SIGNALS, SYSTEMS, and INFERENCE

—

Class Notes for 6.011: Introduction to Communication, Control and Signal Processing Spring 2010

Alan V. Oppenheim and George C. Verghese

Massachusetts Institute of Technology

Contents

CHAPTER 2

Signals and Systems

This text assumes a basic background in the representation of linear, time-invariant systems and the associated continuous-time and discrete-time signals, through convolution, Fourier analysis, Laplace transforms and $\mathcal{Z}$-transforms. In this chapter we briefly summarize and review this assumed background, in part to establish notation that we will be using throughout the text, and also as a convenient reference for the topics in the later chapters. We follow closely the notation, style and presentation in *Signals and Systems*, Oppenheim and Willsky with Nawab, 2nd Edition, Prentice Hall, 1997.

2.1 SIGNALS, SYSTEMS, MODELS, PROPERTIES

Throughout this text we will be considering various classes of signals and systems, developing models for them and studying their properties.

Signals for us will generally be real or complex functions of some independent variables (almost always time and/or a variable denoting the outcome of a probabilistic experiment, for the situations we shall be studying). Signals can be:

- 1-dimensional or multi-dimensional
- continuous-time (CT) or discrete-time (DT)
- deterministic or stochastic (random, probabilistic)

Thus, a DT deterministic time-signal may be denoted by a function $x[n]$ of the integer time (or clock or counting) variable n.

Systems are collections of software or hardware elements, components, subsystems. A system can be viewed as mapping a set of input signals to a set of output or response signals. A more general view is that a system is an entity imposing constraints on a designated set of signals, where the signals are not necessarily labeled as inputs or outputs. Any specific set of signals that satisfies the constraints is termed a behavior of the system.

Models are (usually approximate) mathematical or software or hardware or linguistic or other representations of the constraints imposed on a designated set of

signals by a system. A model is itself a system, because it imposes constraints on the set of signals represented in the model, so we often use the words "system" and "model" interchangeably, although it can sometimes be important to preserve the distinction between something truly physical and our representations of it mathematically or in a computer simulation. We can thus talk of the behavior of a model.

A mapping model of a system comprises the following: a set of input signals $\{x_i(t)\}$, each of which can vary within some specified range of possibilities; similarly, a set of output signals $\{y_j(t)\}$, each of which can vary; and a description of the mapping that uniquely defines the output signals as a function of the input signals. As an example, consider the following single-input, single-output system:

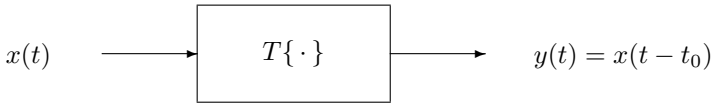

FIGURE 2.1 Name-Mapping Model

Given the input $x(t)$ and the mapping $T\{\cdot\}$, the output $y(t)$ is unique, and in this example equals the input delayed by t_0.

A behavioral model for a set of signals $\{w_i(t)\}$ comprises a listing of the constraints that the $w_i(t)$ must satisfy. The constraints on the voltages across and currents through the components in an electrical circuit, for example, are specified by Kirchhoff's laws, and the defining equations of the components. There can be infinitely many combinations of voltages and currents that will satisfy these constraints.

2.1.1 System/Model Properties

For a system or model specified as a mapping, we have the following definitions of various properties, all of which we assume are familiar. They are stated here for the DT case but easily modified for the CT case. (We also assume a single input signal and a single output signal in our mathematical representation of the definitions below, for notational convenience.)

- **Memoryless** or **Algebraic** or **Non-Dynamic**: The outputs at any instant do not depend on values of the inputs at any other instant: $y[n_0] = T\{x[n_0]\}$ for all n_0.
- **Linear**: The response to an arbitrary linear combination (or "superposition") of inputs signals is always the same linear combination of the individual responses to these signals: $T\{ax_A[n] + bx_B[n]\} = aT\{x_A[n]\} + bT\{x_B[n]\}$, for all x_A, x_B, a and b.

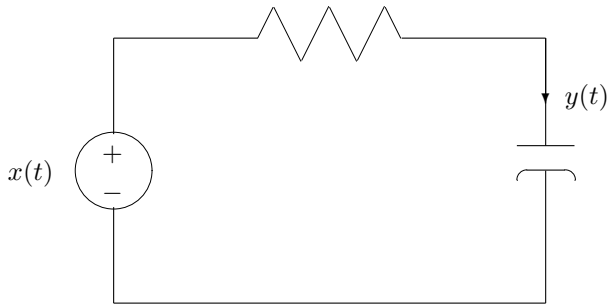

FIGURE 2.2 RLC Circuit

- **Time-Invariant**: The response to an arbitrarily translated set of inputs is always the response to the original set, but translated by the same amount: If $x[n] \rightarrow y[n]$ then $x[n-n_0] \rightarrow y[n-n_0]$ for all x and n_0.
- **Linear and Time-Invariant (LTI)**: The system, model or mapping is both linear and time-invariant.
- **Causal**: The output at any instant does not depend on future inputs: for all n_0, $y[n_0]$ does not depend on $x[n]$ for $n > n_0$. Said another way, if $\widehat{x}[n], \widehat{y}[n]$ denotes another input-output pair of the system, with $\widehat{x}[n] = x[n]$ for $n \leq n_0$, then it must be also true that $\widehat{y}[n] = y[n]$ for $n \leq n_0$. (Here n_0 is arbitrary but fixed.)
- **BIBO Stable**: The response to a bounded input is always bounded: $|x[n]| \leq M_x < \infty$ for all n implies that $|y[n]| \leq M_y < \infty$ for all n.

EXAMPLE 2.1 System Properties

Consider the system with input $x[n]$ and output $y[n]$ defined by the relationship

$$y[n] = x[4n+1] \tag{2.1}$$

We would like to determine whether or not the system has each of the following properties: memoryless, linear, time-invariant, causal, and BIBO stable.

memoryless: a simple counter example suffices. For example, $y[0] = x[1]$, i.e. the output at $n = 0$ depends on input values at times other than at $n = 0$. Therefore it is not memoryless.

linear: To check for linearity, we consider two different inputs, $x_A[n]$ and $x_B[n]$, and compare the output of their linear combination to the linear combination of

their outputs.

$$\begin{aligned} x_A[n] &\rightarrow x_A[4n+1] = y_A[n] \\ x_B[n] &\rightarrow x_B[4n+1] = y_B[n] \\ x_C[n] = (ax_A[n] + bx_B[n]) &\rightarrow (ax_A[4n+1] + bx_B[4n+1]) = y_C[n] \end{aligned}$$

If $y_C[n] = ay_A[n] + by_B[n]$, then the system is linear. This clearly happens in this case.

time-invariant: To check for time-invariance, we need to compare the output due to a time-shifted version of $x[n]$ to the time-shifted version of the output due to $x[n]$.

$$\begin{aligned} x[n] &\rightarrow x[4n+1] = y[n] \\ x_B[n] = x[n+n_0] &\rightarrow x[4n+n_0+1] = y_B[n] \end{aligned}$$

We now need to compare $y[n]$ time-shifted by n_0 (i.e. $y[n+n_0]$) to $y_B[n]$. If they're not equal, then the system is not time-invariant.

$$\begin{aligned} y[n+n_0] &= x[4n+4n_0+1] \\ \text{but} \quad y_B[n] &= x[4n+n_0+1] \end{aligned}$$

Consequently, the system is not time-invariant. To illustrate with a specific counter-example, suppose that $x[n]$ is an impulse, $\delta[n]$, at $n = 0$. In this case, the output, $y_\delta[n]$, would be $\delta[4n+1]$, which is zero for all values of n, and $y[n+n_0]$ would likewise always be zero. However, if we consider $x[n+n_0] = \delta[n+n_0]$, the output will be $\delta[4n+1+n_0]$, which for $n_0 = 3$ will be one at $n = -4$ and zero otherwise.

causal: Since the output at $n = 0$ is the input value at $n = 1$, the system is not causal.

BIBO stable: Since $|y[n]| = |x[4n+1]|$ and the maximum value for all n of $x[n]$ and $x[4n+1]$ is the same, the system is BIBO stable.

2.2 LINEAR, TIME-INVARIANT SYSTEMS

2.2.1 Impulse-Response Representation of LTI Systems

Linear, time-invariant (LTI) systems form the basis for engineering design in many situations. They have the advantage that there is a rich and well-established theory for analysis and design of this class of systems. Furthermore, in many systems that are nonlinear, small deviations from some nominal steady operation are approximately governed by LTI models, so the tools of LTI system analysis and design can be applied incrementally around a nominal operating condition.

A very general way of representing an LTI mapping from an input signal x to an output signal y is through convolution of the input with the system impulse

response. In CT the relationship is

$$y(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau \tag{2.2}$$

where $h(t)$ is the unit impulse response of the system. In DT, we have

$$y[n] = \sum_{k=-\infty}^{\infty} x[k]\,h[n-k] \tag{2.3}$$

where $h[n]$ is the unit sample (or unit "impulse") response of the system.

A common notation for the convolution integral in (2.2) or the convolution sum in (2.3) is as

$$y(t) = x(t) * h(t) \tag{2.4}$$

$$y[n] = x[n] * h[n] \tag{2.5}$$

While this notation can be convenient, it can also easily lead to misinterpretation if not well understood.

The characterization of LTI systems through the convolution is obtained by representing the input signal as a superposition of weighted impulses. In the DT case, suppose we are given an LTI mapping whose impulse response is $h[n]$, i.e., when its input is the unit sample or unit "impulse" function $\delta[n]$, its output is $h[n]$. Now a general input $x[n]$ can be assembled as a sum of scaled and shifted impulses, as follows:

$$x[n] = \sum_{k=-\infty}^{\infty} x[k]\,\delta[n-k] \tag{2.6}$$

The response $y[n]$ to this input, by linearity and time-invariance, is the sum of the similarly scaled and shifted impulse responses, and is therefore given by (2.3). What linearity and time-invariance have allowed us to do is write the response to a general input in terms of the response to a special input. A similar derivation holds for the CT case.

It may seem that the preceding derivation shows all LTI mappings from an input signal to an output signal can be represented via a convolution relationship. However, the use of infinite integrals or sums like those in (2.2), (2.3) and (2.6) actually involves some assumptions about the corresponding mapping. We make no attempt here to elaborate on these assumptions. Nevertheless, it is not hard to find "pathological" examples of LTI mappings — not significant for us in this course, or indeed in most engineering models — where the convolution relationship does not hold because these assumptions are violated.

It follows from (2.2) and (2.3) that a necessary and sufficient condition for an LTI system to be BIBO stable is that the impulse response be absolutely integrable (CT) or absolutely summable (DT), i.e.,

$$\text{BIBO stable (CT)} \iff \int_{-\infty}^{\infty} |h(t)|dt < \infty$$

$$\text{BIBO stable (DT)} \iff \sum_{n=-\infty}^{\infty} |h[n]| < \infty$$

It also follows from (2.2) and (2.3) that a necessary and sufficient condition for an LTI system to be causal is that the impulse response be zero for $t < 0$ (CT) or for $n < 0$ (DT)

2.2.2 Eigenfunction and Transform Representation of LTI Systems

Exponentials are eigenfunctions of LTI mappings, i.e., when the input is an exponential for all time, which we refer to as an "everlasting" exponential, the output is simply a scaled version of the input, so computing the response to an exponential reduces to just multiplying by the appropriate scale factor. Specifically, in the CT case, suppose

$$x(t) = e^{s_0 t} \tag{2.7}$$

for some possibly complex value s_0 (termed the complex frequency). Then from (2.2)

$$\begin{aligned} y(t) &= h(t) * x(t) \\ &= \int_{-\infty}^{\infty} h(\tau) x(t-\tau) d\tau \\ &= \int_{-\infty}^{\infty} h(\tau) e^{s_0(t-\tau)} d\tau \\ &= H(s_0) e^{s_0 t} \end{aligned} \tag{2.8}$$

where

$$H(s) = \int_{-\infty}^{\infty} h(\tau) e^{-s\tau} d\tau \tag{2.9}$$

provided the above integral has a finite value for $s = s_0$ (otherwise the response to the exponential is not well defined). Note that this integral is precisely the bilateral Laplace transform of the impulse response, or the transfer function of the system, and the (interior of the) set of values of s for which the above integral takes a finite value constitutes the region of convergence (ROC) of the transform.

From the preceding discussion, one can recognize what special property of the everlasting exponential causes it to be an eigenfunction of an LTI system: it is the fact that time-shifting an everlasting exponential produces the same result as scaling it by a constant factor. In contrast, the one-sided exponential $e^{s_0 t}u(t)$ — where $u(t)$ denotes the unit step — is in general not an eigenfunction of an LTI mapping: time-shifting a one-sided exponential does not produce the same result as scaling this exponential.

When $x(t) = e^{j\omega t}$, corresponding to having s_0 take the purely imaginary value $j\omega$ in (2.7), the input is bounded for all positive and negative time, and the corresponding output is

$$y(t) = H(j\omega) e^{j\omega t} \tag{2.10}$$

where

$$H(j\omega) = \int_{-\infty}^{\infty} h(t)e^{-j\omega t}\,dt \tag{2.11}$$

EXAMPLE 2.2 Eigenfunctions of LTI Systems

While as demonstrated above, the everlasting complex exponential, $e^{j\omega t}$, is an eigenfunction of any stable LTI system, it is important to recognize that $e^{j\omega t}u(t)$ is not. Consider, as a simple example, a time delay, i.e.

$$y(t) = x(t - t_0) \tag{2.12}$$

The output due to the input $e^{j\omega t}u(t)$ is

$$e^{-j\omega t_0}e^{+j\omega t}u(t - t_0)$$

This is not a simple scaling of the input, so $e^{j\omega t}u(t)$ is not in general an eigenfunction of LTI systems.

The function $H(j\omega)$ in (2.10) is the system frequency response, and is also the continuous-time Fourier transform (CTFT) of the impulse response. The integral that defines the CTFT has a finite value (and can be shown to be a continuous function of ω) if $h(t)$ is absolutely integrable, i.e. provided

$$\int_{-\infty}^{+\infty} |h(t)|\,dt < \infty$$

We have noted that this condition is equivalent to the system being bounded-input, bounded-output (BIBO) stable. The CTFT can also be defined for signals that are not absolutely integrable, e.g., for $h(t) = (\sin t)/t$ whose CTFT is a rectangle in the frequency domain, but we defer examination of conditions for existence of the CTFT.

We can similarly examine the eigenfunction property in the DT case. A DT everlasting "exponential" is a geometric sequence or signal of the form

$$x[n] = z_0^n \tag{2.13}$$

for some possibly complex z_0 (termed the complex frequency). With this DT exponential input, the output of a convolution mapping is (by a simple computation that is analogous to what we showed above for the CT case)

$$y[n] = h[n] * x[n] = H(z_0)z_0^n \tag{2.14}$$

where

$$H(z) = \sum_{k=-\infty}^{\infty} h[k]z^{-k} \tag{2.15}$$

provided the above sum has a finite value when $z = z_0$. Note that this sum is precisely the bilateral $\mathcal{Z}$-transform of the impulse response, and the (interior of the) set of values of z for which the sum takes a finite value constitutes the ROC of the $\mathcal{Z}$-transform. As in the CT case, the one-sided exponential $z_0^n u[n]$ is not in general an eigenfunction.

Again, an important case is when $x[n] = (e^{j\Omega})^n = e^{j\Omega n}$, corresponding to z_0 in (2.13) having unit magnitude and taking the value $e^{j\Omega}$, where Ω — the (real) "frequency" — denotes the angular position (in radians) around the unit circle in the z-plane. Such an $x[n]$ is bounded for all positive and negative time. Although we use a different symbol, Ω, for frequency in the DT case, to distinguish it from the frequency ω in the CT case, it is not unusual in the literature to find ω used in both CT and DT cases for notational convenience. The corresponding output is

$$y[n] = H(e^{j\Omega})e^{j\Omega n} \tag{2.16}$$

where

$$H(e^{j\Omega}) = \sum_{n=-\infty}^{\infty} h[n]e^{-j\Omega n} \tag{2.17}$$

The function $H(e^{j\Omega})$ in (2.17) is the frequency response of the DT system, and is also the discrete-time Fourier transform (DTFT) of the impulse response. The sum that defines the DTFT has a finite value (and can be shown to be a continuous function of Ω) if $h[n]$ is absolutely summable, i.e., provided

$$\sum_{n=-\infty}^{\infty} |\, h[n] \,| < \infty \tag{2.18}$$

We noted that this condition is equivalent to the system being BIBO stable. As with the CTFT, the DTFT can be defined for signals that are not absolutely summable; we will elaborate on this later.

Note from (2.17) that the frequency response for DT systems is always periodic, with period 2π. The "high-frequency" response is found in the vicinity of $\Omega = \pm\pi$, which is consistent with the fact that the input signal $e^{\pm j\pi n} = (-1)^n$ is the most rapidly varying DT signal that one can have.

When the input of an LTI system can be expressed as a linear combination of bounded eigenfunctions, for instance (in the CT case),

$$x(t) = \sum_{\ell} a_\ell e^{j\omega_\ell t} \tag{2.19}$$

then, by linearity, the output is the same linear combination of the responses to the individual exponentials. By the eigenfunction property of exponentials in LTI systems, the response to each exponential involves only scaling by the system's frequency response. Thus

$$y(t) = \sum_{\ell} a_\ell H(j\omega_\ell)e^{j\omega_\ell t} \tag{2.20}$$

Similar expressions can be written for the DT case.

2.2.3 Fourier Transforms

A broad class of input signals can be represented as linear combinations of bounded exponentials, through the Fourier transform. The synthesis/analysis formulas for the CTFT are

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(j\omega)\, e^{j\omega t} d\omega \quad \text{(synthesis)} \tag{2.21}$$

$$X(j\omega) = \int_{-\infty}^{\infty} x(t)\, e^{-j\omega t} dt \quad \text{(analysis)} \tag{2.22}$$

Note that (2.21) expresses $x(t)$ as a linear combination of exponentials — but this weighted combination involves a continuum of exponentials, rather than a finite or countable number. If this signal $x(t)$ is the input to an LTI system with frequency response $H(j\omega)$, then by linearity and the eigenfunction property of exponentials the output is the same weighted combination of the responses to these exponentials, so

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} H(j\omega)X(j\omega)\, e^{j\omega t} d\omega \tag{2.23}$$

By viewing this equation as a CTFT synthesis equation, it follows that the CTFT of $y(t)$ is

$$Y(j\omega) = H(j\omega)X(j\omega) \tag{2.24}$$

Correspondingly, the convolution relationship (2.2) in the time domain becomes multiplication in the transform domain. Thus, to find the response Y at a particular frequency point, we only need to know the input X at that single frequency, and the frequency response of the system at that frequency. This simple fact serves, in large measure, to explain why the frequency domain is virtually indispensable in the analysis of LTI systems.

The corresponding DTFT synthesis/analysis pair is defined by

$$x[n] = \frac{1}{2\pi}\int_{<2\pi>} X(e^{j\Omega})\, e^{j\Omega n} d\Omega \quad \text{(synthesis)} \tag{2.25}$$

$$X(e^{j\Omega}) = \sum_{n=-\infty}^{\infty} x[n]\, e^{-j\Omega n} \quad \text{(analysis)} \tag{2.26}$$

where the notation $< 2\pi >$ on the integral in the synthesis formula denotes integration over any contiguous interval of length 2π, since the DTFT is always periodic in Ω with period 2π, a simple consequence of the fact that $e^{j\Omega}$ is periodic with period 2π. Note that (2.25) expresses $x[n]$ as a weighted combination of a continuum of exponentials.

As in the CT case, it is straightforward to show that if $x[n]$ is the input to an LTI mapping, then the output $y[n]$ has DTFT

$$Y(e^{j\Omega}) = H(e^{j\Omega})X(e^{j\Omega}) \tag{2.27}$$

2.3 DETERMINISTIC SIGNALS AND THEIR FOURIER TRANSFORMS

In this section we review the DTFT of deterministic DT signals in more detail, and highlight the classes of signals that can be guaranteed to have well-defined DTFTs. We shall also devote some attention to the energy density spectrum of signals that have DTFTs. The section will bring out aspects of the DTFT that may not have been emphasized in your earlier signals and systems course. A similar development can be carried out for CTFTs.

2.3.1 Signal Classes and their Fourier Transforms

The DTFT synthesis and analysis pair in (2.25) and (2.26) hold for at least the three large classes of DT signals described below.

Finite-Action Signals. Finite-action signals, which are also called absolutely summable signals or ℓ_1 ("ell-one") signals, are defined by the condition

$$\sum_{k=-\infty}^{\infty} \Big|x[k]\Big| < \infty \tag{2.28}$$

The sum on the left is called the 'action' of the signal. For these ℓ_1 signals, the infinite sum that defines the DTFT is well behaved and the DTFT can be shown to be a continuous function for all Ω (so, in particular, the values at $\Omega = +\pi$ and $\Omega = -\pi$ are well-defined and equal to each other — which is often not the case when signals are not ℓ_1).

Finite-Energy Signals. Finite-energy signals, which are also called square summable or ℓ_2 ("ell-two") signals, are defined by the condition

$$\sum_{k=-\infty}^{\infty} \Big|x[k]\Big|^2 < \infty \tag{2.29}$$

The sum on the left is called the 'energy' of the signal.

In discrete-time, an absolutely summable (i.e., ℓ_1) signal is always square summable (i.e., ℓ_2). (In continuous-time, the story is more complicated: an absolutely integrable signal need not be square integrable, e.g., consider $x(t) = 1/\sqrt{t}$ for $0 < t \leq 1$ and $x(t) = 0$ elsewhere; the source of the problem here is that the signal is not bounded.) However, the reverse is not true. For example, consider the signal $(\sin \Omega_c n)/\pi n$ for $0 < \Omega_c < \pi$, with the value at $n = 0$ taken to be Ω_c/π, or consider the signal $(1/n)u[n-1]$, both of which are ℓ_2 but not ℓ_1. If $x[n]$ is such a signal, its DTFT $X(e^{j\Omega})$ can be thought of as the limit for $N \to \infty$ of the quantity

$$X_N(e^{j\Omega}) = \sum_{k=-N}^{N} x[k]e^{-j\Omega k} \tag{2.30}$$

and the resulting limit will typically have discontinuities at some values of Ω. For instance, the transform of $(\sin \Omega_c n)/\pi n$ has discontinuities at $\Omega = \pm\Omega_c$.

Signals of Slow Growth. Signals of 'slow' growth are signals whose magnitude grows no faster than polynomially with the time index, e.g., $x[n] = n$ for all n. In this case $X_N(e^{j\Omega})$ in (2.30) does not converge in the usual sense, but the DTFT still exists as a generalized (or singularity) function; e.g., if $x[n] = 1$ for all n, then $X(e^{j\Omega}) = 2\pi\delta(\Omega)$ for $|\Omega| \leq \pi$.

Within the class of signals of slow growth, those of most interest to us are bounded (or ℓ_∞) signals:

$$\Big|x[k]\Big| \leq M < \infty \tag{2.31}$$

i.e., signals whose amplitude has a fixed and finite bound for all time. Bounded everlasting exponentials of the form $e^{j\Omega_0 n}$, for instance, play a key role in Fourier transform theory. Such signals need not have finite energy, but will have finite average power over any time interval, where average power is defined as total energy over total time.

Similar classes of signals are defined in continuous-time. Specifically, finite-action (or L_1) signals comprise those that are absolutely integrable, i.e.,

$$\int_{-\infty}^{\infty} \Big|x(t)\Big| dt < \infty \tag{2.32}$$

Finite-energy (or L_2) signals comprise those that are square summable, i.e.,

$$\int_{-\infty}^{\infty} \Big|x(t)\Big|^2 < \infty \tag{2.33}$$

And signals of slow growth are ones for which the magnitude grows no faster than polynomially with time. Bounded (or L_∞) continuous-time signals are those for which the magnitude never exceeds a finite bound M (so these are slow-growth signals as well). These may again not have finite energy, but will have finite average power over any time interval.

In both continuous-time and discrete-time there are many important Fourier transform pairs and Fourier transform properties developed and tabulated in basic texts on signals and systems (see, for example, Chapters 4 and 5 of Oppenheim and Willsky). For convenience, we include here a brief table of DTFT pairs. Other pairs are easily derived from these by applying various DTFT properties. (Note that the δ's in the left column denote unit samples, while those in the right column are unit impulses!)

$$\textbf{DT Signal} \longleftrightarrow \textbf{DTFT for } -\pi < \Omega \leq \pi$$

$$\begin{aligned}
\delta[n] &\longleftrightarrow 1 \\
\delta[n-n_0] &\longleftrightarrow e^{-j\Omega n_0} \\
1 \text{ (for all } n) &\longleftrightarrow 2\pi\delta(\Omega) \\
e^{j\Omega_0 n} \quad (-\pi < \Omega_0 \leq \pi) &\longleftrightarrow 2\pi\delta(\Omega-\Omega_0) \\
a^n u[n]\ ,\ |a|<1 &\longleftrightarrow \frac{1}{1-ae^{-j\Omega}} \\
u[n] &\longleftrightarrow \frac{1}{1-e^{-j\Omega}} + \pi\delta(\Omega) \\
\frac{\sin \Omega_c n}{\pi n} &\longleftrightarrow \begin{cases} 1, & -\Omega_c < \Omega < \Omega_c \\ 0, & \text{otherwise} \end{cases} \\
\left.\begin{array}{ll} 1, & -M \leq n \leq M \\ 0, & \text{otherwise} \end{array}\right\} &\longleftrightarrow \frac{\sin[\Omega(2M+1)/2]}{\sin(\Omega/2)}
\end{aligned}$$

In general it is important and useful to be fluent in deriving and utilizing the main transform pairs and properties. In the following subsection we discuss a particular property, Parseval's identity, which is of particular significance in our later discussion.

There are, of course, other classes of signals that are of interest to us in applications, for instance growing one-sided exponentials. To deal with such signals, we utilize $\mathcal{Z}$-transforms in discrete-time and Laplace transforms in continuous-time.

2.3.2 Parseval's Identity, Energy Spectral Density, Deterministic Autocorrelation

An important property of the Fourier transform is Parseval's identity for ℓ_2 signals. For discrete time, this identity takes the general form

$$\sum_{n=-\infty}^{\infty} x[n]y^*[n] = \frac{1}{2\pi}\int_{<2\pi>} X(e^{j\Omega})Y^*(e^{j\Omega})\,d\Omega \tag{2.34}$$

and for continuous time,

$$\int_{-\infty}^{\infty} x(t)y^*(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(j\omega)Y^*(j\omega)\,d\omega \tag{2.35}$$

where the * denotes the complex conjugate. Specializing to the case where $y[n] = x[n]$ or $y(t) = x(t)$, we obtain

$$\sum_{n=-\infty}^{\infty} |x[n]|^2 = \frac{1}{2\pi}\int_{<2\pi>} |X(e^{j\Omega})|^2\,d\Omega \tag{2.36}$$

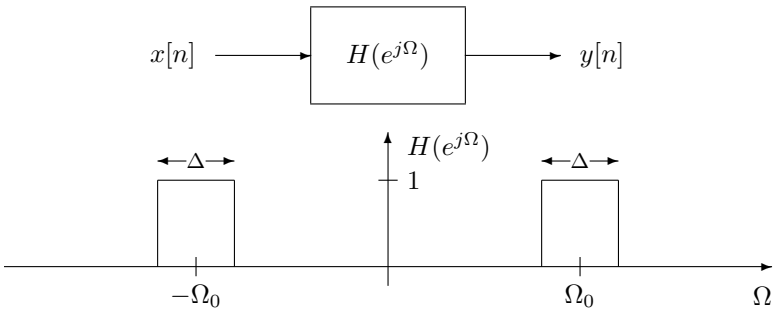

FIGURE 2.3 Ideal bandpass filter.

$$\int_{-\infty}^{\infty} |x(t)|^2 \;=\; \frac{1}{2\pi}\int_{-\infty}^{\infty} |X(j\omega)|^2\, d\omega \tag{2.37}$$

Parseval's identity allows us to evaluate the energy of a signal by integrating the squared magnitude of its transform. What the identity tells us, in effect, is that the energy of a signal equals the energy of its transform (scaled by $1/2\pi$).

The real, even, nonnegative function of Ω defined by

$$\overline{S}_{xx}(e^{j\Omega}) = |X(e^{j\Omega})|^2 \tag{2.38}$$

or

$$\overline{S}_{xx}(j\omega) = |X(j\omega)|^2 \tag{2.39}$$

is referred to as the energy spectral density (ESD), because it describes how the energy of the signal is distributed over frequency. To appreciate this claim more concretely, for discrete-time, consider applying $x[n]$ to the input of an ideal bandpass filter of frequency response $H(e^{j\Omega})$ that has narrow passbands of unit gain and width Δ centered at $\pm\Omega_0$ as indicated in Figure 2.3. The energy of the output signal must then be the energy of $x[n]$ that is contained in the passbands of the filter. To calculate the energy of the output signal, note that this output $y[n]$ has the transform

$$Y(e^{j\Omega}) = H(e^{j\Omega})X(e^{j\Omega}) \tag{2.40}$$

Consequently the output energy, by Parseval's identity, is given by

$$\begin{aligned}\sum_{n=-\infty}^{\infty} |y[n]|^2 &= \frac{1}{2\pi}\int_{<2\pi>} |Y(e^{j\Omega})|^2\, d\Omega \\ &= \frac{1}{2\pi}\int_{\text{passband}} \overline{S}_{xx}(e^{j\Omega})\, d\Omega\end{aligned} \tag{2.41}$$

Thus the energy of $x[n]$ in any frequency band is given by integrating $\overline{S}_{xx}(e^{j\Omega})$ over that band (and scaling by $1/2\pi$). In other words, the energy density of $x[n]$ as a

function of Ω is $\overline{S}_{xx}(\Omega)/(2\pi)$ per radian. An exactly analogous discussion can be carried out for continuous-time signals.

Since the ESD $\overline{S}_{xx}(e^{j\Omega})$ is a real function of Ω, an alternate notation for it could perhaps be $\mathcal{E}_{xx}(\Omega)$, for instance. However, we use the notation $\overline{S}_{xx}(e^{j\Omega})$ in order to make explicit that it is the squared magnitude of $X(e^{j\Omega})$ and also the fact that the ESD for a DT signal is periodic with period 2π.

Given the role of the magnitude squared of the Fourier transform in Parseval's identity, it is interesting to consider what signal it is the Fourier transform of. The answer for DT follows on recognizing that with $x[n]$ real-valued

$$|X(e^{j\Omega})|^2 = X(e^{j\Omega})X(e^{-j\Omega}) \tag{2.42}$$

and that $X(e^{-j\Omega})$ is the transform of the time-reversed signal, $x[-k]$. Thus, since multiplication of transforms in the frequency domain corresponds to convolution of signals in the time domain, we have

$$\overline{S}_{xx}(e^{j\Omega}) = |X(e^{j\Omega})|^2 \iff x[k] * x[-k] = \sum_{n=-\infty}^{\infty} x[n+k]x[n] = \overline{R}_{xx}[k] \tag{2.43}$$

The function $\overline{R}_{xx}[k] = x[k]*x[-k]$ is referred to as the deterministic autocorrelation function of the signal $x[n]$, and we have just established that the transform of the deterministic autocorrelation function is the energy spectral density $\overline{S}_{xx}(e^{j\Omega})$. A basic Fourier transform property tells us that $\overline{R}_{xx}[0]$ — which is the signal energy $\sum_{n=-\infty}^{\infty} x^2[n]$ — is the area under the Fourier transform of $R_{xx}[k]$, scaled by $1/(2\pi)$, namely the scaled area under $\overline{S}_{xx}(e^{j\Omega}) = |X(e^{j\Omega})|^2$; this is just Parseval's identity, of course.

The deterministic autocorrelation function measures how alike a signal and its time-shifted version are, in a total-squared-error sense. More specifically, in discrete-time the total squared error between the signal and its time-shifted version is given by

$$\begin{aligned}\sum_{n=-\infty}^{\infty} (x[n+k] - x[n])^2 &= \sum_{n=-\infty}^{\infty} |x[n+k]|^2 \\ &\quad + \sum_{n=-\infty}^{\infty} |x[n]|^2 - 2\sum_{n=-\infty}^{\infty} x[n+k]x[n] \\ &= 2(\overline{R}_{xx}[0] - \overline{R}_{xx}[k])\end{aligned} \tag{2.44}$$

Since the total squared error is always nonnegative, it follows that $\overline{R}_{xx}[k] \leq \overline{R}_{xx}[0]$, and that the larger the deterministic autocorrelation $\overline{R}_{xx}[k]$ is, the closer the signal $x[n]$ and its time-shifted version $x[n+k]$ are.

Corresponding results hold in continuous time, and in particular

$$\overline{S}_{xx}(j\omega) = |X(j\omega)|^2 \iff x(\tau) * x(-\tau) = \int_{-\infty}^{\infty} x(t+\tau)x(t)dt = \overline{R}_{xx}(\tau) \tag{2.45}$$

where $\overline{R}_{xx}(t)$ is the deterministic autocorrelation function of $x(t)$.

2.4 THE BILATERAL LAPLACE AND $\mathcal{Z}$-TRANSFORMS

The Laplace and $\mathcal{Z}$-transforms can be thought of as extensions of Fourier transforms and are useful for a variety of reasons. They permit a transform treatment of certain classes of signals for which the Fourier transform does not converge. They also augment our understanding of Fourier transforms by moving us into the complex plane, where the theory of complex functions can be applied. We begin in Section 2.4.1 with a detailed review of the bilateral $\mathcal{Z}$-transform. In Section 2.4.3 we give a briefer review of the bilateral Laplace transform, paralleling the discussion in Section 2.4.1.

2.4.1 The Bilateral $\mathcal{Z}$-Transform

The bilateral $\mathcal{Z}$-transform is defined as:

$$X(z) = \mathcal{Z}\{x[n]\} = \sum_{n=-\infty}^{\infty} x[n]z^{-n} \tag{2.46}$$

Here z is a complex variable, which we can also represent in polar form as

$$z = re^{j\Omega}, \quad r \geq 0, \ -\pi < \Omega \leq \pi \tag{2.47}$$

so

$$X(z) = \sum_{n=-\infty}^{\infty} x[n]r^{-n}e^{-j\Omega n} \tag{2.48}$$

The DTFT corresponds to fixing $r = 1$, in which case z takes values on the unit circle. However there are many useful signals for which the infinite sum does not converge (even in the sense of generalized functions) for z confined to the unit circle. The term z^{-n} in the definition of the $\mathcal{Z}$-transform introduces a factor r^{-n} into the infinite sum, which permits the sum to converge (provided r is appropriately restricted) for interesting classes of signals, many of which do not have discrete-time Fourier transforms.

More specifically, note from (2.48) that $X(z)$ can be viewed as the DTFT of $x[n]r^{-n}$. If $r > 1$, then r^{-n} decays geometrically for positive n and grows geometrically for negative n. For $0 < r < 1$, the opposite happens. Consequently, there are many sequences for which $x[n]$ is not absolutely summable but $x[n]r^{-n}$ is, for some range of values of r.

For example, consider $x_1[n] = a^n u[n]$. If $|a| > 1$, this sequence does not have a DTFT. However, for any a, $x[n]r^{-n}$ is absolutely summable provided $r > |a|$. In particular, for example,

$$X_1(z) = 1 + az^{-1} + a^2z^{-2} + \cdots \tag{2.49}$$

$$= \frac{1}{1 - az^{-1}}, \quad |z| = r > |a| \tag{2.50}$$

As a second example, consider $x_2[n] = -a^n u[-n-1]$. This signal does not have a DTFT if $|a| < 1$. However, provided $r < |a|$,

$$X_2(z) = -a^{-1}z - a^{-2}z^2 - \cdots \tag{2.51}$$

$$= \frac{-a^{-1}z}{1-a^{-1}z}, \quad |z| = r < |a| \tag{2.52}$$

$$= \frac{1}{1-az^{-1}}, \quad |z| = r < |a| \tag{2.53}$$

The $\mathcal{Z}$-transforms of the two distinct signals $x_1[n]$ and $x_2[n]$ above get condensed to the same rational expressions, but for different regions of convergence. Hence the ROC is a critical part of the specification of the transform.

When $x[n]$ is a sum of left-sided and/or right-sided DT exponentials, with each term of the form illustrated in the examples above, then $X(z)$ will be rational in z (or equivalently, in z^{-1}):

$$X(z) = \frac{Q(z)}{P(z)} \tag{2.54}$$

with $Q(z)$ and $P(z)$ being polynomials in z.

Rational $\mathcal{Z}$-transforms are typically depicted by a pole-zero plot in the z-plane, with the ROC appropriately indicated. This information uniquely specifies the signal, apart from a constant amplitude scaling. Note that there can be no poles in the ROC, since the transform is required to be finite in the ROC. $\mathcal{Z}$-transforms are often written as ratios of polynomials in z^{-1}. However, the pole-zero plot in the z-plane refers to the polynomials in z. Also note that if poles or zeros at $z = \infty$ are counted, then any ratio of polynomials always has exactly the same number of poles as zeros.

Region of Convergence. To understand the complex-function properties of the $\mathcal{Z}$-transform, we split the infinite sum that defines it into non-negative-time and negative-time portions: The non-negative-time or one-sided $\mathcal{Z}$-transform is defined by

$$\sum_{n=0}^{\infty} x[n]z^{-n} \tag{2.55}$$

and is a power series in z^{-1}. The convergence of the finite sum $\sum_{n=0}^{N} x[n]z^{-n}$ as $N \to \infty$ is governed by the radius of convergence $R_1 \geq 0$, of the power series, i.e. the series converges for each z such that $|z| > R_1$. The resulting function of z is an analytic function in this region, i.e., has a well-defined derivative with respect to the complex variable z at each point in this region, which is what gives the function its nice properties. The infinite sum diverges for $|z| < R_1$. The behavior of the sum on the circle $|z| = R_1$ requires closer examination, and depends on the particular series; the series may converge (but may not converge absolutely) at all points, some points, or no points on this circle. The region $|z| > R_1$ is referred to as the region of convergence (ROC) of the power series.

Next consider the negative-time part:

$$\sum_{n=-\infty}^{-1} x[n]z^{-n} = \sum_{m=1}^{\infty} x[-m]z^{m} \tag{2.56}$$

which is a power series in z, and has a radius of convergence R_2. The series converges (absolutely) for $|z| < R_2$, which constitutes its ROC; the series is an analytic function in this region. The sum diverges for $|z| > R_2$; the behavior for the circle $|z| = R_2$ takes closer examination, and depends on the particular series; the series may converge (but may not converge absolutely) at all points, some points, or no points on this circle. If $R_1 < R_2$ then the $\mathcal{Z}$-transform converges (absolutely) for $R_1 < |z| < R_2$; this annular region is its ROC, and is denoted by $\mathcal{R}_X$. The transform is analytic in this region. The sum that defines the transform diverges for $|z| < R_1$ and $|z| > R_2$. If $R_1 > R_2$, then the $\mathcal{Z}$-transform does not exist (e.g., for $x[n] = 0.5^n u[-n-1] + 2^n u[n]$). If $R_1 = R_2$, then the transform may exist in a technical sense, but is not useful as a $\mathcal{Z}$-transform because it has no ROC. However, if $R_1 = R_2 = 1$, then we may still be able to compute and use a DTFT (e.g., for $x[n] = 3$, all n; or for $x[n] = (\sin \omega_0 n)/(\pi n)$).

Relating the ROC to Signal Properties. For an absolutely summable signal (such as the impulse response of a BIBO-stable system), i.e., an ℓ_1-signal, the unit circle must lie in the ROC or must be a boundary of the ROC. Conversely, we can conclude that a signal is ℓ_1 if the ROC contains the unit circle because the transform converges absolutely in its ROC. If the unit circle constitutes a boundary of the ROC, then further analysis is generally needed to determine if the signal is ℓ_1. Rational transforms always have a pole on the boundary of the ROC, as elaborated on below, so if the unit circle is on the boundary of the ROC of a rational transform, then there is a pole on the unit circle, and the signal cannot be ℓ_1.

For a right-sided signal it is the case that $R_2 = \infty$, i.e., the ROC extends everywhere in the complex plane outside the circle of radius R_1, up to (and perhaps including) ∞. The ROC includes ∞ if the signal is 0 for negative time.

We can state a converse result if, for example, we know the signal comprises only sums of one-sided exponentials, of the form obtained when inverse transforming a rational transform. In this case, if $R_2 = \infty$, then the signal must be right-sided; if the ROC includes ∞, then the signal must be causal, i.e., zero for $n < 0$.

For a left-sided signal, one has $R_1 = 0$, i.e., the ROC extends inwards from the circle of radius R_2, up to (and perhaps including) 0. The ROC includes 0 if the signal is 0 for positive time.

In the case of signals that are sums of one-sided exponentials, we have a converse: if $R_1 = 0$, then the signal must be left-sided; if the ROC includes 0, then the signal must be anti-causal, i.e., zero for $n > 0$.

It is also important to note that the ROC cannot contain poles of the $\mathcal{Z}$-transform, because poles are values of z where the transform has infinite magnitude, while the ROC comprises values of z where the transform converges. For signals with

rational transforms, one can use the fact that such signals are sums of one-sided exponentials to show that the possible boundaries of the ROC are in fact precisely determined by the locations of the poles. Specifically:

(a) the outer bounding circle of the ROC in the rational case contains a pole and/or has radius ∞. If the outer bounding circle is at infinity, then (as we have already noted) the signal is right-sided, and is in fact causal if there is no pole at ∞;

(b) the inner bounding circle of the ROC in the rational case contains a pole and/or has radius 0. If the inner bounding circle reduces to the point 0, then (as we have already noted) the signal is left-sided, and is in fact anti-causal if there is no pole at 0.

2.4.2 The Inverse $\mathcal{Z}$-Transform

One way to invert a rational $\mathcal{Z}$-transform is through the use of a partial fraction expansion, then either directly "recognizeing" the inverse transform of each term in the partial fraction representation, or expanding the term in a power series that converges for z in the specified ROC. For example, a term of the form

$$\frac{1}{1-az^{-1}} \tag{2.57}$$

can be expanded in a power series in az^{-1} if $|a| < |z|$ for z in the ROC, and expanded in a power series in $a^{-1}z$ if $|a| > |z|$ for z in the ROC. Carrying out this procedure for each term in a partial fraction expansion, we find that the signal $x[n]$ is a sum of left-sided and/or right-sided exponentials. For non-rational transforms, where there may not be a partial fraction expansion to simplify the process, it is still reasonable to attempt the inverse transformation by expansion into a power series consistent with the given ROC.

Although we will generally use partial fraction or power series methods to invert $\mathcal{Z}$-transforms, there is an explicit formula that is similar to that of the inverse DTFT, specifically,

$$x[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(z)z^n\,d\omega\Big|_{z=\overline{r}e^{j\omega}} \tag{2.58}$$

where the constant $\overline{r}$ is chosen to place z in the ROC, $\mathcal{R}_X$. This is not the most general inversion formula, but is sufficient for us, and shows that $x[n]$ is expressed as a weighted combination of discrete-time exponentials.

As is the case for Fourier transforms, there are many useful $\mathcal{Z}$-transform pairs and properties developed and tabulated in basic texts on signals and systems. Appropriate use of transform pairs and properties is often the basis for obtaining the $\mathcal{Z}$-transform or the inverse $\mathcal{Z}$-transform of many other signals.

2.4.3 The Bilateral Laplace Transform

As with the $\mathcal{Z}$-transform, the Laplace transform is introduced in part to handle important classes of signals that don't have CTFT's, but also enhances our understanding of the CTFT. The definition of the Laplace transform is

$$X(s) = \int_{-\infty}^{\infty} x(t)\, e^{-st}\, dt \tag{2.59}$$

where s is a complex variable, $s = \sigma + j\omega$. The Laplace transform can thus be thought of as the CTFT of $x(t)\, e^{-\sigma t}$. With σ appropriately chosen, the integral (2.59) can exist even for signals that have no CTFT.

The development of the Laplace transform parallels closely that of the $\mathcal{Z}$-transform in the preceding section, but with e^{σ} playing the role that r did in Section 2.4.1. The (interior of the) set of values of s for which the defining integral converges, as the limits on the integral approach $\pm\infty$, comprises the region of convergence (ROC) for the transform $X(s)$. The ROC is now determined by the minimum and maximum allowable values of σ, say σ_1 and σ_2 respectively. We refer to σ_1, σ_2 as the abscissa of convergence. The corresponding ROC is a vertical strip between σ_1 and σ_2 in the complex plane, $\sigma_1 < Re(s) < \sigma_2$. Equation (2.59) converges absolutely within the ROC; convergence at the left and right bounding vertical lines of the strip has to be separately examined. Furthermore, the transform is analytic (i.e., differentiable as a complex function) throughout the ROC. The strip may extend to $\sigma_1 = -\infty$ on the left, and to $\sigma_2 = +\infty$ on the right. If the strip collapses to a line (so that the ROC vanishes), then the Laplace transform is not useful (except if the line happens to be the $j\omega$ axis, in which case a CTFT analysis may perhaps be recovered).

For example, consider $x_1(t) = e^{at}u(t)$; the integral in (2.59) evaluates to $X_1(s) = 1/(s-a)$ provided $\mathrm{Re}\{s\} > a$. On the other hand, for $x_2(t) = -e^{at}u(-t)$, the integral in (2.59) evaluates to $X_2(s) = 1/(s-a)$ provided $\mathrm{Re}\{s\} < a$. As with the $\mathcal{Z}$-transform, note that the expressions for the transforms above are identical; they are distinguished by their distinct regions of convergence.

The ROC may be related to properties of the signal. For example, for absolutely integrable signals, also referred to as L_1 signals, the integrand in the definition of the Laplace transform is absolutely integrable on the $j\omega$ axis, so the $j\omega$ axis is in the ROC or on its boundary. In the other direction, if the $j\omega$ axis is strictly in the ROC, then the signal is L_1, because the integral converges absolutely in the ROC. Recall that a system has an L_1 impulse response if and only if the system is BIBO stable, so the result here is relevant to discussions of stability: if the $j\omega$ axis is strictly in the ROC of the system function, then the system is BIBO stable.

For right-sided signals, the ROC is some right-half-plane (i.e. all s such that $\mathrm{Re}\{s\} > \sigma_1$). Thus the system function of a causal system will have an ROC that is some right-half-plane. For left-sided signals, the ROC is some left-half-plane. For signals with rational transforms, the ROC contains no poles, and the boundaries of the ROC will have poles. Since the location of the ROC of a transfer function relative to the imaginary axis relates to BIBO stability, and since the poles

identify the boundaries of the ROC, the poles relate to stability. In particular, a system with a right-sided impulse response (e.g., a causal system) will be stable if and only if all its poles are in the left-half-plane, because this is precisely the condition that allows the ROC to contain the imaginary axis. Also note that a signal with a rational transform is causal if and only if it is right-sided.

A further property worth recalling is connected to the fact that exponentials are eigenfunctions of LTI systems. If we denote the Laplace transform of the impulse response $h(t)$ of an LTI system by $H(s)$, referred to as the system function or transfer function, then $e^{s_0 t}$ at the input of the system yields $H(s_0)\, e^{s_0 t}$ at the output, provided s_0 is in the ROC of the transfer function.

2.5 DISCRETE-TIME PROCESSING OF CONTINUOUS-TIME SIGNALS

Many modern systems for applications such as communication, entertainment, navigation and control are a combination of continuous-time and discrete-time subsystems, exploiting the inherent properties and advantages of each. In particular, the discrete-time processing of continuous-time signals is common in such applications, and we describe the essential ideas behind such processing here. As with the earlier sections, we assume that this discussion is primarily a review of familiar material, included here to establish notation and for convenient reference from later chapters in this text. In this section, and throughout this text, we will often be relating the CTFT of a continuous-time signal and the DTFT of a discrete-time signal obtained from samples of the continuous-time signal. We will use the subscripts c and d when necessary to help keep clear which signals are CT and which are DT.

2.5.1 Basic Structure for DT Processing of CT Signals

The basic structure is shown in Figure 2.4. As indicated, the processing involves continuous-to-discrete or C/D conversion to obtain a sequence of samples of the CT signal, then DT filtering to produce a sequence of samples of the desired CT output, then discrete-to-continuous or D/C conversion to reconstruct this desired CT signal from the sequence of samples. We will often restrict ourselves to conditions such that the overall system in Figure 2.4 is equivalent to an LTI continuous-time system. The necessary conditions for this typically include restricting the DT filtering to be LTI processing by a system with frequency response $H_d(e^{j\Omega})$, and also requiring that the input $x_c(t)$ be appropriately bandlimited. To satisfy the latter requirement, it is typical to precede the structure in the figure by a filter whose purpose is to ensure that $x_c(t)$ is essentially bandlimited. While this filter is often referred to as an anti-aliasing filter, we can often allow some aliasing in the C/D conversion if the discrete-time system removes the aliased components; the overall system can then still be a CT LTI system.

The ideal C/D converter in Figure 2.4 has as its output a sequence of samples of $x_c(t)$ with a specified sampling interval T_1, so that the DT signal is $x_d[n] = x_c(nT_1)$. Conceptually, therefore, the ideal C/D converter is straightforward. A practical analog-to-digital (or A/D) converter also quantizes the signal to one of a finite set

of output levels. However, in this text we do not consider the additional effects of quantization.

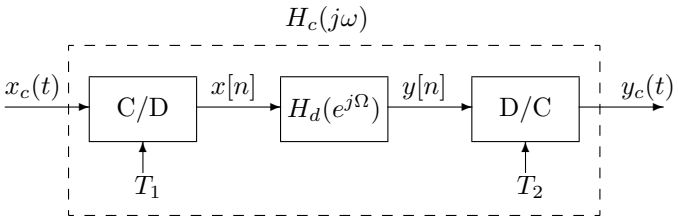

FIGURE 2.4 DT processing of CT signals.

In the frequency domain, the CTFT of $x_c(t)$ and the DTFT of $x_d[n]$ are related by

$$X_d\left(e^{j\Omega}\right)\Big|_{\Omega=\omega T_1} = \frac{1}{T_1}\sum_k X_c\left(j\omega - jk\frac{2\pi}{T_1}\right). \tag{2.60}$$

When $x_c(t)$ is sufficiently bandlimited so that

$$X_c(j\omega) = 0\,, \qquad |\omega| \geq \frac{\pi}{T_1} \tag{2.61}$$

then (2.60) can be rewritten as

$$X_d\left(e^{j\Omega}\right)\Big|_{\Omega=\omega T_1} = \frac{1}{T_1}X_c(j\omega) \qquad |\omega| < \pi/T_1 \tag{2.62a}$$

or equivalently

$$X_d\left(e^{j\Omega}\right) = \frac{1}{T_1}X_c\left(j\frac{\Omega}{T_1}\right) \qquad |\Omega| < \pi\,. \tag{2.62b}$$

Note that $X_d(e^{j\Omega})$ is extended periodically outside the interval $|\Omega| < \pi$. The fact that the above equalities hold under the condition (2.61) is the content of the sampling theorem.

The ideal D/C converter in Figure 2.4 is defined through the interpolation relation

$$y_c(t) = \sum_n y_d[n]\frac{\sin\left(\pi\left(t - nT_2\right)/T_2\right)}{\pi(t - nT_2)/T_2} \tag{2.63}$$

which shows that $y_c(nT_2) = y_d[n]$. Since each term in the above sum is bandlimited to $|\omega| < \pi/T_2$, the CT signal $y_c(t)$ is also bandlimited to this frequency range, so this D/C converter is more completely referred to as the ideal bandlimited interpolating converter. (The C/D converter in Figure 2.4, under the assumption (2.61), is similarly characterized by the fact that the CT signal $x_c(t)$ is the ideal bandlimited interpolation of the DT sequence $x_d[n]$.)

Because $y_c(t)$ is bandlimited and $y_c(nT_2) = y_d[n]$, analogous relations to (2.62) hold between the DTFT of $y_d[n]$ and the CTFT of $y_c(t)$:

$$Y_d\left(e^{j\Omega}\right)\Big|_{\Omega=\omega T_2} = \frac{1}{T_2}Y_c(j\omega) \qquad |\omega| < \pi/T_2 \tag{2.64a}$$

or equivalently

$$Y_d\left(e^{j\Omega}\right) = \frac{1}{T_2}Y_c\left(j\frac{\Omega}{T_2}\right) \qquad |\Omega| < \pi \tag{2.64b}$$

One conceptual representation of the ideal D/C converter is given in Figure 2.5. This figure interprets (2.63) to be the result of evenly spacing a sequence of impulses at intervals of T_2 — the reconstruction interval — with impulse strengths given by the $y_d[n]$, then filtering the result by an ideal low-pass filter $L(j\omega)$ of amplitude T_2 in the passband $|\omega| < \pi/T_2$. This operation produces the bandlimited continuous-time signal $y_c(t)$ that interpolates the specified sequence values $y_d[n]$ at the instants $t = nT_2$, i.e., $y_c(nT_2) = y_d[n]$.

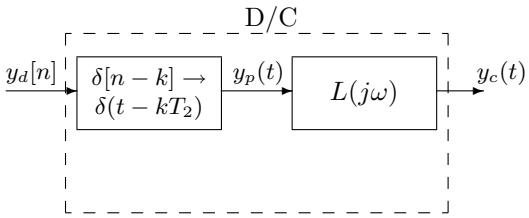

FIGURE 2.5 Conceptual representation of processes that yield ideal D/C conversion, interpolating a DT sequence into a bandlimited CT signal using reconstruction interval T_2.

2.5.2 DT Filtering, and Overall CT Response

Suppose from now on, unless stated otherwise, that $T_1 = T_2 = T$. If in Figure 2.4 the bandlimiting constraint of (2.61) is satisfied, and if we set $y_d[n] = x_d[n]$, then $y_c(t) = x_c(t)$. More generally, when the DT system in Figure 2.4 is an LTI DT filter with frequency response $H_d\left(e^{j\Omega}\right)$, so

$$Y_d(e^{j\Omega}) = H_d(e^{j\Omega})X_d(e^{j\Omega}) \tag{2.65}$$

and provided any aliased components of $x_c(t)$ are eliminated by $H_d(e^{j\Omega})$, then assembling (2.62), (2.64) and (2.65) yields:

$$Y_c(j\omega) = H_d\left(e^{j\Omega}\right)\Big|_{\Omega=\omega T} X_c(j\omega) \quad |\omega| < \pi/T \tag{2.66}$$

The action of the overall system is thus equivalent to that of a CT filter whose frequency response is

$$H_c(j\omega) = H_d\left(e^{j\Omega}\right)\Big|_{\Omega=\omega T} \qquad |\omega| < \pi/T. \tag{2.67}$$

In other words, under the bandlimiting and sampling rate constraints mentioned above, the overall system behaves as an LTI CT filter, and the response of this filter is related to that of the embedded DT filter through a simple frequency scaling. The sampling rate can be lower than the Nyquist rate, provided that the DT filter eliminates any aliased components.

If we wish to use the system in Figure 2.4 to implement a CT LTI filter with frequency response $H_c(j\omega)$, we choose $H_d\left(e^{j\Omega}\right)$ according to (2.67), provided that $x_c(t)$ is appropriately bandlimited.

If $H_c(j\omega) = 0$ for $|\omega| \geq \pi/T$, then (2.67) also corresponds to the following relation between the DT and CT impulse responses:

$$h_d[n] = T\, h_c(nT) \tag{2.68}$$

The DT filter is therefore termed an impulse-invariant version of the CT filter. When $x_c(t)$ and $H_d(e^{j\Omega})$ are not sufficiently bandlimited to avoid aliased components in $y_d[n]$, then the overall system in Figure 2.4 is no longer time invariant. It is, however, still linear since it is a cascade of linear subsystems.

The following two important examples illustrate the use of (2.67) as well as Figure 2.4, both for DT processing of CT signals and for interpretation of an important DT system, whether or not this system is explicitly used in the context of processing CT signals.

EXAMPLE 2.3 Digital Differentiator

In this example we wish to implement a CT differentiator using a DT system in the configuration of Figure 2.4 . We need to choose $H_d\left(e^{j\Omega}\right)$ so that $y_c(t) = \frac{dx_c(t)}{dt}$, assuming that $x_c(t)$ is bandlimited to π/T. The desired overall CT frequency response is therefore

$$H_c(j\omega) = \frac{Y_c(j\omega)}{X_c(j\omega)} = j\omega \tag{2.69}$$

Consequently, using (2.67) we choose $H_d(e^{j\Omega})$ such that

$$H_d\left(e^{j\Omega}\right)\Big|_{\Omega=\omega T} = j\omega \qquad |\omega| < \frac{\pi}{T} \tag{2.70a}$$

or equivalently

$$H_d\left(e^{j\Omega}\right) = j\Omega/T \qquad |\Omega| < \pi \tag{2.70b}$$

A discrete-time system with the frequency response in (2.70b) is commonly referred to as a digital differentiator. To understand the relation between the input $x_d[n]$

and output $y_d[n]$ of the digital differentiator, note that $y_c(t)$ — which is the bandlimited interpolation of $y_d[n]$ — is the derivative of $x_c(t)$, and $x_c(t)$ in turn is the bandlimited interpolation of $x_d[n]$. It follows that $y_d[n]$ can, in effect, be thought of as the result of sampling the derivative of the bandlimited interpolation of $x_d[n]$.

EXAMPLE 2.4 Half-Sample Delay

It often arises in designing discrete-time systems that a phase factor of the form $e^{-j\alpha\Omega}$, $|\Omega| < \pi$, is included or required. When α is an integer, this has a straightforward interpretation, since it corresponds simply to an integer shift by α of the time sequence.

When α is not an integer, the interpretation is not as straightforward, since a DT sequence can only be directly shifted by integer amounts. In this example we consider the case of $\alpha = 1/2$, referred to as a half-sample delay. To provide an interpretation, we consider the implications of choosing the DT system in Figure 2.4 to have frequency response

$$H_d(e^{j\Omega}) = e^{-j\Omega/2} \qquad |\Omega| < \pi \tag{2.71}$$

Whether or not $x_d[n]$ explicitly arose by sampling a CT signal, we can associate with $x_d[n]$ its bandlimited interpolation $x_c(t)$ for any specified sampling or reconstruction interval T. Similarly, we can associate with $y_d[n]$ its bandlimited interpolation $y_c(t)$ using the reconstruction interval T. With $H_d\left(e^{j\Omega}\right)$ given by (2.71), the equivalent CT frequency response relating $y_c(t)$ to $x_c(t)$ is

$$H_c(j\omega) = e^{-j\omega T/2} \tag{2.72}$$

representing a time delay of $T/2$, which is half the sample spacing; consequently, $y_c(t) = x_c(t - T/2)$. We therefore conclude that for a DT system with frequency response given by (2.71), the DT output $y_d[n]$ corresponds to samples of the half-sample delay of the bandlimited interpolation of the input sequence $x_d[n]$. Note that in this interpretation the choice for the value of T is immaterial. (Even if $x_d[n]$ had been the result of regular sampling of a CT signal, that specific sampling period is not required in the interpretation above.)

The preceding interpretation allows us to find the unit-sample (or impulse) response of the half-sample delay system through a simple argument. If $x_d[n] = \delta[n]$, then $x_c(t)$ must be the bandlimited interpolation of this (with some T that we could have specified to take any particular value), so

$$x_c(t) = \frac{\sin(\pi t/T)}{\pi t/T} \tag{2.73}$$

and therefore

$$y_c(t) = \frac{\sin\Big(\pi(t - (T/2))/T\Big)}{\pi(t - (T/2))/T} \tag{2.74}$$

which shows that the desired unit-sample response is

$$y_d[n] = h_d[n] = \frac{\sin\Big(\pi(n-(1/2))\Big)}{\pi(n-(1/2))} \tag{2.75}$$

This discussion of a half-sample delay also generalizes in a straightforward way to any integer or non-integer choice for the value of α.

2.5.3 Non-Ideal D/C converters

In Section 2.5.1 we defined the ideal D/C converter through the bandlimited interpolation formula (2.63); see also Figure 2.5, which corresponds to processing a train of impulses with strengths equal to the sequence values $y_d[n]$ through an ideal low-pass filter. A more general class of D/C converters, which includes the ideal converter as a particular case, creates a CT signal $y_c(t)$ from a DT signal $y_d[n]$ according to the following:

$$y_c(t) = \sum_{n=-\infty}^{\infty} y_d[n]\, p(t-nT) \tag{2.76}$$

where $p(t)$ is some selected basic pulse shape and T is the reconstruction interval or pulse repetition interval. This too can be seen as the result of processing an impulse train of sequence values through a filter, but a filter that has impulse response $p(t)$ rather than that of the ideal low-pass filter. The CT signal $y_c(t)$ is thus constructed by adding together shifted and scaled versions of the basic pulse shape; the number $y_d[n]$ scales $p(t-nT)$, which is the basic pulse delayed by nT. Note that the ideal bandlimited interpolating converter of (2.63) is obtained by choosing

$$p(t) = \frac{\sin(\pi t/T)}{(\pi t/T)} \tag{2.77}$$

We shall be talking in more detail in Chapter 12 about the interpretation of (2.76) as pulse amplitude modulation (PAM) for communicating DT information over a CT channel.

The relationship (2.76) can also be described quite simply in the frequency domain. Taking the CTFT of both sides, denoting the CTFT of $p(t)$ by $P(j\omega)$, and using the fact that delaying a signal by t_0 in the time domain corresponds to multiplication by $e^{-j\omega t_0}$ in the frequency domain, we get

$$\begin{aligned} Y_c(j\omega) &= \Big(\sum_{n=-\infty}^{\infty} y_d[n]\, e^{-jn\omega T}\Big) P(j\omega) \\ &= Y_d(e^{j\Omega})\Big|_{\Omega=\omega T} P(j\omega) \end{aligned} \tag{2.78}$$

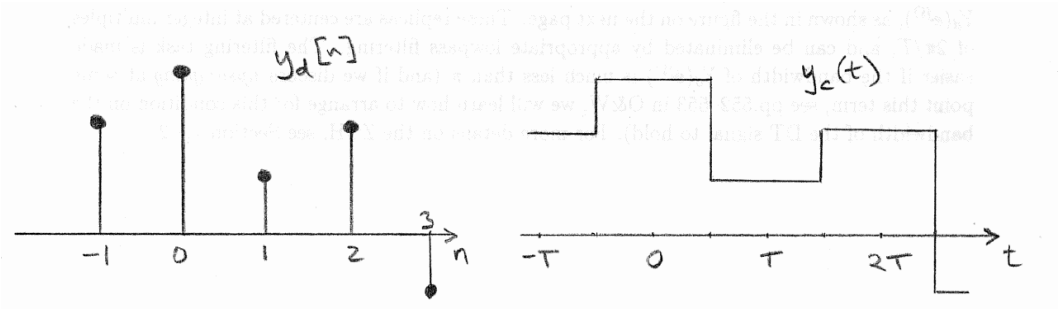

FIGURE 2.6 A centered zero-order hold (ZOH)

In the particular case where $p(t)$ is the sinc pulse in (2.77), with transform $P(j\omega)$ corresponding to an ideal low-pass filter of amplitude T for $|\omega| < \pi/T$ and 0 outside this band, we recover the relation (2.64).

In practice an ideal low-pass filter can only be approximated, with the accuracy of the approximation closely related to cost of implementation. A commonly used simple approximation is the (centered) zero-order hold (ZOH), specified by the choice

$$p(t) = \begin{cases} 1 & \text{for } |t| < (T/2) \\ 0 & \text{elsewhere} \end{cases} \tag{2.79}$$

This D/C converter holds the value of the DT signal at time n, namely the value $y_d[n]$, for an interval of length T centered at nT in the CT domain, as illustrated in Figure 2.6. Such ZOH converters are very commonly used. Another common choice is a centered first-order hold (FOH), for which $p(t)$ is triangular as shown in Figure 2.7. Use of the FOH represents linear interpolation between the sequence values.

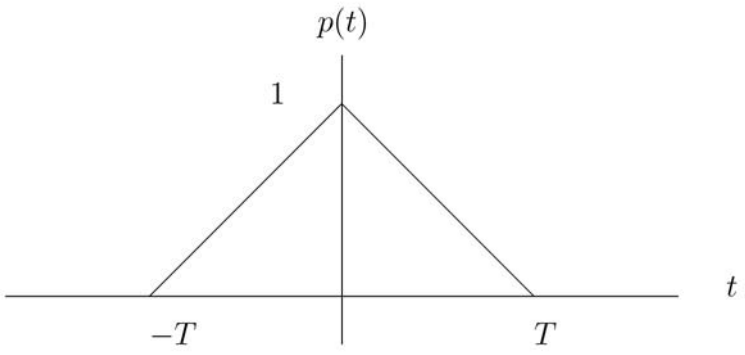

FIGURE 2.7 A centered first order hold (FOH)

CHAPTER 3

Transform Representation of Signals and LTI Systems

As you have seen in your prior studies of signals and systems, and as emphasized in the review in Chapter 2, transforms play a central role in characterizing and representing signals and LTI systems in both continuous and discrete time. In this chapter we discuss some specific aspects of transform representations that will play an important role in later chapters. These aspects include the interpretation of Fourier transform phase through the concept of group delay, and methods — referred to as spectral factorization — for obtaining a Fourier representation (magnitude and phase) when only the Fourier transform magnitude is known.

3.1 FOURIER TRANSFORM MAGNITUDE AND PHASE

The Fourier transform of a signal or the frequency response of an LTI system is in general a complex-valued function. A magnitude-phase representation of a Fourier transform $X(j\omega)$ takes the form

$$X(j\omega) = |X(j\omega)|e^{j\angle X(j\omega)} . \tag{3.1}$$

In eq. (3.1), $|X(j\omega)|$ denotes the (non-negative) magnitude and $\angle X(j\omega)$ denotes the (real-valued) phase. For example, if $X(j\omega)$ is the sinc function, $\sin(\omega)/\omega$, then $|X(j\omega)|$ is the absolute value of this function, while $\angle X(j\omega)$ is 0 in frequency ranges where the sinc is positive, and π in frequency ranges where the sinc is negative. An alternative representation is an amplitude-phase representation

$$A(\omega)e^{j\angle_A X(j\omega)} \tag{3.2}$$

in which $A(\omega) = \pm|X(j\omega)|$ is real but can be positive for some frequencies and negative for others. Correspondingly, $\angle_A X(j\omega) = \angle X(j\omega)$ when $A(\omega) = +|X(j\omega)|$, and $\angle_A X(j\omega) = \angle X(j\omega) \pm \pi$ when $A(\omega) = -|X(j\omega)|$. This representation is often preferred when its use can eliminate discontinuities of π radians in the phase as $A(\omega)$ changes sign. In the case of the sinc function above, for instance, we can pick $A(\omega) = \sin(\omega)/\omega$ and $\angle_A = 0$. It is generally convenient in the following discussion for us to assume that the transform under discussion has no zeros on the $j\omega$-axis, so that we can take $A(\omega) = |X(j\omega)|$ for all ω (or, if we wish, $A(\omega) = -|X(j\omega)|$ for all ω). A similar discussion applies also, of course, in discrete-time.

In either a magnitude-phase representation or an amplitude-phase representation, the phase is ambiguous, as any integer multiple of 2π can be added at any frequency

without changing $X(j\omega)$ in (3.1) or (3.2). A typical phase computation resolves this ambiguity by generating the phase modulo 2π, i.e., as the phase passes through $+\pi$ it "wraps around" to $-\pi$ (or from $-\pi$ wraps around to $+\pi$). In Section 3.2 we will find it convenient to resolve this ambiguity by choosing the phase to be a continuous function of frequency. This is referred to as the unwrapped phase, since the discontinuities at $\pm\pi$ are unwrapped to obtain a continuous phase curve. The unwrapped phase is obtained from $\angle X(j\omega)$ by adding steps of height equal to $\pm\pi$ or $\pm 2\pi$ wherever needed, in order to produce a continuous function of ω. The steps of height $\pm\pi$ are added at points where $X(j\omega)$ passes through 0, to absorb sign changes as needed; the steps of height $\pm 2\pi$ are added wherever else is needed, invoking the fact that such steps make no difference to $X(j\omega)$, as is evident from (3.1). We shall proceed as though $\angle X(j\omega)$ is indeed continuous (and differentiable) at the points of interest, understanding that continuity can indeed be obtained in all cases of interest to us by adding in the appropriate steps of height $\pm\pi$ or $\pm 2\pi$.

Typically, our intuition for the time-domain effects of frequency response magnitude or amplitude on a signal is rather well-developed. For example, if the Fourier transform magnitude is significantly attenuated at high frequencies, then we expect the signal to vary slowly and without sharp discontinuities. On the other hand, a signal in which the low frequencies are attenuated will tend to vary rapidly and without slowly varying trends.

In contrast, visualizing the effect on a signal of the phase of the frequency response of a system is more subtle, but equally important. We begin the discussion by first considering several specific examples which are helpful in then considering the more general case. Throughout this discussion we will consider the system to be an all-pass system with unity gain, i.e. the amplitude of the frequency response $A(j\omega) = 1$ (continuous time) or $A(e^{j\Omega}) = 1$ (discrete time) so that we can focus entirely on the effect of the phase. The unwrapped phase associated with the frequency response will be denoted as $\angle_A H(j\omega)$ (continuous time) and $\angle_A H(e^{j\Omega})$ (discrete time).

EXAMPLE 3.1 Linear Phase

Consider an all-pass system with frequency response

$$H(j\omega) = e^{-j\alpha\omega} \tag{3.3}$$

i.e. in an amplitude/phase representation $A(j\omega) = 1$ and $\angle_A H(j\omega) = -\alpha\omega$. The unwrapped phase for this example is linear with respect to ω, with slope of $-\alpha$. For input x(t) with Fourier transform $X(j\omega)$, the Fourier transform of the output is $Y(j\omega) = X(j\omega)e^{-j\alpha\omega}$ and correspondingly the output $y(t)$ is $x(t-\alpha)$. In words, linear phase with a slope of $-\alpha$ corresponds to a time delay of α (or a time advance if α is negative).

For a discrete time system with

$$H(e^{j\Omega}) = e^{-j\alpha\Omega} \qquad |\Omega| < \pi \tag{3.4}$$

the phase is again linear with slope $-\alpha$. When α is an integer, the time domain interpretation of the effect on an input sequence $x[n]$ is again straightforward and is

a simple delay (α positive) or advance (α negative) of $|\alpha|$. When α is not an integer, the effect is still commonly referred to as "a delay of α", but the interpretation is more subtle. If we think of $x[n]$ as being the result of sampling a band-limited, continuous-time signal $x(t)$ with sampling period T, the output $y[n]$ will be the result of sampling the signal $y(t) = x(t - \alpha T)$ with sampling period T. In fact we saw this result in Example 2.4 of chapter 2 for the specific case of a half-sample delay, i.e. $\alpha = \frac{1}{2}$.

EXAMPLE 3.2 Constant Phase Shift

As a second example, we again consider an all-pass system with $A(j\omega) = 1$ and unwrapped phase

$$\angle_A H(j\omega) = \begin{cases} -\phi_0 & \text{for } \omega > 0 \\ +\phi_0 & \text{for } \omega < 0 \end{cases}$$

as indicated in Figure 3.1

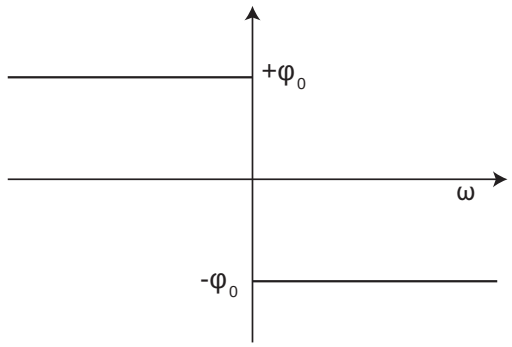

FIGURE 3.1 Phase plot of all-pass system with constant phase shift, ϕ_0.

Note that the phase is required to be an odd function of ω if we assume that the system impulse response is real valued. In this example, we consider $x(t)$ to be of the form

$$x(t) = s(t)\cos(\omega_0 t + \theta) \tag{3.5}$$

i.e. an amplitude-modulated signal at a carrier frequency of ω_0. Consequently, $X(j\omega)$ can be expressed as

$$X(j\omega) = \frac{1}{2}S(j\omega - j\omega_0)e^{j\theta} + \frac{1}{2}S(j\omega + j\omega_0)e^{-j\theta} \tag{3.6}$$

where $S(j\omega)$ denotes the Fourier transform of $s(t)$.

For this example, we also assume that $S(j\omega)$ is bandlimited to $|\omega| < \Delta$, with Δ sufficiently small so that the term $S(j\omega - j\omega_0)e^{j\theta}$ is zero for $\omega < 0$ and the term $S(j\omega + j\omega_0)e^{-j\theta}$ is zero for $\omega > 0$, i.e. that $(\omega_0 - \Delta) > 0$. The associated spectrum of $x(t)$ is depicted in Figure 3.2.

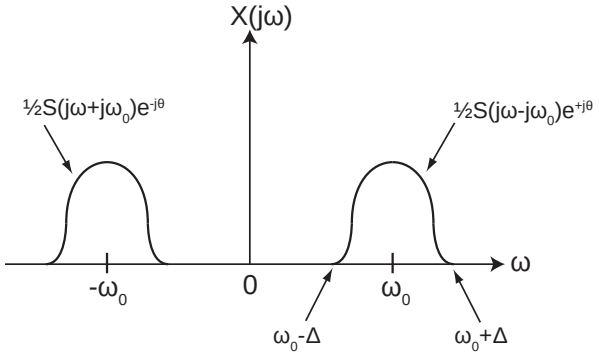

FIGURE 3.2 Spectrum of $x(t)$ with $s(t)$ narrowband

With these assumptions on $x(t)$, it is relatively straightforward to determine the output $y(t)$. Specifically, the system frequency response $H(j\omega)$ is

$$H(j\omega) = \begin{cases} e^{-j\phi_0} & \omega > 0 \\ e^{+j\phi_0} & \omega < 0 \end{cases} \tag{3.7}$$

Since the term $S(j\omega - j\omega_0)e^{j\theta}$ in eq. (3.6) is non-zero only for $\omega > 0$, it is simply multiplied by $e^{-j\phi}$, and similarly the term $S(j\omega + j\omega_0)e^{-j\theta}$ is multiplied only by $e^{+j\phi}$. Consequently, the output frequency response, $Y(j\omega)$, is given by

$$\begin{aligned} Y(j\omega) &= X(j\omega)H(j\omega) \\ &= \frac{1}{2}S(j\omega - j\omega_0)e^{+j\theta}e^{-j\phi_0} + \frac{1}{2}S(j\omega + j\omega_0)e^{-j\theta}e^{+j\phi_0} \end{aligned} \tag{3.8}$$

which we recognize as a simple phase shift by ϕ_0 of the carrier in eq. (3.5), i.e. replacing θ in eq. (3.6) by $\theta - \phi_0$. Consequently,

$$y(t) = s(t)\cos(\omega_0 t + \theta - \phi_0) \tag{3.9}$$

This change in phase of the carrier can also be expressed in terms of a time delay for the carrier by rewriting eq. (3.9) as

$$y(t) = s(t)\cos\left[\omega_0\left(t - \frac{\phi_0}{\omega_0}\right) + \theta\right] \tag{3.10}$$

3.2 GROUP DELAY AND THE EFFECT OF NONLINEAR PHASE

In Example 3.1, we saw that a phase characteristic that is linear with frequency corresponds in the time domain to a time shift. In this section we consider the

effect of a nonlinear phase characteristic. We again assume the system is an all-pass system with frequency response

$$H(j\omega) = A(j\omega)e^{j\angle_A[H(j\omega)]} \tag{3.11}$$

with $A(j\omega) = 1$. A general nonlinear unwrapped phase characteristic is depicted in Figure 3.3

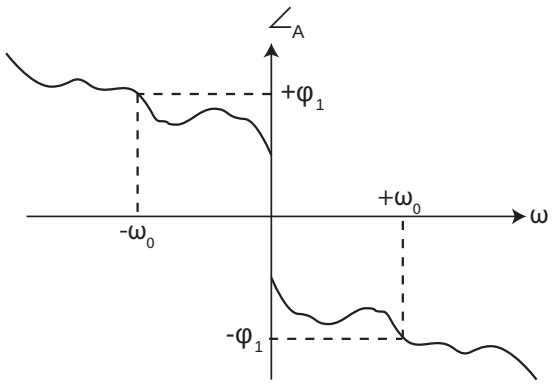

FIGURE 3.3 Nonlinear Unwrapped Phase Characteristic

As we did in Example 3.2, we again assume that $x(t)$ is narrowband of the form of equation (3.5) and as depicted in Figure 3.2. We next assume that Δ in Figure 3.2 is sufficiently small so that in the vicinity of $\pm\omega_0$, $\angle_A H(j\omega)$ can be approximated sufficiently well by the zeroth and first order terms of a Taylor's series expansion, i.e.

$$\angle_A H(j\omega) \approx \angle_A H(j\omega_0) + (\omega - \omega_0)\left[\frac{\mathrm{d}}{\mathrm{d}\omega}\angle_A H(j\omega)\right]_{\omega=\omega_0} \tag{3.12}$$

Defining $\tau_g(\omega)$ as

$$\tau_g(\omega) = -\frac{\mathrm{d}}{\mathrm{d}\omega}\angle_A H(j\omega) \tag{3.13}$$

our approximation to $\angle_A H(j\omega)$ in a small region around $\omega = \omega_0$ is expressed as

$$\angle_A H(j\omega) \approx \angle_A H(j\omega_0) - (\omega - \omega_0)\tau_g(\omega_0) \tag{3.14}$$

Similarly in a small region around $\omega = -\omega_0$, we make the approximation

$$\angle_A H(j\omega) \approx \angle_A H(j\omega_0) - (\omega + \omega_0)\tau_g(-\omega_0) \tag{3.15}$$

As we will see shortly, the quantity $\tau_g(\omega)$ plays a key role in our interpretation of the effect on a signal of a nonlinear phase characteristic.

With the Taylor's series approximation of eqs. (3.14) and (3.15) and for input signals with frequency content for which the approximation is valid, we can replace Figure 3.3 with Figure 3.4.

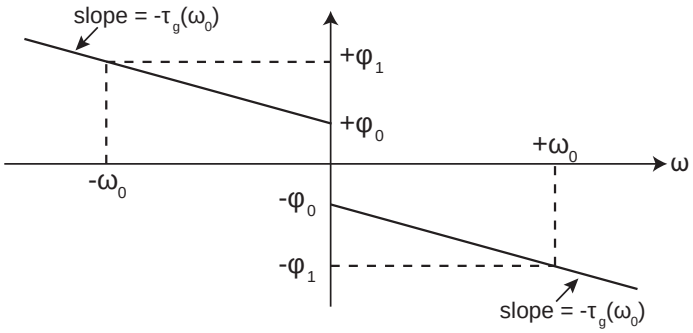

FIGURE 3.4 Taylor's series approximation of nonlinear phase in the vicinity of $\pm\omega_0$

where

$$-\phi_1 = \angle_A H(j\omega_0)$$

and

$$-\phi_0 = \angle_A H(j\omega_0) + \omega_0 \tau_g(\omega_0)$$

Since for LTI systems in cascade, the frequency responses multiply and correspondingly the phases add, we can represent the all-pass frequency response $H(j\omega)$ as the cascade of two all-pass systems, $H_I(j\omega)$ and $H_{II}(j\omega)$, with unwrapped phase as depicted in Figure 3.5.

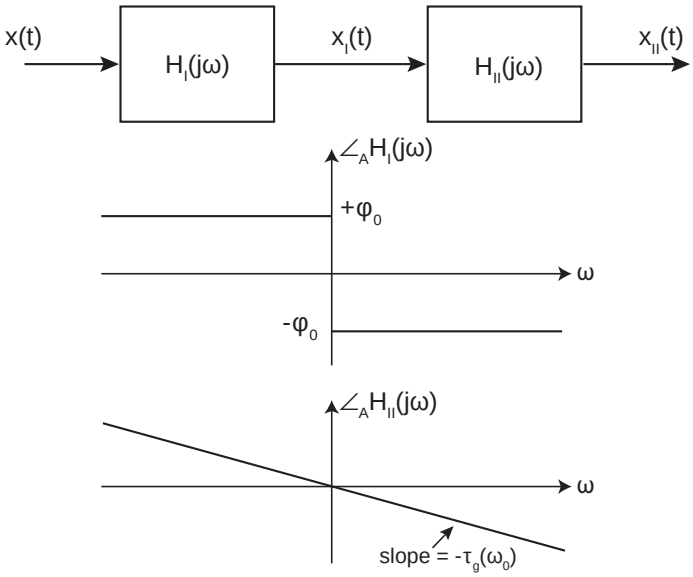

FIGURE 3.5 An all-pass system frequency response, $H(j\omega)$, represented as the cascade of two all-pass systems, $H_I(j\omega)$ and $H_{II}(j\omega)$.

We recognize $H_I(j\omega)$ as corresponding to Example 3.2. Consequently, with $x(t)$ narrowband, we have

$$\begin{aligned} x(t) &= s(t)\cos(\omega_0 t + \theta) \\ x_I(t) &= s(t)\cos\left[\omega_0\left(t - \frac{\phi_0}{\omega_0}\right) + \theta\right] \end{aligned} \tag{3.16}$$

Next we recognize $H_{II}(j\omega)$ as corresponding to Example 3.1 with $\alpha = \tau_g(\omega_0)$. Consequently,

$$x_{II}(t) = x_I\left(t - \tau_g(\omega_0)\right) \tag{3.17}$$

or equivalently

$$x_{II}(t) = s(t - \tau_g(\omega_0))\cos\left[\omega_0\left(t - \frac{\phi_0 + \omega_0\tau_g(\omega_0)}{\omega_0}\right) + \theta\right] \tag{3.18}$$

Since, from Figure 3.4, we see that

$$\phi_1 = \phi_0 + \omega_0\tau_g(\omega_0)$$

equation (3.18) can be rewritten as

$$x_{II}(t) = s(t - \tau_g(\omega_0))\cos\left[\omega_0\left(t - \frac{\phi_1}{\omega_0}\right) + \theta\right] \tag{3.19a}$$

or

$$x_{II}(t) = s(t - \tau_g(\omega_0))\cos\left[\omega_0\left(t - \tau_p(\omega_0)\right) + \theta\right] \tag{3.19b}$$

where τ_p, referred to as the phase delay, is defined as $\tau_p = \frac{\phi_1}{\omega_0}$.

In summary, according to eqs. (3.18) and (3.19a), the time-domain effect of the nonlinear phase for the narrowband group of frequencies around the frequency ω_0 is to delay the narrowband signal by the group delay, $\tau_g(\omega_0)$, and apply an additional phase shift of $\frac{\phi_1}{\omega_0}$ to the carrier. An equivalent, alternate interpretation is that the time-domain envelope of the frequency group is delayed by the group delay and the carrier is delayed by the phase delay.

The discussion has been carried out thus far for narrowband signals. To extend the discussion to broadband signals, we need only recognize that any broadband signal can be viewed as a superposition of narrowband signals. This representation can in fact be developed formally by recognizing that the system in Figure 3.6 is an identity system, i.e. $r(t) = x(t)$ as long as

$$\sum_{i=0}^{\infty} H_i(j\omega) = 1 \tag{3.20}$$

By choosing the filters $H_i(j\omega)$ to satisfy eq. (3.20) and to be narrowband around center frequencies ω_i, each of the output signals, $y_i(t)$, is a narrowband signal. Consequently the time-domain effect of the phase of $G(j\omega)$ is to apply the group

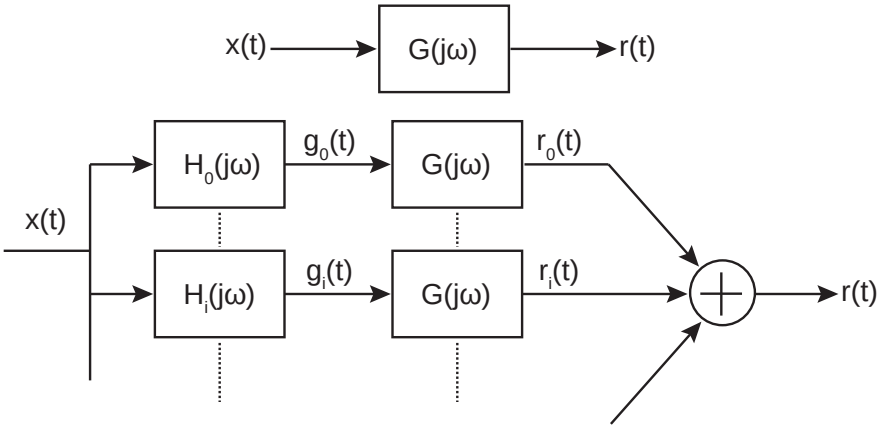

FIGURE 3.6 Continuous-time all-pass system with frequency response amplitude, phase and group delay as shown in Figure 3.7

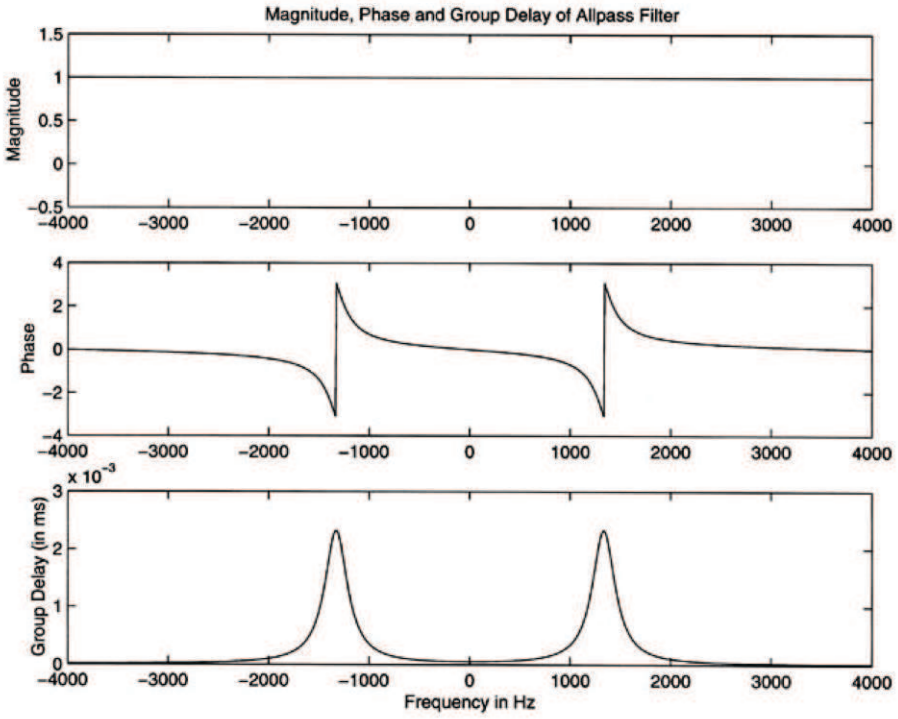

FIGURE 3.7 Magnitude, (nonlinear) phase, and group delay of an all-pass filter.

delay and phase delay to each of the narrowband components (i.e. frequency groups) $y_i(t)$. If the group delay is different at the different center (i.e. carrier) frequencies

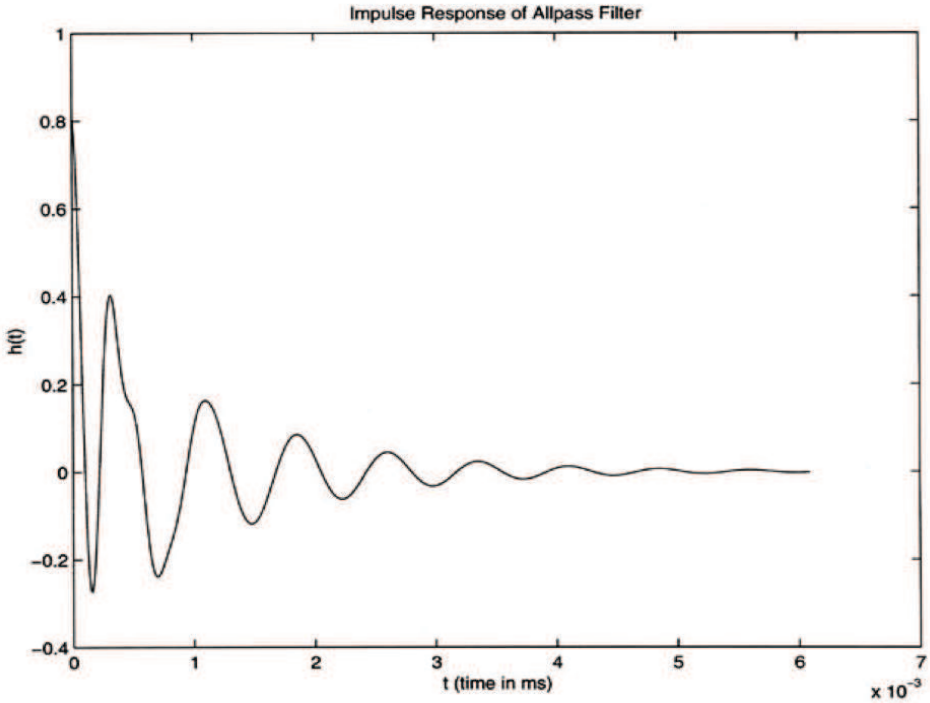

FIGURE 3.8 Impulse response for all-pass filter shown in Figure 3.7

ω_i, then the time domain effect is for different frequency groups to arrive at the output at different times.

As an illustration of this effect, consider $G(j\omega)$ in Figure 3.6 to be the continuous time all-pass system with frequency response amplitude, phase and group delay as shown in Figure 3.7. The corresponding impulse response is shown in Figure 3.8.

If the phase of $G(j\omega)$ were linear with frequency, the impulse response would simply be a delayed impulse, i.e. all the narrowband components would be delayed by the same amount and correspondingly would add up to a delayed impulse. However, as we see in Figure 3.7, the group delay is not constant since the phase is nonlinear. In particular, frequencies around 1200 Hz are delayed significantly more than around other frequencies. Correspondingly, in Figure 3.8 we see that frequency group appearing late in the impulse response.

A second example is shown in Figure 3.9, in which $G(j\omega)$ is again an all-pass system with nonlinear phase and consequently non-constant group delay. With this example, we would expect to see different delays in the frequency groups around $\omega = 2\pi \cdot 50$, $\omega = 2\pi \cdot 100$, and $\omega = 2\pi \cdot 300$ with the group at $\omega = 2\pi \cdot 50$ having the maximum delay and therefore appearing last in the impulse response.

In both of these examples, the input is highly concentrated in time (i.e. an impulse) and the response is dispersed in time because of the non-constant group delay, i.e.

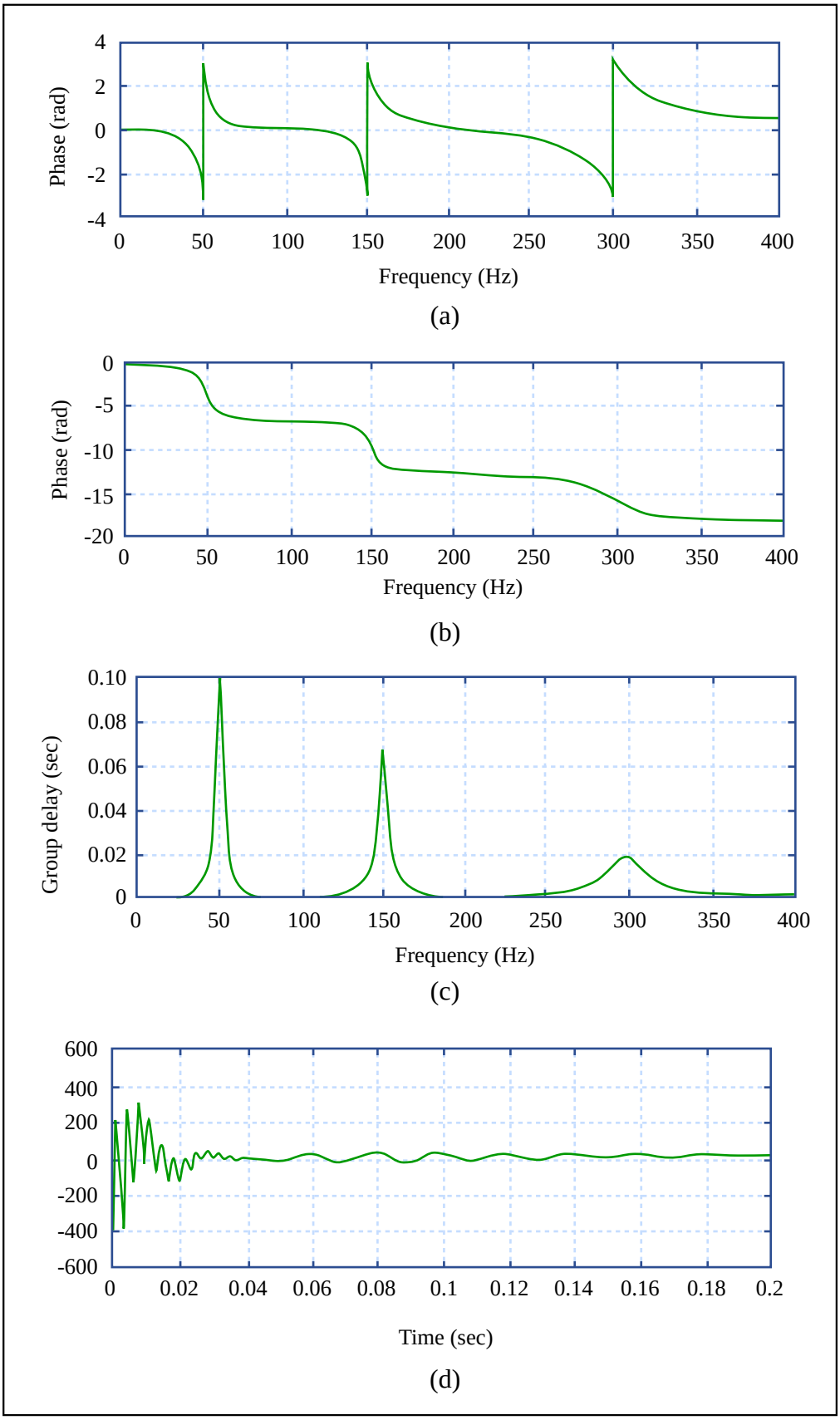

Image by MIT OpenCourseWare, adapted from *Signals and Systems*, Alan Oppenheim and Alan Willsky. Prentice Hall, 1996.

FIGURE 3.9 Phase, group delay, and impulse response for an all-pass system: (a) principal phase; (b) unwrapped phase; (c) group delay; (d) impulse response. (From Oppenheim and Willsky, *Signals and Systems*, Prentice Hall, 1997, Figure 6.5.)

the nonlinear phase. In general, the effect of nonlinear phase is referred to as dispersion. In communication systems and many other application contexts even when a channel has a relatively constant frequency response magnitude characteristic, nonlinear phase can result in significant distortion and other negative consequences because of the resulting time dispersion. For this reason, it is often essential to incorporate phase equalization to compensate for non-constant group-delay.

As a third example, we consider an all-pass system with phase and group delay as shown in Figure 3.10[1]. The input for this example is the touch-tone digit "five" which consists of two very narrowband tones at center frequencies 770 and 1336 Hz. The time-domain signal and its two narrowband component signals are shown in Figure 3.11.

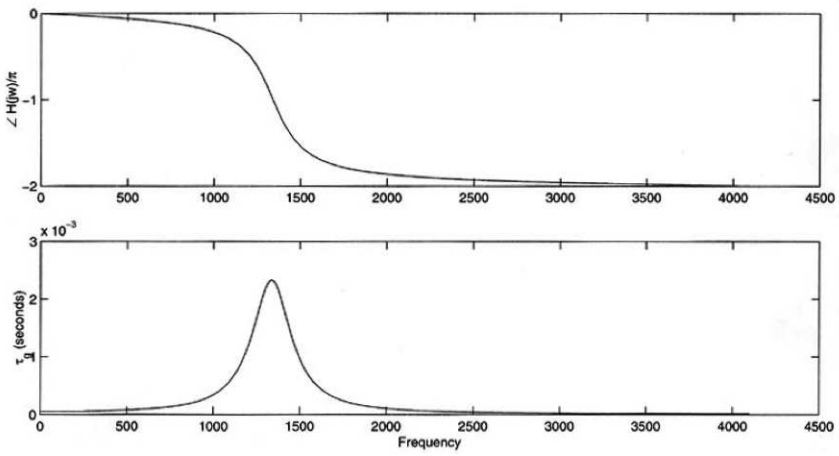

FIGURE 3.10 Phase and group delay for all-pass filter for touch-tone signal example.

The touch-tone signal is processed with multiple passes through the all-pass system of Figure 3.10. From the group delay plot, we expect that, in a single pass through the all-pass filter, the tone at 1336 Hz would be delayed by about 2.5 milliseconds relative to the tone at 770 Hz. After 200 passes, this would accumulate to a relative delay of about 0.5 seconds.

In Figure 3.12, we show the result of multiple passes through filters and the accumulation of the delays.

3.3 ALL-PASS AND MINIMUM-PHASE SYSTEMS

Two particularly interesting classes of stable LTI systems are all-pass systems and minimum-phase systems. We define and discuss them in this section.

[1]This example was developed by Prof. Bernard Lesieutre of the University of Wisconsin, Madison, when he taught the course with us at MIT

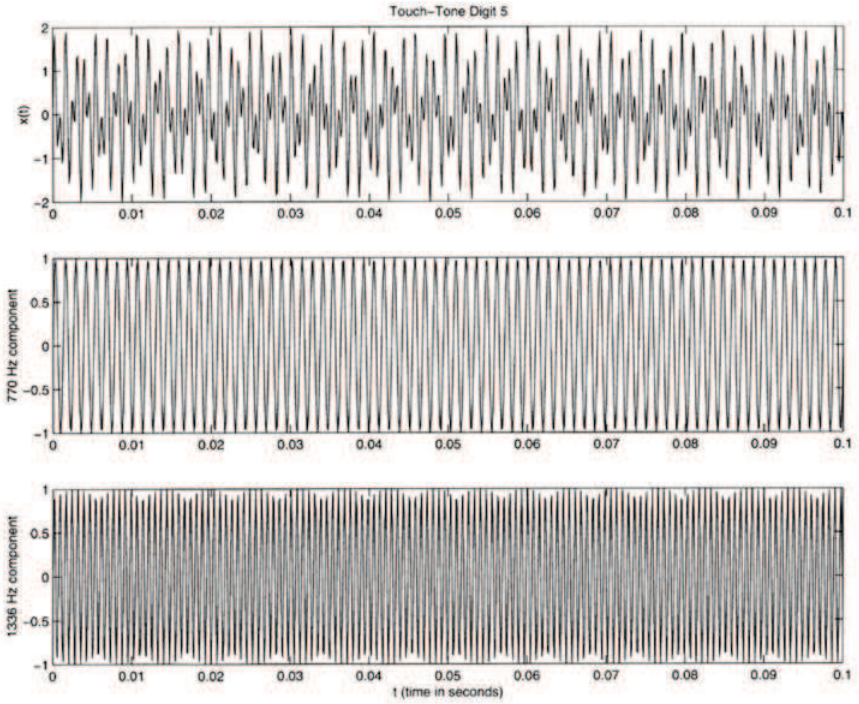

FIGURE 3.11 Touch-tone signal with its two narrowband component signals.

3.3.1 All-Pass Systems

An all-pass system is a stable system for which the magnitude of the frequency response is a constant, independent of frequency. The frequency response in the case of a continuous-time all-pass system is thus of the form

$$H_{ap}(j\omega) = Ae^{j\angle H_{ap}(j\omega)} , \tag{3.21}$$

where A is a constant, not varying with ω. Assuming the associated transfer function $H(s)$ is rational in s, it will correspondingly have the form

$$H_{ap}(s) = A\prod_{k=1}^{M}\frac{s+a_k^*}{s-a_k} . \tag{3.22}$$

Note that for each pole at $s = +a_k$ this has a zero at the mirror image across the imaginary axis, namely at $s = -a_k^*$; and if a_k is complex and the system impulse response is real-valued, every complex pole and zero will occur in a conjugate pair, so there will also be a pole at $s = +a_k^*$ and a zero at $s = -a_k$. An example of a pole-zero diagram (in the s-plane) for a continuous-time all-pass system is shown in Figure (3.13). It is straightforward to verify that each of the M factors in (3.22) has unit magnitude for $s = j\omega$.

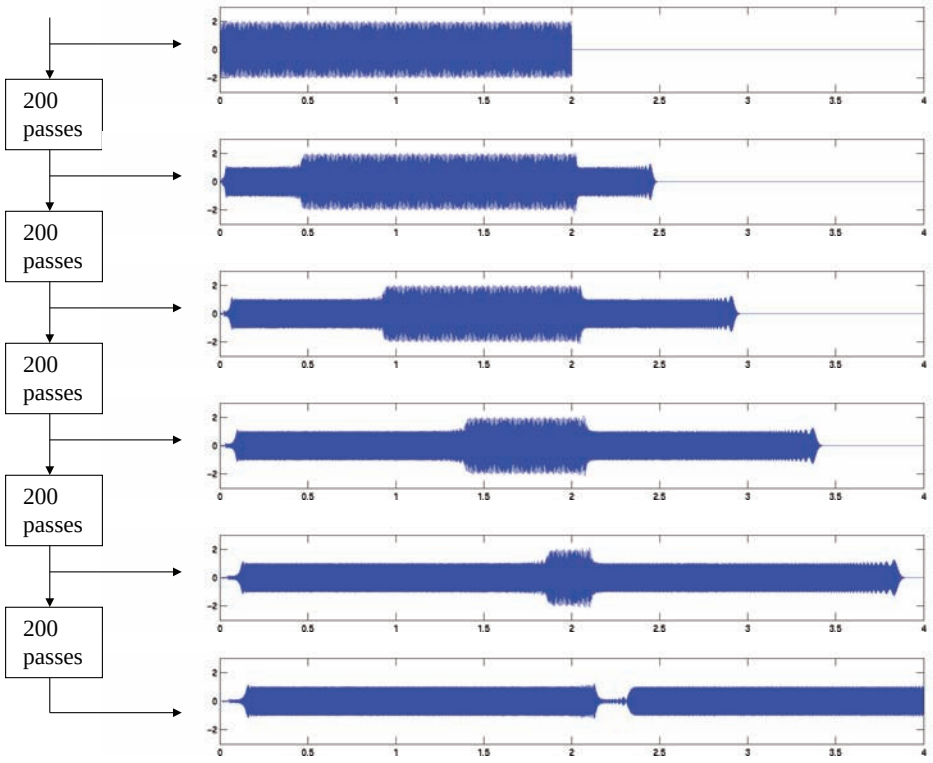

FIGURE 3.12 Effect of passing touchtone signal (Figure 3.11) through multiple passes of an all-pass filter and the accumulation of delays

.

For a discrete-time all-pass system, the frequency response is of the form

$$H_{ap}(e^{j\Omega}) = Ae^{j\angle H_{ap}(e^{j\Omega})} . \tag{3.23}$$

If the associated transfer function $H(z)$ is rational in z, it will have the form

$$H_{ap}(z) = A\prod_{k=1}^{M}\frac{z^{-1}-b_k^*}{1-b_kz^{-1}} . \tag{3.24}$$

The poles and zeros in this case occur at conjugate reciprocal locations: for each pole at $z = b_k$ there is a zero at $z = 1/b_k^*$. A zero at $z = 0$ (and associated pole at ∞) is obtained by setting $b_k = \infty$ in the corresponding factor above, after first dividing both the numerator and denominator by b_k; this results in the corresponding factor in (3.24) being just z. Again, if the impulse response is real-valued then every complex pole and zeros will occur in a conjugate pair, so there will be a pole at $z = b_k^*$ and a zero at $z = 1/b_k$. An example of a pole-zero diagram (in the z plane) for a discrete-time all-pass system is shown in Figure (3.14). It is once more

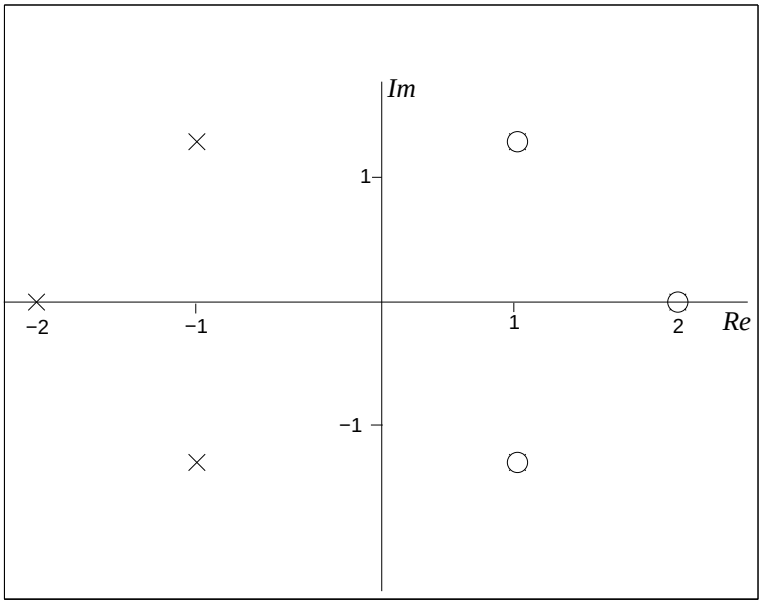

FIGURE 3.13 Typical pole-zero plot for a continuous-time all-pass system.

straightforward to verify that each of the M factors in (3.24) has unit magnitude for $z = e^{j\Omega}$.

The phase of a continuous-time all-pass system will be the sum of the phases associated with each of the M factors in (3.22). Assuming the system is causal (in addition to being stable), then for each of these factors $\mathrm{Re}\{a_k\} < 0$. With some algebra it can be shown that each factor of the form $\frac{s+a_k^*}{s-a_k}$ now has positive group delay at all frequencies, a property that we will make reference to shortly. Similarly, assuming causality (in addition to stability) for the discrete-time all-pass system in (3.24), each factor of the form $\frac{z^{-1}-b_k^*}{1-b_k z^{-1}}$ with $|b_k| < 1$ contributes positive group delay at all frequencies (or zero group delay in the special case of $b_k = 0$). Thus, in both continuous- and discrete-time, the frequency response of a causal all-pass system has constant magnitude and positive group delay at all frequencies.

3.3.2 Minimum-Phase Systems

In discrete-time, a stable system with a rational transfer function is called minimum-phase if its poles and zeros are all inside the unit circle, i.e., have magnitude less than unity. This is equivalent in the DT case to the statement that the system is stable and causal, and has a stable and causal inverse.

A similar definition applies in the case of a stable continuous-time system with a rational transfer function. Such a system is called minimum-phase if its poles and

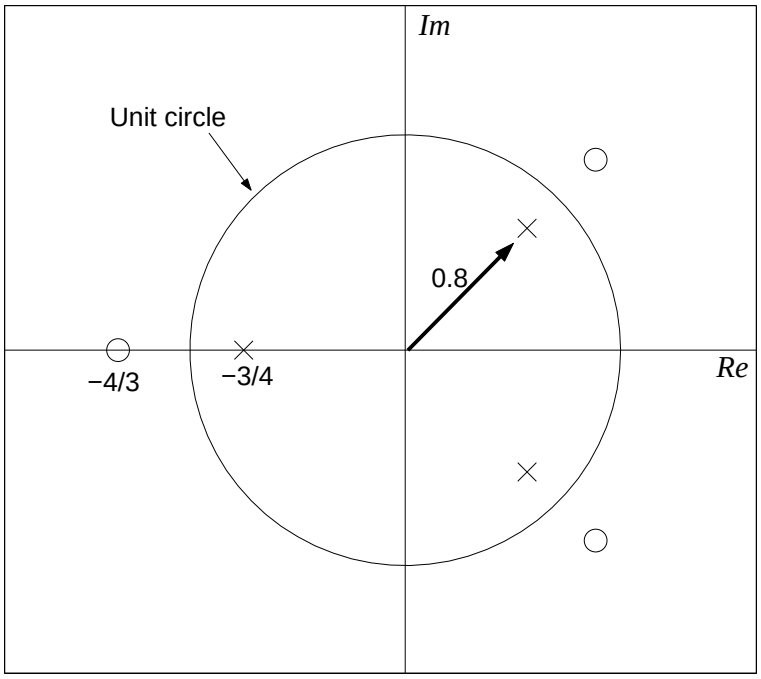

FIGURE 3.14 Typical pole-zero plot for a discrete-time all-pass system.

finite zeros are in the left-half-plane, i.e., have real parts that are negative. The system is therefore necessarily causal. If there are as many finite zeros as there are poles, then a CT minimum-phase system can equivalently be characterized by the statement that both the system and its inverse are stable and causal, just as we had in the DT case. However, it is quite possible — and indeed common — for a CT minimum-phase system to have fewer finite zeros than poles. (Note that a stable CT system must have all its poles at finite locations in the s-plane, since poles at infinity would imply that the output of the system involves derivatives of the input, which is incompatible with stability. Also, whereas in the DT case a zero at infinity is clearly outside the unit circle, in the CT case there is no way to tell if a zero at infinity is in the left half plane or not, so it should be no surprise that the CT definition involves only the finite zeros.)

The use of the term 'minimum phase' is historical, and the property should perhaps more appropriately be termed 'minimum group delay', for reasons that we will bring out next. To do this, we need a fact that we shall shortly establish: that any causal and stable CT system with a rational transfer function $H_{cs}(s)$ and no zeros on the imaginary axis can be represented as the cascade of a minimum-phase system and an all-pass system,

$$H_{cs}(s) = H_{min}(s)H_{ap}(s) \ . \tag{3.25}$$

Similarly, in the DT case, provided the transfer function $H_{cs}(z)$ has no zeros on

the unit circle, it can be written as

$$H_{cs}(z) = H_{min}(z)H_{ap}(z)\ . \tag{3.26}$$

The frequency response magnitude of the all-pass factor is constant, independent of frequency, and for convenience let us set this constant to unity. Then from (3.25)

$$|H_{cs}(j\omega)| = |H_{min}(j\omega)|\ , \quad \text{and} \tag{3.27a}$$

$$\text{grpdelay}[H_{cs}(j\omega)] = \text{grpdelay}[H_{min}(j\omega)] + \text{grpdelay}[H_{ap}(j\omega)] \tag{3.27b}$$

and similar equations hold in the DT case.

We will see in the next section that the minimum-phase term in (3.25) or (3.26) can be uniquely determined from the magnitude of $H_{cs}(j\omega)$, respectively $H_{cs}(e^{j\Omega})$. Consequently all causal, stable systems with the same frequency response magnitude differ only in the choice of the all-pass factor in (3.25) or (3.26). However, we have shown previously that all-pass factors must contribute positive group delay. Therefore we conclude from (3.27b) that among all causal, stable systems with the same CT frequency response magnitude, the one with no all-pass factors in (3.25) will have the minimum group delay. The same result holds in the DT case.

We shall now demonstrate the validity of (3.25); the corresponding result in (3.26) for discrete time follows in a very similar manner. Consider a causal, stable transfer function $H_{cs}(s)$ expressed in the form

$$H_{cs}(s) = A\frac{\prod_{k=1}^{M_1}(s-l_k)\prod_{i=1}^{M_2}(s-r_i)}{\prod_{n=1}^{N}(s-d_n)} \tag{3.28}$$

where the d_n's are the poles of the system, the l_k's are the zeros in the left-half plane and the r_i's are the zeros in the right-half plane. Since $H_{cs}(s)$ is stable and causal, all of the poles are in the left-half plane and would be associated with the factor $H_{min}(s)$ in (3.25), as would be all of the zeros l_k. We next represent the right-half-plane zeros as

$$\prod_{i=1}^{M_2}(s-r_i) = \prod_{i=1}^{M_2}(s+r_i)\prod_{i=1}^{M_2}\frac{(s-r_i)}{(s+r_i)} \tag{3.29}$$

Since $\text{Re}\{r_i\}$ is positive, the first factor in (3.29) represents left-half-plane zeros. The second factor corresponds to all-pass terms with left-half-plane poles, and with zeros at mirror image locations to the poles. Thus, combining (3.28) and (3.29), $H_{cs}(s)$ has been decomposed according to (3.25) where

$$H_{min}(s) = A\frac{\prod_{k=1}^{M_1}(s-l_k)\prod_{i=1}^{M_2}(s+r_i)}{\prod_{n=1}^{N}(s-d_n)} \tag{3.30a}$$

$$H_{ap}(s) = \prod_{i=1}^{M_2}\frac{(s-r_i)}{(s+r_i)} \tag{3.30b}$$

EXAMPLE 3.3 Causal, stable system as cascade of minimum-phase and all-pass

Consider a causal, stable system with transfer function

$$H_{cs} = \frac{(s-1)}{(s+2)(s+3)} \tag{3.31}$$

The corresponding minimum-phase and all-pass factors are

$$H_{min}(s) = \frac{(s+1)}{(s+2)(s+3)} \tag{3.32}$$

$$H_{ap}(s) = \frac{s-1}{s+1} \tag{3.33}$$

3.4 SPECTRAL FACTORIZATION

The minimum-phase/all-pass decomposition developed above is useful in a variety of contexts. One that is of particular interest to us in later chapters arises when we we are given or have measured the magnitude of the frequency response of a stable system with a rational transfer function $H(s)$ (and real-valued impulse response), and our objective is to recover $H(s)$ from this information. A similar task may be posed in the DT case, but we focus on the CT version here. We are thus given

$$|H(j\omega)|^2 = H(j\omega)H^*(j\omega) \tag{3.34}$$

or, since $H^*(j\omega) = H(-j\omega)$,

$$|H(j\omega)|^2 = H(j\omega)H(-j\omega)\ . \tag{3.35}$$

Now $H(j\omega)$ is $H(s)$ for $s = j\omega$, and therefore

$$|H(j\omega)|^2 = H(s)H(-s)\Big|_{s=j\omega} \tag{3.36}$$

For any numerator or denominator factor $(s-a)$ in $H(s)$, there will be a corresponding factor $(-s-a)$ in $H(s)H(-s)$. Thus $H(s)H(-s)$ will consist of factors in the numerator or denominator of the form $(s-a)(-s-a) = -s^2 + a^2$, and will therefore be a rational function of s^2. Consequently $|H(j\omega)|^2$ will be a rational function of ω^2. Thus, if we are given or can express $|H(j\omega)|^2$ as a rational function of ω^2, we can obtain the product $H(s)H(-s)$ by making the substitution $\omega^2 = -s^2$.

The product $H(s)H(-s)$ will always have its zeros in pairs that are mirrored across the imaginary axis of the s-plane, and similarly for its poles. For any pole or zero of $H(s)H(-s)$ at the real value a, there will be another at the mirror image $-a$, while for any pole or zero at the complex value q, there will be others at q^*, $-q$ and $-q^*$,

forming a complex conjugate pair (q, q^*) and its mirror image $(-q^*, -q)$. We then need to assign one of each mirrored real pole and zero and one of each mirrored conjugate pair of poles and zeros to $H(s)$, and the mirror image to $H(-s)$.

If we assume (or know) that $H(s)$ is causal, in addition to being stable, then we would assign the left-half plane poles of each pair to $H(s)$. With no further knowledge or assumption we have no guidance on the assignment of the zeros other than the requirement of assigning one of each mirror image pair to $H(s)$ and the other to $H(-s)$. If we further know or assume that the system is minimum-phase, then the left-half-plane zeros from each mirrored pair are assigned to $H(s)$, and the right-half-plane zeros to $H(-s)$. This process of factoring $H(s)H(-s)$ to obtain $H(s)$ is referred to as spectral factorization.

EXAMPLE 3.4 Spectral factorization

Consider a frequency response magnitude that has been measured or approximated as

$$|H(j\omega)|^2 = \frac{\omega^2 + 1}{\omega^4 + 13\omega^2 + 36} = \frac{\omega^2 + 1}{(\omega^2 + 4)(\omega^2 + 9)} \tag{3.37}$$

Making the substitution $\omega^2 = -s^2$, we obtain

$$H(s)H(-s) = \frac{-s^2 + 1}{(-s^2 + 4)(-s^2 + 9)} \tag{3.38}$$

which we further factor as

$$H(s)H(-s) = \frac{(s + 1)(-s + 1)}{(s + 2)(-s + 2)(s + 3)(-s + 3)} \tag{3.39}$$

It now remains to associate appropriate factors with $H(s)$ and $H(-s)$. Assuming the system is causal in addition to being stable, the two left-half plane poles at $s = -2$ and $s = -3$ must be associated with $H(s)$. With no further assumptions, either one of the numerator factors can be associated with $H(s)$ and the other with $H(-s)$. However, if we know or assume that $H(s)$ is minimum phase, then we would assign the left-half plane zero to $H(s)$, resulting in the choice

$$H(s) = \frac{(s + 1)}{(s + 2)(s + 3)} \tag{3.40}$$

In the discrete-time case, a similar development leads to an expression for $H(z)H(1/z)$ from knowledge of $|H(e^{j\Omega})|^2$. The zeros of $H(z)H(1/z)$ occur in conjugate reciprocal pairs, and similarly for the poles. We again have to split such conjugate reciprocal pairs, assigning one of each to $H(z)$, the other to $H(1/z)$, based on whatever additional knowledge we have. For instance, if $H(z)$ is known to be causal in addition to being stable, then all the poles of $H(z)H(1/z)$ that are in the unit circle are assigned to $H(z)$; and if $H(z)$ is known to be minimum phase as well, then all the zeros of $H(z)H(1/z)$ that are in the unit circle are assigned to $H(z)$.

CHAPTER 4

State-Space Models

4.1 INTRODUCTION

In our discussion of system descriptions up to this point, we have emphasized and utilized system models that represent the transformation of input signals into output signals. In the case of linear and time-invariant (LTI) models, our focus has been on the impulse response, frequency response and transfer function. Such input-output models do not directly consider the internal behavior of the systems they model.

In this chapter we begin a discussion of system models that considers the internal dynamical behavior of the system as well as the input-output characteristics. Internal behavior can be important for a variety of reasons. For example, in examining issues of stability, a system can be stable from an input-output perspective but hidden internal variables may be unstable, yielding what we would want to think of as unstable system behavior.

We introduce in this chapter an important model description that highlights internal behavior of the system and is specially suited to representing causal systems for real-time applications such as control. Specifically, we introduce state-space models for finite-memory (or lumped) causal systems. These models exist for both continuous-time (CT) and discrete-time (DT) systems, and for nonlinear, time-varying systems — although our focus will be on the LTI case.

Having a state-space model for a causal DT system (similar considerations apply in the CT case) allows us to answer a question that gets asked about such systems in many settings: Given the input value $x[n]$ at some arbitrary time n, how much information do we really need about past inputs, i.e., about $x[k]$ for $k < n$, in order to determine the present output $y[n]$? As the system is causal, we know that having all past $x[k]$ (in addition to $x[n]$) will suffice, but do we actually need this much information? This question addresses the issue of memory in the system, and is a worthwhile question for a variety of reasons.

For example, the answer gives us an idea of the complexity, or number of degrees of freedom, associated with the dynamic behavior of the system. The more information we need about past inputs in order to determine the present output, the richer the variety of possible output behaviors, i.e., the more ways we can be surprised in the absence of information about the past.

Furthermore, in a control application, the answer to the above question suggests the required degree of complexity of the controller, because the controller has to

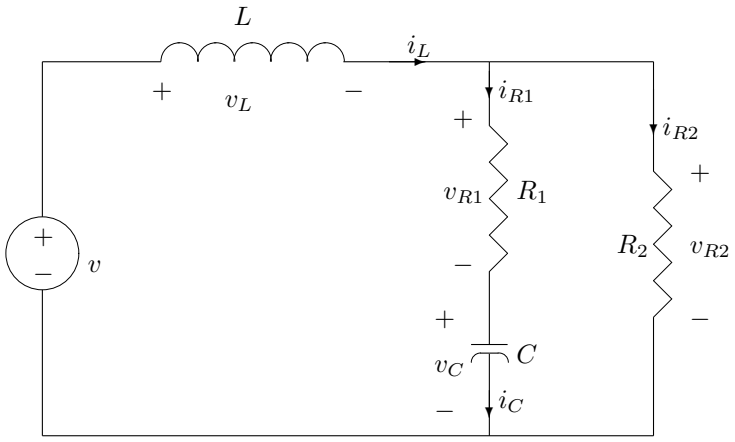

FIGURE 4.1 RLC circuit.

remember enough about the past to determine the effects of present control actions on the response of the system. In addition, for a computer algorithm that acts causally on a data stream, the answer to the above question suggests how much memory will be needed to run the algorithm.

With a state-space description, everything about the past that is relevant to the present and future is summarized in the present state, i.e., in the present values of a set of state variables. The number of state variables, which we refer to as the order of the model, thus indicates the amount of memory or degree of complexity associated with the system or model.

4.2 INPUT-OUTPUT AND INTERNAL DESCRIPTIONS

As a prelude to developing the general form of a state-space model for an LTI system, we present two examples, one in CT and the other in DT.

4.2.1 An RLC circuit

Consider the RLC circuit shown in Figure 4.1. We have labeled all the component voltages and currents in the figure.

The defining equations for the components are:

$$\begin{aligned} L\frac{di_L(t)}{dt} &= v_L(t) \\ C\frac{dv_C(t)}{dt} &= i_C(t) \\ v_{R1}(t) &= R_1 i_{R1}(t) \\ v_{R2}(t) &= R_2 i_{R2}(t)\ , \end{aligned} \tag{4.1}$$

while the voltage source is defined by the condition that its voltage is $v(t)$ regardless of its current $i(t)$. Kirchhoff's voltage and current laws yield

$$\begin{aligned} v(t) &= v_L(t) + v_{R2}(t) \\ v_{R2}(t) &= v_{R1}(t) + v_C(t) \\ i(t) &= i_L(t) \\ i_L(t) &= i_{R1}(t) + i_{R2}(t) \\ i_{R1}(t) &= i_C(t) \ . \end{aligned} \tag{4.2}$$

All these equations together constitute a detailed and explicit representation of the circuit.

Let us take the voltage source $v(t)$ as the input to the circuit; we shall also denote this by $x(t)$, our standard symbol for inputs. Choose any of the circuit voltages or currents as the output — let us choose $v_{R_2}(t)$ for this example, and also denote it by $y(t)$, our standard symbol for outputs. We can then combine (4.1) and (4.2) using, for example, Laplace transforms, in order to obtain a transfer function or a linear constant-coefficient differential equation relating the input and output. The coefficients in the transfer function or differential equation will, of course be functions of the values of the components in the circuit. The resulting transfer function $H(s)$ from input to output is

$$H(s) = \frac{Y(s)}{X(s)} = \frac{\alpha\Big(\frac{R_1}{L}s + \frac{1}{LC}\Big)}{s^2 + \alpha\Big(\frac{1}{R_2C} + \frac{R_1}{L}\Big)s + \alpha\frac{1}{LC}} \tag{4.3}$$

where α denotes the ratio $R_2/(R_1 + R_2)$. The corresponding input-output differential equation is

$$\frac{d^2y(t)}{dt^2} + \alpha\Big(\frac{1}{R_2C} + \frac{R_1}{L}\Big)\frac{dy(t)}{dt} + \alpha\Big(\frac{1}{LC}\Big)y(t) = \alpha\Big(\frac{R_1}{L}\Big)\frac{dx(t)}{dt} + \alpha\Big(\frac{1}{LC}\Big)x(t) \ . \tag{4.4}$$

An important characteristic of a circuit such as in Figure 4.1 is that the behavior for a time interval beginning at some t is completely determined by the input trajectory in that interval as well as the inductor currents and capacitor voltages at time t. Thus, for the specific circuit in Figure 4.1, in determining the response for times $\geq t$, the relevant past history of the system is summarized in $i_L(t)$ and $v_C(t)$. The inductor currents and capacitor voltages in such a circuit at any time t are commonly referred to as state variables, and the particular set of values they take constitutes the state of the system at time t. This state, together with the input from t onwards, are sufficient to completely determine the response at and beyond t.

The concept of state for dynamical systems is an extremely powerful one. For the RLC circuit of Figure 4.1 it motivates us to reduce the full set of equations (4.1) and (4.2) into a set of equations involving just the input, output, and internal variables $i_L(t)$ and $v_C(t)$. Specifically, a description of the desired form can be found by appropriately eliminating the other variables from (4.1) and (4.2), although some

attention is required in order to carry out the elimination efficiently. With this, we arrive at a condensed description, written here using matrix notation, and in a format that we shall encounter frequently in this chapter and the next two:

$$\begin{pmatrix} di_L(t)/dt \\ dv_C(t)/dt \end{pmatrix} = \begin{pmatrix} -\alpha R_1/L & -\alpha/L \\ \alpha/C & -1/(R_1+R_2)C \end{pmatrix} \begin{pmatrix} i_L(t) \\ v_C(t) \end{pmatrix} + \begin{pmatrix} 1/L \\ 0 \end{pmatrix} v(t) \,. \tag{4.5}$$

The use of matrix notation is a convenience; we could of course have simply written the above description as two separate but coupled first-order differential equations with constant coefficients.

We shall come to appreciate the properties and advantages of a description in the form of (4.5), referred to as a CT (and, in this case, LTI) state-space form. Its key feature is that it expresses the rates of change of the state variables at any time t as functions (in this case, LTI functions) of their values and those of the input at that same time t.

As we shall see later, the state-space description can be used to solve for the state variables $i_L(t)$ and $v_C(t)$, given the input $v(t)$ and appropriate auxiliary information (specifically, initial conditions on the state variables). Furthermore, knowledge of $i_L(t)$, $v_C(t)$ and $v(t)$ suffices to reconstruct all the other voltages and currents in the circuit at time t. In particular, any output variable can be written in terms of the retained variables. For instance, if the output of interest for this circuit is the voltage $v_{R2}(t)$ across R_2, we can write (again in matrix notation)

$$v_{R2}(t) = \begin{pmatrix} \alpha R_1 & \alpha \end{pmatrix} \begin{pmatrix} i_L(t) \\ v_C(t) \end{pmatrix} + (\,0\,)\, v(t) \,. \tag{4.6}$$

For this particular example, the output does not involve the input $v(t)$ directly — hence the term $(\,0\,)\, v(t)$ in the above output equation — but in the general case the output equation will involve present values of any inputs in addition to present values of the state variables.

4.2.2 A delay-adder-gain system

For DT systems, the role of state variables is similar to the role discussed in the preceding subsection for CT systems. We illustrate this with the system described by the delay-adder-gain block diagram shown in Figure 4.2.2. The corresponding detailed equations relating the indicated signals are

$$\begin{aligned} q_1[n+1] &= q_2[n] \\ q_2[n+1] &= p[n] \\ p[n] &= x[n] - (1/2)q_1[n] + (3/2)q_2[n] \\ y[n] &= q_2[n] + p[n] \,. \end{aligned} \tag{4.7}$$

The equations in (4.7) can be combined together using, for example, z-transform methods, to obtain the transfer function or linear constant-coefficient difference equation relating input and output:

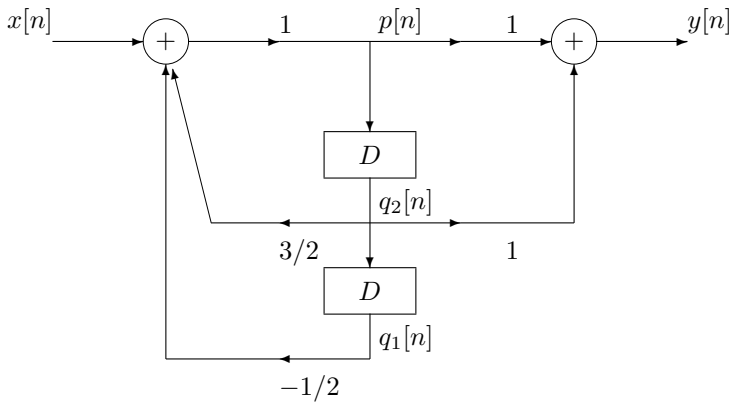

FIGURE 4.2 Delay-adder-gain block diagram.

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1 + z^{-1}}{1 - \frac{3}{2}z^{-1} + \frac{1}{2}z^{-2}} \tag{4.8}$$

and

$$y[n] - \frac{3}{2}y[n-1] + \frac{1}{2}y[n-2] = x[n] + x[n-1]\,. \tag{4.9}$$

The response of the system in an interval of time $\geq n$ is completely determined by the input for times $\geq n$ and the values $q_1[n]$ and $q_2[n]$ that are stored at the outputs of the delay elements at time n. Thus, as with the energy storage elements in the circuit of Figure 4.1, the delay elements in the delay-adder-gain system capture the state of the system at any time, i.e., summarize all the past history with respect to how it affects the present and future response of the system. Consequently, we condense (4.7) in terms of only the input, output and state variables to obtain the following matrix equations:

$$\begin{pmatrix} q_1[n+1] \\ q_2[n+1] \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1/2 & 3/2 \end{pmatrix} \begin{pmatrix} q_1[n] \\ q_2[n] \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} x[n] \tag{4.10}$$

$$y[n] = \begin{pmatrix} -1/2 & 5/2 \end{pmatrix} \begin{pmatrix} q_1[n] \\ q_2[n] \end{pmatrix} + (1)x[n]\,. \tag{4.11}$$

In this case it is quite easy to see that, if we are given the values $q_1[n]$ and $q_2[n]$ of the state variables at some time n, and also the input trajectory from n onwards, i.e., $x[n]$ for times $\geq n$, then we can compute the values of the state variables for all times $> n$, and the output for all times $\geq n$. All that is needed is to iteratively apply (4.10) to find $q_1[n+1]$ and $q_2[n+1]$, then $q_1[n+2]$ and $q_2[n+2]$, and so on for increasing time arguments, and to use (4.11) at each time to find the output.

4.3 STATE-SPACE MODELS

As illustrated in Sections 4.2.1 and 4.2.2, it is often natural and convenient, when studying or modeling physical systems, to focus not just on the input and output signals but rather to describe the interaction and time-evolution of several key variables or signals that are associated with the various component processes internal to the system. Assembling the descriptions of these components and their interconnections leads to a description that is richer than an input–output description. In particular, in Sections 4.2.1 and 4.2.2 the description is in terms of the time evolution of variables referred to as the state variables, which completely capture at any time the past history of the system as it affects the present and future response. We turn now to a more formal definition of state-space models in the DT and CT cases, followed by a discussion of two defining characteristics of such models.

4.3.1 DT State-Space Models

A state-space model is built around a set of state variables; the number of state variables in a model or system is referred to as its order. Although we shall later cite examples of distributed or infinite-order systems, we shall only deal with state-space models of finite order, which are also referred to as lumped systems. For an Lth-order model in the DT case, we shall generically denote the values of the L state variables at time n by $q_1[n], q_2[n], \cdots, q_L[n]$. It is convenient to gather these variables into a state vector:

$$\mathbf{q}[n] = \begin{pmatrix} q_1[n] \\ q_2[n] \\ \vdots \\ q_L[n] \end{pmatrix} . \tag{4.12}$$

The value of this vector constitutes the state of the model or system at time n.

A DT LTI state-space model with single (i.e., scalar) input $x[n]$ and single output $y[n]$ takes the following form, written in compact matrix notation:

$$\mathbf{q}[n+1] = \mathbf{A}\mathbf{q}[n] + \mathbf{b}x[n] , \tag{4.13}$$

$$y[n] = \mathbf{c}^T\mathbf{q}[n] + \mathtt{d}x[n] . \tag{4.14}$$

In (4.13), $\mathbf{A}$ is an $L \times L$ matrix, $\mathbf{b}$ is an $L \times 1$ matrix or column-vector, and $\mathbf{c}^T$ is a $1 \times L$ matrix or row-vector, with the superscript ^{T} denoting transposition of the column vector $\mathbf{c}$ into the desired row vector. The quantity $\mathtt{d}$ is a 1×1 matrix, i.e., a scalar. The entries of all these matrices in the case of an LTI model are numbers or constants or parameters, so they do not vary with n. Note that the model we arrived at in (4.10) and (4.11) of Section 4.2.2 has precisely the above form. We refer to (4.13) as the state evolution equation, and to (4.14) as the output equation. These equations respectively express the next state and the current output at any time as an LTI combination of the current state variables and current input.

Generalizations of the DT LTI State-Space Model. There are various nat-

ural generalizations of the above DT LTI single-input, single-output state-space model. A multi-input DT LTI state-space model replaces the single term $\mathbf{b}x[n]$ in (4.13) by a sum of terms, $\mathbf{b}_1 x_1[n] + \cdots + \mathbf{b}_M x_M[n]$, where M is the number of inputs. This corresponds to replacing the scalar input $x[n]$ by an M-component vector $\mathbf{x}[n]$ of inputs, with a corresponding change of $\mathbf{b}$ to a matrix $\mathbf{B}$ of dimension $L \times M$. Similarly, for a multi-output DT LTI state-space model, the single output equation (4.14) is replaced by a collection of such output equations, one for each of the P outputs. Equivalently, the scalar output $y[n]$ is replaced by a P-component vector $\mathbf{y}[n]$ of outputs, with a corresponding change of $\mathbf{c}^T$ and $\mathtt{d}$ to matrices $\mathbf{C}^T$ and $\mathbf{D}$ of dimension $P \times L$ and $P \times M$ respectively.

A linear but time-varying DT state-space model takes the same form as in (4.13) and (4.14) above, except that some or all of the matrix entries are time-varying. A linear but periodically varying model is a special case of this, with matrix entries that all vary periodically with a common period. A nonlinear, time-invariant model expresses $\mathbf{q}[n+1]$ and $\mathbf{y}[n]$ as nonlinear but time-invariant functions of $\mathbf{q}[n]$ and $\mathbf{x}[n]$, rather than as the LTI functions embodied by the matrix expressions on the right-hand-sides of (4.13) and (4.14). A nonlinear, time-varying model expresses $\mathbf{q}[n+1]$ and $\mathbf{y}[n]$ as nonlinear, time-varying functions of $\mathbf{q}[n]$ and $\mathbf{x}[n]$, and one can also define nonlinear, periodically varying models as a particular case in which the time-variations are periodic with a common period.

4.3.2 CT State-Space Models

Continuous-time state-space descriptions take a very similar form to the DT case. We denote the state variables as $q_i(t)$, $i = 1, 2, ..., L$, and the state vector as

$$\mathbf{q}(t) = \begin{pmatrix} q_1(t) \\ q_2(t) \\ \vdots \\ q_L(t) \end{pmatrix} . \tag{4.15}$$

Whereas in the DT case the state evolution equation expresses the state vector at the next time step in terms of the current state vector and input values, in CT the state evolution equation expresses the rates of change (i.e., derivatives) of each of the state variables as functions of the present state and inputs. The general Lth-order CT LTI state-space representation thus takes the form

$$\frac{d\mathbf{q}(t)}{dt} = \dot{\mathbf{q}}(t) = \mathbf{A}\mathbf{q}(t) + \mathbf{b}x(t) , \tag{4.16}$$

$$y(t) = \mathbf{c}^T\mathbf{q}(t) + \mathtt{d}x(t) , \tag{4.17}$$

where $d\mathbf{q}(t)/dt = \dot{\mathbf{q}}(t)$ denotes the vector whose entries are the derivatives, $dq_i(t)/dt$, of the corresponding entries, $q_i(t)$, of $\mathbf{q}(t)$. Note that the model in (4.5) and (4.6) of Section 4.2.1 is precisely of the above form.

Generalizations to multi-input and multi-output models, and to linear and nonlinear time-varying or periodic models, can be described just as in the case of DT systems, by appropriately relaxing the restrictions on the form of the right-hand sides of (4.16), (4.17). We shall see an example of a nonlinear time-invariant state-space model in Section 1.

4.3.3 Characteristics of State-Space Models

The designations of "state" for $\mathbf{q}[n]$ or $\mathbf{q}(t)$, and of "state-space description" for (4.13), (4.14) and (4.16), (4.17) — or for the various generalizations of these equations — follow from the following two key properties of such models.

State Evolution Property: The state at any initial time, along with the inputs over any interval from that initial time onwards, determine the state over that entire interval. Everything about the past that is relevant to the future state is embodied in the present state.

Instantaneous Output Property: The outputs at any instant can be written in terms of the state and inputs at that same instant.

The state evolution property is what makes state-space models particularly well suited to describing causal systems. In the DT case, the validity of this state evolution property is evident from the state evolution equation (4.13), which allows us to update $\mathbf{q}[n]$ iteratively, going from time n to time $n+1$ using only knowledge of the present state and input. The same argument can also be applied to the generalizations of DT LTI models that we outlined earlier.

The state evolution property should seem intuitively reasonable in the CT case as well. Specifically, knowledge of both the state and the rate of change of the state at any instant allows us to compute the state after a small increment in time. Taking this small step forward, we can re-evaluate the rate of change of the state, and step forward again. A more detailed proof of this property in the general nonlinear and/or time-varying CT case essentially proceeds this way, and is treated in texts that deal with the existence and uniqueness of solutions of differential equations. These more careful treatments also make clear what additional conditions are needed for the state evolution property to hold in the general case. However, the CT LTI case is much simpler, and we shall demonstrate the state evolution property for this class of state-space models in the next chapter, when we show how to explicitly solve for the behavior of such systems.

The instantaneous output property is immediately evident from the output equations (4.14), (4.17). It also holds for the various generalizations of basic single-input, single-output LTI models that we listed earlier.

The two properties above may be considered the defining characteristics of a state-space model. In effect, what we do in setting up a state-space model is to introduce the additional vector of state variables $\mathbf{q}[n]$ or $\mathbf{q}(t)$, to supplement the input variables $x[n]$ or $x(t)$ and output variables $y[n]$ or $y(t)$. This supplementation is done precisely in order to obtain a description that satisfies the two properties above.

Often there are natural choices of state variables suggested directly by the particular context or application. In both DT and CT cases, state variables are related to the "memory" of the system. In many physical situations involving CT models, the state variables are associated with energy storage, because this is what is carried over from the past to the future. Natural state variables for electrical circuits are thus the inductor currents and capacitor voltages, as turned out to be the case in Section 4.2.1. For mechanical systems, natural state variables are the positions and velocities of all the masses in the system (corresponding respectively to potential energy and kinetic energy variables), as we will see in later examples. In the case of a CT integrator-adder-gain block diagram, the natural state variables are associated with the outputs of the integrators, just as in the DT case the natural state variables of a delay-adder-gain model are the outputs of the delay elements, as was the case in the example of Section 4.2.2.

In any of the above contexts, one can choose any alternative set of state variables that together contain exactly the same information. There are also situations in which there is no particularly natural or compelling choice of state variables, but in which it is still possible to define supplementary variables that enable a valid state-space description to be obtained.

Our discussion of the two key properties above — and particularly of the role of the state vector in separating past and future — suggests that state-space models are particularly suited to describing causal systems. In fact, state-space models are almost never used to describe non-causal systems. We shall always assume here, when dealing with state-space models, that they represent causal systems. Although causality is not a central issue in analyzing many aspects of communication or signal processing systems, particularly in non-real-time contexts, it is generally central to simulation and control design for dynamic systems. It is accordingly in such dynamics and control settings that state-space descriptions find their greatest value and use.

4.4 EQUILIBRIA AND LINEARIZATION OF NONLINEAR STATE-SPACE MODELS

An LTI state-space model most commonly arises as an approximate description of the local (or "small-signal") behavior of a nonlinear time-invariant model, for small deviations of its state variables and inputs from a set of constant equilibrium values. In this section we present the conditions that define equilibrium, and describe the role of linearization in obtaining the small-signal model at this equilibrium.

4.4.1 Equilibrium

To make things concrete, consider a DT 3rd-order nonlinear time-invariant state-space system, of the form

$$\begin{aligned} q_1[n+1] &= f_1\Big(q_1[n], q_2[n], q_3[n], x[n]\Big) \\ q_2[n+1] &= f_2\Big(q_1[n], q_2[n], q_3[n], x[n]\Big) \\ q_3[n+1] &= f_3\Big(q_1[n], q_2[n], q_3[n], x[n]\Big) \ , \end{aligned} \tag{4.18}$$

with the output $y[n]$ defined by the equation

$$y[n] = g\Big(q_1[n], q_2[n], q_3[n], x[n]\Big) \ . \tag{4.19}$$

The state evolution functions $f_i(\,\cdot\,)$, for $i = 1, 2, 3$, and the output function $g(\,\cdot\,)$ are all time-invariant nonlinear functions of the three state variables $q_i[n]$ and the input $x[n]$. (Time-invariance of the functions simply means that they combine their arguments in the same way, regardless of the time index n.) The generalization to an Lth-order description should be clear. In vector notation, we can simply write

$$\mathbf{q}[n+1] = \mathbf{f}\Big(\mathbf{q}[n], x[n]\Big) \ , \qquad y[n] = g\Big(\mathbf{q}[n], x[n]\Big) \ , \tag{4.20}$$

where for our 3rd-order case

$$\mathbf{f}(\,\cdot\,) = \begin{bmatrix} f_1(\,\cdot\,) \\ f_2(\,\cdot\,) \\ f_3(\,\cdot\,) \end{bmatrix} \ . \tag{4.21}$$

Suppose now that the input $x[n]$ is constant at the value $\overline{x}$ for all n. The corresponding state equilibrium is a state value $\overline{\mathbf{q}}$ with the property that if $\mathbf{q}[n] = \overline{\mathbf{q}}$ with $x[n] = \overline{x}$, then $\mathbf{q}[n+1] = \overline{\mathbf{q}}$. Equivalently, the point $\overline{\mathbf{q}}$ in the state space is an equilibrium (or equilibrium point) if, with $x[n] \equiv \overline{x}$ for all n and with the system initialized at $\overline{\mathbf{q}}$, the system subsequently remains fixed at $\overline{\mathbf{q}}$. From (4.20), this is equivalent to requiring

$$\overline{\mathbf{q}} = \mathbf{f}(\overline{\mathbf{q}}, \overline{x}) \ . \tag{4.22}$$

The corresponding equilibrium output is

$$\overline{y} = g(\overline{\mathbf{q}}, \overline{x}) \ . \tag{4.23}$$

In defining an equilibrium, no consideration is given to what the system behavior is in the vicinity of the equilibrium point, i.e., of how the system will behave if initialized close to — rather than exactly at — the point $\overline{\mathbf{q}}$. That issue is picked up when one discusses local behavior, and in particular local stability, around the equilibrium.

In the 3rd-order case above, and given $\overline{x}$, we would find the equilibrium by solving the following system of three simultaneous nonlinear equations in three unknowns:

$$\begin{aligned}\overline{q_1} &= f_1(\overline{q_1}, \overline{q_2}, \overline{q_3}, \overline{x}) \\ \overline{q_2} &= f_2(\overline{q_1}, \overline{q_2}, \overline{q_3}, \overline{x}) \\ \overline{q_3} &= f_3(\overline{q_1}, \overline{q_2}, \overline{q_3}, \overline{x}) \ .\end{aligned} \tag{4.24}$$

There is no guarantee in general that an equilibrium exists for the specified constant input $\overline{x}$, and there is no guarantee of a unique equilibrium when an equilibrium does exist.

We can apply the same idea to CT nonlinear time-invariant state-space systems. Again consider the concrete case of a 3rd-order system:

$$\begin{aligned}\dot{q}_1(t) &= f_1\Big(q_1(t), q_2(t), q_3(t), x(t)\Big) \\ \dot{q}_2(t) &= f_1\Big(q_1(t), q_2(t), q_3(t), x(t)\Big) \\ \dot{q}_3(t) &= f_1\Big(q_1(t), q_2(t), q_3(t), x(t)\Big) \ ,\end{aligned} \tag{4.25}$$

with

$$y(t) = g\Big(q_1(t), q_2(t), q_3(t), x(t)\Big) \ , \tag{4.26}$$

or in vector notation,

$$\dot{\mathbf{q}}(t) = \mathbf{f}\Big(\mathbf{q}(t), x(t)\Big) \ , \qquad y(t) = g\Big(\mathbf{q}(t), x(t)\Big) \ . \tag{4.27}$$

Define the equilibrium $\overline{\mathbf{q}}$ again as a state value that the system does not move from when initialized there, and when the input is fixed at $x(t) = \overline{x}$. In the CT case, what this requires is that the rate of change of the state, namely $\dot{\mathbf{q}}(t)$, is zero at the equilibrium, which yields the condition

$$\mathbf{0} = \mathbf{f}(\overline{\mathbf{q}}, \overline{x}) \ . \tag{4.28}$$

For the 3rd-order case, this condition takes the form

$$\begin{aligned}0 &= f_1(\overline{q_1}, \overline{q_2}, \overline{q_3}, \overline{x}) \\ 0 &= f_2(\overline{q_1}, \overline{q_2}, \overline{q_3}, \overline{x}) \\ 0 &= f_3(\overline{q_1}, \overline{q_2}, \overline{q_3}, \overline{x}) \ ,\end{aligned} \tag{4.29}$$

which is again a set of three simultaneous nonlinear equations in three unknowns, with possibly no solution for a specified $\overline{x}$, or one solution, or many.

4.4.2 Linearization

We now examine system behavior in the vicinity of an equilibrium. Consider once more the 3rd-order DT nonlinear system (4.18), and suppose that instead of $x[n] \equiv \overline{x}$, we have $x[n]$ perturbed or deviating from this by a value $\widetilde{x}[n]$, so

$$\widetilde{x}[n] = x[n] - \overline{x} \ . \tag{4.30}$$

The state variables will correspondingly be perturbed from their respective equilibrium values by amounts denoted by

$$\widetilde{q}_i[n] = q_i[n] - \overline{q}_i \tag{4.31}$$

for $i = 1, 2, 3$ (or more generally $i = 1, \cdots, L$), and the output will be perturbed by

$$\widetilde{y}[n] = y[n] - \overline{y} \, . \tag{4.32}$$

Our objective is to find a model that describes the behavior of these various perturbations from equilibrium.

The key to finding a tractable description of the perturbations or deviations from equilibrium is to assume they are small, thereby permitting the use of truncated Taylor series to provide good approximations to the various nonlinear functions. Truncating the Taylor series to first order, i.e., to terms that are linear in the deviations, is referred to as linearization, and produces LTI state-space models in our setting.

To linearize the original DT 3rd-order nonlinear model (4.18), we rewrite the variables appearing in that model in terms of the perturbations, using the quantities defined in (4.30), (4.31), and then expand in Taylor series to first order around the equilibrium values:

$$\begin{aligned} \overline{q}_i + \widetilde{q}_i[n+1] &= f_i\Big(\overline{q}_1 + \widetilde{q}_1[n], \overline{q}_2 + \widetilde{q}_2[n], \overline{q}_3 + \widetilde{q}_3[n], \overline{x} + \widetilde{x}[n]\Big) \qquad \text{for } i = 1, 2, 4 \\ &\approx f_i(\overline{q}_1, \overline{q}_2, \overline{q}_3, \overline{x}) + \frac{\partial f_i}{\partial q_1}\widetilde{q}_1[n] + \frac{\partial f_i}{\partial q_2}\widetilde{q}_2[n] + \frac{\partial f_i}{\partial q_3}\widetilde{q}_3[n] + \frac{\partial f_i}{\partial x}\widetilde{x}[n] \, . \end{aligned} \tag{4.33}$$

All the partial derivatives above are evaluated at the equilibrium values, and are therefore constants, not dependent on the time index n. (Also note that the partial derivatives above are with respect to the continuously variable state and input arguments; there are no "derivatives" taken with respect to n, the discretely varying time index!) The definition of the equilibrium values in (4.24) shows that the term $\overline{q}_i$ on the left of the above set of expressions exactly equals the term $f_i(\overline{q}_1, \overline{q}_2, \overline{q}_3, \overline{x})$ on the right, so what remains is the approximate relation

$$\widetilde{q}_i[n+1] \approx \frac{\partial f_i}{\partial q_1}\widetilde{q}_1[n] + \frac{\partial f_i}{\partial q_2}\widetilde{q}_2[n] + \frac{\partial f_i}{\partial q_3}\widetilde{q}_3[n] + \frac{\partial f_i}{\partial x}\widetilde{x}[n] \tag{4.34}$$

for $i = 1, 2, 3$. Replacing the approximate equality sign ($\approx$) by the equality sign ($=$) in this set of expressions produces what is termed the linearized model at the equilibrium point. This linearized model approximately describes small perturbations away from the equilibrium point.

We may write the linearized model in matrix form:

$$\underbrace{\begin{bmatrix} \widetilde{q}_1[n+1] \\ \widetilde{q}_2[n+1] \\ \widetilde{q}_3[n+1] \end{bmatrix}}_{\widetilde{\mathbf{q}}[n+1]} = \underbrace{\begin{bmatrix} \frac{\partial f_1}{\partial q_1} & \frac{\partial f_1}{\partial q_2} & \frac{\partial f_1}{\partial q_3} \\ \frac{\partial f_2}{\partial q_1} & \frac{\partial f_2}{\partial q_2} & \frac{\partial f_2}{\partial q_3} \\ \frac{\partial f_3}{\partial q_1} & \frac{\partial f_3}{\partial q_2} & \frac{\partial f_3}{\partial q_3} \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} \widetilde{q}_1[n] \\ \widetilde{q}_2[n] \\ \widetilde{q}_3[n] \end{bmatrix}}_{\widetilde{\mathbf{q}}[n]} + \underbrace{\begin{bmatrix} \frac{\partial f_1}{\partial x} \\ \frac{\partial f_2}{\partial x} \\ \frac{\partial f_3}{\partial x} \end{bmatrix}}_{\mathbf{b}} \widetilde{x}[n] \, . \tag{4.35}$$

We have therefore arrived at a standard DT LTI state-space description of the state evolution of our linearized model, with state and input variables that are the respective deviations from equilibrium of the underlying nonlinear model. The corresponding output equation is derived similarly, and takes the form

$$\widetilde{y}[n] = \underbrace{\left[\begin{array}{ccc} \frac{\partial g}{\partial q_1} & \frac{\partial g}{\partial q_2} & \frac{\partial g}{\partial q_3} \end{array}\right]}_{\mathbf{c}^T} \widetilde{\mathbf{q}}[n] + \underbrace{\frac{\partial g}{\partial x}}_{\mathtt{d}} \widetilde{x}[n] \, . \tag{4.36}$$

The matrix of partial derivatives denoted by $\mathbf{A}$ in (4.35) is also called a Jacobian matrix, and denoted in matrix-vector notation by

$$\mathbf{A} = \left[\frac{\partial \mathbf{f}}{\partial \mathbf{q}}\right]_{\overline{\mathbf{q}},\overline{x}} \, . \tag{4.37}$$

The entry in its ith row and jth column is the partial derivative $\partial f_i(\,\cdot\,)/\partial q_j$, evaluated at the equilibrium values of the state and input variables. Similarly,

$$\mathbf{b} = \left[\frac{\partial \mathbf{f}}{\partial x}\right]_{\overline{\mathbf{q}},\overline{x}} \, , \quad \mathbf{c}^T = \left[\frac{\partial g}{\partial \mathbf{q}}\right]_{\overline{\mathbf{q}},\overline{x}} \, , \quad \mathtt{d} = \left[\frac{\partial g}{\partial x}\right]_{\overline{\mathbf{q}},\overline{x}} \, . \tag{4.38}$$

The derivation of linearized state-space models in CT follows exactly the same route, except that the CT equilibrium condition is specified by the condition (4.28) rather than (4.22).

EXAMPLE 4.1 A Hoop-and-Beam System

As an example to illustrate the determination of equilibria and linearizations, we consider in this section a nonlinear state-space model for a particular hoop-and-beam system.

The system in Figure 4.3 comprises a beam pivoted at its midpoint, with a hoop that is constrained to maintain contact with the beam but free to roll along it, without slipping. A torque can be applied to the beam, and acts as the control input. Our eventual objective might be to vary the torque in order to bring the hoop to — and maintain it at — a desired position on the beam. We assume that the only measured output that is available for feedback to the controller is the position of the hoop along the beam.

Natural state variables for such a mechanical system are the position and velocity variables associated with each of its degrees of freedom, namely:

- the position $q_1(t)$ of the point of contact of the hoop relative to the center of the beam;
- the angular position $q_2(t)$ of the beam relative to horizontal;
- the translational velocity $q_3(t) = \dot{q}_1(t)$ of the hoop along the beam;
- the angular velocity $q_4(t) = \dot{q}_2(t)$ of the beam.

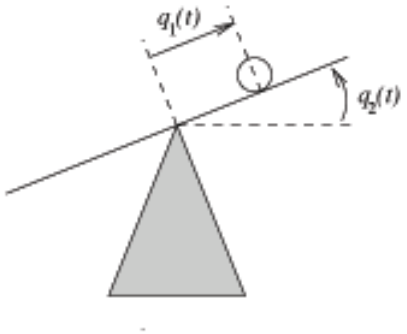

FIGURE 4.3 A hoop rolling on a beam that is free to pivot on its support. The variable $q_1(t)$ is the position of the point of contact of the hoop relative to the center of the beam. The variable $q_2(t)$ is the angle of the beam relative to horizontal.

The measured output is

$$y(t) = q_1(t) \; . \tag{4.39}$$

To specify a state-space model for the system, we express the rate of change of each of these state variables at time t as a function of these variables at t, and as a function of the torque input $x(t)$. We arbitrarily choose the direction of positive torque to be that which would tend to increase the angle $q_2(t)$. The required expressions, which we do not derive here, are most easily obtained using Lagrange's equations of motion, but can also be found by applying the standard and rotational forms of Newton's second law to the system, taking account of the constraint that the hoop rolls without slipping. The resulting nonlinear time-invariant state-space model for the system, with the time argument dropped from the state variables q_i and input x to avoid notational clutter, are:

$$\begin{aligned}
\frac{dq_1}{dt} &= q_3 \\
\frac{dq_2}{dt} &= q_4 \\
\frac{dq_3}{dt} &= \frac{1}{2}\left(q_1 q_4^2 - g\sin(q_2)\right) \\
\frac{dq_4}{dt} &= \frac{mgr\sin(q_2) - mgq_1\cos(q_2) - 2mq_1q_3q_4 + x}{J + mq_1^2} \; .
\end{aligned} \tag{4.40}$$

Here g represents the acceleration due to gravity, m is the mass of the hoop, r is its radius, and J is the moment of inertia of the beam.

Equilibrium values of the model. An equilibrium state of a system is one that

can (ideally) be maintained indefinitely without the action of a control input, or more generally with only constant control action. Our control objective might be to design a feedback control system that regulates the hoop-and-beam system to its equilibrium state, with the beam horizontal and the hoop at the center, i.e., with $q_1(t) \equiv 0$ and $q_2(t) \equiv 0$. The possible zero-control equilibrium positions for any CT system described in state-space form can be found by setting the control input and the state derivatives to 0, and then solving for the state variable values.

For the model above, we see that the only zero-control equilibrium position (with the realistic constraint that $-\frac{\pi}{2} < q_2 < \frac{\pi}{2}$) corresponds to a horizontal beam with the hoop at the center, i.e., $q_1 = q_2 = q_3 = q_4 = 0$. If we allow a constant but nonzero control input, it is straightforward to see from (4.40) that it is possible to have an equilibrium state (i.e., unchanging state variables) with a nonzero q_1, but still with q_2, q_3 and q_4 equal to 0.

Linearization for small perturbations. It is generally quite difficult to elucidate in any detail the global or large-signal behavior of a nonlinear model such as (4.40). However, small deviations of the system around an equilibrium, such as might occur in response to small perturbations of the control input from 0, are quite well modeled by a linearized version of the nonlinear model above. As already described in the previous subsection, a linearized model is obtained by approximating all nonlinear terms using first-order Taylor series expansions around the equilibrium. Linearization of a time-invariant model around an equilibrium point always yields a model that is time invariant, as well as being linear. Thus, even though the original nonlinear model may be difficult to work with, the linearized model around an equilibrium point can be analyzed in great detail, using all the methods available to us for LTI systems. Note also that if the original model is in state-space form, the linearization will be in state-space form too, except that its state variables will be the deviations from equilibrium of the original state variables.

Since the equilibrium of interest to us in the hoop-and-beam example corresponds to all state variables being 0, small deviations from this equilibrium correspond to all state variables being small. The linearization is thus easy to obtain without formal expansion into Taylor series. Specifically, as we discard from the nonlinear model (4.40) all terms of higher order than first in any nonlinear combinations of terms, $\sin(q_2)$ gets replaced by q_2, $\cos(q_2)$ gets replaced by 1, and the terms $q_1 q_4^2$ and $q_1 q_3 q_4$ and q_1^2 are eliminated. The result is the following linearized model in state-space form:

$$\begin{aligned}
\frac{dq_1}{dt} &= q_3 \\
\frac{dq_2}{dt} &= q_4 \\
\frac{dq_3}{dt} &= -\frac{g}{2}q_2 \\
\frac{dq_4}{dt} &= \frac{mg(rq_2 - q_1) + x}{J}
\end{aligned} \tag{4.41}$$

This model, along with the defining equation (4.39) for the output (which is already linear and therefore needs no linearization), can be written in the standard matrix form (4.16) and (4.17) for LTI state-space descriptions, with

$$\begin{aligned}
\mathbf{A} &= \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -g/2 & 0 & 0 \\ -mg/J & mgr/J & 0 & 0 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1/J \end{bmatrix} \\
\mathbf{c}^T &= \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}
\end{aligned} \tag{4.42}$$

The LTI model is much more tractable than the original nonlinear time-invariant model, and consequently controllers can be designed more systematically and confidently. If the resulting controllers, when applied to the system, manage to ensure that deviations from equilibrium remain small, then our use of the linearized model for design will have been justified.

4.5 STATE-SPACE MODELS FROM INPUT–OUTPUT MODELS

State-space representations can be very naturally and directly generated during the modeling process in a variety of settings, as the examples in Sections 4.2.1 and 4.2.2 suggest. Other — and perhaps more familiar — descriptions can then be derived from them; again, these previous examples showed how input–output descriptions could be obtained from state-space descriptions.

It is also possible to proceed in the reverse direction, constructing state-space descriptions from impulse responses or transfer functions or input–output difference equations, for instance. This is often worthwhile as a prelude to simulation, or filter implementation, or in control design, or simply in order to understand the initial description from another point of view. The following two examples illustrate this reverse process, of synthesizing state-space descriptions from input–output descriptions.

4.5.1 Determining a state-space model from an impulse response or transfer function

Consider the impulse response $h[n]$ of a causal DT LTI system. Causality requires of course that $h[n] = 0$ for $n < 0$. The output $y[n]$ can be related to past and

present inputs $x[k]$, $k \leq n$, through the convolution sum

$$y[n] = \sum_{k=-\infty}^{n} h[n-k]\, x[k] \tag{4.43}$$

$$= \Big(\sum_{k=-\infty}^{n-1} h[n-k]\, x[k] \Big) + h[0]x[n] \,. \tag{4.44}$$

The first term above, namely

$$q[n] = \sum_{k=-\infty}^{n-1} h[n-k]\, x[k] \,, \tag{4.45}$$

represents the effect of the past on the present, at time n, and would therefore seem to have some relation to the notion of a state variable. Updating $q[n]$ to the next time step, we obtain

$$q[n+1] = \sum_{k=-\infty}^{n} h[n+1-k]\, x[k] \,. \tag{4.46}$$

In general, if the impulse response has no special form, the successive values of $q[n]$ have to be recomputed from (4.46) for each n. When we move from n to $n+1$, none of the past inputs $x[k]$ for $k \leq n$, can be discarded, because all of the past will again be needed to compute $q[n+1]$. In other words, the memory of the system is infinite.

However, consider the class of systems for which $h[n]$ has the essentially exponential form

$$h[n] = \beta\, \lambda^{n-1} u[n-1] + \mathtt{d}\, \delta[n] \,, \tag{4.47}$$

where β, λ and $\mathtt{d}$ are constants. The corresponding transfer function is

$$H(z) = \frac{\beta}{z-\lambda} + \mathtt{d} \tag{4.48}$$

(with ROC $|z| > |\lambda|$). What is important about this impulse response is that a time-shifted version of it is simply related to a scaled version of it, because of its DT-exponential form. For this case,

$$q[n] = \beta \sum_{k=-\infty}^{n-1} \lambda^{n-1-k} x[k] \tag{4.49}$$

and

$$q[n+1] = \beta \sum_{k=-\infty}^{n} \lambda^{n-k} x[k] \tag{4.50}$$

$$= \lambda \Big(\beta \sum_{k=-\infty}^{n-1} \lambda^{n-1-k} x[k] \Big) + \beta x[n]$$

$$= \lambda q[n] + \beta x[n] \,. \tag{4.51}$$

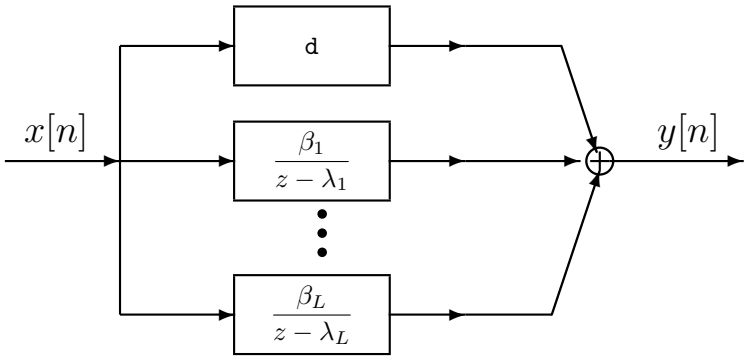

FIGURE 4.4 Decomposition of rational transfer function with distinct poles.

Gathering (4.44) and (4.49) with (4.51) results in a pair of equations that together constitute a state-space description for this system:

$$q[n+1] = \lambda q[n] + \beta x[n] \tag{4.52}$$

$$y[n] = q[n] + \mathtt{d}x[n] \; . \tag{4.53}$$

Let us consider next a similar but higher order system with impulse response:

$$h[n] = (\,\beta_1\lambda_1^{n-1} + \beta_2\lambda_2^{n-1} + \cdots + \beta_L\lambda_L^{n-1}\,)u[n-1] + \mathtt{d}\,\delta[n] \tag{4.54}$$

with the β_i and $\mathtt{d}$ being constants. The corresponding transfer function is

$$H(z) = \Big(\sum_{i=1}^{L} \frac{\beta_i}{z-\lambda_i}\Big) + \mathtt{d} \; . \tag{4.55}$$

By using a partial fraction expansion, the transfer function $H(z)$ of any causal LTI DT system with a rational transfer function can be written in this form, with appropriate choices of the β_i, λ_i, $\mathtt{d}$ and L, provided $H(z)$ has non-repeated — i.e., distinct — poles. Note that although we only treat rational transfer functions $H(z)$ whose numerator and denominator polynomials have real coefficients, the poles of $H(z)$ may include some complex λ_i (and associated β_i), but in each such case its complex conjugate λ_i^* will also be a pole (with associated weighting factor β_i^*), and the sum

$$\beta_i(\lambda_i)^n + \beta_i^*(\lambda_i^*)^n \tag{4.56}$$

will be real.

The block diagram in Figure 4.5.1 shows that this system can be considered as being obtained through the parallel interconnection of subsystems corresponding to the simpler case of (4.47). Motivated by this structure and the treatment of the first-order example, we define a state variable for each of the L subsystems:

$$q_i[n] = \beta_i \sum_{-\infty}^{n-1} \lambda_i^{n-1-k} x[k] \; , \quad i = 1, 2, \ldots, L \; . \tag{4.57}$$

With this, we obtain the following state-evolution equations for the subsystems:

$$q_i[n+1] = \lambda_i q_i[n] + \beta_i x[n] \ , \quad i = 1, 2, \ldots, L \ . \tag{4.58}$$

Also, combining (4.45), (4.53) and (4.54) with the definitions in (4.57), we obtain the output equation

$$y[n] = q_1[n] + q_2[n] + \cdots + q_L[n] + \mathtt{d}\, x[n] \ . \tag{4.59}$$

Equations (4.58) and (4.59) together comprise an Lth-order state-space description of the given system. We can write this state-space description in our standard matrix form (4.13) and (4.14), with

$$\mathbf{A} = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \lambda_L \end{pmatrix} \ , \quad \mathbf{b} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_L \end{pmatrix} \tag{4.60}$$

$$\mathbf{c}^T = \begin{pmatrix} 1 & 1 & \cdots & \cdots & \cdots & 1 \end{pmatrix} \ . \tag{4.61}$$

The diagonal form of $\mathbf{A}$ in (4.60) reflects the fact that the state evolution equations in this example are decoupled, with each state variable being updated independently according to (4.58). We shall see later how a general description of the form (4.13), (4.14), with a distinct-eigenvalue condition that we shall impose, can actually be transformed to a completely equivalent description in which the new $\mathbf{A}$ matrix is diagonal, as in (4.60). (Note, however, that when there are complex eigenvalues, this diagonal state-space representation will have complex entries.)

4.5.2 Determining a state-space model from an input–output difference equation

Let us examine some ways of representing the following input-output difference equation in state-space form:

$$y[n] + a_1 y[n-1] + a_2 y[n-2] = b_1 x[n-1] + b_2 x[n-2] \ . \tag{4.62}$$

One approach, building on the development in the preceding subsection, is to perform a partial fraction expansion of the 2-pole transfer function associated with this system, and thereby obtain a 2nd-order realization in diagonal form. (If the real coefficients a_1 and a_2 are such that the roots of $z^2 + a_1 z + a_2$ are not real but form a complex conjugate pair, then this diagonal 2nd-order realization will have complex entries.)

For a more direct attempt (and to guarantee a real-valued rather than complex-valued state-space model), consider using as state vector the quantity

$$\mathbf{q}[n] = \begin{pmatrix} y[n-1] \\ y[n-2] \\ x[n-1] \\ x[n-2] \end{pmatrix} \ . \tag{4.63}$$

The corresponding 4th-order state-space model would take the form

$$\begin{aligned}\mathbf{q}[n+1] = \begin{pmatrix} y[n] \\ y[n-1] \\ x[n] \\ x[n-1] \end{pmatrix} &= \begin{pmatrix} -a_1 & -a_2 & b_1 & b_2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} y[n-1] \\ y[n-2] \\ x[n-1] \\ x[n-2] \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} x[n] \\ y[n] &= \begin{pmatrix} -a_1 & -a_2 & b_1 & b_2 \end{pmatrix} \begin{pmatrix} y[n-1] \\ y[n-2] \\ x[n-1] \\ x[n-2] \end{pmatrix} \end{aligned} \tag{4.64}$$

If we are somewhat more careful about our choice of state variables, it is possible to get more economical models. For a 3rd-order model, suppose we pick as state vector

$$\mathbf{q}[n] = \begin{pmatrix} y[n] \\ y[n-1] \\ x[n-1] \end{pmatrix} . \tag{4.65}$$

The corresponding 3rd-order state-space model takes the form

$$\begin{aligned}\mathbf{q}[n+1] = \begin{pmatrix} y[n+1] \\ y[n] \\ x[n] \end{pmatrix} &= \begin{pmatrix} -a_1 & -a_2 & b_2 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} y[n] \\ y[n-1] \\ x[n-1] \end{pmatrix} + \begin{pmatrix} b_1 \\ 0 \\ 1 \end{pmatrix} x[n] \\ y[n] &= \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} y[n] \\ y[n-1] \\ x[n-1] \end{pmatrix} \end{aligned} \tag{4.66}$$

A still more subtle choice of state variables yields a 2nd-order state-space model by picking

$$\mathbf{q}[n] = \begin{pmatrix} y[n] \\ -a_2 y[n-1] + b_2 x[n-1] \end{pmatrix} . \tag{4.67}$$

The corresponding 2nd-order state-space model takes the form

$$\begin{aligned}\begin{pmatrix} y[n+1] \\ -a_2 y[n] + b_2 x[n] \end{pmatrix} &= \begin{pmatrix} -a_1 & 1 \\ -a_2 & 0 \end{pmatrix} \begin{pmatrix} y[n] \\ -a_2 y[n-1] + b_2 x[n-1] \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} x[n] \\ y[n] &= \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} y[n] \\ -a_2 y[n-1] + b_2 x[n-1] \end{pmatrix} \end{aligned} \tag{4.68}$$

It turns out to be impossible in general to get a state-space description of order lower than 2 in this case. This should not be surprising, in view of the fact that (4.63) is a 2nd-order difference equation, which we know requires two initial conditions in order to solve forwards in time. Notice how, in each of the above cases, we have incorporated the information contained in the original difference equation (4.63) that we started with.

CHAPTER 5

Properties of LTI State-Space Models

5.1 INTRODUCTION

In Chapter 4 we introduced state-space models for dynamical systems. In this chapter we study the structure and solutions of LTI state-space models. Throughout the discussion we restrict ourselves to the single-input, single-output Lth-order CT LTI state-space model

$$\dot{\mathbf{q}}(t) = \mathbf{A}\mathbf{q}(t) + \mathbf{b}x(t) \tag{5.1}$$

$$y(t) = \mathbf{c}^T\mathbf{q}(t) + \mathtt{d}x(t)\ , \tag{5.2}$$

or the DT LTI state-space model

$$\mathbf{q}[n+1] = \mathbf{A}\mathbf{q}[n] + \mathbf{b}x[n] \tag{5.3}$$

$$y[n] = \mathbf{c}^T\mathbf{q}[n] + \mathtt{d}x[n]\ . \tag{5.4}$$

Equation (5.1) constitutes a representation of CT LTI system dynamics in the form of a set of coupled, first-order, linear, constant-coefficient differential equations for the L variables in $\mathbf{q}(t)$, driven by the input $x(t)$. Equation (5.3) gives a similar difference-equation representation of DT LTI system dynamics.

The basic approach to analyzing LTI state-space models parallels what you should already be familiar with from solving linear constant-coefficient differential or difference equations (of any order) in one variable. Specifically, we first consider the zero-input response to nonzero initial conditions at some starting time, and then augment that with the response due to the nonzero input when the initial conditions are zero. Understanding the full solution from the starting time onwards will give us insight into system stability, and into how the internal behavior relates to the input-output characteristics of the system.

5.2 THE ZERO-INPUT RESPONSE AND MODAL REPRESENTATION

We take our starting time to be 0, without loss of generality (since we are dealing with time-invariant models). Consider the response of the undriven system corresponding to (5.1), i.e., the response with $x(t) \equiv 0$ for $t \geq 0$, but with some nonzero initial condition $\mathbf{q}(0)$. This is the zero-input-response (ZIR) of the system (5.1),

and is a solution of the undriven (or unforced or homogeneous) system

$$\dot{\mathbf{q}}(t) = \mathbf{A}\mathbf{q}(t) \; . \tag{5.5}$$

It is natural when analyzing an undriven LTI system to look for a solution in exponential form (essentially because exponentials have the unique property that shifting them is equivalent to scaling them, and undriven LTI systems are characterized by invariance to shifting and scaling of solutions). We accordingly look for a nonzero solution of the form

$$\mathbf{q}(t) = \mathbf{v}e^{\lambda t} \; , \quad \mathbf{v} \neq \mathbf{0} \; , \tag{5.6}$$

where each state variable is a scalar multiple of the same exponential $e^{\lambda t}$, with these scalar multiples assembled into the vector $\mathbf{v}$. (The boldface $\mathbf{0}$ at the end of the preceding equation denotes an L-component column vector whose entries are all 0 — we shall use $\mathbf{0}$ for any vectors or matrices whose entries are all 0, with the correct dimensions being apparent from the context. Writing $\mathbf{v} \neq \mathbf{0}$ signifies that at least one component of $\mathbf{v}$ is nonzero.)

Substituting (5.6) into (5.5) results in the equation

$$\lambda \mathbf{v}e^{\lambda t} = \mathbf{A}\mathbf{v}e^{\lambda t} \; , \tag{5.7}$$

from which we can conclude that the vector $\mathbf{v}$ and scalar λ must satisfy

$$\lambda \mathbf{v} = \mathbf{A}\mathbf{v} \quad \text{or equivalently} \quad (\lambda \mathbf{I} - \mathbf{A})\mathbf{v} = \mathbf{0} \; , \quad \mathbf{v} \neq \mathbf{0} \; , \tag{5.8}$$

where $\mathbf{I}$ denotes the identity matrix, in this case of dimension $L \times L$. The above equation has a nonzero solution $\mathbf{v}$ if and only if the coefficient matrix $(\lambda \mathbf{I} - \mathbf{A})$ is not invertible, i.e., if and only if its determinant is 0:

$$\det(\lambda \mathbf{I} - \mathbf{A}) = 0 \; . \tag{5.9}$$

For an Lth-order system, it turns out that the above determinant is a monic polynomial of degree L, called the characteristic polynomial of the system or of the matrix $\mathbf{A}$:

$$\det(\lambda I - \mathbf{A}) = a(\lambda) = \lambda^L + a_{L-1}\lambda^{L-1} + \cdots + a_0 \tag{5.10}$$

(The word "monic" simply means that the coefficient of the highest-degree term is 1.) It follows that (5.6) is a nonzero solution of (5.5) if and only if λ is one of the L roots $\{\lambda_i\}_{i=1}^{L}$ of the characteristic polynomial. These roots are referred to as characteristic roots of the system, and as eigenvalues of the matrix $\mathbf{A}$.

The vector $\mathbf{v}$ in (5.6) is correspondingly a nonzero solution $\mathbf{v}_i$ of the system of equations

$$(\lambda_i \mathbf{I} - \mathbf{A})\mathbf{v}_i = \mathbf{0} \; , \quad \mathbf{v}_i \neq \mathbf{0} \; , \tag{5.11}$$

and is termed the characteristic vector or eigenvector associated with λ_i. Note from (5.11) that multiplying any eigenvector by a nonzero scalar again yields an eigenvector, so eigenvectors are only defined up to a nonzero scaling. Any convenient scaling or normalization can be used.

In summary, the undriven system has a solution of the assumed exponential form in (5.6) if and only if λ equals some characteristic value or eigenvalue of $\mathbf{A}$, and the nonzero vector $\mathbf{v}$ is an associated characteristic vector or eigenvector.

We shall only be dealing with state-space models for which all the signals and the coefficient matrices $\mathbf{A}$, $\mathbf{b}$, $\mathbf{c}^T$ and $\mathtt{d}$ are real-valued (though we may subsequently transform these models into the diagonal forms seen in the previous chapter, which may then have complex entries, but occurring in very structured ways). The coefficients a_i defining the characteristic polynomial $a(\lambda)$ in (5.10) are therefore real, and thus the complex roots of this polynomial occur in conjugate pairs. Also, it is straightforward to show that if $\mathbf{v}_i$ is an eigenvector associated with a complex eigenvalue λ_i, then $\mathbf{v}_i^*$ —i.e., the vector whose entries are the complex conjugates of the corresponding entries of $\mathbf{v}_i$ — is an eigenvector associated with λ_i^*, the complex conjugate of λ_i.

We refer to a nonzero solution of the form (5.6) for $\lambda = \lambda_i$ and $\mathbf{v} = \mathbf{v}_i$ as the ith mode of the system (5.1) or (5.5); the associated λ_i is termed the ith modal frequency or characteristic frequency or natural frequency of the system, and $\mathbf{v}_i$ is termed the ith mode shape. Note that if

$$\mathbf{q}(t) = \mathbf{v}_i e^{\lambda_i t} \tag{5.12}$$

then the corresponding initial condition must have been $\mathbf{q}(0) = \mathbf{v}_i$. It can be shown (though we don't do so here) that the system (5.5) — and similarly the system (5.1) — can only have one solution for a given initial condition, so it follows that for the initial condition $\mathbf{q}(0) = \mathbf{v}_i$, only the ith mode will be excited.

It can also be shown that eigenvectors associated with distinct eigenvalues are linearly independent, i.e., none of them can be written as a weighted linear combination of the remaining ones. For simplicity, we shall restrict ourselves throughout to the case where all L eigenvalues of $\mathbf{A}$ are distinct, which will guarantee that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_L$ form an independent set. (In some cases in which $\mathbf{A}$ has repeated eigenvalues, it is possible to find a full set of L independent eigenvectors, but this is not generally true.) We shall repeatedly use the fact that any vector in an L-dimensional space, such as our state vector $\mathbf{q}(t)$ at any specified time $t = t_0$, can be written as a unique linear combination of any L independent vectors in that space, such as our L eigenvectors.

5.2.1 Modal representation of the ZIR

Because (5.5) is linear, a weighted linear combination of modal solutions of the form (5.12), one for each eigenvalue, will also satisfy (5.5). Consequently a more general solution for the zero-input response with distinct eigenvalues is

$$\mathbf{q}(t) = \sum_{i=1}^{L} \alpha_i \mathbf{v}_i e^{\lambda_i t} \tag{5.13}$$

The expression in (5.13) can easily be verified to be a solution of (5.5) for arbitrary weights α_i, with initial condition

$$\mathbf{q}(0) = \sum_{i=1}^{L} \alpha_i \mathbf{v}_i \ . \tag{5.14}$$

Since the L eigenvectors $\mathbf{v}_i$ are independent under our assumption of L distinct eigenvalues, the right side of (5.14) can be made equal to any desired $\mathbf{q}(0)$ by proper choice of the coefficients α_i, and these coefficients are unique. Hence specifying the initial condition of the undriven system (5.5) specifies the α_i via (5.14), and thus specifies the full response of (5.5) via (5.13). In other words, (5.13) is actually a general expression for the ZIR of (5.1) — under our assumption of distinct eigenvalues. We refer to the expression on the right side of (5.13) as the modal decomposition of the ZIR.

The contribution to the modal decomposition from a conjugate pair of eigenvalues $\lambda_i = \sigma_i + j\omega_i$ and $\lambda_i^* = \sigma_i - j\omega_i$, with associated complex conjugate eigenvectors $\mathbf{v}_i = \mathbf{u}_i + j\mathbf{w}_i$ and $\mathbf{v}_i = \mathbf{u}_i - j\mathbf{w}_i$ respectively, will be a real term of the form

$$\alpha_i \mathbf{v}_i e^{\lambda_i t} + \alpha_i^* \mathbf{v}_i^* e^{\lambda_i^* t} \ . \tag{5.15}$$

With a little algebra, the real expression in (5.15) can be reduced to the form

$$\alpha_i \mathbf{v}_i e^{\lambda_i t} + \alpha_i^* \mathbf{v}_i^* e^{\lambda_i^* t} = K_i e^{\sigma_i t}[\mathbf{u}_i \cos(\omega_i t + \theta_i) - \mathbf{w}_i \sin(\omega_i t + \theta_i)] \tag{5.16}$$

for some constants K_i and θ_i that are determined by the initial conditions in the process of matching the two sides of (5.14). The above component of the modal solution therefore lies in the plane spanned by the real and imaginary parts, $\mathbf{u}_i$ and $\mathbf{w}_i$ respectively, of the eigenvector $\mathbf{v}_i$. The associated motion of the component of state trajectory in this plane involves an exponential spiral, with growth or decay of the spiral determined by whether $\sigma_i = \mathrm{Re}\{\lambda_i\}$ is positive or negative respectively (corresponding to the eigenvalue λ_i — and its conjugate λ_i^* — lying in the open right- or left-half-plane respectively). If $\sigma_i = 0$, i.e., if the conjugate pair of eigenvalues lies on the imaginary axis, then the spiral degenerates to a closed loop. The rate of rotation of the spiral is determined by $\omega_i = \mathrm{Im}\{\lambda_i\}$.

A similar development can be carried out in the DT case for the ZIR of (5.3). In that case (5.6) is replaced by a solution of the form

$$\mathbf{q}[n] = \mathbf{v}\lambda^n \tag{5.17}$$

and we find that when $\mathbf{A}$ has L distinct eigenvalues, the modal decomposition of the general ZIR solution takes the form

$$\mathbf{q}[n] = \sum_{i=1}^{L} \alpha_i \mathbf{v}_i \lambda_i^n \ . \tag{5.18}$$

5.2.2 Asymptotic stability

The stability of an LTI system is directly related to the behavior of the modes, and more specifically to the values of the λ_i, the roots of the characteristic polynomial. An LTI state-space system is termed asymptotically stable or internally stable if its ZIR decays to zero for all initial conditions. We see from (5.13) that the condition $\mathrm{Re}\{\lambda_i\} < 0$ for all $1 \leq i \leq L$ is necessary and sufficient for asymptotic stability in the CT case. Thus, all eigenvalues of $\mathbf{A}$ in (5.1) — or natural frequencies of (5.1) — must be in the open left-half-plane.

In the DT case, (5.18) shows that a necessary and sufficient condition for asymptotic stability is $|\lambda_i| < 1$ for all $1 \leq i \leq L$, i.e., all eigenvalues of $\mathbf{A}$ in (5.3) — or natural frequencies of (5.3) — must be strictly within the unit circle.

We used the modal decompositions (5.13) and (5.18) to make these claims regarding stability conditions, but these modal decompositions were obtained under the assumption of distinct eigenvalues. Nevertheless, it can be shown that the stability conditions in the general case are identical to those above.

5.3 COORDINATE TRANSFORMATIONS

We have so far only described the zero-input response of LTI state-space systems. Before presenting the general response, including the effects of inputs, it will be helpful to understand how a given state-space representation can be transformed to an equivalent representation that might be simpler to analyze. Our development is carried out for the CT case, but an entirely similar development can be done for DT.

It is often useful to examine the behavior of a state-space system by rewriting the original description in terms of a transformed set of variables. A particularly important case involves the transformation of the state vector $\mathbf{q}(t)$ to a new state vector $\mathbf{r}(t)$ that decomposes the behavior of the system into its components along each of the eigenvectors $\mathbf{v}_i$:

$$\mathbf{q}(t) = \sum_{i=1}^{L} \mathbf{v}_i r_i(t) = \mathbf{V}\mathbf{r}(t) \,, \tag{5.19}$$

where the ith column of the $L \times L$ matrix $\mathbf{V}$ is the ith eigenvector, $\mathbf{v}_i$:

$$\mathbf{V} = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_L \end{pmatrix} \,. \tag{5.20}$$

We refer to $\mathbf{V}$ as the modal matrix. Under our assumption of distinct eigenvalues, the eigenvectors are independent, which guarantees that $\mathbf{V}$ is invertible, so

$$\mathbf{r}(t) = \mathbf{V}^{-1}\mathbf{q}(t) \,. \tag{5.21}$$

The transformation from the original system description involving $\mathbf{q}(t)$ to one written in terms of $\mathbf{r}(t)$ is called a modal transformation, and the new state variables $r_i(t)$ defined through (5.19) are termed modal variables or modal coordinates.

More generally, a coordinate transformation corresponds to choosing a new state vector $\mathbf{z}(t)$ related to the original state vector $\mathbf{q}(t)$ through the relationship

$$\mathbf{q}(t) = \mathbf{M}\mathbf{z}(t) \tag{5.22}$$

where the constant matrix $\mathbf{M}$ is chosen to be invertible. (The ith column of $\mathbf{M}$ is the representation of the ith unit vector of the new $\mathbf{z}$ coordinates in terms of the old $\mathbf{q}$ coordinates.) Substituting (5.22) in (5.1) and (5.2), and solving for $\dot{\mathbf{z}}(t)$, we obtain

$$\dot{\mathbf{z}}(t) = (\mathbf{M}^{-1}\mathbf{A}\mathbf{M})\mathbf{z}(t) + (\mathbf{M}^{-1}\mathbf{b})x(t) \tag{5.23}$$

$$y(t) = (\mathbf{c}^T\mathbf{M})\mathbf{z}(t) + \mathtt{d}x(t)\ . \tag{5.24}$$

Equations (5.23) and (5.24) are still in state-space form, but with state vector $\mathbf{z}(t)$, and with modified coefficient matrices. This model is entirely equivalent to the original one, since (5.22) permits $\mathbf{q}(t)$ to be obtained from $\mathbf{z}(t)$, and the invertibility of $\mathbf{M}$ permits $\mathbf{z}(t)$ to be obtained from $\mathbf{q}(t)$. It is straightforward to verify that the eigenvalues of $\mathbf{A}$ are identical to those of $\mathbf{M}^{-1}\mathbf{A}\mathbf{M}$, and consequently that the natural frequencies of the transformed system are the same as those of the original system; only the eigenvectors change, with $\mathbf{v}_i$ transforming to $\mathbf{M}^{-1}\mathbf{v}_i$.

We refer to the transformation (5.22) as a similarity transformation, and say that the model (5.23), (5.24) is similar to the model (5.1), (5.2).

Note that the input $x(t)$ and output $y(t)$ are unaffected by this state transformation. For a given input, and assuming an initial state $\mathbf{z}(0)$ in the transformed system that is related to $\mathbf{q}(0)$ via (5.22), we obtain the same output as we would have from (5.1), (5.2). In particular, the transfer function from input to output is unaffected by a similarity transformation.

Similarity transformations can be defined in exactly the same way for the DT case in (5.3), (5.4).

5.3.1 Transformation to Modal Coordinates

What makes the modal similarity transformation (5.19) interesting and useful is the fact that the state evolution matrix $\mathbf{A}$ transforms to a diagonal matrix $\mathbf{\Lambda}$:

$$\mathbf{V}^{-1}\mathbf{A}\mathbf{V} = \text{diagonal}\ \{\lambda_1, \cdots, \lambda_L\} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_L \end{bmatrix} = \mathbf{\Lambda}\ . \tag{5.25}$$

The easiest way to verify this is to establish the equivalent fact that $\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}$, which in turn is simply the equation (5.11), written for $i = 1, \cdots, L$ and stacked up in matrix form.

The diagonal form of $\mathbf{\Lambda}$ causes the corresponding state equations in the new coordinate system to be decoupled. Under this modal transformation, the undriven

system (5.5) is transformed into L decoupled, scalar equations:

$$\dot{r}_i(t) = \lambda_i r_i(t) \quad \text{for i} = 1, 2, \ldots, \text{L} \; . \tag{5.26}$$

Each of these is easy to solve:

$$r_i(t) = e^{\lambda_i t} r_i(0) \; . \tag{5.27}$$

Combining this with (5.19) yields (5.13) again, with $\alpha_i = r_i(0)$.

5.4 THE COMPLETE RESPONSE

Applying the modal transformation (5.19) to the full driven system (5.1), (5.2), we see that the transformed system (5.23), (5.24) takes the following form, which is decoupled into L parallel scalar subsystems:

$$\dot{r}_i(t) = \lambda_i r_i(t) + \beta_i x(t) \; , \quad i = 1, 2, \ldots, L \tag{5.28}$$

$$y(t) = \xi_1 r_1(t) + \cdots + \xi_L r_L(t) + \mathtt{d} x(t) \; , \tag{5.29}$$

where the β_i and ξ_i are defined via

$$\mathbf{V}^{-1}\mathbf{b} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_L \end{bmatrix} = \beta \; , \quad \mathbf{c}^T\mathbf{V} = \begin{bmatrix} \xi_1 & \xi_2 & \cdots & \xi_L \end{bmatrix} = \xi \; . \tag{5.30}$$

The second equation in (5.30) shows that

$$\xi_i = \mathbf{c}^T \mathbf{v}_i \; . \tag{5.31}$$

To find an interpretation of the β_i, note that the first equation in (5.30) can be rewritten as $\mathbf{b} = \mathbf{V}\beta$. Writing out the product $\mathbf{V}\beta$ in detail, we find

$$\mathbf{b} = \mathbf{v}_1\beta_1 + \mathbf{v}_2\beta_2 + \cdots + \mathbf{v}_L\beta_L \; . \tag{5.32}$$

In other words, the coefficients β_i are the coefficients needed to express the input vector $\mathbf{b}$ as a linear combination of the eigenvectors $\mathbf{v}_i$.

Each of the scalar equations in (5.28) is a first-order LTI differential equation, and can be solved explicitly for $t \geq 0$, obtaining

$$r_i(t) = \underbrace{e^{\lambda_i t} r_i(0)}_{\text{ZIR}} + \underbrace{\int_0^t e^{\lambda_i (t-\tau)} \beta_i x(\tau) \, d\tau}_{\text{ZSR}} \; , \quad t \geq 0 \; , \quad 1 \leq i \leq L \; . \tag{5.33}$$

Expressed in this form, we easily recognize the separate contributions to the solution made by: (i) the response due to the initial state (the zero-input response or ZIR); and (ii) the response due to the system input (the zero-state response or ZSR). From the preceding expression and (5.29), one can obtain an expression for $y(t)$.

Introducing the natural "matrix exponential" notation

$$e^{\mathbf{\Lambda}t} = \text{diagonal}\,\{e^{\lambda_1 t}, \cdots, e^{\lambda_L t}\} = \begin{bmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_L t} \end{bmatrix} \tag{5.34}$$

allows us to combine the L equations in (5.33) into the following single matrix equation:

$$\mathbf{r}(t) = e^{\mathbf{\Lambda}t}\mathbf{r}(0) + \int_0^t e^{\mathbf{\Lambda}(t-\tau)}\beta x(\tau)\,d\tau\ , \quad t \geq 0 \tag{5.35}$$

(where the integral of a vector is interpreted as the component-wise integral). Combining this equation with the expression (5.19) that relates $\mathbf{r}(t)$ to $\mathbf{q}(t)$, we finally obtain

$$\mathbf{q}(t) = \left(\mathbf{V}e^{\mathbf{\Lambda}t}\mathbf{V}^{-1}\right)\mathbf{q}(0) + \int_0^t \left(\mathbf{V}e^{\mathbf{\Lambda}(t-\tau)}\mathbf{V}^{-1}\right)\mathbf{b}x(\tau)\,d\tau \tag{5.36}$$

$$= e^{\mathbf{A}t}\mathbf{q}(0) + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{b}x(\tau)\,d\tau\ , \quad t \geq 0\ , \tag{5.37}$$

where, by analogy with (5.25), we have defined the matrix exponential

$$e^{\mathbf{A}t} = \mathbf{V}e^{\mathbf{\Lambda}t}\mathbf{V}^{-1}\ . \tag{5.38}$$

Equation (5.37) gives us, in compact matrix notation, the general solution of the CT LTI system (5.1).

An entirely parallel development can be carried out for the DT LTI case. The corresponding expression for the solution of (5.3) is

$$\mathbf{q}[n] = \left(\mathbf{V}\mathbf{\Lambda}^n\mathbf{V}^{-1}\right)\mathbf{q}[0] + \sum_{k=0}^{n-1}\left(\mathbf{V}\mathbf{\Lambda}^{n-k-1}\mathbf{V}^{-1}\right)\mathbf{b}x[k] \tag{5.39}$$

$$= \mathbf{A}^n\mathbf{q}[0] + \sum_{k=0}^{n-1}\mathbf{A}^{n-k-1}\mathbf{b}x[k]\ , \quad n \geq 0\ . \tag{5.40}$$

Equation (5.40) is exactly the expression one would get by simply iterating (5.3) forward one step at a time, to get $\mathbf{q}[n]$ from $\mathbf{q}[0]$. However, we get additional insight from writing the expression in the modally decomposed form (5.39), because it brings out the role of the eigenvalues of $\mathbf{A}$, i.e., the natural frequencies of the DT system, in determining the behavior of the system, and in particular its stability properties.

5.5 TRANSFER FUNCTION, HIDDEN MODES, REACHABILITY, OBSERVABILITY

The transfer function $H(s)$ of the transformed model (5.28), (5.29) describes the zero-state input-output relationship in the Laplace transform domain, and is straightforward to find because the equations are totally decoupled. Taking the Laplace

transforms of those equations, with zero initial conditions in (5.28), results in

$$R_i(s) = \frac{\beta_i}{s-\lambda_i} X(s) \tag{5.41}$$

$$Y(s) = \Big(\sum_1^L \xi_i R_i(s)\Big) + \mathtt{d}X(s)\ . \tag{5.42}$$

Since $Y(s) = H(s)X(s)$, we obtain

$$H(s) = \Big(\sum_1^L \frac{\xi_i \beta_i}{s-\lambda_i}\Big) + \mathtt{d} \tag{5.43}$$

which can be rewritten in matrix notation as

$$H(s) = \xi^T (s\mathbf{I} - \mathbf{\Lambda})^{-1}\beta + \mathtt{d}\ . \tag{5.44}$$

This is also the transfer function of the original model in (5.1), (5.2), as similarity transformations do not change transfer functions. An alternative expression for the transfer function of (5.1), (5.2) follows from examination of the Laplace transformed version of (5.1), (5.2). We omit the details, but the resulting expression is

$$H(s) = \mathbf{c}^T (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} + \mathtt{d} \tag{5.45}$$

We see from (5.43) that $H(s)$ will have L poles in general. However, if $\beta_j = 0$ for some j — i.e., if $\mathbf{b}$ can be expressed as a linear combination of the eigenvectors other than $\mathbf{v}_j$, see (5.32) — then λ_j fails to appear as a pole of the transfer function, even though it is still a natural frequency of the system and appears in the ZIR for almost all initial conditions. The underlying cause for this hidden mode — an internal mode that is hidden from the input/output transfer function — is evident from (5.28) or (5.41): with $\beta_j = 0$, the input fails to excite the jth mode. We say that the mode associated with λ_j is an unreachable mode in this case. In contrast, if $\beta_k \neq 0$, we refer to the kth mode as reachable. (The term controllable is also used for reachable — although strictly speaking there is a slight difference in the definitions of the two concepts in the DT case.)

If all L modes of the system are reachable, then the system itself is termed reachable, otherwise it is called unreachable. In a reachable system, the input can fully excite the state (and in fact can transfer the state vector from any specified initial condition to any desired target state in finite time). In an unreachable system, this is not possible. The notion of reachability arises in several places in systems and control theory.

The dual situation happens when $\xi_j = 0$ for some j — i.e., if $\mathbf{c}^T\mathbf{v}_j = 0$, see (5.31). In this case again, (5.43) shows that λ_j fails to appear as a pole of the transfer function, even though it is still a natural frequency of the system. Once again, we have a hidden mode. This time, the cause is evident in (5.29) or (5.42): with $\xi_j = 0$, the jth mode fails to appear at the output, even when it is present in the

state response. We say that the mode associated with λ_j is unobservable in this case. In contrast, if $\xi_k \neq 0$, then we call the kth mode observable.

If all L modes of the system are observable, the system itself is termed observable, otherwise it is called unobservable. In an observable system, the behavior of the state vector can be unambiguously inferred from measurements of the input and output over some interval of time, whereas this is not possible for an unobservable system. The concept of observability also arises repeatedly in systems and control theory.

Hidden modes can cause difficulty, especially if they are unstable. However, if all we are concerned about is representing a transfer function, or equivalently the input–output relation of an LTI system, then hidden modes may be of no significance. We can obtain a reduced-order state-space model that has the same transfer function by simply discarding all the equations in (5.28) that correspond to unreachable or unobservable modes, and discarding the corresponding terms in (5.29).

The converse also turns out to be true: if a state-space model is reachable and observable, then there is no lower order state-space system that has the same transfer function; in other words, a state-space model that is reachable and observable is minimal.

Again, an entirely parallel development can be carried out for the DT case, as the next example illustrates.

EXAMPLE 5.1 A discrete-time non-minimal system

In this example we consider the DT system represented by the state equations

$$\begin{pmatrix} q_1[n+1] \\ q_2[n+1] \end{pmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ -1 & \frac{5}{2} \end{bmatrix}}_{\mathbf{A}} \begin{pmatrix} q_1[n] \\ q_2[n] \end{pmatrix} + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\mathbf{b}} x[n] \tag{5.46}$$

$$y[n] = \underbrace{\begin{pmatrix} -1 & \frac{1}{2} \end{pmatrix}}_{\mathbf{c}^T} \begin{pmatrix} q_1[n] \\ q_2[n] \end{pmatrix} + x[n] \tag{5.47}$$

A delay-adder-gain block diagram representing (5.46) and (5.47) is shown in Figure 5.1 below.

The modes of the system correspond to the roots of the characteristic polynomial given by

$$\det\,(\lambda\mathbf{I} - \mathbf{A}) = \lambda^2 - \frac{5}{2}\lambda + 1\,. \tag{5.48}$$

These roots are therefore

$$\lambda_1 = 2\,, \qquad \lambda_2 = \frac{1}{2}\,. \tag{5.49}$$

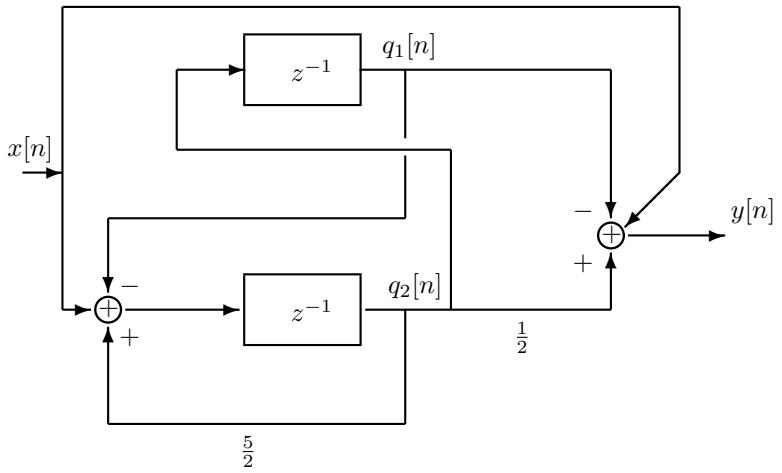

FIGURE 5.1 Delay-adder-gain block diagram for the system in Example 5.1, equations (5.46) and (5.47).

Since it is not the case here that both eigenvalues have magnitude strictly less than 1, the system is not asymptotically stable. The corresponding eigenvectors are found by solving

$$(\lambda \mathbf{I} - \mathbf{A})\mathbf{v} = \begin{pmatrix} \lambda & -1 \\ 1 & \lambda - \frac{5}{2} \end{pmatrix} \mathbf{v} = \mathbf{0} \tag{5.50}$$

with $\lambda = \lambda_1 = 2$, and then again with $\lambda = \lambda_2 = \frac{1}{2}$. This yields

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}. \tag{5.51}$$

The input-output transfer function of the system is given by

$$H(z) = \mathbf{c}^T (z\mathbf{I} - \mathbf{A})^{-1} \mathbf{b} + \mathtt{d} \tag{5.52}$$

$$(z\mathbf{I} - \mathbf{A})^{-1} = \frac{1}{z^2 - \frac{5}{2}z + 1} \begin{bmatrix} z - \frac{5}{2} & 1 \\ -1 & z \end{bmatrix} \tag{5.53}$$

$$\begin{aligned} H(z) &= \frac{1}{z^2 - \frac{5}{2}z + 1} \left\{ \begin{bmatrix} -1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} z - \frac{5}{2} & 1 \\ -1 & z \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\} + 1 \\ &= \frac{1}{2} \ \frac{z-2}{z^2 - \frac{5}{2}z + 1} + 1 = \frac{1}{2} \ \frac{1}{z - \frac{1}{2}} + 1 \\ &= \frac{1}{1 - \frac{1}{2}z^{-1}} \end{aligned} \tag{5.54}$$

Since the transfer function has only one pole and this pole is inside the unit circle, the system is input-output stable. However, the system has two modes, so one of them is a hidden mode, i.e., does not appear in the input-output transfer function. Hidden modes are either unreachable from the input or unobservable in the output, or both. To explicitly check which is the case in this example, we change to modal coordinates, so the original description

$$\mathbf{q}[n+1] = \mathbf{A}\mathbf{q}[n] + \mathbf{b}x[n] \tag{5.55}$$

$$y[n] = \mathbf{c}^T\mathbf{q}[n] + \mathtt{d}x[n] \tag{5.56}$$

gets transformed via

$$\mathbf{q}[n] = \mathbf{V}\mathbf{r}[n] \tag{5.57}$$

to the form

$$\mathbf{r}[n+1] = \underbrace{\mathbf{V}^{-1}\mathbf{A}\mathbf{V}}_{\widehat{\mathbf{A}}=\mathbf{\Lambda}}\,\mathbf{r}[n] \quad + \quad \underbrace{\mathbf{V}^{-1}\mathbf{b}}_{\widehat{\mathbf{b}}=\beta}\,x[n] \tag{5.58}$$

$$y[n] = \underbrace{\mathbf{c}^T\mathbf{V}}_{\widehat{\mathbf{c}}=\xi}\;\mathbf{r}[n] \quad + \quad \mathtt{d}x[n] \tag{5.59}$$

where

$$\mathbf{V} = \begin{bmatrix} | & | \\ \mathbf{v}_1 & \mathbf{v}_2 \\ | & | \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} . \tag{5.60}$$

The new state evolution matrix $\widehat{\mathbf{A}}$ will then be diagonal:

$$\widehat{\mathbf{A}} = \mathbf{\Lambda} = \begin{bmatrix} 2 & 0 \\ \\ 0 & \frac{1}{2} \end{bmatrix} \tag{5.61}$$

and the modified $\mathbf{b}$ and $\mathbf{c}$ matrices will be

$$\widehat{\mathbf{b}} = \beta = \begin{bmatrix} \frac{2}{3} \\ \\ -\frac{1}{3} \end{bmatrix} , \tag{5.62}$$

$$\widehat{\mathbf{c}}^T = \xi = \begin{bmatrix} 0 & -\dfrac{3}{2} \end{bmatrix} , \qquad \mathtt{d} = 1 , \tag{5.63}$$

from which it is clear that the system is reachable (because β has no entries that are 0), but that its eigenvalue $\lambda_1 = 2$ is unobservable (because ξ has a 0 in the first position). Note that if we had mistakenly applied this test in the original coordinates rather than modal coordinates, we would have erroneously decided the first mode is not reachable because the first entry of $\mathbf{b}$ is 0, and that the system is observable because $\mathbf{c}^T$ has no nonzero entries.

In the new coordinates the state equations are

$$\begin{pmatrix} r_1[n+1] \\ r_2[n+1] \end{pmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{pmatrix} r_1[n] \\ r_2[n] \end{pmatrix} + \begin{pmatrix} \frac{2}{3} \\ -\frac{1}{3} \end{pmatrix} x[n] \tag{5.64}$$

$$y[n] = \begin{pmatrix} 0 & -\frac{3}{2} \end{pmatrix} \begin{pmatrix} r_1[n] \\ r_2[n] \end{pmatrix} + x[n] \tag{5.65}$$

or equivalently

$$r_1[n+1] = 2r_1[n] + \frac{2}{3}x[n] \tag{5.66}$$

$$r_2[n+1] = \frac{1}{2}r_2[n] - \frac{1}{3}x[n] \tag{5.67}$$

$$y[n] = -\frac{3}{2}r_2[n] + x[n] \tag{5.68}$$

The delay-adder-gain block diagram represented by (5.64) and (5.65) is shown in Figure 5.2.

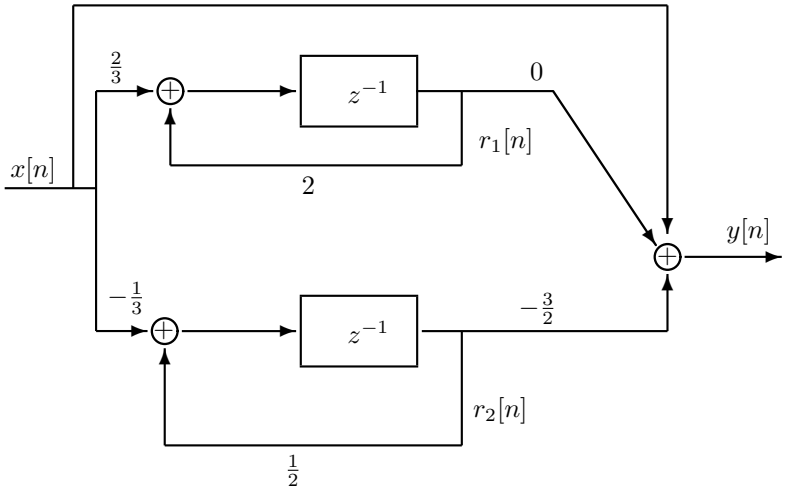

FIGURE 5.2 Delay-adder-gain block diagram for Example 5.1 after a coordinate transformation to display the modes.

In the block diagram of Figure 5.2 representing the state equations in modal coordinates, the modes are individually recognizable. This corresponds to the fact that the original $\mathbf{A}$ matrix has been diagonalized by the coordinate change. From this block diagram we can readily see by inspection that the unstable mode is not observable in the output, since the gain connecting that mode to the output is zero. However, it is reachable from the input.

Note that the block diagram in Figure 5.3 has the same modes and input-output transfer function as that in Figure 5.2. However, in this case the unstable mode is observable but not reachable.

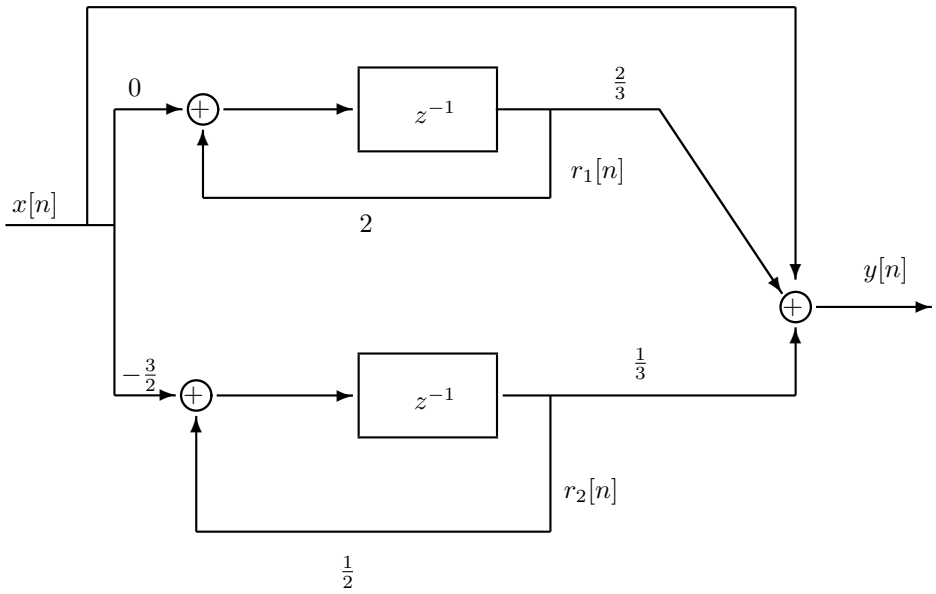

FIGURE 5.3 Delay-adder-gain block diagram for Example 5.1 realizing the same transfer function. In this case the unstable mode is observable but not reachable.

EXAMPLE 5.2 Evaluating asymptotic stability of a linear, periodically varying system

The stability of linear periodically varying systems can be analyzed by methods that are close to those used for LTI systems. Suppose, for instance, that

$$\mathbf{q}[n+1] = \mathbf{A}[n]\mathbf{q}[n] \quad , \quad \mathbf{A}[n] = \mathbf{A}_0 \text{ for even } n, \; \mathbf{A}[n] = \mathbf{A}_1 \text{ for odd } n.$$

Then

$$\mathbf{q}[n+2] = \mathbf{A}_1\mathbf{A}_0\mathbf{q}[n]$$

for even n, so the dynamics of the even samples is governed by an LTI model, and the stability of the even samples is accordingly determined by the eigenvalues of the constant matrix $\mathcal{A}_{even} = \mathbf{A}_1\mathbf{A}_0$. The stability of the odd samples is similarly governed by the eigenvalues of the matrix $\mathcal{A}_{odd} = \mathbf{A}_0\mathbf{A}_1$; it turns out that the nonzero eigenvalues of this matrix are the same as those of $\mathcal{A}_{even}$, so either one can be used for a stability check.

As an example, suppose

$$\mathbf{A}_0 = \begin{pmatrix} 0 & 1 \\ 0 & 3 \end{pmatrix} \quad , \quad \mathbf{A}_1 = \begin{pmatrix} 0 & 1 \\ 4.25 & -1.25 \end{pmatrix} \quad , \tag{5.69}$$

whose respective eigenvalues are $(0\,,\,3)$ and $(1.53\,,\,-2.78)$, so both matrices have eigenvalues of magnitude greater than 1. Now

$$\mathcal{A}_{even} = A_1A_0 = \begin{pmatrix} 0 & 3 \\ 0 & 0.5 \end{pmatrix} \quad , \tag{5.70}$$

and its eigenvalues are $(0\,,\,0.5)$, which corresponds to a stable system!

CHAPTER 6

State Observers and State Feedback

Our study of the modal solutions of LTI state-space models made clear in complete analytical detail that the state at any given time summarizes everything about the past that is relevant to future behavior of the model. More specifically, given the value of the state vector at some initial instant, and given the entire input trajectory over some interval of time extending from the initial instant into the future, one can determine the entire future state and output trajectories of the model over that interval. The same general conclusion holds for nonlinear and time-varying state-space models, although they are generally far less tractable analytically. Our focus will be on LTI models.

It is typically the case that we do not have any direct measurement of the initial state of a system, and will have to make some guess or estimate of it. This uncertainty about the initial state generates uncertainty about the future state trajectory, even if our model for the system is perfect, and even if we have accurate knowledge of the inputs to the system.

The first part of this chapter is devoted to addressing the issue of state trajectory estimation, given uncertainty about the initial state of the system. We shall see that the state can actually be asymptotically determined under appropriate conditions, by means of a so-called state observer. The observer uses a model of the system along with past measurements of both the input and output trajectories of the system.

The second part of the chapter examines how the input to the system should be controlled in order to yield desirable system behavior. We shall see that having knowledge of the present state of the system provides a powerful basis for designing feedback control to stabilize or otherwise improve the behavior of the resulting closed-loop system. When direct measurements of the state are not available, the asymptotic state estimate provided by an observer turns out to suffice.

6.1 PLANT AND MODEL

It is important now to make a distinction between the actual, physical (and causal) system we are interested in studying or working with or controlling — what is often termed the plant (as in "physical plant") — and our idealized model for the plant. The plant is usually a complex, highly nonlinear and time-varying object, typically requiring an infinite number (or a continuum) of state variables and parameters to represent it with ultimate fidelity. Our model, on the other hand, is an idealized and simplified (and often LTI) representation, of relatively low order, that aims to

capture the behavior of the plant in some limited regime of its operation, while remaining tractable for analysis, computation, simulation and design.

The inputs to the model represent the inputs acting on or driving the actual plant, and the outputs of the model represent signals in the plant that are accessible for measurement. In practice we will typically not know all the driving inputs to the plant exactly. Apart from those driving inputs that we have access to, there will also generally be additional unmeasured disturbance inputs acting on the plant that we are only able to characterize in some general way, perhaps as random processes. Similarly, the measured outputs of the plant will differ from what we might predict on the basis of our limited model, partly because of measurement noise.

6.2 STATE ESTIMATION BY REAL-TIME SIMULATION

Suppose the plant of interest to us is correctly described by the following equations, which constitute an Lth-order LTI state-space representation of the plant:

$$\mathbf{q}[n+1] = \mathbf{A}\mathbf{q}[n] + \mathbf{b}x[n] + \mathbf{w}[n] \ , \tag{6.1}$$

$$y[n] = \mathbf{c}^T\mathbf{q}[n] + \mathtt{d}x[n] + \zeta[n] \ . \tag{6.2}$$

Here $x[n]$ denotes the known (scalar) control input, and $\mathbf{w}[n]$ denotes the vector of unknown disturbances that drive the plant, not necessarily through the same channels as the input $x[n]$. For example, we might have $\mathbf{w}[n] = \mathbf{f}v[n]$, where $v[n]$ is a scalar disturbance signal and $\mathbf{f}$ is a vector describing how this scalar disturbance drives the system (just as $\mathbf{b}$ describes how $x[n]$ drives the system). The quantity $y[n]$ denotes the known or measured (scalar) output, and $\zeta[n]$ denotes the unknown noise in this measured output. We refer to $\mathbf{w}[n]$ as plant disturbance or plant noise, and to $\zeta[n]$ as measurement noise. We focus mainly on the DT case now, but essentially everything carries over in a natural way to the CT case.

With the above equations representing the true plant, what sort of model might we use to study or simulate the behavior of the plant, given that we know $x[n]$ and $y[n]$? If nothing further was known about the disturbance variables in $\mathbf{w}[n]$ and the measurement noise $\zeta[n]$, or if we only knew that they could be represented as zero-mean random processes, for instance, then one strategy would be to simply ignore these variables when studying or simulating the plant. If everything else about the plant was known, our representation of the plant's behavior would be embodied in an LTI state-space model of the form

$$\widehat{\mathbf{q}}[n+1] = \mathbf{A}\widehat{\mathbf{q}}[n] + \mathbf{b}x[n] \ , \tag{6.3}$$

$$\widehat{y}[n] = \mathbf{c}^T\widehat{\mathbf{q}}[n] + \mathtt{d}x[n] \ . \tag{6.4}$$

The $x[n]$ that drives our model is the same known $x[n]$ that is an input (along with possibly other inputs) to the plant. However, the state $\widehat{\mathbf{q}}[n]$ and output $\widehat{y}[n]$ of the model will generally differ from the corresponding state $\mathbf{q}[n]$ and output $y[n]$ of the plant, because in our formulation the plant state and output are additionally perturbed by $\mathbf{w}[n]$ and $\zeta[n]$ respectively. The assumption that our model has correctly captured the dynamics of the plant and the relationships among the variables is

what allows us to use the same $\mathbf{A}$, $\mathbf{b}$, $\mathbf{c}^T$ and $\mathtt{d}$ in our model as occur in the "true" plant.

It bears repeating that in reality there are several sources of uncertainty we are ignoring here. At the very least, there will be discrepancies between the actual and assumed parameter values — i.e., between the actual entries of $\mathbf{A}$, $\mathbf{b}$, $\mathbf{c}^T$ and $\mathtt{d}$ in (6.1), (6.2) and the assumed entries of these matrices in (6.3), (6.4) respectively. Even more troublesome is the fact that the actual system is probably more accurately represented by a nonlinear, time-varying model of much higher order than that of our assumed LTI model, and with various other disturbance signals acting on it. We shall not examine the effects of all these additional sources of uncertainty.

With a model in hand, it is natural to consider obtaining an estimate of the current plant state by running the model forward in real time, as a simulator. For this, we initialize the model (6.3) at some initial time (which we take to be $n = 0$ without loss of generality), picking its initial state $\widehat{\mathbf{q}}[0]$ to be some guess or estimate of the initial state of the plant. We then drive the model with the known input $x[n]$ from time $n = 0$ onwards, generating an estimated or predicted state trajectory $\widehat{\mathbf{q}}[n]$ for $n > 0$. We could then also generate the predicted output $\widehat{y}[n]$ using the prescription in (6.4).

In order to examine how well this real-time simulator performs as a state estimator, we examine the error vector

$$\widetilde{\mathbf{q}}[n] = \mathbf{q}[n] - \widehat{\mathbf{q}}[n] \; . \tag{6.5}$$

Note that $\widetilde{\mathbf{q}}[n]$ is the difference between the actual and estimated (or predicted) state trajectories. By subtracting (6.3) from (6.1), we see that this difference, the estimation error or prediction error $\widetilde{\mathbf{q}}[n]$, is itself governed by an LTI state-space equation:

$$\widetilde{\mathbf{q}}[n+1] = \mathbf{A}\widetilde{\mathbf{q}}[n] + \mathbf{w}[n] \tag{6.6}$$

with initial condition

$$\widetilde{\mathbf{q}}[0] = \mathbf{q}[0] - \widehat{\mathbf{q}}[0] \; . \tag{6.7}$$

This initial condition is our uncertainty about the initial state of the plant.

What (6.6) shows is that, if the original system (6.1) is unstable (i.e., if $\mathbf{A}$ has eigenvalues of magnitude greater than 1), or has otherwise undesirable dynamics, and if either $\widetilde{\mathbf{q}}[0]$ or $\mathbf{w}[n]$ is nonzero, then the error $\widetilde{\mathbf{q}}[n]$ between the actual and estimated state trajectories will grow exponentially, or will have otherwise undesirable behavior, see Figure 6.1. Even if the plant is not unstable, we see from (6.6) that the error dynamics are driven by the disturbance process $\mathbf{w}[n]$, and we have no means to shape the effect of this disturbance on the estimation error. The real-time simulator is thus generally an inadequate way of reconstructing the state.

6.3 THE STATE OBSERVER

To do better than the real-time simulator (6.3), we must use not only the input $x[n]$ but also the measured output $y[n]$. The key idea is to use the discrepancy between

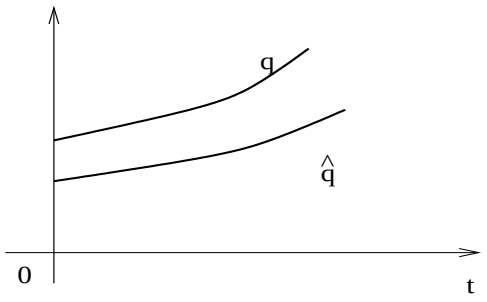

FIGURE 6.1 Schematic representation of the effect of an erroneous initial condition on the state estimate produced by the real-time simulator for an unstable plant.

actual and predicted outputs, $y[n]$ in (6.2) and $\widehat{y}[n]$ in (6.4) respectively — i.e., to use the output prediction error — as a correction term for the real-time simulator. The resulting system is termed a state observer (or state estimator) for the plant, and in our setting takes the form

$$\begin{aligned}\widehat{\mathbf{q}}[n+1] = \mathbf{A}\widehat{\mathbf{q}}[n] + \mathbf{b}x[n] \\ - \ell\Big(y[n] - \widehat{y}[n]\Big) \ .\end{aligned} \tag{6.8}$$

The observer equation above has been written in a way that displays its two constituent parts: a part that simulates as closely as possible the plant whose states we are trying to estimate, and a part that feeds the correction term $y[n]-\widehat{y}[n]$ into this simulation. This correction term is applied through the L-component vector ℓ, termed the observer gain vector, with ith component ℓ_i. (The negative sign in front of ℓ in (6.8) is used only to simplify the appearance of some later expressions). Figure 6.2 is a block-diagram representation of the resulting structure.

Now subtracting (6.8) from (6.1), we find that the state estimation error or observer error satisfies

$$\begin{aligned}\widetilde{\mathbf{q}}[n+1] &= \mathbf{A}\widetilde{\mathbf{q}}[n] + \mathbf{w}[n] + \ell\Big(y[n] - \mathbf{c}^T\widehat{\mathbf{q}}[n] - \mathtt{d}x[n]\Big) \\ &= (\mathbf{A} + \ell\mathbf{c}^T)\widetilde{\mathbf{q}}[n] + \mathbf{w}[n] + \ell\zeta[n] \ .\end{aligned} \tag{6.9}$$

If the observer gain ℓ is 0, then the error dynamics are evidently just the dynamics of the real-time simulator (6.6). More generally, the dynamics are governed by the system's natural frequencies, namely the eigenvalues of $\mathbf{A}+\ell\mathbf{c}^T$ or the roots of the characteristic polynomial

$$\kappa(\lambda) = \det\Big(\lambda\mathbf{I} - (\mathbf{A} + \ell\mathbf{c}^T)\Big) \tag{6.10}$$

$$= \lambda^L + \kappa_{L-1}\lambda^{L-1} + \cdots + \kappa_0 \ . \tag{6.11}$$

(This polynomial, like all the characteristic polynomials we deal with, has real coefficients and is monic, i.e., its highest-degree term is scaled by 1 rather than some non-unit scalar.)

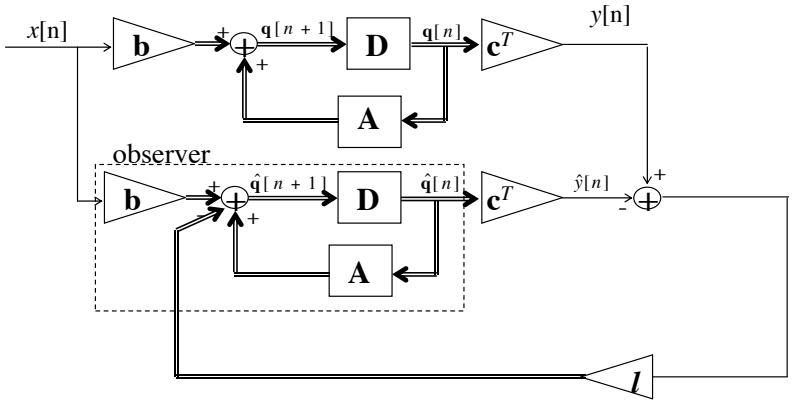

FIGURE 6.2 An observer for the plant in the upper part of the diagram comprises a real-time simulation of the plant, driven by the same input, and corrected by a signal derived from the output prediction error.

Two questions immediately arise:

(i) How much freedom do we have in placing the observer eigenvalues, i.e., the eigenvalues of $\mathbf{A} + \ell\mathbf{c}^T$ or the roots of $\kappa(\lambda)$, by appropriate choice of the observer gain ℓ?

(ii) How does the choice of ℓ shape the effects of the disturbance and noise terms $\mathbf{w}[n]$ and $\zeta[n]$ on the observer error?

Brief answers to these questions are respectively as follows:

(i) At $\ell = 0$ the observer eigenvalues, namely the eigenvalues of $\mathbf{A} + \ell\mathbf{c}^T$, are those of the real-time simulator, which are also those of the given system or plant. By varying the entries of ℓ away from 0, it turns out we can move all the eigenvalues that correspond to observable eigenvalues of the plant (which may number as many as L eigenvalues), and those are the only eigenvalues we can move. Moreover, appropriate choice of ℓ allows us, in principle, to move these observable eigenvalues to any arbitrary set of self-conjugate points in the complex plane. (A self-conjugate set is one that remains unchanged by taking the complex conjugate of the set. This is equivalent to requiring that if a complex point is in such a set, then its complex conjugate is as well.) The self-conjugacy restriction is necessary because we are working with real

parameters and gains.

The unobservable eigenvalues of the plant remain eigenvalues of the observer, and cannot be moved. (This claim can be explicitly demonstrated by transformation to modal coordinates, but we omit the details.) The reason for this is that information about these unobservable modes does not make its way into the output prediction error that is used in the observer to correct the real-time simulator.

It follows from the preceding statements that a stable observer can be designed if and only if all unobservable modes of the plant are stable (a property that is termed detectability). Also, the observer can be designed to have an arbitrary characteristic polynomial $\kappa(\lambda)$ if and only if the plant is observable.

We shall not prove the various claims above. Instead, we limit ourselves to proving, later in this chapter, a closely analogous set of results for the case of state feedback control.

In designing observers analytically for low-order systems, one way to proceed is by specifying a desired set of observer eigenvalues $\epsilon_1, \cdots \epsilon_L$, thus specifying the observer characteristic polynomial $\kappa(\lambda)$ as

$$\kappa(\lambda) = \prod_{i=1}^{L} (\lambda - \epsilon_i) \ . \tag{6.12}$$

Expanding this out and equating it to $\det\Big(\lambda\mathbf{I} - (\mathbf{A} + \ell\mathbf{c}^T)\Big)$, as in (6.10), yields L simultaneous linear equations in the unknown gains $\ell_1, \cdots, \ell_L$. These equations will be consistent and solvable for the observer gains if and only if all the unobservable eigenvalues of the plant are included among the specified observer eigenvalues $\{\epsilon_i\}$.

The preceding results also suggest an alternative way to determine the unobservable eigenvalues of the plant: the roots of $\det\Big(\lambda\mathbf{I} - (\mathbf{A} + \ell\mathbf{c}^T)\Big)$ that cannot be moved, no matter how ℓ is chosen, are precisely the unobservable eigenvalues of the plant. This approach to exposing unobservable modes can be easier in some problems than the approach used in the previous chapter, which required first computing the eigenvectors $\{\mathbf{v}_i\}$ of the system, and then checking for which i we had $\mathbf{c}^T\mathbf{v}_i = 0$.

(ii) We now address how the choice of ℓ shapes the effects of the disturbance and noise terms $\mathbf{w}[n]$ and $\zeta[n]$ on the observer error. The first point to note is that if the error system (6.9) is made asymptotically stable by appropriate choice of observer gain ℓ, then bounded plant disturbance $\mathbf{w}[n]$ and bounded measurement noise $\zeta[n]$ will result in the observer error being bounded. This is most easily proved by transforming to modal coordinates, but we omit the details.

The observer error equation (6.9) shows that the observer gain ℓ enters in two places, first in causing the error dynamics to be governed by the state evolution matrix $\mathbf{A} + \ell\mathbf{c}^T$ rather than $\mathbf{A}$, and again as the input vector for the measurement noise $\zeta[n]$. This highlights a basic tradeoff between error

decay and noise immunity. The observer gain can be used to obtain fast error decay, as might be needed in the presence of plant disturbances $\mathbf{w}[n]$ that continually perturb the system state away from where we think it is — but large entries in ℓ may be required to accomplish this (certainly in the CT case, but also in DT if the model is a sampled-data version of some underlying CT system, as in the following example), and these large entries in ℓ will have the undesired result of accentuating the effect of the measurement noise. A large observer gain may also increase the susceptibility of the observer design to mod! eling errors and other discrepancies. In practice, such considerations would lead us design somewhat conservatively, not attempting to obtain overly fast error-decay dynamics.

Some aspects of the tradeoffs above can be captured in a tractable optimization problem. Modeling $\mathbf{w}[n]$ and $\zeta[n]$ as stationary random processes (which are introduced in a later chapter), we can formulate the problem of picking ℓ to minimize some measure of the steady-state variances in the components of the state estimation error $\widetilde{\mathbf{q}}[n]$. The solution to this and a range of related problems is provided by the so-called Kalman filtering framework. We will be in a position to work through some elementary versions of this once we have developed the machinery for dealing with stationary random processes.

EXAMPLE 6.1 Ship Steering

Consider the following simplified sampled-data model for the steering dynamics of a ship traveling at constant speed, with a rudder angle that is controlled in a piecewise-constant fashion by a computer-based controller:

$$\begin{aligned}\mathbf{q}[n+1] = \begin{bmatrix} q_1[n+1] \\ q_2[n+1] \end{bmatrix} &= \begin{bmatrix} 1 & \sigma \\ 0 & \alpha \end{bmatrix} \begin{bmatrix} q_1[n] \\ q_2[n] \end{bmatrix} + \begin{bmatrix} \epsilon \\ \sigma \end{bmatrix} x[n] \\ &= \mathbf{A}\mathbf{q}[n] + \mathbf{b}x[n] \,. \end{aligned} \tag{6.13}$$

The state vector $\mathbf{q}[n]$ comprises the sampled heading error $q_1[n]$ (which is the direction the ship points in, relative to the desired direction of motion) and the sampled rate of turn $q_2[n]$ of the ship, both sampled at time $t = nT$; $x[n]$ is the constant value of the rudder angle (relative to the direction in which the ship points) in the interval $nT \le t < nT + T$ (we pick positive rudder angle to be that which would tend to increase the heading error). The positive parameters α, σ and ϵ are determined by the type of ship, its speed, and the sampling interval T. In particular, α is generally smaller than 1, but can be larger than 1 for a large tanker; in any case, the system (6.13) is not asymptotically stable. The constant σ is approximately equal to the sampling interval T.

Suppose we had (noisy) measurements of the rate of turn, so

$$\mathbf{c}^T = \begin{pmatrix} 0 & 1 \end{pmatrix} \,. \tag{6.14}$$

Then

$$\mathbf{A} + \ell\mathbf{c}^T = \begin{pmatrix} 1 & \sigma + \ell_1 \\ 0 & \alpha + \ell_2 \end{pmatrix} \,. \tag{6.15}$$

Evidently one natural frequency of the error equation is fixed at 1, no matter what ℓ is. This natural frequency corresponds to a mode of the original system that is unobservable from rate-of-turn measurements. Moreover, it is not an asymptotically stable mode, so the corresponding observer error will not decay. Physically, the problem is that the rate of turn contains no input from or information about the heading error itself.

If, instead, we have (noisy) measurements of the heading error, so

$$\mathbf{c}^T = \begin{pmatrix} 1 & 0 \end{pmatrix} . \tag{6.16}$$

In this case

$$\mathbf{A} + \ell\mathbf{c}^T = \begin{pmatrix} 1+\ell_1 & \sigma \\ \ell_2 & \alpha \end{pmatrix} . \tag{6.17}$$

The characteristic polynomial of this matrix is

$$\kappa(\lambda) = \lambda^2 - \lambda(1+\ell_1+\alpha) + \alpha(1+\ell_1) - \ell_2\sigma . \tag{6.18}$$

This can be made into an arbitrary monic polynomial of degree 2 by choice of the gains ℓ_1 and ℓ_2, which also establishes the observability of our plant model.

One interesting choice of observer gains in this case is $\ell_1 = -1-\alpha$ and $\ell_2 = -\alpha^2/\sigma$ (which, for typical parameter values, results in ℓ_2 being large). With this choice,

$$\mathbf{A} + \ell\mathbf{c}^T = \begin{pmatrix} -\alpha & \sigma \\ -\alpha^2/\sigma & \alpha \end{pmatrix} . \tag{6.19}$$

The characteristic polynomial of this matrix is $\kappa(\lambda) = \lambda^2$, so the natural frequencies of the observer error equation are both at 0.

A DT LTI system with all natural frequencies at 0 is referred to as deadbeat, because its zero-input response settles exactly to the origin in finite time. (This finite-time settling is possible for the zero-input response of an LTI DT system, but not for an LTI CT system, though of course it is possible for an LTI CT system to have an arbitrarily small zero-input response after any specified positive time.) We have not discussed how to analyze LTI state-space models with non-distinct eigenvalues, but to verify the above claim of finite settling for our observer, it suffices to confirm from (6.19) that $(\mathbf{A} + \ell\mathbf{c}^T)^2 = 0$ when the gains ℓ_i are chosen to yield $\kappa(\lambda) = \lambda^2$. This implies that in the absence of plant disturbance and measurement noise, the observer error goes to 0 in at most two steps.

In the presence of measurement noise, one may want to choose a slower error decay, so as to keep the observer gain ℓ — and ℓ_2 in particular — smaller than in the deadbeat case, and thereby not accentuate the effects of measurement noise on the estimation error.

6.4 STATE FEEDBACK CONTROL

For a causal system or plant with inputs that we are able to manipulate, it is natural to ask how the inputs should be chosen in order to cause the system to

behave in some desirable fashion. Feedback control of such a system is based on sensing its present or past behavior, and using the measurements of the sensed variables to generate control signals to apply to it. Feedback control is also referred to as closed-loop control.

Open-loop control, by contrast, is not based on continuous monitoring of the plant, but rather on using only information available at the time that one starts interacting with the system. The trouble with open-loop control is that errors, even if recognized, are not corrected or compensated for. If the plant is poorly behaved or unstable, then uncorrected errors can lead to bad or catastrophic consequences.

Feedforward control refers to schemes incorporating measurements of signals that currently or in the future will affect the plant, but that are not themselves affected by the control. For example, in generating electrical control signals for the positioning motor of a steerable radar antenna, the use of measurements of wind velocity would correspond to feedforward control, whereas the use of measurements of antenna position would correspond to feedback control. Controls can have both feedback and feedforward components.

Our focus in this section is on feedback control. To keep our development streamlined, we assume the plant is well modeled by the following Lth-order LTI state-space description:

$$\mathbf{q}[n+1] = \mathbf{A}\mathbf{q}[n] + \mathbf{b}x[n] \tag{6.20}$$

$$y[n] = \mathbf{c}^T\mathbf{q}[n] \tag{6.21}$$

rather than the more elaborate description (6.1), (6.2). As always, $x[n]$ denotes the control input and $y[n]$ denotes the measured output, both taken to be scalar functions of time. We shall also refer to this as the open-loop system. Again, we treat the DT case, but essentially everything carries over naturally to CT. Also, for notational simplicity, we omit from (6.21) the direct feedthrough term $\mathtt{d}x[n]$ that has appeared in our system descriptions until now, because this term can complicate the appearance of some of the expressions we derive, without being of much significance in itself; it is easily accounted for if necessary.

Denote the characteristic polynomial of the matrix $\mathbf{A}$ in (6.20) by

$$a(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A}) = \prod_{i=1}^{L}(\lambda - \lambda_i)\ . \tag{6.22}$$

The transfer function $H(z)$ of the system (6.20), (6.21) is given by

$$H(z) = \mathbf{c}^T(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} \tag{6.23}$$

$$= \frac{\eta(z)}{a(z)}\ . \tag{6.24}$$

(The absence of the direct feedthrough term in (6.21) causes the degree of the polynomial $\eta(z)$ to be strictly less than L. If the feedthrough term was present, the transfer function would simply have $\mathtt{d}$ added to the $H(z)$ above.) Note that there

may be pole-zero cancelations involving common roots of $a(z)$ and $\eta(z)$ in (6.24), corresponding to the presence of unreachable and/or unobservable modes of the system. Only the uncanceled roots of $a(z)$ survive as poles of $H(z)$, and similarly only the uncanceled roots of $\eta(z)$ survive as zeros of the transfer function.

We reiterate that the model undoubtedly differs from the plant in many ways, but we shall not examine the effects of various possible sources of discrepancy and uncertainty. A proper treatment of such issues constitutes the field of robust control, which continues to be an active area of research.

Since the state of a system completely summarizes the relevant past of the system, we should expect that knowledge of the state at every instant gives us a powerful basis for designing feedback control signals. In this section we consider the use of state feedback for the system (6.20), assuming that we have access to the entire state vector at each time. Though this assumption is unrealistic in general, it will allow us to develop some preliminary results as a benchmark. We shall later consider what happens when we treat the more realistic situation, where the state cannot be measured but has to be estimated instead. It will turn out in the LTI case that the state estimate provided by an observer will actually suffice to accomplish much of what can be achieved when the actual state is used for feedback.

The particular case of LTI state feedback is represented in Figure 6.3, in which the feedback part of the input $x[n]$ is a constant linear function of the state $\mathbf{q}[n]$ at that instant:

$$x[n] = p[n] + \mathbf{g}^T\mathbf{q}[n] \tag{6.25}$$

where the L-component row vector $\mathbf{g}^T$ is the state feedback gain vector (with ith component g_i), and $p[n]$ is some external input signal that can be used to augment the feedback signal. Thus $x[n]$ is $p[n]$ plus a weighted linear combination of the state variables $q_i[n]$, with constant weights g_i.

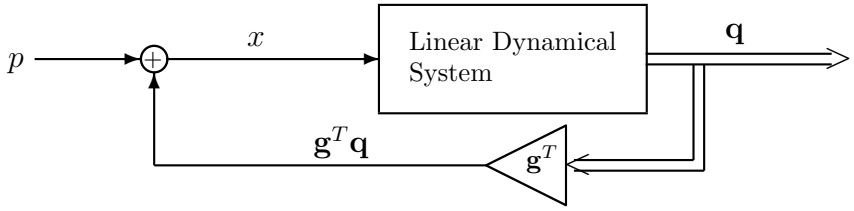

FIGURE 6.3 Linear dynamical system with LTI state feedback. The single lines denote scalar signals and the double lines denote vector signals.

With this choice for $x[n]$, the system (6.20) becomes

$$\begin{aligned}\mathbf{q}[n+1] &= \mathbf{A}\mathbf{q}[n] + \mathbf{b}\Big(p[n] + \mathbf{g}^T\mathbf{q}[n]\Big) \\ &= \Big(\mathbf{A} + \mathbf{b}\mathbf{g}^T\Big)\mathbf{q}[n] + \mathbf{b}p[n]\ .\end{aligned} \tag{6.26}$$

The behavior of this closed-loop system, and in particular its stability, is governed by its natural frequencies, namely by the L eigenvalues of the matrix $\mathbf{A}+\mathbf{b}\mathbf{g}^T$ or the roots of the characteristic polynomial

$$\nu(\lambda) = \det\Big(\lambda\mathbf{I} - (\mathbf{A}+\mathbf{b}\mathbf{g}^T)\Big) \tag{6.27}$$

$$= \lambda^L + \nu_{L-1}\lambda^{L-1} + \cdots + \nu_0 \; . \tag{6.28}$$

Some questions immediately arise:

(i) How much freedom do we have in placing the closed-loop eigenvalues, i.e., the eigenvalues of $\mathbf{A}+\mathbf{b}\mathbf{g}^T$ or the roots of $\nu(\lambda)$, by appropriate choice of the state feedback gain $\mathbf{g}^T$?

(ii) How does state feedback affect reachability, observability and the transfer function of the system?

(iii) How does the choice of $\mathbf{g}^T$ affect the state behavior and the control effort that is required?

Brief answers to these (inter-related) questions are respectively as follows:

(i) By varying the entries of $\mathbf{g}^T$ away from 0, we can move all the reachable eigenvalues of the system (which may number as many as L), and only those eigenvalues. Moreover, appropriate choice of $\mathbf{g}^T$ allows us, in principle, to move the reachable eigenvalues to any arbitrary set of self-conjugate points in the complex plane.

The unreachable eigenvalues of the open-loop system remain eigenvalues of the closed-loop system, and cannot be moved. (This can be explicitly demonstrated by transformation to modal coordinates, but we omit the details.) The reason for this is that the control input cannot access these unreachable modes.

It follows from the preceding claims that a stable closed-loop system can be designed if and only if all unreachable modes of the open-loop system are stable (a property that is termed stabilizability). Also, state feedback can yield an arbitrary closed-loop characteristic polynomial $\nu(\lambda)$ if and only if the open-loop system (6.20) is reachable.

The proof for the above claims is presented in Section 6.4.1.

In designing state feedback control analytically for low-order examples, one way to proceed is by specifying a desired set of closed-loop eigenvalues $\mu_1, \cdots \mu_L$, thus specifying $\nu(\lambda)$ as

$$\nu(\lambda) = \prod_{i=1}^{L}(\lambda - \nu_i) \; . \tag{6.29}$$

Expanding this out and equating it to $\det\Big(\lambda\mathbf{I} - (\mathbf{A}+\mathbf{b}\mathbf{g}^T)\Big)$, as in (6.27), yields L simultaneous linear equations in the unknown gains $g_1, \cdots, g_L$. These equations will be consistent and solvable for the state feedback gains if and

only if all the unreachable eigenvalues of the plant are included among the specified closed-loop eigenvalues $\{\mu_i\}$.

The preceding results also suggest an alternative way to determine the unreachable eigenvalues of the given plant: the roots of $\det\left(\lambda\mathbf{I}-(\mathbf{A}+\mathbf{b}\mathbf{g}^T)\right)$ that cannot be moved, no matter how $\mathbf{g}^T$ is chosen, are precisely the unreachable eigenvalues of the plant. This approach to exposing unreachable modes can be easier in some problems than the approach used in the previous chapter, which required first computing the eigenvectors $\{\mathbf{v}_i\}$ of the plant, and then checking which of these eigenvectors were not needed in writing $\mathbf{b}$ as a linear combination of the eigenvectors.

[The above discussion has closely paralleled our discussion of observers, except that observability statements have been replaced by reachability statements throughout. The underlying reason for this "duality" is that the eigenvalues of $\mathbf{A}+\mathbf{b}\mathbf{g}^T$ are the same as those of its transpose, namely $\mathbf{A}^T+\mathbf{g}\mathbf{b}^T$. The latter matrix has exactly the structure of the matrix $\mathbf{A}+\ell\mathbf{c}^T$ that was the focus of our discussion of observers, except that $\mathbf{A}$ is now replaced by $\mathbf{A}^T$, and $\mathbf{c}^T$ is replaced by $\mathbf{b}^T$. It is not hard to see that the structure of observable and unobservable modes determined by the pair $\mathbf{A}^T$ and $\mathbf{b}^T$ is the same as the structure of reachable and unreachable modes determined by the pair $\mathbf{A}$ and $\mathbf{b}$.]

(ii) The results in part (i) above already suggest the following fact: that whether or not an eigenvalue is reachable from the external input — i.e., from $x[n]$ for the open-loop system and $p[n]$ for the closed-loop system — is unaffected by state feedback. An unreachable eigenvalue of the open-loop system cannot be excited from the input $x[n]$, no matter how the input is generated, and therefore cannot be excited even in closed loop (which also explains why it cannot be moved by state feedback). Similarly, a reachable eigenvalue of the open-loop system can also be excited in the closed-loop system, because any $x[n]$ that excites it in the open-loop system may be generated in the closed-loop system by choosing $p[n]=x[n]-\mathbf{g}^T\mathbf{q}[n]$.

The proof in Section 6.4.1 of the claims in (i) will also establish that the transfer function of the closed-loop system, from $p[n]$ to $y[n]$, is now

$$H_{cl}(z)=\mathbf{c}^T\left(z\mathbf{I}-(\mathbf{A}+\mathbf{b}\mathbf{g}^T)\right)^{-1}\mathbf{b} \tag{6.30}$$

$$=\frac{\eta(z)}{\nu(z)}\,. \tag{6.31}$$

Thus the zeros of the closed-loop transfer function are still drawn from the roots of the same numerator polynomial $\eta(z)$ in (6.24) that contains the zeros of the open-loop system; state feedback does not change $\eta(z)$. However, the actual zeros of the closed-loop system are those roots of $\eta(z)$ that are not canceled by roots of the new closed-loop characteristic polynomial $\nu(z)$, and may therefore differ from the zeros of the open-loop system.

We know from the previous chapter that hidden modes in a transfer function are the result of the modes being unreachable and/or unobservable. Because

state feedback cannot alter reachability properties, it follows that any changes in cancelations of roots of $\eta(z)$, in going from the original open-loop system to the closed-loop one, must be the result of state feedback altering the observability properties of the original modes. If an unobservable (but reachable) eigenvalue of the open-loop system is moved by state feedback and becomes observable, then a previously canceled root of $\eta(z)$ is no longer canceled and now appears as a zero of the closed-loop system. Similarly, if an observable (and reachable) eigenvalue of the open-loop system is moved by state feedback to a location where it now cancels a root of $\eta(z)$, then this root is no longer a zero of the closed-loop system, and this hidden mode corresponds to a mode that has been made unobservable by state feed! back.

(iii) We turn now to the question of how the choice of $\mathbf{g}^T$ affects the state behavior and the control effort that is required. Note first that if $\mathbf{g}^T$ is chosen such that the closed-loop system is asymptotically stable, then a bounded external signal $p[n]$ in (6.26) will lead to a bounded state trajectory in the closed-loop system. This is easily seen by considering the transformation of (6.26) to modal coordinates, but we omit the details.

The state feedback gain $\mathbf{g}^T$ affects the closed-loop system in two key ways, first by causing the dynamics to be governed by the eigenvalues of $\mathbf{A}+\mathbf{b}\mathbf{g}^T$ rather than those of $\mathbf{A}$, and second by determining the scaling of the control input $x[n]$ via the relationship in (6.25). This highlights a basic tradeoff between the response rate and the control effort. The state feedback gain can be used to obtain a fast response, to bring the system state from its initially disturbed value rapidly back to the origin — but large entries in $\mathbf{g}^T$ may be needed to do this (certainly in the CT case, but also in DT if the model is a sampled-data version of some underlying CT system), and these large entries in $\mathbf{g}^T$ result in large control effort being expended. Furthermore, the effects of any errors in measuring or estimating the state vector, or of modeling errors and other discrepancies, are likely to be accentuated with large feedback gains. In practice, these considerations would lead us design somewhat conservatively, not attempting to obtain overly fast closed-loop dynamics. Again, some aspects of the tradeoffs involved can be captured in tractable optimization problems, but these are left to more advanced courses.

We work through a CT example first, partly to make clear that our development carries over directly from the DT to the CT case.

EXAMPLE 6.2 Inverted Pendulum with Torque Control

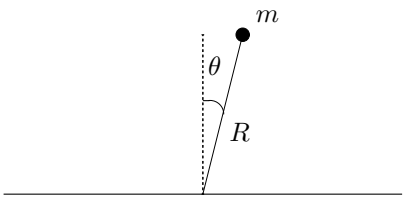

FIGURE 6.4 Inverted pendulum.

Consider the inverted pendulum shown in Figure 6.4, comprising a mass m at the end of a light, hinged rod of length R. For small deviations $\theta(t)$ from the vertical,

$$\frac{d^2\theta(t)}{dt^2} = K\theta(t) + \sigma x(t) \; , \tag{6.32}$$

where $K = g/R$ (g being the acceleration due to gravity), $\sigma = 1/(mR^2)$, and a torque input $x(t)$ is applied at the point of support of the pendulum. Define $q_1(t) = \theta(t)$, $q_2(t) = \dot{\theta}(t)$; then

$$\dot{\mathbf{q}}(t) = \begin{bmatrix} 0 & 1 \\ K & 0 \end{bmatrix} \mathbf{q}(t) + \begin{bmatrix} 0 \\ \sigma \end{bmatrix} x(t) \; . \tag{6.33}$$

We could now determine the system eigenvalues and eigenvectors to decide whether the system is reachable. However, this step is actually not necessary in order to assess reachability and compute a state feedback. Instead, considering directly the effect of the state feedback, we find

$$x(t) = \mathbf{g}^T \mathbf{q}(t) \tag{6.34}$$

$$\dot{\mathbf{q}}(t) = \begin{bmatrix} 0 & 1 \\ K & 0 \end{bmatrix} \mathbf{q}(t) + \begin{bmatrix} 0 \\ \sigma \end{bmatrix} \begin{bmatrix} g_1 & g_2 \end{bmatrix} \mathbf{q}(t) \tag{6.35}$$

$$= \begin{bmatrix} 0 & 1 \\ K + \sigma g_1 & \sigma g_2 \end{bmatrix} \mathbf{q}(t) \; . \tag{6.36}$$

The corresponding characteristic polynomial is

$$\nu(\lambda) = \lambda^2 - \lambda\sigma g_2 - (K + \sigma g_1) \; . \tag{6.37}$$

Inspection of this expression shows that by appropriate choice of the real gains g_1 and g_2 we can make this polynomial into any desired monic second-degree polynomial. In other words, we can obtain any self-conjugate set of closed-loop eigenvalues. This also establishes that the original system is reachable.

Suppose we want the closed-loop eigenvalues at particular numbers μ_1, μ_2, which is equivalent to specifying the closed-loop characteristic polynomial to be

$$\nu(\lambda) = (\lambda - \mu_1)(\lambda - \mu_2) = \lambda^2 - \lambda(\mu_1 + \mu_2) + \mu_1\mu_2 \; . \tag{6.38}$$

Equating this to the polynomial in (6.37) shows that

$$g_1 = -\frac{\mu_1\mu_2 + K}{\sigma} \qquad \text{and} \qquad g_2 = \frac{\mu_1 + \mu_2}{\sigma} \; . \tag{6.39}$$

Both gains are negative when μ_1 and μ_2 form a self-conjugate set in the open left-half plane.

We return now to the ship steering example introduced earlier.

EXAMPLE 6.3 Ship Steering (continued)

Consider again the DT state-space model in Example 6.1, repeated here for convenience:

$$\begin{aligned} \mathbf{q}[n+1] = \begin{bmatrix} q_1[n+1] \\ q_2[n+1] \end{bmatrix} &= \begin{bmatrix} 1 & \sigma \\ 0 & \alpha \end{bmatrix} \begin{bmatrix} q_1[n] \\ q_2[n] \end{bmatrix} + \begin{bmatrix} \epsilon \\ \sigma \end{bmatrix} x[n] \\ &= \mathbf{A}\mathbf{q}[n] + \mathbf{b}x[n] \; . \end{aligned} \tag{6.40}$$

(A model of this form is also obtained for other systems of interest, for instance the motion of a DC motor whose input is a voltage that is held constant over intervals of length T by a computer-based controller. In that case, for $x[n]$ in appropriate units, we have $\alpha = 1$, $\sigma = T$, and $\epsilon = T^2/2$.)

For the purposes of this example, take

$$\mathbf{A} = \begin{bmatrix} 1 & \frac{1}{4} \\ 0 & 1 \end{bmatrix} \; , \quad \mathbf{b} = \begin{bmatrix} \frac{1}{32} \\ \frac{1}{4} \end{bmatrix} \tag{6.41}$$

and set

$$x[n] = g_1 q_1[n] + g_2 q_2[n] \tag{6.42}$$

to get the closed-loop matrix

$$\mathbf{A} + \mathbf{b}\mathbf{g}^T = \begin{bmatrix} 1 + \frac{g_1}{32} & \frac{1}{4} + \frac{g_2}{32} \\ \frac{g_1}{4} & 1 + \frac{g_2}{4} \end{bmatrix} \; . \tag{6.43}$$

The fastest possible closed-loop response in this DT model is the deadbeat behavior described earlier in Example 6.1, obtained by placing both closed-loop natural frequencies at 0, i.e., choosing the closed-loop characteristic polynomial to be $\nu(\lambda) = \lambda^2$. A little bit of algebra shows that g_1 and g_2 need to satisfy the following equations for this to be achieved:

$$\begin{aligned} \frac{g_1}{32} + \frac{g_2}{4} &= -2 \\ -\frac{g_1}{32} + \frac{g_2}{4} &= -1 \; . \end{aligned} \tag{6.44}$$

Solving these simultaneously, we get $g_1 = -16$ and $g_2 = -6$. We have not shown how to analyze system behavior when there are repeated eigenvalues, but in the particular instance of repeated eigenvalues at 0, it is easy to show that the state will die to 0 in a finite number of steps — at most two steps, for this second-order system. To establish this, note that with the above choice of $\mathbf{g}$ we get

$$\mathbf{A} + \mathbf{b}\mathbf{g}^T = \begin{bmatrix} \frac{1}{2} & \frac{1}{16} \\ -4 & -\frac{1}{2} \end{bmatrix} , \tag{6.45}$$

so

$$\left(\mathbf{A} + \mathbf{b}\mathbf{g}^T\right)^2 = 0 , \tag{6.46}$$

which shows that any nonzero initial condition will vanish in two steps. In practice, such deadbeat behavior may not be attainable, as unduly large control effort — rudder angles, in the case of the ship — would be needed. One is likely therefore to aim for slower decay of the error.

Typically, we do not have direct measurements of the state variables, only knowledge of the control input, along with noisy measurements of the system output. The state may then be reconstructed using an observer that produces asymptotically convergent estimates of the state variables, under the assumption that the system (6.20), (6.21) is observable. We shall see in more detail shortly that one can do quite well using the state estimates produced by the observer, in place of direct state measurements, in a feedback control scheme.

6.4.1 Proof of Eigenvalue Placement Results

This subsection presents the proof of the main result claimed earlier for state feedback, namely that it can yield any (monic, real-coefficient) closed-loop characteristic polynomial $\nu(\lambda)$ that includes among its roots all the unreachable eigenvalues of the original system. We shall also demonstrate that the closed-loop transfer function is given by the expression in (6.31).

First transform the open-loop system (6.20), (6.21) to modal coordinates; this changes nothing essential in the system, but simplifies the derivation. Using the same notation for modal coordinates as in the previous chapter, the closed-loop system is now defined by the equations

$$r_i[n+1] = \lambda_i r_i[n] + \beta_i x[n] , \quad i = 1, 2, \ldots, L \tag{6.47}$$

$$x[n] = \gamma_1 r_1[n] + \cdots + \gamma_L r_L[n] + p[n] , \tag{6.48}$$

where

$$\begin{pmatrix} \gamma_1 & \cdots & \gamma_L \end{pmatrix} = \mathbf{g}^T\mathbf{V} , \tag{6.49}$$

and $\mathbf{V}$ is the modal matrix, whose columns are the eigenvectors of the open-loop system. The γ_i are therefore just the state-feedback gains in modal coordinates.

Now using (6.47) and (6.48) to evaluate the transfer function from $p[n]$ to $x[n]$, we get

$$\frac{X(z)}{P(z)} = \Big(1 - \sum_{1}^{L} \frac{\gamma_i \beta_i}{z - \lambda_i}\Big)^{-1} = \frac{a(z)}{\nu(z)} \,. \tag{6.50}$$

To obtain the second equality in the above equation, we have used the following facts: (ii) the open-loop characteristic polynomial $a(z)$ is given by (6.22), and this is what appears in the numerator of (6.50; (ii) the poles of this transfer function must be the closed-loop poles of the system, and its denominator degree must equal its numerator degree, so the denominator of this expression must be the closed-loop characteristic polynomial $\nu(z)$. Then using (6.24), we find that the overall transfer function from the input $p[n]$ of the closed-loop system to the output $y[n]$ is

$$\frac{Y(z)}{P(z)} = \frac{Y(z)}{X(z)} \frac{X(z)}{P(z)} \tag{6.51}$$

$$= \frac{\eta(z)}{a(z)} \frac{a(z)}{\nu(z)} \tag{6.52}$$

$$= \frac{\eta(z)}{\nu(z)} \,. \tag{6.53}$$

The conclusion from all this is that state feedback has changed the denominator of the input-output transfer function expression from $a(z)$ in the open-loop case to $\nu(z)$ in the closed-loop case, and has accordingly modified the characteristic polynomial and poles. State feedback has left unchanged the numerator polynomial $\eta(z)$ from which the zeros are selected; all roots of $\eta(z)$ that are not canceled by roots of $\nu(z)$ will appear as zeros of the closed-loop transfer function.

Inverting (6.50), we find

$$\frac{\nu(z)}{a(z)} = 1 - \sum_{1}^{L} \frac{\gamma_i \beta_i}{z - \lambda_i} \,. \tag{6.54}$$

Hence, given the desired closed-loop characteristic polynomial $\nu(\lambda)$, we can expand $\nu(z)/a(z)$ in a partial fraction expansion, and determine the state feedback gain γ_i (in modal coordinates) for each i by dividing the coefficient of $1/(z - \lambda_i)$ by $-\beta_i$, assuming this is nonzero, i.e., assuming the ith mode is reachable. If the jth mode is unreachable, so $\beta_j = 0$, then λ_j does not appear as a pole on the right side of (6.54), which must mean that $\nu(z)$ has to contain $z - \lambda_j$ as a factor (in order for this factor to cancel out on the left side of the equation), i.e., every unreachable natural frequency of the open-loop system has to remain as a natural frequency of the closed-loop system.

6.5 OBSERVER-BASED FEEDBACK CONTROL

The obstacle to state feedback is the general unavailability of direct measurements of the state. All we typically have are knowledge of what control signal $x[n]$ we are applying, along with (possibly noise-corrupted) measurements of the output $y[n]$, and a nominal model of the system. We have already seen how to use this

information to estimate the state variables, using an observer or state estimator. Let us therefore consider what happens when we use the state estimate provided by the observer, rather than the (unavailable) actual state, in the feedback control law (6.25). With this substitution, (6.25) is modified to

$$\begin{aligned} x[n] &= p[n] + \mathbf{g}^T \widehat{\mathbf{q}}[n] \\ &= p[n] + \mathbf{g}^T (\mathbf{q}[n] - \widetilde{\mathbf{q}}[n]) \; . \end{aligned} \tag{6.55}$$

The overall closed-loop system is then as shown in Figure 6.5, and is governed by the following state-space model, obtained by combining the representations of the subsystems that make up the overall system, namely the plant (6.1), observer error dynamics (6.9), and feedback control law (6.55):

$$\begin{bmatrix} \mathbf{q}[n+1] \\ \widetilde{\mathbf{q}}[n+1] \end{bmatrix} = \begin{bmatrix} \mathbf{A} + \mathbf{b}\mathbf{g}^T & -\mathbf{b}\mathbf{g}^T \\ 0 & \mathbf{A} + \ell \mathbf{c}^T \end{bmatrix} \begin{bmatrix} \mathbf{q}[n] \\ \widetilde{\mathbf{q}}[n] \end{bmatrix} + \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix} p[n] + \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} \mathbf{w}[n] + \begin{bmatrix} 0 \\ \ell \end{bmatrix} \zeta[n] \; . \tag{6.56}$$

Note that we have reverted here to the more elaborate plant representation in (6.1), (6.2) rather than the streamlined one in (6.20), (6.21), in order to display the effect of plant disturbance and measurement error on the overall closed-loop system. (Instead of choosing the state vector of the overall system to comprise the state vector $\mathbf{q}[n]$ of the plant and the state vector $\widetilde{\mathbf{q}}[n]$ of the error equation, we could equivalently have picked $\mathbf{q}[n]$ and $\widehat{\mathbf{q}}[n]$. The former choice leads to more transparent expressions.)

The (block) triangular structure of the state matrix in (6.56) allows us to conclude that the natural frequencies of the overall system are simply the eigenvalues of $\mathbf{A} + \mathbf{b}\mathbf{g}^T$ along with those of $\mathbf{A} + \ell \mathbf{c}^T$. (This is not hard to demonstrate, either based on the definition of eigenvalues and eigenvectors, or using properties of determinants, but we omit the details.) In other words, our observer-based feedback control law results in a nicely behaved closed-loop system, with natural frequencies that are the union of those obtained with perfect state feedback and those obtained for the observer error equation. Both sets of natural frequencies can be arbitrarily selected, provided the open-loop system is reachable and observable. One would normally pick the modes that govern observer error decay to be faster than those associated with state feedback, in order to have reasonably accurate estimates available to the feedback control law before the plant state can wander too far away from what is desired.

The other interesting fact is that the transfer function from $p[n]$ to $y[n]$ in the new closed-loop system is exactly what would be obtained with perfect state feedback, namely the transfer function in (6.46). The reason is that the condition under which the transfer function is computed — as the input-output response when starting from the zero state — ensures that the observer starts up from the same initial condition as the plant. This in turn ensures that there is no estimation error, so the estimated state is as good as the true state. Another way to see this is to note that the observer error modes are unobservable from the available measurements.

The preceding observer-based compensator is the starting point for a very general and powerful approach to control design, one that carries over to the multi-input,

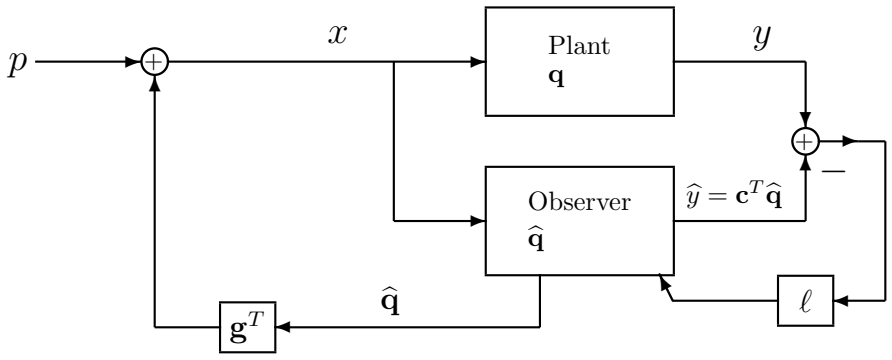

FIGURE 6.5 Observer-based compensator, feeding back an LTI combination of the estimated state variables.

multi-output case. With the appropriate embellishments around this basic structure, one can obtain every possible stabilizing LTI feedback controller for the system (6.20), (6.21). Within this class of controllers, we can search for those that have good robustness properties, in the sense that they are relatively immune to the uncertainties in our models. Further exploration of all this has to be left to more advanced courses.

CHAPTER 7

Probabilistic Models

INTRODUCTION

In the preceding chapters our emphasis has been on deterministic signals. In the remainder of this text we expand the class of signals considered to include those that are based on probabilistic models, referred to as random or stochastic processes. In introducing this important class of signals, we begin in this chapter with a review of the basics of probability and random variables. We assume that you have encountered this foundational material in a previous course, but include a review here for convenient reference and to establish notation. In the following chapter and beyond, we apply these concepts to define and discuss the class of random signals.

7.1 THE BASIC PROBABILITY MODEL

Associated with a basic probability model are the following three components, as indicated in Figure 7.1:

1. **Sample Space** The sample space Ψ is the set of all possible outcomes ψ of the probabilistic experiment that the model represents. We require that one and only one outcome be produced in each experiment with the model.
2. **Event Algebra** An event algebra is a collection of subsets of the sample space — referred to as events in the sample space — chosen such that unions of events and complements of events are themselves events (i.e., are in the collection of subsets). We say that a particular event has occurred if the outcome of the experiment lies in this event subset; thus Ψ is the "certain event" because it always occurs, and the empty set $\emptyset$ is the "impossible event" because it never occurs. Note that intersections of events are also events, because intersections can be expressed in terms of unions and complements.
3. **Probability Measure** A probability measure associates with each event A a number $P(A)$, termed the probability of A, in such a way that:
   - **(a)** $P(A) \geq 0$;
   - **(b)** $P(\Psi) = 1$;
   - **(c)** If $A \cap B = \emptyset$, i.e., if events A and B are mutually exclusive, then

$$P(A \cup B) = P(A) + P(B) .$$

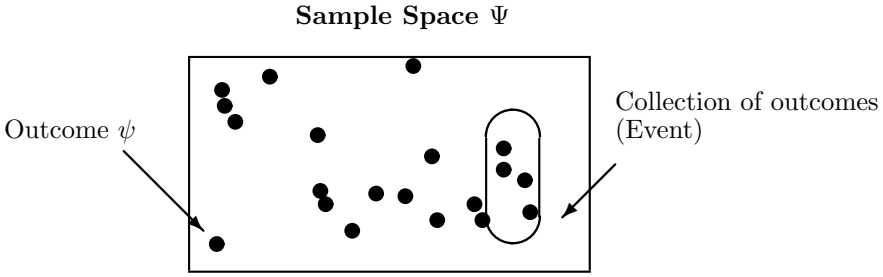

FIGURE 7.1 Sample space and events.

Note that for any particular case we often have a range of options in specifying what constitutes an outcome, in defining an event algebra, and in assigning a probability measure. It is generally convenient to have as few elements or outcomes as possible in a sample space, but we need enough of them to enable specification of the events of interest to us. It is typically convenient to pick the smallest event algebra that contains the events of interest. We also require that there be an assignment of probabilities to events that is consistent with the above conditions. This assignment may be made on the basis of symmetry arguments or in some other way that is suggested by the particular application.

7.2 CONDITIONAL PROBABILITY, BAYES' RULE, AND INDEPENDENCE

The probability of event A, given that event B has occurred, is denoted by $P(A|B)$. Knowing that B has occurred in effect reduces the sample space to the outcomes in B, so a natural definition of the conditional probability is

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)} \text{ if } P(B) > 0 \, . \tag{7.1}$$

It is straightforward to verify that this definition of conditional probability yields a valid probability measure on the sample space B. The preceding equation can also be rearranged to the form

$$P(A \cap B) = P(A|B)P(B) \, . \tag{7.2}$$

We often write $P(AB)$ or $P(A, B)$ for the joint probability $P(A \cap B)$. If $P(B) = 0$, then the conditional probability in (7.1) is undefined.

By symmetry, we can also write

$$P(A \cap B) = P(B|A)P(A) \tag{7.3}$$

Combining the preceding two equations, we obtain one form of Bayes' rule (or theorem), which is at the heart of much of what we'll do with signal detection,

classification, and estimation:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{7.4}$$

A more detailed form of Bayes' rule can be written for the conditional probability of one of a set of events $\{B_j\}$ that are mutually exclusive and collectively exhaustive, i.e. $B_\ell \cap B_m = \emptyset$ if $\ell \neq m$, and $\bigcup_j B_j = \Psi$. In this case,

$$P(A) = \sum_j P(A \cap B_j) = \sum_j P(A|B_j)P(B_j) \tag{7.5}$$

so that

$$P(B_\ell|A) = \frac{P(A|B_\ell)P(B_\ell)}{\sum_j P(A|B_j)P(B_j)} \tag{7.6}$$

Events A and B are said to be independent if

$$P(A|B) = P(A) \tag{7.7}$$

or equivalently if the joint probability factors as

$$P(A \cap B) = P(A)P(B)\ . \tag{7.8}$$

More generally, a collection of events is said to be mutually independent if the probability of the intersection of events from this collection, taken any number at a time, is always the product of the individual probabilities. Note that pairwise independence is not enough. Also, two sets of events $\mathcal{A}$ and $\mathcal{B}$ are said to be independent of each other if the probability of an intersection of events taken from these two sets always factors into the product of the joint probability of those events that are in $\mathcal{A}$ and the joint probability of those events that are in $\mathcal{B}$.

EXAMPLE 7.1 Transmission errors in a communication system

A communication system transmits symbols labeled A, B, and C. Because of errors (noise) introduced by the channel, there is a nonzero probability that for each transmitted symbol, the received symbol differs from the transmitted one. Table 7.1 describes the joint probability for each possible pair of transmitted and received symbols under a certain set of system conditions.

	Symbol received		
Symbol sent	A	B	C
A	0.05	0.10	0.09
B	0.13	0.08	0.21
C	0.12	0.07	0.15

TABLE 7.1 Joint probability for each possible pair of transmitted and received symbols

For notational convenience let's use A_s, B_s, C_s to denote the events that A, B or C respectively is sent, and A_r, B_r, C_r to denote A, B or C respectively being received. So, for example, $P(A_r, B_s) = 0.13$ and $P(C_r, C_s) = 0.15$. To determine the marginal probability $P(A_r)$, we sum the probabilities for all the mutually exclusive ways that A is received. So, for example,

$$\begin{aligned} P(A_r) &= P(A_r, A_s) + P(A_r, B_s) + P(A_r, C_s) \\ &= .05 + .13 + .12 = 0.3 \, . \end{aligned} \tag{7.9}$$

Similarly we can determine the marginal probability $P(A_s)$ as

$$P(A_s) = P(A_r, A_s) + P(B_r, A_s) + P(C_r, A_s) = 0.24 \tag{7.10}$$

In a communication context, it may be important to know the probability, for example, that C was sent, given that B was received, i.e., $P(C_s|B_r)$. That information is not entered directly in the table but can be calculated from it using Bayes' rule. Specifically, the desired conditional probability can be expressed as

$$P(C_s|B_r) = \frac{P(C_s, B_r)}{P(B_r)} \tag{7.11}$$

The numerator in (7.11) is given directly in the table as .07. The denominator is calculated as $P(B_r) = P(B_r, A_s) + P(B_r, B_s) + P(B_r, C_s) = 0.25$. The result then is that $P(C_s|B_r) = 0.28$.

In communication systems it is also often of interest to measure or calculate the probability of a transmission error. Denoting this by P_t it would correspond to any of the following mutually exclusive events happening:

$$(A_s \cap B_r), (A_s \cap C_r), (B_s \cap A_r), (B_s \cap C_r), (C_s \cap A_r), (C_s \cap B_r) \tag{7.12}$$

P_t is therefore the sum of the probabilities of these six mutually exclusive events, and all these probabilities can be read directly from the table in the off-diagonal locations, yielding $P_t = 0.72$.

7.3 RANDOM VARIABLES

A real-valued random variable $X(\,\cdot\,)$ is a function that maps each outcome ψ of a probabilistic experiment to a real number $X(\psi)$, which is termed the *realization* of (or value taken by) the random variable in that experiment. An additional technical requirement imposed on this function is that the set of outcomes $\{\psi\}$ that maps to the interval $X \leq x$ must be an event in Ψ, for all real numbers x. We shall typically just write the random variable as X instead of $X(\,\cdot\,)$ or $X(\psi)$.

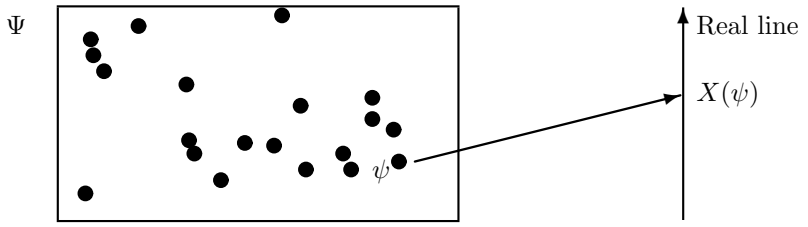

FIGURE 7.2 A random variable.

It is often also convenient to consider random variables taking values that are not specified as real numbers but rather a finite or countable set of labels, say $L_0, L_1, L_2, \ldots$. For instance, the random status of a machine may be tracked using the labels Idle, Busy, and Failed. Similarly, the random presence of a target in a radar scan can be tracked using the labels Absent and Present. We can think of these labels as comprising a set of mutually exclusive and collectively exhaustive events, where each such event comprises all the outcomes that carry that label. We refer to such random variables as random events, mapping each outcome ψ of a probabilistic experiment to the label $L(\psi)$, chosen from the possible values $L_0, L_1, L_2, \ldots$. We shall typically just write L instead of $L(\psi)$.

7.4 CUMULATIVE DISTRIBUTION, PROBABILITY DENSITY, AND PROBABILITY MASS FUNCTION FOR RANDOM VARIABLES

Cumulative Distribution Functions For a (real-valued) random variable X, the probability of the event comprising all ψ for which $X(\psi) \leq x$ is described using the cumulative distribution function (CDF) $F_X(x)$:

$$F_X(x) = P(X \leq x) \, . \tag{7.13}$$

We can therefore write

$$P(a < X \leq b) = F_X(b) - F_X(a) \, . \tag{7.14}$$

In particular, if there is a nonzero probability that X takes a specific value x_1, i.e. if $P(X = x_1) > 0$, then $F_X(x)$ will have a jump at x_1 of height $P(X = x_1)$, and $F_X(x_1) - F_X(x_1-) = P(X = x_1)$. The CDF is nondecreasing as a function of x; it starts from $F_X(-\infty) = 0$ and rises to $F_X(\infty) = 1$.

A related function is the conditional CDF $F_{X|L}(x|L_i)$, used to describe the distribution of X conditioned on some random event L taking the specific value L_i, and assuming $P(L = L_i) > 0$:

$$F_{X|L}(x|L_i) = P(X \leq x|L = L_i) = \frac{P(X \leq x, L = L_i)}{P(L = L_i)} \, . \tag{7.15}$$

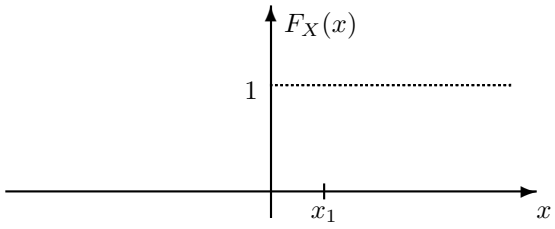

FIGURE 7.3 Example of a CDF.

Probability Density Functions The probability density function (PDF) $f_X(x)$ of the random variable X is the derivative of $F_X(x)$:

$$f_X(x) = \frac{dF_X(x)}{dx} \ . \tag{7.16}$$

It is of course always non-negative because $F_X(x)$ is nondecreasing. At points of discontinuity in $F_X(x)$, corresponding to values of x that have non-zero probability of occurring, there will be (Dirac) impulses in $f_X(x)$, of strength or area equal to the height of the discontinuity. We can write

$$P(a < X \leq b) = \int_a^b f_X(x)\, dx \ . \tag{7.17}$$

(Any impulse of $f_X(x)$ at b would be included in the integral, while any impulse at a would be left out — i.e. the integral actually goes from $a+$ to $b+$.) We can heuristically think of $f_X(x)\, dx$ as giving the probability that X lies in the interval $(x - dx, x]$:

$$P(x - dx < X \leq x) \approx f_X(x)\, dx \ . \tag{7.18}$$

Note that at values of x where $f_X(x)$ does not have an impulse, the probability of X having the value x is zero, i.e., $P(X = x) = 0$.

A related function is the conditional PDF $f_{X|L}(x|L_i)$, defined as the derivative of $F_{X|L}(x|L_i)$ with respect to x.

Probability Mass Function A real-valued discrete random variable X is one that takes only a finite or countable set of real values, $\{x_1, x_2, \cdots\}$. (Hence this is actually a random event — as defined earlier — but specified numerically rather than via labels.) The CDF in this case would be a "staircase" function, while the PDF would be zero everywhere, except for impulses at the x_j, with strengths corresponding to the respective probabilities of the x_j. These strengths/probabilities are conveniently described by the probability mass function (PMF) $p_X(x)$, which gives the probability of the event $X = x_j$:

$$P(X = x_j) = p_X(x_j) \ . \tag{7.19}$$

7.5 JOINTLY DISTRIBUTED RANDOM VARIABLES

We almost always use models involving multiple (or compound) random variables. Such situations are described by joint probabilities. For example, the joint CDF of two random variables X and Y is

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) \; . \tag{7.20}$$

The corresponding joint PDF is

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \, \partial y} \tag{7.21}$$

and has the heuristic interpretation that

$$P(x - dx < X \leq x \; , y - dy < Y \leq y) \approx f_{X,Y}(x,y) \, dx \, dy \; . \tag{7.22}$$

The marginal PDF $f_X(x)$ is defined as the PDF of the random variable X considered on its own, and is related to the joint density $f_{X,Y}(x,y)$ by

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) \, dy \; . \tag{7.23}$$

A similar expression holds for the marginal PDF $f_Y(y)$.

We have already noted that when the model involves a random variable X and a random event L, we may work with the conditional CDF

$$F_{X|L}(x|L_i) = P(X \leq x | L = L_i) = \frac{P(X \leq x, L = L_i)}{P(L = L_i)} \; , \tag{7.24}$$

provided $P(L = L_i) > 0$. The derivative of this function with respect to x gives the conditional PDF $f_{X|L}(x|L_i)$. When the model involves two continuous random variables X and Y, the corresponding function of interest is the conditional PDF $f_{X|Y}(x|y)$ that describes the distribution of X, given that $Y = y$. However, for a continuous random variable Y, $P(Y = y) = 0$, so even though the following definition may seem natural, its justification is more subtle:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \; . \tag{7.25}$$

To see the plausibility of this definition, note that the conditional PDF $f_{X|Y}(x|y)$ must have the property that

$$f_{X|Y}(x|y) \, dx \approx P(x - dx < X \leq x \mid y - dy < Y \leq y) \tag{7.26}$$

but by Bayes' rule the quantity on the right in the above equation can be rewritten as

$$P(x - dx < X \leq x \mid y - dy < Y \leq y) \approx \frac{f_{X,Y}(x,y) \, dx \, dy}{f_Y(y) dy} \; . \tag{7.27}$$

Combining the latter two expressions yields the definition of $f_{X|Y}(x|y)$ given in (7.25).

Using similar reasoning, we can obtain relationships such as the following:

$$P(L = L_i|X = x) = \frac{f_{X|L}(x|L_i)P(L = L_i)}{f_X(x)} \ . \tag{7.28}$$

Two random variables X and Y are said to be independent or statistically independent if their joint PDF (or equivalently their joint CDF) factors into the product of the individual ones:

$$\begin{array}{rcll} f_{X,Y}(x,y) & = & f_X(x)f_Y(y) \ , & \text{or} \\ F_{X,Y}(x,y) & = & F_X(x)F_Y(y) \ . & \end{array} \tag{7.29}$$

This condition turns out to be equivalent to having any collection of events defined in terms of X be independent of any collection of events defined in terms of Y.

For a set of more than two random variables to be independent, we require that the joint PDF (or CDF) of random variables from this set factors into the product of the individual PDFs (respectively, CDFs). One can similarly define independence of random variables and random events.

EXAMPLE 7.2 Independence of events

To illustrate some of the above definitions and concepts in the context of random variables and random events, consider two independent random variables X and Y for which the marginal PDFs are uniform between zero and one:

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

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

Because X and Y are independent, the joint PDF $f_{X,Y}(x,y)$ is given by

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

We define the events A, B, C and D as follows:

$$A = \left\{y > \frac{1}{2}\right\}, \quad B = \left\{y < \frac{1}{2}\right\}, \quad C = \left\{x < \frac{1}{2}\right\},$$
$$D = \left\{x < \frac{1}{2} \text{ and } y < \frac{1}{2}\right\} \cup \left\{x > \frac{1}{2} \text{ and } y > \frac{1}{2}\right\}.$$

These events are illustrated pictorially in Figure 7.4

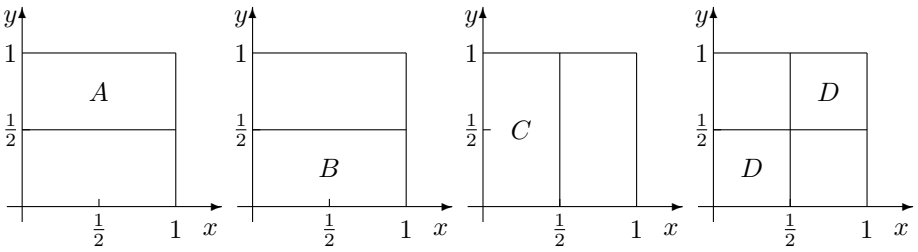

FIGURE 7.4 Illustration of events A, B, C, and D, for Example 7.2

Questions that we might ask include whether these events are pairwise independent, e.g. whether A and C are independent. To answer such questions, we consider whether the joint probability factors into the product of the individual probabilities. So, for example,

$$P(A \cap C) = P\left(y > \frac{1}{2}, x < \frac{1}{2}\right) = \frac{1}{4}$$

$$P(A) = P(C) = \frac{1}{2}$$

Since $P(A \cap C) = P(A)P(C)$, events A and C are independent. However,

$$P(A \cap B) = P\left(y > \frac{1}{2}, y < \frac{1}{2}\right) = 0$$

$$P(A) = P(B) = \frac{1}{2}$$

Since $P(A \cap B) \neq P(A)P(B)$, events A and B are not independent.

Note that $P(A \cap C \cap D) = 0$ since there is no region where all three sets overlap. However, $P(A) = P(C) = P(D) = \frac{1}{2}$, so $P(A \cap C \cap D) \neq P(A)P(C)P(D)$ and the events A, C, and D are not mutually independent, even though they are easily seen to be pairwise independent. For a collection of events to be independent, we require the probability of the intersection of any of the events to equal the product of the probabilities of each individual event. So for the 3–event case, pairwise independence is a necessary but not sufficient condition for independence.

7.6 EXPECTATIONS, MOMENTS AND VARIANCE

For many purposes it suffices to have a more aggregated or approximate description than the PDF provides. The expectation — also termed the expected or mean or average value, or the first-moment — of the real-valued random variable X is

denoted by $E[X]$ or $\overline{X}$ or μ_X, and defined as

$$E[X] = \overline{X} = \mu_X = \int_{-\infty}^{\infty} x f_X(x)\, dx \ . \tag{7.30}$$

In terms of the probability "mass" on the real line, the expectation gives the location of the center of mass. Note that the expected value of a sum of random variables is just the sum of the individual expected values:

$$E[X+Y] = E[X] + E[Y] \ . \tag{7.31}$$

Other simple measures of where the PDF is centered or concentrated are provided by the median, which is the value of x for which $F_X(x) = 0.5$, and by the mode, which is the value of x for which $f_X(x)$ is maximum (in degenerate cases one or both of these may not be unique).

The variance or centered second-moment of the random variable X is denoted by σ_X^2 and defined as

$$\begin{aligned} \sigma_X^2 &= E[(X-\mu_X)^2] = \text{expected squared deviation from the mean} \\ &= \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x) dx \\ &= E[X^2] - \mu_X^2 \ , \end{aligned} \tag{7.32}$$

where the last equation follows on writing $(X-\mu_X)^2 = X^2 - 2\mu_X X + \mu_X^2$ and taking the expectation term by term. We refer to $E[X^2]$ as the second-moment of X. The square root of the variance, termed the standard deviation, is a widely used measure of the spread of the PDF.

The focus of many engineering models that involve random variables is primarily on the means and variances of the random variables. In some cases this is because the detailed PDFs are hard to determine or represent or work with. In other cases, the reason for this focus is that the means and variances completely determine the PDFs, as with the Gaussian (or normal) and uniform PDFs.

EXAMPLE 7.3 Gaussian and uniform random variables

Two common PDF's that we will work with are the Gaussian (or normal) density and the uniform density:

$$\begin{aligned} &\text{Gaussian:} \quad f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2} \\ &\text{Uniform:} \quad f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{7.33}$$

The two parameters m and σ that define the Gaussian PDF can be shown to be its mean and standard deviation respectively. Similarly, though the uniform density can be simply parametrized by its lower and upper limits a and b as above, an

equivalent parametrization is via its mean $m = (a+b)/2$ and standard deviation $\sigma = \sqrt{(b-a)^2/12}$.

There are useful statements that can be made for general PDFs on the basis of just the mean and variance. The most familiar of these is the Chebyshev inequality:

$$P\Big(\frac{|X-\mu_X|}{\sigma_X} \geq k\Big) \leq \frac{1}{k^2} \, . \tag{7.34}$$

This inequality implies that, for any random variable, the probability it lies at or more than 3 standard deviations away from the mean (on either side of the mean) is not greater than $(1/3^2) = 0.11$. Of course, for particular PDFs, much more precise statements can be made, and conclusions derived from the Chebyshev inequality can be very conservative. For instance, in the case of a Gaussian PDF, the probability of being more than 3 standard deviations away from the mean is only 0.0026, while for a uniform PDF the probability of being more than even 2 standard deviations away from the mean is precisely 0.

For much of our discussion we shall make do with evaluating the means and variances of the random variables involved in our models. Also, we will be highlighting problems whose solution only requires knowledge of means and variances.

The conditional expectation of the random variable X, given that the random variable Y takes the value y, is the real number

$$E[X|Y=y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)dx = g(y) \, , \tag{7.35}$$

i.e., this conditional expectation takes some value $g(y)$ when $Y = y$. We may also consider the random variable $g(Y)$, namely the function of the random variable Y that, for each $Y = y$, evaluates to the conditional expectation $E[X|Y=y]$. We refer to this random variable $g(Y)$ as the conditional expectation of X "given Y" (as opposed to "given $Y = y$"), and denote $g(Y)$ by $E[X|Y]$. Note that the expectation $E[g(Y)]$ of the random variable $g(Y)$, i.e. the iterated expectation $E[E[X|Y]]$, is well defined. What we show in the next paragraph is that this iterated expectation works out to something simple, namely $E[X]$. This result will be of particular use in the next chapter.

Consider first how to compute $E[X]$ when we have the joint PDF $f_{X,Y}(x,y)$. One way is to evaluate the marginal density $f_X(x)$ of X, and then use the definition of expectation in (7.30):

$$E[X] = \int_{-\infty}^{\infty} x\Big(\int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy\Big)\, dx \, . \tag{7.36}$$

However, it is often simpler to compute the conditional expectation of X, given $Y = y$, then average this conditional expectation over the possible values of Y, using the marginal density of Y. To derive this more precisely, recall that

$$f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y) \tag{7.37}$$

and use this in (7.36) to deduce that

$$E[X] = \int_{-\infty}^{\infty} f_Y(y)\Big(\int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,dx\Big)\,dy = E_Y[E_{X|Y}[X|Y]]\ . \tag{7.38}$$

We have used subscripts on the preceding expectations in order to make explicit which densities are involved in computing each of them. More simply, one writes

$$E[X] = E[E[X|Y]]\ . \tag{7.39}$$

The preceding result has an important implication for the computation of the expectation of a function of a random variable. Suppose $X = h(Y)$, then $E[X|Y] = h(Y)$, so

$$E[X] = E[E[X|Y]] = \int_{-\infty}^{\infty} h(y) f_Y(y) dy\ . \tag{7.40}$$

This shows that we only need $f_Y(y)$ to calculate the expectation of a function of Y; to compute the expectation of $X = h(Y)$, we do not need to determine $f_X(x)$.

Similarly, if X is a function of *two* random variables, $X = h(Y, Z)$, then

$$E[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(y,z) f_{Y,Z}(y,z) dy\,dz\ . \tag{7.41}$$

It is easy to show from this that if Y and Z are independent, and if $h(y,z) = g(y)\ell(z)$, then

$$E[g(Y)\ell(Z)] = E[g(Y)]E[\ell(Z)]\ . \tag{7.42}$$

7.7 CORRELATION AND COVARIANCE FOR BIVARIATE RANDOM VARIABLES

Consider a pair of jointly distributed random variables X and Y. Their marginal PDFs are simply obtained by projecting the probability mass along the y-axis and x-axis directions respectively:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\ , \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\ . \tag{7.43}$$

In other words, the PDF of X is obtained by integrating the joint PDF over all possible values of the other random variable Y — and similarly for the PDF of Y.

It is of interest, just as in the single-variable case, to be able to capture the location and spread of the bivariate PDF in some aggregate or approximate way, without having to describe the full PDF. And again we turn to notions of mean and variance. The mean value of the bivariate PDF is specified by giving the mean values of each of its two component random variables: the mean value has an x component that is $E[X]$, and a y component that is $E[Y]$, and these two numbers can be evaluated from the respective marginal densities. The center of mass of the bivariate PDF is thus located at

$$(x, y) = (E[X], E[Y])\ . \tag{7.44}$$

A measure of the spread of the bivariate PDF in the x direction may be obtained from the standard deviation σ_X of X, computed from $f_X(x)$; and a measure of the spread in the y direction may be obtained from σ_Y, computed similarly from $f_Y(y)$. However, these two numbers clearly only offer a partial view. We would really like to know what the spread is in a general direction rather than just along the two coordinate axes. We can consider, for instance, the standard deviation (or, equivalently, the variance) of the random variable Z defined as

$$Z = \alpha X + \beta Y \tag{7.45}$$

for arbitrary constants α and β. Note that by choosing α and β appropriately, we get $Z = X$ or $Z = Y$, and therefore recover the special coordinate directions that we have already considered; but being able to analyze the behavior of Z for arbitary α and β allows us to specify the behavior in all directions.

To visualize how Z behaves, note that $Z = 0$ when $\alpha x+\beta y = 0$. This is the equation of a straight line through the origin in the (x, y) plane, a line that indicates the precise combinations of values x and y that contribute to determining $f_Z(0)$, by projection of $f_{X,Y}(x, y)$ along the line. Let us call this the reference line. If Z now takes a nonzero value z, the corresponding set of (x, y) values lies on a line offset from but parallel to the reference line. We project $f_{X,Y}(x, y)$ along this new offset line to determine $f_Z(z)$.

Before seeing what computations are involved in determining the variance of Z, note that the mean of Z is easily found in terms of quantities we have already computed, namely $E[X]$ and $E[Y]$:

$$E[Z] = \alpha E[X] + \beta E[Y] \, . \tag{7.46}$$

As for the variance of Z, it is easy to establish from (7.45) and (7.46) that

$$\sigma_Z^2 = E[Z^2] - (E[Z])^2 = \alpha^2\sigma_X^2 + \beta^2\sigma_Y^2 + 2\alpha\beta\,\sigma_{X,Y} \tag{7.47}$$

where σ_X^2 and σ_Y^2 are the variances already computed along the coordinate directions x and y, and $\sigma_{X,Y}$ is the covariance of X and Y, also denoted by $\mathrm{cov}(X, Y)$ or $C_{X,Y}$, and defined as

$$\sigma_{X,Y} = \mathrm{cov}(X, Y) = C_{X,Y} = E[(X - E[X])(Y - E[Y])] \tag{7.48}$$

or equivalently

$$\sigma_{X,Y} = E[XY] - E[X]E[Y] \, . \tag{7.49}$$

where (7.49) follows from multiplying out the terms in parentheses in (7.48) and then taking term-by-term expectations. Note that when $Y = X$ we recover the familiar expressions for the variance of X. The quantity $E[XY]$ that appears in (7.49), i.e., the expectation of the product of the random variables, is referred to as the correlation or second cross-moment of X and Y (to distinguish it from the second self-moments $E[X^2]$ and $E[Y^2]$), and will be denoted by $R_{X,Y}$:

$$R_{X,Y} = E[XY] \, . \tag{7.50}$$

It is reassuring to note from (7.47) that the covariance $\sigma_{X,Y}$ is the only new quantity needed when going from mean and spread computations along the coordinate axes to such computations along any axis; we do not need a new quantity for each new direction. In summary, we can express the location of $f_{X,Y}(x,y)$ in an aggregate or approximate way in terms of the 1st-moments, $E[X]\,, E[Y]$; and we can express the spread around this location in an aggregate or approximate way in terms of the (central) 2nd-moments, $\sigma_X^2\,, \sigma_Y^2\,, \sigma_{X,Y}$.

It is common to work with a normalized form of the covariance, namely the correlation coefficient $\rho_{X,Y}$:

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}\,. \tag{7.51}$$

This normalization ensures that the correlation coefficient is unchanged if X and/or Y is multiplied by any nonzero constant or has any constant added to it. For instance, the centered and normalized random variables

$$V = \frac{X - \mu_X}{\sigma_X}\,, \qquad W = \frac{Y - \mu_Y}{\sigma_Y}\,, \tag{7.52}$$

each of which has mean 0 and variance 1, have the same correlation coefficient as X and Y. The correlation coefficient might have been better called the covariance coefficient, since it is defined in terms of the covariance and not the correlation of the two random variables, but this more helpful name is not generally utilized.

Invoking the fact that σ_Z^2 in (7.47) must be non-negative, and further noting from this equation that σ_Z^2/β^2 is quadratic in α, it can be proved by elementary analysis of the quadratic expression that

$$|\rho_{X,Y}| \leq 1\,. \tag{7.53}$$

From the various preceding definitions, a positive correlation $R_{X,Y} > 0$ suggests that X and Y tend to take the same sign, on average, whereas a positive covariance $\sigma_{X,Y} > 0$ — or equivalently a positive correlation coefficient $\rho_{X,Y} > 0$ — suggests that the deviations of X and Y from their respective means tend to take the same sign, on average. Conversely, a negative correlation suggests that X and Y tend to take opposite signs, on average, while a negative covariance or correlation coefficient suggests that the deviations of X and Y from their means tend to take opposite signs, on average.

Since the correlation coefficient of X and Y captures some features of the relation between their deviations from their respective means, we might expect that the correlation coefficient can play a role in constructing an estimate of Y from measurements of X, or vice versa. We shall see in the next chapter, where linear minimum mean-square error (LMMSE) estimation is studied, that this is indeed the case.

The random variables X and Y are said to be uncorrelated (or linearly independent, a less common and potentially misleading term) if

$$E[XY] = E[X]E[Y]\,, \tag{7.54}$$

or equivalently if

$$\sigma_{X,Y} = 0 \quad \text{or} \quad \rho_{X,Y} = 0 \, . \tag{7.55}$$

Thus uncorrelated does not mean zero correlation (unless one of the random variables has an expected value of zero). Rather, uncorrelated means zero covariance. Again, a better term for uncorrelated might have been non-covariant, but this term is not widely used.

Note also that independent random variables X and Y, i.e., those for which

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \, , \tag{7.56}$$

are always uncorrelated, but the converse is not generally true: uncorrelated random variables may not be independent. If X and Y are independent, then $E[XY] = E[X]E[Y]$ so X and Y are uncorrelated. The converse does *not* hold in general. For instance, consider the case where the combination (X, Y) takes only the values $(1, 0)$, $(-1, 0)$. $(0, 1)$ and $(0, -1)$, each with equal probability $\frac{1}{4}$. Then X and Y are easily seen to be uncorrelated but dependent, i.e., not independent.

A final bit of terminology that we will shortly motivate and find useful occurs in the following definition: Two random variables X and Y are **orthogonal** if $E[XY] = 0$.

EXAMPLE 7.4 Perfect correlation, zero correlation

Consider the degenerate case where Y is given by a deterministic linear function of a random variable X (so Y is also a random variable, of course):

$$Y = \xi X + \zeta \, , \tag{7.57}$$

where ξ and ζ are constants. Then it is easy to show that $\rho_{X,Y} = 1$ if $\xi > 0$ and $\rho = -1$ if $\xi < 0$. Note that in this case the probability mass is entirely concentrated on the line defined by the above equation, so the bivariate PDF — if we insist on talking about it! — is a two-dimensional impulse (but this fact is not important in evaluating $\rho_{X,Y}$).

You should also have no difficulty establishing that $\rho_{X,Y} = 0$ if

$$Y = \xi X^2 + \zeta \tag{7.58}$$

and X has a PDF $f_X(x)$ that is even about 0, i.e., $f_X(-x) = f_X(x)$.

EXAMPLE 7.5 Bivariate Gaussian density

The random variables X and Y are said to be bivariate Gaussian or bivariate normal if their joint PDF is given by

$$f_{X,Y}(x,y) = c \, \exp\Big\{ -q\Big(\frac{x - \mu_X}{\sigma_X}, \frac{y - \mu_Y}{\sigma_Y} \Big) \Big\} \tag{7.59}$$

where c is a normalizing constant (so that the PDF integrates to 1) and $q(v,w)$ is a quadratic function of its two arguments v and w, expressed in terms of the correlation coefficient ρ of X and Y:

$$c = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \tag{7.60}$$

$$q(v,w) = \frac{1}{2(1-\rho^2)}(v^2 - 2\rho vw + w^2) \tag{7.61}$$

This density is the natural bivariate generalization of the familiar Gaussian density, and has several nice properties:

- The marginal densities of X and Y are Gaussian.
- The conditional density of Y, given $X = x$, is Gaussian with mean ρx and variance $\sigma_Y^2(1-\rho^2)$ (which evidently does not depend on the value of x); and similary for the conditional density of X, given $Y = y$.
- If X and Y are uncorrelated, i.e., if $\rho = 0$, then X and Y are actually independent, a fact that is not generally true for other bivariate random variables, as noted above.
- Any two affine (i.e., linear plus constant) combinations of X and Y are themselves bivariate Gaussian (e.g., $Q = X + 3Y + 2$ and $R = 7X + Y - 3$ are bivariate Gaussian).

The bivariate Gaussian PDF and indeed the associated notion of correlation were essentially discovered by the statistician Francis Galton (a first-cousin of Charles Darwin) in 1886, with help from the mathematician Hamilton Dickson. Galton was actually studying the joint distribution of the heights of parents and children, and found that the marginals and conditionals were well represented as Gaussians. His question to Dickson was: what joint PDF has Gaussian marginals and conditionals? The answer: the bivariate Gaussian! It turns out that there is a 2-dimensional version of the central limit theorem, with the bivariate Gaussian as the limiting density, so this is a reasonable model for two jointly distributed random variables in many settings. There are also natural generalization to many variables.

Some of the generalizations of the preceding discussion from two random variables to many random variables are fairly evident. In particular, the mean of a joint PDF

$$f_{X_1,X_2,\cdots,X_\ell}(x_1,x_2,\cdots,x_\ell) \tag{7.62}$$

in the ℓ-dimensional space of possible values has coordinates that are the respective individual means, $E[X_1],\cdots,E[X_\ell]$. The spreads in the coordinate directions are deduced from the individual (marginal) spreads, $\sigma_{X_1},\cdots,\sigma_{X_\ell}$. To be able to compute the spreads in *arbitrary* directions, we need all the additional $\ell(\ell-1)/2$ central 2nd moments, namely σ_{X_i,X_j} for all $1 \le i < j \le \ell$ (note that $\sigma_{X_j,X_i} = \sigma_{X_i,X_j}$) — but nothing more.

7.8 A VECTOR-SPACE PICTURE FOR CORRELATION PROPERTIES OF RANDOM VARIABLES

A vector-space picture is often useful as an aid to recalling the second-moment relationships between two random variables X and Y. This picture is not just a mnemonic: there is a very precise sense in which random variables can be thought of (or are) vectors in a vector space (of infinite dimensions), as long as we are only interested in their second-moment properties. Although we shall not develop this correspondence in any depth, it can be very helpful in conjecturing or checking answers in the linear minimum mean-square-error (LMMSE) estimation problems that we shall treat.

To develop this picture, we represent the random variables X and Y as vectors $\mathbf{X}$ and $\mathbf{Y}$ in some abstract vector space. For the squared lengths of these vectors, we take the second-moments of the associated random variables, $E[X^2]$ and $E[Y^2]$ respectively. Recall that in Euclidean vector space the squared length of a vector is the inner product of the vector with itself. This suggests that perhaps in our vector-space interpretation the inner product $< \mathbf{X}, \mathbf{Y} >$ between two general vectors $\mathbf{X}$ and $\mathbf{Y}$ should be defined as the correlation (or second cross-moment) of the associate random variables:

$$< \mathbf{X}, \mathbf{Y} >= E[XY] = R_{X,Y} \ . \tag{7.63}$$

This indeed turns out to be the definition that's needed. With this definition, the standard properties required of an inner product in a vector space are satisfied, namely:

Symmetry: $< \mathbf{X}, \mathbf{Y} >=< \mathbf{Y}, \mathbf{X} >$.

Linearity: $< \mathbf{X}, a_1\mathbf{Y_1} + a_2\mathbf{Y_2} >= a_1 < \mathbf{X}, \mathbf{Y_1} > + a_2 < \mathbf{X}, \mathbf{Y_2} >$

Positivity: $< \mathbf{X}, \mathbf{X} >$ is positive for $\mathbf{X} \neq \mathbf{0}$, and $\mathbf{0}$ otherwise.

This definition of inner product is also consistent with the fact that we often refer to two random variables as orthogonal when $E[XY] = 0$.

The centered random variables $X - \mu_X$ and $Y - \mu_Y$ can similary be represented as vectors $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ in this abstract vector space, with squared lengths that are now the variances of the random variables X and Y:

$$\sigma_X^2 = E[(X - \mu_X)^2] \ , \qquad \sigma_Y^2 = E[(Y - \mu_Y)^2] \tag{7.64}$$

respectively. The lengths are therefore the standard deviations of the associated random variables, σ_X and σ_Y respectively. The inner product of the vectors $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ becomes

$$< \widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}} >= E[(X - \mu_X)(Y - \mu_Y)] = \sigma_{X,Y} \ , \tag{7.65}$$

namely the covariance of the random variables.

In Euclidean space the inner product of two vectors is given by the product of the lengths of the individual vectors and the cosine of the angle between them:

$$< \widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}} >= \sigma_{X,Y} = \sigma_X \sigma_Y \cos(\theta) \ , \tag{7.66}$$

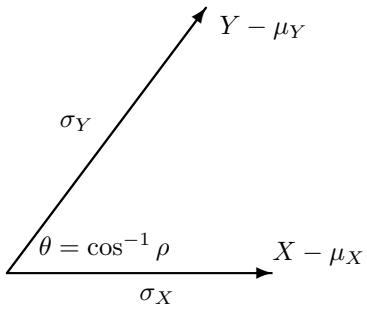

FIGURE 7.5 Random Variables as Vectors.

so the quantity

$$\theta = \cos^{-1}\Big(\frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}\Big) = \cos^{-1}\rho \tag{7.67}$$

can be thought of as the angle between the vectors. Here ρ is the correlation coefficient of the two random variables, so evidently

$$\rho = \cos(\theta)\ . \tag{7.68}$$

Thus, the correlation coefficient is the cosine of the angle between the vectors. It is therefore not surprising at all that

$$-1 \leq \rho \leq 1\ . \tag{7.69}$$

When ρ is near 1, the vectors are nearly aligned in the same direction, whereas when ρ is near -1 they are close to being oppositely aligned. The correlation coefficient is zero when these vectors $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ (which represent the centered random variables) are orthogonal, or equivalently, the corresponding random variables have zero covariance,

$$\sigma_{X,Y} = 0\ . \tag{7.70}$$

CHAPTER 8

Estimation with Minimum Mean Square Error

INTRODUCTION

A recurring theme in this text and in much of communication, control and signal processing is that of making systematic estimates, predictions or decisions about some set of quantities, based on information obtained from measurements of other quantities. This process is commonly referred to as inference. Typically, inferring the desired information from the measurements involves incorporating models that represent our prior knowledge or beliefs about how the measurements relate to the quantities of interest.

Inference about continuous random variables and ultimately about random processes is the topic of this chapter and several that follow. One key step is the introduction of an error criterion that measures, in a probabilistic sense, the error between the desired quantity and our estimate of it. Throughout our discussion in this and the related subsequent chapters, we focus primarily on choosing our estimate to minimize the expected or mean value of the square of the error, referred to as a minimum mean-square-error (MMSE) criterion. In Section 8.1 we consider the MMSE estimate without imposing any constraint on the form that the estimator takes. In Section 8.3 we restrict the estimate to be a linear combination of the measurements, a form of estimation that we refer to as linear minimum mean-square-error (LMMSE) estimation.

Later in the text we turn from inference problems for continuous random variables to inference problems for discrete random quantities, which may be numerically specified or may be non-numerical. In the latter case especially, the various possible outcomes associated with the random quantity are often termed hypotheses, and the inference task in this setting is then referred to as hypothesis testing, i.e., the task of deciding which hypothesis applies, given measurements or observations. The MMSE criterion may not be meaningful in such hypothesis testing problems, but we can for instance aim to minimize the probability of an incorrect inference regarding which hypothesis actually applies.

8.1 ESTIMATION OF A CONTINUOUS RANDOM VARIABLE

To begin the discussion, let us assume that we are interested in a random variable Y and we would like to estimate its value, knowing only its probability density function. We will then broaden the discussion to estimation when we have a measurement or observation of another random variable X, together with the joint probability density function of X and Y.

Based only on knowledge of the PDF of Y, we wish to obtain an estimate of Y — which we denote as $\widehat{y}$ — so as to minimize the mean square error between the actual outcome of the experiment and our estimate $\widehat{y}$. Specifically, we choose $\widehat{y}$ to minimize

$$E[(Y-\widehat{y})^2] = \int (y-\widehat{y})^2 f_Y(y)\,dy\ . \tag{8.1}$$

Differentiating (8.1) with respect to $\widehat{y}$ and equating the result to zero, we obtain

$$-2\int (y-\widehat{y}) f_Y(y)\,dy = 0 \tag{8.2}$$

or

$$\int \widehat{y} f_Y(y)\,dy = \int y f_Y(y)\,dy \tag{8.3}$$

from which

$$\widehat{y} = E[Y]\ . \tag{8.4}$$

The second derivative of $E[(Y-\widehat{y})^2]$ with respect to $\widehat{y}$ is

$$2\int f_Y(y)\,dy = 2\ , \tag{8.5}$$

which is positive, so (8.4) does indeed define the minimizing value of $\widehat{y}$. Hence the MMSE estimate of Y in this case is simply its mean value, $E[Y]$.

The associated error — the actual MMSE — is found by evaluating the expression in (8.1) with $\widehat{y} = E[Y]$. We conclude that the MMSE is just the variance of Y, namely σ_Y^2:

$$\min E[(Y-\widehat{y})^2] = E[(Y-E[Y])^2] = \sigma_Y^2\ . \tag{8.6}$$

In a similar manner, it is possible to show that the median of Y, which has half the probability mass of Y below it and the other half above, is the value of $\widehat{y}$ that minimizes the mean absolute deviation, $E[\,|Y-\widehat{y}|\,]$. Also, the mode of Y, which is the value of y at which the PDF $f_Y(y)$ is largest, turns out to minimize the expected value of an all-or-none cost function, i.e., a cost that is unity when the error is outside of a vanishingly small tolerance band, and is zero within the band. We will not be pursuing these alternative error metrics further, but it is important to be aware that our choice of mean square error, while convenient, is only one of many possible error metrics.

The insights from the simple problem leading to (8.4) and (8.6) carry over directly to the case in which we have additional information in the form of the measured or

observed value x of a random variable X that is related somehow to Y. The only change from the previous discussion is that, given the additional measurement, we work with the conditional or *a posteriori* density $f_{Y|X}(y|x)$, rather than the unconditioned density $f_Y(y)$, and now our aim is to minimize

$$E[\{Y - \widehat{y}(x)\}^2|X = x] = \int \{y - \widehat{y}(x)\}^2 f_{Y|X}(y|x)\, dy\ . \tag{8.7}$$

We have introduced the notation $\widehat{y}(x)$ for our estimate to show that in general it will depend on the specific value x. Exactly the same calculations as in the case of no measurements then show that

$$\widehat{y}(x) = E[Y|X = x]\ , \tag{8.8}$$

the conditional expectation of Y, given $X = x$. The associated MMSE is the variance $\sigma^2_{Y|X}$ of the conditional density $f_{Y|X}(y|x)$, i.e., the MMSE is the conditional variance. Thus, the only change from the case of no measurements is that we now condition on the obtained measurement.

Going a further step, if we have multiple measurements, say $X_1 = x_1, X_2 = x_2, \cdots, X_L = x_L$, then we work with the *a posteriori* density

$$f_{Y\,|\,X_1,X_2,\cdots,X_L}(y\,|\,x_1, x_2, \cdots, x_L)\ . \tag{8.9}$$

Apart from this modification, there is no change in the structure of the solutions. Thus, without further calculation, we can state the following:

$$\boxed{\begin{array}{c}\text{The MMSE estimate of } Y,\\ \text{given } X_1 = x_1, \cdots, X_L = x_L,\\ \text{is the } \textbf{conditional expectation} \text{ of } Y:\\ \\ \widehat{y}(x_1, \ldots, x_L) = E[Y\,|\,X_1 = x_1, \cdots, X_L = x_L]\end{array}} \tag{8.10}$$

For notational convenience, we can arrange the measured random variables into a column vector $\mathbf{X}$, and the corresponding measurements into the column vector $\mathbf{x}$. The dependence of the MMSE estimate on the measurements can now be indicated by the notation $\widehat{y}(\mathbf{x})$, with

$$\widehat{y}(\mathbf{x}) = \int_{-\infty}^{\infty} y\, f_{Y|\mathbf{X}}(y\,|\,\mathbf{X} = \mathbf{x})\, dy = E[\,Y\,|\,\mathbf{X} = \mathbf{x}\,]\ . \tag{8.11}$$

The minimum mean square error (or MMSE) for the given value of $\mathbf{X}$ is again the conditional variance, i.e., the variance $\sigma^2_{Y|\mathbf{X}}$ of the conditional density $f_{Y|\mathbf{X}}(y\,|\,\mathbf{x})$.

EXAMPLE 8.1 MMSE Estimate for Discrete Random Variables

A discrete-time discrete-amplitude sequence $s[n]$ is stored on a noisy medium. The retrieved sequence is $r[n]$. Suppose at some particular time instant $n = n_0$ we have

$s[n_0]$ and $r[n_0]$ modeled as random variables, which we shall simply denote by S and R respectively. From prior measurements, we have determined that S and R have the joint probability mass function (PMF) shown in Figure 8.1.

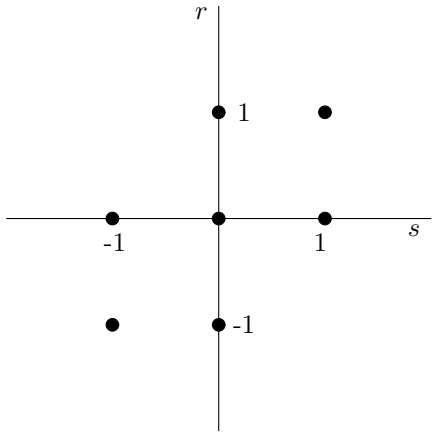

FIGURE 8.1 Joint PMF of S and R.

Based on receiving the value $R = 1$, we would like to make an MMSE estimate $\widehat{s}$ of S. From (8.10), $\widehat{s} = E(S|R = 1)$, which can be determined from the conditional PMF $P_{S|R}(s|R = 1)$, which in turn we can obtain as

$$P_{S|R}(s|R = 1) = \frac{P_{R,S}(R = 1, s)}{P_R(R = 1)} . \tag{8.12}$$

From Figure 8.1,

$$P_R(1) = \frac{2}{7} \tag{8.13}$$

and

$$P_{R,S}(1, s) = \begin{cases} 0 & s = -1 \\ 1/7 & s = 0 \\ 1/7 & s = +1 \end{cases}$$

Consequently,

$$P_{S|R}(s|R = 1) = \begin{cases} 1/2 & s = 0 \\ 1/2 & s = +1 \end{cases}$$

Thus, the MMSE estimate is $\widehat{s} = \frac{1}{2}$. Note that although this estimate minimizes the mean square error, we have not constrained it to take account of the fact that S can only have the discrete values of $+1$, 0 or -1. In a later chapter we will return to this example and consider it from the perspective of hypothesis testing, i.e., determining which of the three known possible values will result in minimizing

a suitable error criterion.

EXAMPLE 8.2 MMSE Estimate of Signal in Additive Noise

A discrete-time sequence $s[n]$ is transmitted over a noisy channel and retrieved. The received sequence $r[n]$ is modeled as $r[n] = s[n] + w[n]$ where $w[n]$ represents the noise. At a particular time instant $n = n_0$, suppose $r[n_0]$, $s[n_0]$ and $w[n_0]$ are random variables, which we denote as R, S and W respectively. We assume that S and W are independent, that W is uniformly distributed between $+\frac{1}{2}$ and $-\frac{1}{2}$, and S is uniformly distributed between -1 and $+1$. The specific received value is $R = \frac{1}{4}$, and we want the MMSE estimate $\widehat{s}$ for S. From (8.10),

$$\widehat{s} = E(S|R = \frac{1}{4}) \tag{8.14}$$

which can be determined from $f_{S|R}(s|R = \frac{1}{4})$:

$$f_{S|R}(s|R = \frac{1}{4}) = \frac{f_{R|S}(\frac{1}{4}|s)f_S(s)}{f_R(\frac{1}{4})} \ . \tag{8.15}$$

We evaluate separately the numerator and denominator terms in (8.15). The PDF $f_{R|S}(r|s)$ is identical in shape to the PDF of W, but with the mean shifted to s, as indicated in Figure 8.2 below. Consequently, $f_{R|S}(\frac{1}{4}|s)$ is as shown in Figure 8.3, and $f_{R|S}(\frac{1}{4}|s)f_S(s)$ is shown in Figure 8.4.

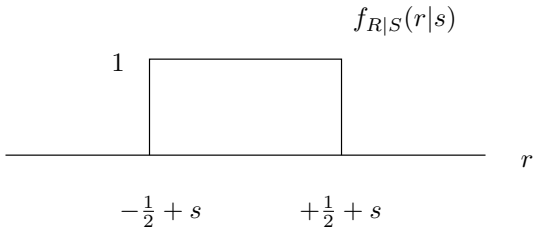

FIGURE 8.2 Conditional PDF of R given S, $f_{R|S}(r|s)$.

To obtain $f_{S|R}(s|R = \frac{1}{4})$ we divide Figure 8.4 by $f_R(\frac{1}{4})$, which can easily be obtained by evaluating the convolution of the PDF's of S and W at the argument $\frac{1}{4}$. More simply, since $f_{S|R}(s|R = \frac{1}{4})$ must have total area of unity and it is the same as Figure 8.4 but scaled by $f_R(\frac{1}{4})$, we can easily obtain it by just normalizing Figure 8.4 to have an area of 1. The resulting value for $\widehat{s}$ is the mean associated with the PDF $f_{S|R}(s|R = \frac{1}{4})$, which will be

$$\widehat{s} = \frac{1}{4} \ . \tag{8.16}$$

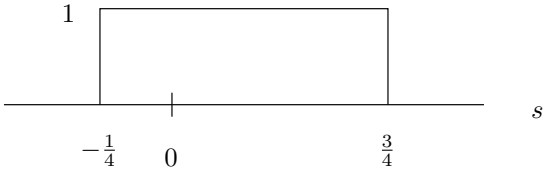

FIGURE 8.3 Plot of $f_{R|S}(\frac{1}{4}|s)$.

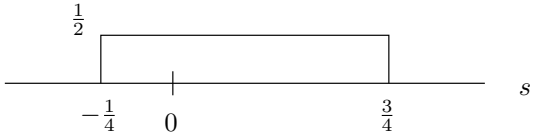

FIGURE 8.4 Plot of $f_{R|S}(\frac{1}{4}|s)f_S(s)$.

The associated MMSE is the variance of this PDF, namely $\frac{1}{12}$.

EXAMPLE 8.3 MMSE Estimate for Bivariate Gaussian Random Variables

Two random variables X and Y are said to have a bivariate Gaussian joint PDF if the joint density of the centered (i.e. zero-mean) and normalized (i.e. unit-variance) random variables

$$V = \frac{X - \mu_X}{\sigma_X} \ , \ \ W = \frac{Y - \mu_Y}{\sigma_Y} \tag{8.17}$$

is given by

$$f_{V,W}(v,w) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{(v^2 - 2\rho vw + w^2)}{2(1-\rho^2)}\right\} . \tag{8.18}$$

Here μ_X and μ_Y are the means of X and Y respectively, and σ_X, σ_Y are the respective standard deviations of X and Y. The number ρ is the correlation coefficient of X and Y, and is defined by

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \ , \ \text{ with } \ \sigma_{XY} = E[XY] - \mu_X \mu_Y \tag{8.19}$$

where σ_{XY} is the covariance of X and Y.

Now, consider $\widehat{y}(x)$, the MMSE estimate of Y given $X = x$, when X and Y are bivariate Gaussian random variables. From (8.10),

$$\widehat{y}(x) = E[Y \,|\, X = x] \tag{8.20}$$

or, in terms of the zero-mean normalized random variables V and W,

$$\begin{aligned}\widehat{y}(x) &= E\left[(\sigma_Y W + \mu_Y)\,|\,V = \frac{x-\mu_X}{\sigma_X}\right] \\ &= \sigma_Y E\left[W\,|\,V = \frac{x-\mu_X}{\sigma_X}\right] + \mu_Y\ . \end{aligned} \tag{8.21}$$

It is straightforward to show with some computation that $f_{W|V}(w\,|\,v)$ is Gaussian with mean ρv, and variance $1-\rho^2$, from which it follows that

$$E\left[W\,|\,V = \frac{x-\mu_X}{\sigma_X}\right] = \rho\left[\frac{x-\mu_X}{\sigma_X}\right]\ . \tag{8.22}$$

Combining (8.21) and (8.22),

$$\begin{aligned}\widehat{y}(x) &= E[\,Y\,|\,\mathbf{X} = x\,] \\ &= \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\end{aligned} \tag{8.23}$$

The MMSE estimate in the case of bivariate Gaussian variables has a nice linear (or more correctly, affine, i.e., linear plus a constant) form.

The minimum mean square error is the variance of the conditional PDF $f_{Y|\mathbf{X}}(y|\mathbf{X} = x)$:

$$E[\,(Y-\widehat{y}(x))^2\,|\,\mathbf{X} = x\,] = \sigma_Y^2\,(1-\rho^2)\ . \tag{8.24}$$

Note that σ_Y^2 is the mean square error in Y in the absence of any additional information. Equation (8.24) shows what the residual mean square error is after we have a measurement of X. It is evident and intuitively reasonable that the larger the magnitude of the correlation coefficient between X and Y, the smaller the residual mean square error.

8.2 FROM ESTIMATES TO AN ESTIMATOR

The MMSE estimate in (8.8) is based on knowing the specific value x that the random variable X takes. While X is a random variable, the specific value x is not, and consequently $\widehat{y}(x)$ is also not a random variable.

As we move forward in the discussion, it is important to draw a distinction between the estimate of a random variable and the procedure by which we form the estimate. This is completely analogous to the distinction between the value of a function at a point and the function itself. We will refer to the procedure or function that produces the estimate as the estimator.

For instance, in Example 8.1 we determined the MMSE estimate of S for the specific value of $R = 1$. We could more generally determine an estimate of S for each of the possible values of R, i.e., $-1, 0,$ and $+1$. We could then have a tabulation of these results available in advance, so that when we retrieve a specific value of R

we can look up the MMSE estimate. Such a table or more generally a function of R would correspond to what we term the MMSE estimator. The input to the table or estimator would be the specific retrieved value and the output would be the estimate associated with that retrieved value.

We have already introduced the notation $\widehat{y}(x)$ to denote the estimate of Y given $X = x$. The function $\widehat{y}(\cdot)$ determines the corresponding estimator, which we will denote by $\widehat{y}(X)$, or more simply by just $\widehat{Y}$, if it is understood what random variable the estimator is operating on. Note that the estimator $\widehat{Y} = \widehat{y}(X)$ is a random variable. We have already seen that the MMSE estimate $\widehat{y}(x)$ is given by the conditional mean, $E[Y|X = x]$, which suggests yet another natural notation for the MMSE estimator:

$$\widehat{Y} = \widehat{y}(X) = E[Y|X] \, . \tag{8.25}$$

Note that $E[Y|X]$ denotes a random variable, not a number.

The preceding discussion applies essentially unchanged to the case where we observe several random variables, assembled in the vector $\mathbf{X}$. The MMSE estimator in this case is denoted by

$$\widehat{Y} = \widehat{y}(\mathbf{X}) = E[Y|\mathbf{X}] \, . \tag{8.26}$$

Perhaps not surprisingly, the MMSE estimator for Y given $\mathbf{X}$ minimizes the mean square error, averaged over all Y and $\mathbf{X}$. This is because the MMSE estimator minimizes the mean square error for each particular value $\mathbf{x}$ of $\mathbf{X}$. More formally,

$$\begin{aligned} E_{Y,\mathbf{X}}\Big([Y - \widehat{y}(\mathbf{X})]^2 \Big) &= E_{\mathbf{X}}\Big(E_{Y|\mathbf{X}}\Big([Y - \widehat{y}(\mathbf{X})]^2 \,|\, \mathbf{X} \Big)\Big) \\ &= \int_{-\infty}^{\infty} \Big(E_{Y|\mathbf{X}}\Big([Y - \widehat{y}(\mathbf{x})]^2 \,|\, \mathbf{X} = \mathbf{x} \Big)\, f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \, . \end{aligned} \tag{8.27}$$

(The subscripts on the expectation operators are used to indicate explicitly which densities are involved in computing the associated expectations; the densities and integration are multivariate when $\mathbf{X}$ is not a scalar.) Because the estimate $\widehat{y}(\mathbf{x})$ is chosen to minimize the inner expectation $E_{Y|\mathbf{X}}$ for each value $\mathbf{x}$ of $\mathbf{X}$, it also minimizes the outer expectation $E_{\mathbf{X}}$, since $f_{\mathbf{X}}(\mathbf{X})$ is nonnegative.

EXAMPLE 8.4 MMSE Estimator for Bivariate Gaussian Random Variables

We have already, in Example 8.3, constructed the MMSE estimate of one member of a pair of bivariate Gaussian random variables, given a measurement of the other. Using the same notation as in that example, it is evident that the MMSE estimator is simply obtained on replacing x by X in (8.23):

$$\widehat{Y} = \widehat{y}(X) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X) \, . \tag{8.28}$$

The conditional MMSE given $X = x$ was found in the earlier example to be $\sigma_Y^2(1 - \rho^2)$, which did not depend on the value of x, so the MMSE of the estimator, averaged

over all X, ends up still being $\sigma_Y^2(1-\rho^2)$.

EXAMPLE 8.5 MMSE Estimator for Signal in Additive Noise

Suppose the random variable X is a noisy measurement of the angular position Y of an antenna, so $X = Y + W$, where W denotes the additive noise. Assume the noise is independent of the angular position, i.e., Y and W are independent random variables, with Y uniformly distributed in the interval $[-1, 1]$ and W uniformly distributed in the interval $[-2, 2]$. (Note that the setup in this example is essentially the same as in Example 8.2, though the context, notation and parameters are different.)

Given that $X = x$, we would like to determine the MMSE estimate $\widehat{y}(x)$, the resulting mean square error, and the overall mean square error averaged over all possible values x that the random variable X can take. Since $\widehat{y}(x)$ is the conditional expectation of Y given $X = x$, we need to determine $f_{Y|X}(y|x)$. For this, we first determine the joint density of Y and W, and from this the required conditional density.

From the independence of Y and W:

$$f_{Y,W}(y,w) = f_Y(y)f_W(w) = \begin{cases} \frac{1}{8} & -2 \le w \le 2, -1 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

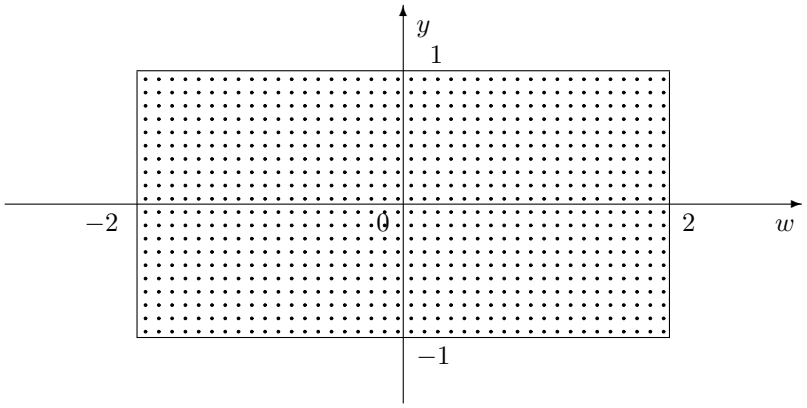

FIGURE 8.5 Joint PDF of Y and W for Example 8.5.

Conditioned on $Y = y$, X is the same as $y + W$, uniformly distributed over the interval $[y-2, y+2]$. Now

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) = (\frac{1}{4})(\frac{1}{2}) = \frac{1}{8}$$

for $-1 \leq y \leq 1$, $y-2 \leq x \leq y+2$, and zero otherwise. The joint PDF is therefore uniform over the parallelogram shown in the Figure 8.6.

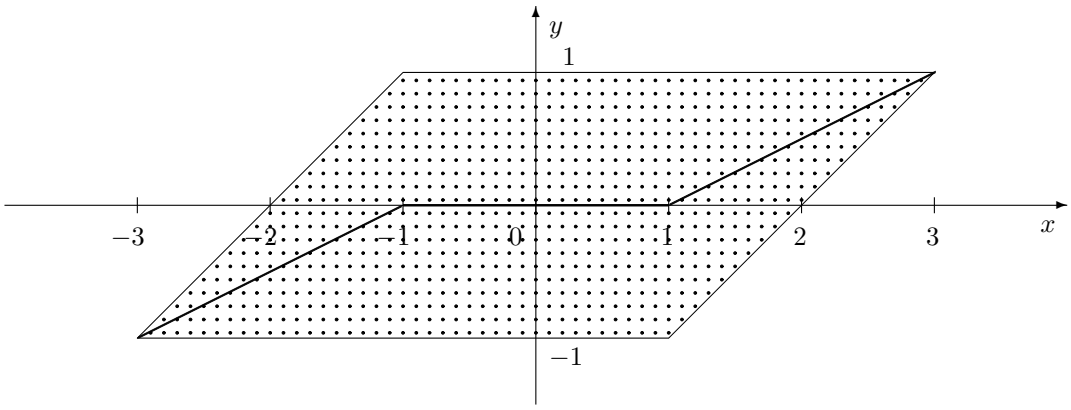

FIGURE 8.6 Joint PDF of X and Y and plot of the MMSE estimator of Y from X for Example 8.5.

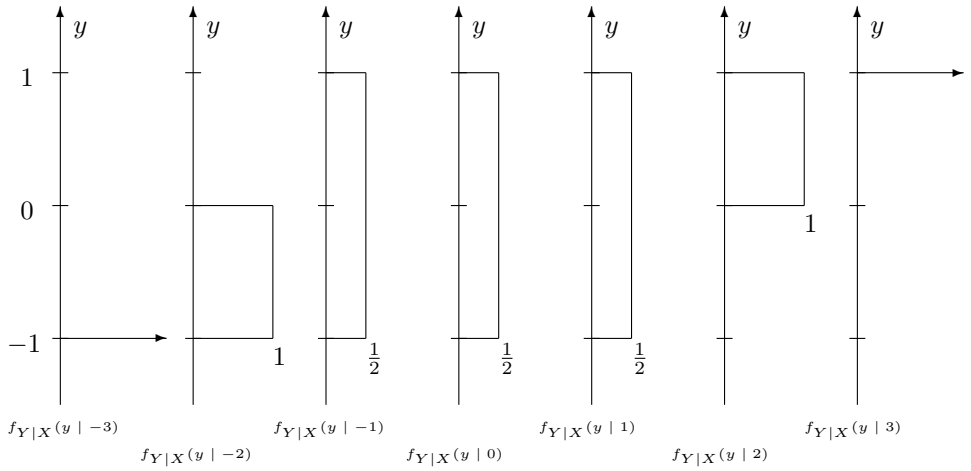

FIGURE 8.7 Conditional PDF $f_{Y|X}$ for various realizations of X for Example 8.5.

Given $X = x$, the conditional PDF $f_{Y|X}$ is uniform on the corresponding vertical section of the parallelogram:

$$f_{Y|X}(y,x) = \begin{cases} \dfrac{1}{3+x} & -3 \leq x \leq -1\,, -1 \leq y \leq x+2 \\ \dfrac{1}{2} & -1 \leq x \leq 1\,, -1 \leq y \leq 1 \\ \dfrac{1}{3-x} & 1 \leq x \leq 3\,, x-2 \leq y \leq 1 \end{cases} \tag{8.29}$$

The MMSE estimate $\widehat{y}(x)$ is the conditional mean of Y given $X = x$, and the conditional mean is the midpoint of the corresponding vertical section of the parallelogram. The conditional mean is displayed as the heavy line on the parallelogram in the second plot. In analytical form,

$$\widehat{y}(x) = E[Y \mid X = x] = \begin{cases} \frac{1}{2} + \frac{1}{2}x & -3 \le x < -1 \\ 0 & -1 \le x < 1 \\ -\frac{1}{2} + \frac{1}{2}x & 1 \le x \le 3 \end{cases} \tag{8.30}$$

The minimum mean square error associated with this estimate is the variance of the uniform distribution in eq. (8.29), specifically:

$$E[\{Y - \widehat{y}(x)\}^2 \mid X = x] \quad \begin{cases} \frac{(3+x)^2}{12} & -3 \le x < -1 \\ \frac{1}{3} & -1 \le x < 1 \\ \frac{(3-x)^2}{12} & 1 \le x \le 3 \end{cases} \tag{8.31}$$

Equation (8.31) specifies the mean square error that results for any specific value x of the measurement of X. Since the measurement is a random variable, it is also of interest to know what the mean square error is, averaged over all possible values of the measurement, i.e. over the random variable X. To determine this, we first determine the marginal PDF of X:

$$f_X(x) = \frac{f_{X,Y}(x,y)}{f_{Y|X}(y \mid x)} = \begin{cases} \frac{3+x}{8} & -3 \le x < -1 \\ \frac{1}{4} & -1 \le x < 1 \\ \frac{3-x}{8} & 1 \le x \le 3 \\ 0 & \text{otherwise} \end{cases}$$

This could also be found by convolution, $f_X = f_Y * f_W$, since Y and W are statistically independent. Then,

$$\begin{aligned} E_X[E_{Y|X}\{(Y - \widehat{y}(x)\}^2 \mid X = x]] &= \int_{-\infty}^{\infty} E[(Y - \widehat{y}(x))^2 \mid X = x] f_X(x) \mathrm{d}x \\ &= \int_{-3}^{-1} (\frac{(3+x)^2}{12})(\frac{3+x}{8})\mathrm{d}x + \int_{-1}^{1} (\frac{1}{3})(\frac{1}{4})\mathrm{d}x + \int_{1}^{3} (\frac{(3-x)^2}{12})(\frac{3-x}{8})\mathrm{d}x \\ &= \frac{1}{4} \end{aligned}$$

Compare this with the mean square error if we just estimated Y by its mean, namely 0. The mean square error would then be the variance σ_Y^2:

$$\sigma_Y^2 = \frac{[1-(-1)]^2}{12} = \frac{1}{3} ,$$

so the mean square error is indeed reduced by allowing ourselves to use knowledge of X and of the probabilistic relation between Y and X.

8.2.1 Orthogonality

A further important property of the MMSE estimator is that the residual error $Y - \widehat{y}(\mathbf{X})$ is orthogonal to any function $h(\mathbf{X})$ of the measured random variables:

$$E_{Y,X}[\{Y - \widehat{y}(\mathbf{X})\}h(\mathbf{X})] = 0 , \tag{8.32}$$

where the expectation is computed over the joint density of Y and $\mathbf{X}$. Rearranging this, we have the equivalent condition

$$E_{Y,X}[\widehat{y}(\mathbf{X})h(\mathbf{X})] = E_{Y,X}[Yh(\mathbf{X})] , \tag{8.33}$$

i.e., the MMSE estimator has the same correlation as Y does with any function of X. In particular, choosing $h(\mathbf{X}) = 1$, we find that

$$E_{Y,X}[\widehat{y}(\mathbf{X})] = E_Y[Y] . \tag{8.34}$$

The latter property results in the estimator being referred to as unbiased: its expected value equals the expected value of the random variable being estimated. We can invoked the unbiasedness property to interpret (8.32) as stating that the estimation error of the MMSE estimator is uncorrelated with any function of the random variables used to construct the estimator.

The proof of the correlation matching property in (8.33) is in the following sequence of equalities:

$$\begin{aligned} E_{Y,X}[\widehat{y}(\mathbf{X})h(\mathbf{X})] &= E_X[E_{Y|X}[Y|\mathbf{X}]h(\mathbf{X})] && (8.35)\\ &= E_X[E_{Y|X}[Yh(\mathbf{X})|\mathbf{X}]] && (8.36)\\ &= E_{Y,X}[Yh(\mathbf{X})] . && (8.37) \end{aligned}$$

Rearranging the final result here, we obtain the orthogonality condition in (8.32).

8.3 LINEAR MINIMUM MEAN SQUARE ERROR ESTIMATION

In general, the conditional expectation $E(Y|\mathbf{X})$ required for the MMSE estimator developed in the preceding sections is difficult to determine, because the conditional density $f_{Y|\mathbf{X}}(y|\mathbf{x})$ is not easily determined. A useful and widely used compromise

is to restrict the estimator to be a fixed linear (or actually affine, i.e., linear plus a constant) function of the measured random variables, and to choose the linear relationship so as to minimize the mean square error. The resulting estimator is called the linear minimum mean square error (LMMSE) estimator. We begin with the simplest case.

Suppose we wish to construct an estimator for the random variable Y in terms of another random variable X, restricting our estimator to be of the form

$$\widehat{Y}_\ell = \widehat{y}_\ell(X) = aX + b \,, \tag{8.38}$$

where a and b are to be determined so as to minimize the mean square error

$$E_{Y,X}[(Y - \widehat{Y}_\ell)^2] = E_{Y,X}[\{Y - (aX + b)\}^2] \,. \tag{8.39}$$

Note that the expectation is taken over the joint density of Y and X; the linear estimator is picked to be optimum when averaged over all possible combinations of Y and X that may occur. We have accordingly used subscripts on the expectation operations in (8.39) to make explicit for now the variables whose joint density the expectation is being computed over; we shall eventually drop the subscripts.

Once the optimum a and b have been chosen in this manner, the estimate of Y, given a particular x, is just $\widehat{y}_\ell(x) = ax + b$, computed with the already designed values of a and b. Thus, in the LMMSE case we construct an optimal linear estimator, and for any particular x this estimator generates an estimate that is not claimed to have any individual optimality property. This is in contrast to the MMSE case considered in the previous sections, where we obtained an optimal MMSE estimate for each x, namely $E[Y|X = x]$, that minimized the mean square error conditioned on $X = x$. The distinction can be summarized as follows: in the unrestricted MMSE case, the optimal estimator is obtained by joining together all the individual optimal estimates, whereas in the LMMSE case the (generally non-optimal) individual estimates are obtained by simply evaluating the optimal linear estimator.

We turn now to minimizing the expression in (8.39), by differentiating it with respect to the parameters a and b, and setting each of the derivatives to 0. (Consideration of the second derivatives will show that we do indeed find minimizing values in this fashion, but we omit the demonstration.) First differentiating (8.39) with respect to b, taking the derivative inside the integral that corresponds to the expectation operation, and then setting the result to 0, we conclude that

$$E_{Y,X}[Y - (aX + b)] = 0 \,, \tag{8.40}$$

or equivalently

$$E[Y] = E[aX + b] = E[\widehat{Y}_\ell] \,, \tag{8.41}$$

from which we deduce that

$$b = \mu_Y - a\mu_X \,, \tag{8.42}$$

where $\mu_Y = E[Y] = E_{Y,X}[Y]$ and $\mu_X = E[X] = E_{Y,X}[X]$. The optimum value of b specified in (8.42) in effect serves to make the linear estimator unbiased, i.e., the

expected value of the estimator becomes equal to the expected value of the random variable we are trying to estimate, as (8.41) shows.

Using (8.42) to substitute for b in (8.38), it follows that

$$\widehat{Y}_\ell = \mu_Y + a(X - \mu_X) \; . \tag{8.43}$$

In other words, to the expected value μ_Y of the random variable Y that we are estimating, the optimal linear estimator adds a suitable multiple of the difference $X - \mu_X$ between the measured random variable and its expected value. We turn now to finding the optimum value of this multiple, a.

First rewrite the error criterion (8.39) as

$$E[\{(Y - \mu_Y) - (\widehat{Y}_\ell - \mu_Y)\}^2] = E[(\widetilde{Y} - a\widetilde{X})^2] \; , \tag{8.44}$$

where

$$\widetilde{Y} = Y - \mu_Y \quad \text{and} \quad \widetilde{X} = X - \mu_X \; , \tag{8.45}$$

and where we have invoked (8.43) to obtain the second equality in (8.44). Now taking the derivative of the error criterion in (8.44) with respect to a, and setting the result to 0, we find

$$E[(\widetilde{Y} - a\widetilde{X})\widetilde{X}] = 0 \; . \tag{8.46}$$

Rearranging this, and recalling that $E[\widetilde{Y}\widetilde{X}] = \sigma_{YX}$, i.e., the covariance of Y and X, and that $E[\widetilde{X}^2] = \sigma_X^2$, we obtain

$$a = \frac{\sigma_{YX}}{\sigma_X^2} = \rho_{YX}\frac{\sigma_Y}{\sigma_X} \; , \tag{8.47}$$

where ρ_{YX} — which we shall simply write as ρ when it is clear from context what variables are involved — denotes the correlation coefficient between Y and X.

It is also enlightening to understand the above expression for a in terms of the vector-space picture for random variables developed in the previous chapter.

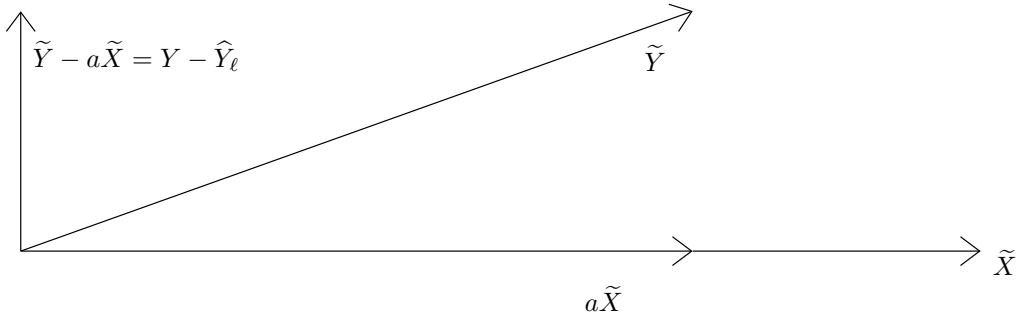

FIGURE 8.8 Expression for a from Eq. (8.47) illustrated in vector space.

The expression (8.44) for the error criterion shows that we are looking for a vector $a\widetilde{X}$, which lies along the vector $\widetilde{X}$, such that the squared length of the error vector

$\widetilde{Y} - a\widetilde{X}$ is minimum. It follows from familiar geometric reasoning that the optimum choice of $a\widetilde{X}$ must be the orthogonal projection of $\widetilde{Y}$ on $\widetilde{X}$, and that this projection is

$$a\widetilde{X} = \frac{<\widetilde{Y}, \widetilde{X}>}{<\widetilde{X}, \widetilde{X}>} \widetilde{X} \ . \tag{8.48}$$

Here, as in the previous chapter, $< U, V >$ denotes the inner product of the vectors U and V, and in the case where the "vectors" are random variables, denotes $E[UV]$. Our expression for a in (8.47) follows immediately. Figure 8.8 shows the construction associated with the requisite calculations. Recall from the previous chapter that the correlation coefficient ρ denotes the cosine of the angle between the vectors $\widetilde{Y}$ and $\widetilde{X}$.

The preceding projection operation implies that the error $\widetilde{Y} - a\widetilde{X}$, which can also be written as $Y - \widehat{Y}_\ell$, must be orthogonal to $\widetilde{X} = X - \mu_X$. This is precisely what (8.46) says. In addition, invoking the unbiasedness of $\widehat{Y}_\ell$ shows that $Y - \widehat{Y}_\ell$ must be orthogonal to μ_X (or any other constant), so $Y - \widehat{Y}_\ell$ is therefore orthogonal to X itself:

$$E[(Y - \widehat{Y}_\ell)X] = 0 \ . \tag{8.49}$$

In other words, the optimal LMMSE estimator is unbiased and such that the estimation error is orthogonal to the random variable on which the estimator is based. (Note that the statement in the case of the MMSE estimator in the previous section was considerably stronger, namely that the error was orthogonal to any function $h(X)$ of the measured random variable, not just to the random variable itself.)

The preceding development shows that the properties of (i) unbiasedness of the estimator, and (ii) orthogonality of the error to the measured random variable, completely characterize the LMMSE estimator. Invoking these properties yields the LMMSE estimator.

Going a step further with the geometric reasoning, we find from Pythagoras's theorem applied to the triangle in Figure 8.8 that the minimum mean square error (MMSE) obtained through use of the LMMSE estimator is

$$\text{MMSE} = E[(\widetilde{Y} - a\widetilde{X})^2] = E[\widetilde{Y}^2](1 - \rho^2) = \sigma_Y^2(1 - \rho^2) \ . \tag{8.50}$$

This result could also be obtained purely analytically, of course, without recourse to the geometric interpretation. The result shows that the mean square error σ_Y^2 that we had prior to estimation in terms of X is reduced by the factor $1 - \rho^2$ when we use X in an LMMSE estimator. The closer that ρ is to $+1$ or -1 (corresponding to strong positive or negative correlation respectively), the more our uncertainty about Y is reduced by using an LMMSE estimator to extract information that X carries about Y.

Our results on the LMMSE estimator can now be summarized in the following expressions for the estimator, with the associated minimum mean square error being given by (8.50):

$$\widehat{Y}_\ell = \widehat{y}_\ell(X) = \mu_Y + \frac{\sigma_{YX}}{\sigma_X^2}(X - \mu_X) = \mu_Y + \rho\,\frac{\sigma_Y}{\sigma_X}(X - \mu_X) \ , \tag{8.51}$$

or the equivalent but perhaps more suggestive form

$$\frac{\widehat{Y}_\ell - \mu_Y}{\sigma_Y} = \rho \frac{X - \mu_X}{\sigma_X} \ . \tag{8.52}$$

The latter expression states that the normalized deviation of the estimator from its mean is ρ times the normalized deviation of the observed variable from its mean; the more highly correlated Y and X are, the more closely we match the two normalized deviations.

Note that our expressions for the LMMSE estimator and its mean square error are the same as those obtained in Example 8.4 for the MMSE estimator in the bivariate Gaussian case. The reason is that the MMSE estimator in that case turned out to be linear (actually, affine), as already noted in the example.

EXAMPLE 8.6 LMMSE Estimator for Signal in Additive Noise

We return to Example 8.5, for which we have already computed the MMSE estimator, and we now design an LMMSE estimator. Recall that the random variable X denotes a noisy measurement of the angular position Y of an antenna, so $X = Y + W$, where W denotes the additive noise. We assume the noise is independent of the angular position, i.e., Y and W are independent random variables, with Y uniformly distributed in the interval $[-1, 1]$ and W uniformly distributed in the interval $[-2, 2]$.

For the LMMSE estimator of Y in terms of X, we need to determine the respective means and variances, as well as the covariance, of these random variables. It is easy to see that

$$\mu_Y = 0 \ , \quad \mu_W = 0 \ , \quad \mu_X = 0 \ , \quad \sigma_Y^2 = \frac{1}{3} \ , \quad \sigma_W^2 = \frac{4}{3} \ ,$$

$$\sigma_X^2 = \sigma_Y^2 + \sigma_W^2 = \frac{5}{3} \ , \quad \sigma_{YX} = \sigma_Y^2 = \frac{1}{3} \ , \quad \rho_{YX} = \frac{1}{\sqrt{5}} \ .$$

The LMMSE estimator is accordingly

$$\widehat{Y}_\ell = \frac{1}{5} X \ ,$$

and the associated MMSE is

$$\sigma_Y^2 (1 - \rho^2) = \frac{4}{15} \ .$$

This MMSE should be compared with the (larger) mean square error of $\frac{1}{3}$ obtained if we simply use $\mu_Y = 0$ as our estimator for Y, and the (smaller) value $\frac{1}{4}$ obtained using the MMSE estimator in Example 8.5.

EXAMPLE 8.7 Single-Point LMMSE Estimator for Sinusoidal Random Process

Consider a sinusoidal signal of the form

$$X(t) = A\cos(\omega_0 t + \Theta) \tag{8.53}$$

where ω_0 is assumed known, while A and Θ are statistically independent random variables, with the PDF of Θ being uniform in the interval $[0, 2\pi]$. Thus $X(t)$ is a random signal, or equivalently a set or "ensemble" of signals corresponding to the various possible outcomes for A and Θ in the underlying probabilistic experiment. We will discuss such signals in more detail in the next chapter, where we will refer to them as random processes. The value that $X(t)$ takes at some particular time $t = t_0$ is simply a random variable, whose specific value will depend on which outcomes for A and Θ are produced by the underlying probabilistic experiment.

Suppose we are interested in determining the LMMSE estimator for $X(t_1)$ based on a measurement of $X(t_0)$, where t_0 and t_1 are specified sampling times. In other words, we want to choose a and b in

$$\widehat{X}(t_1) = aX(t_0) + b \tag{8.54}$$

so as to minimize the mean square error between $X(t_1)$ and $\widehat{X}(t_1)$.

We have established that b must be chosen to ensure the estimator is unbiased:

$$E[\widehat{X}(t_1)] = aE[X(t_0)] + b = E[X(t_1)] \ .$$

Since A and Θ are independent,

$$E[X(t_0)] = E\{A\}\int_0^{2\pi} \frac{1}{2\pi}\cos(\omega_0 t_0 + \theta)\, d\theta = 0$$

and similarly $E[X(t_1)] = 0$, so we choose $b = 0$.

Next we use the fact that the error of the LMMSE estimator is orthogonal to the data:

$$E[(\widehat{X}(t_1) - X(t_1))X(t_0)] = 0$$

and consequently

$$aE[X^2(t_0)] = E[X(t_1)X(t_0)]$$

or

$$a = \frac{E[X(t_1)X(t_0)]}{E[X^2(t_0)]} \ . \tag{8.55}$$

The numerator and denominator in (8.55) are respectively

$$\begin{aligned} E[X(t_1)X(t_0)] &= E[A^2]\int_0^{2\pi} \frac{1}{2\pi}\cos(\omega_0 t_1 + \theta)\cos(\omega_0 t_0 + \theta)\, d\theta \\ &= \frac{E[A^2]}{2}\cos\{\omega_0(t_1 - t_0)\} \end{aligned}$$

and $E[X^2(t_0)] = \frac{E[A^2]}{2}$. Thus $a = \cos\{\omega_0(t_1 - t_0)\}$, so the LMMSE estimator is

$$\widehat{X}(t_1) = X(t_0)\cos\{\omega_0(t_1 - t_0)\} \; . \tag{8.56}$$

It is interesting to observe that the distribution of A doesn't play a role in this equation.

To evaluate the mean square error associated with the LMMSE estimator, we compute the correlation coefficient between the samples of the random signal at t_0 and t_1. It is easily seen that$\rho = a = \cos\{\omega_0(t_1 - t_0)\}$, so the mean square error is

$$\frac{E[A^2]}{2}\Big(1 - \cos^2\{\omega_0(t_1 - t_0)\}\Big) = \frac{E[A^2]}{2}\sin^2\{\omega_0(t_1 - t_0)\} \; . \tag{8.57}$$

We now extend the LMMSE estimator to the case where our estimation of a random variable Y is based on observations of multiple random variables, say $X_1, \ldots, X_L$, gathered in the vector $\mathbf{X}$. The affine estimator may then be written in the form

$$\widehat{Y}_\ell = \widehat{y}_\ell(\mathbf{X}) = a_0 + \sum_{j=1}^{L} a_j X_j \; . \tag{8.58}$$

As we shall see, the coefficient a_i of this LMMSE estimator can be found by solving a linear system of equations that is completely defined by the first and second moments (i.e., means, variances and covariances) of the random variables Y and X_j. The fact that the model (8.58) is linear in the parameters a_i is what results in a linear system of equations; the fact that the model is affine in the random variables is what makes the solution only depend on their first and second moments. Linear equations are easy to solve, and first and second moments are generally easy to determine, hence the popularity of LMMSE estimation.

The development below follows along the same lines as that done earlier in this section for the case where we just had a single observed random variable X, but we use the opportunity to review the logic of the development and to provide a few additional insights.

We want to minimize the mean square error

$$E\Big[\Big(Y - (a_0 + \sum_{j=1}^{L} a_j X_j)\Big)^2\Big] \; , \tag{8.59}$$

where the expectation is computed using the joint density of Y and $\mathbf{X}$. We use the joint density rather than the conditional because the parameters are not going to be picked to be best for a particular set of measured values $\mathbf{x}$ — otherwise we could do as well as the nonlinear estimate in this case, by setting $a_0 = E[Y \,|\, \mathbf{X} = \mathbf{x}]$ and setting all the other a_i to zero. Instead, we are picking the parameters to be the best averaged over all possible $\mathbf{X}$. The linear estimator will in general not be as good

as the unconstrained estimator, except in special cases (some of them important, as in the case of bivariate Gaussian random variables) but this estimator has the advantage that it is easy to solve for, as we now show.

To minimize the expression in (8.59), we differentiate it with respect to a_i for $i = 0, 1, \cdots, L$, and set each of the derivatives to 0. (Again, calculations involving second derivatives establish that we do indeed obtain minimizing values, but we omit these calculation here.) First differentiating with respect to a_0 and setting the result to 0, we conclude that

$$E[Y] = E[\, a_0 + \sum_{j=1}^{L} a_j X_j \,] = E[\widehat{Y}_\ell] \tag{8.60}$$

or

$$a_0 = \mu_Y - \sum_{j=1}^{L} a_j \, \mu_{X_j} \, , \tag{8.61}$$

where $\mu_Y = E[Y]$ and $\mu_{X_j} = E[X_j]$. This optimum value of a_0 serves to make the linear estimator unbiased, in the sense that (8.60) holds, i.e., the expected value of the estimator is the expected value of the random variable we are trying to estimate.

Using (8.61) to substitute for a_0 in (8.58), it follows that

$$\widehat{Y}_\ell = \mu_Y + \sum_{j=1}^{L} a_j (X_j - \mu_{X_j}) \, . \tag{8.62}$$

In other words, the estimator corrects the expected value μ_Y of the variable we are estimating, by a linear combination of the deviations $X_j - \mu_{X_j}$ between the measured random variables and their respective expected values.

Taking account of (8.62), we can rewrite our mean square error criterion (8.59) as

$$E[\{(Y - \mu_Y) - (\widehat{Y}_\ell - \mu_Y)\}^2] = E\Big[\Big(\widetilde{Y} - \sum_{j=1}^{L} a_j \widetilde{X}_j)\Big)^2\Big] \, , \tag{8.63}$$

where

$$\widetilde{Y} = Y - \mu_Y \quad \text{and} \quad \widetilde{X}_j = X_j - \mu_{X_j} \, . \tag{8.64}$$

Differentiating this with respect to each of the remaining coefficients $a_i, i = 1, 2, ...L$, and setting the result to zero produces the equations

$$E[(\widetilde{Y} - \sum_{j=1}^{L} a_j \widetilde{X}_j)\widetilde{X}_i] = 0 \quad i = 1, 2, ..., L \, . \tag{8.65}$$

or equivalently, if we again take account of (8.62),

$$E[(Y - \widehat{Y}_\ell)\widetilde{X}_i] = 0 \quad i = 1, 2, ..., L \, . \tag{8.66}$$

Yet another version follows on noting from (8.60) that $Y - \widehat{Y}_\ell$ is orthogonal to all constants, in particular to μ_{X_i}, so

$$E[(Y - \widehat{Y}_\ell)X_i] = 0 \quad i = 1, 2, ..., L \; . \tag{8.67}$$

All three of the preceding sets of equations express, in slightly different forms, the orthogonality of the estimation error to the random variables used in the estimator. One moves between these forms by invoking the unbiasedness of the estimator. The last of these, (8.67), is the usual statement of the orthogonality condition that governs the LMMSE estimator. (Note once more that the statement in the case of the MMSE estimator in the previous section was considerably stronger, namely that the error was orthogonal to any function $h(\mathbf{X})$ of the measured random variables, not just to the random variables themselves.) Rewriting this last equation as

$$E[YX_i] = E[\widehat{Y}_\ell X_i] \quad i = 1, 2, ..., L \tag{8.68}$$

yields an equivalent statement of the orthogonality condition, namely that the LMMSE estimator $\widehat{Y}_\ell$ has the same correlations as Y with the measured variables X_i.

The orthogonality and unbiasedness conditions together determine the LMMSE estimator completely. Also, the preceding developments shows that the first and second moments of Y and the X_i are exactly matched by the corresponding first and second moments of $\widehat{Y}_\ell$ and the X_i. It follows that Y and $\widehat{Y}_\ell$ cannot be told apart on the basis of only first and second moments with the measured variables X_i.

We focus now on (8.65), because it provides the best route to a solution for the coefficients a_j, $j = 1, \ldots, L$. This set of equations can be expressed as

$$\sum_{j=1}^{L} \sigma_{X_i X_j} a_j = \sigma_{X_i Y} \; , \tag{8.69}$$

where $\sigma_{X_i X_j}$ is the covariance of X_i and X_j (so $\sigma_{X_i X_i}$ is just the variance $\sigma^2_{X_i}$), and $\sigma_{X_i Y}$ is the covariance of X_i and Y. Collecting these equations in matrix form, we obtain

$$\begin{bmatrix} \sigma_{X_1 X_1} & \sigma_{X_1 X_2} & \cdots & \sigma_{X_1 X_L} \\ \sigma_{X_2 X_1} & \sigma_{X_2 X_2} & \cdots & \sigma_{X_2 X_L} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_L X_1} & \sigma_{X_L X_2} & \cdots & \sigma_{X_L X_L} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_L \end{bmatrix} = \begin{bmatrix} \sigma_{X_1 Y} \\ \sigma_{X_2 Y} \\ \vdots \\ \sigma_{X_L Y} \end{bmatrix} \; . \tag{8.70}$$

This set of equations is referred to as the normal equations. We can rewrite the normal equations in more compact matrix notation:

$$(\mathbf{C_{XX}})\,\mathbf{a} = \mathbf{C_{XY}} \tag{8.71}$$

where the definitions of $\mathbf{C_{XX}}$, $\mathbf{a}$, and $\mathbf{C_{XY}}$ should be evident on comparing the last two equations. The solution of this set of L equations in L unknowns yields the

$\{a_j\}$ for $j = 1, \cdots, L$, and these values may be substituted in (8.62) to completely specify the estimator. In matrix notation, the solution is

$$\mathbf{a} = (\mathbf{C_{XX}})^{-1}\mathbf{C_{XY}} \ . \tag{8.72}$$

It can be shown quite straightforwardly (though we omit the demonstration) that the minimum mean square error obtained with the LMMSE estimator is

$$\sigma_Y^2 - \mathbf{C}_{Y\mathbf{X}}(\mathbf{C_{XX}})^{-1}\mathbf{C_{XY}} = \sigma_Y^2 - \mathbf{C}_{Y\mathbf{X}}\mathbf{a} \ , \tag{8.73}$$

where $\mathbf{C}_{Y\mathbf{X}}$ is the transpose of $\mathbf{C_{XY}}$.

EXAMPLE 8.8 Estimation from Two Noisy Measurements

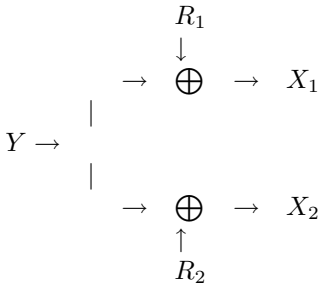

FIGURE 8.9 Illustration of relationship between random variables from Eq. (8.75) for Example 8.8.

Assume that Y, R_1 and R_2 are mutually uncorrelated, and that R_1 and R_2 have zero means and equal variances. We wish to find the linear MMSE estimator for Y, given measurements of X_1 and X_2. This estimator takes the form $\widehat{Y}_\ell = a_0 + a_1X_1 + a_2X_2$. Our requirement that $\widehat{Y}_\ell$ be unbiased results in the constraint

$$a_0 = \mu_Y - a_1\mu_{X_1} - a_2\mu_{X_2} = \mu_Y(1 - a_1 - a_2) \tag{8.74}$$

Next, we need to write down the normal equations, for which some preliminary calculations are required. Since

$$\begin{aligned} X_1 &= Y + R_1 \\ X_2 &= Y + R_2 \end{aligned} \tag{8.75}$$

and Y, R_1 and R_2 are mutually uncorrelated, we find

$$\begin{aligned} E[X_i^2] &= E[Y^2] + E[R_i^2] \ , \\ E[X_1X_2] &= E[Y^2] \ , \\ E[X_iY] &= E[Y^2] \ . \end{aligned} \tag{8.76}$$

The normal equations for this case thus become

$$\begin{bmatrix} \sigma_Y^2 + \sigma_R^2 & \sigma_Y^2 \\ \sigma_Y^2 & \sigma_Y^2 + \sigma_R^2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sigma_Y^2 \\ \sigma_Y^2 \end{bmatrix} \tag{8.77}$$

from which we conclude that

$$\begin{aligned} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} &= \frac{1}{(\sigma_Y^2 + \sigma_R^2)^2 - \sigma_Y^4} \begin{bmatrix} \sigma_Y^2 + \sigma_R^2 & -\sigma_Y^2 \\ -\sigma_Y^2 & \sigma_Y^2 + \sigma_R^2 \end{bmatrix} \begin{bmatrix} \sigma_Y^2 \\ \sigma_Y^2 \end{bmatrix} \\ &= \frac{\sigma_Y^2}{2\sigma_Y^2 + \sigma_R^2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} . \end{aligned} \tag{8.78}$$

Finally, therefore,

$$\widehat{Y}_\ell = \frac{1}{2\sigma_Y^2 + \sigma_R^2}(\sigma_R^2 \mu_Y + \sigma_Y^2 X_1 + \sigma_Y^2 X_2) \tag{8.79}$$

and applying (8.73) we get that the associated minimum mean square error (MMSE) is

$$\frac{\sigma_Y^2 \sigma_R^2}{2\sigma_Y^2 + \sigma_R^2} . \tag{8.80}$$

One can easily check that both the estimator and the associated MMSE take reasonable values at extreme ranges of the signal-to-noise ratio σ_Y^2/σ_R^2.

CHAPTER 9

Random Processes

INTRODUCTION

Much of your background in signals and systems is assumed to have focused on the effect of LTI systems on deterministic signals, developing tools for analyzing this class of signals and systems, and using what you learned in order to understand applications in communication (e.g., AM and FM modulation), control (e.g., stability of feedback systems), and signal processing (e.g., filtering). It is important to develop a comparable understanding and associated tools for treating the effect of LTI systems on signals modeled as the outcome of probabilistic experiments, i.e., a class of signals referred to as random signals (alternatively referred to as random processes or stochastic processes). Such signals play a central role in signal and system design and analysis, and throughout the remainder of this text. In this chapter we define random processes via the associated ensemble of signals, and begin to explore their properties. In successive chapters we use random processes as models for random or uncertain signals that arise in communication, control and signal processing applications.

9.1 DEFINITION AND EXAMPLES OF A RANDOM PROCESS

In Section 7.3 we defined a random variable X as a function that maps each outcome of a probabilistic experiment to a real number. In a similar manner, a real-valued CT or DT random process, $X(t)$ or $X[n]$ respectively, is a function that maps each outcome of a probabilistic experiment to a real CT or DT signal respectively, termed the realization of the random process in that experiment. For any fixed time instant $t = t_0$ or $n = n_0$, the quantities $X(t_0)$ and $X[n_0]$ are just random variables. The collection of signals that can be produced by the random process is referred to as the ensemble of signals in the random process.

EXAMPLE 9.1 Random Oscillators

As an example of a random process, imagine a warehouse containing N harmonic oscillators, each producing a sinusoidal waveform of some specific amplitude, frequency, and phase, all of which may be different for the different oscillators. The probabilistic experiment that results in the ensemble of signals consists of selecting an oscillator according to some probability mass function (PMF) that assigns a probability to each of the numbers from 1 to N, so that the ith oscillator is picked

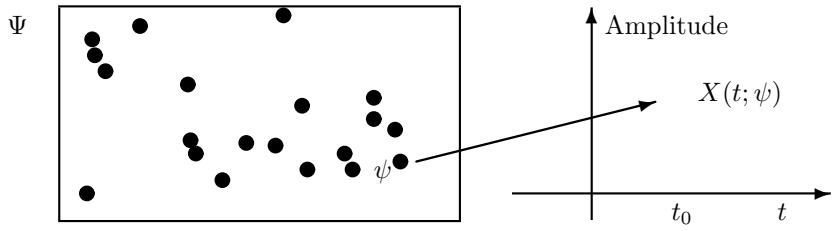

FIGURE 9.1 A random process.

with probability p_i. Associated with each outcome of this experiment is a specific sinusoidal waveform.

In Example 9.1, before an oscillator is chosen, there is uncertainty about what the amplitude, frequency and phase of the outcome of the experiment will be. Consequently, for this example, we might express the random process as

$$X(t) = A\sin(\omega t + \phi)$$

where the amplitude A, frequency ω and phase ϕ are all random variables. The value $X(t_1)$ at some specific time t_1 is also a random variable. In the context of this experiment, knowing the PMF associated with each of the numbers 1 to N involved in choosing an oscillator, as well as the specific amplitude, frequency and phase of each oscillator, we could determine the probability distributions of any of the underlying random variables A, ω, ϕ or $X(t_1)$ mentioned above.

Throughout this and later chapters, we will be considering many other examples of random processes. What is important at this point, however, is to develop a good mental picture of what a random process is. A random process is not just one signal but rather an ensemble of signals, as illustrated schematically in Figure 9.2 below, for which the outcome of the probabilistic experiment could be any of the four waveforms indicated. Each waveform is deterministic, but the process is probabilistic or random because it is not known *a priori* which waveform will be generated by the probabilistic experiment. Consequently, prior to obtaining the outcome of the probabilistic experiment, many aspects of the signal are unpredictable, since there is uncertainty associated with which signal will be produced. After the experiment, or *a posteriori*, the outcome is totally determined.

If we focus on the values that a random process $X(t)$ can take at a particular instant of time, say t_1 — i.e., if we look down the entire ensemble at a fixed time — what we have is a random variable, namely $X(t_1)$. If we focus on the ensemble of values taken at an arbitrary collection of ℓ fixed time instants $t_1 < t_2 < \cdots < t_\ell$ for some arbitrary integer ℓ, we are dealing with a set of ℓ jointly distributed random variables $X(t_1)$, $X(t_2)$, $\cdots$, $X(t_\ell)$, all determined together by the outcome of the underlying probabilistic experiment. From this point of view, a random process

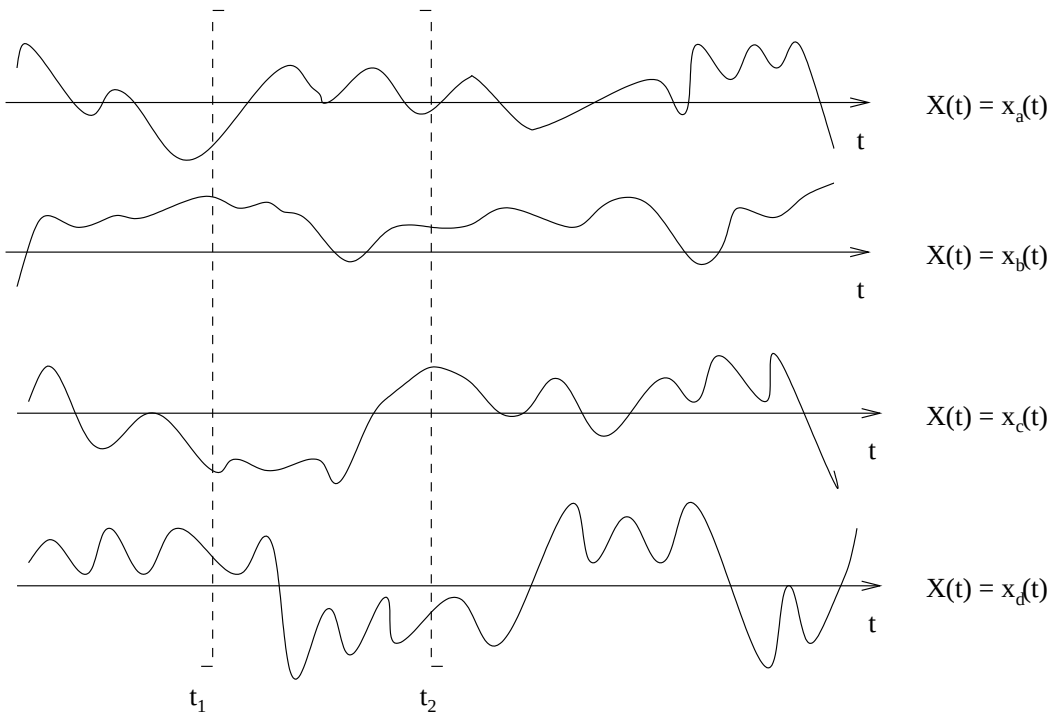

FIGURE 9.2 Realizations of the random process $X(t)$

can be thought of as a family of jointly distributed random variables indexed by t (or n in the DT case). A full probabilistic characterization of this collection of random variables would require the joint PDFs of multiple samples of the signal, taken at arbitrary times:

$$f_{X(t_1),X(t_2),\cdots,X(t_\ell)}(x_1, x_2, \cdots, x_\ell)$$

for all ℓ and all $t_1, t_2, \cdots, t_\ell$.

An important set of questions that arises as we work with random processes in later chapters of this book is whether, by observing just part of the outcome of a random process, we can determine the complete outcome. The answer will depend on the details of the random process, but in general the answer is no. For some random processes, having observed the outcome in a given time interval might provide sufficient information to know exactly which ensemble member was determined. In other cases it would not be sufficient. We will be exploring some of these aspects in more detail later, but we conclude this section with two additional examples that

further emphasize these points.

EXAMPLE 9.2 Ensemble of batteries

Consider a collection of N batteries, each providing one voltage out of a given finite set of voltage values. The histogram of voltages (i.e., the number of batteries with a given voltage) is given in Figure 9.3. The probabilistic experiment is to choose

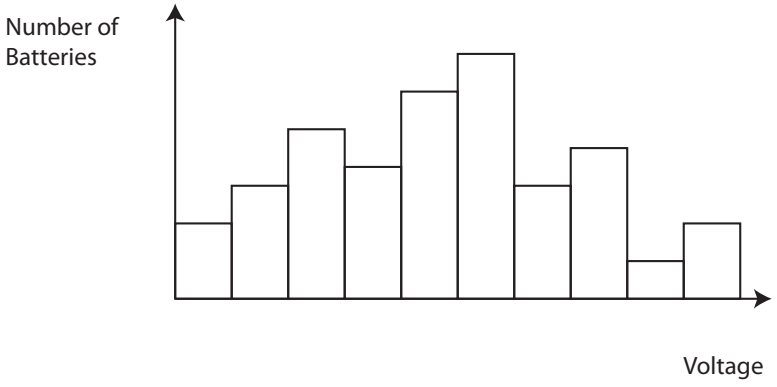

FIGURE 9.3 Histogram of battery distribution for Example 9.2.

one of the batteries, with the probability of picking any specific one being $\frac{1}{N}$, i.e., they are all equally likely to be picked. A little reflection should convince you that if we multiply the histogram in Figure 9.3 by $\frac{1}{N}$, this normalized histogram will represent (or approximate) the PMF for the battery voltage at the outcome of the experiment. Since the battery voltage is a constant signal, this corresponds to a random process, and in fact is similar to the oscillator example discussed earlier, but with $\omega = 0$ and $\phi = 0$, so that only the amplitude is random.

For this example observation of $X(t)$ at any one time is sufficient information to determine the outcome for all time.

EXAMPLE 9.3 Ensemble of coin tossers

Consider N people, each independently having written down a long random string of ones and zeros, with each entry chosen independently of any other entry in their string (similar to a sequence of independent coin tosses). The random process now comprises this ensemble of strings. A realization of the process is obtained by randomly selecting a person (and therefore one of the N strings of ones and zeros), following which the specific ensemble member of the random process is totally determined. The random process described in this example is often referred to as

the Bernoulli process because of the way in which the string of ones and zeros is generated (by independent coin flips).

Now suppose that person shows you only the tenth entry in the string. Can you determine (or predict) the eleventh entry from just that information? Because of the manner in which the string was generated, the answer clearly is no. Similarly if the entire past history up to the tenth entry was revealed to you, could you determine the remaining sequence beyond the tenth? For this example, the answer is again clearly no.

While the entire sequence has been determined by the nature of the experiment, partial observation of a given ensemble member is in general not sufficient to fully specify that member.

Rather than looking at the nth entry of a single ensemble member, we can consider the random variable corresponding to the values from the entire ensemble at the nth entry. Looking down the ensemble at $n = 10$, for example, we would would see ones and zeros with equal probability.

In the above discussion we indicated and emphasized that a random process can be thought of as a family of jointly distributed random variables indexed by t or n. Obviously it would in general be extremely difficult or impossible to represent a random process this way. Fortunately, the most widely used random process models have special structure that permits computation of such a statistical specification. Also, particularly when we are processing our signals with linear systems, we often design the processing or analyze the results by considering only the first and second moments of the process, namely the following functions:

$$\textit{Mean:} \qquad \mu_X(t_i) = E[X(t_i)], \tag{9.1}$$

$$\textit{Auto-correlation:} \quad R_{XX}(t_i, t_j) = E[X(t_i)X(t_j)], \text{ and} \tag{9.2}$$

$$\textit{Auto-covariance:} \quad C_{XX}(t_i, t_j) = E[(X(t_i) - \mu_X(t_i))(X(t_j) - \mu_X(t_j))]$$
$$= R_{XX}(t_i, t_j) - \mu_X(t_i)\mu_X(t_j). \tag{9.3}$$

The word "auto" (which is sometime written without the hyphen, and sometimes dropped altogether to simplify the terminology) here refers to the fact that both samples in the correlation function or the covariance function come from the same process; we shall shortly encounter an extension of this idea, where the samples are taken from two different processes.

One case in which the first and second moments actually suffice to completely specify the process is in the case of what is called a Gaussian process, defined as a process whose samples are always jointly Gaussian (the generalization of the bivariate Gaussian to many variables).

We can also consider multiple random processes, e.g., two processes, $X(t)$ and $Y(t)$. For a full stochastic characterization of this, we need the PDFs of all possible combinations of samples from $X(t), Y(t)$. We say that $X(t)$ and $Y(t)$ are independent if every set of samples from $X(t)$ is independent of every set of samples from $Y(t)$,

so that the joint PDF factors as follows:

$$\begin{aligned} &f_{X(t_1),\cdots,X(t_k),Y(t'_1),\cdots,Y(t'_\ell)}(x_1,\cdots,x_k,y_1,\cdots,y_\ell) \\ &= f_{X(t_1),\cdots,X(t_k)}(x_1,\cdots,x_k).f_{Y(t'_1),\cdots,Y(t'_\ell)}(y_1,\cdots,y_\ell)\ . \end{aligned} \tag{9.4}$$

If only first and second moments are of interest, then in addition to the individual first and second moments of $X(t)$ and $Y(t)$ respectively, we need to consider the cross-moment functions:

$$\textit{Cross-correlation: } R_{XY}(t_i,t_j) = E[X(t_i)Y(t_j)], \text{ and} \tag{9.5}$$

$$\begin{aligned} \textit{Cross-covariance: } C_{XY}(t_i,t_j) &= E[(X(t_i)-\mu_X(t_i))(Y(t_j)-\mu_Y(t_j))] \\ &= R_{XY}(t_i,t_j) - \mu_X(t_i)\mu_Y(t_j). \end{aligned} \tag{9.6}$$

If $C_{XY}(t_1,t_2) = 0$ for all t_1, t_2, we say that the processes $X(t)$ and $Y(t)$ are uncorrelated. Note again that the term "uncorrelated" in its common usage means that the processes have zero covariance rather than zero correlation.

Note that everything we have said above can be carried over to the case of DT random processes, except that now the sampling instants are restricted to be discrete time instants. In accordance with our convention of using square brackets $[\cdot]$ around the time argument for DT signals, we will write $\mu_X[n]$ for the mean of a random process $X[\cdot]$ at time n; similarly, we will write $R_{XX}[n_i,n_j]$ for the correlation function involving samples at times n_i and n_j; and so on.

9.2 STRICT-SENSE STATIONARITY

In general, we would expect that the joint PDFs associated with the random variables obtained by sampling a random process at an arbitrary number k of arbitrary times will be time-dependent, i.e., the joint PDF

$$f_{X(t_1),\cdots,X(t_k)}(x_1,\cdots,x_k)$$

will depend on the specific values of $t_1,\cdots,t_k$. If all the joint PDFs stay the same under arbitrary time *shifts*, i.e., if

$$f_{X(t_1),\cdots,X(t_k)}(x_1,\cdots,x_k) = f_{X(t_1+\tau),\cdots,X(t_k+\tau)}(x_1,\cdots,x_k) \tag{9.7}$$

for arbitrary τ, then the random process is said to be strict-sense stationary (SSS). Said another way, for a strict-sense stationary process, the statistics depend only on the relative times at which the samples are taken, not on the absolute times.

EXAMPLE 9.4 Representing an i.i.d. process

Consider a DT random process whose values $X[n]$ may be regarded as independently chosen at each time n from a fixed PDF $f_X(x)$, so the values are independent and identically distributed, thereby yielding what is called an i.i.d. process. Such processes are widely used in modeling and simulation. For instance, if a particular

DT communication channel corrupts a transmitted signal with added noise that takes independent values at each time instant, but with characteristics that seem unchanging over the time window of interest, then the noise may be well modeled as an i.i.d. process. It is also easy to generate an i.i.d. process in a simulation environment, provided one can arrange a random-number generator to produce samples from a specified PDF (and there are several good ways to do this). Processes with more complicated dependence across time samples can then be obtained by filtering or other operations on the i.i.d. process, as we shall see in the next chapter.

For such an i.i.d. process, we can write the joint PDF quite simply:

$$f_{X[n_1],X[n_2],\cdots,X[n_\ell]}(x_1,x_2,\cdots,x_\ell) = f_X(x_1)f_X(x_2)\cdots f_X(x_\ell) \tag{9.8}$$

for any choice of ℓ and $n_1,\cdots,n_\ell$. The process is clearly SSS.

9.3 WIDE-SENSE STATIONARITY

Of particular use to us is a less restricted type of stationarity. Specifically, if the mean value $\mu_X(t_i)$ is independent of time and the autocorrelation $R_{XX}(t_i,t_j)$ or equivalently the autocovariance $C_{XX}(t_i,t_j)$ is dependent only on the time difference $(t_i - t_j)$, then the process is said to be wide-sense stationary (WSS). Clearly a process that is SSS is also WSS. For a WSS random process $X(t)$, therefore, we have

$$\mu_X(t) = \mu_X \tag{9.9}$$

$$\begin{aligned} R_{XX}(t_1,t_2) &= R_{XX}(t_1+\alpha,t_2+\alpha) \text{ for every } \alpha \\ &= R_{XX}(t_1-t_2,0)\ . \end{aligned} \tag{9.10}$$

(Note that for a Gaussian process (i.e., a process whose samples are always jointly Gaussian) WSS implies SSS, because jointly Gaussian variables are entirely determined by the their joint first and second moments.)

Two random processes $X(t)$ and $Y(t)$ are jointly WSS if their first and second moments (including the cross-covariance) are stationary. In this case we use the notation $R_{XY}(\tau)$ to denote $E[X(t+\tau)Y(t)]$.

EXAMPLE 9.5 Random Oscillators Revisited

Consider again the harmonic oscillators as introduced in Example 9.1, i.e.

$$X(t;A,\Theta) = A\cos(\omega_0 t + \Theta)$$

where A and Θ are independent random variables, and now ω_0 is fixed at some known value.

If Θ is actually fixed at the constant value θ_0, then every outcome is of the form $x(t) = A\cos(\omega_0 t + \theta_0)$, and it is straightforward to see that this process is not WSS

(and hence not SSS). For instance, if A has a nonzero mean value, $\mu_A \neq 0$, then the expected value of the process, namely $\mu_A \cos(\omega_0 t + \theta_0)$, is time varying. To argue that the process is not WSS even when $\mu_A = 0$, we can examine the autocorrelation function. Note that $x(t)$ is fixed at the value 0 for all values of t such that $\omega_0 t + \theta_0$ is an odd multiple of $\pi/2$, and takes the values $\pm A$ half-way between such points; the correlation between such samples taken π/ω_0 apart in time can correspondingly be 0 (in the former case) or $-E[A^2]$ (in the latter). The process is thus not WSS.

On the other hand, if Θ is distributed uniformly in $[-\pi, \pi]$, then

$$\mu_X(t) = \mu_A \int_{-\pi}^{\pi} \frac{1}{2\pi} \cos(\omega_0 t + \theta) d\theta = 0 \,, \tag{9.11}$$

$$\begin{aligned} C_{XX}(t_1, t_2) &= R_{XX}(t_1, t_2) \\ &= E[A^2] E[\cos(\omega_0 t_1 + \Theta) \cos(\omega_0 t_2 + \Theta)] \\ &= \frac{E[A^2]}{2} \cos(\omega_0 (t_2 - t_1)) \,, \end{aligned} \tag{9.12}$$

so the process is WSS. It can also be shown to be SSS, though this is not totally straightforward to show formally.

To simplify notation for a WSS process, we write the correlation function as $R_{XX}(t_1 - t_2)$; the argument $t_1 - t_2$ is referred to as the lag at which the correlation is computed. For the most part, the random processes that we treat will be WSS processes. When considering just first and second moments and not entire PDFs or CDFs, it will be less important to distinguish between the random process $X(t)$ and a specific realization $x(t)$ of it — so we shall go one step further in simplifying notation, by using lower case letters to denote the random process itself. We shall thus talk of the random process $x(t)$, and — in the case of a WSS process — denote its mean by μ_x and its correlation function $E\{x(t+\tau)x(t)\}$ by $R_{xx}(\tau)$. Correspondingly, for DT we'll refer to the random process $x[n]$ and (in the WSS case) denote its mean by μ_x and its correlation function $E\{x[n+m]x[n]\}$ by $R_{xx}[m]$.

9.3.1 Some Properties of WSS Correlation and Covariance Functions

It is easily shown that for real-valued WSS processes $x(t)$ and $y(t)$ the correlation and covariance functions have the following symmetry properties:

$$R_{xx}(\tau) = R_{xx}(-\tau) \,, \qquad C_{xx}(\tau) = C_{xx}(-\tau) \tag{9.13}$$

$$R_{xy}(\tau) = R_{yx}(-\tau) \,, \qquad C_{xy}(\tau) = C_{yx}(-\tau) \,. \tag{9.14}$$

We see from (9.13) that the autocorrelation and autocovariance have even symmetry. Similar properties hold for DT WSS processes.

Another important property of correlation and covariance functions follows from noting that the correlation coefficient of two random variables has magnitude not

exceeding 1. Applying this fact to the samples $x(t)$ and $x(t+\tau)$ of the random process $x(\cdot)$ directly leads to the conclusion that

$$-C_{xx}(0) \leq C_{xx}(\tau) \leq C_{xx}(0) . \tag{9.15}$$

In other words, the autocovariance function never exceeds in magnitude its value at the origin. Adding μ_x^2 to each term above, we find the following inequality holds for correlation functions:

$$-R_{xx}(0) + 2\mu_x^2 \leq R_{xx}(\tau) \leq R_{xx}(0) . \tag{9.16}$$

In Chapter 10 we will demonstrate that correlation and covariance functions are characterized by the property that their Fourier transforms are real and nonnegative at all frequencies, because these transforms describe the frequency distribution of the expected power in the random process. The above symmetry constraints and bounds will then follow as natural consequences, but they are worth highlighting here already.

9.4 SUMMARY OF DEFINITIONS AND NOTATION

In this section we summarize some of the definitions and notation we have previously introduced. As in Section 9.3, we shall use lower case letters to denote random processes, since we will only be dealing with expectations and not densities. Thus, with $x(t)$ and $y(t)$ denoting (real) random processes, we summarize the following definitions:

$$\textit{mean}: \quad \mu_x(t) \triangleq E\{x(t)\} \tag{9.17}$$

$$\textit{autocorrelation}: \quad R_{xx}(t_1,t_2) \triangleq E\{x(t_1)x(t_2)\} \tag{9.18}$$

$$\textit{cross}-\textit{correlation}: \quad R_{xy}(t_1,t_2) \triangleq E\{x(t_1)y(t_2)\} \tag{9.19}$$

$$\begin{aligned} \textit{autocovariance}: \quad & C_{xx}(t_1,t_2) \triangleq E\{[x(t_1)-\mu_x(t_1)][x(t_2)-\mu_x(t_2)]\} \\ & = R_{xx}(t_1,t_2) - \mu_x(t_1)\mu_x(t_2) \end{aligned} \tag{9.20}$$

$$\begin{aligned} \textit{cross}-\textit{covariance}: \quad & C_{xy}(t_1,t_2) \triangleq E\{[x(t_1)-\mu_x(t_1)][y(t_2)-\mu_y(t_2)]\} \\ & = R_{xy}(t_1,t_2) - \mu_x(t_1)\mu_y(t_2) \end{aligned} \tag{9.21}$$

strict-sense stationary (SSS):	all joint statistics for $x(t_1), x(t_2), \ldots, x(t_\ell)$ for all $\ell > 0$ and all choices of sampling instants $t_1, \cdots, t_\ell$ depend only on the *relative* locations of sampling instants.
wide-sense stationary (WSS):	$\mu_x(t)$ is constant at some value μ_x, and $R_{xx}(t_1, t_2)$ is a function of $(t_1 - t_2)$ only, denoted in this case simply by $R_{xx}(t_1 - t_2)$; hence $C_{xx}(t_1, t_2)$ is a function of $(t_1 - t_2)$ only, and written as $C_{xx}(t_1 - t_2)$.
jointly wide-sense stationary:	$x(t)$ and $y(t)$ are individually WSS and $R_{xy}(t_1, t_2)$ is a function of $(t_1 - t_2)$ only, denoted simply by $R_{xy}(t_1 - t_2)$; hence $C_{xy}(t_1, t_2)$ is a function of $(t_1 - t_2)$ only, and written as $C_{xy}(t_1 - t_2)$.

For WSS processes we have, in continuous-time and with simpler notation,

$$R_{xx}(\tau) = E\{x(t+\tau)x(t)\} = E\{x(t)x(t-\tau)\} \tag{9.22}$$
$$R_{xy}(\tau) = E\{x(t+\tau)y(t)\} = E\{x(t)y(t-\tau)\}, \tag{9.23}$$

and in discrete-time,

$$R_{xx}[m] = E\{x[n+m]x[n]\} = E\{x[n]x[n-m]\} \tag{9.24}$$
$$R_{xy}[m] = E\{x[n+m]y[n]\} = E\{x[n]y[n-m]\}. \tag{9.25}$$

We use corresponding (centered) definitions and notation for *covariances*:

$$C_{xx}(\tau), C_{xy}(\tau), C_{xx}[m], \text{ and } C_{xy}[m] \ .$$

It is worth noting that an alternative convention used elsewhere is to define $R_{xy}(\tau)$ as $R_{xy}(\tau) \triangleq E\{x(t)y(t+\tau)\}$. In our notation, this expectation would be denoted by $R_{xy}(-\tau)$. It's important to be careful to take account of what notational convention is being followed when you read this material elsewhere, and you should also be clear about what notational convention we are using in this text.

9.5 FURTHER EXAMPLES

EXAMPLE 9.6 Bernoulli process

The Bernoulli process, a specific example of which was discussed previously in Example 9.3, is an example of an i.i.d. DT process with

$$\mathrm{P}(x[n] = 1) = p \tag{9.26}$$
$$\mathrm{P}(x[n] = -1) = (1 - p) \tag{9.27}$$

and with the value at each time instant n independent of the values at all other

time instants. A simple calculation results in

$$E\{x[n]\} = 2p - 1 = \mu_x \tag{9.28}$$

$$E\{x[n+m]x[n]\} = \begin{cases} 1 & m = 0 \\ (2p-1)^2 & m \neq 0 \end{cases} \tag{9.29}$$

$$C_{xx}[m] = E\{(x[n+m] - \mu_x)(x[n] - \mu_x)\} \tag{9.30}$$

$$= \{1 - (2p-1)^2\}\delta[m] = 4p(1-p)\delta[m]\ . \tag{9.31}$$

EXAMPLE 9.7 Random telegraph wave

A useful example of a CT random process that we'll make occasional reference to is the random telegraph wave. A representative sample function of a random telegraph wave process is shown in Figure 9.4. The random telegraph wave can be defined through the following two properties:

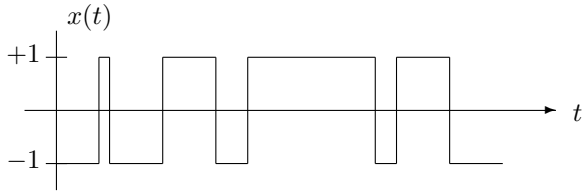

FIGURE 9.4 One realization of a random telegraph wave.

1. $X(0) = \pm 1$ with probability 0.5.
2. $X(t)$ changes polarity at Poisson times, i.e., the probability of k sign changes in a time interval of length T is

$$\mathrm{P}(k \text{ sign changes in an interval of length } T) = \frac{(\lambda T)^k e^{-\lambda T}}{k!}\ . \tag{9.32}$$

Property 2 implies that the probability of a non-negative, even number of sign changes in an interval of length T is

$$\mathrm{P}(\text{a non-negative even \# of sign changes}) = \sum_{\substack{k=0 \\ k \text{ even}}}^{\infty} \frac{(\lambda T)^k e^{-\lambda T}}{k!} = e^{-\lambda T} \sum_{k=0}^{\infty} \frac{1 + (-1)^k}{2} \frac{(\lambda T)^k}{k!} \tag{9.33}$$

Using the identity

$$e^{\lambda T} = \sum_{k=0}^{\infty} \frac{(\lambda T)^k}{k!}$$

equation (9.33) becomes

$$\begin{aligned}\mathrm{P}(\text{a non-negative even \# of sign changes}) &= e^{-\lambda T}\,\frac{(e^{\lambda T}+e^{-\lambda T})}{2}\\ &= \frac{1}{2}(1+e^{-2\lambda T})\,. \end{aligned} \tag{9.34}$$

Similarly, the probability of an odd number of sign changes in an interval of length T is $\frac{1}{2}(1-e^{-2\lambda T})$. It follows that

$$\begin{aligned}\mathrm{P}(X(t)=1) &= \mathrm{P}(X(t)=1|X(0)=1)\mathrm{P}(X(0)=1)\\ &\quad+\mathrm{P}(X(t)=1|X(0)=-1)\mathrm{P}(X(0)=-1)\\ &= \frac{1}{2}\mathrm{P}(\text{even \# of sign changes in } [0,t])\\ &\quad+\frac{1}{2}\mathrm{P}(\text{odd \# of sign changes in } [0,t])\\ &= \frac{1}{2}\left\{\frac{1}{2}(1+e^{-2\lambda t})\right\}+\frac{1}{2}\left\{\frac{1}{2}(1-e^{-2\lambda t})\right\}=\frac{1}{2}\,. \end{aligned} \tag{9.35}$$

Note that because of Property I, the expression in the last line of Eqn. (9.35) is not needed, since the line before that already allows us to conclude that the answer is $\frac{1}{2}$: since the number of sign changes in any interval must be either even or odd, their probabilities add up to 1, so $\mathrm{P}\left(X(t)=1\right)=\frac{1}{2}$. However, if Property 1 is relaxed to allow $\mathrm{P}(X(0)=1)=p_0\neq\frac{1}{2}$, then the above computation must be carried through to the last line, and yields the result

$$\mathrm{P}(X(t)=1)=p_0\left\{\frac{1}{2}(1+e^{-2\lambda t})\right\}+(1-p_0)\left\{\frac{1}{2}(1-e^{-2\lambda t})\right\}=\frac{1}{2}\left\{1+(2p_0-1)e^{-2\lambda t}\right\}\,. \tag{9.36}$$

Returning to the case where Property 1 holds, so $\mathrm{P}(X(t)=1)$, we get

$$\mu_X(t)=0,\ \text{and} \tag{9.37}$$

$$\begin{aligned}R_{XX}(t_1,t_2) &= E[X(t_1)X(t_2)]\\ &= 1\times\mathrm{P}\left(X(t_1)=X(t_2)\right)+(-1)\times\mathrm{P}\left(X(t_1)\neq X(t_2)\right)\\ &= e^{-2\lambda|t_2-t_1|}\,. \end{aligned} \tag{9.38}$$

In other words, the process is exponentially correlated and WSS.

9.6 ERGODICITY

The concept of ergodicity is sophisticated and subtle, but the essential idea is described here. We typically observe the outcome of a random process (e.g., we record a noise waveform) and want to characterize the statistics of the random process by measurements on one ensemble member. For instance, we could consider the time-average of the waveform to represent the mean value of the process (assuming this

mean is constant for all time). We could also construct histograms that represent the fraction of time (rather than the probability-weighted fraction of the ensemble) that the waveform lies in different amplitude bins, and this could be taken to reflect the probability density across the ensemble of the value obtained at a particular sampling time. If the random process is such that the behavior of almost every particular realization over time is representative of the behavior down the ensemble, then the process is called ergodic.

A simple example of a process that is not ergodic is Example 9.2, an ensemble of batteries. Clearly, for this example, the behavior of any realization is not representative of the behavior down the ensemble.

Narrower notions of ergodicity may be defined. For example, if the time average

$$\langle x \rangle = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t)\, dt \tag{9.39}$$

almost always (i.e. for almost every realization or outcome) equals the ensemble average μ_X, then the process is termed *ergodic in the mean*. It can be shown, for instance, that a WSS process with finite variance at each instant and with a covariance function that approaches 0 for large lags is ergodic in the mean. Note that a (nonstationary) process with time-varying mean cannot be ergodic in the mean.

In our discussion of random processes, we will primarily be concerned with first- and second-order moments of random processes. While it is extremely difficult to determine in general whether a random process is ergodic, there are criteria (specified in terms of the moments of the process) that will establish ergodicity in the mean and in the autocorrelation. Frequently, however, such ergodicity is simply assumed for convenience, in the absence of evidence that the assumption is not reasonable. Under this assumption, the mean and autocorrelation can be obtained from time-averaging on a single ensemble member, through the following equalities:

$$E\{x(t)\} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t) dt \tag{9.40}$$

and

$$E\{x(t)x(t+\tau)\} = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t)x(t+\tau) dt \tag{9.41}$$

A random process for which (9.40) and (9.41) are true is referred as second-order ergodic.

9.7 LINEAR ESTIMATION OF RANDOM PROCESSES

A common class of problems in a variety of aspects of communication, control and signal processing involves the estimation of one random process from observations

of another, or estimating (predicting) future values from the observation of past values. For example, it is common in communication systems that the signal at the receiver is a corrupted (e.g., noisy) version of the transmitted signal, and we would like to estimate the transmitted signal from the received signal. Other examples lie in predicting weather and financial data from past observations. We will be treating this general topic in much more detail in later chapters, but a first look at it here can be beneficial in understanding random processes.

We shall first consider a simple example of linear prediction of a random process, then a more elaborate example of linear FIR filtering of a noise-corrupted process to estimate the underlying random signal. We conclude the section with some further discussion of the basic problem of linear estimation of one random variable from measurements of another.

9.7.1 Linear Prediction

As a simple illustration of linear prediction, consider a discrete-time process $x[n]$. Knowing the value at time n_0 we may wish to predict what the value will be m samples into the future, i.e. at time $n_0 + m$. We limit the prediction strategy to a linear one, i.e., with $\hat{x}[n_0 + m]$ denoting the predicted value, we restrict $\hat{x}[n_0 + m]$ to be of the form

$$\hat{x}[n_0 + m] = ax[n_0] + b \tag{9.42}$$

and choose the prediction parameters a and b to minimize the expected value of the square of the error, i.e., choose a and b to minimize

$$\epsilon = E\{(x[n_0 + m] - \hat{x}[n_0 + m])^2\} \tag{9.43}$$

or

$$\epsilon = E\{(x[n_0 + m] - ax[n_0] - b)^2\}. \tag{9.44}$$

To minimize ϵ we set to zero its partial derivative with respect to each of the two parameters and solve for the parameter values. The resulting equations are

$$E\{(x[n_0 + m] - ax[n_0] - b)x[n_0]\} = E\{(x[n_0 + m] - \widehat{x}[n_0 + m])x[n_0]\} = 0 \tag{9.45a}$$

$$E\{x[n_0 + m] - ax[n_0] - b\} = E\{x[n_0 + m] - \widehat{x}[n_0 + m]\} = 0\,. \tag{9.45b}$$

Equation (9.45a) states that the error $x[n_0 + m] - \widehat{x}[n_0 + m]$ associated with the optimal estimate is orthogonal to the available data $x[n_0]$. Equation (9.45b) states that the estimate is unbiased.

Carrying out the multiplications and expectations in the preceding equations results in the following equations, which can be solved for the desired constants.

$$R_{xx}[n_0 + m, n_0] - aR_{xx}[n_0, n_0] - b\mu_x[n_0] = 0 \tag{9.46a}$$

$$\mu_x[n_0 + m] - a\mu_x[n_0] - b = 0. \tag{9.46b}$$

If we assume that the process is WSS so that $R_{xx}[n_0+m, n_0] = R_{xx}[m]$, $R_{xx}[n_0, n_0] = R_{xx}[0]$, and also assume that it is zero mean, ($\mu_x = 0$), then equations (9.46) reduce to

$$a = R_{xx}[m]/R_{xx}[0] \tag{9.47}$$

$$b = 0 \tag{9.48}$$

so that

$$\widehat{x}[n_0 + m] = \frac{R_{xx}[m]}{R_{xx}[0]} x[n_0]. \tag{9.49}$$

If the process is not zero mean, then it is easy to see that

$$\widehat{x}[n_0 + m] = \mu_x + \frac{C_{xx}[m]}{C_{xx}[0]} (x[n_0] - \mu_x) \ . \tag{9.50}$$

An extension of this problem would consider how to do prediction when measurements of several past values are available. Rather than pursue this case, we illustrate next what to do with several measurements in a slightly different setting.

9.7.2 Linear FIR Filtering

As another example, which we will treat in more generality in chapter 11 on Wiener filtering, consider a discrete-time signal $s[n]$ that has been corrupted by additive noise $d[n]$. For example, $s[n]$ might be a signal transmitted over a channel and $d[n]$ the noise introduced by the channel. The received signal $r[n]$ is then

$$r[n] = s[n] + d[n]. \tag{9.51}$$

Assume that both $s[n]$ and $d[n]$ are zero-mean random processes and are uncorrelated. At the receiver we would like to process $r[n]$ with a causal FIR (finite impulse response) filter to estimate the transmitted signal $s[n]$.

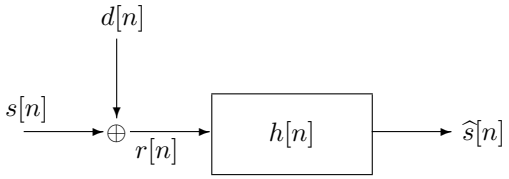

FIGURE 9.5 Estimating the noise corrupted signal.

If $h[n]$ is a causal FIR filter of length L, then

$$\widehat{s}[n] = \sum_{k=0}^{L-1} h[k] r[n-k]. \tag{9.52}$$

We would like to determine the filter coefficients $h[k]$ to minimize the mean square error between $\widehat{s}[n]$ and $s[n]$, i.e., minimize ϵ given by

$$\begin{aligned}\epsilon &= E(s[n]-\widehat{s}[n])^2 \\ &= E(s[n]-\sum_{k=0}^{L-1} h[k]r[n-k])^2. \end{aligned} \tag{9.53}$$

To determine h, we set $\frac{\partial \epsilon}{\partial h[m]} = 0$ for each of the L values of m. Taking this derivative, we get

$$\begin{aligned}\frac{\partial \epsilon}{\partial h[m]} &= -E\{2(s[n]-\sum_{k} h[k]r[n-k])r[n-m]\} \\ &= -E\{2(s[n]-\widehat{s}[n])r[n-m]\} \\ &= 0 \qquad\qquad m=0,1,\cdots,L-1 \end{aligned} \tag{9.54}$$

which is the orthogonality condition we should be expecting: the error $(s[n]-\widehat{s}[n])$ associated with the optimal estimate is orthogonal to the available data, $r[n-m]$.

Carrying out the multiplications in the above equations and taking expectations results in

$$\sum_{k=0}^{L-1} h[k]R_{rr}[m-k] = R_{sr}[m]\ , \quad m=0,1,\cdots,L-1 \tag{9.55}$$

Eqns. (9.55) constitute L equations that can be solved for the L parameters $h[k]$. With $r[n] = s[n] + d[n]$, it is straightforward to show that $R_{sr}[m] = R_{ss}[m] + R_{sd}[m]$ and since we assumed that $s[n]$ and $d[n]$ are uncorrelated, then $R_{sd}[m] = 0$. Similarly, $R_{rr}[m] = R_{ss}[m] + R_{dd}[m]$.

These results are also easily modified for the case where the processes no longer have zero mean.

9.8 THE EFFECT OF LTI SYSTEMS ON WSS PROCESSES

Your prior background in signals and systems, and in the earlier chapters of these notes, has characterized how LTI systems affect the input for deterministic signals.

We will see in later chapters how the correlation properties of a random process, and the effects of LTI systems on these properties, play an important role in understanding and designing systems for such tasks as filtering, signal detection, signal estimation and system identification. We focus in this section on understanding in the time domain how LTI systems shape the correlation properties of a random process. In Chapter 10 we develop a parallel picture in the frequency domain, after establishing that the frequency distribution of the expected power in a random signal is described by the Fourier transform of the autocorrelation function.

Consider an LTI system whose input is a sample function of a WSS random process $x(t)$, i.e., a signal chosen by a probabilistic experiment from the ensemble that constitutes the random process $x(t)$; more simply, we say that the input is the random

process $x(t)$. The WSS input is characterized by its mean and its autocovariance or (equivalently) autocorrelation function.

Among other considerations, we are interested in knowing when the output process $y(t)$ — i.e., the ensemble of signals obtained as responses to the signals in the input ensemble — will itself be WSS, and want to determine its mean and autocovariance or autocorrelation functions, as well as its cross-correlation with the input process. For an LTI system whose impulse response is $h(t)$, the output $y(t)$ is given by the convolution

$$y(t) = \int_{-\infty}^{+\infty} h(v)x(t-v)dv = \int_{-\infty}^{+\infty} x(v)h(t-v)dv \tag{9.56}$$

for any specific input $x(t)$ for which the convolution is well-defined. The convolution is well-defined if, for instance, the input $x(t)$ is bounded and the system is bounded-input bounded-output (BIBO) stable, i.e. $h(t)$ is absolutely integrable. Figure 9.6 indicates what the two components of the integrand in the convolution integral may look like.

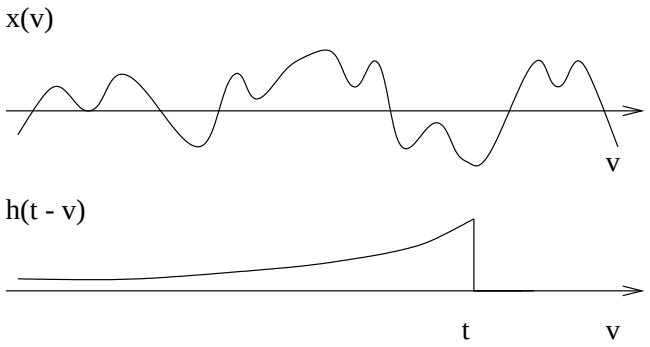

FIGURE 9.6 Illustration of the two terms in the integrand of Eqn. (9.56)

Rather than requiring that every sample function of our input process be bounded, it will suffice for our convolution computations below to assume that $E[x^2(t)] = R_{xx}(0)$ is finite. With this assumption, and also assuming that the system is BIBO stable, we ensure that $y(t)$ is a well-defined random process, and that the formal manipulations we carry out below — for instance, interchanging expectation and convolution — can all be justified more rigorously by methods that are beyond our scope here. In fact, the results we obtain can also be applied, when properly interpreted, to cases where the input process does not have a bounded second moment, e.g., when $x(t)$ is so-called CT white noise, for which $R_{xx}(\tau) = \delta(\tau)$. The results can also be applied to a system that is not BIBO stable, as long as it has a well-defined frequency response $H(j\omega)$, as in the case of an ideal lowpass filter, for example.

We can use the convolution relationship (9.56) to deduce the first- and second-order properties of $y(t)$. What we shall establish is that $y(t)$ is itself WSS, and that

$x(t)$ and $y(t)$ are in fact jointly WSS. We will also develop relationships for the autocorrelation of the output and the cross-correlation between input and output.

First, consider the mean value of the output. Taking the expected value of both sides of (9.56), we find

$$\begin{aligned} E[y(t)] &= E\left\{\int_{-\infty}^{+\infty} h(v)x(t-v)\,dv\right\} \\ &= \int_{-\infty}^{+\infty} h(v)E[x(t-v)]\,dv \\ &= \int_{-\infty}^{+\infty} h(v)\mu_x\,dv \\ &= \mu_x \int_{-\infty}^{+\infty} h(v)\,dv \\ &= H(j0)\,\mu_x = \mu_y\ . \end{aligned} \tag{9.57}$$

In other words, the mean of the output process is *constant*, and equals the mean of the input scaled by the the DC gain of the system. This is also what the response of the system would be if its input were held constant at the value μ_x.

The preceding result and the linearity of the system also allow us to conclude that applying the *zero-mean* WSS process $x(t)-\mu_x$ to the input of the stable LTI system would result in the *zero-mean* process $y(t)-\mu_y$ at the output. This fact will be useful below in converting results that are derived for correlation functions into results that hold for covariance functions.

Next consider the cross-correlation between output and input:

$$\begin{aligned} E\{y(t+\tau)x(t)\} &= E\left\{\left[\int_{-\infty}^{+\infty} h(v)x(t+\tau-v)dv\right]x(t)\right\} \\ &= \int_{-\infty}^{+\infty} h(v)E\{x(t+\tau-v)x(t)\}dv\ . \end{aligned} \tag{9.58}$$

Since $x(t)$ is WSS, $E\{x(t+\tau-v)x(t)\} = R_{xx}(\tau-v)$, so

$$\begin{aligned} E\{y(t+\tau)x(t)\} &= \int_{-\infty}^{+\infty} h(v)R_{xx}(\tau-v)dv \\ &= h(\tau) * R_{xx}(\tau) \\ &= R_{yx}(\tau)\ . \end{aligned} \tag{9.59}$$

Note that the cross-correlation depends only on the lag τ between the sampling instants of the output and input processes, not on both τ and the absolute time location t. Also, this cross-correlation between the output and input is deterministically related to the autocorrelation of the input, and can be viewed as the signal that would result if the system input were the autocorrelation function, as indicated in Figure 9.7.

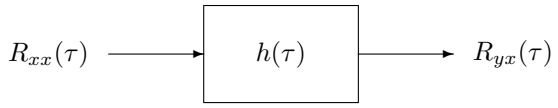

FIGURE 9.7 Representation of Eqn. (9.59)

We can also conclude that

$$R_{xy}(\tau) = R_{yx}(-\tau) = R_{xx}(-\tau) * h(-\tau) = R_{xx}(\tau) * h(-\tau) \ , \tag{9.60}$$

where the second equality follows from Eqn. (9.59) and the fact that time-reversing the two functions in a convolution results in time-reversal of the result, while the last equality follows from the symmetry Eqn. (9.13) of the autocorrelation function.

The above relations can also be expressed in terms of covariance functions, rather than in terms of correlation functions. For this, simply consider the case where the input to the system is the zero-mean WSS process $x(t) - \mu_x$, with corresponding zero-mean output $y(t) - \mu_y$. Since the correlation function for $x(t) - \mu_x$ is the same as the covariance function for $x(t)$, i.e., since

$$R_{x-\mu_x, x-\mu_x}(\tau) = C_{xx}(\tau) \ , \tag{9.61}$$

the results above hold unchanged when every correlation function is replaced by the corresponding covariance function. We therefore have, for instance, that

$$C_{yx}(\tau) = h(\tau) * C_{xx}(\tau) \tag{9.62}$$

Next we consider the autocorrelation of the output $y(t)$:

$$E\{y(t+\tau)y(t)\} = E\left\{ \left[\int_{-\infty}^{+\infty} h(v)x(t+\tau-v)dv \right] y(t) \right\}$$

$$= \int_{-\infty}^{+\infty} h(v) \underbrace{E\{x(t+\tau-v)y(t)\}}_{R_{xy}(\tau-v)} dv$$

$$\begin{aligned} &= \int_{-\infty}^{+\infty} h(v) R_{xy}(\tau - v) dv \\ &= h(\tau) * R_{xy}(\tau) \\ &= R_{yy}(\tau) \ . \end{aligned} \tag{9.63}$$

Note that the autocorrelation of the output depends only on τ, and not on both τ and t. Putting this together with the earlier results, we conclude that $x(t)$ and $y(t)$ are jointly WSS, as claimed.

The corresponding result for covariances is

$$C_{yy}(\tau) = h(\tau) * C_{xy}(\tau) \; . \tag{9.64}$$

Combining (9.63) with (9.60), we find that

$$R_{yy}(\tau) = R_{xx}(\tau) * \underbrace{h(\tau) * h(-\tau)}_{h(\tau)*h(-\tau) \stackrel{\triangle}{=} \overline{R}_{hh}(\tau)} = R_{xx}(\tau) * \overline{R}_{hh}(\tau) \; . \tag{9.65}$$

The function $\overline{R}_{hh}(\tau)$ is typically referred to as the **deterministic autocorrelation function** of $h(t)$, and is given by

$$\overline{R}_{hh}(\tau) = h(\tau) * h(-\tau) = \int_{-\infty}^{+\infty} h(t+\tau)h(t)dt \; . \tag{9.66}$$

For the covariance function version of (9.65), we have

$$C_{yy}(\tau) = C_{xx}(\tau) * \underbrace{h(\tau) * h(-\tau)}_{h(\tau)*h(-\tau) \stackrel{\triangle}{=} \overline{R}_{hh}(\tau)} = C_{xx}(\tau) * \overline{R}_{hh}(\tau) \; . \tag{9.67}$$

Note that the deterministic correlation function of $h(t)$ is still what we use, even when relating the covariances of the input and output. Only the means of the input and output processes get adjusted in arriving at the present result; the impulse response is untouched.

The correlation relations in Eqns. (9.59), (9.60), (9.63) and (9.65), as well as their covariance counterparts, are very powerful, and we will make considerable use of them. Of equal importance are their statements in the Fourier and Laplace transform domains. Denoting the Fourier and Laplace transforms of the correlation function $R_{xx}(\tau)$ by $S_{xx}(j\omega)$ and $S_{xx}(s)$ respectively, and similarly for the other correlation functions of interest, we have:

$$\begin{aligned} S_{yx}(j\omega) &= S_{xx}(j\omega)H(j\omega), & S_{yy}(j\omega) &= S_{xx}(j\omega)|H(j\omega)|^2, \\ S_{yx}(s) &= S_{xx}(s)H(s), & S_{yy}(s) &= S_{xx}(s)H(s)H(-s) \; . \end{aligned} \tag{9.68}$$

We can denote the Fourier and Laplace transforms of the covariance function $C_{xx}(\tau)$ by $D_{xx}(j\omega)$ and $D_{xx}(s)$ respectively, and similarly for the other covariance functions of interest, and then write the same sorts of relationships as above.

Exactly parallel results hold in the DT case. Consider a stable discrete-time LTI system whose impulse response is $h[n]$ and whose input is the WSS random process $x[n]$. Then, as in the continuous-time case, we can conclude that the output process $y[n]$ is *jointly WSS* with the input process $x[n]$, and

$$\mu_y = \mu_x \sum_{-\infty}^{\infty} h[n] \tag{9.69}$$

$$R_{yx}[m] = h[m] * R_{xx}[m] \tag{9.70}$$

$$R_{yy}[m] = R_{xx}[m] * \overline{R}_{hh}[m] \; , \tag{9.71}$$

where $\overline{R}_{hh}[m]$ is the deterministic autocorrelation function of $h[m]$, defined as

$$\overline{R}_{hh}[m] = \sum_{n=-\infty}^{+\infty} h[n+m]h[n] \; . \tag{9.72}$$

The corresponding Fourier and $\mathcal{Z}$-transform statements of these relationships are:

$$\begin{array}{lll} \mu_y = H(e^{j0})\mu_x \; , & S_{yx}(e^{j\Omega}) = S_{xx}(e^{j\Omega})H(e^{j\Omega}) \; , & S_{yy}(e^{j\Omega}) = S_{xx}(e^{j\Omega})|H(e^{j\Omega})|^2, \\ \mu_y = H(1)\mu_x \; , & S_{yx}(z) = S_{xx}(z)H(z) \; , & S_{yy}(z) = S_{xx}(z)H(z)H(1/z). \end{array} \tag{9.73}$$

All of these expressions can also be rewritten for covariances and their transforms.

The basic relationships that we have developed so far in this chapter are extremely powerful. In Chapter 10 we will use these relationships to show that the Fourier transform of the autocorrelation function describes how the expected power of a WSS process is distributed in frequency. For this reason, the Fourier transform of the autocorrelation function is termed the **power spectral density (PSD)** of the process.

The relationships developed in this chapter are also very important in using random processes to measure or identify the impulse response of an LTI system. For example, from (9.70), if the input $x[n]$ to a DT LTI system is a WSS random process with autocorrelation function $R_{xx}[m] = \delta[m]$, then by measuring the cross-correlation between the input and output we obtain a measurement of the system impulse response. It is easy to construct an input process with autocorrelation function $\delta[m]$, for example an i.i.d. process that is equally likely to take the values $+1$ and -1 at each time instant.

As another example, suppose the input $x(t)$ to a CT LTI system is a random telegraph wave, with changes in sign at times that correspond to the arrivals in a Poisson process with rate λ, i.e.,

$$\mathrm{P}(k \text{ switches in an interval of length } T) = \frac{(\lambda T)^k e^{-\lambda T}}{k!} \; . \tag{9.74}$$

Then, assuming $x(0)$ takes the values ± 1 with equal probabilities, we can determine that the process $x(t)$ has zero mean and correlation function $R_{xx}(\tau) = e^{-2\lambda|\tau|}$, so it is WSS (for $t \geq 0$). If we determine the cross-correlation $R_{yx}(\tau)$ with the output $y(t)$ and then use the relation

$$R_{yx}(\tau) = R_{xx}(\tau) * h(\tau) \; , \tag{9.75}$$

we can obtain the system impulse response $h(\tau)$. For example, if $S_{yx}(s), S_{xx}(s)$ and $H(s)$ denote the associated Laplace transforms, then

$$H(s) = \frac{S_{yx}(s)}{S_{xx}(s)} \; . \tag{9.76}$$

Note that $S_{xx}(s)$ is a rather well-behaved function of the complex variable s in this case, whereas any particular sample function of the process $x(t)$ would not have such a well-behaved transform. The same comment applies to $S_{yx}(s)$.

As a third example, suppose that we know the autocorrelation function $R_{xx}[m]$ of the input $x[n]$ to a DT LTI system, but do not have access to $x[n]$ and therefore cannot determine the cross-correlation $R_{yx}[m]$ with the output $y[n]$, but *can* determine the output autocorrelation $R_{yy}[m]$. For example, if

$$R_{xx}[m] = \delta[m] \tag{9.77}$$

and we determine $R_{yy}[m]$ to be $R_{yy}[m] = \left(\frac{1}{2}\right)^{|m|}$, then

$$R_{yy}[m] = \left(\frac{1}{2}\right)^{|m|} = \overline{R}_{hh}[m] = h[m] * h[-m]. \tag{9.78}$$

Equivalently, $H(z)H(z^{-1})$ can be obtained from the $\mathcal{Z}$-transform $S_{yy}(z)$ of $R_{yy}[m]$. Additional assumptions or constraints, for instance on the stability and causality of the system and its inverse, may allow one to recover $H(z)$ from knowledge of $H(z)H(z^{-1})$.

CHAPTER 10

Power Spectral Density

INTRODUCTION

Understanding how the strength of a signal is distributed in the frequency domain, relative to the strengths of other ambient signals, is central to the design of any LTI filter intended to extract or suppress the signal. We know this well in the case of deterministic signals, and it turns out to be just as true in the case of random signals. For instance, if a measured waveform is an audio signal (modeled as a random process since the specific audio signal isn't known) with additive disturbance signals, you might want to build a lowpass LTI filter to extract the audio and suppress the disturbance signals. We would need to decide where to place the cutoff frequency of the filter.

There are two immediate challenges we confront in trying to find an appropriate frequency-domain description for a WSS random process. First, individual sample functions typically don't have transforms that are ordinary, well-behaved functions of frequency; rather, their transforms are only defined in the sense of generalized functions. Second, since the particular sample function is determined as the outcome of a probabilistic experiment, its features will actually be random, so we have to search for features of the transforms that are representative of the whole class of sample functions, i.e., of the random process as a whole.

It turns out that the key is to focus on the *expected power* in the signal. This is a measure of signal strength that meshes nicely with the second-moment characterizations we have for WSS processes, as we show in this chapter. For a process that is second-order ergodic, this will also correspond to the time average power in any realization. We introduce the discussion using the case of CT WSS processes, but the DT case follows very similarly.

10.1 EXPECTED INSTANTANEOUS POWER AND POWER SPECTRAL DENSITY

Motivated by situations in which $x(t)$ is the voltage across (or current through) a unit resistor, we refer to $x^2(t)$ as the *instantaneous power* in the signal $x(t)$. When $x(t)$ is WSS, the *expected* instantaneous power is given by

$$E[x^2(t)] = R_{xx}(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{xx}(j\omega)\,d\omega\,, \tag{10.1}$$

where $S_{xx}(j\omega)$ is the CTFT of the autocorrelation function $R_{xx}(\tau)$. Furthermore, when $x(t)$ is ergodic in correlation, so that time averages and ensemble averages are equal in correlation computations, then (10.1) also represents the time-average power in any ensemble member. Note that since $R_{xx}(\tau) = R_{xx}(-\tau)$, we know $S_{xx}(j\omega)$ is always *real* and *even* in ω; a simpler notation such as $P_{xx}(\omega)$ might therefore have been more appropriate for it, but we shall stick to $S_{xx}(j\omega)$ to avoid a proliferation of notational conventions, and to keep apparent the fact that this quantity is the Fourier transform of $R_{xx}(\tau)$.

The integral above suggests that we might be able to consider the expected (instantaneous) power (or, assuming the process is ergodic, the time-average power) in a frequency band of width $d\omega$ to be given by $(1/2\pi)S_{xx}(j\omega)\,d\omega$. To examine this thought further, consider extracting a band of frequency components of $x(t)$ by passing $x(t)$ through an ideal bandpass filter, shown in Figure 10.1.

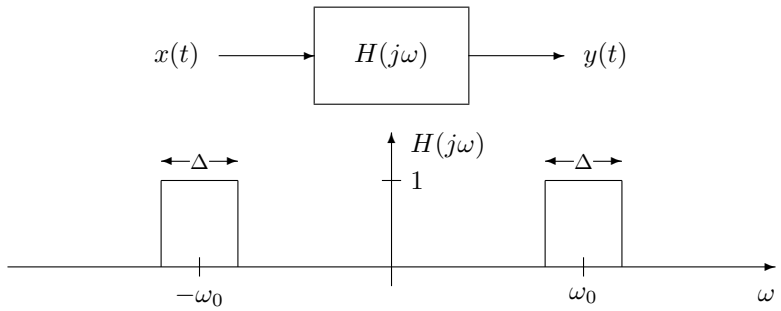

FIGURE 10.1 Ideal bandpass filter to extract a band of frequencies from input, $x(t)$.

Because of the way we are obtaining $y(t)$ from $x(t)$, the expected power in the output $y(t)$ can be interpreted as the expected power that $x(t)$ has in the selected passband. Using the fact that

$$S_{yy}(j\omega) = |H(j\omega)|^2 S_{xx}(j\omega)\ , \tag{10.2}$$

we see that this expected power can be computed as

$$E\{y^2(t)\} = R_{yy}(0) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} S_{yy}(j\omega)\,d\omega = \frac{1}{2\pi}\int_{\text{passband}} S_{xx}(j\omega)\,d\omega\ . \tag{10.3}$$

Thus

$$\frac{1}{2\pi}\int_{\text{passband}} S_{xx}(j\omega)\,d\omega \tag{10.4}$$

is indeed the expected power of $x(t)$ in the passband. It is therefore reasonable to call $S_{xx}(j\omega)$ the **power spectral density (PSD)** of $x(t)$. Note that the instantaneous power of $y(t)$, and hence the expected instantaneous power $E[y^2(t)]$, is always nonnegative, no matter how narrow the passband, It follows that, in addition to being real and even in ω, the PSD is always nonnegative, $S_{xx}(j\omega) \geq 0$ for all ω. While the PSD $S_{xx}(j\omega)$ is the Fourier transform of the autocorrelation function, it

is useful to have a name for the *Laplace* transform of the autocorrelation function; we shall refer to $S_{xx}(s)$ as the *complex* PSD.

Exactly parallel results apply for the DT case, leading to the conclusion that $S_{xx}(e^{j\Omega})$ is the power spectral density of $x[n]$.

10.2 EINSTEIN-WIENER-KHINCHIN THEOREM ON EXPECTED TIME-AVERAGED POWER

The previous section defined the PSD as the transform of the autocorrelation function, and provided an interpretation of this transform. We now develop an alternative route to the PSD. Consider a random realization $x(t)$ of a WSS process. We have already mentioned the difficulties with trying to take the CTFT of $x(t)$ directly, so we proceed indirectly. Let $x_T(t)$ be the signal obtained by windowing $x(t)$, so it equals $x(t)$ in the interval $(-T\,,\,T)$ but is 0 outside this interval. Thus

$$x_T(t) = w_T(t)\,x(t)\;, \tag{10.5}$$

where we define the *window* function $w_T(t)$ to be 1 for $|t| < T$ and 0 otherwise. Let $X_T(j\omega)$ denote the Fourier transform of $x_T(t)$; note that because the signal $x_T(t)$ is nonzero only over the *finite* interval $(-T, T)$, its Fourier transform is typically well defined. We know that the **energy spectral density (ESD)** $\overline{S}_{xx}(j\omega)$ of $x_T(t)$ is given by

$$\overline{S}_{xx}(j\omega) = |X_T(j\omega)|^2 \tag{10.6}$$

and that this ESD is actually the Fourier transform of $x_T(\tau)*x_T^{\leftarrow}(\tau)$, where $x_T^{\leftarrow}(t) = x_T(-t)$. We thus have the CTFT pair

$$x_T(\tau) * x_T^{\leftarrow}(\tau) = \int_{-\infty}^{\infty} w_T(\alpha)w_T(\alpha-\tau)x(\alpha)x(\alpha-\tau)\,d\alpha \Leftrightarrow |X_T(j\omega)|^2\;, \tag{10.7}$$

or, dividing both sides by $2T$ (which is valid, since scaling a signal by a constant scales its Fourier transform by the same amount),

$$\frac{1}{2T}\int_{-\infty}^{\infty} w_T(\alpha)w_T(\alpha-\tau)x(\alpha)x(\alpha-\tau)\,d\alpha \Leftrightarrow \frac{1}{2T}\,|X_T(j\omega)|^2\;. \tag{10.8}$$

The quantity on the right is what we defined (for the DT case) as the **periodogram** of the finite-length signal $x_T(t)$.

Because the Fourier transform operation is linear, the Fourier transform of the expected value of a signal is the expected value of the Fourier transform. We may therefore take expectations of both sides in the preceding equation. Since $E[x(\alpha)x(\alpha-\tau)] = R_{xx}(\tau)$, we conclude that

$$R_{xx}(\tau)\Lambda(\tau) \Leftrightarrow \frac{1}{2T}\,E[|X_T(j\omega)|^2]\;, \tag{10.9}$$

where $\Lambda(\tau)$ is a triangular pulse of height 1 at the origin and decaying to 0 at $|\tau| = 2T$, the result of carrying out the convolution $w_T * w_T^{\leftarrow}(\tau)$ and dividing by

$2T$. Now taking the limit as T goes to ∞, we arrive at the result

$$R_{xx}(\tau) \Leftrightarrow S_{xx}(j\omega) = \lim_{T\to\infty} \frac{1}{2T} E[|X_T(j\omega)|^2] \,. \tag{10.10}$$

This is the **Einstein-Wiener-Khinchin** theorem (proved by Wiener, and independently by Khinchin, in the early 1930's, but — as only recently recognized — stated by Einstein in 1914).

The result is important to us because it underlies a basic method for estimating $S_{xx}(j\omega)$: with a given T, compute the periodogram for several realizations of the random process (i.e., in several independent experiments), and average the results. Increasing the number of realizations over which the averaging is done will reduce the noise in the estimate, while repeating the entire procedure for larger T will improve the frequency resolution of the estimate.

10.2.1 System Identification Using Random Processes as Input

Consider the problem of determining or "identifying" the impulse response $h[n]$ of a stable LTI system from measurements of the input $x[n]$ and output $y[n]$, as indicated in Figure 10.2.

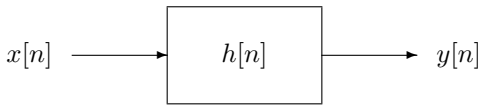

FIGURE 10.2 System with impulse response $h[n]$ to be determined.

The most straightforward approach is to choose the input to be a unit impulse $x[n] = \delta[n]$, and to measure the corresponding output $y[n]$, which by definition is the impulse response. It is often the case in practice, however, that we do not wish to — or are unable to — pick this simple input.

For instance, to obtain a reliable estimate of the impulse response in the presence of measurement errors, we may wish to use a more "energetic" input, one that excites the system more strongly. There are generally limits to the amplitude we can use on the input signal, so to get more energy we have to cause the input to act over a longer time. We could then compute $h[n]$ by evaluating the inverse transform of $H(e^{j\Omega})$, which in turn could be determined as the ratio $Y(e^{j\Omega})/X(e^{j\Omega})$. Care has to be taken, however, to ensure that $X(e^{j\Omega}) \neq 0$ for any Ω; in other words, the input has to be sufficiently "rich". In particular, the input cannot be just a finite linear combination of sinusoids (unless the LTI system is such that knowledge of its frequency response at a finite number of frequencies serves to determine the frequency response at all frequencies — which would be the case with a lumped system, i.e., a finite-order system, except that one would need to know an upper bound on the order of the system so as to have a sufficient number of sinusoids combined in the input).

The above constraints might suggest using a *randomly generated input* signal. For instance, suppose we let the input be a *Bernoulli process*, with $x[n]$ for each n taking the value $+1$ or -1 with equal probability, independently of the values taken at other times. This process is (strict- and) wide-sense stationary, with mean value 0 and autocorrelation function $R_{xx}[m] = \delta[m]$. The corresponding power spectral density $S_{xx}(e^{j\Omega})$ is flat at the value 1 over the entire frequency range $\Omega \in [-\pi, \pi]$; evidently the expected power of $x[n]$ is distributed evenly over all frequencies. A process with flat power spectrum is referred to as a **white process** (a term that is motivated by the rough notion that white light contains all visible frequencies in equal amounts); a process that is not white is termed **colored**.

Now consider what the DTFT $X(e^{j\Omega})$ might look like for a typical sample function of a Bernoulli process. A typical sample function is not absolutely summable or square summable, and so does not fall into either of the categories for which we know that there are nicely behaved DTFTs. We might expect that the DTFT exists in some generalized-function sense (since the sample functions are bounded, and therefore do not grow faster than polynomially with n for large $|n|$), and this is indeed the case, but it is not a simple generalized function; not even as "nice" as the impulses or impulse trains or doublets that we are familiar with.

When the input $x[n]$ is a Bernoulli process, the output $y[n]$ will also be a WSS random process, and $Y(e^{j\Omega})$ will again not be a pleasant transform to deal with. However, recall that

$$R_{yx}[m] = h[m] * R_{xx}[m] \ , \tag{10.11}$$

so if we can estimate the cross-correlation of the input and output, we can determine the impulse response (for this case where $R_{xx}[m] = \delta[m]$) as $h[m] = R_{yx}[m]$. For a more general random process at the input, with a more general $R_{xx}[m]$, we can solve for $H(e^{j\Omega})$ by taking the Fourier transform of (10.11), obtaining

$$H(e^{j\Omega}) = \frac{S_{yx}(e^{j\Omega})}{S_{xx}(e^{j\Omega})} \ . \tag{10.12}$$

If the input is not accessible, and only its autocorrelation (or equivalently its PSD) is known, then we can still determine the *magnitude* of the frequency response, as long as we can estimate the autocorrelation (or PSD) of the output. In this case, we have

$$|H(e^{j\Omega})|^2 = \frac{S_{yy}(e^{j\Omega})}{S_{xx}(e^{j\Omega})} \ . \tag{10.13}$$

Given additional constraints or knowledge about the system, one can often determine a lot more (or even everything) about $H(e^{j\omega})$ from knowledge of its magnitude.

10.2.2 Invoking Ergodicity

How does one estimate $R_{yx}[m]$ and/or $R_{xx}[m]$ in an example such as the one above? The usual procedure is to assume (or prove) that the signals x and y are **ergodic**. What ergodicity permits — as we have noted earlier — is the replacement of an expectation or *ensemble average* by a *time average*, when computing the expected

value of various functions of random variables associated with a stationary random process. Thus a WSS process $x[n]$ would be called *mean-ergodic* if

$$E\{x[n]\} = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{k=-N}^{N} x[k] \; . \tag{10.14}$$

(The convergence on the right hand side involves a sequence of random variables, so there are subtleties involved in defining it precisely, but we bypass these issues in 6.011.) Similarly, for a pair of *jointly-correlation-ergodic* processes, we could replace the cross-correlation $E\{y[n+m]x[n]\}$ by the time average of $y[n+m]x[n]$.

What ergodicity generally requires is that values taken by a typical sample function over time be representative of the values taken across the ensemble. Intuitively, what this requires is that the correlation between samples taken at different times falls off fast enough. For instance, a sufficient condition for a WSS process $x[n]$ with finite variance to be mean-ergodic turns out to be that its autocovariance function $C_{xx}[m]$ tends to 0 as $|m|$ tends to ∞, which is the case with most of the examples we deal with in these notes. A more precise (necessary and sufficient) condition for mean-ergodicity is that the time-averaged autocovariance function $C_{xx}[m]$, weighted by a triangular window, be 0:

$$\lim_{L\to\infty} \frac{1}{2L+1} \sum_{m=-L}^{L} \left(1 - \frac{|m|}{L+1}\right) C_{xx}[m] = 0 \; . \tag{10.15}$$

A similar statement holds in the CT case. More stringent conditions (involving fourth moments rather than just second moments) are needed to ensure that a process is second-order ergodic; however, these conditions are typically satisfied for the processes we consider, where the correlations decay exponentially with lag.

10.2.3 Modeling Filters and Whitening Filters

There are various detection and estimation problems that are relatively easy to formulate, solve, and analyze when some random process that is involved in the problem — for instance, the set of measurements — is white, i.e., has a flat spectral density. When the process is colored rather than white, the easier results from the white case can still often be invoked in some appropriate way if:

(a) the colored process is the result of passing a white process through some LTI *modeling* or *shaping* filter, which shapes the white process at the input into one that has the spectral characteristics of the given colored process at the output; or

(b) the colored process is transformable into a white process by passing it through an LTI *whitening* filter, which flattens out the spectral characteristics of the colored process presented at the input into those of the white noise obtained at the output.

Thus, a modeling or shaping filter is one that converts a white process to some colored process, while a whitening filter converts a colored process to a white process.

An important result that follows from thinking in terms of modeling filters is the following (stated and justified rather informally here — a more careful treatment is beyond our scope):

Key Fact: A real function $R_{xx}[m]$ is the autocorrelation function of a real-valued WSS random process if and only if its transform $S_{xx}(e^{j\Omega})$ is real, even and nonnegative. The transform in this case is the PSD of the process.

The *necessity* of these conditions on the transform of the candidate autocorrelation function follows from properties we have already established for autocorrelation functions and PSDs.

To argue that these conditions are also *sufficient*, suppose $S_{xx}(e^{j\Omega})$ has these properties, and assume for simplicity that it has no impulsive part. Then it has a real and even square root, which we may denote by $\sqrt{S_{xx}(e^{j\Omega})}$. Now construct a (possibly noncausal) *modeling filter* whose frequency response $H(e^{j\Omega})$ equals this square root; the unit-sample reponse of this filter is found by inverse-transforming $H(e^{j\Omega}) = \sqrt{S_{xx}(e^{j\Omega})}$. If we then apply to the input of this filter a (zero-mean) unit-variance white noise process, e.g., a Bernoulli process that has equal probabilities of taking $+1$ and -1 at each DT instant independently of every other instant, then the output will be a WSS process with PSD given by $|H(e^{j\Omega})|^2 = S_{xx}(e^{j\Omega})$, and hence with the specified autocorrelation function.

If the transform $S_{xx}(e^{j\Omega})$ had an impulse at the origin, we could capture this by adding an appropriate constant (determined by the impulse strength) to the output of a modeling filter, constructed as above by using only the non-impulsive part of the transform. For a pair of impulses at frequencies $\Omega = \pm\Omega_o \neq 0$ in the transform, we could similarly add a term of the form $A\cos(\Omega_o n + \Theta)$, where A is deterministic (and determined by the impulse strength) and Θ is independent of all other other variables, and uniform in $[0, 2\pi]$.

Similar statements can be made in the CT case.

We illustrate below the logic involved in designing a whitening filter for a particular example; the logic for a modeling filter is similar (actually, inverse) to this.

Consider the following discrete-time system shown in Figure 10.3.

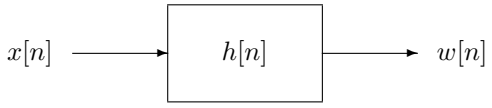

FIGURE 10.3 A discrete-time whitening filter.

Suppose that $x[n]$ is a process with autocorrelation function $R_{xx}[m]$ and PSD $S_{xx}(e^{j\Omega})$, i.e., $S_{xx}(e^{j\Omega}) = \mathcal{F}\{R_{xx}[m]\}$. We would like $w[n]$ to be a white noise output with variance σ_w^2.

We know that

$$S_{ww}(e^{j\Omega}) = |H(e^{j\Omega})|^2 \, S_{xx}(e^{j\Omega}) \tag{10.16}$$

or,

$$|H(e^{j\Omega})|^2 = \frac{\sigma_w^2}{S_{xx}(e^{j\Omega})}. \tag{10.17}$$

This then tells us what the squared magnitude of the frequency response of the LTI system must be to obtain a white noise output with variance σ_w^2. If we have $S_{xx}(e^{j\Omega})$ available as a rational function of $e^{j\Omega}$ (or can model it that way), then we can obtain $H(e^{j\Omega})$ by appropriate factorization of $|H(e^{j\Omega})|^2$.

EXAMPLE 10.1 Whitening filter

Suppose that

$$S_{xx}(e^{j\Omega}) = \frac{5}{4} - \cos(\Omega). \tag{10.18}$$

Then, to whiten $x(t)$, we require a stable LTI filter for which

$$|H(e^{j\Omega})|^2 = \frac{1}{(1 - \frac{1}{2}e^{j\Omega})(1 - \frac{1}{2}e^{-j\Omega})}, \tag{10.19}$$

or equivalently,

$$H(z)H(1/z) = \frac{1}{(1 - \frac{1}{2}z)(1 - \frac{1}{2}z^{-1})} \, . \tag{10.20}$$

The filter is constrained to be stable in order to produce a WSS output. One choice of $H(z)$ that results in a causal filter is

$$H(z) = \frac{1}{1 - \frac{1}{2}z^{-1}} \, , \tag{10.21}$$

with region of convergence (ROC) given by $|z| > \frac{1}{2}$. This system function could be multiplied by the system function $A(z)$ of any *allpass* system, i.e., a system function satisfying $A(z)A(z^{-1}) = 1$, and still produce the same whitening action, because $|A(e^{j\Omega})|^2 = 1$.

10.3 SAMPLING OF BANDLIMITED RANDOM PROCESSES

A WSS random process is termed *bandlimited* if its PSD is bandlimited, i.e., is zero for frequencies outside some finite band. For deterministic signals that are bandlimited, we can sample at or above the Nyquist rate and recover the signal exactly. We examine here whether we can do the same with bandlimited random processes.

In the discussion of sampling and DT processing of CT signals in your prior courses, the derivations and discussion rely heavily on picturing the effect in the frequency

domain, i.e., tracking the Fourier transform of the continuous-time signal through the C/D (sampling) and D/C (reconstruction) process. While the arguments can alternatively be carried out directly in the time domain, for deterministic finite-energy signals the frequency domain development seems more conceptually clear.

As you might expect, results similar to the deterministic case hold for the reconstruction of bandlimited random processes from samples. However, since these stochastic signals do not possess Fourier transforms except in the generalized sense, we carry out the development for random processes directly in the time domain. An essentially parallel argument could have been used in the time domain for deterministic signals (by examining the total energy in the reconstruction error rather than the expected instantaneous power in the reconstruction error, which is what we focus on below).

The basic sampling and bandlimited reconstruction process should be familiar from your prior studies in signals and systems, and is depicted in Figure 10.4 below. In this figure we have explicitly used bold upper-case symbols for the signals to underscore that they are random processes.

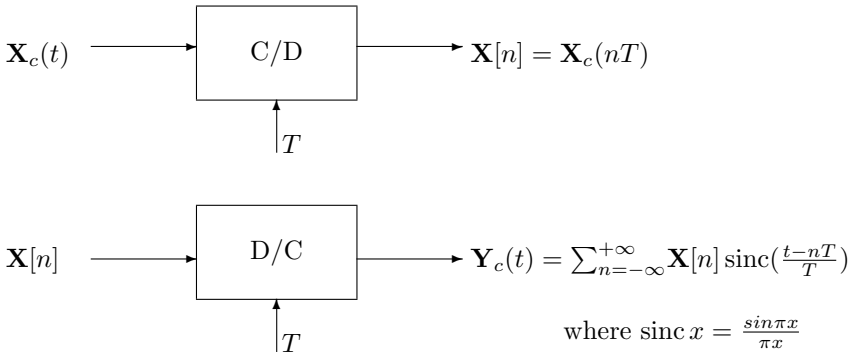

FIGURE 10.4 C/D and D/C for random processes.

For the deterministic case, we know that if $x_c(t)$ is bandlimited to less than $\frac{\pi}{T}$, then with the D/C reconstruction defined as

$$y_c(t) = \sum_n x[n] \operatorname{sinc}(\frac{t-nT}{T}) \tag{10.22}$$

it follows that $y_c(t) = x_c(t)$. In the case of random processes, what we show below is that, under the condition that $S_{x_c x_c}(j\omega)$, the power spectral density of $\mathbf{X}_c(t)$, is bandlimited to less that $\frac{\pi}{T}$, the mean square value of the error between $\mathbf{X}_c(t)$ and $\mathbf{Y}_c(t)$ is zero; i.e., if

$$S_{x_c x_c}(j\omega) = 0 \qquad |w| \geq \frac{\pi}{T}\,, \tag{10.23}$$

then

$$\mathcal{E} \triangleq E\{[\mathbf{X}_c(t) - \mathbf{Y}_c(t)]^2\} = 0 \, . \tag{10.24}$$

This, in effect, says that there is "zero power" in the error. (An alternative proof to the one below is outlined in Problem 13 at the end of this chapter.)

To develop the above result, we expand the error and use the definitions of the C/D (or sampling) and D/C (or ideal bandlimited interpolation) operations in Figure 10.4 to obtain

$$\mathcal{E} = E\{\mathbf{X}_c^2(t)\} + E\{\mathbf{Y}_c^2(t)\} - 2E\{\mathbf{Y}_c(t)\mathbf{X}_c(t)\} \, . \tag{10.25}$$

We first consider the last term, $E\{\mathbf{Y}_c(t)\mathbf{X}_c(t)\}$:

$$\begin{aligned} E\{\mathbf{Y}_c(t)\mathbf{X}_c(t)\} &= E\{ \sum_{n=-\infty}^{+\infty} \mathbf{X}_c(nT) \operatorname{sinc}(\frac{t-nT}{T}) \, \mathbf{X}_c(t)\} \\ &= \sum_{n=-\infty}^{+\infty} R_{x_c x_c}(nT - t) \operatorname{sinc}(\frac{nT-t}{T}) \qquad (10.26) \\ & \qquad (10.27) \end{aligned}$$

where, in the last expression, we have invoked the symmetry of sinc(.) to change the sign of its argument from the expression that precedes it.

Equation (10.26) can be evaluated using Parseval's relation in discrete time, which states that

$$\sum_{n=\infty}^{+\infty} v[n]w[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} V(e^{j\Omega}) W^*(e^{j\Omega}) d\Omega \tag{10.28}$$

To apply Parseval's relation, note that $R_{x_c x_c}(nT - t)$ can be viewed as the result of the C/D or sampling process depicted in Figure 10.5, in which the input is considered to be a function of the variable τ:

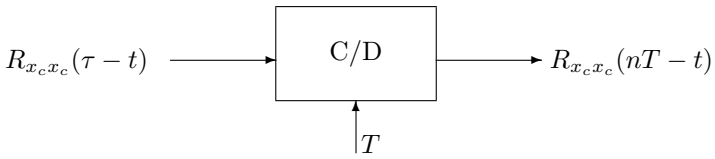

FIGURE 10.5 C/D applied to $R_{x_c x_c}(\tau - t)$.

The CTFT (in the variable τ) of $R_{x_c x_c}(\tau - t)$ is $e^{-j\omega t} S_{x_c x_c}(j\omega)$, and since this is bandlimited to $|\omega| < \frac{\pi}{T}$, the DTFT of its sampled version $R_{x_c x_c}(nT - t)$ is

$$\frac{1}{T} e^{\frac{-j\Omega t}{T}} S_{x_c x_c}(j\frac{\Omega}{T}) \tag{10.29}$$

in the interval $|\Omega| < \pi$. Similarly, the DTFT of $\text{sinc}(\frac{nT-t}{T})$ is $e^{\frac{-j\Omega t}{T}}$. Consequently, under the condition that $S_{x_c x_c}(j\omega)$ is bandlimited to $|\omega| < \frac{\pi}{T}$,

$$\begin{aligned} E\{\mathbf{Y}_c(t)\mathbf{X}_c(t)\} &= \frac{1}{2\pi T}\int_{-\pi}^{\pi} S_{x_c x_c}(\frac{j\Omega}{T})d\Omega \\ &= \frac{1}{2\pi}\int_{-(\pi/T)}^{(\pi/T)} S_{x_c x_c}(j\omega)d\omega \\ &= R_{x_c x_c}(0) = E\{\mathbf{X}_c^2(t)\} \end{aligned} \tag{10.30}$$

Next, we expand the middle term in equation (10.25):

$$\begin{aligned} E\{\mathbf{Y}_c^2(t)\} &= E\{\sum_n \sum_m \mathbf{X}_c(nT)\mathbf{X}_c(mT)\,\text{sinc}(\frac{t-nT}{T})\,\text{sinc}(\frac{t-mT}{T})\} \\ &= \sum_n \sum_m R_{x_c x_c}(nT-mT)\,\text{sinc}(\frac{t-mT}{T})\,\text{sinc}(\frac{t-mT}{T})\ . \end{aligned} \tag{10.31}$$

With the substitution $n - m = r$, we can express 10.31 as

$$E\{\mathbf{Y}_c^2(t)\} = \sum_r R_{x_c x_c}(rT)\ \sum_m \text{sinc}(\frac{t-mT}{T})\,\text{sinc}(\frac{t-mT-rT}{T})\ . \tag{10.32}$$

Using the identity

$$\sum_n \text{sinc}(n-\theta_1)\text{sinc}(n-\theta_2) = \text{sinc}(\theta_2-\theta_1)\ , \tag{10.33}$$

which again comes from Parseval's theorem (see Problem 12 at the end of this chapter), we have

$$\begin{aligned} E\{\mathbf{Y}_c^2(t)\} &= \sum_r R_{x_c x_c}(rT)\,\text{sinc}(r) \\ &= R_{x_c x_c}(0) = E\{\mathbf{X}_c^2\} \end{aligned} \tag{10.34}$$

since $\text{sinc}(r) = 1$ if $r = 0$ and zero otherwise. Substituting 10.31 and 10.34 into 10.25, we obtain the result that $\mathcal{E} = 0$, as desired.

CHAPTER 11

Wiener Filtering

INTRODUCTION

In this chapter we will consider the use of LTI systems in order to perform minimum mean-square-error (MMSE) estimation of a WSS random process of interest, given measurements of another related process. The measurements are applied to the input of the LTI system, and the system is designed to produce as its output the MMSE estimate of the process of interest.

We first develop the results in discrete time, and for convenience assume (unless otherwise stated) that the processes we deal with are *zero-mean*. We will then show that exactly analogous results apply in continuous time, although their derivation is slightly different in certain parts.

Our problem in the DT case may be stated in terms of Figure 11.1.

Here $x[n]$ is a WSS random process that we have measurements of. We want to determine the unit sample response or frequency response of the above LTI system such that the filter output $\widehat{y}[n]$ is the minimum-mean-square-error (MMSE) estimate of some "target" process $y[n]$ that is jointly WSS with $x[n]$. Defining the error $e[n]$ as

$$e[n] \triangleq \widehat{y}[n] - y[n] \ , \tag{11.1}$$

we wish to carry out the following minimization:

$$\min_{h[\,\cdot\,]} \epsilon = E\{e^2[n]\} \ . \tag{11.2}$$

The resulting filter $h[n]$ is called the **Wiener filter** for estimation of $y[n]$ from $x[n]$.

In some contexts it is appropriate or convenient to restrict the filter to be an FIR (finite-duration impulse response) filter of length N, e.g. $h[n] = 0$ except in the interval $0 \leq n \leq N-1$. In other contexts the filter impulse response can be of infinite duration and may either be restricted to be causal or allowed to be noncausal. In the next section we discuss the FIR and general noncausal IIR

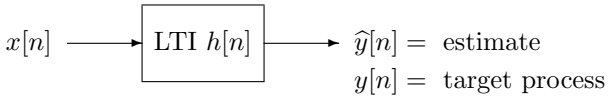

FIGURE 11.1 DT LTI filter for linear MMSE estimation.

(infinite-duration impulse response) cases. A later section deals with the more involved case where the filter is IIR but restricted to be causal.

If $x[n] = y[n] + v[n]$ where $y[n]$ is a signal and $v[n]$ is noise (both random processes), then the above estimation problem is called a *filtering* problem. If $y[n] = x[n + n_0]$ with n_0 positive, and if $h[n]$ is restricted to be causal, then we have a *prediction* problem. Both fit within the same general framework, but the solution under the restriction that $h[n]$ be causal is more subtle.

11.1 NONCAUSAL DT WIENER FILTER

To determine the optimal choice for $h[n]$ in (11.2), we first expand the error criterion in (11.2):

$$\epsilon = E\left\{\left(\sum_{k=-\infty}^{+\infty} h[k]x[n-k] - y[n]\right)^2\right\}. \tag{11.3}$$

The impulse response values that minimize ϵ can then be obtained by setting $\dfrac{\partial \epsilon}{\partial h[m]} = 0$ for all values of m for which $h[m]$ is not restricted to be zero (or otherwise pre-specified):

$$\frac{\partial \epsilon}{\partial h[m]} = E\left\{2\underbrace{\left(\sum_k h[k]x[n-k] - y[n]\right)}_{e[n]} x[n-m]\right\} = 0\,. \tag{11.4}$$

The above equation implies that

$$\begin{aligned} &E\{e[n]x[n-m]\} = 0, \text{ or} \\ &\qquad R_{ex}[m] = 0, \text{ for all } m \text{ for which } h[m] \text{ can be freely chosen.} \end{aligned} \tag{11.5}$$

You may recognize the above equation (or constraint) on the relation between the input and the error as the familiar *orthogonality principle*: for the optimal filter, *the error is orthogonal to all the data that is used to form the estimate*. Under our assumption of zero-mean $x[n]$, orthogonality is equivalent to uncorrelatedness. As we will show shortly, the orthogonality principle also applies in continuous time.

Note that

$$\begin{aligned} R_{ex}[m] &= E\{e[n]x[n-m]\} \\ &= E\{(\widehat{y}[n] - y[n])x[n-m]\} \\ &= R_{\widehat{y}x}[m] - R_{yx}[m]\,. \end{aligned} \tag{11.6}$$

Therefore, an alternative way of stating the orthogonality principle (11.5) is that

$$R_{\widehat{y}x}[m] = R_{yx}[m] \text{ for all appropriate } m\,. \tag{11.7}$$

In other words, for the optimal system, the cross-correlation between the input and output of the estimator equals the cross-correlation between the input and target output.

To actually find the impulse response values, observe that since $\widehat{y}[n]$ is obtained by filtering $x[n]$ through an LTI system with impulse response $h[n]$, the following relationship applies:

$$R_{\widehat{y}x}[m] = h[m] * R_{xx}[m] \; . \tag{11.8}$$

Combining this with the alternative statement of the orthogonality condition, we can write

$$h[m] * R_{xx}[m] = R_{yx}[m] \; , \tag{11.9}$$

or equivalently,

$$\sum_k h[k] R_{xx}[m-k] = R_{yx}[m] \tag{11.10}$$

Equation (11.10) represents a set of linear equations to be solved for the impulse response values. If the filter is FIR of length N, then there are N equations in the N unrestricted values of $h[n]$. For instance, suppose that $h[n]$ is restricted to be zero except for $n \in [0, N-1]$. The condition (11.10) then yields as many equations as unknowns, which can be arranged in the following matrix form, which you may recognize as the appropriate form of the *normal equations* for LMMSE estimation, which we introduced in Chapter 8:

$$\begin{bmatrix} R_{xx}[0] & R_{xx}[-1] & \cdots & R_{xx}[1-N] \\ R_{xx}[1] & R_{xx}[0] & \cdots & R_{xx}[2-N] \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}[N-1] & R_{xx}[N-2] & \cdots & R_{xx}[0] \end{bmatrix} \begin{bmatrix} h[0] \\ h[1] \\ \vdots \\ h[N-1] \end{bmatrix} = \begin{bmatrix} R_{yx}[0] \\ R_{yx}[1] \\ \vdots \\ R_{yx}[N-1] \end{bmatrix} . \tag{11.11}$$

These equations can now be solved for the impulse response values. Because of the particular structure of these equations, there are efficient methods for solving for the unknown parameters, but further discussion of these methods is beyond the scope of our course.

In the case of an IIR filter, equation (11.10) must hold for an infinite number of values of m and, therefore, cannot simply be solved by the methods used for a finite number of linear equations. However, if $h[n]$ is not restricted to be causal or FIR, the equation (11.10) must hold for all values of m from $-\infty$ to $+\infty$, so the z-transform can be applied to equation (11.10) to obtain

$$H(z) S_{xx}(z) = S_{yx}(z) \tag{11.12}$$

The optimal transfer function, i.e. the transfer function of the resulting (Wiener) filter, is then

$$\boxed{H(z) = S_{yx}(z)/S_{xx}(z)} \tag{11.13}$$

If either of the correlation functions involved in this calculation does not possess a z-transform but if both possess Fourier transforms, then the calculation can be carried out in the Fourier transform domain.

Note the similarity between the above expression for the optimal filter and the expression we obtained in Chapters 5 and 7 for the gain σ_{YX}/σ_{XX} that multiplies a zero-mean random variable X to produce the LMMSE estimator for a zero-mean random variables Y. In effect, by going to the transform domain or frequency domain, we have decoupled the design into a problem that — at each frequency — is as simple as the one we solved in the earlier chapters.

As we will see shortly, in continuous time the results are exactly the same:

$$R_{\widehat{yx}}(\tau) = R_{yx}(\tau), \tag{11.14}$$

$$h(\tau) * R_{xx}(\tau) = R_{yx}(\tau), \tag{11.15}$$

$$H(s)S_{xx}(s) = S_{yx}(s), \text{ and} \tag{11.16}$$

$$\boxed{H(s) = S_{yx}(s)/S_{xx}(s)} \tag{11.17}$$

The mean-square-error corresponding to the optimum filter, i.e. the *minimum* MSE, can be determined by straightforward computation. We leave you to show that

$$R_{ee}[m] = R_{yy}[m] - R_{\widehat{yy}}[m] = R_{yy}[m] - h[m] * R_{xy}[m] \tag{11.18}$$

where $h[m]$ is the impulse response of the optimal filter. The MMSE is then just $R_{ee}[0]$. It is illuminating to rewrite this in the frequency domain, but dropping the argument $e^{j\Omega}$ on the power spectra $S_{**}(e^{j\Omega})$ and frequency response $H(e^{j\Omega})$ below to avoid notational clutter:

$$\begin{aligned}
\text{MMSE} = R_{ee}[0] &= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{ee}\, d\Omega \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} (S_{yy} - HS_{xy})\, d\Omega \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{yy}\Big(1 - \frac{S_{yx}S_{xy}}{S_{yy}S_{xx}}\Big)\, d\Omega \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{yy}\Big(1 - \rho_{yx}\rho_{yx}^*\Big)\, d\Omega\,.
\end{aligned} \tag{11.19}$$

The function $\rho_{yx}(e^{j\Omega})$ defined by

$$\rho_{yx}(e^{j\Omega}) = \frac{S_{yx}(e^{j\Omega})}{\sqrt{S_{yy}(e^{j\Omega})S_{xx}(e^{j\Omega})}} \tag{11.20}$$

evidently plays the role of a frequency-domain correlation coefficient (compare with our earlier definition of the correlation coefficient between two random variables). This function is sometimes referred to as the *coherence function* of the two processes. Again, note the similarity of this expression to the expression $\sigma_{YY}(1-\rho_{YX}^2)$ that we obtained in a previous lecture for the (minimum) mean-square-error after LMMSE

estimation of a random variable Y using measurements of a random variable X.

EXAMPLE 11.1 Signal Estimation in Noise (Filtering)

Consider a situation in which $x[n]$, the sum of a target process $y[n]$ and noise $v[n]$, is observed:

$$x[n] = y[n] + v[n] \; . \tag{11.21}$$

We would like to estimate $y[n]$ from our observations of $x[n]$. Assume that the signal and noise are uncorrelated, i.e. $R_{vy}[m] = 0$. Then

$$R_{xx}[m] = R_{yy}[m] + R_{vv}[m], \tag{11.22}$$

$$R_{yx}[m] = R_{yy}[m], \tag{11.23}$$

$$H(e^{j\Omega}) = \frac{S_{yy}(e^{j\Omega})}{S_{yy}(e^{j\Omega}) + S_{vv}(e^{j\Omega})} \; . \tag{11.24}$$

At values of Ω for which the signal power is much greater than the noise power, $H(e^{j\Omega}) \approx 1$. Where the noise power is much greater than the signal power, $H(e^{j\Omega}) \approx 0$. For example, when

$$S_{yy}(e^{j\Omega}) = (1 + e^{-j\Omega})(1 + e^{j\Omega}) = 2(1 + \cos\Omega) \tag{11.25}$$

and the noise is white, the optimal filter will be a low-pass filter with a frequency response that is appropriately shaped, shown in Figure 11.2. Note that the filter in

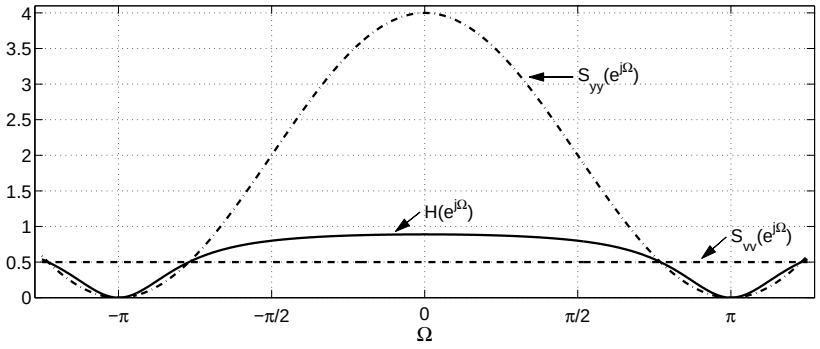

FIGURE 11.2 Optimal filter frequency response, $H(e^{j\Omega})$, input signal PSD signal, $S_{yy}(e^{j\Omega})$, and PSD of white noise, $S_{vv}(e^{j\Omega})$.

this case must have an impulse response that is an even function of time, since its frequency response is a real – and hence even – function of frequency.

Figure 11.3 shows a simulation example of such a filter in action (though for a different $S_{yy}(e^{j\Omega})$. The top plot is the PSD of the signal of interest; the middle plot shows both the signal $s[n]$ and the measured signal $x[n]$; and the bottom plot compares the estimate of $s[n]$ with $s[n]$ itself.

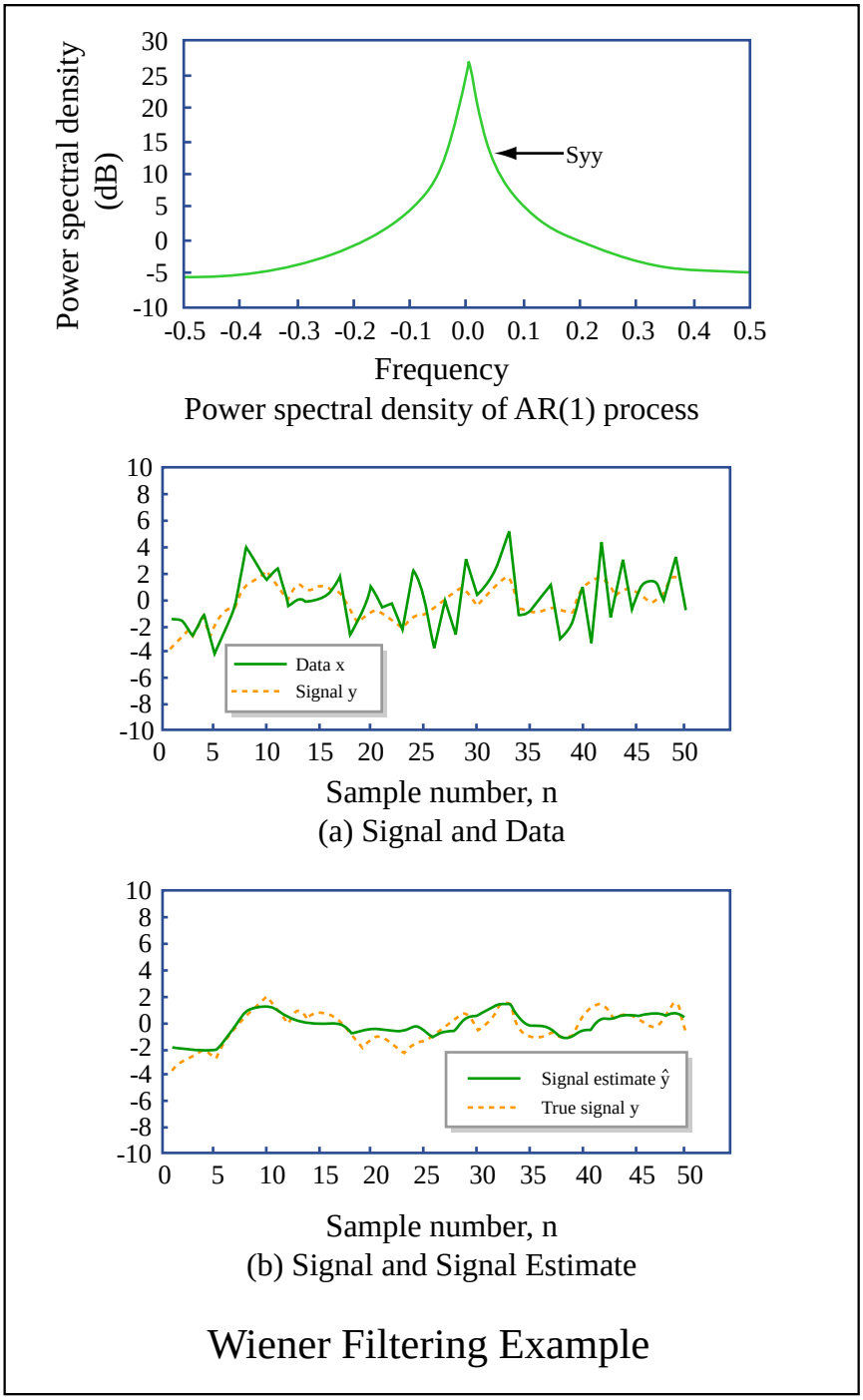

Image by MIT OpenCourseWare, adapted from *Fundamentals of Statistical Signal Processing: Estimation Theory*, Steven Kay. Prentice Hall, 1993.

FIGURE 11.3 Wiener filtering example. (From S.M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, 1993. Figures 11.9 and 11.10.)

EXAMPLE 11.2 Prediction

Suppose we wish to predict the measured process n_0 steps ahead, so

$$y[n] = x[n + n_0] \; . \tag{11.26}$$

Then

$$R_{yx}[m] = R_{xx}[m + n_0] \tag{11.27}$$

so the optimum filter has system function

$$H(z) = z^{n_0} \; . \tag{11.28}$$

This is of course not surprising: since we're allowing the filter to be noncausal, prediction is not a difficult problem! Causal prediction is much more challenging and interesting, and we will examine it later in this chapter.

EXAMPLE 11.3 Deblurring (or Deconvolution)

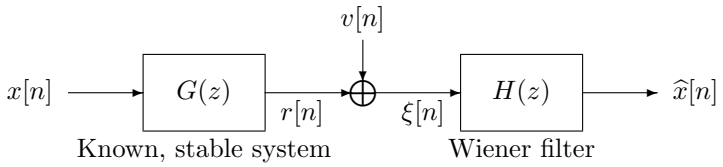

FIGURE 11.4 Wiener filtering of a blurred and noisy signal.

In the Figure 11.4, $r[n]$ is a filtered or "blurred" version of the signal of interest, $x[n]$, while $v[n]$ is additive noise that is uncorrelated with $x[n]$. We wish to design a filter that will deblur the noisy measured signal $\xi[n]$ and produce an estimate of the input signal $x[n]$. Note that in the absence of the additive noise, the inverse filter $1/G(z)$ will recover the input exactly. However, this is not a good solution when noise is present, because the inverse filter accentuates precisely those frequencies where the measurement power is small relative to that of the noise. We shall therefore design a Wiener filter to produce an estimate of the signal $x[n]$.

We have shown that the cross-correlation between the measured signal, which is the input to the Wiener filter, and the estimate produced at its output is equal to the cross-correlation between the measurement process and the target process. In the transform domain, the statement of this condition is

$$S_{\widehat{x}\xi}(z) = S_{x\xi}(z) \tag{11.29}$$

or

$$S_{\xi\xi}(z)H(z) = S_{\widehat{x}\xi}(z) = S_{x\xi}(z) \; . \tag{11.30}$$

We also know that

$$S_{\xi\xi}(z) = S_{vv}(z) + S_{xx}(z)G(z)G(1/z) \tag{11.31}$$
$$S_{x\xi}(z) = S_{xr}(z) \tag{11.32}$$
$$= S_{xx}(z)G(1/z), \tag{11.33}$$

where we have (in the first equality above) used the fact that $S_{vr}(z) = G(1/z)S_{vx}(z) = 0$. We can now write

$$H(z) = \frac{S_{xx}(z)G(1/z)}{S_{vv}(z) + S_{xx}(z)G(z)G(1/z)} \ . \tag{11.34}$$

We leave you to check that this system function assumes reasonable values in the limiting cases where the noise power is very small, or very large. It is also interesting to verify that the same overall filter is obtained if we first find an MMSE estimate $\widehat{r}[n]$ from $\xi[n]$ (as in Example 11.1), and then pass $\widehat{r}[n]$ through the inverse filter $1/G(z)$.

EXAMPLE 11.4 "De-Multiplication"

A message $s[n]$ is transmitted over a multiplicative channel (e.g. a fading channel) so that the received signal $r[n]$ is

$$r[n] = s[n]f[n] \ . \tag{11.35}$$

Suppose $s[n]$ and $f[n]$ are zero mean and independent. We wish to estimate $s[n]$ from $r[n]$ using a Wiener filter.

Again, we have

$$R_{sr}[m] = R_{\widehat{sr}}[m]$$
$$= h[m] * \underbrace{R_{rr}[m]}_{R_{ss}[m]R_{ff}[m]} \ . \tag{11.36}$$

But we also know that $R_{sr}[m] = 0$. Therefore $h[m] = 0$. This example emphasizes that the optimality of a filter satisfying certain constraints and minimizing some criterion does not necessarily make the filter a good one. The constraints on the filter and the criterion have to be relevant and appropriate for the intended task. For instance, if $f[n]$ was known to be i.i.d. and $+1$ or -1 at each time, then simply squaring the received signal $r[n]$ at any time would have at least given us the value of $s^2[n]$, which would seem to be more valuable information than what the Wiener filter produces in this case.

11.2 NONCAUSAL CT WIENER FILTER

In the previous discussion we derived and illustrated the discrete-time Wiener filter for the FIR and noncausal IIR cases. In this section we derive the continuous-time counterpart of the result for the noncausal IIR Wiener filter. The DT derivation involved taking derivatives with respect to a (countable) set of parameters $h[m]$, but in the CT case the impulse response that we seek to compute is a *CT function* $h(t)$, so the DT derivation cannot be directly copied. However, you will see that the results take the same form as in the DT case; furthermore, the derivation below has a natural DT counterpart, which provides an alternate route to the results in the preceding section.

Our problem is again stated in terms of Figure 11.5.

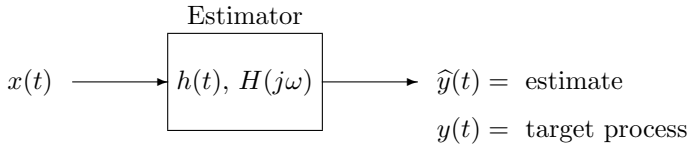

FIGURE 11.5 CT LTI filter for linear MMSE estimation.

Let $x(t)$ be a (zero-mean) WSS random process that we have measurements of. We want to determine the impulse response or frequency response of the above LTI system such that the filter output $\widehat{y}(t)$ is the LMMSE estimate of some (zero-mean) "target" process $y(t)$ that is jointly WSS with $x(t)$. We can again write

$$\begin{aligned} e(t) &\triangleq y(t) - \widehat{y}(t) \\ \min_{h(\cdot)} \epsilon &= E\{e^2(t)\} \ . \end{aligned} \tag{11.37}$$

Assuming the filter is stable (or at least has a well-defined frequency response), the process $\widehat{y}(t)$ is jointly WSS with $x(t)$. Furthermore,

$$E[\widehat{y}(t+\tau)y(t)] = h(\tau) * R_{xy}(\tau) = R_{\widehat{y}y}(\tau) \ , \tag{11.38}$$

The quantity we want to minimize can again be written as

$$\epsilon = E\{e^2(t)\} = R_{ee}(0) \ , \tag{11.39}$$

where the error autocorrelation function $R_{ee}(\tau)$ is — using the definition in (11.37) — evidently given by

$$R_{ee}(\tau) = R_{yy}(\tau) + R_{\widetilde{y}\widetilde{y}}(\tau) - R_{y\widehat{y}}(\tau) - R_{\widehat{y}y}(\tau) \ . \tag{11.40}$$

Thus

$$\begin{aligned}\epsilon = E\{e^2(t)\} = R_{ee}(0) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{ee}(j\omega)\,d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\Big(S_{yy}(j\omega) + S_{\widehat{y}\widehat{y}}(j\omega) - S_{y\widehat{y}}(j\omega) - S_{\widehat{y}y}(j\omega)\Big)\,d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}(S_{yy} + HH^*S_{xx} - H^*S_{yx} - HS_{xy})\,d\omega\ , \quad (11.41)\end{aligned}$$

where we have dropped the argument $j\omega$ from the PSDs in the last line above for notational simplicity, and have used H^* to denote the complex conjugate of $H(j\omega)$, namely $H(-j\omega)$. The expression in this last line is obtained by using the fact that $x(t)$ and $\widehat{y}(t)$ are the WSS input and output, respectively, of a filter whose frequency response is $H(j\omega)$. Note also that because $R_{yx}(\tau) = R_{xy}(-\tau)$ we have

$$S_{yx} = S_{yx}(j\omega) = S_{xy}(-j\omega) = S_{xy}^*\ . \quad (11.42)$$

Our task is now to choose $H(j\omega)$ to minimize the integral in (11.41). We can do this by minimizing the integrand for each ω. The first term in the integrand does not involve or depend on H, so in effect we need to minimize

$$HH^*S_{xx} - H^*S_{yx} - HS_{xy} = HH^*S_{xx} - H^*S_{yx} - HS_{yx}^*\ . \quad (11.43)$$

If all the quantities in this equation were real, this minimization would be straightforward. Even with a complex H and S_{yx}, however, the minimization is not hard.

The key to the minimization is an elementary technique referred to as *completing the square*. For this, we write the quantity in (11.43) in terms of the squared magnitude of a term that is linear in H. This leads to the following rewriting of (11.43):

$$\Big(H\sqrt{S_{xx}} - \frac{S_{yx}}{\sqrt{S_{xx}}}\Big)\Big(H^*\sqrt{S_{xx}} - \frac{S_{yx}^*}{\sqrt{S_{xx}}}\Big) - \frac{S_{yx}S_{yx}^*}{S_{xx}}\ . \quad (11.44)$$

In writing $\sqrt{S_{xx}}$, we have made use of the fact that $S_{xx}(j\omega)$ is real and nonnegative. We have also felt free to divide by $\sqrt{S_{xx}(j\omega)}$ because for any ω where this quantity is 0 it can be shown that $S_{yx}(j\omega) = 0$ also. The optimal choice of $H(j\omega)$ is therefore arbitrary at such ω, as evident from (11.43). We thus only need to compute the optimal H at frequencies where $\sqrt{S_{xx}(j\omega)} > 0$.

Notice that the second term in parentheses in (11.44) is the complex conjugate of the first term, so the product of these two terms in parentheses is real and *nonnegative*. Also, the last term does not involve H at all. To cause the terms in parentheses to vanish and their product to thereby become 0, which is the best we can do, we evidently must choose as follows (assuming there are no additional constraints such as causality on the estimator):

$$H(j\omega) = \frac{S_{yx}(j\omega)}{S_{xx}(j\omega)} \quad (11.45)$$

This expression has the same form as in the DT case. The formula for $H(j\omega)$ causes it to inherit the symmetry properties of $S_{yx}(j\omega)$, so $H(j\omega)$ has a real part that is

even in ω, and an imaginary part that is odd in ω. Its inverse transform is thus a *real* impulse response $h(t)$, and the expression in (11.45) is the frequency response of the optimum (Wiener) filter.

With the choice of optimum filter frequency response in (11.45), the mean-square-error expression in (11.41) reduces (just as in the DT case) to:

$$\begin{aligned}\text{MMSE} = R_{ee}(0) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{ee}\,d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} (S_{yy} - H S_{xy})\,d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{yy}\Big(1 - \frac{S_{yx}S_{xy}}{S_{yy}S_{xx}}\Big)\,d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{yy}(1-\rho\rho^*)\,d\omega \end{aligned} \tag{11.46}$$

where the function $\rho(j\omega)$ is defined by

$$\rho(j\omega) = \frac{S_{yx}(j\omega)}{\sqrt{S_{yy}(j\omega)S_{xx}(j\omega)}} \tag{11.47}$$

and evidently plays the role of a (complex) frequency-by-frequency correlation coefficient, analogous to that played by the correlation coefficient of random variables Y and X.

11.2.1 Orthogonality Property

Rearranging the equation for the optimal Wiener filter, we find

$$H\,S_{xx} = S_{yx} \tag{11.48}$$

or

$$S_{\widehat{yx}} = S_{yx}\ , \tag{11.49}$$

or equivalently

$$R_{\widehat{yx}}(\tau) = R_{yx}(\tau) \text{ for all } \tau\ . \tag{11.50}$$

Again, for the optimal system, the cross-correlation between the input and output of the estimator equals the cross-correlation between the input and target output.

Yet another way to state the above result is via the following orthogonality property:

$$R_{ex}(\tau) = R_{\widehat{yx}}(\tau) - R_{yx}(\tau) = 0 \text{ for all } \tau\ . \tag{11.51}$$

In other words, for the optimal system, the error is orthogonal to the data.

11.3 CAUSAL WIENER FILTERING

In the preceding discussion we developed the Wiener filter with no restrictions on the filter frequency response $H(j\omega)$. This allowed us to minimize a frequency-domain integral by choosing $H(j\omega)$ at each ω to minimize the integrand. However,

if we constrain the filter to be *causal*, then the frequency response cannot be chosen arbitrarily at each frequency, so the previous approach needs to be modified. It can be shown that for a causal system the real part of $H(j\omega)$ can be determined from the imaginary part, and vice versa, using what is known as a Hilbert transform. This shows that $H(j\omega)$ is constrained in the causal case. (We shall not need to deal explicitly with the particular constraint relating the real and imaginary parts of $H(j\omega)$, so we will not pursue the Hilbert transform connection here.) The development of the Wiener filter in the causal case is therefore subtler than the unrestricted case, but you know enough now to be able to follow the argument.

Recall our problem, described in terms of Figure 11.6.

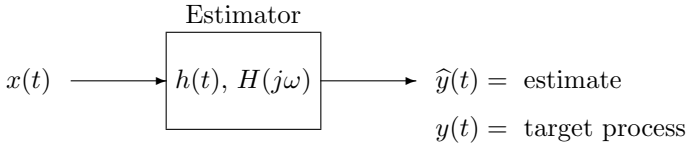

FIGURE 11.6 Representation of LMMSE estimation using an LTI system.

The input $x(t)$ is a (zero-mean) WSS random process that we have measurements of, and we want to determine the impulse response or frequency response of the above LTI system such that the filter output $\widehat{y}(t)$ is the LMMSE estimate of some (zero-mean) "target" process $y(t)$ that is jointly WSS with $x(t)$:

$$\begin{aligned} e(t) &\triangleq y(t) - \widehat{y}(t) \\ \min_{h(\,\cdot\,)} \epsilon &= E\{e^2(t)\}\,. \end{aligned} \tag{11.52}$$

We shall now require, however, that the filter be *causal*. This is essential in, for example, the problem of prediction, where $y(t) = x(t + t_0)$ with $t_0 > 0$.

We have already seen that the quantity we want to minimize can be written as

$$\begin{aligned} \epsilon = E\{e^2(t)\} = R_{ee}(0) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{ee}(j\omega)\,d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \Big(S_{yy}(j\omega) + S_{\widetilde{yy}}(j\omega) - S_{y\widehat{y}}(j\omega) - S_{\widehat{yy}}(j\omega)\Big)\,d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} (S_{yy} + HH^*S_{xx} - H^*S_{yx} - HS_{xy})\,d\omega \qquad (11.53) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \Big|H\sqrt{S_{xx}} - \frac{S_{yx}}{\sqrt{S_{xx}}}\Big|^2 d\omega + \frac{1}{2\pi}\int_{-\infty}^{\infty} \Big(S_{yy} - \frac{S_{yx}S_{yx}^*}{S_{xx}}\Big)\,d\omega\,. \qquad (11.54) \end{aligned}$$

The last equality was the result of "completing the square" on the integrand in the preceding integral. In the case where H is unrestricted, we can set the first integral of the last equation to 0 by choosing

$$H(j\omega) = \frac{S_{yx}(j\omega)}{S_{xx}(j\omega)} \tag{11.55}$$

at each frequency. The second integral of the last equation is unaffected by our choice of H, and determines the MMSE.

If the Wiener filter is required to be *causal*, then we have to deal with the integral

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\left|H\sqrt{S_{xx}}-\frac{S_{yx}}{\sqrt{S_{xx}}}\right|^2 d\omega \tag{11.56}$$

as a whole when we minimize it, because causality imposes constraints on $H(j\omega)$ that prevent it being chosen freely at each ω. (Because of the Hilbert transform relationship mentioned earlier, we could for instance choose the real part of $H(j\omega)$ freely, but then the imaginary part would be totally determined.) We therefore have to proceed more carefully.

Note first that the expression we obtained for the integrand in (11.56) by completing the square is actually not quite as general as we might have made it. Since we may need to use all the flexibility available to us when we tackle the constrained problem, we should explore how generally we can complete the square. Specifically, instead of using the real square root $\sqrt{S_{xx}}$ of the PSD S_{xx}, we could choose a *complex* square root M_{xx}, defined by the requirement that

$$S_{xx}=M_{xx}M_{xx}^* \qquad \text{or} \qquad S_{xx}(j\omega)=M_{xx}(j\omega)M_{xx}(-j\omega)\ , \tag{11.57}$$

and correspondingly rewrite the criterion in (11.56) as

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\left|HM_{xx}-\frac{S_{yx}}{M_{xx}^*}\right|^2 d\omega\ , \tag{11.58}$$

which is easily verified to be the same criterion, although written differently. The quantity $M_{xx}(j\omega)$ is termed a spectral factor of $S_{xx}(j\omega)$ or a modeling filter for the process x. The reason for the latter name is that passing (zero-mean) unit-variance white noise through a filter with frequency response $M_{xx}(j\omega)$ will produce a process with the PSD $S_{xx}(j\omega)$, so we can model the process x as being the result of such a filtering operation. Note that the real square root $\sqrt{S_{xx}}(j\omega)$ we used earlier is a special case of a spectral factor, but others exist. In fact, multiplying $\sqrt{S_{xx}}(j\omega)$ by an all-pass frequency response $A(j\omega)$ will yield a modeling filter:

$$A(j\omega)\sqrt{S_{xx}}(j\omega)=M_{xx}(j\omega)\ , \qquad A(j\omega)A(-j\omega)=1\ . \tag{11.59}$$

Conversely, it is easy to show that the frequency response of any modeling filter can be written as the product of an all-pass frequency response and $\sqrt{S_{xx}}(j\omega)$.

It turns out that under fairly mild conditions (which we shall not go into here) a PSD is guaranteed to have a spectral factor that is the frequency response of a *stable and causal* system, and whose *inverse* is also the frequency response of a *stable and causal* system. (To simplify how we talk about such factors, we shall adopt an abuse of terminology that is common when talking about Fourier transforms, referring to the factor itself — rather than the system whose frequency response is this factor — as being stable and causal, with a stable and causal inverse.) For instance, if

$$S_{xx}(j\omega)=\frac{\omega^2+9}{\omega^2+4}\ , \tag{11.60}$$

then the required factor is

$$M_{xx}(j\omega) = \frac{j\omega + 3}{j\omega + 2} \, . \tag{11.61}$$

We shall limit ourselves entirely to S_{xx} that have such a spectral factor, and assume for the rest of the derivation that the M_{xx} introduced in the criterion (11.58) is such a factor. (Keep in mind that wherever we ask for a stable system here, we can actually make do with a system with a well-defined frequency response, even if it's not BIBO stable, except that our results may then need to be interpreted more carefully.)

With these understandings, it is evident that the term HM_{xx} in the integrand in (11.58) is causal, as it is the cascade of two causal terms. The other term, S_{yx}/M_{xx}^*, is generally not causal, but we may separate its causal part out, denoting the transform of its causal part by $[S_{yx}/M_{xx}^*]_+$, and the transform of its anti-causal part by $[S_{yx}/M_{xx}^*]_-$. (In the DT case, the latter would actually denote the transform of the *strictly* anti-causal part, i.e., at times -1 and earlier; the value at time 0 would be retained with the causal part.)

Now consider rewriting (11.58) in the time domain, using Parseval's theorem. If we denote the inverse transform operation by $\mathcal{I}\{\,\cdot\,\}$, then the result is the following rewriting of our criterion:

$$\int_{-\infty}^{\infty} \left| \mathcal{I}\{HM_{xx}\} - \mathcal{I}\{[S_{yx}/M_{xx}^*]_+ - \mathcal{I}\{[S_{yx}/M_{xx}^*]_-\} \right|^2 dt \tag{11.62}$$

Since the term $\mathcal{I}\{HM_{xx}\}$ is causal (i.e., zero for negative time), the best we can do with it, as far as minimizing this integral is concerned, is to cancel out all of $\mathcal{I}\{[S_{yx}/M_{xx}^*]_+\}$. In other words, our best choice is

$$HM_{xx} = [S_{yx}/M_{xx}^*]_+ \, , \tag{11.63}$$

or

$$H(j\omega) = \frac{1}{M_{xx}(j\omega)} \Big[\frac{S_{yx}(j\omega)}{M_{xx}(-j\omega)} \Big]_+ \, . \tag{11.64}$$

Note that the stability and causality of the *inverse* of M_{xx} guarantee that this last step preserves stability and causality, respectively, of the solution.

The expression in (11.64) is the solution of the Wiener filtering problem under the causality constraint. It is also evident now that the MMSE is larger than in the unconstrained (noncausal) case by the amount

$$\Delta\text{MMSE} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Big| \Big[\frac{S_{yx}}{M_{xx}^*} \Big]_- \Big|^2 d\omega \, . \tag{11.65}$$

EXAMPLE 11.5 DT Prediction

Although the preceding results were developed for the CT case, exactly analogous expressions with obvious modifications (namely, using the DTFT instead of the

CTFT, with integrals from $-\pi$ to π rather than $-\infty$ to ∞, etc.) apply to the DT case.

Consider a process $x[n]$ that is the result of passing (zero-mean) white noise of unit variance through a (modeling) filter with frequency response

$$M_{xx}(e^{j\Omega}) = \alpha_0 + \alpha_1 e^{-j\Omega} \ , \tag{11.66}$$

where both α_0 and α_1 are assumed nonzero. This filter is stable and causal, and if $|\alpha_1| < |\alpha_0|$ then the inverse is stable and causal too. We assume this condition holds. (If it doesn't, we can always find another modeling filter for which it does, by multiplying the present filter by an appropriate allpass filter.)

Suppose we want to do causal one-step prediction for this process, so $y[n] = x[n+1]$. Then $R_{yx}[m] = R_{xx}[m+1]$, so

$$S_{yx} = e^{j\Omega} S_{xx} = e^{j\Omega} M_{xx} M_{xx}^* \ . \tag{11.67}$$

Thus

$$\left[\frac{S_{yx}}{M_{xx}^*}\right]_+ = [e^{j\Omega} M_{xx}]_+ = \alpha_1 \ , \tag{11.68}$$

and so the optimum filter, according to (11.64), has frequency response

$$H(e^{j\Omega}) = \frac{\alpha_1}{\alpha_0 + \alpha_1 e^{-j\Omega}} \ . \tag{11.69}$$

The associated MMSE is evaluated by the expression in (11.65), and turns out to be simply α_0^2 (which can be compared with the value of $\alpha_0^2 + \alpha_1^2$ that would have been obtained if we estimated $x[n+1]$ by just its mean value, namely zero).

11.3.1 Dealing with Nonzero Means

We have so far considered the case where both x and y have zero means (and the practical consequence has been that we haven't had to worry about their PSDs having impulses at the origin). If their means are nonzero, then we can do a better job of estimating $y(t)$ if we allow ourselves to adjust the estimates produced by the LTI system, by adding appropriate constants (to make an *affine* estimator). For this, we can first consider the problem of estimating $y - \mu_y$ from $x - \mu_x$, illustrated in Figure 11.7

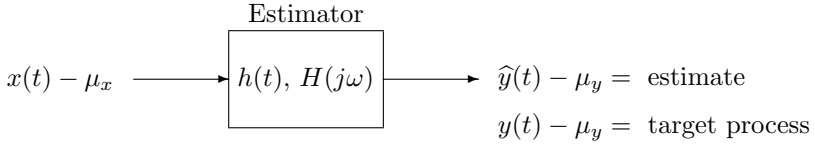

FIGURE 11.7 Wiener filtering with non-zero means.

Denoting the transforms of the *covariances* $C_{xx}(\tau)$ and $C_{yx}(\tau)$ by $D_{xx}(j\omega)$ and $D_{yx}(j\omega)$ respectively (these transforms are sometimes referred to as *covariance*

PSDs), the optimal unconstrained Wiener filter for our task will evidently have a frequency response given by

$$H(j\omega) = \frac{D_{yx}(j\omega)}{D_{xx}(j\omega)} \, . \tag{11.70}$$

We can then add μ_y to the output of this filter to get our LMMSE estimate of $y(t)$.

CHAPTER 12

Pulse Amplitude Modulation (PAM), Quadrature Amplitude Modulation (QAM)

12.1 PULSE AMPLITUDE MODULATION

In Chapter 2, we discussed the discrete-time processing of continuous-time signals, and in that context reviewed and discussed D/C conversion for reconstructing a continuous-time signal from a discrete-time sequence. Another common context in which it is useful and important to generate a continuous-time signal from a sequence is in communication systems, in which discrete data — for example, digital or quantized data — is to be transmitted over a channel in the form of a continuous-time signal. In this case, unlike in the case of DT processing of CT signals, the resulting continuous-time signal will be converted back to a discrete-time signal at the receiving end. Despite this difference in the two contexts, we will see that the same basic analysis applies to both.

As examples of the communication of DT information over CT channels, consider transmitting a binary sequence of 1's and 0's from one computer to another over a telephone line or cable, or from a digital cell phone to a base station over a high-frequency electromagnetic channel. These instances correspond to having analog channels that require the transmitted signal to be continuous in time, and to also be compatible with the bandwidth and other constraints of the channel. Such requirements impact the choice of continuous-time waveform that the discrete sequence is modulated onto.

The translation of a DT signal to a CT signal appropriate for transmission, and the translation back to a DT signal at the receiver, are both accomplished by devices referred to as modems (modulators/demodulators). Pulse Amplitude Modulation (PAM) underlies the operation of a wide variety of modems.

12.1.1 The Transmitted Signal

The basic idea in PAM for communication over a CT channel is to transmit a sequence of CT pulses of some pre-specified shape $p(t)$, with the sequence of pulse amplitudes carrying the information. The associated baseband signal at the transmitter (which is then usually modulated onto some carrier to form a bandpass signal

before actual transmission — but we shall ignore this aspect for now) is given by

$$x(t) = \sum_n a[n]\, p(t - nT) \tag{12.1}$$

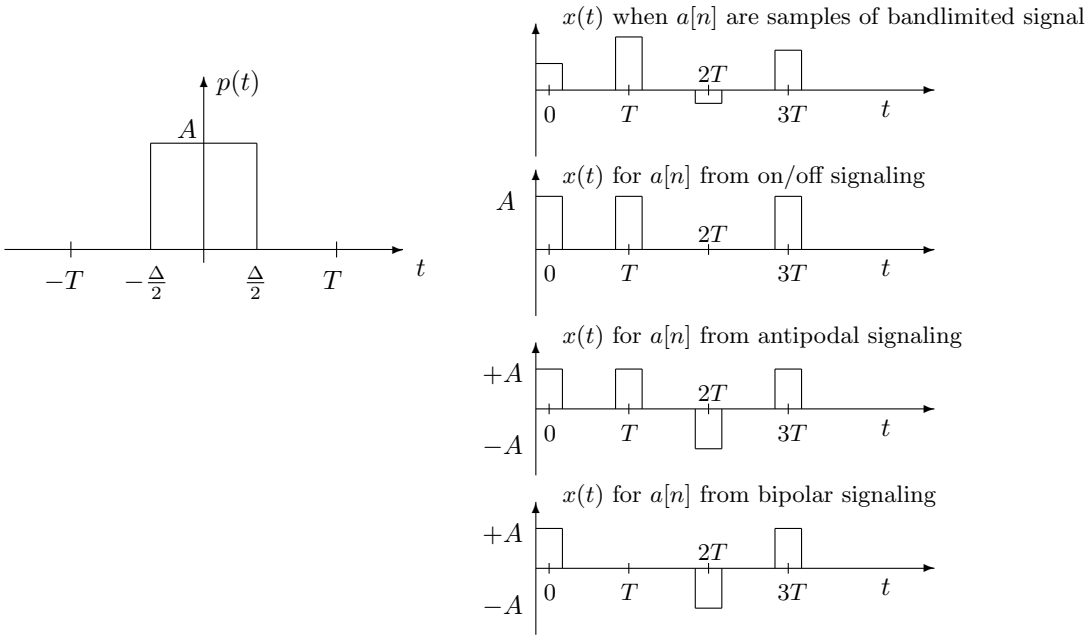

FIGURE 12.1 Baseband signal at the transmitter in Pulse Amplitude Modulation (PAM).

where the numbers $a[n]$ are the pulse amplitudes, and T is the pulse repetition interval or the inter-symbol spacing, so $1/T$ is the symbol rate (or "baud" rate). An individual pulse may be confined to an interval of length T, as shown in Figure 12.1, or it may extend over several intervals, as we will see in several examples shortly. The DT signal $a[n]$ may comprise samples of a bandlimited analog message (taken at the Nyquist rate or higher, and generally quantized to a specified set of levels, for instance 32 levels); or 1 and 0 for on/off or "unipolar" signaling; or 1 and -1 for antipodal or "polar" signaling; or 1, 0 and -1 for "bipolar" signaling; each of these possibilities is illustrated in Figure 12.1.

The particular pulse shape in Figure 12.1 is historically referred to as an RZ (return-to-zero) pulse when $\Delta < T$ and an NRZ (non-return-to-zero) pulse when $\Delta = T$. These pulses would require substantial channel bandwidth (of the order of $1/\Delta$) in order to be transmitted without significant distortion, so we may wish to find alternative choices that use less bandwidth, to accommodate the constraints of the channel. Such considerations are important in designing appropriate pulse shapes, and we shall elaborate on them shortly.

If $p(t)$ is chosen such that $p(0) = 1$ and $p(nT) = 0$ for $n \neq 0$, then we could recover the amplitudes $a[n]$ from the PAM waveform $x(t)$ by just sampling $x(t)$ at times nT, since $x(nT) = a[n]$ in this case. However, our interest is in recovering the amplitudes from the signal at the *receiver*, rather than directly from the transmitted signal, so we need to consider how the communication channel affects $x(t)$. Our objective will be to recover the DT signal in as simple a fashion as possible, while compensating for distortion and noise in the channel.

12.1.2 The Received Signal

When we transmit a PAM signal through a channel, the characteristics of the channel will affect our ability to accurately recover the pulse amplitudes $a[n]$ from the received signal $r(t)$. We might model $r(t)$ as

$$r(t) = h(t) * x(t) + \eta(t) \tag{12.2}$$

corresponding to the channel being modeled as LTI with impulse response $h(t)$, and channel noise being represented through the additive noise signal $\eta(t)$. We would still typically try to recover the pulse amplitudes $a[n]$ from samples of $r(t)$ — or from samples of an appropriately filtered version of $r(t)$ — with the samples taken at intervals of T.

The overall model is shown in Figure 12.2, with $f(t)$ representing the impulse response of an LTI filter at the receiver. This receiver filter will play a key role in filtering out the part of the noise that lies outside the frequency bands in which the signal information is concentrated. Here, we first focus on the noise-free case (for which one would normally set $f(t) = \delta(t)$, corresponding to no filtering before sampling at the receiver end), but for generality we shall take account of the effect of the filter $f(t)$ as well.

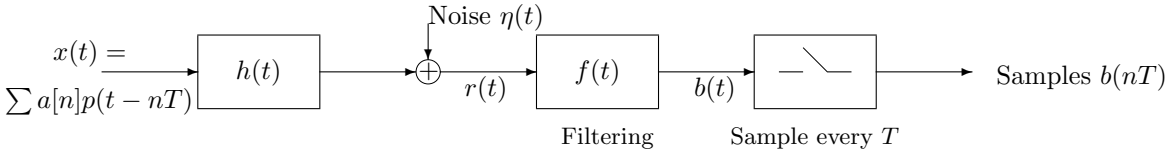

FIGURE 12.2 Transmitter, channel and receiver model for a PAM system.

12.1.3 Frequency-Domain Characterizations

Denote the CTFT of the pulse $p(t)$ by $P(j\omega)$, and similarly for the other CT signals in Figure 12.2. If the frequency response $H(j\omega)$ of the channel is unity over the frequency range where $P(j\omega)$ is significant, then a single pulse $p(t)$ is transmitted essentially without distortion. In this case, we might invoke the linearity and time invariance of our channel model to conclude that $x(t)$ in (12.1) is itself transmitted essentially without distortion, in which case $r(t) \approx x(t)$ in the noise-free case

that we are considering. However, this conclusion leaves the possiblity that distortions which are insignificant when a single pulse is transmitted accumulate in a non-negligible way when a succession of pulses is transmitted. We should therefore directly examine $x(t)$, $r(t)$, and their corresponding Fourier transforms. The understanding we obtain from this is a prerequisite for designing $P(j\omega)$ and picking the inter-symbol time T for a given channel, and also allows us to determine the influence of the DT signal $a[n]$ on the CT signals $x(t)$ and $r(t)$.

To compute $X(j\omega)$, we take the transform of both sides of (12.1):

$$\begin{aligned} X(j\omega) &= \Big(\sum_n a[n]\, e^{-j\omega nT}\Big)\, P(j\omega) \\ &= A(e^{j\Omega})|_{\Omega=\omega T}\; P(j\omega) \end{aligned} \tag{12.3}$$

where $A(e^{j\Omega})$ denotes the DTFT of the sequence $a[n]$. The quantity $A(e^{j\Omega})|_{\Omega=\omega T}$ that appears in the above expression is simply a uniform re-scaling of the frequency axis of the DTFT; in particular, the point $\Omega = \pi$ in the DTFT is mapped to the point $\omega = \pi/T$ in the expression $A(e^{j\Omega})|_{\Omega=\omega T}$.

The expression in (12.3) therefore describes $X(j\omega)$ for us, assuming the DTFT of the sequence $a[n]$ is well defined. For example, if $a[n] = 1$ for all n, corresponding to periodic repetition of the basic pulse waveform $p(t)$, then $A(e^{j\Omega}) = 2\pi\delta(\Omega)$ for $|\Omega| \leq \pi$, and repeats with period 2π outside this range. Hence $X(j\omega)$ comprises a train of impulses spaced apart by $2\pi/T$; the strength of each impulse is $2\pi/T$ times the value of $P(j\omega)$ at the location of the impulse (note that the scaling property of impulses yields $\delta(\Omega) = \delta(\omega T) = (1/T)\delta(\omega)$ for positive T).

In the absence of noise, the received signal $r(t)$ and the signal $b(t)$ that results from filtering at the receiver are both easily characterized in the frequency domain:

$$R(j\omega) = H(j\omega)X(j\omega)\,, \qquad B(j\omega) = F(j\omega)H(j\omega)X(j\omega)\,. \tag{12.4}$$

Some important constraints emerge from (12.3) and (12.4). Note first that for a general DT signal $a[n]$, necessary information about the signal will be distributed in its DTFT $A(e^{j\Omega})$ at frequencies Ω throughout the interval $|\Omega| \leq \pi$; knowing $A(e^{j\Omega})$ only in a smaller range $|\Omega| \leq \Omega_a < \pi$ will in general be *insufficient* to allow reconstruction of the DT signal. Now, setting $\Omega = \omega T$ as specified in (12.3), we see that $A(e^{j\omega T})$ will contain necessary information about the DT signal at frequencies ω that extend throughout the interval $|\omega| \leq \pi/T$. Thus, if $P(j\omega) \neq 0$ for $|\omega| \leq \pi/T$ then $X(j\omega)$ preserves the information in the DT signal; and if $H(j\omega)P(j\omega) \neq 0$ for $|\omega| \leq \pi/T$ then $R(j\omega)$ preserves the information in the DT signal; and if $F(j\omega)H(j\omega)P(j\omega) \neq 0$ for $|\omega| \leq \pi/T$ then $B(j\omega)$ preserves the information in the DT signal.

The above constraints have some design implications. A pulse for which $P(j\omega)$ was nonzero only in a strictly smaller interval $|\omega| \leq \omega_p < \pi/T$ would cause loss of information in going from the DT signal to the PAM signal $x(t)$, and would not be a suitable pulse for the chosen symbol rate $1/T$ (but could become a suitable pulse if the symbol rate was reduced appropriately, to ω_p/π or less).

Similarly, even if the pulse was appropriately designed so that $x(t)$ preserved the information in the DT signal, if we had a lowpass channel for which $H(j\omega)$ was nonzero only in a strictly smaller interval $|\omega| \leq \omega_c < \pi/T$ (so ω_c is the cutoff frequency of the channel), then we would lose information about the DT signal in going from $x(t)$ to $r(t)$; the chosen symbol rate $1/T$ would be inappropriate for this channel, and would need to be reduced to ω_c/π in order to preserve the information in the DT signal.

12.1.4 Inter-Symbol Interference at the Receiver

In the absence of any channel impairments, the signal values can be recovered from the transmitted pulse trains shown in Figure 12.1 by re-sampling at the times which are integer multiples of T. However, these pulses, while nicely time localized, have infinite bandwidth. Since any realistic channel will have a limited bandwidth, one effect of a communication channel on a PAM waveform is to "de-localize" or disperse the energy of each pulse through low-pass filtering. As a consequence, pulses that may not have overlapped (or that overlapped only benignly) at the transmitter may overlap at the receiver in a way that impedes the recovery of the pulse amplitudes from samples of $r(t)$, i.e. in a way that leads to *inter-symbol interference* (ISI). We now make explicit what condition is required in order for ISI to be eliminated

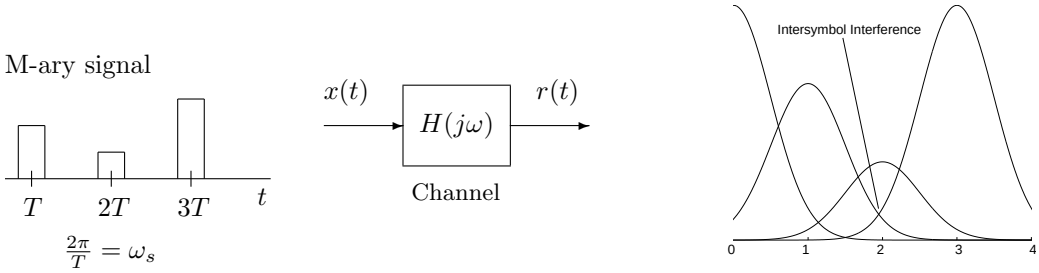

FIGURE 12.3 Illustration of Inter-symbol Interference (ISI).

from the filtered signal $b(t)$ at the receiver. When this no-ISI condition is met, we will again be able to recover the DT signal by simply sampling $b(t)$. Based on this condition, we can identify the additional constraints that must be satisfied by the pulse shape $p(t)$ and the impulse response $f(t)$ of the filter (or channel compensator or *equalizer*) at the receiver so as to eliminate or minimize ISI.

With $x(t)$ as given in (12.1), and noting that $b(t) = f(t) * h(t) * x(t)$ in the noise-free case, we can write

$$b(t) = \sum_n a[n]\, g(t - nT) \tag{12.5}$$

where

$$g(t) = f(t) * h(t) * p(t) \tag{12.6}$$

We assume that $g(t)$ is continuous (i.e., has no discontinuity) at the sampling times

nT. Our requirement for no ISI is then that

$$g(0) = c\,, \quad \text{and} \quad g(nT) = 0 \ \text{ for nonzero integers } n, \tag{12.7}$$

where c is some nonzero constant. If this condition is satisfied, then if follows from (12.5) that $b(nT) = c.a[n]$, and consequently the DT signal is exactly recovered (to within the known scale factor c).

As an example, suppose that $g(t)$ in (12.6) is

$$g(t) = \frac{\sin \omega_c t}{\omega_c t}\,, \tag{12.8}$$

with corresponding $G(j\omega)$ given by

$$\begin{aligned} G(j\omega) &= \frac{\pi}{\omega_c} \quad \text{for} \quad |\omega| < \omega_c \\ &= 0 \quad \text{otherwise}\,. \end{aligned} \tag{12.9}$$

Then choosing the inter-symbol spacing to be $T = \dfrac{\pi}{\omega_c}$, we can avoid ISI in the received samples, since $g(t) = 1$ at $t = 0$ and is zero at other integer multiples of T, as illustrated in Figure 12.4.

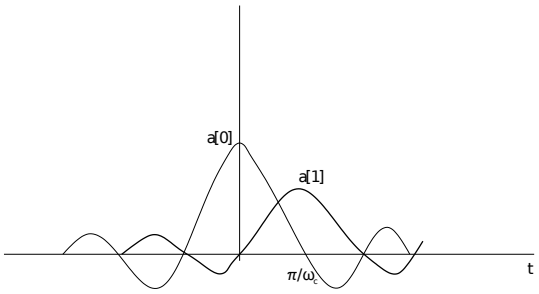

FIGURE 12.4 Illustration of the no-ISI property for PAM when $g(0) = 1$ and $g(t) = 0$ at other integer multiples of the inter-symbol time T.

We are thereby able to transmit at a symbol rate that is twice the cutoff frequency of the channel. From what was said earlier, in the discussion following (12.3) on constraints involving the symbol rate and the channel cutoff frequency, we cannot expect to do better in general.

More generally, in the next section we translate the no-ISI time-domain condition in (12.7) to one that is useful in designing $p(t)$ and $f(t)$ for a given channel. The approach is based on the frequency-domain translation of the no-ISI condition, leading to a result that was first articulated by Nyquist.

12.2 NYQUIST PULSES

The frequency domain interpretation of the no-ISI condition of (12.7) was explored by Nyquist in 1924 (and extended by him in 1928 to a statement of the sampling theorem — this theorem then waited almost 20 years to be brought to prominence by Gabor and Shannon).

Consider sampling $g(t)$ with a periodic impulse train:

$$\widehat{g}(t) = g(t) \sum_{n=-\infty}^{+\infty} \delta(t - nT) \; . \tag{12.10}$$

Then our requirements on $g(t)$ in (12.7) imply that $\widehat{g}(t) = c\,\delta(t)$, an impulse of strength c, whose transform is $\widehat{G}(j\omega) = c$. Taking transforms of both sides of (12.10), and utilizing the fact that multiplication in the time domain corresponds to convolution in the frequency domain, we obtain

$$\widehat{G}(j\omega) = c = \frac{1}{T} \sum_{m=-\infty}^{+\infty} G(j\omega - jm\frac{2\pi}{T}) \; . \tag{12.11}$$

The expression on the right hand side of (12.11) represents a replication of $G(j\omega)$ (scaled by $1/T$) at every integer multiple of $2\pi/T$ along the frequency axis. The Nyquist requirement is thus that $G(j\omega)$ and its replications, spaced $2\pi m/T$ apart for all integer m, add up to a constant. Some examples of $G(j\omega) = F(j\omega)H(j\omega)P(j\omega)$ that satisfy this condition are given below.

The particular case of the sinc function of (12.8) and (12.9) certainly satisfies the Nyquist condition of (12.11).

If we had an ideal lowpass channel $H(j\omega)$ with bandwidth ω_c or greater, then choosing $p(t)$ to be the sinc pulse of (12.8) and not doing any filtering at the receiver — so $F(j\omega) = 1$ — would result in no ISI. However, there are two problems with the sinc characteristic. First, the signal extends indefinitely in time in both directions. Second, the sinc has a very slow roll-off in time (as $1/t$). This slow roll-off in time is coupled to the sharp cut-off of the transform of the sinc in the frequency domain. This is a familiar manifestation of time-frequency duality: quick transition in one domain means slow transition in the other.

It is highly desirable in practice to have pulses that taper off more quickly in time than a sinc. One reason is that, given the inevitable inaccuracies in sampling times due to timing jitter, there will be some unavoidable ISI, and this ISI will propagate for unacceptably long times if the underlying pulse shape decays too slowly. Also, a faster roll-off allows better approximation of a two-sided signal by a one-sided signal, as would be required for a causal implementation. The penalty for more rapid pulse roll-off in time is that the transition in the frequency domain has to be more gradual, necessitating a larger bandwidth for a given symbol rate (or a reduced symbol rate for a given bandwidth).

The two examples in Figure 12.5 have smoother transitions than the previous case, and correspond to pulses that fall off as $1/t^2$. It is evident that both can be made

to satisfy the Nyquist condition by appropriate choice of T.

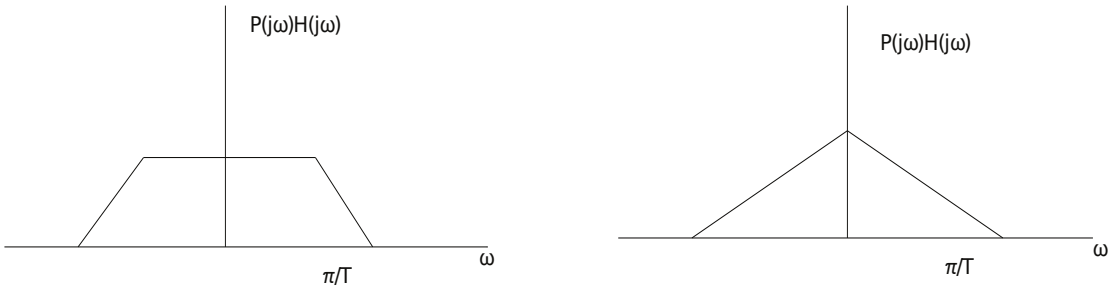

FIGURE 12.5 Two possible choices for the Fourier transform of pulses that decay in time as $1/t^2$ and satisfy the Nyquist zero-ISI condition for appropriate choice of T.

Still smoother transitions can be obtained with a family of frequency-domain characteristics in which there is a *cosine* transition from 1 to 0 over the frequency range $\omega = \frac{\pi}{T}(1-\beta)$ to $\omega = \frac{\pi}{T}(1+\beta)$, where β is termed the roll-off parameter. The corresponding formula for the received and filtered pulse is

$$f(t) * h(t) * p(t) \;=\; \frac{\sin \frac{\pi}{T}t}{\frac{\pi}{T}t} \, \frac{\cos \beta \frac{\pi}{T}t}{1-(2\beta t/T)^2} \tag{12.12}$$

which falls off as $1/t^3$ for large t.

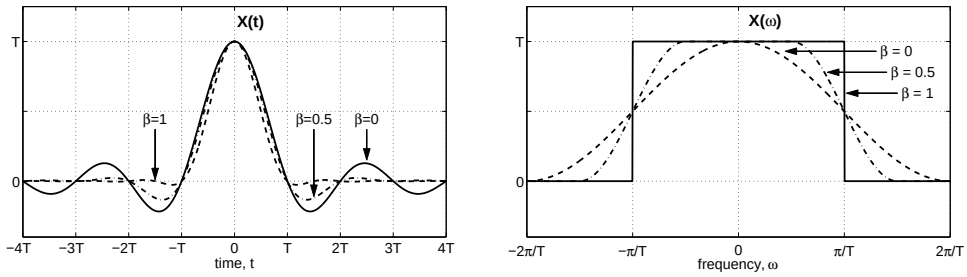

FIGURE 12.6 Time and frequency characteristics of the family of pulses in Eq. (12.12)

Once $G(j\omega)$ is specified, knowledge of the channel characteristic $H(j\omega)$ allows us to determine the corresponding pulse transform $P(j\omega)$, if we fix $F(j\omega) = 1$. In the presence of channel noise that corrupts the received signal $r(t)$, it turns out that it is best to only do part of the pulse shaping at the transmitter, with the rest done at the receiver prior to sampling. For instance, if the channel has no distortion in the passband (i.e., if $H(j\omega) = 1$ in the passband) and if the noise intensity is

TABLE 5.4: Selected CCITT International Telephone Line Modem Standards

Bit Rate	Symbol Rate	Modulation	CCITT Standard
330	300	2FSK	V.21
1,200	600	QPSK	V.22
2,400	600	16QAM	V.22bis
1,200	1,200	2FSK	V.23
2,400	1,200	QPSK	V.26
4,800	1,600	8PSK	V.27
9,600	2,400	Fig. 3.15(a)	V.29
4,800	2,400	QPSK	V.32
9,600	2,400	16QAM	V.32ALT
14,400	2,400	128QAM,TCM	V.32bis
28,800	3,429	1024QAM,TCM	V.fast(V.34)

FIGURE 12.7 From *Digital Transmission Engineering* by J.B.Anderson, IEEE Press 1999. The reference to Fig. 3.15 a is a particular QAM constellation.

uniform in this passband, then the optimal choice of pulse is $P(j\omega) = \sqrt{G(j\omega)}$, assuming that $G(j\omega)$ is purely real, and this is also the optimal choice of receiver filter $F(j\omega)$. We shall say a little more about this sort of issue when we deal with matched filtering in a later chapter.

12.3 CARRIER TRANSMISSION

The previous discussion centered around the design of baseband pulses. For transmission over phone lines, wireless links, satellites, etc. the baseband signal needs to be modulated onto a carrier, i.e. converted to a passband signal. This also opens opportunities for augmentation of PAM. The table in Figure 12.7 shows the evolution of telephone line digital modem standards. FSK refers to frequency-shift-keying, PSK to phase-shift-keying, and QAM to quadrature amplitude modulation, each of which we describe in more detail below. The indicated increase in symbol rate (or *baud* rate) and bit rates over the years corresponds to improvements in signal processing, to better modulation schemes, to the use of better conditioned channels, and to more elaborate coding (and correspondingly complex decoding, but now well within real-time computational capabilities of digital receivers).

For baseband PAM, the transmitted signal is of the form of equation (12.1) i.e.

$$x(t) = \sum_n a[n]\, p(t - nT) \tag{12.13}$$

where $p(t)$ is a lowpass pulse. When this is amplitude-modulated onto a carrier, the transmitted signal takes the form

$$s(t) = \sum_n a[n]\, p(t - nT) \cos(\omega_c t + \theta_c) \tag{12.14}$$

where ω_c and θ_c are the carrier frequency and phase.

In the simplest form of equation (12.14), specifically with ω_c and θ_c fixed, equation (12.14) corresponds to using amplitude modulation to shift the frequency content from baseband to a band centered at the carrier frequency ω_c. However, since two additional parameters have been introduced (i.e. ω_c and θ_c) this opens additional possibilities for embedding data in $s(t)$. Specifically, in addition to changing the amplitude in each symbol interval, we can consider changing the carrier frequency and/or the phase in each symbol interval. These alternatives lead to frequency-shift-keying (FSK) and phase-shift-keying (PSK).

12.3.1 FSK

With frequency shift keying (12.14) takes the form

$$s(t) = \sum_n a[n]\, p(t - nT) \cos((\omega_0 + \Delta_n)t + \theta_c) \tag{12.15}$$

where ω_0 is the nominal carrier frequency and Δ_n is the shift in the carrier frequency in symbol interval n. In principle in FSK both $a[n]$ and Δ_n can incorporate data although it is typically the case that in FSK the amplitude does not change.

12.3.2 PSK

In phase shift keying (12.14) takes the form

$$s(t) = \sum_n a[n]\, p(t - nT) \cos(\omega_c t + \theta_n) \tag{12.16}$$

In each symbol interval, information can then be incorporated in both the pulse amplitude $a[n]$ and the carrier phase θ_n. In what is typically referred to as PSK, information is only incorporated in the phase, i.e. $a[n] = a = constant$.

For example, with

$$\theta_n = \frac{2\pi b_n}{M} \text{ ; } b_n \text{ a non-negative integer} \tag{12.17}$$

one of M symbols can be encoded in the phase in each symbol interval. For $M = 2$, $\theta_n = 0$ or π, commonly referred to as binary PSK (BPSK). With $M = 4$, θ_n takes on one of the four values 0, $\frac{\pi}{2}$, π, or $\frac{3\pi}{2}$.

To interpret PSK somewhat differently and as a prelude to expanding the discussion to a further generalization (quadrature amplitude modulation or QAM) it is convenient to express equation (12.16) in some alternate forms. For example,

$$s(t) = \sum_n Re\{a e^{j\theta_n} p(t - nT) e^{j\omega_c t}\} \tag{12.18}$$

and equivalently

$$s(t) = I(t)\cos(\omega_c t) - Q(t)\sin(\omega_c t) \tag{12.19}$$

with

$$I(t) = \sum_n a_i[n]\, p(t - nT) \tag{12.20}$$

$$Q(t) = \sum_n a_q[n]\, p(t - nT) \tag{12.21}$$

and

$$a_i[n] = a\cos(\theta_n) \tag{12.22}$$

$$a_q[n] = a\sin(\theta_n) \tag{12.23}$$

Equation 12.19 is referred to as the quadrature form of equation 12.16 and $I(t)$ and $Q(t)$ are referred to as the in-phase and quadrature components. For BPSK, $a_i[n] = \pm a$ and $a_q[n] = 0$.

For PSK with θ_n in the form of equation 12.17 and $M = 4$, θ_n can take on any of the four values 0, $\frac{\pi}{2}$, π, or $\frac{3\pi}{2}$. In the form of equations 12.22 and 12.23 $a_i[n]$ will then be either $+a, -a$, or zero and $a_q[n]$ will be either $+a, -a$, or zero. However, clearly QPSK can only encode four symbols in the phase not nine, i.e. the various possibilities for $a_i[n]$ and $a_q[n]$ are not independent. For example, for $M = 4$, if $a_i[n] = +a$ then $a_q[n]$ must be zero since $a_i[n] = +a$ implies that $\theta_n = 0$. A convenient way of looking at this is through what's referred to as an I-Q constellation as shown in Figure 12.8.

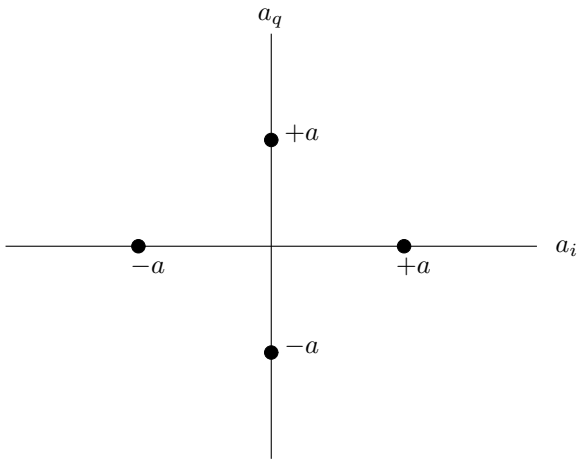

FIGURE 12.8 I-Q Constellation for QPSK.

Each point in the constellation represents a different symbol that can be encoded, and clearly with the constellation of Figure 12.8 one of four symbols can be encoded in each symbol interval (recall that for now, the amplitude $a[n]$ is constant. This will change when we expand the discussion shortly to QAM).

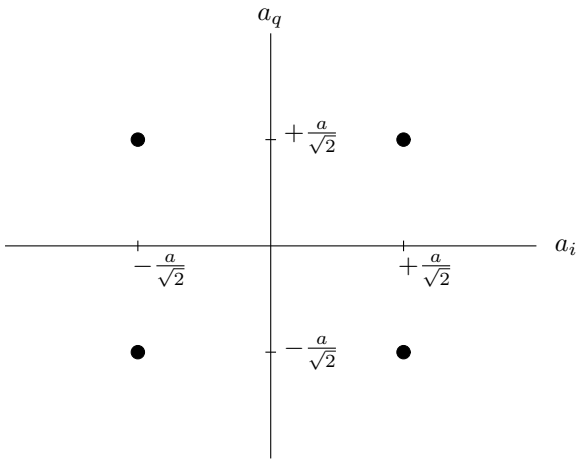

FIGURE 12.9 I-Q Constellation for quadrature phase-shift-keying (QPSK).

An alternative form with four-phase PSK is to choose

$$\theta_n = \frac{2\pi b_n}{4} + \frac{\pi}{4} \; ; \; b_n \text{ a non-negative integer} \tag{12.24}$$

in which case $a_i[n] = \pm\frac{a}{\sqrt{2}}$ and $a_q[n] = \pm\frac{a}{\sqrt{2}}$ resulting in the constellation in Figure 12.9.

In this case, the amplitude modulation of $I(t)$ and $Q(t)$ (equations 12.20 and 12.21) can be done independently. Modulation with this constellation is commonly referred to as QPSK (quadrature phase-shift keying).

In PSK as described above, $a[n]$ was assumed constant. By incorporating encoding in both the amplitude $a[n]$ and phase θ_n in equation 12.16 we are led to a richer form of modulation referred to as quadrature amplitude modulation (QAM). In the form of equations (12.19 - 12.21) we now allow $a_i[n]$ and $a_q[n]$ to be chosen from a richer constellation.

12.3.3 QAM

The QAM constellation diagram is shown in Figure 12.10 for the case where each set of amplitudes can take the values $\pm a$ and $\pm 3a$. The 16 different combinations that are available in this case can be used to code 4 bits, as shown in the figure. This particular constellation is what is used in the V.32ALT standard shown in the table of Figure 12.7. In this standard, the carrier frequency is 1,800 Hz, and the symbol frequency or baud rate $(1/T)$ is 2,400 Hz. With 4 bits per symbol, this works out to the indicated 9,600 bits/second. One baseband pulse shape $p(t)$ that may be used is the square root of the cosine-transition pulse mentioned earlier, say with $\beta = 0.3$. This pulse contains frequencies as high as $1.3 \times 1,200 = 1,560$ Hz.

After modulation of the 1,800 Hz carrier, the signal occupies the band from 240 Hz to 3,360 Hz, which is right in the passband of the voice telephone channel.

The two faster modems shown in the table use more elaborate QAM-based schemes. The V.32bis standard involves 128QAM, which could in principle convey 7 bits per symbol, but at the price of greater sensitivity to noise (because the constellation points are more tightly clustered for a given signal power). However, the QAM in this case is actually combined with so-called *trellis-coded modulation* (TCM), which in effect codes in some redundancy (by introducing dependencies among the modulating amplitudes), leading to greater noise immunity and an effective rate of 6 bits per symbol (think of the TCM as in effect reserving a bit for error checking). The symbol rate here is still 2,400 Hz, so the transmission is at $6 \times 2,400 = 14,400$ bits/second. Similarly, the V.34 standard involves 1024QAM, which could convey 10 bits per symbol, although with more noise sensitivity. The combination with TCM introduces redundancy for error control, and the resulting bit rate is 28,800 bits/second (9 effective bits times a symbol frequency of 3,200 Hz).

Demodulation of Quadrature Modulated PAM signals:
The carrier modulated signals in the form of equations (12.19 - 12.23) can carry encoded data in both the I and Q components $I(t)$ and $Q(t)$. Therefore in demodulation we must be able to extract these seperately. This is done through quadrature demodulation as shown in Figure 12.11

In both the modulation and demodulation, it is assumed that the bandwidth of $p(t)$ is low compared with the carrier frequency w_c so that the bandwidth of $I(t)$ and $Q(t)$ are less than ω_c. The input signal $r_i(t)$ is

$$r_i(t) = I(t)cos^2(\omega_c t) - Q(t)sin(\omega_c t)cos(\omega_c t) \tag{12.25}$$

$$= \frac{1}{2}I(t) - \frac{1}{2}I(t)cos(2\omega_c t) - \frac{1}{2}Q(t)sin(2\omega_c t) \tag{12.26}$$

Similarly

$$r_q(t) = I(t)cos(\omega_c t)sin(\omega_c t) - Q(t)sin^2(\omega_c t) \tag{12.27}$$

$$= \frac{1}{2}I(t)sin(2\omega_c t) + \frac{1}{2}Q(t) - \frac{1}{2}Q(t)cos(2\omega_c t) \tag{12.28}$$

Choosing the cutoff frequency of the lowpass filters to be greater than the bandwidth of $p(t)$ (and therefore also greater than the bandwidth of $I(t)$ and $Q(t)$) but low enough to eliminate the components in $r_i(t)$ and $r_q(t)$ around $2\omega_c$, the outputs will be the quadrature signals $I(t)$ and $Q(t)$.

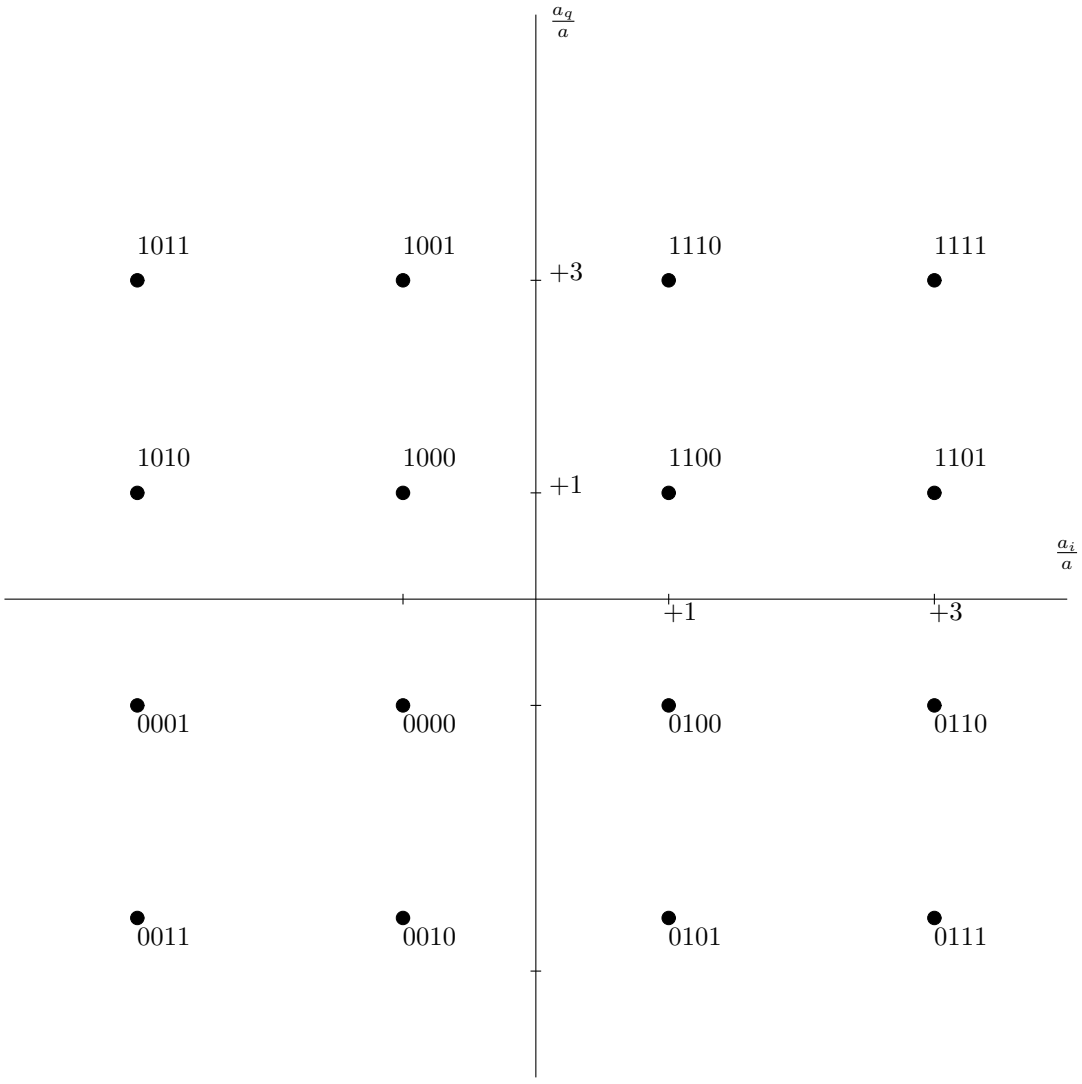

FIGURE 12.10 16 QAM constellation. (From *Digital Transmission Engineering* by J.B. Anderson, IEEE Press, 1999, p.96)

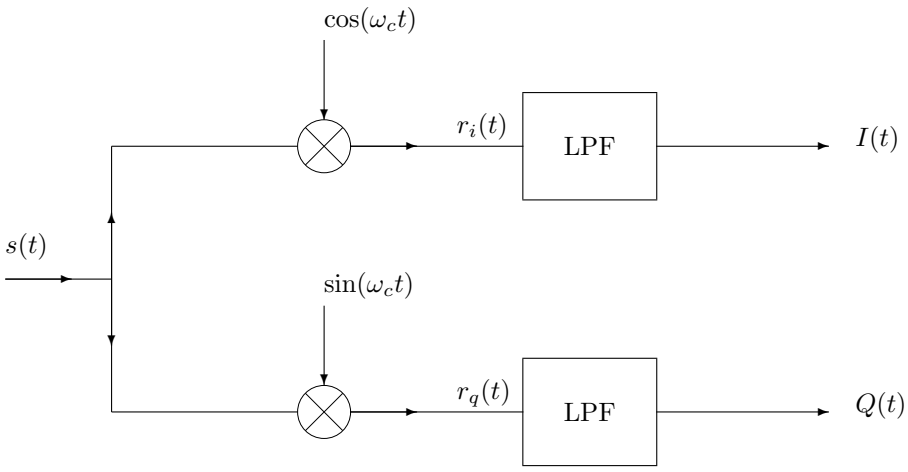

FIGURE 12.11 Demodulation scheme for a Quadrature Modulated PAM Signal.

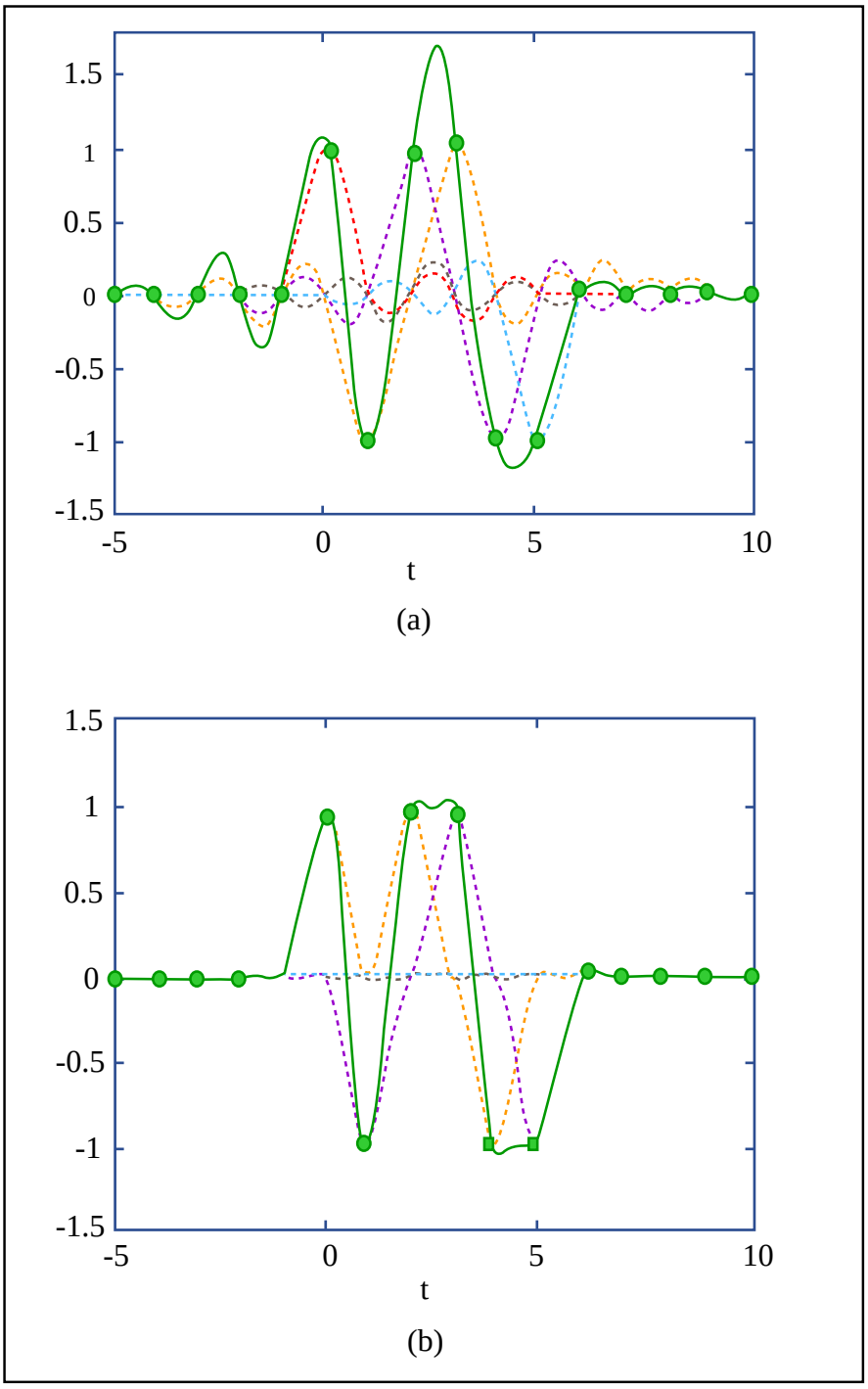

Image by MIT OpenCourseWare, adapted from *Digital Transmission Engineering*, John Anderson. IEEE Press, 1999.

FIGURE 12.12 (a) PAM signal with sinc pulse. (b) PAM signal with 'raised cosine' pulse. Note much larger tails and excursions in narrow band pulse of (a); tails may not be truncated without widening the bandwidth. (From J.B. Anderson, *Digital Transmission Engineering*, IEEE Press, 1999.)

CHAPTER 13

Hypothesis Testing

INTRODUCTION

The topic of hypothesis testing arises in many contexts in signal processing and communications, as well as in medicine, statistics and other settings in which a choice among multiple options or hypotheses is made on the basis of limited and noisy data. For example, from tests on such data, we may need to determine: whether a person does or doesn't have a particular disease; whether or not a particular radar return indicates the presence of an aircraft; which of four values was transmitted at a given time in a PAM system; and so on.

Hypothesis testing provides a framework for selecting among M possible choices or hypotheses in some principled or optimal way. In our discussion we will initially focus on $M = 2$, i.e., on binary hypothesis testing, to illustrate the key concepts. Though Section 13.1 introduces the discussion in the context of binary pulse amplitude modulation in noise, the presentation and results in Section 13.2 apply to the general problem of binary hypothesis testing. In Sections 13.3 and 13.4 we explicitly treat the case of more than two hypotheses.

13.1 BINARY PULSE AMPLITUDE MODULATION IN NOISE

In Chapter 12 we introduced the basic principles of pulse amplitude modulation, and considered the effects of pulse rate, pulse shape, and channel and receiver filtering in PAM systems. We also developed and discussed the condition for no inter-symbol interference (the no-ISI condition). Under the assumption of no ISI, we want to now examine the effect of noise in the channel. Toward this end, we again consider the overall PAM model in Figure 13.1, with the channel noise $v(t)$ represented as an additive term.

For now we will assume no post-filtering at the receiver, i.e., assume $f(t) = \delta(t)$. In Chapter 14 we will see how performance is improved with the use of filtering in the receiver. The basic pulse $p(t)$ going through the channel with impulse response $h(t)$ produces a signal at the channel output that we represent by $s(t) = p(t) * h(t)$. Figure 13.1 thus reduces to the overall system shown in Figure 13.2.

Since we are assuming no ISI, we can carry out our discussion for just a single pulse index n, which we will choose as $n = 0$ for convenience. We therefore focus, in the system of Figure 13.2, on

$$b[0] = r(0) = a[0]s(0) + v(0) \; . \tag{13.1}$$

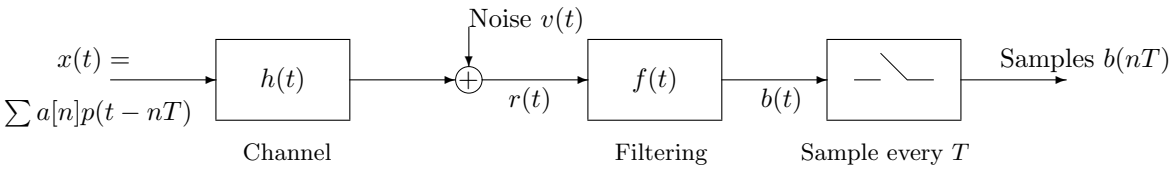

FIGURE 13.1 Overall model of a PAM system.

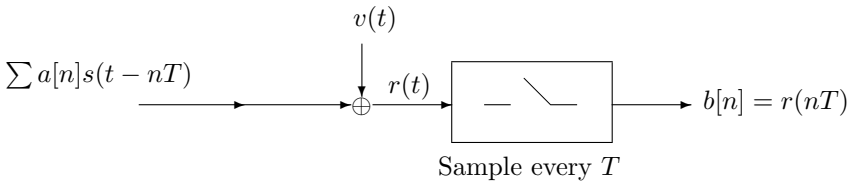

FIGURE 13.2 Simplified representation of a PAM system.

Writing $r(0)$, $a[0]$ and $v(0)$ simply as r, a and v respectively, and setting $s(0) = 1$ without loss of generality, the relation of interest to us is

$$r = a + v \, . \tag{13.2}$$

Our broad objective is to determine the value of a as well as possible, given the measured value r. There are several variations of this problem, depending on the nature of the transmitted sequence $a[n]$ and the characteristics of the noise. The amplitude $a[n]$ may span a continuous range or it may be discrete (e.g., binary). The amplitude may correspondingly be modeled as a random variable A with a known PDF or PMF; then a is the specific value that A takes in a particular outcome or instance of the probabilistic model. The contribution of the noise also is typically represented as a random variable V, usually continuous, with v being the specific value that it takes. We may thus model the quantity r at the receiver as the observation of a random variable R, with

$$R = A + V \, , \tag{13.3}$$

and we want to estimate the value that the random variable A takes, given that $R = r$. Consequently, we need to add a further processing step to our receiver, in which an estimate of A is obtained.

In the case where the pulse amplitude can be only one of two values, i.e., in the case of binary signaling, finding an estimate of A reduces to deciding, on the basis of the observed value r of R, which of the two possible amplitudes was transmitted. Two common forms of binary signaling in PAM systems are on/off signaling and

antipodal signaling. Letting a_1 and a_0 denote the two possible amplitudes (representing for example a binary "one" or "zero"), in on/off signaling we have $a_0 = 0$, $a_1 \neq 0$, whereas in antipodal signaling $a_0 = -a_1 \neq 0$.

Thus, in binary signaling, the required post-processing corresponds to deciding between two alternatives or hypotheses, where the available information may include some prior information along with a measurement r of the single continuous random variable R. (The extension to multiple hypotheses and multiple measurements will be straightforward once the two-hypothesis case is understood.) The hypotheses are listed below:

Hypothesis H_0: the transmitted amplitude A takes the value a_0, so $R = a_0 + V$.

Hypothesis H_1: the transmitted amplitude A takes the value a_1, so $R = a_1 + V$.

Our task now is to decide, given the measurement $R = r$, whether H_0 or H_1 is responsible for the measurement. The next section develops a framework for this sort of hypothesis testing task.

13.2 BINARY HYPOTHESIS TESTING

Our general binary hypothesis testing task is to decide, on the basis of a measurement r of a random variable R, which of two hypotheses — H_0 or H_1 — is responsible for the measurement. We shall indicate these decisions by 'H_0' and 'H_1' respectively (where the quotation marks are intended to suggest the announcement of a decision). An alternative notation is $\widehat{H} = H_0$ and $\widehat{H} = H_1$ respectively, where $\widehat{H}$ denotes our estimate of, or decision on, the hypothesis H.

Suppose H is modeled as a random quantity, and assume we know the *a priori* (i.e., prior) probabilities

$$P(H_0 \text{ is true}) = P(H = H_0) = P(H_0) = p_0 \tag{13.4}$$

and

$$P(H_1 \text{ is true}) = P(H = H_1) = P(H_1) = p_1 \tag{13.5}$$

(where the last two equalities in each case simply define streamlined notation that we will be using). We shall also require the conditional densities $f_{R|H}(r|H_0)$ and $f_{R|H}(r|H_1)$ that tell us how the measured variable is distributed under the two respective hypotheses. These conditional densities in effect constitute the relevant specifications of how the measured data relates to the two hypotheses. For example, in the PAM setting, with R defined as in (13.3) and assuming V is independent of A under each hypothesis, these conditional densities are simply

$$f_{R|H}(r|H_0) = f_V(r - a_0) \quad \text{and} \quad f_{R|H}(r|H_1) = f_V(r - a_1)\,. \tag{13.6}$$

It is natural in many settings, as in the case of digital communication by PAM, to want to minimize the probability of picking the wrong hypothesis, i.e., to choose with minimum probability of error between the hypotheses, given the measurement $R = r$. We will, for most of our discussion of hypothesis testing, focus on this criterion of minimum probability of error.

13.2.1 Deciding with Minimum Probability of Error: The MAP Rule

Consider first how one would choose between H_0 and H_1 with minimum probability of error in the absence of any measurement of R. If we make the choice 'H_0', then we make an error precisely when H_0 does not hold, so the probability of error with this choice is $1 - P(H_0) = 1 - p_0$. Similarly, if we chose 'H_1', then the probability of error is $1 - P(H_1) = 1 - p_1 = p_0$. Thus, for minimum probability of error, we should decide in favor of whichever hypothesis has maximum probability — an intuitively reasonable conclusion. (The preceding reasoning extends in the same way to choosing one from among many hypotheses, and leads to the same conclusion.)

What changes when we aim to choose between H_0 and H_1 with minimum probability of error, knowing that $R = r$? The same reasoning applies as in the preceding paragraph, except that all probabilities now need to be conditioned on the measurement $R = r$. We conclude that to minimize the conditional probability of error, $P(\text{error}|R = r)$, we need to decide in favor of whichever hypothesis has maximum conditional probability, conditioned on the measurement $R = r$. (If there were several random variables for which we had measurements, rather than just the single random variable R, we would simply condition on all the available measurements.) Thus, if $P(H_1|R = r) > P(H_0|R = r)$, we decide '$H_1$', and if $P(H_1|R = r) < P(H_0|R = r)$, we decide '$H_0$'. This may be compactly written as

$$P(H_1|R = r) \underset{\text{'}H_0\text{'}}{\overset{\text{'}H_1\text{'}}{\gtrless}} P(H_0|R = r) \; . \tag{13.7}$$

(If the two conditional probabilities happen to be equal, we get the same conditional probability of error whether we choose 'H_0' or 'H_1'.) The corresponding conditional probability of error is

$$P(\text{error}|R = r) = \min\{1 - P(H_0|R = r), 1 - P(H_1|R = r)\} \; . \tag{13.8}$$

The overall probability of error, P_e, associated with the use of the above decision rule (but before knowing what specific value of R is measured) is obtained by averaging the conditional probability of error in (13.8) over all possible values of r that might be measured, using the PDF $f_R(r)$ as a weighting function. We shall study P_e in more detail shortly.

The conditional probabilities $P(H_0|R = r)$ and $P(H_1|R = r)$ that appear in the expression (13.7) are referred to as the *a posteriori* or posterior probabilities of the hypotheses, to distinguish them from the *a priori* or prior probabilities, $P(H_0)$ and $P(H_1)$. The decision rule in (13.7) is accordingly referred to as the maximum *a posteriori* probability rule, usually abbreviated as the "MAP" rule.

To actually evaluate the posterior probabilities in (13.7), we use Bayes' rule to

rewrite them in terms of known quantities, so the decision rule becomes

$$\frac{p_1 f_{R|H}(r|H_1)}{f_R(r)} \begin{array}{c} \text{`}H_1\text{'} \\ > \\ < \\ \text{`}H_0\text{'} \end{array} \frac{p_0 f_{R|H}(r|H_0)}{f_R(r)} , \tag{13.9}$$

under the reasonable assumption that $f_R(r) > 0$, i.e., that the PDF of R is positive at the value r that was actually measured. (In any case, we only need to specify our decision rule at values of r for which $f_R(r) > 0$, because the choices made at other values of r do not affect the overall probability of error, P_e.) Since the denominator is the same and positive on both sides of the above expression, we may further simplify it to

$$p_1 f_{R|H}(r|H_1) \begin{array}{c} \text{`}H_1\text{'} \\ > \\ < \\ \text{`}H_0\text{'} \end{array} p_0 f_{R|H}(r|H_0) . \tag{13.10}$$

This now provides us with an easily visualized and implemented decision rule. We first use the prior probabilities $p_i = P(H_i)$ to scale the PDFs $f_{R|H}(r|H_i)$ that describe how the measured quantity R is distributed under each of the hypotheses. We then decide in favor of the hypothesis associated with whichever scaled PDF is largest at the measured value r. (The preceding description also applies to choosing with minimum probability of error among multiple hypotheses, rather than just two, and given measurements of several associated random variables, rather than just one — the reasoning is identical.)

13.2.2 Understanding P_e: False Alarm, Miss and Detection

The sample space that is relevant to evaluating a decision rule consists of the following four mutually exclusive and collectively exhaustive possibilities: H_i is true and we declare 'H_j', $i, j = 1, 2$. Of the four possible outcomes, the two that represent errors are (H_0,'H_1') and (H_1,'H_0'). Therefore, the probability of error P_e — averaged over all possible values of the measured random variable — is given by

$$\begin{aligned} P_e &= P(H_0, \text{`}H_1\text{'}) + P(H_1, \text{`}H_0\text{'}) \\ &= p_0 P(\text{`}H_1\text{'}|H_0) + p_1 P(\text{`}H_0\text{'}|H_1) . \end{aligned} \tag{13.11}$$

The conditional probability $P(\text{`}H_1\text{'}|H_0)$ is referred to as the conditional probability of a false alarm, and denoted by $P_{FA}$. The conditional probability $P(\text{`}H_0\text{'}|H_1)$ is referred to as the conditional probability of a miss, and denoted by P_M. The word "conditional" is usually omitted from these terms in normal use, but it is important to keep in mind that the probability of a false alarm and the probability of a miss are defined as conditional probabilities, and are furthermore conditioned on different events.

The preceding terminology is historically motivated by the radar context, in which H_1 represents the presence of a target and H_0 the absence of a target. A false

alarm then occurs if you declare that a target is present when it actually isn't, and a miss occurs if you declare that a target is absent when it actually isn't. We will also make reference to the conditional probability of detection,

$$P_D = P(\text{`}H_1\text{'}|H_1) \; . \tag{13.12}$$

In the radar context, this is the probability of declaring a target is present when it is actually present. As with P_{FA} and P_M, the word "conditional" is usually omitted in normal use, but it is important to keep in mind that the probability of detection is a conditional probability.

Expressing the probability of error in terms of P_{FA} and P_M, (13.11) becomes

$$P_e = p_0 P_{FA} + p_1 P_M \; . \tag{13.13}$$

Also note that

$$P(\text{`}H_0\text{'}|H_1) + P(\text{`}H_1\text{'}|H_1) = 1 \tag{13.14}$$

or

$$P_M = 1 - P_D \; . \tag{13.15}$$

To explicitly relate P_{FA} and P_M to whatever the corresponding decision rule is, it is helpful to introduce the notion of a decision region in measurement space. In the case of a decision rule based on measurement of a single random variable R, specifying the decision rule corresponds to choosing a range of values D_1 on the real line such that, when the measured value r of R falls in D_1, we declare 'H_1', and when r falls outside D_1 — a region that we shall denote by D_0 — then we declare 'H_0'. This is illustrated in Figure 13.3, for some arbitrary choice of D_1. (There is a direct generalization of this notion to the case where multiple random variables are measured.)

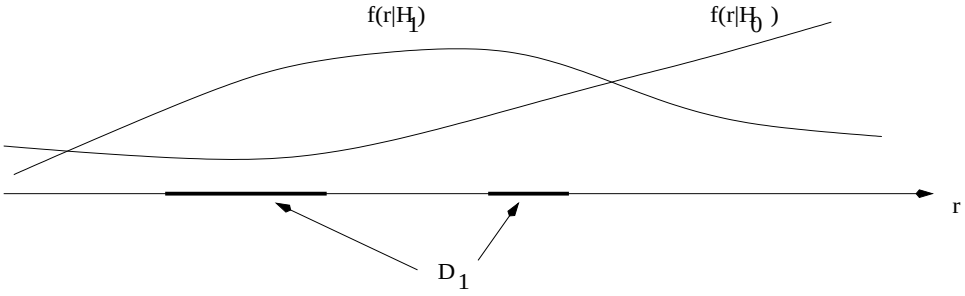

FIGURE 13.3 Decision regions. The choice of D_1 marked here is arbitrary, not the optimal choice for minimum probability of error.

With the preceding definitions, we can write

$$P_{FA} = \int_{D_1} f_{R|H}(r|H_0)dr \tag{13.16}$$

and

$$P_M = \int_{D_0} f_{R|H}(r|H_1)dr \ . \tag{13.17}$$

13.2.3 The Likelihood Ratio Test

Rewriting (13.10), we can state the minimum-P_e decision rule in the form

$$\Lambda(r) = \frac{f_{R|H}(r|H_1)}{f_{R|H}(r|H_0)} \begin{array}{c} \text{`}H_1\text{'} \\ > \\ < \\ \text{`}H_0\text{'} \end{array} \frac{p_0}{p_1} \tag{13.18}$$

or

$$\Lambda(r) \begin{array}{c} \text{`}H_1\text{'} \\ > \\ < \\ \text{`}H_0\text{'} \end{array} \eta \ , \tag{13.19}$$

where $\Lambda(r)$ is referred to as the likelihood ratio, and η is referred to as the threshold. This particular way of writing our decision rule is of interest because other formulations of the binary hypothesis testing problem — with criteria other than minimization of P_e — also often lead to a decision rule that involves comparing the likelihood ratio with a threshold. The only difference is that the threshold is picked differently in these other formulations. We describe two of these alternate formulations — the Neyman-Pearson approach, and minimum risk decisions — in later sections of this chapter.

13.2.4 Other Scenarios

While the above discussion of binary hypothesis testing was introduced in the context of binary PAM, it applies in many other scenarios. For example, in the medical literature, clinical tests are described using a hypothesis testing framework similar to that used here for communication and signal detection problems, with H_0 generally denoting the absence of a medical condition and H_1 its presence. The terminology in the medical context is slightly different, but still suggestive of the intent, as the following examples show:

- P_D is the sensitivity of the clinical test.
- $P(\text{`}H_1\text{'}|H_0)$ is the probability of a false positive (rather than of a false alarm).
- $1 - P_{FA}$ is the specificity of the test.
- $P(H_1)$ is the prevalence of the condition that the test is aimed at.
- $P(H_1 \,|\, \text{`}H_1\text{'})$ is the positive predictive value of the test, and $P(H_0 \,|\, \text{`}H_0\text{'})$ is the negative predictive value.

Some easy exploration using Bayes' rule and the above terminology will lead you to recognize how small the positive predictive value of a test can be if the prevalence of the targeted medical condition is low, even if the test is highly sensitive and specific.

Another important context for binary hypothesis testing is in target detection, such as aircraft detection and tracking, in which a radar pulse is transmitted and the decision on the presence or absence of an aircraft is based on the presence or absence of reflected energy.

13.2.5 Neyman-Pearson Detection and Receiver Operating Characteristics

A difficulty with using the minimization of P_e as the decision criterion in many of these other contexts is that it relies heavily on knowing the *a priori* probabilities p_0 and p_1, and in many situations there is little basis for coming up with these numbers. One alternative that often makes sense is to maximize the probability of detection P_D, while keeping P_{FA} below some specified tolerable level. These conditional probabilities are determined by the measurement models under the different hypotheses, and by the decision rule, but not by the probabilities governing the selection of hypotheses. Such a formulation of the hypothesis testing problem again leads to a decision rule that involves comparing the likelihood ratio with a threshold; the only difference now is that the threshold is picked differently in this formulation. This approach is referred to as Neyman-Pearson detection, and is elaborated on below.

Consider a context in which we want to maximize the probability of detection,

$$P_D = P(\text{`}H_1\text{'}|H_1) = \int_{D_1} f_{R|H}(r|H_1)dr \ , \tag{13.20}$$

while keeping the probability of false alarm,

$$P_{FA} = P(\text{`}H_1\text{'}|H_0) = \int_{D_1} f_{R|H}(r|H_0)dr \ , \tag{13.21}$$

below a pre-specified level. (Both integrals are over the decision region D_1, and augmenting D_1 by adding more of the real axis to it will not decrease either probability.) As we show shortly, we can achieve our objective by picking the decision region D_1 to comprise those values of r for which the likelihood ratio $\Lambda(r)$ exceeds a certain threshold η, so

$$\Lambda(r) = \frac{f_{R|H}(r|H_1)}{f_{R|H}(r|H_0)} \begin{array}{c} \text{`}H_1\text{'} \\ > \\ < \\ \text{`}H_0\text{'} \end{array} \eta \ . \tag{13.22}$$

The threshold η is picked to provide the largest possible P_D while ensuring that P_{FA} is not larger than the pre-specified level. The smaller the η, the larger the decision region D_1 and the value of P_D become, but the larger P_{FA} grows as well, so one would pick the smallest η that is consistent with the given bound on P_{FA}.

To understand why the decision rule in this setting takes the form of (13.22), note that our objective is to include in D_1 values of r that contribute as much as possible to the integral that defines P_D, and as little as possible to the integral that defines P_{FA}. If we start with a high value of the threshold η, we will be including in D_1 those r for which $\Lambda(r)$ is large, and therefore where the contribution to P_D is relatively large compared to the contribution to P_{FA}. Moving η lower, we increase both P_D and P_{FA}, but the rate of increase of P_D drops, while the rate of increase of P_{FA} rises. These increases in P_D and P_{FA} may not be continuous in η. (Reducing η from infinitesimally above some value $\overline{\eta}$ to infinitesimally below this value will give rise to a finite upward jump in both P_D and P_{FA} if $f_{R|H}(r|H_1) = \overline{\eta}\, f_{R|H}(r|H_0)$ throughout some interval of r where both these PDFs are positive.) Typically, though, the variation of P_D and P_{FA} with η is indeed continuous, so as η is lowered we reach a point where the specified bound on P_{FA} is attained, or $P_D = 1$ is reached. This is the value of η used in the Neyman-Pearson test. (In the rare situation where P_{FA} jumps discontinuously from a value below its tolerable level to one above its tolerable level as η is lowered through some value $\overline{\eta}$, it turns out that a randomized decision rule allows one to come right up to the tolerable P_{FA} level, and ! thereby maximize P_D. A case like this is explored in a problem at the end of this chapter.)

The following argument shows in a little more detail, though still informally, why the Neyman-Pearson criterion is equivalent to a likeliood ratio test. If the decision region D_1 is optimal for the Neyman-Pearson criterion, then any change in D_1 that keeps P_{FA} the same cannot lead to an improvement in P_D. So suppose we take a infinitesimal segment of width dr at a point r in the optimal D_1 region and convert it to be part of D_0. In order to keep P_{FA} unchanged, we must correspondingly take an infinitesimal segment of width dr' at an arbitrary point r' in the optimal D_0 region, and convert it to be a part of D_1.

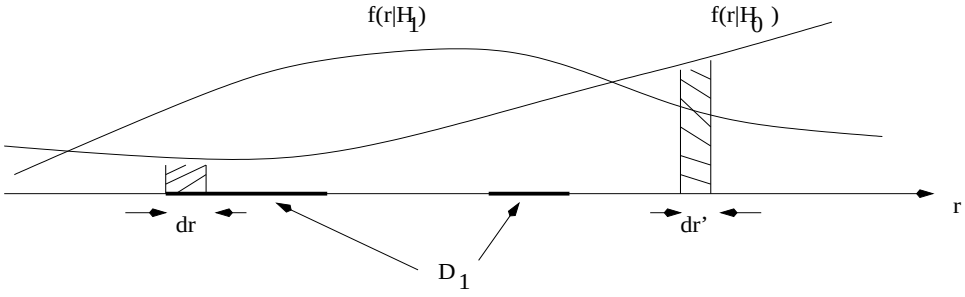

FIGURE 13.4 Illustrating the construction used in deriving the likelihood ratio test for the Neyman-Pearson criterion.

The requirement that P_{FA} be unchanged then imposes the condition

$$f_{R|H}(r'|H_0)\, dr' = f_{R|H}(r|H_0)\, dr\ , \tag{13.23}$$

while the requirement that the new P_D not be larger than the old implies that

$$f_{R|H}(r'|H_1)\,dr' \leq f_{R|H}(r|H_1)\,dr\ . \tag{13.24}$$

Combining (13.23) and (13.24), we find

$$\Lambda(r') \leq \Lambda(r)\ . \tag{13.25}$$

What (13.25) shows is that the likelihood ratio cannot be less inside D_1 than it is in D_0. We can therefore conclude that the optimum solution to the Neyman-Pearson formulation is in fact based on a threshold test on the likelihood ratio:

$$\Lambda(r) = \frac{f_{R|H}(r|H_1)}{f_{R|H}(r|H_0)} \begin{array}{c} \text{`}H_1\text{'} \\ > \\ < \\ \text{`}H_0\text{'} \end{array} \eta\ , \tag{13.26}$$

where the threshold η is picked to obtain the largest possible P_D while ensuring that P_{FA} is not larger than the pre-specified bound.

The above derivation has made various implicit assumptions. However, our purpose is only to convey the essence of how one arrives at a likelihood ratio test in this case.

Receiver Operating Characteristic. In considering which value of P_{FA} to choose as a bound in the Neyman-Pearson test, it is often useful to look at a curve of P_D versus P_{FA} as the parameter η is varied. This is referred to as the Receiver Operating Characteristic (ROC). More generally, such an ROC can be defined for any decision rule that causes P_D to be uniquely fixed, once P_{FA} is specified. The ROC can be used to identify whether, for instance, modifying the variable parameters in a given test to permit a slightly higher P_{FA} results in a significantly higher P_D. The ROC can also be used to compare different tests.

EXAMPLE 13.1 Detection and ROC for Signal in Gaussian Noise

Consider a scenario in which a radar pulse is emitted from a ground station. If an aircraft is located in the propagation path, a reflected pulse will travel back towards the radar station. We assume that the received signal will then consist of noise alone if no aircraft is present, and noise plus the reflected pulse if an aircraft is present. The processing of the received signal results in a number that we model as the realization of a random variable R. If an aircraft is not present, then $R = W$, where W is a random variable denoting the result of processing just the noise. If an aircraft is present, then $R = s + W$, where the constant s is due to processing of the reflected pulse, and is assumed here to be a known value. We thus have the following two hypotheses:

$$H_0:\ \ R = W \tag{13.27}$$

$$H_1:\ \ R = s + W\ . \tag{13.28}$$

Assume that the additive noise term W is Gaussian with zero mean and unit variance, i.e.,

$$f_W(w) = \frac{1}{\sqrt{2\pi}} e^{-w^2/2}. \tag{13.29}$$

Consequently,

$$f_{R|H}(r|H_0) = \frac{1}{\sqrt{2\pi}} e^{-r^2/2} \tag{13.30}$$

$$f_{R|H}(r|H_1) = \frac{1}{\sqrt{2\pi}} e^{-(r-s)^2/2} \,. \tag{13.31}$$

The likelihood ratio as defined in (13.18) is then

$$\begin{aligned} \Lambda(r) &= \exp\Big[-\frac{(r-s)^2}{2} + \frac{r^2}{2}\Big] \\ &= \exp\Big[sr - \frac{s^2}{2}\Big] \,. \end{aligned} \tag{13.32}$$

For detection with minimum probability of error, the decision rule corresponds to evaluating this likelihood ratio at the received value r, and comparing the result against the threshold p_0/p_1, as stated in (13.18):

$$\exp\Big[sr - \frac{s^2}{2}\Big] \underset{\text{`}H_0\text{'}}{\overset{\text{`}H_1\text{'}}{\gtrless}} \eta = \frac{p_0}{p_1} \tag{13.33}$$

It is interesting and important to note that, for this case, the threshold test on the likelihood ratio can be rewritten as a threshold test on the received value r. Specifically, (13.33) can equivalently be expressed as

$$\Big[sr - \frac{s^2}{2}\Big] \underset{\text{`}H_0\text{'}}{\overset{\text{`}H_1\text{'}}{\gtrless}} \ln \eta \,, \tag{13.34}$$

or, if $s > 0$,

$$r \underset{\text{`}H_0\text{'}}{\overset{\text{`}H_1\text{'}}{\gtrless}} \frac{1}{s}\Big[\frac{s^2}{2} + \ln \eta\Big] = \gamma \,, \tag{13.35}$$

where γ denotes the threshold on r. (If $s < 0$, the inequalities in (13.35) are simply reversed.) For example, if both hypotheses are equally likely *a priori*, so that $p_0 = p_1$, then $\ln \eta = 0$ and the decision rule for minimum probability of error when $s > 0$ is simply

$$r \underset{\text{`}H_0\text{'}}{\overset{\text{`}H_1\text{'}}{\gtrless}} \frac{s}{2} = \gamma \,. \tag{13.36}$$

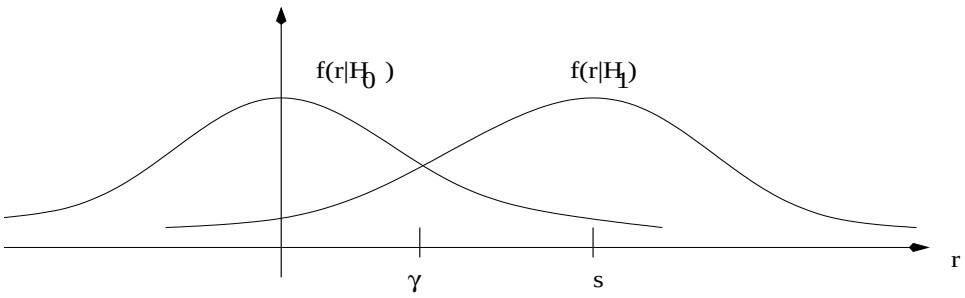

FIGURE 13.5 Threshold γ on measured value r.

The situation is represented in Figure 13.5.

The receiver operating characteristic displays P_D versus P_{FA} as η is varied, and is sketched in Figure 13.6.

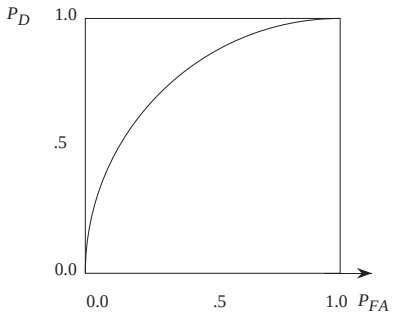

FIGURE 13.6 Receiver operating characteristic.

In a more general setting than the Gaussian case in Example 13.1, a threshold test on the likelihood ratio would not simply translate to a threshold test on the measurement r. Nevertheless, we could still decide to use a simple threshold test on r as our decision rule, and then generate and evaluate the associated receiver operating characteristic.

13.3 MINIMUM RISK DECISIONS

This section briefly describes a decision criterion, called minimum risk, that includes minimum probability of error as a special case, and that in the binary case again leads to a likelihood ratio test. We describe it for the general case of M hypotheses.

Let the available measurement be the value r of the random variable R (the same

development holds if we have measurements of several random variables). Suppose we associate a cost c_{ij} with each combination of model H_j and decision 'H_i' for $0 \leq i, j \leq M-1$, reflecting the costs of actions and consequences that follow from this combination of model and decision. Our objective now is to pick whichever decision has minimum expected cost, or minimum "risk", given the measurement.

The expected cost of deciding 'H_i', conditioned on $R = r$, is given by

$$E[\text{Cost}|R = r, \text{`}H_i\text{'}] = \sum_{j=0}^{M-1} c_{ij} P(H_j|R = r, \text{`}H_i\text{'}) = \sum_{j=0}^{M-1} c_{ij} P(H_j|R = r) \ , \quad (13.37)$$

where the last equality is a consequence of the fact that, given the received measurement $R = r$, the output of the decision rule conveys no additional information about which hypothesis actually holds. The next step is to compare these conditional expected costs for all i, and decide in favor of the hypothesis with minimum conditional expected cost. Specifying our decision for each possible r, we obtain the decision rule that minimizes the overall expected cost or risk.

[It is in this setting that hypothesis testing comes closest to the estimation problems for continuous random variables that we considered in our chapter on minimum mean-square-error estimation. We noted there that a variety of such estimation problems can be formulated in terms of minimizing an expected cost function. Establishing an estimate for a random variable is like carrying out a hypothesis test for a continuum of numerically specified hypotheses (rather than just M general hypotheses), with a cost function that penalizes some measure of the numerical distance between the actual hypothesis and the one we decide on.]

Note that if $c_{ii} = 0$ for all i and if $c_{ij} = 1$ for $j \neq i$, so we penalize all errors equally, then the conditional expected cost in (13.37) becomes

$$E[\text{Cost}|R = r, \text{`}H_i\text{'}] = \sum_{j \neq i} P(H_j|r) = 1 - P(H_i|r) \ . \quad (13.38)$$

This conditional expected cost is thus precisely the conditional probability of error associated with deciding 'H_i', conditioned on $R = r$. The right side of the equation then shows that to minimize this conditional probability of error we should decide in favor of the hypothesis with largest conditional probability. In other words, with this choice of costs, the risk (when the expectation is taken over all possible values of r) is exactly the probability of error P_e, and the optimum decision rule for minimizing this criterion is again seen to be the MAP rule.

Using Bayes' rule in (13.37) and noting that $f_R(r)$ — assumed positive — is common to all the quantities involved in our comparison, we see that an equivalent but more directly implementable procedure is to pick the hypothesis for which

$$\sum_{j=0}^{M-1} c_{ij} f(r|H_j) P(H_j) \quad (13.39)$$

is minimum. In the case of two hypotheses, and assuming $c_{01} > c_{11}$, it is easy to

see that the decision rule based on (13.39) can be rewritten as

$$\Lambda(r) = \frac{f(r|H_1)}{f(r|H_0)} \begin{array}{c} \text{`}H_1\text{'} \\ > \\ < \\ \text{`}H_0\text{'} \end{array} \frac{P(H_0)(c_{10} - c_{00})}{P(H_1)(c_{01} - c_{11})} = \eta \,, \tag{13.40}$$

where $\Lambda(r)$ denotes the likelihood ratio, and η is the threshold. We have therefore again arrived at a decision rule that involves comparing a likelihood ratio with a threshold. If $c_{ii} = 0$ for $i = 0, 1$ and if $c_{ij} = 1$ for $j \neq i$, then we obtain the threshold associated with the MAP decision rule for minimum P_e, as expected.

The trouble with the above minimum risk approach to classification, and with the minimum error probability formulation that we have examined a few times already, is the requirement that the prior probabilities $P(H_i)$ be known.

It is often unrealistic to assume that prior probabilities are known, so we are led to consider alternative criteria. Most important among these alternatives is the Neyman-Pearson approach treated earlier, where the decision is based on the conditional probabilities P_D and P_{FA}, thereby avoiding the need for prior probabilities on the hypotheses.

13.4 HYPOTHESIS TESTING IN CODED DIGITAL COMMUNICATION

In our discussion of PAM earlier in this chapter, we considered binary hypothesis testing on a single received pulse. In modern communication systems, an alphabet of symbols may be transmitted, with each symbol encoded into a binary sequence of "ones" and "zeroes". Consequently, in addition to making a binary decision on each received pulse, we may need to further decode a string of bits to make our best judgement of the transmitted symbol, and perhaps yet further processing to decide on the sequence of symbols that constitutes the entire message. It would in principle be better to take all the raw measurements and then make optimal decisions about the entire sequence of symbols that was transmitted, but this would be a hugely more complex task. In practice, therefore, the task is commonly broken down into three stages, as here, with locally optimal decisions made at the single-pulse level to decode sequences of "ones" and "zeros", then further decisions made to decode at the symbol level, and still further decisions made at the symbol sequence level. In this section we illustrate the second of these decoding stages.

For concreteness, we center our discussion on the system in Figure 13.7. Suppose the transmitter randomly selects for transmission one of four possible symbols, which we label A, B, C and D. The probabilities with which these are selected will be denoted by $P(A)$, $P(B)$, $P(C)$ and $P(D)$ respectively. Whatever symbol the transmitter selects is now coded appropriately for transmission over the binary channel. The coding adds some redundancy to provide a basis for error correction at the receiver, in order to combat errors introduced by channel noise that may corrupt the individual bits. The resulting signal is then sent to the receiver. After

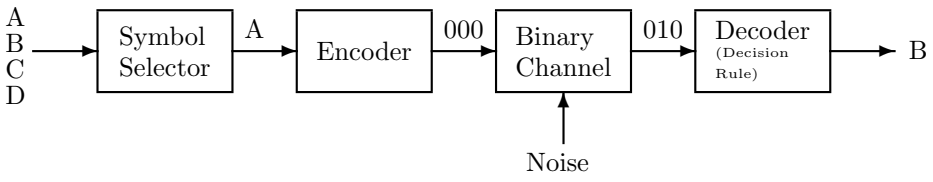

FIGURE 13.7 Communication over a binary channel.

the receiver decodes the received pulses, attempting to correct for channel noise in the process, it has to arrive at a decision as to which symbol was transmitted.

A natural criterion for measuring the performance of the receiver, with whatever decision process or decision rule it applies, is again the probability of error, P_e. It is natural, in a communications setting, to want minimum probability of error, and this is the criterion we adopt.

In the development below, rather than simply invoking the MAP rule we derived earlier, we repeat in this higher-level setting the line of reasoning that led to the MAP rule. We do this partly because there are some differences from what we considered earlier: we now have multiple hypotheses (four in our example), not just a pair of hypotheses; and the measured quantity is a discrete random symbol (more exactly, the received and possibly noise corrupted binary code for a transmitted symbol), rather than a continuous random variable. However, it will be clear that the problem here is not fundamentally different or harder.

13.4.1 Optimal a priori Decision

Consider, first of all, what the minimum-probability-of-error decision rule would be for the receiver if the channel was down, i.e., if the receiver had to decide on the transmitted signal without the benefit of any received signal, using only on *a priori* information. If the receiver guesses that the transmitter selected the symbol A, then the receiver is correct if A was indeed the transmitted symbol, and the receiver has made an error if A was not the transmitted symbol. Hence the receiver's probability of error with this choice is $1-P(A)$. Similar reasoning applies for the other symbols. So the minimum-probability-of-error decision rule for the receiver is to decide in favor of whichever symbol has maximum probability. This seems quite obvious for this simple case, and the general case (i.e., with the channel functioning) is not really any harder. We turn now to this general case, where the receiver actually receives the result of sending the transmitted signal through the noisy channel.

13.4.2 The Transmission Model

Let us model the channel as a binary channel, which accepts 1's and 0's from the transmitter, and delivers 1's and 0's to the receiver. Suppose that because of the noise in the channel there is a probability $p > 0$ that a transmitted 1 is received as a 0, and that a transmitted 0 is received as a 1. Because the probability is the same for both types of errors, this binary channel is called symmetric (we could treat the non-symmetric case as easily, apart from some increased notational burden). Implicit in our definition of this channel is the assumption that it is memoryless, i.e., its characteristics during any particular transmission slot are independent of what has been transmitted in other time slots. The channel is also assumed time-invariant, i.e., its characteristics do not vary with time.

Given such a channel, the transmitter needs to code the selected symbol into binary form. Suppose the transmitter uses 3 bits to code each symbol, as follows:

$$A : 000 \ , \quad B : 011 \ , \quad C : 101 \ , \quad D : 110 \ . \tag{13.41}$$

Because of the finite probability of bit-errors introduced by the channel, the received sequence for any of these transmissions could be any 3-bit binary number:

$$\begin{aligned} &R_0 = 000 \ , \quad R_1 = 001 \ , \quad R_2 = 010 \ , \quad R_3 = 011 \ , \\ &R_4 = 100 \ , \quad R_5 = 101 \ , \quad R_6 = 110 \ , \quad R_7 = 111 \ . \end{aligned} \tag{13.42}$$

The redundancy introduced by using 3 bits — rather than the 2 bits that would suffice to communicate our set of four symbols — is intended to provide some protection against channel noise. Notice that with our particular 3-bits/symbol code, a single bit-error would be recognized at the receiver as an error, because it would result in an invalid codeword. It takes two bit-errors (which are rarer than single bit-errors) to convert any valid codeword into another valid one, and thereby elude recognition of the error by the receiver.

There are now various probabilities that it might potentially be of interest to evaluate, such as:

- $P(R_1 \mid D)$, the probability that R_1 is received, given that D was sent;
- $P(D \mid R_1)$, the probability that D was sent, given that R_1 was received — this is the *a posteriori* probability of D, in contrast to $P(D)$, which is the *a priori* probability of D;
- $P(D, R_1)$, the probability that D is sent and R_1 is received;
- $P(R_1)$, the probability that R_1 is received.

The sample space of our probabilistic experiment can be described by Table 13.1, which contains an entry corresponding to every possible combination of transmitted symbol and received sequence. In the jth row of column A, we enter the probability $P(A, R_j)$ that A was transmitted and R_j received, and similarly for

columns B, C, and D. The simplest way to actually compute this probability is by recognizing that $P(A, R_j) = P(R_j|A)P(A)$; the characterization of the channel permits computation of $P(R_j|A)$, while the characterization of the information source at the transmitter yields the prior probability $P(A)$. Note that we can also write $P(A, R_j) = P(A|R_j)P(R_j)$. Examples of these three ways of writing the probabilities of the outcomes of our experiment are shown in the table.

13.4.3 Optimal a posteriori Decision

We now want to design the decision rule for the receiver, i.e., the rule by which it decides or hypothesizes what symbol was transmitted, after the reception of a particular sequence. We would like to do this in such a way that the probability of error, P_e, is minimized.

Since a decision rule in our example selects one of the four possible symbols (or hypotheses), namely A, B, C, or D, for each possible R_j, it can be represented in Table 13.1 by selecting one (and only one) entry in each row; we shall mark the selected entry by a box. For instance, a particular decision rule may declare D to be the transmitted signal whenever it receives R_4; this is indicated on the table by putting a box around the entry in row R_4, column D, as shown. Each possible decision rule is therefore associated with a table of the preceding form, with precisely one entry boxed in each row.

Now, for a given decision rule, the probability of being correct is the sum of the probabilities in all the boxed entries, because this sum gives the total probability that the decision rule declares in favor of the same symbol that was transmitted. The probability of error, P_e, is therefore 1 minus the probability of being correct.

It follows that to specify the decision rule for minimum probability of error or maximum probability of being correct, we must pick in each row the box that has the maximum entry. (If more than one entry has the maximum value, we are free to pick one of these arbitrarily — P_e is not affected by which of these we pick.) For row R_j in Table 13.1, we should pick for the optimum decision rule the symbol for which we maximize

$$\begin{aligned} P(\text{symbol}, R_j) &= P(R_j|\,\text{symbol})P(\text{symbol}) \\ &= P(\text{symbol}\,|R_j)P(R_j)\ . \end{aligned} \tag{13.43}$$

Table 13.2 displays some examples of the required computation in a particular numerical case. The computation in this example is carried out according to the prescription on the right side in the first of the above pair of equations. As noted earlier, this is generally the form that yields the most direct computation in practice, because the characterization of the channel usually permits direct computation of $P(R_j|\,\text{symbol})$, while the characterization of the information source at the transmitter yields the prior probabilities $P(\text{symbol})$.

The right side of the second equation in (13.43) permits a nice, intuitive interpretation of what the optimum decision rule does. Since our comparison is being done across the row, for a given R_j the term $P(R_j)$ in the second equation stays the

	$A: 000$	$B: 011$	$C: 101$	$D: 110$
$R_0 = 000$	$P(A, R_0)$	$P(B, R_0)$ $= P(R_0\|B)P(B)$ $= p^2(1-p)P(B)$	$P(C, R_0)$ $= P(C\|R_0)P(R_0)$	$P(D, R_0)$
$R_1 = 001$				
$R_2 = 010$				
$R_3 = 011$				
$R_4 = 100$	$P(A, R_4)$	$P(B, R_4)$	$P(C, R_4)$	$\boxed{P(D, R_4)}$
$R_5 = 101$				
$R_6 = 110$				
$R_7 = 111$				

TABLE 13.1 Each entry corresponds to a transmitted symbol and a received sequence.

same, so actually all that we need to compare are the *a posteriori* probabilities, $P(\text{symbol}\,|R_j)$, i.e. the probabilities of the various symbols, given the data. The optimum decision rule therefore picks the symbol with the maximum *a posteriori* probability. This is again the MAP decision rule that we derived previously in the binary hypothesis case. To summarize the important result we have arrived at here, and which we shall encounter again in more elaborate hypothesis testing contexts:

For minimum error probability P_e, decide in favor of the choice that has maximum *a posteriori* probability, i.e., the choice whose probability, conditioned on the available data, is maximum.

Note that the only difference from the minimum-P_e *a priori* decision rule we arrived at earlier, for the case where the channel was down, is the computation now has to involve conditional or *a posteriori* probabilities — conditioned on the received information — rather than the *a priori* probabilities. The receiver still decides in favor of the most probable choice, but now incorporating (i.e., conditioning on) the received information.

	000 A	011 B	101 C	110 D	Decision
R_0 000					
R_1 001					
R_2 010	$\left(\frac{3}{4}\right)^2 \frac{1}{4}\frac{1}{2}$	$\left(\frac{3}{4}\right)^2 \frac{1}{4}\frac{1}{4}$	$\left(\frac{1}{4}\right)^3 \frac{1}{8}$	$\left(\frac{3}{4}\right)^2 \frac{1}{4}\frac{1}{8}$	'A'
R_3 011					
R_4 100					
R_5 101					
R_6 110	$\left(\frac{1}{4}\right)^2 \frac{3}{4}\frac{1}{2}$	$\left(\frac{1}{4}\right)^2 \frac{3}{4}\frac{1}{4}$	$\left(\frac{1}{4}\right)^2 \frac{3}{4}\frac{1}{8}$	$\left(\frac{3}{4}\right)^3 \frac{1}{8}$	'D'
R_7 111					

TABLE 13.2 Designing the optimal decision rule, with $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{4}$, $P(C) = \frac{1}{8}$, $P(D) = \frac{1}{8}$, $p = \frac{1}{4}$. The MAP rule chooses the symbol that maximizes the *a posteriori* probability, $P(\text{symbol} \mid \text{data})$.

CHAPTER 14

Signal Detection

14.1 SIGNAL DETECTION AS HYPOTHESIS TESTING

In Chapter 13 we considered hypothesis testing in the context of random variables. The detector resulting in the minimum probability of error corresponds to the MAP test as developed in section 13.2.1 or equivalently the likelihood ratio test in section 13.2.3.

In this chapter we extend those results to a class of detection problems that are central in radar, sonar and communications, involving measurements of signals over time. The generic signal detection problem that we consider corresponds to receiving a signal $r(t)$ over a noisy channel. $r(t)$ either contains a known deterministic pulse $s(t)$ or it does not contain the pulse. Thus our two hypotheses are

$$\begin{aligned} H_1 &: r(t) = s(t) + w(t) \\ H_0 &: r(t) = w(t), \end{aligned} \tag{14.1}$$

where $w(t)$ is a wide-sense stationary random process. One example of a scenario in which this problem arises is in binary communication using pulse amplitude modulation. In that context the presence or absence of the pulse $s(t)$ represents the transmission of a "one" or a "zero". As another example, radar and sonar systems are based on transmitting a pulse and detecting the presence or absence of an echo.

In our treatment in this chapter we first consider the case in which the noise is white and carry out the formulation and analysis in discrete-time which avoids some of the subtler issues associated with continuous-time white noise. We also initially treat the case in which the noise is Gaussian. In Section 14.3.4 we extend the discussion to discrete-time Gaussian colored noise. In Section 14.3.2 we discuss the implications when the noise is not Gaussian and in Section 14.3.3 we discuss how the results generalize to the continuous-time case.

14.2 OPTIMAL DETECTION IN WHITE GAUSSIAN NOISE

In the signal detection task outlined above, our hypothesis test is no longer based on the measurement of a single (scalar) random variable R, but instead involves a collection of L (scalar) random variables $R_1, R_2, \ldots, R_L$.

Specifically, we receive the (finite-length) DT signal $r[n]$, $n = 1, 2, \cdots, L$, regarded as the realization of a random process. More simply, the signal $r[n]$ is modeled as

the values taken by a set of random variables $R[n]$. Let H_0 denote the hypothesis that the random waveform is only white Gaussian noise, i.e.

$$H_0: \quad R[n] = W[n] \tag{14.2}$$

where the $W[n]$ for $n = 1, 2, \cdots, L$ are independent, zero-mean, Gaussian random variables, with variance σ^2. Similarly, let H_1 denote the hypothesis that the waveform $R[n]$ is the sum of white Gaussian noise $W[n]$ and a known, deterministic signal $s[n]$, i.e.

$$H_1: \quad R[n] = s[n] + W[n] \tag{14.3}$$

where the $W[n]$ are again distributed as above. Our task is to decide in favor of H_0 or H_1 on the basis of the measurements $r[n]$.

The nature and derivation of the solutions to such decision problems are similar to those in Chapter 13, except that we now use posterior probabilities conditioned on the entire collection of measurements, i.e. $P(H_i|r[1], r[2], \cdots, r[L])$ rather than $P(H_i|r)$. Similarly, we use compound (or joint) PDF's, such as $f(r[1], r[2], \cdots, r[L]|H_i)$ instead of $f(r|H_i)$. The associated decision regions D_i are now regions in an L-dimensional space, rather than segments of the real line.

For detection with minimum probability of error, we again use the MAP rule or equivalently compare the values of

$$f(r[1], r[2], \ldots, r[L] \,|\, H_i)\, P(H_i) \tag{14.4}$$

for $i = 0, 1$, and decide in favor of whichever hypothesis yields the maximum value of this expression, i.e. the form of equation (13.7) for the case of multiple measurements is

$$\boxed{f(r[1], r[2], \ldots, r[L] \,|\, H_1)\, P(H_1) \underset{\text{`}H_0\text{'}}{\overset{\text{`}H_1\text{'}}{\gtrless}} f(r[1], r[2], \ldots, r[L] \,|\, H_0)\, P(H_0)} \tag{14.5}$$

which also can easily be put into the form of equation (13.18) corresponding to the likelihood ratio test.

With $W[n]$ white and Gaussian, the conditional densities in (14.5) are easy to evaluate, and take the form

$$\begin{aligned} f(r[1], r[2], \ldots, r[L] \,|\, H_0) &= \frac{1}{(2\pi\sigma^2)^{(L/2)}} \prod_{n=1}^{L} \exp - \left\{ \frac{(r[n])^2}{2\sigma^2} \right\} \\ &= \frac{1}{(2\pi\sigma^2)^{(L/2)}} \exp - \left\{ \sum_{n=1}^{L} \frac{(r[n])^2}{2\sigma^2} \right\} \end{aligned} \tag{14.6}$$

and

$$f(r[1], r[2], \ldots, r[L] \,|\, H_1) = \frac{1}{(2\pi\sigma^2)^{(L/2)}} \prod_{n=1}^{L} \exp - \left\{ \frac{(r[n] - s[n])^2}{2\sigma^2} \right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{(L/2)}} \exp - \left\{ \sum_{n=1}^{L} \frac{(r[n] - s[n])^2}{2\sigma^2} \right\} \tag{14.7}$$

The inequality in equation (14.5) (or any inequality in general) will, of course still hold if a nonlinear, strictly increasing function is applied to both sides. Because of the form of equations (14.6) and (14.7) it is particularly convenient to replace equation (14.5) by applying the natural logarithm to both sides of the inequality. The resulting inequality, in the case of (14.6) and (14.7), is:

$$g = \sum_{n=1}^{L} r[n]s[n] \quad \begin{matrix} \text{“}H_1\text{”} \\ > \\ < \\ \text{“}H_0\text{”} \end{matrix} \quad \sigma^2 \ln \left(\frac{P(H_0)}{P(H_1)} \right) + \frac{1}{2} \sum_{n=1}^{L} s^2[n] \tag{14.8}$$

The sum on the left-hand side of Eq. (14.8) is referred to as the deterministic correlation between $r[n]$ and $s[n]$, which we denote as g. The second sum on the right-hand side is the energy in the deterministic signal $s[n]$ which we denote by $\mathcal{E}$. For convenience we denote the threshold represented by the entire right hand side of (14.8) as γ, i.e., equation (14.8) becomes

$$g \quad \begin{matrix} \text{“}H_1\text{”} \\ > \\ < \\ \text{“}H_0\text{”} \end{matrix} \quad \gamma \tag{14.9a}$$

$$\text{where } \gamma = \sigma^2 \ln(\frac{P(H_0)}{P(H_1)}) + \frac{\mathcal{E}}{2} \tag{14.9b}$$

If the Neyman-Pearson formulation is used, then the optimal decision rule is still of the form of equation (14.8), except that the right hand side of the inequality is determined by the specified bound on P_{FA}.

If hypothesis H_0 is true, i.e. if the signal $s[n]$ is absent, then $r[n]$ on the left hand side of equation (14.8) will be Gaussian white noise only, i.e. g will be the random variable

$$G = \sum_{n=1}^{L} W[n]s[n] \tag{14.10}$$

Since $W[n]$ at each value of n is Gaussian, with zero mean and variance σ^2, and since a weighted, linear combination of Gaussian random variables is also Gaussian, the random variable G is Gaussian with mean zero and variance $\sigma^2 \sum_{n=1}^{L} s^2[n] = \sigma^2\mathcal{E}$.

When the signal is actually present, i.e., when H_1 holds, the random variable is the realisation of a Gaussian random variable with mean $\mathcal{E}$ and still with variance $\mathcal{E}\sigma^2$ or standard deviation $\sigma\sqrt{\mathcal{E}}$. The optimal test in (14.8) is therefore described by Figure 14.1 which is of course similar to that in Figure 13.5 :

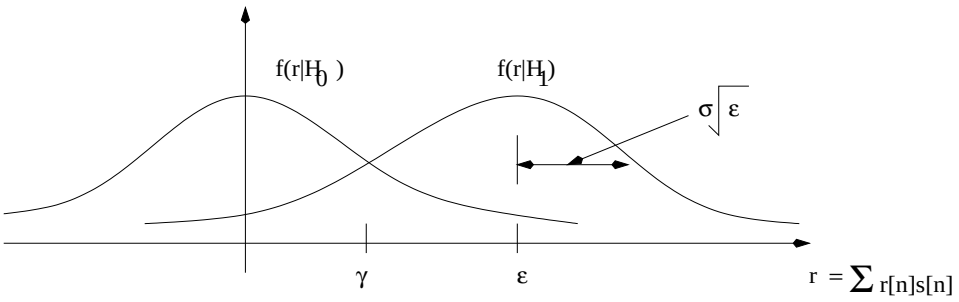

FIGURE 14.1 Optimal test for two hypotheses with equal variances and different means.

Using the facts summarized in this figure, and given a detection threshold γ on the correlation (e.g. with γ picked equal to the right side of (14.8), or in some other way), we can compute P_{FA}, P_D, P_e, and other probabilities of interest.

Figure 14.1 makes evident that the performance of the detection strategy is determined entirely by the ratio $\mathcal{E}/(\sigma\sqrt{\mathcal{E}})$, or equivalently by the signal-to-noise ratio $\mathcal{E}/\sigma^2$, i.e. the ratio of the signal energy $\mathcal{E}$ to the noise variance σ^2.

14.2.1 Matched Filtering

Since the correlation sum in (14.8) constitutes a linear operation on the measured signal, we can consider computing the sum through the use of an LTI filter and the output sampled at an appropriate time to form the correlation sum g. Specifically, with $h[n]$ as the impulse response and $r[n]$ as the input, the output will be the convolution sum

$$\sum_{k=-\infty}^{\infty} r[k]h[n-k] \tag{14.11}$$

For $r[n] = 0$ except for $1 \leq n \leq L$ and with $h[n]$ chosen as $s[-n]$, the filter output at $n = 0$ is $\sum_{k=1}^{L} r[k]s[k] = g$ as required. In other words, we choose the filter impulse response to be a time-reversed version of the target signal for $n = 1, 2, \ldots, L$, with $h[n] = 0$ elsewhere. This filter is said to be the matched filter for the target signal. The structure of the optimum detector for a finite-length signal in white Gaussian noise is therefore as shown below:

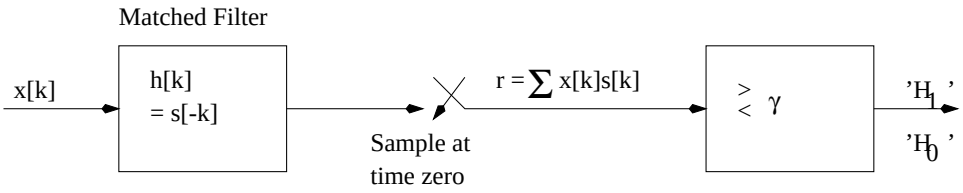

FIGURE 14.2 Optimum detector

14.2.2 Signal Classification

We can easily extend the previous two-hypothesis problem to the multiple hypothesis case, where H_i, $i = 0, 1, \cdots, M-1$ denotes the hypothesis that the signal $R[n]$, $n = 1, 2, \cdots, L$, is a noise-corrupted version of the ith deterministic signal $s_i[n]$, selected from a possible set of M deterministic signals:

$$H_i: \quad R[n] = s_i[n] + W[n] \tag{14.12}$$

with the $W[n]$ denoting independent, zero-mean, Gaussian random variables with variance σ^2. This scenario arises, for example, in radar signature analysis. Different aircraft reflect a radar pulse differently, typically with a distinct signature that can be used to identify not only its presence, but the type of aircraft. In this case, each of the signals $s_i[n]$ and correspondingly each hypothesis H_i would correspond to the presence of a particular type of aircraft. Thus, our task is to decide in favor of one of the hypotheses, given a set of measurements $r[n]$ of $R[n]$. For minimum error probability, the required test involves comparison of the quantities

$$\left(\sum_{n=1}^{L} r[n]s_i[n]\right) - \frac{\mathcal{E}_i}{2} + \sigma^2 \ln P(H_i) \tag{14.13}$$

where $\mathcal{E}_i$ denotes the energy of the ith signal. The largest of the expressions in (14.13), for $i = 0, 1, \cdots, M-1$, determines which hypothesis is selected. If the signals have equal energies and equal prior probabilities, then the above comparison reduces to deciding in favor of the signal with the highest deterministic correlation

$$\sum_{n=1}^{L} r[n]s_i[n] \; . \tag{14.14}$$

14.3 A GENERAL DETECTOR STRUCTURE

The matched filter developed in Section 14.2 extends to the case where we have an infinite number of measurements rather than just L measurements. As we will see in Section 14.3.4, it also extends to the case of colored noise. We shall, for simplicity, treat these extensions by assuming the general detector structure, shown in Figure

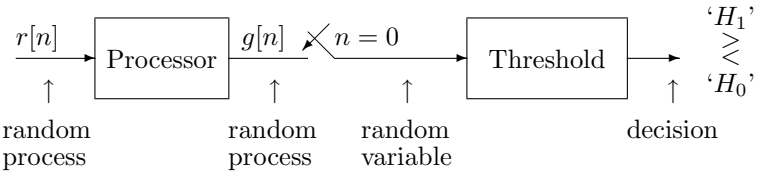

FIGURE 14.3 A general detector structure.

11.7, and determine an optimum choice of processor and of detection threshold for each scenario.

We are assuming that the transmitter and receiver are synchronized, so that we test $g[n]$ at a known (fixed) time, which we choose here as $n = 0$. The choice of 0 as the sampling instant is for convenience; any other instant may be picked, with a corresponding time-shift in the operation of the processor. Although the processor could in general be nonlinear, we shall assume the processing will be done with an LTI filter. Thus the system to be considered is shown in Figure 14.4; a corresponding system can be considered for continuous time.

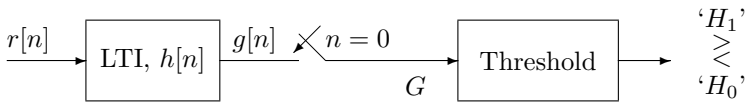

FIGURE 14.4 Detector structure of Figure 14.3 with the processor as an LTI system.

It can be shown formally, but is also intuitively reasonable, that scaling $h[n]$ by a constant gain will not affect the overall performance of the detector if the threshold is correspondingly adjusted since a constant overall gain scales the signal and noise identically.

For convenience, we normalize the gain of the LTI system so as to have

$$\sum_{n=-\infty}^{+\infty} h^2[n] = 1 \, . \tag{14.15}$$

If $r[n]$ is a Gaussian random process, then so is $g[n]$, because it is obtained by linear processing of $r[n]$, and therefore G is a Gaussian random variable in this case.

14.3.1 Pulse Detection in White Noise

To suggest the approach we consider a very simple choice of LTI processor, namely with $h[n] = \delta[n]$, so

$$\begin{aligned} H_1 &: G = g[0] = s[0] + w[0] \\ H_0 &: G = g[0] = w[0] \, . \end{aligned} \tag{14.16}$$

Also for convenience we assume that $s[0]$ is positive.

Thus, under each hypothesis, $g[0]$ is Gaussian:

$$H_1 : f_{G|H}(g|H_1) = N(s[0], \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(g-s[0])^2}{2\sigma^2}\right]$$
$$H_0 : f_{G|H}(g|H_0) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{g^2}{2\sigma^2}\right] . \tag{14.17}$$

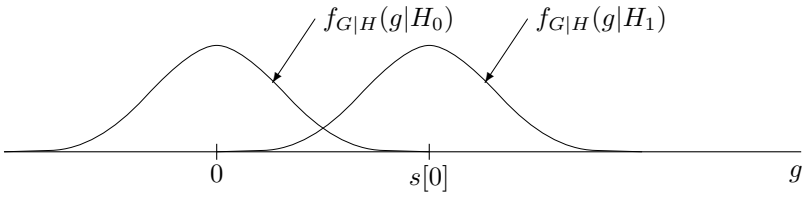

FIGURE 14.5 PDF's for the two hypotheses in Eq. (14.16).

This is just the binary hypothesis testing problem on the random variable G treated in Section 13.2 and correspondingly the MAP rule for detection with minimum probability of error is given by

$$P(H_1 \mid G = g) \underset{\text{`}H_0\text{'}}{\overset{\text{`}H_1\text{'}}{\gtrless}} P(H_0 \mid G = g) ,$$

or, equivalently, the likelihood ratio test:

$$\frac{f_{G|H}(g \mid H_1)}{f_{G|H}(g \mid H_0)} \underset{\text{`}H_0\text{'},}{\overset{\text{`}H_1\text{'}}{\gtrless}} \frac{P(H_0)}{P(H_1)} = \eta . \tag{14.18}$$

Evaluating equation (14.18) using equation (14.17) leads to the relationship

$$\exp\left\{\left[-\frac{(g-s[0])^2}{2\sigma^2}\right] + \left[\frac{g^2}{2\sigma^2}\right]\right\} \underset{\text{`}H_0\text{'},}{\overset{\text{`}H_1\text{'}}{\gtrless}} \frac{P(H_0)}{P(H_1)} \tag{14.19}$$

and equivalently

$$\exp\left[\frac{gs[0]}{\sigma^2} - \frac{s^2[0]}{2\sigma^2}\right] \underset{\text{`}H_0\text{'},}{\overset{\text{`}H_1\text{'}}{\gtrless}} \frac{P(H_0)}{P(H_1)} \tag{14.20}$$

or, taking the natural logarithm of both sides of the likelihood ratio test as we did in Section 14.2, equation (14.20) is replaced by

$$g \underset{\text{`}H_0\text{'},}{\overset{\text{`}H_1\text{'}}{\gtrless}} \frac{s[0]}{2} + \frac{\sigma^2}{s[0]} \ln \frac{P(H_0)}{P(H_1)} \tag{14.21}$$

We may not know the *a priori* probabilities $\mathrm{P}(H_0)$ and $\mathrm{P}(H_1)$ or, for other reasons, may want to modify the threshold, but still using a threshold test on the likelihood ratio, or a threshold test of the form

$$g \underset{\text{`}H_0\text{'}}{\overset{\text{`}H_1\text{'}}{\gtrless}} \lambda \,. \tag{14.22}$$

Sweeping the threshholds over all possible values leads to the receiver operating characteristics as discussed in Section 13.2.5.

We next consider the more general case in which $h[n]$ is not the identity system. Then, under the two hypothesis we have:

$$\begin{array}{lrcl} H_1: & g[n] & = & s[n]*h[n]+w[n]*h[n] \\ H_0: & g[n] & = & w[n]*h[n]\,, \end{array} \tag{14.23}$$

The term $w[n]*h[n]$ still represents noise but is no longer white, i.e. its spectral shape is changed by the filter $h[n]$. Denoting $w[n]*h[n]$ as $v[n]$, the autocorrelation function of $v[n]$ is

$$R_{vv}[m] = R_{ww}[m] * \overline{R}_{hh}[m] \tag{14.24}$$

and in particular the mean $v[n]$ is zero and its variance is

$$\mathrm{var}\{v[n]\} = \sigma^2 \sum_{n=-\infty}^{\infty} h^2[n]. \tag{14.25}$$

Because of the normalization in equation (14.15) the variance of $v[n]$ is the same as that of the white noise, i.e. $\mathrm{var}\{v[n]\} = \sigma^2$. Furthermore, since $w[n]$ is Gaussian so is $v[n]$. Consequently the value $g[0]$ is again a Gaussian random variable with variance σ^2. The mean of $g[0]$ under the two hypotheses is now:

$$\begin{array}{lrcl} H_1: & E\{g[n]\} & = & \displaystyle\sum_{n=-\infty}^{\infty} h[n]s[-n] \triangleq \mu \\ H_0: & E\{g[n]\} & = & 0, \end{array} \tag{14.26}$$

Therefore equation (14.17) is replaced by

$$\begin{aligned} &H_1 : f_{G|H}(g|H_1) = N(\mu, \sigma^2) \\ &H_0 : f_{G|H}(g|H_0) = N(0, \sigma^2). \end{aligned} \tag{14.27}$$

The probability density functions representing the two hypothesis are shown in Figure 14.6 below. On this figure we have also indicated the threshold γ of equation (14.27) above which we would declare H_1 to be true and below which we would declare H_0 to be true. Also indicated by the shaded areas are the areas under the PDF's that would correspond to P_{FA} and P_D.

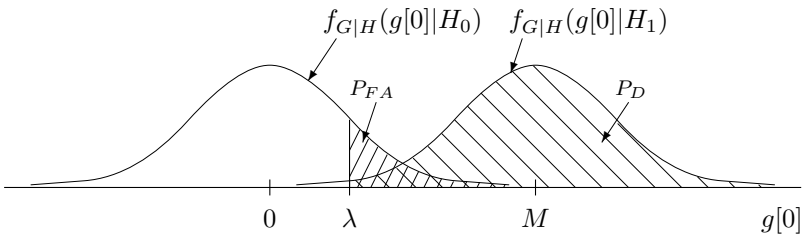

FIGURE 14.6 Indication of the areas representing P_{FA} and P_D.

The value of P_{FA} is fixed by the shape of $f_{G|H}(g[0]|H_0)$ and the value of the threshold γ. Since $f_{G|H}(g[0]|H_0)$ is not dependent on $h[n]$, the choice of $h[n]$ will not affect P_{FA}. The variance of $f_G(g[0]|H_1)$ is also not influenced by the choice of $h[n]$ but its mean μ is. In particular, as we see from Figure 14.6, the value of P_D is given by

$$P_D = \int_{\gamma}^{\infty} f_G(g[0]|H_1)dg \tag{14.28}$$

which increases as μ increases. Consequently, to minimize $P(\text{error})$, or alternatively to maximize P_D for a given P_{FA}, we want to maximize the value of μ. To determine the choice of $h[n]$ to maximize μ we use the Schwarz inequality:

$$\left|\sum h[n]s[-n]\right|^2 \leq \sum h^2[n] \sum s^2[-n] \tag{14.29}$$

with equality if and only if $h[n] = cs[-n]$ for some constant c. Since we normalized the energy in $h[n]$, the optimum filter is $h[n] = (\frac{1}{\sqrt{\mathcal{E}}})s[-n]$, which is again the matched filter. (This is as expected, since the optimum detector for a known finite-length pulse in white Gaussian noise has already been shown in Section 14.2.1 to have the form we assumed here, with the impulse response of the LTI filter being matched to the signal.) The filter output $g[n]$ due to the pulse is then $\frac{1}{\sqrt{\mathcal{E}}}\overline{R}_{ss}[n]$ and the output due to the noise is the colored noise $v[n]$ with variance σ^2. Since $g[0]$ is a random variable with mean $\frac{1}{\sqrt{\mathcal{E}}}\sum_{n=-\infty}^{\infty} s^2[n]$ and variance σ^2, only the energy in the pulse and not its specific shape, affects the performance of the detector.

14.3.2 Maximizing SNR

If $w[n]$ is white but not Gaussian, then $g[0]$ is not Gaussian. However, $g[0]$ is still distributed the same under each hypothesis, except that its mean under H_0 is 0 while the mean under H_1 is μ as given in equation (14.26). The matched filter in this case still maximizes the output signal-to-noise ratio (SNR) in the specified structure (namely, LTI filtering followed by sampling), where the SNR is defined as $E\{g[0]|H_1\}^2/\sigma^2$. The square root of the SNR is the relative separation between the means of the two distributions, measured in standard deviations. In some intuitive sense, therefore, maximizing the SNR tries to separate the two distributions as well

as possible. However, this does not in general necessarily correspond to minimizing the probability of error.

14.3.3 Continuous-Time Matched Filters

All of the matched filter results developed in this section carry over in a direct way to continuous-time. In Figure 14.7 we show the continuous-time counterpart to Figure 14.4 As before, we normalize the gain of $h(t)$ so that

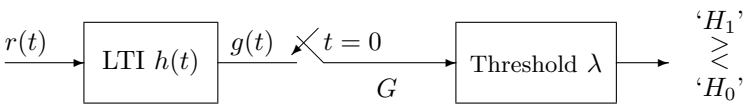

FIGURE 14.7 Continuous-time matched filtering.

$$\int_{-\infty}^{\infty} h^2(t)dt = 1 \tag{14.30}$$

with $r(t)$ a Gaussian random process, $g(t)$ is also Gaussian and G is a Gaussian random variable. Under the two hypotheses the PDF of G is then given by

$$\begin{aligned} H_1 &: f_{G|H}(g|H_1) = N(\mu, \sigma_G^2) \\ H_0 &: f_{G|H}(g|H_0) = N(0, \sigma_G^2) \ , \end{aligned} \tag{14.31}$$

where

$$\sigma_G^2 = N_0 \int_{-\infty}^{\infty} h^2(t)dt = N_0 \tag{14.32}$$

and

$$\mu = \int_{-\infty}^{\infty} h(t)s(-t)dt \tag{14.33}$$

Consequently, as in the discrete-time case, the probability of error is minimized by choosing $h(t)$ to separate the two PDF's in equation (14.31) as much as possible. With the continuous-time version of the Cauchy-Schwarz inequality applied to equation (14.33) we then conclude that the optimum choice for $h(t)$ is proportional to $s(-t)$, i.e. again the matched filter

---

EXAMPLE 14.1 PAM with Matched Filter

Figure 14.8(a) shows an example of a typical noise-free binary PAM signal as represented by Eq. (13.1). The pulse $p(t)$ is a rectangular pulse of length 50 sec. The binary sequence $a[n]$ over the time interval shown is indicated above the waveform. In the absence of noise, the optimal threshold detector of the form of Figure 14.4

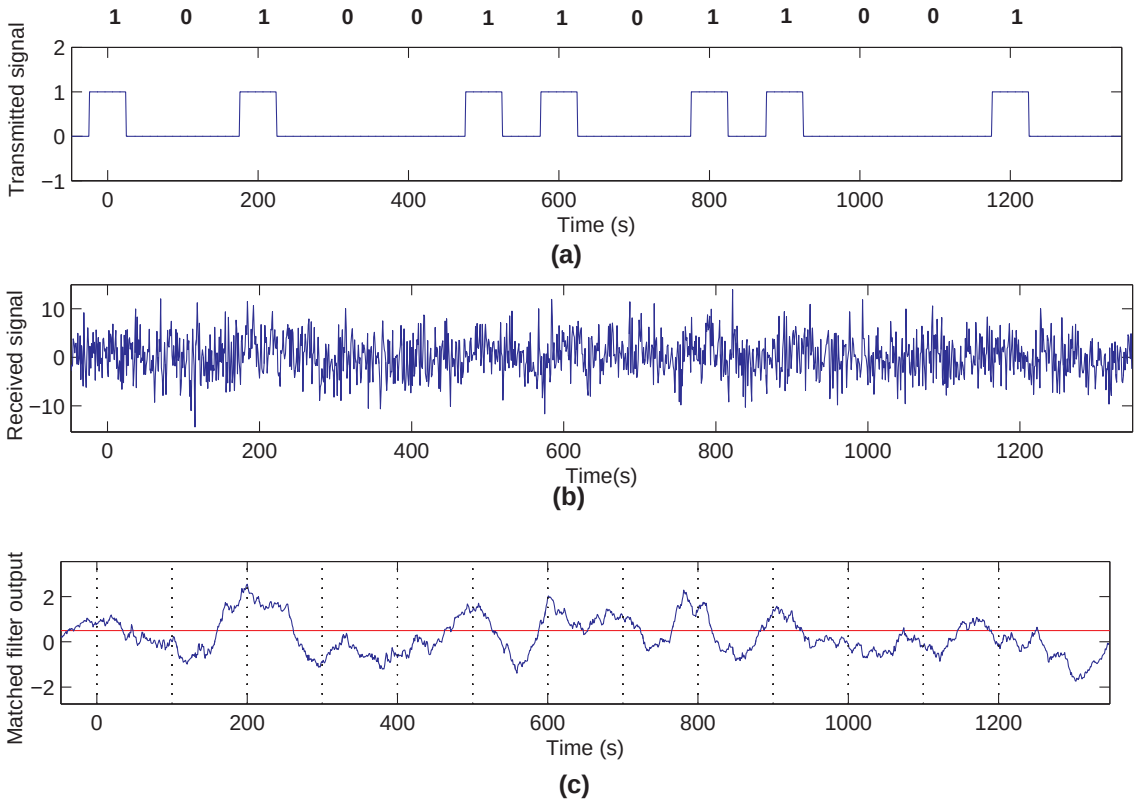

FIGURE 14.8 Binary detection with on/off signaling

would simply test at integer multiples of T whether the received signal is positive or zero. Clearly the probability of error in this noise-free case would be zero.

In Figure 14.8(b) we show the same PAM signal but with wideband Gaussian noise added. If $h(t)$ is the identity system and the threshold of the detector is chosen according to Eq. (14.18) with $P(H_0) = P(H_1)$ i.e. using the likelihood ratio test but without the matched filter, the decoded binary sequence is 0100111111011 which has 6 bit errors. Figure 14.8(c) shows the output of the matched filter, i.e. with $h(t) = s(-t)$. The detector threshold is again chosen based on the likelihood ratio test. The resulting decoded binary sequence is 1010011111000 which has 2 bit errors

In Figure 14.9 we show the corresponding results when antipodal rather than on-off signaling is used. Figure 14.9(a) depicts the transmitted waveform with the same binary sequence as was used in Figure 14.8, and Figure 14.9(b) the received signal including additive noise. If $h(t) = \delta(t)$ and $P(H_0) = P(H_1)$, then the choice of threshold for the likelihood ratio test is zero. The decoded binary sequence is

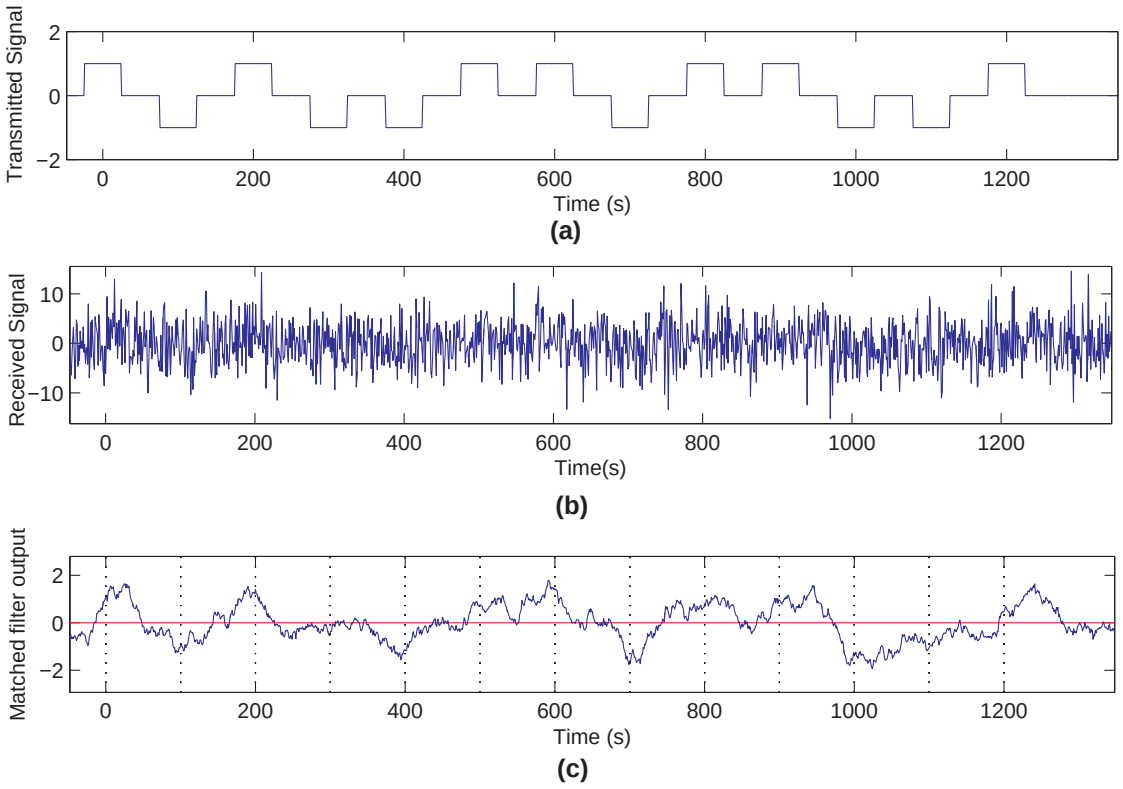

FIGURE 14.9 Binary Detection with antipodal signaling

0001001011001, resulting in 4 bit errors. With $h(t)$ chosen as the matched filter the signal before the threshold detector is that shown in Figure 14.9(c). The resulting decoded binary sequence is 1010011011001 with no bit errors. In Table 14.1 we summarize the results for this specific example based on a simulation with a binary sequence of length 10^4.

	No matched filter	W/ matched Filter
On/Off Signaling	0.4808	0.3752
Antipodal Signaling	0.4620	0.2457

TABLE 14.1 Bit error rate for a PAM signal illustrating effect of matched filter for two different signaling schemes.

14.3.4 Pulse Detection in Colored Noise

In Sections 14.2 and 14.3 the optimal detector was developed under the assumption that the noise is white. When the noise is colored , i.e. when its spectral density is not flat, the results are easily modified. We again assume a detector of the form of Figure 14.4. The two hypotheses are now:

$$\begin{aligned} H_1 &: r[n] = s[n] + v[n], \\ H_0 &: r[n] = v[n] \ , \end{aligned} \tag{14.34}$$

where $v[n]$ is again a zero-mean Gaussian process but in general, not white. The autocorrelation function of $v[n]$ is denoted by $R_{vv}[m]$ and the power spectral density by $S_{vv}(e^{j\Omega})$. The basic approach is to transform the problem to that dealt with in the previous section by first processing $r[n]$ with a whitening filter as was discussed in Section 10.2.3 , which is always possible as long as $S_{vv}(e^{j\Omega})$ is strictly positive, i.e. it is not zero at any value of Ω. This first stage of filtering is depicted in Figure 14.10.

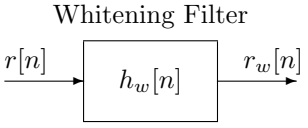

FIGURE 14.10 First stage of filtering

The impulse response $h_w[n]$ is chosen so that its output due to the input noise $v[n]$ is white, with variance σ^2 and, of course, will also be Gaussian. With this pre-processing the signal $r_w[n]$ now has the form assumed in Section 14.3.4 with the white noise $w[n]$ corresponding to $v[n] * h_w[n]$ and the pulse $s[n]$ replaced by $p[n] = s[n] * h_w[n]$. The detector structure now takes the form shown in Figure 14.11 where $h[n]$ is again the matched filter, but in this case matched to the pulse $p[n]$, i.e. $h_m[n]$ is proportional to $p[-n]$.

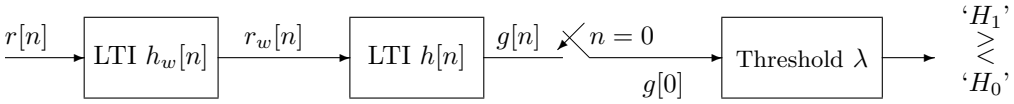

FIGURE 14.11 Detector structure with colored noise.

Assuming that $h_w[n]$ is invertible (i.e. its $\mathcal{Z}$-transform has no zeros on the unit circle) there is no loss of generality in having first applied a whitening filter. To see this concretely denote the combined LTI filter from $r[n]$ to $g[n]$ as $h_c[n]$ and assume that if whitening had not first been applied, the optimum choice for the filter from $r[n]$ to $g[n]$ is $h_{opt}[n]$. Since

$$h_c[n] = h_w[n] * h_m[n] \tag{14.35}$$

where $h_m[n]$ denotes the matched filter after whitening. If the performance with $h_{opt}[n]$ is better than with $h_c[n]$, this would imply that choosing $h_m[n]$ as $h_{opt}[n] * h_w^{inv}[n]$ would lead to better performance on the whitened signal. However, as seen in Section 14.3, $h_m[n] = p[-n]$ is the optimum choice after the whitening and consequently we conclude that

$$h_m[n] = p[-n] = h_{opt}[n] * h_w^{inv}[n] \tag{14.36}$$

or equivalently

$$h_{opt}[n] = h_w[n] * p[-n] \tag{14.37}$$

In the following example we illustrate the determination of the optimum detector in the case of colored noise.

EXAMPLE 14.2 Pulse Detection in Colored Noise

Consider a pulse $s[n]$ in colored noise $v[n]$, with

$$s[n] = \delta[n] \; . \tag{14.38}$$

and

$$\begin{aligned} R_{vv}[m] &= (\frac{1}{2})^{|m|}, \;\; \text{so} \;\; \sigma_v^2 = 1 \\ \text{then } S_{vv}(z) &= \frac{3/4}{(1-\frac{1}{2}z^{-1})(1-\frac{1}{2}z)} \; . \end{aligned} \tag{14.39}$$

The noise component $w[n]$ of desired output of the whitening filter has autocorrelation function $R_{ww}[m] = \sigma^2\delta[m]$ and consequently we require that

$$\begin{aligned} S_{vv}(z)H_w(z)H_w(1/z) &= \sigma^2 \\ \text{Thus } H_w(z)H_w(1/z) &= \frac{\sigma^2}{S_{vv}(z)} = \sigma^2\frac{4}{3}(1-\frac{1}{2}z^{-1})(1-\frac{1}{2}z) \; . \end{aligned} \tag{14.40}$$

We can of course choose σ arbitrarily (since it will only impact the overall gain). Choosing $\sigma^2 = 1$, either

$$\begin{aligned} H_w(z) &= (1-\frac{1}{2}z^{-1}), \;\; \text{or} \\ H_w(z) &= (1-\frac{1}{2}z) \end{aligned} \tag{14.41}$$

Note that the second of these choices is non-causal. There are also other possibile choices since we can cascade either choice with an all-pass $H_{ap}(z)$ such that $H_{ap}(z)H_{ap}(1/z) = 1$.

With the first choice for $H_w(z)$ from above, we have

$$\begin{aligned} H_w(z) &= (1 - \frac{1}{2}z^{-1}), \\ h_w[n] &= \delta[n] - \frac{1}{2}\delta[n-1], \\ \sigma^2 &= 3/4, \\ p[n] &= s[n] - \frac{1}{2}s[n-1], \text{ and} \\ h[n] &= Ap[-n] \text{ for any convenient choice of } A. \end{aligned} \tag{14.42}$$

In our discussion in Section 14.3 of the detection of a pulse in white noise, we observed that the energy in the pulse affects performance of the detector but not the specific pulse shape. This was a consequence of the fact that the filter is chosen to maximize the quantity $\frac{1}{\sqrt{\mathcal{E}}}R_{ss}[0]$ where $s[n]$ is the pulse to be detected. For the case of a pulse in colored noise, we correspondingly want to maximize the energy $\mathcal{E}_p$ in $p[n]$ where

$$p[n] = h_w[n] * s[n] \tag{14.43}$$

Expressed in the frequency domain,

$$P(e^{j\Omega}) = H_w(e^{j\Omega})S(e^{j\Omega}) \tag{14.44}$$

and from Parseval's relation

$$\mathcal{E}_p = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H_w(e^{j\Omega})|^2 |S(e^{j\Omega})|^2 d\Omega \tag{14.45a}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|S(e^{j\Omega})|^2}{S_{vv}(e^{j\Omega})} d\Omega \tag{14.45b}$$

Based only on Eq. (14.45b), $\mathcal{E}_p$ can be maximized by placing all of the energy of the transmitted signal $s[n]$ at the frequency at which $S_{vv}(e^{j\Omega})$ is minimum. However, in many situations the transmitted signal is constrained in other ways, such as peak amplitude and/or time duration. The task then is to choose $s[n]$ to maximize the integral in Eq. (14.45b) under these constraints. There is generally no closed-form solution to this optimization problem, but roughly speaking a good solution will distribute the signal energy so that it is more concentrated where the power $S_{vv}(e^{j\Omega})$ of the colored noise is less.

6.011 Introduction to Communication, Control, and Signal Processing
Spring 2010