Introduction to Random Processes and Applications

Don H. Johnson
Rice University

2009

Contents

Chapter 1

Probability

1.1 Foundations of Probability Theory

The basis of probability theory is a set of events — sample space — and a systematic set of numbers — probabilities — assigned to each event. What is an "event" and what kind of mathematical structure can collections of events — sets — have?

1.1.1 Mathematical Structure of Events

Letting A and B denote *events*, each of which consist of a collection of indecomposable elementary events ω_i. Events can be manipulated according to the union, intersection and complement operations.

$$\begin{aligned} A \bigcup B &= \{\omega : \omega \in A \text{ or } \omega \in B\} \text{ (union)} \\ A \bigcap B &= \{\omega : \omega \in A \text{ and } \omega \in B\} \text{ (intersection)} \\ \overline{A} &= \{\omega : \omega \notin A\} \text{ (complement)} \\ \overline{A \cup B} &= \overline{A} \cap \overline{B}. \end{aligned}$$

The null set $\emptyset$ is the complement of Ω, the universal set containing all events. "Indecomposable" or elementary events have nothing in common: $\omega_i \bigcap \omega_j = \emptyset$. Events, on the other hand, may share elements. Events are said to be *mutually exclusive* if there is no element common to both events: $A \bigcap B = \emptyset$.

For a collection of events $\mathscr{A}$ to be an *algebra*,

- $\emptyset \in \mathscr{A}$ and $\Omega \in \mathscr{A}$.
- If the events $A \in \mathscr{A}$ and $B \in \mathscr{A}$, then both the union and intersection of these events are in $\mathscr{A}$: $A \bigcup B \in \mathscr{A}$ and $A \bigcap B \in \mathscr{A}$. This property implies that all finite unions and intersections of events are also contained in the algebra.

$$\text{If } A_1, \ldots, A_N \in \mathscr{A},\ \bigcup_{i=1}^{N} A_n \in \mathscr{A} \quad \text{and} \ \bigcap_{i=1}^{N} A_n \in \mathscr{A}$$

We say that $\mathscr{F}$ is a *σ-algebra* if the algebra is closed under all countable intersections and unions. Note that this means that

$$\text{If } A_1, \ldots, \in \mathscr{F}, \quad \bigcup_{i=1}^{\infty} A_n \in \mathscr{F} \text{ and } \bigcap_{i=1}^{\infty} A_n \in \mathscr{F} \ .$$

In probability theory, a *sample space* is the set Ω of all possible elementary outcomes ω_i of an experiment, which can be collected into event sets.

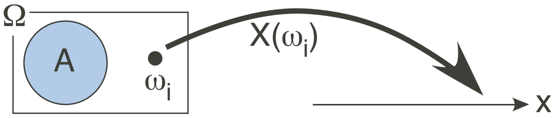


Figure 1.1: A random variable X is a function having a domain on the σ-algebra of events and a range lying somewhere on the real line. Random variables need not be one-to-one or onto.

1.1.2 The Probability of an Event

The key aspect of the theory is the system of assigning probabilities to events. Associated with each event A_i is a *probability measure* $\Pr[A_i]$, sometimes denoted by π_i, that obeys the *axioms of probability*.

- $\Pr[A_i] \geq 0$
- $\Pr[\Omega] = 1$
- If $A \bigcap B = \emptyset$, then $\Pr[A \bigcup B] = \Pr[A] + \Pr[B]$.

The consistent set of probabilities $\Pr[\cdot]$ assigned to events are known as the *a priori probabilities*. From the axioms, probability assignments for Boolean expressions can be computed. For example, simple Boolean manipulations ($A \bigcup B = A \bigcup (\overline{A}B)$ and $AB \bigcup \overline{A}B = B$) lead to

$$\Pr[A \bigcup B] = \Pr[A] + \Pr[B] - \Pr[A \bigcap B] .$$

Suppose $\Pr[B] \neq 0$. Suppose we know that the event B has occurred; what is the probability that event A also occurred? This calculation is known as the *conditional probability* of A given B and is denoted by $\Pr[A|B]$. To evaluate conditional probabilities, consider B to be the sample space rather than Ω. To obtain a probability assignment under these circumstances consistent with the axioms of probability, we must have

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} .$$

The event is said to be *statistically independent* of B if $\Pr[A|B] = \Pr[A]$: the occurrence of the event B does not change the probability that A occurred. When independent, the probability of their intersection $\Pr[A \cap B]$ is given by the product of the *a priori* probabilities $\Pr[A] \cdot \Pr[B]$. This property is necessary and sufficient for the independence of the two events. As $\Pr[A|B] = \Pr[A \cap B] / \Pr[B]$ and $\Pr[B|A] = \Pr[A \cap B] / \Pr[A]$, we obtain *Bayes' Rule*.

$$\Pr[B|A] = \frac{\Pr[A|B] \cdot \Pr[B]}{\Pr[A]}$$

All situations demanding a stochastic model are defined by what is known as the ordered-triple of probability theory $(\Omega, \mathscr{F}, P)$. The universal set Ω defines the set of events, $\mathscr{F}$ defines elementary events (by imposing the structure of union and intersection), and P the probability assignment that conforms to the laws of probability. In most applications, only the probability law needs precise definition. In some advanced situations, precisely defining the σ-algebra is also required.

1.2 Random Variables and Probability Density Functions

A *random variable* X is the assignment of a number—real or complex—to each sample point in sample space; mathematically, $X : \Omega \mapsto \mathbb{R}$ (see Figure 1.1). Thus, a random variable can be considered a function whose domain is a set and whose range are, most commonly, a subset of the real line. This range could be discrete-valued (especially when the domain Ω is discrete). In this case, the random variable is said to be *symbolic-valued*. In some cases, the symbols can be related to the integers, and then the values of the random variable can be ordered. When the range is continuous, an interval on the real-line say, we have a *continuous-valued* random variable. In some cases, the random variable is a *mixed* random variable: it is both discrete- and continuous-valued.

The *probability distribution function* or *cumulative* can be defined for continuous, discrete (only if an ordering exists), and mixed random variables.

$$P_X(x) \equiv \Pr[X \leq x] \,.$$

Note that X denotes the random variable and x denotes the argument of the distribution function. Probability distribution functions are increasing functions: if $A = \{\omega : X(\omega) \leq x_1\}$ and $B = \{\omega : x_1 < X(\omega) \leq x_2\}$, $\Pr[A \cup B] = \Pr[A] + \Pr[B] \implies P_X(x_2) = P_X(x_1) + \Pr[x_1 < X \leq x_2]$,* which means that $P_X(x_2) \geq P_X(x_1)$, $x_1 \leq x_2$.

The *probability density function* $p_X(x)$ is defined to be that function when integrated yields the distribution function.

$$P_X(x) = \int_{-\infty}^{x} p_X(\alpha)\, d\alpha$$

As distribution functions may be discontinuous when the random variable is discrete or mixed, we allow density functions to contain impulses. Furthermore, density functions must be non-negative since their integrals are increasing.

1.2.1 Function of a Random Variable

When random variables are real-valued, we can consider applying a real-valued function. Let $Y = f(X)$; in essence, we have the sequence of maps $f\colon\ \Omega \mapsto \mathbb{R} \mapsto \mathbb{R}$, which is equivalent to a simple mapping from sample space Ω to the real line. Mappings of this sort constitute the definition of a random variable, leading us to conclude that Y is a random variable. Now the question becomes "What are Y's probabilistic properties?". The key to determining the probability density function, which would allow calculation of the mean and variance, for example, is to use the probability *distribution* function.

For the moment, assume that $f(\cdot)$ is a monotonically increasing function. The probability distribution of Y we seek is

$$\begin{aligned} P_Y(y) &= \Pr[Y \leq y] \\ &= \Pr[f(X) \leq y] \\ &= \Pr[X \leq f^{-1}(y)] \qquad (*) \\ &= P_X\big(f^{-1}(y)\big) \end{aligned}$$

Equation (*) is the key step; here, $f^{-1}(y)$ is the inverse function. Because $f(\cdot)$ is a strictly increasing function, the underlying portion of sample space corresponding to $Y \leq y$ must be the same as that corresponding to $X \leq f^{-1}(y)$. We can find Y's density by evaluating the derivative.

$$p_y(y) = \frac{df^{-1}(y)}{dy} p_X\big(f^{-1}(y)\big)$$

The derivative term amounts to $1/f'(x)|_{x=y}$.

The style of this derivation applies to monotonically decreasing functions as well. The difference is that the set corresponding to $Y \leq y$ now corresponds to $X \geq f^{-1}(x)$. Now, $P_Y(y) = 1 - P_X\big(f^{-1}(y)\big)$. The probability density function of a monotonic—increasing or decreasing—function of a random variable is found according to the formula

$$\boxed{p_y(y) = \left|\frac{1}{f'\big(f^{-1}(y)\big)}\right| p_X\big(f^{-1}(y)\big)}\,.$$

*What property do the sets A and B have that makes this expression correct?

Example

Suppose X has an exponential probability density: $p_X(x) = e^{-x}\mathrm{u}(x)$, where $\mathrm{u}(x)$ is the unit-step function. We have $Y = X^2$. Because the square-function is monotonic over the positive real line, our formula applies. We find that

$$p_Y(y) = \frac{1}{2\sqrt{y}}e^{-\sqrt{y}},\ y > 0\ .$$

Although difficult to show, this density indeed integrates to one.

1.2.2 Expected Values

The *expected value* of a function $f(\cdot)$ of a random variable X is defined to be

$$\mathscr{E}[f(X)] = \int_{-\infty}^{\infty} f(x)p_X(x)\,dx\,.$$

Several important quantities are expected values, with specific forms for the function $f(\cdot)$.

- $f(X) = X$.
  The *expected value* or *mean* of a random variable is the center-of-mass of the probability density function. We shall often denote the expected value by m_X or just m when the meaning is clear. Note that the expected value can be a number never assumed by the random variable ($p_X(m)$ can be zero). An important property of the expected value of a random variable is *linearity*: $\mathscr{E}[aX] = a\mathscr{E}[X]$, a being a scalar.
- $f(X) = X^2$.
  $\mathscr{E}[X^2]$ is known as the *mean squared value* of X and represents the "power" in the random variable.
- $f(X) = (X - m_X)^2$.
  The so-called second central difference of a random variable is its *variance*, usually denoted by σ_X^2. This expression for the variance simplifies to $\sigma_X^2 = \mathscr{E}[X^2] - \mathscr{E}^2[X]$, which expresses the variance operator $\mathscr{V}[\cdot]$. The square root of the variance σ_X is the *standard deviation* and measures the spread of the distribution of X. Among all possible second differences $(X - c)^2$, the minimum value occurs when $c = m_X$ (simply evaluate the derivative with respect to c and equate it to zero).
- $f(X) = X^n$.
  $\mathscr{E}[X^n]$ is the n^{th} *moment* of the random variable and $\mathscr{E}\left[(X - m_X)^n\right]$ the n^{th} central moment.
- $f(X) = e^{juX}$.
  The *characteristic function* of a random variable is essentially the Fourier Transform of the probability density function.

  $$\mathscr{E}\left[e^{j\nu X}\right] \equiv \Phi_X(j\nu) = \int_{-\infty}^{\infty} p_X(x)e^{j\nu x}\,dx$$

  The moments of a random variable can be calculated from the derivatives of the characteristic function evaluated at the origin.

  $$\mathscr{E}[X^n] = j^{-n}\frac{d^n\Phi_X(j\nu)}{d\nu^n}\bigg|_{\nu=0}$$

1.2.3 Jointly Distributed Random Variables

Two (or more) random variables can be defined over the same sample space: $X : \Omega \mapsto \mathbb{R}, Y : \Omega \mapsto \mathbb{R}$. More generally, we can have a random vector (dimension N) $\mathbf{X} : \Omega \mapsto \mathbb{R}^N$. First, let's consider the two-dimensional case: $\mathbf{X} = \{X, Y\}$. Just as with jointly defined events, the *joint distribution function* is easily defined.

$$P_{X,Y}(x,y) \equiv \Pr[\{X \le x\} \cap \{Y \le y\}]$$

The *joint probability density function* $p_{X,Y}(x,y)$ is related to the distribution function via double integration.

$$P_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} p_{X,Y}(\alpha,\beta)\,d\alpha\,d\beta \quad \text{or} \quad p_{X,Y}(x,y) = \frac{\partial^2 P_{X,Y}(x,y)}{\partial x \partial y}$$

Since $\lim_{y\to\infty} P_{X,Y}(x,y) = P_X(x)$, the so-called *marginal density functions* can be related to the joint density function.

$$p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x,\beta)\,d\beta \text{ and } p_Y(y) = \int_{-\infty}^{\infty} p_{X,Y}(\alpha,y)\,d\alpha$$

Extending the ideas of conditional probabilities, the *conditional probability density function* $p_{X|Y}(x|Y=y)$ is defined (when $p_Y(y) \neq 0$) as

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

For jointly defined random variables, expected values are defined similarly as with single random variables. Probably the most important joint moment is the *covariance*:

$$\text{cov}[X,Y] \equiv \mathscr{E}[XY] - \mathscr{E}[X]\cdot\mathscr{E}[Y], \quad \text{where } \mathscr{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy p_{X,Y}(x,y)\,dx\,dy\,.$$

Related to the covariance is the (confusingly named) *correlation coefficient*: the covariance normalized by the standard deviations of the component random variables.

$$\rho_{X,Y} = \frac{\text{cov}[X,Y]}{\sigma_X \sigma_Y}$$

Because of the Cauchy-Schwarz inequality, the correlation coefficient's value ranges between -1 and 1.

A *conditional expected value* is the mean of the conditional density.

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

Note that the conditional expected value is now a function of Y and is therefore a random variable. Consequently, it too has an expected value, which is easily evaluated to be the expected value of X.

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

More generally, the expected value of a function of two random variables can be shown to be the expected value of a conditional expected value: $\mathscr{E}\left[f(X,Y)\right] = \mathscr{E}\left[\mathscr{E}[f(X,Y)|Y]\right]$. This kind of calculation is frequently simpler to evaluate than trying to find the expected value of $f(X,Y)$ "all at once." A particularly interesting example of this simplicity is the *random sum of random variables*. Let L be a random variable and $\{X_l\}$ a sequence of random variables. We will find occasion to consider the quantity $\sum_{l=1}^{L} X_l$. Assuming that the each component of the sequence has the same expected value $\mathscr{E}[X]$, the expected value of the sum is found to be

$$\begin{aligned}\mathscr{E}[S_L] &= \mathscr{E}\left[\mathscr{E}\left[\textstyle\sum_{l=1}^{L} X_l|L\right]\right]\\ &= \mathscr{E}\left[L\cdot\mathscr{E}[X]\right]\\ &= \mathscr{E}[L]\cdot\mathscr{E}[X]\end{aligned}$$

1.2.4 Notions of Statistical Dependence

Statistical dependence is *not* consequent of the structure or nature of the underlying event space. Rather, *statistical* dependence captures how two (or more) random variables are interrelated through the probability law P.

Two random variables are *statistically independent* when $p_{X|Y}(x|Y=y) = p_X(x)$, which is equivalent to the condition that the joint density function is separable: $p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$. Thus, no matter what the conditioning value y, the probabilistic properties of X remain unchanged from what they would be if we were ignorant of Y's value.

A weaker form of independence is *mean-square* independence wherein the conditional mean equals the expected value: $\mathscr{E}[X|Y=y] = \mathscr{E}[X]$ for all values of y. Clearly, random variables that are statistically independent are also mean-square independent, but not the other way around. The essential reason for mean-square independence being weaker is because expected values are integrals.

When two random variables are *uncorrelated*, their covariance and correlation coefficient equals zero so that $\mathscr{E}[XY] = \mathscr{E}[X]\,\mathscr{E}[Y]$. Statistically independent and mean-square independent random variables are always uncorrelated, but uncorrelated random variables can be dependent. For example, let X be uniformly distributed over $[-1,1]$ and let $Y = X^2$. The two random variables are uncorrelated, but are clearly not statistically independent. The correlation coefficient equals zero when two randfom variables are uncorrelated.

$$\text{statistical independence} \Longrightarrow \text{mean-square independence} \Longrightarrow \text{uncorrelated}$$

Despite being the weakest, the correlation coefficient is typically used to assess from data the statistical dependence between two random variables. The more stringent notions would have one determine if a function, in the case of statistical dependence, or a scalar, in the case of mean-square dependence, varied with a random variable's value.* The correlation coefficient represents a single quantity that not only can assess whether two random variables are uncorrelated (if the correlation coefficient zero?), but also measure the degree of correlation. The correlation coefficient quantifies the degree to which the statistical relationship between two random variables can be summarized by a straight line. When $Y = aX$, $\rho_{X,Y} = 1$ if $a > 0$ and $\rho_{X,Y} = -1$ if $a < 0$. The smaller the magnitude of ρ, the less correlated the two random variables.

1.2.5 Random Vectors

A *random vector* $\mathbf{X}$ is an ordered sequence of random variables $\mathbf{X} = \text{col}[X_1, \ldots, X_L]$. The density function of a random vector is defined in a manner similar to that for pairs of random variables. The expected value of a random vector is the vector of expected values.

$$\mathscr{E}[\mathbf{X}] = \int_{-\infty}^{\infty} \mathbf{x} p_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} = \text{col}\left[\mathscr{E}[X_1], \ldots, \mathscr{E}[X_L]\right]$$

The *covariance matrix* $\mathbf{K}_X$ is an $L \times L$ matrix consisting of all possible covariances among the random vector's components.

$$\mathbf{K}^X_{ij} = \text{cov}[X_i, X_j] = \mathscr{E}[X_i X_j^*] - \mathscr{E}[X_i]\,\mathscr{E}[X_j^*] \quad i,j = 1,\ldots,L$$

Using matrix notation, the covariance matrix can be written as $\mathbf{K}_X = \mathscr{E}\left[(\mathbf{X} - \mathscr{E}[\mathbf{X}])(\mathbf{X} - \mathscr{E}[\mathbf{X}])'\right]$. Using this expression, the covariance matrix is seen to be a symmetric matrix and, when the random vector has no zero-variance component, its covariance matrix is positive-definite. Note in particular that when the random variables are real-valued, the diagonal elements of a covariance matrix equal the variances of the components: $\mathbf{K}^X_{ii} = \sigma^2_{X_i}$. *Circular* random vectors are complex-valued with uncorrelated, identically distributed, real and imaginary parts. In this case, $\mathscr{E}\left[|X_i|^2\right] = 2\sigma^2_{X_i}$ and $\mathscr{E}\left[X_i^2\right] = 0$. By convention, $\sigma^2_{X_i}$ denotes the variance of the real (or imaginary) part. The characteristic function of a real-valued random vector is defined to be

$$\Phi_{\mathbf{X}}(j\nu) = \mathscr{E}\left[e^{j\nu^t \mathbf{X}}\right].$$

*Information theoretic quantities, like entropy, can be used to assess statistical dependence. If and only if X,Y are statistically independent does the joint entropy $\mathscr{H}(X,Y)$ equal $\mathscr{H}(X) \cdot \mathscr{H}(Y)$. Techniques exist for measuring entropy without needing to estimate the probability function.

1.2.6 Single function of a random vector

Just as shown in §1.2.1, the key tool is the distribution function. When $Y = f(\mathbf{X})$, a scalar-valued function of a vector, we need to find that portion of the domain that corresponds to $f(\mathbf{X}) \leq y$. Once this region is determined, the density can be found.

For example, the maximum of a random vector is a random variable whose probability density is usually quite different than the distributions of the vector's components. The probability that the maximum is less than some number μ is equal to the probability that *all* of the components are less than μ.

$$\Pr[\max \mathbf{X} < \mu] = P_{\mathbf{X}}(\mu, \ldots, \mu)$$

Assuming that the components of $\mathbf{X}$ are statistically independent, this expression becomes

$$\Pr[\max \mathbf{X} < \mu] = \prod_{i=1}^{\dim \mathbf{X}} P_{X_i}(\mu)\,,$$

and the density of the maximum has an interesting answer.

$$p_{\max \mathbf{X}}(\mu) = \sum_{j=1}^{\dim \mathbf{X}} p_{X_j}(\mu) \prod_{i \neq j} P_{X_i}(\mu)$$

When the random vector's components are identically distributed, we have

$$p_{\max \mathbf{X}}(\mu) = (\dim \mathbf{X}) p_X(\mu) P_X^{(\dim \mathbf{X})-1}(\mu)\ .$$

1.2.7 Several functions of a random vector

When we have a vector-valued function of a vector (and the input and output dimensions don't necessarily match), finding the joint density of the function can be quite complicated, but the recipe of using the joint distribution function still applies. In some (intersting) cases, the derivation flows nicely. Consider the case where $\mathbf{Y} = \mathbf{A}\mathbf{X}$, where $\mathbf{A}$ is an invertible matrix.

$$\begin{aligned} P_{\mathbf{Y}}(\mathbf{y}) &= \Pr[\mathbf{A}\mathbf{X} \leq \mathbf{y}] \\ &= \Pr\left[\mathbf{X} \leq \mathbf{A}^{-1}\mathbf{y}\right] \\ &= P_{\mathbf{X}}\left(\mathbf{A}^{-1}\mathbf{y}\right) \end{aligned}$$

To find the density, we need to evaluate the N^{th}-order mixed derivative (N is the dimension of the random vectors). The Jacobian appears and in this case, the Jacobian is the determinant of the matrix $\mathbf{A}$.

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|\det \mathbf{A}|} p_{\mathbf{X}}\left(\mathbf{A}^{-1}\mathbf{y}\right)$$

1.3 Sequences of Random Variables

Sequences of random variables $X_1, X_2, \ldots$ denotes a sequence of functions defined on the probability space. We care how this sequence of random variables behaves, in particular does the sequence *converge* to some well-defined random variable?

$$\lim_{n\to\infty} X_n \overset{?}{=} X$$

One could simply extend the definition of a convergent sequence of real-valued functions: does $f_n(x) \to f(x)$? Here, convergence means that the sequence of real numbers $f_n(x_0)$ converges to $f(x_0)$ for all choices of x_0. Were it so simple. This kind of convergence is known as *point-wise convergence*. It is well-known that Fourier series do not converge point-wise (points of discontinuity cause problems). Consequently, we need weaker forms of convergence, which amounts to defining what "=" means. For random variables, it is even more complicated because we also need to include the definition of probability for all the random variables involved. Consequently, many forms of convergence have been defined.

Sure convergence. The random sequence X_n converges surely to the random variable X if the sequence $X(n,\omega)$ converges to the function $X(\omega)$ as $n \to \infty$ for all $\omega \in \Omega$. Sure convergence amounts to point-wise convergence of nonrandom functions. This most restrictive form of convergence requires that the sequence converges even on sets that have probability zero of occurring. Consequently, this form of convergence is usually too demanding.

Almost-sure convergence. The sequence X_n converges *a.s.* to X on all sets that have non-zero probability.

$$\Pr\left[\lim_{n\to\infty} X_n = X\right] = 1$$

This form of convergence is also known as *probability-one convergence*.

Mean-square convergence. A sequence of random variables converges in the mean-square sense if

$$\lim_{n\to\infty} \mathscr{E}[|X_n - X|^2] = 0 \ .$$

This kind of convergence depends only on the second-order properties of the random variable (all second moments must be finite, of course) and thus is a weak form of convergence.

Convergence in probability. This even weaker form of convergence than in mean-square demands that the probability the sequence deviates from the limit be zero.

$$\lim_{n\to\infty} \Pr[|X_n - X| > \varepsilon] = 0 \quad \forall\, \varepsilon > 0$$

"Weaker" means that we can show that all sequences converging in mean-square also converge in probability but not vice-versa. The proof relies on the *Chebyshev inequality*.

$$\Pr[|Y| > \varepsilon] \le \frac{\mathscr{E}[|Y|^2]}{\varepsilon^2}$$

Showing this result is easy.

$$\begin{aligned}
\mathscr{E}[|Y|^2] &= \int_{-\infty}^{\infty} y^2 p_Y(y)\,dy \\
&\ge \int_{|y|\ge\varepsilon} y^2 p_Y(y)\,dy \\
&\ge \varepsilon^2 \int_{|y|\ge\varepsilon} p_Y(y)\,dy \\
&= \varepsilon^2 \Pr[|Y| > \varepsilon]
\end{aligned}$$

To apply the Chebyshev inequality, we let $Y = X_n - X$. Assuming $X_n \to X$ in the mean-square sense, $\lim_{n\to\infty} \mathscr{E}[|Y|^2] = 0$. Consequently, for any $\varepsilon > 0$, $\lim_{n\to\infty} \Pr[|X_n - X| > \varepsilon] = 0$. To show that the converse does not apply, we need only create a sequence without second moments, like Cauchy random variables, that converges in probability. For example, Let X be Cauchy and define $X_n = X + 1/n$.

Convergence in distribution. Let the random variables X_n have a probability *distribution* function $P_{X_n}(\cdot)$. The sequence formed by these random variables converges in distribution to the random variable X if

$$\lim_{n\to\infty} P_{X_n}(x) = P_X(x)$$

for all points of continuity of $P_X(x)$. This is the weakest form of convergence of those described here since it only concerns the probability assignments, not the inherent properties of the random variables.

The hierarchy of convergence modes of random sequences is shown in Figure 1.2.

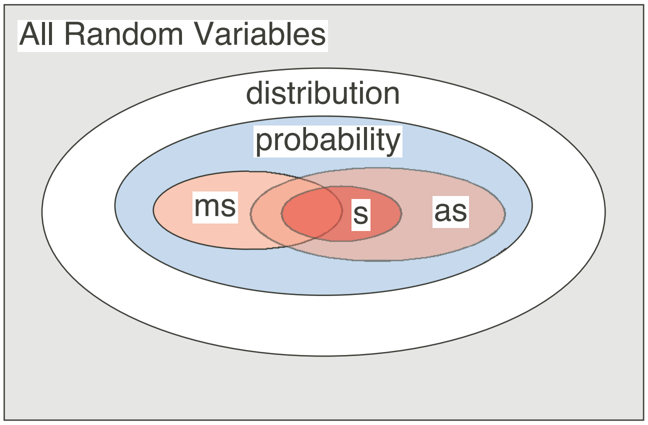


Figure 1.2: The implication hierarchy of notions of convergence are depicted. The weakest form (in probability) encompasses more situations while the most restrictive—surely—applies to the fewest.

Perhaps the most common application of notions of convergence is the "Law of the Unconscious Statistician:" the sample average of statistically independent, identically distributed random variables converges to the mean.

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathscr{E}[X]$$

Here, the sequence of random variables is the sample average and the convergent random variable only assumes the value of the mean with non-zero probability. The Strong Law of Large Numbers uses the notion of almost sure convergence and the Weak Law of Large Numbers uses convergence in probability.

1.4 Special Random Variables

The Appendix describes the properties of many kinds of random variables. A few of the most important ones are described here.

1.4.1 The Gaussian Random Variable

The random variable X is said to be a *Gaussian random variable*[*] if its probability density function has the form

$$\boxed{p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}}.$$

The mean of such a Gaussian random variable is m and its variance σ^2. As a shorthand notation, this information is denoted by $x \sim \mathscr{N}(m,\sigma^2)$. The characteristic function $\Phi_X(\cdot)$ of a Gaussian random variable is given by

$$\Phi_X(j\nu) = e^{jm\nu} \cdot e^{-\sigma^2\nu^2/2}.$$

No closed form expression exists for the probability distribution function of a Gaussian random variable. For a zero-mean, unit-variance, Gaussian random variable $\big(\mathscr{N}(0,1)\big)$, the probability that it *exceeds* the value x is denoted by $Q(x)$.

$$\Pr[X > x] = 1 - P_X(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\alpha^2/2}\, d\alpha \equiv Q(x)$$

A plot of $Q(\cdot)$ is shown in Fig. 1.3. When the Gaussian random variable has non-zero mean and/or non-unit

[*] Gaussian random variables are also known as *normal* random variables.

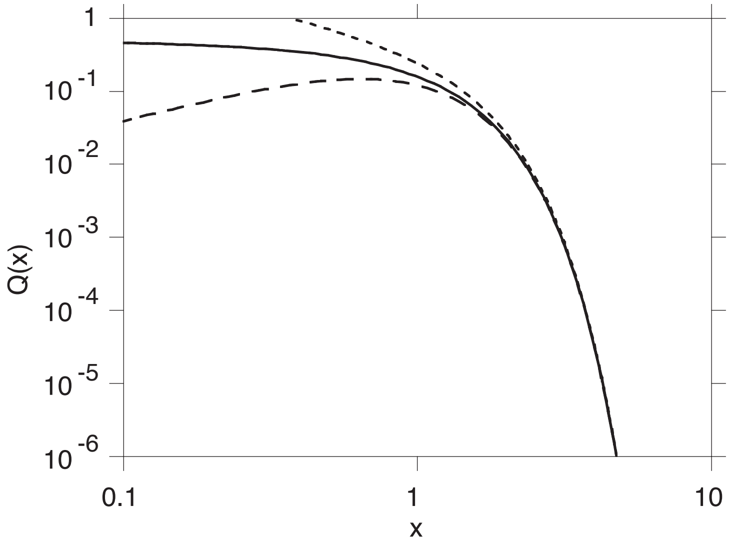


Figure 1.3: The function $Q(\cdot)$ is plotted on logarithmic coordinates. Beyond values of about two, this function decreases quite rapidly. Two approximations are also shown that correspond to the upper and lower bounds given by Eq. 1.1.

variance, the probability of it exceeding x can also be expressed in terms of $Q(\cdot)$.

$$\Pr[X > x] = Q\left(\frac{x-m}{\sigma}\right), \quad X \sim \mathcal{N}(m, \sigma^2)$$

Integrating by parts, $Q(\cdot)$ is bounded (for $x > 0$) by

$$\frac{1}{\sqrt{2\pi}} \cdot \frac{x}{1+x^2} e^{-x^2/2} \leq Q(x) \leq \frac{1}{\sqrt{2\pi}x} e^{-x^2/2}. \tag{1.1}$$

As x becomes large, these bounds approach each other and either can serve as an approximation to $Q(\cdot)$; the upper bound is usually chosen because of its relative simplicity. The lower bound can be improved; noting that the term $x/(1+x^2)$ decreases for $x < 1$ and that $Q(x)$ increases as x decreases, the term can be replaced by its value at $x = 1$ without affecting the sense of the bound for $x \leq 1$.

$$\frac{1}{2\sqrt{2\pi}} e^{-x^2/2} \leq Q(x), \quad x \leq 1 \tag{1.2}$$

We will have occasion to evaluate the expected value of $\exp\{aX + bX^2\}$ where $X \sim \mathcal{N}(m, \sigma^2)$ and a, b are constants. By definition,

$$\mathcal{E}[e^{aX+bX^2}] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\{ax + bx^2 - (x-m)^2/(2\sigma^2)\}\, dx$$

The argument of the exponential requires manipulation (*i.e.*, completing the square) before the integral can be evaluated. This expression can be written as

$$-\frac{1}{2\sigma^2}\{(1-2b\sigma^2)x^2 - 2(m+a\sigma^2)x + m^2\}\ .$$

Completing the square, this expression can be written

$$-\frac{1-2b\sigma^2}{2\sigma^2}\left(x - \frac{m+a\sigma^2}{1-2b\sigma^2}\right)^2 + \frac{1-2b\sigma^2}{2\sigma^2}\left(\frac{m+a\sigma^2}{1-2b\sigma^2}\right)^2 - \frac{m^2}{2\sigma^2}$$

We are now ready to evaluate the integral. Using this expression,

$$\mathscr{E}[e^{aX+bX^2}] = \exp\left\{\frac{1-2b\sigma^2}{2\sigma^2}\left(\frac{m+a\sigma^2}{1-2b\sigma^2}\right)^2 - \frac{m^2}{2\sigma^2}\right\} \times$$
$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1-2b\sigma^2}{2\sigma^2}\left(x-\frac{m+a\sigma^2}{1-2b\sigma^2}\right)^2\right\}dx\,.$$

Let

$$\alpha = \frac{x-\frac{m+a\sigma^2}{1-2b\sigma^2}}{\frac{\sigma}{\sqrt{1-2b\sigma^2}}},$$

which implies that we must require that $1-2b\sigma^2>0$ $\bigl(\text{or } b<1/(2\sigma^2)\bigr)$. We then obtain

$$\mathscr{E}\left[e^{aX+bX^2}\right] = \exp\left\{\frac{1-2b\sigma^2}{2\sigma^2}\left(\frac{m+a\sigma^2}{1-2b\sigma^2}\right)^2-\frac{m^2}{2\sigma^2}\right\}\frac{1}{\sqrt{1-2b\sigma^2}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{\alpha^2}{2}}\,d\alpha\,.$$

The integral equals unity, leaving the result

$$\boxed{\mathscr{E}[e^{aX+bX^2}] = \frac{\exp\left\{\frac{1-2b\sigma^2}{2\sigma^2}\left(\frac{m+a\sigma^2}{1-2b\sigma^2}\right)^2-\frac{m^2}{2\sigma^2}\right\}}{\sqrt{1-2b\sigma^2}}\,,b<\frac{1}{2\sigma^2}}$$

Important special cases are

1. $a=0, X\sim\mathscr{N}(m,\sigma^2)$.
$$\mathscr{E}[e^{bX^2}] = \frac{\exp\left\{\frac{bm^2}{1-2b\sigma^2}\right\}}{\sqrt{1-2b\sigma^2}}$$

2. $a=0, X\sim\mathscr{N}(0,\sigma^2)$.
$$\mathscr{E}[e^{bX^2}] = \frac{1}{\sqrt{1-2b\sigma^2}}$$

3. $X\sim\mathscr{N}(0,\sigma^2)$.
$$\mathscr{E}[e^{aX+bX^2}] = \frac{\exp\left\{\frac{a^2\sigma^2}{2(1-2b\sigma^2)}\right\}}{1-2b\sigma^2}$$

The real-valued random vector $\mathbf{X}$ is said to be a *Gaussian random vector* if its joint distribution function has the form

$$\boxed{p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{\det[2\pi\mathbf{K}]}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^t\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})\right\}}\,.$$

If complex-valued, the joint distribution of a circular Gaussian random vector is given by

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{\det[\pi\mathbf{K}]}}\exp\left\{-(\mathbf{x}-\mathbf{m}_X)'\mathbf{K}_X^{-1}(\mathbf{x}-\mathbf{m}_X)\right\}\,. \tag{1.3}$$

The vector $\mathbf{m}_X$ denotes the expected value of the Gaussian random vector and $\mathbf{K}_X$ its covariance matrix.

$$\mathbf{m}_X = \mathscr{E}[\mathbf{X}] \qquad \mathbf{K}_X = \mathscr{E}[\mathbf{X}\mathbf{X}'] - \mathbf{m}_X\mathbf{m}_X'$$

As in the univariate case, the Gaussian distribution of a random vector is denoted by $\mathbf{X} \sim \mathscr{N}(\mathbf{m}_X, \mathbf{K}_X)$. Note that if the covariance matrix is diagonal, which would occur if the components of the random vector were pairwise uncorrelated, the joint probability density factors into the marginal distributions. Thus, for Gaussian random vectors, if all components are pairwise uncorrelated, the random variables are statistically independent. The weakest form of statistical independence implies the strongest.

After applying a linear transformation to Gaussian random vector, such as $\mathbf{Y} = \mathbf{A}\mathbf{X}$, the result is also a Gaussian random vector (a random variable if the matrix is a row vector): $\mathbf{Y} \sim \mathscr{N}(\mathbf{A}\mathbf{m}_X, \mathbf{A}\mathbf{K}_X\mathbf{A}')$.

The characteristic function of a Gaussian random vector is given by

$$\Phi_{\mathbf{X}}(j\nu) = \exp\left\{+j\nu^t\mathbf{m}_X - \frac{1}{2}\nu^t\mathbf{K}_X\nu\right\}.$$

From this formula, the N^{th}-order moment formula for jointly distributed Gaussian random variables is easily derived.*

$$\mathscr{E}[X_1\cdots X_N] = \begin{cases} \sum_{\text{all}\,\mathscr{P}_N} \mathscr{E}[X_{\mathscr{P}_N(1)}X_{\mathscr{P}_N(2)}]\cdots\mathscr{E}[X_{\mathscr{P}_N(N-1)}X_{\mathscr{P}_N(N)}], & N \text{ even} \\ \sum_{\text{all}\,\mathscr{P}_N} \mathscr{E}[X_{\mathscr{P}_N(1)}]\,\mathscr{E}[X_{\mathscr{P}_N(2)}X_{\mathscr{P}_N(3)}]\cdots\mathscr{E}[X_{\mathscr{P}_N(N-1)}X_{\mathscr{P}_N(N)}], & N \text{ odd,} \end{cases}$$

where $\mathscr{P}_N$ denotes a permutation of the first N integers and $\mathscr{P}_N(i)$ the i^{th} element of the permutation. For example, $\mathscr{E}[X_1X_2X_3X_4] = \mathscr{E}[X_1X_2]\,\mathscr{E}[X_3X_4] + \mathscr{E}[X_1X_3]\,\mathscr{E}[X_2X_4] + \mathscr{E}[X_1X_4]\,\mathscr{E}[X_2X_3]$.

1.4.2 The Central Limit Theorem

Let $\{X_l\}$ denote a sequence of independent, identically distributed, random variables. Assuming they have zero means and finite variances (equaling σ^2), the Central Limit Theorem states that the sum $\sum_{l=1}^{L} X_l/\sqrt{L}$ converges in distribution to a Gaussian random variable.

$$\frac{1}{\sqrt{L}}\sum_{l=1}^{L} X_l \xrightarrow{L\to\infty} \mathscr{N}(0,\sigma^2)$$

Because of its generality, this theorem is often used to simplify calculations involving *finite* sums of non-Gaussian random variables. However, attention is seldom paid to the *convergence rate* of the Central Limit Theorem. Kolmogorov, the famous twentieth century mathematician, is reputed to have said "The Central Limit Theorem is a dangerous tool in the hands of amateurs." Let's see what he meant.

Taking $\sigma^2 = 1$, the key result is that the magnitude of the difference between $P(x)$, defined to be the probability that the sum given above exceeds x, and $Q(x)$, the probability that a unit-variance Gaussian random variable exceeds x, is bounded by a quantity inversely related to the square root of L [7: Theorem 24].

$$|P(x) - Q(x)| \le c \cdot \frac{\mathscr{E}\left[|X|^3\right]}{\sigma^3} \cdot \frac{1}{\sqrt{L}}$$

The constant of proportionality c is a number known to be about 0.8 [11: p. 6]. The ratio of absolute third moment of X_l to the cube of its standard deviation, known as the skew and denoted by γ_X, depends only on the distribution of X_l and is independent of scale. This bound on the absolute error has been shown to be tight [7: pp. 79ff]. Using our lower bound for $Q(\cdot)$ (Eq. 1.2 $\{10\}$), we find that the relative error in the Central Limit Theorem approximation to the distribution of finite sums is bounded for $x > 0$ as

$$\boxed{\frac{|P(x) - Q(x)|}{Q(x)} \le c\gamma_X\sqrt{\frac{2\pi}{L}}e^{+x^2/2} \cdot \begin{cases} 2, & x \le 1 \\ \frac{1+x^2}{x}, & x > 1 \end{cases}}.$$

[*] $\mathscr{E}[X_1\cdots X_N] = j^{-N}\,\frac{\partial^N}{\partial\nu_1\cdots\partial\nu_N}\Phi_{\mathbf{X}}(j\nu)\Big|_{\nu=\mathbf{0}}$.

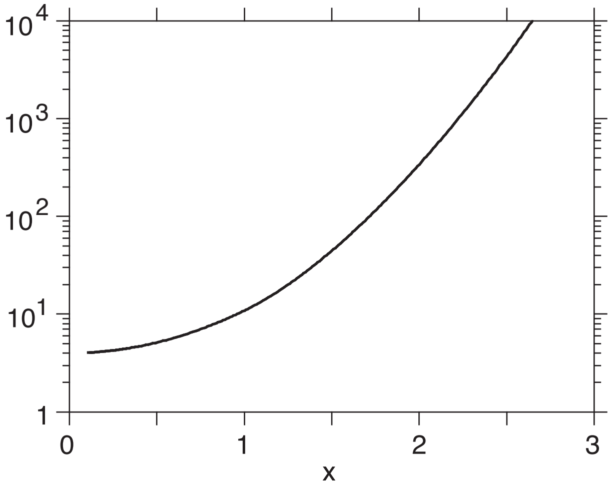


Figure 1.4: The quantity which governs the limits of validity for numerically applying the Central Limit Theorem on finite numbers of data is shown over a portion of its range. To judge these limits, we must compute the quantity $L\varepsilon^2/2\pi c^2\gamma_X$, where ε denotes the desired percentage error in the Central Limit Theorem approximation and L the number of observations. Selecting this value on the vertical axis and determining the value of x yielding it, we find the normalized ($x = 1$ implies unit variance) upper limit on an L-term sum to which the Central Limit Theorem is guaranteed to apply. Note how rapidly the curve increases, suggesting that large amounts of data are needed for accurate approximation.

Suppose we require that the relative error not exceed some specified value ε. The normalized (by the standard deviation) boundary x at which the approximation is evaluated must not violate

$$\frac{L\varepsilon^2}{2\pi c^2\gamma_X^2} \geq e^{x^2}\cdot\begin{cases}4 & x\leq 1\\ \left(\frac{1+x^2}{x}\right)^2 & x>1\end{cases}.$$

As shown in Fig. 1.4, the right side of this equation is a monotonically increasing function.

Example

For example, if $\varepsilon = 0.1$ and taking $c\gamma_X$ arbitrarily to be unity (a reasonable value), the upper limit of the preceding equation becomes $1.6\times 10^{-3}L$. Examining Fig. 1.4, we find that for $L = 10,000$, x must not exceed 1.17. Because we have normalized to unit variance, this example suggests that the Gaussian approximates the distribution of a ten-thousand term sum only over a range corresponding to an 76% area about the mean. Consequently, the Central Limit Theorem, as a finite-sample distributional approximation, is only guaranteed to hold near the mode of the Gaussian, with *huge* numbers of observations needed to specify the tail behavior. Realizing this fact will keep us from being ignorant amateurs.

1.4.3 The Exponential Random Variable

The exponential random variable is positive-valued and has a probability density given by

$$p_X(x) = \lambda e^{\lambda x}\mathrm{u}(x)\ .$$

The expected value of the exponential random variable is $1/\lambda$ and the variance is $1/\lambda^2$. This makes the exponential random variable's coefficient of variation equal to one.

1.4.4 The Bernoulli Random Variable

Perhaps the simplest example of a random variable is the *Bernoulli* random variable. Sometimes called a binary random variable, it only assumes the values 0 and 1.

$$\begin{aligned}\Pr[X=1] &= p\\ \Pr[X=0] &= 1-p\end{aligned}$$

Thus, a Bernoulli random variable can be considered either a discrete- or continuous-valued random variable. In the latter case, the Bernoulli random variable's probability density is given by

$$p_x(x) = (1-p)\delta(x) + p\delta(x-1)\ .$$

The expected value of a Bernoulli random variable equals the probability the random variable equals one: $\mathscr{E}[X] = p$. The sum of N statistically independent Bernoulli random variables is known as a *binomial* random variable because its probability mass function has the form

$$\Pr\left[\sum_{n=1}^{N} X_n = k\right] = \binom{N}{k} p^k (1-p)^{N-k}, \quad k = 0,\ldots,N\ .$$

When Bernoulli random variables are statistically *dependent*, the correlation among random variable pairs is no longer sufficient to describe the joint probability function. A more detailed statistical structure than that imposed by pairwise correlation occurs with most non-Gaussian random variables. The only cases in which correlation determines the dependence structure occurs when we can write the joint distribution as

$$p_{\mathbf{X}}(\mathbf{x}) = f\left((\mathbf{X}-\mathbf{m})'\mathbf{K}^{-1}(\mathbf{X}-\mathbf{m})\right)\ ,$$

where $f(\cdot)$ is some function of a scalar that can yield a joint density function. One such example is $f(x) \propto 1/(1+x)$. More generally, all joint density functions can be expanded in terms of a set of orthogonal functions as

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{n=1}^{N} p_{X_n}(x_n) \cdot \left[1 + \sum_{k=2}^{N} \sum_{j=1}^{\binom{N}{k}} \sum_{i_1,\ldots,i_N \in \mathscr{I}_j^N(k)} a_{i_1,\ldots,i_N} \prod_{n=1}^{N} \psi_{i_n}(x_n)\right]$$

This expression reflects just how complicated dependence can be. First of all, the set $\mathscr{I}_j^N(k)$ denotes the integers $i_1,\ldots,i_N$ that reflect the j^{th} subset of arrangements of the integers $1,\ldots,N$ of order k. For example, $\mathscr{I}_1^N(2) = \{110, 120, 130, \ldots, 210, 220, 230, \ldots\}$. Furthermore, $\psi_0(x) = 1$. Thus, this representation of joint distributions shown that pairs, triples, etc. of random variables can be individually dependent.

1.4.5 Stable Random Variables

Stable random variables play and interesting niche role in probability theory. X is a stable random variable if the weighted sum of two statistically independent instances has the "same" (within a scaling and shift) probability distribution. For example, Gaussian random variables are stable. What is interesting about non-Gaussian stable random variables is that they disobey the Central Limit Theorem. For example, the sum of two Cauchy random variables is also Cauchy, which has a probability density function of the form

$$p_x(x) = \frac{1}{\pi} \cdot \frac{\sigma}{\sigma^2 + x^2}\ .$$

Clearly, the sum of any number of Cauchy random variables will never "converge" to the Gaussian. The Central Limit Theorem requirement violated by stable random variables, save for the Gaussian, is that they have infinite variance. All stable random variables have a characteristic function of the form

$$\Phi_x(j\nu) = e^{-a|\nu|^\alpha},\ 0 \le \alpha \le 2,\ a \text{ a constant}$$

The Gaussian case occurs when $\alpha = 2$.

Problems

1.1 Space Exploration and MTV
Joe is an astronaut for project Pluto. The mission success or failure depends only on the behavior of three major systems. Joe feels that the following assumptions are valid and apply to the performance of the entire mission:

- The mission is a failure only if two or more major systems fail.
- System I, the Gronk system, fails with probability 0.1.
- System II, the Frab system, fails with probability 0.5 if at least one other system fails. If no other system fails, the probability the Frab system fails is 0.1.
- System III, the beer cooler (obviously, the most important), fails with probability 0.5 if the Gronk system fails. Otherwise the beer cooler cannot fail.

(a) What is the probability that the mission succeeds but that the beer cooler fails?

(b) What is the probability that all three systems fail?

(c) Given that more than one system failed, determine the probability that:

(i) The Gronk did not fail.
(ii) The beer cooler failed.
(iii) Both the Gronk and the Frab failed.

(d) About the time Joe was due back on Earth, you overhear a radio broadcast about Joe while watching MTV. You are not positive what the radio announcer said, but you decide that it is twice as likely that you heard ”mission a success” as opposed to ”mission a failure”. What is the probability that the Gronk failed?

1.2 Lost Boyfriend
Rachel has lost her boyfriend Al in either Duncan Hall (with *a priori* probability 0.4) or in Abercrombie (with *a priori* probability 0.6). If Al retains interest in Rachel but is not found by the N^{th} day of the search, he will find another girlfriend that evening with probability $\frac{N}{N+2}$ and remain uninterested in Rachel forever. If Al is in Duncan Hall (interested or not) and Rachel spends the day searching for him in Duncan, the conditional probability of finding Al that day is 0.25. Similarly, if Al is in Abercrombie and Rachel spends a day searching for him there, she will fuind him that day with probability 0.15. Al does not move between the buildings and Rachel can only search in the daytime and moves between the buildings after breakfast.

(a) In which building should Rachel look to maximize the probability that she finds Al on the first day of the search?

(b) Given that Rachel looked in Duncan on the first day but did not find Al, what is the probability that Al was in Duncan?

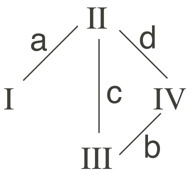
(c) If Rachel flips a fair coin to determine where to look the first day and she find Al on the first day, what is the probability she looked in Duncan?

(d) Rachel has decided to look in Duncan for the first two days. What is the *a priori* probability that she will find and interested boyfriend for the first time on the second day?

(e) Rachel has decided to look in Duncan for the first two days. Given that she is unsuccessful on the first day, determine the probability that she does not find an uninterested Al on the second day.

(f) Racel *finally* locates Al on the fourth day of the search. She looked in Duncan Hall for three days and in Abercrombie the fourth. What is the probability she found him still infatuated with her?

1.3 Communication Links
A communication network consists of four nodes (I–IV) connected via four links (*a*–*d*).



However, all the links may not be available. Let p denote the probability that any link is available and assume that the availability of a link is statistically independent of any other link's state. Two terminal can communicate only if they are connected by at least one chain of links.

(a) Let $A = \{\omega\colon \text{I and IV can communicate}\}$. Calculate $\Pr[A]$.

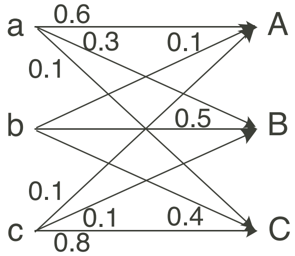
(b) Let $B = \{\omega\colon \text{II and III can communicate}\}$. Calculate $\Pr[B]$.

(c) Calculate $\Pr[AB]$. Are the events A and B statistically independent?

(d) Prove that $\Pr[A]$ would be increased if link c were connected between I and III as opposed to II and III.

1.4 Communication Channels

A noisy discrete communication channel is available. Once each microsecond, one letter from the three-letter alphabet $\{a, b, a\}$ is transmitted and one letter from the three-letter alphabet $\{A, B, C\}$ is received. The conditional probability of each received letter given the transmission letter is provided by the following transition diagram.



The *a priori* probability of each letter being transmitted is $\Pr[a] = 0.3, \Pr[b] = 0.5, \Pr[c] = 0.2$.

(a) What decision rule—an algorithm for relating a received letter to a transmitted letter—has the largest probability of being correct.

(b) What is the probability of error for this decision rule?

(c) What is the maximum probability of error that could be obtained without the the use of the channel? In other words, the receiver must decide what is transmitted without receiving *anything*!

1.5 Probability Density Functions?

Which of the following are probability density functions? Indicate your reasoning. For those that are valid, what is the mean and variance of the random variable?

(a) $p_X(x) = \dfrac{e^{-|x|}}{2}$ **(b)** $p_X(x) = \dfrac{\sin 2\pi x}{\pi x}$

(c) $p_X(x) = \begin{cases} 1-|x| & |x| \le 1 \\ 0 & \text{otherwise} \end{cases}$ **(d)** $p_X(x) = \begin{cases} 1 & |x| \le 1 \\ 0 & \text{otherwise} \end{cases}$

(e) $p_X(x) = \dfrac{1}{4}\delta(x+1) + \dfrac{1}{2}\delta(x) + \dfrac{1}{4}\delta(x-1)$ **(f)** $p_X(x) = \begin{cases} e^{-(x-1)} & x \ge 1 \\ 0 & \text{otherwise} \end{cases}$

1.6 Generating Random Variables

A crucial skill in developing simulations of systems subject to random influences is *random variable generation*. Most computers (and environments like MATLAB) have software that generates statistically independent, uniformly distributed, random sequences. In MATLAB, the function is `rand`. We want to change the probability distribution to one required by the problem at hand. One technique is known as the *distribution method*.

(a) If $P_X(x)$ is the desired distribution, show that $U = P_X(X)$ (applying the distribution function to a random variable having that distribution) is uniformly distributed over $[0,1)$. This result means that $X = P_X^{-1}(U)$ has the desired distribution. Consequently, to generate a random variable having any distribution we want, we only need the inverse function of the distribution function.

(b) Why is the Gaussian not in the class of "nice" probability distribution functions?

(c) How would you generate random variables having the hyperbolic secant $p_X(x) = (1/2)\text{sech}(\pi x/2)$, the Laplacian and the Cauchy densities?

(d) Write MATLAB functions that generate these random variables. Again use `hist` to plot the probability function. What do you notice about these random variables?

1.7 Cauchy Random Variables

The random variables X_1 and X_2 have the joint pdf

$$p_{X_1,X_2}(x_1,x_2) = \frac{1}{\pi^2}\frac{b_1 b_2}{(b_1^2+x_1^2)(b_2^2+x_2^2)}\;, b_1, b_2 > 0\;.$$

(a) Show that X_1 and X_2 are statistically independent random variables with Cauchy density functions.

(b) Show that $\Phi_{X_1}(j\nu) = e^{-b_1|\nu|}$.

(c) Define $Y = X_1 + X_2$. Determine $p_Y(y)$.

(d) Let $\{Z_i\}$ be a set of N statistically independent Cauchy random variables with $b_i = b, i = 1,\dots,N$. Define

$$Z = \frac{1}{N}\sum_{i=1}^{N} Z_i.$$

Determine $p_Z(z)$. Is Z—the sample mean—a good estimate of the expected value $\mathscr{E}[Z_i]$?

1.8 Correlation Coefficients

The random variables X, Y have the joint probability density $p_{X,Y}(x,y)$. The *correlation coefficient* $\rho_{X,Y}$ is defined to be

$$\rho_{X,Y} \equiv \frac{\mathscr{E}\big[(X-m_X)(Y-m_y)\big]}{\sigma_X\sigma_Y}\;.$$

(a) Using the Cauchy-Schwarz inequality, show that correlation coefficients always have a magnitude less than to equal to one.

(b) We would like find an affine estimate of one random variable's value from the other. So, if we wanted to estimate X from Y, our estimate $\widehat{X}$ has the form $\widehat{X} = aY + b$, where a, b are constants to be found. Our criterion is the mean-squared estimation error: $\varepsilon^2 = \mathscr{E}\big[(\widehat{X}-X)^2\big]$. First of all, let $a = 0$: we want to estimate X without using Y at all. Find the optimal value of b.

(c) Find the optimal values for both constants. Express your result using the correlation coefficient.

(d) What is the expected value of your estimate?

(e) What is the smallest possible mean-squared error? What influence does the correlation coefficient have on the estimate's accuracy?

1.9 Probabilistic Football

A football team, which shall remain nameless, likes to mix passing and running plays. The yardage gained on any running play is a random variable uniformly distributed between zero and ten yards regardless of the yardage gained on any other play. The team's quarterback, Bob Linguini, has a strange quirk: the yardage gained on a passing play depends on the previous play. If the previous play was a running play, the yardage gained passing is a random variable uniformly distributed betwee zero and twenty yards. If the previous play that gained Y yards, the yardage gained is a random variable uniformly distributed between $-Y$ and $20-Y$ yards. On any play, the team is equally likely to run or pass.

(a) What is the probability density function of the random variable defined to be the total yardage agained on a running play followed by a passing play?

(b) A running play is executed followed by two passing plays. Find the probability density function of the yardage gained on the second passing play.

(c) What is the probability that a total of at least ten yards is gained in the two passing plays mentioned in part (b)?

1.10 Order Statistics

Let $X_1, \ldots, X_N$ be independent, identically distributed random variables. The density of each random variable is $p_X(x)$. The *order statistics* $X_0(1), \ldots, X_0(N)$ of this set of random variables is the set that results when the original one is ordered (sorted).

$$X_0(1) \leq X_0(2) \leq \ldots \leq X_0(N)$$

(a) What is the joint density of the original set of random variables?

(b) What is the density of $X_0(N)$, the largest of the set?

(c) Show that the joint density of the ordered random variables is

$$p_{X_0(1),\ldots,X_0(N)}(x_1, \ldots, x_N) = N! p_X(x_1) \cdots p_X(x_N)$$

(d) Consider a Poisson process having constant intensity λ_0. N events are observed to occur in the interval $[0, T)$. Show that the joint density of the times of occurrence $W_1, \ldots, W_N$ is the same as the order statistics of a set of random variables. Find the common density of these random variables.

1.11 Estimating Characteristic Functions

Suppose you have a sequence of statistically independent, identically distributed random variables $X_1, \ldots, X_N$. From these, we want to estimate the characteristic function of the underlying random variable. One way to estimate it is to compute

$$\widehat{\Phi}_X(j\nu) = \frac{1}{N} \sum_{n=1}^{N} e^{j\nu X_n}$$

(a) What is the expected value of the estimate?

(b) Does this estimate converge to the actual characteristic function? If yes, demonstrate how; if not, why not?

Chapter 2

Stochastic Processes

2.1 Stochastic Processes

2.1.1 Basic Definitions

A *random* or *stochastic* process is the assignment of a function of a real variable to each sample point ω in sample space (see Figure 2.1). Thus, the process $X(\omega, t)$ can be considered a function of two variables. For each ω, the time function must be well-behaved and may or may not look random to the eye. Each time function of the process is called a *sample function* and must be defined over the entire domain of interest. For each t, we have a function of ω, which is precisely the definition of a random variable. Hence the *amplitude* of a random process is a random variable. The *amplitude distribution* of a process refers to the probability density function of the amplitude: $p_{X(t)}(x)$. By examining the process's amplitude at several instants, the joint amplitude distribution can also be defined. For the purposes of this book, a process is said to be *stationary* when the joint amplitude distribution depends on the differences between the selected time instants.

The *expected value* or *mean* of a process is the expected value of the amplitude at each t.

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

For the most part, we take the mean to be zero. The *correlation function* is the first-order joint moment between the process's amplitudes at two times.

$$R_X(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 p_{X(t_1),X(t_2)}(x_1, x_2)\, dx_1\, dx_2$$

Since the joint distribution for stationary processes depends only on the time difference, correlation functions of stationary processes depend only on $|t_1 - t_2|$. In this case, correlation functions are really functions of a single variable (the time difference) and are usually written as $R_X(\tau)$ where $\tau = t_1 - t_2$. Related to the correlation function is the *covariance function* $K_X(\tau)$, which equals the correlation function minus the square

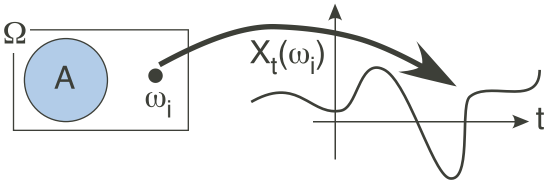


Figure 2.1: A stochastic process is defined much like a random variable (Figure 1.1) but with a time function assigned to each element of event space. The collection of time function is known as the *ensemble*.

of the mean.

$$K_X(\tau) = R_X(\tau) - m_X^2$$

The variance of the process equals the covariance function evaluated as the origin. The *power spectrum* of a stationary process is the Fourier Transform of the correlation function.

$$\mathscr{S}_X(f) = \int_{-\infty}^{\infty} R_X(\tau) e^{-j2\pi f\tau}\, d\tau$$

A particularly important example of a random process is *white noise*. The process $X(t)$ is said to be white if it has zero mean and a correlation function proportional to an impulse.

$$\mathscr{E}[X(t)] = 0 \quad R_X(\tau) = \frac{N_0}{2}\delta(\tau)$$

The power spectrum of white noise is constant for all frequencies, equaling $N_0/2$. which is known as the *spectral height*.*

When a stationary process $X(t)$ is passed through a stable linear, time-invariant filter, the resulting output $Y(t)$ is also a stationary process having power density spectrum

$$\mathscr{S}_Y(f) = |H(f)|^2 \mathscr{S}_X(f)\,,$$

where $H(f)$ is the filter's transfer function.

2.1.2 The Gaussian Process

A random process $X(t)$ is Gaussian if the joint density of the N amplitudes $X(t_1), \ldots, X(t_N)$ comprise a Gaussian random vector. The elements of the required covariance matrix equal the covariance between the appropriate amplitudes: $K_{ij} = K_X(t_i, t_j)$. Assuming the mean is known, the entire structure of the Gaussian random process is specified once the correlation function or, equivalently, the power spectrum are known. As linear transformations of Gaussian random processes yield another Gaussian process, linear operations such as differentiation, integration, linear filtering, sampling, and summation with other Gaussian processes result in a Gaussian process.

2.1.3 Sampling and Random Sequences

The usual Sampling Theorem applies to random processes, with the spectrum of interest being the power spectrum. If stationary process $X(t)$ is bandlimited—$\mathscr{S}_X(f) = 0$, $|f| > W$, as long as the sampling interval T satisfies the classic constraint $T < \pi/W$ the sequence $X(lT)$ represents the original process. A sampled process is itself a random process defined over discrete time. Hence, all of the random process notions introduced in the previous section apply to the random sequence $\widetilde{X}(l) \equiv X(lT)$. The correlation functions of these two processes are related as

$$R_{\widetilde{X}}(k) = \mathscr{E}[\widetilde{X}(l)\widetilde{X}(l+k)] = R_X(kT)\,.$$

We note especially that for distinct samples of a random process to be uncorrelated, the correlation function $R_X(kT)$ must equal zero for all non-zero k. This requirement places severe restrictions on the correlation function (hence the power spectrum) of the original process. One correlation function satisfying this property is derived from the random process which has a bandlimited, constant-valued power spectrum over precisely the frequency region needed to satisfy the sampling criterion. *No other power spectrum satisfying the sampling criterion has this property*. Hence, sampling does not normally yield uncorrelated amplitudes, meaning that *discrete-time white noise* is a rarity. White noise has a correlation function given by $R_{\widetilde{X}}(k) = \sigma^2\delta(k)$, where $\delta(\cdot)$ is the unit sample. The power spectrum of white noise is a constant: $\mathscr{S}_{\widetilde{X}}(f) = \sigma^2$.

*The curious reader can track down why the spectral height of white noise has the fraction one-half in it. This definition is the convention.

2.2 Structural Aspects of Waveform Processes

2.2.1 Stationarity

The stationarity of a waveform process is often assumed; otherwise, the process's temporal variations must be specified, either explicitly or through a model. When stationarity holds, the process's joint amplitude distribution does not depend on absolute time: The density depends only on the intervals among temporal samples. Consequently, specification of the joint amplitude distribution amounts to defining the underlying process's stationarity.

When system models are handy, time-invariant models are required to produce stationary outputs. In this context, *stability* is closely linked to stationarity [39]. The stability theory of difference equations, both linear and nonlinear, is technically delicate because of the possibility of chaos. Ignoring techncial conditions for the moment to get to the heart of the matter, iteration of a stable system with no input from any initial condition should lead to an output that lies in some compact set defined on the real line. A system is strongly stable if this compact set consists of a single point. For a strongly stable linear system, this limit set corresponds to the origin if all the poles lie within the unit circle.

To develop a parallel notion, iterating a system is equivalent to passing the output from one probability distribution to another, with the "initial condition" corresponding to an initial distribution for the the system's state. Stationarity thus means that the output's probability distribution asymptotically falls in a restricted set of distributions across any distributions of initial conditions. We seek conditions under which this restricted set consists of a single distribution, the so-called *stationary distribution*. For example, when the system is linear and time-invariant, a stationary distribution results if the system is strongly stable regardless of the white noise input's Gaussianity or the initial distribution. Assuming the distribution of the initial condition equals this stationary distribution, examination of the transfer functions pole locations or calculation of the system matrix's eigenvalues thus suffices as a stationarity test.

We concentrate here on *Markovian* systems: The output is computed from the past p outputs $\mathbf{X}_{l-1} = \text{col}[X_{l-1},\dots,X_{l-p}]$ and the input at a single time instant.

$$\boxed{X_l = \mathrm{G}[\mathbf{X}_{l-1};W_l]} \tag{2.1}$$

Many results, arising from the extensive literature on Markov chains, exist for this model and not for the more general one given previously. Some special-case results are known for the situation in which the output depends on several past input values.

In applications, the input frequently appears additively, yielding the *additive-input* Markovian model. Here, the system model $\mathrm{G}[\cdot;\cdot]$ equals the sum of a "state-dependent" part $\mathrm{G}_s[\mathbf{X}_{l-1}]$ and an input.*

$$\boxed{\mathrm{G}[\mathbf{X}_{l-1};W_l] = \mathrm{G}_s[\mathbf{X}_{l-1}] + W_l} \tag{2.2}$$

When the system is linear, for example, this relation expresses the well-known autoregressive model: $\mathrm{G}[\mathbf{X}_{l-1};W_l] = \sum_{k=1}^{p} a_k X_{l-k} + W_l$.

For Markovian systems (additive-input or not), a type of dynamic system model can be found for the p^{th}-order multivariate density of the output process X_l. We begin by noting that the conditional density of the output at time l given the values of the previous p states is easily expressed in terms of the input's amplitude distribution. For a given value $\mathbf{X}$ of $\mathbf{X}_{l-1}$, an output equaling X_0 could have arisen from one of several input values, depending on the nonlinear nature of $\mathrm{G}[\mathbf{X};\cdot]$. Notation demands that we represent the i^{th} possibility as an index on the system's transformation rather than on the input: $x_0 = \mathrm{G}_{(i)}[\mathbf{X};w]$, $i = 1,\dots$. The density associated with the output conditioned on state thus equals

$$p_{X_l|\mathbf{X}_{l-1}}(X_0 \mid \mathbf{X}) = \sum_i \left| \frac{\partial}{\partial X_0} \mathrm{G}_{(i)}^{-1}[\mathbf{X};X_0] \right| p_W\left(\mathrm{G}_{(i)}^{-1}[\mathbf{X};X_0]\right)$$

*This equation might suggest that no memoryless transformations can be applied to the input. Such transformations would modify the input's amplitude distribution, but not its whiteness. To keep the notation under control, we use W_l to represent the transformed input. Do note, however, that subsequent results place requirements on this distribution.

$\mathrm{G}_{(i)}^{-1}[\mathbf{X};\cdot]$ denotes the inverse function of $\mathrm{G}_{(i)}[\mathbf{X};\cdot]$. Multiplying this conditional density by the p^{th}-order density of the state gives the $(p+1)^{th}$-order density, which, when integrated over the most distant state, yields the output's multivariate density. The resulting integral equation describes the structural evolution of the system's state.

$$\boxed{p_{\mathbf{X}_l}(\mathbf{X}_0) = \int_{-\infty}^{\infty} p_{X_l|\mathbf{X}_{l-1}}(X_0 \mid \mathbf{X}_{-1})\, p_{\mathbf{X}_{l-1}}(\mathbf{X}_{-1})\, dX_{-p}} \tag{2.3}$$

Here, $\mathbf{X}_k = \mathrm{col}[X_k,\ldots,X_{k-p+1}]$. When the process is stationary, the p^{th}-order joint densities appearing on each side of this equation are equal. To determine when such a stationary output exists, we seek conditions under which this equation has a unique solution. Such conditions revolve around the properties of $p_{X_l|\mathbf{X}_{l-1}}(X_0 \mid \mathbf{X}_{-1})$, which depends on *both* the system's characteristics and the input's amplitude distribution.

For nonlinear systems, stability tests are also equivalent to stationarity tests. Stability of nonlinear difference equations would lead us too far afield; we concentrate on stationarity results here. The Markov system model of Eq. (2.1) can be tested for stability using Lyapunov functions.

Theorem *A stationary distribution exists for a Markovian system if the output is* weakly continuous *(the conditional expected value $\mathscr{E}[g(\mathbf{X}_l) \mid \mathbf{X}_{l-1} = \mathbf{X}]$ is continuous for all bounded, continuous functions $g(\cdot)$) and if there exists a continuous, non-negative function $L(\cdot)$, a* Lyapunov function, *that satisfies $L(\mathbf{X}) \to \infty$ as $||\mathbf{X}|| \to \infty$ and, for some bounded positive constant K,*

$$\begin{aligned} \mathscr{E}[L(\mathbf{X}_{l+1}) - L(\mathbf{X}_l) \mid \mathbf{X}_l = \mathbf{X}] &\leq K, \mathbf{X} \in \mathscr{S} \\ \mathscr{E}[L(\mathbf{X}_{l+1}) - L(\mathbf{X}_l) \mid \mathbf{X}_l = \mathbf{X}] &\leq 0, \mathbf{X} \notin \mathscr{S} \end{aligned}$$

$\mathscr{S}$ denotes a compact set in $\mathbb{R}^p$. Here, $\mathbf{X}_l$ denotes the state of the Markov system at time l: $\mathbf{X}_l =$ $\mathrm{col}[X_l,\ldots,X_{l-p+1}]$.

The quantity $\dot{L}(\mathbf{X})$ denotes the expected change $\mathscr{E}[L(\mathbf{X}_{l+1}) - L(\mathbf{X}_l) \mid \mathbf{X}_l] = \mathbf{X}$ in the system's "energy" as time goes on. The continuity condition is satisfied when the system $\mathrm{G}[\cdot;\cdot]$ is continuous in each state variable and in the input. In some cases, a smooth amplitude distribution for the input W_l suffices.

Example

Consider a first-order linear Markovian system expressed by

$$X_l = aX_{l-1} + W_l$$

where W_l is a zero-mean, white input. Let the Lyapunov function be $L(X) = X^2$. We calculate change $\dot{L}(X)$ in system energy from sample to sample to be

$$\begin{aligned} \dot{L}(X) &= \mathscr{E}[(aX_l + W_{l+1})^2 - (X_l)^2 \mid X_l = X] \\ &= (a^2-1)X^2 + \mathscr{E}[W_l^2] \end{aligned}$$

For the quantity $(a^2-1)X^2 + \mathscr{E}[W_l^2]$ to remain bounded as a function of X, we must require $|a| \leq 1$. For the system energy to increase (exceed zero) *only* within a compact set, we must further require $|a|$ to be strictly less than one. Thus, the output of a simple first-order, linear, Markov system is guaranteed so long as the input's mean-square-value is bounded.

What about inputs having "infinite" variance? We know, for example, that the output of a linear system excited by white noise having a stable distribution (Cauchy, for example) also has the same distribution. To show that in such cases stationarity can result, we must choose a different Lyapunov function. In the Cauchy case, choosing $L(X) = \sqrt{|X|}$ provides an affirmative result.

As this example shows, the theorem requires the existence of only one Lyapunov function to demonstrate stationarity. To show that stationarity does *not* obtain can be much more difficult: We would need to show that no Lyapunov function can exist. When we do demonstrate stationarity, some restriction on system parameters must usually be enforced. For a given Lyapunov function choice, such restrictions are sufficient, but may not be necessary.

Example

Consider the bilinear Markov system described by

$$X_l = (a + bW_l)X_{l-1} + W_l\,,$$

where the input has zero-mean and finite variance σ_W^2. When we choose $L(X) = x^2$ as before, the parameters must satisfy $a^2 + b^2\sigma_W^2 < 1$. If we use $L(X) = |X|$ instead, the condition $\mathscr{E}[|a + bW_l|] < 1$ results. What parameter ranges satisfy both conditions depends on the input's amplitude distribution. For example, when the input has a Laplacian distribution, the latter condition becomes

$$|a| + \left(|b|\sigma_W/\sqrt{2}\right)\exp\left\{-\left|\frac{a}{b}\right|\frac{1}{\sqrt{\sigma_W^2/2}}\right\} < 1\,.$$

In this case, the parabolic choice defines a more restrictive set.

The set over which the energy change equals zero defines the output's stationarity. For instance, in the first example, $\dot{L}(X) = 0$ occurs when $X^2 = \mathscr{E}[W_l^2]/(1-a^2)$, which precisely equals the output variance under stationary conditions. Because of this observation, we can distill an intuitive feel for what the theorem means. Only over a restricted range of state does the system "expand;" over a much larger set the system contracts, tending toward no energy change as "stationary" behavior dominates. This lack of expansion is equivalent to stability. When the distribution if the system's initial condition equals the stationary one, the output's distribution is unchanging.

For a stationary process to be produced by passing white noise through some system, the production must have started in the distant, unremembered past or at some finite time with a particular choice of initial condition dsitribution. Here a quandry arises: How does one distinguish a specific initial condition as being "bad" or "good?" Assuming the stationary distribution is never zero, *any* initial condition could have arisen from a given distribution, the stationary one in particular, thereby resulting in stationary behavior. In other words, transients never occur! While this argument may hold theoretically, the authors prefer starting systems at the Big Bang and concentrate on observing and processing signals long afterwards.

Assuming the theorem's conditions are satisfied, we obtain a fundamental relation that a stationary process's joint amplitude distribution must satisfy when we generate it by passing white noise through a single-input system.

$$\boxed{p_{\mathbf{X}_l^{(p)}}(X_0,\ldots,X_{-(p-1)}) = \int_{-\infty}^{\infty} p_{X_l|\mathbf{X}_{l-1}^{(p)}}(X_0 \mid X_{-1},\ldots,X_{-p})\,p_{\mathbf{X}_l^{(p)}}(X_{-1},\ldots,X_{-p})\,dX_{-p}} \tag{2.4}$$

From one viewpoint, the p^{th}-order amplitude distribution is an eigenfunction having eigenvalue one of the "kernel" $p_{X_l|\mathbf{X}_{l-1}}(X_0 \mid \mathbf{X}_{-1})$. Finding this eigenfunction, either analytically or numerically, seems feasible only for low-order (small p) systems [31],[39: §4.2.4].

Work is simpler in the additive-input case, in which this integral equation becomes

$$p_{\mathbf{X}}(X_0,\ldots,X_{-(p-1)}) = \int_{-\infty}^{\infty} p_W(X_0 - \mathrm{G}_s[X_{-1},\ldots,X_{-p}])\,p_{\mathbf{X}}(X_{-1},\ldots,X_{-p})\,dX_{-p}$$

Here, the kernel's dependence on both the input's amplitude distribution and the system's characteristics become explicit. From this relationship, we can easily see that if the input has an even probability density, so too will the output if and only if the system's input-output relation is odd: $\mathrm{G}_s[-\mathbf{X}] = -\mathrm{G}_s[\mathbf{X}]$. We can also find more explicit relationships using this equation. Take the first-order ($p = 1$) case for example. Evaluating the Fourier transform yields

$$\Phi_X(\nu) = \Phi_W(\nu)\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{j\nu(\mathrm{G}_s[X_{-1}]-X_{-1})}\Phi_X(\nu)\,d\nu dX_{-1}$$

We see that what remains is a kind of Fourier transform in which the system's strictly nonlinear part $\mathrm{G}_s[X]-X$ plays a central role: The more complicated this term is, the more difficult this equation is to use.

When the system is linear, $X_l = aX_{l-1} + W_l$, we obtain the simplest possible result.

$$\Phi_W(\nu) = \frac{\Phi_X(\nu)}{\Phi_X(a\nu)}$$

Given the input distribution, only with difficulty can the output distribution be found from this formula. Curiously, if we have the *output's* amplitude distribution, we can use this result to calculate the input distribution. For example, assuming the output has a stable distribution, which is defined as having a characteristic function of the form $\exp\{-|\nu|^r\}$, $0 < r \leq 2$, we find that the input must also possess a stable distribution of the same degree r. Note, however, that in general that there is no guarantee that the ratio is a characteristic function. We note that, because characteristic functions are positive-definite, they obey $|\Phi(\nu)| \leq \Phi(0) = 1$. Because of the denominator, arbitrary substitution of valid characteristic functions into this ratio may well produce a quantity that exceeds one. In such cases, we have just proven that certain distributions *cannot* describe first-order Markov, linear processes.

Example

The most famous example of this phenomenon is due to Rosenblatt [34: p. 52]. Let's try to force the output of a first-order linear system to have a uniform amplitude distribution. The corresponding characteristic function has the functional form of $\sin\nu/\nu$. When we calculate the ratio $a\sin\nu/\sin a\nu$, we find that the zeros of this function occuring in the demoninator cause difficulty: Unless they are cancelled by zeros in the numerator, the potential characteristic function will be infinite, a property characteristic functions do not possess. Cancellation *only* occurs when the parameter a equals the reciprocal of an integer. Thus, we conclude that a process having a uniform amplitude distribution and a linear, first-order dependence structure can only occur when the correlation coefficient of successive values equals $\pm 1/2, \pm 1/3, \ldots$. Fig. 2.2 portrays an example of this process.

2.2.2 Time-Reversibility

A process's time-reversibility characteristics comprise a special form of dependence structure. A stationary process X_l is *time-reversible* if the multivariate density of the amplitudes at the ordered times $l_1, \ldots, l_N$ equals that of the amplitudes at the times $-l_1, \ldots, -l_N$.

$$p_{X_{l_1}\cdots X_{l_N}}(X_1, \ldots, X_N) = p_{X_{-l_1}\cdots X_{-l_N}}(X_1, \ldots, X_N)$$

Assuming the stationarity equation 2.4 {23} has a solution, the result can be examined to determine its time-reversibility. Examining this evolution equation does not reveal any general structure that the input or the system must possess to produce a time-reversible output. In fact, cases exist wherein the input's amplitude distribution *solely* determines the output's time-reversibility: Non-Gaussian, time-reversible processes cannot be produced by causal linear filters excited by white noise [42]. Only in the Gaussian case is the output time-reversible.

Figure 2.2: A portion of the sample function taken from a first-order linear Markov process is shown in the bottom panel. A histogram estimate of this process's amplitude distribution (10,000 samples) is shown in the top panel. Here, the process is generated according to $X_l = \frac{1}{2}X_{l-1} + W_l$, where $\{W_l\}$ is an IID sequence with each element equaling $\pm\frac{1}{2}$ with probability $1/2$.

Theorem [42] *Assume a linear system is governed by the difference equation*

$$X_l = \sum_{i=1}^{p} a_i X_{l-i} + \sum_{j=0}^{q} b_j W_{l-j}$$

where W_l represents the white noise input. The output is time-reversible if and only if the input is Gaussian or if $a_i = 0$ and the coefficients $\{b_j\}$ obey the symmetry property $b_j = \pm b_{q-j}$, $j = 0, \ldots, \lfloor q/2 \rfloor$.

Only when the system's transfer function has linear phase can a non-Gaussian input produce a time-reversible, non-Gaussian linear process. In all other cases, the filter's output conveys its causality, making measurements of the filter's phase characteristics much easier. Note that a process's time-reversibility depends on *both* the generation system's characteristics and on the amplitude distribution of the white input.

Multivariate distributions that correspond to the joint distribution of non-Gaussian, time-reversible processes are easily found. For example, the so-called elliptically symmetric distributions fall into this class [25]. On the other hand, even conjuring examples of densities for time-irreversible processes can be quite difficult:

The bivariate amplitude density of time-irreversible processes are not symmetric functions—$p_{X_{l_1},X_{l_2}}(\alpha,\beta) \neq p_{X_{l_1},X_{l_2}}(\beta,\alpha)$—but have equal marginals $\left(p_{X_{l_1}}(\alpha) = p_{X_{l_2}}(\alpha)\right)$. In either case, if we can specify the output's multivariate distribution, a system generation model for it can be found from the conditional density of the stationarity equation's kernel $p_{X_l|\mathbf{X}_{l-1}}(X_0 \mid \mathbf{X})$.

The importance of time-irreversible stationary processes rests on their physical existence. Thermodynamic arguments demonstrate that only in very carefully controlled circumstances can time-reversible processes describe physical measurements. Thus, physically plausible models *must* produce time-irreversible processes. What models produce time-irreversible outputs can only be assessed by calculating the multivariate distribution; to emphasize what was mentioned previously, changing the input's amplitude distribution can change the output from a time-reversible to a time-irreversible one.

2.2.3 Statistical Dependence

The dependence structure of a non-Gaussian process both illuminates what system describes its generation and what form the multivariate distribution must take. In fact, modeling physical situations *demands* that a dependence structure be imposed. The authors would be remiss not to point out that if the process was obtained by periodically sampling a continuous-time one, the resulting measurement *cannot* be white. To show this fact, consider the correlation function $R_X(\tau) = \mathscr{E}[X_{l+\tau}X_l]$ of the sequence. It equals the sampled values on the continuous-time process's correlation function: $R_X(\tau) = R_{\widetilde{X}}(\tau T_s)$, where $\widetilde{X}(t)$ denotes the continuous-time process and T_s the sampling interval. For X_l to be white, we at least need the correlation function to correpond to a unit sample. This condition means that the analog signal's correlation function must have zero-crossings uniformly separated by T_s. Assuming that we wish to bey the Sampling Theorem, this situation *only* occurs when the analog signal is ideally bandlimited to precisely the Nyquist frequency. Interestingly, oversampling only increases correlation; to achieve maximal decorrelation, we must in general undersample to try to effect a white sampling sequence.

Broad categories for dependence have been defined based on analytic and theoretical considerations. Unfortunately, testing data against these can be virtually impossible, which makes assuming they apply somewhat tenuous.

Definition *Two (or more) amplitudes are* independent *if they are statistically independent: X_{l_1} is independent of X_{l_2} if $p_{X_{l_1}|X_{l_2}}(X_1 \mid X_2) = p_{X_{l_1}}(X_1)$.*

Note this notion's symmetry: If X_{l_1} is independent of X_{l_2}, then so is X_{l_2} of X_{l_1}. This symmetry contrasts with the asymmetry of the next dependence category.

Definition *The amplitude X_{l_1} is* mean-square independent *of X_{l_2} if the conditional expected value of the first with respect to the second does not depend on the conditioning value.*

$$\mathscr{E}[X_{l_1} \mid X_{l_2} = X] = \mathscr{E}[X_{l_1}]$$

Mean-square independence is *not* a symmetric notion: $\mathscr{E}[X_{l_1} \mid X_{l_2} = X] = \mathscr{E}[X_{l_1}] \not\Longrightarrow \mathscr{E}[X_{l_2} \mid X_{l_1} = X] = \mathscr{E}[X_{l_2}]$. As an example, consider a time-irreversible process defined by $X_l = \left(\frac{1}{2} - \frac{1}{2}X_{l-1}\right)W_l$, where $W_l = \pm\frac{1}{2}$ with equal probability. Clearly, $\mathscr{E}[X_l \mid X_{l-1}] = 0$, which also means that $\mathscr{E}[X_l] = 0$. The time-reversed system has the form $X_{l-1} = 1 - 2|X_l|$ [26]. Now, $\mathscr{E}[X_{l-1} \mid X_l] = 1 - 2|X_l|$. Thus, we say that X_l is mean-square independent of X_{l-1}, but X_{l-1} is not mean-square independent of X_l. It is mean-square independence's asymmetry that underlies the conditional expected value's utility in exploring a signal's time-reversibility.

Definition *Two amplitudes are* strictly uncorrelated *if the expected value of the product of arbitrary functions of each amplitude equals the product of the expected values of each function for all reasonable functional choices.*

$$\mathscr{E}[g_1(X_{l_1})g_2(X_{l_2})] = \mathscr{E}[g_1(X_{l_1})]\,\mathscr{E}[g_2(X_{l_2})]\,,\forall\left|\mathscr{E}[g_i(X_l)]\right| < \infty$$

Weakly uncorrelated (uncorrelated) *amplitudes occur when we consider the expected values of the amplitudes directly.*

$$\mathscr{E}[X_{l_1}X_{l_2}] = \mathscr{E}[X_{l_1}] \cdot \mathscr{E}[X_{l_2}]$$

The notion of correlation is symmetric, meaning that we can say two amplitudes are uncorrelated without ambiguity.

These dependence categories form a clear progression, wherein each implies others.

$$\boxed{\text{independence} \Longrightarrow \text{m.s. independence} \Longrightarrow \text{strongly uncorrelated} \Longrightarrow \text{uncorrelated}}$$

Independence is simultaneously the simplest, the most powerful, and the most difficult to verify empirically. Because the minimum mean-squared error predictor of a random variable X given another, say Y is the conditional expected value $\mathscr{E}[X \mid Y = y]$, mean-square independence means that providing an amplitude value does not reduce the mean-squared error in predicting another. The notion of strongly uncorrelated is not used frequently; uncorrelated and correlated amplitudes correspond to dependence categories prevalent in the second-order radom process theory.

Spanning these categories, succint dependence structures have been defined that ease the transistion to theoretical development and incorporating model-based notions.

Markov dependence. When the conditional distribution of a process's amplitude at time l given the process's entire past $\{X_{l-1}, X_{l-2}, \ldots\}$ functionally depends *only* on the p most recent values, the process is *Markovian* of order p.

$$p_{X_l|X_{l-1},\ldots}(X_0 \mid X_{-1}, \ldots) = p_{X_l|X_{l-1},\ldots,X_{l-p}}(X_0 \mid X_{-1}, \ldots, X_{-p})$$

From this definition, it easily follows that Markov dependence extends to any set of past *adjacent* p values.

$$p_{X_l|X_{l-\tau-1},\ldots}(X_0 \mid X_{-(l+1)}, \ldots) = p_{X_l|X_{l-\tau-1},\ldots,X_{l-\tau-p}}(X_0 \mid X_{-(l+1)}, \ldots, X_{-(l+p)}), \tau \geq 0$$

Be that as it may, this dependence structure does *not* mean that the process depends only on a *subset* of adjacent p values. For example, if $p > 1$, the conditional density $p_{X_l|X_{l-\tau}}(X_0 \mid X_{-\tau})$ usually depends on $X_{-\tau}$ for all τ, even if $\tau > p$. Thus, a Markov process's memory—the time-span over which a value depends on a previous one—is usually infinite.

Systems that produce Markov dependence structures have the form expressed by Eq. (2.1) $\{21\}$. One important special case of which is the *linear autoregressive* process, described by the input-output relation $X_l = \sum_{k=1}^{p} a_k X_{l-k} + W_l$. Including more than one input amplitude value in the system's input-output relation usually means that Markovian dependence does *not* result.

The Markov structure for an additive-input, stationary process (Eq. 2.2) $\{21\}$ allows the explicit evaluation of the process's multivariate amplitude distribution for *any* order. Let $\mathbf{X}_\tau^{(k)}$ denote $\operatorname{col} X_\tau, X_{\tau-1}, \ldots, X_{i-k+1}$, the point at which we evaluate the k^{th}-order density. The index τ indicates the lag relative to the time l of the temporal origin of the associated amplitude vector. Assume the process has Markovian order p and that we want the k^{th}-order amplitude distribution. For the moment, $k > p$. The joint distribution of the amplitude vector $\mathbf{X}_l^{(k)}$ equals

$$p_{\mathbf{X}_l^{(k)}}(\mathbf{X}_0^{(k)}) = p_{X_l|\mathbf{X}_{l-1}^{(k-1)}}(X_0 \mid \mathbf{X}^{(k-1)}) p_{\mathbf{X}_{l-1}^{(k-1)}}(\mathbf{X}_{-1}^{(k-1)})$$

Because of the Markovian structure, the $(k-1)^{th}$-order conditional density in this expression equals the p^{th}-order conditional density. The additive input structure allows explicit calculation of this conditional density.

$$p_{X_l|\mathbf{X}_{l-1}^{(p)}}(X_0 \mid \mathbf{X}_{-1}^{(p)}) = p_W(X_0 - \mathrm{G}_s[\mathbf{X}_{-1}^{(p)}])$$

This reduction of the multivariate density into the product of a known conditional density and a lower order multivariate density can be repeated until a p^{th}-order multivariate density results. *Assuming* we can calculate this density, the final expression for the desired multivariate density is

$$\boxed{p_{\mathbf{X}_l^{(k)}}(\mathbf{X}^{(k)}) = p_{\mathbf{X}_{l-k+p}^{(p)}}(\mathbf{X}_{-k+p}^{(p)}) \prod_{i=1}^{k-p} p_W(X_{-i+1} - \mathrm{G}_s[\mathbf{X}_{-i}^{(p)}])}$$

When the order of the multivariate density is less than the Markovian order p, we can simply integrate the p^{th} density over the unwanted components. Thus, regardless of the number of and selection of process values for which we want the joint probability distribution, we can calculate the multivariate density if we know the p^{th} order joint distribution and the amplitude distribution of the white noise input. Note how this expression exemplifies the infinite-duration dependence structure of the Markov process: No matter how remote the lag τ, the pair $X_l, X_{l-\tau}$ are dependent through the telescope-like product terms.

Example

Consider the first-order, linear autoregressive case: $X_l = aX_{l-1} + W_l$. The k^{th}-order multivariate density has the expression

$$p_{\mathbf{X}_l^{(k)}}(\mathbf{X}^{(k)}) = p_X(X_{-k+1}) \prod_{i=1}^{k-1} p_W(X_{-i+1} - aX_{-i})$$

When the white input has a stable amplitude distribution, the output has the same distribution with a different "variance": When a stationary distribution exists, the variance of the output equals that of the input divided by $1 - a^2$. Note that among stable distributions, only the Gaussian has finite variance. Be that as it may, we can derive the multivariate distribution for a linear, first-order Markov Cauchy process to be

$$p_{\mathbf{X}_l^{(k)}}(\mathbf{X}^{(k)}) = \frac{\sqrt{1-a^2}}{\pi^k \sigma^k} \frac{\sigma^2}{\sigma^2 + (1-a^2)X_{l-k+1}^2} \prod_{i=1}^{k-1} \frac{\sigma^2}{\sigma^2 + \left(X_{l-i+1} - aX_{l-i}\right)^2}$$

Note the asymmetry in this joint distribution, which indicates the process's time-irreversibility.

q-dependence. This property applies to processes generated by systems that depend only on the most recent q input values.

$$X_l = \mathrm{G}_i[W_l, \ldots, W_{l-q+1}]$$

Thus, amplitude values separated by q or more samples are independent. For statisticians, q-dependence refines the definition of *moving average* processes. In the signal processing terminology of linear systems, such finite-memory systems are said to be *FIR* (have Finite-duration Impulse Responses). Correlation matrices of q-dependent processes are banded, with nonzero correlation extending only over $q - 1$ diagonals above and below the main one.

Mixing. Mixing structures rival Markovian structures in theoretical importance; in those situations where these structures both occur, very powerful results obtain. For a formal definition of mixing, define $\mathscr{X}_a^b$ to be the σ-field generated by the amplitudes $\{X_l, a \leq l \leq b\}$. With these σ-fields, we can (conceptually) define individual, conditional, and joint probabilities of the occurence of particular sets defined over time-restricted portions of the process. The fundamental notion of mixing is that the joint probability of two sets of process amplitudes asymptotically factors—the two sets become independent—as the temporal separation between the sets increases. Formally, let $A \in \mathscr{X}_a^b$ and $A' \in \mathscr{X}_{a'}^{b'}$ be sets that are members of σ-fields defined over time intervals. Let τ denote the *shift operator* that operates on set so that $\tau A \in \mathscr{X}_{a+1}^{b+1}$.

Definition *The process X_l is said to be* mixing *if, for all A, A',*

$$\lim_{\tau\to\infty} \Pr[A \cap \tau^\tau A'] = \Pr[A] \cdot \Pr[A']$$

where τ^τ denotes a shift by τ (applying the shift operator τ times).

Unfortunately, many special measures of mixing have been developed over the years [3], mudding the waters somewhat since Rosenblatt's original formulation of the idea [32]. These variations assess, using different criteria, the rate by which the dependence of two temporally separated amplitude sets decreases as the sets become more widely separated. For the following definitions, we define the sets A, A' over the special σ-fields $\mathscr{X}_{-\infty}^{0}$ and $\mathscr{X}_{0}^{\infty}$, respectively. Emphasis here on time zero is arbitrary; any finite value could be used.

Definition *A process is said to be* strongly mixing *if the mixing coefficient $\alpha(\tau)$ asymptotically equals zero.*

$$\sup_{A,A'} \left| \Pr[A \cap \tau^\tau A'] - \Pr[A]\Pr[A'] \right| = \alpha(\tau) \xrightarrow{\tau\to\infty} 0$$

A stationary process is said to be ψ-mixing *if*

$$\sup_{A,A'} \frac{\left| \Pr[A \cap \tau^\tau A'] - \Pr[A]\Pr[A'] \right|}{\Pr[A]\Pr[A']} = \psi(\tau) \xrightarrow{\tau\to\infty} 0$$

A process is said to be uniformly strongly mixing, uniformly mixing, *or* ϕ-mixing *if the two sets are asympotically independent in a slightly different way than in strong mixing.*

$$\sup_{A,A'} \left| \Pr[\tau^\tau A' \mid A] - \Pr[A] \right| = \phi(\tau) \xrightarrow{\tau\to\infty} 0$$

A process is said to be ρ-mixing *if the maximal correlation coefficient between the two σ-fields is asympotically zero.*

$$\sup_{U \in L_2(\mathscr{X}_{-\infty}^{0}),\, V \in L_2(\mathscr{X}_{\tau}^{\infty})} \frac{\left|\operatorname{cov}(U,V)\right|}{(\mathscr{V}U)^{1/2}(\mathscr{V}V)^{1/2}} = \rho(\tau) \xrightarrow{\tau\to\infty} 0$$

Here $L_2(\mathscr{X}_a^b)$ denotes the collection of all second-order (finite variance) random variables measurable with respect to $\mathscr{X}_a^b$.

The various mixing coefficients obey interesting inequalities [3, 14].

$$\begin{aligned}
&\alpha(\tau) \le \frac{1}{4} \quad \phi(\tau) \le 1 \quad \rho(\tau) \le 1 \\
&4\alpha(\tau) \le 2\phi(\tau) \le \psi(\tau) \\
&4\alpha(\tau) \le \rho(\tau) \le \psi(\tau) \\
&\rho(\tau) \le 2\phi^{1/2}(\tau)
\end{aligned}$$

For stationary Gaussian processes, $\alpha(\tau) \le \rho(\tau) \le 2\pi\alpha(\tau)$, which means that Gaussian processes are strongly mixing if and only if they are ρ-mixing [18]. From these inequalities, we glean the following maximal (no other implications exist) relations among the various types of mixing conditions [3].

ψ-mixing	$\Longrightarrow$	uniform mixing	$\Longrightarrow$	strongly mixing
ψ-mixing	$\Longrightarrow$	ρ-mixing	$\Longrightarrow$	strongly mixing
strongly mixing	$\Longrightarrow$	mixing		

Strong mixing was the first defined condition and claims the high ground; the more recent definitions yield more stringent criteria, leaving the adverb in "strongly mixing" somewhat inappropriately chosen.

If stationary processes can be shown to be strongly mixing, many interesting properties can result. Among process classes have been shown to be strongly mixing are white noise, Markov processes (if they are purely nondeterministic)* [33: p. 195], and Gaussian processes having continuous and positive power spectra [18]. A Markov process cannot be ϕ-mixing without its mixing coefficients decreasing exponentially [8]: $\phi(\tau) \leq ca^{\tau}$, $0 < a < 1$ and c a positive constant. If X_l is strongly mixing, then the process $Y_l = \mathrm{G}_i[X_l, \ldots, X_{l-q+1}]$ is strongly mixing [34: p. 79]. This result means that all q-dependent processes are strongly mixing.

We defer discussing the interaction of mixing and ergodic theory to §2.2.4 {30}. For now, we comment on an issue that should worry the signal processor: "When can I produce estimates from dependent data and how does the dependence affect my estimate's accuracy?" Suffice it that if a stationary sequence is strongly mixing, then estimation procedures, such as kernel estimates for densities and regression functions and distributional parameters, converge, but at a slower rate. Even the Central Limit Theorem survives!

Theorem [15: p. 316] *Let $S_k = \sum_{l=0}^{k-1} X_l$ be the cumulative sum of a strongly mixing process having mixing coefficient $\alpha(k)$. Let $P_k(\cdot)$ denote the distribution function of the cumulative sum normalized as $N_k^{-1}S_k - M_k$, where $\lim_{k\to\infty} N_k = \infty$. If $P_k(\cdot)$ converges to a non-degenerate distribution function $P(\cdot)$, then $P(\cdot)$ is stable. If this stable distribution's parameter equals β, then $N_k = k^{1/\beta} g(k)$, where $g(k)$ is slowly varying as $k \to \infty$.*

Theorem [15: pp. 346–7] *Let X_l be a zero-mean, strongly mixing stationary process with mixing coefficient $\alpha(\tau)$ and variance σ^2 that has the property $\mathscr{E}[|X_l|^{2+\delta}] < \infty$ for some $\delta > 0$. If*

$$\sum_{\tau=1}^{\infty} \alpha(\tau)^{\delta/(2+\delta)} < \infty\,,$$

then $(k^{1/2}\sigma)^{-1} \sum_{l=0}^{k-1} X_l$ converges to the unit normal.

Putting these results in context, we found that distributions of sums of random variables are never far away from densities of infinitely divisible random variables. For those situations for which these asymptotic densities approach a limiting distribution, amplitude averages taken from observing strongly mixing processes also obey a Central Limit Theorem.

2.2.4 Ergodicity

Perhaps the most subtle structural component of a random process is its ergodicity. Intuitively, an ergodic process is one in which estimates of its characteristics—expected value, covariance, amplitude distribution, etc.—have meaning. Nonergodic processes have the somewhat counterintuitive property that estimates *cannot* converge; their dependence structure is such that convergence, no matter how many observations are available, never occurs. Presumably, nature is not so capricious to not allow us to learn from experiment its underlying structure: We *believe* that physically relevant models should produce ergodic processes. On the other hand, we can produce nonergodic models without difficulty. For example, all spherically symmetric joint amplitude distributions (save the Gaussian) correspond to nonergodic processes [41]. Thus, testing models for ergodic behavior should be high on the checklist for reasonableness and applicability.

Definition *The process X_l is said to be* ergodic *if temporal averages of a function $g(\cdot)$ of some finite collection C_X observations converges to its expected value.*

$$\lim_{\tau\to\infty} \frac{1}{\tau} \sum_{\tau=0}^{\tau-1} g(\tau^{\tau} C_X) = \mathscr{E}[g(C_X)] \text{ a.s.}$$

This collection consists of a selection of process amplitudes $\{X_{l_0}, X_{l_1}, \ldots\}$. We clearly need to require that $\mathscr{E}[|g(C_X)|] < \infty$.

* A process is *purely nondeterministic* if a set derived from currently occuring amplitudes is aymptotically independent of a set derived from amplitudes occuring in the distant past. Formally, if $A \in \mathscr{X}_a^b$, $\Pr[A \mid \mathscr{X}_{-\infty}^{\tau}] \to \Pr[A]$ as $\tau \to -\infty$ almost everywhere.

This definition formalizes our notion of meaningful measurements, but does not directly relate ergodicity to a process's structural properties. To obtain results, two important theorems relate ergodicity to joint-amplitude-distribution and system descriptions of random processes.

Theorem *Let A and A' be sets contained in the σ-algebras $\mathscr{X}_a^b$ and $\mathscr{X}_{a'}^{b'}$ respectively, which are generated from the stationary random process X_l over the interval expressed by the subscripts and superscripts. Let τ denote the shift operator $\{28\}$ that can be applied to these sets and τ^{-1} be the operator's inverse image: $\tau^{-1}A = \{\omega\colon \tau\omega \in A\}$. X_l is ergodic if and only if**

$$\lim_{\tau\to\infty}\frac{1}{\tau}\sum_{\tau=0}^{\tau-1}\Pr[A\cap\tau^{-\tau}A'] = \Pr[A]\Pr[A']$$

for all choices of A, A' and intervals $[a,b], [a',b']$.

Clearly, this theorem is satisfied for all mixing processes; thus, if one can show that a process is mixing (Def. 2.2.3 $\{29\}$), estimates computed from it will be meaningful. Thus, white noise and purely nondeterministic Markov processes are ergodic because they are strongly mixing (hence mixing). Note, however, that mixing more than satisfies the theorem's requirements.

In any case, this theorem can be used to determine if a process is not ergodic from its multivariate amplitude distribution. Let the sets A, A' denote small intervals in $\mathbb{R}^n$ centered at $\mathbf{X}$ and $\mathbf{X}'$, respectively. The theorem says, after making appropriate smoothness assumptions on the amplitude distribution, that if the process is ergodic, then

$$\boxed{\lim_{\tau\to\infty}\frac{1}{\tau}\sum_{\tau=0}^{\tau-1} p_{\{X_a\cdots X_b\},\{X_{a'+\tau}\cdots X_{b'+\tau}\}}(\mathbf{X},\mathbf{X}') = p_{\{X_a\cdots X_b\}}(\mathbf{X})\,p_{\{X_{a'}\cdots X_{b'}\}}(\mathbf{X}')}$$

In particular, for the bivariate density, we have the ergodicity test

$$\boxed{\lim_{\tau\to\infty}\frac{1}{\tau}\sum_{\tau=0}^{\tau-1} p_{X_0,X_\tau}(X,X') = p_X(X)p_X(X')}$$

An equivalent test of the bivariate distribution for mixing is

$$\lim_{\tau\to\infty} p_{X_0,X_\tau}(X,X') = p_X(X)p_X(X')$$

Note that if an amplitude distribution "passes" these tests, that does *not* mean that they are ergodic or mixing. Because the theorem demands *all* sets selected from the σ-algebras satisfy the condition, a passing distribution is only *consistent* with ergodicity. On the other hand, a failing distribution is not ergodic, presumably dismissing the corresponding process from consideration as a viable model for reality.

Example

Consider the multivariate Gaussian density expressed by $\mathscr{N}(\mathbf{0},\mathbf{K}_X)$. When the covariance function $K_X(\tau)$, which corresponds to entries in the covariance matrix $\mathbf{K}_X$, approaches zero for large lags, the multivariate density factors: $\lim_{\tau\to\infty}\mathscr{N}(\mathbf{0},\mathbf{K}_X) = \prod\mathscr{N}(0,\sigma_X^2)$. When we consider two groups of amplitudes separated by lag τ, increasing the separation creates a covariance matrix that asymptotically has two square matrices on the diagonal and zero-valued entries elsewhere. Again, the multivariate density factors and the ergodicity test is passed for *all* choices of groups. Thus, stationary Gaussian processes are mixing, hence ergodic.

*The seemingly odd appearance of $\tau^{-1}A'$ allows us to focus on a fixed set A' that corresponds to more and more remotely shifted set.

Example

Now consider the elliptically symmetric bivariate distribution having Laplacian marginals [25].

$$p_{X_0,X_\tau}(X,X') = \frac{1}{\pi\sigma^2(1-\rho_\tau^2)^{1/2}} K_0\left(\frac{2(X^2+(X')^2-2\rho_\tau XX')}{\sigma^2(1-\rho_\tau^2)}\right)$$

Here, $K_0(\cdot)$ denotes the modifed Bessel function of the second kind. All densities in this class are parameterized by the correlation coefficient ρ_τ between X_0 and X_τ. Letting this coefficient approach zero (as in the Gaussian example just described), the density does *not* factor, meaning that the process is not mixing. Applying the more direct, but harder to use, sum-of-densities test, this density still does not pass. Thus, as expected (the only ergodic elliptically symmetric process is the Gaussian), the process corresponding to this density is not ergodic.

Theorem *Assume the conditions apply for a white-noise driven Markov system to produce a stationary output (Thm. 2.2.1 {22}). Furthermore, assume that*

1. *zero is an equilibrium point of the zero-input system* ($0 = \mathrm{G}[0;0]$);
2. *the amplitude density of the input* W_l *is non-zero in some open interval that includes the origin;*
3. *the transformation* $\mathrm{G}[\cdot;\cdot]$ *is continuous everywhere and continuously differentiable in a neighborhood of the origin;*
4. *for some* $\varepsilon > 0$, $\mathscr{E}[||\mathrm{G}[\mathbf{X};W_l]||] \leq \varepsilon$ *for all values of* $\mathbf{X}$ *lying in the system's state space.*

Under these conditions, the output X_l *produced by* $\mathrm{G}[\mathbf{X}_{l-1};W_l]$ *has exponentially decreasing mixing coefficients* $\rho(\tau)$*, and hence is ergodic.*

Thus, stable systems that, like linear systems, produce zero output with no input and satisfy continuity properties yield an ergodic output. All stable linear systems fall into this category; chaotic systems do not (their output to zero-input never dies). Because the theorem does not provide necessary and sufficient conditions, one wonders how nonlinear systems fit. In the case of additive-input Markov systems, if $\mathrm{G}_s[\mathbf{X}_{l-1}]$ is bounded and the input's amplitude distribution function is continuous, then the output is ϕ-mixing [9]. Thus if the output is stationary, it is ergodic.

It is unfortunate these two theorem's conditions differ. The theorem that applies to the joint amplitude distribution is more direct and exhaustive. The system-based one is not necessarily comprehensive: other systems and inputs may exist that produce ergodic outputs. More work is needed in this area to produce a comprehensive ergodic theorem.

2.3 Simple Waveform Processes

Special classes of processes have found application to engineering problems, both practical and theoretical. Results even for these admittedly special cases are spotty; this situation indicates the unevenness of the non-Gaussian terrain.

White noise. The most elementary waveform process—it has the simplest structure—is white noise. This process consists of a sequence independent, identically distributed, amplitudes that can have any probability distribution. Note that this definition, with the specification of *independent* amplitudes, is stronger than some definitions that demand only uncorrelated values. In most cases, we take the process to have zero mean. Clearly, white noise is stationary and ergodic. As mentioned previously, white noise cannot be obtained by sampling (in a realistic fashion) a continuous-time process {26}. This model represents a convenient fiction, especially for the input to system models for generating processes.

A white-noise-type process having subtly more dependence occurs in least-squares estimation. The least squares predictor of X_l based on the observation of $\{X_b,\dots,X_a\}$, $a \leq b < l$, is the conditional expected value:

$\widehat{X_l} = \mathscr{E}[X_l \mid X_b, \ldots, X_a]$. The estimation error forms a mean-squared independent sequence $\{26\}$, which of course is not a white sequence. This next-of-kin to white noise that is produced by least-squares estimation errors is known as an *innovations* sequence.

Linear processes. From a system's viewpoint, the simplest nontrivial dependence structure is expressed by passing white noise through a linear system. When the system is strictly stable (all poles lie inside the unit circle), the resultant is a stationary *linear process*. For a finite variance input to yield a finite variance output, we must impose a somewhat more restrictive condition: $\sum hl^2 < \infty$, where hl denotes the system's unit-sample response. In this case, a linear system generates a linear process from white noise according to the convolution sum.

$$X_l = \sum_k hl - kW_k$$

Note that the filter need not be causal to produce a well-defined linear process.

The most frequently used linear process is the *ARMA*—autoregressive-moving average—process. Here, the system's input-output relation is expressed by the difference equation

$$X_l = \sum_{i=1}^{p} a_i X_{l-i} + \sum_{j=0}^{q} b_j W_{l-j}$$

The dependence structure of such processes is represented by the notation ARMA(p, q). Here, p equals the number of poles in the system's transfer function, q the number of zeros. A more specialized, but frequently used, case results when no zeros occur ($q = 0$). In this case, we have a pure AR process, which is symbolized by AR(p). In both ARMA and AR processes when all poles lie inside the unit circle, the process is strongly mixing, which means that the output is ergodic. When no poles occur, the ARMA model becomes a MA(q) model, which produces a linear q-dependent output. This process is clearly ergodic so long as its order q is finite $\{30\}$. As described previously, the only time-reversible linear processes are linear Gaussian ones, produced when the white-noise input is Gaussian, and when the system has a linear phase transfer function (no poles and either even- or odd-symmetric moving-average coefficients $\{b_j\}$) (Thm. 2.2.2 $\{25\}$). Typically, linear non-Gaussian processes are time-irreversible.*

Stable processes. Having more theroretical than practical interest are stable processes. Here, the white-noise input to a linear system has an amplitude distribution drawn from the collection of stable probability densities. Because a linear combination of stable random variables produces a stable random variable having the same "form": weighted sums of Gaussians are Gaussian, weighted sums of Cauchy random variables has a Cauchy distribution. Thus, a linear system's output, when driven by stable white noise, is also stable and has a marginal amplitude distribution of the same form—the parameters can differ—as the input's. All stable processes are linear, and these processes can be used to explore the dependence structures of linear processes. However, the *only* stable density having a finite variance is the Gaussian, which means that non-Gaussian stable processes *all* have the somewhat unrealistic property of infinite power.

Chaotic processes. Using the word "process" along side "chaotic" may seem to clash. However, chaotic signals, which are produced by zero-input, nonlinear systems, do share some common properties with random processes.

Consider the zero-input input-output relation expressed by

$$X_l = \mathrm{G}_s[X_{l-1}, \ldots, X_{l-p}]$$

Assume this system's *initial condition*, values for the states $X_{l-1}, \ldots, X_{l-p}$, have some joint distribution. What is the output's asymptotic distribution, if it exists? In other words, we ask when repeated solutions the state-evolution equation (2.3) $\{22\}$ converge. In some cases, no limit exists, and no such *stationary density* can be

*It is somewhat curious that the dependence structure expressed by elliptically symmetric distributions contains *all* least-squares predictors that turn out to be linear, and yet, because the corresponding processes are time-reversible, they cannot be produced by linear systems having a white-noise input [2]!

defined. If the limit exists, then the evolution equation defines the limiting distributions.

$$p_{\mathbf{X}^{(p)}}(\mathbf{X}_0^{(p)}) = \int_{-\infty}^{\infty} p_{X_l|\mathbf{X}_{l-1}^{(p)}}(X_0 \mid \mathbf{X}_{-1}^{(p)}) p_{\mathbf{X}^{(p)}}(\mathbf{X}_{-1}^{(p)})\, dX_{-p}$$

Recall the notation $\mathbf{X}_\tau^{(p)} = \operatorname{col} X_\tau, \dots, X_{\tau-p+1}$. The conditional density expresses the input-output relation $\mathrm{G}_s[\cdot]$.

Because stable linear systems settle to zero from any initial condition, linear processes have quite boring limiting distributions: $p_{\mathbf{X}^{(p)}}(\mathbf{X}_0^{(p)}) = \prod_\tau \delta(X_\tau)$. Nontrivial results can emerge in the nonlinear case.

Example

Consider the over-used example of the input-output relation expressed by the Hénon map.

$$X_l = 4X_{l-1}\left(1 - X_{l-1}\right)$$

If the initial condition lies outside the interval $[0,1]$, the system's output is unbounded: The system acts as if it is unstable. If, however, the system is started within this interval, the output's amplitude remains in $[0,1]$ forever. To investigate solutions to the evolution equation, the conditional distribution specifying the system has the simple form

$$p_{X_l|X_{l-1}}(X_0 \mid X_{-1}) = \delta\left(X_0 - 4X_{-1}(1 - X_{-1})\right)$$

To perform the required integration, we need to express X_{l-1} in terms of X_l. Using the quadratic formula, we find that $X_{l-1} = \frac{1}{2}\left(1 \pm \sqrt{1 - X_l}\right)$. A relation like $\delta\left(y - f(x)\right)$ equals $\sum_i \left|df^{-1}(y_i)/dy\right| \delta\left(x - f^{-1}(y_i)\right)$, where $\{y_i\}$ denotes the solution set of $x = f(y)$. The evolution equation thus constrains any stationary distribution to satisfy

$$p_X(X) = \frac{1}{4\sqrt{1-X}}\left[p_X\left(\tfrac{1}{2}(1 + \sqrt{1-X})\right) + p_X\left(\tfrac{1}{2}(1 - \sqrt{1-X})\right)\right]$$

Substitution shows that

$$p_X(X) = \frac{1}{\pi\sqrt{X(1-X)}}$$

solves this equation. This result means that the output of the zero-input system expressed by the Hénon map can rattle around inside the interval $[0,1]$ forever. Fig. 2.3 demonstrates a typical signal generated according to this difference equation. We should note that not all initial conditions lead to interesting rattles. If X_0 equals either 0 or $3/4$, the output equals these values forever; furthermore, initial conditions that produce these values later in time yield an output that gets "stuck." These initial conditions are sparse (the probability of such a value being chosen from the stationary amplitude distribution is zero.)

In such situations, when a nontrivial amplitude distribution satisfies the evolution equation corresponding to a zero-input, hence deterministic, system, the signal thus produced is said to be *chaotic*.

Chaotic signals have been segregated from stochastic ones. A relation between chaotic signals, which are generated deterministically, and stochastic ones, generated by passing white noise through some system, would seem only remotely possible. When we consider the *time-reversed* system, the one that generates signals in the opposite temporal direction, a random process characterization must result in at least some cases. Take the just presented Hénon map example; the systems that generate the *same signal values* forward and backward in time are

$$\begin{aligned} X_l &= 4X_{l-1}\left(1 - X_{l-1}\right) \\ X_{l-1} &= \frac{1}{2}\left(1 + W_l\sqrt{1 - X_l}\right),\ \Pr[W_l = \pm 1] = \frac{1}{2} \end{aligned}$$

Figure 2.3: Using the initial condition $X_0 = 1/\sqrt{2}$ in the Hénon difference equation produces the depicted signal. Visually, this signal seems odd somehow, having some structure that does not correspond to intuitive notions of what random signals should appear to be.

The second equation results from the sign ambiguity of the quadratic formula applied to the first equation to find X_{l-1} in terms of X_l. The equally likely choice for the probability assignment creates an amplitude distribution for the signal that agrees with the chaotic system's stationary distribution. Because of this kind of constraint, a nonlinear Markov model for the signal *must* be used to describe how to generate the chaotic signal in the opposite temporal direction. The authors chose temporal "direction" arbitrarily in this example; nature is not so arbitrary. If we impose causality as a constraint to help define the "right" model for a given set of observations, one of these models—either the stochastic or deterministic choice—becomes the preferred model. Thus, the observations' time-reversibility structure determines which model describes natural phenomena!

Example

Another, more interesting, example of this duality is the process discussed by Rosenblatt [34: p. 52]. Here, with K some nonzero integer (positive or negative), two systems produce the identical signal values.

$$\begin{aligned} X_l &= \frac{1}{K}X_{l-1} + W_l,\ \Pr[W_l = k/K] = \frac{1}{K},\ k = 0,\ldots,|K|-1 \\ X_{l-1} &= \big(KX_l\big) \bmod 1 \end{aligned}$$

The first difference equation describes a linear autoregressive (hence Markov) process. Using the techniques used in a previous example $\{24\}$ of this process, we find that the generated signal's amplitude distribution is uniform over $[0,1]$. The (deterministic) equation describing how to generate this signal in the opposite temporal direction is also known as a congruential uniform random number generator. Apparently, chaotic signals have been used for decades to model stochastic phenomena! The stochastic counterpart indicates that the successive outputs of this random number generator are correlated (correlation coefficient $1/K$). This correlation is one reason why K is usually quite large in applications.

The authors emphasize that these examples demonstrate that *randomness and determinism are not dichotomous concepts*. Either model can describe the same set of observations in these cases. Tests that attempt to distinguish chaotic signals from random ones cannot succeed unless they take into account the direction of time. Signal processing researchers are intensely investigating the full picture of how chaos and randomness are related.

2.4 Structure of Point Processes

This book has concentrated so far on waveform processes—time series—wherein the observations are a stochastic sequence derived by somehow sampling a waveform. This mindset limits models and algorithms, in which the non-Gaussian world only consists of difficult to characterize waveforms. We turn now to point processes, a class of processes that have *no* waveform. Such processes cannot be Gaussian, meaning that the non-Gaussian process category must embrace them. Loosely speaking, a point process consists of a sequence of events occuring with respect to some continuous parameter, with time being the primary example here; the times at which events occur completely characterizes the process. In some point process models, a value (possibly random) is associated with each event; in this case, both event times and event values comprise the point process. Such nonwaveform processes occur in interesting applications, the primary one being neuroscience. For example, many neurons in the central nervous system work by assimilating their inputs to produce a sequence of identical-waveform pulses—action potentials—that propagate actively (without waveform dispersion) through their output cable—axon—to serve as the input to one or several neurons. In many neural groups, the times at which action potentials occur seem random. In sensory systems, for example, repeated presentations of the same external input do not produce the same sequence of action potentials. Clearly, information is conveyed by action potential sequences, else our brains would be incapable of assimilating information about the environment. Analyzing and modeling neural data amounts to developing a point process characterization and then exploiting it. How do we describe a process having no waveform, and how can information be encoded into and extracted from such processes?

2.4.1 Definitions

Point processes are described from the viewpoint of observing them. The process is born at $t=0$. We denote by N_{t_1,t_2} the number of events occuring in the *observation interval* $[t_1,t_2)$. Observations have no meaning that occur prior to the process's birth; we require $t_1 \geq 0$.[*] A shorthand notation for the total number $N_{0,t}$ of events occuring up to time t is N_t, which is known as the *counting process*. The *occurence time* of the n^{th} event we denote by W_n and the N_t event times we denote by the *event vector* $\mathbf{W} = \operatorname{col} W_1,\dots,W_{N_t}$. Note that the number of elements in this vector can be variable. The n^{th} *interevent interval* τ_n equals the time between the n^{th} and $(n-1)^{th}$ events: $\tau_n = w_n - w_{n-1}$.[†] The *history* of a point process is the couple $\{N_t,\mathbf{W}\}$: the number of events that have occurred up to time t and when they occurred.

Definition *A* regular *point process has the property that the probability of an event occuring in the interval $[t,t+\delta)$ conditioned on the process's history is asymptotically ($\delta \to 0$) proportional to the length of the interval.*

$$\Pr\left[N_{t,t+\delta} = 1 \mid N_t = n, \mathbf{W} = \mathbf{w}\right] = \lambda\left(t;n;\mathbf{w}\right)\delta + O\left(\delta\right)$$

Here, $O(\cdot)$ denotes a quantity that approaches zero faster than its argument: $\lim_{x\to 0} O(x)/x = 0$. The quantity $\lambda(t;n;\mathbf{w})$ denotes the point process's intensity, *an expression of the instantaneous rate at which events occur and how this rate depends on process history.*

In waveform processes, the stochastic process is defined by the multivariate distribution of arbitrarly selected amplitudes. For a regular point process, the intensity defines it: We distinguish point processes *solely* by their intensity definitions. Save for the Poisson process, the intensity depends on history, and expresses the process's dependence structure. Note that because the intensity is proportional to probability, *all* intensities are non-negative and bounded, and have units of events/s. When an intensity equals zero, *no* events occur. Because the intensity defines how the event occurrence rate varies with previous event occurrences, which are governed by the point process's probability law, the intensity itself is a random process when viewed as a waveform. When the point process is ergodic, the intensity can be estimated from observations, meaning that, as opposed to waveform processes, we have a chance of *deriving* a model for an observed point process.

Often, we want the rate of event occurrence to depend on more than just the process's history. To model seasonal variations of rain storm occurrences, for example, the intensity should depend on some periodic

[*] The notion of the Big Bang comes to mind here.

[†] We define the occurrence time w_0 to equal t_1, the beginning of the observation interval.

waveform. In such cases, the intensity depends on time through more than its history. The waveform controlling event occurrence in addition to history can be deterministic, *or* it can be a sample function of a waveform process. When deterministic, we have a non-stationary point process; when stochastic, we have what is known as a *doubly stochastic point process*. The intensity of such processes is indicated by $\lambda\,(s(t);t;N_t;\mathbf{W})$, where $s(\cdot)$ represents a signal that somehow modulates the event occurrence rate.

In applications, we may want to associate with each event a value or sequence of values. For example, Californians not only care when earthquakes occur, but also how strong they are.

Definition *A* marked *point process is regular and has associated with each event a* mark $\mathbf{U}$*: a vector of random variables* $\operatorname{col}U_1,\dots,U_m$ *having a joint probability distribution dependent on process history (which now includes previous marks).*

The intensity of a marked point process has the complicated general form $\lambda\left(t;N_t;\mathbf{W};\mathbf{U}_1,\dots,\mathbf{U}_{N_t}\right)$. In marked processes, the rate at which events occur as well as mark values can depend on previous marks and when they occurred. For example, earthquake rate increases (temporarily) after a "big one" occurs, and these aftershocks thankfully have decreased mark values.

The definition of stationarity is somewhat tricky because of the explicit inclusion of the process's birth at time $t=0$. Frequently, point processes have a nontrivial dependence structure: Rate of event occurrence does depend on process history. Because no history exists prior to the process's birthday, how the intensity evolves from having no history to having one becomes an important mathematical detail that we would like to ignore in applications: We would like to assume that the start-up transient has dissipated, leaving the process in kind of "steady-state." The only point processes—Poisson processes—immune from this transient have intensities that do not depend on process history.

Definition *The* sample function density $p_{N_{t_1,t_2},\mathbf{W}}(n;\mathbf{w})$ *equals the joint probability density of the number and occurence times of events that occur in a given interval* $[t_1,t_2)$.

Note that this density completely characterizes a point process during the stated time interval. A reasonable definition of stationarity must involve this density's properties as the observation interval becomes more distant from process initiation.

Definition *A point process is said to be* stationary *if the sample function density asymptotically depends only on the interval* $t-W_{N_T}$ *and on the interevent interval vector* τ *equivalent to* $\mathbf{W}$.

This definition recalls the existence of a stationary distribution for a waveform process produced by a system provided by a white input $\{23\}$. The point process equivalent of a generation model is the intensity; as described subsequently $\{42\}$, the sample function density completely depends on the intensity and, given the intensity, we can generate the point process.

2.4.2 The Poisson Process

Some signals have no waveform. Consider the measurement of when lightning strikes occur within some region; the random process is the sequence of event times, which has no intrinsic waveform. Such processes are termed *point processes*, and have been shown [37] to have a simple mathematical structure. Define some quantities first. Let N_t be the number of events that have occurred up to time t (observations are by convention assumed to start at $t=0$). This quantity is termed the counting process, and has the shape of a staircase function: The counting function consists of a series of plateaus always equal to an integer, with jumps between plateaus occurring when events occur. $N_{t_1,t_2}=N_{t_2}-N_{t_1}$ corresponds to the number of events in the interval $[t_1,t_2)$. Consequently, $N_t=N_{0,t}$. The event times comprise the random vector $\mathbf{W}$; the dimension of this vector is N_t, the number of events that have occurred. The occurrence of events is governed by a quantity known as the *intensity* $\lambda\,(t;N_t;\mathbf{W})$ of the point process through the probability law

$$\Pr[N_{t,t+\Delta t}=1\mid N_t;\mathbf{W}]=\lambda\,(t;N_t;\mathbf{W})\,\Delta t$$

for sufficiently small Δt. Note that this probability is a conditional probability; it can depend on how many events occurred previously and when they occurred. The intensity can also vary with time to describe nonstationary point processes. The intensity has units of events/s, and it can be viewed as the instantaneous rate at which events occur.

The simplest point process from a structural viewpoint, the Poisson process, has no dependence on process history. A stationary Poisson process results when the intensity equals a constant: $\lambda\,(t;N_t;\mathbf{W}) = \lambda_0$. Thus, in a Poisson process, a coin is flipped every Δt seconds, with a constant probability of heads (an event) occurring that equals $\lambda_0\Delta t$ and is independent of the occurrence of past (and future) events. When this probability varies with time, the intensity equals $\lambda\,(t)$, a non-negative signal, and a nonstationary Poisson process results.*

From the Poisson process's definition, we can derive the probability laws that govern event occurrence. These fall into two categories: the *count* statistics $\Pr[N_{t_1,t_2} = n]$, the probability of obtaining n events in an interval $[t_1,t_2)$, and the *time of occurrence* statistics $p_{\mathbf{W}^{(n)}}(\mathbf{w})$, the joint distribution of the first n event times in the observation interval. These times form the vector $\mathbf{W}^{(n)}$, the occurrence time vector of dimension n. From these two probability distributions, we can derive the sample function density.

Count statistics. We derive a differentio-difference equation that $\Pr[N_{t_1,t_2} = n]$, $t_1 < t_2$, must satisfy for event occurrence in an interval to be regular and independent of event occurrences in disjoint intervals. Let t_1 be fixed and consider event occurrence in the intervals $[t_1,t_2)$ and $[t_2,t_2+\delta)$, and how these contribute to the occurrence of n events in the union of the two intervals. If k events occur in $[t_1,t_2)$, then $n-k$ must occur in $[t_2,t_2+\delta)$. Furthermore, the scenarios for different values of k are mutually exclusive. Consequently,

$$\begin{aligned}
\Pr[N_{t_1,t_2+\delta} = n] &= \sum_{k=0}^{n} \Pr[N_{t_1,t_2} = k, N_{t_2,t_2+\delta} = n-k] \\
&= \Pr[N_{t_2,t_2+\delta} = 0 | N_{t_1,t_2} = n]\,\Pr[N_{t_1,t_2} = n] \\
&\quad + \Pr[N_{t_2,t_2+\delta} = 1 | N_{t_1,t_2} = n-1]\,\Pr[N_{t_1,t_2} = n-1] \\
&\quad + \cdots + \sum_{k=2}^{n} \Pr[N_{t_2,t_2+\delta} = k | N_{t_1,t_2} = n-k]\,\Pr[N_{t_1,t_2} = n-k]
\end{aligned}$$

Because of the independence of event occurrence in disjoint intervals, the conditional probabilities in this expression equal the unconditional ones. When δ is small, only the first two will be significant to first order in δ. Rearranging and taking the obvious limit, we have the equation defining the count statistics.

$$\frac{d\Pr[N_{t_1,t_2} = n]}{dt_2} = -\lambda\,(t_2)\Pr[N_{t_1,t_2} = n] + \lambda\,(t_2)\Pr[N_{t_1,t_2} = n-1]$$

To solve this equation, we apply a z-transform to both sides. Defining the transform of $\Pr[N_{t_1,t_2} = n]$ to be $P(t_2,z)$,† we have

$$\frac{\partial P(t_2,z)}{\partial t_2} = -\lambda\,(t_2)(1-z^{-1})P(t_2,z)$$

Applying the boundary condition that $P(t_1,z) = 1$, this simple first-order differential equation has the solution

$$P(t_2,z) = \exp\left\{-(1-z^{-1})\int_{t_1}^{t_2}\lambda\,(\alpha)\,d\alpha\right\}$$

To evaluate the inverse z-transform, we simply exploit the Taylor series expression for the exponential, and we find that a Poisson probability mass function governs the count statistics for a Poisson process.

$$\boxed{\Pr[N_{t_1,t_2} = n] = \frac{\left(\int_{t_1}^{t_2}\lambda\,(\alpha)\,d\alpha\right)^n}{n!}\exp\left\{-\int_{t_1}^{t_2}\lambda\,(\alpha)\,d\alpha\right\}} \tag{2.5}$$

*In the literature, stationary Poisson processes are sometimes termed homogeneous, nonstationary ones inhomogeneous.

†Remember, t_1 is fixed and can be suppressed notationally.

The integral of the intensity occurs frequently, and we succinctly denote it by $\Lambda_{t_1}^{t_2}$. When the Poisson process is stationary, the intensity equals a constant, and the count statistics depend only on the difference $t_2 - t_1$.

Time of occurrence statistics. To derive the multivariate distribution of $\mathbf{W}$, we use the count statistics and the independence properties of the Poisson process. The density we seek satisfies

$$\int_{w_1}^{w_1+\delta_1} \cdots \int_{w_n}^{w_n+\delta_n} p_{\mathbf{W}^{(n)}}(\upsilon)\, d\upsilon = \Pr\big[W_1 \in [w_1, w_1+\delta_1), \ldots, W_n \in [w_n, w_n+\delta_n)\big]$$

The expression on the right equals the probability that no events occur in $[t_1, w_1)$, one event in $[w_1, w_1+\delta_1)$, no event in $[w_1+\delta_1, w_2)$, etc.. Because of the independence of event occurrence in these disjoint intervals, we can multiply together the probability of these event occurrences, each of which is given by the count statistics.

$$\begin{aligned}
&\Pr\big[W_1 \in [w_1, w_1+\delta_1), \ldots, W_n \in [w_n, w_n+\delta_n)\big] \\
&= e^{-\Lambda_{t_1}^{w_1}} \cdot \Lambda_{w_1}^{w_1+\delta_1} e^{-\Lambda_{w_1}^{w_1+\delta_1}} \cdot e^{-\Lambda_{w_1+\delta_1}^{w_2}} \cdot \Lambda_{w_2}^{w_2+\delta_2} e^{-\Lambda_{w_2}^{w_2+\delta_2}} \cdots \Lambda_{w_n}^{w_n+\delta_n} e^{-\Lambda_{w_n}^{w_n+\delta_n}} \\
&\approx \left(\prod_{k=1}^{n} \lambda(w_k)\delta_k\right) e^{-\Lambda_{t_1}^{w_n}} \quad \text{for small } \delta_k
\end{aligned}$$

From this approximation, we find that the joint distribution of the first n event times equals

$$\boxed{p_{\mathbf{W}^{(n)}}(\mathbf{w}) = \begin{cases} \left(\displaystyle\prod_{k=1}^{n} \lambda(w_k)\right) \exp\left\{-\displaystyle\int_{t_1}^{w_n} \lambda(\alpha)\, d\alpha\right\}, & t_1 \le w_1 \le w_2 \le \cdots \le w_n \\ 0, & \text{otherwise} \end{cases}}$$

Sample function density. For Poisson processes, the sample function density describes the joint distribution of counts and event times within a specified time interval. Thus, it can be written as

$$p_{N_{t_1,t_2},\mathbf{W}}(n; \mathbf{w}) = \Pr[N_{t_1,t_2} = n | W_1 = w_1, \ldots, W_n = w_n]\, p_{\mathbf{W}^{(n)}}(\mathbf{w})$$

The second term in the product equals the distribution derived previously for the time of occurrence statistics. The conditional probability equals the probability that no events occur between w_n and t_2; from the Poisson process's count statistics, this probability equals $\exp\{-\Lambda_{w_n}^{t_2}\}$. Consequently, the sample function density for the Poisson process, be it stationary or not, equals

$$\boxed{p_{N_{t_1,t_2},\mathbf{W}}(n; \mathbf{w}) = \left(\prod_{k=1}^{n} \lambda(w_k)\right) \exp\left\{-\int_{t_1}^{t_2} \lambda(\alpha)\, d\alpha\right\}} \tag{2.6}$$

Properties. From the probability distributions derived on the previous pages, we can discern many structural properties of the Poisson process. These properties set the stage for delineating other point processes from the Poisson. They, as described subsequently, have much more structure and are much more difficult to handle analytically.

The counting process N_t is an independent increment process. For a Poisson process, the number of events in disjoint intervals are statistically independent of each other, meaning that we have an independent increment process. When the Poisson process is stationary, increments taken over equi-duration intervals are identically distributed as well as being statistically independent. Two important results obtain from this property. First, the counting process's covariance function $K_N(t,u)$ equals $\sigma^2 \min(t,u)$. This close relation to the Wiener waveform process indicates the fundamental nature of the Poisson process in the world of point processes. Note, however, that the Poisson counting process is *not* continuous almost surely. Second, the sequence of counts forms an ergodic process, meaning we can estimate the intensity parameter from observations.

The mean and variance of the number of events in an interval can be easily calculated from the Poisson distribution. Alternatively, we can calculate the characteristic function and evaluate its derivatives. The characteristic function of an increment equals

$$\Phi_{N_{t_1,t_2}}(\nu) = \exp\left\{\left(e^{j\nu}-1\right)\Lambda_{t_1}^{t_2}\right\}$$

The first two moments and variance of an increment of the Poisson process, be it stationary or not, equal

$$\boxed{\begin{aligned}
&\mathscr{E}[N_{t_1,t_2}] = \Lambda_{t_1}^{t_2} \\
&\mathscr{E}[N_{t_1,t_2}^2] = \Lambda_{t_1}^{t_2} + \left(\Lambda_{t_1}^{t_2}\right)^2 \\
&\mathscr{V}[N_{t_1,t_2}] = \Lambda_{t_1}^{t_2}
\end{aligned}}$$

Note that the mean equals the variance here, a trademark of the Poisson process.

Poisson process event times form a Markov process. Consider the conditional density $p_{W_n|W_{n-1},\ldots,W_1}(w_n|w_{n-1},\ldots,w_1)$. This density equals the ratio of the event time densities for the n- and $(n-1)$-dimensional event time vectors. Simple substitution yields

$$p_{W_n|W_{n-1},\ldots,W_1}(w_n|w_{n-1},\ldots,w_1) = \lambda(w_n)\exp\left\{-\int_{w_{n-1}}^{w_n}\lambda(\alpha)\,d\alpha\right\}, w_n \geq w_{n-1}$$

Thus, the n^{th} event time depends only on when the $(n-1)^{th}$ event occurs, meaning that we have a Markov process. Note that event times are ordered: The n^{th} event must occur after the $(n-1)^{th}$, etc.. Thus, the values of this Markov process keep increasing, meaning that from this viewpoint, the event times form a *nonstationary* Markovian sequence. When the process is stationary, the evolutionary density is exponential. It is this special form of event occurrence time density that defines a Poisson process.

Inter-event intervals in a Poisson process form a white sequence. Exploiting the previous property, the duration of the n^{th} interval $\tau_n = w_n - w_{n-1}$ does not depend on the lengths of previous (or future) intervals. Consequently, the sequence of inter-event intervals forms a "white" sequence. The sequence may not be identically distributed unless the process is stationary. In the stationary case, inter-event intervals are truly white—they form an IID sequence—and have an exponential distribution.

$$p_{\tau_n}(\tau) = \lambda_0 e^{-\lambda_0\tau}, \tau \geq 0$$

To show that the exponential density for a white sequence corresponds to the most "random" distribution, Parzen [30] proved that the *ordered* times of n events sprinkled independently and uniformly over a given interval form a stationary Poisson process. If the density of event sprinkling is not uniform, the resulting ordered times constitute a nonstationary Poisson process with an intensity proportional to the sprinkling density.

Doubly stochastic Poisson processes. Here, the intensity $\lambda(t)$ equals a sample function drawn from some waveform process. In waveform processes, the analogous concept does not have nearly the impact it does here. Because intensity waveforms must be non-negative, the intensity process *must* be nonzero mean and non-Gaussian. Assume throughout that the intensity process is stationary for simplicity. This model arises in those situations in which the event occurrence rate clearly varies unpredictably with time. Such processes have the property that the variance-to-mean ratio of the number of events in any interval exceeds one. In the process of deriving this last property, we illustrate the typical way of analyzing doubly stochastic processes: Condition on the intensity equaling a particular sample function, use the statistical characteristics of nonstationary Poisson processes, then "average" with respect to the intensity process. To calculate the expected number N_{t_1,t_2} of events in a interval, we use conditional expected values:

$$\begin{aligned} \mathscr{E}[N_{t_1,t_2}] &= \mathscr{E}\big[\mathscr{E}[N_{t_1,t_2}|\lambda(t),t_1 \le t < t_2]\big] \\ &= \mathscr{E}[\int_{t_1}^{t_2} \lambda(\alpha)\,d\alpha] \\ &= (t_2 - t_1)\cdot \mathscr{E}[\lambda(t)] \end{aligned}$$

This result can also be written as the expected value of the integrated intensity: $\mathscr{E}[N_{t_1,t_2}] = \mathscr{E}[\Lambda_{t_1}^{t_2}]$. Similar calculations yield the increment's second moment and variance.

$$\begin{aligned} \mathscr{E}[(N_{t_1,t_2})^2] &= \mathscr{E}[\Lambda_{t_1}^{t_2}] + \mathscr{E}[\left(\Lambda_{t_1}^{t_2}\right)^2] \\ \mathscr{V}[N_{t_1,t_2}] &= \mathscr{E}[\Lambda_{t_1}^{t_2}] + \mathscr{V}[\Lambda_{t_1}^{t_2}] \end{aligned}$$

Using the last result, we find that the variance-to-mean ratio in a doubly stochastic process always exceeds unity, equaling one plus the variance-to-mean ratio of the intensity process.

The approach of sample-function conditioning can also be used to derive the density of the number of events occurring in an interval for a doubly stochastic Poisson process. Conditioned on the occurrence of a sample function, the probability of n events occurring in the interval $[t_1,t_2)$ equals (Eq. 2.5, $\{38\}$)

$$\Pr\left[N_{t_1,t_2} = n|\lambda(t),t_1 \le t < t_2\right] = \frac{\left(\Lambda_{t_1}^{t_2}\right)^n}{n!}\exp\left\{-\Lambda_{t_1}^{t_2}\right\}$$

Because $\Lambda_{t_1}^{t_2}$ is a random variable, the unconditional distribution equals this conditional probability averaged with respect to this random variable's density. This average is known as the Poisson Transform of the random variable's density.

$$\boxed{\Pr\left[N_{t_1,t_2} = n\right] = \int_0^\infty \frac{\alpha^n}{n!}e^{-\alpha}p_{\Lambda_{t_1}^{t_2}}(\alpha)\,d\alpha}$$

2.4.3 Non-Poisson Processes

In the light of the Poisson process's structural characteristics and Def. 2.4.1 $\{36\}$, any regular point process is *conditionally* Poisson. The intensity expresses the dependence of the probability of an event occurring on the process's past and, if nonstationary, on a separate function of time. Thus, probabilities related to occurrence of the next event *all* have formulae identical to corresponding ones for the Poisson case, save for conditioning on history.

$$\boxed{\begin{aligned} &\Pr\left[N_{t,w_{N_t}} = 0 \mid N_t = n, \mathbf{W}^{(n)} = \mathbf{w}\right] = \exp\left\{-\int_{w_n}^{t} \lambda\left(\alpha;n;\mathbf{w}\right)d\alpha\right\} \\ &p_{\boldsymbol{\tau}_{n+1}|N_t,\mathbf{W}^{(n)}}(\boldsymbol{\tau} \mid n,\mathbf{w}) = \lambda\left(w_n+\boldsymbol{\tau};n;\mathbf{w}\right)\exp\left\{-\int_{w_n}^{w_n+\boldsymbol{\tau}} \lambda\left(w_n+\alpha;n;\mathbf{w}\right)d\alpha\right\} \end{aligned}}$$

These expressions encompass both stationary and nonstationary cases. When stationary, the intensity depends only on interevent intervals, and this dependence is frequently expressed by rewriting the intensity in terms of intervals instead of time: $\lambda\left(t;n;\mathbf{w}\right) \Leftrightarrow \widetilde{\lambda}\left(\boldsymbol{\tau};n;\boldsymbol{\tau}\right)$. Using this re-expression, the last equation becomes

$$p_{\boldsymbol{\tau}_{n+1}|N_t,\boldsymbol{\tau}^{(n)}}(\boldsymbol{\tau}_{n+1} \mid n,\boldsymbol{\tau}) = \widetilde{\lambda}\left(\boldsymbol{\tau}_{n+1};n;\boldsymbol{\tau}\right)\exp\left\{-\int_0^{\boldsymbol{\tau}_{n+1}} \widetilde{\lambda}\left(\alpha;n;\boldsymbol{\tau}\right)d\alpha\right\} \tag{2.7}$$

Here, $\boldsymbol{\tau}_{n+1}$ is defined to be the time until the next event: $\boldsymbol{\tau}_{n+1} = t - w_{N_t}$.

From these relations, we can derive a system's model for generating regular point processes [17, 28]. We exploit here a property of the conditonal distribution function. For any random variable X, its distribution function maps its range *uniformly* into the interval $[0,1]$. Furthermore, the *inverse distributionfunction* $P_X^{-1}(\cdot)$ maps the unit interval into the random variable's domain. Thus, applying the inverse distribution function to a uniformly distributed random variable $U \sim \mathscr{U}(0,1)$ results in the random variable X: $P_X^{-1}(U) = X$. This property underlies many a random variable generation technique. Here, we apply it to the conditional interval distribution, which equals $\exp\left\{-\int_0^{\tau_{n+1}} \widetilde{\lambda}\,(\alpha; n; \tau)\, d\alpha\right\}$.

$$-\ln U_{n+1} = \int_0^{\tau_{n+1}} \widetilde{\lambda}\,(\alpha; n; \tau)\, d\alpha$$

The negative logarithm of a uniform random variable equals a unit-parameter exponential random variable, which has density $\exp\{-x\}\mathrm{u}(x)$. Denoting such a random variable as E, we find that

$$E_{n+1} = \int_0^{\tau_{n+1}} \widetilde{\lambda}\,(\alpha; n; \tau)\, d\alpha$$

When applied to the sequence of (dependent) interevent intervals, this mapping generates a sequence of *independent*, identically distributed random variables—white noise. Thus, the intervals in a regular point process can be generated by passing white noise (distributed exponentially) through the inverse function of the above integral.*

$$\boxed{\begin{aligned} \tau_{n+1} &= \mathrm{G}[\tau; E_{n+1}] \\ \mathrm{G}^{-1}[\tau; \tau_{n+1}] &= \int_0^{\tau_{n+1}} \widetilde{\lambda}\,(\alpha; n; \tau)\, d\alpha \end{aligned}} \tag{2.8}$$

When the intensity depends only on a *finite* number of events that occurred prior to time t, the sequence of interevent intervals constitute a Markov process. Thus, *the Markov process structural characterizations developed in previous sections for waveform processes apply as well to the sequence of intervevent intervals in a stationary point process*. When the point process is nonstationary, the generating system varies with time, which must be expressed as the sum of interval durations: $t = \sum \tau_n$.

Example

The simplest possible example is the stationary Poisson process. Assume that it has intensity equaling λ_0. The integral can be calculated explicitly as $\lambda_0 \tau$, and its inverse is, of course, also linear. We find that

$$\tau_n = \frac{1}{\lambda_0} E_n$$

This "system" yields a sequence of independent, exponentially distributed random variables having parameter λ_0.

The next several examples illustrate frequently encountered intensities. These are all stationary; they can be made nonstationary by including temporal variations into the intensity and doubly stochastic by making these variations dependent on a random process.

Renewal processes. Second only to the Poisson process in simplicity, stationary renewal processes are characterized by independent interevent intervals that are not exponentially distributed. Here, the probability of an event depends on time since the last event occurrence.

$$\widetilde{\lambda}\,(\tau_{n+1}; n; \tau) = \lambda_0 \cdot r(\tau_{n+1})$$

*Note that this inverse function is *always* well-defined. When the intensity is positive, the integral is strictly increasing. When zero, integral is a constant over some contiguous range of interevent intervals. In this case, the inverse function is taken to be the range's rightmost edge.

Here, $r(\cdot)$ denotes the recovery function, which is normalized so that $\lim_{\tau\to\infty} r(\tau) = 1$. The Poisson process is a renewal process with a recovery function equaling one for all inervals. The normalization isolates the dependence of event occurrence on interval from the implicit occurrence rate λ_0. Note that this rate does *not* equal the average occurrence rate except in the Poisson case.

In a renewal process, the interval distribution can be directly calculated from the intensity and vice versa. Using Eq. 2.7 $\{41\}$,

$$\begin{aligned} p_\tau(\tau) &= \lambda_0 r(\tau) \exp\left\{-\lambda_0 \int_0^\tau r(\alpha)\, d\alpha\right\} \\ \lambda_0 r(\tau) &= \frac{p_\tau(\tau)}{1 - P_\tau(\tau)} \end{aligned}$$

The ratio in the last equation is known in point process literature as the *hazard function* and the *age-specfic failure rate*. This terminology comes from considering events as component failures of some sort and noting that the ratio can be interpreted as the probability of a failure in instant occurring at τ given that the failure interval exceeds τ.

Example

One example recovery function is a delayed step: $r(\tau) = \mathrm{u}(\tau - \Delta)$. Here, Δ is known as the deadtime: Because the intensity equals zero for Δ seconds after each event, events cannot occur during this time. This model has been used to approximate "latching" of a photomultiplier tube after a recorded incident photon and to describe discharge patterns of single auditory neurons [16]. The average rate at which events occur in this process equals $\lambda_0/(1+\lambda_0\Delta)$. This result can be easily derived by considering how to generate it: Using the generation equation (2.8), we find that $\tau_n = \frac{1}{\lambda_0}E_n + \Delta$.

Passing from a stationary renewal process model to a nonstationary one can be done in several ways. One approach used in modeling neural discharges [16] is to express the intensity as the product of a recovery function and a time-varying rate.

$$\lambda\,(t;n;\mathbf{w}) = s(t)\cdot \mathrm{u}(t - w_n - \Delta)$$

Example

One renewal process that exhibits positive-aging—the probability of an event increases with time since the last event—has a linear recovery function: $r(\tau) = a\tau$. Here, the interevent intervals have a Rayleigh density.

$$p_\tau(\tau) = \lambda_0 a\tau \exp\left\{-\lambda_0 a\tau^2/2\right\}$$

First-order Markovian point processes. More complicated point process dependence structures have been found in recordings from single neurons [17]. Here, the probability of an event depends not only on time since the last event, but also on time since the pentultimate one. Thus, the sequence of intervals forms a first-order Markov process in which the intensity has the form

$$\widetilde{\lambda}\,(\tau;n;\tau) = \lambda_0 \cdot r\big(\tau_{n+1} - s(\tau_n)\big)$$

where $s(\cdot)$ is a positive-valued shifting function that essentially delays the recovery function to longer interval durations in a way that depends on the previous interval's duration. When the shifting function is a decreasing function, the delay is less for longer preceding intervals than shorter ones. Thus, in this process, long intervals tend to be followed by short ones and vice versa. Simple calculations from the generation equation

(Eq. 2.8 {42}) show that the conditional expected value, equivalent to the least-squares predictor, equals the shifting function plus a constant.

$$\mathscr{E}[\tau_{n+1} \mid \tau_n] = s(\tau_n) + C$$

The constant depends in a complicated way on both the recovery function and the shifting function.

Hawkes' process [12]. This process demonstrates that regular point processes need not be Markovian. Here, the intensity depends on the output of a linear filter that has an input equal to impulses occurring at event times.

$$\lambda\,(t;n;\mathbf{w}) = \lambda_0 + \int_{-\infty}^{t} h(t-\alpha)\,dN_\alpha$$

The Steiljes integral in this expression simply equals the summed impulse responses delayed by all event times occurring prior to time t:

$$\lambda\,(t;n;\mathbf{w}) = \lambda_0 + \sum_{i=1}^{n} h(t-w_i)$$

Thus, the intensity depends on all past event times. Note that not all impulse responses can occur in this expression. Fir instance, intensities are always positive quantities, meaning that the summed impulse responses cannot be more negative than $-\lambda_0$. Further restrictions on the impulse response result if we demand the Hawkes' process be stationary. Defining $\bar{\lambda}$ as the average occurrence rate, the intensity must satisfy

$$\bar{\lambda} = \lambda_0 + \bar{\lambda} \int_{-\infty}^{t} h(t-\alpha)\,d\alpha$$

From this constraint we find the average occurrence rate equals $\lambda_0/\,(1-\int_0^\infty h(\alpha)\,d\alpha)$, which means that the filter's impulse response must satisfy $\int_0^\infty h(\alpha)\,d\alpha < 1$. This constraint means that the filter's gain at zero frequency must be less than unity.

2.5 Linear Vector Spaces

One of the more powerful tools in statistical communication theory is the abstract concept of a linear vector space. The key result that concerns us is the *representation theorem*: a deterministic time function can be uniquely represented by a sequence of numbers. The stochastic version of this theorem states that a process can be represented by a sequence of uncorrelated random variables. These results will allow us to exploit the theory of hypothesis testing to derive the *optimum* detection strategy.

2.5.1 Basics

Definition *A* linear vector space $\mathscr{S}$ *is a collection of elements called* vectors *having the following properties:*

1. *The vector-addition operation can be defined so that if $x,y,z \in \mathscr{S}$:*
   (a) *$x+y \in \mathscr{S}$ (the space is closed under addition)*
   (b) *$x+y = y+x$ (Commutivity)*
   (c) *$(x+y)+z = x+(y+z)$ (Associativity)*
   (d) *The zero vector exists and is always an element of $\mathscr{S}$. The zero vector is defined by $x+0=x$.*
   (e) *For each $x \in \mathscr{S}$, a unique vector $(-x)$ is also an element of $\mathscr{S}$ so that $x+(-x)=0$, the zero vector.*
2. *Associated with the set of vectors is a set of scalars which constitute an algebraic field. A field is a set of elements which obey the well-known laws of associativity and commutivity for both addition and multiplication. If a,b are scalars, the elements x,y of a linear vector space have the properties that:*
   (a) *$a \cdot x$ (multiplication by scalar a) is defined and $a \cdot x \in \mathscr{S}$.*

(b) *$a \cdot (b \cdot x) = (ab) \cdot x$.*

(c) *If "1" and "0" denotes the multiplicative and additive identity elements respectively of the field of scalars; then $1 \cdot x = x$ and $0 \cdot x = 0$*

(d) *$a(x+y) = ax + ay$ and $(a+b)x = ax + bx$.*

There are many examples of linear vector spaces. A familiar example is the set of column vectors of length N. In this case, we define the sum of two vectors to be:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_N + y_N \end{bmatrix}$$

and scalar multiplication to be $a \cdot \mathrm{col}[x_1 x_2 \cdots x_N] = \mathrm{col}[ax_1\, ax_2 \cdots ax_N]$. All of the properties listed above are satisfied.

A more interesting (and useful) example is the collection of square integrable functions. A square-integrable function $x(t)$ satisfies:

$$\int_{T_i}^{T_f} |x(t)|^2 dt < \infty .$$

One can verify that this collection constitutes a linear vector space. In fact, this space is so important that it has a special name—$L^2(T_i, T_f)$ (read this as *el-two*); the arguments denote the range of integration.

Definition *Let $\mathscr{S}$ be a linear vector space. A* subspace *$\mathscr{T}$ of $\mathscr{S}$ is a subset of $\mathscr{S}$ which is closed. In other words, if $x, y \in \mathscr{T}$, then $x, y \in \mathscr{S}$ and all elements of $\mathscr{T}$ are elements of $\mathscr{S}$, but some elements of $\mathscr{S}$ are not elements of $\mathscr{T}$. Furthermore, the linear combination $ax + by \in \mathscr{T}$ for all scalars a, b. A subspace is sometimes referred to as a* closed linear manifold.

2.5.2 Inner Product Spaces

A structure needs to be defined for linear vector spaces so that definitions for the length of a vector and for the distance between any two vectors can be obtained. The notions of length and distance are closely related to the concept of an inner product.

Definition *An* inner product *of two real vectors $x, y \in \mathscr{S}$, is denoted by $\langle x, y \rangle$ and is a* scalar *assigned to the vectors x and y which satisfies the following properties:*

1. *$\langle x, y \rangle = \langle y, x \rangle$*
2. *$\langle ax, y \rangle = a\langle x, y \rangle$, a is a scalar*
3. *$\langle x+y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$, z a vector.*
4. *$\langle x, x \rangle > 0$ unless $x = 0$. In this case, $\langle x, x \rangle = 0$.*

As an example, an inner product for the space consisting of column matrices can be defined as

$$\langle x, y \rangle = \mathbf{x}^t \mathbf{y} = \sum_{i=1}^{N} x_i y_i .$$

The reader should verify that this is indeed a valid inner product (*i.e.*, it satisfies all of the properties given above). It should be noted that this definition of an inner product is not unique: there are other inner product definitions which also satisfy all of these properties. For example, another valid inner product is

$$\langle x, y \rangle = \mathbf{x}^t \mathbf{K} \mathbf{y} .$$

where $\mathbf{K}$ is an $N \times N$ positive-definite matrix. Choices of the matrix $\mathbf{K}$ which are not positive definite do not yield valid inner products (property 4 is not satisfied). The matrix $\mathbf{K}$ is termed the *kernel* of the inner product. When this matrix is something other than an identity matrix, the inner product is sometimes written as $\langle x, y\rangle_{\mathbf{K}}$ to denote explicitly the presence of the kernel in the inner product.

Definition *The* norm *of a vector $x \in \mathscr{S}$ is denoted by $\|x\|$ and is defined by:*

$$\|x\| = \langle x, x\rangle^{1/2} \tag{2.9}$$

Because of the properties of an inner product, the norm of a vector is always greater than zero unless the vector is identically zero. The norm of a vector is related to the notion of the *length* of a vector. For example, if the vector x is multiplied by a constant scalar a, the norm of the vector is also multiplied by a.

$$\|ax\| = \langle ax, ax\rangle^{1/2} = |a|\|x\|$$

In other words, "longer" vectors $(a > 1)$ have larger norms. A norm can also be defined when the inner product contains a kernel. In this case, the norm is written $\|x\|_{\mathbf{K}}$ for clarity.

Definition *An* inner product *space is a linear vector space in which an inner product can be defined for all elements of the space and a norm is given by equation 2.9. Note in particular that* every *element of an inner product space must satisfy the axioms of a valid inner product.*

For the space $\mathscr{S}$ consisting of column matrices, the norm of a vector is given by (consistent with the first choice of an inner product)

$$\|x\| = \left(\sum_{i=1}^{N} x_i^2\right)^{1/2} .$$

This choice of a norm corresponds to the Cartesian definition of the length of a vector.

One of the fundamental properties of inner product spaces is the *Schwarz inequality.*

$$|\langle x, y\rangle| \leq \|x\|\,\|y\| \tag{2.10}$$

This is one of the most important inequalities we shall encounter. To demonstrate this inequality, consider the norm squared of $x + ay$.

$$\|x+ay\|^2 = \langle x+ay, x+ay\rangle = \|x\|^2 + 2a\langle x, y\rangle + a^2\|y\|^2$$

Let $a = -\langle x, y\rangle / \|y\|^2$. In this case:

$$\begin{aligned}\|x+ay\|^2 &= \|x\|^2 - 2\frac{|\langle x,y\rangle|^2}{\|y\|^2} + \frac{|\langle x,y\rangle|^2}{\|y\|^4}\|y\|^2 \\ &= \|x\|^2 - \frac{|\langle x,y\rangle|^2}{\|y\|^2}\end{aligned}$$

As the left hand side of this result is non-negative, the right-hand side is lower-bounded by zero. The Schwarz inequality of Eq. 2.10 is thus obtained. Note that equality occurs *only* when $x = -ay$, or equivalently when $x = cy$, where c is any constant.

Definition *Two vectors are said to be* orthogonal *if the inner product of the vectors is zero: $\langle x, y\rangle = 0$.*

Consistent with these results is the concept of the "angle" between two vectors. The cosine of this angle is defined by:

$$\cos(x, y) = \frac{\langle x, y\rangle}{\|x\|\,\|y\|}$$

Because of the Schwarz inequality, $|\cos(x, y)| \leq 1$. The angle between orthogonal vectors is $\pm\pi/2$ and the angle between vectors satisfying Eq. 2.10 with equality $(x \propto y)$ is zero (the vectors are parallel to each other).

Definition *The distance d between two vectors is taken to be the norm of the difference of the vectors.*

$$d(x,y) = ||x-y||$$

In our example of the normed space of column matrices, the distance between x and y would be

$$||x-y|| = \left[\sum_{i=1}^{N}(x_i-y_i)^2\right]^{1/2},$$

which agrees with the Cartesian notion of distance. Because of the properties of the inner product, this distance measure (or metric) has the following properties:

- $d(x,y) = d(y,x)$ (Distance does not depend on how it is measured.)
- $d(x,y) = 0 \implies x = y$ (Zero distance means equality)
- $d(x,z) \leq d(x,y) + d(y,z)$ (Triangle inequality)

We use this distance measure to define what we mean by convergence. When we say the sequence of vectors $\{x_n\}$ converges to x $(x_n \to x)$, we mean

$$\lim_{n\to\infty} ||x_n - x|| = 0$$

2.5.3 Hilbert Spaces

Definition *A Hilbert space $\mathscr{H}$ is a closed, normed linear vector space which contains all of its limit points: if $\{x_n\}$ is any sequence of elements in $\mathscr{H}$ that converges to x, then x is also contained in $\mathscr{H}$. x is termed the* limit point *of the sequence.*

Example

Let the space consist of all rational numbers. Let the inner product be simple multiplication: $\langle x,y\rangle = xy$. However, the limit point of the sequence $x_n = 1+1+1/2!+\cdots+1/n!$ is not a rational number. Consequently, this space is *not* a Hilbert space. However, if we define the space to consist of all finite numbers, we have a Hilbert space.

Definition *If $\mathscr{Y}$ is a subspace of $\mathscr{H}$, the vector x is* orthogonal *to the subspace $\mathscr{Y}$ for every $y \in \mathscr{Y}$, $\langle x,y\rangle = 0$.*

We now arrive at a fundamental theorem.

Theorem *Let $\mathscr{H}$ be a Hilbert space and $\mathscr{Y}$ a subspace of it. Any element $x \in H$ has the unique decomposition $x = y+z$, where $y \in \mathscr{Y}$ and z is orthogonal to $\mathscr{Y}$. Furthermore, $||x-y|| = \min_{v\in\mathscr{Y}} ||x-v||$: the distance between x and all elements of $\mathscr{Y}$ is minimized by the vector y. This element y is termed the* projection *of x onto $\mathscr{Y}$.*

Geometrically, $\mathscr{Y}$ is a line or a plane passing through the origin. Any vector x can be expressed as the linear combination of a vector lying in $\mathscr{Y}$ and a vector orthogonal to y. This theorem is of extreme importance in linear estimation theory and plays a fundamental role in detection theory.

2.5.4 Separable Vector Spaces

Definition *A Hilbert space $\mathscr{H}$ is said to be* separable *if there exists a set of vectors $\{\phi_i\}$, $i = 1,\ldots$, elements of $\mathscr{H}$, that express* every *element $x \in \mathscr{H}$ as*

$$x = \sum_{i=1}^{\infty} x_i\phi_i, \tag{2.11}$$

where x_i are scalar constants associated with ϕ_i and x and where "equality" is taken to mean that the distance between each side becomes zero as more terms are taken in the right.

$$\lim_{m\to\infty}\left\| x - \sum_{i=1}^{m} x_i\phi_i \right\| = 0$$

The set of vectors $\{\phi_i\}$ are said to form a *complete* set if the above relationship is valid. A complete set is said to form a *basis* for the space $\mathscr{H}$. Usually the elements of the basis for a space are taken to be linearly independent. *Linear independence* implies that the expression of the zero vector by a basis can only be made by zero coefficients.

$$\sum_{i=1}^{\infty} x_i\phi_i = 0 \Leftrightarrow x_i = 0\,,\; i = 1,\ldots$$

The *representation theorem* states simply that separable vector spaces exist. The representation of the vector x is the sequence of coefficients $\{x_i\}$.

Example

The space consisting of column matrices of length N is easily shown to be separable. Let the vector ϕ_i be given a column matrix having a one in the i^{th} row and zeros in the remaining rows: $\phi_i = \mathrm{col}[0,\ldots,0,1,0,\ldots,0]$. This set of vectors $\{\phi_i\}$, $i = 1,\ldots,N$ constitutes a basis for the space. Obviously if the vector x is given by $x = \mathrm{col}[x_1\, x_2 \ldots x_N]$, it may be expressed as:

$$x = \sum_{i=1}^{N} x_i\phi_i$$

using the basis vectors just defined.

In general, the upper limit on the sum in Eq. 2.11 is infinite. For the previous example, the upper limit is finite. The number of basis vectors that is *required* to express every element of a separable space in terms of Eq. 2.11 is said to be the *dimension* of the space. In this example, the dimension of the space is N. There exist separable vector spaces for which the dimension is infinite.

Definition *The basis for a separable vector space is said to be an* orthonormal *basis if the elements of the basis satisfy the following two properties:*

- *The inner product between distinct elements of the basis is zero (i.e., the elements of the basis are mutually orthogonal).*

$$\langle\phi_i,\phi_j\rangle = 0\,,\; i \neq j$$

- *The norm of each element of a basis is one (normality).*

$$\|\phi_i\| = 1\,,\; i = 1,\ldots$$

For example, the basis given above for the space of N-dimensional column matrices is orthonormal. For clarity, two facts must be explicitly stated. First, not every basis is orthonormal. If the vector space is separable, a complete set of vectors can be found; however, this set does not have to be orthonormal to be a basis. Secondly, not every set of orthonormal vectors can constitute a basis. When the vector space L^2 is discussed in detail, this point will be illustrated.

Despite these qualifications, an orthonormal basis exists for every separable vector space. There is an explicit algorithm—the *Gram-Schmidt procedure*—for deriving an orthonormal set of functions from a complete set. Let $\{\phi_i\}$ denote a basis; the orthonormal basis $\{\psi_i\}$ is sought. The Gram-Schmidt procedure is:

1. $\psi_1 = \phi_1/||\phi_1||$.
   This step makes ψ_1 have unit length.
2. $\psi_2' = \phi_2 - \langle\psi_1, \phi_2\rangle\psi_1$.
   Consequently, the inner product between ψ_2' and ψ_1 is zero. We obtain ψ_2 from ψ_2' forcing the vector to have unit length.

$2'$. $\psi_2 = \psi_2'/||\psi_2'||$.
   The algorithm now generalizes.

k. $\psi_k' = \phi_k - \sum_{i=1}^{k-1}\langle\psi_i, \phi_k\rangle\psi_i$

k'. $\psi_k = \psi_k'/||\psi_k'||$

By construction, this new set of vectors is an orthonormal set. As the original set of vectors $\{\phi_i\}$ is a complete set, and, as each ψ_k is just a linear combination of ϕ_i, $i = 1,\ldots,k$, the derived set $\{\psi_i\}$ is also complete. Because of the existence of this algorithm, a basis for a vector space is usually assumed to be orthonormal.

A vector's representation with respect to an orthonormal basis $\{\phi_i\}$ is easily computed. The vector x may be expressed by:

$$x = \sum_{i=1}^{\infty} x_i\phi_i \tag{2.12}$$

$$x_i = \langle x, \phi_i\rangle \tag{2.13}$$

This formula is easily confirmed by substituting Eq. 2.12 into Eq. 2.13 and using the properties of an inner product. Note that the exact element values of a given vector's representation depends upon both the vector *and* the choice of basis. Consequently, a meaningful specification of the representation of a vector must include the definition of the basis.

The mathematical representation of a vector (expressed by equations 2.12 and 2.13) can be expressed geometrically. This expression is a generalization of the Cartesian representation of numbers. Perpendicular axes are drawn; these axes correspond to the orthonormal basis vector used in the representation. A given vector is representation as a point in the ”plane” with the value of the component along the ϕ_i axis being x_i.

An important relationship follows from this mathematical representation of vectors. Let x and y be any two vectors in a separable space. These vectors are represented with respect to an orthonormal basis by $\{x_i\}$ and $\{y_i\}$, respectively. The inner product $\langle x, y\rangle$ is related to these representations by:

$$\langle x, y\rangle = \sum_{i=1}^{\infty} x_i y_i$$

This result is termed *Parseval's Theorem*. Consequently, the inner product between any two vectors can be computed from their representations. A special case of this result corresponds to the Cartesian notion of the length of a vector; when $x = y$, Parseval's relationship becomes:

$$||x|| = \left[\sum_{i=1}^{\infty} x_i^2\right]^{1/2}$$

These two relationships are key results of the representation theorem. The implication is that any inner product computed from vectors can also be computed from their representations. There are circumstances in which the latter computation is more manageable than the former and, furthermore, of greater theoretical significance.

2.5.5 The Vector Space L^2

Special attention needs to be paid to the vector space $L^2(T_i, T_f)$: the collection of functions $x(t)$ which are square-integrable over the interval (T_i, T_f):

$$\int_{T_i}^{T_f} |x(t)|^2\,dt < \infty$$

An inner product can be defined for this space as:

$$\langle x,y\rangle = \int_{T_i}^{T_f} x(t)y(t)\,dt \tag{2.14}$$

Consistent with this definition, the length of the vector $x(t)$ is given by

$$||x|| = \left[\int_{T_i}^{T_f} |x(t)|^2\,dt\right]^{1/2}$$

Physically, $||x||^2$ can be related to the energy contained in the signal over (T_i, T_f). This space is a Hilbert space. If T_i and T_f are both finite, an orthonormal basis is easily found which spans it. For simplicity of notation, let $T_i = 0$ and $T_f = T$. The set of functions defined by:

$$\begin{aligned}\phi_{2i-1}(t) &= \left(\frac{2}{T}\right)^{1/2}\cos\frac{2\pi(i-1)t}{T}\\ \phi_{2i}(t) &= \left(\frac{2}{T}\right)^{1/2}\sin\frac{2\pi it}{T}\end{aligned} \tag{2.15}$$

is complete over the interval $(0,T)$ and therefore constitutes a basis for $L^2(0,T)$. By demonstrating a basis, we conclude that $L^2(0,T)$ is a separable vector space. The representation of functions with respect to this basis corresponds to the well-known Fourier series expansion of a function. As most functions require an infinite number of terms in their Fourier series representation, this space is infinite dimensional.

There also exist orthonormal sets of functions that do *not* constitute a basis. For example, the set $\{\phi_i(t)\}$ defined by:

$$\phi_i(t) = \begin{cases}\frac{1}{T} & iT \le t < (i+1)T\\ 0 & \text{otherwise}\end{cases} \qquad i = 0,1,\ldots$$

over $L^2(0,\infty)$. The members of this set are normal (unit norm) and are mutually orthogonal (no member overlaps with any other). Consequently, this set is an orthonormal set. However, it does not constitute a basis for $L^2(0,\infty)$. Functions piecewise constant over intervals of length T are the only members of $L^2(0,\infty)$ which can be represented by this set. Other functions such as $e^{-t}\mathrm{u}(t)$ cannot be represented by the $\{\phi_i(t)\}$ defined above. Consequently, *orthonormality of a set of functions does not guarantee completeness.*

While $L^2(0,T)$ is a separable space, examples can be given in which the representation of a vector in this space is not precisely equal to the vector. More precisely, let $x(t) \in L^2(0,T)$ and the set $\{\phi_i(t)\}$ be defined by Eq. (2.15). The fact that $\{\phi_i(t)\}$ constitutes a basis for the space implies:

$$\left\| x(t) - \sum_{i=1}^{\infty} x_i\phi_i(t) \right\| = 0$$

where

$$x_i = \int_0^T x(t)\phi_i(t)\,dt.$$

In particular, let $x(t)$ be:

$$x(t) = \begin{cases}1 & 0 \le t \le T/2\\ 0 & T/2 < t < T\end{cases}$$

Obviously, this function is an element of $L^2(0,T)$. However, the representation of this function is not equal to 1 at $t = T/2$. In fact, the peak error never decreases as more terms are taken in the representation. In the special case of the Fourier series, the existence of this "error" is termed the *Gibbs phenomenon*. However, this "error" has zero norm in $L^2(0,T)$; consequently, the Fourier series expansion of this function is equal to the function in the sense that the function and its expansion have zero distance between them. However, one of the axioms of a valid inner product is that if $||e|| = 0 \implies e = 0$. The condition is satisfied, but the conclusion does not seem to be valid. Apparently, valid elements of $L^2(0,T)$ can be defined which are nonzero but have zero norm. An example is

$$e = \begin{cases} 1 & t = T/2 \\ 0 & \text{otherwise} \end{cases}$$

So as not to destroy the theory, the most common method of resolving the conflict is to weaken the definition of equality. The essence of the problem is that while two vectors x and y can differ from each other and be zero distance apart, the difference between them is "trivial". This difference has zero norm which, in L^2, implies that the magnitude of $(x-y)$ integrates to zero. Consequently, the vectors are essentially equal. This notion of equality is usually written as $x = y$ a.e. (x equals y *almost everywhere*). With this convention, we have:

$$||e|| = 0 \implies e = 0 \text{ a.e.}$$

Consequently, the error between a vector and its representation is zero almost everywhere.

Weakening the notion of equality in this fashion might seem to compromise the utility of the theory. However, if one suspects that two vectors in an inner product space are equal (*e.g.*, a vector and its representation), it is quite difficult to prove that they are strictly equal (and as has been seen, this conclusion may not be valid). Usually, proving they are equal almost everywhere is much easier. While this weaker notion of equality does not imply strict equality, one can be assured that any difference between them is insignificant. The measure of "significance" for a vector space is expressed by the definition of the norm for the space.

2.5.6 A Hilbert Space for Stochastic Processes

The result of primary concern here is the construction of a Hilbert space for stochastic processes. The space consisting of random variables X having a finite mean-square value is (almost) a Hilbert space with inner product $\mathscr{E}[XY]$. Consequently, the distance between two random variables X and Y is

$$d(X,Y) = \left\{\mathscr{E}[(X-Y)^2]\right\}^{1/2}$$

Now $d(X,Y) = 0 \implies \mathscr{E}[(X-Y)^2] = 0$. However, this does not imply that $X = Y$. Those sets with probability zero appear again. Consequently, we do not have a Hilbert space unless we agree $X = Y$ means $\Pr[X = Y] = 1$.

Let $X(t)$ be a process with $\mathscr{E}[X^2(t)] < \infty$. For each $t, X(t)$ is an element of the Hilbert space just defined. Parametrically, $X(t)$ is therefore regarded as a "curve" in a Hilbert space. This curve is continuous if

$$\lim_{t\to u} \mathscr{E}\left[\left(X(t) - X(u)\right)^2\right] = 0$$

Processes satisfying this condition are said to be *continuous in the quadratic mean*. The vector space of greatest importance is analogous to $L^2(T_i, T_f)$ previously defined. Consider the collection of real-valued stochastic processes $X(t)$ for which

$$\int_{T_i}^{T_f} \mathscr{E}[X(t)^2]\,dt < \infty$$

Stochastic processes in this collection are easily verified to constitute a linear vector space. Define an inner product for this space as:

$$\mathscr{E}[\langle X(t), Y(t)\rangle] = \mathscr{E}\left[\int_{T_i}^{T_f} X(t)Y(t)\,dt\right]$$

While this equation is a valid inner product, the left-hand side will be used to denote the inner product instead of the notation previously defined. We take $\langle X(t), Y(t)\rangle$ to be the *time-domain inner product* as in

Eq. (2.14). In this way, the deterministic portion of the inner product and the expected value portion are explicitly indicated. This convention allows certain theoretical manipulations to be performed more easily.

One of the more interesting results of the theory of stochastic processes is that the normed vector space for processes previously defined is separable. Consequently, there exists a complete (and, by assumption, orthonormal) set $\{\phi_i(t)\}, i = 1, \ldots$ of deterministic (nonrandom) functions which constitutes a basis. A process in the space of stochastic processes can be represented as

$$X(t) = \sum_{i=1}^{\infty} X_i \phi_i(t), \quad T_i \le t \le T_f,$$

where $\{X_i\}$, the representation of $X(t)$, is a sequence of random variables given by

$$X_i = \langle X(t), \phi_i(t) \rangle \quad \text{or} \quad X_i = \int_{T_i}^{T_f} X(t)\phi_i(t)\, dt.$$

Strict equality between a process and its representation cannot be assured. Not only does the analogous issue in $L^2(0,T)$ occur with respect to representing individual sample functions, but also sample functions assigned a zero probability of occurrence can be troublesome. In fact, the ensemble of any stochastic process can be augmented by a set of sample functions that are not well-behaved (e.g., a sequence of impulses) but have probability zero. In a practical sense, this augmentation is trivial: such members of the process cannot occur. Therefore, one says that two processes $X(t)$ and $Y(t)$ are equal almost everywhere if the distance between $||X(t) - Y(t)||$ is zero. The implication is that any lack of strict equality between the processes (strict equality means the processes match on a sample-function-by-sample-function basis) is "trivial".

2.5.7 Karhunen-Loève Expansion

The representation of the process, $X(t)$, is the sequence of random variables X_i. The choice basis of $\{\phi_i(t)\}$ is unrestricted. Of particular interest is to restrict the basis functions to those which make the $\{X_i\}$ *uncorrelated* random variables. When this requirement is satisfied, the resulting representation of $X(t)$ is termed the *Karhunen-Loève* expansion. Mathematically, we require $\mathscr{E}[X_iX_j] = \mathscr{E}[X_i]\,\mathscr{E}[X_j], i \neq j$. This requirement can be expressed in terms of the correlation function of $X(t)$.

$$\begin{aligned}\mathscr{E}[X_iX_j] &= \mathscr{E}\left[\int_0^T X(\alpha)\phi_i(\alpha)\,d\alpha \int_0^T X(\beta)\phi_j(\beta)\,d\beta\right] \\ &= \int_0^T\int_0^T \phi_i(\alpha)\phi_j(\beta)R_X(\alpha,\beta)\,d\alpha\,d\beta\end{aligned}$$

As $\mathscr{E}[X_i]$ is given by

$$\mathscr{E}[X_i] = \int_0^T m_X(\alpha)\phi_i(\alpha)\,d\alpha,$$

our requirement becomes

$$\int_0^T\int_0^T \phi_i(\alpha)\phi_j(\beta)R_X(\alpha,\beta)\,d\alpha\,d\beta = \int_0^T m_X(\alpha)\phi_i(\alpha)\,d\alpha \int_0^T m_X(\beta)\phi_j(\beta)\,d\beta,\ i \neq j.$$

Simple manipulations result in the expression

$$\int_0^T \phi_i(\alpha)\left[\int_0^T K_X(\alpha,\beta)\phi_j(\beta)\,d\beta\right]d\alpha = 0,\ i \neq j.$$

When $i = j$, the quantity $\mathscr{E}[X_i^2] - \mathscr{E}^2[X_i]$ is just the variance of X_i. Our requirement is obtained by satisfying

$$\int_0^T \phi_i(\alpha)\left[\int_0^T K_X(\alpha,\beta)\phi_j(\beta)\,d\beta\right]d\alpha = \lambda_i\delta_{ij}$$

or

$$\int_0^T \phi_i(\alpha) g_j(\alpha)\, d\alpha = 0\,,\; i \neq j\,,$$

where

$$g_j(\alpha) = \int_0^T K_X(\alpha,\beta)\phi_j(\beta)\, d\beta\,.$$

Furthermore, this requirement must hold for each j which differs from the choice of i. A choice of a function $g_j(\alpha)$ satisfying this requirement is a function which is proportional to $\phi_j(\alpha)$: $g_j(\alpha) = \lambda_j \phi_j(\alpha)$. Therefore,

$$\boxed{\int_0^T K_X(\alpha,\beta)\phi_j(\beta)\, d\beta = \lambda_j \phi_j(\alpha)}\,.$$

The $\{\phi_i\}$ which allow the representation of $X(t)$ to be a sequence of uncorrelated random variables must satisfy this integral equation. This type of equation occurs often in applied mathematics; it is termed the *eigenequation*. The sequences $\{\phi_i\}$ and $\{\lambda_i\}$ are the eigenfunctions and eigenvalues of $K_X(\alpha,\beta)$, the covariance function of $X(t)$. It is easily verified that:

$$K_X(t,u) = \sum_{i=1}^{\infty} \lambda_i \phi_i(t)\phi_i(u)$$

This result is termed *Mercer's Theorem*.

The approach to solving for the eigenfunction and eigenvalues of $K_X(t,u)$ is to convert the integral equation into an ordinary differential equation which can be solved. This approach is best illustrated by an example.

Example

$K_X(t,u) = \sigma^2 \min(t,u)$. The eigenequation can be written in this case as

$$\sigma^2 \left[\int_0^t u\phi(u)\, du + t\int_t^T \phi(u)\, du\right] = \lambda\phi(t)\,.$$

Evaluating the first derivative of this expression,

$$\sigma^2 t\phi(t) + \sigma^2 \int_t^T \phi(u)\, du - \sigma^2 t\phi(t) = \lambda \frac{d\phi(t)}{dt}$$

$$\text{or} \quad \sigma^2 \int_t^T \phi(u) du = \lambda \frac{d\phi}{dt}\,.$$

Evaluating the derivative of the last expression yields the simple equation

$$-\sigma^2 \phi(t) = \lambda \frac{d^2\phi}{dt^2}\,.$$

This equation has a general solution of the form $\phi(t) = A\sin\frac{\sigma}{\sqrt{\lambda}}t + B\cos\frac{\sigma}{\sqrt{\lambda}}t$. It is easily seen that B must be zero. The amplitude A is found by requiring $||\phi|| = 1$. To find λ, one must return to the original integral equation. Substituting, we have

$$\sigma^2 A\int_0^t u\sin\frac{\sigma}{\sqrt{\lambda}}u\, du + \sigma^2 tA\int_t^T \sin\frac{\sigma}{\sqrt{\lambda}}u\, du = \lambda A\sin\frac{\sigma}{\sqrt{\lambda}}t$$

After some manipulation, we find that

$$A\lambda\sin\frac{\sigma}{\sqrt{\lambda}}t - A\sigma t\sqrt{\lambda}\cos\frac{\sigma}{\sqrt{\lambda}}T = \lambda A\sin\frac{\sigma}{\sqrt{\lambda}}t\ \forall t \in [0,T)\,.$$

$$\text{or} \quad A\sigma t\sqrt{\lambda}\cos\frac{\sigma}{\sqrt{\lambda}}T = 0\forall t \in [0,T)\,.$$

Therefore, $\frac{\sigma}{\sqrt{\lambda}}T = (n-1/2)\pi, n = 1,2,\ldots$ and we have

$$\lambda_n = \frac{\sigma^2 T^2}{(n-1/2)^2\pi^2}$$

$$\phi_n(t) = \left(\frac{2}{T}\right)^{1/2} \sin\frac{(n-1/2)\pi t}{T}\,.$$

The Karhunen-Loève expansion has several important properties.

- The eigenfunctions of a positive-definite covariance function constitute a complete set. One can easily show that these eigenfunctions are also mutually orthogonal with respect to both the usual inner product and with respect to the inner product derived from the covariance function.
- If $X(t)$ Gaussian, X_i are Gaussian random variables. As the random variables $\{X_i\}$ are uncorrelated and Gaussian, the $\{X_i\}$ comprise a sequence of statistically independent random variables.
- Assume $K_X(t,u) = \frac{N_0}{2}\delta(t-u)$: the stochastic process $X(t)$ is white. Then

$$\int \frac{N_0}{2}\delta(t-u)\phi(u)du = \lambda\phi(t)$$

for all $\phi(t)$. Consequently, if $\lambda_i = N_0/2$, this constraint equation is satisfied *no matter what choice is made for the orthonormal set* $\{\phi_i(t)\}$. Therefore, the representation of white, Gaussian processes consists of a sequence of statistically independent, identically-distributed (mean zero and variance $N_0/2$) Gaussian random variables. This example constitutes the simplest case of the Karhunen-Loève expansion.

Problems

2.1 Simple Processes

Determine the mean, correlation function and first-order amplitude distribution of each of the processes defined below.

(a) X_t is defined by the following equally likely sample functions.

$$X_t(\omega_1) = 1 \qquad X_t(\omega_3) = \sin\pi t$$
$$X_t(\omega_2) = -2 \qquad X_t(\omega_4) = \cos\pi t$$

(b) X_t is defined by $X_t = \cos(At+\theta)$, where A and θ are statistically independent random variables. θ is uniformly distributed over $[0,2\pi)$ and A has the density function

$$p_A(A) = \frac{1}{\pi(1+A^2)}$$

2.2 An Elementary Process

The joint density of the amplitudes of a stochastic process X_t at the specific times $t = t_1$ and $t = t_2$ $(t_2 > t_1)$ is found to be

$$p_{X_{t_1},X_{t_2}}(x_1,x_2) = \begin{cases} \text{constant} & x_1 > x_2\ , 0 < x_1, x_2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

This joint density is found to be a valid joint density for X_t and X_u when $|t-u| = |t_2 - t_1|$.

(a) Find the correlation function $R_X(t,u)$ at the times $t = t_1$ and $u = t_2$.

(b) Find the expected value of X_t for all t.

(c) Is this process wide-sense stationary?

2.3 Freshman Chemistry

A student in freshman chemistry lab is frustrated; an experiment is not going well and limited time is available to perform the experiment again. The outcome of each experiment is a Gaussian random variable X having a mean equal to the value being sought and variance σ^2. The student decides to average his experimental outcomes.

$$Y_l = \frac{1}{l}\sum_{k=1}^{l} X_k \,, l = 1,2,\ldots$$

Each outcome X_i is uncorrelated with all other outcomes X_j, $j \neq i$.

(a) Find the mean and correlation function of the stochastic sequence Y_l.

(b) Is Y_l stationary? Indicate your reasoning.

(c) How large must n be to ensure that the probability of the relative error $\left|(Y_l - \mathscr{E}[Y_l])/\sigma\right|$ being less than 0.1 is 0.95?

2.4 The Morning After the Night Before

Sammy has had a raucous evening and is trying to return home after being dropped at the RMC. As he has had a little too much, he walks in a random fashion. His friends observe that at each second after he leaves the RMC, he has moved a distance Δ since the previous observation. The distance Δ is a Gaussian random variable with zero mean and standard deviation one meter. In addition, this distance can be reasonably assumed to be statistically independent of his movements at all other times (he is *really* out of it!). We wish to predict, to some degree, Sammy's position with time. For simplicity, we assume his movements are in one dimension.

(a) Define a stochastic process X_t that describes Sammy's position relative to the RMC at each observation time.

(b) What is the mean and variance of this process?

(c) What is the probability that sammy is more than ten meters from the RMC after two minutes of observations? First put your answer in terms of $Q(\cdot)$, then find a numeric answer.

(d) After ten minutes of wandering, Sammy bumps into a tree 150 meters from the RMC. What is the probability that Sammy comes within one meter of the same tree ten seconds after his collision. Again, express your answer first in terms of $Q(\cdot)$ and then find a numeric answer.

2.5 Not-So-Random Sample Functions

Consider the process defined by $X_t = A\cos(2\pi f_0 t)$ where A is a random variable and f_0 is a constant.

(a) Find the first-order density $p_{X_t}(x)$.

(b) Find the mean and correlation function of X_t.

(c) Can $\mathscr{S}_X(f)$ be calculated? If so, calculate it; if not, why not?

(d) Now let $X_t = A\cos(2\pi f_0 t) + B\sin(2\pi f_0 t)$, where A and B are random variables and f_0 constant.

(e) Find necessary and sufficient conditions for X_t to be wide-sense stationary.

(f) Show that necessary and sufficient conditions for the stochastic process X_t defined by $X_t = \cos(2\pi f_0 t + \theta)$ with f_0 a constant to be wide-sense stationary is that the characteristic function $\Phi_\theta(j\nu)$ satisfy

$$\Phi_\theta(j1) = 0 = \Phi_\theta(j2).$$

2.6 Random Frequency and Phase

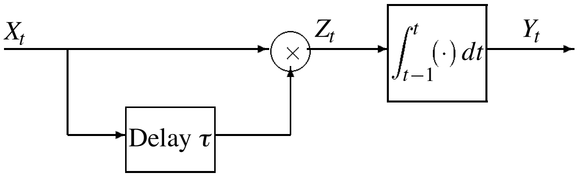
A stochastic process X_t is defined to be

$$X_t = \cos(2\pi F t + \theta)$$

where F and θ are statistically independent random variables. The quantity θ is uniformly distributed over $[-\pi, \pi)$ and F can assume one of the values $1, 2$, or 3 with equal probability.

(a) Compute the mean and correlation function of X_t.

This process serves as the input to the following system.



Note: The signals are *multiplied*, not summed, at the node located just before the integrator. Y_t and Z_t are related by

$$Y_t = \int_{t-1}^{t} Z_\alpha \, d\alpha$$

(b) Is the process Z_t wide-sense stationary?

(c) What is the expected value and correlation function of Y_t?

2.7 Random Amplitude, Frequency and Phase

Let the stochastic process X_t be defined by

$$X_t = A \cos(2\pi F t + \theta)$$

where A, F, and θ are statistically independent random variables. The random variable θ is uniformly distributed over the interval $[-\pi, \pi)$. The densities of the other random variables are to be determined.

(a) Show that X_t is a wide-sense stationary process.

(b) Is X_t strict-sense stationary? Why or why not?

(c) The inventor of this process claims that X_t can have *any* correlation function one desires by manipulating the densities of A and F. Demonstrate the validity of this result and the requirements these densities must satisfy.

(d) The inventor also claims that X_t can have *any* first-order density one desires so long as the desired density is bounded. Show that this claim is also valid. Furthermore, show that the requirements placed on the densities of A and F are consistent with those found in the previous part.

(e) Could this process be fruitfully used in simulations to emulate a process having a specific correlation function and first-order density? In other words, would statistics computed from the simulation results be meaningful? Why or why not?

2.8 Quadrature Representation of Stochastic Signals

Let X_t be a zero-mean process having the quadrature representation

$$X_t = X_{c,t} \cos 2\pi f_o t - X_{s,t} \sin 2\pi f_o t\ ,$$

where $X_{c,t}$ and $X_{s,t}$ are jointly wide-sense stationary, Gaussian processes.

(a) Show that X_t is a Gaussian process.

(b) In terms of the correlation functions $R_{X_c}(\tau)$ and $R_{X_s}(\tau)$ and the cross-correlation function $R_{X_cX_s}(\tau)$, determine sufficient conditions for X_t to be wide-sense stationary.

(c) Under these conditions, what are the joint statistics of teh random variables $X_{c,t}$ and $X_{s,u}$? In particular , so that $R_{X_cX_s}(\tau) = 0$ for $\tau = 0$.

(d) Show that if the power spectrum of X_t is bandpass and symmetric about $f = \pm f_o$, then $R_{X_cX_s}(\tau) = 0$ for all τ. What does this result say about the joint statistics of the processes $X_{c,t}$ and $X_{s,t}$?

2.9 Random Telegraph Wave

One form of the *random telegraph wave* X_t is derived from a stationary Poisson process $N_{0,t}$ having constant event rate λ.

$$X_t = \begin{cases} +1 & N_{0,t} \text{ even} \\ -1 & N_{0,t} \text{ odd} \end{cases}$$

For $t < 0$, the process is undefined. $N_{0,t}$ denotes the number of events that have occurred in the interval $[0,t)$ and has a probability distribution given by

$$\Pr[N_{0,t} = n] = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad t \geq 0 \; .$$

Note that $\Pr[N_{0,0} = 0] \equiv 1$.

(a) What is the probability distribution of X_t?

(b) Find the mean and correlation function of X_t.

(c) Is this process wide-sense stationary? stationary in the stricter sense? Provide your reasoning.

2.10 Independent Increment Processes

A stochastic process X_t is said to have *stationary, independent increments* if, for $t_1 < t_2 < t_3 < t_4$:

- The random variable $X_{t_2} - X_{t_1}$ is statistically independent of the random variable $X_{t_4} - X_{t_3}$.
- The pdf of $X_{t_2} - X_{t_1}$ is equal to the pdf of $X_{t_2+T} - X_{t_1+T}$ for all t_1, t_2, T.

The process is identically equal to zero at $t = 0$ ($\Pr[X_0 = 0] = 1$).

(a) What is the expected value and variance of $X_{t_1+t_2}$?
Hint: Write $X_{t_1+t_2} = [X_{t_1+t_2} - X_{t_1}] + [X_{t_1} - X_0]$.

(b) Using the result of part (a), find expressions for $\mathscr{E}[X_t]$ and $\mathscr{V}[X_t]$.

(c) Define $\Phi_{X_t}(j\nu)$ to be the characteristic function of the first-order density of the process X_t. Show that this characteristic function must be of the form:

$$\Phi_{X_t}(j\nu) = e^{t \cdot f(\nu)}$$

where $f(\nu)$ is a conjugate-symmetric function of ν.

(d) Compute $K_X(t,u)$.

(e) The process X_t is passed through a linear, time-invariant filter having the transfer function $H(f) = j2\pi f$. Letting Y_t denote the output, determine $K_Y(t,u)$.

2.11 Martingales

A process X_t is said to be a *martingale* if it satisfies the relationship

$$\mathscr{E}[X_t | X_{u'}, u' \leq u] = X_u \; , u \leq t \; .$$

In words, the expected value of a martingale at time t given all values of the process that have been observed up to time u ($u \leq t$) is equal to the most recently observed value of the process (X_u).

(a) Show that the mean of a martingale is a constant.

(b) Show that all zero-mean, independent, stationary-increment processes are martingales. Are all martingales independent increment processes?

(c) If X_t is a zero-mean martingale having a possibly time-varying variance $\sigma^2(t)$, show that its correlation function is given by

$$R_X(t,u) = \sigma^2(\min(t,u))$$

2.12 Time-Rescaling

In this problem, assume that the process X_t has stationary, independent increments.

(a) Define Y_t to be:

$$Y_t = g(t) \cdot X_{h(t)/g(t)}$$

where $g(t)$ and $h(t)$ are deterministic functions and $h(t)/g(t)$ is a strictly increasing function. Find the mean and covariance functions of Y_t.

(b) A stochastic process X_t is said to be a *Markov* process if, for $t_1 < t_2 < \ldots < t_{n-1} < t_n$, the conditional density of X_{t_n} satisfies:

$$p_{X_{t_n}|X_{t_1},\ldots,X_{t_{n-1}}}(X_n|X_1,\ldots,X_{n-1}) = p_{X_{t_n}|X_{t_{n-1}}}(X_n|X_{n-1}).$$

Show that all independent-increment processes are Markov processes.

2.13 Mean-Square Continuity

Let X_t be a stochastic process with correlation function $R_X(t,u)$. X_t is said to be *mean-square continuous* if

$$\lim_{t\to u} \mathscr{E}[(X_t - X_u)^2] = 0 \text{ , forall } t,u \text{ .}$$

(a) Show that X_t is mean-square continuous if and only if the correlation function $R_X(t,u)$ is continuous at $t = u$.

(b) Show that if $R_X(t,u)$ is continuous at $t = u$, it is continuous for all t and u.

(c) Show that a zero-mean, independent-increment process with stationary increments is mean-square continuous.

(d) Show that a stationary Poisson process is mean-square continuous. Note that this process has no continuous sample functions, but is continuous in the mean-square sense.

2.14 Properties of Correlation Functions

(a) Show that correlation and covariance functions have the following properties:

1. $R_X(t,u) = R_X(u,t)$
2. $R_X(\tau) = R_X(-\tau)$
3. $K_X^2(t,u) \leq K_X(t,t) \cdot K_X(u,u)$
4. $|K_X(t,u)| \leq \frac{1}{2}[K_X(t,t) + K_X(u,u)]$
5. $|R_X(\tau)| \leq R_X(0)$

(b) Let X_t be a wide-sense stationary random process. If $s(t)$ is a deterministic function and we define $Y_t = X_t + s(t)$, what is the expected value and correlation function of Y_t?

2.15 Correlation Functions and Power Spectra

(a) Which of the following are valid correlation functions? Indicate your reasoning.

1. $R_X(t,u) = e^{-(t-u)^2}$
2. $R_X(t,u) = \sigma^2 \max(t,u)$
3. $R_X(t,u) = e^{-(t-u)}$
4. $R_X(t,u) = \cos t \cdot \cos u$
5. $R_X(\tau) = e^{-|\tau|} - e^{-2|\tau|}$
6. $R_X(\tau) = \frac{5 \sin 1000\tau}{\tau}$
7. $R_X(\tau) = \begin{cases} 1 - \frac{|\tau|}{T} & |\tau| \leq T \\ 0 & \text{otherwise} \end{cases}$
8. $R_X(\tau) = \begin{cases} 1 & |\tau| \leq T \\ 0 & \text{otherwise} \end{cases}$
9. $R_X(\tau) = \delta(\tau) + 25$
10. $R_X(\tau) = \delta(\tau+1) + \delta(\tau) + \delta(\tau-1)$

(b) Which of the following are valid power density spectra? Indicate your reasoning.

1. $\mathscr{S}_X(f) = \frac{\sin \pi f}{\pi f}$
2. $\mathscr{S}_X(f) = \left(\frac{\sin \pi f}{\pi f}\right)^2$
3. $\mathscr{S}_X(f) = \exp\left\{-\frac{(f-f_0)^2}{4}\right\}$
4. $\mathscr{S}_X(f) = e^{-|f|} - e^{-2|f|}$
5. $\mathscr{S}_X(f) = 1 + 0.25 e^{-j2\pi f}$
6. $\mathscr{S}_X(f) = \begin{cases} 1 & |f| \leq 1/T \\ 0 & \text{otherwise} \end{cases}$

2.16 The Bispectrum

The idea of a correlation function can be extended to higher order moments. For example, the third order "correlation" function $R_X^{(3)}(t_1,t_2,t_3)$ of a random process X_t is defined to be

$$R_X^{(3)}(t_1,t_2,t_3) = \mathscr{E}[X_{t_1} X_{t_2} X_{t_3}]$$

(a) Show that if X_t is strict-sense stationary, then the third-order correlation function depends only on the time differences $t_2 - t_1$ and $t_3 - t_1$.

(b) Find the third-order correlation function of $X_t = A\cos(2\pi f_0 t + \Theta)$, where $\Theta \sim U[-\pi,\pi)$ and A, f_0 are constants.

(c) Let $Z_t = X_t + Y_t$, where X_t is Gaussian, Y_t is non-Gaussian, and X_t, Y_t are statistically independent, zero-mean processes. Find the third-order correlation function of Z_t.

2.17 Joint Statistics of a Process and its Derivative

Let X_t be a wide-sense stationary stochastic process. Let $\dot{X}_t$ denote the derivative of X_t.

(a) Compute the expected value and correlation function of $\dot{X}_t$ in terms of the expected value and correlation function of X_t.

(b) Under what conditions are $\dot{X}_t$ and X_t orthogonal? In other words, when does $\langle \dot{X}, X \rangle = 0$ where $\langle \dot{X}, X \rangle = \mathscr{E}[\dot{X}_t X_t]$?

(c) Compute the mean and correlation function of $Y_t = X_t - \dot{X}_t$.

(d) The bandwidth of the process X_t can be defined by

$$B_X^2 = \frac{\displaystyle\int_{-\infty}^{\infty} f^2 \mathscr{S}_X(f)\,df}{\displaystyle\int_{-\infty}^{\infty} \mathscr{S}_X(f)\,df}$$

Express this definition in terms of the mean and correlation functions of X_t and $\dot{X}_t$.

(e) The statistic U is used to count the average number of excursions of the stochastic process X_t across the level $X_t = A$ in the interval $[0,T]$. One form of this statistic is

$$U = \frac{1}{T} \int_0^T \left| \frac{d}{dt} \mathrm{u}(X_t - A) \right| dt$$

where $\mathrm{u}(\cdot)$ denotes the unit step function. Find the expected value of U, using in your final expression the formula for B_X. Assume that the conditions found in part (b) are met and X_t is a Gaussian process.

2.18 A Non-Gaussian Process

Let $\{X_l\}$ denote a sequence of independent, identically distributed random variables. This sequence serves as the input to a discrete-time system having an input-output relationship given by the difference equation

$$Y_l = aY_{l-1} + X_l$$

(a) If $X_l \sim \mathscr{N}(0, \sigma^2)$, find the probability density function of each element of the output sequence $\{Y_l\}$.

(b) Show that $|\Phi_{X_l}(j\nu)| \leq \Phi_{X_l}(j0)$ for all choices of ν no matter what the amplitude distribution of X_l may be.

(c) If X_l is non-Gaussian, the computation of the probability density of Y_l can be difficult. On the other hand, if the density of Y_l is known, the density of X_l can be found. How is the characteristic function of X_l related to the characteristic function of Y_l?

(d) Show that if Y_l is uniformly distributed over $[-1, 1)$, the only allowed values of the parameter a are those equalling $1/m$, $m = \pm 2, \pm 3, \pm 4, \ldots$.

2.19 Nonlinearities and Processes

The Gaussian wide-sense stationary random process X_t, having mean zero and correlation function $R_X(\tau)$, is squared during signal processing: $Y_t = X_t^2$.

(a) Is Y_t a wide-sense stationary process? Why or why not?

(b) What is the probability density of Y_t?

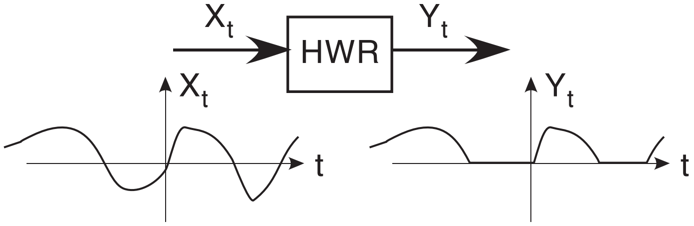
(c) If it exists, find the power density spectrum of Y_t.

2.20 Nonlinear Processing

A stationary, zero-mean Gaussian process is passed through a half-wave rectifier, which has an input-output relationship given by

$$Y_t = \begin{cases} X_t & X_t \geq 0 \\ 0 & X_t < 0 \end{cases}$$

In other words, the output equals the positive-valued amplitudes of the input and is zero otherwise.



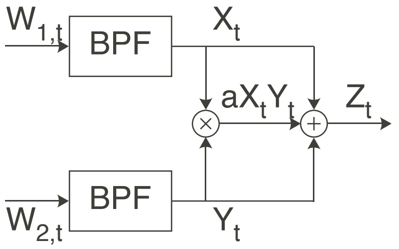
(a) What is the mean and variance of Y_t? Express your answer in terms of the correlation function of X_t.

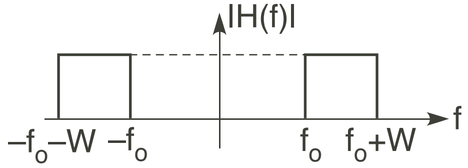
(b) Is the output Y_t a stationary process? Indicate why or why not.

(c) What is the cross-correlation function between input and output? Express your answer in terms of the correlation function of X_t.

2.21 Intermodulation Distortion

X_t and Y_t are the outputs of identical bandpass filters having statistically independent, identically distributed white-noise processes $W_{1,t}$ and $W_{2,t}$ as inputs.





The frequency f_0 is much larger than the bandwidth W.

(a) Find the correlation function of X_t.

Z_t consists of the sum of X_t, Y_t and an *intermodulation distortion* component aX_tY_t, where a is an unknown constant.

(b) What is the cross-correlation function of Y_t and Z_t?

(c) Find the correlation function of Z_t.

(d) You want to remove the intermodulation distortion component from Z_t. Can this removal be accomplished by operating *only* on Z_t? If so, how; if not, why not.

2.22 Predicting the Stock Market

The price of a certain stock can fluctuate during the day while the true value is rising or falling. To facilitate financial decisions, a Wall Street broker decides to use stochastic process theory. The price P_t of a stock is described by

$$P_t = Kt + N_t \,, 0 \leq t < 1$$

where K is the constant our knowledgeable broker is seeking and N_t is a stochastic process describing the random fluctuations. N_t is a white, Gaussian process having spectral height $N_0/2$. The broker decides to estimate K according to:

$$\widehat{K} = \int_0^1 P_t g(t)\, dt$$

where the best function $g(t)$ is to be found.

(a) Find the probability density function of the estimate $\widehat{K}$ for any $g(t)$ the broker might choose.

(b) A simple-minded estimate of K is to use simple averaging (*i.e.*, set $g(t) = \text{constant}$). Find the value of this constant which results in $\mathscr{E}[\widehat{K}] = K$. What is the resulting percentage error as expressed by $\sqrt{\mathscr{V}[\widehat{K}]}/|\mathscr{E}[\widehat{K}]|$.

(c) Find $g(t)$ which minimizes the percentage error and yields $\mathscr{E}[\widehat{K}] = K$. How much better is this optimum choice than simple averaging?

2.23 Constant or no Constant?

To determine the presence or absence of a constant voltage measured in the presence of additive, white Gaussian noise (spectral height $N_0/2$), an engineer decide to compute the average $\bar{V}$ of the measured voltage V_t.

$$\bar{V} = \frac{1}{T}\int_0^T V_t\, dt$$

The value of the constant voltage, if present, is V_0. The presence and absence of the voltage are equally likely to occur.

(a) Derive a good method by which the engineer can use the average to determine the presence or absence of the constant voltage.

(b) Determine the probability that the voltage is present when the engineer's method announces it is.

(c) The engineer decides to improve the performance of his technique by computing the more complicated quantity V given by

$$V = \int_0^T f(t) V_t \, dt$$

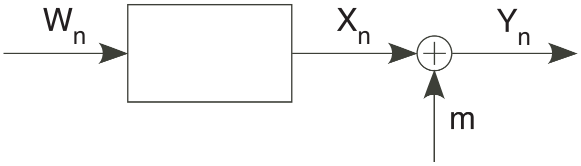
What function $f(t)$ maximizes the probability found in part (b)?

2.24 Estimating the Mean

Suppose you have stochastic process Y_n produced by the depicted system. The input W_n is discrete-time white noise (not necessarily Gaussian) having zero mean and correlation function $R_W(l) = \sigma_W^2 \delta(l)$. The system relating X_n to the white-noise input is governed by the difference equation

$$X_n = aX_{n-1} + W_n, \quad |a| < 1 \ .$$

The quantity m is an unknown constant.



(a) Is the process Y_n stationary? Why or why not?

(b) What is the correlation function of Y_n?

(c) We want to estimate the constant m by averaging Y_n over a finite time interval: $\widehat{m} = \frac{1}{N}\sum_{n=0}^{N-1} Y_n$. What is the expected value and variance of this estimate?

2.25 Generating Random Processes

It is desired to generate a wide-sense stationary process with correlation function

$$R_X(\tau) = e^{-|\tau|}.$$

Two methods are proposed.

1. Let $X_t = A\cos(2\pi Ft + \theta)$ where A, F, and θ are statistically independent random variables.
2. Define X_t by:

$$X_t = \int_0^\infty h(\alpha) N_{t-\alpha} \, d\alpha$$

where N_t is white and $h(t)$ is the impulse response of the appropriate filter.

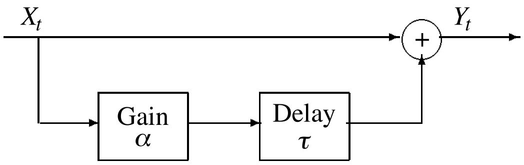
(a) Find at least two impulse responses $h(t)$ that will work in method 2.

(b) Specify the densities for A, F, θ in method 1 that yield the desired results.

(c) Sketch sample functions generated by each method. Interpret your result. What are the technical differences between these processes?

2.26 Multipath Channels

Let X_t be a Gaussian random process with mean $m_X(t)$ and covariance function $K_X(t,u)$. The process is passed through the depicted system.



(a) Is Y_t a Gaussian process? If so, compute the pdf of Y_t.

(b) What are the mean and covariance functions of Y_t?

(c) If X_t is stationary, is Y_t stationary?

(d) Compute the cross-correlation function between X_t and Y_t.

2.27 A Simple Filter

The white process X_t serves as the input to a system having output Y_t. The input-output relationship of this system is determined by the differential equation

$$\dot{Y}_t + 2Y_t = X_t$$

(a) Find the mean and correlation function of Y_t.

(b) Compute the cross-correlation function between X_t and Y_t.

(c) Show that the correlation function of Y_t obeys a homogeneous version of the differential equation governing the system for positive values of τ.

$$\dot{R}_Y(\tau) + 2R_Y(\tau) = 0\ , \tau > 0$$

Do **not** use your answer to part (a) to work this part. Rather, show the validity of this result in a more general fashion.

2.28 Noise Reduction Filters

Noise reduction filters are used to reduce, as much as possible, the noise component of a noise-corrupted signal. Let the signal of interest be described as a wide-sense stationary process X_t. The observed signal is given by $Y_t = X_t + N_t$, where N_t is a process modeling the noise that is statistically independent of the signal.

(a) Assuming that the noise is white, find a relationship that the transfer function of the filter must satisfy to maximize the signal-to-noise ratio (i.e., the ratio of the signal power to the noise power) in the filtered output.

(b) Compute the resulting signal-to-noise ratio when the signal correlation function is

$$R_X(\tau) = \sigma^2 e^{-a|\tau|}.$$

2.29 Analog Communication

The message M_t is to be transmitted using amplitude modulation. The transmitted signal has the form

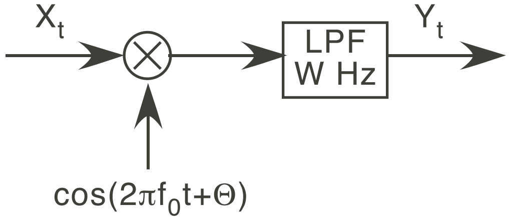
$$X_t = (1 + M_t)\cos(2\pi f_0 t + \Theta),\ f_0 = \text{constant}, \Theta \sim U(-\pi, \pi]$$

The message is a wide-sense stationary Gaussian process having zero mean. It is bandlimited to W Hz and statistically independent of the random phase Θ. A few technical details are not quite worked out...

(a) For technical reasons, the transmitter needs to *clip* the message so that it never exceeds 1 in magnitude. What is the probability that the magnitude of M_t at a randomly chosen time exceeds 1? Express your answer in terms of the mean and correlation function of M_t.

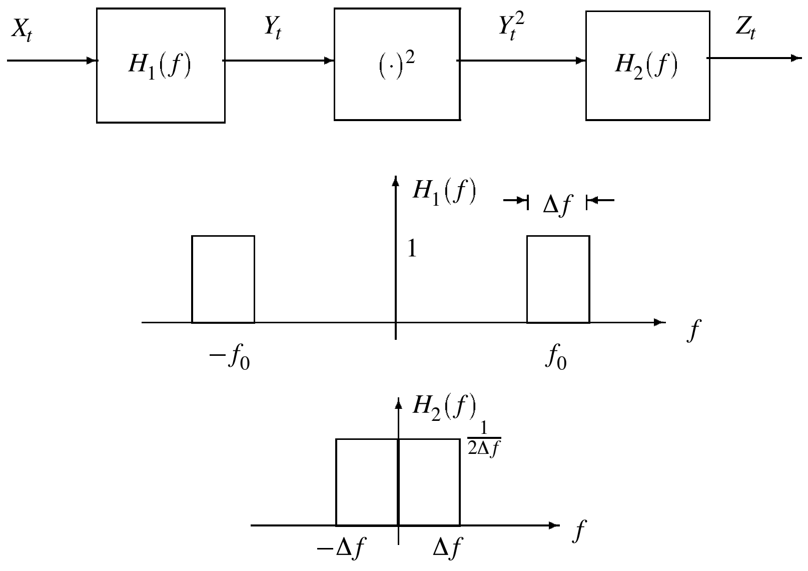
(b) What is the power spectrum of X_t? Assume that the probability of clipping the message is negligible.

(c) A standard receiver multiplies the incoming signal by a sinusoid having a frequency and phase equal to that of the carrier, then passes the product through an ideal lowpass filter having a cutoff frequency of W Hz. What is the power spectrum of the receiver's output? Assume $f_0 \gg W$.



2.30 Measuring the Power Spectrum with Analog Means

In practice, one often wants to measure the power density of a stochastic process. For the purposes of this problem, assume the process X_t is wide-sense stationary, zero mean, and Gaussian. The following measurement system is proposed.



where $H_1(f)$ is the transfer function of an ideal bandpass filter and $H_2(f)$ is an ideal lowpass. Assume that Δf is small compared to range of frequencies over which $\mathscr{S}_X(f)$ varies.

(a) Find the mean and correlation function of Y_t^2 in terms of the second-order statistics of X_t.

(b) Compute the power density spectrum of the process Z_t.

(c) Compute the expected value of Z_t.

(d) By considering the variance of Z_t, comment on the accuracy of this measurement of the power density of the process X_t.

2.31 Three Filters

Let X_t be a stationary, zero-mean random process that serves as the input to three linear, time-invariant filters. The power density spectrum of X_t is $\mathscr{S}_X(f) = N_0/2$. The impulse responses of the filters are

$$h_1(t) = \begin{cases} 1 & 0 \le t \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$h_2(t) = \begin{cases} 2e^{-t} & t \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

$$h_3(t) = \begin{cases} \sqrt{2}\sin 2\pi t & 0 \le t \le 2 \\ 0 & \text{otherwise} \end{cases}$$

The output of filter i is denoted by $Y_i(t)$.

(a) Compute $\mathscr{E}[Y_i(t)]$ and $\mathscr{E}[Y_i^2(t)]$ for $i = 1, 2, 3$.

(b) Compute $R_{XY_2}(t, u)$. Interpret your result.

(c) Is there any pair of processes for which $\mathscr{E}[Y_i(t) \cdot Y_j(t)] = 0$ for all t?

(d) Is there any pair of processes for which $\mathscr{E}[Y_i(t)\cdot Y_j(u)] = 0$ for all t and u?

2.32 Generation of Processes
Let X_t be a wide-sense stationary process having correlation function $R_x(\tau) = \delta(\tau)$. X_t serves as the input to a linear, time-invariant system having impulse response $h(t)$.

(a) Determine an $h(t0$ so that the linear system is stable and causal, and yields an output Y_t having the correlation function
$$R_Y(\tau) = e^{-|\tau|} + e^{-|\tau|} .$$

(b) Show that your answer is not unique by finding at least three other alternatives.

2.33 Time-Bandwidth Product
It is frequently claimed that the relation between noise bandwidth and reciprocal duration of the observation interval play a key role in determining whether DFT values are approximately uncorrelated. While the statements sound plausible, their veracity should be checked. Let the covariance function of the observation noise be $K_N(l) = a^{|l|}$.

(a) How is the bandwidth (defined by the half-power point) of this noise's power spectrum related to the parameter a? How is the duration (defined to be two time constants) of the covariance function related to a?

(b) Find the variance of the length-L DFT of this noise process as a function of the frequency index k. This result should be compared with the power spectrum calculated in part (a); they *should* resemble each other when the "memory" of the noise—the duration of the covariance function—is much less than L while demonstrating differences as the memory becomes comparable to or exceeds L.

(c) Calculate the covariance between adjacent frequency indices. Under what conditions will they be approximately uncorrelated? Relate your answer to the relations of a to L found in the previous part.

2.34 More Time-Bandwidth Product
The results derived in Problem 2.33 assumed that a length-L Fourier Transform was computed from a length-L segment of the noise process. What will happen if the transform has length $2L$ with the observation interval remaining unchanged?

(a) Find the variance of DFT values at index k.

(b) Assuming the conditions in Problem 2.33 for uncorrelated adjacent samples, now what is the correlation between adjacent DFT values?

2.35 Sampling Stochastic Processes

(a) Let X_t be a wide-sense stationary process bandlimited to W Hz. The sampling interval T_s satisfies $T_s \leq \frac{1}{2W}$. What is the covariance of successive samples?

(b) Now let X_t be Gaussian. What conditions on T_s will insure that successive samples will be statistically independent?

(c) Now assume the process is *not* strictly bandmilited to W Hz. This process serves as the input to an ideal lowpass filter having cutoff frequency W to produce the process Y_t. This output is sampled every $\frac{1}{2W}$ seconds to yield an approximate representation Z_t of the original signal X_t. Show that the mean-squared value of the sampling error, defined to be $\varepsilon^2 = \mathscr{E}[(X_t - Z_t)^2]$, is given by $\varepsilon^2 = 2\int_W^\infty \mathscr{S}_x(f)\,df$.

2.36 Properties of the Poisson Process
Let N_t be a Poisson process with intensity $\lambda(t)$.

(a) What is the expected value and variance of the number of events occurring in the time interval $[t,u)$?

(b) Under what conditions is N_t a stationary, independent increment process?

(c) Compute $R_N(t,u)$.

(d) Assume that $\lambda(t) = \lambda_0$, a constant. What is the conditional density $p_{W_n|W_{n-1}}(w_n|w_{n-1})$? From this relationship, find the density of τ_n, the time interval between W_n and W_{n-1}.

2.37 Optical Communications

In optical communication systems, a photomultiplier tube is used to convert the arrival of photons into electric pulses so that each arrival can be counted by other electronics. Being overly clever, a clever Rice engineer bought a photomultiplier tube from AGGIE PMT, Inc. The AGGIE PMT device is unreliable. When it is working, each photon is properly converted to an electric pulse. When not working, it has "dead-time" effects: the conversion of a photon arrival blocks out the conversion of the next photon. After a photon arrival has been missed, the device converts the next arrival properly. To detect whether the Aggie device is working properly or not, the clever Rice engineer decides to use results from a statistical signal processing course he is taking to help him. A calibrated light source is used to give an average arrival rate of λ photons/sec on the surface of the photomultiplier tube. Photon arrivals are described by a Poisson process.

(a) Find the density of the time between electric pulses if the AGGIE device has these dead-time effects.

(b) Can the times of occurrence of electric pulses be well-described by a Poisson process when the dead-time effects are present? If so, find the parameters of the process; if not, state why.

(c) Assuming the device is as likely to not be working as it is to be working, find a procedure to determine its mode of operation based on the observation of the time between two successive electric pulses.

2.38 Shot Noise

Shot noise is noise measured in vacuum tube circuits which is due to the spontaneous emission of electrons from each tube's cathode. The electron emission is assumed to be described as a stationary Poisson process of intensity λ. The impact of each electron on a tube's anode causes a current to flow in the attached circuit equal to the impulse response of the circuit. Thus, shot noise is often modeled as a sequence of impulses (whose times of occurrence are a Poisson process) passing through a linear, time-invariant system.

(a) What is the correlation function of X_t?
Hint: Relate X_t to the counting process N_t.

(b) Show that for any wide-sense stationary process X_t for which $\lim_{\tau\to\infty} R_X(\tau) = 0$, the mean of X_t is zero. Use this result to show that if $\lim_{\tau\to\infty} R_X(\tau)$ exists, the value of the limit equals the square of the mean of the process.

(c) Find the power density spectrum of the shot noise process Y_t.

(d) Evaluate the mean and variance of Y_t.

2.39 Filtering Poisson Processes

An impulse is associated with the occurrence of each event in a stationary Poisson process. This derived process serves as the input to a linear, time-invariant filter having transfer function $H(f)$, which is given by

$$H(f) = 1 - e^{-j\pi fT} + e^{-j2\pi fT}\ , T = \text{constant}\,.$$

(a) What is the mean and covariance function of the input to the filter?

(b) What is the mean and covariance function of the output of the filter?

(c) Now let the filter have any impulse response that has duration T (i.e., $h(t) = 0, t < 0$ and $t > T$). Find the impulse response that yields the smallest possible coefficient of variation $v(t)$. The coefficient of variation, defined to be the ratio of the process's standard deviation to its mean at time t, measures the percentage variation of a positive-valued process.

2.40 Linear Vector Spaces

Do the following classes of stochastic processes constitute a linear vector space? If so, indicate the proof; if not, show why not.

1. All stochastic processes.
2. All wide-sense stationary processes.
3. All nonstationary stochastic processes.

2.41 Inner Products

Show that the inner product of two vectors satisfies the following relationships.

(a) $|\langle x,y\rangle| \leq ||x|| \cdot ||y||$, the Schwarz inequality.

(b) $||x+y|| \leq ||x|| + ||y||$, the triangle inequality.

(c) $||x+y||^2 + ||x-y||^2 = 2||x||^2 + 2||y||^2$, the parallelogram equality.

(d) Find an expression for the inner product $\langle x,y\rangle$ using norms *only*.

2.42 Defining a Linear Vector Space

Let x and y be elements of a normed, linear vector space.

(a) Determine whether the following are valid inner products for the indicated space.

1. $\langle x,y\rangle = \mathbf{x}^t\mathbf{A}\mathbf{y}$ where $\mathbf{A}$ is a nonsingular, $N \times N$ matrix and $\mathbf{x}$, $\mathbf{y}$ are elements of the space of N-dimensional column matrices.
2. $\langle x,y\rangle = \mathbf{x}\mathbf{y}^t$ where $\mathbf{x}, \mathbf{y}$ are elements of the space of N-dimensional column matrices.
3. $\langle x,y\rangle = \int_0^T x(t)y(T-t)\,dt$ where x, y are finite-energy signals defined over $[0,T]$.
4. $\langle x,y\rangle = \int_0^T w(t)x(t)y(t)\,dt$ where $w(t)$ is a non-negative function and x, y are finite-energy signals defined over $[0,T]$.
5. $\mathscr{E}[XY]$ where X and Y are real-valued random variables having finite mean-square values.
6. $\text{cov}(X,Y)$, the covariance of the real-valued random variables X and Y. Assume that the random variables have finite mean-square values.

(b) Under what conditions is

$$\int_0^T\int_0^T Q(t,u)x(t)y(u)\,dt\,du$$

a valid inner product for the set of finite-energy functions defined over $[0,T]$?

2.43 Inner Products with Kernels

Let an inner product be defined with respect to the positive-definite, symmetric kernel Q.

$$\langle x,y\rangle_Q = xQy$$

where xQy is the abstract notation for the mapping of the two vectors to a scalar. For example, if $\mathbf{x}$ and $\mathbf{y}$ are column matrices, $\mathbf{Q}$ is a positive-definite square matrix and

$$\langle x,y\rangle_Q = \mathbf{x}^t\mathbf{Q}\mathbf{y}\ .$$

If x and y are defined in L^2, then

$$\langle x,y\rangle_Q = \int\!\!\int x(t)Q(t,u)y(u)\,dt\,du.$$

Let v denote an eigenvector of Q: $Qv = \lambda v$.

(a) Show that the eigenvectors of a positive-definite, symmetric kernel are orthogonal.

$$\langle \mathbf{v}_i, \mathbf{v}_j\rangle = 0\ , i \neq j\ .$$

(b) Show that these eigenvectors are orthogonal with respect to the inner product generated by Q. Consequently, the eigenvectors are orthogonal with respect to two different inner products.

(c) Let $\widetilde{Q}$ be the inverse kernel associated with Q. If $\mathbf{Q}$ is a matrix, then $\mathbf{Q}\widetilde{\mathbf{Q}} = \mathbf{I}$. If Q is a continuous-time kernel, then

$$\int Q(t,u)\widetilde{Q}(u,v)\,du = \delta(t-v)\ .$$

Show that the eigenvectors of the inverse kernel are equal to those of the kernel. How are the associated eigenvalues of these kernels related to each other?

2.44 A Karhunen-Loève Expansion

Let the covariance function of a wide-sense stationary process be

$$K_X(\tau) = \begin{cases} 1-|\tau| & |\tau| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the eigenfunctions and eigenvalues associated with the Karhunen-Loève expansion of X_t over $(0,T)$ with $T < 1$.

2.45 Karhunen-Loève Expansions

The purpose of this problem is to derive a general result which describes conditions for an orthonormal basis to result in an uncorrelated representation of a process. Let X denote a stochastic process which has the expansion

$$X = \sum_{i=1}^{\infty} \langle X, \phi_i \rangle \phi_i$$

where $\{\phi_i\}$ denotes a complete, orthonormal basis with respect to the inner product $\langle \cdot, \cdot \rangle$.

$$\langle \phi_i, \phi_j \rangle = \delta_{ij}$$

The process X may be nonstationary and may have non-zero mean.

(a) To require that the representation be an uncorrelated sequence is equivalent to requiring:

$$\mathscr{E}[\langle X, \phi_i\rangle\langle X, \phi_j\rangle] - \mathscr{E}[\langle X, \phi_i\rangle]\,\mathscr{E}[\langle X, \phi_j\rangle] = \lambda_i \delta_{ij}$$

Show that this requirement implies:

$$\mathscr{E}[X\langle X - m_X, \phi_i\rangle] = \lambda_i \phi_i$$

where $m_X = \mathscr{E}[X]$.

(b) Let X be a finite-length stochastic sequence so that it can be considered a random vector. Define the inner product $\langle X, Y\rangle$ to be $\mathbf{X}^t\mathbf{Y}$. Show that above equation is equivalent to

$$\mathbf{K}_X \phi = \lambda \phi\ .$$

(c) Let X be a continuous parameter process so that

$$\langle X, Y\rangle = \int_0^T X_t Y_t\,dt$$

Show that this inner product implies

$$\int_0^T K_X(t,u)\phi(u)\,du = \lambda \phi(t).$$

(d) Again let X be a continuous parameter process. However, define the inner product to be

$$\langle X, Y\rangle = \int_0^T \int_0^T Q(t,u) X_t Y_u\,dt\,du.$$

where $Q(t,u)$ is a non-negative definite function. Find the equivalent relationship implied by the requirements of the Karhunen-Loève expansion. Under what conditions will the ϕs satisfying this relationship not depend on the covariance function of X?

Chapter 3

Estimation Theory

In searching for methods of extracting information from noisy observations, this chapter describes *estimation theory*, which has the goal of *extracting from noise-corrupted observations the values of disturbance parameters (noise variance, for example), signal parameters (amplitude or propagation direction), or signal waveforms*. Estimation theory assumes that the observations contain an information-bearing quantity, thereby tacitly assuming that detection-based preprocessing has been performed (in other words, do I have something in the observations worth estimating?). Conversely, detection theory often requires estimation of unknown parameters: Signal presence is assumed, parameter estimates are incorporated into the detection statistic, and consistency of observations and assumptions tested. Consequently, detection and estimation theory form a symbiotic relationship, each requiring the other to yield high-quality signal processing algorithms.

Despite a wide variety of error criteria and problem frameworks, the optimal detector is characterized by a single result: the likelihood ratio test. Surprisingly, optimal detectors thus derived are usually easy to implement, not often requiring simplification to obtain a feasible realization in hardware or software. In contrast to detection theory, no fundamental result in estimation theory exists to be summoned to attack the problem at hand. The choice of error criterion and its optimization heavily influences the form of the estimation procedure. Because of the variety of criterion-dependent estimators, arguments frequently rage about which of several optimal estimators is "better." Each procedure is optimum for its assumed error criterion; thus, the argument becomes which error criterion best describes some intuitive notion of quality. When more *ad hoc*, noncriterion-based procedures[*] are used, we cannot assess the quality of the resulting estimator relative to the best achievable. As shown later, bounds on the estimation error do exist, but their tightness and applicability to a given situation are always issues in assessing estimator quality. At best, estimation theory is less structured than detection theory. Detection is science, estimation art. Inventiveness coupled with an understanding of the problem (what types of errors are critically important, for example) are key elements to deciding which estimation procedure "fits" a given problem well.

3.1 Terminology in Estimation Theory

More so than detection theory, estimation theory relies on jargon to characterize the properties of estimators. Without knowing any estimation technique, let's use parameter estimation as our discussion prototype. The parameter estimation problem is to determine from a set of L observations, represented by the L-dimensional vector $\mathbf{X}$, the values of parameters denoted by the vector θ. We write the *estimate* of this parameter vector as $\widehat{\theta}(\mathbf{X})$, where the "hat" denotes the estimate, and the functional dependence on $\mathbf{X}$ explicitly denotes the dependence of the estimate on the observations. This dependence is always present,[†] but we frequently denote the estimate compactly as $\widehat{\theta}$. Because of the probabilistic nature of the problems considered in this chapter, a parameter estimate is itself a random vector, having its own statistical characteristics. The *estimation error*

[*] This governmentese phrase concisely means guessing.

[†] Estimating the value of a parameter given no data may be an interesting problem in clairvoyance, but not in estimation theory.

$\varepsilon(\mathbf{X})$ equals the estimate minus the actual parameter value: $\varepsilon(\mathbf{X}) = \widehat{\theta}(\mathbf{X}) - \theta$. It too is a random quantity and is often used in the criterion function. For example, the *mean-squared error* is given by $\mathscr{E}[\varepsilon^t \varepsilon]$; the minimum mean-squared error estimate would minimize this quantity. The mean-squared error matrix is $\mathscr{E}[\varepsilon\varepsilon^t]$; on the main diagonal, its entries are the mean-squared estimation errors for each component of the parameter vector, whereas the off-diagonal terms express the correlation between the errors. The mean-squared estimation error $\mathscr{E}[\varepsilon^t \varepsilon]$ equals the trace of the mean-squared error matrix $\operatorname{tr}\{\mathscr{E}[\varepsilon\varepsilon^t]\}$.

Bias. An estimate is said to be *unbiased* if the expected value of the estimate equals the true value of the parameter: $\mathscr{E}[\widehat{\theta}|\theta] = \theta$. Otherwise, the estimate is said to be *biased*: $\mathscr{E}[\widehat{\theta}|\theta] \neq \theta$. The *bias* $\mathbf{b}(\theta)$ is usually considered to be additive, so that $\mathbf{b}(\theta) = \mathscr{E}[\widehat{\theta}|\theta] - \theta$. When we have a biased estimate, the bias usually depends on the number of observations L. An estimate is said to be *asymptotically unbiased* if the bias tends to zero for large L: $\lim_{L\to\infty} \mathbf{b} = \mathbf{0}$. An estimate's variance equals the mean-squared estimation error *only* if the estimate is unbiased.

An unbiased estimate has a probability distribution where the mean equals the actual value of the parameter. Should the lack of bias be considered a desirable property? If many unbiased estimates are computed from statistically independent sets of observations having the same parameter value, the average of these estimates will be close to this value. This property does *not* mean that the estimate has less error than a biased one; there exist biased estimates whose mean-squared errors are smaller than unbiased ones. In such cases, the biased estimate is usually asymptotically unbiased. Lack of bias is good, but that is just one aspect of how we evaluate estimators.

Consistency. We term an estimate *consistent* if the mean-squared estimation error tends to zero as the number of observations becomes large: $\lim_{L\to\infty} \mathscr{E}[\varepsilon^t \varepsilon] = 0$. Thus, a consistent estimate must be at least asymptotically unbiased. Unbiased estimates do exist whose errors never diminish as more data are collected: Their variances remain nonzero no matter how much data are available. Inconsistent estimates may provide reasonable estimates when the amount of data is limited, but have the counterintuitive property that the quality of the estimate does not improve as the number of observations increases. Although appropriate in the proper circumstances (smaller mean-squared error than a consistent estimate over a pertinent range of values of L), consistent estimates are usually favored in practice.

Efficiency. As estimators can be derived in a variety of ways, their error characteristics must always be analyzed and compared. In practice, many problems and the estimators derived for them are sufficiently complicated to render analytic studies of the errors difficult, if not impossible. Instead, numerical simulation and comparison with lower bounds on the estimation error are frequently used instead to assess the estimator performance. An *efficient* estimate has a mean-squared error that equals a particular lower bound: the Cramér-Rao bound. If an efficient estimate exists (the Cramér-Rao bound is the greatest lower bound), it is optimum in the mean-squared sense: No other estimate has a smaller mean-squared error (see §3.2.4 {79} for details).

For many problems no efficient estimate exists. In such cases, the Cramér-Rao bound remains a lower bound, but its value is smaller than that achievable by any estimator. How much smaller is usually not known. However, practitioners frequently use the Cramér-Rao bound in comparisons with numerical error calculations. Another issue is the choice of mean-squared error as the estimation criterion; it may not suffice to pointedly assess estimator performance in a particular problem. Nevertheless, every problem is usually subjected to a Cramér-Rao bound computation and the existence of an efficient estimate considered.

3.2 Parameter Estimation

Determining signal parameter values or a probability distribution's parameters are the simplest estimation problems. Their fundamental utility in signal processing is unquestioned. How do we estimate noise power? What is the best estimator of signal amplitude? Examination of useful estimators, and evaluation of their properties and performances constitute a case study of estimation problems. As expected, many of these issues are interrelated and serve to highlight the intricacies that arise in estimation theory.

All parameters of concern here have unknown values; we classify parameter estimation problems according to whether the parameter is stochastic or not. If so, then the parameter has a probability density known as the *prior density* (one that applies before the data become available). Choosing the prior, as we have said so often, narrows the problem considerably, suggesting that measurement of the parameter's density would yield something like what was assumed! Said another way, if a prior is not chosen from fundamental considerations (such as the physics of the problem) but from *ad hoc* assumptions, the results could tend to resemble the assumptions you placed on the problem. On the other hand, if the density is not known, the parameter is termed "nonrandom," and its values range unrestricted over some interval. The resulting nonrandom-parameter estimation problem differs greatly from the random-parameter problem. We consider first the latter problem, letting θ be a scalar parameter having the prior density $p_\theta(\theta)$. The impact of the *a priori* density becomes evident as various error criteria are established, and an "optimum" estimator is derived.

3.2.1 Minimum Mean-Squared Error Estimators

In terms of the densities involved in scalar random-parameter problems, the mean-squared error is given by

$$\mathcal{E}[\varepsilon^2] = \iint (\theta - \widehat{\theta})^2 p_{\mathbf{X},\theta}(\mathbf{x},\theta)\, d\mathbf{x}\, d\theta$$

where $p_{\mathbf{X},\theta}(\mathbf{X},\theta)$ is the joint density of the observations and the parameter. To minimize this integral with respect to $\widehat{\theta}$, we rewrite it using the laws of conditional probability as

$$\mathcal{E}[\varepsilon^2] = \int p_{\mathbf{X}}(\mathbf{x}) \left(\int [\theta - \widehat{\theta}(\mathbf{X})]^2 p_{\theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \right) d\mathbf{x}$$

The density $p_{\mathbf{X}}(\cdot)$ is nonnegative. To minimize the mean-squared error, we must minimize the inner integral for each value of $\mathbf{X}$ because the integral is weighted by a positive quantity. We focus attention on the inner integral, which is the conditional expected value of the squared estimation error. The condition, a fixed value of $\mathbf{X}$, implies that we seek that constant $[\widehat{\theta}(\mathbf{X})]$ derived from $\mathbf{X}$ that minimizes the second moment of the random parameter θ. A well-known result from probability theory states that the minimum of $\mathcal{E}[(x-c)^2]$ occurs when the constant c equals the expected value of the random variable x (see §1.2.2 {4}). The inner integral and thereby the mean-squared error is minimized by choosing the estimator to be the conditional expected value of the parameter given the observations.

$$\boxed{\widehat{\theta}_{\text{MMSE}}(\mathbf{X}) = \mathcal{E}[\theta|\mathbf{X}]}$$

Thus, a parameter's minimum mean-squared error (*MMSE*) estimate is the parameter's *a posteriori* (after the observations have been obtained) expected value.

The associated conditional probability density $p_{\theta|\mathbf{X}}(\theta|\mathbf{X})$ is not often directly stated in a problem definition and must somehow be derived. In many applications, the likelihood function $p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)$ and the *a priori* density of the parameter are a direct consequence of the problem statement. These densities can be used to find the joint density of the observations and the parameter, enabling us to use Bayes's Rule to find the *a posteriori* density *if* we knew the unconditional probability density of the observations.

$$p_{\theta|\mathbf{X}}(\theta|\mathbf{X}) = \frac{p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)p_\theta(\theta)}{p_{\mathbf{X}}(\mathbf{X})}$$

This density $p_{\mathbf{X}}(\mathbf{X})$ is often difficult to determine. Be that as it may, to find the *a posteriori* conditional expected value, it need not be known. The numerator entirely expresses the *a posteriori* density's dependence on θ; the denominator only serves as the scaling factor to yield a unit-area quantity. The expected value is the center-of-mass of the probability density and does *not* depend directly on the "weight" of the density, bypassing calculation of the scaling factor. If not, the *MMSE* estimate can be exceedingly difficult to compute.

Example

Let L statistically independent observations be obtained, each of which is expressed by $X(l) = \theta + N(l)$. Each $N(l)$ is a Gaussian random variable having zero mean and variance σ_N^2. Thus, the unknown parameter in this problem is the mean of the observations. Assume it to be a Gaussian random variable *a priori* (mean m_θ and variance σ_θ^2). The likelihood function is easily found to be

$$p_{\mathbf{X}|\theta}(\mathbf{X}|\theta) = \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left\{-\frac{1}{2}\left(\frac{X(l)-\theta}{\sigma_N}\right)^2\right\}$$

so that the *a posteriori* density is given by

$$p_{\theta|\mathbf{X}}(\theta|\mathbf{X}) = \frac{\dfrac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp\left\{-\dfrac{1}{2}\left(\dfrac{\theta-m_\theta}{\sigma_\theta}\right)^2\right\} \prod_{l=0}^{L-1} \dfrac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left\{-\dfrac{1}{2}\left(\dfrac{X(l)-\theta}{\sigma_N}\right)^2\right\}}{p_{\mathbf{X}}(\mathbf{X})}$$

In an attempt to find the expected value of this distribution, lump all terms that do not depend *explicitly* on the quantity θ into a proportionality term.

$$p_{\theta|\mathbf{X}}(\theta|\mathbf{X}) \propto \exp\left\{-\frac{1}{2}\left[\frac{\sum(X(l)-\theta)^2}{\sigma_N^2} + \frac{(\theta-m_\theta)^2}{\sigma_\theta^2}\right]\right\}$$

After some manipulation, this expression can be written as

$$p_{\theta|\mathbf{X}}(\theta|\mathbf{X}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[\theta - \sigma^2\left(\frac{m_\theta}{\sigma_\theta^2} + \frac{\sum X(l)}{\sigma_N^2}\right)\right]^2\right\}$$

where σ^2 is a quantity that succinctly expresses the ratio $\sigma_N^2\sigma_\theta^2/(\sigma_N^2 + L\sigma_\theta^2)$. The form of the *a posteriori* density suggests that it too is Gaussian; its mean, and therefore the *MMSE* estimate of θ, is given by

$$\widehat{\theta}_{\mathrm{MMSE}}(\mathbf{X}) = \sigma^2\left(\frac{m_\theta}{\sigma_\theta^2} + \frac{\sum X(l)}{\sigma_N^2}\right)$$

More insight into the nature of this estimate is gained by rewriting it as

$$\widehat{\theta}_{\mathrm{MMSE}}(\mathbf{X}) = \frac{\sigma_N^2/L}{\sigma_\theta^2 + \sigma_N^2/L} m_\theta + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_N^2/L} \cdot \frac{1}{L}\sum_{l=0}^{L-1} X(l)$$

The term σ_N^2/L is the variance of the averaged observations for a given value of θ; it expresses the squared error encountered in estimating the mean by simple averaging. If this error is much greater than the *a priori* variance of θ ($\sigma_N^2/L \gg \sigma_\theta^2$), implying that the observations are noisier than the variation of the parameter, the *MMSE* estimate ignores the observations and tends to yield the *a priori* mean m_θ as its value. If the averaged observations are less variable than the parameter, the second term dominates, and the average of the observations is the estimate's value. This estimate behavior between these extremes is very intuitive. The detailed form of the estimate indicates how the squared error can be minimized by a linear combination of these extreme estimates.

The conditional expected value of the estimate equals

$$\mathscr{E}[\widehat{\theta}_{\mathrm{MMSE}}|\theta] = \frac{\sigma_N^2/L}{\sigma_\theta^2 + \sigma_N^2/L} m_\theta + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_N^2/L}\theta$$

This estimate is biased because its expected value does not equal the value of the sought-after parameter. It is asymptotically unbiased as the squared measurement error σ_N^2/L tends to zero as L becomes large. The consistency of the estimator is determined by investigating the expected value of the squared error. Note that the variance of the *a posteriori* density is the quantity σ^2; as this quantity does not depend on $\mathbf{X}$, it also equals the unconditional variance. As the number of observations increases, this variance tends to zero. In concert with the estimate being asymptotically unbiased, the expected value of the estimation error thus tends to zero, implying that we have a consistent estimate.

3.2.2 Maximum a Posteriori Estimators

In those cases in which the expected value of the *a posteriori* density cannot be computed, a related but simpler estimate, the maximum *a posteriori* (*MAP*) estimate, can usually be evaluated. The estimate $\widehat{\theta}_{\text{MAP}}(\mathbf{X})$ equals the location of the maximum of the *a posteriori* density. Assuming that this maximum can be found by evaluating the derivative of the *a posteriori* density, the *MAP* estimate is the solution of the equation

$$\boxed{\left.\frac{\partial p_{\theta|\mathbf{X}}(\theta|\mathbf{X})}{\partial\theta}\right|_{\theta=\widehat{\theta}_{\text{MAP}}}=0}$$

Any scaling of the density by a positive quantity that depends on $\mathbf{X}$ does not change the location of the maximum. Symbolically, $p_{\theta|\mathbf{X}} = p_{\mathbf{X}|\theta}p_\theta/p_{\mathbf{X}}$; the derivative does not involve the denominator, and this term can be ignored. Thus, the only quantities required to compute $\widehat{\theta}_{\text{MAP}}$ are the likelihood function and the parameter's *a priori* density.

Although not apparent in its definition, the *MAP* estimate does satisfy an error criterion. Define a criterion that is zero over a small range of values about $\varepsilon = 0$ and a positive constant outside that range. Minimization of the expected value of this criterion with respect to $\widehat{\theta}$ is accomplished by centering the criterion function at the maximum of the density. The region having the largest area is thus "notched out," and the criterion is minimized. Whenever the *a posteriori* density is symmetric and unimodal, the *MAP* and *MMSE* estimates coincide. In Gaussian problems, such as the last example, this equivalence is always valid. In more general circumstances, they differ.

Example

Let the observations have the same form as the previous example, but with the modification that the parameter is now uniformly distributed over the interval $[\theta_1, \theta_2]$. The *a posteriori* mean cannot be computed in closed form. To obtain the *MAP* estimate, we need to find the location of the maximum of

$$p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)p_\theta(\theta) = \frac{1}{\theta_2-\theta_1}\prod_{l=0}^{L-1}\frac{1}{\sqrt{2\pi\sigma_N^2}}\exp\left\{-\frac{1}{2}\left(\frac{X(l)-\theta}{\sigma_N}\right)^2\right\}, \quad \theta_1 \le \theta \le \theta_2$$

Evaluating the logarithm of this quantity does not change the location of the maximum and simplifies the manipulations in many problems. Here, the logarithm is

$$\ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)p_\theta(\theta) = -\ln(\theta_2-\theta_1) - \sum_{l=0}^{L-1}\left(\frac{X(l)-\theta}{\sigma_N}\right)^2 + \ln C, \quad \theta_1 \le \theta \le \theta_2$$

where C is a constant with respect to θ. Assuming that the maximum is interior to the domain of the parameter, the *MAP* estimate is found to be the sample average $\sum X(l)/L$. If the average lies outside this interval, the corresponding endpoint of the interval is the location of the maximum. To summarize,

$$\widehat{\theta}_{\text{MAP}}(\mathbf{X}) = \begin{cases}\theta_1, & \sum_l X(l)/L < \theta_1\\ \sum_l X(l)/L, & \theta_1 \le \sum_l X(l)/L \le \theta_2\\ \theta_2, & \theta_2 < \sum_l X(l)/L\end{cases}$$

The *a posteriori* density is not symmetric because of the finite domain of θ. Thus, the *MAP* estimate is not equivalent to the *MMSE* estimate, and the accompanying increase in the mean-squared error is difficult to compute. When the sample average is the estimate, the estimate is unbiased; otherwise it is biased. Asymptotically, the variance of the average tends to zero, with the consequences that the estimate is unbiased and consistent.

3.2.3 Linear Estimators

We derived the minimum mean-squared error estimator in the previous section with no constraint on the form of the estimator. Depending on the problem, the computations could be a linear function of the observations (which is always the case in Gaussian problems) or nonlinear. Deriving this estimator is often difficult, which limits its application. We consider here a variation of *MMSE* estimation by constraining the estimator to be linear while minimizing the mean-squared estimation error. Such *linear estimators* may not be optimum; the conditional expected value may be nonlinear and it *always* has the smallest mean-squared error. Despite this occasional performance deficit, linear estimators have well-understood properties, they interact well with other signal processing algorithms because of linearity, and they can always be derived, no matter what the problem.

Let the parameter estimate $\widehat{\theta}(\mathbf{X})$ be expressed as $\mathscr{L}[\mathbf{X}]$, where $\mathscr{L}[\cdot]$ is a linear operator: $\mathscr{L}[a_1\mathbf{X}_1 + a_2\mathbf{X}_2] = a_1\mathscr{L}[\mathbf{X}_1] + a_2\mathscr{L}[\mathbf{X}_2]$, a_1, a_2 scalars. Although all estimators of this form are obviously linear, the term *linear estimator* denotes that member of this family that minimizes the mean-squared estimation error.

$$\arg\min_{\mathscr{L}[\mathbf{X}]} \mathscr{E}[\varepsilon^t\varepsilon] = \widehat{\theta}_{\mathrm{LIN}}(\mathbf{X})$$

Because of the transformation's linearity, the theory of linear vector spaces can be fruitfully used to derive the estimator and to specify its properties. One result of that theoretical framework is the well-known *Orthogonality Principle* [29: 407–14]: The linear estimator is that particular linear transformation that yields an estimation error orthogonal to all linear transformations of the data. The orthogonality of the error to *all* linear transformations is termed the "universality constraint." This principle provides us not only with a formal definition of the linear estimator but also with the mechanism to derive it. To demonstrate this intriguing result, let $\langle\cdot,\cdot\rangle$ denote the abstract inner product between two vectors and $||\cdot||$ the associated norm.

$$||\mathbf{x}||^2 = \langle\mathbf{x},\mathbf{x}\rangle$$

For example, if $\mathbf{x}$ and $\mathbf{y}$ are each column matrices having only one column,* their inner product might be defined as $\langle\mathbf{x},\mathbf{y}\rangle = \mathbf{x}^t\mathbf{y}$. Thus, the linear estimator as defined by the Orthogonality Principle must satisfy

$$\boxed{\mathscr{E}[\langle\widehat{\theta}_{\mathrm{LIN}}(\mathbf{X}) - \theta, \mathscr{L}[\mathbf{X}]\rangle] = 0, \quad \text{for all linear transformations } \mathscr{L}[\cdot]} \tag{3.1}$$

To see that this principle produces the *MMSE* linear estimator, we express the mean-squared estimation error $\mathscr{E}[\varepsilon^t\varepsilon] = \mathscr{E}[||\varepsilon||^2]$ for *any* choice of linear estimator $\widehat{\theta}$ as

$$\begin{aligned}\mathscr{E}[||\widehat{\theta}-\theta||^2] &= \mathscr{E}[||(\widehat{\theta}_{\mathrm{LIN}}-\theta)-(\widehat{\theta}_{\mathrm{LIN}}-\widehat{\theta})||^2]\\ &= \mathscr{E}[||\widehat{\theta}_{\mathrm{LIN}}-\theta||^2] + \mathscr{E}[||\widehat{\theta}_{\mathrm{LIN}}-\widehat{\theta}||^2] - 2\mathscr{E}[\langle\widehat{\theta}_{\mathrm{LIN}}-\theta, \widehat{\theta}_{\mathrm{LIN}}-\widehat{\theta}\rangle]\end{aligned}$$

As $\widehat{\theta}_{\mathrm{LIN}} - \widehat{\theta}$ is the difference of two linear transformations, it too is linear and is orthogonal to the estimation error resulting from $\widehat{\theta}_{\mathrm{LIN}}$. As a result, the last term is zero and the mean-squared estimation error is the sum

*There is a confusion as to what a vector is. "Matrices having one column" are colloquially termed vectors as are the field quantities such as electric and magnetic fields. "Vectors" and their associated inner products are taken to be much more general mathematical objects than these. Hence the prose in this section is rather contorted.

of two squared norms, each of which is, of course, nonnegative. Only the second norm varies with estimator choice; we minimize the mean-squared estimation error by choosing the estimator $\widehat{\theta}$ to be the estimator $\widehat{\theta}_{\text{LIN}}$, which sets the second term to zero.

The estimation error for the minimum mean-squared linear estimator can be calculated to some degree without knowledge of the form of the estimator. The mean-squared estimation error is given by

$$\begin{aligned}\mathscr{E}[\|\widehat{\theta}_{\text{LIN}}-\theta\|^2] &= \mathscr{E}[\langle\widehat{\theta}_{\text{LIN}}-\theta,\widehat{\theta}_{\text{LIN}}-\theta\rangle] \\ &= \mathscr{E}[\langle\widehat{\theta}_{\text{LIN}}-\theta,\widehat{\theta}_{\text{LIN}}\rangle]+\mathscr{E}[\langle\widehat{\theta}_{\text{LIN}}-\theta,-\theta\rangle]\end{aligned}$$

The first term is zero because of the Orthogonality Principle. Rewriting the second term yields a general expression for the *MMSE* linear estimator's mean-squared error.

$$\boxed{\mathscr{E}[\|\varepsilon\|^2] = \mathscr{E}[\|\theta\|^2]-\mathscr{E}[\langle\widehat{\theta}_{\text{LIN}},\theta\rangle]}$$

This error is the difference of two terms. The first, the mean-squared value of the parameter, represents the largest value that the estimation error can be for any reasonable estimator. That error can be obtained by the estimator that ignores the data and has a value of zero. The second term reduces this maximum error and represents the degree to which the estimate and the parameter agree on the average.

Note that the definition of the minimum mean-squared error *linear* estimator makes no explicit assumptions about the parameter estimation problem being solved. This property makes this kind of estimator attractive in many applications where neither the *a priori* density of the parameter vector nor the density of the observations is known precisely. Linear transformations, however, are homogeneous: A zero-valued input yields a zero output. Thus, the linear estimator is especially pertinent to those problems where the expected value of the parameter is zero. If the expected value is nonzero, the linear estimator would not necessarily yield the best result (see Problem 3.14).

Example

Express the first example $\{72\}$ in vector notation so that the observation vector is written as

$$\mathbf{X} = \mathbf{A}\theta + \mathbf{N}$$

where the matrix $\mathbf{A}$ has the form $\mathbf{A} = \text{col}[1,\ldots,1]$. The expected value of the parameter is zero. The linear estimator has the form $\widehat{\theta}_{\text{LIN}} = \mathbf{LX}$, where $\mathbf{L}$ is a $1\times L$ matrix. The Orthogonality Principle states that the linear estimator satisfies

$$\mathscr{E}[(\mathbf{LX}-\theta)^t\mathbf{MX}] = 0, \quad \text{for all } 1\times L \text{ matrices } \mathbf{M}$$

To use the Orthogonality Principle to derive an equation implicitly specifying the linear estimator, the "for all linear transformations" phrase must be interpreted. Usually, the quantity specifying the linear transformation must be removed from the constraining inner product by imposing a very stringent but equivalent condition. In this example, this phrase becomes one about matrices. The elements of the matrix $\mathbf{M}$ can be such that each element of the observation vector multiplies each element of the estimation error. Thus, in this problem the Orthogonality Principle means that the expected value of the matrix consisting of all pairwise products of these elements must be zero.

$$\mathscr{E}[(\mathbf{LX}-\theta)\mathbf{X}^t] = \mathbf{0}$$

Thus, two terms must equal each other: $\mathscr{E}[\mathbf{LXX}^t] = \mathscr{E}[\theta\mathbf{X}^t]$. The second term equals $\mathscr{E}[\theta^2]\mathbf{A}^t$ as the additive noise and the parameter are assumed to be statistically independent quantities. The quantity $\mathscr{E}[\mathbf{XX}^t]$ in the first term is the correlation matrix of the observations, which is given by $\mathbf{AA}^t\mathscr{E}[\theta^2]+\mathbf{K}_N$. Here, $\mathbf{K}_N$ is the noise covariance matrix, and $\mathscr{E}[\theta^2]$ is the parameter's variance. The quantity

$\mathbf{A}\mathbf{A}^t$ is a $L \times L$ matrix with each element equaling 1. The noise vector has independent components; the covariance matrix thus equals $\sigma_N^2 \mathbf{I}$. The equation that $\mathbf{L}$ must satisfy is therefore given by

$$[\mathbf{L}_1 \cdots \mathbf{L}_L] \cdot \begin{bmatrix} \sigma_N^2 + \sigma_\theta^2 & \sigma_\theta^2 & \cdots & \sigma_\theta^2 \\ \sigma_\theta^2 & \sigma_N^2 + \sigma_\theta^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_\theta^2 \\ \sigma_\theta^2 & \cdots & \sigma_\theta^2 & \sigma_N^2 + \sigma_\theta^2 \end{bmatrix} = [\ \sigma_\theta^2 \ \ \cdots \ \ \sigma_\theta^2\ ]$$

The components of $\mathbf{L}$ are equal and are given by $\mathbf{L}_i = \sigma_\theta^2 / (\sigma_N^2 + L\sigma_\theta^2)$. Thus, the minimum mean-squared error linear estimator has the form

$$\widehat{\theta}_{\mathrm{LIN}}(\mathbf{X}) = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_N^2/L} \frac{1}{L} \sum_l X(l)$$

Note that this result equals the minimum mean-squared error estimate derived earlier under the condition that $\mathscr{E}[\theta] = 0$. Mean-squared error, linear estimators, and Gaussian problems are intimately related to each other. The linear minimum mean-squared error solution to a problem is optimal if the underlying distributions are Gaussian.

3.2.4 Maximum Likelihood Estimators

When the *a priori* density of a parameter is not known or the parameter itself is inconveniently described as a random variable, techniques must be developed that make no presumption about the relative possibilities of parameter values. Lacking this knowledge, we can expect the error characteristics of the resulting estimates to be worse than those which can use it.

The maximum likelihood estimate $\widehat{\theta}_{\mathrm{ML}}(\mathbf{X})$ of a nonrandom parameter is, simply, that value which maximizes the likelihood function (the *a priori* density of the observations). Assuming that the maximum can be found by evaluating a derivative, $\widehat{\theta}_{\mathrm{ML}}(\mathbf{X})$ is defined by

$$\boxed{\left.\frac{\partial p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial \theta}\right|_{\theta=\widehat{\theta}_{\mathrm{ML}}} = 0} \, .$$

The logarithm of the likelihood function may also be used in this maximization.

Example

Let $X(l)$ be a sequence of independent, identically distributed Gaussian random variables having an unknown mean θ but a known variance σ_N^2. Often, we cannot assign a probability density to a parameter of a random variable's density; we simply do not know what the parameter's value is. Maximum likelihood estimates are often used in such problems. In the specific case here, the derivative of the logarithm of the likelihood function equals

$$\frac{\partial \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial \theta} = \frac{1}{\sigma_N^2} \sum_{l=0}^{L-1} [X(l) - \theta] \, .$$

The solution of this equation is the maximum likelihood estimate, which equals the sample average.

$$\widehat{\theta}_{\mathrm{ML}} = \frac{1}{L} \sum_{l=0}^{L-1} X(l)$$

The expected value of this estimate $\mathscr{E}[\widehat{\theta}_{\text{ML}}|\theta]$ equals the actual value θ, showing that the maximum likelihood estimate is unbiased. The mean-square error equals σ_N^2/L and we infer that this estimate is consistent.

Parameter Vectors

The maximum likelihood procedure (as well as the others being discussed) can be easily generalized to situations where more than one parameter must be estimated. Letting θ denote the parameter vector, the likelihood function is now expressed as $p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)$. The maximum likelihood estimate $\widehat{\theta}_{\text{ML}}$ of the parameter vector is given by the location of the maximum of the likelihood function (or equivalently of its logarithm). Using derivatives, the calculation of the maximum likelihood estimate becomes

$$\boxed{\nabla_\theta \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)\Big|_{\theta=\widehat{\theta}_{\text{ML}}} = \mathbf{0}},$$

where ∇_θ denotes the gradient with respect to the parameter vector. This equation means that we must estimate all of the parameters *simultaneously* by setting the partial of the likelihood function with respect to *each* parameter to zero. Given P parameters, we must solve in most cases a set of P nonlinear, simultaneous equations to find the maximum likelihood estimates.

Example

Let's extend the previous example to the situation where neither the mean nor the variance of a sequence of independent Gaussian random variables is known. The likelihood function is, in this case,

$$p_{\mathbf{X}|\theta}(\mathbf{X}|\theta) = \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi\theta_2}} \exp\left\{-\frac{1}{2\theta_2}\left[X(l)-\theta_1\right]^2\right\}.$$

Evaluating the partial derivatives of the logarithm of this quantity, we find the following set of two equations to solve for θ_1, representing the mean, and θ_2, representing the variance.*

$$\frac{1}{\theta_2}\sum_{l=0}^{L-1}[X(l)-\theta_1] = 0$$

$$-\frac{L}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{l=0}^{L-1}[X(l)-\theta_1]^2 = 0$$

The solution of this set of equations is easily found to be

$$\widehat{\theta}_1^{\text{ML}} = \frac{1}{L}\sum_{l=0}^{L-1} X(l)$$

$$\widehat{\theta}_2^{\text{ML}} = \frac{1}{L}\sum_{l=0}^{L-1}\left(X(l)-\widehat{\theta}_1^{\text{ML}}\right)^2$$

The expected value of $\widehat{\theta}_1^{\text{ML}}$ equals the actual value of θ_1; thus, this estimate is unbiased. However, the expected value of the estimate of the variance equals $\theta_2 \cdot (L-1)/L$. The estimate of the variance is biased, but asymptotically unbiased. This bias can be removed by replacing the normalization of L in the averaging computation for $\widehat{\theta}_2^{\text{ML}}$ by $L-1$.

*The variance rather than the standard deviation is represented by θ_2. The mathematics is messier and the estimator has less attractive properties in the latter case. Problem 3.8 illustrates this point.

Cramér-Rao Bound

The mean-square estimation error for *any* estimate of a nonrandom parameter has a lower bound, the *Cramér-Rao bound* [6: pp. 474–477], which defines the ultimate accuracy of *any* estimation procedure. This lower bound, as shown later, is intimately related to the maximum likelihood estimator.

We seek a "bound" on the mean-squared error matrix $\mathbf{M}$ defined to be

$$\mathbf{M} = \mathscr{E}[(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^t] = \mathscr{E}[\varepsilon\varepsilon^t] \,.$$

A matrix is "lower bounded" by a second matrix if the difference between the two is a non-negative definite matrix. Define the column matrix $\mathbf{x}$ to be

$$\mathbf{x} = \begin{bmatrix} \widehat{\theta} - \theta - \mathbf{b}(\theta) \\ \nabla_\theta \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta) \end{bmatrix} ,$$

where $\mathbf{b}(\theta)$ denotes the column matrix of estimator biases. To derive the Cramér-Rao bound, evaluate $\mathscr{E}[\mathbf{x}\mathbf{x}^t]$.

$$\mathscr{E}[\mathbf{x}\mathbf{x}^t] = \begin{bmatrix} \mathbf{M} - \mathbf{b}\mathbf{b}^t & \mathbf{I} + \nabla_\theta \mathbf{b} \\ (\mathbf{I} + \nabla_\theta \mathbf{b})^t & \mathbf{F} \end{bmatrix}$$

where $\nabla_\theta \mathbf{b}$ represents the matrix of partial derivatives of the bias $[\partial b_i / \partial \theta_j]$ and the matrix $\mathbf{F}$ is the *Fisher information matrix*

$$\boxed{\mathbf{F} = \mathscr{E}\left[\left(\nabla_\theta \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)\right)\left(\nabla_\theta \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)\right)^t\right]} \tag{3.2}$$

Note that this matrix can alternatively be expressed as

$$\mathbf{F} = -\,\mathscr{E}\left[\nabla_\theta \nabla_\theta^t \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)\right] \,.$$

The notation $\nabla_\theta \nabla_\theta^t$ means the matrix of all second partials of the quantity it operates on (the gradient of the gradient). This matrix is known as the Hessian. Demonstrating the equivalence of these two forms for the Fisher information is quite easy. Because $\int p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)\,d\mathbf{X} = 1$ for all choices of the parameter vector, the gradient of this expression equals zero. Furthermore, $\nabla_\theta \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta) = \nabla_\theta\, p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)/p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)$. Combining these results yields

$$\int \left(\nabla_\theta \ln p_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\right) p_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\,d\mathbf{x} = \mathbf{0} \,.$$

Evaluating the gradient of this quantity (using the chain rule) also yields zero.

$$\begin{aligned} &\int \left(\nabla_\theta \nabla_\theta^t \ln p_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\right) p_{\mathbf{X}|\theta}(\mathbf{x}|\theta) + \left(\nabla_\theta \ln p_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\right)\left(\nabla_\theta \ln p_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\right)^t p_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\,d\mathbf{x} = 0 \\ \text{or}\quad &\mathscr{E}\left[\left(\nabla_\theta \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)\right)\left(\nabla_\theta \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)\right)^t\right] = -\,\mathscr{E}\left[\nabla_\theta \nabla_\theta^t \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)\right] \end{aligned}$$

Calculating the expected value for the Hessian form is sometimes easier than finding the expected value of the outer product of the gradient with itself. In the scalar case, we have

$$\mathscr{E}\left[\left(\frac{\partial \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial\theta}\right)^2\right] = -\,\mathscr{E}\left[\frac{\partial^2 \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial\theta^2}\right] .$$

Returning to the derivation, the matrix $\mathscr{E}[\mathbf{x}\mathbf{x}^t]$ is non-negative definite because it is a correlation matrix. Thus, for any column matrix α, the quadratic form $\alpha^t\,\mathscr{E}[\mathbf{x}\mathbf{x}^t]\alpha$ is non-negative. Choose a form for α that simplifies the quadratic form. A convenient choice is

$$\alpha = \begin{bmatrix} \beta \\ -\mathbf{F}^{-1}\left(\mathbf{I} + \nabla_\theta \mathbf{b}\right)^t \beta \end{bmatrix} ,$$

where β is an arbitrary column matrix. The quadratic form becomes in this case

$$\alpha^t\,\mathscr{E}[\mathbf{x}\mathbf{x}^t]\alpha = \beta^t\left[\mathbf{M}-\mathbf{b}\mathbf{b}^t-\left(\mathbf{I}+\nabla_\theta\,\mathbf{b}\right)\mathbf{F}^{-1}\left(\mathbf{I}+\nabla_\theta\,\mathbf{b}\right)^t\right]\beta\,.$$

As this quadratic form must be non-negative, the matrix expression enclosed in brackets must be non-negative definite. We thus obtain the well-known Cramér-Rao bound on the mean-square error matrix.

$$\boxed{\mathscr{E}[\varepsilon\varepsilon^t]\geq \mathbf{b}(\theta)\mathbf{b}^t(\theta)+\left(\mathbf{I}+\nabla_\theta\,\mathbf{b}\right)\mathbf{F}^{-1}\left(\mathbf{I}+\nabla_\theta\,\mathbf{b}\right)^t}$$

This form for the Cramér-Rao bound does *not* mean that each term in the matrix of squared errors is greater than the corresponding term in the bounding matrix. As stated earlier, this expression means that the difference between these matrices is non-negative definite. For a matrix to be non-negative definite, each term on the main diagonal must be non-negative. The elements of the main diagonal of $\mathscr{E}[\varepsilon\varepsilon^t]$ are the squared errors of the estimate of the individual parameters. Thus, for each parameter, the mean-squared estimation error can be no smaller than

$$\mathscr{E}[(\widehat{\theta}_i-\theta_i)^2]\geq b_i^2(\theta)+\left[\left(\mathbf{I}+\nabla_\theta\,\mathbf{b}\right)\mathbf{F}^{-1}\left(\mathbf{I}+\nabla_\theta\,\mathbf{b}\right)^t\right]_{ii}.$$

This bound simplifies greatly if the estimator is unbiased ($\mathbf{b}=\mathbf{0}$). In this case, the Cramér-Rao bound becomes

$$\mathscr{E}[(\widehat{\theta}_i-\theta_i)^2]\geq \mathbf{F}_{ii}^{-1}.$$

Thus, the mean-squared error for each parameter in a multiple-parameter, unbiased-estimator problem can be no smaller than the corresponding diagonal term in the *inverse* of the Fisher information matrix. In such problems, the estimate's error characteristics of any parameter become intertwined with the other parameters in a complicated way. Any estimator satisfying the Cramér-Rao bound with equality is said to be *efficient*.

Example

Let's evaluate the Cramér-Rao bound for the example we have been discussing: the estimation of the mean and variance of a length L sequence of statistically independent Gaussian random variables. Let the estimate of the mean θ_1 be the sample average $\widehat{\theta}_1=\sum X(l)/L$; as shown in the last example, this estimate is unbiased. Let the estimate of the variance θ_2 be the unbiased estimate $\widehat{\theta}_2=[\sum(X(l)-\widehat{\theta}_1)^2]/(L-1)$. Each term in the Fisher information matrix $\mathbf{F}$ is given by the expected value of the paired products of derivatives of the logarithm of the likelihood function.

$$\mathbf{F}_{ij}=\mathscr{E}\left[\frac{\partial\ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial\theta_i}\frac{\partial\ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial\theta_j}\right]$$

The logarithm of the likelihood function is

$$\ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)=-\frac{L}{2}\ln 2\pi\theta_2-\frac{1}{2\theta_2}\sum_{l=0}^{L-1}[X(l)-\theta_1]^2\,,$$

its partial derivatives are

$$\frac{\partial\ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial\theta_1}=\frac{1}{\theta_2}\sum_{l=0}^{L-1}[X(l)-\theta_1] \tag{3.3}$$

$$\frac{\partial\ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial\theta_2}=-\frac{L}{2\theta_2}+\frac{1}{2\theta_2^2}\sum_{l=0}^{L-1}[X(l)-\theta_1]^2 \tag{3.4}$$

and its second partials are

$$\begin{aligned}
\frac{\partial^2 \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial\theta_1^2} &= -\frac{L}{\theta_2} & \frac{\partial^2 \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial\theta_1\partial\theta_2} &= -\frac{1}{\theta_2^2}\sum_{l=0}^{L-1}[X(l)-\theta_1] \\
\frac{\partial^2 \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial\theta_2\partial\theta_1} &= -\frac{1}{\theta_2^2}\sum_{l=0}^{L-1}[X(l)-\theta_1] & \frac{\partial^2 \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial\theta_2^2} &== \frac{L}{2\theta_2^2}-\frac{1}{\theta_2^3}\sum_{l=0}^{L-1}[X(l)-\theta_1]^2
\end{aligned}$$

The Fisher information matrix has the surprisingly simple form

$$\mathbf{F} = \begin{bmatrix} \frac{L}{\theta_2} & 0 \\ 0 & \frac{L}{2\theta_2^2} \end{bmatrix};$$

its inverse is also a diagonal matrix with the elements on the main diagonal equalling the reciprocal of those in the original matrix. Because of the zero-valued off-diagonal entries in the Fisher information matrix, the errors between the corresponding estimates are not inter-dependent. In this problem, the mean-square estimation errors can be no smaller than

$$\begin{aligned}
\mathcal{E}[(\widehat{\theta}_1-\theta_1)^2] &\geq \frac{\theta_2}{L} \\
\mathcal{E}[(\widehat{\theta}_2-\theta_2)^2] &\geq \frac{2\theta_2^2}{L}
\end{aligned}$$

Note that *nowhere* in the preceding example did the form of the estimator enter into the computation of the bound. The only quantity used in the computation of the Cramér-Rao bound is the logarithm of the likelihood function, which is a consequence of the problem statement, not how it is solved. *Only in the case of unbiased estimators is the bound independent of the estimators used.*[*] Because of this property, the Cramér-Rao bound is frequently used to assess the performance limits that can be obtained with an unbiased estimator in a particular problem. When bias is present, the exact form of the estimator's bias explicitly enters the computation of the bound. All too frequently, the unbiased form is used in situations where the *existence* of an unbiased estimator can be questioned. As we shall see, one such problem is time delay estimation, presumably of some importance to the reader. This misapplication of the unbiased Cramér-Rao arises from desperation: the estimator is so complicated and nonlinear that computing the bias is nearly impossible. As shown in Problem 3.9, biased estimators can yield mean-squared errors smaller as well as larger than the unbiased version of the Cramér-Rao bound. Consequently, desperation can yield misinterpretation when a general result is misapplied.

In the single-parameter estimation problem, the Cramér-Rao bound incorporating bias has the well-known form[†]

$$\boxed{\mathcal{E}[\varepsilon^2] \geq b^2 + \frac{\left(1+\frac{db}{d\theta}\right)^2}{\mathcal{E}\left[\left(\frac{\partial \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial\theta}\right)^2\right]}}\,.$$

Note that the sign of the bias's derivative determines whether this bound is larger or potentially smaller than the unbiased version, which is obtained by setting the bias term to zero.

[*] That's why we assumed in the example that we used an unbiased estimator for the variance.

[†] Note that this bound differs somewhat from that originally given by Cramér [6: p. 480]; his derivation ignores the additive bias term $\mathbf{bb}'$.

Efficiency

An interesting question arises: when, if ever, is the bound satisfied with equality? Recalling the details of the derivation of the bound, equality results when the quantity $\mathscr{E}[\alpha^t \mathbf{x}\mathbf{x}^t \alpha]$ equals zero. As this quantity is the expected value of the square of $\alpha^t \mathbf{x}$, it can only equal zero if $\alpha^t \mathbf{x} = 0$. Substituting in the form of the column matrices α and $\mathbf{x}$, equality in the Cramér-Rao bound results whenever

$$\boxed{\nabla_\theta \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta) = \left[\mathbf{I} + (\nabla_\theta \mathbf{b})^t\right]^{-1} \mathbf{F}[\widehat{\theta}(\mathbf{X}) - \theta - \mathbf{b}]}\,. \tag{3.5}$$

This complicated expression means that only if estimation problems (as expressed by the *a priori* density) have the form of the right side of this equation can the mean-square estimation error equal the Cramér-Rao bound. In particular, the gradient of the log likelihood function can *only* depend on the observations through the estimator. In all other problems, the Cramér-Rao bound is a lower bound but not a tight one; *no* estimator can have error characteristics that equal it. In such cases, we have limited insight into ultimate limitations on estimation error size with the Cramér-Rao bound. However, consider the case where the estimator is unbiased ($\mathbf{b} = \mathbf{0}$). In addition, note the maximum likelihood estimate occurs when the gradient of the logarithm of the likelihood function equals zero: $\nabla_\theta \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta) = 0$ when $\theta = \widehat{\theta}_{\mathrm{ML}}$. In this case, the condition for equality in the Cramér-Rao bound becomes

$$\mathbf{F}[\widehat{\theta} - \widehat{\theta}_{\mathrm{ML}}] = \mathbf{0}\,.$$

As the Fisher information matrix is positive-definite, we conclude that if the estimator equals the maximum likelihood estimator, equality in the Cramér-Rao bound can be achieved. To summarize, if the Cramér-Rao bound *can* be satisfied with equality, *only* the maximum likelihood estimate will achieve it. To use estimation theoretic terminology, *if an efficient estimate exists, it is the maximum likelihood estimate*. This result stresses the importance of maximum likelihood estimates, despite the seemingly *ad hoc* manner by which they are defined.

Example

Consider the Gaussian example being examined so frequently in this section. The components of the gradient of the logarithm of the likelihood function were given earlier by equations (3.3) {79}. These expressions can be rearranged to reveal

$$\begin{bmatrix} \dfrac{\partial \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial \theta_1} \\[2ex] \dfrac{\partial \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} \frac{L}{\theta_2}\left[\left(\frac{1}{L}\sum_l X(l)\right) - \theta_1\right] \\[1ex] -\frac{L}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_l [X(l) - \theta_1]^2 \end{bmatrix}.$$

The first component, which corresponds to the estimate of the mean, *is* expressed in the form required for the existence of an efficient estimate. The second component—the partial with respect to the variance θ_2—*cannot* be rewritten in a similar fashion. No unbiased, efficient estimate of the variance exists in this problem. The mean-squared error of the variance's unbiased estimate, but not the maximum likelihood estimate, is lower-bounded by $2\theta_2^2/(L-1)^2$. This error is strictly greater than the Cramér-Rao bound of $2\theta_2^2/L^2$. As no unbiased estimate of the variance can have a mean-squared error equal to the Cramér-Rao bound (no efficient estimate exists for the variance in the Gaussian problem), one presumes that the closeness of the error of our unbiased estimator to the bound implies that it possesses the smallest squared-error of any estimate. This presumption may, of course, be incorrect.

Properties of the Maximum Likelihood Estimator

The maximum likelihood estimate is the most used estimation technique for nonrandom parameters. Not only because of its close linkage to the Cramér-Rao bound, but also because it has desirable asymptotic properties in the context of *any* Problem [6: pp. 500–6].

1. *The maximum likelihood estimate is at least asymptotically unbiased.* It may be unbiased for any number of observations (as in the estimation of the mean of a sequence of independent random variables) for some problems.
2. *The maximum likelihood estimate is consistent.*
3. *The maximum likelihood estimate is asymptotically efficient.* As more and more data are incorporated into an estimate, the Cramér-Rao bound accurately projects the best attainable error and the maximum likelihood estimate has those optimal characteristics.
4. *Asymptotically, the maximum likelihood estimate is distributed as a Gaussian random variable.* Because of the previous properties, the mean asymptotically equals the parameter and the covariance matrix is $[L\mathbf{F}(\theta)]^{-1}$.

Most would agree that a "good" estimator should have these properties. What these results do not provide is an assessment of how many observations are needed for the asymptotic results to apply to some specified degree of precision. Consequently, they should be used with caution; for instance, some other estimator may have a smaller mean-square error than the maximum likelihood for a modest number of observations.

3.3 Signal Parameter Estimation

One extension of parametric estimation theory necessary for its application to signal processing is the estimation of signal parameters. We assume that we observe a signal $s(l,\theta)$, whose characteristics are known save a few parameters θ, in the presence of noise. Signal parameters, such as amplitude, time origin, and frequency if the signal is sinusoidal, must be determined in some way. In many cases of interest, we would find it difficult to justify a particular form for the unknown parameters' *a priori* density. Because of such uncertainties, the minimum mean-squared error and maximum *a posteriori* estimators *cannot* be used in many cases. The minimum mean-squared error *linear* estimator does not require this density, but it is most fruitfully used when the unknown parameter appears in the problem in a linear fashion (such as signal amplitude as we shall see).

3.3.1 Linear Minimum Mean-Squared Error Estimator

The only parameter that is linearly related to a signal is the amplitude. Consider, therefore, the problem where the observations are modeled as

$$X(l) = \theta s(l) + N(l), \quad l = 0,\dots,L-1$$

The signal waveform $s(l)$ is known and its energy normalized to be unity ($\sum s^2(l) = 1$). The linear estimate of the signal's amplitude is assumed to be of the form $\widehat{\theta} = \sum h(l)X(l)$, where $h(l)$ minimizes the mean-squared error. To use the Orthogonality Principle expressed by Eq. 3.1 $\{74\}$, an inner product must be defined for scalars. Little choice avails itself but multiplication as the inner product of two scalars. The Orthogonality Principle states that the estimation error must be orthogonal to all linear transformations defining the kind of estimator being sought.

$$\mathscr{E}\left[\left(\sum_{l=0}^{L-1} h_{\mathrm{LIN}}(l)X(l) - \theta\right)\sum_{k=0}^{L-1} h(k)X(k)\right] = 0 \quad \text{for all } h(\cdot)$$

Manipulating this equation to make the universality constraint more transparent results in

$$\sum_{k=0}^{L-1} h(k)\cdot\mathscr{E}\left[\left(\sum_{l=0}^{L-1} h_{\mathrm{LIN}}(l)X(l) - \theta\right)X(k)\right] = 0 \quad \text{for all } h(\cdot)$$

Written in this way, the expected value must be 0 for each value of k to satisfy the constraint. Thus, the quantity $h_{\mathrm{LIN}}(\cdot)$ of the estimator of the signal's amplitude must satisfy

$$\sum_{l=0}^{L-1} h_{\mathrm{LIN}}(l)\,\mathscr{E}[X(l)X(k)] = \mathscr{E}[\theta X(k)] \quad \text{for all } k$$

Assuming that the signal's amplitude has zero mean and is statistically independent of the zero-mean noise, the expected values in this equation are given by

$$\begin{aligned} \mathscr{E}[X(l)X(k)] &= \sigma_\theta^2 s(l)s(k) + K_N(k,l) \\ \mathscr{E}[\theta X(k)] &= \sigma_\theta^2 s(k) \end{aligned}$$

where $K_N(k,l)$ is the covariance function of the noise. The equation that must be solved for the unit-sample response $h_{\text{LIN}}(\cdot)$ of the optimal linear *MMSE* estimator of signal amplitude becomes

$$\boxed{\sum_{l=0}^{L-1} h_{\text{LIN}}(l)K_N(k,l) = \sigma_\theta^2 s(k)\left[1 - \sum_{l=0}^{L-1} h_{\text{LIN}}(l)s(l)\right] \quad \text{for all } k}$$

This equation is easily solved once phrased in matrix notation. Letting $\mathbf{K}_N$ denote the covariance matrix of the noise, $\mathbf{s}$ the signal vector, and $\mathbf{h}_{\text{LIN}}$ the vector of coefficients, this equation becomes

$$\mathbf{K}_N\mathbf{h}_{\text{LIN}} = \sigma_\theta^2[1 - \mathbf{s}^t\mathbf{h}_{\text{LIN}}]\mathbf{s}$$

The matched filter for colored-noise problems consisted of the dot product between the vector of observations and $\mathbf{K}_N^{-1}\mathbf{s}$ (see the detector result $\{127\}$). Assume that the solution to the linear estimation problem is proportional to the detection theoretical one: $\mathbf{h}_{\text{LIN}} = c\mathbf{K}_N^{-1}\mathbf{s}$, where c is a scalar constant. This proposed solution satisfies the equation; the *MMSE* estimate of signal amplitude corresponds to applying a matched filter to the observations with

$$\boxed{\mathbf{h}_{\text{LIN}} = \frac{\sigma_\theta^2}{1 + \sigma_\theta^2\mathbf{s}^t\mathbf{K}_N^{-1}\mathbf{s}}\mathbf{K}_N^{-1}\mathbf{s}}$$

The mean-squared estimation error of signal amplitude is given by

$$\mathscr{E}[\varepsilon^2] = \sigma_\theta^2 - \mathscr{E}\left[\theta\sum_{l=0}^{L-1} h_{\text{LIN}}(l)X(l)\right]$$

Substituting the vector expression for $\mathbf{h}_{\text{LIN}}$ yields the result that the mean-squared estimation error equals the proportionality constant c defined earlier.

$$\mathscr{E}[\varepsilon^2] = \frac{\sigma_\theta^2}{1 + \sigma_\theta^2\mathbf{s}^t\mathbf{K}_N^{-1}\mathbf{s}}$$

Thus, the linear filter that produces the optimal estimate of signal amplitude is equivalent to the matched filter used to detect the signal's presence. We have found this situation to occur when estimates of unknown parameters are needed to solve the detection problem. If we had not assumed the noise to be Gaussian, however, this detection-theoretic result would be different, but the estimator would be unchanged. To repeat, this invariance occurs because the linear *MMSE* estimator requires *no* assumptions on the noise's amplitude characteristics.

Example

Let the noise be white so that its covariance matrix is proportional to the identity matrix ($\mathbf{K}_N = \sigma_N^2\mathbf{I}$). The weighting factor in the minimum mean-squared error linear estimator is proportional to the signal waveform.

$$h_{\text{LIN}}(l) = \frac{\sigma_\theta^2}{\sigma_N^2 + \sigma_\theta^2}s(l) \qquad \widehat{\theta}_{\text{LIN}} = \frac{\sigma_\theta^2}{\sigma_N^2 + \sigma_\theta^2}\sum_{l=0}^{L-1} s(l)X(l)$$

This proportionality constant depends only on the relative variances of the noise and the parameter. *If* the noise variance can be considered to be much smaller than the *a priori* variance of the amplitude, then this constant does not depend on these variances and equals unity. Otherwise, the variances must be known.

We find the mean-squared estimation error to be

$$\mathscr{E}[\varepsilon^2] = \frac{\sigma_\theta^2}{1+\sigma_\theta^2/\sigma_N^2}$$

This error is significantly reduced from its nominal value σ_θ^2 only when the variance of the noise is small compared with the *a priori* variance of the amplitude. Otherwise, this admittedly optimum amplitude estimate performs poorly, and we might as well as have ignored the data and "guessed" that the amplitude was zero.*

3.3.2 Maximum Likelihood Estimators

Many situations are either not well suited to linear estimation procedures, or the parameter is not well described as a random variable. For example, signal delay is observed nonlinearly and usually no *a priori* density can be assigned. In such cases, maximum likelihood estimators are more frequently used. Because of the Cramér-Rao bound, fundamental limits on parameter estimation performance can be derived for *any* signal parameter estimation problem where the parameter is not random.

Assume that the data are expressed as a signal observed in the presence of additive Gaussian noise.

$$X(l) = s(l,\theta) + N(l), \quad l = 0,\ldots,L-1$$

The vector of observations $\mathbf{X}$ is formed from the data in the obvious way. Evaluating the logarithm of the observation vector's joint density,

$$\ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta) = -\frac{1}{2}\ln\det[2\pi\mathbf{K}_N] - \frac{1}{2}[\mathbf{X}-\mathbf{s}(\theta)]^t\mathbf{K}_N^{-1}[\mathbf{X}-\mathbf{s}(\theta)]$$

where $\mathbf{s}(\theta)$ is the signal vector having P unknown parameters, and $\mathbf{K}_N$ is the covariance matrix of the noise. The partial derivative of this likelihood function with respect to the i^{th} parameter θ_i is, for real-valued signals,

$$\frac{\partial \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial \theta_i} = [\mathbf{X}-\mathbf{s}(\theta)]^t\mathbf{K}_N^{-1}\frac{\partial \mathbf{s}(\theta)}{\partial \theta_i}$$

and, for complex-valued ones,

$$\frac{\partial \ln p_{\mathbf{X}|\theta}(\mathbf{X}|\theta)}{\partial \theta_i} = \mathrm{Re}\left[[\mathbf{X}-\mathbf{s}(\theta)]'\mathbf{K}_N^{-1}\frac{\partial \mathbf{s}(\theta)}{\partial \theta_i}\right]$$

If the maximum of the likelihood function can be found by setting its gradient to $\mathbf{0}$, the maximum likelihood estimate of the parameter vector is the solution of the set of equations

$$\boxed{[\mathbf{X}-\mathbf{s}(\theta)]^t\mathbf{K}_N^{-1}\frac{\partial \mathbf{s}(\theta)}{\partial \theta_i}\bigg|_{\theta=\widehat{\theta}_{\mathrm{ML}}} = 0, \quad i = 1,\ldots,P}$$

The Cramér-Rao bound depends on the evaluation of the Fisher information matrix $\mathbf{F}$. The elements of this matrix are found to be

$$F_{ij} = \frac{\partial \mathbf{s}^t(\theta)}{\partial \theta_i}\mathbf{K}_N^{-1}\frac{\partial \mathbf{s}(\theta)}{\partial \theta_j}, \quad i,j = 1,\ldots,P \tag{3.6}$$

*In other words, the problem is difficult in this case.

Further computation of the Cramér-Rao bound's components is problem dependent if more than one parameter is involved, and the off-diagonal terms of $\mathbf{F}$ are nonzero. If only one parameter is unknown, the Cramér-Rao bound is given by

$$\boxed{\mathscr{E}[\varepsilon^2] \geq b^2(\theta) + \frac{\left(1+\dfrac{db(\theta)}{d\theta}\right)^2}{\dfrac{\partial \mathbf{s}^t(\theta)}{\partial\theta}\mathbf{K}_N^{-1}\dfrac{\partial \mathbf{s}(\theta)}{\partial\theta}}}$$

When the signal depends on the parameter nonlinearly (which constitute the interesting cases), the maximum likelihood estimate is usually biased. Thus, the numerator of the expression for the bound cannot be ignored. One interesting special case occurs when the noise is white. The Cramér-Rao bound becomes

$$\mathscr{E}[\varepsilon^2] \geq b^2(\theta) + \frac{\sigma_N^2\left(1+\dfrac{db(\theta)}{d\theta}\right)^2}{\displaystyle\sum_{l=0}^{L-1}\left(\frac{\partial s(l,\theta)}{\partial\theta}\right)^2}$$

The derivative of the signal with respect to the parameter can be interpreted as the sensitivity of the signal to the parameter. The mean-squared estimation error depends on the "integrated" squared sensitivity: The greater this sensitivity, the smaller the bound.

For an efficient estimate of a signal parameter to exist, the estimate must satisfy the condition we derived earlier (Eq. 3.5 {81}).

$$[\nabla_\theta \mathbf{s}(\theta)]^t\mathbf{K}_N^{-1}[\mathbf{X}-\mathbf{s}(\theta)] \overset{?}{=} \left[\mathbf{I}+(\nabla_\theta \mathbf{b})^t\right]^{-1}[\nabla_\theta \mathbf{s}(\theta)]^t\mathbf{K}_N^{-1}[\nabla_\theta \mathbf{s}(\theta)][\widehat{\theta}(\mathbf{X})-\theta-\mathbf{b}]$$

Because of the complexity of this requirement, we quite rightly question the existence of any efficient estimator, especially when the signal depends nonlinearly on the parameter (see Problem 3.15).

Example

Let the unknown parameter be the signal's amplitude; the signal is expressed as $\theta s(l)$ and is observed in the presence of additive noise. The maximum likelihood estimate of the amplitude is the solution of the equation

$$[\mathbf{X}-\widehat{\theta}_{\mathrm{ML}}\mathbf{s}]^t\mathbf{K}_N^{-1}\mathbf{s}=0$$

The form of this equation suggests that the maximum likelihood estimate is efficient. The amplitude estimate is given by

$$\widehat{\theta}_{\mathrm{ML}} = \frac{\mathbf{X}^t\mathbf{K}_N^{-1}\mathbf{s}}{\mathbf{s}^t\mathbf{K}_N^{-1}\mathbf{s}}$$

The form of this estimator is precisely that of the matched filter derived in the colored-noise situation (see Eq. 4.9 {127}). The expected value of the estimate equals the actual amplitude. Thus the bias is zero and the Cramér-Rao bound is given by

$$\mathscr{E}[\varepsilon^2] \geq \left(\mathbf{s}^t\mathbf{K}_N^{-1}\mathbf{s}\right)^{-1}$$

The condition for an efficient estimate becomes

$$\mathbf{s}^t\mathbf{K}_N^{-1}(\mathbf{X}-\theta\mathbf{s}) \overset{?}{=} \mathbf{s}^t\mathbf{K}_N^{-1}\mathbf{s}\cdot(\widehat{\theta}_{\mathrm{ML}}-\theta)$$

whose veracity we can easily verify.

In the special case where the noise is white, the estimator has the form $\widehat{\theta}_{\mathrm{ML}}=\mathbf{X}^t\mathbf{s}$, and the Cramér-Rao bound equals σ_N^2 (the nominal signal is assumed to have unit energy). The maximum likelihood

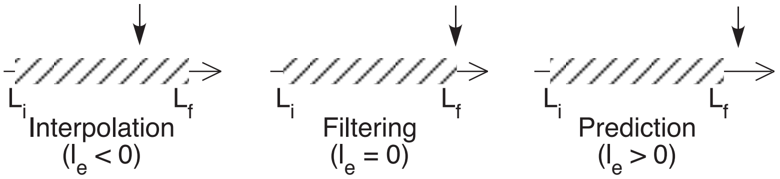


Figure 3.1: The three classical categories of linear signal waveform estimation are defined by the observation interval's relation to the time at which we want to estimate the signal value. As time evolves, so does the observation interval so that l_e, the interval between the last observation and the estimation time, is fixed.

estimate of the amplitude has *fixed* error characteristics that do not depend on the actual signal amplitude. A signal-to-noise ratio for the estimate, defined to be $\theta^2/\mathscr{E}[\varepsilon^2]$, equals the signal-to-noise ratio of the observed signal.

When the amplitude is well described as a random variable, its linear minimum mean-squared error estimator has the form

$$\widehat{\theta}_{\text{LIN}} = \frac{\sigma_\theta^2 \mathbf{X}^t \mathbf{K}_N^{-1} \mathbf{s}}{1 + \sigma_\theta^2 \mathbf{s}^t \mathbf{K}_N^{-1} \mathbf{s}}$$

which we found in the white-noise case becomes a weighted version of the maximum likelihood estimate (see the example $\{83\}$).

$$\widehat{\theta}_{\text{LIN}} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_N^2} \mathbf{X}^t \mathbf{s}$$

Seemingly, these two estimators are being used to solve the same problem: Estimating the amplitude of a signal whose waveform is known. They make very different assumptions, however, about the nature of the unknown parameter; in one it is a random variable (and thus it has a variance), whereas in the other it is not (and variance makes no sense). Despite this fundamental difference, the computations for each estimator are equivalent. It is reassuring that different approaches to solving similar problems yield similar procedures.

3.4 Linear Signal Waveform Estimation

When the details of a signal's waveform are unknown, describing the signal parametrically is usually unsatisfactory. We need techniques that estimate waveforms rather than numbers. For example, we may want to know the propagating signal's waveform contained in the noise-corrupted array output. Without some *a priori* information, this task is impossible; if neither the signal nor the noise is known, how can anyone discriminate one from the other? The key to waveform estimation is how much prior information we have about the signal and the noise, and how valid that information is. Given noisy observations of a signal throughout the interval $[L_i, L_f]$, the waveform estimation problem is to estimate accurately the value of the signal at some moment $L_f + l_e$. In most situations, the observation interval evolves with the passage of time while the estimation time is fixed relative to the occurrence of the most recent observation (in other words, l_e is a constant). Linear waveform estimation results when we apply a linear filter to the observations.

Waveform estimation problems are usually placed into one of three categories [1: 9–11] based on the value of l_e (see Fig. 3.1):

Interpolation. The interpolation or smoothing problem is to estimate the signal at some moment within the observation interval ($l_e < 0$). Observations are thus considered before and after the time at which the signal needs to be estimated. In practice, applying interpolation filtering means that the estimated signal waveform is produced some time *after* it occurred.

Filtering. We estimate the signal at the end of the observation interval ($l_e = 0$). Thus, a waveform estimate is produced as soon as the signal is observed. The filtering problem arises when we want to remove noise (as much as possible) from noise-corrupted signal observations as they are obtained.

Prediction. Here, we attempt to predict the signal's value at some future time ($l_e > 0$). The signal's structure must be well known to enable us to predict what values the signal obtains. Prediction filters have obvious applications in sonar/radar tracking and stock market analysis. Of all the waveform estimation problems, this one produces the largest errors.

Waveform estimation algorithms are *not* defined by this categorization; each technique can be applied to each type of problem (in most cases). Instead, the algorithms are defined according to the signal model. Correctness of the signal model governs the utility of a given technique. Because the signal usually appears *linearly* in the expression for the observations (the noise is usually additive), *linear* waveform estimation methods—filters—are frequently employed.

3.4.1 General Considerations

In the context of *linear* waveform estimation, the signal as well as the noise is considered to be a stochastic sequence. Furthermore, the signal component $\widetilde{S}$ in the observations is assumed to only be *related* to the signal s to be estimated and not necessarily equal to it: $X(l) = \widetilde{S}(l) + N(l)$. For example, the observations may contain a filtered version of the signal when we require an estimate of the prefiltered waveform. In this situation, the signal filter is usually known. The noise and signal components are zero-mean random sequences statistically independent of each other. The optimum filter that provides the waveform estimate $\widehat{S}(l)$ can be time invariant (Wiener filters), time varying (Kalman filters), or data dependent (adaptive filters). Choosing an estimation strategy is determined by the signal's characteristics and the degree to which these characteristics are known. For generality, we allow the optimum filter's unit-sample response $h_\diamond(l,k)$ to be time varying: It depends directly on the values of l, the "time variable" and k, the "time" at which the unit sample is presented. When the filter is time invariant, the unit-sample response would be a function of the interval $l-k$, time since presentation of the unit sample. The fundamental form of the observations and the estimated signal in all linear waveform estimators is

$$\boxed{\begin{array}{rcl} X(l) & = & \widetilde{S}(l) + N(l) \\ \widehat{S}(L_f + l_e) & = & \sum_{k=L_i}^{L_f} h_\diamond(L_f, k) X(k) \end{array}}$$

The estimate of the signal's value at $L_f + l_e$ is thus produced at time L_f in the filter's output. The duration of the filter's unit-sample response extends over the entire observation interval $[L_i, L_f]$.

The Orthogonality Principle that proved so useful in linear parameter estimation can be applied here. It states that the estimation error must be orthogonal to all linear transformations of the observations (see Eq. 3.1 {74}). For the waveform estimation problem, this requirement implies that

$$\mathscr{E}\left[\{S(L_f + l_e) - \widehat{S}(L_f + l_e)\} \sum_{k=L_i}^{L_f} h(L_f, k) X(k)\right] = 0 \quad \text{for all } h(\cdot,\cdot)$$

This expression implies that each observed value must be orthogonal to the estimation error at time $L_f + l_e$.

$$\mathscr{E}\left[\{S(L_f + l_e) - \sum_{j=L_i}^{L_f} h_\diamond(L_f, j) X(j)\} X(k)\right] = 0 \quad \text{for all } k \text{ in } [L_i, L_f]$$

Simplifying this expression, the fundamental equation that determines the unit-sample response of the linear minimum mean-squared error filter is

$$\boxed{K_{S\widetilde{S}}(L_f+l_e,k)=\sum_{j=L_i}^{L_f}K_X(j,k)h_\diamond(L_f,j)\quad\text{for all } k \text{ in } [L_i,L_f]}$$

where $K_X(k,l)$ is the covariance function of the observations, equaling $\mathcal{E}[X(k)X(l)]$, and $K_{S\widetilde{S}}(L_f+l_e,k)$ is the cross-covariance between the signal at L_f+l_e and the signal-related component of the observation at k. When the signal and noise are uncorrelated, $K_X(k,l)=K_{\widetilde{S}}(k,l)+K_N(k,l)$. Given these quantities, the preceding equation must then be solved for the unit-sample response of the optimum filter. This equation is known as the *generalized Wiener-Hopf equation*.

From the general theory of linear estimators, the mean-squared estimation error at index l equals the variance of the quantity being estimated minus the estimate's projection onto the signal.

$$\mathcal{E}[\varepsilon^2(l)]=K_S(l,l)-\mathcal{E}[\widehat{S}(l)S(l)]$$

Expressing the signal estimate as a linear filtering operation on the observations, this expression becomes

$$\mathcal{E}[\varepsilon^2(l)]=K_S(l,l)-\sum_{k=L_i}^{L_f}h_\diamond(L_f,k)K_{S\widetilde{S}}(l,k)$$

Further reduction of this expression is usually problem dependent, as succeeding sections illustrate.

3.4.2 Wiener Filters

Wiener filters are the solutions of the linear minimum mean-squared waveform estimation problem for the special case in which the noise and the signal are *stationary* random sequences [13: 100–18];[40: 481–515];[43]. The covariance functions appearing in the generalized Wiener-Hopf equation thus depend on the difference of their arguments. Considering the form of this equation, one would expect the unit-sample response of the optimum filter to depend on its arguments in a similar fashion. This presumption is in fact valid, and Wiener filters are always time invariant.

$$\widehat{S}(L_f+l_e)=\sum_{k=L_i}^{L_f}h_\diamond(L_f-k)X(k)$$

We consider first the case in which the initial observation time L_i equals $-\infty$. The resulting filter uses all of the observations available at any moment.* The errors that result from using this filter are smaller than those obtained when the filter is constrained to use a finite number of observations (such as some number of recent samples). The choice of $L_i=-\infty$ corresponds to an infinite-duration impulse response (*IIR*) Wiener filter; in a succeeding section, L_i is finite and a finite-duration impulse response (*FIR*) Wiener filter results. The error characteristics of the *IIR* Wiener filter generally bound those of *FIR* Wiener filters because more observations are used. We write the generalized Wiener-Hopf equation for the *IIR* case as

$$K_{S\widetilde{S}}(L_f+l_e-k)=\sum_{j=-\infty}^{L_f}K_X(j-k)\cdot h_\diamond(L_f-j)\quad\text{for all } k \text{ in } (-\infty,L_f]$$

Changing summation variables results in the somewhat simpler expression known as the Wiener-Hopf equation. It and the expression for the mean-squared estimation error are given by

$$\boxed{\begin{aligned}K_{S\widetilde{S}}(l+l_e)&=\sum_{k=0}^{\infty}K_X(l-k)h_\diamond(k)\quad\text{for all } l \text{ in } [0,\infty)\\ \mathcal{E}[\varepsilon^2]&=K_S(0)-\sum_{k=0}^{\infty}h_\diamond(k)K_{S\widetilde{S}}(l_e+k)\end{aligned}}\tag{3.7}$$

*Presumably, observations have been continuously available since the beginning of the universe.

The first term in the error expression is the signal variance. The mean-squared error of the signal estimate cannot be greater than this quantity; this error results when the estimate always equals 0.

In many circumstances, we want to estimate the signal directly contained in observations: $X = S + N$. This situation leads to a somewhat simpler form for the Wiener-Hopf equation.

$$K_S(l+l_e) = \sum_{k=0}^{\infty} \left[K_S(l-k) + K_N(l-k)\right] h_\diamond(k) \quad \text{for all } l \text{ in } [0,\infty)$$

It is this form we solve, but the previous one is required in its solution.

Solving the Wiener-Hopf equation. The Wiener-Hopf equation at first glance appears to be a convolution integral, implying that the optimum filter's frequency response could be easily found. The constraining condition—the equation applies only for the variable l in the interval $[0,\infty)$—means, however, that Fourier techniques *cannot* be used for the general case. If the Fourier Transform of the left side of the Wiener-Hopf equation were evaluated only over the constraining interval, the covariance function on the left would be *implicitly* assumed 0 outside the interval, which is usually not the case. Simply stated but mathematically complicated, the covariance function of the signal outside this interval is not to be considered in the solution of the equation.

One set of circumstances does allow Fourier techniques. Let the Wiener filter be noncausal with $L_f = +\infty$. In this case, the Wiener-Hopf equation becomes

$$K_S(l) = \sum_{k=-\infty}^{\infty} K_X(l-k) h_\diamond(k) \qquad \text{for all } l$$

As this equation must be valid for all values of l, a convolution sum emerges. The frequency response $H_\diamond(f)$ of the optimum filter is thus given by

$$H_\diamond(f) = \frac{\mathcal{S}_S(f)}{\mathcal{S}_S(f) + \mathcal{S}_N(f)}$$

where $\mathcal{S}_S(f)$ and $\mathcal{S}_N(f)$ are, respectively, the signal and the noise power spectra. Because this expression is real and even, the unit-sample response of the optimum filter is also real and even. The filter is therefore noncausal and usually has an infinite duration unit-sample response. This result is not often used in temporal signal processing but may find applications in spatial problems. Be that as it may, because this filter can use the entire set of observations to estimate the signal's value at any moment, it yields the smallest estimation error of *any* linear filter. Computing this error thus establishes a bound on how well any causal or *FIR* Wiener filter performs. The mean-squared estimation error of the noncausal Wiener filter can be expressed in the time domain or frequency domain.

$$\begin{aligned} \mathcal{E}[\varepsilon^2] &= K_S(0) - \sum_{l=-\infty}^{\infty} h_\diamond(l) K_S(l) \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\mathcal{S}_S(f)\mathcal{S}_N(f)}{\mathcal{S}_S(f) + \mathcal{S}_N(f)}\, df \end{aligned}$$

The causal solution to the Wiener-Hopf equation, the frequency response of the causal Wiener filter, is the product of two terms: the frequency response of a *whitening* filter and the frequency response of the signal estimation filter based on whitened observations [40: 482–93].

$$\boxed{H_\diamond(f) = \frac{1}{[\mathcal{S}_S + \mathcal{S}_N]^{+}(f)} \cdot \left[\frac{e^{+j2\pi f l_e}\mathcal{S}_S(f)}{[\mathcal{S}_S + \mathcal{S}_N]^{+*}(f)}\right]_{+}}$$

$[\mathscr{S}(f)]_+$ means the Fourier Transform of a covariance function's causal part, which corresponds to its values at nonnegative indices and $\mathscr{S}^+(f)$ the stable, causal, and minimum-phase square root of $\mathscr{S}(f)$. Evaluation of this expression therefore involves both forms of causal-part extraction. This solution is clearly much more complicated than anticipated when we first gave the Wiener-Hopf equation. How to solve it is best seen by example, which we provide once we determine an expression for the mean-squared estimation error.

Error characteristics of Wiener filter output. Assuming that $\widetilde{S}$ equals S, the expression for the mean-squared estimation error given in Eq. 3.7 {88}, can be simplified with the result

$$\mathscr{E}[\varepsilon^2] = K_S(0) - \sum_{k=0}^{\infty} h_\diamond(k) K_S(l_e+k) \tag{3.8}$$

Applying Parseval's Theorem to the summation, this expression can also be written in the frequency domain as

$$\mathscr{E}[\varepsilon^2] = K_S(0) - \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{[\mathscr{S}_S+\mathscr{S}_N]^+(f)} \left[\frac{e^{+j2\pi f l_e}\mathscr{S}_S(f)}{[\mathscr{S}_S+\mathscr{S}_N]^{+*}(f)}\right]_+ \left[e^{-j2\pi f l_e}\mathscr{S}_S(f)\right]_+ df$$

Noting that the first and third terms in the integral can be combined, the mean-squared error can also be written as

$$\mathscr{E}[\varepsilon^2] = K_S(0) - \int_{-\frac{1}{2}}^{\frac{1}{2}} \left|\left[\frac{e^{+j2\pi f l_e}\mathscr{S}_S(f)}{[\mathscr{S}_S+\mathscr{S}_N]^{+*}(f)}\right]_+\right|^2 df$$

The expression within the magnitude bars equals the frequency response of the second component of the Wiener filter's frequency response. Again using Parseval's Theorem to return to the time domain, the mean-squared error can be expressed directly either in terms of the filter's unit-sample response or in terms of signal and noise quantities by

$$\mathscr{E}[\varepsilon^2] = K_S(0) - \sum_{k=0}^{\infty} K_{SS_w}^2(l_e+k)$$

where the latter quantity is the cross-covariance function between the signal and the signal after passage through the whitening filter.

Example

Let's estimate the value of $S(L_f+l_e)$ with a Wiener filter using the observations obtained up to and including time L_f. The additive noise in the observations is white, having variance $8/7$. The power spectrum of the signal is given by

$$\begin{aligned}\mathscr{S}_S(f) &= \frac{1}{5/4-\cos 2\pi f}\\ &= \frac{1}{1-0.5e^{-j2\pi f}} \cdot \frac{1}{1-0.5e^{+j2\pi f}}\end{aligned}$$

The variance of the signal equals the value of the covariance function (found by the inverse Fourier Transform of this expression) at the origin. In this case, the variance equals $4/3$; the signal-to-noise ratio of the observations, taken to be the ratio of their variances, equals $7/6$.

The power spectrum of the observations is the sum of the signal and noise power spectra.

$$\begin{aligned}\mathscr{S}_S(f)+\mathscr{S}_N(f) &= \frac{1}{1-0.5e^{-j2\pi f}}\frac{1}{1-0.5e^{+j2\pi f}}+\frac{8}{7}\\ &= \frac{16}{7}\frac{\left(1-0.25e^{-j2\pi f}\right)\left(1-0.25e^{+j2\pi f}\right)}{\left(1-0.5e^{-j2\pi f}\right)\left(1-0.5e^{+j2\pi f}\right)}\end{aligned}$$

The noncausal Wiener filter has the frequency response

$$\frac{\mathcal{S}_S(f)}{\mathcal{S}_S(f)+\mathcal{S}_N(f)} = \frac{7}{16}\frac{1}{(1-0.25e^{-j2\pi f})(1-0.25e^{+j2\pi f})}$$

The unit-sample response corresponding to this frequency response and the covariance function of the signal are found to be

$$h_\diamond(l) = \frac{7}{15}\left(\frac{1}{4}\right)^{|l|} \quad \text{and} \quad K_S(l) = \frac{4}{3}\left(\frac{1}{2}\right)^{|l|}$$

Using Eq. 3.8 $\{90\}$, we find that the mean-squared estimation error for the noncausal estimator equals $4/3-4/5=8/15$.

The convolutionally causal part of signal-plus-noise power spectrum consists of the first terms in the numerator and denominator of the signal-plus-noise power spectrum.

$$[\mathcal{S}_S+\mathcal{S}_N]^+(f) = \frac{4}{\sqrt{7}}\frac{1-0.25e^{-j2\pi f}}{1-0.5e^{-j2\pi f}}$$

The second term in the expression for the frequency response of the optimum filter is given by

$$\begin{aligned}\frac{e^{+j2\pi f l_e}\mathcal{S}_S(f)}{[\mathcal{S}_S+\mathcal{S}_N]^{+*}(f)} &= \frac{\frac{e^{+j2\pi f l_e}}{(1-0.5e^{-j2\pi f})(1-0.5e^{+j2\pi f})}}{\frac{4}{\sqrt{7}}\frac{1-0.25e^{+j2\pi f}}{1-0.5e^{+j2\pi f}}} \\ &= \frac{\sqrt{7}}{4}\frac{e^{+j2\pi f l_e}}{(1-0.5e^{-j2\pi f})(1-0.25e^{+j2\pi f})}\end{aligned}$$

The additively causal part of this Fourier Transform is found by evaluating its partial fraction expansion.

$$\frac{\sqrt{7}}{4}\frac{e^{+j2\pi f l_e}}{(1-0.5e^{-j2\pi f})(1-0.25e^{+j2\pi f})} = \frac{e^{+j2\pi f l_e}}{2\sqrt{7}}\left[\frac{4}{1-0.5e^{-j2\pi f}} - \frac{2e^{+j2\pi f}}{1-0.25e^{+j2\pi f}}\right]$$

The simplest solution occurs when l_e equals zero: Estimate the signal value at the moment of the most recent observation. The first term on the right side of the preceding expression corresponds to the additively causal portion.

$$\left[\frac{\mathcal{S}_S(f)}{[\mathcal{S}_S+\mathcal{S}_N]^{+*}(f)}\right]_+ = \frac{2}{\sqrt{7}}\frac{1}{1-0.5e^{-j2\pi f}}$$

The frequency response of the Wiener filter is found to be

$$\begin{aligned}H_\diamond(f) &= \frac{\sqrt{7}}{4}\frac{1-0.5e^{-j2\pi f}}{1-0.25e^{-j2\pi f}}\cdot\frac{2}{\sqrt{7}}\frac{1}{1-0.5e^{-j2\pi f}} \\ &= \frac{1}{2}\frac{1}{1-0.25e^{-j2\pi f}}\end{aligned}$$

The Wiener filter has the form of a simple first-order filter with the pole arising from the whitening filter. The difference equation corresponding to this frequency response is

$$\widehat{S}(l) = \frac{1}{4}\widehat{S}(l-1)+\frac{1}{2}X(l)$$

The waveforms that result in this example are exemplified in Fig. 3.2.

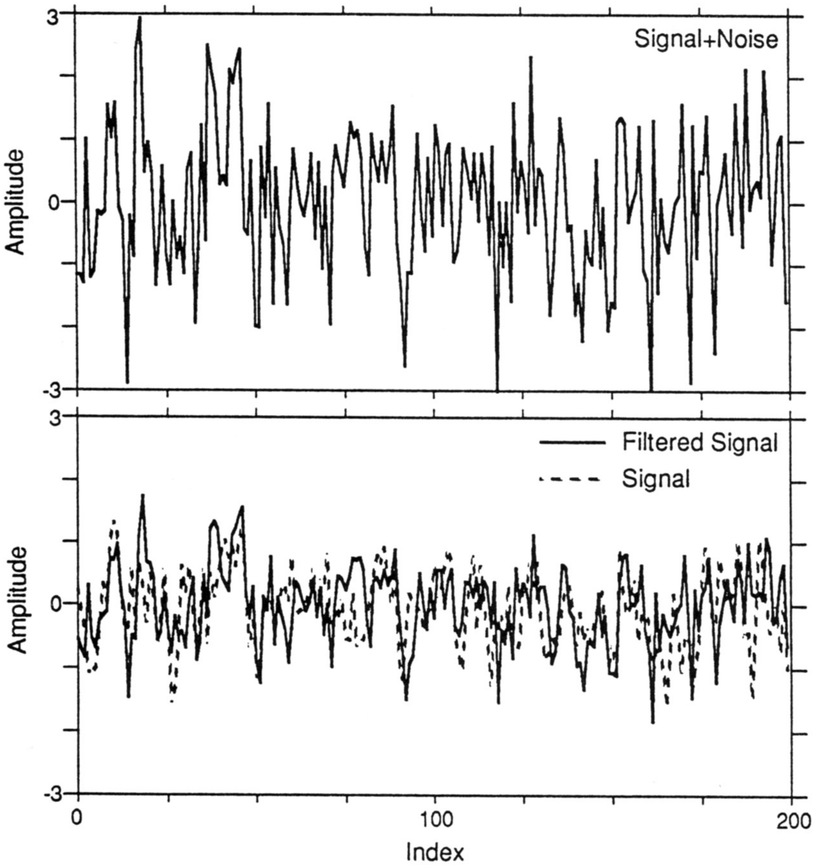


Figure 3.2: The upper panel displays observations having statistic characteristics corresponding to those given in the accompanying example. The output of the causal Wiener filter is shown in the bottom panel along with the actual signal, which is shown as a dashed line.

To find the mean-squared estimation error, the cross-covariance between the signal and its whitened counterpart is required. This quantity equals the inverse transform of the Wiener filter's second component and thus equals $(2/\sqrt{7})\,(1/2)^l$, $l \geq 0$. The mean-squared estimation error is numerically equal to

$$\begin{aligned}\mathscr{E}[\varepsilon^2] &= \frac{4}{3} - \sum_{l=0}^{\infty}\left[\frac{2}{\sqrt{7}}\left(\frac{1}{2}\right)^l\right]^2 \\ &= \frac{4}{7} = 0.57\end{aligned}$$

which compares with the smallest possible value of 0.53 provided by the noncausal Wiener filter. Thus, little is lost by using the causal filter. The signal-to-noise ratio of the estimated signal is equal to $K_S(0)/\mathscr{E}[\varepsilon^2]$. The causal filter yields a signal-to-noise ratio of 2.33, which should be compared with the ratio of 1.17 in the observations. The ratio of the signal-to-noise ratios at the output and input of a signal processing operation is usually referred to as the *processing gain*. The best possible processing gain is 2.14 and equals 2.0 in the causal case. These rather modest gains are because of the close similarity between the power spectra of the signal and the noise. As the parameter of this signal's power spectrum is increased, the two become less similar, and the processing gain increases.

Now consider the case in which $l_e > 0$: We want to predict the signal's future value. The whitening filter portion of the solution does not depend on the value of l_e and is therefore identical to that just given. The second component of the Wiener filter's frequency response does depend on l_e and is given for positive values of l_e by

$$\left[\frac{e^{+j2\pi f l_e}\mathscr{S}_S(f)}{[\mathscr{S}_S+\mathscr{S}_N]^{+*}(f)}\right]_+ = \left[\frac{\frac{2}{\sqrt{7}}e^{+j2\pi f l_e}}{1-0.5e^{-j2\pi f}}\right]_+$$

The causal portion of this frequency response is found by shifting the unit-sample response to the *left* and retaining the positive-time portion. Because this frequency response has only one pole, this manipulation is expressed simply as a scaling.

$$\left[\frac{e^{+j2\pi f l_e}\mathscr{S}_S(f)}{[\mathscr{S}_S+\mathscr{S}_N]^{+*}(f)}\right]_+ = \frac{2}{\sqrt{7}}\frac{\left(\frac{1}{2}\right)^{l_e}}{1-0.5e^{-j2\pi f}}$$

The frequency response of the prediction filter is thus given by

$$H_\diamond(f) = \frac{1}{2}\frac{2^{-l_e}}{1-0.25e^{-j2\pi f}}$$

The optimum linear predictor is a *scaled* version of the signal estimator. The mean-squared error increases as the desired time of the predicted value exceeds the time of the last observation. In particular, the signal-to-noise ratio of the predicted value is given by

$$\frac{K_S(0)}{\mathscr{E}[\varepsilon^2]} = \frac{1}{1-\frac{4}{7}\left(\frac{1}{2}\right)^{2l_e}}$$

The signal-to-noise ratio decreases rapidly as the prediction time extends into the future. This decrease is directly related to the reduced correlation between the signal and its future values in this example. This correlation is described by the absolute value of the signal's covariance function relative to its maximum value at the origin. As a covariance function broadens (corresponding to a lower frequency signal), the prediction error decreases. If a covariance function oscillates, the mean-squared prediction error varies in a similar fashion.

Finite-duration Wiener filters. Another useful formulation of Wiener filter theory is to constrain the filter's unit-sample response to have finite duration. To find this solution to the Wiener-Hopf equation, the values of L_f and L_i are chosen to be finite. Letting L represent the duration of the filter's unit-sample response ($L = L_f - L_i + 1$), the Wiener-Hopf equation becomes

$$K_{S\tilde{S}}(l+l_e) = \sum_{k=0}^{L-1} K_X(k-l)h_\diamond(k), \quad \text{for all } l \text{ in } [0, L-1]$$

This system of equations can be written in matrix form as $\mathbf{k}_{S\tilde{S}}(l_e) = \mathbf{K}_X\mathbf{h}_\diamond$.

$$\begin{bmatrix} K_{S\tilde{S}}(l_e) \\ K_{S\tilde{S}}(l_e+1) \\ \vdots \\ K_{S\tilde{S}}(l_e+L-1) \end{bmatrix} = \begin{bmatrix} K_X(0) & K_X(1) & \cdots & K_X(L-1) \\ K_X(-1) & K_X(0) & \cdots & K_X(L-2) \\ \vdots & K_X(-1) & \ddots & \vdots \\ K_X(-L+1) & \cdots & K_X(-1) & K_X(0) \end{bmatrix} \cdot \begin{bmatrix} h_\diamond(0) \\ h_\diamond(1) \\ \vdots \\ h_\diamond(L-1) \end{bmatrix}$$

When the signal component of the observations equals the signal being estimated ($\widetilde{S} = S$), the Wiener-Hopf equation becomes $\mathbf{k}_S(l_e) = \mathbf{K}_X\mathbf{h}_\diamond$. The $L\times L$ matrix $\mathbf{K}_X$ is the covariance matrix of the sequence of L observations. In the simple case of uncorrelated signal and noise components, this covariance matrix is the sum of those of the signal and the noise ($\mathbf{K}_X = \mathbf{K}_S + \mathbf{K}_N$). This matrix has an inverse in all but unusual circumstances with the result that the unit-sample response of the *FIR* Wiener filter is given by

$$\boxed{\mathbf{h}_\diamond = \mathbf{K}_X^{-1}\mathbf{k}_S(l_e)}$$

Because this covariance matrix is Toeplitz and Hermitian, its inverse can be efficiently computed using a variety of algorithms [23: 80–90]. The mean-squared error of the estimate is given by

$$\begin{aligned}\mathscr{E}[\varepsilon^2] &= K_S(0) - \sum_{k=0}^{L-1} h_\diamond(k)K_S(l_e+k)\\ &= K_S(0) - \mathbf{k}_S^t(l_e)\mathbf{K}_X^{-1}\mathbf{k}_S(l_e)\end{aligned}$$

Linear prediction. One especially important variation of the *FIR* Wiener filter occurs in the unique situation in which no observation noise is present, and the signal generation model contains only poles [21, 22]. Thus, the signal $S(l)$ is generated by passing white noise $w(l)$ through a linear system given by the difference equation

$$S(l) = a_1S(l-1) + a_2S(l-2) + \cdots + a_pS(l-p) + w(l)$$

The coefficients $a_1,\ldots,a_p$ are unknown. This signal modeling approach is frequently used to estimate the signal's spectrum.

As no noise is present in the observations, the filtered estimate of the signal ($l_e = 0$) equals $S(l)$ and the estimation error is exactly 0. The concern of linear prediction is not this trivial problem, but the so-called *one-step prediction problem* ($l_e = 1$): Predict the value of the signal at index l given values of $S(l-1)$,$S(l-2),\ldots$. Thus, we seek a *FIR* Wiener filter predictor, which has the form

$$\widehat{S}(l) = h(0)S(l-1) + h(1)S(l-2) + \cdots + h(p-1)S(l-p)$$

Comparing the signal model equation to that for the Wiener filter predictor, we see that the model parameters $\{a_1,\ldots,a_p\}$ equal the Wiener filter's unit-sample response $h(\cdot)$ because the input $w(l)$ is uncorrelated from sample to sample. In linear prediction, the signal model parameters are used notationally to express the filter coefficients.

The Orthogonality Principle can be used to find the minimum mean-squared error predictor of the next signal value. By requiring orthogonality of the prediction error to each of the observations used in the estimate, the following set of equations results.

$$\begin{aligned}a_1K_S(0) + a_2K_S(1) + \cdots + a_pK_S(p-1) &= K_S(1)\\ a_1K_S(1) + a_2K_S(0) + \cdots + a_pK_S(p-2) &= K_S(2)\\ \vdots\qquad\qquad &\qquad \vdots\\ a_1K_S(p-1) + a_2K_S(p-2) + \cdots + a_pK_S(0) &= K_S(p)\end{aligned}$$

In linear prediction, these are known as the *Yule-Walker equations*. Expressing them concisely in matrix form $\mathbf{K}_S\mathbf{a} = \mathbf{k}_S$, the solution is $\mathbf{a} = \mathbf{K}_S^{-1}\mathbf{k}_S$.

From the signal model equation, we see that the mean-squared prediction error $\mathscr{E}[\{S(l) - \widehat{S}(l)\}^2]$ equals the variance σ_w^2 of the white-noise input to the model. Computing the mean-squared estimation error according to Eq. 3.7 {88}, this variance is expressed by

$$\sigma_w^2 = K_S(0) - a_1K_S(1) - \cdots - a_pK_S(p)$$

This result can be combined with the previous set of equations to yield a unified set of equations for the unknown parameters and the mean-squared error of the optimal linear predictive filter.

$$\boxed{\begin{bmatrix} K_S(0) & K_S(1) & \cdots & K_S(p) \\ K_S(1) & K_S(0) & \cdots & K_S(p-1) \\ \vdots & K_S(1) & \ddots & \vdots \\ K_S(p) & \cdots & K_S(1) & K_S(0) \end{bmatrix} \cdot \begin{bmatrix} 1 \\ -a_1 \\ \vdots \\ -a_p \end{bmatrix} = \begin{bmatrix} \sigma_w^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}} \tag{3.9}$$

To solve this set of equations for the model coefficients and the input-noise variance conceptually, we compute the preliminary result $\mathbf{K}_S^{-1}\delta$. The first element of this vector equals the reciprocal of σ_w^2; normalizing $\mathbf{K}_S^{-1}\delta$ so that its leading term is unity yields the coefficient vector $\mathbf{a}$. Levinson's algorithm can be used to solve these equations efficiently and simultaneously obtain the noise variance [23: 211–16].

3.5 Probability Density Estimation

Many signal processing algorithms, implicitly or explicitly, assume that the signal and the observation noise are each well described as Gaussian random sequences. Virtually all linear estimation and prediction filters minimize the mean-squared error while not explicitly assuming any form for the amplitude distribution of the signal or noise. In many formal waveform estimation theories where probability density is, for better or worse, specified, the mean-squared error arises from Gaussian assumptions. A similar situation occurs explicitly in detection theory. The matched filter is provably the optimum detection rule *only* when the observation noise is Gaussian. When the noise is non-Gaussian, the detector assumes some other form. Much of what has been presented in this chapter is based *implicitly* on a Gaussian model for both the signal and the noise. When non-Gaussian distributions are assumed, the quantities upon which optimal linear filtering theory are based, covariance functions, no longer suffice to characterize the observations. While the joint amplitude distribution of any zero-mean, stationary Gaussian stochastic process is entirely characterized by its covariance function; non-Gaussian processes require more. Optimal linear filtering results can be applied in non-Gaussian problems, but we should realize that other informative aspects of the process are being ignored.

This discussion would seem to be leading to a formulation of optimal filtering in a non-Gaussian setting. Would that such theories were easy to use; virtually all of them require knowledge of process characteristics that are difficult to measure and the resulting filters are typically nonlinear [20: chapter 8]. Rather than present preliminary results, we take the tack that knowledge is better than ignorance: At least the first-order amplitude distribution of the observed signals should be considered during the signal processing design. If the signal is found to be Gaussian, then linear filtering results can be applied with the knowledge than no other filtering strategy will yield better results. If non-Gaussian, the linear filtering can still be used and the engineer must be aware that future systems might yield "better" results.*

3.5.1 Types

When the observations are discrete-valued or made so by digital-to-analog converters, estimating the probability mass function is straightforward: Count the relative number of times each value occurs. Let $X(0),\dots,X(L-1)$ denote a sequence of observations, each of which takes on values from the set $\mathscr{A} = \{a_1,\dots,a_N\}$. This set is known as an *alphabet* and each a_n is a letter in that alphabet. We estimate the probability that an observation equals one of the letters according to

$$\widehat{P}_X(a_n) = \frac{1}{L}\sum_{l=0}^{L-1} I\bigl(X(l) = a_n\bigr) \ ,$$

*Note that linear filtering optimizes the mean-squared error whether the signals involved are Gaussian or not. Other error criteria might better capture unexpected changes in signal characteristics and non-Gaussian processes contain internal statistical structure beyond that described by the covariance function.

where $I(\cdot)$ is the indicator function, equaling one if its argument is true and zero otherwise. This kind of estimate is known in information theory as a *type* [5: Chap. 12], and types have remarkable properties. For example, if the observations are statistically independent, the probability that a given sequence occurs equals

$$\Pr[\mathbf{X}=\{X(0),\ldots,X(L-1)\}]=\prod_{l=0}^{L-1} P_X(X(l)) \ .$$

Evaluating the logarithm, we find that

$$\log \Pr[\mathbf{X}] = \sum_{l=0}^{L-1} \log P_X(X(l))$$

Note that the number of times each letter occurs equals $L\widehat{P}_X(a_n)$. Using this fact, we can convert this sum to a sum over letters.

$$\begin{aligned}
\log \Pr[\mathbf{X}] &= \sum_{n=0}^{N-1} L\widehat{P}_X(a_n) \log P_X(a_n) \\
&=L \sum_{n=0}^{N-1} \widehat{P}_X(a_n) \left[\log \widehat{P}_X(a_n) - \log \frac{\widehat{P}_X(a_n)}{P_X(a_n)}\right] \\
\log \Pr[\mathbf{X}] &= -L\left[\mathscr{H}(\widehat{P}_X) + \mathscr{D}(\widehat{P}_X \| P_X)\right]
\end{aligned}$$

which yields

$$\Pr[\mathbf{X}] = e^{-L\left[\mathscr{H}(\widehat{P}_X) + \mathscr{D}(\widehat{P}_X \| P_X)\right]} \tag{3.10}$$

We introduce the *entropy* [5: §2.1] and *Kullback-Leibler distance*.

$$\begin{aligned}
\mathscr{H}(P) &= -\sum_{n=0}^{N-1} P(a_n) \log P(a_n) \\
\mathscr{D}(P_1 \| P_0) &= \sum_{n=0}^{N-1} P_1(a_n) \log \frac{P_1(a_n)}{P_0(a_n)}
\end{aligned}$$

Because the Kullback-Leibler distance is non-negative, equaling zero *only* when the two probability distributions equal each other, we maximize Eq. (3.10) with respect to P by choosing $P=\widehat{P}$: The type estimator is the maximum likelihood estimator of P_X.

The number of length-L observation sequences having a given type $\widehat{P}$ approximately equals $e^{-L\mathscr{H}(\widehat{P})}$. The probability that a given sequence has a given type approximately equals $e^{-L\mathscr{D}(\widehat{P}\|P)}$, which means that the probability a given sequence has a type *not* equal to the true distribution decays exponentially with the number of observations. Thus, while the coin flip sequences {H,H,H,H,H} and {T,T,H,H,T} are equally likely (assuming a fair coin), the second is more *typical* because its type is closer to the true distribution.

3.5.2 Histogram Estimators

By far the most used technique for estimating the probability distribution of a continuous-valued random variable is the *histogram*; more sophisticated techniques are discussed in [36]. For real-valued data, subdivide the real line into N intervals $(X_i, X_{i+1}]$ having widths $\delta_i = X_{i+1} - X_i$, $i=1,\ldots,N$. These regions are called "bins" and they should encompass the range of values assumed by the data. For large values, the "edge bins" can extend to infinity to catch the overflows. Given L observations of a stationary random sequence $X(l)$,

$l = 0, \ldots, L-1$, the histogram estimate $h(i)$ is formed by simply forming a type from the number L_i of these observations that fall into the i^{th} bin and dividing by the binwidth δ_i.

$$\hat{p}_X(X) = \begin{cases} h(1) = \frac{L_1}{L\delta_1} & X_1 < X \le X_2 \\ h(2) = \frac{L_2}{L\delta_2} & X_2 < X \le X_3 \\ \vdots & \\ h(N) = \frac{L_N}{L\delta_N} & X_N < X \le X_{N+1} \end{cases}$$

The histogram estimate resembles a rectangular approximation to the density. Unless the underlying density has the same form (a rare event), the histogram estimate does *not* converge to the true density as the number L of observations grows. Presumably, the value of the histogram at each bin converges to the probability that the observations lie in that bin.

$$\lim_{L\to\infty} \frac{L_i}{L} = \int_{X_i}^{X_{i+1}} p_X(X)\,dX$$

To demonstrate this intuitive feeling, we compactly denote the histogram estimate by using *indicator functions*. An indicator function $I_i[X(l)]$ for the i^{th} bin equals one if the observation $X(l)$ lies in the bin and is zero otherwise. The estimate is simply the average of the indicator functions across the observations.

$$h(i) = \frac{1}{L\delta_i} \sum_{l=0}^{L-1} I_i[X(l)]$$

The expected value of $I_i[X(l)]$ is simply the probability P_i that the observation lies in the i^{th} bin. Thus, the expected value of each histogram value equals the integral of the actual density over the bin, showing that the histogram is an unbiased estimate of this integral. Convergence can be tested by computing the variance of the estimate. The variance of one bin in the histogram is given by

$$\mathscr{V}[h(i)] = \frac{P_i - P_i^2}{L\delta_i^2} + \frac{1}{L^2\delta_i^2} \sum_{k \neq l} \left(\mathscr{E}\{I_i[X(k)]I_i[X(l)]\} - P_i^2 \right)$$

To simplify this expression, the correlation between the observations must be specified. If the values are statistically independent (we have white noise), each term in the sum becomes zero and the variance is given by $\mathscr{V}[h(i)] = (P_i - P_i^2)/(L\delta_i^2)$. Thus, the variance tends to zero as $L \to \infty$ and the histogram estimate is consistent, converging to P_i/δ_i. If the observations are not white, convergence becomes problematical. Assume, for example, that $I_i[X(k)]$ and $I_i[X(l)]$ are correlated in a first-order, geometric fashion.

$$\mathscr{E}\{I_i[X(k)]I_i[X(l)]\} - P_i^2 = P_i^2 \rho^{|k-l|}$$

The variance does increase with this presumed correlation until, at the extreme ($\rho = 1$), the variance is a constant independent of L! In summary, if the observations are mutually correlated and the histogram estimate converges, the estimate converges to the proper value but more slowly than if the observations were white. The estimate may not converge if the observations are heavily dependent from index to index. This type of dependence structure occurs when the power spectrum of the observations is lowpass with an extremely low cutoff frequency.

Convergence to the density rather than its integral over a region can occur if, as the amount of data grows, we reduce the binwidth δ_i and increase N, the number of bins. However, if we choose the binwidth too small for the amount of available data, few bins contain data and the estimate is inaccurate. Letting X' denote the midpoint of a bin, using a Taylor expansion about this point reveals that the mean-squared error between the histogram and the density at that point is [38: 44–59]

$$\mathscr{E}\{[p_X(X') - h(i)]^2\} = \frac{p_X(X')}{2L\delta_i} + \frac{\delta_i^4}{36} \left[\left. \frac{d^2 p_X(X)}{dX^2} \right|_{X=X'} \right]^2 + O\left(\frac{1}{L}\right) + O(\delta_i^5)$$

This mean-squared error becomes zero *only* if $L \to \infty$, $L\delta_i \to \infty$, *and* $\delta_i \to 0$. Thus, the binwidth must decrease *more slowly* than the rate of increase of the number of observations. We find the "optimum" compromise between the decreasing binwidth and the increasing amount of data to be*

$$\delta_i = \left[\frac{9 p_X(X')}{2\left[d^2 p_X(X)/dX^2 |_{X=X'} \right]^2} \right]^{1/5} L^{-1/5}$$

Using this binwidth, we find the the mean-squared error to be proportional to $L^{-4/5}$. We have thus discovered the famous "4/5" rule of density estimation; this is one of the few cases where the variance of a convergent statistic decreases more slowly than the reciprocal of the number of observations. In practice, this optimal binwidth cannot be used because the proportionality constant depends of the unknown density being estimated. Roughly speaking, wider bins should be employed where the density is changing slowly. How the optimal binwidth varies with L can be used to adjust the histogram estimate as more data become available.

3.5.3 Density Verification

Once a density estimate is produced, the class of density that best coincides with the estimate remains an issue: Is the density just estimated statistically similar to a Gaussian? The histogram estimate can be used directly in a hypothesis test to determine similarity with any proposed density. Assume that the observations are obtained from a white, stationary, stochastic sequence. Let $\mathcal{M}_0$ denote the hypothesis that the data have an amplitude distribution equal to the presumed density and $\mathcal{M}_1$ the dissimilarity hypothesis. If $\mathcal{M}_0$ is true, the estimate for each bin should not deviate greatly from the probability of a randomly chosen datum lying in the bin. We determine this probability from the presumed density by integrating over the bin. Summing these deviations over the entire estimate, the result should not exceed a threshold. The theory of standard hypothesis testing requires us to produce a specific density for the alternative hypothesis $\mathcal{M}_1$. We cannot rationally assign such a density; consistency is being tested, not whether either of two densities provides the best fit. However, taking inspiration from the Neyman-Pearson approach to hypothesis testing (§4.1.2 {113}), we can develop a test statistic and require its statistical characteristics *only* under $\mathcal{M}_0$. The typically used, but *ad hoc* test statistic $S(L,N)$ is related to the histogram estimate's mean-squared error [6: 416–41].

$$S(L,N) = \sum_{i=1}^{N} \frac{(L_i - LP_i)^2}{LP_i} = \sum_{i=1}^{N} \frac{L_i^2}{LP_i} - L$$

This statistic sums over the various bins the squared error of the number of observations relative to the expected number. For large L, $S(L,N)$ has a χ^2 probability distribution with $N-1$ degrees of freedom [6: 417]. Thus, for a given number of observations L we establish a threshold η_N by picking a false-alarm probability P_F and using tables to solve $\Pr[\chi^2_{N-1} > \eta_N] = P_F$. To enhance the validity of this approximation, statisticians recommend selecting the binwidth so that each bin contains at least ten observations. In practice, we fulfill this criterion by merging adjacent bins until a sufficient number of observations occur in the new bin and defining its binwidth as the sum of the merged bins' widths. Thus, the number of bins is reduced to some number N', which determines the degrees of freedom in the hypothesis test. The similarity test between the histogram estimate of a probability density function and an assumed ideal form becomes

$$\boxed{S(L,N') \underset{\mathcal{M}_0}{\overset{\mathcal{M}_1}{\gtrless}} \eta_{N'}}$$

In many circumstances, the formula for the density is known but not some of its parameters. In the Gaussian case, for example, the mean or variance are usually unknown. These parameters must be determined

*This result assumes that the second derivative of the density is nonzero. If it is not, either the Taylor series expansion brings higher order terms into play or, if all the derivatives are zero, no optimum binwidth can be defined for minimizing the mean-squared error.

from the same data used in the consistency test before the test can be used. Doesn't the fact that we use estimates rather than actual values affect the similarity test? The answer is "yes," but in an interesting way: The similarity test changes only in that the number of degrees of freedom of the χ^2 random variable used to establish the threshold is reduced by one for each estimated parameter. If a Gaussian density is being tested, for example, the mean and variance usually need to be found. The threshold should then be determined according to the distribution of a $\chi^2_{N'-3}$ random variable.

Example

Three sets of observations are considered: Two are drawn from a Gaussian distribution and the other not. The first Gaussian example is white noise, a signal whose characteristics match the assumptions of this section. The second is non-Gaussian, which should not pass the test. Finally, the last test consists of colored Gaussian noise that, because of dependent samples, does not have as many degrees of freedom as would be expected. The number of data available in each case is 2000. The histogram estimator uses fixed-width bins and the χ^2 test demands at least ten observations per merged bin. The mean and variance estimates are used in constructing the nominal Gaussian density. The histogram estimates and their approximation by the nominal density whose mean and variance were computed from the data are shown in Fig. 3.3. The chi-squared test ($P_F = 0.1$) yielded the following results.

Density	N'	$\chi^2_{N'-3}$	$S(2000, N')$
White Gaussian	70	82.2	78.4
White sech	65	76.6	232.6
Colored Gaussian	65	76.6	77.8

The white Gaussian noise example clearly passes the χ^2 test. The test correctly evaluated the non-Gaussian example, but declared the colored Gaussian data to be non-Gaussian, yielding a value near the threshold. Failing in the latter case to correctly determine the data's Gaussianity, we see that the χ^2 test is sensitive to the statistical independence of the observations.

Problems

3.1 Estimates of identical parameters are heavily dependent on the assumed underlying probability densities. To understand this sensitivity better, consider the following variety of problems, each of which asks for estimates of quantities related to variance. Determine the bias and consistency in each case.

(a) Compute the maximum *a posteriori* and maximum likelihood estimates of θ based on L statistically independent observations of a Maxwellian random variable X.

$$p_{X|\theta}(X|\theta) = \sqrt{\frac{2}{\pi}}\,\theta^{-3/2} X^2 e^{-\frac{1}{2}X^2/\theta} \qquad X > 0, \theta > 0$$

$$p_\theta(\theta) = \lambda e^{-\lambda\theta}, \qquad \theta > 0$$

(b) Find the maximum *a posteriori* estimate of the variance σ^2 from L statistically independent observations having the exponential density

$$p_X(X) = \frac{1}{\sqrt{\sigma^2}} e^{-X/\sqrt{\sigma^2}} \qquad X > 0$$

where the variance is uniformly distributed over the interval $[0, \sigma^2_{\max})$.

(c) Find the maximum likelihood estimate of the variance of L identically distributed, but dependent Gaussian random variables. Here, the covariance matrix is written $\mathbf{K}_X = \sigma^2 \widetilde{\mathbf{K}}_X$, where the normalized covariance matrix has trace $\mathrm{tr}[\widetilde{\mathbf{K}}_X] = L$. Assume the random variables have zero mean.

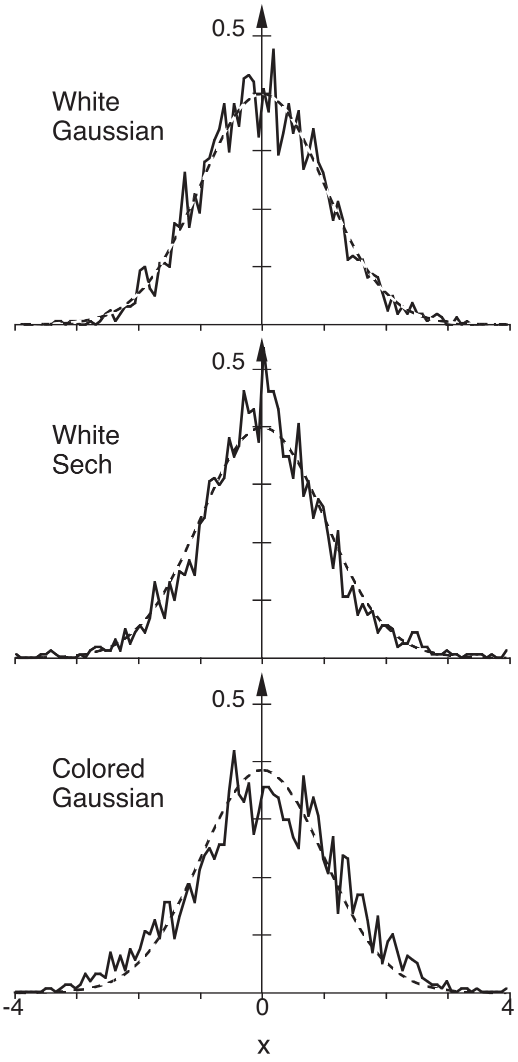


Figure 3.3: Three histogram density estimates are shown and compared with Gaussian densities having the same mean and variance. The histogram on the top is obtained from Gaussian data that are presumed to be white. The middle one is obtained from a non-Gaussian distribution related to the hyperbolic secant $[p_X(X) = \frac{1}{2\sigma}\text{sech}^2(\pi X/2\sigma)]$. This density resembles a Gaussian about the origin but decreases exponentially in the tails. The bottom histogram is taken from a first-order autoregressive Gaussian signal. Thus, these data are correlated, but yield a histogram resembling the true amplitude distribution. In each case, 2000 data points were used and the histogram contained 100 bins.

3.2 Imagine yourself idly standing on the corner in a large city when you note the serial number of a passing beer truck. Because you are idle, you wish to estimate (guess may be more accurate here) how many beer trucks the city has from this single observation.

(a) Making appropriate assumptions, the beer truck's number is drawn from a uniform probability density ranging between zero and some unknown upper limit, find the maximum likelihood estimate of the upper limit.

(b) Show that this estimate is biased.

(c) In one of your extraordinarily idle moments, you observe throughout the city L beer trucks. Assuming them to be independent observations, now what is the maximum likelihood estimate of the total?

(d) Is this estimate of θ biased? asymptotically biased? consistent?

3.3 Estimating a Bit
To send a bit, a discrete-time communications system transmits either $+1$ or -1 for L successive indices. The channel adds white Gaussian noise and the receiver must determine which bit was sent from these noise-corrupted observations. The bit's values are equally likely.

(a) What is the MAP estimate of the bit's value?

(b) Determine the bias, if any, of the MAP estimate?

(c) Is the MAP estimate in this case consistent?

(d) Find the minimum mean-squared error estimator of the bit.

3.4 We make L observations $X_1,\dots,X_L$ of a parameter θ corrupted by additive noise ($X_l = \theta + N_l$). The parameter θ is a Gaussian random variable $[\theta \sim \mathcal{N}(0,\sigma_\theta^2)]$ and N_l are statistically independent Gaussian random variables $[N_l \sim \mathcal{N}(0,\sigma_N^2)]$.

(a) Find the *MMSE* estimate of θ.

(b) Find the maximum *a posteriori* estimate of θ.

(c) Compute the resulting mean-squared error for each estimate.

(d) Consider an alternate procedure based on the same observations X_l. Using the *MMSE* criterion, we estimate θ immediately after each observation. This procedure yields the sequence of estimates $\widehat{\theta}_1(X_1), \widehat{\theta}_2(X_1,X_2),\dots,\widehat{\theta}_L(X_1,\dots,X_L)$. Express $\widehat{\theta}_l$ as a function of $\widehat{\theta}_{l-1}$, σ_{l-1}^2, and X_l. Here, σ_l^2 denotes the variance of the estimation error of the l^{th} estimate. Show that

$$\frac{1}{\sigma_l^2} = \frac{1}{\sigma_\theta^2} + \frac{l}{\sigma_N^2}$$

3.5 Estimating Phase, Amplitude and Frequency
You are given the discrete-time signal $A\cos(2\pi f_0 l - \theta)$ observed in the presence of white Gaussian noise having zero-mean and variance σ^2. Eventually, we want to find all the parameters, but let's build up to that by estimating first the phase θ, then the amplitude A, finally incorporating the frequency f_0. Throughout assume the number of observations L contains an integer number of periods of the sinusoidal signal.

(a) What is the maximum likelihood estimate for the phase assuming the amplitude and frequency are known?

(b) Find the Cramér-Rao bound for your estimate.

(c) Create a MATLAB simulation of your estimation procedure. Let $A = 1$ and $f_0 = 100/L$, with $L = 1024$. Run $1{,}000$-trial simulations to estimate the phase for $\theta = \pi/4$ for signal-to-noise ratios A^2/σ^2 of 1.0 and 0.04. Calculate the empirical mean and standard deviation of your estimates. Do they agree with theory?

(d) What are the maximum likelihood estimates of the phase and the amplitude? Find the Cramér-Rao bound for this case.

(e) Using the same simulations as in part (c), do these estimates have the predicted statistics?

(f) Find the joint maximum likelihood estimates for all three parameters. Calculate the Cramér-Rao lower bound for each parameter's estimation error.

Note: The analytic maximum likelihood estimate of the frequency is difficult to apply to data. Find an equivalent empirical estimate.

(g) For the data file `sineinnoise.mat`, estimate the amplitude, frequency and phase. Indicate the possible range of values for your estimates (*i.e.*, provide error bars).

3.6 Although the maximum likelihood estimation procedure was not clearly defined until early in the 20th century, Gauss showed in 1805 that the Gaussian density* was the *sole* density for which the maximum likelihood estimate of the mean equaled the sample average. Let $\{X_0, \dots, X_{L-1}\}$ be a sequence of statistically independent, identically distributed random variables.

- **(a)** What equation defines the maximum likelihood estimate $\widehat{m}_{\mathrm{ML}}$ of the mean m when the common probability density function of the data has the form $p(X-m)$?
- **(b)** The sample average is, of course, $\sum_l X_l/L$. Show that it minimizes the mean-squared error $\sum_l [X_l - m]^2$.
- **(c)** Equating the sample average to $\widehat{m}_{\mathrm{ML}}$, combine this equation with the maximum likelihood equation to show that the Gaussian density uniquely satisfies the equations.

Note: Because both equations equal 0, they can be equated. Use the fact that they must hold for *all L* to derive the result. Gauss thus showed that mean-squared error and the Gaussian density were closely linked, presaging ideas from modern robust estimation theory.

3.7 What's In-Between the Samples?
We sample a stationary random process X_t every T seconds, ignoring whether the process is bandlimited or not. To reconstruct the signal from the samples, we use *linear interpolation*.

$$\widehat{X}_t = aX_{(n-1)T} + bX_{nT}, \quad (n-1)T \le t \le nT$$

- **(a)** Find the minimum mean-squared error linear interpolator. In other words, what are the best values for a and b?
- **(b)** Show that the maximum likelihood interpolator is also linear when X_t is a wide-sense stationary, zero-mean, Gaussian process. In other words, if $X_{(n-1)T}$ and X_{nT} comprise your observations, show that the maximum likelihood estimate of X_t has the linear form given above. For this part, you do not need to find a and b.
- **(c)** Find the Cramér-Rao bound for the interpolation estimate of X_t.

3.8 In an example {77}, we derived the maximum likelihood estimate of the mean and variance of a Gaussian random vector. You might wonder why we chose to estimate the variance σ^2 rather than the standard deviation σ. Using the same assumptions provided in the example, let's explore the consequences of estimating a *function* of a parameter [40: Probs. 2.4.9, 2.4.10].

- **(a)** Assuming that the mean is known, find the maximum likelihood estimates of first the variance, then the standard deviation.
- **(b)** Are these estimates biased?
- **(c)** Describe how these two estimates are related. Assuming that $f(\cdot)$ is a monotonic function, how are $\widehat{\theta}_{\mathrm{ML}}$ and $\widehat{f(\theta)}_{\mathrm{ML}}$ related in general?

  These results suggest a general question. Consider the problem of estimating some function of a parameter θ, say $f_1(\theta)$. The observed quantity is X and the conditional density $p_{X|\theta}(X|\theta)$ is known. Assume that θ is a nonrandom parameter.
- **(d)** What are the conditions for an efficient estimate $\widehat{f_1(\theta)}$ to exist?
- **(e)** What is the lower bound on the variance of the error of any unbiased estimate of $f_1(\theta)$?
- **(f)** Assume an efficient estimate of $f_1(\theta)$ exists; when can an efficient estimate of some other function $f_2(\theta)$ exist?

3.9 Let the observations $X(l)$ consist of statistically independent, identically distributed Gaussian random variables having zero mean but unknown variance. We wish to estimate σ^2, their variance.

*It wasn't called the Gaussian density in 1805; this result is one of the reasons why it is.

(a) Find the maximum likelihood estimate $\widehat{\sigma^2}_{\text{ML}}$ and compute the resulting mean-squared error.

(b) Show that this estimate is efficient.

(c) Consider a new estimate $\widehat{\sigma^2}_{\text{NEW}}$ given by $\widehat{\sigma^2}_{\text{NEW}} = \alpha\widehat{\sigma^2}_{\text{ML}}$, where α is a constant. Find the value of α that minimizes the mean-squared error for $\widehat{\sigma^2}_{\text{NEW}}$. Show that the mean-squared error of $\widehat{\sigma^2}_{\text{NEW}}$ is less than that of $\widehat{\sigma^2}_{\text{ML}}$. Is this result compatible with part b?

3.10 Optimal and Simple Communications

A multiplexed communication system needs to be designed that sends two numbers simultaneously. Perhaps the simplest design represents the numbers as the amplitudes of two carrier signals. The received signal has the form

$$R_l = A_1 c_1(l) + A_2 c_2(l) + N_l, \quad l = 0, \dots, L-1$$

where N_l is ubiquitous additive (not necessarily white) Gaussian noise. The carrier signals $c_1(l)$ and $c_2(l)$ have unit energy; their detailed waveforms need to be selected to provide the best possible system design.

(a) What is the maximum likelihood estimate of the amplitudes?

(b) Is the maximum likelihood estimate biased or not? If it is biased, what are the most general conditions on the carrier signals and the noise would it make it unbiased?

(c) Under what conditions are the amplitude estimation errors uncorrelated and as small as possible?

3.11 MIMO Channels

Two parameters θ_1, θ_2 are transmitted over a MIMO (Multiple-Input, Multiple-Output) channel. The two parameters constitute the channel's two-dimensional vector input θ, and the channel output is $\mathbf{H}\theta$. $\mathbf{H}$ is the *non-square* "transfer function" matrix that represents the set of linear combinations of the parameters found in the output. The observations consist of

$$\mathbf{R} = \mathbf{H}\theta + \mathbf{N},$$

where the noise vector $\mathbf{N}$ is Gaussian, having zero mean and covariance matrix $\mathbf{K}$.

(a) What is the maximum likelihood estimate of θ?

(b) Find this estimate's total mean-squared error.

(c) Is this estimate biased? Is it efficient?

3.12 Prediction

A signal $s(l)$ can be described as a stochastic process that has zero mean and covariance function $K_s(\ell) = \sigma_s^2 a^{|\ell|}$. This signal is observed in additive white Gaussian noise having variance σ^2. The signal and noise are statistically independent of each other.

(a) Find the optimal predictor $\widehat{s}(l+1)$ that is based on observations that end at time l and begin at time $l-L+1$.

(b) How does this predictor change if we want to estimate $s(l+k)$ based on observations made over $[l, \dots, l+L-1]$?

(c) How does the predictor's mean-squared error vary with k?

3.13 Let the observations be of the form $\mathbf{X} = \mathbf{H}\theta + \mathbf{n}$ where θ and $\mathbf{n}$ are statistically independent Gaussian random vectors.

$$\theta \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_\theta) \qquad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_n)$$

The vector θ has dimension M; the vectors $\mathbf{X}$ and $\mathbf{n}$ have dimension N.

(a) Derive the minimum mean-squared error estimate of θ, $\widehat{\theta}_{\text{MMSE}}$, from the relationship $\widehat{\theta}_{\text{MMSE}} = \mathscr{E}[\theta|\mathbf{X}]$.

(b) Show that this estimate and the optimum linear estimate $\widehat{\theta}_{\text{LIN}}$ derived by the Orthogonality Principle are equal.

(c) Find an expression for the mean-squared error when these estimates are used.

3.14 Suppose we consider an estimate of the parameter θ having the form $\widehat{\theta} = \mathscr{L}(\mathbf{X}) + C$, where $\mathbf{X}$ denotes the vector of the observables and $\mathscr{L}(\cdot)$ is a linear operator. The quantity C is a constant. This estimate is *not* a linear function of the observables unless $C = 0$. We are interested in finding applications for which it is advantageous to allow $C \neq 0$. Estimates of this form we term "quasi-linear."

(a) Show that the optimum (minimum mean-squared error) quasi-linear estimate satisfies

$$\mathscr{E}[\langle \mathscr{L}_\diamond(\mathbf{X}) + C_\diamond - \theta, \mathscr{L}(\mathbf{X}) + C \rangle] = 0, \quad \text{for all } \mathscr{L}(\cdot) \text{ and } C$$

where $\widehat{\theta}_{\text{QLIN}} = \mathscr{L}_\diamond(\mathbf{X}) + C_\diamond$.

(b) Find a general expression for the mean-squared error incurred by the optimum quasi-linear estimate.

(c) Such estimates yield a smaller mean-squared error when the parameter θ has a nonzero mean. Let θ be a scalar parameter with mean m. The observables comprise a vector $\mathbf{X}$ having components given by $X_l = \theta + N_l, l = 1, \ldots, L$ where N_l are statistically independent Gaussian random variables $[N_l \sim \mathscr{N}(0, \sigma_N^2)]$ independent of θ. Compute expressions for $\widehat{\theta}_{\text{QLIN}}$ and $\widehat{\theta}_{\text{LIN}}$. Verify that $\widehat{\theta}_{\text{QLIN}}$ yields a smaller mean-squared error when $m \neq 0$.

3.15 On Page 85, we questioned the existence of an efficient estimator for signal parameters. We found in the succeeding example that an unbiased efficient estimator exists for the signal amplitude. Can a nonlinearly represented parameter, such as time delay, have an efficient estimator?

(a) Simplify the condition for the existence of an efficient estimator by assuming it to be unbiased. Note carefully the dimensions of the matrices involved.

(b) Show that the only solution in this case occurs when the signal depends "linearly" on the parameter vector.

3.16 Cramér-Rao Bound for Signal Parameters
In many problems, the signal as well as the noise are sometimes modeled as Gaussian processes. Let's explore what differences arise in the Cramér-Rao bounds for the stochastic and deterministic signal cases. Assume that the signal contains unknown parameters θ, that it is statistically independent of the noise, and that the noise covariance matrix is known.

(a) What forms do the conditional densities of the observations take under the two assumptions? What are the two covariance matrices?

(b) As a preliminary, show that

$$\frac{\partial \mathbf{A}^{-1}}{\partial \theta} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \theta} \mathbf{A}^{-1}.$$

(c) Assuming the stochastic signal model, show that each element of the Fisher information matrix has the form

$$F_{ij} = \frac{1}{2} \operatorname{tr}\left[\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_i} \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_j}\right]$$

where $\mathbf{K}$ denotes the covariance matrix of the observations. Specialize this expression by assuming the noise component has no unknown parameters.

3.17 Estimating the Amplitude of a Sinusoid
Suppose you observe a discrete-time sinusoid in additive *Laplacian* white noise having variance per sample of σ^2.

$$X_l = A \sin(2\pi f_0 l) + N_l, \; l = 0, \ldots, L-1$$

The frequency is known and is harmonic with the observation interval ($f_0 = n/L$ for some integer n).

(a) What equation determines the maximum likelihood amplitude estimate?

(b) Because no closed form solution for the estimate is evident, write a MATLAB program that simulates the observations and finds the estimate. Set $L = 1024$ and $f_0 = 100/L$. Let $A = 1$ and $\sigma^2 = 1$. Compute and plot the derivative of the log likelihood function for values of A close to the true amplitude. What do you conclude from this result?

(c) For your dataset, find the maximum likelihood estimate of A.

(d) Find the Cramér-Rao bound for the error variance.

(e) In practice, using Gaussian-noise assumption provides far simpler answers than using a model more closely describing the data. Simulate 1,000 observational trials using the Laplacian noise description. Use two values for the noise variance—1 and 25—in each set of trials. Determine the mean and variance of the Gaussian-derived amplitude estimate. What is the theoretical value for the mean-squared error when the Gaussian-based estimate for the amplitude is used? Does the simulated variance agree with this result?

3.18 In Poisson problems, the number of events n occurring in the interval $[0,T)$ is governed by the probability distribution {38}

$$\Pr[n] = \frac{(\lambda_0 T)^n}{n!} e^{-\lambda_0 T} ,$$

where λ_0 is the average rate at which events occur.

(a) What is the maximum likelihood estimate of average rate?

(b) Does this estimate satisfy the Cramér-Rao bound?

(c) Now suppose the rate varies sinusoidally according to

$$\lambda(t) = \lambda_0 \exp\{a \cos 2\pi f_0 t\} \quad 0 \le t < T$$

where the frequency f_0 is a harmonic of $1/T$. What are the maximum likelihood estimates of λ_0 and a in this case?

Note: The following facts will prove useful.

$I_0(a) = \dfrac{1}{2\pi}\displaystyle\int_0^{2\pi} e^{a\cos\theta}\, d\theta$ is the modified Bessel function of the first kind, order 0.

$I_0'(a) = I_1(a)$, the modified Bessel function of the first kind, order 1.

$$I_0''(a) = \frac{I_2(a) + I_0(a)}{2}$$

(d) Find the Cramér-Rao bounds for the mean-squared estimation errors for $\widehat{\lambda_0}$ and $\widehat{a}$ assuming unbiased estimators.

3.19 In the "classic" radar problem, not only is the time of arrival of the radar pulse unknown but also the amplitude. In this problem, we seek methods of simultaneously estimating these parameters. The received signal $X(l)$ is of the form

$$X(l) = \theta_1 s(l - \theta_2) + N(l)$$

where θ_1 is Gaussian with zero mean and variance σ_1^2 and θ_2 is uniformly distributed over the observation interval. Find the receiver that computes the maximum *a posteriori* estimates of θ_1 and θ_2 jointly. Draw a block diagram of this receiver and interpret its structure.

3.20 We can derive the Cramér-Rao bound for estimating a signal's delay.

(a) The parameter θ is the delay of the signal $s(\cdot)$ observed in additive, white Gaussian noise: $X(l) = s(l-\theta) + N(l), l = 0, \dots, L-1$. Derive the Cramér-Rao bound for this problem.

(b) This bound is claimed to be given by $\sigma_N^2/E\beta^2$, where β^2 is the mean-squared bandwidth. Derive this result from your general formula. Does the bound make sense for all values of signal-to-noise ratio E/σ_N^2?

(c) Using optimal detection theory, derive the expression for the probability of error incurred when trying to distinguish between a delay of τ and a delay of $\tau+\Delta$. Consistent with the problem posed for the Cramér-Rao bound, assume the delayed signals are observed in additive, white Gaussian noise.

3.21 Estimating Model Probabilities
We want to estimate the *a priori* probability π_0 based on data obtained over N statistically independent observation intervals. During each length-L observation interval, the observations consist of white Gaussian noise having variance of either σ_0^2 or σ_1^2 ($\sigma_1^2 > \sigma_0^2$). π_0 is the probability that the observations have variance σ_0^2 and we do not know which model applies for any observation interval.

(a) One approach is to classify each observation interval according to its variance, count the number of times the variance equals σ_0^2, and divide this count by N. What is the classification algorithm that lies at the heart of this estimator?

(b) What classifier threshold yields an unbiased estimate of π_0? Comment on the feasibility of this approach.

(c) Rather than use this *ad hoc* approach, let's use a systematic estimation approach: what is the maximum likelihood estimate of π_0 based on the N observation intervals?

3.22 The signal has a power spectrum given by

$$\mathscr{S}_s(f) = \frac{17}{20} \cdot \frac{1+\frac{8}{17}\cos 2\pi f}{1-\frac{4}{5}\cos 2\pi f}$$

This signal is observed in additive white noise having variance equaling 6.

(a) Find the unit-sample response of the noncausal Wiener filter.

(b) Find the difference equation governing the causal Wiener filter ($l_e = 0$).

(c) Calculate the signal processing gain of Wiener filter.

3.23 The *Eckhart filter* is an optimum linear filter that maximizes the signal-to-noise ratio of its output [10]. To find the unit-sample response of the *FIR* Eckhart filter, consider observations of the form $\mathbf{X} = \mathbf{s} + \mathbf{N}$ where the covariance matrix of the noise is known. The signal-to-noise ratio is computed according to $\mathscr{E}[||\mathbf{h}^t\mathbf{s}||^2]/\mathscr{E}[||\mathbf{h}^t\mathbf{N}||^2]$, where $\mathbf{h}$ is the desired unit-sample response.

(a) Assuming the signal is nonrandom, find the Eckhart filter's unit-sample response.

(b) What is the signal-to-noise ratio produced by the Eckhart filter? How does it compare with that produced by the corresponding Wiener filter?

(c) Now assume the signal is random, having covariance matrix $\mathbf{K}_s$. Characterize the Eckhart filter.

3.24 Optimal Spectral Estimation
While many spectral estimation procedures are found in the literature, few take into account the presence of additive noise. Assume that the observations consist of a signal s and statistically independent, additive, zero-mean noise N.

$$X(l) = s(l) + N(l),\ l = 0, \ldots, L-1$$

Treat finding the optimal estimate of the signal's spectrum as an optimal *FIR* filtering problem, where the quantity to be estimated is $\sum_l s(l)e^{-j2\pi fl}$.

(a) Find the spectral estimate that minimizes the mean-squared estimation error.

(b) Find this estimate's mean-squared error.

(c) Under what conditions is this estimate unbiased?

3.25 The covariance function estimate is claimed to be biased because the number of terms used at each lag varied without a corresponding variation in the normalization. Let's explore that claim closely. Assume that we now estimate the covariance function according to

$$\widehat{K}_X(m) = \frac{1}{D-|m|} \sum_{n=0}^{D-|m|-1} X(n)X(n+m)\,, \quad 0 \le |m| \le D-1$$

(a) Find the expected value of this revised estimator, and show that it is indeed unbiased.

(b) To derive the variance of this estimate, we need the fourth moment of the observations, which is conveniently given in Chapter 1 (§1.4.1 {12}). Derive the covariance estimate's variance and determine whether it is consistent or not.

(c) Evaluate the expected value and variance of the spectral estimate corresponding to this covariance estimate.

(d) Does spectral estimate consistency become a reality with this new estimation procedure?

3.26 Let's see how spectral estimators work on a "real" dataset. The file `spectral.mat` contains a signal comprised of a sinusoid and additive noise.

(a) Evaluate the periodogram for this signal. What is your best guess for the sine wave's frequency and amplitude based on the periodogram? Are your estimates "good" in any sense?

(b) Use the Barlett spectral estimation procedure instead of the periodogram. Use a section length of 500, a Hanning window, and half-section overlap. Now what are your estimates of the sinusoid's frequency and amplitude?

(c) The default transform size in MATLAB's `fft` function is the data's length. Instead of using the section length in your Bartlett estimate, use a longer transform to determine the frequency resolution of the spectral peak presumably corresponding to the sinusoid. Compare the resolutions that section lengths of 500 and 1000 provide. Also compare the resolutions the Hanning and rectangular windows provide for these section lengths.

3.27 Optimal Spectral Estimation
While many spectral estimation procedures are found in the literature, few take into account the presence of additive noise. Assume that the observations consist of a signal s and statistically independent, additive, zero-mean noise N.

$$X(l) = s(l) + N(l),\ l = 0,\ldots,L-1$$

Treat finding the optimal estimate of the signal's spectrum as an optimal *FIR* filtering problem, where the quantity to be estimated is $\sum_l s(l)e^{-j2\pi fl}$.

(a) Find the spectral estimate that minimizes the mean-squared estimation error.

(b) Find this estimate's mean-squared error.

(c) Under what conditions is this estimate unbiased?

3.28 Filter Coefficient Estimation
White Gaussian noise $W(l)$ serves as the input to a simple digital filter governed by the difference equation

$$X_l = aX_{l-1} + W_l\,.$$

We want to estimate the filter's coefficient a by processing the output observed over $l = 0,\ldots,L-1$. Prior to $l = 0$, the filter's input is zero.

(a) Find an estimate of a.

(b) What is the Cramér-Rao bound for your estimate?

3.29 The histogram probability density estimator is a special case of a more general class of estimators known as *kernel estimators*.

$$\hat{p}_X(x) = \frac{1}{L}\sum_{l=0}^{L-1} k\big(x - X(l)\big)$$

Here, the kernel $k(\cdot)$ is usually taken to be a density itself.

(a) What is the kernel for the histogram estimator?

(b) Interpret the kernel estimator in signal processing terminology. Predict what the most time consuming computation of this estimate might be. Why?

(c) Show that the sample average equals the expected value of a random variable having the density $\hat{p}_X(x)$ *regardless* of the choice of kernel.

Chapter 4

Detection Theory

4.1 Elementary Hypothesis Testing

In statistics, hypothesis testing is some times known as decision theory or simply testing. Here, one of several *models* $\mathscr{M}$ are presumed to describe a set of observed data, and you want to find which model best describes the observations. The key result around which all decision theory revolves is the likelihood ratio test.

4.1.1 The Likelihood Ratio Test

In a binary hypothesis testing problem, four possible outcomes can result. Model $\mathscr{M}_0$ did in fact represent the best model for the data and the decision rule said it was (a correct decision) or said it wasn't (an erroneous decision). The other two outcomes arise when model $\mathscr{M}_1$ was in fact true with either a correct or incorrect decision made. The decision process operates by segmenting the range of observation values into two disjoint *decision regions* $\Re_0$ and $\Re_1$. All values of $\mathbf{X}$ fall into either $\Re_0$ or $\Re_1$. If a given $\mathbf{X}$ lies in $\Re_0$, for example, we will announce our decision "model $\mathscr{M}_0$ was true"; if in $\Re_1$, model $\mathscr{M}_1$ would be proclaimed. To derive a rational method of deciding which model best describes the observations, we need a criterion to assess the quality of the decision process. Optimizing this criterion will specify the decision regions.

The *Bayes' decision criterion* seeks to minimize a cost function associated with making a decision. Let C_{ij} be the cost of mistaking model j for model i ($i \neq j$) and C_{ii} the presumably smaller cost of correctly choosing model i: $C_{ij} > C_{ii}, i \neq j$. Let π_i be the *a priori* probability of model i. The so-called *Bayes' cost* $\overline{C}$ is the average cost of making a decision.

$$\begin{aligned}
\overline{C} &= \sum_{i,j} C_{ij} \Pr[\text{say } \mathscr{M}_i \text{ when } \mathscr{M}_j \text{ true}] \\
&= \sum_{i,j} C_{ij} \pi_j \Pr[\text{say } \mathscr{M}_i | \mathscr{M}_j \text{ true}]
\end{aligned}$$

The Bayes' cost can be expressed as

$$\begin{aligned}
\overline{C} &= \sum_{i,j} C_{ij} \pi_j \Pr[\mathbf{X} \in \Re_i | \mathscr{M}_j \text{ true}] \\
&= \sum_{i,j} C_{ij} \pi_j \int_{\Re_i} p_{\mathbf{X}|\mathscr{M}_j}(\mathbf{X}|\mathscr{M}_j)\, d\mathbf{X} \\
&= \int_{\Re_0} \{C_{00}\pi_0 p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0) + C_{01}\pi_1 p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1)\}\, d\mathbf{X} \\
&+ \int_{\Re_1} \{C_{10}\pi_0 p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0) + C_{11}\pi_1 p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1)\}\, d\mathbf{X}
\end{aligned}$$

$p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{X}|\mathscr{M}_i)$ is the conditional probability density function of the observed data $\mathbf{X}$ given that model $\mathscr{M}_i$ was true. To minimize this expression with respect to the decision regions $\Re_0$ and $\Re_1$, ponder which integral would yield the smallest value if its integration domain included a specific observation vector. This selection process defines the decision regions; for example, we choose $\mathscr{M}_0$ for those values of $\mathbf{X}$ which yield a smaller value for the first integral.

$$\Re_0 = \{\mathbf{X}\colon \pi_0 C_{00} p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0) + \pi_1 C_{01} p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1) < \pi_0 C_{10} p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0) + \pi_1 C_{11} p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1)\}$$

We choose $\mathscr{M}_1$ when the inequality is reversed. This expression is easily manipulated to obtain the decision rule known as the *likelihood ratio test*.

$$\boxed{\frac{p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1)}{p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0)} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{\pi_0(C_{10}-C_{00})}{\pi_1(C_{01}-C_{11})}} \tag{4.1}$$

The comparison relation means selecting model $\mathscr{M}_1$ if the left-hand ratio exceeds the value on the right; otherwise, $\mathscr{M}_0$ is selected. Thus, the *likelihood ratio* $p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1)/p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0)$, symbolically represented by $\Lambda(\mathbf{X})$, is computed from the observed value of $\mathbf{X}$ and then compared with a *threshold* η equaling $\left[\pi_0(C_{10}-C_{00})\right]/\left[\pi_1(C_{01}-C_{11})\right]$. Thus, when two models are hypothesized, the likelihood ratio test can be succinctly expressed as the comparison of the likelihood ratio with a threshold.

$$\boxed{\Lambda(\mathbf{X}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \eta} \tag{4.2}$$

The data processing operations are captured entirely by the likelihood ratio $p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1)/p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0)$. Furthermore, note that only the value of the likelihood ratio *relative* to the threshold matters; to simplify the computation of the likelihood ratio, we can perform *any* positively monotonic operations simultaneously on the likelihood ratio and the threshold without affecting the comparison. We can multiply the ratio by a positive constant, add any constant, or apply a monotonically increasing function which simplifies the expressions. We single one such function, the logarithm, because it simplifies likelihood ratios that commonly occur in signal processing applications. Known as the log-likelihood, we explicitly express the likelihood ratio test with it as

$$\boxed{\ln\Lambda(\mathbf{X}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \ln\eta}\,. \tag{4.3}$$

Useful simplifying transformations are problem-dependent; by laying bare that aspect of the observations essential to the model testing problem, we reveal the *sufficient statistic* $\Upsilon(\mathbf{X})$: the scalar quantity which best summarizes the data [19: pp. 18-22]. The likelihood ratio test is best expressed in terms of the sufficient statistic.

$$\boxed{\Upsilon(\mathbf{X}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma} \tag{4.4}$$

We will denote the threshold value by γ when the sufficient statistic is used or by η when the likelihood ratio appears prior to its reduction to a sufficient statistic.

As we shall see, if we use a different criterion other than the Bayes' criterion, the decision rule often involves the likelihood ratio. The likelihood ratio is comprised of the quantities $p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{X}|\mathscr{M}_i)$, termed the *likelihood function*, which is also important in estimation theory. It is this conditional density that portrays the probabilistic model describing data generation. The likelihood function completely characterizes the kind

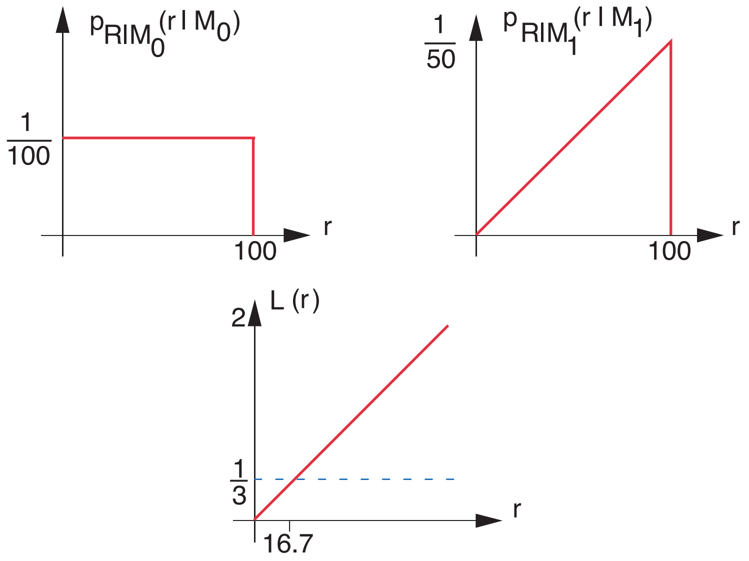


Figure 4.1: Conditional densities for the grade distributions assuming that a student did not study ($\mathcal{M}_0$) or did ($\mathcal{M}_1$) are shown in the top row. The lower portion depicts the likelihood ratio formed from these densities.

of "world" assumed by each model; for each model, we must specify the likelihood function so that we can solve the hypothesis testing problem.

A complication, which arises in some cases, is that the sufficient statistic may not be monotonic. If monotonic, the decision regions $\Re_0$ and $\Re_1$ are simply connected (all portions of a region can be reached without crossing into the other region). If not, the regions are not simply connected and decision region islands are created (see Problem 4.2). Such regions usually complicate calculations of decision performance. Monotonic or not, the decision rule proceeds as described: the sufficient statistic is computed for each observation vector and compared to a threshold.

Example

An instructor in a course in detection theory wants to determine if a particular student studied for his last test. The observed quantity is the student's grade, which we denote by X. Failure may not indicate studiousness: conscientious students may fail the test. Define the models as

$$\begin{aligned} \mathcal{M}_0&: \quad \text{did not study} \\ \mathcal{M}_1&: \quad \text{studied} \end{aligned}$$

The conditional densities of the grade are shown in Fig. 4.1. Based on knowledge of student behavior, the instructor assigns *a priori* probabilities of $\pi_0 = 1/4$ and $\pi_1 = 3/4$. The costs C_{ij} are chosen to reflect the instructor's sensitivity to student feelings: $C_{01} = 1 = C_{10}$ (an erroneous decision either way is given the same cost) and $C_{00} = 0 = C_{11}$. The likelihood ratio is plotted in Fig. 4.1 and the threshold value η, which is computed from the *a priori* probabilities and the costs to be $1/3$, is indicated. The calculations of this comparison can be simplified in an obvious way.

$$\frac{X}{50} \underset{\mathcal{M}_0}{\overset{\mathcal{M}_1}{\gtrless}} \frac{1}{3} \quad \text{or} \quad X \underset{\mathcal{M}_0}{\overset{\mathcal{M}_1}{\gtrless}} \frac{50}{3} = 16.7$$

The multiplication by the factor of 50 is a simple illustration of the reduction of the likelihood ratio to a sufficient statistic. Based on the assigned costs and *a priori* probabilities, the optimum decision rule says the instructor must assume that the student did not study if the student's grade is less than 16.7; if greater, the student is assumed to have studied despite receiving an abysmally low grade such as 20. Note that as the densities given by each model overlap entirely: the possibility of making the wrong interpretation *always* haunts the instructor. However, no other procedure will be better!

4.1.2 Criteria in Hypothesis Testing

The criterion used in the previous section—minimize the average cost of an incorrect decision—may seem to be a contrived way of quantifying decisions. Well, often it is. For example, the Bayesian decision rule depends explicitly on the *a priori* probabilities; a rational method of assigning values to these—either by experiment or through true knowledge of the relative likelihood of each model—may be unreasonable. In this section, we develop alternative decision rules that try to answer such objections. One essential point will emerge from these considerations: *the fundamental nature of the decision rule does not change with choice of optimization criterion*. Even criteria remote from error measures can result in the likelihood ratio test (see Problem 4.4). Such results do not occur often in signal processing and underline the likelihood ratio test's significance.

Maximum Probability of a Correct Decision

As only one model can describe any given set of data (the models are mutually exclusive), the probability of being correct P_c for distinguishing two models is given by

$$P_c = \Pr[\text{say } \mathscr{M}_0 \text{ when } \mathscr{M}_0 \text{ true}] + \Pr[\text{say } \mathscr{M}_1 \text{ when } \mathscr{M}_1 \text{ true}]\,.$$

We wish to determine the optimum decision region placement by maximizing P_c. Expressing the probability correct in terms of the likelihood functions $p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{X}|\mathscr{M}_i)$, the *a priori* probabilities, and the decision regions,

$$P_c = \int_{\Re_0} \pi_0 p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0)\,d\mathbf{X} + \int_{\Re_1} \pi_1 p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1)\,d\mathbf{X}\,.$$

We want to maximize P_c by selecting the decision regions $\Re_0$ and $\Re_1$. The probability correct is maximized by associating each value of $\mathbf{X}$ with the largest term in the expression for P_c. Decision region $\Re_0$, for example, is defined by the collection of values of $\mathbf{X}$ for which the first term is largest. As all of the quantities involved are non-negative, the decision rule maximizing the probability of a correct decision is

$$\boxed{\text{Given } \mathbf{X},\ \text{choose } \mathscr{M}_i \text{ for which the product } \pi_i p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{X}|\mathscr{M}_i) \text{ is largest.}}$$

Simple manipulations lead to the likelihood ratio test.

$$\frac{p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1)}{p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0)} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{\pi_0}{\pi_1}$$

Note that if the Bayes' costs were chosen so that $C_{ii} = 0$ and $C_{ij} = C$, $(i \neq j)$, we would have the same threshold as in the previous section.

To evaluate the quality of the decision rule, we usually compute the *probability of error* P_e rather than the probability of being correct. This quantity can be expressed in terms of the observations, the likelihood ratio, and the sufficient statistic.

$$\boxed{\begin{aligned} P_e &= \pi_0 \int_{\Re_1} p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0)\,d\mathbf{X} + \pi_1 \int_{\Re_0} p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1)\,d\mathbf{X} \\ &= \pi_0 \int_{\Lambda>\eta} p_{\Lambda|\mathscr{M}_0}(\Lambda|\mathscr{M}_0)\,d\Lambda + \pi_1 \int_{\Lambda<\eta} p_{\Lambda|\mathscr{M}_1}(\Lambda|\mathscr{M}_1)\,d\Lambda \\ &= \pi_0 \int_{\Upsilon>\gamma} p_{\Upsilon|\mathscr{M}_0}(\Upsilon|\mathscr{M}_0)\,d\Upsilon + \pi_1 \int_{\Upsilon<\gamma} p_{\Upsilon|\mathscr{M}_1}(\Upsilon|\mathscr{M}_1)\,d\Upsilon \end{aligned}} \tag{4.5}$$

When the likelihood ratio is non-monotonic, the first expression is most difficult to evaluate. When monotonic, the middle expression proves the most difficult. Furthermore, these expressions point out that the likelihood ratio and the sufficient statistic can be considered a function of the observations $\mathbf{X}$; hence, they are random variables and have probability densities for each model. Another aspect of the resulting probability of error is that *no other decision rule can yield a lower probability of error*. This statement is obvious as we minimized the probability of error in deriving the likelihood ratio test. The point is that these expressions represent a lower bound on performance (as assessed by the probability of error). This probability will be non-zero if the conditional densities overlap over some range of values of $\mathbf{X}$, such as occurred in the previous example. In this region of overlap, the observed values are ambiguous: either model is consistent with the observations. Our "optimum" decision rule operates in such regions by selecting that model which is most likely (has the highest probability) of generating any particular value.

Neyman-Pearson Criterion

Situations occur frequently where assigning or measuring the *a priori* probabilities P_i is unreasonable. For example, just what is the *a priori* probability of a supernova occurring in any particular region of the sky? We clearly need a model evaluation procedure which can function without *a priori* probabilities. This kind of test results when the so-called Neyman-Pearson criterion is used to derive the decision rule. The ideas behind and decision rules derived with the Neyman-Pearson criterion [27] will serve us well in sequel; their result is important!

Using nomenclature from radar, where model $\mathscr{M}_1$ represents the presence of a target and $\mathscr{M}_0$ its absence, the various types of correct and incorrect decisions have the following names [44: pp. 127–9].*

$P_D = \Pr[\text{say } \mathscr{M}_1 \vert \mathscr{M}_1 \text{ true}]$	*Detection*–we say it's there when it is
$P_F = \Pr[\text{say } \mathscr{M}_1 \vert \mathscr{M}_0 \text{ true}]$	*False-alarm*–we say it's there when it's not
$P_M = \Pr[\text{say } \mathscr{M}_0 \vert \mathscr{M}_1 \text{ true}]$	*Miss*–we say it's not there when it is

The remaining probability $\Pr[\text{say } \mathscr{M}_0|\mathscr{M}_0 \text{ true}]$ has historically been left nameless and equals $1-P_F$. We should also note that the detection and miss probabilities are related by $P_M = 1-P_D$. As these are conditional probabilities, they do not depend on the *a priori* probabilities and the two probabilities P_F and P_D characterize the errors when *any* decision rule is used.

These two probabilities are related to each other in an interesting way. Expressing these quantities in terms of the decision regions and the likelihood functions, we have

$$P_F = \int_{\Re_1} p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0)\,d\mathbf{X}, \qquad P_D = \int_{\Re_1} p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1)\,d\mathbf{X}\,.$$

As the region $\Re_1$ shrinks, *both* of these probabilities tend toward zero; as $\Re_1$ expands to engulf the entire range of observation values, they both tend toward unity. This rather direct relationship between P_D and P_F does not mean that they equal each other; in most cases, as $\Re_1$ expands, P_D increases more rapidly than P_F (we had better be right more often than we are wrong!). However, the "ultimate" situation where a rule is always right and never wrong ($P_D = 1, P_F = 0$) cannot occur when the conditional distributions overlap. Thus, to increase the detection probability we must also allow the false-alarm probability to increase. This behavior represents the fundamental tradeoff in hypothesis testing and detection theory.

One can attempt to impose a performance criterion that depends only on these probabilities with the consequent decision rule not depending on the *a priori* probabilities. The Neyman-Pearson criterion assumes that the false-alarm probability is constrained to be less than or equal to a specified value α while we attempt to maximize the detection probability P_D.

$$\max_{\Re_1} P_D \;\text{ subject to } P_F \leq \alpha$$

*In hypothesis testing, a false-alarm is known as a type I error and a miss a type II error.

A subtlety of the succeeding solution is that the underlying probability distribution functions may not be continuous, with the result that P_F can never equal the constraining value α. Furthermore, an (unlikely) possibility is that the optimum value for the false-alarm probability is somewhat less than the criterion value. Assume, therefore, that we rephrase the optimization problem by requiring that the false-alarm probability equal a value α' that is less than or equal to α.

This optimization problem can be solved using Lagrange multipliers; we seek to find the decision rule that maximizes

$$F = P_D + \lambda(P_F - \alpha'),$$

where λ is the Lagrange multiplier. This optimization technique amounts to finding the decision rule that maximizes F, then finding the value of the multiplier that allows the criterion to be satisfied. As is usual in the derivation of optimum decision rules, we maximize these quantities with respect to the decision regions. Expressing P_D and P_F in terms of them, we have

$$\begin{aligned} F &= \int_{\Re_1} p_{\mathbf{X}|\mathcal{M}_1}(\mathbf{X}|\mathcal{M}_1)\,d\mathbf{X} + \lambda\left(\int_{\Re_1} p_{\mathbf{X}|\mathcal{M}_0}(\mathbf{X}|\mathcal{M}_0)d\mathbf{X} - \alpha'\right) \\ &= -\lambda\alpha' + \int_{\Re_1}\left[p_{\mathbf{X}|\mathcal{M}_1}(\mathbf{X}|\mathcal{M}_1) + \lambda p_{\mathbf{X}|\mathcal{M}_0}(\mathbf{X}|\mathcal{M}_0)\right] d\mathbf{X}. \end{aligned}$$

To maximize this quantity with respect to $\Re_1$, we need only to integrate over those regions of $\mathbf{X}$ where the integrand is positive. The region $\Re_1$ thus corresponds to those values of $\mathbf{X}$ where $p_{\mathbf{X}|\mathcal{M}_1}(\mathbf{X}|\mathcal{M}_1) > -\lambda p_{\mathbf{X}|\mathcal{M}_0}(\mathbf{X}|\mathcal{M}_0)$ and the resulting decision rule is

$$\frac{p_{\mathbf{X}|\mathcal{M}_1}(\mathbf{X}|\mathcal{M}_1)}{p_{\mathbf{X}|\mathcal{M}_0}(\mathbf{X}|\mathcal{M}_0)} \underset{\mathcal{M}_0}{\overset{\mathcal{M}_1}{\gtrless}} -\lambda$$

The ubiquitous likelihood ratio test again appears; it *is* indeed the fundamental quantity in hypothesis testing. Using the logarithm of the likelihood ratio or the sufficient statistic, this result can be expressed as either

$$\ln\Lambda(\mathbf{X}) \underset{\mathcal{M}_0}{\overset{\mathcal{M}_1}{\gtrless}} \ln(-\lambda) \quad \text{or} \quad \Upsilon(\mathbf{X}) \underset{\mathcal{M}_0}{\overset{\mathcal{M}_1}{\gtrless}} \gamma.$$

We have not as yet found a value for the threshold. The false-alarm probability can be expressed in terms of the Neyman-Pearson threshold in two (useful) ways.

$$\boxed{\begin{aligned} P_F &= \int_{-\lambda}^{\infty} p_{\Lambda|\mathcal{M}_0}(\Lambda|\mathcal{M}_0)\,d\Lambda \\ &= \int_{\gamma}^{\infty} p_{\Upsilon|\mathcal{M}_0}(\Upsilon|\mathcal{M}_0)\,d\Upsilon \end{aligned}} \tag{4.6}$$

One of these implicit equations must be solved for the threshold by setting P_F equal to α'. The selection of which to use is usually based on pragmatic considerations: the easiest to compute. From the previous discussion of the relationship between the detection and false-alarm probabilities, we find that to maximize P_D we must allow α' to be as large as possible while remaining less than α. Thus, we want to find the *smallest* value of $-\lambda$ (note the minus sign) consistent with the constraint. Computation of the threshold is problem-dependent, but a solution always exists.

Example

An important application of the likelihood ratio test occurs when $\mathbf{X}$ is a Gaussian random vector for each model. Suppose the models correspond to Gaussian random vectors having different mean values but sharing the same identity covariance.

$$\mathscr{M}_0 : \mathbf{X} \sim \mathscr{N}(\mathbf{0}, \sigma^2\mathbf{I})$$
$$\mathscr{M}_1 : \mathbf{X} \sim \mathscr{N}(\mathbf{m}, \sigma^2\mathbf{I})$$

Thus, $\mathbf{X}$ is of dimension L and has statistically independent, equal variance components. The vector of means $\mathbf{m} = \text{col}[m_0, \ldots, m_{L-1}]$ distinguishes the two models. The likelihood functions associated this problem are

$$p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0) = \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{X_l}{\sigma}\right)^2\right\}$$
$$p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1) = \prod_{l=0}^{L-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{X_l - m_l}{\sigma}\right)^2\right\}$$

The likelihood ratio $\Lambda(\mathbf{X})$ becomes

$$\Lambda(\mathbf{X}) = \frac{\displaystyle\prod_{l=0}^{L-1} \exp\left\{-\frac{1}{2}\left(\frac{X_l - m_l}{\sigma}\right)^2\right\}}{\displaystyle\prod_{l=0}^{L-1} \exp\left\{-\frac{1}{2}\left(\frac{X_l}{\sigma}\right)^2\right\}}$$

This expression for the likelihood ratio is complicated. In the Gaussian case (and many others), we use the logarithm the reduce the complexity of the likelihood ratio and form a sufficient statistic.

$$\begin{aligned}\ln\Lambda(\mathbf{X}) &= \sum_{l=0}^{L-1}\left\{-\frac{1}{2}\frac{(X_l - m_l)^2}{\sigma^2} + \frac{1}{2}\frac{X_l^2}{\sigma^2}\right\} \\ &= \frac{1}{\sigma^2}\sum_{l=0}^{L-1} m_l X_l - \frac{1}{2\sigma^2}\sum_{l=0}^{L-1} m_l^2\end{aligned}$$

The likelihood ratio test then has the much simpler, but equivalent form

$$\sum_{l=0}^{L-1} m_l X_l \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln\eta + \frac{1}{2}\sum_{l=0}^{L-1} m_l^2 .$$

To focus on the model evaluation aspects of this problem, let's assume means be equal to a positive constant: $m_l = m\ (> 0)$.*

$$\sum_{l=0}^{L-1} X_l \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \frac{\sigma^2}{m}\ln\eta + \frac{Lm}{2}$$

Note that all that need be known about the observations $\{X_l\}$ is their sum. This quantity is the sufficient statistic for the Gaussian problem: $\Upsilon(\mathbf{X}) = \sum X_l$ and $\gamma = \sigma^2 \ln\eta/m + Lm/2$.

When trying to compute the probability of error or the threshold in the Neyman-Pearson criterion, we must find the conditional probability density of one of the decision statistics: the likelihood ratio, the log-likelihood, or the sufficient statistic. The log-likelihood and the sufficient statistic are quite

*Why did the author assume that the mean was positive? What would happen if it were negative?

x	$Q^{-1}(x)$
10^{-1}	1.281
10^{-2}	2.396
10^{-3}	3.090
10^{-4}	3.719
10^{-5}	4.265
10^{-6}	4.754

Table 4.1: The table displays interesting values for $Q^{-1}(\cdot)$ that can be used to determine thresholds in the Neyman-Pearson variant of the likelihood ratio test. Note how little the inverse function changes for decade changes in its argument; $Q(\cdot)$ is indeed *very* nonlinear.

similar in this problem, but clearly we should use the latter. One practical property of the sufficient statistic is that it usually simplifies computations. For this Gaussian example, the sufficient statistic is a Gaussian random variable under each model.

$$\begin{aligned}\mathscr{M}_0 &: \Upsilon(\mathbf{X}) \sim \mathscr{N}(0, L\sigma^2)\\ \mathscr{M}_1 &: \Upsilon(\mathbf{X}) \sim \mathscr{N}(Lm, L\sigma^2)\end{aligned}$$

To find the probability of error from the expressions found on Page 112, we must evaluate the area under a Gaussian probability density function. These integrals are succinctly expressed in terms of $Q(x)$, which denotes the probability that a unit-variance, zero-mean Gaussian random variable exceeds x (see chapter 1 {9}). As $1-Q(x)=Q(-x)$, the probability of error can be written as

$$P_e = \pi_1 Q\left(\frac{Lm-\gamma}{\sqrt{L}\sigma}\right) + \pi_0 Q\left(\frac{\gamma}{\sqrt{L}\sigma}\right).$$

An interesting special case occurs when $\pi_0 = 1/2 = \pi_1$. In this case, $\gamma = Lm/2$ and the probability of error becomes

$$P_e = Q\left(\frac{\sqrt{L}m}{2\sigma}\right).$$

As $Q(\cdot)$ is a monotonically decreasing function, the probability of error decreases with increasing values of the ratio $\sqrt{L}m/2\sigma$. However, as shown in appendix Fig. 1.3 {10}, $Q(\cdot)$ decreases in a nonlinear fashion. Thus, increasing m by a factor of two may decrease the probability of error by a larger *or* a smaller factor; the amount of change depends on the initial value of the ratio.

To find the threshold for the Neyman-Pearson test from the expressions given on Page 114, we need the area under a Gaussian density.

$$P_F = Q\left(\frac{\gamma}{\sqrt{L\sigma^2}}\right) = \alpha' \tag{4.7}$$

As $Q(\cdot)$ is a monotonic and continuous function, we can now set α' equal to the criterion value α with the result

$$\gamma = \sqrt{L}\sigma Q^{-1}(\alpha).$$

where $Q^{-1}(\cdot)$ denotes the inverse function of $Q(\cdot)$. The solution of this equation cannot be performed analytically as no closed form expression exists for $Q(\cdot)$ (much less its inverse function); the criterion value must be found from tables or numerical routines. Because Gaussian problems arise frequently, the accompanying table provides numeric values for this quantity at the decade points. The detection probability is given by

$$P_D = Q\left(Q^{-1}(\alpha) - \frac{\sqrt{L}m}{\sigma}\right).$$

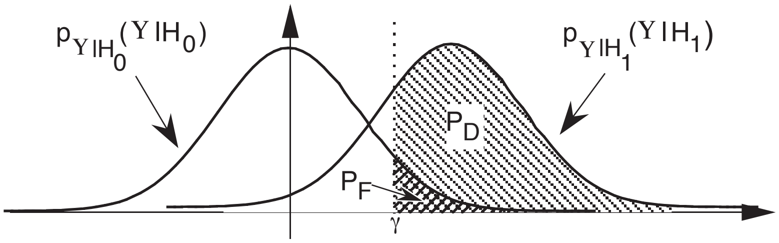


Figure 4.2: The densities of the sufficient statistic $\Upsilon(\mathbf{X})$ conditioned on two hypotheses are shown for the Gaussian example. The threshold γ used to distinguish between the two models is indicated. The false-alarm probability is the area under the density corresponding to $\mathscr{M}_0$ to the right of the threshold; the detection probability is the area under the density corresponding to $\mathscr{M}_1$.

4.1.3 Performance Evaluation

We alluded earlier {113} to the relationship between the false-alarm probability P_F and the detection probability P_D as one varies the decision region. Because the Neyman-Pearson criterion depends on specifying the false-alarm probability to yield an acceptable detection probability, we need to examine carefully how the detection probability is affected by a specification of the false-alarm probability. The usual way these quantities are discussed is through a parametric plot of P_D versus P_F: the *receiver operating characteristic* or ROC.

As we discovered in the Gaussian example {115}, the sufficient statistic provides the simplest way of computing these probabilities; thus, they are usually considered to depend on the threshold parameter γ. In these terms, we have

$$P_D = \int_\gamma^\infty p_{\Upsilon|\mathscr{M}_1}(\Upsilon|\mathscr{M}_1)\,d\Upsilon \quad \text{and} \quad P_F = \int_\gamma^\infty p_{\Upsilon|\mathscr{M}_0}(\Upsilon|\mathscr{M}_0)\,d\Upsilon\,. \tag{4.8}$$

These densities and their relationship to the threshold γ are shown in Fig. 4.2. We see that the detection probability is greater than or equal to the false-alarm probability. Since these probabilities must decrease monotonically as the threshold is increased, the ROC curve must be concave-down and must *always* exceed the equality line (Fig. 4.3). The degree to which the ROC departs from the equality line $P_D = P_F$ measures the relative "distinctiveness" between the two hypothesized models for generating the observations. In the limit, the two models can be distinguished perfectly if the ROC is discontinuous and consists of the point $(1,0)$. The two are totally confused if the ROC lies on the equality line (this would mean, of course, that the two models are identical); distinguishing the two in this case would be "somewhat difficult".

Example

Consider the Gaussian example we have been discussing where the two models differ only in the means of the conditional distributions. In this case, the two model-testing probabilities are given by

$$P_F = Q\left(\frac{\gamma}{\sqrt{L}\sigma}\right) \quad \text{and} \quad P_D = Q\left(\frac{\gamma - Lm}{\sqrt{L}\sigma}\right)\,.$$

By re-expressing γ as $\frac{\sigma^2}{m}\gamma' + \frac{Lm}{2}$, we discover that these probabilities depend only on the ratio $\sqrt{L}m/\sigma$.

$$P_F = Q\left(\frac{\gamma'}{\sqrt{L}m/\sigma} + \frac{\sqrt{L}m}{2\sigma}\right), \qquad P_D = Q\left(\frac{\gamma'}{\sqrt{L}m/\sigma} - \frac{\sqrt{L}m}{2\sigma}\right)$$

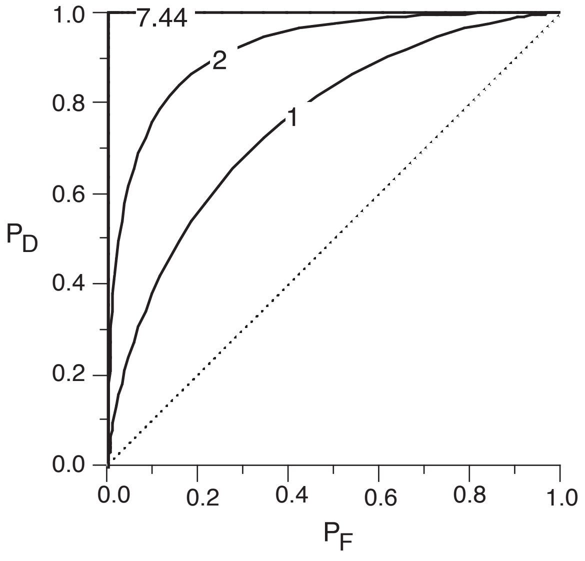


Figure 4.3: A plot of the receiver operating characteristic for the densities shown in the previous figure. Three ROC curves are shown corresponding to different values for the parameter $\sqrt{L}m/\sigma$.

As this signal-to-noise ratio increases, the ROC curve approaches its "ideal" form: the northwest corner of a square as illustrated in Fig. 4.3 by the value of 7.44 for $\sqrt{L}m/\sigma$, which corresponds to a signal-to-noise ratio of $7.44^2 \approx 17$ dB. If a small false-alarm probability (say 10^{-4}) is specified, a large detection probability (0.9999) can result. Such values of signal-to-noise ratios can thus be considered "large" and the corresponding model evaluation problem relatively easy. If, however, the signal-to-noise ratio equals 4 (6 dB), the figure illustrates the worsened performance: a 10^{-4} specification on the false-alarm probability would result in a detection probability of essentially zero. Thus, in a fairly small signal-to-noise ratio range, the likelihood ratio test's performance capabilities can vary dramatically. However, no other decision rule can yield better performance.

Specification of the false-alarm probability for a new problem requires experience. Choosing a "reasonable" value for the false-alarm probability in the Neyman-Pearson criterion depends strongly on the problem difficulty. Too small a number will result in small detection probabilities; too large and the detection probability will be close to unity, suggesting that fewer false alarms could have been tolerated. Problem difficulty is assessed by the degree to which the conditional densities $p_{\mathbf{X}|\mathscr{M}_0}(\mathbf{X}|\mathscr{M}_0)$ and $p_{\mathbf{X}|\mathscr{M}_1}(\mathbf{X}|\mathscr{M}_1)$ overlap, a problem dependent measurement. If we are testing whether a distribution has one of two possible mean values as in our Gaussian example, a quantity like a signal-to-noise ratio will probably emerge as determining performance. The performance in this case can vary drastically depending on whether the signal-to-noise ratio is large or small. In other kinds of problems, the best possible performance provided by the likelihood ratio test can be poor. For example, consider the problem of determining which of two zero-mean probability densities describes a given set of data consisting of statistically independent observations (Problem 4.2). Presumably, the variances of these two densities are equal as we are trying to determine which density is most appropriate. In this case, the performance probabilities can be quite low, especially when the general shapes of the densities are similar. Thus a single quantity, like the signal-to-noise ratio, does *not* emerge to characterize problem difficulty in all hypothesis testing problems. In sequel, we will analyze each model evaluation and detection problem in a standard way. After the sufficient statistic has been found, we will seek a value for the threshold that attains a specified false-alarm probability. The detection probability will then be determined as a function of "problem difficulty", the measure of which is problem-dependent. We can control the choice of false-alarm probability; we cannot control over problem difficulty. Confusingly, the detection probability will vary with *both* the specified false-alarm probability and the problem difficulty.

We are implicitly assuming that we have a rational method for choosing the false-alarm probability criterion value. In signal processing applications, we usually make a sequence of decisions and pass them to systems making more global determinations. For example, in digital communications problems the model evaluation formalism could be used to "receive" each bit. Each bit is received in sequence and then passed to the decoder which invokes error-correction algorithms. The important notions here are that the decision-making process occurs at a given *rate* and that the decisions are presented to other signal processing systems. The rate at which errors occur in system input(s) greatly influences system design. Thus, the selection of a false-alarm probability is usually governed by the *error rate* that can be tolerated by succeeding systems. If the decision rate is one per day, then a moderately large (say 0.1) false-alarm probability might be appropriate. If the decision rate is a million per second as in a one megabit communication channel, the false-alarm probability should be much lower: 10^{-12} would suffice for the one-tenth per day error rate.

4.1.4 Beyond Two Models

Frequently, more than two viable models for data generation can be defined for a given situation. The *classification* problem is to determine which of several models best "fits" a set of measurements. For example, determining the type of airplane from its radar returns forms a classification problem. The model evaluation framework has the right structure if we can allow more than two model. We happily note that in deriving the likelihood ratio test we did not need to assume that only two possible descriptions exist. Go back and examine the expression for the maximum probability correct decision rule $\{112\}$. If K models seem appropriate for a specific problem, the decision rule maximizing the probability of making a correct choice is

$$\boxed{\text{Choose the largest of } \pi_i p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{X}|\mathscr{M}_i), \quad i=1,\dots,K.}$$

To determine the largest of K quantities, exactly $K-1$ numeric comparisons need be made. When we have two possible models ($K=2$), this decision rule reduces to the computation of the likelihood ratio and its comparison to a threshold. In general, $K-1$ likelihood ratios need to be computed and compared to a threshold. Thus the likelihood ratio test can be viewed as a specific method for determining the largest of the decision statistics $\pi_i p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{X}|\mathscr{M}_i)$.

Since we need only the relative ordering of the K decision statistics to make a decision, we can apply any transformation $T(\cdot)$ to them that does not affect ordering. In general, possible transformations must be positively monotonic to satisfy this condition. For example, the needless common additive components in the decision statistics can be eliminated, even if they depend on the observations. Mathematically, "common" means that the quantity does not depend on the model index i. The transformation in this case would be of the form $T(z_i)=z_i-a$, clearly a monotonic transformation. A *positive* multiplicative factor can also the "canceled"; if negative, the ordering would be reversed and that cannot be allowed. The simplest resulting expression becomes the sufficient statistic $\Upsilon_i(\mathbf{X})$ for the model. Expressed in terms of the sufficient statistic, the maximum probability correct or the Bayesian decision rule becomes

$$\boxed{\text{Choose the largest of } C_i+\Upsilon_i(\mathbf{X}), \quad i=1,\dots,K},$$

where C_i summarizes all additive terms that do not depend on the observation vector $\mathbf{X}$. The quantity $\Upsilon_i(\mathbf{X})$ is termed the *sufficient statistic associated with model i*. In many cases, the functional form of the sufficient statistic varies little from one model to another and expresses the necessary operations that summarize the observations. The constants C_i are usually lumped together to yield the threshold against which we compare the sufficient statistic. For example, in the binary model situation, the decision rule becomes

$$\Upsilon_1(\mathbf{X})+C_1 \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \Upsilon_0(\mathbf{X})+C_0 \quad \text{or} \quad \Upsilon_1(\mathbf{X})-\Upsilon_0(\mathbf{X}) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} C_0-C_1.$$

Thus, the sufficient statistic for the decision rule is $\Upsilon_1(\mathbf{X})-\Upsilon_0(\mathbf{X})$ and the threshold γ is C_0-C_1.

Example

In the Gaussian problem just discussed, the logarithm of the likelihood function is

$$\ln p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{X}|\mathscr{M}_i) = -\frac{L}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{l=0}^{L-1}(X_l - m^{(i)})^2\,,$$

where $m^{(i)}$ is the mean under model i. After appropriate simplification that retains the ordering, we have

$$\Upsilon_i(\mathbf{X}) = \frac{m^{(i)}}{\sigma^2}\sum_{l=0}^{L-1} X_l \qquad C_i = -\frac{1}{2}\frac{Lm^{(i)^2}}{\sigma^2} + c_i\,.$$

The term c_i is a constant defined by the error criterion; for the maximum probability correct criterion, this constant is $\ln \pi_i$.

When employing the Neyman-Pearson test, we need to specify the various error probabilities $\Pr[\text{say } \mathscr{M}_i|\mathscr{M}_j \text{ true}]$. These specifications amount to determining the constants c_i when the sufficient statistic is used. Since $K-1$ comparisons will be used to home in on the optimal decision, only $K-1$ error probabilities need be specified. Typically, the quantities $\Pr[\text{say } \mathscr{M}_i|\mathscr{M}_0 \text{ true}]$, $i = 1,\dots,K-1$, are used, particularly when the model $\mathscr{M}_0$ represents the situation when no signal is present (see Problem 4.7).

4.2 Detection of Signals in Gaussian Noise

For the moment, we assume we know the joint distribution of the noise values. In most cases, the various models for the form of the observations—the hypotheses—do not differ because of noise characteristics. Rather, the signal component determines model variations and the noise is statistically independent of the signal; such is the specificity of detection problems in contrast to the generality of model evaluation. For example, we may want to determine whether a signal characteristic of a particular ship is present in a sonar array's output (the signal is known) or whether no ship is present (zero-valued signal).

To apply optimal hypothesis testing procedures previously derived, we first obtain a finite number L of observations—$X(l)$, $l = 0,\dots,L-1$. These observations are usually obtained from continuous-time observations in one of two ways. Two commonly used methods for passing from continuous-time to discrete-time are known: *integrate-and-dump* and *sampling*. These techniques are illustrated in figure 4.4.

Integrate-and-Dump

In this procedure, no attention is paid to the bandwidth of the noise in selecting the sampling rate. Instead, the sampling interval Δ is selected according to the characteristics of the signal set. Because of the finite duration of the integrator, successive samples are statistically independent when the noise bandwidth exceeds $1/\Delta$. Consequently, the sampling rate can be varied to some extent while retaining this desirable analytic property.

Sampling

Traditional engineering considerations governed the selection of the sampling filter and the sampling rate. As in the integrate-and-dump procedure, the sampling rate is chosen according to signal properties. Presumably, changes in sampling rate would force changes in the filter. As we shall see, this linkage has dramatic implications on performance.

With either method, the continuous-time detection problem of selecting between models (a binary selection is used here for simplicity)

$$\begin{aligned}
\mathscr{M}_0 &: \quad X(t) = s^0(t) + N(t) \quad 0 \le t < T \\
\mathscr{M}_1 &: \quad X(t) = s^1(t) + N(t) \quad 0 \le t < T
\end{aligned}$$

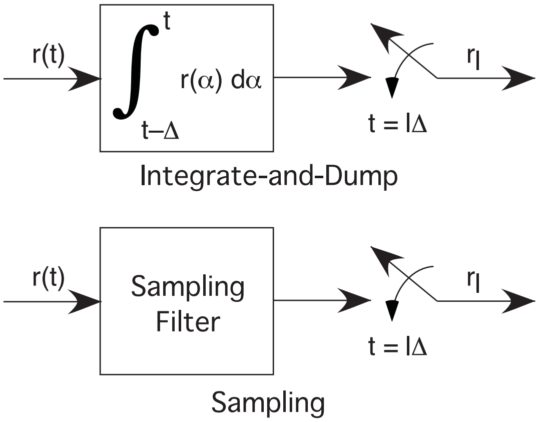


Figure 4.4: The two most common methods of converting continuous-time observations into discrete-time ones are shown. In the upper panel, the integrate-and-dump method is shown: the input is integrated over an interval of duration Δ and the result sampled. In the lower panel, the sampling method merely samples the input every Δ seconds.

where $\{s_i(t)\}$ denotes the known signal set and $N(t)$ denotes additive noise modeled as a stationary stochastic process* is converted into the discrete-time detection problem

$$\begin{array}{lll} \mathscr{M}_0: & X_l = s_l^0 + N_l & 0 \le l < L \\ \mathscr{M}_1: & X_l = s_l^1 + N_l & 0 \le l < L \end{array}$$

where the sampling interval is *always* taken to divide the observation interval T: $L = T/\Delta$. We form the discrete-time observations into a vector: $\mathbf{X} = \text{col}[X(0),\ldots,X(L-1)]$. The binary detection problem is to distinguish between two possible signals present in the noisy output waveform.

$$\begin{array}{l} \mathscr{M}_0: \mathbf{X} = \mathbf{s}_0 + \mathbf{N} \\ \mathscr{M}_1: \mathbf{X} = \mathbf{s}_1 + \mathbf{N} \end{array}$$

To apply our model evaluation results, we need the probability density of $\mathbf{X}$ under each model. As the only probabilistic component of the observations is the noise, the required density for the detection problem is given by

$$\boxed{p_{\mathbf{X}|\mathscr{M}_i}(\mathbf{X}|\mathscr{M}_i) = p_{\mathbf{N}}(\mathbf{X} - \mathbf{s}_i)}$$

and the corresponding likelihood ratio by

$$\boxed{\Lambda(\mathbf{X}) = \frac{p_{\mathbf{N}}(\mathbf{X} - \mathbf{s}_1)}{p_{\mathbf{N}}(\mathbf{X} - \mathbf{s}_0)}}.$$

Much of detection theory revolves about interpreting this likelihood ratio and deriving the detection threshold (either *threshold* or γ).

4.2.1 White Gaussian Noise

By far the easiest detection problem to solve occurs when the noise vector consists of statistically independent, identically distributed, Gaussian random variables. In this book, a "white" sequence consists of statistically independent random variables. The white sequence's mean is usually taken to be zero* and each component's

*We are *not* assuming the amplitude distribution of the noise to be Gaussian.

*The zero-mean assumption is realistic for the detection problem. If the mean were non-zero, simply subtracting it from the observed sequence results in a zero-mean noise component.

variance is σ^2. The equal-variance assumption implies the noise characteristics are unchanging throughout the entire set of observations. The probability density of the zero-mean noise vector evaluated at $\mathbf{X}-\mathbf{s}_i$ equals that of Gaussian random vector having independent components ($\mathbf{K}=\sigma^2\mathbf{I}$) with mean $\mathbf{s}_i$.

$$p_{\mathbf{N}}(\mathbf{X}-\mathbf{s}_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{L/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{X}-\mathbf{s}_i)^t(\mathbf{X}-\mathbf{s}_i)\right\}$$

The resulting detection problem is similar to the Gaussian example examined so frequently in the hypothesis testing sections, with the distinction here being a non-zero mean under both models. The logarithm of the likelihood ratio becomes

$$(\mathbf{X}-\mathbf{s}_0)^t(\mathbf{X}-\mathbf{s}_0) - (\mathbf{X}-\mathbf{s}_1)^t(\mathbf{X}-\mathbf{s}_1) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} 2\sigma^2 \ln\eta$$

and the usual simplifications yield in

$$\left(\mathbf{X}^t\mathbf{s}_1 - \frac{\mathbf{s}_1^t\mathbf{s}_1}{2}\right) - \left(\mathbf{X}^t\mathbf{s}_0 - \frac{\mathbf{s}_0^t\mathbf{s}_0}{2}\right) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln\eta\,.$$

The quantities in parentheses express the signal processing operations for each model. If more than two signals were assumed possible, quantities such as these would need to be computed for each signal and the largest selected. This decision rule is optimum for the additive, white Gaussian noise problem.

Each term in the computations for the optimum detector has a signal processing interpretation. When expanded, the term $\mathbf{s}_i^t\mathbf{s}_i$ equals $\sum_{l=0}^{L-1} s_i^2(l)$, which is the *signal energy* E_i. The remaining term—$\mathbf{X}^t\mathbf{s}_i$—is the only one involving the observations and hence constitutes the sufficient statistic $\Upsilon_i(\mathbf{X})$ for the additive white Gaussian noise detection problem.

$$\boxed{\Upsilon_i(\mathbf{X}) = \mathbf{X}^t\mathbf{s}_i}$$

An abstract, but physically relevant, interpretation of this important quantity comes from the theory of linear vector spaces. There, the quantity $\mathbf{X}^t\mathbf{s}_i$ would be termed the *dot product* between $\mathbf{X}$ and $\mathbf{s}_i$ or the *projection* of $\mathbf{X}$ onto $\mathbf{s}_i$. By employing the Schwarz inequality, the largest value of this quantity occurs when these vectors are proportional to each other. Thus, a dot product computation measures how much alike two vectors are: they are completely alike when they are parallel (proportional) and completely dissimilar when orthogonal (the dot product is zero). More precisely, the dot product removes those components from the observations which are orthogonal to the signal. The dot product thereby generalizes the familiar notion of filtering a signal contaminated by broadband noise. In filtering, the signal-to-noise ratio of a bandlimited signal can be drastically improved by lowpass filtering; the output would consist only of the signal and "in-band" noise. The dot product serves a similar role, ideally removing those "out-of-band" components (the orthogonal ones) and retaining the "in-band" ones (those parallel to the signal).

Expanding the dot product, $\mathbf{X}^t\mathbf{s}_i = \sum_{l=0}^{L-1} X(l)s_i(l)$, another signal processing interpretation emerges. The dot product now describes a finite impulse response (FIR) filtering operation evaluated at a specific index. To demonstrate this interpretation, let $h(l)$ be the unit-sample response of a linear, shift-invariant filter where $h(l)=0$ for $l<0$ and $l \geq L$. Letting $X(l)$ be the filter's input sequence, the convolution sum expresses the output.

$$X(k)\star h(k) = \sum_{l=k-(L-1)}^{k} X(l)h(k-l)\,,$$

Letting $k=L-1$, the index at which the unit-sample response's last value overlaps the input's value at the origin, we have

$$X(k)\star h(k)\big|_{k=L-1} = \sum_{l=0}^{L-1} X(l)h(L-1-l).$$

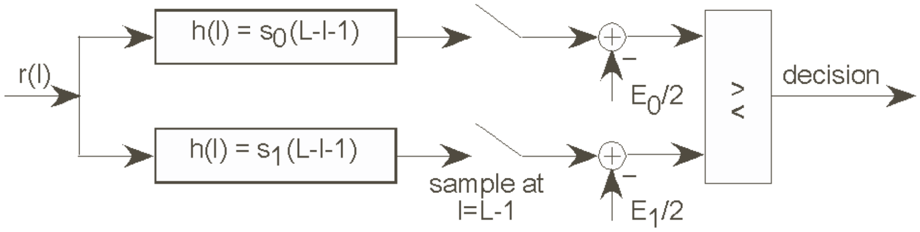


Figure 4.5: The detector for signals contained in additive, white Gaussian noise consists of a matched filter, whose output is sampled at the duration of the signal and half of the signal energy is subtracted from it. The optimum detector incorporates a matched filter for each signal compares their outputs to determine the largest.

If we set the unit-sample response equal to the index-reversed, then delayed signal $\big(h(l)=s_i(L-1-l)\big)$, we have

$$X(k)\star s_i(L-1-k)\big|_{k=L-1}=\sum_{l=0}^{L-1}X(l)s_i(l),$$

which equals the observation-dependent component of the optimal detector's sufficient statistic. Fig. 4.5 depicts these computations graphically. The sufficient statistic for the i^{th} signal is thus expressed in signal processing notation as $X(k)\star s_i(L-1-k)\big|_{k=L-1}-E_i/2$. The filtering term is called a *matched filter* because the observations are passed through a filter whose unit-sample response "matches" that of the signal being sought. We sample the matched filter's output at the precise moment when all of the observations fall within the filter's memory and then adjust this value by half the signal energy. The adjusted values for the two assumed signals are subtracted and compared to a threshold.

To compute the performance probabilities, the expressions should be simplified in the ways discussed in the hypothesis testing sections. As the energy terms are known *a priori*, they can be incorporated into the threshold with the result

$$\sum_{l=0}^{L-1}X(l)[s_1(l)-s_0(l)]\underset{\mathcal{M}_0}{\overset{\mathcal{M}_1}{\gtrless}}\sigma^2\ln\eta+\frac{E_1-E_0}{2}.$$

The left term constitutes the sufficient statistic for the binary detection problem. Because the additive noise is presumed Gaussian, the sufficient statistic is a Gaussian random variable no matter which model is assumed. Under $\mathcal{M}_i$, the specifics of this probability distribution are

$$\sum_{l=0}^{L-1}X(l)[s_1(l)-s_0(l)]\sim\mathcal{N}\Big(\sum s_i(l)[s_1(l)-s_0(l)],\sigma^2\sum[s_1(l)-s_0(l)]^2\Big).$$

The false-alarm probability is given by

$$P_F=Q\left(\frac{\sigma^2\ln\eta+(E_1-E_0)/2-\sum s_0(l)[s_1(l)-s_0(l)]}{\sigma\cdot\left\{\sum[s_1(l)-s_0(l)]^2\right\}^{1/2}}\right).$$

The signal-related terms in the numerator of this expression can be manipulated with the false-alarm probability (and the detection probability) for the optimal white Gaussian noise detector succinctly expressed by

$$\boxed{\begin{aligned}P_F&=Q\left(\frac{\ln\eta+\frac{1}{2\sigma^2}\sum[s_1(l)-s_0(l)]^2}{\frac{1}{\sigma}\left\{\sum[s_1(l)-s_0(l)]^2\right\}^{1/2}}\right)\\P_D&=Q\left(\frac{\ln\eta-\frac{1}{2\sigma^2}\sum[s_1(l)-s_0(l)]^2}{\frac{1}{\sigma}\left\{\sum[s_1(l)-s_0(l)]^2\right\}^{1/2}}\right)\end{aligned}}$$

Note that the *only* signal-related quantity affecting this performance probability (and all of the others) is the *ratio of energy in the difference signal to the noise variance*. The larger this ratio, the better (smaller) the performance probabilities become. Note that the details of the signal waveforms do not greatly affect the energy of the difference signal. For example, consider the case where the two signal energies are equal ($E_0 = E_1 = E$); the energy of the difference signal is given by $2E - 2\sum s_0(l)s_1(l)$. The largest value of this energy occurs when the signals are negatives of each other, with the difference-signal energy equaling $4E$. Thus, equal-energy but opposite-signed signals such as sine waves, square-waves, Bessel functions, etc. *all* yield exactly the same performance levels. The essential signal properties that do yield good performance values are elucidated by an alternate interpretation. The term $\sum[s_1(l) - s_0(l)]^2$ equals $||\mathbf{s}_1 - \mathbf{s}_0||^2$, the L^2 norm of the difference signal. Geometrically, the difference-signal energy is the same quantity as the square of the Euclidean distance between the two signals. In these terms, a larger distance between the two signals will mean better performance.

Example

A common detection problem in array processing is to determine whether a signal is present ($\mathscr{M}_1$) or not ($\mathscr{M}_0$) in the array output. In this case, $s_0(l) = 0$. The optimal detector relies on filtering the array output with a matched filter having an impulse response based on the assumed signal. Letting the signal under $\mathscr{M}_1$ be denoted simply by $s(l)$, the optimal detector consists of

$$X(l) \star s(L-1-l)|_{l=L-1} - E/2 \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \sigma^2 \ln \eta$$

$$\text{or} \quad X(l) \star s(L-1-l)|_{l=L-1} \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \gamma\,.$$

The false-alarm and detection probabilities are given by

$$P_F = Q\left(\frac{\gamma}{E^{1/2}/\sigma}\right) \quad P_D = Q\left(Q^{-1}(P_F) - \sqrt{\frac{E}{\sigma}}\right)\,.$$

Fig. 4.6 displays the probability of detection as a function of the signal-to-noise ratio E/σ^2 for several values of false-alarm probability. Given an estimate of the expected signal-to-noise ratio, these curves can be used to assess the trade-off between the false-alarm and detection probabilities.

The important parameter determining detector performance derived in this example is the *signal-to-noise ratio* E/σ^2: the larger it is, the smaller the false-alarm probability is (generally speaking). Signal-to-noise ratios can be measured in many different ways. For example, one measure might be the ratio of the rms signal amplitude to the rms noise amplitude. Note that the important one for the detection problem is much different. The signal portion is the *sum* of the squared signal values over the *entire* set of observed values—the signal energy; the noise portion is the variance of *each* noise component—the noise power. Thus, energy can be increased in two ways that increase the signal-to-noise ratio: the signal can be made larger *or* the observations can be extended to encompass a larger number of values.

To illustrate this point, two signals having the same energy are shown in Fig. 4.7. When these signals are shown in the presence of additive noise, the signal is visible on the left because its amplitude is larger; the one on the right is much more difficult to discern. The instantaneous signal-to-noise ratio—the ratio of signal amplitude to average noise amplitude—is the important visual cue. However, the kind of signal-to-noise ratio that determines detection performance belies the eye. The matched filter outputs have similar maximal values, indicating that total signal energy rather than amplitude determines the performance of a matched filter detector.

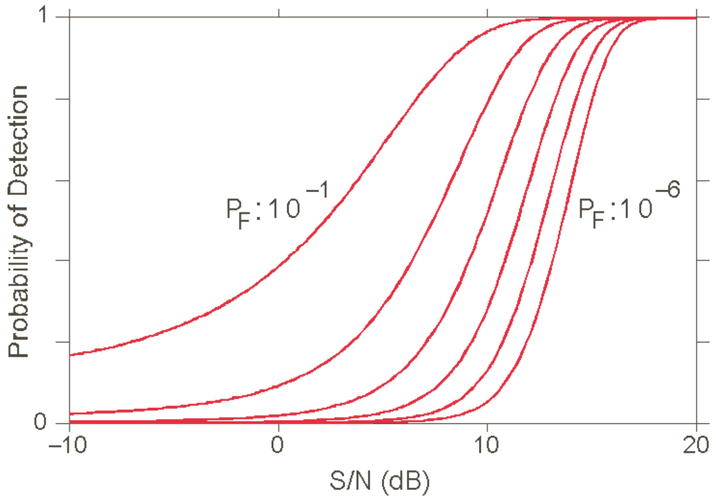


Figure 4.6: The probability of detection is plotted versus signal-to-noise ratio for various values of the false-alarm probability P_F. False-alarm probabilities range from 10^{-1} down to 10^{-6} by decades. The matched filter receiver was used since the noise is white and Gaussian. Note how the range of signal-to-noise ratios over which the detection probability changes shrinks as the false-alarm probability decreases. This effect is a consequence of the non-linear nature of the function $Q(\cdot)$.

Validity of the White Noise Assumption

The optimal detection paradigm for the additive, white Gaussian noise problem has a relatively simple solution: construct FIR filters whose unit-sample responses are related to the presumed signals and compare the filtered outputs with a threshold. We may well wonder which assumptions made in this problem are most questionable in "real-world" applications. Noise is additive in most cases. In many situations, the additive noise present in observed data is Gaussian. Because of the Central Limit Theorem, if numerous noise sources impinge on a measuring device, their superposition will be Gaussian to a great extent. As we know from the discussion in §1.4.2 {12}, glibly appealing to the Central Limit Theorem is not without hazards; the non-Gaussian detection problem will be discussed in some detail later. Interestingly, the weakest assumption is the "whiteness" of the noise. Note that the observation sequence is obtained as a result of *sampling* the sensor outputs. Assuming white noise samples does *not* mean that the continuous-time noise was white. White noise in continuous time has infinite variance and cannot be sampled; discrete-time white noise has a finite variance with a constant power spectrum. The Sampling Theorem suggests that a signal is represented accurately by its samples only if we choose a sampling frequency commensurate with the signal's bandwidth. One should note that fidelity of representation does *not* mean that the sample values are independent. In most cases, satisfying the Sampling Theorem means that the samples are correlated. As shown in §2.1.3 {20}, the correlation function of sampled noise equals samples of the original correlation function. For the sampled noise to be white, $\mathcal{E}[N(l_1T)N(l_2T)] = 0$ for $l_1 \neq l_2$: the samples of the correlation function at locations other than the origin must all be zero. While some correlation functions have this property, *many examples satisfy the sampling theorem but do not yield uncorrelated samples*. In many practical situations, *undersampling* the noise will reduce inter-sample correlation. Thus, we obtain uncorrelated samples either by deliberately undersampling, which wastes signal energy, or by imposing anti-aliasing filters that have a bandwidth larger than the signal and sampling at the signal's Nyquist rate. Since the noise power spectrum usually extends to higher frequencies than the signal, this intentional undersampling can result in larger noise variance. In either case, by trying to make the problem at hand match the solution, we are actually reducing performance! We need a *direct* approach to attacking the correlated noise issue that arises in virtually *all* sampled-data detection problems rather than trying to work around it.

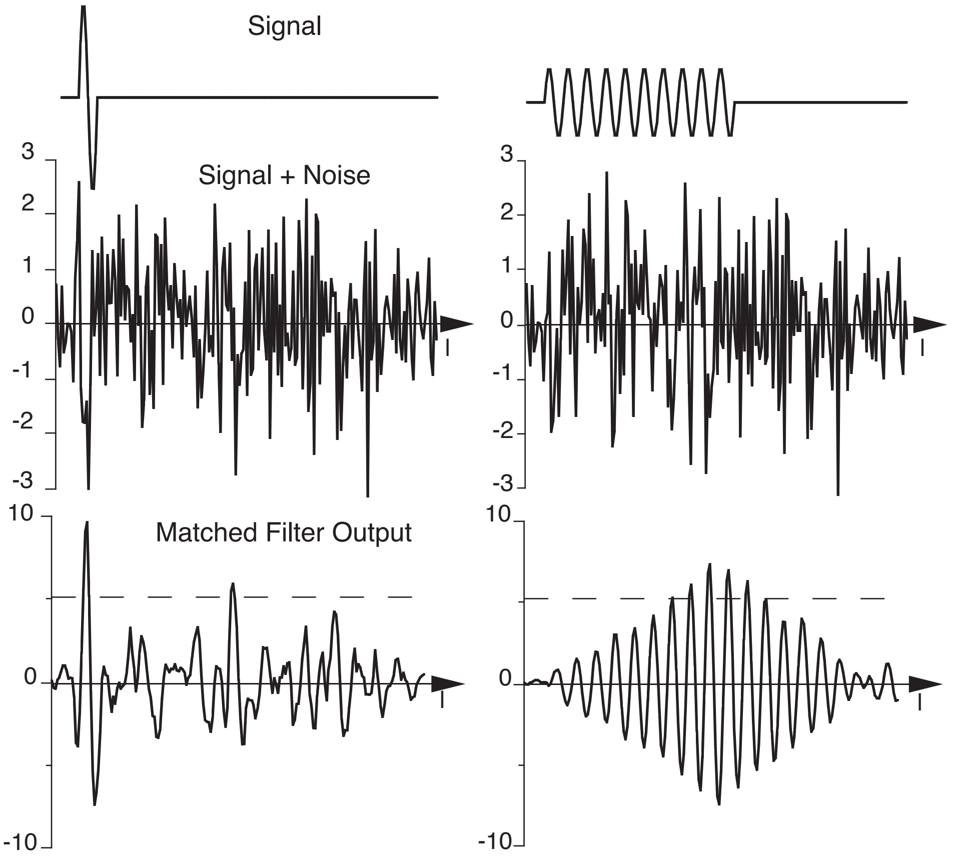


Figure 4.7: Two signals having the same energy are shown at the top of the figure. The one on the left equals one cycle of a sinusoid having ten samples/period ($\sin(2\pi f_o l)$ with $f_o = 0.1$). On the right, ten cycles of similar signal is shown, with an amplitude a factor of $\sqrt{10}$ smaller. The middle portion of the figure shows these signals with the same noise signal added; the duration of this signal is 200 samples. The lower portion depicts the outputs of matched filters for each signal. The detection threshold was set by specifying a false-alarm probability of 10^{-2}.

4.2.2 Colored Gaussian Noise

When the additive Gaussian noise in the sensors' outputs is colored (i.e., the noise values are correlated in some fashion), the linearity of beamforming algorithms means that the array processing output X also contains colored noise. The solution to the colored-noise, binary detection problem remains the likelihood ratio, but differs in the form of the *a priori* densities. The noise will again be assumed zero mean, but the noise vector has a non-trivial covariance matrix $\mathbf{K}$: $\mathbf{N} \sim \mathscr{N}(\mathbf{0}, \mathbf{K})$.

$$p_{\mathbf{N}}(\mathbf{N}) = \frac{1}{\sqrt{\det[2\pi\mathbf{K}]}} \exp\left\{-\frac{1}{2}\mathbf{N}^t\mathbf{K}^{-1}\mathbf{N}\right\}$$

In this case, the logarithm of the likelihood ratio is

$$(\mathbf{X}-\mathbf{s}_1)^t\mathbf{K}^{-1}(\mathbf{X}-\mathbf{s}_1) - (\mathbf{X}-\mathbf{s}_0)^t\mathbf{K}^{-1}(\mathbf{X}-\mathbf{s}_0) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} 2\ln\eta$$

which, after the usual simplifications, is written

$$\left[\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_1 - \frac{\mathbf{s}_1^t\mathbf{K}^{-1}\mathbf{s}_1}{2}\right] - \left[\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_0 - \frac{\mathbf{s}_0^t\mathbf{K}^{-1}\mathbf{s}_0}{2}\right] \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \ln\eta\,.$$

The sufficient statistic for the colored Gaussian noise detection problem is

$$\boxed{\Upsilon_i(\mathbf{X}) = \mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_i}\,. \tag{4.9}$$

The quantities computed for each signal have a similar, but more complicated interpretation than in the white noise case. $\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_i$ is a dot product, but with respect to the so-called *kernel* $\mathbf{K}^{-1}$. The effect of the kernel is to weight certain components more heavily than others. A positive-definite symmetric matrix (the covariance matrix is one such example) can be expressed in terms of its eigenvectors and eigenvalues.

$$\mathbf{K}^{-1} = \sum_{k=1}^{L} \frac{1}{\lambda_k}\mathbf{v}_k\mathbf{v}_k^t$$

The sufficient statistic can thus be written as the complicated summation

$$\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_i = \sum_{k=1}^{L} \frac{1}{\lambda_k}(\mathbf{X}^t\mathbf{v}_k)(\mathbf{v}_k^t\mathbf{s}_i)\,,$$

where λ_k and $\mathbf{v}_k$ denote the k^{th} eigenvalue and eigenvector of the covariance matrix $\mathbf{K}$. Each of the constituent dot products is largest when the signal and the observation vectors have strong components parallel to $\mathbf{v}_k$. However, the product of these dot products is weighted by the reciprocal of the associated eigenvalue. Thus, components in the observation vector parallel to the signal will tend to be accentuated; those components parallel to the eigenvectors having the *smaller* eigenvalues will receive greater accentuation than others. The usual notions of parallelism and orthogonality become "skewed" because of the presence of the kernel. A covariance matrix's eigenvalue has "units" of variance; these accentuated directions thus correspond to small noise variances. We can therefore view the weighted dot product as a computation that is simultaneously trying to select components in the observations similar to the signal, but concentrating on those where the noise variance is small.

The second term in the expressions constituting the optimal detector are of the form $\mathbf{s}_i^t\mathbf{K}^{-1}\mathbf{s}_i$. This quantity is a special case of the dot product just discussed. The two vectors involved in this dot product are identical; they are parallel by definition. The weighting of the signal components by the reciprocal eigenvalues remains. Recalling the units of the eigenvectors of $\mathbf{K}$, $\mathbf{s}_i^t\mathbf{K}^{-1}\mathbf{s}_i$ has the units of a signal-to-noise ratio, which is computed in a way that enhances the contribution of those signal components parallel to the "low noise" directions.

To compute the performance probabilities, we express the detection rule in terms of the sufficient statistic.

$$\mathbf{X}^t\mathbf{K}^{-1}(\mathbf{s}_1 - \mathbf{s}_0) \underset{\mathscr{M}_0}{\overset{\mathscr{M}_1}{\gtrless}} \ln\eta + \frac{1}{2}(\mathbf{s}_1^t\mathbf{K}^{-1}\mathbf{s}_1 - \mathbf{s}_0^t\mathbf{K}^{-1}\mathbf{s}_0)$$

The distribution of the sufficient statistic on the left side of this equation is Gaussian because it consists as a linear transformation of the Gaussian random vector $\mathbf{X}$. Assuming the i^{th} model to be true,

$$\mathbf{X}^t\mathbf{K}^{-1}(\mathbf{s}_1 - \mathbf{s}_0) \sim \mathscr{N}\left(\mathbf{s}_i^t\mathbf{K}^{-1}(\mathbf{s}_1 - \mathbf{s}_0), (\mathbf{s}_1 - \mathbf{s}_0)^t\mathbf{K}^{-1}(\mathbf{s}_1 - \mathbf{s}_0)\right).$$

The false-alarm probability for the optimal Gaussian colored noise detector is given by

$$\boxed{P_F = Q\left(\frac{\ln\eta + \frac{1}{2}(\mathbf{s}_1 - \mathbf{s}_0)^t\mathbf{K}^{-1}(\mathbf{s}_1 - \mathbf{s}_0)}{[(\mathbf{s}_1 - \mathbf{s}_0)^t\mathbf{K}^{-1}(\mathbf{s}_1 - \mathbf{s}_0)]^{1/2}}\right)}\,. \tag{4.10}$$

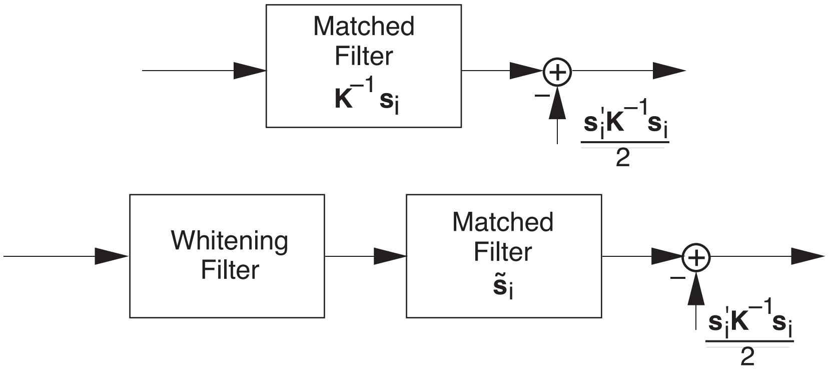

Matched Filter $K^{-1}s_i$
−
$\frac{s_i' K^{-1} s_i}{2}$
Whitening Filter
Matched Filter $\tilde{s}_i$
−
$\frac{s_i' K^{-1} s_i}{2}$

Figure 4.8: These diagrams depict the signal processing operations involved in the optimum detector when the additive noise is not white. The upper diagram shows a matched filter whose unit-sample response depends both on the signal and the noise characteristics. The lower diagram is often termed the whitening filter structure, where the noise components of the observed data are first whitened, then passed through a matched filter whose unit-sample response is related to the "whitened" signal.

As in the white noise case, the important signal-related quantity in this expression is the signal-to-noise ratio of the difference signal. The distance interpretation of this quantity remains, but the distance is now warped by the kernel's presence in the dot product.

The sufficient statistic computed for each signal can be given two signal processing interpretations in the colored noise case. Both of these rest on considering the quantity $\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_i$ as a simple dot product, but with different ideas on grouping terms. The simplest is to group the kernel with the signal so that the sufficient statistic is the dot product between the observations and a *modified* version of the signal $\tilde{\mathbf{s}}_i = \mathbf{K}^{-1}\mathbf{s}_i$. This modified signal thus becomes the equivalent to the unit-sample response of the matched filter. In this form, the observed data are unaltered and passed through a matched filter whose unit-sample response depends on both the signal and the noise characteristics. The size of the noise covariance matrix, equal to the number of observations used by the detector, is usually large: hundreds if not thousands of samples are possible. Thus, computation of the inverse of the noise covariance matrix becomes an issue. This problem needs to be solved only once if the noise characteristics are static; the inverse can be precomputed on a general purpose computer using well-established numerical algorithms. The signal-to-noise ratio term of the sufficient statistic is the dot product of the signal with the modified signal $\tilde{\mathbf{s}}_i$. This view of the receiver structure is shown in Fig. 4.8.

A second and more theoretically powerful view of the computations involved in the colored noise detector emerges when we *factor* covariance matrix. The *Cholesky factorization* of a positive-definite, symmetric matrix (such as a covariance matrix or its inverse) has the form $\mathbf{K} = \mathbf{LDL}^t$. With this factorization, the sufficient statistic can be written as

$$\mathbf{X}^t\mathbf{K}^{-1}\mathbf{s}_i = \left(\mathbf{D}^{-1/2}\mathbf{L}^{-1}\mathbf{X}\right)^t \left(\mathbf{D}^{-1/2}\mathbf{L}^{-1}\mathbf{s}_i\right).$$

The components of the dot product are multiplied by the same matrix ($\mathbf{D}^{-1/2}\mathbf{L}^{-1}$), which is lower-triangular. *If* this matrix were also Toeplitz, the product of this kind between a Toeplitz matrix and a vector would be equivalent to the convolution of the components of the vector with the first column of the matrix. If the matrix is not Toeplitz (which, inconveniently, is the typical case), a convolution also results, but with a unit-sample response that varies with the index of the output—a time-varying, linear filtering operation. The variation of the unit-sample response corresponds to the different rows of the matrix $\mathbf{D}^{-1/2}\mathbf{L}^{-1}$ running *backwards* from the main-diagonal entry. What is the physical interpretation of the action of this filter? The covariance of the random vector $\mathbf{x} = \mathbf{AX}$ is given by $\mathbf{K}_x = \mathbf{AK}_X\mathbf{A}^t$. Applying this result to the current situation, we set $\mathbf{A} = \mathbf{D}^{-1/2}\mathbf{L}^{-1}$ and $\mathbf{K}_X = \mathbf{K} = \mathbf{LDL}^t$ with the result that the covariance matrix $\mathbf{K}_x$ is the identity matrix! Thus, the matrix $\mathbf{D}^{-1/2}\mathbf{L}^{-1}$ corresponds to a (possibly time-varying) *whitening filter*: we have converted the colored-noise component of the observed data to white noise! As the filter is always linear, the Gaussian observation noise remains Gaussian at the output. Thus, the colored noise problem is converted into a simpler one with the whitening filter: the whitened observations are first match-filtered with the "whitened" signal $\mathbf{s}_i^+ = \mathbf{D}^{-1/2}\mathbf{L}^{-1}\mathbf{s}_i$ (whitened with respect to noise characteristics only) then half the energy of the whitened signal is subtracted (Fig. 4.8).

Example

To demonstrate the interpretation of the Cholesky factorization of the covariance matrix as a time-varying whitening filter, consider the covariance matrix

$$\mathbf{K} = \begin{bmatrix} 1 & a & a^2 & a^3 \\ a & 1 & a & a^2 \\ a^2 & a & 1 & a \\ a^3 & a^2 & a & 1 \end{bmatrix}.$$

This covariance matrix indicates that the noise was produced by passing white Gaussian noise through a first-order filter having coefficient a: $N(l) = aN(l-1) + \left(1-a^2\right)^{1/2} w(l)$, where $w(l)$ is unit-variance white noise. Thus, we would expect that if a whitening filter emerged from the matrix manipulations (derived just below), it would be a first-order FIR filter having an unit-sample response proportional to

$$h(l) = \begin{cases} 1, & l=0 \\ -a, & l=1 \\ 0, & \text{otherwise}. \end{cases}$$

Simple arithmetic calculations of the Cholesky decomposition suffice to show that the matrices **L** and **D** are given by

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ a & 1 & 0 & 0 \\ a^2 & a & 1 & 0 \\ a^3 & a^2 & a & 1 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-a^2 & 0 & 0 \\ 0 & 0 & 1-a^2 & 0 \\ 0 & 0 & 0 & 1-a^2 \end{bmatrix}$$

and that their inverses are

$$\mathbf{L}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -a & 1 & 0 & 0 \\ 0 & -a & 1 & 0 \\ 0 & 0 & -a & 1 \end{bmatrix} \quad \mathbf{D}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{1-a^2} & 0 & 0 \\ 0 & 0 & \frac{1}{1-a^2} & 0 \\ 0 & 0 & 0 & \frac{1}{1-a^2} \end{bmatrix}.$$

Because **D** is diagonal, the matrix $\mathbf{D}^{-1/2}$ equals the term-by-term square root of the inverse of **D**. The product of interest here is therefore given by

$$\mathbf{D}^{-1/2}\mathbf{L}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{-a}{\sqrt{1-a^2}} & \frac{1}{\sqrt{1-a^2}} & 0 & 0 \\ 0 & \frac{-a}{\sqrt{1-a^2}} & \frac{1}{\sqrt{1-a^2}} & 0 \\ 0 & 0 & \frac{-a}{\sqrt{1-a^2}} & \frac{1}{\sqrt{1-a^2}} \end{bmatrix}.$$

Let $\tilde{\mathbf{X}}$ express the product $\mathbf{D}^{-1/2}\mathbf{L}^{-1}\mathbf{X}$. This vector's elements are given by

$$\tilde{X}_0 = X_0, \quad \tilde{X}_1 = \frac{1}{\sqrt{1-a^2}}[X_1 - aX_0], \quad etc.$$

Thus, the expected FIR whitening filter emerges after the first term. The expression could *not* be of this form as no observations were assumed to precede X_0. This edge effect is the source of the time-varying aspect of the whitening filter. If the system modeling the noise generation process has only poles, this whitening filter will always stabilize—not vary with time—once sufficient data are present within the memory of the FIR inverse filter. In contrast, the presence of zeros in the generation system would imply an IIR whitening filter. With finite data, the unit-sample response would then change on each output sample.

4.3 Continuous-time detection

In previous sections, we used a sampling approach to detect which of several signals was present in additive noise. While less general, an alternate approach can be used in situations where the additive noise is Gaussian. In such cases, the problem can be solved entirely in continuous-time without requiring sampling. This approach relies on the *Karhunen-Loève expansion*, which results in a representation of the received process $X(t)$. In general, this representation is an infinite dimensional vector; the critical result of continuous-time detection is that a finite-dimensional representation can always be found so that hypothesis testing results can be applied.

4.3.1 Matched Filter Receiver for White Gaussian Noise Channels

The received signal $X(t)$ is assumed to have one of K forms

$$X(t) = s_i(t) + N(t)\,,\ i = 0,\dots,K-1\,, 0 \le t < T$$

where the $\{s_i(t)\}$ comprise the signal set. $N(t)$ is usually assumed to be statistically independent of the transmitted signal and a white, Gaussian process having spectral height $N_0/2$. We represent the received signal with a Karhunen-Loève expansion.

$$X(t) = \sum_{j=1}^{\infty} X_j \phi_j(t) = \sum_{j=1}^{\infty} (s_{ij} + N_j)\phi_j(t)$$

where $\{s_{ij}\}$ and $\{N_j\}$ are the representations of the signal $s_i(t)$ and the noise $N(t)$, respectively. To have a Karhunen-Loève expansion, it suffices to choose $\{\phi_j(t)\}$ so that the $\{N_j\}$ are pairwise uncorrelated. As $N(t)$ is white, we may choose *any* $\{\phi_j(t)\}$ we want! In particular, choose $\{\phi_j(t)\}$ to be the set of functions which yield a finite-dimensional representation for the signals $s_i(t)$. A complete, but not necessarily orthonormal, set of functions that does this is

$$s_0(t), \dots, s_{K-1}(t), \psi_0(t), \psi_1(t), \dots$$

where $\{\psi_j(t)\}$ denotes any complete set of functions. We form the set $\{\phi_j(t)\}$ by applying the *Gram-Schmidt procedure* to this set. With this basis, $s_{ij} = 0\,,\ j \ge K$. In this case, the representation of $X(t)$ becomes

$$X_j = \begin{cases} s_{ij} + N_j & j = 0,\dots,K-1 \\ N_j & j \ge K \end{cases}$$

so that we may write the model evaluation problem we are attempting to solve as

$$\mathcal{M}_0:\ X(t) = (s_{00} + N_0)\phi_0(t) + \cdots + (s_{0,K-1} + N_{K-1})\phi_{K-1}(t) + \sum_{j \ge K} N_j \phi_j(t)$$
$$\mathcal{M}_1:\ X(t) = (s_{10} + N_0)\phi_0(t) + \cdots + (s_{1,K-1} + N_{K-1})\phi_{K-1}(t) + \sum_{j \ge K} N_j \phi_j(t)$$

We make two observations:

- We can consider the model evaluation problem that operates on the *representation* of the received signal rather than the signal itself. Recall that using the representation is equivalent to using the original process. We have thus created an equivalent model evaluation problem. For the binary signal set case,

$$\begin{aligned}\mathscr{M}_0 &: \mathbf{X} = \mathbf{s}_0 + \mathbf{N}\\ \mathscr{M}_1 &: \mathbf{X} = \mathbf{s}_1 + \mathbf{N}\end{aligned}$$

  where $\mathbf{N}$ contains statistically independent Gaussian components, each of which has variance $N_0/2$.
- Note that components are statistically independent of each other and that, for $j \geq K$, the representation contains *no* signal-related information. Because these components are extraneous and will not contribute to improved performance, we can reduce the dimension of the problem to no more than K by ignoring these components. By rejecting these noise-only components, we are effectively filtering out "out-of-band" noise, retaining those components related to the signals. Using eigenfunctions related to the signals defines *signal space*, allowing us to ideally reject pure-noise components.

As a consequence of these observations, we have a model evaluation problem of the form

$$\mathbf{X} = \begin{bmatrix} X_0 \\ \vdots \\ X_{K-1} \end{bmatrix} = \begin{bmatrix} s_{i,0} \\ \vdots \\ s_{i,K-1} \end{bmatrix} + \begin{bmatrix} N_0 \\ \vdots \\ N_{K-1} \end{bmatrix}$$

We know how to solve this problem; we compute

$$\Upsilon_i(\mathbf{X}) = \frac{N_0}{2}\ln\pi_i + \langle \mathbf{s}_i, \mathbf{X}\rangle - \frac{\|\mathbf{s}_i\|^2}{2}\ , i = 0, \ldots, K-1$$

and choose the largest. The components of the signal and received vectors are given by

$$s_{ij} = \int_0^T s_i(t)\phi_j(t)\,dt \qquad X_j = \int_0^T X(t)\phi_j(t)\,dt$$

Because of Parseval's Theorem, the inner product between representations equals the time-domain inner product between the represented signals.

$$\langle \mathbf{s}_i, \mathbf{X}\rangle = \int_0^T s_i(t)X(t)\,dt$$

Furthermore, $\|\mathbf{s}_i\|^2 = \int_0^T s_i^2(t)\,dt = E_i$, the energy in the i^{th} signal. Thus, the sufficient statistic for the optimal detector has a closed form time-domain expression.

$$\boxed{\Upsilon_i(X) = \frac{N_0}{2}\ln\pi_i + \int_0^T s_i(t)X(t)\,dt - \frac{E_i}{2}}$$

This form of the minimum probability of error receiver is termed a *correlation receiver*; see Fig. 4.9. Each transmitted signal and the received signal are correlated to obtain the sufficient statistic. These operations project the received signal onto signal space.

An alternate structure which computes the same quantities can be derived by noting that if $f(t)$ and $g(t)$ are nonzero only over $[0,T]$, the inner product (correlation) operation can be written as a convolution followed by a sampler.

$$\int_0^T f(t)g(t)\,dt = f(t) \star g(T-t)\Big|_{t=T}$$

Consequently, we can restructure the "correlation" operation as a filtering-and-sampling operation. The impulse responses of the linear filters are time-reversed, delayed versions of the signals in the signal set. This

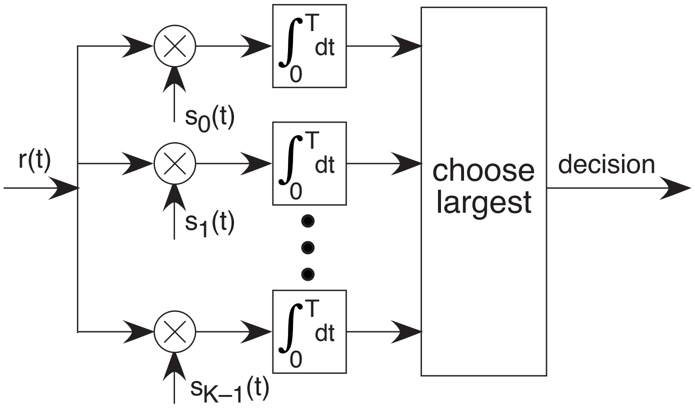


Figure 4.9: Correlation receiver structure for the optimum detector. When unequally likely and/or unequal-energy signals are used, the correction term $N_0/2\ln\pi_i - E_i/2$ must be added to each integrator's output.

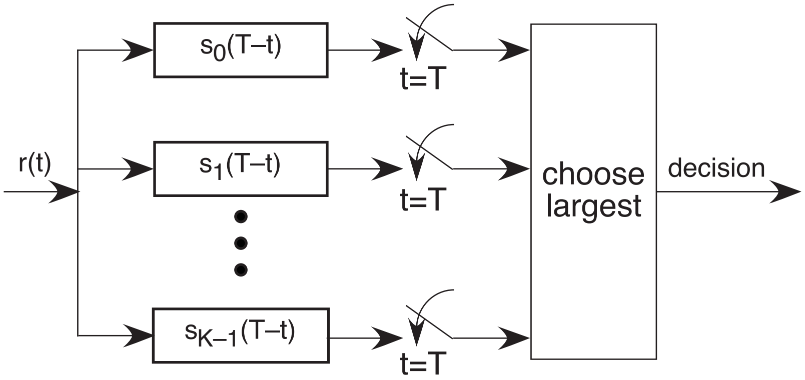


Figure 4.10: Matched filter receiver structure for the optimum detector. When unequally likely, unequal signals are used, the correction term $N_0/2\ln\pi_i - E_i/2$ must be added to each sampler's output.

structure for the minimum probability of error receiver is known as the *matched-filter receiver*; see Fig. 4.10. Each type of receiver has the same performance; however, the matched filter receiver is usually easier to construct because the correlation receiver requires an analog multiplier.

As we know, receiver performance is judged by the probability of error, which, for equally likely signals in a binary signal set, is given by

$$\boxed{P_e = Q\left(\frac{||\mathbf{s}_0 - \mathbf{s}_1||}{2\sqrt{N_0/2}}\right)}\,.$$

The computation of the probability of error and the dimensionality of the problem can be assessed by considering *signal space*: The representation of the signals with respect to a basis. The number of basis elements required to represent the signal set defines dimensionality. The geometric configuration of the signals in this space is known as the *signal constellation*. Once this constellation is found, computing intersignal distances is easy.

4.3.2 Binary Signaling Schemes

The following series of examples are important as they constitute the most popular signaling schemes in binary digital communication. For all of these examples, the elements of each signal set are assumed to be equally likely. Under this assumption, the $N_0/2\ln\pi_i$ term in the expression for $\Upsilon_i(\mathbf{X})$ cancels with the result that the computations simplify to

$$\Upsilon_i(\mathbf{X}) = \langle \mathbf{X}, \mathbf{s}_i \rangle - \frac{\|\mathbf{s}_i\|^2}{2} \quad \text{for all } i.$$

Note especially that under these conditions, the optimum receiver does not require knowledge of the spectral height $N_0/2$ of the channel noise, an important simplification in practice.

Example

Let the binary signal set be

$$s_0(t) = 0\,, \qquad s_1(t) = \sqrt{\frac{E}{T}} \quad 0 \le t < T$$

The receiver is a single correlator, with the output compared to the threshold $E/2$. The distance between the signals is easily seen to be $\|s_0 - s_1\| = \sqrt{E}$. Consequently, the probability of error which results from employing this signal set equals $P_e = Q\big(\sqrt{E/(2N_0)}\big)$. This signaling scheme is termed *amplitude-shift keying* (ASK) or *on-off keying* (OOK).

Example

Let the binary signal set be

$$s_0(t) = \begin{cases} \sqrt{E/T}, & 0 \le t < T/2 \\ -\sqrt{E/T}, & T/2 \le t < T \end{cases} \qquad s_1(t) = \sqrt{E/T}\ , 0 \le t < T$$

When these signals are equally likely to be sent, the sufficient statistic for this problem becomes $\Upsilon_i(X) = \langle X, s_i \rangle$. Note that the energy term $\|s_i\|^2/2$ does not occur: For *any* signal set containing equal-energy components, this term is common and need not be computed. Consequently, the receiver for signal sets having this property need not know the energy of the received signals. In practical applications, the energy of the signal portion of the received waveform may not be known precisely; for example, the physical distance between the transmitter and the receiver, which determines how much the signal is attenuated, may be unknown. A signal set which does require knowledge of the received signal energy is shown in the first example (ASK).

From the signal constellation, the distance between the signals is $\|s_1 - s_2\| = \sqrt{2E}$, resulting in a probability of error equal to

$$P_e = Q\left(\sqrt{\frac{E}{N_0}}\right).$$

This particular example has no specific name; however, note that $\langle s_0, s_1 \rangle = 0$, meaning that the signals are orthogonal to each other. Such signal sets are said to be *orthogonal* signal sets.

Example

Let the signal set be defined as

$$s_0(t) = \sqrt{\frac{E}{T}} \qquad s_1(t) = -\sqrt{\frac{E}{T}} \quad 0 \le t \le T\,.$$

Note that this signal set is another example of one having equal-energy components; therefore, the receiver need not contain information concerning the energy of the received signals. The distance between the signals is $||s_1 - s_2|| = 2\sqrt{E}$ so that

$$P_e = Q\left(\sqrt{\frac{2E}{N_0}}\right) .$$

This signal set is termed an *antipodal* (opposite-signed) signal set. If the energy of each component of a signal set is constrained to be less than a given value, the signal set having the largest distance between its components is the antipodal signal set.

A greater distance between the components of the signal set implies a better performance (*i.e.*, smaller P_e) for the same signal energy. Note that these probabilities of error are monotonic functions of the ratio of signal energy to channel-noise spectral height. In designing a digital communications system on the basis of performance only, maximum performance is obtained by increasing signal energy and choosing the "best" signal set: the antipodal signal set. Furthermore, note that performance does not depend on the detailed waveforms of the signals. *Signal sets having the same signal constellation have the same performance.*

The previous examples are in the class of "baseband" signal sets: The spectra of the signals is concentrated at low frequencies. Modulated signal sets, those having their spectra concentrated at high frequencies, can be analyzed in a similar fashion. Note that since the following examples have constellations identical with their baseband counterparts, their performances are also the same. The signal set consisting of

$$s_0(t) = 0 \qquad s_1(t) = \sqrt{\frac{2E}{T}}\sin 2\pi f_0 t \ , 0 \le t < T$$

(where $f_0 T$ is an integer) is an example of a modulated ASK signal set. An orthogonal signal set is exemplified by *frequency-shift keying* (FSK):

$$s_0(t) = \sqrt{\frac{2E}{T}}\sin 2\pi f_0 t \qquad s_1(t) = \sqrt{\frac{2E}{T}}\sin 2\pi f_1 t \ , 0 \le t < T$$

where $f_0 T$ and $f_1 T$ are distinct integers. Finally, *phase-shift keying* (PSK) corresponds to an antipodal signal set.

$$s_0(t) = \sqrt{\frac{2E}{T}}\sin 2\pi f_0 t \qquad s_1(t) = -\sqrt{\frac{2E}{T}}\sin 2\pi f_0 t \ , 0 \le t < T$$

4.3.3 *K*-ary Signal Sets

To generalize these results to *K*-ary signal sets is obvious. The optimum receiver computes

$$\Upsilon_i(X) = N_0/2 \ln \pi_i + \langle s_i, X\rangle - \frac{||s_i||^2}{2} \ , i = 0, \ldots, K-1 \, .$$

for each i and chooses the largest. Conceptually, these are no more complicated than binary signal sets. The minimum probability of error receiver remains a matched filter and has a similar structure to those shown previously. However, the computation of the probability of error may not be simple.

Problems

4.1 Consider the following two-model evaluation problem [40: Prob. 2.2.1].

$$\mathcal{M}_0: X = N$$
$$\mathcal{M}_1: X = s + N,$$

where s and N are statistically independent, positively valued, random variables having the densities

$$p_s(s) = ae^{-as} \quad \text{and} \quad p_N(N) = be^{-bN}.$$

(a) Prove that the likelihood ratio test reduces to

$$X \underset{\mathcal{M}_0}{\overset{\mathcal{M}_1}{\gtrless}} \gamma$$

(b) Find γ for the minimum probability of error test as a function of the *a priori* probabilities.

(c) Now assume that we need a Neyman-Pearson test. Find γ as a function P_F, the false-alarm probability.

4.2 Two models describe different equi-variance statistical models for the observations [40: Prob. 2.2.11].

$$\mathcal{M}_0: p_X(X) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|X|}$$
$$\mathcal{M}_1: p_X(X) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}X^2}$$

(a) Find the likelihood ratio.

(b) Compute the decision regions for various values of the threshold in the likelihood ratio test.

(c) Assuming these two densities are equally likely, find the probability of making an error in distinguishing between them.

4.3 **Cautious Decision Making**

Wanting to be very cautious before making a decision, an ELEC 531 student wants to explicitly allow no decision to be made if the data don't warrant a "firm" decision. The Bayes cost criterion is used to derive the cautious detector. Let the cost of a wrong decision be $C_{\text{wd}} > 0$, the cost of making no decision be $C_? > 0$ and the cost of making a correct decision be zero. Two signal models are possible and they have *a priori* probabilities π_0 and π_1.

(a) Derive the detector that minimizes the average Bayes cost, showing that it is a likelihood ratio detector.

(b) For what choices of C_{wd} and $C_?$ is the decision rule well-defined?

(c) Let the observations consist of L samples, with model 0 corresponding to white Gaussian noise and model 1 corresponding to a known signal in additive white Gaussian noise. Find the decision rule in terms the sufficient statistic.

4.4 A hypothesis testing criterion radically different from those discussed in §4.1.1 and §4.1.2 is *minimum equivocation*. In this information theoretic approach, the two-model testing problem is modeled as a digital channel (Fig. 4.11). The channel's inputs, generically represented by the $\mathbf{x}$, are the models and the channel's outputs, denoted by $\mathbf{y}$, are the decisions.

The quality of such information theoretic channels is quantified by the *mutual information* $I(\mathbf{x};\mathbf{y})$ defined to be difference between the entropy of the inputs and the *equivocation* [5: §2.3, 2.4].

$$I(\mathbf{x};\mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y})$$
$$H(\mathbf{x}) = -\sum_i P(x_i)\log P(x_i)$$
$$H(\mathbf{x}|\mathbf{y}) = -\sum_{i,j} P(x_i,y_j)\log\frac{P(x_i,y_j)}{P(y_j)}$$

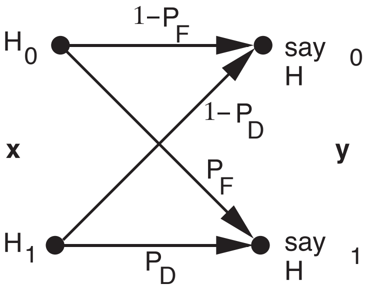


Figure 4.11: The two-model testing problem can be abstractly described as a communication channel where the inputs are the models and the outputs are the decisions. The transition probabilities are related to the false-alarm (P_F) and detection (P_D) probabilities.

Here, $P(x_i)$ denotes the *a priori* probabilities, $P(y_j)$ the output probabilities, and $P(x_i,y_j)$ the joint probability of input x_i resulting in output y_j. For example, $P(x_0,y_0) = P(x_0)(1-P_F)$ and $P(y_0) = P(x_0)(1-P_F)+P(x_1)(1-P_D)$. For a fixed set of *a priori* probabilities, show that the decision rule that maximizes the mutual information is the likelihood ratio test. What is the threshold when this criterion is employed?

Note: This problem is relatively difficult. The key to its solution is to exploit the concavity of the entropy function.

4.5 Detection and Estimation Working Together
Detectors are frequently used to determine if a signal is even present before applying estimators to tease out the signal. The same, but unknown signal having duration L may or may not be present in additive white Gaussian noise during the i^{th} observation interval, $i = 1,\dots$.

$$\mathcal{M}_0\colon \mathbf{R}_i = \mathbf{N}_i$$
$$\mathcal{M}_1\colon \mathbf{R}_i = \mathbf{s}+\mathbf{N}_i$$

π_0, π_1 denote the *a priori* probabilities. Once M intervals have been determined by the front-end detector to contain a signal, we apply the maximum likelihood estimator to measure the signal.

(a) What is the maximum likelihood signal estimate?

(b) What is the front-end detector's algorithm?

(c) Even if we use an optimal front-end detector, it can make errors, saying a signal is present when it isn't. What is the mean-squared error of the combined detector-estimator in terms of the detector's detection and false-alarm probabilities?

4.6 Non-Gaussian statistical models sometimes yield surprising results in comparison to Gaussian ones. Consider the following hypothesis testing problem where the observations have a Laplacian probability distribution.

$$\mathcal{M}_0\colon p_X(X) = \frac{1}{2}e^{-|X+m|}$$
$$\mathcal{M}_1\colon p_X(X) = \frac{1}{2}e^{-|X-m|}$$

(a) Find the sufficient statistic for the optimal decision rule.

(b) What decision rule guarantees that the miss probability will be less than 0.1?

4.7 Developing a Neyman-Pearson decision rule for more than two models has not been detailed because a mathematical quandry arises. The issue is that we have several performance probabilities we want to optimize. In essence, we are optimizing a vector of performance probabilities, which requires us to specify a norm. Many norms can be chosen; we select one in this problem.

Assume K distinct models are required to account for the observations. We seek to maximize the *sum* of the probabilities of correctly announcing $\mathcal{M}_i$, $i = 1,\dots,K$. This choice amounts to maximizing the L^1 norm of the detection probabilities. We constrain the probability of announcing $\mathcal{M}_i$ when model $\mathcal{M}_0$ was indeed true to not exceed a specified value.

(a) Formulate the optimization problem that simultaneously maximizes $\sum_{i'} \Pr[\text{say } \mathscr{M}_i | \mathscr{M}_i]$ under the constraint $\Pr[\text{say } \mathscr{M}_i | \mathscr{M}_0] \leq \alpha_i$. Find the solution using Lagrange multipliers.

(b) Can you find the Lagrange multipliers?

(c) Can your solution can be expressed as choosing the largest of the sufficient statistics $\Upsilon_i(\mathbf{X}) + C_i$?

4.8 Pattern recognition relies heavily on ideas derived from the principles of statistical model testing. Measurements are made of a test object and these are compared with those of "standard" objects to determine which the test object most closely resembles. Assume that the measurement vector $\mathbf{X}$ is jointly Gaussian with mean $\mathbf{m}_i$ $(i = 1, \dots, K)$ and covariance matrix $\sigma^2 \mathbf{I}$ (i.e., statistically independent components). Thus, there are K possible objects, each having an "ideal" measurement vector $\mathbf{m}_i$ and probability π_i of being present.

(a) How is the minimum probability of error choice of object determined from the observation of $\mathbf{X}$?

(b) Assuming that only two equally likely objects are possible $(K = 2)$, what is the probability of error of your decision rule?

(c) The expense of making measurements is always a practical consideration. Assuming each measurement costs the same to perform, how would you determine the effectiveness of a measurement vector's component?

4.9 Define y to be

$$y = \sum_{k=0}^{L} x_k$$

where the x_k are statistically independent random variables, each having a Gaussian density $\mathscr{N}(0, \sigma^2)$. The number L of variables in the sum is a random variable with a Poisson distribution.

$$\Pr[L = l] = \frac{\lambda^l}{l!} e^{-\lambda}, \quad l = 0, 1, \dots$$

Based upon the observation of y, we want to decide whether $L \leq 1$ or $L > 1$. Write an expression for the minimum P_e likelihood ratio test.

4.10 One observation of the random variable X is obtained. This random variable is either uniformly distributed between -1 and $+1$ or expressed as the sum of statistically independent random variables, each of which is also uniformly distributed between -1 and $+1$.

(a) Suppose there are two terms in the aforementioned sum. Assuming that the two models are equally likely, find the minimum probability of error decision rule.

(b) Compute the resulting probability of error of your decision rule.

(c) Show that the decision rule found in part (a) applies no matter how many terms are assumed present in the sum.

4.11 The observed random variable X has a Gaussian density on each of five models.

$$p_{X|\mathscr{M}_i}(X|\mathscr{M}_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(X - m_i)^2}{2\sigma^2} \right\}, \quad i = 1, 2, \dots, 5$$

where $m_1 = -2m$, $m_2 = -m$, $m_3 = 0$, $m_4 = +m$, and $m_5 = +2m$. The models are equally likely and the criterion of the test is to minimize P_e.

(a) Draw the decision regions on the X-axis.

(b) Compute the probability of error.

(c) Let $\sigma = 1$. Sketch accurately P_e as a function of m.

4.12 The goal is to choose which of the following four models is true upon the reception of the three-dimensional vector $\mathbf{X}$ [40: Prob. 2.6.6].

$$\begin{aligned}\mathscr{M}_0 &: \mathbf{X} = \mathbf{m}_0 + \mathbf{N}\\ \mathscr{M}_1 &: \mathbf{X} = \mathbf{m}_1 + \mathbf{N}\\ \mathscr{M}_2 &: \mathbf{X} = \mathbf{m}_2 + \mathbf{N}\\ \mathscr{M}_3 &: \mathbf{X} = \mathbf{m}_3 + \mathbf{N}\end{aligned}$$

where

$$\mathbf{m}_0 = \begin{bmatrix} a \\ 0 \\ b \end{bmatrix}, \quad \mathbf{m}_1 = \begin{bmatrix} 0 \\ a \\ b \end{bmatrix}, \quad \mathbf{m}_2 = \begin{bmatrix} -a \\ 0 \\ b \end{bmatrix}, \quad \mathbf{m}_3 = \begin{bmatrix} 0 \\ -a \\ b \end{bmatrix}.$$

The noise vector $\mathbf{N}$ is a Gaussian random vector having statistically independent, identically distributed components, each of which has zero mean and variance σ^2. We have L independent observations of the received vector $\mathbf{X}$.

(a) Assuming equally likely models, find the minimum P_e decision rule.

(b) Calculate the resulting error probability.

(c) Show that neither the decision rule nor the probability of error do not depend on b. Intuitively, why is this fact true?

4.13 Discrete Estimation

Estimation theory focuses on deriving effective (minimum error) techniques for determining the value of *continuous-valued* quantities. When the quantity is discrete-valued (integer-valued, for example), the usual approaches don't work well since they usually produce estimates not in the set of known values. This problem explores applying decision-theoretic approaches to yield a framework for discrete estimation.

Let's explore a specific example. Let a sequence of statistically independent, identically distributed observations be Gaussian having mean m and variance σ^2. The mean m can only assume the values $-1, 0$, and 1, and these are equally likely. The mean, whatever its value, is constant throughout the L observations.

(a) What is the minimum probability of error decision rule? What is the resulting probability of error?

(b) What is the MAP estimate of the mean?

(c) The problem with using the detection approach of part (a) is that the probability of error is not a standard error metric in estimation theory. Suppose we want to find the minimum mean-squared error, discrete-valued estimate. Show that by defining an appropriate Bayes cost function, you can create a detection problem that minimizes the mean-squared error. What is the Bayes cost function that works?

(d) Find the minimum mean-squared error estimate using the minimum Bayes cost detector.

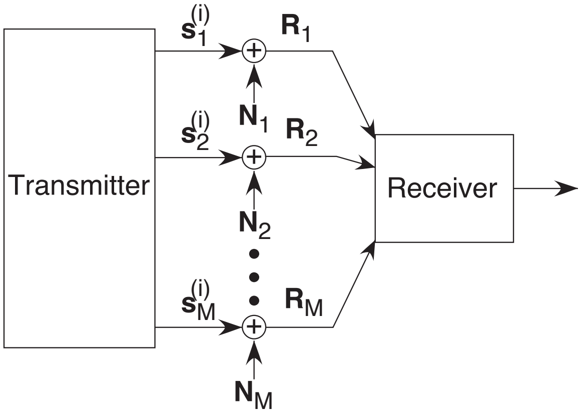
(e) What is the resulting mean-squared error?

4.14 Diversity Communication

In diversity signaling, one of two equally likely signal groups is transmitted, with each member $\mathbf{s}_m$ of the group $\{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_M\}$ sent through one of M parallel channels simultaneously. The receiver has access to all channels and can use them to make a decision as to which signal group was transmitted. The received vector (dimension L) emerging from the m^{th} channel has the form

$$\mathbf{X}_m = \mathbf{s}_m^{(i)} + \mathbf{N}_m, \quad i = 0, 1$$

The noise is colored, Gaussian, and independent from channel to channel. The statistical properties of the noise are known and they are the same in each channel.



(a) What is the optimal diversity receiver?

In an endeavor to make communications secure, bright engineers decide to modify the usual diversity signaling scenario. Rather than send signals down each channel, only one channel will be used for each transmission, with the chosen channel selected randomly from transmission to transmission. The identity of the used channel is kept secret, even from the receiver. In essence, each group now mostly consists of zero-valued signals save for the one $\mathbf{s}_m$ the transmitter chose. The receiver has access to all channels and must determine which of the signal groups was sent. In this scenario, the channel actually used contains no information about the signal group.

(b) How would the receiver determine which channel was used?

(c) What is the optimal decision rule?

4.15 To gain some appreciation of some of the issues in implementing a detector, this problem asks you to program (preferably in MATLAB) a simple detector and numerically compare its performance with theoretical predictions. Let the observations consist of a signal contained in additive Gaussian white noise.

$$\begin{aligned}\mathscr{M}_0 &: X(l) = N(l), l = 0,\dots,L-1\\ \mathscr{M}_1 &: X = A\sin(2\pi l/L) + N(l), l = 0,\dots,L-1\end{aligned}$$

The variance of each noise value equals σ^2.

(a) What is the theoretical false-alarm probability of the minimum P_e detector when the hypotheses are equally likely?

(b) Write a MATLAB program that estimates the false-alarm probability. How many simulation trials are needed to accurately estimate the false-alarm probability? Choose values for A and σ^2 that will result in values for P_F of 0.1 and 0.01. Estimate the false-alarm probability and compare with the theoretical value in each case.

4.16 Drug Testing

In testing drugs, the variability among patients makes judging effectiveness difficult, but not impossible. The number of people N a drug cures has a geometric probability distribution.

$$\Pr[N = n] = (1-a)a^n,\ n = 0, 1, \dots$$

You perform a drug trial over a very large population (large enough so that the approximation of the geometric probability distribution remains valid). Either the drug is ineffective, in which case the distribution's parameter equals a_0, or is effective and the parameter equals a_*, $a_* > a_0$. The *a priori* probability that the drug will be effective is π_*.

(a) Construct the minimum probability of error test that decides drug effectiveness.

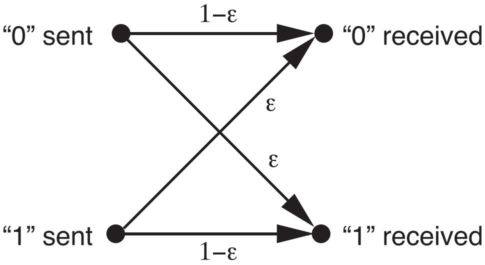


Figure 4.12: A binary symmetric digital communications channel.

(b) What is the probability of error the test will achieve?

(c) Now suppose the drug trial is repeated in several countries, each of which has a large population. Because conducting such tests is expensive, you want a test that will reach a conclusion as quickly as possible. Find a test that will achieve false-positive and false-negative rates of α and $1-\beta$, respectively, as quickly as possible.

4.17 The optimum reception of binary information can be viewed as a model testing problem. Here, equally-likely binary data (a "zero" or a "one") is transmitted through a binary symmetric channel. The indicated parameters denote the probabilities of receiving a binary digit given that a particular digit was sent. Assume that $\varepsilon = 0.1$.

(a) Assuming a single transmission for each digit, what is the minimum probability of error receiver and what is the resulting probability of error?

(b) One method of improving the probability of error is to repeat the digit to be transmitted L times. This transmission scheme is equivalent to the so-called *repetition code*. The receiver uses all of the received L digits to decide what digit was actually sent. Assume that the results of each transmission are statistically independent of all others. Construct the minimum probability of error receiver and find an expression for P_e in terms of L.

(c) Assume that we desire the probability of error to be 10^{-6}. How long a repetition code is required to achieve this goal for the channel given above? Assume that the leading term in the probability of error expression found in part (b) dominates.

4.18 In some cases it might be wise to *not* make a decision when the data do not justify it. Thus, in addition to declaring that one of two models occurred, we might declare "no decision" when the data are indecisive. Assume you observe L statistically independent observations X_l, each of which is Gaussian and has a variance of two. Under one model the mean is zero, and under the other the mean is one. The models are equally likely to occur.

(a) Construct a hypothesis testing rule that yields a probability of no-decision no larger than some specified value α, maximizes the probabilities of making correct decisions when they are made, and makes these correct-decision probabilities equal.

(b) What is the probability of a correct decision for your rule?

4.19 You decide to flip coins with Sleazy Sam. If heads is the result of a coin flip, you win one dollar; if tails, Sam wins a dollar. However, Sam's reputation has preceded him. You suspect that the probability of tails, p, may not be $1/2$. You want to determine whether a biased coin is being used or not after observing the results of three coin tosses.

(a) You suspect that $p = 3/4$. Assuming that the probability of a biased coin equals that of an unbiased coin, how would you decide whether a biased coin is being used or not in a "good" fashion?

(b) Using your decision rule, what is the probability that your determination is incorrect?

(c) One potential flaw with your decision rule is that a specific value of p was assumed. Can a reasonable decision rule be developed without knowing p? If so, demonstrate the rule; if not, show why not.

4.20 When a patient is screened for the presence of a disease in an organ, a section of tissue is viewed under a microscope and a count of abnormal cells made. Even under healthy conditions, a small number of abnormal cells will be present. Presumably a much larger number will be present if the organ is diseased. Assume that the number L of abnormal cells in a section is geometrically distributed.

$$\Pr[L = l] = (1-\alpha)\alpha^l, l = 0, 1, \ldots$$

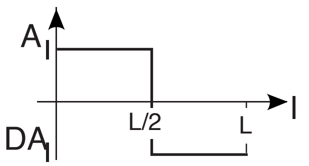
The parameter α of a diseased organ will be larger than that of a healthy one. The probability of a randomly selected organ being diseased is p.

(a) Assuming that the value of the parameter α is known in each situation, find the best method of deciding whether an organ is diseased.

(b) Using your method, a patient was said to have a diseased organ. In this case, what is the probability that the organ is diseased?

(c) Assume that α is known only for healthy organs. Find the disease screening method that minimizes the maximum possible value of the probability that the screening method will be in error.

4.21 Interference
Wireless communication seldomly occurs without the presence of interference that originates from other communications signals. Suppose a binary communication system uses the antipodal signal set $s_i(l) = (-1)^{i+1}A$, $l = 0, \ldots, L-1$, with $i = 0, 1$. An interfering communication system also uses an antipodal signal set having the depicted basic signal. Its amplitude A_I is unknown. The received signal consists of the sum of our signal, the interfering signal, and white Gaussian noise.



(a) What receiver would be used if the bit intervals of the two communication systems were aligned?

(b) How would this receiver change if the bit intervals did *not* align, with the time shift not known? Assume that the receiver knows the time origin of our communications.

4.22 Assume we have N sensors each determining whether a signal is present in white Gaussian noise or not. The identical signal and noise models apply at each sensor, with the signal having energy E.

(a) What is each sensor's receiver?

(b) Assuming the signal is as likely as not, what is the optimal fusion rule?

(c) Does this distributed detection system yield the same error probabilities as the optimal detector that assimilates all the observations directly?

4.23 Data are often processed "in the field," with the results from several systems sent a central place for final analysis. Consider a detection system wherein each of N field radar systems detects the presence or absence of an airplane. The detection results are collected together so that a final judgment about the airplane's presence can be made. Assume each field system has false-alarm and detection probabilities P_F and P_D respectively.

(a) Find the optimal detection strategy for making a final determination that maximizes the probability of making a correct decision. Assume that the *a priori* probabilities π_0, π_1 of the airplane's absence or presence, respectively, are known.

(b) How does the airplane detection system change when the *a priori* probabilities are not known? Require that the central judgment have a false-alarm probability no bigger than $(P_F)^N$.

4.24 Mathematically, a preconception is a model for the "world" that you believe applies over a broad class of circumstances. Clearly, you should be vigilant and continually judge your assumption's correctness. Let $\{X_l\}$ denote a sequence of random variables that you believe to be independent and identically distributed with a Gaussian distribution having zero mean and variance σ^2. Elements of this sequence arrive one after the other, and you decide to use the sample average M_l as a test statistic.

$$M_l = \frac{1}{l}\sum_{i=1}^{l} X_i$$

(a) Based on the sample average, develop a procedure that tests for each l whether the preconceived model is correct. This test should be designed so that it continually monitors the validity of the assumptions, and indicates at each l whether the preconception is valid or not. Establish this test so that it yield a constant probability of judging the model incorrect when, in fact, it is actually valid.

(b) To judge the efficacy of this test, assume the elements of the actual sequence have the assumed distribution, but that they are correlated with correlation coefficient ρ. Determine the probability (as a function of l) that your test correctly invalidates the preconception.

(c) Is the test based on the sample average optimal? If so, prove it so; if not, find the optimal one.

4.25 Assume that observations of a sinusoidal signal $s(l) = A\sin(2\pi fl)$, $l = 0,\ldots,L-1$, are contaminated by first-order colored noise as described in the example $\{129\}$.

(a) Find the unit-sample response of the whitening filter.

(b) Assuming that the alternative model is the sole presence of the colored Gaussian noise, what is the probability of detection?

(c) How does this probability vary with signal frequency f when the first-order coefficient is positive? Does your result make sense? Why?

4.26 In space-time coding systems, a common bit stream is transmitted over several channels simultaneously but using different signals. $\mathbf{X}^{(k)}$ denotes the signal received from the k^{th} channel, $k = 1,\ldots,K$, and the received signal equals $\mathbf{s}^{(k,i)} + \mathbf{N}^{(k)}$. Here, i equals 0 or 1, corresponding to the bit being transmitted. Each signal has length L. $\mathbf{N}^{(k)}$ denotes a Gaussian random vector with statistically independent components having mean zero and variance σ_k^2 (the variance depends on the channel).

(a) Assuming equally likely bit transmissions, find the minimum probability of error decision rule.

(b) What is the probability that your decision rule makes an error?

(c) Suppose each channel has its *own* decision rule, which is designed to yield the same miss probability as the others. Now what is the minimum probability of error decision rule of the system that combines the individual decisions into one?

4.27 The performance for the optimal detector in white Gaussian noise problems depends only on the distance between the signals. Let's confirm this result experimentally. Define the signal under one hypothesis to be a unit-amplitude sinusoid having one cycle within the 50-sample observation interval. Observations of this signal are contaminated by additive white Gaussian noise having variance equal to 1.5. The hypotheses are equally likely.

(a) Let the second hypothesis be a cosine of the same frequency. Calculate and estimate the detector's false-alarm probability.

(b) Now let the signals correspond to square-waves constructed from the sinusoids used in the previous part. Normalize them so that they have the same energy as the sinusoids. Calculate and estimate the detector's false-alarm probability.

(c) Now let the noise be Laplacian with variance 1.5. Although no analytic expression for the detector performance can be found, do the simulated performances for the sinusoid and the square-wave signals change significantly?

(d) Finally, let the second signal be the negative of the sinusoid. Repeat the calculations and the simulation for Gaussian noise.

4.28 Physical constraints imposed on signals can change what signal set choices result in the best detection performance. Let one of two equally likely discrete-time signals be observed in the presence of white Gaussian noise (variance/sample equals σ^2).

$$\begin{aligned} \mathscr{M}_0 &: X(l) = s^{(0)}(l) + N(l) \\ \mathscr{M}_1 &: X(l) = s^{(1)}(l) + N(l) \end{aligned} \quad l = 0, \ldots, L-1$$

We are free to choose any signals we like, but there are constraints. Average signal power equals $\sum_l s^2(l)/L$, and peak power equals $\max_l s^2(l)$.

(a) Assuming the average signal power must be less than P_{ave}, what are the optimal signal choices? Is your answer unique?

(b) When the peak power P_{peak} is constrained, what are the optimal signal choices?

(c) If $P_{\text{ave}} = P_{\text{peak}}$, which constraint yields the best detection performance?

4.29 One of the more interesting problems in detection theory is determining when the probability distribution of the observations differs from that in other portions of the observation interval. The most common form of the problem is that over the interval $[0, C)$, the observations have one form, and that in the remainder of the observation interval $[C, L-1]$ have a different probability distribution. The change detection problem is to determine whether in fact a change has occurred and, if so, estimate when that change occurs.

To explore the change detection problem, let's explore the simple situation where the mean of white Gaussian noise changes at the C^{th} sample.

$$\begin{aligned} \mathscr{M}_0 &: X(l) \sim \mathscr{N}(0, \sigma^2), l = 0, \ldots, L-1 \\ \mathscr{M}_1 &: X(l) \sim \begin{cases} \mathscr{N}(0, \sigma^2), & l = 0, \ldots, C-1 \\ \mathscr{N}(m, \sigma^2), & l = C, \ldots, L-1 \end{cases} \end{aligned}$$

The observations in each case are statistically independent of the others.

(a) Find a detector for this change problem when m is a known positive number.

(b) Find an expression for the threshold in the detector using the Neyman-Pearson criterion.

(c) How does the detector change when the value of m is not known?

sinusoids in noise. present. the frequencies of which are integer fractions of the number of observations: sinusoids is present while ignoring the second? signal's squared amplitude to the noise variance at the frequency of the signal? the observations. the input.

4.30 A sampled signal is suspected of consisting of a periodic component and additive Gaussian noise. The signal, if present, has a known period L. The number of samples equals N, a multiple of L. The noise is white and has known variance σ^2. A consultant (you!) has been asked to determine the signal's presence.

(a) Assuming the signal is a sinusoid with unknown phase and amplitude, what should be done to determine the presence of the sinusoid so that a false-alarm probability criterion of 0.1 is met?

(b) Other than its periodic nature, now assume that the signal's waveform is unknown. What computations must the optimum detector perform?

4.31 The QAM (Quadrature Amplitude Modulation) signal set consists of signals of the form

$$s_i(l) = A_i^c \cos(2\pi f_c l) + A_i^s \sin(2\pi f_c l) ,$$

where A_i^c and A_i^s are amplitudes that define each element of the signal set. These are chosen according to design constraints. Assume the signals are observed in additive Gaussian noise.

(a) What is the optimal amplitude choice for the binary and quaternary (four-signal) signal sets when the noise is white and the signal energy is constrained ($\sum_l s_i^2(l) < E$)? Comment on the uniqueness of your answers.

(b) Describe the optimal binary QAM signal set when the noise is colored.

(c) Now suppose the *peak* amplitude ($\max_l |s_i(l)| < A_{\max}$) is constrained. What are the optimal signal sets (both binary and quaternary) for the white noise case? Again, comment on uniqueness.

4.32 Checking for Repetitions

Consider the following detection problem.

$$\mathscr{M}_0: \begin{array}{rcl} \mathbf{X}_1 &=& \mathbf{s}+\mathbf{N}_1 \\ \mathbf{X}_2 &=& \mathbf{s}+\mathbf{N}_2 \end{array}$$

$$\mathscr{M}_1: \begin{array}{rcl} \mathbf{X}_1 &=& \mathbf{s}_1+\mathbf{N}_1 \\ \mathbf{X}_2 &=& \mathbf{s}_2+\mathbf{N}_2 \end{array}$$

Here, the two observations either contain the same signal or they contain different ones. The noise vectors N_1 and N_2 are statistically independent of each other and identically distributed, with each being Gaussian with zero mean and covariance matrix $\mathbf{K} = \sigma^2\mathbf{I}$.

(a) Find the decision rule that minimizes the false-alarm probability when the miss probability is required to be less than $1-\beta$.

(b) Now suppose *none* of the signals is known. All is that is known is that under $\mathscr{M}_0$, the signals are the same and that under $\mathscr{M}_1$ they are different. What is the optimal decision rule under these conditions?

4.33 A sampled signal is suspected of consisting of a periodic component and additive Gaussian noise. The signal, if present, has a known period N. The number of samples equals L, a multiple of N. The noise is white and has known variance σ^2. A consultant (you!) has been asked to determine the signal's presence.

(a) Assuming the signal is a sinusoid with unknown phase and amplitude, what should be done to determine the presence of the sinusoid so that a false-alarm probability criterion of 0.1 is met?

(b) Other than its periodic nature, now assume that the signal's waveform is unknown. What computations must the optimum detector perform?

4.34 Delegating Responsibility

Modern management styles tend to want decisions to be made locally (by people at the scene) rather than by "the boss." While this approach might be considered more democratic, we should understand how to make decisions under such organizational constraints and what the performance might be.

Let three "local" systems separately make observations. Each local system's observations are identically distributed and statistically independent of the others, and based on the observations, each system decides which of two models applies best. The judgments are relayed to the central manager who must make the final decision. Assume the local observations consist either of white Gaussian noise or of a signal having energy E to which the same white Gaussian noise has been added. The signal energy is the same at each local system. Each local decision system must meet a performance standard on the probability it declares the presence of a signal when none is present.

(a) What decision rule should each local system use?

(b) Assuming the observation models are equally likely, how should the central management make its decision so as to minimize the probability of error?

(c) Is this decentralized decision system optimal (*i.e.*, the probability of error for the final decision is minimized)? If so, demonstrate optimality; if not, find the optimal system.

Appendix
Probability Distributions

Name	*Probability*	*Mean*	*Variance*	*Relationships*
Discrete Uniform	$\frac{1}{N-M+1} \quad M \leq n \leq N$ $0 \quad$ otherwise	$\frac{M+N}{2}$	$\frac{(N-M+2)(N-M)}{12}$	
Bernoulli or Binary	$\Pr(n=0) = 1-p$ $\Pr(n=1) = p$	p	$p(1-p)$	
Binomial	$\binom{N}{n} p^n (1-p)^{N-n}$, $n = 0, \ldots, N$	Np	$Np(1-p)$	Sum of N IID Bernoulli
Geometric	$(1-p)p^n$, $n \geq 0$	$p/1-p$	$p/(1-p)^2$	
Negative Binomial	$\binom{n-1}{N-1} p^N (1-p)^{n-N}, n \geq N$	N/p	$N(1-p)/p^2$	
Poisson	$\frac{\lambda^n e^{-\lambda}}{n!}, n \geq 0$	λ	λ	
Hypergeometric	$\frac{\binom{a}{n}\binom{b}{N-n}}{\binom{a+b}{N}}, n = 0, \ldots, N;$ $0 \leq n \leq a+b; 0 \leq N \leq a+b$	$Na/(a+b)$	$\frac{Nab(a+b-N)}{(a+b)^2(a+b-1)}$	
Logarithmic	$\frac{-p^n}{n \log(1-p)}$	$\frac{-p}{(1-p)\log(1-p)}$	$\frac{-p[p+\log(1-p)]}{(1-p)\log q}$	

Table 2: Discrete probability distributions.

Name	*Density*	*Mean*	*Variance*	*Relationships*
Gaussian (Normal)	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$	m	σ^2	
Bivariate Gaussian	$\frac{1}{2\pi(1-\rho^2)^{1/2}\sigma_x\sigma_y}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-m_x}{\sigma_x}\right)^2-2\rho\left(\frac{x-m_x}{\sigma_x}\right)\left(\frac{y-m_y}{\sigma_y}\right)+\left(\frac{y-m_y}{\sigma_y}\right)^2\right]\right\}$	$\mathscr{E}[x]=m_x$, $\mathscr{E}[y]=m_y$	$\mathscr{V}[x]=\sigma_x^2$, $\mathscr{V}[y]=\sigma_y^2$, $\mathscr{E}[xy]=$ $m_xm_y+\rho\sigma_x\rho\sigma_y$	ρ: correlation coefficient
Conditional Gaussian	$p(x\mid y)=\frac{1}{\sqrt{2\pi(1-\rho)^2\sigma_x^2}}\exp\left\{-\frac{\left(x-m_x-\frac{\rho\sigma_x}{\sigma_y}(y-m_y)\right)^2}{2\sigma_x^2(1-\rho^2)}\right\}$	$m_x + \frac{\rho\sigma_x}{\sigma_y}(y-m_y)$	$\sigma_x^2(1-\rho^2)$	
Multivariate Gaussian	$\frac{1}{(\det(2\pi\mathbf{K}))^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^t\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})\right\}$	$\mathbf{m}$	$\mathbf{K}$	
Generalized Gaussian	$\frac{1}{2\Gamma(1+1/r)A(r)}e^{-\left\lvert\frac{x-m}{A(r)}\right\rvert^r}$	m	σ^2	$A(r)=$ $\left[\frac{\sigma^2\Gamma(1/r)}{\Gamma(3/r)}\right]^{1/2}$
Chi-Squared (χ_ν^2)	$\frac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{\frac{\nu}{2}-1}e^{-x/2}, 0\leq x$	ν	2ν	$\chi_\nu^2=\sum_{i=1}^\nu X_i^2$, X_i IID $\mathscr{N}(0,1)$
Noncentral Chi-Squared ($\chi_\nu'^2(\lambda)$)	$\frac{1}{2}(x/\lambda)^{(\nu-2)/4}I_{(\nu-2)/2}(\sqrt{\lambda}x)e^{-1/2(\lambda+x)}$	$\nu+\lambda$	$2(\nu+2\lambda)$	$\chi_\nu'^2=\sum_{i=1}^\nu X_i^2$, X_i IID $\mathscr{N}(m_i,1)$ $\lambda=\sum_{i=1}^\nu m_i^2$
Chi χ_ν	$\frac{x^{\nu-1}e^{-x^2/2}}{2^{\nu/2-1}\Gamma\left(\frac{\nu}{2}\right)}, 0\leq x$	$\sqrt{2}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{1}{2}\nu\right)}$	$\frac{2\left[\Gamma\left(\frac{1}{2}\nu\right)\Gamma\left(\frac{1}{2}\nu+1\right)-\Gamma^2\left(\frac{\nu+1}{2}\right)\right]}{\Gamma^2\left(\frac{1}{2}\nu\right)}$	$\chi_\nu=\sqrt{\chi_\nu^2}$
Student's t	$\frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$	0	$\frac{\nu}{\nu-2}, 2<\nu$	
Beta $\beta_{m,n}$	$\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}x^{m/2-1}(1-x)^{n/2-1}, 0<x<1, 0<a,b$	$\frac{m}{m+n}$	$\frac{2mn}{(m+n)^2(m+n+2)}$	$\beta_{m,n}=\frac{\chi_m^2}{\chi_m^2+\chi_n^2}$
F Distribution	$\frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)}\left(\frac{m}{n}\right)^{m/2}\frac{x^{(m-2)/2}}{[1+(m/n)x]^{(m+n)/2}}, 0\leq x; 1\leq m,n$	$\frac{n}{n-2}, n>2$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, n>4$	$F_{m,n}=\frac{\chi_m^2/m}{\chi_n^2/n}$
Non-central F $F'_{m,n}(\lambda)$	$\sum_{k=0}^{\infty}\frac{\left(\frac{\lambda}{2}\right)^k}{k!}e^{-\frac{\lambda}{2}}p_{\beta_{\frac{m}{2}+k,\frac{n}{2}}}\left(\frac{mx}{mx+n}\right)$	$\frac{n}{n-2}, n>2$	$2\left(\frac{n}{m}\right)^2\frac{(m+\lambda)^2+(m+2\lambda)(n-2)}{(n-2)^2(n-4)}$, $n>4$	$F'_{m,n}(\lambda)=\frac{\chi_m'^2(\lambda)/m}{\chi_n^2/n}$
Wishart $\mathbf{W}_M(N,\mathbf{K})$	$\frac{(\det[\mathbf{w}])^{\frac{N-M-1}{2}}}{2^{\frac{NM}{2}}\Gamma_M(\frac{N}{2})(\det[\mathbf{K}])^{\frac{K}{2}}}e^{-\frac{\mathrm{tr}[\mathbf{K}^{-1}\mathbf{w}]}{2}}$, $\Gamma_M(\frac{N}{2})=\pi^{M(M-1)/4}\times\prod_{m=0}^{M-1}\Gamma\left(\frac{N}{2}-\frac{m}{2}\right)$	$N\mathbf{K}$	$\mathrm{cov}[\mathbf{W}_{ij}\mathbf{W}_{kl}]=$ $N\cdot(\mathbf{K}_{ik}\mathbf{K}_{jl}+\mathbf{K}_{il}\mathbf{K}_{jk})$	$\mathbf{W}_M(N,\mathbf{K})=\sum_{n=1}^N\mathbf{X}_n\mathbf{X}'_n$, $\mathbf{X}_n\sim\mathscr{N}(\mathbf{0},\mathbf{K})$ $\dim[\mathbf{X}]=M$

Table 3: Distributions related to the Gaussian.

Name	*Density*	*Mean*	*Variance*	*Relationships*
Uniform	$\frac{1}{b-a}, a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
Triangular	$2x/a \quad 0 \leq x \leq a$ $2(1-x)/(1-a) \quad a \leq x \leq 1$	$\frac{1+a}{3}$	$\frac{1-a+a^2}{18}$	
Exponential	$\lambda e^{-\lambda x}, 0 \leq x$	$1/\lambda$	$1/\lambda^2$	
Lognormal	$\frac{1}{\sqrt{2\pi\sigma^2 x^2}} e^{-\frac{1}{2}\left(\frac{\log x - m}{\sigma}\right)^2}, 0 < x$	$e^{m+\frac{\sigma^2}{2}}$	$e^{2m}\left(e^{2\sigma^2} - e^{\sigma^2}\right)$	
Maxwell	$\sqrt{\frac{2}{\pi}} a^{3/2} x^2 e^{-ax^2/2}, 0 < x$	$\sqrt{\frac{8}{\pi a}}$	$\left(3 - \frac{8}{\pi}\right) a^{-1}$	
Laplacian	$\frac{1}{\sqrt{2\sigma^2}} e^{-\frac{\lvert x-m\rvert}{\sqrt{\sigma^2/2}}}$	m	σ^2	
Gamma	$\frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, 0 < x, 0 < a, b$	$\frac{a}{b}$	$\frac{a}{b^2}$	
Rayleigh	$2axe^{-ax^2}, 0 \leq x$	$\sqrt{\frac{\pi}{4a}}$	$\frac{1}{a}\left(1 - \frac{\pi}{4}\right)$	
Weibull	$abx^{b-1} e^{-ax^b}, 0 < x, 0 < a, b$	$(1/a)^{1/b} \cdot$ $\Gamma(1+1/b)$	$a^{-2/b} \cdot \left[\Gamma(1+2/b) - \Gamma^2(1+1/b)\right]$	
Arc-Sine	$\frac{1}{\pi\sqrt{x(1-x)}}, 0 < x < 1$	$\frac{1}{2}$	$\frac{1}{8}$	
Sine Amplitude	$\frac{1}{\pi\sqrt{1-x^2}}, \lvert x\rvert < 1$	0	$\frac{1}{2}$	
Circular Normal	$\frac{e^{a\cos(x-m)}}{2\pi I_0(a)}, -\pi < x \leq \pi$	m		
Cauchy	$\frac{a/\pi}{(x-m)^2+a^2}$	m (from symmetry arguments)	∞	
Logistic	$\frac{e^{-(x-m)/a}}{a[1+e^{-(x-m)/a}]^2}, 0 < a$	m	$\frac{a^2\pi^2}{3}$	
Gumbel	$\frac{e^{-(x-m)/a}}{a} \exp\left\{-e^{-(x-m)/a}\right\},$ $0 < a$	$m + a\gamma$	$\frac{a^2\pi^2}{6}$	
Pareto	$\frac{ab^a}{x^{1-a}}, 0 < a; 0 < b \leq x$	$\frac{ab}{a-1}, a > 1$	$\frac{ab^2}{(a-2)(a-1)^2}, a > 2$	

Table 4: Non-Gaussian distributions.

Bibliography

1. B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice Hall, Englewood Cliffs, NJ, 1979.
2. I. F. Blake and J. B. Thomas. On a class of processes arising in linear estimation theory. *IEEE Trans. Info. Th.*, IT–14:12–16, January 1968.
3. R. Bradley. Basic properties of strong mixing conditions. In E. Eberlein and M. S. Taqqu, editors, *Dependence in Probability and Statistics*, pages 165–192. Birkhäuser, Boston, MA, 1986.
4. J. P. Burg. The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics*, 37:375–376, Apr. 1972.
5. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.
6. H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ, 1946.
7. H. Cramér. *Random Variables and Probability Distributions*. Cambridge University Press, third edition, 1970.
8. Y. A. Davydov. Mixing conditions for Markov chains. *Th. Prob. and its Applications*, 18:312–328, 1973.
9. P. Doukhan and M. Ghindes. Estimation dans le processus $x_{n+1} = f(x_n) + \varepsilon_n$. *C. R. Acad. Science A*, 297:61–64, 1980.
10. D. J. Edelblute, J. M. Fisk, and G. L. Kinnison. Criteria for optimum-signal-detection theory for arrays. *J. Acoust. Soc. Am.*, 41:199–205, Jan. 1967.
11. P. Hall. *Rates of convergence in the central limit theorem*, volume 62 of *Research Notes in Mathematics*. Pitman Advanced Publishing Program, 1982.
12. A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58:83–90, 1971.
13. S. Haykin. *Adaptive Filter Theory*. Prentice Hall, Englewood Cliffs, NJ, 1986.
14. I. A. Ibragimov. A note on the central limit theorem for dependent random variables. *Th. Prob. and its Applications*, 20:135–141, 1975.
15. I. A. Ibragimov and Yu.V. Linnik. *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff Publishing, Groningen, 1971.
16. D. H. Johnson and A. Swami. The transmission of signals by auditory-nerve fiber discharge patterns. *J. Acoust. Soc. Am.*, 74:493–501, Aug 1983.
17. D. H. Johnson, C. Tsuchitani, D. Linebarger, and M. Johnson. The application of a point process model to the single unit responses of the cat lateral superior olivae to ipsilaterally presented tones. *Hearing Res.*, 21:135–159, 1986.
18. A. N. Kolmogorov and Y. A. Rozanov. On strong mixing conditions for stationary gaussian processes. *Th. Prob. and its Applications*, 5:204–208, 1960.
19. E. L. Lehmann. *Testing Statistical Hypotheses*. John Wiley & Sons, New York, second edition, 1986.
20. R. S. Lipster and A. N. Shiryayev. *Statistics of Random Processes I: General Theory*. Springer-Verlag, New York, 1977.
21. J. Makhoul. Linear prediction: A tutorial review. *Proc. IEEE*, 63:561–580, Apr. 1975.
22. J. D. Markel and A. H. Gray, Jr. *Linear Prediction of Speech*. Springer-Verlag, New York, 1976.
23. S. L. Marple, Jr. *Digital Spectral Analysis*. Prentice Hall, Englewood Cliffs, NJ, 1987.

24. D. P. McGinn and D. H. Johnson. Estimation of all-pole model parameters from noise-corrupted sequences. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-37:433–436, Mar. 1989.
25. D. K. McGraw and J. F. Wagner. Elliptically symmetric distributions. *IEEE Trans. Info. Th.*, IT-14:110–120, 1968.
26. P. A. P. Moran. Some experiments on the prediction of sunspot numbers. *J. Royal Stat. Soc. B*, 16:112–117, 1954.
27. J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. Ser. A*, 231:289–337, Feb. 1933.
28. T. Ozaki. Maximum likelihood estimation of hawkes' self-exciting point processes. *Ann. Inst. Math. Stat.*, 31:145–155, 1979. Part B.
29. A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, second edition, 1984.
30. E. Parzen. *Stochastic Processes*. Holden-Day, San Francisco, 1962.
31. P. S. Rao and D. H. Johnson. Generation and analysis of non-Gaussian Markov time series. *IEEE Trans. Signal Processing*, 40:845–856, 1992.
32. M. Rosenblatt. A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. U. S. A.*, 42:43–47, 1956.
33. M. Rosenblatt. *Markov Processes. Structure and Asymptotic Behavior*. Springer-Verlag, New York, 1971.
34. M. Rosenblatt. *Stationary Sequences and Random Fields*. Birkhauser, Boston, MA, 1985.
35. H. Sakai and H. Tokumaru. Statistical analysis of a spectral estimator for ARMA processes. *IEEE Trans. Auto. Control*, AC-25:122–124, Feb. 1980.
36. B. W. Silverman. *Density Estimation*. Chapman & Hall, London, 1986.
37. D. L. Snyder. *Random Point Processes*. Wiley, New York, 1975.
38. J. R. Thompson and R. A. Tapia. *Nonparametric Function Estimation, Modeling, and Simulation*. SIAM, Philadelphia, PA, 1990.
39. H. Tong. *Non-linear Time Series*. Clarendon Press, Oxford, 1990.
40. H. L. van Trees. *Detection, Estimation, and Modulation Theory, Part I*. John Wiley & Sons, New York, 1968.
41. A. M. Vershik. Some characteristic properties of Gaussian stochastic process. *Th. Prob. and its Applications*, 9:353–356, 1964.
42. G. H. Weiss. Time reversibility of linear stochastic processes. *J. Appl. Prob.*, 12:831–836, 1975.
43. N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press, Cambridge, MA, 1949.
44. P. M. Woodward. *Probability and Information Theory, with Applications to Radar*. Pergamon Press, Oxford, second edition, 1964.