Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory. Harry L. Van Trees

ISBNs: 0-471-09517-6 (Paperback); 0-471-22108-2 (Electronic)

3

Representations of Random Processes

3.1 INTRODUCTION

In this chapter we discuss briefly some of the methods of characterizing random processes that we need for the remainder of the book. The essential idea that we want to emphasize is straightforward. There are many alternate ways of characterizing waveforms and random processes, but the best depends heavily on the problem that we are trying to solve. An intelligent characterization frequently makes the problem solution almost trivial.

Several methods of characterizing signals come immediately to mind. The first is a time-domain characterization. A typical signal made up of pulses of various heights is shown in Fig. 3.1. A time-domain characterization describes the signal shape clearly.

Is it a good representation? To answer this question we must specify what we are going to do with the signal. In Fig. 3.2 we illustrate two possible cases. In the first we pass the signal through a limiter and want to

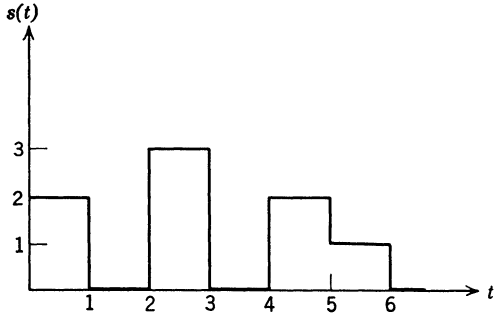

Fig. 3.1 A typical signal.

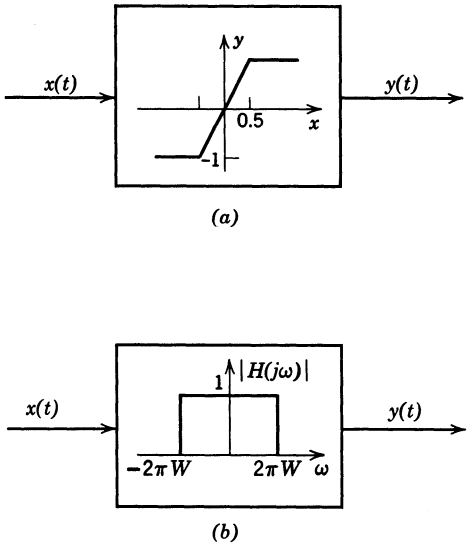

Fig. 3.2 Operations on signals

calculate the output. The time domain characterization enables us to find the output by inspection. In the second we pass the signal through an ideal low-pass filter and want to calculate the energy in the output. In this case a time-domain approach is difficult. If, however, we take the Fourier transform of $s(t)$,

$$S(j\omega) = \int_{-\infty}^{\infty} s(t)e^{-j\omega t}\,dt, \tag{1}$$

the resulting problem is straightforward. The energy in $y(t)$ is E_y, where

$$E_y = 2\int_0^{2\pi W} |S(j\omega)|^2 \frac{d\omega}{2\pi}. \tag{2}$$

Thus, as we well know, both the time-domain and frequency-domain descriptions play an important role in system analysis. The point of the example is that the most efficient characterization depends on the problem of interest.

To motivate another method of characterization consider the simple communication systems shown in Fig. 3.3. When hypothesis 1 is true, the deterministic signal $s_1(t)$ is transmitted. When hypothesis 0 is true, the signal $s_2(t)$ is transmitted. The particular transmitted waveforms are different in systems *A*, *B*, and *C*. The noise in each idealized system is constructed by multiplying the two deterministic waveforms by independent, zero-mean, Gaussian random variables and adding the resulting waveforms. The noise waveform will have a different shape in each system.

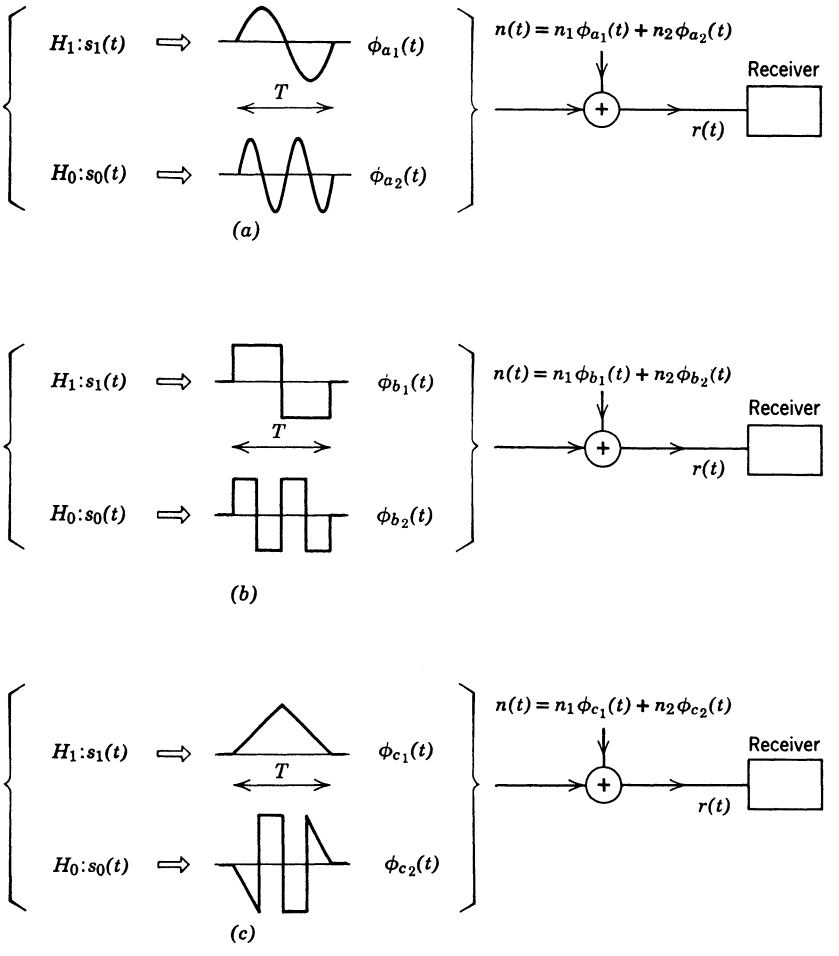

Fig. 3.3 Three hypothetical communication systems.

The receiver wants to decide which hypothesis is true. We see that the transmitted signal and additive noise are appreciably different waveforms in systems *A*, *B*, and *C*. In all cases, however, they can be written as

$$\begin{aligned} s_1(t) &= s_1\phi_1(t), & 0 \le t \le T, \\ s_2(t) &= s_2\phi_2(t), & 0 \le t \le T, \\ n(t) &= n_1\phi_1(t) + n_2\phi_2(t), & 0 \le t \le T, \end{aligned} \tag{3}$$

where the functions $\phi_1(t)$ and $\phi_2(t)$ are *orthonormal*; that is,

$$\int_0^T \phi_i(t)\,\phi_j(t)\,dt = \delta_{ij}, \qquad i, j = 1, 2. \tag{4}$$

The functions $\phi_1(t)$ and $\phi_2(t)$ are different in the three systems. It is clear that because

$$\begin{aligned} r(t) &= (s_1 + n_1)\,\phi_1(t) + n_2\,\phi_2(t), \qquad 0 \le t \le T : H_1, \\ r(t) &= n_1\,\phi_1(t) + (s_2 + n_2)\,\phi_2(t), \qquad 0 \le t \le T : H_0, \end{aligned} \tag{5}$$

we must base our decision on the observed value of the coefficients of the two functions. Thus the test can be viewed as

$$\begin{aligned} \mathbf{r} \triangleq \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} &= \begin{bmatrix} s_1 \\ 0 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \end{bmatrix}, \qquad : H_1, \\ \mathbf{r} \triangleq \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} &= \begin{bmatrix} 0 \\ s_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \end{bmatrix}, \qquad : H_0. \end{aligned} \tag{6}$$

This, however, is just a problem in classical detection that we encountered in Chapter 2.

The important observation is that any pair of orthonormal functions $\phi_1(t)$ and $\phi_2(t)$ will give the same detection performance. Therefore either a time-domain or frequency-domain characterization will tend to obscure the significant features of this particular problem. We refer to this third method of characterization as an *orthogonal series* representation.

We develop this method of characterizing both deterministic signals and random processes in this chapter. In the next section we discuss deterministic signals.

3.2 DETERMINISTIC FUNCTIONS: ORTHOGONAL REPRESENTATIONS

Consider the function $x(t)$ which is defined over the interval $[0, T]$ as shown in Fig. 3.4. We assume that the energy in the function has some finite value E_x.

$$E_x = \int_0^T x^2(t)\,dt < \infty. \tag{7}$$

Now the sketch implies one way of specifying $x(t)$. For every t we know the value of the function $x(t)$. Alternately, we may wish to specify $x(t)$ by a countable set of numbers.

The simple example in the last section suggests writing

$$x(t) = \sum_{i=1}^{\infty} x_i\,\phi_i(t), \tag{8}$$†

† Throughout most of our discussion we are concerned with expanding real waveforms using real orthonormal functions and real coefficients. The modifications to include complex orthonormal functions and coefficients are straightforward.

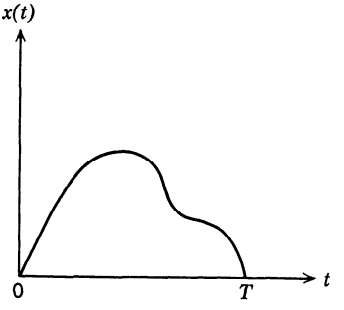

Fig. 3.4 A time-limited function.

where the $\phi_i(t)$ are some set of orthonormal functions. For example, we could choose a set of sines and cosines

$$\begin{aligned}
\phi_1(t) &= \left(\frac{1}{T}\right)^{1/2}, \\
\phi_2(t) &= \left(\frac{2}{T}\right)^{1/2} \cos\left(\frac{2\pi}{T}t\right), \\
\phi_3(t) &= \left(\frac{2}{T}\right)^{1/2} \sin\left(\frac{2\pi}{T}t\right), \\
&\vdots \\
\phi_{2n}(t) &= \left(\frac{2}{T}\right)^{1/2} \cos\left(\frac{2\pi}{T}nt\right).
\end{aligned} \qquad 0 \le t \le T. \tag{9}$$

Several mathematical and practical questions come to mind. The mathematical questions are the following:

1. Because it is only practical to use a finite number (N) of coefficients, how should we choose the coefficients to minimize the mean-square approximation (or representation) error?
2. As N increases, we would like the mean-square approximation error to go to zero. When does this happen?

The practical question is this:

If we receive $x(t)$ as a voltage waveform, how can we generate the coefficients experimentally?

First we consider the mathematical questions. The representation error is

$$e_N(t) = x(t) - \sum_{i=1}^{N} x_i \phi_i(t), \tag{10}$$

when we use N terms. The energy in the error is

$$E_e(N) \triangleq \int_0^T e_N^2(t)\, dt = \int_0^T \left[x(t) - \sum_{i=1}^{N} x_i \phi_i(t)\right]^2 dt. \tag{11}$$

We want to minimize this energy for any N by choosing the x_i appropriately. By differentiating with respect to some particular x_j, setting the result equal to zero, and solving, we obtain

$$x_j = \int_0^T x(t)\,\phi_j(t)\,dt. \tag{12}$$

Because the second derivative is a positive constant, the x_j given by (12) provides an absolute minimum. The choice of coefficient does not change as N is increased because of the orthonormality of the functions.

Finally, we look at the energy in the representation error as $N \to \infty$.

$$\begin{aligned} E_e(N) &\triangleq \int_0^T e_N^2(t)\,dt = \int_0^T \left[x(t) - \sum_{i=1}^{N} x_i\,\phi_i(t)\right]^2 dt \\ &= E_x - 2\sum_{i=1}^{N} \int_0^T x(t)x_i\,\phi_i(t)\,dt + \int_0^T \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j\,\phi_i(t)\,\phi_j(t)\,dt \\ &= E_x - \sum_{i=1}^{N} x_i^2. \end{aligned} \tag{13}$$

Because the x_i^2 are nonnegative, the error is a monotone-decreasing function of N.

If,

$$\lim_{N\to\infty} E_e(N) = 0 \tag{14}$$

for all $x(t)$ with finite energy, we say that the $\phi_i(t)$, $i = 1, \ldots,$ are a *complete orthonormal* (CON) *set* over the interval $[0, T]$ for the class of functions with finite energy. The importance of completeness is clear. If we are willing to use more coefficients, the representation error decreases. In general, we want to be able to decrease the energy in the error to any desired value by letting N become large enough.

We observe that for CON sets

$$E_x = \sum_{i=1}^{\infty} x_i^2. \tag{15}$$

Equation 15 is just Parseval's theorem. We also observe that x_i^2 represents the energy in a particular component of the signal.

Two possible ways of generating the coefficients are shown in Fig. 3.5. In the first system we multiply $x(t)$ by $\phi_i(t)$ and integrate over $[0, T]$. This is referred to as a correlation operation. In the second we pass $x(t)$ into a set of linear filters with impulse responses $h_i(\tau) = \phi_i(T - \tau)$ and observe the outputs at time T. We see that the sampled output of the ith filter is

$$\int_0^T x(\tau)\,h_i(T - \tau)\,d\tau.$$

For the particular impulse response used this is x_i,

$$x_i = \int_0^T x(\tau)\,\phi_i(\tau)\,d\tau, \qquad i = 1, 2, \ldots, N. \tag{16}$$

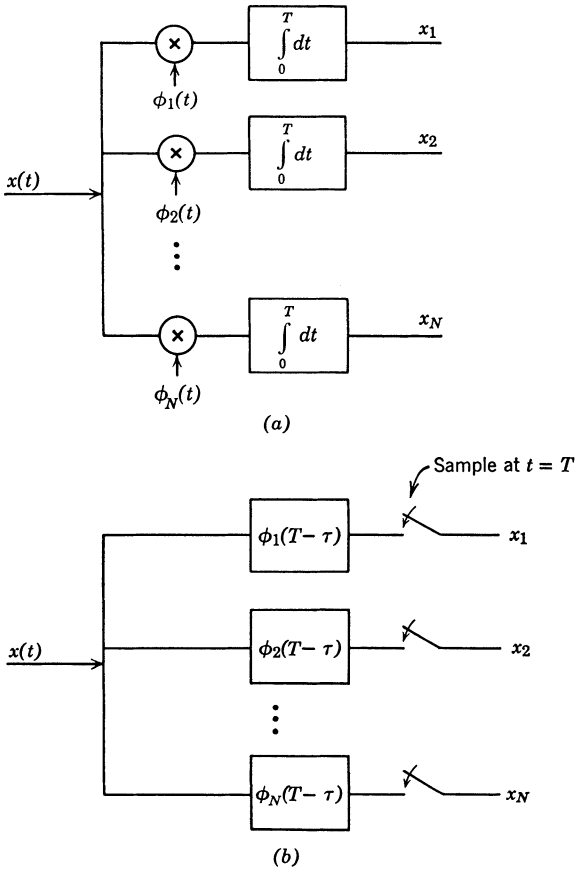

Fig. 3.5 Generation of expansion coefficients: (*a*) correlation operation; (*b*) filter operation.

In Chapter 2 we saw that it was convenient to consider N observations as a point in an N-dimensional space. We shall find that it is equally useful to think of the N coefficients as defining a point in a space. For arbitrary signals we may need an infinite dimensional space. Thus any finite energy signal can be represented as a vector. In Fig. 3.6 we show two signals—$s_1(t)$ and $s_2(t)$:

$$s_1(t) = \sum_{i=1}^{3} s_{1i}\,\phi_i(t),$$
$$s_2(t) = \sum_{i=1}^{3} s_{2i}\,\phi_i(t). \tag{17a}$$

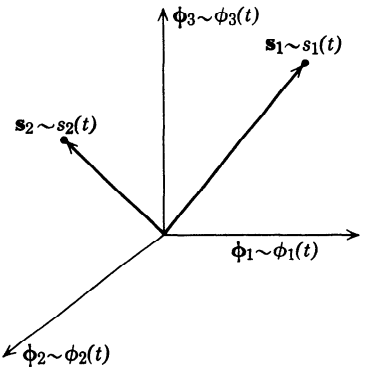

Fig. 3.6 Representation of a signal as a vector.

The corresponding signal vectors are

$$\mathbf{s}_1 \triangleq \begin{bmatrix} s_{11} \\ s_{12} \\ s_{13} \end{bmatrix} = \sum_{i=1}^{3} s_{1i}\boldsymbol{\phi}_i,$$

$$\mathbf{s}_2 \triangleq \begin{bmatrix} s_{21} \\ s_{22} \\ s_{23} \end{bmatrix} = \sum_{i=1}^{3} s_{2i}\boldsymbol{\phi}_i. \tag{17b}$$

Several observations follow immediately:

1. The length of the signal vector squared equals the energy in the signal.

$$\begin{aligned} |\mathbf{s}_1|^2 &= E_1, \\ |\mathbf{s}_2|^2 &= E_2. \end{aligned} \tag{18}$$

2. The correlation coefficient between two signals is defined as

$$\rho_{12} \triangleq \frac{\int_0^T s_1(t)\, s_2(t)\, dt}{\sqrt{E_1 E_2}}. \tag{19}$$

Substituting (17*a*) into (19), we have

$$\rho_{12} = \frac{\int_0^T \left[\sum_{i=1}^{3} s_{1i}\, \phi_i(t)\right]\left[\sum_{j=1}^{3} s_{2j}\, \phi_j(t)\right] dt}{\sqrt{E_1 E_2}}. \tag{20}$$

Using the orthonormality of the coordinate functions the integral reduces to

$$\rho_{12} = \frac{\sum_{i=1}^{3} s_{1i}s_{2i}}{\sqrt{E_1E_2}}. \tag{21}$$

The numerator is just the dot product of $\mathbf{s}_1$ and $\mathbf{s}_2$. Using (18) in the denominator, we obtain,

$$\rho_{12} = \frac{\mathbf{s}_1 \cdot \mathbf{s}_2}{|\mathbf{s}_1|\,|\mathbf{s}_2|}. \tag{22}$$

The obvious advantage of the vector space interpretation is that it enables us to use familiar geometric ideas in dealing with waveforms.

We now extend these ideas to random waveforms.

3.3 RANDOM PROCESS CHARACTERIZATION

We begin our discussion in this section by reviewing briefly how random processes are conventionally defined and characterized.

3.3.1 Random Processes: Conventional Characterizations

The basic idea of a random process is familiar. Each time we conduct an experiment the outcome is a function over an interval of time instead of just a single number. Our mathematical model is illustrated in Fig. 3.7. Each point in the sample space Ω maps into a time function. We could write the function that came from ω_i as $x(t, \omega_i)$ to emphasize its origin, but it is easier to denote it simply as $x(t)$. The collection of waveforms generated from the points in Ω are referred to as an *ensemble*. If we look down the ensemble at any one time, say t_1, we will have a random variable $x_{t_1} \triangleq x(t_1, \omega)$. Similarly, at other times t_i we have random variables x_{t_i}.

Clearly, we could characterize any particular random variable x_{t_i} by its probability density. A more difficult question is how to characterize the entire process. There is an obvious property that this characterization should have. If we consider a set of times $t_1, t_2, \ldots, t_n$ in the interval in which the process is defined, there are n random variables $x_{t_1}, x_{t_2}, x_{t_3}, \ldots, x_{t_n}$. Any *complete* characterization should be able to specify the joint density $p_{x_{t_1}x_{t_2}\cdots x_{t_n}}(X_1, X_2, \ldots, X_n)$. Furthermore, it should be able to specify this density for any set of n times in the interval (for any finite n).

Unfortunately, it is not obvious that a characterization of this kind will be adequate to answer all questions of interest about a random process. Even if it does turn out to be adequate, there is a practical difficulty in actually specifying these densities for an arbitrary random process.

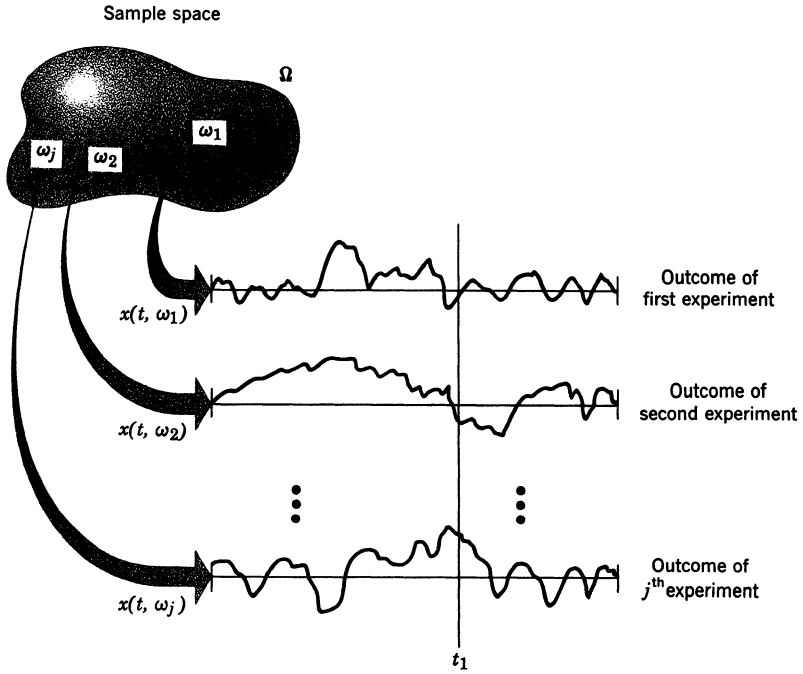

Fig. 3.7 An ensemble of sample functions.

There are two common ways of handling this difficulty in specifying the nth-order density.

Structured Processes. We consider *only* those processes in which any nth-order density has a certain structure that can be produced by using some low-order density and a known algorithm.

Example. Consider the probability density at the ordered set of times

$$t_1 < t_2 < t_3 < \cdots < t_{n-1} < t_n.$$

If

$$p_{x_{t_n}|x_{t_{n-1}}\cdots x_{t_1}}(X_{t_n}|X_{t_{n-1}}\cdots X_{t_1}) = p_{x_{t_n}|x_{t_{n-1}}}(X_{t_n}|X_{t_{n-1}}), \tag{23}$$

the process is called a Markov process. Here knowledge of the second-order density enables us to construct the nth order density (e.g., [2, p. 44] or Problems 3.3.9 and 3.3.10). Other structured processes will appear naturally as our discussion proceeds.

Partial Characterization. We now consider operations on the random process that can be studied without actually completely characterizing the process. For these operations we need only a partial characterization. A

large number of partial characterizations are possible. Two of the most widely used are the following:

1. Single-time characterizations.
2. Second-moment characterizations.

In a single-time characterization we specify only $p_{x_t}(X)$, the first-order probability density at time t. In general, it will be a function of time. A simple example illustrates the usefulness of this characterization.

Example. Let

$$r(t) = x(t) + n(t). \tag{24}$$

Assume that x_t and n_t are statistically independent and $p_{x_t}(X)$ and $p_{n_t}(N)$ are known. We operate on $r(t)$ with a no-memory nonlinear device to obtain a minimum mean square error estimate of $x(t)$ which we denote by $\hat{x}(t)$.

From Chapter 2, $\hat{x}(t)$ is just the conditional mean. Because we are constrained to a no-memory operation, we can use only $r(t)$. Then

$$\hat{x}(t) = \int_{-\infty}^{\infty} X_t p_{x_t|r_t}(X_t|R_t)\, dX_t \triangleq f(R_t). \tag{25}$$

If x_t is Gaussian, $N(0, \sigma_x)$, and n_t is Gaussian, $N(0, \sigma_n)$, it is a simple exercise (cf. Problem 3.3.2) to show that

$$f(R_t) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_n^2} R_t, \tag{26}$$

so that the no-memory device happens to be linear. Observe that because we allowed only a no-memory device a complete characterization of the process was not necessary.

In a second-moment characterization we specify only the first and second moments of the process. We define the mean-value function of the process as

$$m_x(t) \triangleq E(x_t) = \int_{-\infty}^{\infty} X_t p_{x_t}(X_t)\, dX_t. \tag{27}$$

In general, this is a function of time. The correlation function is defined as

$$R_x(t, u) \triangleq E(x_t x_u) = \int_{-\infty}^{\infty}\int X_t X_u p_{x_t x_u}(X_t, X_u)\, dX_t\, dX_u. \tag{28}$$

The covariance function is defined as

$$\begin{aligned} K_x(t, u) &\triangleq E\{[x_t - m_x(t)][x_u - m_x(u)]\} \\ &= R_x(t, u) - m_x(t)\, m_x(u). \end{aligned} \tag{29}$$

This partial characterization is well suited to *linear* operations on random processes. This type of application is familiar (e.g., [1], pp. 171–185).

The covariance function has several properties of interest to us. Looking at the definition in (29), we see that it is symmetric:

$$K_x(t, u) = K_x(u, t). \tag{30}$$

If we multiply a sample function $x(t)$ by some deterministic square-integrable function $f(t)$ and integrate over the interval $[0, T]$, we obtain a random variable:

$$x_f \triangleq \int_0^T x(t)\, f(t)\, dt. \tag{31}$$

The mean of this random variable is

$$E(x_f) = \bar{x}_f \triangleq E\int_0^T x(t)\, f(t)\, dt = \int_0^T m_x(t)\, f(t)\, dt, \tag{32}$$

and the variance is

$$\begin{aligned}\text{Var}\,(x_f) &\triangleq E[(x_f - \bar{x}_f)^2] \\ &= E\left\{\int_0^T [x(t) - m_x(t)]\, f(t)\, dt \int_0^T [x(u) - m_x(u)]\, f(u)\, du\right\}. \end{aligned} \tag{33}$$

Bringing the expectation inside the integral, we have

$$\text{Var}\,(x_f) = \int_0^T\int f(t)\, K_x(t, u)\, f(u)\, dt\, du. \tag{34}$$

The variance must be greater than or equal to zero. Thus, we have shown that

$$\boxed{\int_0^T\int f(t)\, K_x(t, u)\, f(u)\, dt\, du \geq 0} \tag{35}$$

for any $f(t)$ with finite energy. We call this property nonnegative definiteness. If the inequality is strict for every $f(t)$ with nonzero finite energy, we say that $K_x(t, u)$ is positive definite. We shall need the two properties in (30) and (35) in the next section.

If the process is defined over an infinite interval and the covariance function depends only on $|t - u|$ and not t or u individually, we say that the process is *covariance-stationary* and write

$$K_x(t, u) = K_x(t - u) = K_x(\tau). \tag{36}\dagger$$

Similarly, if the correlation function depends only on $|t - u|$, we say that the process is *correlation-stationary* and write

$$R_x(t, u) = R_x(t - u) = R_x(\tau). \tag{37}$$

† It is important to observe that although $K_x(t, u)$ is a function of two variables and $K_x(\tau)$ of only one variable, we use the same notation for both. This economizes on symbols and should cause no confusion.

For stationary processes, a characterization using the power density spectrum $S_x(\omega)$ is equivalent to the correlation function characterization

$$S_x(\omega) \triangleq \int_{-\infty}^{\infty} R_x(\tau) e^{-j\omega\tau}\, d\tau$$

and

$$R_x(\tau) = \int_{-\infty}^{\infty} S_x(\omega) e^{j\omega\tau} \frac{d\omega}{2\pi}. \tag{38}$$

As already pointed out, these partial characterizations are useful only when the operations performed on the random process are constrained to have a certain form. A much more useful representation for the problems of interest to us is a characterization in terms of an orthogonal series expansion. In the next section we use a series expansion to develop a second-moment characterization. In the succeeding section we extend it to provide a complete characterization for a particular process of interest. It is worthwhile to observe that we have yet to commit ourselves in regard to a complete characterization of a random process.

3.3.2 Series Representation of Sample Functions of Random Processes

In Section 3.2 we saw how we could represent a deterministic waveform with finite energy in terms of a series expansion. We now want to extend these ideas to include sample functions of a random process. We start off by choosing an arbitrary complete orthonormal set: $\phi_1(t), \phi_2(t), \ldots$. For the moment we shall not specify the exact form of the $\phi_i(t)$. To expand $x(t)$ we write

$$x(t) = \lim_{N\to\infty} \sum_{i=1}^{N} x_i \phi_i(t), \qquad 0 \le t \le T, \tag{39}$$

where

$$x_i \triangleq \int_0^T x(t)\, \phi_i(t)\, dt. \tag{40}$$

We have not yet specified the type of convergence required of the sum on the right-hand side. Various types of convergence for sequences of random variables are discussed in the prerequisite references [1, p. 63] or [29].

An ordinary limit is not useful because this would require establishing conditions on the process to guarantee that *every* sample function could be represented in this manner.

A more practical type of convergence is mean-square convergence:

$$x(t) = \underset{N\to\infty}{\text{l.i.m.}} \sum_{i=1}^{N} x_i\, \phi_i(t), \qquad 0 \le t \le T. \tag{41}$$

The notation "l.i.m." denotes limit in the mean (e.g., [1, p. 63]) which is defined as,

$$\lim_{N\to\infty} E\left[\left(x_t - \sum_{i=1}^{N} x_i\,\phi_i(t)\right)^2\right] = 0, \qquad 0 \le t \le T. \tag{42}$$

For the moment we assume that we can find conditions on the process to guarantee the convergence indicated in (42).

Before doing so we discuss an appropriate choice for the orthonormal set. In our discussions of classical detection theory our observation space was finite dimensional and usually came with a built-in coordinate system. In Section 2.6 we found that problems were frequently easier to solve if we used a new coordinate system in which the random variables were uncorrelated (if they happened to be Gaussian variables, they were also statistically independent). In dealing with continuous waveforms we have the advantage that there is no specified coordinate system, and therefore we can choose one to suit our purposes. From our previous results a logical choice is a set of $\phi_i(t)$ that leads to *uncorrelated* coefficients.

If

$$E(x_i) \triangleq m_i, \tag{43}$$

we would like

$$E[(x_i - m_i)(x_j - m_j)] = \lambda_i \delta_{ij}. \tag{44}$$

For simplicity we assume that $m_i = 0$ for all i. Several observations are worthwhile:

1. The value x_i^2 has a simple physical interpretation. It corresponds to the *energy* along the coordinate function $\phi_i(t)$ in a particular sample function.
2. Similarly, $E(x_i^2) = \lambda_i$ corresponds to the *expected* value of the energy along $\phi_i(t)$, assuming that $m_i = 0$. Clearly, $\lambda_i \ge 0$ for all i.
3. If $K_x(t, u)$ is positive definite, every λ_i is greater than zero. This follows directly from (35). A little later it will be easy to show that if $K_x(t, u)$ is not positive definite, at least one λ_i must equal zero.

We now want to determine what the requirement in (44) implies about the complete orthogonal set. Substituting (40) into (44) and bringing the expectation inside the integral, we obtain

$$\begin{aligned}\lambda_i \delta_{ij} = E(x_i x_j) &= E\left[\int_0^T x(t)\,\phi_i(t)\,dt \int_0^T x(u)\,\phi_j(u)\,du\right] \\ &= \int_0^T \phi_i(t)\,dt \int_0^T K_x(t, u)\,\phi_j(u)\,du, \qquad \text{for all } i \text{ and } j. \end{aligned} \tag{45}$$

In order that (45) may hold for all choices of i and a particular j, it is necessary and sufficient that the inner integral equal $\lambda_j \phi_j(t)$:

$$\lambda_j \phi_j(t) = \int_0^T K_x(t, u) \phi_j(u)\, du, \qquad 0 \le t \le T. \tag{46}$$

The functions $\phi_i(t)$ are called eigenfunctions and the numbers λ_i are called eigenvalues.

Therefore we want to demonstrate that for some useful class of random processes there exist solutions to (46) with the desired properties. The form of (46) is reminiscent of the equation that specified the eigenvectors and eigenvalues in Section 2.6 (2–363),

$$\lambda \boldsymbol{\phi} = \mathbf{K}_x \boldsymbol{\phi}, \tag{47}$$

where $\mathbf{K}_x$ was a symmetric, nonnegative definite matrix. This was a set of N simultaneous homogeneous linear equations where N was the dimensionality of the observation space. Using results from linear equation theory, we saw that there were N real, nonnegative values of λ for which (47) had a nontrivial solution. Now the coordinate space is infinite and we have a homogeneous linear integral equation to solve.

The function $K_x(t, u)$ is called the kernel of the integral equation, and because it is a covariance function it is symmetric and nonnegative definite. We restrict our attention to processes with a finite mean-square value $[E(x^2(t)) < \infty]$. Their covariance functions satisfy the restriction

$$\int_0^T\int K_x^2(t, u)\, dt\, du \le \left[\int_0^T E[x^2(t)]\, dt\right]^2 < \infty, \tag{48}$$

where T is a finite number.

The restrictions in the last paragraph enable us to employ standard results from linear integral equation theory† (e.g., Courant and Hilbert [3], Chapter 3; Riesz and Nagy [4]; Lovitt [5]; or Tricomi [6]).

Properties of Integral Equations

1. There exist *at least one* square-integrable function $\phi(t)$ and real number $\lambda \neq 0$ that satisfy (46).

It is clear that there may not be more than one solution. For example,

$$K_x(t, u) = \sigma_f^2 f(t) f(u), \qquad 0 \le t, u \le T. \tag{49}$$

has only one nonzero eigenvalue and one normalized eigenfunction.

2. By looking at (46) we see that if $\phi_j(t)$ is a solution then $c\phi_j(t)$ is also a solution. Therefore we can always normalize the eigenfunctions.

† Here we follow Davenport and Root [1], p. 373.

3. If $\phi_1(t)$ and $\phi_2(t)$ are eigenfunctions associated with the same eigenvalue λ, then $c_1\phi_1(t) + c_2\phi_2(t)$ is also an eigenfunction associated with λ.

4. The eigenfunctions corresponding to different eigenvalues are orthogonal.

5. There is at most a countably infinite set of eigenvalues and all are bounded.

6. For any particular λ there is at most a *finite* number of linearly independent eigenfunctions. [Observe that we mean algebraic linear independence; $f(t)$ is linearly independent of the set $\phi_i(t)$, $i = 1, 2, \ldots, K$, if it cannot be written as a weighted sum of the $\phi_i(t)$.] These can always be orthonormalized (e.g., by the Gram-Schmidt procedure; see Problem 4.2.7 in Chapter 4).

7. Because $K_x(t, u)$ is nonnegative definite, the kernel $K_x(t, u)$ can be expanded in the series

$$K_x(t, u) = \sum_{i=1}^{\infty} \lambda_i \phi_i(t) \phi_i(u), \qquad 0 \leq t, u \leq T, \tag{50}$$

where the convergence is uniform for $0 \leq t, u \leq T$. (This is called Mercer's theorem.)

8. If $K_x(t, u)$ is positive definite, the eigenfunctions form a complete orthonormal set. From our results in Section 3.2 this implies that we can expand any deterministic function with finite energy in terms of the eigenfunctions.

9. If $K_x(t, u)$ is not positive definite, the eigenfunctions cannot form a complete orthonormal set. [This follows directly from (35) and (40).] Frequently, we augment the eigenfunctions with enough additional orthogonal functions to obtain a complete set. We occasionally refer to these additional functions as eigenfunctions with zero eigenvalues.

10. The sum of the eigenvalues is the expected value of the energy of the process in the interval $(0, T)$, that is,

$$E\left[\int_0^T x^2(t)\, dt\right] = \int_0^T K_x(t, t)\, dt = \sum_{i=1}^{\infty} \lambda_i.$$

(Recall that $x(t)$ is assumed to be zero-mean.)

These properties guarantee that we can find a set of $\phi_i(t)$ that leads to uncorrelated coefficients. It remains to verify the assumption that we made in (42). We denote the expected value of the error if $x(t)$ is approximated by the first N terms as $\xi_N(t)$:

$$\xi_N(t) \triangleq E\left[\left(x(t) - \sum_{i=1}^{N} x_i \phi_i(t)\right)^2\right]. \tag{51}$$

Evaluating the expectation, we have

$$\xi_N(t) = K_x(t, t) - 2E\left[x(t)\sum_{i=1}^{N} x_i\,\phi_i(t)\right] + E\left[\sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j\,\phi_i(t)\,\phi_j(t)\right], \tag{52}$$

$$\xi_N(t) = K_x(t, t) - 2E\left[x(t)\sum_{i=1}^{N}\left(\int_0^T x(u)\,\phi_i(u)\,du\right)\phi_i(t)\right] + \sum_{i=1}^{N}\lambda_i\,\phi_i(t)\,\phi_i(t), \tag{53}$$

$$\xi_N(t) = K_x(t, t) - 2\sum_{i=1}^{N}\left(\int_0^T K_x(t, u)\,\phi_i(u)\,du\right)\phi_i(t) + \sum_{i=1}^{N}\lambda_i\,\phi_i(t)\,\phi_i(t), \tag{54}$$

$$\xi_N(t) = K_x(t, t) - \sum_{i=1}^{N}\lambda_i\,\phi_i(t)\,\phi_i(t). \tag{55}$$

Property 7 guarantees that the sum will converge uniformly to $K_x(t, t)$ as $N \to \infty$. Therefore

$$\lim_{N\to\infty}\ \xi_N(t) = 0, \qquad 0 \le t \le T, \tag{56}$$

which is the desired result. (Observe that the convergence in Property 7 implies that for any $\epsilon > 0$ there exists an N_1 *independent of* t such that $\xi_N(t) < \epsilon$ for all $N > N_1$).

The series expansion we have developed in this section is generally referred to as the Karhunen-Loève expansion. (Karhunen [25], Loève [26], p. 478, and [30].) It provides a second-moment characterization in terms of uncorrelated random variables. This property, by itself, is not too important. In the next section we shall find that for a particular process of interest, the Gaussian random process, the coefficients in the expansion are statistically independent Gaussian random variables. It is in this case that the expansion finds its most important application.

3.3.3 Gaussian Processes

We now return to the question of a suitable complete characterization of a random process. We shall confine our attention to Gaussian random processes. To find a suitable definition let us recall how we defined jointly Gaussian random variables in Section 2.6. We said that the random variables $x_1, x_2, \ldots, x_N$ were jointly Gaussian if

$$y = \sum_{i=1}^{N} g_i x_i \tag{57}$$

was a Gaussian random variable for any set of g_i. In 2.6 N was finite and we required the g_i to be finite. If N is countably infinite, we require the g_i to be such that $E[y^2] < \infty$. In the random process, instead of a linear

transformation on a set of random variables, we are interested in a linear functional of a random function. This suggests the following definition:

Definition. Let $x(t)$ be a random process defined over some interval $[T_\alpha, T_\beta]$ with a mean-value $m_x(t)$ and covariance function $K_x(t, u)$. If every linear functional of $x(t)$ is a Gaussian random variable, then $x(t)$ is a Gaussian random process. In other words, if

$$y = \int_{T_\alpha}^{T_\beta} g(u)\, x(u)\, du, \tag{58}$$

and $g(u)$ is any function such that $E[y^2] < \infty$. Then, in order for $x(u)$ to be a Gaussian random process, y must be a Gaussian random variable for every $g(u)$ in the above class.

Several properties follow immediately from this definition.

Property 1. The output of a linear system is a particular linear functional of interest. We denote the impulse response as $h(t, u)$, the output at time t due to a unit impulse input at time u. If the input is $x(t)$ which is a sample function from a Gaussian random process, the output $y(t)$ is also.

Proof:

$$y(t) = \int_{T_\alpha}^{T_\beta} h(t, u)\, x(u)\, du, \qquad T_\gamma \leq t \leq T_\Delta. \tag{59}$$

The interval $[T_\gamma, T_\Delta]$ is simply the range over which $y(t)$ is defined. We assume that $h(t, u)$ is such that $E[y^2(t)] < \infty$ for all t in $[T_\gamma, T_\Delta]$. From the definition it is clear that y_t is a Gaussian random variable. To show that $y(t)$ is a Gaussian random process we must show that any linear functional of it is a Gaussian random variable. Thus,

$$z \triangleq \int_{T_\gamma}^{T_\Delta} g_y(t)\, y(t)\, dt, \tag{60}$$

or

$$z = \int_{T_\gamma}^{T_\Delta} g_y(t)\, dt \int_{T_\alpha}^{T_\beta} h(t, u)\, x(u)\, du, \tag{61}$$

must be Gaussian for every $g_y(t)$ [such that $E[z^2] < \infty$]. Integrating with respect to t and defining the result as

$$g(u) \triangleq \int_{T_\gamma}^{T_\Delta} g_y(t)\, h(t, u)\, dt, \tag{62}$$

we have

$$z = \int_{T_\alpha}^{T_\beta} g(u)\, x(u)\, du, \tag{63}$$

which is Gaussian by definition.

Thus we have shown that if the input to a linear system is a Gaussian random process the output is a Gaussian random process.

Property 2. If

$$y_1 = \int_{T_\alpha}^{T_\beta} g_1(u)\, x(u)\, du \tag{64}$$

and

$$y_2 = \int_{T_\alpha}^{T_\beta} g_2(u)\, x(u)\, du, \tag{65}$$

where $x(u)$ is a Gaussian random process, then y_1 and y_2 are jointly Gaussian. (The proof is obvious in light of (57).)

Property 3. If

$$x_i = \int_{T_\alpha}^{T_\beta} \phi_i(u)\, x(u)\, du \tag{66}$$

and

$$x_j = \int_{T_\alpha}^{T_\beta} \phi_j(u)\, x(u)\, du, \tag{67}$$

where $\phi_i(u)$ and $\phi_j(u)$ are orthonormalized eigenfunctions of (46) [now the interval of interest is (T_α, T_β) instead of $(0, T)$] then x_i and x_j are statistically independent Gaussian random variables ($i \neq j$). Thus,

$$p_{x_i}(X_i) = \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left[-\frac{(X_i - m_i)^2}{2\lambda_i}\right], \tag{68}$$

where

$$m_i \triangleq \int_{T_\alpha}^{T_\beta} m_x(t)\, \phi_i(t)\, dt. \tag{69}$$

This property follows from Property 2 and (45).

Property 4. For any set of times $t_1, t_2, t_3, \ldots, t_n$ in the interval $[T_\alpha, T_\beta]$ the random variables $x_{t_1}, x_{t_2}, \ldots, x_{t_n}$ are jointly Gaussian random variables.

Proof: If we denote the set by the vector $\mathbf{x}_t$,

$$\mathbf{x}_t \triangleq \begin{bmatrix} x_{t_1} \\ x_{t_2} \\ \vdots \\ x_{t_n} \end{bmatrix}, \tag{70}$$

whose mean is $\mathbf{m}_x$,

$$\mathbf{m}_x \triangleq E\begin{bmatrix} x_{t_1} \\ x_{t_2} \\ \vdots \\ x_{t_n} \end{bmatrix} = \begin{bmatrix} m_x(t_1) \\ m_x(t_2) \\ \vdots \\ m_x(t_n) \end{bmatrix}, \tag{71}$$

then the joint probability density is

$$p_{\mathbf{x}_t}(\mathbf{X}) = [(2\pi)^{n/2}|\mathbf{\Lambda}_\mathbf{x}|^{1/2}]^{-1} \exp\left[-\tfrac{1}{2}(\mathbf{X} - \mathbf{m}_\mathbf{x})^T\mathbf{\Lambda}_\mathbf{x}^{-1}(\mathbf{X} - \mathbf{m}_\mathbf{x})\right] \tag{72}$$

and the joint characteristic function is

$$M_{\mathbf{x}_t}(j\mathbf{v}) = \exp\left(j\mathbf{v}^T\mathbf{m}_\mathbf{x} - \tfrac{1}{2}\mathbf{v}^T\mathbf{\Lambda}_\mathbf{x}\mathbf{v}\right), \tag{73}$$

where $\mathbf{\Lambda}_\mathbf{x}$ is the covariance matrix of the random variables $x_{t_1}, x_{t_2}, \ldots, x_{t_n}$. (We assume $\mathbf{\Lambda}_\mathbf{x}$ is nonsingular.) The ij element is

$$\mathbf{\Lambda}_{\mathbf{x},ij} = E[(x_{t_i} - m_x(t_i))(x_{t_j} - m_x(t_j))]. \tag{74}$$

This property follows by using the function

$$g(u) = \sum_{i=1}^{n} g_i\,\delta(u - t_i) \tag{75}$$

in (58) and the result in (57). Thus we see that our definition has the desirable property suggested in Section 3.3.1, for it uniquely specifies the joint density at any set of times. Frequently Property 4 is used as the basic definition. The disadvantage of this approach is that it is more difficult to prove that our definition and Properties 1–3 follow from (72) than vice-versa.

The Gaussian process we have defined has two main virtues:

1. The physical mechanisms that produce many processes are such that a Gaussian model is appropriate.
2. The Gaussian process has many properties that make analytic results feasible.

Discussions of physical mechanisms that lead logically to Gaussian processes are available in [7] and [8]. Other properties of the Gaussian process which are not necessary for our main discussion, are developed in the problems (cf. Problems 3.3.12–3.3.18).

We shall encounter multiple processes that are jointly Gaussian. The definition is a straightforward extension of the preceding one.

Definition. Let $x_1(t), x_2(t), \ldots, x_N(t)$ be a set of random processes defined over the intervals $(T_{\alpha_1}, T_{\beta_1}), (T_{\alpha_2}, T_{\beta_2}), \ldots, (T_{\alpha_N}, T_{\beta_N})$, respectively. If every sum of arbitrary functionals of $x_i(t)$, $i = 1, \ldots, N$, is a Gaussian random variable, then the processes $x_1(t), x_2(t), \ldots, x_N(t)$ are defined to be jointly Gaussian random processes. In other words,

$$y = \sum_{i=1}^{N} \int_{T_{\alpha_i}}^{T_{\beta_i}} g_i(u)\, x_i(u)\, du$$

must be Gaussian for every set of $g_i(u)$ such that $E[y^2] < \infty$.

Other properties of jointly Gaussian processes are discussed in the problems.

Property 3 is the reason for our emphasis on the Karhunen-Loève expansion. It enables us to characterize a Gaussian process in terms of an at most countably infinite set of statistically independent Gaussian random variables. The significance of this will perhaps be best appreciated when we see how easy it makes our ensuing work. Observe that if we had chosen to emphasize Markov processes the orthogonal expansion method of characterization would not have been particularly useful. In Section 6.3 we discuss characterizations that emphasize the Markovian structure.

The Karhunen-Loève expansion is useful in two ways:

1. Many of our theoretical derivations use it as a tool. In the majority of these cases the eigenfunctions and eigenvalues do not appear in the final result. The integral equation that specifies them (46) need never be solved.
2. In other cases the result requires an explicit solution for one or more eigenfunctions and eigenvalues. Here we must be able to solve the equation exactly or find good approximate solutions.

In the next section we consider some useful situations in which solutions can be obtained.

3.4 HOMOGENEOUS INTEGRAL EQUATIONS AND EIGENFUNCTIONS

In this section we shall study in some detail the behavior of the solutions to (46). In addition to the obvious benefit of being able to solve for an eigenfunction when it is necessary, the discussion serves several other purposes:

1. By looking at several typical cases and finding the eigenvalues and eigenfunctions the idea of a coordinate expansion becomes somewhat easier to visualize.
2. In many cases we shall have to make approximations to get to the final result. We need to develop some feeling for what can be neglected and what is important.
3. We want to relate the behavior of the eigenvalues and eigenfunctions to more familiar ideas such as the power density spectrum.

In Section 3.4.1 we illustrate a technique that is useful whenever the random process is stationary and has a rational power density spectrum. In Section 3.4.2 we consider bandlimited stationary processes, and in Section 3.4.3 we look at an important nonstationary process. Next in Section 3.4.4 we introduce the idea of a "white" process. In Section 3.4.5, we derive the optimum linear filter for estimating a message corrupted by

noise. Finally, in Section 3.4.6, we examine the asymptotic behavior of the eigenfunctions and eigenvalues for large time intervals.

3.4.1 Rational Spectra

The first set of random processes of interest are stationary and have spectra that can be written as a ratio of two polynomials in ω^2.

$$S_x(\omega) = \frac{N(\omega^2)}{D(\omega^2)}, \tag{76}$$

where $N(\omega^2)$ is a polynomial of order q in ω^2 and $D(\omega^2)$ is a polynomial of order p in ω^2. Because we assume that $x(t)$ has a finite mean-square value, $q < p$. We refer to these spectra as rational. There is a routine but tedious method of solution. The basic idea is straightforward. We convert the integral equation to a differential equation whose solution can be easily found. Then we substitute the solution back into the integral equation to satisfy the boundary conditions. We first demonstrate the technique by considering a simple example and then return to the general case and formalize the solution procedure. (Detailed discussions of similar problems are contained in Slepian [9], Youla [10], Davenport and Root [1], Laning and Battin [11], Darlington [12], Helstrom [13, 14], or Zadeh and Ragazzini [22].)

Example. Let

$$S_x(\omega) = \frac{2\alpha P}{\omega^2 + \alpha^2}, \qquad -\infty < \omega < \infty, \tag{77}$$

or

$$R_x(\tau) = P \exp(-\alpha|\tau|), \qquad -\infty < \tau < \infty. \tag{78}$$

The mean-square value of $x(t)$ is P. The integral equation of interest is

$$\int_{-T}^{T} P \exp(-\alpha|t - u|)\phi(u)du = \lambda\,\phi(t), \qquad -T \le t \le T. \tag{79}$$

(The algebra becomes less tedious with a symmetric interval.)

As indicated above, we solve the integral equation by finding the corresponding differential equation, solving it, and substituting it back into the integral equation. First, we rewrite (79) to eliminate the magnitude sign.

$$\lambda\,\phi(t) = \int_{-T}^{t} P \exp[-\alpha(t - u)]\,\phi(u)du + \int_{t}^{T} P \exp[-\alpha(u - t)]\,\phi(u)du. \tag{80}$$

Differentiating once, we have

$$\lambda\,\dot{\phi}(t) = -P\alpha e^{-\alpha t}\int_{-T}^{t} e^{+\alpha u}\,\phi(u)du + P\alpha e^{+\alpha t}\int_{t}^{T} e^{-\alpha u}\,\phi(u)du. \tag{81}$$

Differentiating a second time gives

$$\lambda \ddot{\phi}(t) = P\alpha^2 \int_{-T}^{T} e^{-\alpha|t-u|}\, \phi(u)\, du - 2P\alpha\, \phi(t); \tag{82}$$

but the first term on the right-hand side is just $\alpha^2 \lambda\, \phi(t)$. Therefore

$$\lambda \ddot{\phi}(t) = \alpha^2 \lambda\, \phi(t) - 2P\alpha\, \phi(t) \tag{83}$$

or, for $\lambda \neq 0$,

$$\ddot{\phi}(t) = \frac{\alpha^2(\lambda - 2P/\alpha)}{\lambda}\, \phi(t). \tag{84}$$

The solution to (83) has four possible forms corresponding to

$$\begin{aligned} &\text{(i)} && \lambda = 0; \\ &\text{(ii)} && 0 < \lambda < \frac{2P}{\alpha}; \\ &\text{(iii)} && \lambda = \frac{2P}{\alpha}; \\ &\text{(iv)} && \lambda > \frac{2P}{\alpha}. \end{aligned} \tag{85}$$

We can show that the integral equation cannot be satisfied for (i), (iii), and (iv). (Cf. Problem 3.4.1.)
For (ii) we may write

$$b^2 = \frac{-\alpha^2(\lambda - 2P/\alpha)}{\lambda}, \qquad 0 < b^2 < \infty. \tag{86}$$

Then

$$\phi(t) = c_1 e^{jbt} + c_2 e^{-jbt}. \tag{87}$$

Substituting (87) into (80) and performing the integration, we obtain

$$0 = e^{-\alpha t}\left[\frac{c_1 e^{-(\alpha + jb)T}}{\alpha + jb} + \frac{c_2 e^{-(\alpha - jb)T}}{\alpha - jb}\right] - e^{+\alpha t}\left[\frac{c_1 e^{-(\alpha - jb)T}}{-\alpha + jb} + \frac{c_2 e^{-(\alpha + jb)T}}{-\alpha - jb}\right]. \tag{88}$$

We can easily verify that if $c_1 \neq \pm c_2$, (88) cannot be satisfied for all time. For $c_1 = -c_2$ we require that $\tan bT = -b/\alpha$. For $c_1 = c_2$ we require $\tan bT = \alpha/b$. Combining these two equations, we have

$$\left(\tan bT + \frac{b}{\alpha}\right)\left(\tan bT - \frac{\alpha}{b}\right) = 0. \tag{89}$$

The values of b that satisfy (89) can be determined graphically as shown in Fig. 3.8. The upper set of intersections correspond to the second term in (89) and the lower set to the first term. The corresponding eigenvalues are

$$\lambda_i = \frac{2P\alpha}{\alpha^2 + b_i^2}, \qquad i = 1, 2, \ldots. \tag{90}$$

Observe that we have ordered the solutions to (89), $b_1 < b_2 < b_3 < \cdots$. From (90) we see that this orders the eigenvalues $\lambda_1 > \lambda_2 > \lambda_3 \cdots$. The odd-numbered solutions correspond to $c_1 = c_2$ and therefore

$$\phi_i(t) = \frac{1}{T^{1/2}\left(1 + \dfrac{\sin 2b_i T}{2b_i T}\right)^{1/2}} \cos b_i t, \qquad -T \leq t \leq T \qquad (i \text{ odd}). \tag{91}$$

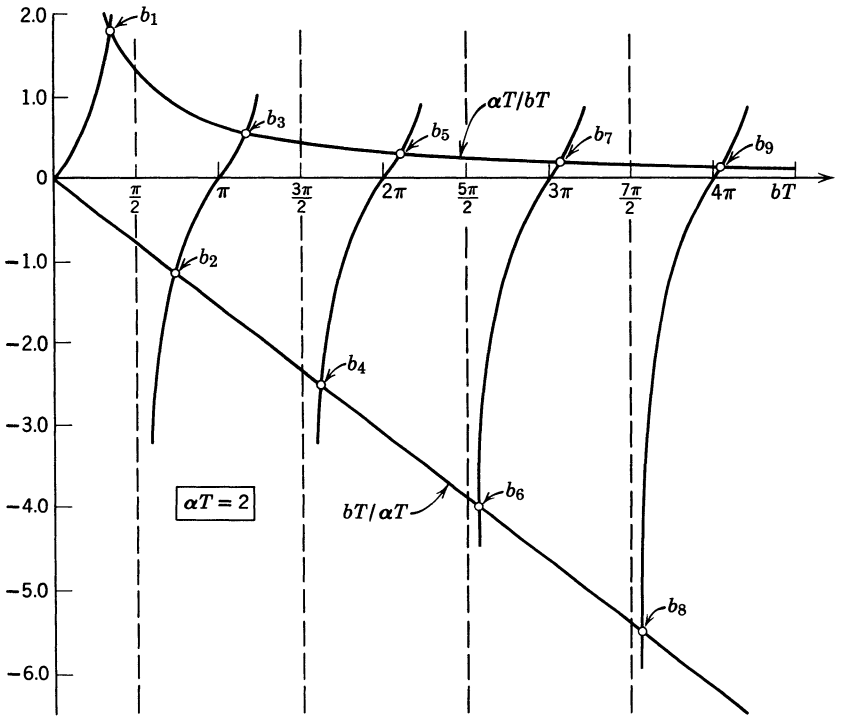

Fig. 3.8 Graphical solution of transcendental equation.

The even-numbered solutions correspond to $c_1 = -c_2$ and therefore

$$\phi_i(t) = \frac{1}{T^{1/2}\left(1 - \dfrac{\sin 2b_iT}{2b_iT}\right)^{1/2}} \sin b_i t, \qquad -T \leq t \leq T \qquad (i \text{ even}). \tag{92}$$

We see that the eigenfunctions are cosines and sines whose frequencies are *not* harmonically related.

Several interesting observations may be made with respect to this example:

1. The eigenvalue corresponding to a particular eigenfunction is equal to the height of the power density spectrum at that frequency.
2. As T increases, b_n decreases monotonically and therefore λ_n increases monotonically.
3. As bT increases, the upper intersections occur at approximately $(i - 1)\,\pi/2$ [i odd] and the lower intersections occur at approximately

$(i-1)\pi/2$ [i even]. From (91) and (92) we see that the higher index eigenfunctions are approximately a set of periodic sines and cosines.

$$\phi_i(t) \cong \begin{cases} \dfrac{1}{T^{1/2}\left(1+\dfrac{\sin 2b_iT}{2b_iT}\right)^{1/2}} \cos\left[\dfrac{(i-1)\pi}{2T}t\right], & -T \le t \le T \quad (i \text{ odd}). \\ \dfrac{1}{T^{1/2}\left(1-\dfrac{\sin 2b_iT}{2b_iT}\right)^{1/2}} \sin\left[\dfrac{(i-1)\pi}{2T}t\right], & -T \le t \le T \quad (i \text{ even}). \end{cases}$$

This behavior is referred to as the asymptotic behavior.

The first observation is not true in general. In a later section (p. 204) we shall show that the λ_n are always monotonically increasing functions of T. We shall also show that the asymptotic behavior seen in this example is typical of stationary processes.

Our discussion up to this point has dealt with a particular spectrum. We now return to the general case.

It is easy to generalize the technique to arbitrary rational spectra. First we write $S_x(\omega)$ as a ratio of two polynomials,

$$S_x(\omega) = \frac{N(\omega^2)}{D(\omega^2)}. \tag{93}$$

Looking at (83), we see that the differential equation does *not* depend explicitly on T. This independence is true whenever the spectrum has the form in (93). Therefore we would obtain the same differential equation if we started with the integral equation

$$\lambda\,\phi(t) = \int_{-\infty}^{\infty} K_x(t-u)\,\phi(u)\,du, \qquad -\infty < t < \infty. \tag{94}$$

By use of Fourier transforms a formal solution to this equation follows immediately:

$$\lambda\Phi(j\omega) = S_x(\omega)\,\Phi(j\omega) = \frac{N(\omega^2)}{D(\omega^2)}\,\Phi(j\omega) \tag{95}$$

or

$$0 = [\lambda D(\omega^2) - N(\omega^2)]\,\Phi(j\omega). \tag{96}$$

There are $2p$ homogeneous solutions to the differential equation corresponding to (96) for every value of λ (corresponding to the roots of the polynomial in the bracket). We denote them as $\phi_{h_i}(t, \lambda)$, $i = 1, \ldots, 2p$. To find the solution to (46) we substitute

$$\phi(t) = \sum_{i=1}^{2p} a_i \phi_{h_i}(t, \lambda) \tag{97}$$

into the integral equation and solve for those values of λ and a_i that lead

to a solution. There are no conceptual difficulties, but the procedure is tedious.†

One particular family of spectra serves as a useful model for many physical processes and also leads to tractable solutions to the integral equation for the problem under discussion. This is the family described by the equation

$$S_x(\omega:n) = \left(\frac{2nP}{\alpha}\right)\frac{\sin(\pi/2n)}{1 + (\omega/\alpha)^{2n}}. \tag{98}$$

It is referred to as the Butterworth family and is shown in Figure 3.9. When $n = 1$, we have the simple one-pole spectrum. As n increases, the attenuation versus frequency for $\omega > \alpha$ increases more rapidly. In the

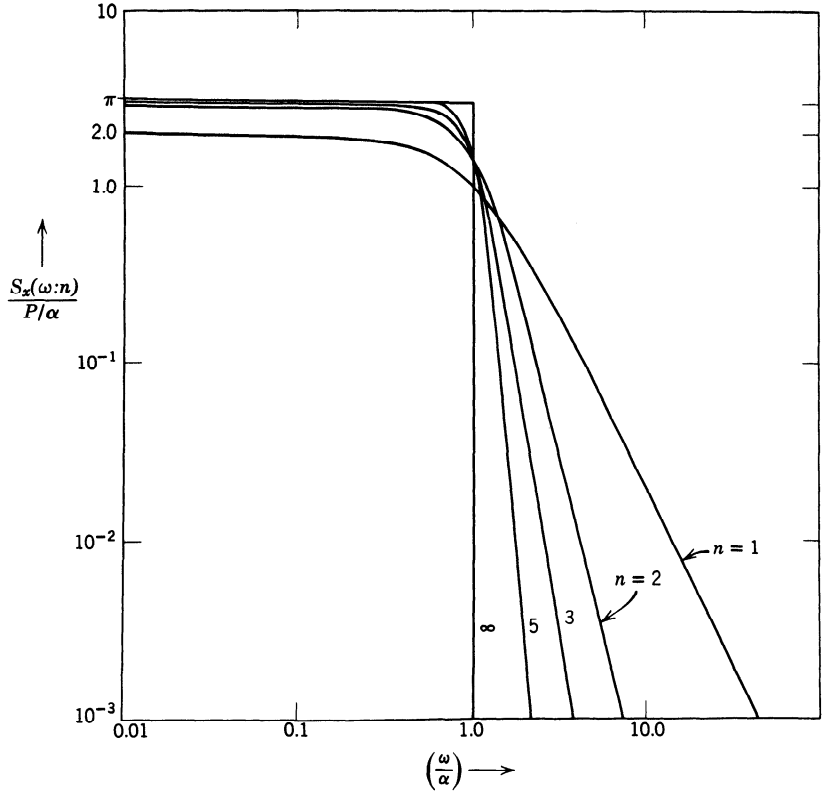

Fig. 3.9 Butterworth spectra.

† In Appendix I of Part II we shall develop a technique due to A. Baggeroer [32, 33] that is more efficient. At the present point in our discussion we lack the necessary background for the development.

limit, as $n \to \infty$, we have an ideal bandlimited spectrum. In the next section we discuss the eigenfunctions and eigenvalues for the bandlimited spectrum.

3.4.2 Bandlimited Spectra

When the spectrum is not rational, the differential equation corresponding to the integral equation will usually have time-varying coefficients. Fortunately, in many cases of interest the resulting differential equation is some canonical type whose solutions have been tabulated. An example in this category is the bandlimited spectrum shown in Fig. 3.10. In this case

$$S_x(\omega) = \begin{cases} \dfrac{\pi P}{\alpha}, & |\omega| \leq \alpha, \\ 0, & |\omega| > \alpha, \end{cases} \tag{99}$$

or, in cycles per second,

$$S_x(\omega) = \begin{cases} \dfrac{P}{2W}, & |f| \leq W, \\ 0, & |f| > W, \end{cases} \tag{100}$$

where

$$2\pi W = \alpha. \tag{101}$$

The corresponding covariance function is

$$K_x(t, u) = P \frac{\sin \alpha(t - u)}{\alpha(t - u)}. \tag{102}$$

The integral equation of interest becomes

$$\lambda \phi(t) = \int_{-T/2}^{+T/2} P \frac{\sin \alpha(t - u)}{\alpha(t - u)} \phi(u)\, du. \tag{103}$$

[This is just (46) with the interval shifted to simplify notation.]

Once again the procedure is to find a related differential equation and to examine its solution. We are, however, more interested in the results

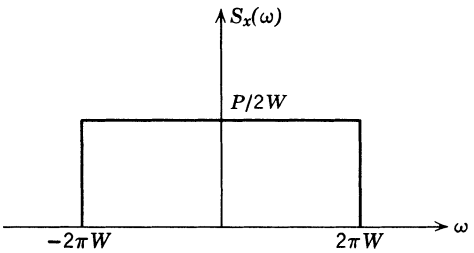

Fig. 3.10 Bandlimited spectrum.

than the detailed techniques; therefore we merely state them ([9], [15], [16], [17], and [18] are useful for further study).

The related differential equation over a normalized interval is

$$(1 - t^2)\ddot{f}(t) - 2t\dot{f}(t) + (\mu - c^2t^2)f(t) = 0, \qquad -1 < t < 1, \quad (104)$$

where

$$c = \frac{\alpha T}{2} = \pi WT \quad (105)$$

and μ is the eigenvalue. This equation has continuous solutions for certain values of $\mu(c)$. These solutions are called *angular prolate spheroidal functions* and are denoted by $S_{0n}(c, t)$, $n = 0, 1, 2, \ldots$. A plot of typical $S_{0n}(c, t)$ is contained in [15], [16], and [17]. These functions also satisfy the integral equation

$$2[R_{0n}^{(1)}(c, 1)]^2 S_{0n}(c, t) = \int_{-1}^{+1} \frac{\sin c(t - u)}{c(t - u)} S_{0n}(c, u)\, du, \qquad -1 \leq t \leq 1, \quad (106)$$

or changing variables

$$PT\left[R_{0n}^{(1)}\left(\frac{\alpha T}{2}, 1\right)\right]^2 S_{0n}\left(\frac{\alpha T}{2}, \frac{2t}{T}\right) = \int_{-T/2}^{T/2} \frac{P \sin \alpha(t - u)}{\alpha(t - u)} S_{0n}\left(\frac{\alpha T}{2}, \frac{2u}{T}\right) du, \qquad -\frac{T}{2} \leq t \leq \frac{T}{2}, \quad (107)$$

where $R_{0n}^{(1)}(\alpha T/2, 1)$ is a *radial prolate spheroidal function.* Thus the eigenvalues are

$$\lambda_n = PT\left[R_{0n}^{(1)}\left(\frac{\alpha T}{2}, 1\right)\right]^2, \qquad n = 0, 1, 2, \ldots.$$

These functions are tabulated in several references (e.g. [18] or [19]).

The first several eigenvalues for various values of WT are shown in Figs. 3.11 and 3.12. We observe a very interesting phenomenon. For values of $n > (2WT + 1)$ the values of λ_n rapidly approach zero. We can check the total energy in the remaining eigenvalues, for

$$\sum_{i=0}^{\infty} \lambda_i = \int_{-T/2}^{+T/2} K_x(t, t)\, dt = PT. \quad (108)$$

In Fig. 3.11, $2WT = 2.55$ and the first four eigenvalues sum to (2.54/2.55) PT. In Fig. 3.12, $2WT = 5.10$ and the first six eigenvalues sum to (5.09/5.10)PT. This behavior is discussed in detail in [17]. Our example suggests that the following statement is plausible. When a bandlimited process $[-W, W$ cps] is observed over a T-second interval, there are only $(2TW + 1)$ significant eigenvalues. This result will be important to us in later chapters (specifically Chapters II-2 and II-3) when we obtain approximate solutions

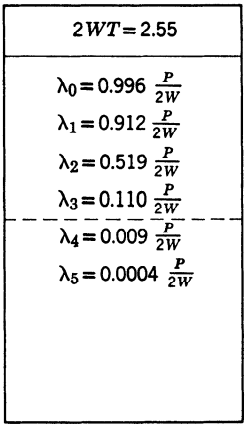

Fig. 3.11 Eigenvalues for a bandlimited spectrum ($2WT = 2.55$).

by neglecting the higher eigenfunctions. More precise statements about the behavior are contained in [15], [16], and [17].

3.4.3 Nonstationary Processes

The process of interest is the simple Wiener process. It was developed as a model for Brownian motion and is discussed in detail in [20] and [21]. A typical sample function is shown in Fig. 3.13.

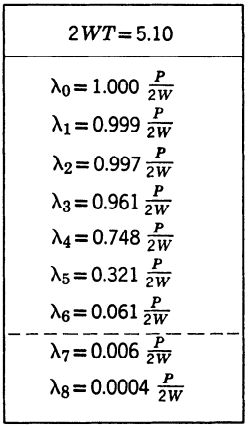

Fig. 3.12 Eigenvalues of a bandlimited spectrum ($2WT = 5.10$).

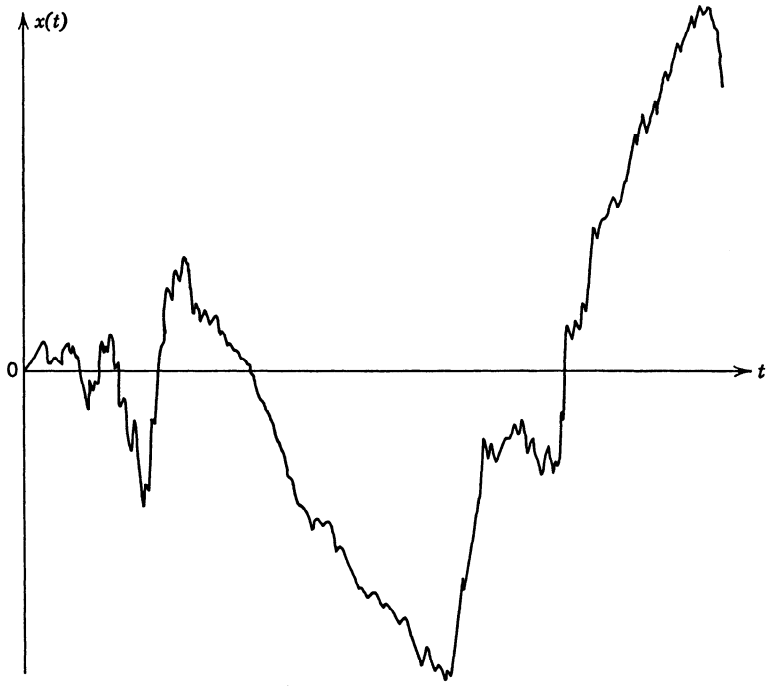

Fig. 3.13 Sample function of a Wiener process.

This process is defined for $t \geq 0$ and is characterized by the following properties:

$$x(0) = 0,$$

$$E[x(t)] = 0, \tag{109}$$

$$E[x^2(t)] = \sigma^2 t, \tag{110}$$

$$p_{x_t}(X_t) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left(-\frac{X_t^2}{2\sigma^2 t}\right). \tag{111}$$

The increment variables are independent; that is, if $t_3 > t_2 > t_1$, then $(x_{t_3} - x_{t_2})$ and $(x_{t_2} - x_{t_1})$ are statistically independent. In the next example we solve (46) for the Wiener process.

Example. Wiener Process. Using the properties of the Wiener process, we can show that

$$K_x(t, u) = \sigma^2 \min(u,t) = \begin{cases} \sigma^2 u, & u \leq t, \\ \sigma^2 t & t \leq u, \end{cases} \tag{112}$$

In this case (46) becomes

$$\lambda \phi(t) = \int_0^T K_x(t, u)\, \phi(u)\, du, \qquad 0 \leq t \leq T. \tag{113}$$

Substituting (112) into (113)

$$\lambda \phi(t) = \sigma^2 \int_0^t u \phi(u)\, du + \sigma^2 t \int_t^T \phi(u)\, du. \tag{114}$$

Proceeding as in Section 3.4.1, we differentiate (114) and obtain,

$$\lambda \dot{\phi}(t) = \sigma^2 \int_t^T \phi(u)\, du. \tag{115}$$

Differentiating again, we obtain

$$\lambda \ddot{\phi}(t) = -\sigma^2 \phi(t), \tag{116}$$

or, for $\lambda \neq 0$,

$$\ddot{\phi}(t) + \frac{\sigma^2}{\lambda} \phi(t) = 0. \tag{117}$$

There are three possible ranges for λ:

$$\begin{array}{ll} \text{(i)} & \lambda < 0, \\ \text{(ii)} & \lambda = 0, \\ \text{(iii)} & \lambda > 0. \end{array} \tag{118}$$

We can easily verify (cf. Problem 3.4.3) that (i) and (ii) do not provide solutions that will satisfy the integral equation. For $\lambda > 0$ we proceed exactly as in the preceding section and find

$$\lambda_n = \frac{\sigma^2 T^2}{(n - \frac{1}{2})^2 \pi^2}, \qquad n = 1, 2, \cdots \tag{119}$$

and

$$\phi_n(t) = \left(\frac{2}{T}\right)^{1/2} \sin\left[\left(n - \frac{1}{2}\right)\frac{\pi}{T} t\right] \qquad 0 \leq t \leq T. \tag{120}$$

Once again the eigenfunctions are sinusoids.

The Wiener process is important for several reasons.

1. A large class of processes can be transformed into the Wiener process
2. A large class of processes can be generated by passing a Wiener process through a linear or nonlinear system. (We discuss this in detail later.)

3.4.4 White Noise Processes

Another interesting process can be derived from the Wiener process. Using (41) and (120) we can expand $x(t)$ in a series.

$$x(t) = \underset{K \to \infty}{\text{l.i.m.}} \sum_{n=1}^{K} x_n \left(\frac{2}{T}\right)^{1/2} \sin\left[\left(n - \frac{1}{2}\right)\frac{\pi}{T} t\right], \tag{121}$$

where the mean-square value of the coefficient is given by (119):

$$E[x_n^2] = \frac{\sigma^2 T^2}{(n - \frac{1}{2})^2 \pi^2}. \tag{122}$$

We denote the K-term approximation as $x_K(t)$.

Now let us determine what happens when we differentiate $x_K(t)$:

$$\dot{x}_K(t) = \sum_{n=1}^{K} x_n\left(n - \frac{1}{2}\right)\frac{\pi}{T}\left\{\left(\frac{2}{T}\right)^{1/2} \cos\left[\left(n - \frac{1}{2}\right)\frac{\pi}{T}t\right]\right\}. \tag{123}$$

We see that the time function inside the braces is still normalized. Thus we may write

$$\dot{x}_K(t) = \sum_{n=1}^{K} w_n\left(\frac{2}{T}\right)^{1/2} \cos\left(n - \frac{1}{2}\right)\frac{\pi}{T}t, \tag{124}$$

where

$$E(w_n^2) = \sigma^2.$$

We observe that we have generated a process in which every eigenvalue is *equal.* Clearly, if we let $K \to \infty$, the series would not converge. If it did, it would correspond to a process with infinite energy over $[0, T]$.

We can *formally* obtain the covariance function of this resulting process by differentiating $K_x(t, u)$:

$$K_{\dot{x}}(t, u) = \frac{\partial^2}{\partial t\, \partial u} K_x(t, u) = \frac{\partial^2}{\partial t\, \partial u}[\sigma^2 \min(t, u)]$$

$$= \sigma^2 \delta(t - u), \qquad 0 \le t, u \le T. \tag{125}$$

We see that the covariance function is an impulse. Still proceeding formally, we can look at the solution to the integral equation (46) for an impulse covariance function:

$$\lambda \phi(t) = \sigma^2 \int_0^T \delta(t - u)\, \phi(u)\, du, \qquad 0 < t < T. \tag{126}$$

The equation is satisfied for any $\phi(t)$ with $\lambda = \sigma^2$. Thus *any* set of orthonormal functions is suitable for decomposing this process. The reason for the nonuniqueness is that the impulse kernel is not square-integrable. The properties stated on pp. 180–181 assumed square-integrability.

We shall find the resulting process to be a useful artifice for many models. We summarize its properties in the following definitions.

Definition. A *Gaussian* white noise process is a Gaussian process whose covariance function is $\sigma^2\, \delta(t - u)$. It may be decomposed over the interval $[0, T]$ by using *any* set of orthonormal functions $\phi_i(t)$. The coefficients along *each* coordinate function are statistically independent Gaussian variables with equal variance σ^2.

Some related notions follow easily.

Property. We can write, formally

$$\sigma^2 \delta(t-u) = \sum_{i=1}^{\infty} \sigma^2 \phi_i(t) \phi_i(u), \qquad 0 \leq t, u \leq T. \tag{127}$$

or, equivalently,

$$\delta(t-u) = \sum_{i=1}^{\infty} \phi_i(t) \phi_i(u), \qquad 0 \leq t, u \leq T. \tag{128}$$

Property. If the coefficients are uncorrelated, with equal variances, but *not* Gaussian, the process is referred to as a white process.

Property. If the process is defined over the infinite interval, its spectrum is

$$S_x(\omega) = \sigma^2; \tag{129}$$

that is, it is constant over all frequencies. The value of each eigenvalue corresponds to the spectral height σ^2.

The utility of a white noise process is parallel to that of an impulse input in the analysis of linear systems. Just as we can observe an impulse only after it has been through a system with some finite bandwidth, we can observe white noise only after it has passed through a similar system. Therefore, as long as the bandwidth of the noise is appreciably larger than that of the system, it can be considered as having an infinite bandwidth.

To illustrate a typical application of eigenfunction expansions we consider a simple problem.

3.4.5 The Optimum Linear Filter

In this section we consider the problem of trying to estimate a message in the presence of interfering noise. Our treatment at this point is reasonably brief. We return to this problem and study it in detail in Chapter 6. Here we have three objectives in mind:

1. The introduction of time-varying linear filters and simple minimization techniques.
2. The development of a specific result to be used in subsequent chapters; specifically, the integral equation whose solution is the optimum linear filter.
3. The illustration of how the orthogonal expansion techniques we have just developed will enable us to obtain a formal solution to an integral equation.

The system of interest is shown in Fig. 3.14. The message $a(t)$ is a sample function from a zero-mean random process with a finite mean-square value and a covariance function $K_a(t, u)$. It is corrupted by an uncorrelated

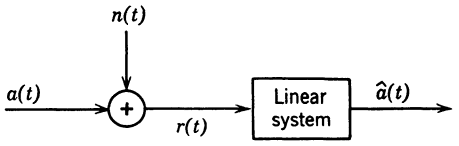

Fig. 3.14 Linear filter problem.

additive zero-mean noise $n(t)$ with covariance function $K_n(t, u)$. We observe the sum of these two processes,

$$r(t) = a(t) + n(t), \qquad 0 \leq t \leq T. \tag{130}$$

We pass $r(t)$ through a linear filter to obtain an estimate of $a(t)$ denoted by $\hat{a}(t)$.

Because $a(t)$ is not necessarily stationary and the observation interval is finite, we anticipate that to obtain the best estimate we may require a time-varying filter. We characterize the filter by its impulse response $h(t, u)$, which is the value of the output at time t when the input is an impulse at time u. If the system is physically realizable, then

$$h(t, u) = 0, \qquad t < u,$$

for the output cannot precede the input. If the system is time-invariant, then $h(t, u)$ depends only on the difference $(t - u)$. We assume that $r(t)$ equals zero for $t < 0$ and $t > T$. Because the system is linear, the output due to $r(t)$, $0 \leq t \leq T$, can be written as

$$\hat{a}(t) = \int_0^T h(t, u)\, r(u)\, du, \tag{131}$$

which is an obvious generalization of the convolution integral.

We want to choose $h(t, u)$ to minimize the mean of the squared error integrated over the interval $[0, T]$. In other words, we want to choose $h(t, u)$ to minimize the quantity

$$\begin{aligned} \xi_I &\triangleq E\left\{\frac{1}{T}\int_0^T [a(t) - \hat{a}(t)]^2\, dt\right\} \\ &= E\left\{\frac{1}{T}\int_0^T \left[a(t) - \int_0^T h(t, u)\, r(u)\, du\right]^2 dt\right\}. \end{aligned} \tag{132}$$

Thus we are minimizing the mean-square error integrated over the interval. We refer to ξ_I as the *interval estimation error.*

Similarly, we can define a *point estimation error*:

$$\xi_P(t) = E\left\{\left[a(t) - \int_0^T h(t, u)\, r(u)\, du\right]^2\right\}, \qquad 0 \leq t \leq T. \tag{133}$$

Clearly, if we minimize the point error at each time, the total interval

error will be minimized. One way to solve this minimization problem is to use standard variational techniques (e.g., [31], Chapter 2). Our approach is less formal and leads directly to a necessary and sufficient condition. We require the filter $h(t, u)$ to be a continuous function in both variables over the area $0 \leq t, u \leq T$ and denote the $h(t, u)$ that minimizes $\xi_P(t)$ as $h_o(t, u)$. Any other filter function $h(t, u)$ in the allowed class can be written as

$$h(t, u) = h_o(t, u) + \epsilon h_\epsilon(t, u), \qquad 0 \leq t, u \leq T, \tag{134}$$

where ϵ is a real parameter and $h_\epsilon(t, u)$ is in the allowable class of filters. Taking the expectation of (133), substituting (134) into the result, and grouping terms according to the power of ϵ, we obtain

$$\begin{aligned}\xi_P(t:\epsilon) = {} & K_a(t, t) - 2\int_0^T h(t, u)\, K_a(t, u)\, du \\ & + \int_0^T dv \int_0^T du\, h(t, v)\, h(t, u)\, K_r(u, v)\end{aligned} \tag{135}$$

or

$$\begin{aligned}\xi_P(t:\epsilon) = {} & K_a(t, t) - 2\int_0^T h_o(t, u)\, K_a(t, u)\, du \\ & + \int_0^T dv \int_0^T du\, h_o(t, u)\, h_o(t, v)\, K_r(u, v) \\ & - 2\epsilon \int_0^T du\, h_\epsilon(t, u)\left[K_a(t, u) - \int_0^T h_o(t, v)\, K_r(u, v)\, dv\right] \\ & + \epsilon^2 \int_0^T\int h_\epsilon(t, v)\, h_\epsilon(t, u)\, K_r(u, v)\, du\, dv.\end{aligned} \tag{136}$$

If we denote the first three terms as $\xi_{P_o}(t)$ and the last two terms as $\Delta\xi(t:\epsilon)$, then (136) becomes

$$\xi_P(t:\epsilon) = \xi_{P_o}(t) + \Delta\xi(t:\epsilon). \tag{137}$$

Now, if $h_o(t, u)$ is the optimum filter, then $\Delta\xi(t:\epsilon)$ must be greater than or equal to zero for all allowable $h_\epsilon(t, u)$ and all $\epsilon \neq 0$. We show that a necessary and sufficient condition for this to be true is that

$$\boxed{K_a(t, u) - \int_0^T h_o(t, v)\, K_r(u, v)\, dv = 0, \qquad \begtab{l} 0 \leq t \leq T \\ 0 < u < T. \end{tab}} \tag{138}$$

The equation for $\Delta\xi(t:\epsilon)$ is

$$\begin{aligned}\Delta\xi(t:\epsilon) = {} & -2\epsilon \int_0^T du\, h_\epsilon(t, u)\left[K_a(t, u) - \int_0^T h_o(t, v)\, K_r(u, v)\, dv\right] \\ & + \epsilon^2 \int_0^T\int h_\epsilon(t, v)\, h_\epsilon(t, u)\, K_r(u, v)\, du\, dv.\end{aligned} \tag{139}$$

Three observations are needed:

1. The second term is nonnegative for any choice of $h_\epsilon(t, v)$ and ϵ because $K_r(t, u)$ is nonnegative definite.
2. Unless

$$\int_0^T h_\epsilon(t, u)\left[K_a(t, u) - \int_0^T h_o(t, v)\, K_r(u, v)\, dv\right] du = 0, \tag{140}$$

there exists for every continuous $h_\epsilon(t, u)$ a range of values of ϵ that will cause $\Delta\xi(t:\epsilon)$ to be negative. Specifically, $\Delta\xi(t:\epsilon) < 0$ for all

$$0 < \epsilon < \frac{2\int_0^T h_\epsilon(t, u)\left[K_a(t, u) - \int_0^T h_o(t, v)\, K_r(u, v)\, dv\right] du}{\int_0^T\int h_\epsilon(t, v)\, h_\epsilon(t, u)\, K_r(u, v)\, du\, dv} \tag{141}$$

if the numerator on the right side of (141) is positive. $\Delta\xi(t:\epsilon)$ is negative for all negative ϵ greater than the right side of (141) if the numerator is negative.

3. In order that (140) may hold, it is necessary and sufficient that the term in the bracket be identically zero for all $0 < u < T$. Thus

$$K_a(t, u) - \int_0^T h_o(t, v)\, K_r(u, v)\, dv = 0, \qquad \begin{matrix} 0 \le t \le T \\ 0 < u < T. \end{matrix} \tag{142}$$

The inequality on u is strict if there is a white noise component in $r(t)$ because the second term is discontinuous at $u = 0$ and $u = T$. If (142) is not true, we can make the left side of (140) positive by choosing $h_\epsilon(t, u) > 0$ for those values of u in which the left side of (142) is greater than zero and $h_\epsilon(t, u) < 0$ elsewhere. These three observations complete the proof of (138).

The result in (138) is fundamental to many of our later problems. For the case of current interest we assume that the additive noise is white. Then

$$K_r(t, u) = \frac{N_0}{2}\,\delta(t - u) + K_a(t, u). \tag{143}$$

Substituting (143) into (138), we obtain

$$\boxed{\frac{N_0}{2}\, h_o(t, u) + \int_0^T h_o(t, v)\, K_a(u, v)\, dv = K_a(t, u), \qquad \begin{matrix} 0 \le t \le T \\ 0 < u < T. \end{matrix}} \tag{144}$$

Observe that $h_o(t, 0)$ and $h_o(t, T)$ are uniquely specified by the continuity requirement

$$h_o(t, 0) = \lim_{u \to 0+} h_o(t, u) \tag{145a}$$

$$h_o(t, T) = \lim_{u \to T-} h_o(t, u). \tag{145b}$$