covariances of the underlying random variables! In other words, only knowledge of second-order statistics are needed to find the LLSE estimate, and not the entire conditional density, as was required to find the BLSE or MAP estimates. Further, notice that the form of the LLSE estimate and MSE are the *same* as that we obtained for the scalar Gaussian case of Example 10.5. In particular, the estimator is the same as that we would obtain if the underlying random variables were jointly Gaussian with the same means and covariances.

Before proceeding to some examples, let us present the LLSE estimate for the more general vector case. In this case the general form of the LLSE estimate is:

$$\widehat{\underline{x}}_{\text{LLSE}}(\underline{y}) \ = \ C_1^T \underline{y} + \underline{c}_2 \tag{10.126}$$

for some constant matrices C_1 and $\underline{c}_2$. As before, we need to choose the constants C_1 and $\underline{c}_2$ to minimize the mean square error cost criterion:

$$E\left[J_{\text{BLSE}}(\underline{x}, C_1^T \underline{y} + \underline{c}_2)\right] \tag{10.127}$$

$$= \ E\left[\left(\underline{x} - C_1^T \underline{y} - \underline{c}_2\right)^T \left(\underline{x} - C_1^T \underline{y} - \underline{c}_2\right)\right] \tag{10.128}$$

$$= \ E\left[\underline{x}^T \underline{x}\right] - 2E\left[\underline{x}^T C_1^T \underline{y}\right] - 2\underline{c}_2^T E\left[\underline{x}\right] + E\left[\underline{y}^T C_1 C_1^T \underline{y}\right] + 2\underline{c}_2^T C_1^T E\left[\underline{y}\right] + \underline{c}_2^T \underline{c}_2 \tag{10.129}$$

$$= \ E\left[\text{tr}\left(\underline{x}\underline{x}^T\right)\right] - 2E\left[\text{tr}\left(C_1^T \underline{y}\underline{x}^T\right)\right] - 2\underline{c}_2^T \underline{m}_x + E\left[\text{tr}\left(C_1^T \underline{y}\underline{y}^T C_1\right)\right] + 2\underline{c}_2^T C_1^T \underline{m}_y + \underline{c}_2^T \underline{c}_2 \tag{10.130}$$

$$= \ \text{tr}\left(E\left[\underline{x}\underline{x}^T\right]\right) - 2\text{tr}\left(C_1^T E\left[\underline{y}\underline{x}^T\right]\right) - 2\underline{c}_2^T \underline{m}_x + \text{tr}\left(C_1^T E\left[\underline{y}\underline{y}^T\right] C_1\right) + 2\underline{c}_2^T C_1^T \underline{m}_y + \underline{c}_2^T \underline{c}_2 \tag{10.131}$$

$$= \ \text{tr}\left(E\left[\underline{x}\underline{x}^T\right]\right) - 2\text{tr}\left(C_1^T E\left[\underline{y}\underline{x}^T\right]\right) - 2\underline{c}_2^T \underline{m}_x + \text{tr}\left(C_1^T E\left[\underline{y}\underline{y}^T\right] C_1\right) + 2\text{tr}\left(C_1^T \underline{m}_y \underline{c}_2^T\right) + \underline{c}_2^T \underline{c}_2 \tag{10.132}$$

where in going from (10.129) to (10.130) and from (10.131) to (10.132) we have used that fact that $\underline{x}^T \underline{y} = \text{tr}(\underline{y}\underline{x}^T)$ for the trace of a matrix (see Appendix C). Now to minimize this expression with respect to C_1 and $\underline{c}_2$ we take derivatives with respect to these two quantities (using the rules in Appendix C) and set them equal to zero:

$$\frac{\partial}{\partial C_1} E\left[J_{\text{BLSE}}(x, C_1^T y + \underline{c}_2)\right] \quad = \quad -2E\left[\underline{y}\underline{x}^T\right] + 2E\left[\underline{y}\underline{y}^T\right] C_1 + 2\underline{m}_y \underline{c}_2^T = 0 \tag{10.133}$$

$$\frac{\partial}{\partial \underline{c}_2} E\left[J_{\text{BLSE}}(x, C_1^T y + \underline{c}_2)\right] \quad = \quad -2\underline{m}_x + 2C_1^T \underline{m}_y + 2\underline{c}_2 = 0 \tag{10.134}$$

Solving (10.134) for $\underline{c}_2$ yields:

$$\underline{c}_2 \quad = \quad \underline{m}_x - C_1^T \underline{m}_y \tag{10.135}$$

Substituting this into (10.133) yields for following equation which the optimal C_1 must satisfy:

$$0 \quad = \quad -2E\left[\underline{y}\underline{x}^T\right] + 2E\left[\underline{y}\underline{y}^T\right] C_1 + 2\underline{m}_y \left(\underline{m}_x - C_1^T \underline{m}_y\right)^T \tag{10.136}$$

$$= \quad -2\left(E\left[\underline{y}\underline{x}^T\right] - \underline{m}_y \underline{m}_x^T\right) + 2\left(E\left[\underline{y}\underline{y}^T\right] - \underline{m}_y \underline{m}_y^T\right) C_1 \tag{10.137}$$

$$= \quad -2\Lambda_{yx} + 2\Lambda_y C_1 \tag{10.138}$$

Solving for C_1 yields:

$$C_1 \quad = \quad \Lambda_y^{-1}\Lambda_{yx} \tag{10.139}$$

Substituting the expressions for C_1 and $\underline{c}_2$ into the definition of the LLSE form we obtain for the vector LLSE estimate:

$$\widehat{\underline{x}}_{\text{LLSE}}(\underline{y}) \quad = \quad C_1^T \underline{y} + \underline{c}_2 = \left(\Lambda_y^{-1}\Lambda_{yx}\right)^T \underline{y} + \underline{m}_x - \left(\Lambda_y^{-1}\Lambda_{yx}\right)^T \underline{m}_y \tag{10.140}$$

$$= \quad \underline{m}_x + \Lambda_{yx}^T \Lambda_y^{-1}\left(\underline{y} - \underline{m}_y\right) \tag{10.141}$$

Since $\Lambda_{yx}^T = \Lambda_{xy}$ we obtain for the LLSE estimate in the vector case:

$$\boxed{\widehat{\underline{x}}_{\text{LLSE}}(\underline{y}) = m_x + \Lambda_{xy}\Lambda_y^{-1}\left(\underline{y} - \underline{m}_y\right)} \tag{10.142}$$

The corresponding error covariance can be obtained by direct substitution as:

$$\Lambda_{\text{LLSE}} \quad = \quad E\left[\left(\underline{x} - \widehat{\underline{x}}_{\text{LLSE}}(\underline{y})\right)\left(\underline{x} - \widehat{\underline{x}}_{\text{LLSE}}(\underline{y})\right)^T\right] = \Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T \tag{10.143}$$

Using the properties of the trace, we see that the MSE is just the trace of the error covariance Λ_{LLSE}:

$$\text{MSE} \quad = \quad E\left[\left(\underline{x} - \widehat{\underline{x}}_{\text{LLSE}}(\underline{y})\right)^T\left(\underline{x} - \widehat{\underline{x}}_{\text{LLSE}}(\underline{y})\right)\right] = \text{tr}\left(\Lambda_{\text{LLSE}}\right) \tag{10.144}$$

Again note that these expressions only depend on the means, covariances, and cross-covariances of the underlying random variables. As for the scalar case, the formula for the LLSE is the same as that obtained for the vector Gaussian case of Example 10.6. Now let us examine some examples:

Example 10.13
In this example we return to Example 10.2, but this time seek the LLSE estimate. Let us apply the LLSE formula (10.142). We need the second-order quantities m_x, m_y, λ_{xy}, λ_x, and λ_y. The means m_x and m_y are zero by symmetry of the density. The covariances are obtained as:

$$\lambda_y = \lambda_x \quad = \quad \int_{-\infty}^{\infty} y^2\, p_Y(y)\, dy = \int_{-1}^{0} y^2(1 + y)\, dy + \int_{0}^{1} y^2(1 - y)\, dy = \frac{1}{6} \tag{10.145}$$

$$\lambda_{xy} \quad = \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, p_{X,Y}(x, y)\, dx\, dy = 2\int_{0}^{1}\int_{0}^{1-y} \frac{xy}{2}\, dx\, dy - 2\int_{0}^{1}\int_{0}^{1-y} \frac{xy}{2}\, dx\, dy = 0 \tag{10.146}$$

where we have used the symmetry of the density in obtaining the expression for λ_{xy}. Thus we obtain for the LLSE:

$$\widehat{x}_{\mathsf{LLSE}}(y) = 0 \tag{10.147}$$

as before. Using the formula for the MSE we obtain

$$\mathsf{MSE} \;=\; \lambda_x - \frac{\lambda_{xy}^2}{\lambda_y} = \frac{1}{6} - \frac{0}{1/6} = \frac{1}{6} \tag{10.148}$$

Example 10.14
For this example let us revisit the problem of Example 10.3, but seek the LLSE estimate. Note that the BLSE for this example happened to be linear. Since the LLSE is nothing more than the minimum MSE estimator restricted to have a linear form, the BLSE and the LLSE are the same for this example. In other words, if the BLSE happens to be linear, we certainly cannot find a different linear estimator with lower MSE! Even though we know the answer, let us find the LLSE via the formula.

Again, we need the second order quantities m_x, m_y, λ_{xy}, λ_x, and λ_y:

$$m_y \;=\; \int_{-\infty}^{\infty} y p_Y(y)\, dy = \int_0^1 3y^3\, dy = \frac{3}{4} \tag{10.149}$$

$$m_x \;=\; \int_{-\infty}^{\infty} x p_X(x)\, dx = \int_0^1 x 6x\,(1-x)\ dx = \frac{1}{2} \tag{10.150}$$

$$\lambda_y \;=\; \int_{-\infty}^{\infty} y^2\, p_Y(y)\, dy - m_y^2 = \int_0^1 3y^4\, dy - \frac{9}{16} = \frac{3}{80} \tag{10.151}$$

$$\lambda_x \;=\; \int_{-\infty}^{\infty} x^2\, p_X(x)\, dx - m_x^2 = \int_0^1 x^2\, 6x\,(1-x)\ dx - \frac{1}{4} = \frac{1}{20} \tag{10.152}$$

$$\lambda_{xy} \;=\; \int_{-\infty}^{\infty}\!\!\int_{-\infty}^{\infty} xy\, p_{X,Y}(x,y)\, dx\, dy - m_x m_y = \int_0^1\!\!\int_0^y 6x^2 y\ dx\, dy - \frac{1}{2}\frac{3}{4} = \frac{1}{40} \tag{10.153}$$

Thus we obtain for the LLSE:

$$\widehat{x}_{\mathsf{LLSE}}(y) \;=\; m_x + \frac{\lambda_{xy}}{\lambda_y}\,(y - m_y) = \frac{1}{2} + \frac{1/40}{3/80}\left(y - \frac{3}{4}\right) = \frac{2}{3}y \tag{10.154}$$

as before. Using the formula for the MSE we obtain

$$\mathsf{MSE} \;=\; \lambda_x - \frac{\lambda_{xy}^2}{\lambda_y} = \frac{1}{20} - \frac{(1/40)^2}{3/80} = \frac{1}{30} \tag{10.155}$$

Example 10.15
For this example let us revisit the problem of Example 10.4. Again, we need the second order quantities m_x, m_y, λ_{xy}, λ_x, and λ_y:

$$m_y \;=\; \int_{-\infty}^{\infty} y p_Y(y)\, dy = \int_0^1 5y^5\, dy = \frac{5}{6} \tag{10.156}$$

$$m_x \;=\; \int_{-\infty}^{\infty}\!\!\int_{-\infty}^{\infty} x p_{X,Y}(x,y)\, dx\, dy = \int_0^1\!\!\int_0^{y^2} x 10x\ dx\, dy = \frac{10}{21} \tag{10.157}$$

$$\lambda_y \;=\; \int_{-\infty}^{\infty} y^2\, p_Y(y)\, dy - m_y^2 = \int_0^1 5y^6\, dy - \left(\frac{5}{6}\right)^2 = \frac{5}{252} \tag{10.158}$$

$$\lambda_x \;=\; \int_{-\infty}^{\infty}\!\!\int_{-\infty}^{\infty} x^2\, p_X(x)\, dx\, dy - m_x^2 = \int_0^1\!\!\int_0^{y^2} x^2\, 10x\ dx\, dy - \left(\frac{10}{21}\right)^2 = \frac{5}{18} - \left(\frac{10}{21}\right)^2 = \frac{5}{98} \tag{10.159}$$

$$\lambda_{xy} \;=\; \int_{-\infty}^{\infty}\!\!\int_{-\infty}^{\infty} xy\, p_{X,Y}(x,y)\, dx\, dy - m_x m_y = \int_0^1\!\!\int_0^{y^2} 10x^2 y\ dx\, dy - \frac{5}{6}\frac{10}{21} = \frac{5}{12} - \frac{5}{6}\frac{10}{21} = \frac{5}{252} \tag{10.160}$$

Thus we obtain for the LLSE:

$$\widehat{x}_{\mathsf{LLSE}}(y) \;=\; m_x + \frac{\lambda_{xy}}{\lambda_y}\,(y - m_y) = \frac{10}{21} + \frac{5/252}{5/252}\left(y - \frac{5}{6}\right) = y - \frac{5}{14} \tag{10.161}$$

as before. Using the formula for the MSE we obtain

$$\text{MSE} \quad = \quad \lambda_x - \frac{\lambda_{xy}^2}{\lambda_y} = \frac{5}{98} - \frac{(5/252)^2}{5/252} = \frac{55}{1764} = 0.0312 \tag{10.162}$$

Note that this MSE is worse than that obtained by the optimal minimum MSE nonlinear BLSE estimator of Example 10.4 – but not much worse.

Example 10.16 (Scalar Gaussian Case)

Let us now examine the problem of Example 10.5 with regard to the LLSE. For this jointly Gaussian problem we can immediately see that the LLSE estimate is identical to the BLSE estimate!

Example 10.17 (Vector Gaussian Case)

Finally, we have the vector Gaussian case. As for the scalar Gaussian case, we can immediately see that the LLSE estimate is identical to the BLSE estimate. Thus for jointly Gaussian problems we have the interesting result that the BLSE, MAP and LLSE estimators are all the same.

Example 10.18

In this example we examine the following problem: Let the random vector $\underline{z}$ be linearly related to the random vector $\underline{x}$ as follows:

$$\underline{z} = F\underline{x} + H\underline{w} + \underline{c} \tag{10.163}$$

where $\underline{c}$ is deterministic, $E[\underline{w}] = \underline{m}_w$, and $\underline{w}$ is correlated with $\underline{x}$. Find the LLSE estimate of $\underline{z}$ based on observation of $\underline{y}$ in terms of the second-order statistics of $\underline{x}$ and $\underline{y}$.

As always for LLSE estimates we need to find the second order quantities: $\underline{m}_z$, Λ_{zy}, and Λ_z:

$$\underline{m}_z \quad = \quad E\left[\underline{z}\right] = F\underline{m}_x + H\underline{m}_w + \underline{c} \tag{10.164}$$

$$\Lambda_z \quad = \quad E\left[\underline{z}\,\underline{z}^T\right] - \underline{m}_z\underline{m}_z^T \tag{10.165}$$

$$= \quad FE\left[\underline{x}\,\underline{x}^T\right]F^T + FE\left[\underline{x}\,\underline{w}^T\right]H^T + HE\left[\underline{w}\,\underline{x}^T\right]F^T \tag{10.166}$$

$$\quad + HE\left[\underline{w}\,\underline{w}^T\right]H^T - F\underline{m}_x\underline{m}_x^T F^T - F\underline{m}_x\underline{m}_w^T H^T - H\underline{m}_w\underline{m}_x^T F^T - H\underline{m}_x\underline{m}_x^T H^T \tag{10.167}$$

$$= \quad F\Lambda_x F^T + H\Lambda_w H^T + F\Lambda_{xw}H^T + H\Lambda_{xw}^T F^T \tag{10.168}$$

$$\Lambda_{zy} \quad = \quad E\left[\underline{z}\,\underline{y}^T\right] - \underline{m}_z\underline{m}_y^T \tag{10.169}$$

$$= \quad E\left[\left(F\underline{x} + H\underline{w} + \underline{c}\right)\underline{y}^T\right] - \left(F\underline{m}_x + H\underline{m}_w + \underline{c}\right)\underline{m}_y^T \tag{10.170}$$

$$= \quad F\Lambda_{xy} + H\Lambda_{wy} \tag{10.171}$$

Thus the LLSE estimate is given by:

$$\widehat{z}_{\mathsf{LLSE}}(y) \quad = \quad \underline{m}_z + \Lambda_{zy}\Lambda_y^{-1}\left(\underline{y} - \underline{m}_y\right) \tag{10.172}$$

$$= \quad F\underline{m}_x + H\underline{m}_w + \underline{c} + \left[F\Lambda_{xy} + H\Lambda_{wy}\right]\Lambda_y^{-1}\left(\underline{y} - \underline{m}_y\right) \tag{10.173}$$

$$= \quad F\left[\underline{m}_x + \Lambda_{xy}\Lambda_y^{-1}\left(\underline{y} - \underline{m}_y\right)\right] + H\left[\underline{m}_w + \Lambda_{wy}\Lambda_y^{-1}\left(\underline{y} - \underline{m}_y\right)\right] + \underline{c} \tag{10.174}$$

$$= \quad F\widehat{x}_{\mathsf{LLSE}}(y) + H\widehat{w}_{\mathsf{LLSE}}(y) + \underline{c} \tag{10.175}$$

Note that the LLSE estimate of $\underline{z}$ can be written in terms of the LLSE estimate of $\underline{x}$ and $\underline{w}$! The corresponding error covariance for this case is given by:

$$\Lambda_{z,\mathsf{LLSE}} \quad = \quad \Lambda_z - \Lambda_{zy}\Lambda_y^{-1}\Lambda_{zy}^T \tag{10.176}$$

$$= \quad F\Lambda_x F^T + F\Lambda_{xw}H^T + H\Lambda_{xw}^T F^T + H\Lambda_w H^T - \left[F\Lambda_{xy} + H\Lambda_{wy}\right]\Lambda_y^{-1}\left[F\Lambda_{xy} + H\Lambda_{wy}\right]^T \tag{10.177}$$

$$= \quad F\left(\Lambda_x - \Lambda_{xy}\Lambda_y^{-1}\Lambda_{xy}^T\right)F^T + H\left(\Lambda_w - \Lambda_{wy}\Lambda_y^{-1}\Lambda_{wy}^T\right)H^T \tag{10.178}$$

$$\quad + F\Lambda_{xy}\left[I - \Lambda_y^{-1}\Lambda_{wy}^T\right]H^T + H\left[I - \Lambda_{wy}\Lambda_y^{-1}\right]\Lambda_{xw}^T F^T \tag{10.179}$$

$$= \quad F\Lambda_{x,\mathsf{LLSE}}F^T + H\Lambda_{w,\mathsf{LLSE}}H^T + F\Lambda_{xy}\left[I - \Lambda_y^{-1}\Lambda_{wy}^T\right]H^T + H\left[I - \Lambda_{wy}\Lambda_y^{-1}\right]\Lambda_{xw}^T F^T \tag{10.180}$$

where $\Lambda_{x,\mathsf{LLSE}}$ is the error covariance associated with the LLSE of $\underline{x}$ based on $\underline{y}$ and $\Lambda_{w,\mathsf{LLSE}}$ is the error covariance associated with the LLSE of $\underline{w}$ based on $\underline{y}$. Thus we can also express the error covariance in terms of the LLSE estimate of $\underline{x}$ and $\underline{w}$ for this example. We will use these expressions in our study of the Kalman filter later in the notes.

Linear Observation Model   Here we examine a case that often is used in practice. In particular, suppose that $\underline{y}$ and $\underline{x}$ are related by the following linear observation equation:

$$\underline{y} \;=\; H\underline{x} + \underline{v} \tag{10.181}$$

where $\underline{x}$ is a random vector with mean $\underline{m}_x$ and covariance matrix Q and $\underline{v}$ is a zero-mean random vector uncorrelated with $\underline{x}$ with covariance matrix R. The linear observation model in (10.181) is a common one in engineering practice. Let us find the LLSE estimate for this case. As usual, we need to find the quantities second-order quantities $\underline{m}_y$, Λ_{xy}, and Λ_y:

$$\underline{m}_y \;=\; H\underline{m}_x \tag{10.182}$$

$$\Lambda_y \;=\; E\left[\left(\underline{y} - \underline{m}_y\right)\left(\underline{y} - \underline{m}_y\right)^T\right] = HQH^T + R \tag{10.183}$$

$$\Lambda_{xy} \;=\; E\left[\left(\underline{x} - \underline{m}_x\right)\left(\underline{y} - \underline{m}_y\right)^T\right] = E\left[\left(\underline{x} - \underline{m}_x\right)\left(H\underline{x} + \underline{v} - \underline{m}_y\right)^T\right] = QH^T \tag{10.184}$$

Thus we obtain for the LLSE estimate and the associated error covariance Λ_{LLSE}:

$$\widehat{\underline{x}}_{\mathrm{LLSE}}(\underline{y}) \;=\; \underline{m}_x + QH^T\left(HQH^T + R\right)^{-1}\left(\underline{y} - H\underline{m}_x\right) \tag{10.185}$$

$$\Lambda_{\mathrm{LLSE}} \;=\; Q - QH^T\left(HQH^T + R\right)^{-1}HQ \tag{10.186}$$

As usual, the MSE is the trace of the error covariance. There are a number of alternate forms associated with the LLSE for this case that are of particular interest. The first is the following alternative form for the error covariance:

$$\Lambda_{\mathrm{LLSE}}^{-1} = Q^{-1} + H^T R^{-1} H \tag{10.187}$$

The inverse of a covariance is often interpreted as a measure of information. In fact, such a covariance inverse is sometimes termed an "information matrix." With this interpretation, we see that (10.187) states that the total information after the incorporation of a measurement equals the prior information Q^{-1} plus the information $H^T R^{-1} H$ available in the measurement. To verify (10.187) we must show that the following identity is true:

$$\left(Q - QH^T\left[H\Sigma_x H^T + R\right]^{-1}HQ\right)\left(Q^{-1} + H^T R^{-1} H\right) = I \tag{10.188}$$

Multiplying out the terms on the left hand side, we find:

$$I - QH^T\left[H\Sigma_x H^T + R\right]^{-1}H + QH^T R^{-1}H - QH^T\left[HQH^T + R\right]^{-1}HQH^T R^{-1}H \tag{10.189}$$

$$= \; I + QH^T\left[HQH^T + R\right]^{-1}\left(-I + \left[HQH^T + R\right]R^{-1} - HQH^T R^{-1}\right)H \tag{10.190}$$

$$= \; I \tag{10.191}$$

so that (10.187) is verified.

   Secondly, we note that there is an alternate expression for the gain term multiplying the observations in (10.185). This gain term is given in (10.185) by:

$$K = QH^T\left[HQH^T + R\right]^{-1} \tag{10.192}$$

The alternate form is given by:

$$K = \Lambda_{\mathrm{LLSE}}H^T R^{-1} = \left(H^T R^{-1}H + Q^{-1}\right)^{-1}H^T R^{-1} \tag{10.193}$$

where Λ_{LLSE} is the error covariance. Give this equivalence, we see that the LLSE estimate (which is also the MAP estimate for the Gaussian case) must satisfy the following implicit relationship, termed the *normal equations*:

$$\left(H^T R^{-1}H + Q^{-1}\right)\left(\widehat{\underline{x}}_{\mathrm{LLSE}}(\underline{y}) - \underline{m}_x\right) \;=\; H^T R^{-1}\left(\underline{y} - \underline{m}_y\right) \tag{10.194}$$

Note that the matrix on the LHS of this equation is inverse of the error covariance, i.e. the "information matrix" for the problem!

To verify the alternate form for the gain K we proceed as follows

$$
\begin{aligned}
\Lambda_{\text{LLSE}} H^T R^{-1} &= \left(Q - Q H^T \left[H Q H^T + R \right]^{-1} H Q \right) H^T R^{-1} & (10.195) \\
&= Q H^T \left[H Q H^T + R \right]^{-1} \left(H Q H^T + R - H Q H^T \right) R^{-1} & (10.196) \\
&= Q H^T \left[H Q H^T + R \right]^{-1} & (10.197)
\end{aligned}
$$

Finally, we can obtain another alternate expression for the error covariance Λ_{LLSE}. To this end, note that we can write the estimation error in the following form:

$$
\begin{aligned}
\underline{e} &= \underline{x} - \widehat{\underline{x}}_{\text{LLSE}}(\underline{y}) & (10.198) \\
&= (I - KH)(\underline{x} - \underline{m}_x) - K\underline{v} & (10.199)
\end{aligned}
$$

From this form we find our alternate expression for Λ_{LLSE}:

$$
\Lambda_{\text{LLSE}} = E\left[\underline{e}\underline{e}^T\right] = (I - KH)\,Q\,(I - KH)^T + KRK^T \tag{10.200}
$$

Geometric Characterization of LLSE Estimates: Before leaving LLSE estimation we present an extremely important characterization of the LLSE estimate based on geometric notions. Specifically, $\widehat{x}_{\text{LLSE}}(y)$ is the unique linear function of y such that the error $e = x - \widehat{x}_{\text{LLSE}}(y)$ is zero mean (i.e. unbiased) and uncorrelated with any linear function of the data y. That is, an equivalent characterization of the LLSE is that it is the estimator that satisfies the following two conditions:

Unbiased: $E\left[x - \widehat{x}_{\text{LLSE}}(y)\right] = 0$

Error $\perp$ Data: $E\left\{\left[x - \widehat{x}_{\text{LLSE}}(y)\right] g(y)\right\} = 0$ for all linear functions $g(\cdot)$.

This geometric situation is depicted in Figure 10.6. The idea is that the optimal estimate is that linear function of the data which has no correlation with the error. Intuitively, if correlation remained between the error and the estimate, there would remain information in the error of help in estimating the signal that we should have extracted. Note that this geometric condition implies that the error is orthogonal (i.e. uncorrelated with) both the data itself (which is obviously a trival function of the data) as well as the LLSE estimate (which is clearly a linear function of the data).
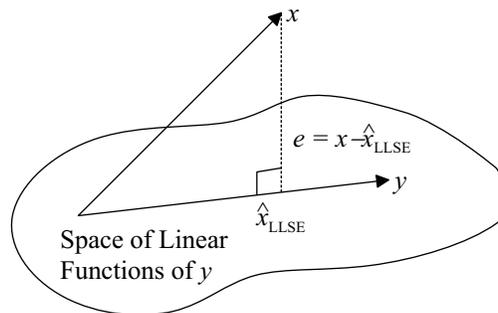
Figure 10.6: Illustration of the projection theorem for LLSE.

Let us close by summarizing the properties of LLSE estimates:

- The LLSE estimate is the minimum MSE estimate over all *linear* functions of the data.

- The LLSE estimate is always unbiased.

- The associated error covariances satisfy: $0 \leq \Lambda_{\text{BLSE}} \leq \Lambda_{\text{LLSE}} \leq \Lambda_x$

- The LLSE estimate equals the BLSE estimate for the jointly Gaussian case.

- The LLSE estimate only requires knowledge of second-order properties.

10.6 Nonrandom Parameter Estimation

In our discussion of Bayes or random parameter estimation we modeled the unknown parameter X as a random variable or vector. This was our "model of nature." In many cases it is not realistic to model X in this way. For example, if we are attempting to estimate the bias of a coin or the orientation of a target, these quantities are not random, but they are still unknown. This leads us to a different model of nature better matched to such problems. In particular, we model X as an *unknown but nonrandom* parameter, so X is just a constant. This seemingly minor change impacts all the elements of our estimation problem. Let us examine the three elements of any estimation problem in this light:

1. **Parameter Model:** As we just discussed X is now modeled as an *unknown deterministic parameter*.

2. **Observation Model:** This is given by $p_{Y|X}(y \mid x)$. Since X is nonrandom, $p_{Y|X}(y \mid x)$ is now a *parameterized density*.

3. **Estimation Rule:** As we discuss in greater detail below, the direct approach to finding a good estimator as the minimizer of a criterion, such as we took in the Bayes case, will present problems. Basically, these can be traced to the fact that we can no longer average over X, since it is no longer random. Instead we take the approach of proposing an estimator and then evaluating its performance. In particular, we will evaluate candidate estimators on the basis of their bias, variance, and mean square error.

In the Bayes case we found our estimators by minimizing the expected value of a cost $E[J(\widehat{x}(y), x)]$. Since this expected value was over the randomness in *both* X and Y, for a given cost structure $J(\cdot, \cdot)$ and a given estimator $\widehat{x}(y)$ the quantity $E[J(\widehat{x}(y), x)]$ was a *constant* – i.e. each estimator produced a single cost, and we could simply search for the one with the smallest cost. Suppose we try this approach in the nonrandom case, e.g. for the square error cost $J(\widehat{x}(y), x) = (\widehat{x}(y) - x)^2$. Since the only randomness in the problem is with respect to Y a rational approach is to try and find the minimum of this cost averaged with respect to the parameterized density $p_{Y|X}(y \mid x)$. This yields:

$$\widehat{x}^* \quad = \quad \arg\min_{\widehat{x}} \int_{-\infty}^{\infty} (\widehat{x} - x)^2 \, p_{Y|X}(y \mid x) \, dy \tag{10.201}$$

$$= \quad x \tag{10.202}$$

Thus the optimal estimate is just the unknown parameter x itself! Note, this result may seem strange, but recall that the estimator is just a mapping from the data y to a corresponding estimate $\widehat{x}(y)$, thus in performing the minimization in (10.201) we are really asking the question "what is the best value of x to assign this particular value of the observation y to." Clearly, since x is fixed, it is the best value! This is right, but not very useful since we are assuming we do not know its value[1]. What we will do instead is to look at the behavior of the estimation error $e(y) = x - \widehat{x}(y)$ and see if we can find estimators with desirable error behavior. The three measures of error behavior we will be interested in are the bias, the variance, and the mean square error (MSE). We examine each of these in the nonrandom context next.

Bias: In the nonrandom case, the bias $b(x)$ of an estimator $\widehat{x}(y)$ is given by:

$$b(x) = E[e \mid X = x] = E[x - \widehat{x}(y) \mid X = x] = x - \int_{-\infty}^{\infty} \widehat{x}(y) \, p_{Y|X}(y \mid x) \, dy \tag{10.203}$$

Unlike the random parameter case, we cannot average over the prior density of X. The consequence is that $b(x)$ is in general a function of X! In particular, we can define 3 broad cases of bias behavior:

---

[1] This argument will hold true for any cost which is nonnegative and zero when $\widehat{x}(y) = x$.

1. $b(x) = 0$ for all values of $X = x$. In this case we can say that the estimate is unbiased.

2. $b(x) = c$ where c is a constant independent of X. Here the estimator has constant bias. If the constant c is known, we can always obtain an unbiased estimator by simply subtracting c from the estimate.

3. $b(x) = f(x)$ for some function $f(\cdot)$. In this, the general case, the bias is unknown (since X itself is unknown) and we cannot simply subtract it out to obtain an unbiased estimate.

Clearly, we desire estimators whose error is small on average, i.e. who have small bias.

Variance: Having small bias is not enough. The average behavior of an estimator may be good, yet its its variability may be high. What we also would like is for the variance of the estimate to be small so we are confident that on any particular instance the estimate is close to the true value. For the nonrandom case the error covariance matrix is given by:

$$\Lambda_e(x) \quad = \quad \text{Cov}\,[e, e] = E\left[[e - b(x)][e - b(x)]^T\right] = E\left[ee^T\right] - b(x)b^T(x) \tag{10.204}$$

This provides a measure of the spread of the error. Again, $\Lambda_e(x)$ is a function of X in general.

MSE: The last measure of estimator quality we will be concerned with is the mean square error or MSE. This is given by:

$$\text{MSE} \quad = \quad E\left[e^T e\right] = \text{tr}\left(E\left[ee^T\right]\right) = \text{tr}\left[\Lambda_e(x) + b(x)b(x)^T\right] \tag{10.205}$$

Thus we see that the MSE is a function of both the variance and the bias. In particular, we do not simply want to minimze the variance if this will lead to a large bias. For example, we could take as our estimate a constant C independent of y. The variance $\Lambda_e(x)$ of this estimator would be identically zero, but the bias would be $x - C$ and could be large.

   In general, we seek unbiased estimators of minimum variance. Unfortunately, there is no straightforward procedure that leads to minimum variance unbiased estimators in the nonrandom case, thus we have to define an estimator and see how well it works. In this search for good estimators it is useful to know how good *any* unbiased estimator can do. Then we have a yardstick against which to measure a given estimator. We provide such a bound next.

10.6.1   Cramer-Rao Bound

Let $\widehat{x}(\underline{y})$ be any *unbiased* estimate of the deterministic but unknown (scalar) parameter X and let $\lambda_e(x) = E[(x - \widehat{x}(\underline{y}))^2]$ be its associated error covariance (which is also the mean square error since its unbiased).

   The Cramer-Rao Bound is a bound on the estimation error covariance of *any* unbiased estimate of the deterministic but unknown parameter X. The result is as follows:

Theorem 10.1 (Cramer-Rao Bound)
If $\widehat{x}(\underline{y})$ is any *unbiased* estimate of the deterministic parameter X, and $\lambda_e(x) = E[(x - \widehat{x}(\underline{y}))^2]$ is its associated error covariance, then:

$$\lambda_e(x) \geq \frac{1}{I_y(x)} \tag{10.206}$$

where $I_y(x)$ is given by:

$$I_y(x) \quad = \quad E\left[\left.\left(\frac{\partial}{\partial x} \ln p_{Y|X}(y \mid x)\right)^2 \right| X = x\right] \tag{10.207}$$

$$= \quad -E\left[\left. \frac{\partial^2}{\partial x^2} \ln p_{Y|X}(y \mid x) \right| X = x\right] \tag{10.208}$$

The quantity $I_y(x)$, which plays a central role in the CRB is called the *Fisher information* in y about x. Any unbiased estimator that achieves the CRB is termed *efficient*. While we do not discuss them here, there are also vector forms of the CRB and extensions to account for biased estimators.

Let us now prove the CRB and its two alternate expressions. Let $\widehat{x}(y)$ be any unbiased estimate of x and define the error in this estimate as $e(y) = \widehat{x}(y) - x$. The error e is a random variable itself since it depends on y. In particular, since the estimate in question is unbiased we know that $E[e] = 0$. Note also that $E[e^2] = \lambda_e(x)$, the error variance associated with the estimate $\widehat{x}(y)$. Since the estimate is unbiased we have:

$$E[e] \quad = \quad \int_{-\infty}^{\infty} (\widehat{x}(y) - x) p_{Y|X}(y \mid x)\, dy = 0 \tag{10.209}$$

Now differentiating with respect to x and using the chain rule we obtain:

$$\frac{\partial}{\partial x} \int_{-\infty}^{\infty} (\widehat{x}(y) - x)\, p_{Y|X}(y \mid x)\, dy \quad = \quad \int_{-\infty}^{\infty} \left[(\widehat{x}(y) - x)\frac{\partial}{\partial x} p_{Y|X}(y \mid x) - p_{Y|X}(y \mid x) \right] dy \tag{10.210}$$

$$= \quad 0 \tag{10.211}$$

Now the second term in the integral integrates to 1 so we have:

$$\int_{-\infty}^{\infty} (\widehat{x}(y) - x)\, \frac{\partial}{\partial x} p_{Y|X}(y \mid x)\, dy \quad = \quad \int_{-\infty}^{\infty} (\widehat{x}(y) - x)\, p_{Y|X}(y \mid x)\frac{\partial}{\partial x}\ln p_{Y|X}(y \mid x)\, dy \tag{10.212}$$

$$= \quad 1 \tag{10.213}$$

Now since the expression is equal to 1 we can square it:

$$1 \quad = \quad \left(\int_{-\infty}^{\infty} (\widehat{x}(y) - x)\, p_{Y|X}(y \mid x)\frac{\partial}{\partial x}\ln p_{Y|X}(y \mid x)\, dy \right)^2 \tag{10.214}$$

$$= \quad \left(\int_{-\infty}^{\infty} \left[(\widehat{x}(y) - x)\sqrt{p_{Y|X}(y \mid x)} \right] \left[\sqrt{p_{Y|X}(y \mid x)}\frac{\partial}{\partial x}\ln p_{Y|X}(y \mid x) \right] dy \right)^2 \tag{10.215}$$

$$\leq \quad \left[\int_{-\infty}^{\infty} (\widehat{x}(y) - x)^2\, p_{Y|X}(y \mid x)\, dy \right] \left[\int_{-\infty}^{\infty} p_{Y|X}(y \mid x) \left[\frac{\partial}{\partial x}\ln p_{Y|X}(y \mid x) \right]^2 dy \right] \tag{10.216}$$

$$= \quad \lambda_e(x)\ E\left[\left(\frac{\partial}{\partial x}\ln p_{Y|X}(y \mid x) \right)^2 \Bigg| X = x \right] \tag{10.217}$$

$$= \quad \lambda_e(x)\ I_y(x) \tag{10.218}$$

where the inequality follows from the Schwartz inequality for functions, which states that:

$$\left(\int_{-\infty}^{\infty} f_1(y)\, f_2(y)\, dy \right)^2 \leq \left(\int_{-\infty}^{\infty} f_1^2(y)\, dy \right) \left(\int_{-\infty}^{\infty} f_2^2(y)\, dy \right)$$

Thus we see that:

$$\lambda_e(x) \quad \geq \quad \frac{1}{I_y(x)} \tag{10.219}$$

as desired.

The second form of the CRB can be shown as follows. Observe that:

$$\int_{-\infty}^{\infty} p_{Y|X}(y \mid x)\, dy \quad = \quad 1 \tag{10.220}$$

Differentiating once with respect to x we obtain

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial x} p_{Y|X}(y \mid x)\, dy = \int_{-\infty}^{\infty} p_{Y|X}(y \mid x)\frac{\partial}{\partial x}\ln p_{Y|X}(y \mid x)\, dy = 0 \tag{10.221}$$

Now differentiating this resulting expression with respect to x yields:

$$0 = \frac{\partial}{\partial x} \int_{-\infty}^{\infty} p_{Y|X}(y \mid x) \frac{\partial}{\partial x} \ln p_{Y|X}(y \mid x) \, dy \tag{10.222}$$

$$= \int_{-\infty}^{\infty} p_{Y|X}(y \mid x) \frac{\partial^2}{\partial x^2} \ln p_{Y|X}(y \mid x) \, dy + \int_{-\infty}^{\infty} \frac{\partial}{\partial x} p_{Y|X}(y \mid x) \frac{\partial}{\partial x} \ln p_{Y|X}(y \mid x) \, dy \tag{10.223}$$

$$= \int_{-\infty}^{\infty} p_{Y|X}(y \mid x) \frac{\partial^2}{\partial x^2} \ln p_{Y|X}(y \mid x) \, dy + \int_{-\infty}^{\infty} p_{Y|X}(y \mid x) \left[\frac{\partial}{\partial x} \ln p_{Y|X}(y \mid x) \right]^2 dy \tag{10.224}$$

$$= E \left\{ \frac{\partial^2}{\partial x^2} \ln p_{Y|X}(y \mid x) \Big| X = x \right\} + E \left\{ \left[\frac{\partial}{\partial x} \ln p_{Y|X}(y \mid x) \right]^2 \Big| X = x \right\} \tag{10.225}$$

Where we have used the fact that:

$$p_{Y|X}(y \mid x) \frac{\partial}{\partial x} \ln p_{Y|X}(y \mid x) = \frac{\partial}{\partial x} p_{Y|X}(y \mid x) \tag{10.226}$$

Thus from (10.225) we see that:

$$E \left\{ \frac{\partial^2}{\partial x^2} \ln p_{Y|X}(y \mid x) \Big| X = x \right\} = -E \left\{ \left[\frac{\partial}{\partial x} \ln p_{Y|X}(y \mid x) \right]^2 \Big| X = x \right\} \tag{10.227}$$

which demonstrates the equivalance.

Now from the definition of the Schwarz inequality on which the CRB is based, equality in the CRB holds if and only if:

$$\widehat{x}(y) - x = k(x) \frac{\partial}{\partial x} \ln p_{Y|X}(y \mid x) \tag{10.228}$$

for some $k(x) > 0$. Now when equality holds in the CRB the variance of both sides of (10.228) must be the same. The variance of the left hand side is given by $\Lambda_e(x) = 1/I_y(x)$ while the variance of the right hand side is given by $k^2 I_y(x)$, thus $k(x) = 1/I_y(x)$. This implies that $\widehat{x}(y)$ is an efficient estimate if and only if

$$\widehat{x}(y) = x + \frac{1}{I_y(x)} \frac{\partial}{\partial x} \ln p_{Y|X}(y \mid x) \tag{10.229}$$

Now since the left hand side of (10.229) is only a function of y, an unbiased efficient estimator will exist if and only if the right hand function is independent of x. This give us the following result:

Theorem 10.2 (Existance of Efficient Estimator)
An unbiased efficient estimator of the nonrandom parameter x exists if and only if

$$x + \frac{1}{I_y(x)} \frac{\partial}{\partial x} \ln p_{Y|X}(y \mid x) \tag{10.230}$$

is independent of x, where $I_y(x)$ is the Fisher information in y about x.

The expression (10.230) does not depend on knowledge of a particular estimator and is computable. This gives us a way of telling if an efficient estimator exists for a given situation without needing to know the estimator.

Example 10.19
Suppose we have the following measurement:

$$y = hx + w \qquad w \sim N(0, r) \tag{10.231}$$

and consider the estimator of x given by:

$$\widehat{x}(y) = \frac{y}{h} \tag{10.232}$$

The bias of this estimator is given by:

$$E[\widehat{x}(y)] = E\left[\frac{y}{h}\right] = E\left[x + \frac{w}{h}\right] = x \tag{10.233}$$

So the estimator is unbiased. Now consider the error variance:

$$\lambda(x) = \mathsf{Var}(\widehat{x}(y)) = E\left[(x - \widehat{x}(y))^2\right] = E\left[\frac{w^2}{h^2}\right] = \frac{r}{h^2} \tag{10.234}$$

Now lets compute the CRB and see how good the estimator is. First note that:

$$p_{Y|X}(y \mid x) \quad = \quad N(y; hx, r) \tag{10.235}$$

Therefore:

$$\ln(p_{Y|X}(y \mid x)) \quad = \quad -\ln\left(\sqrt{2\pi r}\right) - \frac{(y - hx)^2}{2r} \tag{10.236}$$

$$\implies \quad \frac{\partial^2}{\partial x^2}\left[\ln p_{Y|X}(y \mid x)\right] \quad = \quad -\frac{h^2}{r} \tag{10.237}$$

$$\implies \quad I_y(x) \quad = \quad -E\left[\frac{\partial^2}{\partial x^2}\ln p_{Y|X}(y \mid x)\right] = \frac{h^2}{r} \tag{10.238}$$

$$\implies \quad \lambda(x) \quad \geq \quad \frac{1}{I_y(x)} = \frac{r}{h^2} \tag{10.239}$$

So we see that the given estimator is efficient

Example 10.20
In this example suppose we want to estimate x from the following nonlinear observation:

$$y = h(x) + w \qquad w \sim N(0, r) \tag{10.240}$$

With some calculation we can show that

$$I_y(x) \quad = \quad \frac{\left(\frac{\partial h(x)}{\partial x}\right)^2}{r} \tag{10.241}$$

Computing the criterion (10.229) yields:

$$x + \frac{1}{I_y(x)}\frac{\partial}{\partial x}\ln p_{Y|X}(y \mid x) \quad = \quad x + \frac{y - h(x)}{\frac{\partial h}{\partial x}} \tag{10.242}$$

For an efficient unbiased estimator to exist this expression must be indepent of x. Suppose $h(x) = x^3$. In this case we find:

$$x + \frac{y - h(x)}{\frac{\partial h}{\partial x}} \quad = \quad x + \frac{y - x^3}{3x^2} = \frac{2}{3}x - \frac{1}{3}\frac{y}{x^2} \tag{10.243}$$

This is a function of x so we can tell that no efficient estimator exists when $h(x) = x^3$.

10.6.2 Maximum-Likelihood Estimation

One reasonable approach to estimation in the nonrandom parameter case is the *maximum likelihood method*. In general, we denote the function $p_{Y|X}(y \mid x)$ viewed as a function of x as the *likelihood function*.

Definition 10.1 (Maximum-Likelihood Estimate)
The *maximum likelihood estimate* $\widehat{x}_{ML}(y)$ is that value of x for which the likelihood function is maximized:

$$\boxed{\widehat{x}_{ML}(y) = \arg\max_x p_{Y|X}(y \mid x)} \tag{10.244}$$

Let us give a graphical interpretation to this maximization. As shown in Figure 10.7(a), for each value of x we obtain a density for y. In this case $p_{Y|X}(y \mid x)$ is viewed as a function of y for each fixed x. At a particular observation $y = y_0$ we can imagine finding the value of these densities as we change x. The figure shows the values of the densities for two such values of x (x_1 and x_2) for a given observation. If we now plot these values as a function of x we obtain the graph shown in Figure 10.7(b). In this case $p_{Y|X}(y_0 \mid x)$ is viewed as a function of x with $y = y_0$ fixed. Note that $p_{Y|X}(y_0 \mid x)$ (i.e. the function plotted in (b)) is *not* a density, but rather a graph of the density values as the parameter is changed. For example there is no requirement that $p_{Y|X}(y_0 \mid x)$ integrated over x sum to one. The ML estimate is the maximum of this graph. Finally, in Figure 10.7(c) we plot $p_{Y|X}(y \mid x)$ as a function of *both* x and y. In this view, the ML estimate is the maximum of the corresponding surface in the x direction for the given observed value of $y = y_0$.

(a)

(b)

(c)
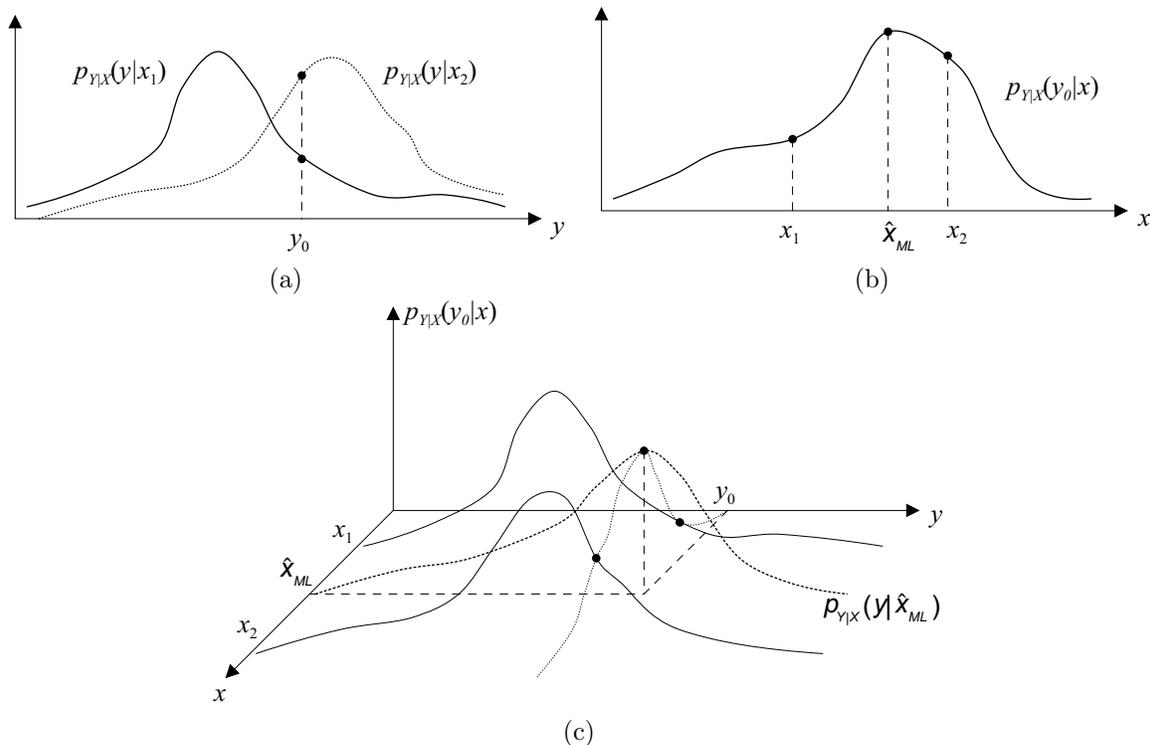
Figure 10.7: Interpretation of the ML Estimator: (a) $p_{Y|X}(y \mid x)$ viewed as a function of y for fixed values of x, (b) $p_{Y|X}(y \mid x)$ viewed as a function of x for fixed y, (c) $p_{Y|X}(y \mid x)$ viewed as a function of both x and y. For a given observation y_0, $\widehat{x}_{ML}(y)$ is the maximum with respect to x for the given $y = y_0$.

In practice, we often work with the logarithm of the likelihood function $\ln p_{Y|X}(y \mid x)$ which is called the log likelihood function. If the maximum is interior to the range of x and the log likelihood function has a continuous first derivative, then the ML estimate must satisfy the following *ML equation:*

$$\left. \frac{\partial \ln p_{Y|X}(y \mid x)}{\partial x} \right|_{x=\widehat{x}_{ML}(y)} = 0 \tag{10.245}$$

To show a fundamental and important property of ML estimates, consider the condition for an efficient estimate given in (10.229). Now if $\widehat{x}(y)$ is any efficient estimate then it must satisfy (10.229) evaluated at any value of x. Suppose we use $x = \widehat{x}_{ML}(y)$. From the definition (10.245) we see that when $x = \widehat{x}_{ML}(y)$ the second term in (10.229) is equal to zero. Thus:

$$\widehat{x}(y) = \widehat{x}_{ML}(y) \tag{10.246}$$

so that *if an efficient unbiased estimator exists it must be an ML estimator.*

Example 10.21
Consider the problem of Example 1 again. Let us find the ML estimator for this case:

$$\ln\left(p_{Y|X}(y \mid x)\right) = -\ln\left(\sqrt{2\pi r}\right) - \frac{(y - hx)^2}{2r} \tag{10.247}$$

$$\implies \frac{\partial}{\partial x}\left[\ln p_{Y|X}(y \mid x)\right] = \frac{2hy - 2h^2x}{2r} = 0 \tag{10.248}$$

$$\implies \widehat{x}_{ML}(y) = \frac{y}{h} \tag{10.249}$$

Thus the estimator we examined in Example 1 was really the ML estimator. We already know it is unbiased and efficient.

Example 10.22
Suppose the observation $y \geq 0$ is given by an exponential distribution with parameter x:

$$p_{Y|X}(y \mid x) = \frac{1}{x}e^{-y/x} \tag{10.250}$$

where $x \geq 0$. The maximum likelihood estimate is obtained from:

$$\frac{\partial}{\partial x}\left[\ln p_{Y|X}(y \mid x)\right] = \frac{\partial}{\partial x}\left[-\ln x - \frac{y}{x}\right] = -\frac{1}{x} + \frac{y}{x^2} = 0 \tag{10.251}$$

$$\implies \widehat{x}_{ML}(y) = y \tag{10.252}$$

Now the bias of this estimate is given by:

$$E[x - \widehat{x}(y)] = E[x - y] = 0 \tag{10.253}$$

so the ML estimate is unbiased. Next lets find the variance:

$$\lambda(x) = \mathsf{Var}(\widehat{x}(y)) = E\left[(x - \widehat{x}(y))^2\right] = E\left[(y - x)^2\right] = x^2 \tag{10.254}$$

since the variance of the exponentially distributed random variable y is x^2. Note that the error variance is a function of x in this problem. Now lets compute the CRB:

$$I_y(x) = E\left[\left(\frac{\partial}{\partial x}\ln p_{Y|X}(y \mid x)\right)^2\right] = E\left[\frac{(y - x)^2}{x^4}\right] = \frac{x^2}{x^4} = \frac{1}{x^2} \tag{10.255}$$

where we have again used the fact that the variance of y is x^2. We find that:

$$\lambda(x) = x^2 = \frac{1}{I_y(x)} \tag{10.256}$$

and thus the ML estimate is also efficient.

10.6.3   Comparison to MAP estimation

Finally let us compare MAP and ML estimation. Recall from our treatment of the detection problem that these two forms of detection were closely related. Specifically, the ML detection rule was presented as a special case of MAP detection wherein the prior probabilities of the hypotheses were the same. We will see that, despite their differences, a similar tie can be made in the estimation context. To this end, consider the estimation problem of Example 1 again, where we desire an estimate of x based on the observation:

$$y = hx + w \quad w \sim N(0, r) \tag{10.257}$$

We have already found the ML estimate and estimation error variance for this problem in (10.249) and (10.234). Now suppose that in addition we have the following prior information on x:

$$x \sim N(m_x, \lambda_x) \tag{10.258}$$

In this case the MAP estimate is given by:

$$\widehat{x}_{MAP}(y) = m_x + \frac{h\lambda_x}{h^2\lambda_x + r}(y - hm_x) = \underbrace{\left[\frac{r/h^2}{\lambda_x + r/h^2}\right]}_{\text{Prior Weight}}\underbrace{m_x}_{\text{Prior Est}} + \underbrace{\left[\frac{\lambda_x}{\lambda_x + r/h^2}\right]}_{\text{Obs Weight}}\underbrace{\frac{y}{h}}_{\widehat{x}_{ML}(y)} \tag{10.259}$$

We can see that the MAP estimate is composed of two parts. A part due to the prior m_x and a part that corresponds precisely to the ML estimate y/h. These two parts are weighted according to their relative reliability. In particular, suppose $\lambda_x \to \infty$, then we see that $\widehat{x}_{MAP}(y) \to \widehat{x}_{ML}(y)$. That is, as the prior information goes to zero the MAP estimate approaches the ML estimate. These observations are summarized in Table 10.1.

MAP Estimation	ML Estimation
$\begin{aligned} y &= hx + v, \quad v \sim N(0, r) \\ x &\sim N(m_x, \lambda_x) \end{aligned}$	$y = hx + v, \quad v \sim N(0, r)$
$\widehat{x}_{MAP}(y) = \underbrace{\left[\dfrac{r/h^2}{\lambda_x + r/h^2}\right]}_{\text{Prior Weight}} \underbrace{m_x}_{\text{Prior Est}} + \underbrace{\left[\dfrac{\lambda_x}{\lambda_x + r/h^2}\right]}_{\text{Obs Weight}} \underbrace{\dfrac{y}{h}}_{\widehat{x}_{ML}(y)}$	$\widehat{x}_{ML}(y) = \dfrac{y}{h}$
$\underbrace{\left(\dfrac{1}{\lambda_{MAP}}\right)}_{\text{Total Info}} = \underbrace{\left(\dfrac{1}{\lambda_x}\right)}_{\text{Prior Info}} + \underbrace{\left(\dfrac{1}{\lambda_{ML}}\right)}_{\text{Obs Info}}$ $\implies \lambda_{MAP} = \dfrac{1}{\frac{1}{\lambda_x} + \frac{1}{\lambda_{ML}}} \le \lambda_{ML}$	$\lambda_{ML} = \dfrac{r}{h^2}$

Table 10.1: Comparison of MAP and ML Estimation for a particular example.

This discussion has focussed on comparison of MAP and ML approaches for a particular example. More generally we can compare the MAP and ML equations that the corresponding estimates must satisfy:

$$\begin{array}{c|c}
\text{MAP Equation} & \text{ML Equation} \\[4pt]
\left[\dfrac{\partial \ln p_{Y|X}(y \mid x)}{\partial x} + \dfrac{\partial \ln p_X(x)}{\partial x}\right]\Bigg|_{x=\widehat{x}_{MAP}(y)} = 0 & \left[\dfrac{\partial \ln p_{Y|X}(y \mid x)}{\partial x}\right]\Bigg|_{x=\widehat{x}_{ML}(y)} = 0
\end{array} \qquad (10.260)$$

We can directly see that $\widehat{x}_{MAP}(y) \to \widehat{x}_{ML}(y)$ as $p_X(x)$ become flatter and flatter over all of x so that $\partial/\partial x \to 0$. Again, this implies there our prior information is going to zero as well, since the pdf for X becomes uniformly distributed over its entire range. Thus we see, as in the detection problem, that, for the same observation model, as the prior becomes more uniform in the MAP case, the MAP estimate approaches the ML estimate.

Chapter 11

LLSE Estimation of Stochastic Processes and Wiener Filtering

11.1 Introduction

In Chapter 10 we studied the estimation of random variables based on observation of other random variables. In this chapter we extend this work to study the estimation of stochastic processes. Our focus will be the problem of finding the best *linear* minimum mean square error estimate (i.e. the linear least squares estimate – LLSE) of the zero-mean random process $X(t)$ based on observation of the zero-mean process $Y(\tau)$ for $T_i \leq \tau \leq T_f$. If the process is not zero mean, then we estimate $\tilde{X}(t) = X(t) - m_x(t)$ based on $\tilde{Y}(\tau) = Y(\tau) - m_y(\tau)$ and substitute the definitions of $\tilde{X}(t)$ and $\tilde{Y}(\tau)$ in at the end, as usual (if this bothers you see Appendix D). In our search for the best estimate we assume that the second order statistics of the processes $K_{XX}(t,s)$, $K_{YX}(t,s)$, and $K_{YY}(t,s)$ are known. Note, since we are assuming the means are zero $K_{XX}(t,s) = R_{XX}(t,s)$, etc. In the development to follow we will present the results for both the continuous and the discrete time cases concurrently, since they are so similar. In summary then, the problem setup is given by the following:

$$X(t) \quad = \quad \text{Process to estimate} \tag{11.1}$$

$$Y(\tau); \quad T_i \leq \tau \leq T_f \quad = \quad \text{Observed Process} \tag{11.2}$$

$$\text{Given:} \quad K_{XX}(t,s), \ K_{YX}(t,s), \ K_{YY}(t,s) \tag{11.3}$$

$$x(t), y(\tau) \text{ zero mean} \tag{11.4}$$

Since we want a linear estimate we know the estimate will be a linear combination of the points in the observation interval. For the discrete case, which we already have experience with, this corresponds to a simple weighted sum, while for the continuous case this operation corresponds to a weighted integral. In particular, the form of the estimator for each case will be:

$$\text{CT:} \quad \widehat{x}(t) \quad = \quad \int_{T_i}^{T_f} h(t,\sigma)\, y(\sigma)\, d\sigma \tag{11.5}$$

$$\text{DT:} \quad \widehat{x}(t) \quad = \quad \sum_{\sigma=T_i}^{T_f} h(t,\sigma)\, y(\sigma) \tag{11.6}$$

Thus the estimate may be viewed as the output of a *time-varying* linear filter whose weighting pattern $h(t,\sigma)$ defines the estimator, as shown in Figure 11.1. When estimating one stochastic process based on observation of another then, the estimator itself may be simply interpreted as a *filter*, and LLSE estimation viewed as a problem of filter design.

There is some standard terminology that is used in describing such problems depending on the time t the estimate $\widehat{x}(t)$ is generated relative to the time interval $T_i \leq \tau \leq T_f$ of the observation. These correspond to
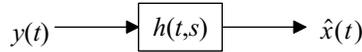
Figure 11.1: Linear Estimator for a Stochastic Process.

whether the estimate is on the boundary, the interior, or the exterior of the observation interval. The most common are:

$$
\begin{aligned}
T_i < t < T_f &\quad : \quad \text{Smoothing, Noncausal filtering} \\
t = T_f &\quad : \quad \text{Filtering} \\
T_f < t &\quad : \quad \text{Prediction}
\end{aligned}
$$

The relationship between the time the estimate is generated and the time interval of the observation is shown in Figure 11.2.
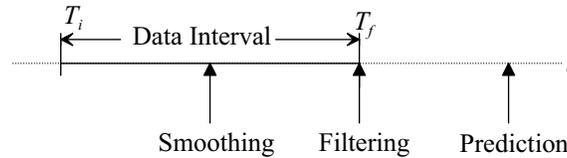
Figure 11.2: Estimation Types Based on Relative Times of Observation and Estimate.

11.2 Historical Context

Before proceding to the mathematical developments leading to the LLSE estimate, it is useful (and perhaps some will think more interesting!) to first consider the historical context for its development, and hence better understand the type of the result we will obtain. The problem of linear least square estimation of one stochastic process based on observation of another stochastic process was originaly done for the continuous case by Norbert Wiener and hence often is called "Wiener Filtering," even when applied to the discrete-time situation. Actually, the discrete case was solved by the Russian mathematician Kolmogorov.

To understand the goal in the mind of these pioneers, let us start with a question. What do you think of when you think of a "filter"? In other words, if asked to design or implement a filter by e.g. your boss or instructor, what would you bring back as your result? (Please take a minute to think about this before reading on!). In the present day, when filter banks rule the DSP journals and wavelets are everywhere, most people would immediately think of a *digital* definition of a filter, and of a filter as being defined e.g. by its vectors of delay coefficients or some such – at least that is what I do. But such was not always the case, and we must think about how Wiener saw the world to understand the form of his solution. Let us begin with a brief history of Norbert Wiener.

Norbert Wiener was born Noverber 26, 1894 in Columbia, MO and died March 18, 1964 in Stockholm Sweden. He was a child prodigy, finishing high school at the age of 11 and getting an undergraduate degree from Tufts in mathematics at age 14. He went on to Harvard and obtained his Ph.D. in Mathematical Logic at age 18! Wiener then went abroad and stuided under the great mathematicians Russel, Hardy, and Hilbert. Finally in 1919 he obtained a teaching appointment at MIT, where he remained for the rest of his life. Wiener contributed to many areas including cybernetics (a term he coined), stochastic processes, and quantum theory.

Thus we can see that Wiener was active during the period from the 1920's to, say, the 1950's. The major world event during this period was World War II, and this event formed the backdrop for Wiener's work. In particular, Wiener worked on gun fire control at MIT – the goal being to direct a gun to shoot down an airplane. With this motivation, Wiener worked through the 1930's on the problem of estimation and

prediction of continuous-time processes – which is all that really mattered at the time! The first general purpose electromechanical digital computer, the "Mark I," was build at Harvard in 1944 (a multiply took 4 seconds) and the first electronic digital computer was built in 1946, as the war drew to a close. The results of Wiener's work on this problem (not declassified till the late 1940's) were written up in an internal MIT technical report entitled "Extrapolation, Interpolation, and Smoothing of Stationary Time Series," known popularly by engineers of the time as "the yellow peril", due to the yellow color of its original cover.

For Wiener, working when he did, a filter was truly an analog device composed of capacitors, resistors, and the like. As such, specification of an realizable filter required specification of its poles and zeros – in other words, what was needed was a *closed form solution* to the problem if it was to be implemented. We will contrast this to the recursive *algorithm* comprising the Kalman filter later. Given this view, the focus of work during Wiener's time and for some time beyond was on an closed form solution to the LLSE estimation problem and the explicit specification of the corresponding filter. Let us now proceed to find this solution.

11.3   LLSE Problem Solution: The Wiener-Hopf Equation

The solution to the problem described in (11.1)–(11.4) can be found through use of the *orthogonality principle* discussed previously. In particular, we know that the optimal LLSE estimate will have the property that it is unbiased and that the error is orthogonal to the estimate:

$$E\left[(x(t) - \widehat{x}(t))\, y(\tau)\right] = 0 \quad \forall \tau \in [T_i,\, T_f] \tag{11.7}$$

Expanding this expression, we obtain the following condition that the optimal LLSE estimate must satisfy:

$$K_{XY}(t,\tau) \;\;=\;\; K_{\widehat{X}Y}(t,\tau) \quad \forall \tau \in [T_i,\, T_f] \tag{11.8}$$

Note that this condition says that the optimal estimate $\widehat{x}(t)$ has the same cross-correlation with the data as the true process $x(t)$.

We can find $K_{\widehat{X}Y}(t,\tau)$ using the definition of cross-correlation and (11.5) or (11.6). Working with the CT case:

$$K_{\widehat{X}Y}(t,\tau) \;\;=\;\; E\left[\widehat{x}(t)y(\tau)\right] \tag{11.9}$$

$$=\;\; \int_{T_i}^{T_f} h(t,\sigma)\, E[y(\sigma)y(\tau)]\, d\sigma \tag{11.10}$$

$$=\;\; \int_{T_i}^{T_f} h(t,\sigma)\, K_{YY}(\sigma,\tau)\, d\sigma \tag{11.11}$$

Similarly in discrete time we have:

$$K_{\widehat{X}Y}(t,\tau) \;\;=\;\; \sum_{\sigma=T_i}^{T_f} h(t,\sigma)\, K_{YY}(\sigma,\tau) \tag{11.12}$$

Now we can substitute (11.11) or (11.12) into (11.8) to obtain the following conditions that the optimal filter $h(t,\sigma)$ must satisfy:

$$\text{CT:} \quad K_{XY}(t,\tau) \;\;=\;\; \int_{T_i}^{T_f} h(t,\sigma)\, K_{YY}(\sigma,\tau)\, d\sigma \quad \forall \tau \in [T_i,\, T_f] \tag{11.13}$$

$$\text{DT:} \quad K_{XY}(t,\tau) \;\;=\;\; \sum_{\sigma=T_i}^{T_f} h(t,\sigma)\, K_{YY}(\sigma,\tau) \quad \forall \tau \in [T_i,\, T_f] \tag{11.14}$$

This equation is called the *Wiener-Hopf* equation, and captures the conditions that the optimal estimate must satisfy.

In addition to the Wiener-Hopf equation for the optimal estimate, we can also get expressions for the estimation error variance $\Lambda_{LSE}(t)$:

$$
\begin{align}
\Lambda_{LSE}(t) &= E\left[(x(t) - \widehat{x}(t))^2\right] = E\left[(x(t) - \widehat{x}(t))(x(t) - \widehat{x}(t))\right] \tag{11.15} \\
&= E\left[x(t)(x(t) - \widehat{x}(t))\right] - E\left[\widehat{x}(t)(x(t) - \widehat{x}(t))\right] \tag{11.16} \\
&= K_{XX}(t,t) - K_{X\widehat{X}}(t,t) \tag{11.17}
\end{align}
$$

where we have used the fact that the second term in (11.16) is zero since the error is orthogonal to linear functions of the data, in this case the estimate itself. Now we can calculate $K_{X\widehat{X}}(t,t)$ from the definition of covariance and (11.5) or (11.6). Doing this yields the following expressions for the error covariance:

$$
\text{CT:} \quad \Lambda_{LSE}(t) = K_{XX}(t,t) - \int_{T_i}^{T_f} h(t,\sigma)\, K_{YX}(\sigma,t)\, d\sigma \tag{11.18}
$$

$$
\text{DT:} \quad \Lambda_{LSE}(t) = K_{XX}(t,t) - \sum_{\sigma=T_i}^{T_f} h(t,\sigma)\, K_{YX}(\sigma,t) \tag{11.19}
$$

Before we proceed to Wiener filtering, we stop to point out that while this development may seem new and intimidating, we have visited some of these issues before in our study of random variables. In particular, consider the discrete-time result (11.14) when the times involved are finite. We can collect the equations represented by this set of conditions into vector form to obtain a matrix equation capturing the entire set:

$$
\left[\begin{array}{ccc} K_{XY}(t,T_i), & \cdots & K_{XY}(t,T_f) \end{array}\right] = \tag{11.20}
$$

$$
\left[\begin{array}{ccc} h(t,T_i), & \cdots & h(t,T_f) \end{array}\right]
\left[\begin{array}{cccc}
K_{YY}(T_i,T_i) & K_{YY}(T_i,T_i+1) & \cdots & K_{YY}(T_i,T_f) \\
K_{YY}(T_i+1,T_i) & & & \\
\vdots & & & \vdots \\
K_{YY}(T_f,T_i) & K_{YY}(T_f,T_i+1) & \cdots & K_{YY}(T_f,T_f)
\end{array}\right]
$$

$$
\implies \quad \Lambda_{X\underline{Y}} = \underline{h}^T \Lambda_{\underline{Y}} \tag{11.21}
$$

where we have made the natural matrix/vector associations in the last equation. As can be seen (11.21) are nothing more than the familiar normal equations. Their solution is given by:

$$
\underline{h}^T = \Lambda_{X\underline{Y}} \Lambda_{\underline{Y}}^{-1} \tag{11.22}
$$

Recall that the associated error covariance for this case was given by

$$
\Lambda_{LSE} = \Lambda_X - \Lambda_{X\underline{Y}} \Lambda_{\underline{Y}}^{-1} \Lambda_{X\underline{Y}}^T = \Lambda_X - \underline{h}^T \Lambda_{X\underline{Y}}^T \tag{11.23}
$$

Notice the similarity between (11.23) and (11.19).

Thus solving the Wiener-Hopf equations in the general (finite observation interval) discrete-time case is equivalent to solving the normal equations and, while straightforward, could be computationally challenging for large problem sizes. In our examination of the Kalman filter we will see that in certain situations we can be more computationally efficient. In the continuous time case we must solve the corresponding *integral equation* given by (11.13). Overall, solving the Wiener-Hopf equation is difficult in general and we must look at special cases to proceed further.

11.4 Wiener Filtering

In this section we discuss what is known as *Wiener Filtering*. Wiener filtering problems are a subclass of LLSE problems where additional assumptions are made. In particular, all Wiener filtering problems satisfy the following assumptions:

Definition 11.1 (Wiener Filtering)
- Find LLSE estimate of $x(t)$ based on $y(\tau)$, $\tau \in [T_i, \, T_f]$

- $x(t)$, $y(t)$ are jointly wide-sense stationary

- $T_i = -\infty$

Thus the additional assumptions over and above the LLSE problem are the stationarity of the processes and the fact that we observe the data starting at $T_i = -\infty$. These two additional assumptions assure that there are no transients in the estimate, in particular that the corresponding filter $h(t, \sigma)$ will be *time-invariant*, i.e.:

$$h(t, \sigma) = h(t - \sigma) \tag{11.24}$$

Substituting this into the expressions for the Wiener-Hopf equation (11.13) and (11.14) and making the changes of variables $v = t - \sigma$ and $u = t - \tau$ yields the following expressions for the Wiener-Hopf equations for this case:

$$\text{CT:} \quad K_{YX}(u) \;=\; \int_{t-T_f}^{\infty} h(v) \, K_{YY}(u-v) \, dv \quad t - T_f \leq u \leq \infty \tag{11.25}$$

$$\text{DT:} \quad K_{YX}(u) \;=\; \sum_{v=t-T_f}^{\infty} h(v) \, K_{YY}(u-v) \quad t - T_f \leq u \leq \infty \tag{11.26}$$

There are common subclasses of Wiener filtering problems based on the choice of the right end point of the observation interval T_f. The two cases we will look at in detail are when $T_f = +\infty$, so we have observations for all of time, and when $T_f = t$ so we produce an estimate based only on past observations.

11.4.1   Noncausal Wiener Filtering (Wiener Smoothing)

The first case we will examine is when we choose $T_f = +\infty$, so that we base our estimate at time t on observations of $y(\tau)$ for *all time*: $\tau \in [-\infty, \, \infty]$. This case is referred to as *Wiener smoothing* or *noncausal Wiener filtering*, since the filter will necessarily be noncausal. For this reason the corresponding filter is sometimes referred to as the *unrealizable Wiener filter*.

Definition 11.2 (Noncausal Wiener Filtering or Wiener Smoothing)
- Find LLSE estimate of $x(t)$ based on $y(\tau)$, $\tau \in [T_i, \, T_f]$

- $x(t)$, $y(t)$ are jointly wide-sense stationary

- $T_i = -\infty$

- $T_f = +\infty$

In this case the Wiener-Hopf equations become:

$$\text{CT:} \quad K_{YX}(u) \;=\; \int_{-\infty}^{\infty} h_{nc}(v) \, K_{YY}(u-v) \, dv \quad -\infty \leq u \leq \infty \tag{11.27}$$

$$\text{DT:} \quad K_{YX}(u) \;=\; \sum_{v=-\infty}^{\infty} h_{nc}(v) \, K_{YY}(u-v) \quad -\infty \leq u \leq \infty \tag{11.28}$$

Thus the noncausal Wiener filter is the time invariant impulse that satisfies (11.27) or (11.28). To solve these equations we only need to recognize the expression on the right as a convolution and use e.g. Fourier or Laplace (or Z) transforms. Taking Laplace transforms of (11.27) and z-transforms of (11.28) gives the following:

$$\text{CT:} \quad S_{YX}(s) \;=\; S_{YY}(s) H_{nc}(s) \tag{11.29}$$

$$\text{DT:} \quad S_{YX}(z) \;=\; S_{YY}(z) H_{nc}(z) \tag{11.30}$$

Solving yields for the optimal noncausal Wiener filter:

$$\text{CT:} \quad H_{nc}(s) \quad = \quad \frac{S_{YX}(s)}{S_{YY}(s)} \tag{11.31}$$

$$\text{DT:} \quad H_{nc}(z) \quad = \quad \frac{S_{YX}(z)}{S_{YY}(z)} \tag{11.32}$$

Since we require stationarity we need for the filters to be stable. The region of convergence must therefore include the $j\omega$ axis in continuous time and the unit circle in discrete time. Also, if $S_{YY}(j\omega) = 0$ for some frequency ω (11.29) (or (11.30) in discrete time) shows that the corresponding value of $H_{nc}(j\omega)$ or $H_{nc}(e^{j\omega})$ does not matter, and the estimate is indeterminant at this frequency.

The corresponding equation for the estimation error covariance for the Non-causal Wiener filter Λ_{NCWF} can be obtained from the general formulas (11.18) or (11.19):

$$\text{CT:} \quad \Lambda_{NCWF} \quad = \quad K_{XX}(0) - \int_{-\infty}^{\infty} h(u)\, K_{YX}(u)\, du \tag{11.33}$$

$$\text{DT:} \quad \Lambda_{NCWF} \quad = \quad K_{XX}(0) - \sum_{u=-\infty}^{\infty} h(u)\, K_{YX}(u) \tag{11.34}$$

Notice that these expressions are independent of time, as we might expect given the stationarity assumptions.

Another expression for the estimation error covariance can be obtained by the following line of reasoning, which we elaborate for the continuous time case. First recall that the error is given by $e(t) = x(t) - \widehat{x}(t)$, and that this error is uncorrelated with the data $y(\tau)$ and thus with the estimate $\widehat{x}(t)$, which itself is a linear function of the data. Thus:

$$x(t) \quad = \quad \widehat{x}(t) + e(t) \tag{11.35}$$

$$\Downarrow \tag{11.36}$$

$$S_{XX}(s) \quad = \quad S_{\widehat{x}\widehat{x}}(s) + S_{EE}(s) \tag{11.37}$$

$$\Downarrow \tag{11.38}$$

$$S_{EE}(s) \quad = \quad S_{XX}(s) - S_{\widehat{X}\widehat{X}}(s) \tag{11.39}$$

This expression is valid for *any* LLSE estimator for which the transforms are valid. Now for the case of the noncausal Wiener we have the following expression for the second term:

$$S_{\widehat{X}\widehat{X}}(s) \quad = \quad H_{nc}(s)H_{nc}(-s)S_{YY}(s) \tag{11.40}$$

$$= \quad \frac{S_{YX}(s)}{S_{YY}(s)} \frac{S_{YX}(-s)}{S_{YY}(s)} S_{YY}(s) \tag{11.41}$$

$$= \quad \frac{S_{YX}(s)S_{YX}(-s)}{S_{YY}(s)} \tag{11.42}$$

Thus, substituting (11.42) into (11.37) and solving for $S_{EE}(s)$ we have:

$$S_{EE}(s) \quad = \quad S_{XX}(s) - \frac{S_{YX}(s)S_{YX}(-s)}{S_{YY}(s)} \tag{11.43}$$

If we let $s = j\omega$ we obtain:

$$S_{EE}(j\omega) \quad = \quad S_{XX}(j\omega) - \frac{|S_{YX}(j\omega)|^2}{S_{YY}(j\omega)} \tag{11.44}$$

This expression, though derived for the continuous time case, is also valid for the discrete time case, with appropriate adjustments to the transform definitions. Finally, we have that:

$$\Lambda_{NCWF} \quad = \quad R_{EE}(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_{EE}(j\omega)\, d\omega \tag{11.45}$$

The mean square error can thus be obtained either by finding $R_{EE}(\tau)$ as the inverse transform of $S_{EE}(j\omega)$ and then evaluating the result at $\tau = 0$ or by directly evaluating the integral in (11.45).

Linear Observations and Additive Noise:

We now consider the important special case of a continuous-time process with linear observations and additive noise. Suppose:

$$y(t) = x(t) + v(t) \tag{11.46}$$

where $x(t)$ and $v(t)$ are uncorrelated zero mean wide-sense stationary random processes. We wish to find the noncausal Wiener filter for this problem.

We start by finding the covariances $K_{YX}(t)$ and $K_{YY}(t)$:

$$K_{YX}(t) = E[y(\tau) x(t+\tau)] = E[(x(t) + v(t)) x(t+\tau)] = K_{XX}(t) \tag{11.47}$$

$$K_{YY}(t) = E[(x(\tau) + v(\tau))(x(t+\tau) + v(t+\tau))] = K_{XX}(t) + K_{VV}(t) \tag{11.48}$$

Taking Laplace transforms of the expressions (11.47), (11.48) we obtain the following the power spectral density relationships:

$$S_{YX}(s) = S_{XX}(s) \tag{11.49}$$

$$S_{YY}(s) = S_{XX}(s) + S_{VV}(s) \tag{11.50}$$

Using these expressions in the formula (11.31) for optimal noncausal Wiener filter yields the following filter:

$$H_{nc}(s) = \frac{S_{YX}(s)}{S_{YY}(s)} = \frac{S_{XX}(s)}{S_{XX}(s) + S_{VV}(s)} \tag{11.51}$$

Note that this expression for the filter is real, even and nonnegative, thus it indeed corresponds to a two-sided or noncausal filter impulse response $h(t)$.

We can also obtain an expression for the power spectral density of the estimation error covariance using (11.43):

$$S_{EE}(s) = S_{XX}(s) - \frac{S_{YX}(s)S_{YX}(-s)}{S_{YY}(s)} = S_{XX}(s) - \frac{S_{XX}^2(s)}{S_{XX}(s) + S_{VV}(s)} = \frac{S_{XX}(s)S_{VV}(s)}{S_{XX}(s) + S_{VV}(s)} \tag{11.52}$$

$$= H_{nc}(s)S_{VV}(s) \tag{11.53}$$

Before proceeding to an example let us interpret the behavior of the filter (11.51). From its form we can see that this filter does a reasonable thing. In particular, at frequencies where the power in the signal $S_{XX}(j\omega)$ is large relative to the power in the additive noise $S_{VV}(j\omega)$, the filter $H(j\omega) \approx 1$, looks like an ideal all pass at that frequency, and thus allows the signal to pass unaltered. Conversely, at those frequencies where the power in the noise $S_{VV}(j\omega)$ is large relative to the power in the signal $S_{XX}(j\omega)$, the filter $H(j\omega) \approx 0$ and attenuates both components. Similar results hold for the discrete-time case as well.

Example 11.1 (Single Pole Spectrum:)
Let us apply the above results to a particular example. Specifically suppose $x(t)$ and $y(t)$ are related by (11.46) and

$$K_{XX}(t) = Qe^{-\alpha|t|}, \quad \alpha > 0 \tag{11.54}$$

$$K_{VV}(t) = R\delta(t) \tag{11.55}$$

and we want to find the noncausal Wiener filter. Taking transforms and applying the formulas (11.49),(11.50) we find:

$$S_{XX}(s) = \frac{2Q\alpha}{\alpha^2 - s^2} \tag{11.56}$$

$$S_{VV}(s) = R \tag{11.57}$$

$$S_{YY}(s) = S_{XX}(s) + S_{VV}(s) = \frac{R(\beta^2 - s^2)}{\alpha^2 - s^2}; \quad \beta^2 = \frac{2Q\alpha}{R} + \alpha^2 \tag{11.58}$$

Now we apply (11.51) to find the optimal estimate:

$$H_{nc}(s) = \frac{2Q\alpha}{(\alpha^2 - s^2)} \frac{(\alpha^2 - s^2)}{R(\beta^2 - s^2)} = \frac{2Q\alpha}{R(\beta^2 - s^2)} \tag{11.59}$$

$$\implies H_{nc}(f) = \frac{Q\alpha}{R\beta} \left[\frac{2/\beta}{1 + \left(\frac{2\pi f}{\beta}\right)^2} \right] \tag{11.60}$$

We can find the corresponding impulse response by taking the inverse transform of $H_{nc}(s)$ in (11.59):

$$h_{nc}(t) \quad = \quad \frac{Q\alpha}{R\beta}e^{-\beta|t|} \tag{11.61}$$

This impulse response is sketched in Figure 11.3.
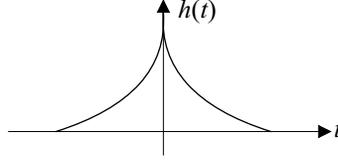
Figure 11.3: Impulse response of noncausal Wiener filter of example.

This is a single pole filter with its bandwidth equal to β. The filter looks at the power in the signal and the power in the noise at each frequency and adjusts its gain accordingly. In Figure 11.4 we show the power spectra of the signal and the noise. As $R \to \infty$, so the signal to noise ratio goes to zero, the bandwidth of $H_{nc}(j\omega)$ approaches the bandwidth of $S_{XX}(j\omega)$ and the overall amplitude goes to zero. As $R \to 0$, so the signal to noise ratio goes to infinity, the bandwidth of $H_{nc}(j\omega)$ approaches infinity and the filter approaches an all pass.
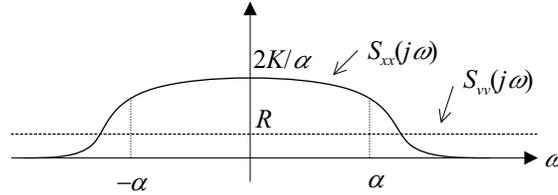
Figure 11.4: Power spectra of signal and noise for example.

Finally, we can obtain the error variance using any of the general expressions we have previously derived. For example using the time domain expression (11.33) we obtain:

$$\Lambda_{NCWF} \quad = \quad K_{XX}(0) - \int_{-\infty}^{\infty} h_{nc}(u)\, K_{YX}(u)\, du = Q - \int_{-\infty}^{\infty} \underbrace{\frac{Q\alpha}{R\beta}e^{-\beta|t|}}_{h_{nc}(u)}\, \underbrace{Qe^{-\alpha|t|}}_{K_{YX}(u)}\, du \tag{11.62}$$

$$= \quad Q - 2Q\frac{Q\alpha}{R\beta}\int_{0}^{\infty} e^{-(\alpha+\beta)t}\, dt \tag{11.63}$$

$$= \quad Q - \frac{2Q^2\alpha}{R\beta(\alpha+\beta)} \tag{11.64}$$

where we have used the fact that $K_{XX}(0) = Q$ and $K_{YX}(u) = K_{XX}(u)$. We could also use the frequency domain expression (11.43) as follows:

$$S_{EE}(s) \quad = \quad S_{XX}(s) - \frac{S_{YX}(s)S_{YX}(-s)}{S_{YY}(s)} = S_{XX}(s) - \frac{S_{XX}^2(s)}{S_{XX}(s) + S_{VV}(s)} = H_{nc}(s)S_{VV}(s) \tag{11.65}$$

$$= \quad \frac{2Q\alpha}{\beta^2 - s^2} \tag{11.66}$$

Now taking the inverse transform we obtain:

$$R_{EE}(t) \quad = \quad \frac{Q\alpha}{\beta}e^{-\beta|t|} \tag{11.67}$$

Thus:

$$\Lambda_{NCWF} \quad = \quad R_{EE}(0) = \frac{Q\alpha}{\beta} \tag{11.68}$$

This expression looks different than (11.64), but recall that $\beta^2 = \frac{2Q\alpha}{R} + \alpha^2$. Using this definition we can show that the these two solutions are actually the same by showing their difference is zero:

$$Q - \frac{2Q^2\alpha}{R\beta(\alpha + \beta)} - \frac{Q\alpha}{\beta} = \frac{Q\left(R\beta(\alpha + \beta)\right) - 2Q^2\alpha}{R\beta(\alpha + \beta)} - \frac{Q\alpha\left(R(\alpha + \beta)\right)}{R\beta(\alpha + \beta)} \tag{11.69}$$

$$= \frac{QR\alpha\beta + QR\beta^2 - 2Q^2\alpha - QR\alpha^2 - QR\alpha\beta}{R\beta(\alpha + \beta)} \tag{11.70}$$

$$= \frac{Q\left(R\beta^2 - 2Q\alpha - R\alpha^2\right)}{R\beta(\alpha + \beta)} \tag{11.71}$$

$$= \frac{Q\left(2Q\alpha + R\alpha^2 - 2Q\alpha - R\alpha^2\right)}{R\beta(\alpha + \beta)} = 0 \tag{11.72}$$

where in the last equality we substituted in the definition for β^2.

11.4.2 Causal Wiener Filtering

The other case of Wiener filtering we will examine is when we choose the end of the observation interval to coincide with the time of the estimate $T_f = t$ so that we produce the estimate based only on *past observations* $y(\tau)$, $\tau \in [-\infty, t]$. This case is referred to as *causal Wiener filtering*.

Definition 11.3 (Causal Wiener Filtering)
- Find LLSE estimate of $x(t)$ based on $y(\tau)$, $\tau \in [T_i, T_f]$
- $x(t)$, $y(t)$ are jointly wide-sense stationary
- $T_i = -\infty$
- $T_f = t$

Since we are basing the estimate on only past values of $y(\tau)$ the filter in this case will be *causal* and thus realizable in real time – a property of considerable practical interest if we are to implement the filter. For this reason the corresponding filter is sometimes referred to as the *realizable Wiener Filter*. In particular, the optimal filter $h_c(t)$ is linear, time-invariant and has the property that $h(t) = 0$ for $t < 0$. The optimal causal estimate will therefore be of the form:

$$\text{CT:} \quad \widehat{x}(t) = \int_{-\infty}^{t} h_c(t - \tau)y(\tau)\,d\tau \tag{11.73}$$

$$\text{DT:} \quad \widehat{x}(t) = \sum_{\tau=-\infty}^{t} h_c(t - \tau)\,y(\tau) \tag{11.74}$$

Applying the general Wiener-Hopf equation (11.25) (or (11.26) in discrete time) to this case we find that the Wiener-Hopf equations for the impulse response of the optimal filter become:

$$\text{CT:} \quad K_{YX}(t - \tau) = \int_{0}^{\infty} h_c(t - \sigma)\,K_{YY}(\sigma - \tau)\,d\sigma \quad -\infty \leq \tau \leq t \tag{11.75}$$

$$\text{DT:} \quad K_{YX}(t - \tau) = \sum_{0}^{\infty} h_c(t - \sigma)\,K_{YY}(\sigma - \tau) \quad -\infty \leq \tau \leq t \tag{11.76}$$

Now by making the change of variables $t - \tau = u$ and $t - \sigma = v$ we obtain:

$$\text{CT:} \quad K_{YX}(u) = \int_{0}^{\infty} h_c(v)\,K_{YY}(u - v)\,dv \quad 0 \leq u \leq \infty \tag{11.77}$$

$$\text{DT:} \quad K_{YX}(u) = \sum_{v=0}^{\infty} h_c(v)\,K_{YY}(u - v) \quad 0 \leq u \leq \infty \tag{11.78}$$

From (11.77) or (11.78) we can see that the Wiener-Hopf equation is only enforced for $u \geq 0$ and that it only constrains $h_c(t)$ for $t \geq 0$. We have the additional constraint that $h_c(t) = 0$ for $t < 0$.

These expressions seem quite similar to those obtained for the case of the noncausal Wiener filter (11.27) or (11.28) and thus we might think of using the same solution methods we used there (i.e. transform techniques). This is not the case, however, as the causality constraint makes things considerably more difficult. While the expressions on the right in (11.77) or (11.78) are still convolution integrals (or sums), $K_{YX}(u)$ need not equal this integral (or sum) for $u < 0$. Thus we cannot simply use bilateral Laplace transforms. We might think of using Unilateral Laplace Transforms, since interval in question is unilateral. But $K_{YY}(\tau)$ in the integral (sum) is an auto-covariance, so it possesses even symmetry and thus cannot be represented using a unilateral Laplace transform. So transform techniques directly applied to (11.77) or (11.78) will not work and we are forced to seek other approaches. In the development to follow we focus on the continuous-time case. As usual, similar arguments may be made for the discrete-time case with the substitution of z-transform for Laplace transform and inside (outside) unit circle for left (right) half complex plane.

Examination of (11.77) or (11.78) shows that the Wiener-Hopf equation for this case is easy to solve if the data $y(\tau)$ are *white* so $K_{YY}(t) = \delta(t)$. We will see a similar situation in our treatment of the Kalman filter and sequential estimation. Our approach there will be to first whiten the data by finding the innovations $\nu(t)$ and then to generate an estimate based on the resulting whitened observations. This is the approach Bode and Shannon took to solving (11.77) and it is the approach we will take. The basic plan of attack is shown in Figure 11.5 From experience we know (or might suspect) that the whitening filter will involve
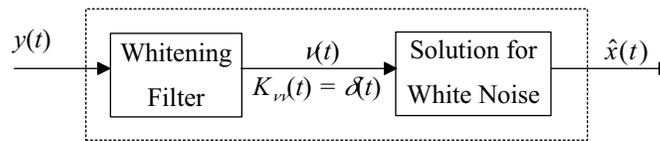
Figure 11.5: Bode-Shannon Whitening Approach to Causal Wiener Filtering.

spectral factorization. For the overall filter to be causal each block must be causal. In particular, the choice we make in the spectral factorization step must yield a corresponding filter that is causal. Further, for there to be no loss of information the whitening filter must also be invertible and thus stable with a stable inverse. With these points in mind we will first find the causal Wiener filter for white noise (the second block in Figure 11.5) then we will find the appropriate whitening filter.

Causal Wiener Filter for White Noise
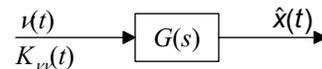
Figure 11.6: Wiener Filter for White Noise Observations.

Here we will consider the problem of designing the optimal causal Wiener filter for $x(t)$ based on observation of a zero-mean wide-sense-stationary unit power white noise, as depicted in Figure 11.6. To avoid confusion we will denote this white observation process as $\nu(t)$ and will denote the corresponding Wiener filter system function as $G(s)$ and its impulse response as $g(t)$. Since the observation process $\nu(t)$ is white we have:

$$K_{\nu\nu}(t) \quad = \quad \delta(t) \tag{11.79}$$

Substituting this expression into the Wiener-Hopf equation (11.77) (or (11.78) in discrete time) yields the following expression for the optimal filter impulse response $g(t)$:

$$K_{\nu X}(u) \quad = \quad \int_0^\infty g(v)\, K_{\nu\nu}(u-v)\, dv = \int_0^\infty g(v)\, \delta(u-v)\, dv = g(u) \quad 0 \le u \le \infty \tag{11.80}$$

The optimal causal Wiener filter for white noise is thus given by:

$$g(t) \quad = \quad \begin{cases} K_{\nu X}(t) & 0 \leq u \leq \infty \\ 0 & t < 0 \end{cases} \tag{11.81}$$

$$= \quad K_{\nu X}(t) u_{-1}(t) \tag{11.82}$$

where $u_{-1}(t)$ is the unit step. The causal Wiener filter for white noise observations is given by the causal (i.e. nonnegative time) part of the cross-covariance between $x(t)$ and the observations.

For its use later, let us introduce the following notation for the positive and negative time portions of signals:

$$\{K_{\nu X}(t)\}_{+} \quad \equiv \quad K_{\nu X}(t) u_{-1}(t) \tag{11.83}$$

$$\{K_{\nu X}(t)\}_{-} \quad \equiv \quad K_{\nu X}(t) u_{-1}(-t) \tag{11.84}$$

Thus we can always decompose $K_{\nu X}(t)$ into its positive time and negative time components as:

$$K_{\nu X}(t) \quad = \quad \{K_{\nu X}(t)\}_{+} + \{K_{\nu X}(t)\}_{-} \tag{11.85}$$

A similar decomposition can obviously be applied to any time function. Let us also introduce a similar notation for the bilateral Laplace transforms of these positive and negative time components:

$$\{S_{\nu X}(s)\}_{+} \equiv \int_{0^-}^{\infty} K_{\nu X}(t) e^{-st}\, dt \quad \longleftrightarrow \quad K_{\nu X}(t) u_{-1}(t) \equiv \{K_{\nu X}(t)\}_{+} \tag{11.86}$$

$$\{S_{\nu X}(s)\}_{-} \equiv \int_{-\infty}^{0^-} K_{\nu X}(t) e^{-st}\, dt \quad \longleftrightarrow \quad K_{\nu X}(t) u_{-1}(-t) \equiv \{K_{\nu X}(t)\}_{-} \tag{11.87}$$

$$S_{\nu X}(s) = \int_{-\infty}^{\infty} K_{\nu X}(t) e^{-st}\, dt \quad = \quad \{S_{\nu X}(s)\}_{+} + \{S_{\nu X}(s)\}_{-} \tag{11.88}$$

For example, $\{S_{\nu X}(s)\}_{+}$ is the bilateral Laplace transform of the positive time portion of $K_{\nu X}(t)$. Note that terms at $t = 0$ (e.g. impulses at $t = 0$) are included in the definition of $\{K_{\nu X}(t)\}_{+}$. Based on this notation the optimal causal Wiener filter for white noise observations is given by:

$$\{K_{\nu X}(t)\}_{+} = g(t) \quad \longleftrightarrow \quad G(s) = \{S_{\nu X}(s)\}_{+} \tag{11.89}$$

The relationship between these quantities is shown in Figure 11.7. If we possess an expression for $K_{\nu X}(t)$ then finding its positive time portion is straightforward, and the filter system function $G(s)$ may then be obtained as the Laplace transform of this quantity, as indicated in the figure. A natural question is whether we may *directly* find $G(s)$ from knowledge of cross-power spectral density $S_{\nu X}(s)$. When $S_{\nu X}(s)$ possess a *rational* spectrum then we can indeed find $G(s)$ directly in the transform domain through use of a partial fraction expansion.

$$K_{\nu X}(t) \quad \longrightarrow \quad g(t) = \{K_{\nu X}(t)\}_{+}$$

$$\text{BLT} \Big\updownarrow \qquad\qquad \text{BLT} \Big\updownarrow$$

$$S_{\nu X}(s) \quad \longrightarrow \quad G(s) = \{S_{\nu X}(s)\}_{+}$$

Figure 11.7: Relationship between time domain and Laplace domain quantities for the Causal Wiener Filter for White Noise Observations.

To understand the relationship between the bilateral Laplace transform of a rational function and the bilateral Laplace transform of its positive time part, consider a general rational function with Laplace transform $F(s)$. Being rational, this function possess a partial fraction expansion as:

$$F(s) \quad = \quad \sum_{i=1}^{m} \sum_{k=1}^{k_i} \frac{A_{ik}}{(s - s_i)^k} \tag{11.90}$$

This partial fraction expansion expression represents $F(s)$ as a sum of terms. The inverse Laplace transform of $F(s)$ will thus be composed of the sum of the inverse transforms of each term. Recall that the expression (11.90) alone does not uniquely specify the a time function – a corresponding *region of convergence* (ROC) must also always be specified. For stability of the associated time function, the ROC associated with each term must include the $j\omega$ axis, as shown in Figure 11.8. The ROC depends on the poles, not the zeros.
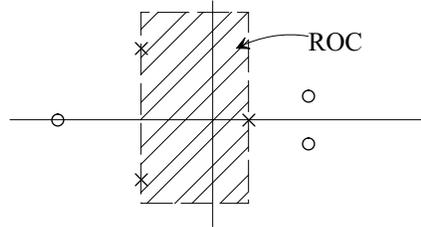
Figure 11.8: Pole-zero plot and associated regions of convergence.

Now recall that right going or positive time signals have right going ROCs and left going or negative time signals have left going ROCs. Thus the terms in the sum (11.90) corresponding to the positive-time portions of $f(t)$ are those corresponding to the left-half plane poles, while the terms (11.90) corresponding to negative-time parts of $f(t)$ are those corresponding to the right-half plane poles (note that the zeros play no direct role in this discussion). Thus we have:

$$\{F(s)\}_+ \;=\; \sum_{\substack{i=1 \\ \mathrm{Re}(s_i)<0}}^{m} \sum_{k=1}^{k_i} \frac{A_{ik}}{(s-s_i)^k} \tag{11.91}$$

$$\{F(s)\}_- \;=\; \sum_{\substack{i=1 \\ \mathrm{Re}(s_i)>0}}^{m} \sum_{k=1}^{k_i} \frac{A_{ik}}{(s-s_i)^k} \tag{11.92}$$

If $F(s)$ is not strictly proper, so the partial fraction expansion includes nonnegative powers of s, these are included in the definition of $\{F(s)\}_+$ since, by definition, we include terms at $t = 0$ in the definition of $\{f(t)\}_+$. In summary, we can directly find the bilateral Laplace transforms of the positive and negative time portions of the function $f(t)$ from knowledge of its rational transform $F(s)$ via partial fraction expansion. A similar argument can be made for the discrete time case with the inside and outside of the unit circle playing the role of the left and right half plane, respectively. We summarize these points for continuous signals and Laplace transforms below:

- A stable time function has an ROC that contains the $j\omega$ axis and vice versa.

- A right-sided time function possesses a right-sided ROC and vice versa.

- A two-sided time function has a bounded ROC (i.e. the ROC is a strip) and vice versa.

- The right-sided part of a two-sided signal corresponds to the left-half plane poles.

- To get the right-sided part of a two-sided signal, perform partial fraction expansion of the transform and keep the poles with the right-sided ROCs together with any positive powers of s (corresponding to singularities at the origin).

- The ROC, and thus stability and left/right sidedness of a signal depends on the *poles*, not the zeros of the signal.

For discrete-time signals, a similar set of conditions applies to the z-transform with the role of the $j\omega$ axis replaced by the unit circle, and left and right s-plane replaced by inside and outside unit circle. Thus for discrete-time signals and z-transforms we have:

- A stable discrete-time function has an ROC that contains the unit circle and vice versa.

- A right-sided discrete-time function possesses an outward going ROC and vice versa.

- A two-sided discrete-time function has a bounded ROC (i.e. the ROC is an annulus) and vice versa.

- The right-sided part of a two-sided discrete-time signal corresponds to the poles in the unit circle.

- To get the right-sided part of a two-sided discrete-time signal, perform partial fraction expansion of the transform and keep the poles with the outward-going ROCs.

- The ROC, and thus stability and left/right sidedness of a signal depends on the *poles*, not the zeros of the signal.

Some examples serve to illustrate these developments for the continuous-time case.

Example 11.2
Suppose $f(t)$ is given by:

$$f(t) = e^{-at}u_{-1}(t) \tag{11.93}$$

where $a > 0$. This is a right-going, stable signal. The bilateral Laplace transform is given by:

$$F(s) = \frac{1}{s+a} \tag{11.94}$$

with ROC $Re(s) > -a$. A sketch of $f(t)$, pole-zero plot, and the corresponding ROC is shown in Figure 11.9
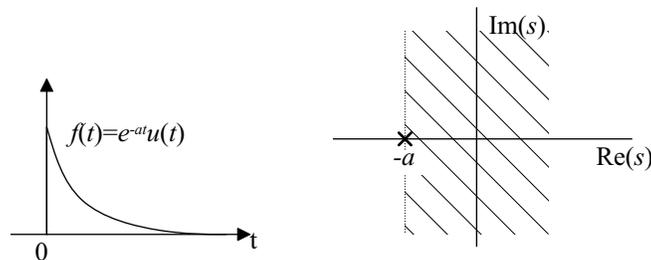
Figure 11.9: Function $f(t)$, the pole-zero plot, and the corresponding ROC.

Example 11.3
Suppose $f(t)$ is now given by:

$$f(t) = -e^{-at}u_{-1}(-t) \tag{11.95}$$

where $a > 0$, so it is a left-going, unstable signal. The corresponding bilateral Laplace transform is given by:

$$F(s) = \frac{1}{s+a} \tag{11.96}$$

with ROC $Re(s) < -a$. A sketch of $f(t)$, pole-zero plot, and the corresponding ROC is shown in Figure 11.10

Example 11.4
Suppose $f(t)$ is now given by:

$$f(t) = \underbrace{Ae^{-at}u_{-1}(t)}_{\{f(t)\}_+} - \underbrace{Be^{bt}u_{-1}(-t)}_{\{f(t)\}_-} \tag{11.97}$$

where $a > 0$, $b > 0$, so it is a two-sided, stable signal. The corresponding bilateral Laplace transform is given by:

$$F(s) = \underbrace{\frac{A}{s+a}}_{\{F(s)\}_+} + \underbrace{\frac{B}{s-b}}_{\{F(s)\}_-} = (A+B)\frac{s + (aB - Ab)/(A+B)}{(s+a)(s-b)} \tag{11.98}$$

with ROC $-a < Re(s) < a$. A sketch of $f(t)$, pole-zero plot, and the corresponding ROC is shown in Figure 11.11:
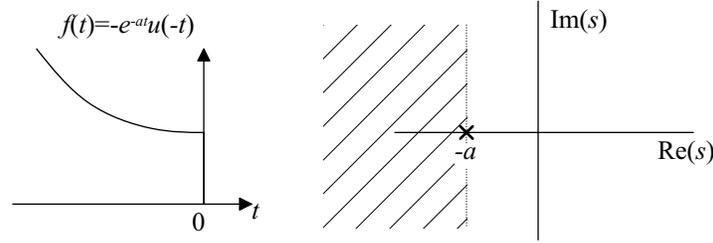
Figure 11.10: Function $f(t)$, the pole-zero plot, and the corresponding ROC.
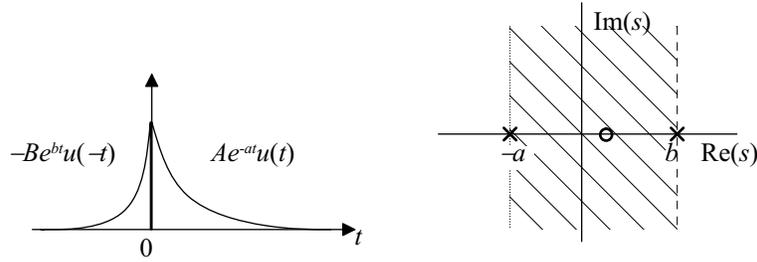
Figure 11.11: Function $f(t)$, the pole-zero plot, and the corresponding ROC.

Example 11.5

Suppose $f(t)$ is given by:

$$f(t) \quad = \quad e^{-a|t|} = \underbrace{e^{-at}u_{-1}(t)}_{\{f(t)\}_+} + \underbrace{e^{at}u_{-1}(-t)}_{\{f(t)\}_-} \tag{11.99}$$

where $a > 0$. Now, starting with the bilateral Laplace transform and performing a partial fraction expansion:

$$F(s) \quad = \quad \frac{2a}{a^2 - s^2} = \underbrace{\frac{1}{s+a}}_{\text{LHP Pole}} - \underbrace{\frac{1}{s-a}}_{\text{RHP Pole}} \tag{11.100}$$

with ROC $-a < Re(s) < b$. A sketch of $f(t)$, pole-zero plot, and the corresponding ROC is shown in Figure 11.12: Now
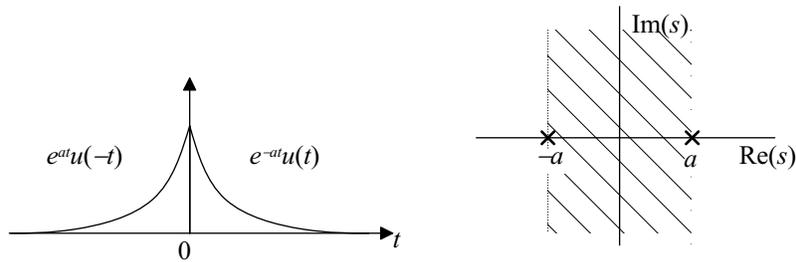
Figure 11.12: Function $f(t)$, the pole-zero plot, and the corresponding ROC.

we have that:

$$\{F(s)\}_+ = \frac{1}{s+a} \quad \longleftrightarrow \quad e^{-at}u_{-1}(t) = \{f(t)\}_+ \tag{11.101}$$

Example 11.6

Now we consider a case with terms at the origin. Suppose $f(t)$ is given by:

$$f(t) \quad = \quad \delta + e^{-a|t|} = \underbrace{\delta(t) + e^{-at}u_{-1}(t)}_{\{f(t)\}_+} + \underbrace{e^{at}u_{-1}(-t)}_{\{f(t)\}_-} \tag{11.102}$$

where again $a > 0$. Starting with the bilateral Laplace transform and performing a partial fraction expansion:

$$F(s) \;=\; 1 + \frac{2a}{a^2 - s^2} = \underbrace{1 + \frac{1}{s+a}}_{\substack{\text{LHP Pole plus Positive}\\\text{Powers of }s}} - \underbrace{\frac{1}{s-a}}_{\text{RHP Pole}} \tag{11.103}$$

$$\{F(s)\}_+ = 1 + \frac{1}{s+a} \quad\longleftrightarrow\quad \delta(t) + e^{-at}u_{-1}(t) = \{f(t)\}_+ \tag{11.104}$$

Example 11.7
Next we consider a non-rational example. Suppose $f(t)$ is given by:

$$f(t) = e^{-a(t+T)}u_{-1}(t+T) \quad\longleftrightarrow\quad F(s) = \frac{e^{sT}}{s+a} \tag{11.105}$$

where again $a > 0$. Unfortunately, $F(s)$ is *not* a rational function of s, so partial fraction techniques will not help. Let us examine this example in more detail. There are two cases to consider, depending on the sign of T.
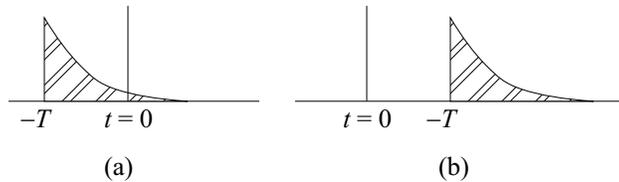
Figure 11.13: Plot of $f(t)$ for $T > 0$ and $T < 0$.

(a) Suppose $T > 0$. This case is shown in Figure 11.13(a). In this case:

$$\{f(t)\}_+ = e^{-a(t+T)}u_{-1}(t+T)u_{-1}(t) = e^{-a(t+T)}u_{-1}(t) = e^{-aT}e^{-at}u_{-1}(t) \quad\longleftrightarrow\quad \frac{e^{-aT}}{s+a} = \{F(s)\}_+ \tag{11.106}$$

(b) Now suppose $T < 0$. This case is shown in Figure 11.13(b). In this case:

$$\{f(t)\}_+ = e^{-a(t+T)}u_{-1}(t+T)u_{-1}(t) = e^{-a(t+T)}u_{-1}(t+T) = f(t) \quad\longleftrightarrow\quad F(s) = \frac{e^{sT}}{s+a} = \{F(s)\}_+ \tag{11.107}$$

Thus we see that there are no simple formulas relating $F(s)$ and $\{F(s)\}_+$ in the non-rational case and we are forced to apply the definitions and work in the time domain.

Causal Whitening Filter

In the previous subsection we found the optimal causal Wiener filter for the case of white noise observations. In this white noise case the causal Wiener filter has a particularly simple form. Unfortunately, in practice our observations are usually *not* white and we must therefore whiten them. This is precisely the purpose of a *whitening filter*, as illustrated in Figure 11.14. We have already seen a closely related process of spectral shaping, wherein white noise is passed through a linear time-invariant system to achieve a desired spectral shape. Here we desire the reverse of this procedure.
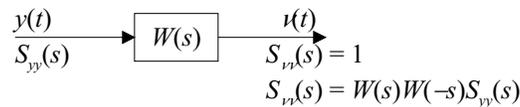
Figure 11.14: Whitening Filter $W(s)$.

In particular, we desire a linear time-invariant filter with system function $W(s)$ that will take a zero-mean wide-sense stationary random process $y(t)$ with spectrum $S_{YY}(s)$ and turn it into a zero-mean wide-sense stationary unit spectrum white noise process $\nu(t)$, termed the *innovations*, because it contains the

unpredictable information in the data. Our idea is that if we can find such a filter, we can use it to whiten our observations, then pass the resulting whitened observations through $G(s)$, the causal Wiener filter for white noise we found in the preceding subsection. To be useful for our estimation problem, however we require a number of additional constraints on $W(s)$ over and above its ability to generate uncorrelated outputs. First, if the processes are to be wide-sense stationary $W(s)$ must be *stable*. Next, if the overall cascade of $W(s)$ and $G(s)$ is to be causal $W(s)$ itself must be *causal*. Further, if the estimate based on the whitened observations $\nu(t)$ is to be the same as the estimate based on the original observations, then the whitening transformation must be invertible. This requires that $W(s)$ be *causally invertible* as well.

From Figure 11.14 it is clear that the filter we seek must satisfy:

$$W(s)W(-s)S_{YY}(s) = 1 \tag{11.108}$$

where $W(s)$ is stable, causal, and causally invertible. For general processes and spectral density functions $S_{YY}(s)$ finding such a filter can be difficult. If $S_{YY}(s)$ is a rational power spectral density however then a straightforward approach exists based on *spectral factorization* of the spectrum of the observation process $S_{YY}(s)$ Let us thus focus on this case of rational $S_{YY}(s)$. Since $S_{YY}(s)$ is the spectral density of a real valued random process it possesses certain symmetry properties. In particular, $S_{YY}(j\omega)$ must be a finite, real-valued, even function of ω. These symmetry properties constrain the behavior of the poles and zeros of $S_{YY}(s)$. In particular, for rational spectral densities, they imply that $S_{YY}(s)$ is the ratio of two polynomials in s^2. Thus if $s = \sigma_i$ is a zero of $S_{YY}(s)$ then $s = -\sigma_i$ must also be a zero and if $s = p_i$ is a pole of $S_{YY}(s)$ then $s = -p_i$ must also be a pole. Further, since $S_{YY}(j\omega)$ is a real valued function, these poles and zeros must occur in complex conjugate pairs. Finally, $S_{YY}(s)$ can have no poles on the $j\omega$ axis and any zero must occur with even multiplicity. These symmetry properties are illustrated in Figure 11.15.
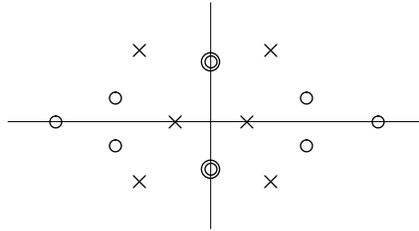
Figure 11.15: Illustration of pole-zero symmetry properties of $S_{YY}(s)$.

Given these symmetry properties we can always write any rational system function $S_{YY}(s)$ in the following form:

$$S_{YY}(s) = \frac{M\prod_i(s-\sigma_i)(-s-\sigma_i)}{\prod_i(s-p_i)(-s-p_i)} = \left(\frac{\sqrt{M}\prod_i(s-\sigma_i)}{\prod_i(s-p_i)}\right)\left(\frac{\sqrt{M}\prod_i(-s-\sigma_i)}{\prod_i(-s-p_i)}\right) \tag{11.109}$$

where p_i are the left-half plane poles and σ_i are the left-half plane zeros, and M is a positive constant. The first factor in (11.109) is composed of just the left-half poles and zeros while the second factor in (11.109) is composed of only the right-half poles and zeros. These two factors are the key to our solution so we define some special notation for them:

$$S_{YY}^+(s) = \left(\frac{\sqrt{M}\prod_i(s-\sigma_i)}{\prod_i(s-p_i)}\right) = \text{Left-half poles and zeros of } S_{YY}(s) \tag{11.110}$$

$$S_{YY}^-(s) = \left(\frac{\sqrt{M}\prod_i(-s-\sigma_i)}{\prod_i(-s-p_i)}\right) = S_{YY}^+(-s) = \text{Right-half poles and zeros of } S_{YY}(s) \tag{11.111}$$

$$S_{YY}(s) = S_{YY}^+(s)S_{YY}^-(s) \tag{11.112}$$

Notice that the $S_{YY}^+(s)$ term is causal and causally invertible, since *both* its poles and zeros are in the left half plane by construction. This decomposition of a rational spectrum, in this case $S_{YY}(s)$, into a causal and causally invertible factor $S_{YY}^+(s)$ and its mirror image $S_{YY}^-(s)$, related as above, is termed spectral factorization.

From the above discussion it should be obvious that this process of spectral factorization will provide us our desired whitening filter $W(s)$. In particular, from (11.108) and (11.110)–(11.112) together with the properties of the factors involved we can see that the desired causal and causally invertible whitening filter can be obtained as:

$$W(s) = \frac{1}{S_{YY}^+(s)} \tag{11.113}$$

The Overall Causal Wiener Filter

We are now ready to assemble the pieces of the general causal Wiener filter solution. To summarize we have found a causal and causally invertible whitening filter, thus LLSE based on $y(\tau)$ or the whitened output $\nu(\tau)$ is equivalent. In addition, we have found the optimal causal Wiener filter for white noise. The overall filter is found by combining these two pieces as shown in Figure 11.16
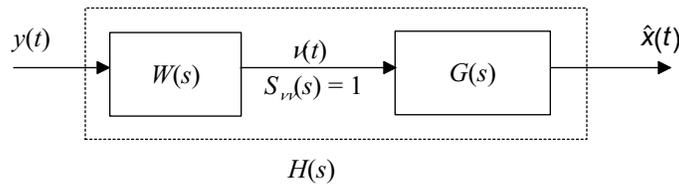
Figure 11.16: Overall causal Wiener Filter.

Now recall that the causal Wiener filter for white noise $G(s)$ was given by:

$$G(s) = \{S_{\nu X}(s)\}_+ \tag{11.114}$$

The innovation process $\nu(t)$ is the output of a linear time-invariant system with system function $W(s)$ and input $y(t)$, thus using our relationships for random processes through linear systems we find:

$$S_{\nu X}(s) = W(-s)S_{YX}(s) = \frac{S_{YX}(s)}{S_{YY}^+(-s)} = \frac{S_{YX}(s)}{S_{YY}^-(s)} \tag{11.115}$$

where in the last equality we have used the fact that $S_{YY}^+(-s) = S_{YY}^-(s)$. Now we can combine (11.115) and (11.114) to get the following expression for $G(s)$:

$$G(s) = \left\{ \frac{S_{YX}(s)}{S_{YY}^-(s)} \right\}_+ \tag{11.116}$$

Finally we can combine the whitening filter $W(s)$ and the causal Wiener filter for the white noise case $G(s)$ to obtain the overall causal Wiener filter $H_c(s)$:

$$H_c(s) = W(s)\,G(s) = \underbrace{\frac{1}{S_{YY}^+(s)}}_{\substack{\text{Whitening} \\ \text{Filter}}} \underbrace{\left\{ \frac{S_{YX}(s)}{S_{YY}^-(s)} \right\}_+}_{\substack{\text{CWF for In-} \\ \text{novations}}} \tag{11.117}$$

We summarize this solution in Figure 11.17. Compare this result to that obtained for the noncausal Wiener filter (11.31). Before moving on a word of caution is in order regarding the notation used in (11.117) – the terms $S_{YY}^+(s)$ and $S_{YY}^-(s)$ are the *spectral factors* of $S_{YY}(s)$ and $\{S_{YX}(s)/S_{YY}^-(s)\}_+$ is the *positive time part*
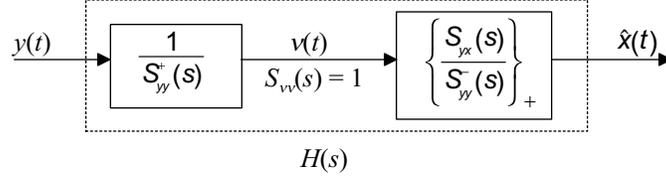
Figure 11.17: Summary of Causal Wiener Filter.

of $S_{YX}(s)/S_{YY}^-(s)$. Be careful not to confuse these! Terms like $S_{YY}^+(s)$ are found by spectral factorization while terms like $\left\{S_{YX}(s)/S_{YY}^-(s)\right\}_+$ are found via partial fraction expansion.

Finally we need to specify the corresponding estimation error variance or mean square error. We have three methods we can use. First we use a direct application of the general formula (11.18) to the case at hand to obtain:

$$\Lambda_{CWF} = K_{XX}(0) - \int_0^\infty h_c(\tau)\, K_{YX}(\tau)\, d\tau \tag{11.118}$$

where $h_c(\tau)$ is the impulse response corresponding to the overall causal Wiener filter $H_c(s)$ given in (11.117).

The second expression for the mean square error is based on the following reasoning. As we have argued, the optimal causal LLSE estimate $\hat{x}(t)$ based on either the original observations $y(\tau)$ or the innovations $\nu(\tau)$ is equivalent. Thus the error covariance based on either observations is equivalent. Thus we can as well apply the formula (11.118) to the innovation-based filter $g(\tau)$ applied to the innovation $\nu(\tau)$ to obtain:

$$\Lambda_{CWF} = K_{XX}(0) - \int_0^\infty g(\tau)\, K_{\nu X}(\tau)\, d\tau \tag{11.119}$$

where $g(t)$ is the impulse response of the causal Wiener filter for the innovations. Now from (11.89) and (11.116), $g(t)$ is given by:

$$g(t) = K_{\nu X}(t) u_{-1}(t) \quad \longleftrightarrow \quad G(s) = \{S_{\nu X}(s)\}_+ = \left\{\frac{S_{YX}(s)}{S_{YY}^-(s)}\right\}_+ \tag{11.120}$$

Using these relationships we get the equivalent expression for the estimation error variance:

$$\Lambda_{CWF} = K_{XX}(0) - \int_0^\infty K_{\nu X}^2(\tau)\, d\tau = K_{XX}(0) - \int_0^\infty g^2(\tau)\, d\tau \tag{11.121}$$

where the impulse $g(t)$ can be obtained as the inverse bilateral Laplace transform of $G(s)$ as specified in (11.120).

Finally, we can obtain a frequency domain expression for the error following the line of argument associated with equations (11.35)-(11.37). In particular we have:

$$S_{EE}(s) = S_{XX}(s) - S_{\hat{x}\hat{x}}(s) \tag{11.122}$$

But, for the causal Wiener filter we obtain for the second term:

$$S_{\hat{x}\hat{x}}(s) = H_c(s)H_c(-s)S_{YY}(s) = W(s)G(s)W(-s)G(-s)S_{YY}(s) \tag{11.123}$$

$$= \frac{S_{YY}(s)}{S_{YY}^+(s)S_{YY}^+(-s)}G(s)G(-s) = \frac{1}{S_{YY}^+(s)}\frac{1}{S_{YY}^-(s)}S_{YY}(s)G(s)G(-s) \tag{11.124}$$

$$= G(s)G(-s) \tag{11.125}$$

Thus, for the causal Wiener filter we get the following expression for the power spectral density of the error:

$$S_{EE}(s) = S_{XX}(s) - G(s)G(-s) \tag{11.126}$$

If we let $s = j\omega$ we obtain:

$$S_{EE}(j\omega) = S_{XX}(j\omega) - |G(j\omega)|^2 \tag{11.127}$$

This expression, though derived for the continuous time case, is also valid for the discrete time case, with appropriate adjustments to the transform definitions. Finally, we have that:

$$\Lambda_{CWF} = R_{EE}(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_{EE}(j\omega)\, d\omega \tag{11.128}$$

The mean square error can thus be obtained either by finding $R_{EE}(\tau)$ as the inverse transform of $S_{EE}(j\omega)$ and then evaluating the result at $\tau = 0$ or by directly evaluating the integral in (11.128).

Now let us consider some examples.

Example 11.8

Suppose the underlying process $x(t)$ is zero mean, wide-sense stationary with $K_{XX}(t) = Qe^{-\alpha|t|}$ and suppose we observe:

$$y(t) = x(t) + v(t) \tag{11.129}$$

where $v(t)$ is a zero mean, wide-sense stationary white noise process, uncorrelated with $x(t)$ and with covariance function $K_{VV}(t) = R\delta(t)$. We wish to find the causal Wiener filter for this problem. We first find $S_{YY}(s)$ as a function of $S_{XX}(s)$ and $S_{VV}(s)$:

$$S_{XX}(s) = \frac{2\alpha Q}{\alpha^2 - s^2} \tag{11.130}$$

$$S_{VV}(s) = R \tag{11.131}$$

$$\implies S_{YY}(s) = S_{XX}(s) + S_{VV}(s) \tag{11.132}$$

$$= \frac{2\alpha Q}{\alpha^2 - s^2} + R = R\left(\frac{\beta^2 - s^2}{\alpha^2 - s^2}\right) \qquad \beta = \left(\frac{2Q\alpha}{R} + \alpha^2\right)^{1/2} \tag{11.133}$$

$$= \underbrace{\left\{R^{1/2}\frac{s+\beta}{s+\alpha}\right\}}_{S_{YY}^+(s)} \underbrace{\left\{R^{1/2}\frac{\beta-s}{\alpha-s}\right\}}_{S_{YY}^-(s)} \tag{11.134}$$

Thus we find for the whitening filter $W(s)$:

$$W(s) = \frac{1}{S_{YY}^+(s)} = \left\{R^{-1/2}\frac{s+\alpha}{s+\beta}\right\} \tag{11.135}$$

Now we need to find $G(s)$, the causal Wiener filter for the innovations. To do this we will need $S_{YX}(s)$:

$$S_{YX}(s) = S_{XX}(s) = \frac{2Q\alpha}{\alpha^2 - s^2} \tag{11.136}$$

where we have used the fact that $x(t)$ and $v(t)$ are uncorrelated so $K_{YX}(t) = K_{XX}(t)$. Now we can find $G(s)$:

$$G(s) = \left\{\frac{S_{YX}(s)}{S_{YY}^-(s)}\right\}_+ \tag{11.137}$$

$$= \left\{\left(\frac{2Q\alpha}{\alpha^2 - s^2}\right)\left(R^{-1/2}\frac{\alpha-s}{\beta-s}\right)\right\}_+ \tag{11.138}$$

$$= \left\{\frac{a}{s+\alpha} + \frac{b}{s-\beta}\right\}_+ \qquad a = \frac{2Q\alpha R^{-1/2}}{\alpha+\beta}, \qquad b = -\frac{2Q\alpha R^{-1/2}}{\alpha+\beta} \tag{11.139}$$

$$= \frac{a}{s+\alpha} \tag{11.140}$$

where we have identified the first term as the positive time part of the signal. Overall we obtain for $H_c(s)$:

$$H_c(s) = W(s)G(s) \tag{11.141}$$

$$= \frac{2Q\alpha}{R(\alpha+\beta)}\left(\frac{1}{s+\beta}\right) \tag{11.142}$$

Taking inverse transforms we obtain for the corresponding causal Wiener filter impulse response:

$$h_c(t) \quad = \quad \frac{2Q\alpha}{R(\alpha + \beta)} e^{-\beta t} u_{-1}(t) \tag{11.143}$$

Finally we find the associated estimation error covariance:

$$K_{XX}(\tau) \quad = \quad Qe^{-\alpha|t|} \longrightarrow K_{XX}(0) = Q \tag{11.144}$$

$$K_{YX}(\tau) \quad = \quad K_{XX}(\tau) = Qe^{\alpha|\tau|} \tag{11.145}$$

$$\Longrightarrow \quad \Lambda_{CWF} \quad = \quad K_{XX}(0) - \int_0^\infty h_c(\tau)\, K_{YX}(\tau)\, d\tau \tag{11.146}$$

$$= \quad Q - \int_0^\infty \frac{2Q\alpha}{R(\alpha + \beta)} e^{-\beta\tau} Q e^{-\alpha|\tau|}\, d\tau \tag{11.147}$$

$$= \quad Q - \frac{2Q^2\alpha}{R(\alpha + \beta)^2} \left. \left(-e^{-(\alpha+\beta)t} \right) \right|_0^\infty \tag{11.148}$$

$$= \quad Q - \frac{2Q^2\alpha}{R(\alpha + \beta)^2} \tag{11.149}$$

We can compare this mean square error to that obtained when the noncausal Wiener filter is used:

$$\Lambda_{NCWF} \quad = \quad Q - \frac{2Q^2\alpha}{R\beta(\alpha + \beta)} \tag{11.150}$$

Note that the noncausal Wiener filter achieves a lower error variance compared to the causal Wiener filter. Indeed this must be the case, since the noncausal Wiener filter uses *all* the data and the causal Wiener filter only part of the data to generate its estimate. In general, this observation is true.

Example 11.9
Now let us consider a discrete-time example. Suppose the underlying process $x(t)$ is a zero mean, wide-sense stationary first order autoregressive process:

$$x(t + 1) = 0.8x(t) + w(t), \qquad K_{WW}(t) = 0.36\delta(t) \tag{11.151}$$

Using our results for first-order autoregressive processes we find that this process has the following autocovariance function:

$$K_{XX}(t) = 0.8^{|t|} \tag{11.152}$$

Suppose we observe:

$$y(t) \quad = \quad x(t) + v(t), \qquad K_{VV}(t) = \delta(t) \tag{11.153}$$

where $v(t)$ is a zero mean, wide-sense stationary white noise process, uncorrelated with $x(t)$ and with the given covariance function. We wish to find the causal Wiener filter for this problem. First lets find the whitening filter $W(z) = 1/S_{YY}^+(z)$. We will need to find $S_{YY}(z)$ as a function of $S_{XX}(z)$ and $S_{VV}(z)$. Using the independence of $x(t)$ and $v(t)$:

$$S_{YY}(z) \quad = \quad S_{XX}(z) + S_{VV}(z) = \frac{0.36}{(1 - 0.8z)(1 - 0.8z^{-1})} + 1 \tag{11.154}$$

$$= \quad \underbrace{\left\{ \frac{\sqrt{8/5}\left(1 - \frac{1}{2}z\right)}{(1 - 0.8z)} \right\}}_{S_{YY}^-(z)} \underbrace{\left\{ \frac{\sqrt{8/5}\left(1 - \frac{1}{2}z^{-1}\right)}{(1 - 0.8z^{-1})} \right\}}_{S_{YY}^+(z)} \tag{11.155}$$

Thus we find for the whitening filter $W(z)$:

$$W(z) \quad = \quad \frac{1}{S_{YY}^+(z)} = \sqrt{\frac{5}{8}} \left(\frac{1 - 0.8z^{-1}}{1 - \frac{1}{2}z^{-1}} \right) \tag{11.156}$$

Now we need to find $G(z)$, the causal Wiener filter for the innovations. To do this we will need $S_{YX}(z)$:

$$S_{YX}(z) \quad = \quad S_{XX}(z) = \frac{0.36}{(1 - 0.8z)(1 - 0.8z^{-1})} \tag{11.157}$$

where we have used the fact that $x(t)$ and $v(t)$ are uncorrelated so $K_{YX}(t) = K_{XX}(t)$. Now we can find $G(z)$:

$$G(z) = \left\{ \frac{S_{YX}(s)}{S_{YY}^-(s)} \right\}_+ \tag{11.158}$$

$$= \left\{ \left(\frac{0.36}{(1-0.8z)(1-0.8z^{-1})} \right) \left(\sqrt{\frac{5}{8}} \frac{(1-0.8z)}{(1-\frac{1}{2}z)} \right) \right\}_+ = \left\{ \frac{0.36\sqrt{5/8}}{(1-\frac{1}{2}z)(1-0.8z^{-1})} \right\}_+ \tag{11.159}$$

$$= \left\{ \frac{-0.72\sqrt{5/8}z^{-1}}{(1-2z^{-1})(1-0.8z^{-1})} \right\}_+ = \left\{ \frac{-18/25\sqrt{5/8}z^{-1}}{(1-2z^{-1})(1-0.8z^{-1})} \right\}_+ \tag{11.160}$$

$$= \left\{ \frac{A}{1-2z^{-1}} + \frac{B}{1-0.8z^{-1}} \right\}_+ \tag{11.161}$$

$$= \frac{B}{1-0.8z^{-1}}, \tag{11.162}$$

$$B = \left. \frac{-0.72\sqrt{5/8}z^{-1}}{1-2z^{-1}} \right|_{z=0.8} = 0.6\sqrt{5/8}, \quad A = \left. \frac{-0.72\sqrt{5/8}z^{-1}}{1-0.8z^{-1}} \right|_{z=2} = -0.6\sqrt{5/8} \tag{11.163}$$

where we have identified the second term in the partial fraction expansion of (11.161) as the positive time part of the signal. Note that while we have provided the value of A in the partial fraction expansion, it is not needed. We obtain for $G(z)$ and $g(t)$:

$$G(z) = \frac{0.6\sqrt{5/8}}{1-0.8z^{-1}} \quad \Longleftrightarrow \quad g(t) = 0.6\sqrt{5/8}(0.8)^t u_{-1}(t) \tag{11.164}$$

Overall we obtain for $H_c(z)$:

$$H_c(z) = G(z)W(z) \tag{11.165}$$

$$= \left(\frac{0.6\sqrt{5/8}}{1-0.8z^{-1}} \right) \left(\frac{\sqrt{5/8}(1-0.8z^{-1})}{1-\frac{1}{2}z^{-1}} \right) \tag{11.166}$$

$$= \left(\frac{0.375}{1-\frac{1}{2}z^{-1}} \right) \tag{11.167}$$

Taking inverse transforms we obtain for the corresponding causal Wiener filter impulse response:

$$h_c(t) = 0.375 \left(\frac{1}{2} \right)^t u_{-1}(t) \tag{11.168}$$

Finally we find the associated estimation error covariance or mean square error:

$$\Lambda_{CWF} = K_{XX}(0) - \sum_{k=0}^{\infty} g^2(k) \tag{11.169}$$

$$= 1 - \sum_{k=0}^{\infty} 0.36(5/8)(0.8)^{2k} \tag{11.170}$$

$$= 1 - 0.36(5/8) \frac{1}{1-(0.8)^2} = 1 - 5/8 \tag{11.171}$$

$$= 0.375 \tag{11.172}$$

We will see in Example 12.1 that both the causal Wiener filter and its mean square error are the same as that obtained for the Kalman filter *in steady state* – i.e. when we apply the Kalman filter to a stationary problem over an infinite time interval, so transients have had a chance to die out.

11.4.3 Summary

We close this discussion of Wiener filters for wide-sense stationary processes by presenting a table summarizing the causal and noncausal Wiener filters.

Wiener Filter Type	Observation Interval	Optimal Filter	Estimation Error Variance/MSE
Noncausal	$[-\infty,\ +\infty]$	$H_{nc}(s) = \dfrac{S_{YX}(s)}{S_{YY}(s)}$	$\Lambda_{NCWF} = K_{XX}(0) - \displaystyle\int_{-\infty}^{\infty} h_{nc}(u)\, K_{YX}(u)\, du$
Causal	$[-\infty,\ t]$	$H_c(s) = \dfrac{1}{S_{YY}^{+}(s)} \left\{ \dfrac{S_{YX}(s)}{S_{YY}^{-}(s)} \right\}_{+}$	$\Lambda_{CWF} = K_{XX}(0) - \displaystyle\int_{0}^{\infty} g^2(\tau)\, d\tau$

Figure 11.18: Summary of Wiener Filter Solutions

Because of its potentially confusing nature, we also give a summary of the notation used with respect to Wiener filter solutions. Suppose $F(s)$ is the bilateral Laplace transform corresponding to the time function $f(t)$, then the following apply:

Transform Domain		Time Domain

$$F(s) \qquad \longleftrightarrow \qquad f(t)$$

$F^{+}(s)$
Spectral Factorization w/ LHP poles and zeros

$F^{-}(s)$
Spectral Factorization w/ RHP poles and zeros

$$\{F(s)\}_{+} \equiv \int_{0^{-}}^{\infty} f(t)e^{-st}\, dt \qquad \longleftrightarrow \qquad f(t)u_{-1}(t) \equiv \{f(t)\}_{+}$$
PFE terms w/ LHP poles $\qquad\qquad\qquad$ Positive time part

$$\{F(s)\}_{-} \equiv \int_{-\infty}^{0^{-}} f(t)e^{-st}\, dt \qquad \longleftrightarrow \qquad f(t)u_{-1}(-t) \equiv \{f(t)\}_{-}$$
PFE terms w/ RHP poles $\qquad\qquad\qquad$ Negative time part

Chapter 12

Recursive LLSE: The Kalman Filter

12.1   Introduction

In Chapter 11 we studied LLSE estimation of stochastic processes based on observation of other, related stochastic processes. The focus in Chapter 11 was on the Wiener filter and its characterization of the optimal estimator as an explicit filter impulse response. In this Chapter we will study the *recursive computation* of the LLSE estimate of a random process. Such recursive computation of the LLSE is the centerpiece of the Kalman Filter, which is used throughout science and engineering. We will restrict our attention to the discrete-time case for simplicity. The flavor of the results for the continuous-time case is similar, but the theoretical development is more complicated. We will start by studying the simpler problem of recursively estimating a (static) random vector. Using the insights we develop there we will tackle the case of recursively estimating a random process. Throughout this chapter we will be concerned with LLSE estimates, which are also the MMSE estimates for the Gaussian case.

12.2   Historical Context

Before proceding to the mathematical developments leading to the Kalman filter, it is again useful to first consider the historical context for its development. Let us begin with a brief history of Rudolf Kalman, the inventor of the Kalman filter.

Rudolf Kalman was born May 19, 1930 in Budapest, Hungary and is currently a Professor of Mathematics (Emeritus) at the Swiss Federal Institute of Technology (ETH) in Switzerland. He emigrated from Hungary to the United States with his family towards the end of the World War II. He received his bachelor's degree in 1953 and his masters degree in 1954, both in Electrical Engineering from MIT. His master's thesis topic was the behavior of the solutions of second-order difference equations. He continued his stidues at Columbia University, where he received his Sc.D. in 1957 working on problems related to control theory. In 1958 he joined the Research Institute for Advanced Study (RIAS) where he worked from 1958 to 1964. It was during this time that the he developed the Kalman filter.

Thus, Kalman was actively involved with the Kalman filter development during the late 1950's and early 1960's. This was the start of the computer revolution and Kalman's view of the LLSE estimation problem must be understood in this context. In particular, Kalman derived his solution by viewing the problem in *state-space* form, which lead him to an associated *dynamic* definition of the optimal filter as an *algorithm*. Contrast this to the explicit expression of the filter provided by Kalman. By taking such a state-space view, non-stationarities in the underlying processes could be dealt with as well. Note that Kalman's solution would not have helped the engineers of Wiener's time, had they even had it, because it requires a computer to implement. Thus, beyond his discovery of the filter bearing his name, Kalman's contribution was showing the field a different way of conceptualizing what is meant as a "solution" to the LLSE problem – a conceptualization matched to the implementational paradigm of the times.

12.3   Recursive Estimation of a Random Vector

We begin our treatment by examining the simpler problem of estimating random *vector* based on observation of a discrete-time vector random process (i.e. a series of random vectors). In particular, consider the following LLSE estimation problem. Let $\underline{y}_0, \underline{y}_1, \ldots$ be a sequence of random vectors, and let $\widehat{\underline{x}}_k$ denote the LLSE estimate of $\underline{x}$ based on observation of $\underline{y}_0, \underline{y}_1, \ldots, \underline{y}_k$. Let Σ_k denote the corresponding error covariance of this estimate (so that e.g. $\mathsf{tr}(\Sigma_k) = \mathrm{MSE}$). What we would like to do is to develop a recursive procedure for computing $\widehat{\underline{x}}_{k+1}$ and Σ_{k+l} from the previous estimate $\widehat{\underline{x}}_k$, Σ_k, the new observation $\underline{y}_{k+1}$ and their joint second-order statistics. Ideally, we would like to use only the new measurement to perform this update.

Discrete-time Innovations Process   We will proceed by using the discrete-time *innovations process*. We saw the value of using an innovations approach in our treatment of the Wiener filter, and a similar approach will aid us here. To this end, let

$$\underline{e}_k = \underline{x} - \widehat{\underline{x}}_k \tag{12.1}$$

Then

$$E\left[\underline{e}_k\right] = 0, \qquad E\left[\underline{e}_k \underline{e}_k^T\right] = \Sigma_k \tag{12.2}$$

Note that from the geometric characterization of the LLSE estimate we have that the error is uncorrelated with all the observations in the past:

$$E\left[\underline{e}_k \underline{y}_j^T\right] = 0, \quad \text{for all} \quad j = 0, 1, \ldots, k. \tag{12.3}$$

We can restate our original problem as follows: Compute the LLSE estimate of

$$\underline{x} = \widehat{\underline{x}}_k + \underline{e}_k \tag{12.4}$$

given the information in the vector

$$\underline{Y} = \begin{bmatrix} \underline{y}_0 \\ \underline{y}_1 \\ \vdots \\ \underline{y}_{k+1} \end{bmatrix} \tag{12.5}$$

To solve this problem we can apply the analysis in Example 10.18. In doing this note that that $\widehat{\underline{x}}_k$ is a deterministic linear function of $\underline{y}_0, \underline{y}_1, \ldots, \underline{y}_k$ so that its LLSE estimate based on these vectors is just the function itself. Further, thanks to (12.3), the error $\underline{e}_k$ is uncorrelated with $\widehat{\underline{x}}_k$. Therefore, applying (10.175), we have that:

$$\widehat{\underline{x}}_{k+1} = \widehat{\underline{x}}_k + \widehat{\underline{e}}_k(\underline{Y}) \tag{12.6}$$

where $\widehat{\underline{e}}_k(\underline{Y})$ is the LLSE estimate (MMSE estimate if the random variables are jointly Gaussian) of $\underline{e}_k$ based on $\underline{Y}$. We can write an explicit expression for this estimate as

$$\widehat{\underline{e}}_k(\underline{Y}) = \Sigma_{eY}\Sigma_Y^{-1}\left(\underline{Y} - \underline{m}_Y\right) \tag{12.7}$$

Note that, thanks to the orthogonality properties in (12.3), we have

$$\Sigma_{eY} = E\{\underline{e}_k \underline{Y}^T\} = \begin{pmatrix} 0 & 0 & \cdots & 0 & E\left[\underline{e}_k \underline{y}_{k+1}^T\right] \end{pmatrix} \tag{12.8}$$

However, this is not enough to guarantee that (12.7) is a function only of $\underline{y}_{k+1}$ (minus its mean). Indeed, if $\underline{y}_0, \underline{y}_1, \ldots, \underline{y}_{k+1}$ are all correlated, then Σ_Y is a full matrix, and in general $\widehat{\underline{e}}_k(\underline{Y})$ is a function of all these measurements.

Suppose, however that $\underline{y}_{k+1}$ is uncorrelated with $\underline{y}_0, \underline{y}_1, \ldots, \underline{y}_k$. Then,

$$\Sigma_Y = \left[\begin{array}{c|c} \mathrm{Cov}\left(\begin{bmatrix} \underline{y}_0 \\ \underline{y}_1 \\ \vdots \\ \underline{y}_k \end{bmatrix}, \begin{bmatrix} \underline{y}_0 \\ \underline{y}_1 \\ \vdots \\ \underline{y}_k \end{bmatrix} \right) & 0 \\ \hline 0 & \Sigma_{y_{k+1}} \end{array} \right] \tag{12.9}$$

where

$$\Sigma_{y_{k+1}} = \text{Cov}\left(\underline{y}_{k+1}, \underline{y}_{k+1}\right) \tag{12.10}$$

Then, from (12.6)-(12.10), we have that

$$\widehat{\underline{x}}_{k+1} = \widehat{\underline{x}}_k + E\left[\underline{e}_k \underline{y}_{k+1}^T\right] \Sigma_{y_{k+1}}^{-1}\left(\underline{y}_{k+1} - \underline{m}_{y_{k+1}}\right) \tag{12.11}$$

Also, in this case, from (10.180), (12.8)–(12.10) and a bit of algebra, we obtain:

$$\Sigma_{k+1} = \Sigma_{e_{k+1}} = \Sigma_k - E\left[\underline{e}_k \underline{y}_{k+1}^T\right] \Sigma_{y_{k+1}}^{-1} E\left[\underline{e}_k \underline{y}_{k+1}^T\right]^T \tag{12.12}$$

A simple way to recognize the correctness of the above formula is to notice that, from (12.11), we have

$$\underline{e}_k = \underline{e}_{k+1} - E\left[\underline{e}_k \underline{y}_{k+1}^T\right] \Sigma_{y_{k+1}}^{-1}\left(\underline{y}_{k+1} - \underline{m}_{y_{k+1}}\right) \tag{12.13}$$

and that the two terms in the right hand side of the above equation are uncorrelated because of (12.3). Thus, the covariance of the left hand side is the sum of the covariances of the two terms on the right-hand side; that is,

$$\Sigma_k = \Sigma_{k+1} + E\left[\underline{e}_k \underline{y}_{k+1}^T\right] \Sigma_{y_{k+1}}^{-1} E\left[\underline{e}_k \underline{y}_{k+1}^T\right]^T \tag{12.14}$$

which yields (12.12).

The consequences of (12.11), (12.12) are substantial:

- In (12.11), we have the recursive equation we desire. The updated estimate $\widehat{\underline{x}}_{k+1}$ equals the previous estimate $\widehat{\underline{x}}_k$ plus the estimate of the previous estimation error based only on the latest measurement $\underline{y}_{k+1}$ (it is here where the lack of correlation with the previous data is needed).

- Indeed, this lack of correlation has reduced our problem to the standard static estimation formula. Specifically, if we regard $\widehat{\underline{x}}_k$ as our prior mean, then equations (12.11), (12.12) are exactly the same as (10.142), (10.143)! That is, we use our latest measurement to estimate the remaining random portion, $\underline{e}_k$, of $\underline{x}$ and reduce the covariance according to the standard LLSE estimation formula for estimating $\underline{e}_k$ based on $\underline{y}_{k+1}$.

While this is quite nice, we don't usually have the luxury of having an uncorrelated measurement sequence. However, what we can imagine doing in this case is the following. First note that if $\underline{\nu} = G\underline{y} + \underline{b}$ is an invertible transformation of $\underline{y}$ with G and $\underline{b}$ deterministic, the information content of $\underline{y}$ and $\underline{\nu}$ are identical, and so are the LLSEs based on the two vectors. Now, suppose we can find such a transformation on the measurement sequence $\underline{y}_0, \underline{y}_1, \ldots, \underline{y}_{k+1}$ so that:

- For each k the transformation is such that $\underline{\nu}_k$ is a linear function of $\underline{y}_0, \underline{y}_1, \ldots, \underline{y}_k$, and the map $\underline{y}_0, \ldots, \underline{y}_k \to \underline{\nu}_0, \ldots, \underline{\nu}_k$ is invertible.

- The $\underline{\nu}_j, j = 0, \ldots, k$ form an uncorrelated sequence of random variables.

Then, since $\widehat{\underline{x}}_k$ is the LLSE of $\underline{x}$ based on either $\underline{y}_0, \ldots, \underline{y}_k$ or $\underline{\nu}_0, \ldots, \underline{\nu}_k$, and thanks to (12.11), (12.12) and the lack of correlation of the sequence of $\underline{\nu}_j$, we have

$$\widehat{\underline{x}}_{k+1} = \widehat{\underline{x}}_k + E\left[\underline{e}_k \underline{\nu}_{k+1}^T\right] \Sigma_{\nu_{k+1}}^{-1}\left[\underline{\nu}_{k+1} - \underline{m}_{\nu_{k+1}}\right] \tag{12.15}$$

$$\Sigma_{k+1} = \Sigma_k - E\left[\underline{e}_k \underline{\nu}_{k+1}^T\right] \Sigma_{\nu_{k+1}}^{-1} E\left[\underline{e}_k \underline{\nu}_{k+1}^T\right]^T \tag{12.16}$$

The process $\underline{\nu}_0, \ldots, \underline{\nu}_k$ is known as the *innovations process*, and can be obtained as a result of the basic properties of LLSE. Specifically, let $\widehat{\underline{y}}(k|k-1)$ denote the LLSE of the vector $\underline{y}_k$ based on observation of the vectors $\underline{y}_0, \ldots, \underline{y}_{k-1}$, and define

$$\underline{\nu}_0 = \underline{y}_0 - \underline{m}_{y_0} \tag{12.17}$$

$$\underline{\nu}_k = \underline{y}_k - \widehat{\underline{y}}(k|k-1), \quad k = 1, \ldots \tag{12.18}$$

Then $\underline{\nu}_k$ is obviously a function of $\underline{y}_0, \ldots, \underline{y}_k$. To show that there is no loss of information, so that we can recover $\underline{y}_0, \ldots, \underline{y}_k$ from the innovations sequence $\underline{\nu}_0, \ldots, \underline{\nu}_k$, note that this is obviously true for $k = 0$. Proceeding by induction, assume that it is also true for all $j \leq k - 1$. Then, $\hat{\underline{y}}(k|k-1)$ is also the least squares estimate of $\underline{y}_k$ based on $\underline{\nu}_0, \ldots, \underline{\nu}_{k-1}$, so it is a linear function of the past innovations, and, from (12.18) we have

$$\underline{y}_k = \underline{\nu}_k + \hat{\underline{y}}(k \mid k - 1) \tag{12.19}$$

which shows that it is true for $j = k$ as well. Note finally that since $\underline{\nu}_k$ is the estimation error in estimating $\underline{y}_k$ based on $\underline{y}_0, \ldots, \underline{y}_{k-1}$, it is also uncorrelated with $\underline{y}_0, \ldots, \underline{y}_{k-1}$ and thus with $\underline{\nu}_0, \ldots, \underline{\nu}_{k-1}$. Thus, the sequence $\underline{\nu}_0, \ldots, \underline{\nu}_k$ satisfies the conditions we were looking for, and is a zero-mean innovations process.

Let us make several comments. Note first that the computation of the $\underline{\nu}_k$ involves the solution of a sequence of LLSE problems. Thus, for the computational efficiency of (12.15)–(12.16) to be of real value, the computation of these LLSE must be simple. While this is not always the case, it is true for the very important class of models discussed in this chapter. Finally, we note that the procedure which we have described for constructing the innovations process is the well-known *Gram-Schmidt Orthogonalization Procedure* for obtaining a set of orthogonal vectors from a set of linearly independent vectors. Indeed, there are strong geometric interpretations associated with the construction of the innovations sequence.

Another insight which we obtain from the innovations sequence is that it represents a particular factorization of the covariance vector of the observations. To illustrate this, consider a random vector $\underline{y}$ with components $y_1, \ldots, y_p$, and with covariance $\Sigma_y = (\Sigma_{ij})$. In this case, constructing the innovations one component at a time, and using (10.142), (10.143) and the uncorrelated property of the innovations, we have the following:

$$\begin{array}{rcll} \nu_1 & = & y_1, & \Sigma_{\nu_1} = \Sigma_{11} \\ \nu_2 & = & y_2 - a_{21}\nu_1, & \Sigma_{\nu_2} = \Sigma_{22} - a_{21}^2 \Sigma_{\nu_1} \end{array} \tag{12.20}$$

and more generally,

$$\nu_k = y_k - a_{k1}\nu_1 - \ldots - a_{k,k-1}\nu_{k-1}, \quad \Sigma_{\nu_k} = \Sigma_{kk} - a_{k1}^2 \Sigma_{\nu_1} - \ldots - a_{k,k-1}^2 \Sigma_{\nu_{k-1}} \tag{12.21}$$

where $a_{kj} = \frac{\Sigma_{y_k \nu_j}}{\Sigma_{\nu_j}}, j = 1, \ldots, k - 1$. These coefficients can be computed recursively, as:

$$\begin{array}{rcl} \Sigma_{y_k \nu_1} & = & \Sigma_{k1} \\ \Sigma_{y_k \nu_j} & = & \Sigma_{kj} - a_{j1}\Sigma_{k1} - \ldots - a_{j,j-1}\Sigma_{k,j-1}, \quad j = 1, \ldots, k - 1. \end{array} \tag{12.22}$$

Note that his procedure uses the elements of the matrix Σ_y to construct a lower triangular matrix

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21} & 1 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 1 & \cdots & 0 \\ \cdots & \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & a_{p3} & \cdots & 1 \end{bmatrix} \tag{12.23}$$

so that

$$\begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} = A \begin{bmatrix} \nu_1 \\ \vdots \\ \nu_p \end{bmatrix} \tag{12.24}$$

Taking covariances of both sides of (12.24) we have

$$\Sigma_y = A\Sigma_\nu A^T \tag{12.25}$$

$$\Sigma_\nu = \text{diag}\left(\Sigma_{\nu_1}, \ldots, \Sigma_{\nu_p}\right) \tag{12.26}$$

Equation 12.25 yields an LDU (lower-triangular, diagonal, upper-triangular) factorization of Σ_y. Since the lower and upper triangular parts of the factorization are transposes of each other, this factorization is known

as a Cholesky factorization in linear algebra. Note that the matrix A is trivially invertible, since it has 1 as its diagonal elements. Once this factorization is available, the inverse of Σ_y can be computed directly, as

$$\Sigma_y^{-1} = \left(A^T\right)^{-1} \Sigma_\nu^{-1} A^{-1} \tag{12.27}$$

Note also that the matrix A^{-1} is also lower-triangular.

What we have so far is a way, based on the innovations associated with data sequence, to recursively estimate a random vector. In the next section we will use these results to find the LLSE estimate a discrete-time *stochastic process* from such a data sequence. The solution this problem is the famed Kalman filter.

12.4 The Discrete-Time Kalman Filter

In this section, we use the innovations theory to solve an estimation problem for a stochastic process evolving in discrete time. The exposition in this section is entirely in terms of a vector-valued stochastic process $\underline{x}$, observed from another vector-valued stochastic process $\underline{y}$. The exposition is presented in terms of finding the MMSE estimate for Gaussian random vectors. However, recall that the same algorithm must obtain the LLSE estimator for any process, given the first-order and second-order statistics of the process.

Now to develop the DT Kalman Filter, consider the dynamic system:

$$\begin{align}
\underline{x}(t+1) &= A(t)\underline{x}(t) + B(t)\underline{u}(t) + G(t)\underline{w}(t) \tag{12.28} \\
\underline{y}(t) &= C(t)\underline{x}(t) + \underline{v}(t) \tag{12.29}
\end{align}$$

where $\underline{x}(t) \in R^n, \underline{y}(t) \in R^p, \underline{u}(t)$ is a known input, and $\underline{w}(t), \underline{v}(t)$ are independent, zero-mean, Gaussian white noise proceses, with

$$\begin{align}
E[\underline{w}(t)\underline{w}^T(s)] &= Q(t)\delta(t-s) \tag{12.30} \\
E[\underline{v}(t)\underline{v}^T(s)] &= R(t)\delta(t-s) \tag{12.31} \\
E[\underline{w}(t)\underline{v}^T(s)] &= 0 \tag{12.32}
\end{align}$$

In addition, assume that the initial conditions $\underline{x}(t_0)$ are Gaussian, with mean $\underline{m}_x(t_0)$ and covariance $P_x(t_0)$, and suppose that the initial conditions are independent of the process noise $\underline{w}(t)$ and the measurement noise $\underline{w}(t)$ for all times $t = 0, \ldots$. We also assume that the measurement noise covariance is positive definite ($R(t) > 0$), and thus invertible (i.e. there are no perfect observations). Note that the assumption that $\underline{x}(t_0), \underline{w}(t), \underline{v}(t)$ are Gaussian implies that all of the random variables are Gaussian. In this context, independence and uncorrelatedness are equivalent concepts. Thus, extension of the innovations concept to this problem is equivalent to requiring that the innovations be independent.

We will use the following notation throughout the development:

$$\begin{align}
\widehat{\underline{x}}(t \mid s) &= \text{LLSE of } \underline{x}(t) \text{ given } y(\tau), \tau \le s \tag{12.33} \\
\underline{e}(t \mid s) &= \underline{x}(t) - \widehat{\underline{x}}(t \mid s) \tag{12.34} \\
P(t \mid s) &= E\left[\underline{e}(t \mid s)\underline{e}(t \mid s)^T\right] \tag{12.35}
\end{align}$$

We are interested in developing a recursive approach for computation of $\widehat{\underline{x}}(t \mid t-1)$ or $\widehat{\underline{x}}(t \mid t)$. This is the problem of performing optimal causal filtering. The solution of this problem is the Kalman Filter. The discrete-time Kalman filter is often presented as a series of three steps. In particular, each step can be considered as an estimation sub-problem in its own right. We consider each of these subproblems next and through their solution find the solution to the overall Kalman filter.

12.4.1 Initialization

First we have an initialization step:

$$\begin{align}
\widehat{\underline{x}}(t_0|t_0 - 1) &= \underline{m}_x(t_0) \tag{12.36} \\
P(t_0|t_0 - 1) &= P_x(t_0) \tag{12.37}
\end{align}$$

That is, before any measurements are taken, the best estimates are based solely on the prior information.

12.4.2   Measurement Update Step

We start the actual estimation process by updating the estimate from the previous step to take into account the observation at the current time. That is:

$$\text{Suppose we have:} \quad \widehat{\underline{x}}(t \mid t-1), \quad P(t \mid t-1)$$
$$\text{And we observe:} \quad \underline{y}(t)$$

$$\text{Now compute:} \quad \widehat{\underline{x}}(t \mid t), \quad P(t \mid t)$$

The solution of the update step is just a direct application of the analysis in the preceding subsections. Specifically, let us define the innovations

$$\nu(t) = \underline{y}(t) - \widehat{\underline{y}}(t \mid t-1) \tag{12.38}$$

Since $\underline{v}(t)$ is uncorrelated with $\underline{y}(0), \ldots, \underline{y}(t-1)$ and with $\underline{x}(t)$, we can readily compute

$$\widehat{\underline{y}}(t \mid t-1) = C(t)\widehat{\underline{x}}(t|t-1), \quad \nu(t) = C(t)\underline{e}(t \mid t-1) + \underline{v}(t) \tag{12.39}$$

and

$$P_{\nu(t)} = E\left[\nu(t)\nu(t)^T\right] = C(t)P(t \mid t-1)C^T(t) + R(t) \tag{12.40}$$

Writing $\underline{x}(t) = \underline{e}(t \mid t-1) + \widehat{\underline{x}}(t|t-1)$, and using the fact that the innovations are zero-mean, we can then apply the estimation formula for Gaussian random vectors with Gaussian observations to obtain the update relations:

$$\widehat{\underline{x}}(t \mid t) = \widehat{\underline{x}}(t \mid t-1) + E\left[\underline{e}(t \mid t-1)\nu(t)^T\right] P_{\nu(t)}^{-1}\nu(t) \tag{12.41}$$

$$P(t \mid t) = P(t \mid t-1) - E\left[\underline{e}(t \mid t-1)\nu(t)^T\right] P_{\nu(t)}^{-1} E\left[\underline{e}(t|t-1)\nu(t)^T\right]^T \tag{12.42}$$

Furthermore,

$$\begin{aligned}
E\left[\underline{e}(t \mid t-1)\nu(t)^T\right] &= E\left[\underline{e}(t \mid t-1)\left(C(t)\underline{e}(t|t-1) + \underline{v}(t)\right)^T\right] \\
&= E\left[\underline{e}(t \mid t-1)\underline{e}(t \mid t-1)^T\right] C^T(t) + E\left[\underline{e}(t \mid t-1)\underline{v}(t)^T\right] \\
&= P(t|t-1)C^T(t) \tag{12.43}
\end{aligned}$$

where we have used the definition of $P(t|t-1)$ and the fact that the measurement noise $\underline{v}(t)$ is uncorrelated (independent) of $\underline{x}(t)$ and $\underline{y}(t_0), \ldots, \underline{y}(t-1)$. Substituting (12.43) into (12.41), (12.42) yields the following set of equations for the update step:

$$\widehat{\underline{x}}(t|t) = \widehat{\underline{x}}(t|t-1) + P(t|t-1)C^T(t)\left[C(t)P(t|t-1)C^T(t) + R(t)\right]^{-1}\left[\underline{y}(t) - C(t)\widehat{\underline{x}}(t|t-1)\right] \tag{12.44}$$

$$P(t|t) = P(t|t-1) - P(t|t-1)C^T(t)\left[C(t)P(t|t-1)C^T(t) + R(t)\right]^{-1}C(t)P(t|t-1) \tag{12.45}$$

Notice that this step simply updates the current estimate to take into account the new observation – the dynamic equation is not used.

12.4.3   Prediction Step

Now we perform a prediction step, where we use the dynamic equation to generate the best estimate at time $t+1$ based only on the data up to time t. That is, we solve the following subproblem:

$$\text{Suppose we have:} \quad \widehat{\underline{x}}(t \mid t), \quad P(t \mid t)$$

$$\text{Now compute:} \quad \widehat{\underline{x}}(t+1 \mid t), \quad P(t+1 \mid t)$$

The solution of the prediction step is simple to derive, since by assumption $\underline{w}(t)$ is independent of (uncorrelated with) $\underline{y}(0), \ldots \underline{y}(t)$, and $B(t)\underline{u}(t)$ is deterministic. Thus,

$$
\begin{aligned}
\widehat{\underline{x}}(t+1|t) &= E\left[\underline{x}(t+1) \mid \underline{y}(0), \ldots \underline{y}(t)\right] && (12.46) \\
&= E\left[A(t)\underline{x}(t) + B\underline{u}(t) + G(t)\underline{w}(t) \mid \underline{y}(0), \ldots \underline{y}(t)\right] \\
&= A(t)E\left[\underline{x}(t) \mid \underline{y}(0), \ldots \underline{y}(t)\right] + B\underline{u}(t) && (12.47) \\
&= A(t)\widehat{\underline{x}}(t \mid t) + B(t)\underline{u}(t) && (12.48)
\end{aligned}
$$

Similarly, we can obtain an expression for the error $\underline{e}(t+1|t)$ as

$$
\underline{e}(t+1 \mid t) = \underline{x}(t+1) - \widehat{\underline{x}}(t+1 \mid t) = A(t)\underline{e}(t \mid t) + G(t)\underline{w}(t) \tag{12.49}
$$

Note that the two terms in the above equation are independent by assumption, because the process noise $\underline{w}(t)$ is independent of all past and current values of the state $\underline{x}$ and past and current values of the measurement noise $\underline{v}$. Thus, we obtain for the predicted error covariance:

$$
P(t+1 \mid t) = E\left[\underline{e}(t+1 \mid t)\underline{e}(t+1 \mid t)^T\right] = A(t)P(t \mid t)A^T(t) + G(t)Q(t)G^T(t) \tag{12.50}
$$

Together we have for the prediction step:

$$
\begin{aligned}
\widehat{\underline{x}}(t+1|t) &= A(t)\widehat{\underline{x}}(t \mid t) + B(t)\underline{u}(t) && (12.51) \\
P(t+1 \mid t) &= A(t)P(t|t)A^T(t) + G(t)Q(t)G^T(t) && (12.52)
\end{aligned}
$$

Note this is simply a one-step prediction and does not involve the observation.

12.4.4   Summary

Combining these steps we obtain the DT Kalman Filter, which we summarize here for convenience. First, the process in question is described by the following autoregressive dynamic equation and observation equation:

$$
\begin{aligned}
\underline{x}(t+1) &= A(t)\underline{x}(t) + B(t)\underline{u}(t) + G(t)\underline{w}(t) && (12.53) \\
\underline{y}(t) &= C(t)\underline{x}(t) + \underline{v}(t) && (12.54)
\end{aligned}
$$

where $\underline{u}(t)$ is a known input, and $\underline{w}(t)$, $\underline{v}(t)$ are independent, zero-mean, white noise processes, with

$$
\begin{aligned}
E[\underline{w}(t)\underline{w}^T(s)] &= Q(t)\delta(t-s) && (12.55) \\
E[\underline{v}(t)\underline{v}^T(s)] &= R(t)\delta(t-s) && (12.56) \\
E[\underline{w}(t)\underline{v}^T(s)] &= 0 && (12.57)
\end{aligned}
$$

where we assume that $R(t) > 0$. Also the second order statistics of the initial condition are given by:

$$
E\left[\underline{x}(t_0)\right] = \underline{m}_x(t_0), \qquad E\left[(\underline{x}(t_0) - \underline{m}_x(t_0))(\underline{x}(t_0) - \underline{m}_x(t_0))^T\right] = P_x(t_0) \tag{12.58}
$$

and the initial conditions are independent of the process noise $\underline{v}(t)$ and the measurement noise $\underline{w}(t)$ for all times $t = 0, \ldots$. The Kalman filter for this system is given by:

Initialization:

$$
\begin{aligned}
\widehat{\underline{x}}(t_0|t_0 - 1) &= \underline{m}_x(t_0) && (12.59) \\
P(t_0|t_0 - 1) &= P_x(t_0) && (12.60)
\end{aligned}
$$

Update Step:

$$
\widehat{\underline{x}}(t|t) = \widehat{\underline{x}}(t|t-1) + P(t|t-1)C^T(t)\left[C(t)P(t|t-1)C^T(t) + R(t)\right]^{-1}\left[\underline{y}(t) - C(t)\widehat{\underline{x}}(t|t-1)\right] \tag{12.61}
$$

$$
P(t \mid t) = P(t \mid t-1) - P(t \mid t-1)C^T(t)\left[C(t)P(t \mid t-1)C^T(t) + R(t)\right]^{-1}C(t)P(t \mid t-1) \tag{12.62}
$$

Prediction Step:

$$\widehat{\underline{x}}(t+1 \mid t) = A(t)\widehat{\underline{x}}(t \mid t) + B(t)\underline{u}(t) \tag{12.63}$$

$$P(t+1 \mid t) = A(t)P(t \mid t)A^T(t) + G(t)Q(t)G^T(t) \tag{12.64}$$

Note that the Kalman filter has an intuitively appealing structure: the filter mimics the noise-free dynamics for prediction (cf (12.63)) and corrects for the difference between the observations $\underline{y}(t)$ and the best prediction of $\underline{y}(t)$ based on the preceding data (cf (12.61)).

12.4.5 Additional Points

First there are several alternate forms for these equations. In particular, we have for the update step:

$$\widehat{\underline{x}}(t \mid t) = \widehat{\underline{x}}(t \mid t-1) + K(t)\underline{\nu}(t) \tag{12.65}$$

$$K(t) = P(t|t-1)C^T(t)\left[C(t)P(t|t-1)C^T(t) + R(t)\right]^{-1} \tag{12.66}$$

$$= P(t|t)C(t)R^{-1}(t) \tag{12.67}$$

where $K(t)$ as defined above is the *Kalman gain*. Also there is the equivalent formula for the error covariance:

$$P(t|t) = [I - K(t)C(t)]\,P(t|t-1)\,[I - K(t)C(t)]^T + K(t)R(t)K^T(t) \tag{12.68}$$

Sometimes this form is preferred from a numerical standpoint, since it represents the error covariance as the sum of two positive semi-definite quantities (which must be positive semi-definite) rather than as the *difference* of two such quantities (which could, with e.g. roundoff, become indefinite).

In addition, for the update step there is the following equivalent "information" form of the error covariance update:

$$P^{-1}(t|t) = P^{-1}(t|t-1) + C^T(t)R^{-1}(t)C(t) \tag{12.69}$$

This last form emphasizes some important insights about the Kalman filter. In particular, note that in the prediction step (12.64) it is the *covariances* or uncertainty that is additive. This makes sense since in this step we have no observation but are only taking into account the effect of the dynamic equation, which is increasing the uncertainty in the problem do to the additive noise $\underline{w}(t)$. In contrast, consider the update step. In this step, we are accounting for the influence of an observation. By viewing the inverse of a covariance as a measure of information, we can see from (12.69) that it is *information* that is additive. In other words, the uncertainty increases during the prediction step and decreases during the update step. The structure of this relationship has deep consequences for our ability to efficiently implement the Kalman filter, which are beyond the scope of this course.

Finally, also note that the recursion (12.62), (12.64) for the error covariance does not depend on the data $\underline{y}(t)$ and thus can be precomputed. Thus, the gain $P(t|t-1)C^T(t)[C(t)P(t|t-1)C^T(t) + R(t)]^{-1}$ can also be precomputed. Equations (12.62), (12.64) are referred together as the discrete-time Riccati equation.

12.4.6 Example

We now consider an example. We will use the same setup used in Example 11.9 in the section on Wiener filtering so we may compare the results.

Example 12.1
The underlying process $x(t)$ is zero mean with its covariance structure *implicitly* described by the following autoregressive model and observation equation:

$$x(t+1) = 0.8x(t) + w(t) \tag{12.70}$$

$$y(t) = x(t) + v(t) \tag{12.71}$$

where $w(t)$ and $v(t)$ are zero mean, wide-sense stationary white noise processes, uncorrelated with each other with covariance functions $K_{ww}(t) = 0.36\delta(t)$ and $K_{vv}(t) = 1\delta(t)$, respectively. We are also given that $E[x(0)] = 0$ and $E[x^2(0)] = 1$, where $x(0)$ is independent of $w(t)$ and $v(t)$. Our goal is to find the Kalman filter for this problem – that is, find the linear least square estimate of $x(t)$ based on the statistical initial condition and the data $y(t)$ observed up to time t. Note that this is a *filtering* problem, in that the estimate only uses data up to the current time.

We simply apply the Kalman filtering equations in Section 12.4.4 with $A = 0.8$, $B = 0$, $C = 1$, $G = 1$:

Initialization:

$$\widehat{x}(0|-1) \quad = \quad m_x(0) = 0 \tag{12.72}$$

$$P(0|-1) \quad = \quad R_{xx}(0) = 1 \tag{12.73}$$

Update Step:

$$\widehat{x}(t|t) = \widehat{x}(t|t-1) + \underbrace{\left[\frac{P(t|t-1)}{P(t|t-1)+1}\right]}_{\text{Kalman Gain } K(t)} [y(t) - \widehat{x}(t|t-1)] \tag{12.74}$$

$$P(t|t) = \frac{P(t|t-1)}{P(t|t-1)+1} \tag{12.75}$$

Prediction Step:

$$\widehat{x}(t+1|t) = 0.8\widehat{x}(t|t) \tag{12.76}$$

$$P(t+1|t) = (0.8)^2 P(t|t) + 0.36 \tag{12.77}$$
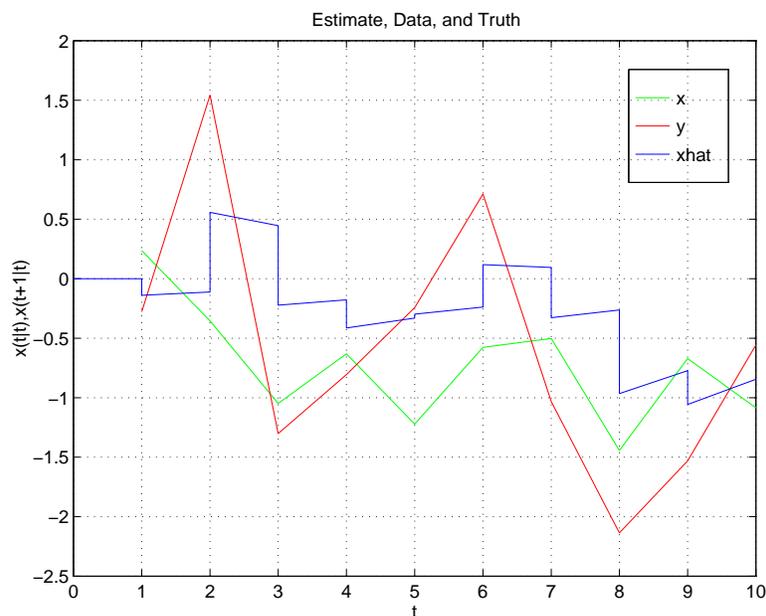
Figure 12.1: Kalman Filtering Example: Estimate

In Figure 12.1 we show an example where a "true" process and noisy observations were generated according to the description of (12.70). This noisy signal was then used as input to the Kalman filtering equations (12.72)–(12.77) to obtain an estimate. The figure shows the original signal in green and the noisy observations in red. In blue are shown both the predicted estimate $\widehat{x}(t+1|t)$ and the updated estimate $\widehat{x}(t|t)$ after the current observation is taken into account. The predicted estimates lead from one discrete time point to the next using the system dynamics to propogate the estimate forward. This part of the filter corresponds to the piecewise linear parts of the blue curve. The predicted estimate is following open loop decay provided by the system dynamics, and thus in the absence of observations will decay to zero as $(0.8)^t$, which is why all these linear segments are "pointed" to 0.

At each time point the predicted estimate is then corrected to take into account the observation at that time. This correction exhibits itself as the "jumps" in the estimated signal at each time. For this example, where we are observing the points themselves in noise, this will tend to pull the estimate in the direction of the current observation, as can be seen in the figure. Overall, we can view the situation as follows: we have *two* estimates at each time – one just prior to the observation (that has not seen it yet) and one just after the observation (which has taken it into account). It is not surprising that they are different!

Now in Figure 12.2 we show the associated error covariance values and the corresponding Kalman gain. First consider Figure 12.2(a), which shows both the predicted error covariance $P(t+1|t)$ and the updated error covariance $P(t|t)$. Can you guess from intuition which must be which? In the prediction step we have no new data, but are only taking the
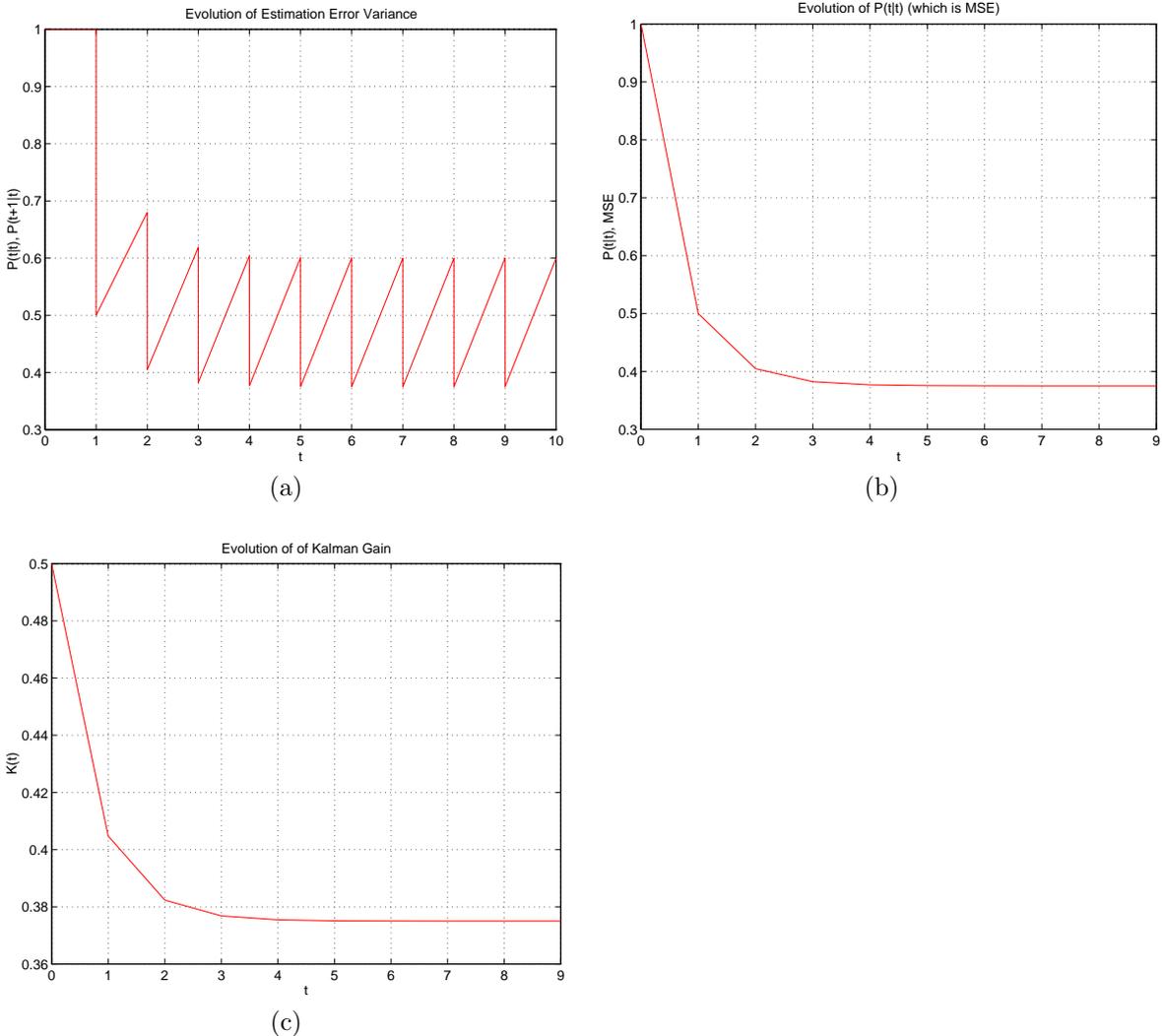
Figure 12.2: Kalman Filtering Example: Covariance and Gain

dynamics into account. Since the dynamic equation is driving by a noise process, our uncertainty will increase. Thus the linear segments connecting one discrete time step with the next are due to the prediction step and the larger value at each time is evidentially $P(t+1|t)$. Next, the observation is taken into account during the update step. The observation helps the estimate, removing uncertainty and accounting for the drop in error covariance at each time. Thus the lower value at each time must be $P(t|t)$. Overall we see the ongoing battle between the dynamics pumping uncertainty into the problem and the observations taking it out, as exhibited in the sawtooth nature of the waveform. In such a display it is difficult to see if any sort of steady state is being reached.

In Figure 12.2(b) we have displayed just the updated covariance $P(t|t)$, which is just the lower envelop of the curve in Figure 12.2(a). This is the mean square error for the problem. Note that this value is appearently approaching a steady state as time going on. In Figure 12.2(c) we have displayed the corresponding Kalman gain as a function of time. It too appears to be approaching a steady state value as time evolves.

The above discussion raises the question of what happens in steady state. If a steady state exists then we would expect that

$$P \;=\; P(t+1|t) = P(t|t-1) \tag{12.78}$$

where P denotes the steady state predicted covariance. Note that we *are not* saying that $P(t+1|t)$ will equal $P(t|t)$, which it is clear from Figure 12.2(a) will never be the case! Now substituting the expression for the updated error covariance

$P(t|t)$ in (12.75) into the expression for the predicted error covariance $P(t+1|t)$ in (12.77) we obtain:

$$P(t+1|t) \quad = \quad (0.8)^2 \frac{P(t|t-1)}{P(t|t-1)+1} + 0.36 \tag{12.79}$$

$$\implies \quad P \quad = \quad (0.8)^2 \frac{P}{P+1} + 0.36 \tag{12.80}$$

where we have made the substitution for the steady state error covariance in obtaining the last equation. Solving this equation for P we obtain:

$$P \quad = \quad 0.6 \tag{12.81}$$

Substituting this steady state value of the error covariance into the expression for the Kalman gain, we obtain a corresponding steady state value for the Kalman gain:

$$K \quad = \quad \frac{P}{P+1} = 0.375 \tag{12.82}$$

We can also find the value for the mean square error in steady state:

$$\mathsf{MSE} \quad = \quad P(t|t) = \frac{P}{P+1} = 0.375 \tag{12.83}$$

We can now find the filter in steady state:

$$\widehat{x}(t|t) \quad = \quad \widehat{x}(t|t-1) + K\left[y(t) - \widehat{x}(t|t-1)\right] \tag{12.84}$$

$$= \quad (1-K)0.8\widehat{x}(t-1|t-1) + Ky(t) \tag{12.85}$$

$$= \quad 0.8(1 - 0.375)\widehat{x}(t-1|t-1) + 0.375y(t) \tag{12.86}$$

$$= \quad \frac{1}{2}\widehat{x}(t-1|t-1) + 0.375y(t) \tag{12.87}$$

Notice that this is nothing more than linear time invariant system with input $y(t)$ (i.e. the data) and output $\widehat{x}(t|t)$ (i.e. the estimate). Since it is an LTI system we can use transform techniques to find its system function, which relates the input and output:

$$H_{kf}(z) \quad = \quad \frac{0.375}{1 - \frac{1}{2}z^{-1}} \tag{12.88}$$

Compare this solution with that obtained in Example 11.9.

Note that the steady state solution obtained in Example 12.1 is precisely the *same* filter that we obtained in Example 11.9 for the causal Wiener filter for this problem! Thus we have discovered the following important result: the Kalman filter for a time invariant system in steady state is the same as the causal Wiener filter. Note that the MSE we found for the steady state Kalman filter (12.83) is also the same as the MSE we found for the causal Wiener filter for this problem, as it must be since both are just finding the LLSE estimate. Let us emphasize that it is the *causal* Wiener filter that the Kalman filter tends to in steady state.

12.4.7 Comparison of the Wiener and Kalman Filter

We close this chapter with a comparison of the causal Wiener and Kalman filters. This comparison is given in Table 12.1. While the causal Wiener filter and the Kalman filter are in general different, if we apply the Kalman filter to an stationary process arising from an LTI system over an infinite time horizon the corresponding steady-state Kalman filter approaches the causal Wiener filter. This makes perfect sense, since in this case they are both solving the same problem. In this case note that the causal Wiener filter is a "frequency domain" solution while the Kalman filter is a "time-domain" solution.

Causal Wiener Filter	Kalman Filter
Infinite Observation Interval	Finite Observation Interval
Stationary Processes	Non-Stationary Processes
LTI Filter	Time Varying Filter
Closed Form Solution	Recursive Algorithm

Table 12.1: Comparison of the causal Wiener filter and the Kalman filter.

Chapter 13

Discrete State Markov Processes

Previously, we have discussed the concept of Markov processes. In this section, we want to focus on continuous-time, discrete-valued Markov processes. A special case of this process is the Poisson processss. Before we proceed to analyze this process, we want to provide an introduction to the theory of discrete-valued, continuous time Markov processes. In particular, by means of a limiting argument, we will obtain a characterization of such processes in terms of transition rates, using a limiting argument for discrete time, discrete-valued random processes.

13.1 Discrete-time, Discrete Valued Markov Processes

Let $x(n), n = 0, \ldots$ be a discrete time, discrete-valued Markov process with possible values $\{1, 2, \ldots\}$. We assume that the process can have an infinite number of states; the case where there are only a finite number of states is a special case, and is often referred to as a Markov Chain. The Markov process is characterized by its initial distribution

$$\underline{p}(0) = \begin{bmatrix} p(x(0) = 1) \\ p(x(0) = 2) \\ \vdots \end{bmatrix}$$

and the one-step transition matrices (with an infinite number of rows and columns) with elements $p_{ij}(n\Delta)$, where

$$p_{ij}(n) \equiv P(x(n+1) = j \mid x(n) = i) \tag{13.1}$$

Note that the transition probabilities are time-dependent, so that this is an inhomogeneous Markov process. Let $P(n)$ denote the transition matrix at time $n\Delta$.

Using the above equations, we can solve recursively for $\underline{p}(n\Delta)$ as

$$\underline{p}(n) \equiv \begin{bmatrix} p(x(n) = 1) \\ p(x(n) = 2) \\ \vdots \end{bmatrix} = P(0, n)^T \underline{p}(0) \tag{13.2}$$

where $P(m, n) = P(m)P(m+1) \cdots P(n-1)$ for $0 \le m < n$. The multistep transition matrix satisfies the Chapman-Kolmogorov equation

$$P(k, n) = P(k, m)P(m, n) \quad \text{for } k \le m < n$$

Note that $P(k, k) = I$.

Discrete-time one-step ($P(n)$) and multi-time step ($P(k, n)$) transition probability matrices must satisfy the laws of conservation of probability. That is, for any i, we must have

$$\sum_{j=1}^{\infty} p_{ij}(n) = 1 \tag{13.3}$$

When the number of states is finite and equal to N, the state transition matrix will be an $N \times N$ matrix $P(n)$, where $P(n)$ is such that all of its rows sum up to 1. Matrices with the property that the rows sum up to 1 are known as *stochastic matrices*. The other key property of transition probabilty matrices is that all of their elements are nonnegative: $p_{ij} \geq 0$. Combined with the fact that the columns add up to 1, one can use the following theorem from linear algebra:

Gershgoren's Theorem: Consider a square matrix A of dimension $n \times n$. Define distances $d_i = \sum_{j=1, j \neq i}^{n} |A_{ij}|$. Define the set $L = \cup_{i=1}^{n} \{|\lambda - A_{ii}| \leq d_i\}$. Then, all of the eigenvalues of A are contained in the set L.

In other words, the magnitudes of the off-diagonal elements can be summed up, and provide a bound for how far away from the diagonal elements are the eigenvalues. The set L consists of the union of circles of radius d_i centered around each of the diagonal elements A_{ii}. The important implication of Gershgoren's theorem for stochastic matrices is that, since the rows must add up to 1 and all of the elements are positive, all of the circles are inside the unit circle, so that all of the eigenvalues of stochastic matrices have magnitude less than or equal to 1! Furthermore, we know that at least one eigenvalue has value equal to 1, because

$$A \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

since each row must add up to 1.

A special class of discrete-time, discrete-valued Markov processes occurs when the transition probability matrix $P(n)$ does not depend on time. In this case, this class of Markov processes is called *homogeneous*, and has the property that $P(n) \equiv P$ for all $n \geq 0$. In this case, $P(0, n) = P^n$, and, since all of the eigenvalues of P are less than or equal to 1, a limit exists as $n \to \infty$. That is,

$$\lim_{n \to \infty} P^n = P_\infty \tag{13.4}$$

and

$$\lim_{n \to \infty} \underline{p}(n) = \lim_{n \to \infty} (P^n)^T \underline{p}(0) = P_\infty^T \underline{p}(0) \tag{13.5}$$

Under some regularity conditions on the Markov process which guarantee that there is a unique eigenvalue of P with magnitude 1 (an example of such a condition is when there is a path of transitions with non-zero probability between every two pairs of states i and j), one can show that the above limit does not depend on $\underline{p}(0)$; that is,

$$\lim_{n \to \infty} \underline{p}(n) = P_\infty^T \underline{p}(0) = \pi$$

where π are known as the steady-state probabilities or the *ergodic* probabilities; these probabilities satisfy the equation

$$P^T \pi = \pi \tag{13.6}$$

and thus, π is the unique eigenvector of P^T corresponding to the eigenvalue 1. Note that π is a vector of probabilities which must sum up to 1, thereby uniquely specifying π.

13.2 Continuous-Time, Discrete Valued Markov Processes

To convert the above results to the continuous-time case, we want to let $\Delta \to 0$. Suppose that, for very small Δ, we have

$$p_{ij}(n\Delta) = \begin{cases} q_{ij}(n\Delta)\Delta + o(\Delta) & \text{if } i \neq j \\ 1 - \sum_{j \neq i} q_{ij}(n\Delta)\Delta + o(\Delta) & \text{if } i = j \end{cases}$$

where $o(\Delta)$ denotes a term for which $\lim_{\Delta ra 0} \frac{o(\Delta)}{\Delta} = 0$. The above equation should be interpreted as indicating that the probability that there is a transition in the value of the state from i to j is linearly

proportional to Δ for small Δ; with probability close to 1, the process will not undergo any transitions. The quantity $q_{ij}(n\Delta)$ is called the transition rate from i to j at time $n\Delta$.

Now, we would like to take the limit of the above equation as $\Delta \to 0$, while holding the product $n\Delta = t$. Define as before the multistage transition matrix $P(s,t) = [p_{ij}(s,t)]$, where

$$p_{ij}(s,t) = P(x(t) = j | x(s) = i)$$

From eq. (13.2), we have

$$\underline{p}(t) \equiv \begin{bmatrix} P(x(t) = 0) \\ P(x(t) = 1) \\ \vdots \end{bmatrix} = P(0,t)^T \underline{p}(0)$$

What is missing to completely specify the continuous-time limit is to determine $P(s,t)$ for any $s \leq t$. In order to do this, we use the limit process. For a Markov process, we know that the Chapman-Kolmogorov equation holds:

$$P(s,t) = P(s, t-\Delta)P(t-\Delta, t)$$

Substituting the definition of the one-step transition probability from eq. (13.1), we obtain

$$P(s,t) = P(s, t-\Delta) \begin{bmatrix} 1-\sum_{j\neq 1} q_{1j}(t-\Delta)\Delta+o(\Delta) & q_{12}(t-\Delta)\Delta + o(\Delta) & q_{13}(t-\Delta)\Delta + o(\Delta) & \cdots \\ q_{21}(t-\Delta)\Delta + o(\Delta) & 1-\sum_{j\neq 2} q_{2j}(t-\Delta)\Delta+o(\Delta) & q_{23}(t-\Delta)\Delta + o(\Delta) & \cdots \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix} \quad (13.7)$$

$$= P(s, t-\Delta)(I + Q(t-\Delta)\Delta + o(\Delta)) \qquad (13.8)$$

where

$$Q(t-\Delta) = \begin{bmatrix} -\sum_{j\neq 1} q_{1j}(t-\Delta) & q_{12}(t-\Delta) & q_{13}(t-\Delta) & \cdots \\ q_{21}(t-\Delta) & -\sum_{j\neq 2} q_{2j}(t-\Delta) & q_{23}(t-\Delta) & \cdots \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix} \qquad (13.9)$$

Thus, rearranging terms and dividing by Δ, we get

$$\frac{P(s,t) - P(s,t-\Delta)}{\Delta} = P(s,t-\Delta)Q(t-\Delta) + P(s,t-\Delta)\frac{o(\Delta)}{\Delta} \qquad (13.10)$$

Taking limits as $\Delta \to 0$ gives

$$\frac{\partial}{\partial t}P(s,t) = P(s,t)Q(t) \quad \text{for } t \geq s \geq 0 \qquad (13.11)$$

with the initial condition $P(s,s) = I$. Note also that $\underline{p}(t) = P(s,t)^T\underline{p}(s)$; thus, taking the transpose of the above equation and multiplying by $\underline{p}(s)$ on the right gives

$$\frac{\partial}{\partial t}P(s,t)^T\underline{p}(s) = \frac{d}{dt}\underline{p}(t) = Q(t)^T P(s,t)^T \underline{p}(s) = Q(t)^T\underline{p}(t) \qquad (13.12)$$

subject to the initial condition $\underline{p}(0)$. The above equation governs how the probability mass function of $x(t)$ evolves over time.

The matrix $Q(t)$ has some special properties: in particular, note that the rows sum up to 0! This implies that $Q(t)$ has at least one zero eigenvalue. Indeed, Gershgoren's theorem implies that all of the eigenvalues of $Q(t)$ have non-positive real part!

As in the discrete-time case, one can consider the special case of *homogeneous* Markov processes where the transition rates $q_{ij}(t) \equiv q_{ij}$ are independent of time. In this case, equation (13.12) becomes

$$\frac{d}{dt}\underline{p}(t) = Q^T\underline{p}(t) \qquad (13.13)$$

$$\underline{p}(t) = e^{Q^T t}\underline{p}(0)$$

Following the analogy with the discrete-time case, in case that Q has a single zero eigenvalue, the above equation will converge as $t \to \infty$ to

$$\pi = \lim_{t \to \infty} \underline{p}(t) \tag{13.14}$$

where π will be the unique eigenvector satisfying

$$Q^T \pi = 0$$

with nonnegative elements summing up to 1.

There is one key property of homogeneous Markov processes which is worth noting. Define the time T_i of state i as

$$T_i = \min\{t | \underline{p}(t) \neq \underline{p}(0), \quad \text{where} \quad P(x(0) = i) = 1\} \tag{13.15}$$

That is, T_i is the time until the process which starts at state i leaves state i. Then, T_i is an exponentially distributed random variable with rate $-q_{ii} = \sum_{j \neq i} q_{ij}$. An equivalent statement is

$$P(T_i < t) = e^{-t/q_{ii}}$$

A simple way to see this is to note that, as long as $x(s) = i$ for all $s < t$, the i-th element of equation (13.13) becomes

$$\frac{d}{dt} \underline{p}(t)_i = -q_{ii} \underline{p}(t)_i \tag{13.16}$$

which gives the above expression.

13.3 Birth-Death Processes

Rather than continuing to focus on general discrete-valued, continuous-time Markov processes, let's examine first a simpler case which serves as the foundation for queuing theory: the case of continuous-time birth-death processes. A continuous-time birth-death process $x(t)$ is a discrete-time, continuous-state Markov process which has the special transition rate matrix defined as

$$[Q(t)]_{ij} = \begin{cases} \lambda_i(t) & \text{if } i = j - 1 \\ \mu_i(t) & \text{if } i = j + 1 \\ -(\lambda_i(t) + \mu_i(t)) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{13.17}$$

where we define $\mu_1(t) = 0$ to indicate that there can be no transitions to a value below 1. In other words, over a small interval Δ, the process $x(t) = i$ has a small probability $\lambda_i(t)\Delta$ that its value increases by 1, a small probability $\mu_i(\Delta)$ that its value decreases by 1, and a probability $1 - (\lambda_i(t) + \mu_i(t))\Delta$ that its value stays the same. The parameters $\lambda_i(t)$ are the birth rates, and the parameters $\mu_i(t)$ are the death rates.

A birth-death process is said to be *homogeneous* if the birth rates and death rates are independent of time. For a homogeneous birth-death process, we have $P(s,t) = P(0, t - s)$ for $0 \leq s \leq t$.

Example 13.1
Assume that $x(t)$ is a homogeneous birth-death process with birth rates

$$\lambda_i = \begin{cases} \lambda & \text{if } i = 1 \\ 0 & \text{otherwise} \end{cases}$$

and death rates

$$\mu_i = \begin{cases} \lambda & \text{if } i = 2 \\ 0 & \text{otherwise} \end{cases}$$

Then, the process $x(t)$ can only take values on $\{1, 2\}$. This is known as a *finite-state* Markov process. Define a new process $y(t) = 2x(t) - 3$; then, $y(t)$ looks like the random telegraph process. Indeed, we have constructed the random telegraph process, as we can verify below! The transition probability density $P(s,t)$ satisfies the equation

$$\frac{\partial}{\partial t} P(s,t) = P(s,t) \begin{bmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{bmatrix} \equiv P(s,t)Q$$

subject to the initial condition $P(s,s) = I$. The solution of this equation is given by

$$P(s,t) = e^{Q(t-s)} = I + Q(t-s) + \frac{Q^2}{2!}(t-s)^2 + \dots$$

Evaluating the exponential, we can compute the probability that $y(t) = 1$, conditioned on $y(0) = 0$, as $P(0,t)_{12} = e^{-\lambda t}\sinh(\lambda t)$, which is the same probability distribution as was computed for the random telegraph process.

Example 13.2

Suppose $x(t)$ is a homogeneous birth process that is 1 with certainty at $t = 0$, and has constant birth rates $\lambda_i(t) = \lambda$. The resulting process is the same as $N(t) + 1$, where $N(t)$ is a standard Poisson process with rate λ. To see this, since the process is homogeneous,

$$P(s,t) = P(0, t-s) \text{ for } 0 \le s \le t$$

Furthermore, since the initial condition is $\underline{p}(0) = \begin{bmatrix} 1 & 0 & 0 & \cdots \end{bmatrix}^T$, we have

$$\underline{p}(t) = \begin{bmatrix} P(0,t)_{11} \\ P(0,t)_{12} \\ \vdots \end{bmatrix}$$

Furthermore, the Chapman-Kolmogorov equation becomes:

$$\frac{d}{dt}P(0,t) = P(0,t)\begin{bmatrix} -\lambda & \lambda & 0 & \cdots \\ 0 & -\lambda & \lambda & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

In particular, looking at the first row of the matrix $P(0,t)$, we find

$$\frac{d}{dt}P(0,t)_{11} = -\lambda P(0,t)_{11}$$

$$\frac{d}{dt}P(0,t)_{12} = \lambda P(0,t)_{11} - \lambda P(0,t)_{12}$$

$$\frac{d}{dt}P(0,t)_{1n} = \lambda P(0,t)_{1(n-1)} - \lambda P(0,t)_{1n}$$

Solving recursively, we find

$$\begin{aligned} P(0,t)_{11} &= e^{-\lambda t} \\ P(0,t)_{12} &= \int_0^t e^{-\lambda(t-s)} e^{-\lambda s}\lambda\, ds = \lambda t e^{-\lambda t} \\ P(0,t)_{13} &= \int_0^t e^{-\lambda(t-s)}\lambda^2 s e^{-\lambda s}\, ds = \frac{(\lambda t)^2}{2!}e^{-\lambda t} \\ P(0,t)_{1n} &= \frac{(\lambda t)^n}{n!}e^{-\lambda t} \end{aligned}$$

which is clearly the Poisson distribution associated with the Poisson process.

One nice property of the birth-death process is that it is straightforward to compute the steady-state distribution π, since the matrix Q is tri-diagonal! Thus, starting with the first equation, one obtains:

$$\begin{aligned} \mu_2\pi_2 - \lambda_1\pi_1 &= 0 \\ \lambda_1\pi_1 + \mu_3\pi_3 - (\mu_2 + \lambda_2)\pi_2 &= 0 \\ \lambda_2\pi_2 + \mu_4\pi_4 - (\mu_3 + \lambda_3)\pi_3 &= 0 \\ \vdots \end{aligned} \qquad (13.18)$$

Solving recursively, one obtains:

$$\begin{aligned} \mu_2\pi_2 &= \lambda_1\pi_1 \\ \mu_3\pi_3 &= \lambda_2\pi_2 \\ \mu_4\pi_4 &= \lambda_3\pi_3 \\ \vdots \end{aligned} \qquad (13.19)$$

The above equations imply that the "probability flow" is balanced between each pair of states at steady state (this is known as *detailed balance*). Thus, we can express all of the probabilities in terms of the first value π_1, to obtain

$$\pi_n = \left(\prod_{j=2}^{n} \frac{\lambda_{j-1}}{\mu_j} \right) \pi_1 \tag{13.20}$$

Since the probabilities must add up to 1, one gets the following expression for π_1:

$$\pi_1 = \frac{1}{1 + \sum_{i=2}^{\infty} \left(\prod_{j=2}^{n} \frac{\lambda_{j-1}}{\mu_j} \right)} \tag{13.21}$$

In some birth-death processes, the birth rate is zero after a certain value, so that there are only a finite number of possible states. These are special cases where the above summation is easy to compute. Another special case is considered in a later section on queuing systems, where the birth rates λ_i and the death rates μ_i are independent of i, except for $\mu_1 = 0$.

13.4   Queuing Systems

A queuing system is an example of a birth-death process, where arrivals to a queue are modeled as Poisson processes, and departures from a queue are also modeled as a separate independent Poisson process. The simplest case of a queuing system is the M/M/1 (the notation stands for Markov arrivals, Markov departures, 1 server): in this system, arrival of customers to a queue is modeled by a Poisson process $N(t)$ with constant arrival rate λ. Thus, it is assumed that, with probability 1, there is at most one arrival at a particular time t. As discussed in the previous section, this corresponds to a birth process, with constant transition rate λ, where the states are now numbered 0, 1, ... to correspond to the number of customers in a queue. In addition to arrivals, there is a separate independent Poisson process $D(t)$ representing the departure process, which has constant rate μ *as long as there are customers in the queue*, and rate 0 otherwise. Again, with probability 1, there is a maximum of one departure at each time. Thus, we see that the M/M/1 queue is a birth-death process with parameters

$$\lambda_i = \lambda, i = 0, 1, \ldots; \quad \mu_0 = 0, \ \mu_i = \mu, i = 1, 2, \ldots \tag{13.22}$$

Using formulas (13.20,13.21), we obtain the steady state distribution for the M/M/1 queue:

$$\pi_n = \left(\prod_{j=1}^{n} \frac{\lambda}{\mu} \right) \pi_0, \quad i = 1, 2, \ldots \tag{13.23}$$

$$\pi_0 = \frac{1}{1 + \sum_{i=1}^{\infty} \left(\prod_{j=1}^{n} \frac{\lambda}{\mu} \right)} = 1 - \frac{\lambda}{\mu} \tag{13.24}$$

The factor $\frac{\lambda}{\mu} = \rho$ is called the *utilization factor*. In terms of this factor, the steady-state distribution becomes

$$\pi_n = \rho^n (1 - \rho) = \frac{\lambda}{\mu - \lambda}, \quad n = 0, 1, \ldots \tag{13.25}$$

Note that the utilization ρ must be less than 1 for a steady-state distribution to exist. Using this distribution, we can compute several key properties of M/M/1 queues, including the expected number of customers in the system, as well as the average waiting time. The expected number of customers is given by:

$$E[N] = \sum_{n=0}^{\infty} n\pi_n = \sum_{n=0}^{\infty} n\rho^n (1 - \rho) = \frac{\rho}{1 - \rho} \tag{13.26}$$

Using this relationship, we compute the expected waiting time of a newly-arrived customer to begin service; this requires that the existing average number of customers be served. Thus, the waiting time for service is computed as the product of the average number of customers in the system times the average departure time per customer. The average departure time per customer is given by $1/\mu$, using the exponential distribution of the interarrival time of Poisson processes. Thus, the average waiting time is

$$E[W] = E[N]\frac{1}{\mu} = \frac{\rho}{\mu - \lambda} \tag{13.27}$$

Thus, the larger the utilization, the longer the waiting time. A final relation is the average amount of time to finish service, which is

$$E[T] = \frac{1}{\mu} + E[W] = \frac{1}{\mu - \lambda}$$

Similar formulas can be developed for queues with more than one server. We provide some examples below.

Example 13.3

Consider the following discrete-space problem: Baybanks has one ATM (automatic teller machine). Customers arrive to use the machine randomly, as a Poisson process, at the rate of 10 customers/hour, and perform one transaction each. Assume that the duration of each transaction is an exponential random variable, independent from transaction to transaction, and independent of the arrival process. The average transaction duration is 5 minutes, so the service rate is 12 transactions/hour, as long as there are enough customers.

1. To formulate this model as an M/M/1 queueing model, we can draw the state transition diagram as in Figure 13.1:

2. Assume that the process is in steady state. What is the probability that there is no one using the ATM when you arrive? We can compute this as the steady state probability $\pi_0 = 1 - \frac{\lambda}{\mu}$, where μ is the departure rate of 12, and λ is the arrival rate of 10, so it is 1/6.

3. What is the steady state probability that at least 5 people are waiting in line, including the one being served, when you arrive? The answer is given by

$$1 - \pi_0 - \pi_1 - \pi_2 - \pi_3 - \pi_4 = 1 - 1/6\left(1 + 5/6 + (5/6)^2 + (5/6)^3 + (5/6)^4\right) = 1 - 1/6\frac{1 - (5/6)^5}{1 - 5/6} = (5/6)^5$$

4. If you get there and there are 3 people ahead of you, how long do you expect to wait until you begin to use the ATM? This is a simple question, since you have to wait until the persons in front of you are served. Thus, the waiting time is 3 times 5 = 15 minutes, since the durations are independent.

5. What is the expected wating time for a new arrival? Using the queuing formula, it becomes $\frac{1/6}{12-10} = 1/12$ hour, or 5 minutes. In essence, the expected number in the queue is 1 customer.
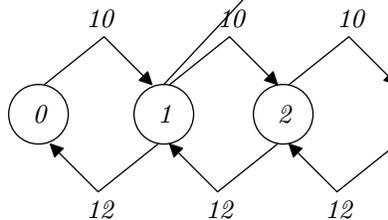
Figure 13.1: Diagram for example

Example 13.4

Consider the following problem: There is a bank with 2 tellers. Customers arrive at the bank and join a **single** line (queue...); whenever a teller is free, the next customer in line will go to that teller and be served. Assume that the arrival process of customers to the bank is modeled as a Poisson process, with arrival rate 8 customers/hour. Assume that the service time of each teller for each customer is independent, exponentially distributed, and with average service time of 1/8

hour, so that the service rate of each teller is 8 customers per hour. Note that there is zero probabillity that two customers arrive simultaneously, or that both tellers finish simultaneously, so this is a birth-death process! Note also that, when both tellers are busy, the service rate is 16 customers/hour, whereas when only one teller is busy, the service rate is only 8 per hour.

1. First, we formulate a continuous time, discrete space Markov chain model of the above problem, and draw the state transition diagram, showing the transition rates. Note that the service rate when both tellers are busy (e.g. $n > 1$ customers in the bank) is 16/hour, whereas the service rate when $n = 0$ is zero, and $n = 1$ is 8/hour. This will be a standard diagram for birth-death process, with arrival rates always 8. However, when there are no customers in the queue, the departure is 0; when there is only 1 customer in the queue, the departure rate will be that of only one server, which is 8. As long as there are 2 or more customers, both servers will be busy and the departure rate will be 16.

2. Next, let's solve for the steady-state probability of this birth-death process by writing an expression for π_n in terms of π_1,and then writing an expression relating π_0 and π_1. Note that conservation of probability flow at state zero yields the equation: $\pi_0 = \pi_1$; the birth-death conservation equations (detailed balance) for every other node yield $16\pi_{n+1} = 8\pi_n$. Thus, for $n > 0$, we have $\pi_{n+1} = 1/2\pi_n = (1/2)^n\pi_1$. Conservation of probability yields:

$$\pi_0 + \pi_1 + (1/2)\pi_1 + (1/2)^2\pi_1 + \ldots = 1 = \pi_1\left(1 + \sum_{n=0}^{\infty}(1/2)^n\right) = 3\pi_1$$

Thus, $\pi_0 = \pi_1 = 1/3, \pi_{n+1} = (1/2)^n 1/3$.

3. Now, let's use the formula $\sum_{n=1}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2}$ to obtain the expected number of customers in the bank at steady state (including those being served...).

$$E[n] = \sum_{n=1}^{\infty} n\pi_n = \sum_{n=1}^{\infty} \pi_1 n(1/2)^{n-1} = 4/3$$

4. Now, consider the same bank, with only a single teller with rate 8 customers/hour, but with half the arrival rate (that is, arrival rate of 4 customers/hour). The expected number of customers in the bank at steady state is given by a standard queue, with utilization $\rho = 1/2$. Thus, the expected number of customers in the bank in steady state is $\frac{\rho}{1-\rho} = 1$.

5. Given the answers to (c) and (d) above, is it better to have 2 small banks with one teller each, splitting the arrival traffic in half, or to have one larger bank with two tellers sharing a single queue? Clearly, it is better to have one large bank, since the average number of customers is 4/3, whereas for two banks, the average number of customers is $1 + 1 = 2$. This is why you see those very long single lines at airports or at banks.

13.5   Inhomogeneous Poisson Processes

In this section, we extend the concept of Poisson processes to processes where the process rate can depend on time. Recall that the construction of the Poisson process began with the definition of a set of independent, identically distributed, exponentially-distributed random variables which represented the interarrival times of the Poisson process. This resulted in an independent increments process, with the property that the event of a jump in any interval $(t, t + \Delta]$ was independent from that of a jump in any other disjoint interval $(t_1, t_1 + \Delta]$, and the probability of this event was defined in terms of the jump rate as:

$$P[N(t + \Delta) - N(t) = k] = \begin{cases} 1 - \lambda\Delta + o(\Delta) & \text{if } k = 0 \\ \lambda\Delta & \text{if } k = 1 \\ o(\Delta) & \text{if } k > 1 \end{cases} \qquad (13.28)$$

In order to generalize this construction, we want to maintain this independent increments property, but to allow the instantaneous jump rate to be time-dependent.

In essence, our construction of an inhomogeneous Poisson process is based on letting the instantaneous jump rate be a time-dependent function $\lambda(t)$; thus, we have

$$P[N(t + \Delta) - N(t) = k] = \begin{cases} 1 - \lambda(t)\Delta + o(\Delta) & \text{if } k = 0 \\ \lambda(t)\Delta & \text{if } k = 1 \\ o(\Delta) & \text{if } k > 1 \end{cases} \qquad (13.29)$$

where the notation $o(\Delta)$ is used to denote a term such that $\lim_{\Delta \to 0} \frac{o(\Delta)}{\Delta} = 0$. Consider, therefore, the probability density associated with the first jump time T_1. By a limiting argument as $\Delta \to 0$, we obtain

$$
\begin{aligned}
P[T_1 \leq T] &= \lim_{\Delta \to 0} \left(1 - \prod_{k=0}^{T/\Delta} 1 - \lambda(k\Delta)\Delta + o(\Delta) \right) \\
&= 1 - \lim_{\Delta \to 0} \left(\prod_{k=0}^{T/\Delta} e^{-\lambda(k\Delta)\Delta + o(\Delta)} \right) \\
&= 1 - \lim_{\Delta \to 0} e^{-\sum_{k=0}^{T/\Delta} \lambda(k\Delta)\Delta} \\
&= 1 - e^{-\int_0^T \lambda(t)dt}
\end{aligned}
\tag{13.30}
$$

Hence, the probability density is given by

$$
p(T) = \frac{d}{dT} P[T_1 \leq T] = \lambda(T) e^{-\int_0^T \lambda(t)dt}
\tag{13.31}
$$

Given T_1, we can construct the conditional probability density of T_2 in an identical manner, using the independence of the probability that jumps occur in disjoint intervals, as:

$$
p(T_2 = t \mid T_1) = \begin{cases} \lambda(t) e^{-\int_{T_1}^t \lambda(s)\,ds} & \text{for } t \geq T_1 \\ 0 & \text{otherwise} \end{cases}
\tag{13.32}
$$

Generalizing, we can obtain the following formula for the conditional density of T_n, the n-th jump time:

$$
p(T_n = t \mid T_1, \ldots, T_{n-1}) = p(T_n = t \mid T_{n-1}) = \begin{cases} \lambda(t) e^{-\int_{T_{n-1}}^t \lambda(s)\,ds} & \text{for } t \geq T_{n-1} \\ 0 & \text{otherwise} \end{cases}
\tag{13.33}
$$

Combining the above equations, we obtain an expression for the joint probability density of $T_1, T_2, \ldots, T_n$:

$$
\begin{aligned}
p(T_1 = t_1, \ldots, T_n = t_n) &= p(T_1 = t_1) \prod_{k=2}^n p(T_k = t_k \mid T_1 = t_1, \ldots, T_{k-1} = t_{k-1}) \\
&= p(T_1 = t_1) \prod_{k=2}^n p(T_k = t_k \mid T_{k-1} = t_{k-1}) \\
&= \begin{cases} \prod_{k=1}^n \lambda(t_k) e^{-\int_{t_{k-1}}^{t_k} \lambda(s)ds} & \text{if } t_0 = 0 \leq t_1 \leq \ldots \leq t_n \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} (\prod_{k=1}^n \lambda(t_k)) e^{-\int_0^{t_n} \lambda(s)ds} & \text{if } t_0 = 0 \leq t_1 \leq \ldots \leq t_n \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
\tag{13.34}
$$

Furthermore, from this probability, we can compute the conditional probability that $T_{k+1} > T$, given the value of T_k, as

$$
P[T_{k+1} > T | T_k] = e^{-\int_{T_k}^T \lambda(s)ds}
\tag{13.35}
$$

Note that this implies that $N(T) = k$. Thus, we can compute the joint probability density of $T_1, \ldots, T_k$ and $N(T) = k$ as

$$
p(T_1 = t_1, \ldots, T_n = t_n, N(T) = n) = \begin{cases} (\prod_{k=1}^n \lambda(t_k)) e^{-\int_0^T \lambda(s)ds} & \text{if } t_0 = 0 \leq t_1 \leq \ldots \leq t_n \leq T \\ 0 & \text{otherwise} \end{cases}
$$

As a final note, we need to compute the probability distribution of $N(t)$, and of the increment $N(t) - N(s)$. We can do this from the previous equation, as:

$$
\begin{aligned}
P(N(t) = n) &= \int_0^t \int 0^{t_n} \cdots \int_0^{t_1} \left(\prod_{k=1}^n \lambda(t_k) dt_k \right) e^{-\int_0^T \lambda(s)\, ds} \\
&= e^{-\int_0^T \lambda(s) ds} \frac{\left(\int_0^t \lambda(s) ds \right)^n}{n!}
\end{aligned}
\tag{13.36}
$$

To see how this last formula comes about, define the auxiliary function $F(t) = \int_0^t \lambda(s) ds$. Then, note the following identity:

$$
\begin{aligned}
\int_0^{t_2} \lambda(t_1) \int_0^{t_1} \lambda(t_0) dt_0 dt_1 &= \int_0^{t_2} F(t_1) \lambda(t_1) dt_1 \\
&= \int_0^{t_2} F(t_1) dF(t_1) = F^2(t_2) - F^2(0) = F^2(t_2).
\end{aligned}
\tag{13.37}
$$

Proceeding with the various integrals in the same fashion yields the result.

Note that we can perform a similar computation for the probability that $N(t) - N(s) = k$, to obtain

$$
P(N(t) - N(s) = k) = e^{-\int_s^t \lambda(\sigma) d\sigma} \frac{\left(\int_s^t \lambda(\sigma) d\sigma \right)^k}{k!}
$$

One of the important properties of inhomogeneous Poisson processes is that they are birth-death processes. In particular, the process $x(t) = N(t) + 1$ is a birth-death process with birth rate $\lambda(t)$ and death rate 0 for all states.

Other properties of inhomogeneous Poisson processes are:

1. The mean of the process, $m_N(t)$, is given by

$$
m_N(t) = \int_0^t \lambda(s) ds \,, t \geq 0.
$$

2. The covariance of the process, $K_N(t, s)$ is given by

$$
K_N(t, s) = \int_0^{\min(t,s)} \lambda(\sigma) d\sigma, \ t, s \geq 0.
$$

3. Consider the conditional probability density of the jump times $T_1, \ldots, T_k$, given the information that $N(t) = k$. This is given by:

$$
\begin{aligned}
p(T_1, \ldots, T_n | N(t) = k) &= \frac{p(T_1, \ldots, T_n, N(t) = k)}{P(N(t) = n)} \\
&= \frac{\left(\prod_{k=1}^n \lambda(T_k) \right) e^{-\int_0^t \lambda(s)\, ds}}{e^{-\int_0^t \lambda(s) ds} \frac{\left(\int_0^t \lambda(s)\, ds \right)^n}{n!}} \\
&= n! \frac{\left(\prod_{k=1}^n \lambda(T_k) \right)}{\left(\int_0^t \lambda(s)\, ds \right)^n} = n! \prod_{k=1}^n \frac{\lambda(T_k)}{\int_0^t \lambda(s)\, ds}
\end{aligned}
\tag{13.38}
$$

4. Now, consider the above ordered sequence of times $T_1, \ldots, T_n$, and apply a random permutation to obtain the unordered times $U_1, \ldots, U_n$. Assume that each random permutation is equally likely, with probability $\frac{1}{n!}$. Then,

$$
p(U_1, \ldots, U_n | N(t) = n) = \prod_{k=1}^n \frac{\lambda(U_k)}{\int_0^t \lambda(s) ds}
$$

since the sum over all permutations must equal the original probability $p(T_1, \ldots, T_n | N(t) = k)$. The unusual fact is that, conditioned on knowing that there are only n transitions up to time t, the unordered event times $U_1, \ldots, U_n$ are conditionally independent and identically distributed with conditional probability density

$$p(U_k \mid N(t) = n) = \begin{cases} \frac{\lambda(U_k)}{\int_0^t \lambda(s)\, ds} & \text{if } i \leq n \\ 0 & \text{otherwise} \end{cases}$$

5. Let $N(t)$ be a homogeneous Poisson process, so that $\lambda(t) \equiv \lambda$. Let T_k denote the time of the k-th jump in the Poisson process. Then, the random variables $\tau_k = T_k - T_{k-1}$ are exponential, independent, identically distributed with rate λ.

13.6 Applications of Poisson Processes

In this subsection, we describe various applications which can be analyzed using the properties of Poisson processes discussed previously.

Example 13.5
At a customer facility, customers arrive at a rate of 3 customers per hour, randomly distributed according to a Poisson process with constant rate 3. Assume that the doors open at 9:00 am. What is the expected time until the arrival of the 10-th customer? What is the probability that, if the doors close at 10:00 am for 15 minutes for a coffee break, one or more customers will arrive during the break? Suppose that, instead of taking a break at 10:00, the store waits until a customer arrives after 10:00 am (and is served instantaneously) before taking a break; what will the probability be that no customers arrive during the break? In order to answer this, all of the above questions can be posed in the context of a Poisson process $N(t)$ with homogeneous rate $\lambda = 3$. The first question is computing $E[T_{10}]$; since T_{10} is the sum of 10 independent, identically distributed random variables, each of which has mean $1/3$, then $E[T_{10}] = 10/3$. The probability that any customers arrive from 10:00 to 10:15 is given by

$$1 - P(N(t + 1/4) - N(t) = 0) = 1 - e^{-0.75}$$

Since the process is homogeneous, and has independent increments, the above probability is also the probability of any arrivals over any $1/4$ hour interval, no matter whether the interval started after a previous arrival. Thus, the probability of no arrivals is $e^{-0.75}$.

Example 13.6
Consider an extension of the previous problem, where, at the end of a visit, a customer tips the store a random amount. Let y_k denote the tip left by the k-th customer, where the sequence y_k is independent, identically distributed with common density function $p(y)$, and is independent of the arrival Poisson process. Define the process $x(t)$ to be the amount of tips received up to time t. What are the mean $m_x(t)$ and variance $\sigma_x(t)^2$? To solve this problem, note that

$$x(t) = \begin{cases} 0 & \text{if } N(t) = 0 \\ \sum_{k=1}^{N(t)} y_k & \text{otherwise} \end{cases}$$

To compute the mean of $x(t)$, we use the smoothing property of expectations, as follows:

$$E[x(t)] = E[E[x(t) \mid N(t) = n]] = E\left[E\left[\sum_{k=1}^{n} y_k \mid N(t) = n\right]\right] = E[N(t)]\, m_y = 3t m_y \qquad (13.39)$$

where we used the independent, identically distributed property of y_k, and the independence of y_k and $N(t)$. Similarly,

$$\begin{aligned} E[x(t)^2] &= E\left[E\left[\left(\sum_{k=1}^{n} y_k\right)^2 \mid N(t) = n\right]\right] \\ &= E\left[E\left[\sum_{k=1}^{n}\sum_{j=1}^{n} y_k y_j \mid N(t) = n\right]\right] \\ &= E\left[N(t)E\left[y_k^2\right] + N(t)(N(t) - 1)m_y^2 \mid N(t) = n\right] \\ &= 3t E\left[y_k^2\right] + m_y^2\left(E\left[N(t)^2\right] - 3t\right) \\ &= 3t E[y_k^2] + m_y^2(9t^2 + 3t - 3t) = 3t E[y_k^2] + (3t m_y)^2 \qquad (13.40) \end{aligned}$$

so that the variance is given by

$$\sigma_x(t)^2 = 3tE[y^2] = 3tm_y^2 + 3t\sigma_y^2$$

Note: The above process $x(t)$ is called a compound Poisson counting process, because the sample functions jump at Poisson times, but with a random amplitude. In general, a compound Poisson process has the following properties:

1. It is an independent increments process.

2. The characteristic function of each increment is given by

$$\Phi_{x(t)-x(s)}(jw) = e^{\int_s^t \lambda(\tau)d\tau(\Phi_y(jw)-1)}$$

where $\Phi_y(jw)$ is the characteristic function of the jump size.

3. The mean is given by

$$m_x(t) = m_y \int_0^t \lambda(\tau)d\tau \text{ for } t \geq 0$$

4. The autocovariance function is given by

$$K_x(t,s) = \left(m_y^2 + \sigma_y^2\right) \int_0^{\min(t,s)} \lambda(\tau)\, d\tau \text{ for } t, s \geq 0$$

Appendix A

Useful Transforms

The Fourier transform and inverse Fourier transform of aperiodic signals are defined in Table A.1. Notice that the discrete-time Fourier transform (DTFT) is periodic with period 2π in radian frequency or period 1 in cycles/sec. For periodic signals, we have the Fourier series and transform relationships shown in Table A.2. Note that the discrete-time Fourier series coefficients are periodic with period N, and thus $X(e^{j\omega})$ is also periodic. In Table A.6 we give useful Laplace transform pairs. In Table A.7 we give z-transform pairs.

	Continuous-Time	Discrete-Time
FT (radians)	$X(j\omega) = \mathcal{F}\left[x(t)\right] = \int_{-\infty}^{\infty} x(t)\, e^{-j\omega t}\, dt$	$X(e^{j\omega}) = \mathcal{F}\left[x(n)\right] = \sum_{n=-\infty}^{\infty} x(n)\, e^{-j\omega n}$
FT (Hertz)	$X(j2\pi f) = \mathcal{F}\left[x(t)\right] = \int_{-\infty}^{\infty} x(t)\, e^{-j2\pi ft}\, dt$	$X(e^{j2\pi f}) = \mathcal{F}\left[x(n)\right] = \sum_{n=-\infty}^{\infty} x(n)\, e^{-j2\pi fn}$
Inverse FT (radians)	$x(t) = \mathcal{F}^{-1}\left[X(j\omega)\right] = \dfrac{1}{2\pi} \int_{-\infty}^{\infty} X(j\omega)\, e^{j\omega t}\, d\omega$	$x(n) = \mathcal{F}^{-1}\left[X(e^{j\omega})\right] = \dfrac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega)\, e^{j\omega n}\, d\omega$
Inverse FT (Hertz)	$x(t) = \mathcal{F}^{-1}\left[X(j2\pi f)\right] = \int_{-\infty}^{\infty} X(f)\, e^{j2\pi ft}\, df$	$x(n) = \mathcal{F}^{-1}\left[X(e^{j2\pi f})\right] = \int_{-1/2}^{1/2} X(e^{j2\pi f})\, e^{j2\pi fn}\, df$

Table A.1: Fourier transform and inverse Fourier transform definitions.

In Table A.3 we summarize some useful Fourier transform properties for both the continuous- and discrete-time cases. For compactness, we slightly bend our notation and represent either the continuous or discrete transform in Hertz as $X(f)$ and the transform in radians as $X(\omega)$. In Table A.4 we present useful continuous-time Fourier transform pairs, while in Table A.5 we present useful discrete-time Fourier transform pairs.

	Continuous-Time Period T, $f_0 = \frac{1}{T}$, $\omega_0 = \frac{2\pi}{T}$	Discrete-Time Period N, $f_0 = \frac{1}{N}$, $\omega_0 = \frac{2\pi}{N}$
Fourier Series Coefficients (radians)	$a_k = \dfrac{1}{T} \displaystyle\int_{t_0}^{t_0+T} x(t)\, e^{-j\omega_0 kt}\, dt$	$a_k = \dfrac{1}{N} \displaystyle\sum_{n=n_0+1}^{n_0+N} x(n)\, e^{-j\omega_0 kt}$
Fourier Series Coefficients (Hertz)	$a_k = \dfrac{1}{T} \displaystyle\int_{t_0}^{t_0+T} x(t)\, e^{-j2\pi k f_0 t}\, dt$	$a_k = \dfrac{1}{N} \displaystyle\sum_{n=n_0+1}^{n_0+N} x(n)\, e^{-j2\pi k f_0 n}$
Fourier Series Representation (radians)	$x(t) = \displaystyle\sum_{k=-\infty}^{\infty} a_k\, e^{j\omega_0 kt}$	$x(n) = \displaystyle\sum_{k=k_0+1}^{k_0+N} a_k\, e^{j\omega_0 kn}$
Fourier Series Representation (Hertz)	$x(t) = \displaystyle\sum_{k=-\infty}^{\infty} a_k\, e^{j2\pi f_0 kt}$	$x(n) = \displaystyle\sum_{k=k_0+1}^{k_0+N} a_k\, e^{j2\pi f_0 kn}$
Fourier Transform (radians)	$X(j\omega) = 2\pi \displaystyle\sum_{k=-\infty}^{\infty} a_k\, \delta(\omega - k\omega_0)$	$X(e^{j\omega}) = 2\pi \displaystyle\sum_{k=-\infty}^{\infty} a_k\, \delta(\omega - k\omega_0)$
Fourier Transform (Hertz)	$X(j2\pi f) = \displaystyle\sum_{k=-\infty}^{\infty} a_k\, \delta(f - k f_0)$	$X(e^{j2\pi f}) = \displaystyle\sum_{k=-\infty}^{\infty} a_k\, \delta(f - k f_0)$

Table A.2: Discrete-time Fourier series and transform relationships.

	$x(t)$ or $x(n)$	$X(f)$	$X(\omega)$
Linearity	$ax(t)+by(t)$	$aX(f)+bY(f)$	$aX(\omega)+bY(\omega)$
Time Shift	$x(t-t_0)$	$e^{-j2\pi ft_0}X(f)$	$e^{-j\omega t_0}X(\omega)$
Modulation	$e^{j2\pi f_0 t}x(t)$	$X(f-f_0)$	$X(\omega-2\pi f_0)$
Modulation (alt.)	$e^{j\omega_0 t}x(t)$	$X(f-\frac{\omega_0}{2\pi})$	$X(\omega-\omega_0)$
Time reversal	$x(-t)$	$X(-f)$	$X(-\omega)$
Conjugate	$x(t)^*$	$X^*(-f)$	$X^*(-\omega)$
Convolution	$x(t)*h(t)$	$X(f)H(f)$	$X(\omega)H(\omega)$
Real functions	$x(t)=x^*(t)$	$X(f)=X^*(-f)$	$X(\omega)=X^*(-\omega)$
	$x(t)$ Only	$X(f)$	$X(\omega)$
Multiplication	$x(t)y(t)$	$X(f)*Y(f)$	$\frac{1}{2\pi}X(\omega)*Y(\omega)$
Time derivative	$\frac{d}{dt}x(t)$	$j2\pi fX(f)$	$j\omega X(\omega)$
Freq derivative	$tx(t)$	$\frac{j}{2\pi}\frac{d}{df}X(f)$	$j\frac{d}{d\omega}X(\omega)$
Scaling	$x(at)$	$\frac{1}{|a|}X(\frac{f}{a})$	$\frac{1}{|a|}X(\frac{\omega}{a})$
Zero value	$\int_{-\infty}^{\infty}x(t)\,dt=X(0)$	$\int_{-\infty}^{\infty}X(f)\,df=x(0)$	$\frac{1}{2\pi}\int_{-\infty}^{\infty}X(\omega)\,d\omega=x(0)$
Parseval's Thm	$\int_{-\infty}^{\infty}|x(t)|^2\,dt$	$=\int_{-\infty}^{\infty}|X(f)|^2\,df$	$=\frac{1}{2\pi}\int_{-\infty}^{\infty}|X(\omega)|^2\,d\omega$
	$x(n)$ Only	$X(f)$	$X(\omega)$
Multiplication	$x(n)y(n)$	$\int_{f_0}^{f_0+1}X(v)Y(f-v)dv$	$\frac{1}{2\pi}\int_{\omega_0}^{\omega_0+2\pi}X(v)Y(\omega-v)dv$
Freq derivative	$nx(n)$	$\frac{j}{2\pi}\frac{d}{df}X(f)$	$j\frac{d}{d\omega}X(\omega)$
Zero value	$X(0)=\sum_{n=-\infty}^{\infty}x(n)$	$\int_{-1/2}^{1/2}X(f)\,df=x(0)$	$\frac{1}{2\pi}\int_{-\pi}^{\pi}X(\omega)\,d\omega=x(0)$
Parseval's Thm	$\sum_{n=-\infty}^{\infty}|x(n)|^2$	$=\int_{-1/2}^{1/2}|X(f)|^2\,df$	$=\frac{1}{2\pi}\int_{-\pi}^{\pi}|X(\omega)|^2\,d\omega$

Table A.3: Fourier Transform Properties.

$x(t)$	$X(f)$	$X(\omega)$
$\delta(t)$	1	1
$\delta(t-t_0)$	$e^{-j2\pi ft_0}$	$e^{-j\omega t_0}$
1	$\delta(f)$	$2\pi\delta(\omega)$
$\cos(2\pi f_0 t)$	$\frac{1}{2}\delta(f-f_0)+\frac{1}{2}\delta(f+f_0)$	$\pi\delta(\omega-2\pi f_0)+\pi\delta(\omega+2\pi f_0)$
$\cos(\omega_0 t)$	$\frac{1}{2}\delta(f-\omega_0/2\pi)+\frac{1}{2}\delta(f+\omega_0/2\pi)$	$\pi\delta(\omega-\omega_0)+\pi\delta(\omega+\omega_0)$
$e^{-\alpha t}u(t),\ \alpha>0$	$\frac{1}{\alpha+j2\pi f}$	$\frac{1}{\alpha+j\omega}$
$e^{-\alpha|t|},\ \alpha>0$	$\frac{2\alpha}{\alpha^2+(2\pi f)^2}$	$\frac{2\alpha}{\alpha^2+\omega^2}$
$te^{-\alpha t}u(t),\ \alpha>0$	$\frac{1}{(\alpha+j2\pi f)^2}$	$\frac{1}{(\alpha+j\omega)^2}$
$|t|e^{-\alpha|t|},\ \alpha>0$	$\frac{2[\alpha^2-(2\pi f)^2]}{[\alpha^2+(2\pi f)^2]^2}$	$\frac{2[\alpha^2-\omega^2]}{[\alpha^2+\omega^2]^2}$
$e^{-\pi t^2}$	$e^{-\pi f^2}$	$e^{-\omega^2/4\pi}$
Box: 1 for $t\in[-T,T]$	$2T\frac{\sin(2\pi fT)}{2\pi fT}$	$2T\frac{\sin(\omega T)}{\omega T}$
$2f_c\frac{\sin(2f_c t)}{2f_c t}$	Box: 1 for $f\in[-f_c,f_c]$	Box: 1 for $\omega\in[-2\pi f_c,2\pi f_c]$
$\frac{\omega_c}{\pi}\frac{\sin(\omega_c t/\pi)}{\omega_c t/\pi}$	Box: 1 for $f\in[-\omega_c/2\pi,\omega_c/2\pi]$	Box: 1 for $\omega\in[-\omega_c,\omega_c]$
Triangle: $1-\frac{|t|}{2T},\ t\in[-2T,\,2T]$	$2T\frac{\sin^2(2\pi fT)}{(2\pi fT)^2}$	$2T\frac{\sin^2(\omega T)}{(\omega T)^2}$
$\sum_m\delta(t-mT)$	$\frac{1}{T}\sum_k\delta(f-k/T)$	$\frac{2\pi}{T}\sum_k\delta(\omega-2\pi k/T)$

Table A.4: Useful Continuous-Time Fourier Transform Pairs

$x(n)$	$X(f)$	$X(\omega)$
$\delta(n)$	1	1
$\delta(n - n_0)$	$e^{-j2\pi f n_0}$	$e^{-j\omega n_0}$
1	$\sum_k \delta(f + k)$	$2\pi \sum_k \delta(\omega + 2\pi k)$
$e^{j\omega_0 n}$	$\sum_k \delta(f - \omega_0/2\pi + k)$	$2\pi \sum_k \delta(\omega - \omega_0 + 2\pi k)$
$e^{j2\pi f_0 n}$	$\sum_k \delta(f - f_0 + k)$	$2\pi \sum_k \delta(\omega - 2\pi f_0 + 2\pi k)$
$\cos(\omega_0 n + \phi)$	$\frac{1}{2} \sum_k [e^{j\phi}\delta(f - \omega_0/2\pi + k)$	$\pi \sum_k [e^{j\phi}\delta(\omega - \omega_0 + 2\pi k)$
	$\quad + e^{-j\phi}\delta(f + \omega_0/2\pi + k)]$	$\quad + e^{-j\phi}\delta(\omega + \omega_0 + 2\pi k)]$
$\cos(2\pi f_0 n + \phi)$	$\frac{1}{2} \sum_k [e^{j\phi}\delta(f - f_0 + k)$	$\pi \sum_k [e^{j\phi}\delta(\omega - 2\pi f_0 + 2\pi k)$
	$\quad + e^{-j\phi}\delta(f + f_0 + k)]$	$\quad + e^{-j\phi}\delta(\omega + 2\pi f_0 + 2\pi k)]$
$a^n u(n),\ |a| < 1$	$\frac{1}{1-ae^{-j2\pi f}}$	$\frac{1}{1-ae^{-j\omega}}$
$(n+1)a^n u(n),\ |a| < 1$	$\frac{1}{(1-ae^{-j2\pi f})^2}$	$\frac{1}{(1-ae^{-j\omega})^2}$
$a^{|n|},\ |a| < 1$	$\frac{1-a^2}{1+a^2-2a\cos(2\pi f)}$	$\frac{1-a^2}{1+a^2-2a\cos(\omega)}$
Box: 1 for $n \in [-N, N]$	$\frac{\sin(\pi f(2N+1))}{\sin(\pi f)}$	$\frac{\sin(\omega(2N+1)/2)}{\sin(\omega/2)}$
$\frac{\sin(Wn)}{\pi n};\ W \in [0,\pi]$	Periodic (1): 1 for $f \in [-W/2\pi, W/2\pi]$	Periodic (2π): 1 for $\omega \in [-W, W]$
$\sum_m \delta(n - mN)$	$\frac{1}{N} \sum_k \delta(f - k/N)$	$\frac{2\pi}{N} \sum_k \delta(\omega - 2\pi k/N)$

Table A.5: Useful Discrete-Time Fourier Transform Pairs

$x(t)$	$X(s)$	ROC
$\delta(t)$	1	All s
$u(t)$	$\frac{1}{s}$	$\mathrm{Re}(s) > 0$
$-u(-t)$	$\frac{1}{s}$	$\mathrm{Re}(s) < 0$
$e^{-\alpha t}u(t)$	$\frac{1}{s + \alpha}$	$\mathrm{Re}(s) > -\alpha$
$-e^{-\alpha t}u(-t)$	$\frac{1}{s + \alpha}$	$\mathrm{Re}(s) < -\alpha$
$e^{-\alpha |t|},\ \alpha > 0$	$\frac{2\alpha}{(\alpha^2 - s^2)}$	$|\mathrm{Re}(s)| < \alpha$
$\frac{t^{n-1}}{(n-1)!}e^{-\alpha t}u(t)$	$\frac{1}{(s+\alpha)^n}$	$\mathrm{Re}(s) > -\alpha$
$-\frac{t^{n-1}}{(n-1)!}e^{-\alpha t}u(-t)$	$\frac{1}{(s+\alpha)^n}$	$\mathrm{Re}(s) < -\alpha$
$\delta(t - T)$	e^{-sT}	All s

Table A.6: Useful Laplace Transform Pairs

$f(k)$	$F(z)$	ROC
$\delta(k)$	1	All z
$u(k)$	$(1 - z^{-1})$	$1 < |z|$
$ku(k)$	$\dfrac{z^{-1}}{(1 - z^{-1})^2}$	$1 < |z|$
$k^n u(k)$	$\left(-z\dfrac{d}{dz}\right)^n \dfrac{1}{(1 - z^{-1})}$	$1 < |z|$
$\begin{pmatrix} k \\ n \end{pmatrix}, \quad n \le k$	$\dfrac{z^{-n}}{(1 - z^{-1})^{n+1}}$	$0 < |z|$
$\begin{pmatrix} n \\ k \end{pmatrix}, \quad 0 \le k \le n$	$(1 + z^{-1})^n$	$0 < |z|$
$\alpha^k u(k)$	$\dfrac{1}{(1 - \alpha z^{-1})}$	$|\alpha| < |z|$
$k^n \alpha^k u(k)$	$\left(-z\dfrac{d}{dz}\right)^n \dfrac{1}{(1 - \alpha z^{-1})}$	$|\alpha| < |z|$
$\alpha^k u(-k - 1)$	$\dfrac{-1}{(1 - \alpha z^{-1})}$	$|z| < |\alpha|$
$k^n \alpha^k u(-k - 1)$	$-\left(-z\dfrac{d}{dz}\right)^n \dfrac{1}{(1 - \alpha z^{-1})}$	$|z| < |\alpha|$
$\alpha^{|k|}$	$\dfrac{1 - \alpha^2}{(1 - \alpha z)(1 - \alpha z^{-1})}$	$|\alpha| < |z| < \left|\dfrac{1}{\alpha}\right|$
$\dfrac{1}{k}u(k - 1)$	$-\ln(1 - z^{-1})$	$1 < |z|$
$\cos(\alpha k)u(k)$	$\dfrac{1 - \cos(\alpha)z^{-1}}{1 - 2\cos(\alpha)z^{-1} + z^{-2}}$	$1 < |z|$
$\sin(\alpha k)u(k)$	$\dfrac{\sin(\alpha)z^{-1}}{1 - 2\cos(\alpha)z^{-1} + z^{-2}}$	$1 < |z|$
$(a\cos(\alpha k) + b\sin(\alpha k))\,u(k)$	$\dfrac{a + (b\sin(\alpha) - a\cos(\alpha))z^{-1}}{1 - 2\cos(\alpha)z^{-1} + z^{-2}}$	$1 < |z|$
$\cosh(\alpha k)u(k)$	$\dfrac{1 - \cosh(\alpha)z^{-1}}{1 - 2\cosh(\alpha)z^{-1} + z^{-2}}$	$\max\{|\alpha|, |1/\alpha|\} < |z|$
$\sinh(\alpha k)u(k)$	$\dfrac{\sinh(\alpha)z^{-1}}{1 - 2\cosh(\alpha)z^{-1} + z^{-2}}$	$\max\{|\alpha|, |1/\alpha|\} < |z|$

Table A.7: Useful Z-Transform Pairs

Appendix B

Partial-Fraction Expansions

In this appendix we examine the tool of partial-fraction expansions. Partial-fraction expansions provide a way of representing a rational polynomial or transform as a *sum* of simpler terms. We first treat the continuous-time problem, then the discrete-time one.

B.1 Continuous-Time Signals

Consider the problem of inverting rational transforms, i.e. those of the form:

$$X(s) = \frac{a_m s^m + a_{m-1} s^{m-1} + \cdots a_1 s + a_0}{s^n + d_{n-1} s^{n-1} + \cdots + d_1 s + d_0} \tag{B.1}$$

If $m \geq n$ we can use long division to reduce $X(s)$ to the sum of a polynomial in s and a strictly proper rational function, as follows:

$$X(s) = c_{m-n} s^{m-n} + c_{m-n-1} s^{m-n-1} + \cdots + c_1 s + c_0 + X_p(s) \tag{B.2}$$

where $X_p(s)$ is a proper rational function of s:

$$X_p(s) = \frac{\alpha_{n-1} s^{n-1} + \alpha_{n-2} s^{n-2} + \cdots \alpha_1 s + \alpha_0}{s^n + d_{n-1} s^{n-1} + \cdots + d_1 s + d_0} \tag{B.3}$$

Thus, the inverse transform $x(t)$ of $X(p)$ is given by:

$$x(t) = c_{m-n} u_{m-n}(t) + c_{m-n-1} u_{m-n-1}(t) + \cdots + c_1 u_1(t) + c_0 u_0(t) + x_p(t) \tag{B.4}$$

where $x_p(t)$ is the inverse transform of $X_p(s)$ and $u_k(t)$ represents the generalized function of order k – i.e. $u_k(t) = \frac{d^k}{dt^k} \delta(t)$. The above is nothing more than statement of the fact that we can always write a rational transform as the sum of a polynomial part – which is easy to invert – and a strictly proper part, whose inverse we discuss next.

To find $x_p(t)$ we may use <u>partial fraction expansion</u>, which allows us to write the $X_p(s)$ as the sum of a number of simpler, single-pole components. This assumes that $X_p(s)$ is a strictly proper rational function. Suppose there are r distinct poles or roots to the denominator p_i, each of muliplicity k_i, and that the denominator is factored as:

$$s^n + d_{n-1} s^{n-1} + \cdots + d_1 s + d_0 = (s - p_1)^{k_1} (s - p_2)^{k_2} \cdots (s - p_r)^{k_r} \tag{B.5}$$

Then $X_p(s)$ can always be rewritten as a partial fraction expansion as follows:

$$X_p(s) \quad = \quad \frac{A_{11}}{(s - p_1)} + \frac{A_{12}}{(s - p_1)^2} + \cdots + \frac{A_{1k_1}}{(s - p_1)^{k_1}} \tag{B.6}$$

$$+ \frac{A_{21}}{(s - p_2)} + \frac{A_{22}}{(s - p_2)^2} + \cdots + \frac{A_{2k_2}}{(s - p_2)^{k_2}}$$

$$\vdots$$

$$+ \frac{A_{r1}}{(s - p_r)} + \frac{A_{r2}}{(s - p_r)^2} + \cdots + \frac{A_{rk_r}}{(s - p_r)^{k_r}}$$

$$= \quad \sum_{i=1}^{r} \sum_{j=1}^{k_i} \frac{A_{ij}}{(s - p_i)^j}$$

The coefficients A_{ij} can be obtained by equating the two expressions (B.6) and (B.3), clearing the denominators and matching like powers of s. Alternatively, a closed form expression for the coefficients is given by:

$$A_{ij} = \frac{1}{(k_i - j)!} \left[\frac{d^{k_i - j}}{ds^{k_i - j}} (s - p_i)^{k_i} X_p(s) \right] \bigg|_{s = p_i} \tag{B.7}$$

The inverse transform of (B.6) can then be obtained on a term-by-term basis, since we have split it into simpler terms.

Example B.1
Suppose the signal transform is given by:

$$X(s) = \frac{s + 2}{(s + 1)^2 (s + 3)} \tag{B.8}$$

The transform is already strictly proper so no long division is needed in this case. The partial fraction expansion is given by:

$$X(s) = \frac{A_{11}}{s + 1} + \frac{A_{12}}{(s + 1)^2} + \frac{A_{21}}{s + 3} \tag{B.9}$$

The coefficients are given by:

$$A_{11} \quad = \quad \frac{1}{(2 - 1)!} \frac{d}{ds} \left[(s + 1)^2 X(s) \right]\big|_{s = -1} = \frac{1}{4} \tag{B.10}$$

$$A_{12} \quad = \quad \left[(s + 1)^2 X(s) \right]\big|_{s = -1} = \frac{1}{2} \tag{B.11}$$

$$A_{21} \quad = \quad \left[(s + 3) X(s) \right]\big|_{s = -3} = -\frac{1}{4} \tag{B.12}$$

Therefore, we have that:

$$X(s) = \frac{s + 2}{(s + 1)^2 (s + 3)} = \frac{\frac{1}{4}}{s + 1} + \frac{\frac{1}{2}}{(s + 1)^2} - \frac{\frac{1}{4}}{s + 3} \tag{B.13}$$

Taking inverse transforms, we find:

$$x(t) = \left[\frac{1}{4} e^{-t} + \frac{1}{2} t e^{-t} - \frac{1}{4} e^{-3t} \right] u_{-1}(t) \tag{B.14}$$

B.2   Discrete-Time Signals

We now turn our attention to the problem of inverting rational z-transforms, i.e. those of the form:

$$X(z) = \frac{a_m z^m + a_{m-1} z^{m-1} + \cdots a_1 z + a_0}{z^q + d_{q-1} z^{n-1} + \cdots + d_1 z + d_0} \tag{B.15}$$

As in continuous time, if $m \geq q$ we can use long division to reduce $X(z)$ to the sum of a polynomial in z and a strictly proper rational function of z, as follows:

$$X(z) = c_{m-q} z^{m-q} + c_{m-q-1} z^{m-q-1} + \cdots + c_1 z + c_0 + X_p(z) \tag{B.16}$$

where $X_p(z)$ is a proper rational function of z:

$$X_p(z) = \frac{\alpha_{q-1}z^{q-1} + \alpha_{q-2}z^{q-2} + \cdots \alpha_1 z + \alpha_0}{z^q + d_{q-1}z^{q-1} + \cdots + d_1 z + d_0} \tag{B.17}$$

Thus, the inverse transform $x(n)$ of $X(z)$ is given by:

$$x(n) = c_{m-q}\delta(n + m - q) + c_{m-q-1}\delta(n + m - q - 1) + \cdots + c_1\delta(n+1) + c_0\delta(n) + x_p(n) \tag{B.18}$$

where $x_p(n)$ is the inverse transform of $X_p(z)$. Notice that in the discrete case, the higher order generalized functions of the continuous case become positive time shifts.

Now to find $x_p(n)$ we may again use partial fraction expansion. This assumes that $X_p(z)$ is a strictly proper rational function, which it is by design. Again, suppose there are r distinct poles or roots to the denominator p_i, each of muliplicity k_i, and that the denominator is factored as:

$$z^q + d_{q-1}z^{q-1} + \cdots + d_1 z + d_0 = (z - p_1)^{k_1}(z - p_2)^{k_2} \cdots (z - p_r)^{k_r} \tag{B.19}$$

Then $X_p(z)$ can always be rewritten as a partial fraction expansion as follows:

$$\begin{aligned} X_p(z) &= \frac{A_{11}}{(z - p_1)} + \frac{A_{12}}{(z - p_1)^2} + \cdots + \frac{A_{1k_1}}{(z - p_1)^{k_1}} \\ &+ \frac{A_{21}}{(z - p_2)} + \frac{A_{22}}{(z - p_2)^2} + \cdots + \frac{A_{2k_2}}{(z - p_2)^{k_2}} \\ &\qquad\qquad \vdots \\ &+ \frac{A_{r1}}{(z - p_r)} + \frac{A_{r2}}{(z - p_r)^2} + \cdots + \frac{A_{rk_r}}{(z - p_r)^{k_r}} \\ &= \sum_{i=1}^{r}\sum_{j=1}^{k_i} \frac{A_{ij}}{(z - p_i)^j} \end{aligned} \tag{B.20}$$

As in the continuous case, the coefficients A_{ij} can be obtained either by equating the two expressions (B.20) and (B.17) clearing the denominators and matching like powers of z or by using the closed form expression for the coefficients given before:

$$A_{ij} = \frac{1}{(k_i - j)!} \left[\frac{d^{k_i-j}}{dz^{k_i-j}}(z - p_i)^{k_i} X_p(z) \right]\Bigg|_{z=p_i} \tag{B.21}$$

The inverse transform of (B.20) can then be obtained on a term-by-term basis, since we have split it into simpler terms. The only additional tricky part is that discrete-time transform expressions are often given in terms of negative powers of z. For example, suppose $m > q$:

$$X(z) = \frac{a_m + a_{m-1}z^{-1} + \cdots a_1 z^{1-m} + a_0 z^{-m}}{z^{q-m} + d_{q-m-1}z^{n-1} + \cdots + d_1 z^{1-m} + d_0 z^{-m}} \tag{B.22}$$

Clearly such expressions can be converted to the form in (B.15), and then the results above applied. Alternatively, the change of variables $\nu = z^{-1}$ can be made, the expansion done, then the variable changed back. We will illustrate both approaches below.

Example B.2
Suppose the signal transform is given by:

$$X(z) = \frac{2}{1 - \frac{3}{4}z^{-1} + \frac{1}{8}z^{-2}} = \frac{2z^2}{z^2 - \frac{3}{4}z^1 + \frac{1}{8}} = \frac{2z^2}{\left(z - \frac{1}{2}\right)\left(z - \frac{1}{4}\right)} \tag{B.23}$$

The partial fraction expansion of the term in brackets is given by:

$$X(z) = z^2 \left[\frac{A_{11}}{z - \frac{1}{2}} + \frac{A_{21}}{z - \frac{1}{4}} \right] \tag{B.24}$$

The coefficients are given by:

$$A_{11} = \left[\left(z - \frac{1}{2}\right)\frac{X(z)}{z^2}\right]\Bigg|_{z=\frac{1}{2}} = \frac{2}{\frac{1}{2} - \frac{1}{4}} = 8 \tag{B.25}$$

$$A_{21} = \left[\left(z - \frac{1}{4}\right)\frac{X(z)}{z^2}\right]\Bigg|_{z=\frac{1}{4}} = \frac{2}{\frac{1}{4} - \frac{1}{2}} = -8 \tag{B.26}$$

Therefore, we have that:

$$X(z) = z^2\left[\frac{8}{z - \frac{1}{2}} - \frac{8}{z - \frac{1}{4}}\right] = z^2\left[\frac{8}{z - \frac{1}{2}} - \frac{8}{z - \frac{1}{4}}\right] = z8\left[\frac{1}{1 - \frac{1}{2}z^{-1}} - \frac{1}{1 - \frac{1}{4}z^{-1}}\right] \tag{B.27}$$

Taking inverse transforms, we find:

$$x(n) = 8\left[\left(\frac{1}{2}\right)^{n+1}u_{-1}(n+1) - \left(\frac{1}{4}\right)^{n+1}u_{-1}(n+1)\right] = 4\left(\frac{1}{2}\right)^n u_{-1}(n) - 2\left(\frac{1}{4}\right)^n u_{-1}(n) \tag{B.28}$$

Now the alternative way to solve this is to make the substitution $\nu = z^{-1}$ at the outset:

$$X(\nu) = \frac{2}{1 - \frac{3}{4}\nu + \frac{1}{8}\nu^2} = \frac{2}{\left(1 - \frac{1}{2}\nu\right)\left(1 - \frac{1}{4}\nu\right)} = \frac{2}{\left(1 - \frac{1}{2}z^{-1}\right)\left(1 - \frac{1}{4}z^{-1}\right)} \tag{B.29}$$

The partial fraction expansion is given by:

$$X(\nu) = \frac{B_{11}}{1 - \frac{1}{2}\nu} + \frac{B_{21}}{1 - \frac{1}{4}\nu} = \frac{B_{11}}{1 - \frac{1}{2}z^{-1}} + \frac{B_{21}}{1 - \frac{1}{4}z^{-1}} \tag{B.30}$$

The coefficients are given by:

$$B_{11} = \left[\left(1 - \frac{1}{2}\nu\right)X(\nu)\right]\Bigg|_{\nu=2} = \frac{2}{1 - \frac{1}{2}} = 4 \tag{B.31}$$

$$B_{21} = \left[\left(1 - \frac{1}{4}\nu\right)X(\nu)\right]\Bigg|_{\nu=4} = \frac{2}{1 - 2} = -2 \tag{B.32}$$

Therefore, we have that:

$$X(\nu) = \frac{4}{1 - \frac{1}{2}\nu} - \frac{2}{1 - \frac{1}{4}\nu} \tag{B.33}$$

Equivalently, making the inverse change of variables $z^{-1} = \nu$:

$$X(z) = \frac{4}{1 - \frac{1}{2}z^{-1}} - \frac{2}{1 - \frac{1}{4}z^{-1}} \tag{B.34}$$

Taking inverse transforms, we find:

$$x(n) = 4\left(\frac{1}{2}\right)^n u_{-1}(n) - 2\left(\frac{1}{4}\right)^n u_{-1}(n) \tag{B.35}$$

as before.

Appendix C

Summary of Linear Algebra

Linear algebra is concerned with the solution of sets of simultaneous systems of linear equations. The linear nature of these sets of equations leads naturally to both a convenient notation and a deep connections with the properties of vectors and matrices. These notes are intended to provide a summary and review of the important concepts and notation that arise.

C.1   Vectors and Matrices

A vector is just an array of numbers stacked together:

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{C.1}$$

where $x_1, x_2, \ldots, x_n$ may be either real or complex numbers. We often denote such column vectors by underlined lowercase letters, as shown in (C.1). The set of all n-dimensional vectors of real numbers is usually denoted by R^n while the set of all n-dimensional vectors of complex numbers is denoted by C^n. The *transpose* of a column vector $\underline{x}$ is a row vector:

$$\underline{x}^T = [x_1, x_2, \ldots, x_n] \tag{C.2}$$

The sum of two vectors is defined on a component-by-component basis

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + x_2 \\ \vdots \\ x_n + y_n \end{bmatrix} \tag{C.3}$$

Similarly, the product of a vector and a scalar is defined componentwise as:

$$\alpha \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{bmatrix} \tag{C.4}$$

where α is a real or complex number.

A set of vectors $\{\underline{x}_1, \ldots, \underline{x}_r\}$ in R^n or C^n is termed *linearly independent* if and only if

$$\alpha_1 \underline{x}_1 + \alpha_2 \underline{x}_2 + \ldots + \alpha_r \underline{x}_r + = \underline{0} \tag{C.5}$$

implies that

$$\alpha_1 = \alpha_2 = \ldots = \alpha_r = 0 \tag{C.6}$$

where $\underline{0}$ is the n-vector of zeros. Otherwise one of the x_i can be written as a *linear combination* of the others and the set of vectors is termed *linearly dependent*. In R^n we can have at most n linearly independent vectors in any given set. Conversely, given any set of n linearly independent vectors $\{\underline{x}_1, \ldots, \underline{x}_n\}$ in R^n, any other vector can be written as a linear combination of the vectors $\underline{x}_1, \ldots, \underline{x}_n$. Any such a set is termed a *basis* for R^n.

Given two vectors of the same length, $x, y \in R^n$, we can define the *dot* or *inner product* between the vectors:

$$\underline{x}^T \underline{y} = \langle x, y \rangle = \sum_{i=1}^{n} x_i y_i = \underline{y}^T \underline{x} \quad \in R \tag{C.7}$$

Two n-vectors $\underline{x}$ and $\underline{y}$ are termed *orthogonal*, denoted $\underline{x} \perp \underline{y}$ if

$$\underline{x}^T \underline{y} = 0 \tag{C.8}$$

A set of nonzero, mutually orthogonal vectors is always linearly independent. The *length* or standard *norm* of the vector $\underline{x} \in R^n$ is

$$||\underline{x}|| = \sqrt{\underline{x}^T \underline{x}} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{C.9}$$

Note that the inner product provides information about the angle between the vectors. In particular, $\underline{x}^T \underline{y} = ||\underline{x}|| \, ||\underline{y}|| \cos(\angle(x, y))$.

As in the case of vectors, *matrices* are simply arrays of numbers in a regular grid:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{C.10}$$

where $a_{11}, \ldots, a_{mn}$ may be either real or complex numbers. We can see that vectors are matrices with a special form – they only have a single column or row. Matrices are often denoted by capital letters. The element in the i-th row and j-th column of A will be denoted by a_{ij}, $[A]_{ij}$, or $(A)_{ij}$, depending on the situation. If A has m rows and n columns we say that A is an $m \times n$ matrix. The set of all $m \times n$ real-valued matrices is denoted $R^{m \times n}$ while the set of all $m \times n$ complex-valued matrices is denoted $C^{m \times n}$.

If $m = n$, A is a *square* matrix. The *transpose* of an $m \times n$ matrix A is the $m \times n$ matrix whose elements are a_{ji}, i.e. that is rows are exchanged for coluns and vice versa. With A defined as in (C.10) we have:

$$A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix} \tag{C.11}$$

A square matrix is said to be *symmetric* if $A^T = A$. A *diagonal* square matrix only has nonzero entries along its diagonal and is of the form

$$A = \begin{bmatrix} a_1 & 0 & 0 & \cdots & 0 \\ 0 & a_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_n \end{bmatrix} \tag{C.12}$$

This is sometimes denoted as $\mathsf{diag}(a_1, \ldots, a_n)$. The *identity matrix* is denoted by I and is the diagonal matrix with ones along its diagonal:

$$I = \mathsf{diag}(1, \ldots, 1) \tag{C.13}$$

On occasion we will write I_n to make clear the size (i.e. $n \times n$) of the identity matrix. The *trace* of a square matrix A is the sum of its diagonal elements:

$$\text{tr}(A) = \sum_{i=1}^{n} a_{ii} \tag{C.14}$$

In particular, we have for a square matrix A that $\text{tr}(A) = \text{tr}(A^T)$. Note that

$$||\underline{x}||^2 = \underline{x}^T \underline{x} = \text{tr}\left(\underline{x}^T \underline{x}\right) = \text{tr}\left(\underline{x}\underline{x}^T\right) \tag{C.15}$$

We now consider operations involving matrices. As with vectors we define addition between matrices and multiplication of a matrix by a scalar on a component by component basis:

$$A + B = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix} \tag{C.16}$$

$$= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{bmatrix} \tag{C.17}$$

$$\alpha A = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \cdots & \alpha a_{1n} \\ \alpha a_{21} & \alpha a_{22} & \cdots & \alpha a_{2n} \\ \vdots & \vdots & & \vdots \\ \alpha a_{m1} & \alpha a_{m2} & \cdots & \alpha a_{mn} \end{bmatrix} \tag{C.18}$$

Let A be an $m \times n$ matrix and B an $n \times p$ matrix. Then the *matrix product* of A and B is denoted by $C = AB$ where C is an $m \times p$ matrix whose elements are given by

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \tag{C.19}$$

Note that A and B must satisfy some dimensional constraints for the above expression to make any sense. In particular, the number of columns of A must equal the number of rows of B for AB to be defined. One implication is that BA may not be defined even if AB is (consider the case of $m = 2$, $n = 3$, $p = 4$). Note also, that even if BA is defined it may not be of the same size as AB. For example, if A is 2×3 and B is 3×2, then AB is 2×2, but BA is 3×3. In general, $AB \neq BA$ so the *order* of matrix multiplication is very important. Some other important relationships are:

$$AI = IA = A \tag{C.20}$$

so the identity matrix is the identity element of matrix multiplication. It can be verified that the transpose operation behaves as follows:

$$(AB)^T = B^T A^T \tag{C.21}$$

Also, if $A \in R^{m \times n}$ and $\underline{x} \in R^n$ then $A\underline{x} \in R^m$. In addition, if both AB and BA are defined,

$$\text{tr}(AB) = \text{tr}(BA) \tag{C.22}$$

Further, note that the tr is a linear operation, so that:

$$\text{tr}(A + B + C) = \text{tr}(A) + \text{tr}(B) + \text{tr}(C) \tag{C.23}$$

Let $\underline{x} \in R^n$ and $\underline{y} \in R^m$. Then the *dyadic* or *outer product* of the vectors $\underline{x}$ and $\underline{y}$ is the $n \times m$ matrix

$$\underline{x}\underline{y}^T = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_m \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_m \\ \vdots & \vdots & & \vdots \\ x_n y_1 & x_n y_2 & \cdots & x_n y_m \end{bmatrix} \tag{C.24}$$

On occasion we will find it useful to deal with matrices written in block form

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array} \right] \tag{C.25}$$

where A_{11} is $m_1 \times n_1$, A_{12} is $m_1 \times n_2$, A_{21} is $m_2 \times n_1$, A_{22} is $m_2 \times n_2$. The product of two matrices in block form is computed in a manner analogous to usual matrix multiplication, only the block become the basic elements. For example

$$\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array} \right] \left[\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array} \right] = \left[\begin{array}{c|c} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ \hline A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{array} \right] \tag{C.26}$$

where the blocks on the left-hand side must be partitioned in a compatible fashion, and the order of multiplication of the various terms on the right-hand side is important.

C.2   Matrix Inverses and Determinants

An $n \times n$ matrix is *invertible* or *nonsingular* if the only solution of the equation $A\underline{x} = 0$ is $\underline{x} = 0$. That is, the only vector producing zero output is the zero vector. If this is the case, then there exists another $n \times n$ matrix A^{-1} called the *inverse* of A, so that

$$AA^{-1} = A^{-1}A = I \tag{C.27}$$

If no such matrix exists A is termed *non-invertible* or *singular*. The property of invertibility is related to the solution of sets of equations. To see this, consider the set of equations

$$A\underline{x} = \underline{y} \tag{C.28}$$

where A is $n \times n$. This equation has a unique solution $\underline{x}$ for any $\underline{y}$ if and only if A is invertible (in which case the solution is $A^{-1}\underline{y}$). Conversely, if A is singular, then there exists a non-zero vector $\underline{x}_N$ such that $A\underline{x}_N = 0$. In this case, we can add any multiple of $\underline{x}_N$ to a solution of (C.28) and produce another solution. Thus if A is singular, the system of equations will not have a unique solution.

The *determinant* of a square matrix A, denoted by $|A|$ or $\det(A)$, can be defined recursively. If A is 1×1, then $|A| = A$. If A is $n \times n$, then we can compute $|A|$ by "expanding by minors" using any row or column. For example, using the i-th row:

$$|A| = a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{in}A_{in} \tag{C.29}$$

or using the j-th column

$$|A| = a_{1j}A_{1j} + a_{2j}A_{2j} + \cdots + a_{nj}A_{nj} \tag{C.30}$$

where

$$A_{ij} = (-l)^{i+j} \det(M_{ij}) \tag{C.31}$$

where M_{ij} is the $(n-1) \times (n-1)$ matrix obtained from A by deleting the i-th row and j-th column. For example

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} \tag{C.32}$$

As a more complex example we compute

$$
\begin{vmatrix}
2 & 0 & 0 & 3 \\
1 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 \\
5 & 1 & 1 & 9
\end{vmatrix}
$$

$$
= \; 2(-1)^{1+1}\begin{vmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 9 \end{vmatrix} + 0(-1)^{1+2}\begin{vmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 5 & 1 & 9 \end{vmatrix} + 0(-1)^{1+3}\begin{vmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 5 & 1 & 9 \end{vmatrix} + 3(-1)^{1+4}\begin{vmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 5 & 1 & 1 \end{vmatrix}
$$

$$
= \; 2(-1)^{1+1}\begin{vmatrix} 1 & 0 \\ 1 & 9 \end{vmatrix} - 3(-1)^{1+1}\begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} - 3(-1)^{1+2}\begin{vmatrix} 1 & 1 \\ 5 & 1 \end{vmatrix}
$$

$$
= \; 2\cdot 9 - 3\cdot 0 + 3\cdot(-4) = 6 \tag{C.33}
$$

Several properties of determinants are

$$
|AB| = |A||B| \tag{C.34}
$$

$$
|\alpha A| = \alpha^n |A| \tag{C.35}
$$

$$
\left|A^T\right| = |A| \tag{C.36}
$$

$$
\left|A^{-1}\right| = \frac{1}{|A|} \tag{C.37}
$$

The invertibility of a matrix A is equivalent to each of the following statements:

1. $|A| \neq 0$

2. All of the columns of A are linearly independent.

3. All of the rows of A are linearly independent.

The inverse of A can be expressed as

$$
A^{-1} = \frac{1}{|A|}C^T \tag{C.38}
$$

where $C_{ij} = A_{ij}$ as defined in (C.31). For example

$$
\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}}\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \tag{C.39}
$$

Some properties of inverses are

$$
\left(A^T\right)^{-1} = \left(A^{-1}\right)^T \tag{C.40}
$$

$$
(AB)^{-1} = B^{-1}A^{-1} \tag{C.41}
$$

$$
A = \mathsf{diag}\,(\mu_1,\ldots,\mu_n) \implies A^{-1} = \mathsf{diag}\left(\frac{1}{\mu_1},\cdots,\frac{1}{\mu_n}\right) \tag{C.42}
$$

A matrix P is *orthogonal* if

$$
P^{-1} = P^T \tag{C.43}
$$

If we think of P as consisting of a set of columns, i.e.

$$
P = \begin{bmatrix} \underline{x}_1 & | & \underline{x}_2 & | & \cdots & | & \underline{x}_n \end{bmatrix} \tag{C.44}
$$

then in general

$$
P^T P = \begin{bmatrix}
\underline{x}_1^T\underline{x}_1 & \underline{x}_1^T\underline{x}_2 & \cdots & \underline{x}_1^T\underline{x}_n \\
\underline{x}_2^T\underline{x}_1 & \underline{x}_2^T\underline{x}_2 & \cdots & \underline{x}_2^T\underline{x}_n \\
\vdots & \vdots & & \vdots \\
\underline{x}_n^T\underline{x}_1 & \underline{x}_n^T\underline{x}_2 & \cdots & \underline{x}_n^T\underline{x}_n
\end{bmatrix} \tag{C.45}
$$

Consequently, we see that P is orthogonal if and only if its columns are *orthogonal*, i.e. $\underline{x}_i \perp \underline{x}_j$, $i \neq j$, and $||\underline{x}_i|| = I$.

There are also some useful results for block matrices. For example, for a block diagonal matrix

$$A = \text{diag}\,(F_1, \ldots, F_r) \Rightarrow A^{-1} = \text{diag}\,\left(F_1^{-1}, \ldots, F_r^{-1}\right) \tag{C.46}$$

Also, the formulas

$$\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right]^{-1} =$$

$$\left[\begin{array}{c|c} \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} & -\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}A_{12}A_{22}^{-1} \\ \hline -A_{22}^{-1}A_{21}\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} & A_{22}^{-1} + A_{22}^{-1}A_{21}\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}A_{12}A_{22}^{-1} \end{array}\right] \tag{C.47}$$

$$\det\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] = \left|A_{11} - A_{12}A_{22}^{-1}A_{21}\right|\,|A_{22}| \tag{C.48}$$

which are valid if A_{22} is nonsingular, are verified by noting that

$$\left[\begin{array}{c|c} I & -A_{12}A_{22}^{-1} \\ \hline 0 & I \end{array}\right]\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right]\left[\begin{array}{c|c} I & 0 \\ \hline -A_{22}^{-1}A_{21} & I \end{array}\right] = \left[\begin{array}{c|c} A_{11} - A_{12}A_{22}^{-1}A_{21} & 0 \\ \hline 0 & A_{22} \end{array}\right] \tag{C.49}$$

Similarly, if A_{11} is nonsingular

$$\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right]^{-1} =$$

$$\left[\begin{array}{c|c} A_{11}^{-1} + A_{11}^{-1}A_{12}\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} \\ \hline -\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1}A_{21}A_{11}^{-1} & \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} \end{array}\right] \tag{C.50}$$

Comparison of the above yields the useful result

$$\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} = A_{11}^{-1} + A_{11}^{-1}A_{12}\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1}A_{21}A_{11}^{-1} \tag{C.51}$$

C.3 Eigenvalues and Eigenvectors

Let A be an $n \times n$ real matrix. A scalar λ is called an *eigenvalue* of A with associated nonzero *eigenvector* $\underline{x}$ if

$$A\underline{x} = \lambda\underline{x} \tag{C.52}$$

The above equation can be rewritten as

$$(\lambda I - A)\underline{x} = \underline{0} \tag{C.53}$$

Thus λ is an eigenvalue of A if and only if (C.53) has a solution $\underline{x} \neq \underline{0}$. This will be the case if and only if $\lambda I - A$ is singular, i.e. if and only if λ is a solution of the *characteristic equation*

$$p_A(\lambda) = |\lambda I - A| = 0 \tag{C.54}$$

Here $p_A(\lambda)$ is the *characteristic polynomial* of A and is of the form

$$p_A(\lambda) = \lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0 = (\lambda - \lambda_1)\cdots(\lambda - \lambda_n) \tag{C.55}$$

Here $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the n eigenvalues, which may or may not be distinct. Some of the λ_i may in general be complex, in which case they occur in complex conjugate pairs. However, if A is symmetric, the λ_i are always real. Also note that

$$|A| = (-1)^n p_A(0) = (-1)^n \alpha_0 = \lambda_1 \cdots \lambda_n \tag{C.56}$$

so that A is invertible if and only if all of the eigenvalues of A are nonzero. In addition one can show that

$$\operatorname{tr}(A) = -\alpha_{n-1} = \lambda_1 + \lambda_2 + \cdots + \lambda_n \tag{C.57}$$

If λ_i is an eigenvalue of A, then we can determine an associated eigenvector by solving the set of linear equations

$$A\underline{x} = \lambda_i \underline{x} \tag{C.58}$$

Note that if $\underline{x}$ is an eigenvector, so is $\alpha\underline{x}$ for any scalar α. Consequently, we can always adjust the length of the eigenvectors arbitrarily. Note that each distinct λ_i has a linearly independent $\underline{x}_i$ corresponding to it. If λ_i has multiplicity $k > 1$, i.e. if λ_i is a k-th order root of $p_A(\lambda)$, then there may be anywhere from 1 to k linearly independent eigenvectors associated with λ_i. If A is symmetric, however, there are *always* a full set of linearly independent eigenvectors. Furthermore, these eigenvectors can be taken to be *orthogonal* and in fact orthonormal.

C.4   Similarity Transformation

Let A be an $n \times n$ matrix, and let P be an invertible matrix of the same size. We can then define a *similarity transformation* of A

$$B = PAP^{-1} \tag{C.59}$$

We sometimes say that "B is similar to A". A similarity transformation corresponds essentially to a change of coordinates. Specifically, suppose

$$\underline{y} = A\underline{x} \tag{C.60}$$

and consider a change of coordinates

$$\underline{u} = P\underline{x}, \qquad \underline{v} = P\underline{y} \tag{C.61}$$

(so that each component of $\underline{u}$, for example, is a weighted sum of components of $\underline{x}$ and vice versa, since $\underline{x} = P^{-1}\underline{u}$). Then

$$\underline{v} = B\underline{u} \tag{C.62}$$

Note that

$$\begin{aligned} p_B(\lambda) &= |\lambda I - B| = |\lambda PP^{-1} - PAP^{-1}| = |P^{-1}(\lambda I - A)P| \\ &= |P^{-1}||\lambda I - A||P| = |\lambda I - A| = p_A(\lambda) \end{aligned} \tag{C.63}$$

so the eigenvalues of B and A are the same. Also

$$\operatorname{tr}(B) = \operatorname{tr}(PAP^{-1} = \operatorname{tr}(P^{-1}PA) = \operatorname{tr}(A) \tag{C.64}$$

Suppose that the $n \times n$ matrix A has a full set of linearly independent eigenvectors $\underline{x}_1, \ldots, \underline{x}_n$, so that

$$A\underline{x}_i = \lambda_i \underline{x}_i, \qquad i = 1, \ldots, n \tag{C.65}$$

The existence of such a complete set of eigenvectors is guaranteed, for example, if the λ_i are all distinct or if A is symmetric.

We can rewrite (C.65) as one equation

$$A\begin{bmatrix} \underline{x}_1 & | & \underline{x}_2 & | & \cdots & | & \underline{x}_n \end{bmatrix} = \begin{bmatrix} \underline{x}_1 & | & \underline{x}_2 & | & \cdots & | & \underline{x}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix} \tag{C.66}$$

Let

$$p^{-1} = \begin{bmatrix} \underline{x}_1 & | & \underline{x}_2 & | & \cdots & | & \underline{x}_n \end{bmatrix}$$

which is invertible, since the columns $\underline{x}_1, \ldots, \underline{x}_n$ are linearly independent. Then (C.66) implies that

$$PAP^{-1} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \tag{C.67}$$

Note that if A is symmetric we can choose the $\underline{x}_i$ to be orthonormal so that $P^{-1} = P^T$.

C.5 Positive-Definite Matrices

A symmetric square matrix A is *positive semidefinite*, written $A \geq 0$, if and only if

$$\underline{x}^T A x \geq 0 \tag{C.68}$$

for all vectors $\underline{x}$. This matrix A is *positive definite*, written $A > 0$, if

$$\underline{x}^T A \underline{x} > 0 \quad \text{for all } \underline{x} \tag{C.69}$$

It is not difficult to see that a positive semidefinite matrix is positive definite if and only if it is invertible.

Some basic facts about positive semidefinite matrices are the following:

(i) If $A \geq 0$ and $B \geq 0$, then $A + B > 0$, since

$$\underline{x}^T (A + B) \underline{x} = \underline{x}^T A \underline{x} + \underline{x}^T B \underline{x} \tag{C.70}$$

(ii) If either A or B in (i) is positive definite, then so is $A + B$. This again follows from (C.70).

(iii) If $A > 0$, then $A^{-1} > 0$, since

$$\underline{x}^T A^{-1} \underline{x} = (A^{-1}\underline{x})^T A (A^{-1}\underline{x}) > 0 \quad \text{if } \underline{x} \neq 0 \tag{C.71}$$

(iv) If $Q \geq 0$ then $F^T Q F \geq 0$ for *any* (not necessarily square) matrix for which $F^T Q F$ is defined. This follows from

$$\underline{x}^T (F^T Q F) \underline{x} = (F\underline{x})^T Q (F\underline{x}) \geq 0 \tag{C.72}$$

(v) If $Q > 0$ and F is invertible, $F^T Q F > 0$. This also follows from (C.72).

One test for positive definiteness is *Sylvester's Test*. Let

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{12} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{bmatrix} \tag{C.73}$$

Then A is positive semidefinite (positive definite) if and only if

$$\begin{aligned} a_{11} &\geq 0 \quad\quad (> 0) \\ \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix} &\geq 0 \quad\quad (> 0) \\ \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{vmatrix} &\geq 0 \quad\quad (> 0) \\ \text{etc.} \end{aligned} \tag{C.74}$$

Let $A = A^T$, and let P be the orthogonal matrix of eigenvectors so that

$$PAP^T = \mathsf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \tag{C.75}$$

Then

$$\underline{x}^T A \underline{x} = \underline{x}^T P^T \left(PAP^T\right) P\underline{x} = \lambda_1 z_1^2 + \lambda_2 z_2^2 + \cdots + \lambda_n z_n^2 \tag{C.76}$$

where

$$\underline{z} = P\underline{x} \tag{C.77}$$

and we have used (C.75). From this we can conclude that

- $A = A^T$ is positive semidefinite if and only if all its eigenvalues are nonnegative.

- $A = A^T$ is positive definite if and only if all its eigenvalues are strictly positive.

Note that we now also show that if $A \geq 0$ then A has a *square root matrix* F so that

$$A = F^T F \tag{C.78}$$

and specifically from (C.75) we see that we can take

$$F = \mathsf{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}\right) P \tag{C.79}$$

Note that F in (C.79) is invertible if and only if $A > 0$. Also note that the square root matrix as defined in (C.78) is far from unique. Specifically, let Q be any orthogonal matrix, and let

$$\widehat{F} = QF \tag{C.80}$$

Then

$$\widehat{F}^T \widehat{F} = F^T Q^T Q F = F^T I F = F^T F = A \tag{C.81}$$

C.6   Subspaces

A subset $S \subseteq R^n$ is a subspace if S is closed under vector addition and scalar multiplication. Examples of subspaces of R^2 are[1]

$$S_1 = \left\{ \begin{bmatrix} a \\ 0 \end{bmatrix} \mid a \in R \right\} \tag{C.82}$$

$$S_2 = \left\{ \begin{bmatrix} a \\ 2a \end{bmatrix} \mid a \in R \right\} \tag{C.83}$$

The *dimension* of a subspace equals the maximum number of vectors in S that can form a linearly independent set.

Let K be any subset of R^n. The *orthogonal complement* of K is defined as follows:

$$K^\perp = \left\{ \underline{x} \in R^n \mid \underline{x} \perp \underline{y} \; \forall \underline{y} \in K \right\} \tag{C.84}$$

$K^\perp$ is a subspace whether or not K is, since if $\underline{x}_1, \underline{x}_2 \in K^\perp$, $\underline{y} \in K$

$$(\underline{x}_1 + \underline{x}_2)^T \underline{y} = \underline{x}_1^T \underline{y} + \underline{x}_2^T \underline{y} = 0 \tag{C.85}$$

$$(\alpha \underline{x}_1)^T \underline{y} = \alpha \underline{x}_1^T \underline{y} = 0 \tag{C.86}$$

so $(\underline{x}_1 + \underline{x}_2) \in K^\perp$ and $\alpha \underline{x}_1 \in K^\perp$.

Let $\underline{d}$ be a single nonzero vector in R^n and consider $\{d\}^\perp$. This is a subspace of dimension $n-1$. For example, as illustrated in Figure C.6, when $n = 2$ the set of $\underline{x}$ such that $\underline{d}^T \underline{x} = 0$ is a line through the origin perpendicular to d. In 3-dimensions this set is a plane through the origin, again perpendicular to d. Note that the subspace $\{\underline{d}\}^\perp$ splits R^n into two *half-spaces*, one corresponding to those $\underline{x}$ for which $\underline{d}^T \underline{x} > 0$, the other to those $\underline{x}$ for which $\underline{d}^T \underline{x} < 0$.

---

[1] Here R denotes the set of real numbers.
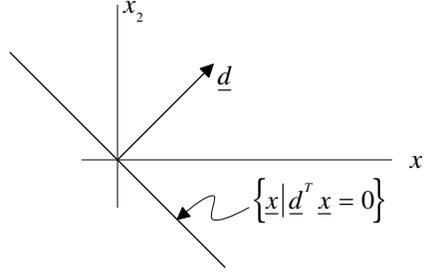
C.7   Vector Calculus

First consider a vector-valued function of a scalar real variable, denoted $\underline{f}(x)$. Calculus operations for functions of this type are defined component-wise, as follows:

$$\frac{d}{dx}\underline{f}(x) = \begin{bmatrix} \frac{d}{dx}f_1(x) \\ \frac{d}{dx}f_2(x) \\ \vdots \\ \frac{d}{dx}f_M(x) \end{bmatrix} \tag{C.87}$$

Conversely, now consider a scalar valued function of a vector argument, i.e. a function of n-real variables

$$f(\underline{x}) = f(x_1, \ldots, x_n) = f\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \tag{C.88}$$

Partial derivatives, integrals, etc., are defined in terms of the vector:

$$\frac{\partial f}{\partial \underline{x}}(\underline{x}) = \underline{f}_{\underline{x}}(\underline{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\underline{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\underline{x}) \end{bmatrix} \tag{C.89}$$

The second-order derivative can also be defined

$$\frac{\partial^2 f}{\partial x^2}(\underline{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\underline{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\underline{x}) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\underline{x}) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\underline{x}) \end{bmatrix} \tag{C.90}$$

Finally, let $f(x)$ be an $m \times 1$ vector-valued function of the $n \times 1$ vector variable x. We can define:

$$\frac{\partial \underline{f}}{\partial \underline{x}}(\underline{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\underline{x}) & \cdots & \frac{\partial f_m}{\partial x_1}(\underline{x}) \\ \vdots & & \vdots \\ \frac{\partial f_1}{\partial x_n}(\underline{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\underline{x}) \end{bmatrix} = f_{\underline{x}}(\underline{x}) \tag{C.91}$$

If $f(X)$ is a scalar function of the $n \times m$ matrix X, then the derivative of $f(X)$ with respect to X is given by:

$$\frac{\partial f}{\partial X} = \begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \cdots & \frac{\partial f}{\partial X_{1n}} \\ \vdots & & \vdots \\ \frac{\partial f}{\partial X_{m1}} & \cdots & \frac{\partial f}{\partial X_{mn}} \end{bmatrix} \tag{C.92}$$

If $F(X)$ is a $p \times q$ matrix and X is an $m \times n$ matrix, then the derivative of $F(X)$ with respect to X is given by:

$$\frac{\partial F}{\partial X} = \begin{bmatrix} \frac{\partial F}{\partial X_{11}} & \cdots & \frac{\partial F}{\partial X_{1n}} \\ \vdots & & \vdots \\ \frac{\partial F}{\partial X_{m1}} & \cdots & \frac{\partial F}{\partial X_{mn}} \end{bmatrix} \tag{C.93}$$

Examples of common calculations involving a vector $\underline{x}$ and matrices A, B, and X include:

$$\frac{\partial \underline{x}}{\partial \underline{x}} = I \tag{C.94}$$

$$\frac{\partial}{\partial \underline{x}} \left(A\underline{x} \right) = A^T \tag{C.95}$$

$$\frac{\partial}{\partial \underline{x}} \left(\underline{x}^T A \right) = A \tag{C.96}$$

$$\frac{\partial}{\partial \underline{x}} \left(\underline{x}^T A \underline{x} \right) = \left(A + A^T \right) \underline{x} \tag{C.97}$$

$$\frac{\partial^2}{\partial \underline{x}^2} \left(\underline{x}^T A \underline{x} \right) = A + A^T \tag{C.98}$$

$$\frac{\partial \mathrm{tr} \left(AX \right)}{\partial X} = A^T \tag{C.99}$$

$$\frac{\partial \mathrm{tr} \left(X^T A \right)}{\partial X} = A \tag{C.100}$$

$$\frac{\partial \mathrm{tr} \left(X^T A X B \right)}{\partial X} = AXB + A^T X B^T \tag{C.101}$$

Appendix D

The non-zero mean case

Here we consider LLSE of stochastic processes in the non-zero mean case and show that the approach of estimating mean subtracted process $\tilde{X}(t) = X(t) - m_x(t)$ based on the mean subtracted observation $\tilde{Y}(\tau) = Y(\tau) - m_y(\tau)$ really does produce the correct results. For simplicity we consider the vector case. Suppose we want to estimate the non-zero mean vector $\underline{X}$ based on the non-zero mean vector $\underline{Y}$. By assumption the estimate will be of the form:

$$\widehat{\underline{x}} = L\underline{y} + \underline{b} \tag{D.1}$$

and our task reduces to finding L and b. The solution is given by the orthogonality conditions:

$$E\left[\widehat{\underline{x}}\right] = E\left[\underline{x}\right] \tag{D.2}$$

$$E\left[(\underline{x} - \widehat{\underline{x}})\underline{y}^T\right] = 0 \tag{D.3}$$

Using the first equation we find that:

$$\underline{b} = m_x - Lm_y \tag{D.4}$$

i.e. that $\underline{b}$ only depends on the means and is zero for the zero-mean case. Further the form of the estimate can be now seen to be:

$$\widehat{\underline{x}} = m_x + L\left(\underline{y} - m_y\right) \tag{D.5}$$

Now applying the second orthogonality constraint, we know that $\underline{e} \perp \underline{y}$, where $\underline{e} = (\underline{x} - \widehat{\underline{x}})$. That is:

$$
\begin{aligned}
0 &= E\left[(\underline{x} - \widehat{\underline{x}})\underline{y}^T\right] \tag{D.6}\\
&= E\left[\{\underline{x} - m_x - L(\underline{y} - m_y)\}\{(\underline{y} - m_y) + m_y\}\right] \tag{D.7}\\
&= E\left[\{\underline{x} - m_x - L(\underline{y} - m_y)\}(\underline{y} - m_y)\right] + \underbrace{E\left[\{\underline{x} - m_x - L(\underline{y} - m_y)\}m_y\right]}_{=0} \tag{D.8}\\
&= K_{xy} - LK_{yy} \tag{D.9}\\
\implies L &= K_{xy}K_{yy}^{-1} \tag{D.10}
\end{aligned}
$$

Thus we have that

$$\widehat{\underline{x}} = m_x + K_{xy}K_{yy}^{-1}(\underline{y} - m_y) \tag{D.11}$$

Now lets find the estimation error covariance. This is given by:

$$
\begin{aligned}
E\left[\underline{e}\,\underline{e}^T\right] &= E\left[(\underline{x} - \widehat{\underline{x}})(\underline{x} - \widehat{\underline{x}})^T\right] = E\left[(\underline{x} - \widehat{\underline{x}})\{\underline{x} - m_x - L(\underline{y} - m_y)\}^T\right] \tag{D.12}\\
&= E\left[(\underline{x} - \widehat{\underline{x}})(\underline{x} - m_x)^T\right] - \underbrace{E\left[(\underline{x} - \widehat{\underline{x}})\underline{y}^T L^T\right]}_{=\,0 \text{ since } e\,\perp\,y} + \underbrace{E\left[(\underline{x} - \widehat{\underline{x}})m_y^T L^T\right]}_{=\,0 \text{ since } m_x\,=\,E[\widehat{\underline{x}}]} \tag{D.13}\\
&= E\left[\{(\underline{x} - m_x) - L(\underline{y} - m_y)\}(\underline{x} - m_x)^T\right] \tag{D.14}\\
&= K_{xx} - LK_{yx} = K_{xx} - K_{xy}K_{yy}^{-1}K_{xy}^T \tag{D.15}
\end{aligned}
$$

Note that this is the same result as for the zero mean case.

Finally let us compare these results to what we would find by estimating $\tilde{X}(t) = X(t) - m_x(t)$ based on $\tilde{Y}(\tau) = Y(\tau) - m_y(\tau)$ and substituting the definitions of $\tilde{X}(t)$ and $\tilde{Y}(\tau)$ in at the end. Note that:

$$\widehat{\underline{\tilde{x}}} = K_{\tilde{x}\tilde{y}} K_{\tilde{y}\tilde{y}}^{-1} \underline{\tilde{y}} \tag{D.16}$$

But

$$K_{\tilde{x}\tilde{y}} = K_{xy}, \quad K_{\tilde{y}\tilde{y}} = K_{yy}, \quad K_{\tilde{x}\tilde{x}} = K_{xx} \tag{D.17}$$

Thus

$$(\widehat{\underline{x}} - m_x) = K_{xy} K_{yy}^{-1} (\underline{y} - m_y) \tag{D.18}$$

Note that this is the same estimate we obtained by direct calculation. Thus indeed estimating $(\underline{x} - m_x)$ based on $(\underline{y} - m_y)$ then adding the means back in produces the same result. Finally, we can calculate the error covariance:

$$\Lambda_L = E[\tilde{e}\tilde{e}^T] = K_{\tilde{x}\tilde{x}} - K_{\tilde{x}\tilde{y}} K_{\tilde{y}\tilde{y}}^{-1} K_{\tilde{x}\tilde{y}} \tag{D.19}$$

$$= K_{xx} - K_{xy} K_{yy}^{-1} K_{xy}^T \tag{D.20}$$

Again this is the same result we obtained via direct calculation.

Index