(c) Under what conditions are the estimates from (a) and (b) approximately equal?

24. Suppose that, given $\Theta = \theta, Y_1, \ldots, Y_n$ are i.i.d. real observations with marginal densities

$$f_\theta(y) = \begin{cases} \theta^{-1}e^{-y/\theta}, & y \geq 0 \\ 0, & y < 0. \end{cases}$$

(a) Find the maximum-likelihood estimate of θ based on $Y_1, \ldots, Y_n$. Compute its mean and variance.

(b) Compute the Cramér-Rao lower bound for the variance of unbiased estimates of θ.

(c) Suppose Θ is uniformly distributed on (0, 1]. Find the MAP estimate of Θ

(d) For $n = 3$, find the MMSE estimate of Θ. Assume the same prior as in part (c).

(e) For $n = 2$, find the MMAE estimate of Θ. Assume the same prior as in part (c).

25. Suppose that, given $\Theta = \theta$, the real observation Y has pdf

$$p_\theta(y) = \begin{cases} \frac{6(y^2+\theta y)}{2+3\theta}, & 0 \leq y \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Suppose Θ is uniformly distributed on [0, 1]. Find the MMSE estimate and corresponding minimum Bayes risk.

(b) With Θ as in (a), find the MAP estimate and the MMAE estimate of Θ.

(c) Find the maximum-likelihood estimate of θ and compute its bias.

(d) Compute the Cramér-Rao lower bound on the variance of unbiased estimates of θ.

V

Elements of Signal Estimation

V.A Introduction

In Chapter IV we discussed methods for designing estimators for static parameters, that is, for parameters that are not changing with time. In many applications we are interested in the related problem of estimating dynamic or time-varying parameters. In the traditional terminology, a dynamic parameter is usually called a *signal*, so the latter problem is known as *signal estimation* or *tracking*.

Such problems arise in many applications. For example, one function of many radar systems is to track targets as they move through the radar's scanning area. This means that the radar must estimate the position of the target (and perhaps its velocity) at successive times. Since the targets of interest are usually moving and the position measurements are noisy, this is a signal estimation problem. Another application is that of analog communications, in which analog information (e.g., audio or video) is transmitted by modulating the amplitude, frequency, or phase of a sinusoidal carrier. The receiver's function in this situation is to determine the transmitted information with as high a fidelity as possible on the basis of a noisy observation of the received waveform. Again, since the transmitted information is time varying, this problem is one of signal estimation.

The dynamic nature of the parameter in signal estimation problems adds a new dimension to the statistical modeling of these problems. In particular, the dynamic properties of the signal (i.e., how fast and in what manner it can change) must be modeled at least statistically in order to obtain meaningful signal estimation procedures. Also, performance expectations for estimators of dynamic parameters should be different from those for static parameters. In particular, unlike the static case, we cannot expect an estimator of a signal to be perfect as the number of observations becomes infinite because of the time variation in the signal.

In this chapter we discuss the basic ideas behind some of the signal estimation techniques used most often in practice. In Section V.B we discuss *Kalman-Bucy filtering*, which provides a very useful algorithm for estimating signals that are generated by finite-dimensional linear dynamical models. In Section V.C the general problem of estimating signals as linear transformations of the observations is developed, and in Section V.D a particular case of linear estimation, *Wiener-Kolmogorov filtering*, which is a method of estimating signals whose statistics are stationary in time, is considered.

V.B Kalman-Bucy Filtering

Many time-varying physical phenomena of interest can be modeled as obeying equations of the type

$$\underline{X}_{n+1} = \underline{f}_n(\underline{X}_n, \underline{U}_n), \quad n = 0, 1, \ldots, \tag{V.B.1}$$

where $\underline{X}_0, \underline{X}_1, \ldots,$ is a sequence of vectors in $\mathbb{R}^m$ representing the phenomenon under study; $\underline{U}_0, \underline{U}_1, \ldots,$ is a sequence of vectors in $\mathbb{R}^s$ "acting" on $\{\underline{X}_n\}_{n=1}^{\infty}$; and where $\underline{f}_0, \underline{f}_1, \ldots,$ is a sequence of functions (or, in other words, a time-varying function), each mapping $\mathbb{R}^m \times \mathbb{R}^s$ to $\mathbb{R}^m$. Equation (V.B.1) is an example of a *dynamical system*, with $\underline{X}_n$ representing the *state* of the system at time n and with $\underline{U}_n$ representing the *input* to the system at time n [see, e.g., Desoer (1970)]. A dynamical system is a system having the property that for any fixed times l and k, $\underline{X}_l$ is determined completely from the state at time k (i.e., $\underline{X}_k$) and the inputs from times k up through $l-1$ (i.e., $\{\underline{U}_n\}_{n=k}^{l-1}$). Note that complete determination of $\{\underline{X}_n\}_{n=1}^{\infty}$ from (V.B.1) requires not only the specification of the input sequence but also the specification of the *initial condition* $\underline{X}_0$. If the input sequence or the initial condition is random, the states $\underline{X}_0, \underline{X}_1, \ldots,$ form a sequence of random vectors and (V.B.1) is referred to as *stochastic system.*

Equation (V.B.1) describes the evolution of the states of a system, so it is usually known as the *state equation* of the system. The system may also have associated with it an *output* sequence $\underline{Z}_0, \underline{Z}_1, \ldots,$ of vectors in $\mathbb{R}^k$, possibly different from the state sequence, and given by the *output equation*

$$\underline{Z}_n = \underline{h}_n(\underline{X}_n), \quad n = 0, 1, \ldots, \tag{V.B.2}$$

where $\underline{h}_n$ maps $\mathbb{R}^m$ to $\mathbb{R}^k$. Thus the overall system is a mapping from the initial condition $\underline{X}_0$ and input sequence $\{\underline{U}_n\}_{n=n}^{\infty}$ to the output sequence $\{\underline{Z}_n\}_{n=0}^{\infty}$.

An example of a system described by equations of the type (V.B.1) and (V.B.2) is the following.

Example V.B.1: One-Dimensional Motion

Suppose that we wish to model the one-dimensional motion of a particle that is subjected to an acceleration A_t for $t \geq 0$. Note that the position, P_t, and velocity, V_t, of the particle at each time t satisfy the equations $V_t = dP_t/dt$ and $A_t = dV_t/dt$. Assume that we look at the position of the particle every T_s seconds, and we wish to write a model of the form (V.B.1) and (V.B.2) describing the particle's motion from observation time to observation time. Assuming that T_s is small, a Taylor series approximation allows us to write

$$P_{(n+1)T_s} \cong P_{nT_s} + T_s V_{nT_s} \tag{V.B.3a}$$

and

$$V_{(n+1)T_s} \cong V_{nT_s} + T_s A_{nT_s}. \tag{V.B.3b}$$

We see from (V.B.3) that two states are needed to describe the motion of the particle, namely, position and velocity. On defining $Z_n = X_{1,n} = P_{nT_s}, X_{2,n} = V_{nT_s}$, and $U_n = A_{nT_s}$, the motion can be described approximately by the state equation

$$\underline{X}_{n+1} = \mathbf{F}\underline{X}_n + \mathbf{G}U_n, \quad n = 0, 1, \ldots, \tag{V.B.4}$$

and the output equation

$$Z_n = \mathbf{H}\underline{X}_n, \quad n = 0, 1, \ldots, \tag{V.B.5}$$

where $\mathbf{F}$ is the 2×2 matrix

$$\mathbf{F} = \begin{pmatrix} 1 & T_s \\ 0 & 1 \end{pmatrix}, \tag{V.B.6}$$

$\mathbf{G}$ is the 2×1 matrix

$$\mathbf{G} = \begin{pmatrix} 0 \\ T_s \end{pmatrix}, \tag{V.B.7}$$

and $\mathbf{H}$ is the 1×2 matrix

$$\mathbf{H} = (1 \vdots 0). \tag{V.B.8}$$

Thus in this case $m = 2, s = 1, k = 1$, and $\underline{f}_n$ and $\underline{h}_n$ are given, respectively, by

$$\underline{f}_n(\underline{X}, \underline{U}) = \mathbf{F}\underline{X} + \mathbf{G}\underline{U} \tag{V.B.9}$$

and

$$\underline{h}_n(\underline{X}) = \mathbf{H}\underline{X}. \tag{V.B.10}$$

This particular model is discussed further below.

In many applications we are faced with the following problem. We observe the output of a stochastic system in the presence of observation noise (or *measurement noise*) up to some time, say t, and we wish to estimate the state of the system at some time u. That is, we have an observation sequence

$$\underline{Y}_n = \underline{Z}_n + \underline{V}_n, \quad n = 0, 1, \ldots, t, \tag{V.B.11}$$

from which we wish to estimate $\underline{X}_u$. In (V.B.11), the sequence $\underline{V}_0, \underline{V}_1, \ldots,$ represents measurement noise, and (V.B.11) is sometimes known as the *measurement equation.* If $u = t$, this estimation problem is known as the *filtering* problem; for $u < t$, it is known as the *smoothing* problem; and for $u > t$, it is known as the *prediction* problem. Also, the term *state estimation* is applied to all such problems.

As noted above, state estimation problems arise in many applications. For example, in so-called track-while-scan (TWS) radar, radar measurements of the position of a target are made on each scan of a scanning radar. These measurements are noisy observations of a stochastic system similar to that of Example V.B.1 (with random acceleration), and the radar on each scan would like to estimate the current position of the target and also to predict the position the target will occupy on the next scan. At each scanning time t, then, a TWS radar estimates states at $u = t$ and $u = t + 1$ based on the past observation record of the position of the target. (This particular application is discussed further below.)

Other applications of state estimation arise in automatic control systems such as those for aircraft flight control or chemical process control. In flight control the states of interest are the positional coordinates of the aircraft and also the attitudinal coordinates (roll, pitch, and yaw) describing the angular orientation of the aircraft. The state equation in this case describes the dynamics of the aircraft, and the inputs may consist of both control forces and random forces (such as turbulence) operating on the aircraft. In chemical process control the states may be quantities such as temperatures and concentrations of various chemicals, and the state equation describes the dynamics of the chemical reactions involved. Of course, many other applications fit within the context of the general model discussed here.

If we adopt the mean-norm-squared-error performance measure $E\{\| \hat{\underline{X}}_u - \underline{X}_u \|^2\}$ for state estimates $\hat{\underline{X}}_u$ in the model above, we know from Chapter IV (see Case IV.B.4) that the optimum estimate is the conditional mean

$$\hat{\underline{X}}_u = E\{\underline{X}_u | \underline{Y}_0, \ldots, \underline{Y}_t\}. \tag{V.B.12}$$

Of course, for fixed u and t, this problem is no different from the vector estimation problems discussed in Chapter IV. However, we are usually interested in producing estimates in real time as t increases. Since the data set grows linearly with t, the conditional-mean estimates of (V.B.12) will not be practical unless the system model has a structure that makes (V.B.12) computationally efficient. Thus before considering (V.B.12) further, we will first place suitable restrictions on the model of (V.B.1), (V.B.2), and (V.B.11).

One such restriction that we now impose is that the system be a *linear stochastic system*; i.e.,that the state and observation equations are of the form

$$\underline{X}_{n+1} = \mathbf{F}_n \underline{X}_n + \mathbf{G}_n \underline{U}_n, \quad n = 0, 1, \ldots, \tag{V.B.13a}$$

$$\underline{Y}_n = \mathbf{H}_n \underline{X}_n + \underline{V}_n, \quad n = 0, 1, \ldots, \tag{V.B.13b}$$

where, for each $n, \mathbf{F}_n, \mathbf{G}_n$, and $\mathbf{H}_n$ are matrices of appropriate dimensions ($m \times m, m \times s$, and $k \times m$, respectively). The linear model of (V.B.13a) is appropriate for many applications. For example, the one-dimensional motion model (and its two- and three-dimensional analogs) of Example V.B.1 gives rise to a linear stochastic system when the acceleration acting on the particle is random. Also, many nonlinear systems can be approximated by linear systems when the states of interest represent deviations of the system trajectory from some nominal trajectory. In particular, many systems can be linearized about a nominal state trajectory by use of Taylor series expansions of the nonlinearities $\underline{f}_n$.

A further assumption that allows great simplification of the estimate (V.B.12) is that the input sequence $\{\underline{U}_n\}_{n=0}^{\infty}$ and the observation noise $\{\underline{V}_n\}_{n=0}^{\infty}$ are independent sequences of independent zero-mean Gaussian random vectors. It is also convenient to assume that the initial condition $\underline{X}_0$ is a Gaussian random vector independent of $\{\underline{U}_n\}_{n=0}^{\infty}$ and $\{\underline{V}_n\}_{n=0}^{\infty}$. As is discussed briefly below, the independence assumptions on the sequences $\{\underline{U}_n\}_{n=0}^{\infty}$ and $\{\underline{V}_n\}_{n=0}^{\infty}$ can be relaxed. Also, the assumption of zero mean is primarily for convenience. The Gaussian assumption, on the other hand, is crucial. However, this assumption is not unrealistic in many models since the observation noise is often due to Gaussian thermal noise in the sensor electronics, and the random inputs to the system are often due to phenomena such as turbulence that can be modeled accurately as having Gaussian statistics. Moreover, it turns out that the Gaussian assumption can be relaxed if one is willing to accept the best estimator among the class of all *linear* estimators, as will be discussed below.

Within the assumptions above, the conditional-mean state estimator (V.B.12) takes on a very nice form from the viewpoint of computational efficiency. Although this form appears in several other state estimation problems, we will consider the particular problems of filtering ($u = t$) and one-step prediction ($u = t + 1$), as these are the most common cases arising in applications. The simultaneous solution to these two problems is given by the following.

Proposition V.B.1: The Discrete-Time Kalman-Bucy Filter

For the linear stochastic system (V.B.13) with $\{\underline{U}_n\}_{n=0}^{\infty}$ and $\{\underline{V}_n\}_{n=0}^{\infty}$ being independent sequences of independent zero-mean Gaussian vectors independent of the Gaussian initial condition $\underline{X}_0$, the estimates

$\hat{\underline{X}}_{t|t} \triangleq E\{\underline{X}_t|\underline{Y}_0^t\}$[1] and $\hat{\underline{X}}_{t+1|t} \triangleq E\{X_{t+1}|\underline{Y}_0^t\}$ are given recursively by the following equations.

$$\hat{\underline{X}}_{t|t} = \hat{\underline{X}}_{t|t-1} + \mathbf{K}_t(\underline{Y}_t - \mathbf{H}_t\hat{\underline{X}}_{t|t-1}), \quad t = 0, 1, \ldots, \tag{V.B.14a}$$

and

$$\hat{\underline{X}}_{t+1|t} = \mathbf{F}_t\hat{\underline{X}}_{t|t}, \quad t = 0, 1, \ldots, \tag{V.B.14b}$$

with the initialization $\hat{\underline{X}}_{0|-1} = \underline{m}_0 \triangleq E\{\underline{X}_0\}$, where the matrix $\mathbf{K}_t$ is given by

$$\mathbf{K}_t = \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^T(\mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^T + \mathbf{R}_t)^{-1} \tag{V.B.15}$$

with $\boldsymbol{\Sigma}_{t|t-1} \triangleq \mathrm{Cov}(\underline{X}_t|\underline{Y}_0^{t-1})$ and $\mathbf{R}_t \triangleq \mathrm{Cov}(\underline{V}_t)$. Note that since $\hat{\underline{X}}_{t|t-1} = E\{\underline{X}_t|\underline{Y}_0^{t-1}\}$, $\boldsymbol{\Sigma}_{t|t-1}$ is the covariance matrix of the prediction error, $\underline{X}_t - \hat{\underline{X}}_{t|t-1}$, conditioned on $\underline{Y}_0^{t-1}$. This matrix can be computed jointly with the filtering error covariance, $\boldsymbol{\Sigma}_{t|t} \triangleq \mathrm{Cov}(\underline{X}_t|Y_0^t)$ from the following recursion.

$$\boldsymbol{\Sigma}_{t|t} = \boldsymbol{\Sigma}_{t|t-1} - \mathbf{K}_t\mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}, \quad t = 0, 1, \ldots, \tag{V.B.16a}$$

$$\boldsymbol{\Sigma}_{t+1|t} = \mathbf{F}_t\boldsymbol{\Sigma}_{t|t}\mathbf{F}_t^T + \mathbf{G}_t\mathbf{Q}_t\mathbf{G}_t^T, \quad t = 0, 1, \ldots, \tag{V.B.16b}$$

with the initialization $\boldsymbol{\Sigma}_{0|-1} = \boldsymbol{\Sigma}_0 \triangleq \mathrm{Cov}(\underline{X}_0)$, where $\mathbf{Q}_t$ is the covariance matrix of the tth-state input $[\mathbf{Q}_t \triangleq \mathrm{Cov}(\underline{U}_t)]$.

Proof: To prove the proposition, we first show (V.B.14b) and (V.B.16b) directly, and then prove (V.B.14a) and (V.B.16a) by induction. To see (V.B.14b), we note from the state equation that

$$\begin{aligned} \hat{\underline{X}}_{t+1|t} &= E\{\underline{X}_{t+1}|\underline{Y}_0^t\} = E\{\mathbf{F}_t\underline{X}_t + \mathbf{G}_t\underline{U}_t|\underline{Y}_0^t\} \\ &= \mathbf{F}_tE\{\underline{X}_t|\underline{Y}_0^t\} + \mathbf{G}_tE\{\underline{U}_t|\underline{Y}_0^t\} \\ &= F_t\hat{\underline{X}}_{t|t} + \mathbf{G}_tE\{\underline{U}_t|\underline{Y}_0^t\} \end{aligned} \tag{V.B.17}$$

where the third equality follows from the linearity of the expectation and the final equality follows from the definition of $\hat{\underline{X}}_{t|t}$. Note that $\underline{Y}_0^t$ is determined by $\underline{X}_0^t$ and $\underline{V}_0^t$ or in turn by $\underline{X}_0, \underline{U}_0^{t-1}$, and $\underline{V}_0^t$, all of which are independent of $\underline{U}_t$. Thus the conditioning in the second term of (V.B.17) is irrelevant and

[1]For compactness of notation we will use the symbol $\underline{Y}_a^b$ to denote the set $\underline{Y}_a, \ldots, \underline{Y}_b$ for $b > a$.

$E\{\underline{U}_t|\underline{Y}_0^t\} = E\{\underline{U}_t\} = \underline{0}$. Equation (V.B.14b) then follows from (V.B.17). Similarly, we have

$$\begin{aligned} \boldsymbol{\Sigma}_{t+1|t} &= \mathrm{Cov}(\underline{X}_{t+1}|\underline{Y}_0^t) \\ &= \mathrm{Cov}(\mathbf{F}_t\underline{X}_t + \mathbf{G}_t\underline{U}_t|\underline{Y}_0^t) \\ &= \mathrm{Cov}(\mathbf{F}_t\underline{X}_t|\underline{Y}_0^t) + \mathrm{Cov}(\mathbf{G}_t\underline{U}_t|\underline{Y}_0^t) \\ &= \mathrm{Cov}(\mathbf{F}_t\underline{X}_t|\underline{Y}_0^t) + \mathrm{Cov}(\mathbf{G}_t\underline{U}_t), \end{aligned} \tag{V.B.18}$$

since $\underline{U}_t$ is independent of $\underline{X}_t$ and $\underline{Y}_0^t$. Using the property that $\mathrm{Cov}(\mathbf{A}\underline{X}) = \mathbf{A}\mathrm{Cov}(\underline{X})\mathbf{A}^T$ and the definitions of $\boldsymbol{\Sigma}_{t|t}$ and $\mathbf{Q}_t$, we have

$$\begin{aligned} \boldsymbol{\Sigma}_{t+1|t} &= \mathbf{F}_t\mathrm{Cov}(\underline{X}_t|Y_0^t)\mathbf{F}_t^T + \mathbf{G}_t\mathrm{Cov}(\underline{U}_t)\mathbf{G}_t^T \\ &= \mathbf{F}_t\boldsymbol{\Sigma}_{t|t}\mathbf{F}_t^T + \mathbf{G}_t\mathbf{Q}_t\mathbf{G}_t^T, \end{aligned} \tag{V.B.19}$$

which is (V.B.16b).

Thus we have shown that (V.B.14b) and (V.B.16b) hold. We now use induction to show that the other two equations [(V.B.14a) and (V.B.16a)] in the recursion are valid. To do this we must show that they are valid for $t = 0$ and that for arbitrary $t_0 > 0$, their validity for $t = t_0 - 1$ implies their validity for $t = t_0$. For $t = 0$ the measurement equation is given by

$$\underline{Y}_0 = \mathbf{H}_0\underline{X}_0 + \underline{V}_0. \tag{V.B.20}$$

Since $\underline{X}_0$ and $\underline{V}_0$ are independent Gaussian vectors, we see that the estimation of $\underline{X}_0$ from $\underline{Y}_0$ fits the linear estimation model discussed as Example IV.B.3. In particular, since $\underline{X}_0 \sim N(\underline{m}_0, \boldsymbol{\Sigma}_0)$ and $\underline{V}_0 \sim \mathcal{N}(0, \mathbf{R}_0)$, we see from (IV.B.53) that

$$\begin{aligned} \hat{\underline{X}}_{0|0} &\triangleq E\{\underline{X}_0|\underline{Y}_0\} \\ &= \underline{m}_0 + \boldsymbol{\Sigma}_0\mathbf{H}_0^T(\mathbf{H}_0\boldsymbol{\Sigma}_0\mathbf{H}_0^T + \mathbf{R}_0)^{-1}(\underline{Y}_0 - \mathbf{H}_0\underline{m}_0) \\ &= \hat{\underline{X}}_{0|-1} + \mathbf{K}_0(\underline{Y}_0 - \mathbf{H}_0\hat{\underline{X}}_{0|-1}), \end{aligned} \tag{V.B.21}$$

where we have used the following definitions from the proposition: $\hat{\underline{X}}_{0|-1} = \underline{m}_0, \mathbf{K}_0 = \boldsymbol{\Sigma}_{0|-1}\mathbf{H}_0^T(\mathbf{H}_0\boldsymbol{\Sigma}_{0|-1}\mathbf{H}_0^T + \mathbf{R}_0)^{-1}$, and $\boldsymbol{\Sigma}_{0|-1} = \boldsymbol{\Sigma}_0$. Equation (V.B.21) is (V.B.14a) for $t = 0$. The error covariance from (V.B.21) is given from (IV.B.54) as

$$\begin{aligned} \boldsymbol{\Sigma}_{0|0} &= \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0\mathbf{H}_0^T(\mathbf{H}_0\boldsymbol{\Sigma}_0\mathbf{H}_0^T + \mathbf{R}_0)^{-1}\mathbf{H}_0\boldsymbol{\Sigma}_0 \\ &= \boldsymbol{\Sigma}_{0|-1} - \mathbf{K}_0\mathbf{H}_0\boldsymbol{\Sigma}_{0|-1}, \end{aligned} \tag{V.B.22}$$

which is (V.B.16a) for $t = 0$.

To complete the proof, we now assume that (V.B.14a) and (V.B.16a) are valid for $t = t_0 - 1$. Note that $\underline{X}_{t_0}$ and $\underline{Y}_0^{t_0-1}$ are derived by linear transformation of the Gaussian vectors $\underline{X}_0, \underline{U}_0^{t_0-1}$, and $\underline{V}_0^{t_0-1}$. This implies that $\underline{X}_{t_0}$ and $\underline{Y}_0^{t_0-1}$ are jointly Gaussian and thus that $\underline{X}_{t_0}$ is conditionally Gaussian given $\underline{Y}_0^{t_0-1}$. In particular, the conditional distribution of $\underline{X}_{t_0}$ given $\underline{Y}_0^{t_0-1}$ is $\mathcal{N}(\hat{\underline{X}}_{t_0|t_0-1}, \mathbf{\Sigma}_{t_0|t_0-1})$. Also note that $\underline{V}_{t_0}$ is Gaussian and independent of $\underline{Y}_0^{t_0-1}$, so it is also conditionally Gaussian given $\underline{Y}_0^{t_0-1}$ with distribution $\mathcal{N}(\underline{0}, \mathbf{R}_{t_0})$. Since $\underline{V}_{t_0}$ is independent of all of $\underline{X}_0, \underline{V}_0^{t_0-1}$ and $\underline{U}_0^{t_0-1}$, it is conditionally independent of $\underline{X}_{t_0}$ given $\underline{Y}_0^{t_0-1}$. From the remarks above we see that, given $\underline{Y}_0^{t_0-1}$, the observation equation

$$\underline{Y}_{t_0} = \mathbf{H}_{t_0}\underline{X}_{t_0} + \underline{V}_{t_0} \tag{V.B.23}$$

is a Gaussian linear equation of the form discussed in Example IV.B.3. Now, if we compute the conditional expectation of $\underline{X}_{t_0}$ given $\underline{Y}_{t_0}$ under the conditional model (V.B.23) given $\underline{Y}_0^{t_0-1}$ we will get $\hat{\underline{X}}_{t_0|t_0}$, the conditional expectation of $\underline{X}_{t_0}$ given $\underline{Y}_0^{t_0}$. From (IV.B.53) we thus have

$$\begin{aligned}\hat{\underline{X}}_{t_0|t_0} &= \hat{\underline{X}}_{t_0|t_0-1} \\ &\quad + \mathbf{\Sigma}_{t_0|t_0-1}\mathbf{H}_{t_0}^T(\mathbf{H}_{t_0}\mathbf{\Sigma}_{t_0|t_0-1}\mathbf{H}_{t_0}^T + \mathbf{R}_{t_0})^{-1} \\ &\quad \times(\underline{Y}_{t_0} - \mathbf{H}_{t_0}\hat{\underline{X}}_{t_0|t_0-1}),\end{aligned} \tag{V.B.24}$$

where we have used the fact that $\underline{X}_{t_0}$ has the $\mathcal{N}(\hat{\underline{X}}_{t_0|t_0-1}, \mathbf{\Sigma}_{t_0|t_0-1})$ distribution conditioned on $\underline{Y}_0^{t_0-1}$. Using the definition of $\mathbf{K}_{t_0}$, we see that (V.B.24) is (V.B.14a) for $t = t_0$. Similarly, by applying (IV.B.54) and the argument above, we arrive at (V.B.16a). We thus have shown that $\hat{\underline{X}}_{t_0|t_0}$ [resp. $\mathbf{\Sigma}_{t_0|t_0}$] is given in terms of $\hat{\underline{X}}_{t_0|t_0-1}$ [resp. $\mathbf{\Sigma}_{t_0|t_0-1}$] by (V.B.14a) [resp. (V.B.16a)]. We have already shown that $\hat{\underline{X}}_{t_0|t_0-1}$ [resp. $\mathbf{\Sigma}_{t_0|t_0-1}$] is obtained from $\hat{\underline{X}}_{t_0-1|t_0-1}$ [resp. $\mathbf{\Sigma}_{t_0-1|t_0-1}$] via (V.B.14b) [resp. (V.B.16b)], and thus assuming the validity of (V.B.14a) [resp. (V.B.16a)] for $t = t_0 - 1$ implies its validity for $t = t_0$. This completes the proof of the proposition. □

The estimator structure described by Proposition V.B.1 is known as the *discrete-time Kalman-Bucy filter* because it is the discrete-time version of a continuous-time recursive state estimator developed principally by R. E. Kalman and R. S. Bucy in the late 1950s. This estimator is depicted in Fig. V.B.1. The computational simplicity of this structure is evident from the figure. In particular, although the estimators $\hat{\underline{X}}_{t+1|t}$ or $\hat{\underline{X}}_{t|t}$ depend on all the data $\underline{Y}_0^t$, they are computed at each stage from only the latest observation $\underline{Y}_t$ and the previous prediction $\hat{\underline{X}}_{t|t-1}$. Thus rather than having to store the $(t+1)k$-dimensional vectors $\underline{Y}_0$ (and hence having a linearly growing memory and computational burden), we need only to store and update

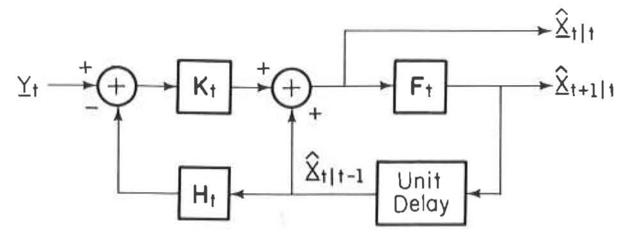

FIGURE V.B.1. The discrete-time Kalman-Bucy filter

the single m-vector $\hat{X}_{t|t-1}$. All other parts of the estimator (including the *Kalman gain matrix*, $\mathbf{K}_t$) are determined completely from the parameters of the model and are independent of the data.

Note that the recursions (V.B.14) and (V.B.16) each consist of two basic steps. The first of these steps is the *measurement update*[(V.B.14a) and (V.B.16a)], which provides the means of updating the estimate and covariance of $\underline{X}_t$ given $\underline{Y}_0^{t-1}$ to incorporate the new observation $\underline{Y}_t$. The second basic step is the *time update* [(V.B.14b)and (V.B.16b)], which provides the means for projecting the state estimate and covariance based on the observation $\underline{Y}_0^t$ to the next time $(t+1)$ before the $(t+1)$st measurement is taken. Examination of the proof of Proposition V.B.1 reveals that the time update is derived exclusively from the state equation, whereas the measurement update is derived from the measurement equation.

It is interesting to consider the measurement update equation (V.B.14a) further. In particular, the estimate $\hat{\underline{X}}_{t|t}$, which is the best estimate of $\underline{X}_t$ based on $\underline{Y}_0^t$, can be viewed as the combination of the best estimate of $\underline{X}_t$ based on the past data, $\hat{\underline{X}}_{t|t-1}$, and a correction term, $\mathbf{K}_t(\underline{Y}_t - \mathbf{H}_t\hat{\underline{X}}_{t|t-1})$. The vector $\underline{I}_t \triangleq (\underline{Y}_t - \mathbf{H}_t\hat{\underline{X}}_{t|t-1})$ appearing in the correction term has an interesting interpretation. In particular, since $\underline{Y}_t = \mathbf{H}_t\underline{X}_t + \underline{V}_t$, we note that $\hat{\underline{Y}}_{t|t-1} \triangleq E\{\underline{Y}_t|\underline{Y}_0^{t-1}\} = \mathbf{H}_t E\{\underline{X}_t|\underline{Y}_0^{t-1}\} + E\{\underline{V}_t|\underline{Y}_0^{t-1}\} = \mathbf{H}_t\hat{\underline{X}}_{t|t-1}$, where we have used the facts that $\underline{V}_t$ is independent of $\underline{Y}_0^{t-1}$ and has zero mean. Thus $\underline{I}_t = \underline{Y}_t - \hat{\underline{Y}}_{t|t-1}$ represents an error signal; it is the error in the prediction of $\underline{Y}_t$ from its past $\underline{Y}_0^{t-1}$. This error is sometimes known as the (prediction) *residual* or the *innovation*. This latter term comes from the fact that we can write $\underline{Y}_t$ as

$$\underline{Y}_t = \hat{\underline{Y}}_{t|t-1} + \underline{I}_t, \tag{V.B.25}$$

with the interpretation that $\hat{\underline{Y}}_{t|t-1}$ is the part of $\underline{Y}_t$ that can be predicted from the past, and $\underline{I}_t$ is the part of $\underline{Y}_t$ that cannot be predicted. Thus $\underline{I}_t$ contains the *new* information that is gained by taking the tth observation; hence the term "innovation." (Recall that this sequence arose in the Gaussian detection problems of Chapter III.)

It is not hard to show that the innovation sequence $\{\underline{I}_t\}_{t=0}^{\infty}$ is a sequence of independent zero-mean Gaussian random vectors. First, the fact that $\{\underline{I}_t\}_{t=0}^{\infty}$ is a Gaussian sequence follows from the fact that $\{\underline{Y}_t\}_{t=0}^{\infty}$ is a Gaussian sequence and that $\{\underline{I}_t\}_{t=0}^{\infty}$ is a linear transformation on $\{\underline{Y}_t\}_{t=0}^{\infty}$. The mean of $\underline{I}_t$ is

$$\begin{aligned} E\{\underline{I}_t\} &= E\{\underline{Y}_t - E\{\underline{Y}_t|\underline{Y}_0^{t-1}\}\} \\ &= E\{\underline{Y}_t\} - E\{\underline{Y}_t\} = 0, \end{aligned}$$

where we have used the iterated expectation property of conditional expectations $(E\{Y\} = E\{E\{Y|X\}\})$. Also, we note that because $E\{\underline{I}_t\} = 0$,

$$\text{Cov}(\underline{I}_t, \underline{I}_s) = E\{\underline{I}_t\ \underline{I}_s^T\}.$$

Assuming that $s < t$, we have

$$E\{\underline{I}_t\underline{I}_s^T\} = E\{E\{\underline{I}_t\underline{I}_s^T|\underline{Y}_0^s\}\} = E\{E\{\underline{I}_t|\underline{Y}_0^s\}\underline{I}_s^T\}, \tag{V.B.26}$$

where the second equality follows from the fact that $\underline{I}_s$ is constant given $\underline{Y}_0^s$. Noting that

$$\begin{aligned} E\{\underline{I}_t|\underline{Y}_0^s\} &= E\{\underline{Y}_t|\underline{Y}_0^s\} - E\{E\{\underline{Y}_t|\underline{Y}_0^{t-1}\}|\underline{Y}_0^s\} \\ &= E\{\underline{Y}_t|\underline{Y}_0^s\} - E\{\underline{Y}_t|\underline{Y}_0^s\} = \underline{0}, \end{aligned}$$

(V.B.26) implies that $\text{Cov}(\underline{I}_t, \underline{I}_s) = \underline{0}$. For $t < s$, a symmetrical argument yields the same result. Thus the innovation vectors are mutually uncorrelated, and since they are jointly Gaussian, this implies that they are mutually independent.

From the discussion above and (V.B.25) we can reiterate the interpretation that $\underline{Y}_t$ consists of a part, $\hat{\underline{Y}}_{t|t-1}$, completely dependent on the past and a part, $\underline{I}_t$, completely independent of the past. This implies that the innovations sequence provides a set of independent observations that is equivalent to the original set $\{\underline{Y}_t\}_{t=0}^{\infty}$. Thus the formation of the innovations sequence is a prewhitening operation as discussed in Chapter III.[2]

The following examples illustrate various properties of the Kalman filter.

Example V.B.2: The Time-Invariant Single-Variable Case

The simplest model with which the Kalman filter can be illustrated is the one-dimensional $(m = k = 1)$ case in which all parameters of the model

[2]Note that the vectors $\underline{I}_t$ are not identically distributed. However, it is easy to see that $\text{Cov}(\underline{I}_t) = \mathbf{H}_t\mathbf{\Sigma}_{t|t-1}\mathbf{H}_t + \mathbf{R}_t \triangleq \mathbf{D}_t$, so $\mathbf{D}_t^{-1/2}\underline{I}_t$ will give a sequence of i.i.d. $\mathcal{N}(\underline{0}, \mathbf{I})$ observations equivalent to $\{\underline{I}_t\}_{t=0}^{\infty}$, where $\mathbf{D}^{1/2}$ denotes the square root of the matrix $\mathbf{D}$ as discussed in Section III.B. Note that the gain $\mathbf{K}_t$ can be written as $\mathbf{\Sigma}_{t|t-1}\mathbf{H}_t\mathbf{D}_t^{-1/2}\mathbf{D}_t^{-1/2}$, so that the Kalman-Bucy filter is actually providing a white (i.i.d.) sequence equivalent to the observation.

are independent of time. In particular, consider the model

$$\begin{aligned} X_{n+1} &= fX_n + U_n, \quad n = 0, 1, \ldots, \\ Y_n &= hX_n + V_n, \quad n = 0, 1, \ldots, \end{aligned} \tag{V.B.27}$$

where $\{U_n\}_{n=0}^{\infty}$ and $\{V_n\}_{n=0}^{\infty}$ are independent sequences of i.i.d. $\mathcal{N}(0, q)$ and $\mathcal{N}(0, r)$ random variables, respectively, $X_0 \sim \mathcal{N}(m_0, \mathbf{\Sigma}_0)$, and where f, h, q, r, and $\mathbf{\Sigma}_0$ are scalars.

The estimation recursions for this case are

$$\hat{X}_{t+1|t} = f\hat{X}_{t|t}, \quad t = 0, 1, \ldots, \tag{V.B.28a}$$

and

$$\hat{X}_{t|t} = \hat{X}_{t|t-1} + K_t(Y_t - h\hat{X}_{t|t-1}), \quad t = 0, 1, \ldots, \tag{V.B.28b}$$

with K_t given by

$$K_t = \frac{\mathbf{\Sigma}_{t|t-1}h}{(h^2\mathbf{\Sigma}_{t|t-1} + r)} = \frac{1}{h}\frac{\mathbf{\Sigma}_{t|t-1}}{\mathbf{\Sigma}_{t|t-1} + r/h^2}. \tag{V.B.29}$$

The role of the Kalman gain in the measurement update (and hence the operation of the Kalman filter) is easily seen from the expression of (V.B.29). In particular, we note that $\mathbf{\Sigma}_{t|t-1}$ is the MSE incurred in the estimation of X_t from $\underline{Y}_0^{t-1}$, and the ratio r/h^2 is a measure of the "noisiness" of the observations. The latter observation follows from the fact that $Y_t/h = X_t + V_t/h$ is an equivalent measurement to Y_t (assuming that $h \neq 0$), and the variance of V_t/h is r/h^2. From these observations on (V.B.29) we see that if the previous prediction of X_t is of much higher quality than the current observation (i.e., $\mathbf{\Sigma}_{t|t-1} \ll r/h^2$), then the gain $K_t \cong 0$ and $\hat{X}_{t|t} \cong \hat{X}_{t|t-1}$. That is, in this case we trust our previous estimate of X_t much more than we trust our observation, so we retain the former estimate. In the opposite situation in which our previous estimate is much noisier than our observation (i.e., $\mathbf{\Sigma}_{t|t-1} \gg r/h^2$), the Kalman gain $K_t \cong 1/h$, and $\hat{X}_{t|t} \cong Y_t/h$. Thus in the second case we simply ignore our previous measurements and invert the current measurement equation. Of course, between these two extremes the measurement update balances these two ways of updating. The update in the vector case has a similar interpretation, although it cannot be parametrized as easily as in this scalar case.

It is interesting to compare the measurement update here with the Bayesian estimation of signal amplitude as discussed in Example IV.B.2. In particular, we can write the measurement update equation as

$$\hat{X}_{t|t} = \frac{v^2d^2\hat{\theta}_1 + \mu}{v^2d^2 + 1}, \tag{V.B.30}$$

where we have identified $\hat{\theta}_1 = Y_t/h, \mu = \hat{X}_{t|t-1}, v^2 = \Sigma_{t|t-1}$, and $d^2 = h^2/r$. Comparing (V.B.30) with (IV.B.34), we see that the distribution of X_t conditioned on Y_0^{t-1} can be interpreted as a prior distribution for X_t [it is $\mathcal{N}(\hat{X}_{t|t-1}, \Sigma_{t|t-1})$], and the update balances this prior knowledge with the knowledge gained by the observation Y_t, according to the value of v^2d^2. (Of course, this fact is the essence of the derivation of the measurement update given in the proof of Proposition V.B.1.)

For this scalar time-invariant model, the time and measurement updates for the estimation covariance become

$$\Sigma_{t+1|t} = f^2\Sigma_{t|t} + q \tag{V.B.31a}$$

and

$$\Sigma_{t|t} = \frac{\Sigma_{t|t-1}}{\frac{h^2}{r}\Sigma_{t|t-1} + 1}. \tag{V.B.31b}$$

Note that we can eliminate the coupling between these equations to get separate recursions for each quantity. For example, inserting (V.B.31b) into (V.B.31a) yields the recursion

$$\Sigma_{t+1|t} = \frac{f^2\Sigma_{t|t-1}}{h^2\Sigma_{t|t-1}/r + 1} + q, \quad t = 0, 1, \ldots. \tag{V.B.32}$$

(Of course the initialization is $\Sigma_{0|-1} = \Sigma_0$.)

In examining (V.B.32), the question arises as to whether the sequence generated by this recursion approaches a constant as t increases. If so, the Kalman gain approaches a constant also and the Kalman-Bucy filter becomes time-invariant asymptotically in t. Note that if $\Sigma_{t+1|t}$ does approach a constant, say Σ_∞, then Σ_∞ must satisfy

$$\Sigma_\infty = \frac{f^2\Sigma_\infty}{h^2\Sigma_\infty/r + 1} + q \tag{V.B.33}$$

since both $\Sigma_{t+1|t}$ and $\Sigma_{t|t-1}$ are approaching Σ_∞. Equation (V.B.33) is a quadratic equation and it has the unique positive solution

$$\begin{aligned}\Sigma_\infty &= \frac{1}{2}\left\{\left[\frac{r}{h^2}(1-f^2) - q\right]^2 + \frac{4rq}{h^2}\right\}^{1/2} \\ &\quad - \frac{r}{2h^2}(1-f^2) + q.\end{aligned} \tag{V.B.34}$$

On combining (V.B.32) and (V.B.33), we have[3]

$$\begin{aligned}|\Sigma_{t+1|t} - \Sigma_\infty| &= f^2\left|\frac{\Sigma_{t|t-1}}{h^2\Sigma_{t|t-1}/r + 1} - \frac{\Sigma_\infty}{h^2\Sigma_\infty/r + 1}\right| \\ &\leq f^2|\Sigma_{t|t-1} - \Sigma_\infty|, \; t = 0, 1, \ldots,\end{aligned} \tag{V.B.35}$$

[3]To see the inequality in (V.B.35), define $g(x) = x/(ax+1)$, with $a = h^2/r$.

which implies that

$$|\Sigma_{t+1|t} - \Sigma_\infty| \leq f^{2(t+1)}|\Sigma_0 - \Sigma_\infty|. \tag{V.B.36}$$

If $|f| < 1$, then (V.B.36) implies that $\Sigma_{t+1|t} \to \Sigma_\infty$ as $t \to \infty$. Thus the condition $|f| < 1$ is sufficient for the Kalman-Bucy filter and its performance to approach a steady state for this model. [Note that $|f| < 1$ is also the condition for asymptotic stability of the original system (V.B.27).]

Example V.B.3: Track-While-Scan (TWS) Radar

A commonly used type of radar is one that regularly scans some area (say an airfield) and keeps track of the trajectories of various targets in the scanning area by processing position measurements taken once each scan. The radar also predicts the positions the targets will occupy on the next scan. Since the maneuver strategies of the targets are usually unknown to the radar, one way of modeling target motion for the purposes of devising optimum tracking schemes is to assume that the targets of interest undergo random accelerations. A simple model for this type of motion is to assume that these accelerations are i.i.d. from scan to scan and are Gaussian. Although the target motion is three-dimensional, it is simpler to discuss this tracking problem in a single dimension only. The assumptions above lead to a state/measurement model of the form described in Example V.B.1. In particular, we can use the model

$$\begin{aligned}\begin{pmatrix} P_{n+1} \\ V_{n+1}\end{pmatrix} &= \begin{pmatrix} 1 & T_s \\ 0 & 1\end{pmatrix}\begin{pmatrix} P_n \\ V_n\end{pmatrix} + \begin{pmatrix} 0 \\ T_s\end{pmatrix} A_n \\ Y_n &= (1 \vdots 0)\begin{pmatrix} P_n \\ V_n\end{pmatrix} + \epsilon_n,\end{aligned} \tag{V.B.37}$$

where P_n and V_n represent the target position and velocity, respectively, on the nth scan, T_s is the time the radar takes to complete each scan, A_n is the target acceleration during the nth scanning period, Y_n is the position measurement at the nth sighting, and ϵ_n is the error in this measurement. (To track in all three dimensions we would have a six-state, three-measurement model. However, if the accelerations and measurement noises in the three dimensions are independent of one another, the three dimensions can be tracked independently.)

By Taylor's theorem, we have for each real x and y,

$$|g(x) - g(y)| = |x - y|\;|g'(\xi)|$$

for some ξ between x and y. We have $g'(\xi) = 1/(a\xi+1)^2$ which satisfies $|g'(\xi)| \leq 1$ for $\xi \geq 0$. Since $\Sigma_{t|t-1} > 0$, and $\Sigma_\infty > 0$, we have $|g(\Sigma_{t|t-1}) - g(\Sigma_\infty)| \leq |\Sigma_{t|t-1} - \Sigma_\infty|$.

Thus assuming that all statistics are Gaussian and time-invariant, the optimum tracker/predictor equations are

$$\begin{pmatrix} \hat{P}_{t+1|t} \\ \hat{V}_{t+1|t} \end{pmatrix} = \begin{pmatrix} \hat{P}_{t|t} + T_s \hat{V}_{t|t} \\ \hat{V}_{t|t} \end{pmatrix} \tag{V.B.38}$$

and

$$\begin{pmatrix} \hat{P}_{t|t} \\ \hat{V}_{t|t} \end{pmatrix} = \begin{pmatrix} \hat{P}_{t|t-1} \\ \hat{V}_{t|t-1} \end{pmatrix} + \begin{pmatrix} K_{t,1} \\ K_{t,2} \end{pmatrix} (Y_t - \hat{P}_{t|t-1}), \tag{V.B.39}$$

where in this case, the gain matrix $\mathbf{K}_t$ is a 2×1 vector. This gain vector is given by

$$\begin{pmatrix} K_{t,1} \\ K_{t,2} \end{pmatrix} = \begin{pmatrix} \Sigma_{t|t-1}(1,1)/(\Sigma_{t|t-1}(1,1) + r) \\ \Sigma_{t|t-1}(2,1)/(\Sigma_{t|t-1}(1,1) + r) \end{pmatrix} \tag{V.B.40}$$

where $\Sigma_{t|t-1}(k,l)$ is the $(k-l)$th component of the matrix $\mathbf{\Sigma}_{t|t-1}$, and where r is the variance of the measurement noise. The matrix $\mathbf{\Sigma}_{t|t-1}$, of course, is computed through the recursions of Proposition V.B.1.

To reduce the computational burden of this tracker, the time-varying filter (V.B.39) is sometimes replaced in practical systems with a time-invariant filter

$$\begin{pmatrix} \hat{P}_{t|t} \\ \hat{V}_{t|t} \end{pmatrix} = \begin{pmatrix} \hat{P}_{t|t-1} \\ \hat{V}_{t|t-1} \end{pmatrix} + \begin{pmatrix} \alpha \\ \beta/T_s \end{pmatrix} (Y_t - \hat{P}_{t|t-1}), \tag{V.B.41}$$

where α and β are constants. The constants α and β can be chosen to trade-off various performance characteristics, such as speed of response and accuracy of track. This type of tracker is sometimes known as an α–β *tracker*.

The TWS radar problem will be discussed further below.

Returning to the general Kalman-Bucy filter of Proposition V.B.1, we note that the coupled recursions in each of (V.B.14) and (V.B.16) can be separated to give recursions for the prediction quantities $\underline{\hat{X}}_{t+1|t}$ and $\mathbf{\Sigma}_{t+1|t}$ not involving the filtering quantities $\underline{\hat{X}}_{t|t}$ and $\mathbf{\Sigma}_{t|t}$, and vice versa (as was noted in Example V.B.2). For example on substituting the measurement updates into the time updates we have

$$\underline{\hat{X}}_{t+1|t} = \mathbf{F}_t \underline{\hat{X}}_{t|t-1} + \mathbf{F}_t \mathbf{K}_t \underline{I}_t, \quad t = 0, 1, \ldots, \tag{V.B.42a}$$

and

$$\begin{aligned} \mathbf{\Sigma}_{t+1|t} \;&=\; \mathbf{F}_t \mathbf{\Sigma}_{t|t-1} \mathbf{F}_t^T - \mathbf{F}_t \mathbf{K}_t \mathbf{H}_t \mathbf{\Sigma}_{t|t-1} \mathbf{F}_t^T \\ &\quad + \mathbf{G}_t \mathbf{Q}_t \mathbf{G}_t, \quad t = 0, 1, \ldots. \end{aligned} \tag{V.B.42b}$$

Note that the prediction filter (V.B.42a) is a linear stochastic system driven by the innovations sequence. This system has the same dynamics (i.e., $\mathbf{F}_t$'s) as the system we are trying to track. Thus to track $\underline{X}_t$ we are building a system comprising a duplicate of the dynamics that govern $\underline{X}_t$ and then driving it with the innovations through the matrix sequence $\mathbf{F}_t\mathbf{K}_t$.

The covariance update (V.B.42b) is a dynamical system with a matrix state. It is a nonlinear system since the $\mathbf{K}_t$ term in the second term on the right depends on $\mathbf{\Sigma}_{t|t-1}$. This equation is known as a (discrete-time) *Riccati equation*. As in the scalar case of Example V.B.2, the time-invariant version of this equation (in which $\mathbf{F}_t, \mathbf{G}_t, \mathbf{H}_t, \mathbf{Q}_t$, and $\mathbf{R}_t$ are all independent of t) can be studied for possible convergence to steady state. A sufficient (but not necessary) condition for $\mathbf{\Sigma}_{t+1|t}$ to converge to a steady state is that all eigenvalues of $\mathbf{F}$ have less than unit magnitude. (This condition is necessary and sufficient for the original system to be asymptotically stable.) Another issue relating to (V.B.42b) is that numerical problems sometimes arise in the computation of the matrix inverse $(\mathbf{H}_t\mathbf{\Sigma}_{t|t-1}\mathbf{H}_t^T + \mathbf{R}_t)^{-1}$ appearing in the $\mathbf{K}_t$ term of this equation. Thus it is sometimes convenient to replace (V.B.42b) with an equivalent equation for propagating the square root of $\mathbf{\Sigma}_{t+1|t}$ which leads to fewer numerical problems. See Anderson and Moore (1979) for a discussion of these and related issues.

All of the assumptions regarding the system and measurement models that we have made here were used in the derivation of the Kalman-Bucy filter. All of these assumptions are necessary, but as mentioned earlier in this section, some of them can be circumvented by appropriately redefining the model or performance objectives. For example, the independence assumptions on the input and noise sequences $\{\underline{U}_k\}_{k=0}^{\infty}$ and $\{\underline{V}_k\}_{k=0}^{\infty}$ can be relaxed by modeling these processes as themselves being derived from linear stochastic systems driven by independent sequences. The states of the original stochastic system can then be augmented with the states of these additional systems to give an overall higher-dimensional model, but one driven by and observed in independent sequences. The standard Kalman-Bucy filter can then be applied to this augmented system. The disadvantage of this approach, of course, is that it requires a higher-dimensional filter because the noise and input states must also be tracked.

To illustrate this approach we consider the following modification of Example V.B.3.

Example V.B.4: TWS Radar with Dependent Acceleration Sequences

In this example we reconsider the track-while-scan (TWS) radar application discussed in Example V.B.3. For the scanning speeds and target types of interest in many applications, it is often unrealistic to assume that the target acceleration is independent from scan to scan. (For example, the inertial characteristics of the target may preclude such motion.) A simple yet useful model for target acceleration that allows for dependence between accelerations on different scans is that the acceleration sequence $\{A_n\}_{n=0}^{\infty}$ is generated by the stochastic system

$$A_{n+1} = \rho A_n + W_n, \quad n = 0, 1, \ldots, \tag{V.B.43}$$

with a Gaussian initial condition A_0 and an i.i.d. Gaussian input sequence $\{W_n\}_{n=0}^{\infty}$, where ρ is a parameter satisfying $0 \leq \rho < 1$. Note that if $\rho = 0$, there is no dependence in the acceleration sequence, whereas larger values of ρ imply more highly correlated accelerations.

With accelerations satisfying (V.B.43), the model of (V.B.37) no longer satisfies the assumptions required for the Kalman-Bucy filter. However, we can augment this model to include the acceleration dynamics (V.B.43) by treating the acceleration as a state rather than as an input. In particular, we have the model

$$\begin{pmatrix} P_{n+1} \\ V_{n+1} \\ A_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & T_s & 0 \\ 0 & 1 & T_s \\ 0 & 0 & \rho \end{pmatrix} \begin{pmatrix} P_n \\ V_n \\ A_n \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} W_n, \quad n = 0, 1, \ldots, \tag{V.B.44a}$$

$$Y_n = (1 \vdots 0 \vdots 0) \begin{pmatrix} P_n \\ V_n \\ A_n \end{pmatrix} + \epsilon_n, \quad n = 0, 1, \ldots, \tag{V.B.44b}$$

which leads to the estimator recursions

$$\begin{pmatrix} \hat{P}_{t+1|t} \\ \hat{V}_{t+1|t} \\ \hat{A}_{t+1|t} \end{pmatrix} = \begin{pmatrix} \hat{P}_{t|t} + T_s \hat{V}_{t|t} \\ \hat{V}_{t|t} + T_s \hat{A}_{t|t} \\ \rho \hat{A}_{t|t} \end{pmatrix} \tag{V.B.45a}$$

and

$$\begin{pmatrix} \hat{P}_{t|t} \\ \hat{V}_{t|t} \\ \hat{A}_{t|t} \end{pmatrix} = \begin{pmatrix} \hat{P}_{t|t-1} \\ \hat{V}_{t|t-1} \\ \hat{A}_{t|t-1} \end{pmatrix} + \begin{pmatrix} K_{t,1} \\ K_{t,2} \\ K_{t,3} \end{pmatrix} (Y_t - \hat{P}_{t|t-1}), \tag{V.B.45b}$$

where the gains are given by

$$\begin{pmatrix} K_{t,1} \\ K_{t,2} \\ K_{t,3} \end{pmatrix} = \begin{pmatrix} \Sigma_{t|t-1}(1,1)/(\Sigma_{t|t-1}(1,1) + r) \\ \Sigma_{t|t-1}(2,1)/(\Sigma_{t|t-1}(1,1) + r) \\ \Sigma_{t|t-1}(3,1)/(\Sigma_{t|t-1}(1,1) + r) \end{pmatrix}. \tag{V.B.46}$$

Note that we now must track the acceleration in addition to position and velocity. As in the lower-order model of Example V.B.3, the gain vector in (V.B.45b) is sometimes replaced in practice with a constant vector, usually denoted by

$$\begin{pmatrix} \alpha \\ \beta/T_s \\ \gamma/T_s^2 \end{pmatrix},$$

in order to reduce computational requirements. The result is known as an *α-β-γ tracker*, and the three parameters α, β, and γ are chosen to given desired performance characteristics.

The example above illustrates how dependence in the input sequence can be handled in the Kalman-Bucy filtering model. For a more detailed discussion of the issue of dependence, the reader is referred to Anderson and Moore (1979). The other principal assumptions in the Kalman model are the linearity of the state and measurement equations and the Gaussianity of the statistics. The latter assumption can be dropped if one is interested in optimizing over all linear filters rather than over all estimators as we have done here. Note that the Kalman-Bucy filter is specified by the second-order statistics (mean and covariances) of the random quantities in the model, and it is in fact the optimum (MMSE) estimator among all linear filters for any initial condition, input and noise sequences with these given second-order statistics (whether they are Gaussian or not). This issue is discussed in Section V.C. The assumption of linearity in the state and observation equations is more difficult to relax than that of Gaussianity. Without this linearity the MMSE state estimation problem becomes quite difficult analytically. Nevertheless, there are several useful techniques for dealing with state estimation in nonlinear systems. Some of these are discussed in Section VII.C in the context of continuous-time signal estimation.

V.C Linear Estimation

In Section V.B we considered optimum estimation in the linear stochastic system model with Gaussian statistics. As noted above, the Kalman-Bucy filter is optimum not only for this model but is also optimum among all linear estimators for the same model with non-Gaussian statistics provided that the second-order statistics of the model (i.e., means and covariances) remain unchanged. The latter result is a particular case of a general theory of optimum linear estimation in which only second-order statistics are needed to specify the optimum procedures. In this section we develop this idea further, and in the following section we apply this theory to a general class of problems known as Wiener-Kolmogorov filtering.

Suppose that we have two sequences of random variables $\{Y_n\}_{n=-\infty}^{\infty}$ and $\{X_n\}_{n=-\infty}^{\infty}$. We observe Y_n some set of times $a \leq n \leq b$ and we wish to estimate X_t from these observations for some particular time t. Of course, the optimum estimator (in the MMSE sense) is the conditional mean, $\hat{X}_t = E\{X_t|Y_a^b\}$, and the computation of this estimate has been discussed previously. However, if the number of observations $(b-a+1)$ is large, this computation can be quite cumbersome unless the problem exhibits special structure (as in the Kalman-Bucy model). Furthermore, the determination of the conditional mean generally requires knowledge of the joint distribution of the variables $X_t, Y_a, \ldots, Y_b$, knowledge that may be impractical (or impossible) to obtain in practice.

One way of circumventing the first of these problems is to constrain the estimators to be considered to be of some computationally convenient form, and then to minimize the MSE over this constrained class. One such constraint that is quite useful in this context is the *linear constraint*, in which we consider estimates $\hat{X}_t$ of the form

$$\hat{X}_t = \sum_{n=a}^{b} h_{t,n} Y_n + c_t, \tag{V.C.1}$$

where $h_{t,a}, \ldots, h_{t,b}$, and c_t are scalars.[4] As we shall see below, this constraint also solves the second problem of having to specify the joint distribution of all variables, since only knowledge of *second-order statistics* will be needed to optimize over linear estimates. Before considering this optimization, we must first note some analytical properties of the sum (V.C.1).

For finite a and b, the meaning of the sum in (V.C.1) is clear. However, we will also be interested in cases in which $a = -\infty, b = +\infty$, or both. Although the meaning of (V.C.1) is clear from a practical viewpoint in such cases, for analytical purposes we must define precisely what we mean by these infinite sums of random variables. The most useful definition in this context is the mean-square sum, in which, for example, for $a = -\infty$ and b finite, the equation (V.C.1) means that

$$\lim_{m\to-\infty} E\left\{\left(\sum_{n=m}^{b} h_{t,n} Y_n + c_t - \hat{X}_t\right)^2\right\} = 0. \tag{V.C.2}$$

The sum in (V.C.1) is defined similarly for $b = +\infty$ with a finite and for $a = -\infty, b = +\infty$. Because of the limiting definition of (V.C.2), the observation set for $a = -\infty$ and b finite should be interpreted as $a < t \le b$ rather than $a \le t \le b$, with a similar interpretation for $b = +\infty$.

[4]Estimates of the form (V.C.1) are more properly termed *affine*. Because of the additive constant c_t, they are not actually linear. However, the term "linear" is fairly standard in this context, so we will use it here. It should be noted that if $X_t, Y_a, \ldots, Y_b$ are jointly Gaussian random variables, then $E\{X_t|Y_a^b\}$ is of the form (V.C.1), so optimization over linear estimates yields globally optimum estimators in this particular case.

In order to proceed further with the linear estimation problem, we assume for the remainder of this section that $\{X_n\}_{n=-\infty}^{\infty}$ and $\{Y_n\}_{n=-\infty}^{\infty}$ are second-order sequences, i.e., that $E\{X_n^2\} < \infty$ and $E\{Y_n^2\} < \infty$ for all n. Also, we denote by $\mathcal{H}_a^b$ the set of all estimates of the form (V.C.1) based on Y_a^b. The following preliminary results concerning $\mathcal{H}_a^b$ will be used later.

Proposition V.C.1:
Suppose that $\hat{X}_t \in \mathcal{H}_a^b$. Then
(i) $E\{(\hat{X}_t)^2\} < \infty$; and
(ii) if Z is a random variable satisfying $E\{Z^2\} < \infty$, then

$$E\{Z\hat{X}_t\} = \sum_{n=a}^{b} h_{t,n} E\{ZY_n\} + c_t E\{Z\}.$$

Proof: These two properties are obvious if a and b are both finite. In this case, property (i) follows from successive application of the inequality, $(x+y)^2 \le 4(x^2+y^2)$, and property (ii) is simply the linearity property of expectation. To prove these properties for the situation in which a, b, or both is infinite, we consider the specific case in which $a = -\infty$ and b is finite. (Proofs for the other two cases are identical to this one.)

To prove property (i) in this case we write, for $m < b$,

$$\hat{X}_t = \sum_{n=m}^{b} h_{t,n} Y_n + c_t + \left(\hat{X}_t - \sum_{n=m}^{b} h_{t,n} Y_n - c_t\right). \tag{V.C.3}$$

Again, using the inequality $(x+y)^2 \le 4(x^2+y^2)$ and taking expectations, we have

$$\begin{aligned} E\{(\hat{X}_t)^2\} &\le 4E\left\{\left(\sum_{n=m}^{b} h_{t,n} Y_n + c_t\right)^2\right\} \\ &+ 4E\left\{\left(\hat{X}_t - \sum_{n=m}^{b} h_{t,n} Y_n - c_t\right)^2\right\}. \end{aligned} \tag{V.C.4}$$

The first term on the right-hand side of (V.C.4) is finite by the validity of property (i) for finite a and b. The second term on the right-hand side of (V.C.4) converges to zero an m approaches -∞. Thus there must be a value of m that makes this term finite, which implies that $E\{(\hat{X}_t)^2\} < \infty$.

To prove property (ii), we consider for $m < b$ the quantity

$$E\{Z\hat{X}_t\} - \sum_{n=m}^{b} h_{t,n}E\{XY_n\} - c_tE\{Z\} =$$
$$E\left\{Z\left(\hat{X}_t - \sum_{n=m}^{b} h_{t,n}Y_n - c_t\right)\right\}. \tag{V.C.5}$$

From the Schwarz inequality we have

$$\left|E\left\{Z\left(\hat{X}_t - \sum_{n=m}^{b} h_{t,n}Y_n - c_t\right)\right\}\right|^2$$
$$\leq E\{Z^2\}E\left\{\left(\hat{X}_t - \sum_{n=m}^{b} h_{t,n}Y_n - c_t\right)^2\right\}. \tag{V.C.6}$$

By assumption $E\{Z^2\} < \infty$ and by definition $E\{(\hat{X}_t - \sum_{n=m}^{b} h_{t,n}Y_n - c_t)^2\} \to 0$ as $m \to -\infty$. Thus (V.C.5) and (V.C.6) imply property (ii). This completes the proof of this proposition. □

Having constrained ourselves to estimators of the form (V.C.1), we would like to find the best such estimate in the minimum-mean-squared-error sense; i.e., we would like to solve the problem

$$\min_{\hat{X}_t \in \mathcal{H}_a^b} E\{(\hat{X}_t - X_t)^2\}. \tag{V.C.7}$$

The solution to this problem can be characterized by the following.

Proposition V.C.2: The Orthogonality Principle
$\hat{X}_t \in \mathcal{H}_a^b$ solves (V.C.7) if and only if

$$E\{(\hat{X}_t - X_t)Z\} = 0 \text{ for all } Z \in \mathcal{H}_a^b. \tag{V.C.8}$$

Proof: First suppose that $\hat{X}_t$ satisfies (V.C.8), and let $\tilde{X}_t$ be any other estimate in $\mathcal{H}_a^b$. Then the MSE associated with $\tilde{X}_t$ is given by

$$\begin{aligned} E\{(X_t - \tilde{X}_t)^2\} &= E\{(X_t - \hat{X}_t + \hat{X}_t - \tilde{X}_t)^2\} \\ &= E\{(X_t - \hat{X}_t)^2\} \\ &\quad + 2E\{(X_t - \hat{X}_t)(\hat{X}_t - \tilde{X}_t)\} \\ &\quad + E\{(\hat{X}_t - \tilde{X}_t)^2\}. \end{aligned} \tag{V.C.9}$$

It is easy to see that $\hat{X}_t \in \mathcal{H}_a^b$ and $\tilde{X}_t \in \mathcal{H}_a^b$ imply that $(\hat{X}_t - \tilde{X}_t) \in \mathcal{H}_a^b$, and thus the second term on the right-hand side of (V.C.9) is zero. This gives

$$\begin{aligned} E\{(X_t - \tilde{X}_t)^2\} &= E\{(X_t - \hat{X}_t)^2\} + E\{(\hat{X}_t - \tilde{X}_t)^2\} \\ &\geq E\{(X_t - \hat{X}_t)^2\}. \end{aligned} \tag{V.C.10}$$

Since $\tilde{X}_t$ was chosen arbitrarily (V.C.10) proves the sufficiency of (V.C.8) for $\hat{X}_t$ to solve (V.C.7).

To prove the necessity of (V.C.8), suppose that $\tilde{X}_t \in \mathcal{H}_a^b$ and that there is a $Z \in \mathcal{H}_a^b$ such that $E\{(X_t - \tilde{X}_t)Z\} \neq 0$. Define a new estimator $\hat{X}_t$ by

$$\hat{X}_t = \tilde{X}_t + \frac{E\{(X_t - \tilde{X}_t)Z\}}{E\{Z^2\}}Z. \tag{V.C.11}$$

(Note that the condition $E\{(X_t - \tilde{X}_t)Z\} \neq 0$ implies that $E\{Z^2\} > 0$.) A straightforward computation gives that

$$\begin{aligned} E\{(X_t - \hat{X}_t)^2\} &= E\{(X_t - \tilde{X}_t)^2\} - \frac{|E\{(X_t - \tilde{X}_t)Z\}|^2}{\{EZ^2\}} \\ &< E\{(X_t - \tilde{X}_t)^2\}. \end{aligned} \tag{V.C.12}$$

Thus $\hat{X}_t$ is a better estimator than $\tilde{X}_t$, so $\tilde{X}_t$ cannot solve (V.C.7). This proves the necessity of (V.C.8) and completes the proof of this proposition. □

Proposition V.C.2 says that $\hat{X}_t$ is a MMSE linear estimator of X_t given Y_a^b if and only if the estimation error, $X_t - \hat{X}_t$, is orthogonal to every linear function of the observations Y_a^b. This result is known as the *orthogonality principle.*[5] This result is a special case of a more general result in analysis known as the *projection theorem*, which has the following familiar form in the particular case of a finite-dimensional vector space.

Suppose that $\underline{x}$ and $\underline{y}$ are two vectors of the same dimension, and suppose that we would like to approximate $\underline{x}$ by a constant, say α, times $\underline{y}$ such that the length of the error vector $\underline{x} - \alpha\underline{y}$ is as small as possible. It is easy to see that α minimizes this length if and only if the error vector is perpendicular

[5]It is interesting to note that the conditional-mean estimator $\hat{X}_t = E\{X_t|Y_a^b\}$ uniquely satisfies the analogous condition

$$E\{(X_t - \hat{X}_t)Z\} = 0 \text{ for all } Z \in \mathcal{G}_a^b,$$

where $\mathcal{G}_a^b$ denotes the set of all random variables of the form $g(Y_a^b)$ satisfying $E\{g^2(Y_a^b)\} < \infty$.

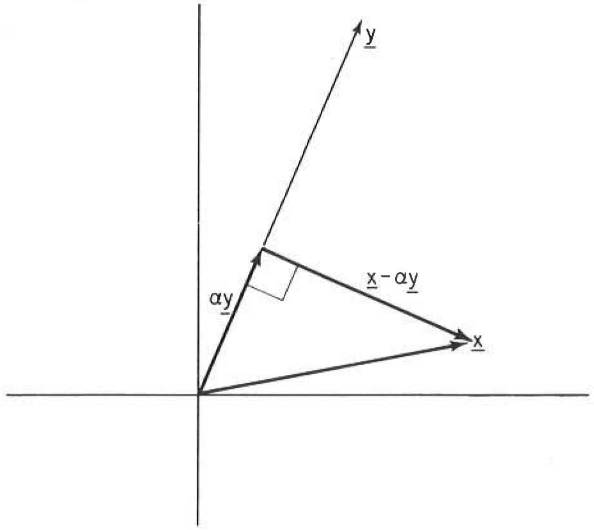

FIGURE V.C.1. Illustration of the orthogonality principle.

(i.e., orthogonal) to the line that is aligned along $\underline{y}$ (see Fig. V.C.1) and hence to every constant multiple of $\underline{y}$. The resulting approximation is the *projection of* $\underline{x}$ *in the* $\underline{y}$ *direction.*

The analogy between the problem and that of linear MMSE estimation is straightforward. The vector $\underline{y}$ is analogous to our observations Y_a^b and the line aligned along $\underline{y}$ is analogous to the set of all linear estimates $\mathcal{H}_a^b$. The vector $\underline{x}$ corresponds to our quantity X_t to be estimated, and the length of the error vector $\| \underline{x} - \alpha \underline{y} \|^2$ is analogous to the MSE, $E\{(X_t - \hat{X}_t)^2\}$. Thus we can think of the linear MMSE estimate as being the *projection* of X_t onto the data Y_a^b.

The result of Proposition V.C.2 characterizes solutions to (V.C.7). A more convenient form of this result for finding such solutions is given by the following result.

Proposition V.C.3: An Alternative Orthogonality Condition
$\hat{X}_t$ solves (V.C.7) if and only if

$$E\{\hat{X}_t\} = E\{X_t\} \tag{V.C.13}$$

and

$$E\{(X_t - \hat{X}_t)Y_l\} = 0, \text{ for all } a \le l \le b. \tag{V.C.14}$$

Proof: The necessity of (V.C.13) and (V.C.14) follows from the application the orthogonality condition (V.C.8) to the particular elements of

$\mathcal{H}_a^b$, $Z = 1$ and $Z = Y_l$, respectively. The sufficiency follows from property (ii) of Proposition V.C.1. In particular, if we assume that $\hat{X}_t$ satisfies (V.C.13) and (V.C.14), then with $Z = \sum_{n=a}^b h_{t,n} Y_n + c_t$, we have

$$\begin{aligned} E\{(X_t - \hat{X}_t)Z\} &= \sum_{n=a}^b h_{t,n} E\{(X_t - \hat{X}_t)Y_n\} \\ &+ c_t E\{X_t - \hat{X}_t\} = 0, \end{aligned} \tag{V.C.15}$$

which gives (V.C.8). This completes the proof. □

Using Proposition V.C.2, we can obtain equations specifying the coefficients of an optimum estimator of the form (V.C.1). In particular, on substituting (V.C.1) into (V.C.13), we have

$$E\left\{\sum_{n=a}^b h_{t,n} Y_n + c_t\right\} = E\{X_t\},$$

from which we have [using property (ii) of Proposition V.C.1 with $Z = 1$]

$$c_t = E\{X_t\} - \sum_{n=a}^b h_{t,n} E\{Y_n\}. \tag{V.C.16}$$

From (V.C.1) and (V.C.14), we have the relationship

$$E\left\{\left(X_t - \sum_{n=a}^b h_{t,n} Y_n - c_t\right) Y_l\right\} = 0, \quad a \le l \le b. \tag{V.C.17}$$

Substituting (V.C.16) into (V.C.17), we get, successively,

$$E\left\{\left[(X_t - E\{X_t\}) - \sum_{n=a}^b h_{t,n}(Y_n - E\{Y_n\})\right] Y_l\right\} = 0, \quad a \le l \le b,$$

$$E\{(X_t - E\{X_t\})Y_l\} = \sum_{n=a}^b h_{t,n} E\{(Y_n - E\{Y_n\})Y_l\}, \quad a \le l \le b,$$

$$\mathrm{Cov}(X_t, Y_l) = \sum_{n=a}^b h_{t,n} \mathrm{Cov}(Y_n, Y_l), \quad a \le l \le b,$$

and finally

$$C_{XY}(t, l) = \sum_{n=a}^b h_{t,n} C_Y(n, l), \quad a \le l \le b, \tag{V.C.18}$$

where $C_{XY}(t,l) \triangleq \text{Cov}(X_t, Y_l)$ is the *cross-covariance function* of the sequences $\{X_n\}_{n=-\infty}^{\infty}$ and $\{Y_n\}_{n=-\infty}^{\infty}$, and where $C_Y(n,l) \triangleq \text{Cov}(Y_n, Y_l)$ is the *autocovariance function* of the sequence $\{Y_n\}_{n=-\infty}^{\infty}$.

Equations (V.C.16) and (V.C.18) give equations that are necessary and sufficient for a set of coefficients $\{h_{t,n}\}_{n=a}^{b}$ and c_t to yield an optimum linear estimator of X_t from Y_a^b. Note that these equations involve only the means, covariances, and cross-covariances (i.e., the *second-order statistics*) of Y_a^b and X_t. This provides a significant practical advantage over the conditional-mean estimator, $E\{X_t|Y_a^b\}$ which in general requires the joint distribution of Y_a^b and X_t, since second-order statistics are much easier to model analytically or to estimate accurately from observed data than are multivariate distribution functions. If $\{h_{t,n}\}_{n=a}^{b}$ can be found to solve (V.C.18), the optimum choice of c_t is immediate from (V.C.16). Examination of (V.C.16) reveals that the role of c_t is to adjust the mean of $\hat{X}_t$ to equal that of X_t, so that an optimum linear estimate will always be of the form

$$\hat{X}_t = E\{X_t\} + \sum_{n=a}^{b} h_{t,n}(Y_n - E\{Y_n\}). \tag{V.C.19}$$

Thus for the purpose of discussion we can, without loss of generality, assume that the means of $\{X_n\}_{n=-\infty}^{\infty}$ and $\{Y_n\}_{n=-\infty}^{\infty}$ are zero, which we henceforth do. With this assumption we always have $c_t = 0$ and (V.C.16) is unnecessary.

Equation (V.C.18) is thus the key equation determining the optimum linear estimator. This equation is known as the *Wiener-Hopf equation.* For finite a and b, this equation is quite easy to solve in principle. In particular, we note that (V.C.18) is a set of $(b-a+1)$ linear equations in $(b-a+1)$ unknowns. This can be rewritten in matrix form as

$$\underline{\sigma}_{XY}(t) = \mathbf{\Sigma}_Y \underline{h}_t, \tag{V.C.20}$$

where $\underline{\sigma}_{XY}(t) \triangleq [C_{XY}(t,a), \ldots, C_{XY}(t,b)]^T$, $\underline{h}_t \triangleq (h_{t,a}, \ldots, h_{t,b})^T$, and $\mathbf{\Sigma}_Y$ is the covariance matrix of the vector $(Y_a, \ldots, Y_b)^T$. Assuming that $\mathbf{\Sigma}_Y$ is positive definite,[6] we see from (V.C.20) that the optimum estimator coefficients are given by

$$\underline{h}_t = \mathbf{\Sigma}_Y^{-1} \underline{\sigma}_{XY}(t). \tag{V.C.21}$$

Thus for finite a and b, the MMSE estimation problem is, in principle, solved. In practice, however, the determination of these coefficients sometimes presents computational difficulties because of the inversion of the

[6]Since $\mathbf{\Sigma}_Y$ is a covariance matrix it must be at least nonnegative definite. If it is not strictly positive definite, then this implies that there are redundant observations as noted in Section III.B. Even in this case, however, (V.C.20) has a solution, although not a unique one.

matrix $\mathbf{\Sigma}_Y$. In general, inversion of a $k \times k$ matrix requires a number of basic computational operations of the order of the k^3. In our case k equals the number of observations which, for many signal estimation applications, grows linearly with time. So, in general, the computation of optimum coefficients from (V.C.21) cannot be accomplished in real time. For this reason the study of linear signal estimation is dominated by the investigation of particular models that allow for more efficient computation of these coefficients. One such model is the Kalman-Bucy model of Section V.B, which we discuss further below. Two other important models of this type are the Levinson model, which yields an efficient computational algorithm for the optimum filter coefficients, and the Wiener-Kolmogorov model, which essentially overcomes this problem by allowing a to be $-\infty$. The Levinson model is discussed in the following example, and the Wiener-Kolmogorov model is discussed in Section V.D.

Example V.C.1: Levinson Filtering

Levinson filtering is concerned with one-step prediction of a random sequence whose second-order statistics are stationary in time. In particular, in this model the autocovariance function of our observation sequence $\{Y_n\}_{n=-\infty}^{\infty}$ is assumed to satisfy the condition

$$C_Y(n,l) = C_Y(n-l, 0) \tag{V.C.22}$$

for all integers n and l. Such a sequence is said to be *covariance stationary* or *wide-sense stationary* (w.s.s.), and for convenience we usually write the autocovariance function of a w.s.s. sequence as a function of single variable, the time difference, by suppressing the 0 in the second argument on the right-hand side of (V.C.22); i.e.,

$$C_Y(n,l) \equiv C_Y(n-l). \tag{V.C.23}$$

Note that, since $\text{Cov}(Y_n, Y_l) = \text{Cov}(Y_l, Y_n)$, the function C_Y is symmetric: $C_Y(n-l) = C_Y(l-n)$.

In the Levinson filtering problem we observe Y_n for $0 \le n \le t$ and we wish to estimate Y_{t+1}. In our previous notation we have $a = 0, b = t$, and $X_t = Y_{t+1}$. The cross-covariance function of $\{X_n\}_{n=-\infty}^{\infty} \equiv \{Y_{n+1}\}_{n=-\infty}^{\infty}$ and $\{Y_n\}_{n=-\infty}^{\infty}$ is thus

$$\begin{aligned} C_{XY}(t,l) &= \text{Cov}(X_t, Y_l) \\ &= \text{Cov}(Y_{t+1}, Y_l) \\ &= C_Y(t+1-l), \end{aligned} \tag{V.C.24}$$

and, of course, the $n-l$th element of $\mathbf{\Sigma}_Y$ is $C_Y(n-l)$. Thus the Wiener-Hopf equation (V.C.20) becomes

$$\begin{pmatrix} C_Y(t+1) \\ C_Y(t) \\ \cdot \\ \cdot \\ \cdot \\ C_Y(1) \end{pmatrix} = \begin{pmatrix} C_Y(0) & C_Y(1) & \cdots & C_Y(t) \\ C_Y(1) & C_Y(0) & \cdots & C_Y(t-1) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ & & & C_Y(1) \\ C_Y(t) & \cdots & C_Y(1) & C_Y(0) \end{pmatrix} \begin{pmatrix} h_{t,0} \\ h_{t,1} \\ \cdot \\ \cdot \\ \cdot \\ h_{t,t} \end{pmatrix} \tag{V.C.25}$$

a set of equations sometimes known as the *Yule-Walker equations.*

Because $\{Y_n\}_{n=-\infty}^{\infty}$ is w.s.s., the matrix $\mathbf{\Sigma}_Y$ on the right-hand side of (V.C.25) is a *Toeplitz* matrix, which means that its entries are constant along the diagonals [since $C_Y(n,l) = C_Y(n-l)$]. Unlike general covariance matrices, $k \times k$ Toeplitz matrices can be inverted in a number of operations that is of the order of k^2. [A well-known algorithm for doing this is due to Trench (1964).] Thus for any linear MMSE problem in which the observations are w.s.s., the complexity of computing the estimator coefficients is reduced by a factor equal to the number of observations. However, in the Levinson problem there is additional structure that allows for a further simplification in computing the estimator coefficients. In particular, the vector $\underline{\sigma}_{XY}(t)$ on the left of (V.C.25) is like the first row of $\mathbf{\Sigma}_Y$ shifted by one time unit. This structure allows the coefficients to be computed recursively in t.

It is conventional in this problem to rewrite the predictor $\hat{Y}_{t+1} = \sum_{n=0}^{t} h_{t,n}Y_n$ as $\hat{Y}_{t+1} = -\sum_{n=0}^{t} a_{t+1,t+1-n}Y_n$. The coefficients $a_{t,1}, \ldots, a_{t,t}$ can then be updated recursively (in t) through the following algorithm, known as the *Levinson algorithm:*

$$a_{t+1,k} = a_{t,k} - k_t a_{t,t+1-k}, \quad k = 1, \ldots, t, \tag{V.C.26}$$

and

$$a_{t+1,t+1} = -k_t, \tag{V.C.27}$$

where k_t is generated recursively with $\epsilon_t \triangleq E\{(Y_t - \hat{Y}_t)^2\}$ via

$$\epsilon_{t+1} = (1 - k_t^2)\epsilon_t \tag{V.C.28}$$

and

$$k_t = \frac{C_Y(t+1) + \sum_{k=1}^{t} a_{t,k} C_Y(t+1-k)}{\epsilon_t}. \tag{V.C.29}$$

This algorithm is initialized by $k_0 = -C_Y(1)/C_Y(0)$ and $\epsilon_0 = C_Y(0)$. Note that this algorithm computes the MMSE, ϵ_t, as a by-product. It also computes the coefficients $k_1, k_2, \ldots$, which are known as *partial correlation* (PARCOR) *coefficients* or *reflection coefficients.* The latter coefficients

are useful for implementing the one-step predictor using a lattice filter. A derivation of the Levinson algorithm is found by Honig and Messerschmitt (1984), together with a discussion of the implementation and several applications of one-step predictors.

The linear estimation problem can be extended straightforwardly to the case in which the observation sequence is a sequence of vectors (say k-dimensional) and the quantity to be estimated is also a vector (say, m-dimensional). In this case, we consider estimates of the form

$$\hat{\underline{X}}_t = \sum_{n=a}^{b} \mathbf{H}_{t,n}\underline{Y}_n + \underline{c}_t, \tag{V.C.30}$$

where $\{\mathbf{H}_{t,n}\}_{n=a}^{b}$ is a sequence of $m \times k$ matrices and $\underline{c}_t \in \mathbb{R}^m$. With a or b infinite (V.C.30) is defined in the mean-norm sense; e.g., with $a = -\infty$ and b finite,

$$\lim_{j \to -\infty} E\left\{ \left\| \sum_{n=j}^{b} \mathbf{H}_{t,n}\underline{Y}_n + \underline{c}_t - \hat{\underline{X}}_t \right\|^2 \right\} = 0,$$

where $\| \underline{x} \|^2 \triangleq \underline{x}^T\underline{x}$. If we wish to choose an estimate to solve

$$\min_{\hat{\underline{X}} \in \mathcal{H}_a^b} E\{\| \underline{X}_t - \hat{\underline{X}}_t \|^2\},$$

where $\mathcal{H}_a^b$ is the set of all estimators of the form (V.C.30), it follows similarly to Proposition V.C.2 that $\hat{\underline{X}}_t$ is optimum if and only if

$$E\{(\underline{X}_t - \hat{\underline{X}}_t)^T \underline{Z}\} = 0, \text{ for all } \underline{Z} \in \mathcal{H}_a^b. \tag{V.C.31}$$

Equation (V.C.31) can be transformed into the equivalent conditions

$$E\{\hat{\underline{X}}_t\} = E\{\underline{X}_t\} \tag{V.C.32a}$$

and

$$E\{(\underline{X}_t - \hat{\underline{X}}_t)\underline{Y}_l^T\} = \mathbf{0}, \quad a \le l \le b, \tag{V.C.32b}$$

where $\mathbf{0}$ denotes the matrix of all zeroes. These equations in turn give an equation for the optimum $\underline{c}_t$ and a vector Wiener-Hopf equation:

$$\mathbf{C}_{XY}(t,l) = \sum_{n=a}^{b} \mathbf{H}_{t,n}\mathbf{C}_Y(n,l), \quad a \le l \le b, \tag{V.C.33}$$

where $\mathbf{C}_{XY}(t,l) \triangleq \mathrm{Cov}(\underline{X}_t, \underline{Y}_l)$ is the (matrix) cross-covariance function of $\{\underline{X}_n\}_{n=-\infty}^{\infty}$ and $\{\underline{Y}_n\}_{n=-\infty}^{\infty}$ and similarly $\mathbf{C}_Y(n,l) \triangleq \mathrm{Cov}(\underline{Y}_n, \underline{Y}_l)$. Note that $\mathbf{C}_{XY}(t,l)$ and $\mathbf{C}_Y(n,l)$ are $m \times k$ and $k \times k$ matrices, respectively.

For finite a and b, the vector Wiener-Hopf equation (V.C.33) gives a set of $(b-a+1) \times m \times k$ linear equations in the same number of unknowns. It can thus be solved by matrix inversion subject to positive definiteness of the covariance matrix of the $(b-a+1)$ k-dimensional vector $(\underline{Y}_a^T, \underline{Y}_{a+1}^T, \ldots, \underline{Y}_b^T)^T$. In fact, the minimization of the mean norm error $E\{\| \underline{X}_t - \hat{\underline{X}}_t \|^2\}$ is equivalent to minimizing the mean-square error on each component of $\underline{X}_t$. So the vector Wiener-Hopf equation is essentially a set of m scalar Wiener-Hopf equations, each with $(b-a+1)k$ observations. Unfortunately, this structure does not simplify the solution since it is the observation dimension that affects the computational burden most. As in the scalar case, computational issues are often dominant in the study of these problems. The Levinson problem can be formulated in the vector case as well as in the scalar case, with an efficient solution algorithm similar to that of Example V.C.1. Moreover, in addition to its role as a global MMSE estimator in the linear-Gaussian model of Section V.B, the Kalman-Bucy filter can also be interpreted as a linear MMSE estimator in a less restrictive model. This result is summarized in the following example.

Example V.C.2: The Kalman-Bucy Filter as a Linear MMSE Estimator

Consider the linear stochastic system model

$$\underline{X}_{n+1} = \mathbf{F}_n \underline{X}_n + \mathbf{G}_n \underline{U}_n, \quad n = 0, 1, \ldots \tag{V.C.34a}$$

$$\underline{Y}_n = \mathbf{H}_n \underline{X}_n + \underline{V}_n, \quad n = 0, 1, \ldots, \tag{V.C.34b}$$

where, for each $n \geq 0$, $\underline{X}_n, \underline{U}_n, \underline{Y}_n$, and $\underline{V}_n$ are random vectors of dimension m, s, k, and k, respectively, and $\mathbf{F}_n$, $\mathbf{G}_n$, and $\mathbf{H}_n$ are matrices of appropriate dimensions. We assume that $\{\underline{U}_n\}_{n=-\infty}^{\infty}$ and $\{\underline{V}_n\}_{n=-\infty}^{\infty}$ are uncorrelated sequences of zero-mean uncorrelated random vectors [i.e., $\mathrm{Cov}(\underline{V}_n, \underline{U}_l) = \mathbf{O}$ for all n and l and $\mathrm{Cov}(\underline{U}_n, \underline{U}_l) = \mathrm{Cov}(\underline{V}_n, \underline{V}_l) = \mathbf{O}$ for all $n \neq l$, where $\mathbf{O}$ denotes a matrix of all zeros], and that the initial condition $\underline{X}_0$ is uncorrelated with both $\{\underline{U}_n\}_{n=-\infty}^{\infty}$ and $\{\underline{V}_n\}_{n=-\infty}^{\infty}$. We also assume that $\underline{U}_n$ and $\underline{V}_n$ have known covariance matrices $\mathbf{Q}_n$ and $\mathbf{R}_n$, respectively, for each n, and that $\underline{X}_0$ has known mean $\underline{m}_0$ and covariance matrix $\mathbf{\Sigma}_0$. Apart from these assumptions, the statistics of the various random quantities are arbitrary (e.g, no Gaussian assumption is made here).

Within the assumptions above it can be shown that the Kalman-Bucy filtering recursions of Proposition V.B.1 give the linear minimum-mean-

norm-error estimators of $\underline{X}_t$ and $\underline{X}_{t+1}$ from the measurements $\underline{Y}_0^t$. Although we will not develop this result in detail here,[7] the application of the orthogonality principle can be illustrated in deriving the estimator time update, $\hat{\underline{X}}_{t+1|t} = \mathbf{F}_t \hat{\underline{X}}_{t|t}$. In particular, suppose that $\hat{\underline{X}}_{t|t}$ is the best linear estimator of $\underline{X}_t$ given $\underline{Y}_0^t$ and consider the quantity

$$E\{(\underline{X}_{t+1} - \mathbf{F}_t \hat{\underline{X}}_{t|t})^T \underline{Z}\} \tag{V.C.35}$$

for $\underline{Z} \in \mathcal{H}_0^t$. Using the state equation (V.C.34a), (V.C.35) becomes

$$\begin{aligned} E\{(\mathbf{F}_t \underline{X}_t \; &+ \; \mathbf{G}_t \underline{U}_t - \mathbf{F}_t \hat{\underline{X}}_{t|t})^T \underline{Z}\} \\ &= \; E\{(\underline{X}_t - \hat{\underline{X}}_{t|t})^T \mathbf{F}_t^T \underline{Z}\} + E\{\underline{U}^T \underline{Z}\} \mathbf{G}_t^T. \end{aligned} \tag{V.C.36}$$

Since $\hat{\underline{X}}_{t|t}$ is assumed to be the best linear estimator of $\underline{X}_t$ from $\underline{Y}_0^t$ and since $\mathbf{F}_t^T \underline{Z} \in \mathcal{H}_0^t$ whenever $\underline{Z} \in \mathcal{H}_0^t$, the first term on the right-hand side of (V.C.36) is zero for any $\underline{Z} \in \mathcal{H}_0^t$. With regard to the second term on the right of (V.C.36), we note that $\underline{Z}$ is a linear transformation of $\underline{Y}_0^t$, which in turn is a linear transformation of $\underline{X}_0, \underline{U}_0^{t-1}$, and $\underline{V}_0^t$, all of which are uncorrelated with $\underline{U}_t$. Thus $\underline{U}_t$ and $\underline{Z}$ are uncorrelated and we have that $E\{(\underline{X}_{t+1} - \mathbf{F}_t \hat{\underline{X}}_{t|t})^T \underline{Z}\} = 0$ for all $\underline{Z} \in \mathcal{H}_0^t$, implying from the orthogonality principle that $\mathbf{F}_t \hat{\underline{X}}_{t|t}$ is the best linear estimator of $\underline{X}_{t+1}$ given $\underline{Y}_0^t$. The proof of the covariance time update is almost identical to that given in Proposition V.B.1 for the Gaussian case.

V.D Wiener-Kolmogorov Filtering

In Section V.C we derived the Wiener-Hopf equation, which specifies the coefficients for optimum linear estimation of one random variable, X_t, from observation of a set of other random variables, $Y_a, \ldots, Y_b$. In most signal estimation applications the number of observations $(b-a+1)$ grows linearly with t, so further assumptions are usually needed in order to compute coefficients of the corresponding estimator efficiently. Two such sets of assumptions are those made by the Levinson and Kalman-Bucy filtering models. Another set of simplifying assumptions, known as the Wiener-Kolmogorov model, leads to the solution of the optimum linear estimation problem for a wide class of signal estimation applications. In this section we develop the latter model in some detail.

[7] Actually, this result can be inferred from the optimality of the Kalman-Bucy filter in the Gaussian case and its linearity. In particular, the model above includes the Gaussian model as a special case. Thus if some other structure were the best linear estimator for this model, then the Kalman-Bucy filter could not be globally optimum for the particular case of Gaussian statistics.

As in the Levinson problem, we assume that the (scalar) observation sequence is wide-sense stationary; i.e., $C_Y(n,l) \triangleq \text{Cov}(Y_n, Y_l) = C_Y(n-l,0) \equiv C_Y(n-l)$ for all integers n and l. We also assume that the observation sequence and the (scalar) sequence $\{X_n\}_{n=-\infty}^{\infty}$ are *jointly wide-sense stationary*; i.e., we assume that $C_{XY}(t,n) \triangleq \text{Cov}(X_t, Y_n) = C_{XY}(t-n,0) \equiv C_{XY}(t-n)$ for all integers t and n. (We continue to assume, without loss of generality, that all X_n's and Y_n's have zero means.) We also assume that the number of observations is infinite, and we will consider two such cases: the so-called *noncausal Wiener-Kolmogorov filtering problem*, in which we take $a = -\infty$ and $b = +\infty$; and the *causal Wiener-Kolmogorov filtering problem*, in which $a = -\infty$ and $b = t$. We treat the noncausal case first because its solution is simpler.

V.D.1 Noncausal Wiener-Kolmogorov Filtering

The noncausal Wiener-Kolmogorov problem is so called because we are estimating at time t based on observations for all times, $-\infty < t < \infty$. Thus the estimate

$$\hat{X}_t = \sum_{n=-\infty}^{\infty} h_{t,n} Y_n, \tag{V.D.1}$$

if thought of as a linear filtering operation on the sequence $\{Y_n\}_{n=-\infty}^{\infty}$, is not necessarily causal; that is, the impulse response $\{h_{t,n}\}_{n=-\infty}^{\infty}$ may not satisfy $h_{t,n} = 0$ for $t > n$. This implies that the estimate at the "present" time t may depend on observations at future times $n > t$. Obviously, for real-time estimation problems one should restrict attention to causal filters; however, for applications in which the data have been stored or in which the index t is a spatial parameter rather than a time parameter (as in image or array processing), this type of causality is not an issue.

The Wiener-Hopf equation (V.C.18) for this problem is

$$C_{XY}(t,l) = \sum_{n=-\infty}^{\infty} h_{t,n} C_Y(n,l), \quad -\infty < l < \infty, \tag{V.D.2}$$

which, from the stationarity assumptions, can be written as

$$C_{XY}(t-l) = \sum_{n=-\infty}^{\infty} h_{t,n} C_Y(n-l), \quad -\infty < l < \infty. \tag{V.D.3}$$

To put (V.D.3) in a more tractable form, let us define a new variable, $\tau = t - l$, from which we have

$$C_{XY}(\tau) = \sum_{n=-\infty}^{\infty} h_{t,n} C_Y(n+\tau-t), \quad -\infty < \tau < \infty. \tag{V.D.4}$$

Now, changing variables in the sum with the substitution $\alpha = t - n$, the Wiener-Hopf equation becomes

$$C_{XY}(\tau) = \sum_{\alpha=-\infty}^{\infty} h_{t,t-\alpha} C_Y(\alpha-\tau), \quad -\infty < \tau < \infty. \tag{V.D.5}$$

Note that the variable t appears in (V.D.5) only in the coefficient sequence $\{h_{t,t-\alpha}\}_{\alpha=-\infty}^{\infty}$. This implies that if the Wiener-Hopf equation has a solution in this case, we can choose that solution independently of t. That is, an optimum $\{h_{t,n}\}_{n=-\infty}^{\infty}$ can be chosen such that $h_{t,t-\alpha}$ depends only on α, or equivalently, that $h_{t,t-\alpha} = h_{\alpha,0}$ for all integers t and α. Thus if a solution exists, it can be chosen to be *time-invariant* (or *shift-invariant*) with coefficient sequence $h_{t,n} = h_{t-n,0} \triangleq h_{t-n}$, where for convenience we suppress the second index. With this observation, and noting that $C_Y(\alpha-\tau) = C_Y(\tau-\alpha)$, the Wiener-Hopf equation becomes

$$C_{XY}(\tau) = \sum_{\alpha=-\infty}^{\infty} h_\alpha C_Y(\tau-\alpha), \quad -\infty < \tau < \infty. \tag{V.D.6}$$

The right-hand side of (V.D.6) is recognized as the discrete-time *convolution* of the sequences $\{h_n\}_{n=-\infty}^{\infty}$ and $\{C_Y(n)\}_{n=-\infty}^{\infty}$. Thus (V.D.6) is a convolution equation, which can be converted to a simple algebraic equation by converting to the frequency domain. In particular, on assuming that the following discrete-time Fourier transforms exist:

$$H(\omega) \triangleq \sum_{n=-\infty}^{\infty} h_n e^{-i\omega n}, \quad -\pi \le \omega \le \pi, \tag{V.D.7}$$

$$\phi_{XY}(\omega) \triangleq \sum_{n=-\infty}^{\infty} C_{XY}(n) e^{-i\omega n}, \quad -\pi \le \omega \le \pi, \tag{V.D.8}$$

and

$$\phi_Y(\omega) \triangleq \sum_{n=-\infty}^{\infty} C_Y(n) e^{-i\omega n}, \quad -\pi \le \omega \le \pi, \tag{V.D.9}$$

the Wiener-Hopf equation becomes

$$\phi_{XY}(\omega) = H(\omega)\phi_Y(\omega), \quad -\pi \le \omega \le \pi. \tag{V.D.10}$$

Note that H is the *transfer function* of the filter $\{h_n\}_{n=-\infty}^{\infty}, \phi_Y$ is the *power spectral density* (or spectrum) of the sequence$\{Y_n\}_{n=-\infty}^{\infty}$, and ϕ_{XY} is the *cross power spectral density* of the sequences $\{X_n\}_{n=-\infty}^{\infty}$ and $\{Y_n\}_{n=-\infty}^{\infty}$.

In the form (V.D.10), the Wiener-Hopf equation is easily solved for the transfer function of the optimum estimator, i.e.,

$$H(\omega) = \frac{\phi_{XY}(\omega)}{\phi_Y(\omega)}, \quad -\pi \le \omega \le \pi, \tag{V.D.11}$$

from which the (time-invariant) filter coefficients become,[8]

$$h_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\phi_{XY}(\omega)}{\phi_Y(\omega)} e^{i\omega n} d\omega, \quad n \in \mathbb{Z}. \tag{V.D.12}$$

Within the assumptions made above (V.D.11) or (V.D.12) specifies the optimum linear estimator in the noncausal Wiener-Kolmogorov problem. Before giving a specific example to illustrate this result, it is of interest to consider the value of the mean squared error incurred by using this filter. In particular, we would like an expression for the minimum value of the MSE,

$$\text{MMSE} \triangleq \min_{\hat{X}_t \in \mathcal{H}_{-\infty}^{\infty}} E\{(X_t - \hat{X}_t)^2\}.$$

We have that

$$\text{MMSE} = E\{(X_t - \hat{X}_t)^2\},$$

where $\hat{X}_t$ is the optimum estimate. We can write

$$\begin{aligned} \text{MMSE} &= E\{(X_t - \hat{X}_t)^2\} \\ &= E\{(X_t - \hat{X}_t)X_t\} - E\{(X_t - \hat{X}_t)\hat{X}_t\} \\ &= E\{(X_t - \hat{X}_t)X_t\} = E\{X_t^2\} - E\{\hat{X}_t X_t\}, \end{aligned} \tag{V.D.13}$$

where the disappearance of the term $E\{(X_t - \hat{X}_t)\hat{X}_t\}$ is due to the orthogonality principle.

Consider first the second term on the right-hand side of (V.D.13). We have

$$\begin{aligned} E\{\hat{X}_t X_t\} &= E\left\{\left(\sum_{n=-\infty}^{\infty} h_{t-n} Y_n\right) X_t\right\} = \sum_{n=-\infty}^{\infty} h_{t-n} E\{Y_n X_t\} \\ &= \sum_{n=-\infty}^{\infty} h_{t-n} C_{XY}(t-n) = \sum_{\alpha=-\infty}^{\infty} h_\alpha C_{XY}(\alpha), \end{aligned} \tag{V.D.14}$$

where we have made the substitution $\alpha = t - n$. Noting that the right-hand side of (V.D.14) is the zeroth term of the convolution of $\{h_n\}_{n=-\infty}^{\infty}$ and

[8]Here, and elsewhere in this book, $\mathbb{Z}$ denotes the set of all integers.

$\{C_{XY}(-n)\}_{n=-\infty}^{\infty}$, we have that

$$E\{\hat{X}_t X_t\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(\omega)\bar{\phi}_{XY}(\omega) d\omega, \tag{V.D.15}$$

where $\bar{\phi}_{XY}(\omega)$ is the discrete-time Fourier transform of the sequence $\{C_{XY}(-n)\}_{n=-\infty}^{\infty}$; i.e.,

$$\bar{\phi}_{XY}(\omega) = \sum_{n=-\infty}^{\infty} C_{XY}(-n) e^{-i\omega n}, \quad -\pi \le \omega \le \pi. \tag{V.D.16}$$

Setting $\alpha = -n$ in (V.D.16), we have

$$\bar{\phi}_{XY}(\omega) = \sum_{\alpha=-\infty}^{\infty} C_{XY}(\alpha) e^{i\omega a} = \phi_{XY}^{\star}(\omega), \tag{V.D.17}$$

where the superscript $\star$ denotes complex conjugation. From (V.D.11) and (V.D.15), we thus have

$$\begin{aligned} E\{\hat{X}_t X_t\} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\phi_{XY}(\omega)}{\phi_Y(\omega)} \phi_{XY}^{\star}(\omega) d\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|\phi_{XY}(\omega)|^2}{\phi_Y(\omega)} d\omega. \end{aligned} \tag{V.D.18}$$

Assuming that the sequence $\{X_n\}_{n=-\infty}^{\infty}$ is w.s.s. with power spectrum ϕ_X, we can write the first term on the right-hand side of (V.D.13) as

$$E\{X_t^2\} = C_X(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi_X(\omega) d\omega,$$

from which we have

$$\text{MMSE} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left[\phi_X(\omega) - \frac{|\phi_{XY}(\omega)|^2}{\phi_Y(\omega)}\right] d\omega. \tag{V.D.19}$$

Equation (V.D.19) can be rewritten as

$$\text{MMSE} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left[1 - \frac{|\phi_{XY}(\omega)|^2}{\phi_X(\omega)\phi_Y(\omega)}\right] \phi_X(\omega) d\omega, \tag{V.D.20}$$

so that the performance of the optimum filter is the integral of the function $[1 - |\phi_{XY}|^2/\phi_X\phi_Y]$ weighted by ϕ_X. A property of the cross-spectrum is that $|\phi_{XY}(\omega)|^2 \le \phi_X(\omega)\phi_Y(\omega)$, $-\pi \le \omega \le \pi$, with equality for all $\omega \in [-\pi, \pi]$ if and only if the sequence $\{X_n\}_{n=-\infty}^{\infty}$ and $\{Y_n\}_{n=-\infty}^{\infty}$ are perfectly correlated (i.e., $X_t \in \mathcal{H}_{-\infty}^{\infty}$ for all $t \in \mathbb{Z}$). Thus (V.D.20) shows that the MMSE ranges from $E\{X_t^2\}$ to zero as the relationship between the sequences $\{X_n\}_{n=-\infty}^{\infty}$

and $\{Y_n\}_{n=-\infty}^{\infty}$ ranges from complete uncorrelatedness $[\phi_{XY}(\omega) = 0, -\pi \leq \omega \leq \pi]$ to perfect correlation. Note that in the first of these two extremes we have $\hat{X}_t = E\{X_t\} \equiv 0$, and in the latter one we have $\hat{X}_t = X_t$. The interesting cases, of course, are between these two extremes, in which case $0 < \text{MMSE} < E\{X_t^2\}$.

To illustrate noncausal Wiener-Kolmogorov filtering, we consider the following example.

Example V.D.1: Signal Estimation in Additive Noise

Consider the observation model

$$Y_n = S_n + N_n, \quad n \in \mathbb{Z}, \tag{V.D.21}$$

where $\{S_n\}_{n=-\infty}^{\infty}$ and $\{N_n\}_{n=-\infty}^{\infty}$ are uncorrelated, zero-mean, w.s.s sequences representing signal and noise, respectively. Suppose that the quantity we wish to estimate at time t is the signal at time $t+\lambda$ for some integer λ; i.e.,

$$X_t = S_{t+\lambda}. \tag{V.D.22}$$

This problem represents filtering ($\lambda = 0$), prediction ($\lambda > 0$), or smoothing ($\lambda < 0$) of the signal. Denoting the power spectra of signal and noise by ϕ_S and ϕ_N, respectively, it is straightforward to show that

$$\phi_Y(\omega) = \phi_S(\omega) + \phi_N(\omega), \quad -\pi \leq \omega \leq \pi, \tag{V.D.23}$$

$$\phi_{XY}(\omega) = e^{i\omega\lambda}\phi_S(\omega), \quad -\pi \leq \omega \leq \pi, \tag{V.D.24}$$

and

$$\phi_X(\omega) = \phi_S(\omega), \quad -\pi \leq \omega \leq \pi. \tag{V.D.25}$$

From (V.D.11), (V.D.23), and (V.D.24) we see that the transfer function of the optimum noncausal filter is

$$H(\omega) = \frac{e^{i\omega\lambda}\phi_S(\omega)}{\phi_S(\omega) + \phi_N(\omega)}, \quad -\pi \leq \omega \leq \pi. \tag{V.D.26}$$

The interpretation of this filter is straightforward. The term $e^{i\omega\lambda}$ is a (unit-magnitude) phase term that corresponds to a shift of λ time units in the time domain. Thus this term merely time-shifts the data sequence to account for the fact that we wish to estimate the signal at time $t+\lambda$. Note that this shifting is causal if $\lambda \leq 0$ and is noncausal if $\lambda > 0$. The remaining term in the filter transfer function is its magnitude, $\phi_S/(\phi_S + \phi_N)$, which represents the gain of the filter. We can rewrite this term as

$$|H(\omega)| = \frac{\phi_S(\omega)/\phi_N(\omega)}{\phi_S(\omega)/\phi_N(\omega) + 1}, \quad -\pi \leq \omega \leq \pi. \tag{V.D.27}$$

Note that this term ranges from zero to unity as the ratio $\phi_S(\omega)/\phi_N(\omega)$ ranges from zero to infinity. (Note that power spectra are real nonnegative functions.) In particular, if $\phi_S(\omega)/\phi_N(\omega) \ll 1$, then $|H(\omega)| \cong 0$, and if $\phi_S(\omega)/\phi_N(\omega) \gg 1$, then $|H(\omega)| \cong 1$. The quantity $\phi_S(\omega)/\phi_N(\omega)$ can be interpreted as a measure of the signal-to-noise power ratio at frequency ω. Thus we see that if the noise is dominant at a given frequency, then the filter gain at that frequency is essentially zero, and if the signal is dominant, the gain is essentially unity. Between these extremes the gain is chosen to balance the effect of distorting the signal (caused by less than unity gain) and the effect of allowing noise to pass through the filter (caused by greater than zero gain).

From (V.D.20) and (V.D.23) through (V.D.25) the performance of the noncausal Wiener-Kolmogorov filter is given in this case by

$$\text{MMSE} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\phi_S(\omega)\phi_N(\omega)}{\phi_S(\omega) + \phi_N(\omega)} d\omega. \tag{V.D.28}$$

(Note that the shift λ is irrelevant to performance, as it should be for this noncausal case.) Since $\phi_S(\omega)/[\phi_S(\omega) + \phi_N(\omega)] \leq 1$ and $\phi_N(\omega)/[\phi_S(\omega) + \phi_N(\omega)] \leq 1$, (V.D.28) implies that the MMSE is never larger than the minimum of the average signal power $[(\frac{1}{2\pi})\int_{-\pi}^{\pi}\phi_S(\omega)d\omega]$ and the average noise power $[(\frac{1}{2\pi})\int_{-\pi}^{\pi}\phi_N(\omega)d\omega]$; i.e.,

$$\text{MMSE} \leq \min\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi}\phi_S(\omega)d\omega, \frac{1}{2\pi}\int_{-\pi}^{\pi}\phi_N(\omega)d\omega\right\}. \tag{V.D.29}$$

It achieves the first of these quantities in the limit as the ratio $\phi_S(\omega)/\phi_N(\omega)$ approaches zero uniformly in $[-\pi, \pi]$, in which case the optimum filter becomes a no-pass $[H(\omega) \cong 0]$ filter; and it achieves the second of these quantities in the limit as $\phi_S(\omega)/\phi_N(\omega)$ increases without bound uniformly in $[-\pi, \pi]$, in which case the optimum filter becomes an all-pass $[H(\omega) \cong 1]$ filter. Equation (V.D.28) also indicates that the MMSE is zero if and only if $\phi_S(\omega)\phi_N(\omega) = 0$ for almost all $\omega \in [-\pi, \pi]$, a condition that holds, for example, when the signal and noise occupy different parts of the frequency band.

V.D.2 Causal Wiener-Kolmogorov Filtering

As noted previously, noncausal linear estimators are not suitable for applications in which real-time estimates are desired. Since many applications do require real-time estimates, it is thus of interest to consider the causal Wiener-Kolmogorov problem, in which we wish to estimate X_t based on observation of the sequence $\{Y_n\}_{n=-\infty}^{\infty}$ only up to time t. This corresponds to the case $a = -\infty$ and $b = t$ in the notation employed previously.

To develop the solution to this problem, we first note that the set $\mathcal{H}_{-\infty}^{t}$ is a subset of $\mathcal{H}_{-\infty}^{\infty}$. Thus if the solution to the noncausal Wiener-Kolmogorov problem happens to be causal, it also solves the causal Wiener-Kolmogorov problem. Unfortunately, except under very special circumstances, the solution to the noncausal problem will in fact be strictly noncausal. However, there is a very definite relationship between the solutions to the causal and noncausal Wiener-Kolmogorov problems. In particular, let $\hat{X}_t$ and $\tilde{X}_t$ denote, respectively, the causal and noncausal Wiener-Kolmogorov estimates for a given model. Since $(X_t - \hat{X}_t) = (\tilde{X}_t - \hat{X}_t) + (X_t - \tilde{X}_t)$, we have for any $Z \in \mathcal{H}_{-\infty}^{t}$ that

$$\begin{aligned} 0 &= E\{(X_t - \hat{X}_t)Z\} \\ &= E\{(\tilde{X}_t - \hat{X}_t)Z\} + E\{(X_t - \tilde{X}_t)Z\}. \end{aligned} \quad \text{(V.D.30)}$$

Note that the term $E\{(X_t - \tilde{X}_t)Z\}$ is zero due to the orthogonality principle applied to $\tilde{X}_t$ and to the fact that $Z \in \mathcal{H}_{-\infty}^{\infty}$ (since $\mathcal{H}_{-\infty}^{t} \subset \mathcal{H}_{-\infty}^{\infty}$). Thus we have from (V.D.30) that

$$E\{(\tilde{X}_t - \hat{X}_t)Z\} = 0, \text{ for all } Z \in \mathcal{H}_{-\infty}^{t}. \quad \text{(V.D.31)}$$

Equation (V.D.31) and the orthogonality principle imply that $\hat{X}_t$ is the MMSE estimate of $\tilde{X}_t$ among all estimates in $\mathcal{H}_{-\infty}^{t}$. In other words, $\hat{X}_t$, which is the projection of X_t onto $\mathcal{H}_{-\infty}^{t}$, can be obtained by first projecting X_t onto $\mathcal{H}_{-\infty}^{\infty}$ to get $\tilde{X}_t$ and then projecting $\tilde{X}_t$ onto $\mathcal{H}_{-\infty}^{t}$. A geometric analogy to this fact can be seen by considering the space $\mathbb{R}^3$ with the standard orthogonal axes labeled x, y, and z. To find the x-projection of a vector in $\mathbb{R}^3$, we can first project the vector onto the x-y plane (or the x-z plane) and then project the result onto the x-axis. This works because the x-axis is a subset of the x-y (or x-z) plane.

In view of the above, a good starting point in seeking a causal MMSE estimator is to consider first the noncausal MMSE estimator

$$\tilde{X}_t = \sum_{n=-\infty}^{\infty} \tilde{h}_{t-n} Y_n \quad \text{(V.D.32)}$$

with $\{\tilde{h}_n\}_{n=-\infty}^{\infty}$ given by (V.D.12). Since we would like to project $\tilde{X}_t$ onto the set of linear estimates generated by $\{Y_n\}_{n=-\infty}^{t}$, we might be tempted to try a simple truncation of (V.D.32); i.e., we might consider the estimate

$$\bar{X}_t \triangleq \sum_{n=-\infty}^{t} \tilde{h}_{t-n} Y_n. \quad \text{(V.D.33)}$$

If $\bar{X}_t$ is the projection of $\tilde{X}_t$ onto $\mathcal{H}_{-\infty}^{t}$, then the error, $\tilde{X}_t - \bar{X}_t = \sum_{n=t+1}^{\infty} \tilde{h}_{t-n} Y_n$, must be orthogonal to Y_m for all $m \leq t$. However, since the sequences $\{Y_n\}_{n=t+1}^{\infty}$ and $\{Y_n\}_{n=-\infty}^{t}$ are usually correlated, and since the coefficient sequence $\{\tilde{h}_n\}_{n=-\infty}^{\infty}$ is chosen to satisfy a different orthogonality condition, it is unlikely that $(\tilde{X}_t - \bar{X}_t)$ will be orthogonal to $\mathcal{H}_{-\infty}^{t}$ for the general case. However, one situation in which $\bar{X}_t$ of (V.D.33) will satisfy the required orthogonality condition is when $\{Y_n\}_{n=-\infty}^{\infty}$ is a sequence of uncorrelated random variables. We then have

$$\begin{aligned} E\{(\tilde{X}_t - \bar{X}_t)Y_m\} &= \sum_{n=t+1}^{\infty} \tilde{h}_{t-n} E\{Y_n Y_m\} \\ &= \sigma^2 \sum_{n=t+1}^{\infty} \tilde{h}_{t-n} \delta_{n,m} = 0, \quad m \leq t, \end{aligned} \quad \text{(V.D.34)}$$

where $\delta_{n,m}$ is the Kronecker delta ($\delta_{n,m} = 1$ if $n = m$ and $\delta_{n,m} = 0$ if $n \neq m$) and where $\sigma^2 = E\{Y_n^2\}$. (Recall that $\{Y_n\}_{n=-\infty}^{\infty}$ is assumed to be w.s.s. and zero-mean.)

Thus from the above we see that if we could first convert $\{Y_n\}_{n=-\infty}^{\infty}$ into an equivalent w.s.s. sequence $\{Z_n\}_{n=-\infty}^{\infty}$ of *uncorrelated* random variables by a causal linear operation, then the causal Wiener-Kolmogorov estimator would be given by simple truncation [as in (V.D.33)] of the optimum noncausal estimator of X_t based on $\{Z_n\}_{n=-\infty}^{\infty}$. Since such a sequence $\{Z_n\}_{n=-\infty}^{\infty}$ would have a constant, say unity spectrum, this latter estimator would be given by

$$\hat{X}_t = \sum_{n=-\infty}^{t} \hat{h}_{t-n} Z_n, \quad \text{(V.D.35)}$$

where

$$\hat{h}_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_{XZ}(\omega) e^{i\omega n} d\omega = C_{XZ}(n), \quad n \geq 0, \quad \text{(V.D.36)}$$

with ϕ_{XZ} and C_{XZ} the cross spectrum and cross covariance, respectively, between $\{X_n\}_{n=-\infty}^{\infty}$ and $\{Z_n\}_{n=-\infty}^{\infty}$.

The idea of causally converting $\{Y_n\}_{n=-\infty}^{\infty}$ into an equivalent white[9] sequence $\{Z_n\}_{n=-\infty}^{\infty}$ is not an unrealistic one in view of similar ideas that have arisen in the Gaussian detection and estimation problems of Sections III.B and V.B. In view of these earlier analyses, let us suppose that $\hat{Y}_{t|t-1}$ is the best linear prediction of Y_t from $\{Y_n\}_{n=-\infty}^{t-1}$; i.e., suppose that $\hat{Y}_{t|t-1}$ minimizes $E\{(Y_t - Z)^2\}$ over all $Z \in \mathcal{H}_{-\infty}^{t}$. Also, let σ_t^2 denote the mean-squared

[9] Here we use the term white to denote a w.s.s. sequence of zero-mean uncorrelated random variables. Such a sequence will have a spectrum that is constant for $\omega \in [-\pi, \pi]$. Thus by analogy with white light, which is light containing equal levels of all visible wavelengths, such a sequence is termed *white*.

error in this prediction; i.e.,

$$\sigma_t^2 = E\{(Y_t - \hat{Y}_{t|t-1})^2\}. \tag{V.D.37}$$

Now define a sequence $\{Z_n\}_{n=-\infty}^{\infty}$ by

$$Z_n = \frac{Y_n - \hat{Y}_{n|n-1}}{\sigma_n}, \quad n \in \mathbb{Z}, \tag{V.D.38}$$

and that $Z_n \in \mathcal{H}_{-\infty}^n$. We have $E\{Z_n^2\} = 1, E\{Z_n\} = 0$, and $\mathrm{Cov}(Z_n, Z_m) = E\{Z_n Z_m\}$. With $m < n$ we have

$$E\{Z_n Z_m\} = (1/\sigma_n) E\{(Y_n - \hat{Y}_{n|n-1}) Z_m\} = 0,$$

by the orthogonality principle since $Z_m \in \mathcal{H}_{-\infty}^{n-1}$. Similarly, $E\{Z_n Z_m\} = 0$ for $m > n$. Thus, this $\{Z_n\}_{n=-\infty}^{\infty}$ is a white sequence obtained by causal linear transformation of $\{Y_n\}_{n=-\infty}^{\infty}$.

Now if we could show that $\{Z_n\}_{n=-\infty}^{\infty}$ is equivalent to $\{Y_n\}_{n=-\infty}^{\infty}$ for the purposes of linear MMSE estimation, then the causal Wiener-Kolmogorov estimation problem is effectively reduced to the problem of one-step linear prediction. This follows since one-step linear prediction can be used to prewhiten the observations via (V.D.38), and then the causal estimator of X_t from the prewhitened data is given straightforwardly by (V.D.35) and (V.D.36). Such an equivalence between $\{Z_n\}_{n=-\infty}^{\infty}$ and $\{Y_n\}_{n=-\infty}^{\infty}$ can in fact be established within a mild condition on the spectrum of $\{Y_n\}_{n=-\infty}^{\infty}$. To show this, we now turn to the analysis of the specific problem of linear prediction.

Linear Prediction

Consider the specific causal Wiener-Kolmogorov problem in which $X_t = Y_{t+\lambda}$, where λ is a positive integer. Note that for $\lambda = 1$, this is similar to the Levinson problem except that our observations now extend back in time to $-\infty$. To seek a solution to this problem, we first give the following result, which is central to the theory of linear prediction.

Proposition V.D.1: The Spectral Factorization Theorem

Suppose that $\{Y_n\}_{n=-\infty}^{\infty}$ has a spectrum satisfying the so-called *Paley-Wiener condition*, given by

$$c_0 \stackrel{\Delta}{=} \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \phi_Y(\omega) d\omega > -\infty. \tag{V.D.39}$$

Then ϕ_Y can be written as $\phi_Y(\omega) = \phi_Y^+(\omega)\phi_Y^-(\omega), -\pi \le \omega \le \pi$, where ϕ_Y^+ and ϕ_Y^- are two functions satisfying $|\phi_Y^+(\omega)|^2 = |\phi_Y^-(\omega)|^2 = \phi_Y(\omega)$,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_Y^+(\omega) e^{in\omega} d\omega = 0, \text{ for all } n < 0 \tag{V.D.40a}$$

and

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_Y^-(\omega) e^{in\omega} d\omega = 0, \text{ for all } n > 0. \tag{V.D.40b}$$

Moreover (V.D. 40a) [resp. (V.D. 40b)] is also satisfied when ϕ_Y^+ [resp. ϕ_Y^-] is replaced by $1/\phi_Y^+$[resp. $1/\phi_Y^-$].[10]

Proof: A complete proof of the result can be found, for example, in Ash and Gardner (1975). Here we will outline the key ideas in this argument.

First we note that since $\log(x)$ is a concave function of x, Jensen's inequality implies that

$$c_0 \le \log\left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_Y(\omega) d\omega\right) = \log(E\{Y^2\}) < \infty.$$

Thus the condition (V.D.39) is equivalent to the condition $|c_0| < \infty$. This allows us to write $\log \phi_Y$ as a discrete-time Fourier transform

$$\log \phi_Y(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{-i\omega n}, \quad -\pi < \omega \le \pi, \tag{V.D.41a}$$

where

$$c_n \stackrel{\Delta}{=} \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{in\omega} \log \phi_Y(\omega), \quad n \in \mathbb{Z}. \tag{V.D.41b}$$

From (V.D.41a) we can write

$$\begin{aligned} \phi_Y(\omega) &= \exp\left\{\sum_{n=-\infty}^{\infty} c_n e^{-i\omega n}\right\} \\ &= \phi_Y^+(\omega)\phi_Y^-(\omega), \quad -\pi \le \omega \le \pi, \end{aligned} \tag{V.D.42}$$

where

$$\phi_Y^+(\omega) \stackrel{\Delta}{=} \exp\left\{\frac{c_0}{2} + \sum_{n=1}^{\infty} c_n e^{-i\omega n}\right\} \tag{V.D.43}$$

and

$$\phi_Y^-(\omega) \stackrel{\Delta}{=} \exp\left\{\frac{c_0}{2} + \sum_{n=-\infty}^{-1} c_n e^{-i\omega n}\right\}. \tag{V.D.44}$$

[10]Note that (V.D.39) implies that ϕ_Y is nonzero. Since $|\phi_Y^+|^2 = |\phi_Y^-|^2 = \phi_Y$, we see that ϕ_Y^+ and ϕ_Y^- must also be nonzero, so $1/\phi_Y^+$ and $1/\phi_Y^-$ are well defined. With $1/\phi_Y^+$ satisfying (V.D. 40a), ϕ_Y^+ is said to be of *minimum phase*.

We would now like to show that the functions ϕ_Y^+ and ϕ_Y^- have the desired properties. To do so, we must first note that all power spectral densities are even-symmetric functions [i.e., $\phi_Y(-\omega) = \phi_Y(\omega)$], a fact that follows straightforwardly from the even symmetry of C_Y. The even symmetry of $\phi_Y(\omega)$ implies that $\log \phi_Y(\omega)$ is also even symmetric, which in turn implies that the coefficients $\{c_n\}_{n=-\infty}^{\infty}$ of (V.D.41b) are real and even symmetric (i.e., $c_{-n} = c_n = c_n^*$). Thus we can write

$$\begin{aligned} \phi_Y^-(\omega) &= \exp\left\{\frac{c_0}{2} + \sum_{n=1}^{\infty} c_{-n} e^{i\omega n}\right\} \\ &= \exp\left\{\frac{c_0}{2} + \sum_{n=1}^{\infty} c_n e^{i\omega n}\right\} = [\phi^+(\omega)]^*. \end{aligned} \qquad \text{(V.D.45)}$$

Equation (V.D.45) implies that

$$|\phi_Y^-(\omega)|^2 = |\phi_Y^+(\omega)|^2 = \phi_Y^+(\omega)[\phi_Y^+(\omega)]^* = \phi_Y^+(\omega)\phi_Y^-(\omega) = \phi_Y(\omega).$$

To verify (V.D. 40a) we first note that e^z has the power series expansion

$$e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}. \qquad \text{(V.D.46)}$$

So we can write

$$\phi_Y^+(\omega) = e^{c_0/2}\left(\sum_{k=0}^{\infty}\left(\sum_{n=1}^{\infty} c_n e^{-i\omega n}\right)^k / k!\right). \qquad \text{(V.D.47)}$$

Inspection of (V.D.47) reveals that its right-hand side contains only non-negative powers of $e^{-i\omega}$. This implies that the Fourier components of $\phi_Y^+(\omega)$ with negative indices are all zero. This is (V.D. 40a). Equation (V.D. 40b) and the analogous conditions for $1/\phi_Y^+$ and $1/\phi_Y^-$ follow by similar arguments. □

It is interesting to note that the spectral factorization of ϕ_Y into a product of causal and anticausal parts is analogous to the Cholesky decomposition of a covariance matrix into lower and upper triangular factors.

We can use the spectral decomposition of Proposition V.D.1 to find linear MMSE predictors. To do so we henceforth assume that ϕ_Y satisfies (V.D.39). Consider the time-invariant linear filter with transfer function $H(\omega) = 1/\phi_Y^+(\omega)$. Note that this is a causal filter by way of condition (V.D. 40a) applied to $1/\phi_Y^+$. Suppose that we apply the sequences $\{Y_n\}_{n=-\infty}^{\infty}$ to this filter and let $\{W_n\}_{n=-\infty}^{\infty}$ denote the output sequence. A well-known

result in the analysis of second-order random sequences is that the output of a time-invariant linear filter driven by a w.s.s. process is also w.s.s. and that the spectrum of the output process is given by $|H(\omega)|^2\phi(\omega)$, where H is the filter transfer function and ϕ is the input spectrum [see, e.g., Wong (1983)]. Thus $\{W_n\}_{n=-\infty}^{\infty}$ is a w.s.s. sequence and its spectrum is

$$\begin{aligned} \phi_W(\omega) &= \left|\frac{1}{\phi_Y^+(\omega)}\right|^2 \phi_Y(\omega) \\ &= \frac{\phi_Y(\omega)}{|\phi_Y^+(\omega)|^2} = 1, \quad -\pi \le \omega \le \pi. \end{aligned} \qquad \text{(V.D.48)}$$

Since a constant spectrum corresponds to a white sequence, we see that the filter $1/\phi_Y^+(\omega)$ is a whitening filter for $\{Y_n\}_{n=-\infty}^{\infty}$. Moreover, $\{Y_n\}_{n=-\infty}^{\infty}$ is obtained causally from $\{W_n\}_{n=-\infty}^{\infty}$ by applying the latter sequence to the filter with transfer function $\phi_Y^+(\omega)$. Thus we can write

$$Y_t = \sum_{n=-\infty}^{t} f_{t-n} W_n, \quad t \in \mathbb{Z}, \qquad \text{(V.D.49)}$$

where

$$f_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi_Y^+(\omega) e^{i\omega n} d\omega, \quad n \ge 0. \qquad \text{(V.D.50)}$$

Equation (V.D.49) gives a representation for $\{Y_n\}_{n=-\infty}^{\infty}$ as the output of a time-invariant linear filter driven by a white sequence. This representation can be used to derive the optimum linear predictor of the sequence $\{Y_n\}_{n=-\infty}^{\infty}$. In particular, we note that for $\lambda > 0$,

$$\begin{aligned} Y_{t+\lambda} &= \sum_{n=-\infty}^{t+\lambda} f_{t+\lambda-n} W_n \\ &= \sum_{n=t+1}^{t+\lambda} f_{t+\lambda-n} W_n + \sum_{n=-\infty}^{t} f_{t+\lambda-n} W_n. \end{aligned} \qquad \text{(V.D.51)}$$

Since $\{W_n\}_{n=-\infty}^{\infty}$ is white, the variables $W_{t+1}, \ldots, W_{t+\lambda}$ are orthogonal to $\{W_n\}_{n-\infty}^{t}$, and the representation (V.D.49) thus implies that they are orthogonal to $\mathcal{H}_{-\infty}^t$. On rearranging (V.D.51) as

$$Y_{t+\lambda} - \sum_{n=-\infty}^{t} f_{t+\lambda-n} W_n = \sum_{n=t+1}^{t+\lambda} f_{t+\lambda-n} W_n,$$

we then have that $Y_{t+\lambda} - \sum_{n=-\infty}^{t} f_{t+\lambda-n} W_n$ is orthogonal to $\mathcal{H}_{-\infty}^t$. Since $\sum_{n=-\infty}^{t} f_{t+\lambda-n} W_n \in \mathcal{H}_{-\infty}^t$, the orthogonality principle implies that the

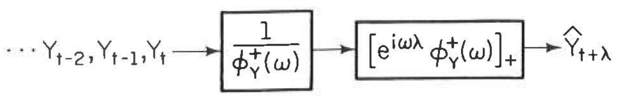

FIGURE V.D.1. Representation of the optimum pure-prediction filter

best linear prediction of $Y_{t+\lambda}$ from $\{Y_n\}_{n=-\infty}^{t}$ is given by

$$\hat{Y}_{t+\lambda} = \sum_{n=-\infty}^{t} f_{t+\lambda-n} W_n. \tag{V.D.52}$$

The linear prediction filter of (V.D.52) can be thought of as a series connection of two time-invariant linear filters, as depicted in Fig. V.D.1. The first of these two filters is the whitening filter with transfer function $1/\phi_Y^+(\omega)$. The second filter has impulse response

$$\begin{cases} f_{n+\lambda} & \text{for } n \geq 0 \\ 0 & \text{for } n < 0. \end{cases} \tag{V.D.53}$$

If we define the operation $[H(\omega)]_+$ by

$$[H(\omega)]_+ = \sum_{n=0}^{\infty} h_n e^{-i\omega n}, \tag{V.D.54}$$

where $h_n = (1/2\pi)\int_{-\pi}^{\pi} H(\omega)e^{i\omega n}d\omega$, the filter of (V.D.53) has the transfer function $[e^{i\omega\lambda}\phi_Y^+(\omega)]_+$. Thus the overall transfer function for optimum linear prediction λ steps into the future can be written as

$$\frac{1}{\phi_Y^+(\omega)}[e^{i\omega\lambda}\phi_Y^+(\omega)]_+ \stackrel{\Delta}{=} H_\lambda(\omega). \tag{V.D.55}$$

The representation of (V.D.49) and (V.D.52) also allows us to write an expression for the mean-squared error incurred in optimum linear prediction. In particular, from (V.D.51) we have that

$$\begin{aligned} \text{MMSE} &= \min_{Z \in \mathcal{H}_{-\infty}^{t}} E\{(Y_{t+\lambda} - Z)^2\} = E\{(Y_{t+\lambda} - \hat{Y}_{t+\lambda})^2\} \\ &= E\left\{\left(\sum_{n=t+1}^{t+\lambda} f_{t+\lambda-n}W_n\right)^2\right\} = \sum_{n=t+1}^{t+\lambda} f_{t+\lambda-n}^2, \end{aligned} \tag{V.D.56}$$

where the last equality follows from the orthogonality of $W_{t+1}, W_{t+2}, \ldots,$ and $W_{t+\lambda}$. A simple change of variables in (V.D.56) gives

$$\text{MMSE} = \sum_{n=0}^{\lambda-1} f_n^2. \tag{V.D.57}$$

The coefficients $\{f_n\}_{n=0}^{\lambda-1}$ can be obtained from (V.D.47) since f_n is the nth Fourier coefficient of ϕ_Y^+, although this is somewhat tedious for large λ. The most interesting case is one-step prediction ($\lambda = 1$). It is easily seen from (V.D.47) that $a_0 = e^{c_0/2}$; thus the value of MMSE for one-step prediction is

$$\text{MMSE} = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log \phi_Y(\omega)d\omega\right\}, \tag{V.D.58}$$

a result known as the *Kolmogorov-Szegö-Krein formula.*

Equation (V.D.51) and the orthogonality of $\{W_n\}_{n=-\infty}^{\infty}$ imply that[11]

$$\sum_{n=0}^{\infty} f_n^2 = E\{(Y_{t+\lambda})^2\} < \infty. \tag{V.D.59}$$

Comparing (V.D.57) and (V.D.59), we see that as we try to predict further into the future (i.e., as λ increases), the minimum mean-squared prediction error approaches the mean-squared value of the quantity we are trying to predict. An equivalent interpretation is that if we fix the time at which we are trying to estimate $\{Y_n\}_{n=-\infty}^{\infty}$ (i.e., fix $t + \lambda$) and let t approach $-\infty$, then in the limit as $t \to -\infty$ the observations are of no use in predicting $Y_{t+\lambda}$. In other words, for fixed n, no part of Y_n can be determined from the infinite past. A sequence with this property is said to be *purely nondeterministic*, and a sequence has this property if and only if it has the representation $\sum_{n=-\infty}^{t} f_{t-n}W_n$, with $\{W_n\}_{n=-\infty}^{\infty}$ white and $\sum_{n=0}^{\infty} f_n^2 < \infty$. Proposition V.D.1 and the analysis following it show that the Paley-Wiener condition is sufficient for the existence of this type of representation. It can be shown that this condition is also a necessary condition for the existence of such a representation [see Ash and Gardner (1975) for further discussion of this notion].

Having determined a mechanism for causally prewhitening a covariance stationary sequence of observations, the solution to the general causal Wiener-Kolmogorov problem follows almost immediately. In particular, assuming that $\{Y_n\}_{n=-\infty}^{\infty}$ satisfies the Paley-Wiener condition, it can be whitened (into, say, $\{Z_n\}_{n=-\infty}^{\infty}$) by passing it through the filter $1/\phi_Y^+$. Then we need only find the cross spectrum ϕ_{XZ} and (V.D.35) and (V.D.36) give us the optimum filter to follow the prewhitener as $[\phi_{XZ}(\omega)]_+$, where the notation $[\cdot]_+$ is as defined in (V.D.54). It is a straightforward exercise to

[11]Equivalently, Parceval's formula gives

$$\sum_{n=0}^{\infty} f_n^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\phi_Y^+(\omega)|^2 d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi_Y(\omega)d\omega = E\{Y_t^2\}, \ t \in \mathbb{Z}.$$

show that if $\{Y_n\}_{n=-\infty}^{\infty}$ is passed through a filter with transfer function H to get the sequence $\{Z_n\}_{n=-\infty}^{\infty}$, then $\phi_{XZ} = \phi_{XY}H^*$. Thus since in our case $H = 1/\phi_Y^+$ we have

$$\phi_{XZ}(\omega) = \frac{\phi_{XY}(\omega)}{[\phi_Y^+(\omega)]^*} = \frac{\phi_{XY}(\omega)}{\phi_Y^-(\omega)}. \tag{V.D.60}$$

We then have that the causal Wiener-Kolmogorov filter for estimating X_t from $\{Y_n\}_{n=-\infty}^{t}$ has transfer function

$$\frac{1}{\phi_Y^+(\omega)}\left[\frac{\phi_{XY}(\omega)}{\phi_Y^-(\omega)}\right]_+. \tag{V.D.61}$$

It is interesting to note that the noncausal Wiener-Kolmogorov filter has transfer function [see (V.D.11)]

$$\frac{\phi_{XY}(\omega)}{\phi_Y(\omega)} = \frac{1}{\phi_Y^+(\omega)}\left[\frac{\phi_{XY}(\omega)}{\phi_Y^-(\omega)}\right]. \tag{V.D.62}$$

Thus both filters can be represented as a series connection of a causal prewhitener with a second filter. In the noncausal filter this second filter is ϕ_{XY}/ϕ_Y^- and in the causal case it is the (additive) causal part of ϕ_{XY}/ϕ_Y^-.

Factorization of Rational Spectra

Note that the key step in designing a causal Wiener-Kolmogorov filter is the factorization of the observation spectrum ϕ_Y. Since ϕ_Y^+ and ϕ_Y^- can be written in terms of the Fourier coefficients $\{c_n\}_{n=-\infty}^{\infty}$ of $\log \phi_Y$ [see (V.D.43) and (V.D.44)] this factorization can be performed numerically by computing the c_n's. (The logarithm of the spectrum is often termed the *cepstrum*.) For a large class of spectra of interest in practice, however, spectral factorization can be viewed as the factorization of complex polynomials. We discuss this issue briefly in the following paragraphs.

The power spectrum of a random sequence $\{Y_n\}_{n=-\infty}^{\infty}$ is said to be *rational* if it can be written as the ratio of two real trigonometric polynomials; i.e., ϕ is rational if we can write

$$\phi_Y(\omega) = \frac{n_0 + 2\sum_{k=1}^{p} n_k \cos k\omega}{d_0 + 2\sum_{k=1}^{m} d_k \cos k\omega}, \tag{V.D.63}$$

where m and p are positive integers and $n_0, \ldots, n_p, d_0, \ldots, d_m$ are real numbers. Many random sequences arising in practice have this type of spectrum, and most covariance stationary sequences have spectra that can be approximated arbitrarily closely by rational spectra with large enough choice of the orders m and p. Since power spectra must be even symmetric about $\omega = 0$, the polynomials in (V.D.62) contain only cosine terms and no sine terms.

Since $2\cos k\omega = e^{ik\omega} + e^{-ik\omega}$, the spectrum of (V.D.63) can be written as

$$\phi_Y(\omega) = \frac{N(e^{i\omega})}{D(e^{i\omega})}, \tag{V.D.64}$$

where N and D are polynomials of a complex variable z definedby

$$N(z) = \sum_{k=-p}^{p} n_{|k|} z^{-k} \tag{V.D.65a}$$

and

$$D(z) = \sum_{k=-m}^{m} d_{|k|} z^{-k}. \tag{V.D.65b}$$

Note that $z^p N(z)$ is a $(2p)$th order polynomial, so it has $2p$ roots $z_1, z_2, \ldots, z_{2p}$, and can be written as $n_p \Pi_{k=1}^{2p}(z - z_k)$. Thus we can write

$$N(z) = n_p z^{-p} \prod_{k=1}^{2p}(z - z_k). \tag{V.D.66}$$

Since $N(z) = N(1/z)$, the roots $z_1, \ldots, z_{2p}$ must be in reciprocal pairs; i.e., for each root z_k there is another root equal to $1/z_k$. Assuming for notational convenience that $z_1, \ldots, z_{2p}$ are ordered such that $|z_1| \geq |z_2| \geq \cdots \geq |z_{2p}|$, it is straightforward to write

$$N(z) = B(z)B(1/z), \tag{V.D.67}$$

where

$$B(z) = [(-1)^p n_p / z_1 z_2 \cdots z_p]^{1/2} \prod_{k=1}^{p}(z^{-1} - z_k). \tag{V.D.68}$$

Since $|z_1| \geq |z_2| \geq \cdots \geq |z_{2p}|$, we must have $|z_{2p}| = 1/|z_1|, |z_{2p-1}| = 1/|z_2|, \cdots, |z_{p+1}| = 1/|z_p|$, from which it follows that $|z_1| \geq |z_2| \geq \cdots \geq |z_p| \geq 1$. Note that $B(z)$ can be expanded into the form

$$B(z) = \sum_{k=0}^{p} b_k z^{-k}. \tag{V.D.69}$$

Similarly, the polynomial $D(z)$ can be written as

$$D(z) = A(z)A(1/z) \tag{V.D.70}$$

with $A(z)$ in the form

$$A(z) = [d_m/p_1p_2\cdots p_m]^{1/2}\prod_{k=1}^{m}(z^{-1}-p_k)$$
$$= \sum_{k=0}^{m} a_k z^{-k}, \qquad \text{(V.D.71)}$$

where $|p_1| \geq |p_2| \geq \cdots \geq |p_m| \geq 1$.

We see from the above that the rational spectrum $\phi_Y(\omega)$ can be written as

$$\phi_Y(\omega) = \frac{B(e^{i\omega})B(e^{-i\omega})}{A(e^{i\omega})A(e^{-i\omega})}. \qquad \text{(V.D.72)}$$

We assume henceforth that none of the roots of $B(z)$ or $A(z)$ is on the unit circle $|z| = 1$ (i.e., we assume that $|z_p| > 1$ and $|p_m| > 1$). [This ensures that $\phi_Y(\omega)$ is bounded from above and is bounded away from zero from below, which in turn implies that it satisfies the Paley-Wiener condition.] It is not hard to show [see, e.g., Oppenheim and Schafer (1975)] that both $B(e^{i\omega})/A(e^{i\omega})$ and $A(e^{i\omega})/B(e^{i\omega})$ are causal stable transfer functions, and that both $B(e^{-i\omega})/A(e^{-i\omega})$ and $A(e^{-i\omega})/B(e^{-i\omega})$ (V.D.28) are anticausal stable transfer functions. It follows from this and (V.D.72) that the spectral factors of ϕ_Y are

$$\phi_Y^+(\omega) = B(e^{i\omega})/A(e^{i\omega}) \qquad \text{(V.D.73a)}$$

and

$$\phi_Y^-(\omega) = [\phi_Y^+(\omega)]^* = B(e^{-i\omega})/A(e^{-i\omega}). \qquad \text{(V.D.73b)}$$

The whitening filter for $\{Y_n\}_{n=-\infty}^{\infty}$ is now given by

$$\frac{1}{\phi_Y^+(\omega)} = \frac{A(e^{i\omega})}{B(e^{i\omega})}. \qquad \text{(V.D.74)}$$

Equivalently, with $\{Z_n\}_{n=-\infty}^{\infty}$ representing the whitened sequence, we can say that the output of the filter $A(e^{i\omega})$ when applied to $\{Y_n\}_{n=-\infty}^{\infty}$ equals the output of the filter $B(e^{i\omega})$ when applied to $\{Z_n\}_{n=-\infty}^{\infty}$. From (V.D.69) and (V.D.71)the impulse responses of $A(e^{i\omega})$ and $B(e^{i\omega})$ are, respectively,

$$\begin{cases} a_n & \text{if } 0 \leq n \leq m \\ 0 & \text{otherwise} \end{cases} \qquad \text{(V.D.75)}$$

and

$$\begin{cases} b_n & \text{if } 0 \leq n \leq p \\ 0 & \text{otherwise.} \end{cases} \qquad \text{(V.D.76)}$$

This implies that $\{Y_n\}_{n=-\infty}^{\infty}$ and $\{Z_n\}_{n=-\infty}^{\infty}$ are related by

$$\sum_{k=0}^{m} a_k Y_{n-k} = \sum_{k=0}^{p} b_k Z_{n-k}, \quad n \in \mathbb{Z}. \qquad \text{(V.D.77)}$$

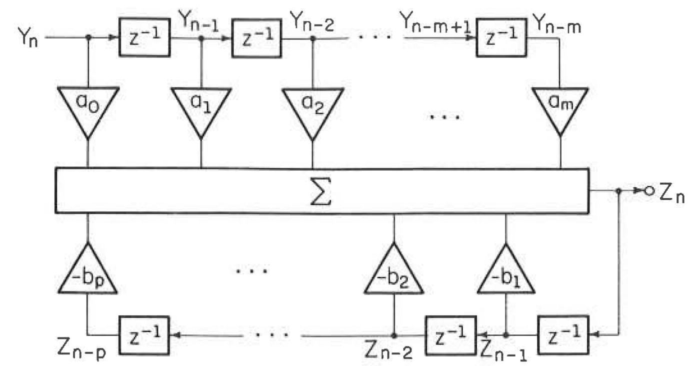

FIGURE V.D.2. Whitening filter for a sequence with a rational spectrum.

Thus the sequence $\{Z_n\}_{n=-\infty}^{\infty}$ satisfies the recursion

$$b_0 Z_n = -\sum_{k=1}^{p} b_k Z_{n-k} + \sum_{k=0}^{m} a_k Y_{n-k}. \qquad \text{(V.D.78)}$$

This recursion represents a finite-dimensional linear digital filter, as illustrated in Fig. V.D.2. (In the figure z^{-1} represents a delay of one time unit.)

Note that (V.D.77) also implies that $\{Y_n\}_{n=-\infty}^{\infty}$ is generated from $\{Z_n\}_{n=-\infty}^{\infty}$ by the recursion

$$a_0 Y_n = -\sum_{k=1}^{m} a_k Y_{n-k} + \sum_{k=0}^{p} b_k Z_{n-k}, \quad n \in \mathbb{Z}. \qquad \text{(V.D.79)}$$

A sequence generated in this fashion from a white sequence is said to be an *autoregressive/moving-average sequence with autoregressive order m and moving-average order p*, or an ARMA (m,p) sequence. An ARMA $(m,0)$ sequence is called an *autoregressive sequence of order m* [AR(m)] and an ARMA$(0,p)$ sequence is called a *moving average of order p* [MA(p)]. [With $m = 0$, the first sum in (V.D.79) is taken to be zero.] ARMA models are closely related to the state-space models arising in Kalman-Bucy filtering, and some of their properties relevant to filtering can be found in Anderson and Moore (1979).

Example V.D.2: Pure Prediction of a Wide-Sense Markov Sequence

A simple but useful model for the correlation structure of covariance stationary random sequences is the so-called *wide-sense Markov model*:

$$C_Y(n) = Pr^{|n|}, \quad n \in \mathbb{Z}, \qquad \text{(V.D.80)}$$

where $|r| < 1$ and $P > 0$. The power spectrum corresponding to (V.D.80) is given by [see, e.g., Thomas (1971)]

$$\phi_Y(\omega) = \frac{P(1-r^2)}{1 - 2r\cos\omega + r^2}. \tag{V.D.81}$$

Note that (V.D.81) is a rational spectrum, and using $2\cos\omega = e^{i\omega} + e^{-i\omega}$, we have

$$\begin{aligned}\phi_Y(\omega) &= \frac{P(1-r^2)}{1 - re^{i\omega} - re^{-i\omega} + r^2} \\ &= \frac{P(1-r^2)}{(1-re^{-i\omega})(1-re^{i\omega})} \\ &= \frac{1}{A(e^{i\omega})A(e^{-i\omega})},\end{aligned} \tag{V.D.82}$$

where

$$A(z) = a_0 + a_1 z^{-1}, \tag{V.D.83}$$

with $a_0 = [P(1-r^2)]^{-1/2}$ and $a_1 = -r[P(1-r^2)]^{-1/2}$.

Suppose that we wish to predict $\{Y_n\}_{n=-\infty}^{\infty}\lambda$ steps into the future. The transfer function of the optimum prediction is given by (V.D.55), which in this case becomes

$$\hat{H}_\lambda(\omega) = A(e^{i\omega})\left[\frac{e^{i\omega\lambda}}{A(e^{i\omega})}\right]_+. \tag{V.D.84}$$

On using the geometric series, $\sum_{k=0}^{\infty} x^k = 1/(1-x)$ for $|x| < 1$, we have

$$\frac{1}{A(z)} = \frac{1}{a_0(1-rz^{-1})} = \frac{1}{a_0}\sum_{n=0}^{\infty} r^n z^{-n}, \tag{V.D.85}$$

which converges for $|z| = 1$ since $|r| < 1$. So $1/A(e^{i\omega}) = (1/a_0)\sum_{n=0}^{\infty} r^n e^{-i\omega n}$ and we have the following steps:

$$\begin{aligned}\left[\frac{e^{i\omega\lambda}}{A(e^{i\omega})}\right]_+ &= \left[\frac{1}{a_0}\sum_{n=0}^{\infty} r^n e^{-i\omega(n-\lambda)}\right]_+ \\ &= \frac{1}{a_0}\sum_{n=\lambda}^{\infty} r^n e^{-i\omega(n-\lambda)} \\ &= \frac{1}{a_0}\sum_{l=0}^{\infty} r^{l+\lambda} e^{-i\omega l} = \frac{r^\lambda}{A(e^{i\omega})}.\end{aligned} \tag{V.D.86}$$

Considering (V.D.84) and (V.D.86), we have that $H_\lambda(\omega) = r^\lambda$; that is, in this case the optimum predictor is a pure gain. The impulse response of the predictor is thus $h_0 = r^\lambda$ and $h_n = 0, n \neq 0$, so we have simply

$$\hat{Y}_{t+\lambda} = r^\lambda Y_t. \tag{V.D.87}$$

The mean-squared prediction error is easily computed from (V.D.87) and (V.D.80) as

$$\begin{aligned}E\{(Y_{t+\lambda} - \hat{Y}_{t+\lambda})^2\} &= E\{Y_{t+\lambda}^2\} - E\{Y_{t+\lambda}\hat{Y}_{t+\lambda}\} \\ &= E\{Y_{t+\lambda}^2\} - r^\lambda E\{Y_{t+\lambda}Y_t\} \\ &= C_Y(0) - r^\lambda C_Y(\lambda) \\ &= P(1-r^{2\lambda}).\end{aligned} \tag{V.D.88}$$

Since $|r| < 1$, the prediction error increases monotonically from $(1-r^2)P$ to P as λ increases from 1 to ∞.

Example V.D.3: Pure Prediction of AR(m) Sequences

In view of (V.D.83), a wide-sense Markov sequence is an AR(1) sequence. In particular, $\{Y_n\}_{n=-\infty}^{\infty}$ is generated by

$$Y_{t+1} = rY_t + [P(1-r^2)]^{1/2} Z_{t+1}, \quad t \in \mathbb{Z}, \tag{V.D.89}$$

where $\{Z_n\}_{n=-\infty}^{\infty}$ is white. Since Z_{t+1} is orthogonal to $\{Z_n\}_{n=-\infty}^{t}$ and hence to $\{Y_n\}_{n=-\infty}^{t}$, we see from (V.D.89) that $(Y_{t+1} - rY_t)$ is orthogonal to $\{Y_n\}_{n=-\infty}^{t}$ and thus the orthogonality principle implies that rY_t is the MMSE linear estimate of Y_{t+1} from $\{Y_n\}_{n=-\infty}^{t}$. This is (V.D.87) for $\lambda = 1$.

Similarly, for any autoregressive sequence

$$Y_{t+1} = -\sum_{k=1}^{m} a_k Y_{t+1-k} + b_0 Z_{t+1}, \quad t \in \mathbb{Z}, \tag{V.D.90}$$

(without loss of generality we take $a_0 = 1$), the quantity $Y_{t+1} + \sum_{k=1}^{m} a_k Y_{t+1-k} b_0 Z_{t+1}$ is orthogonal to $\{Y_n\}_{n=-\infty}^{t}$. So the optimum one-step predictor is

$$\hat{Y}_{t+1} = -\sum_{k=1}^{m} a_k Y_{t+1-k}. \tag{V.D.91}$$

The minimum mean-squared prediction error is simply

$$\begin{aligned}\text{MMSE} &= E\left\{\left(Y_{t+1} + \sum_{k=1}^{m} a_k Y_{t+1-k}\right)^2\right\} \\ &= E\{b_0^2 Z_{t+1}^2\} = b_0^2 E\{Z_{t+1}^2\} = b_0^2.\end{aligned} \tag{V.D.92}$$

For the AR(1) case $b_0^2 = P(1-r^2)$, which agrees with (V.D.88). In general, the Kolmogorov-Szegö-Krein formula (V.D.58) gives

$$b_0^2 = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\phi_Y(\omega)d\omega\right\}. \tag{V.D.93}$$

Example V.D.4: Filtering, Prediction, and Smoothing of Wide-Sense Markov Sequences in White Noise
Consider the observation model

$$Y_n = S_n + N_n, \quad n \in \mathbb{Z}, \tag{V.D.94}$$

where $\{S_n\}_{n=-\infty}^{\infty}$ and $\{N_n\}_{n=-\infty}^{\infty}$ are zero-mean orthogonal wide-sense stationary sequences. Assume that $\{N_n\}_{n=-\infty}^{\infty}$ is white with $E\{N_n^2\} = v_N^2$ and that $\{S_n\}_{n=-\infty}^{\infty}$ is wide-sense Markov with $C_S(n) = Pr^{|n|}, n \in \mathbb{Z}$. Referring to Example V.D.2, and using the orthogonality of $\{S_n\}_{n=-\infty}^{\infty}$ and $\{N_n\}_{n=-\infty}^{\infty}$, the spectrum of the observation is given by

$$\begin{aligned}\phi_Y(\omega) &= \phi_S(\omega) + \phi_N(\omega) = \frac{P(1-r^2)}{(1-re^{-i\omega})(1-re^{i\omega})} + v_N^2 \\ &= \frac{P(1-r^2) + v_N^2(1-re^{-i\omega})(1-re^{i\omega})}{(1-re^{-i\omega})(1-re^{i\omega})},\end{aligned} \tag{V.D.95}$$

which is a rational spectrum.

The denominator polynomial in ϕ_Y is already factored as $A(z)A(1/z)$ with $A(z) = 1-rz^{-1}$. The numerator polynomial is $N(z) = n_1 z + n_0 + n_1 z^{-1}$ with $n_0 = P(1-r^2) + v_N^2(1+r^2)$ and $n_1 = -v_N^2 r$. Using the quadratic formula we can write $N(z)$ as

$$N(z) = n_1 z^{-1}(z - z_1)(z - 1/z_1),$$

where

$$z_1 = -[(n_0^2 - 4n_1^2)^{1/2} + n_0]/2n_1.$$

Note that $|z_1| > 1$, and thus $N(z) = B(z)B(1/z)$, where

$$B(z) = \sqrt{-n_1/z_1}(z^{-1} - z_1) = b_0 + b_1 z^{-1}, \tag{V.D.96}$$

with $b_0 = -z_1\sqrt{-n_1/z_1}$ and $b_1 = \sqrt{-n_1/z_1}$. The whitening filter in this case thus becomes

$$\frac{1}{\phi_Y^+(\omega)} = \frac{A(e^{i\omega})}{B(e^{i\omega})} = \frac{1-re^{-i\omega}}{b_0 + b_1 e^{-i\omega}}. \tag{V.D.97}$$

As in Example V.D.1, suppose that we are interested in estimating the signal sequence $\{S_n\}_{n=-\infty}^{\infty}$ at time $t+\lambda$. Then $X_t = S_{t+\lambda}$ and the required cross spectrum is given [see (V.D.24)] by

$$\phi_{XY}(\omega) = e^{i\omega\lambda}\phi_S(\omega) = \frac{P(1-r^2)e^{i\omega\lambda}}{A(e^{i\omega})A(e^{-i\omega})}. \tag{V.D.98}$$

Applying (V.D.97) and (V.D.98) to (V.D.61), the transfer function of the optimum filter is given by

$$\begin{aligned}H(\omega) &= \frac{A(e^{i\omega})}{B(e^{i\omega})}\left[\frac{P(1-r^2)e^{i\omega\lambda}}{A(e^{i\omega})B(e^{-i\omega})}\right]_+ \\ &= \left\{\frac{1-re^{-i\omega}}{b_0 + b_1 e^{-i\omega}}\right\}\left[\frac{P(1-r^2)e^{+i\omega\lambda}}{(1-re^{-i\omega})(b_0+b_1 e^{i\omega})}\right]_+.\end{aligned} \tag{V.D.99}$$

To simplify (V.D.99), consider the function of a complex variable z given by

$$\hat{H}(z) = \left(\frac{P(1-r^2)}{1-rz^{-1}}\right)\left(\frac{1}{b_0 + b_1 z}\right). \tag{V.D.100}$$

Using a partial fraction expansion, we can write

$$\hat{H}(z) = \frac{k'}{1-rz^{-1}} + \frac{k'}{1-z/z_1}, \tag{V.D.101}$$

where $z_1 = -b_0/b_1$ and $k' = P(1-r^2)/(b_0 + b_1 r)$. Using the geometric series, $\hat{H}$ becomes

$$\begin{aligned}\hat{H}(z) &= k'\sum_{n=0}^{\infty} r^n z^{-1} + k'\sum_{n=0}^{\infty} z_1^{-1} z^n \\ &= k'\sum_{n=0}^{\infty} r^n z^{-1} + k'\sum_{n=-\infty}^{0} z_1^n z^{-1}.\end{aligned} \tag{V.D.102}$$

The impulse response of $\hat{H}$ is then

$$\hat{h}_n = \begin{cases} k' z_1^n & \text{if } n < 0 \\ 2k' & \text{if } n = 0 \\ k' r^n & \text{if } n > 0. \end{cases} \tag{V.D.103}$$

The impulse response of $e^{i\omega\lambda}\hat{H}(\omega)$ thus becomes

$$\hat{h}_{n+\lambda} = \begin{cases} k' z_1^\lambda z_1^n & \text{if } n < -\lambda \\ 2k' & \text{if } n = -\lambda \\ k' r^\lambda r^n & \text{if } n > -\lambda. \end{cases} \tag{V.D.104}$$

The filter $\hat{h}_{n+\lambda}$ is illustrated in Fig. V.D.3. In order to get $\tilde{H}(\omega) \triangleq [e^{i\omega\lambda}\hat{H}(\omega)]_+$, we must truncate $\hat{h}_{n+\lambda}$ to be causal. From (V.D.104) we have that for $\lambda > 0$, the truncated impulse response is

$$\tilde{h}_n = \begin{cases} 0 & \text{if } n < 0 \\ k' r^\lambda r^n & \text{if } n \leq 0, \end{cases} \tag{V.D.105}$$

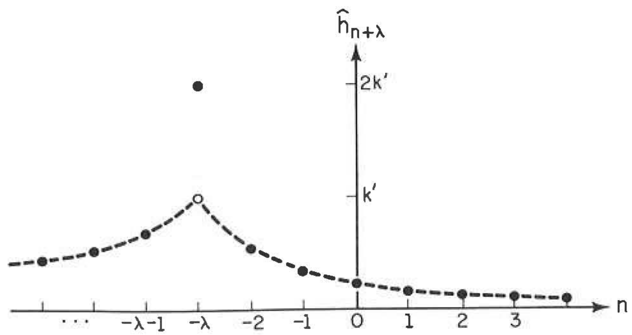

FIGURE V.D.3. Impulse response of the filter $e^{i\omega\lambda}\hat{H}(\omega)$.

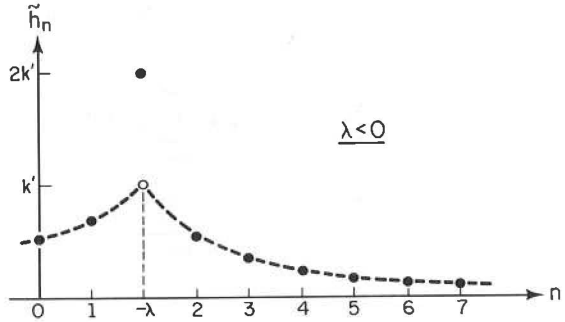

FIGURE V.D.4. Impulse response of the filter $[e^{i\omega\lambda}\hat{H}(\omega)]_+$ for $\lambda < 0$ (smoothing).

and for $\lambda \leq 0$ it becomes

$$\tilde{h}_n = \begin{cases} 0 & \text{if } n < 0 \\ k' z_1^{\lambda} z_1^{n} & \text{if } 0 \leq n < -\lambda \\ 2k' & \text{if } n = -\lambda \\ k' r^{\lambda} r^{n} & \text{if } n > -\lambda. \end{cases} \tag{V.D.106}$$

The two cases $\lambda < 0$ and $\lambda > 0$ are illustrated in Figs. V.D.4 and V.D.5, respectively.

To carry the analysis further, let us consider the case of prediction ($\lambda > 0$). From (V.D.105) we have that

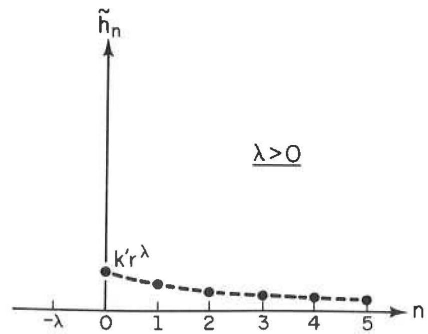

FIGURE V.D.5. Impulse response of the filter $[e^{i\omega\lambda}\hat{H}(\omega)]_+$ for $\lambda > 0$ (prediction).

$$\begin{aligned} \tilde{H}(e^{i\omega}) &= \sum_{n=0}^{\infty} \tilde{h}_n e^{-in\omega} \\ &= \sum_{n=0}^{\infty} k' r^{\lambda} r^{n} e^{-in\omega} = \frac{k' r^{\lambda}}{1 - re^{-i\omega}}. \end{aligned} \tag{V.D.107}$$

Combining (V.D.99) and (V.D.107), we see that the optimum prediction filter has transfer function

$$H(\omega) = \frac{k' r^{\lambda}}{b_0 + b_1 e^{-i\omega}} = \frac{k' r^{\lambda}/b_0}{1 - e^{-i\omega}/z_1}. \tag{V.D.108}$$

The impulse response of the optimum predictor is thus

$$h_n = \begin{cases} 0 & \text{if } n < 0, \\ \frac{k' r^{\lambda}}{b_0} z_1^{-n} & \text{if } n \geq 0. \end{cases} \tag{V.D.109}$$

Alternatively, this optimum predictor can be implemented recursively by

$$\hat{S}_{t+\lambda|t} = \frac{1}{z_1}\hat{S}_{t-1+\lambda|t-1} + \frac{k' r^{\lambda}}{b_0} Y_t, \quad t \in \mathbb{Z} \tag{V.D.110}$$

where $\hat{S}_{t+\lambda|t}$ and $\hat{S}_{t-1+\lambda|t-1}$ denote the optimum predictor of $S_{t+\lambda}$ from $\{Y_n\}_{n=-\infty}^{t}$ and $S_{t-1+\lambda}$ from $\{Y_n\}_{n=-\infty}^{t-1}$, respectively. Note that when $v_N^2 = 0$ (i.e., when there is no noise), $\hat{S}_{t+\lambda|t}$ reduces straightforwardly to the pure predictor derived in Example V.D.2.

It is of interest to consider the case of one-step prediction ($\lambda = 1$) further. Straightforward algebra yields that $z_1^{-1} = (r - k'r/b_0)$, so that (V.D.110) can be rewritten as

$$\hat{S}_{t+1|t} = r\hat{S}_{t|t-1} + \frac{k'r}{b_0}(Y_t - \hat{S}_{t|t-1}). \tag{V.D.111}$$

This form is reminiscent of the Kalman-Bucy prediction filter of Section V.B, which updates the one-step predictor in a state-space model in this same fashion. In fact, since $\{S_n\}_{n=-\infty}^{\infty}$ in this case is an AR (1) sequence, it can be generated via [see (V.D.89)]

$$S_{n+1} = rS_n + [P(1-r^2)]^{1/2}W_n, \quad n \in \mathbb{Z} \tag{V.D.112}$$

where $\{W_n\}_{n=-\infty}^{\infty}$ is white with unit variance ($\{W_n\}_{n=-\infty}^{\infty}$ is the prewhitened signal). The observation model is

$$Y_n = S_n + N_n, \quad n \in \mathbb{Z} \tag{V.D.113}$$

where $\{N_n\}_{n=-\infty}^{\infty}$ is a white sequence with variance v_N^2. Since $\{N_n\}_{n=-\infty}^{\infty}$ and $\{S_n\}_{n=-\infty}^{\infty}$ are orthogonal, so are $\{N_n\}_{n=-\infty}^{\infty}$ and $\{W_n\}_{n=-\infty}^{\infty}$. Thus (V.D.112) and (V.D.113) is a scalar time-invariant Kalman-Bucy model with white orthogonal noises. Thus from Example V.C.2 we know that the Kalman-Bucy filter provides the linear MMSE estimates of S_t and S_{t+1} given $\{Y_n; n \leq t\}$. The basic difference between this case and that treated in Section V.B is that (V.D.112) and (V.D.113) is a stationary or steady-state model. Its Kalman-Bucy prediction filter is thus the steady-state version derived in Example V.B.2, which is identical to (V.D.111) with the appropriate identification of equivalent parameters. In particular, the parameter set (a, b, c, q, r) in the Kalman-Bucy model of Example V.B.2 corresponds to the parameter set $(r, \sqrt{P(1-r^2)}, 1, 1, v_N^2)$ here.

Thus in the scalar time-invariant case, we can think of the Wiener-Kolmogorov filter as a steady-state version of the Kalman-Bucy filter, or, conversely, we can think of the Kalman-Bucy filter as a version of the Wiener-Kolmogorov filter that includes transient behavior. A similar identification can be made between other stable time-invariant Kalman-Bucy models and Wiener-Kolmogorov filters for signals with rational spectra observed in white noise [see, e.g., Anderson and Moore (1979) for further discussion of this issue.] Note, however, that Wiener-Kolmogorov filtering applies to more general spectral models for signals and noise, and that Kalman-Bucy filtering also applies to time-varying and unstable state-space models.

V.E Exercises

1. Show directly (i.e., without using the facts that $\hat{\underline{X}}_{t|t-1} = E\{\underline{X}_t | \underline{Y}_0^{t-1}\}$ and $\hat{\underline{X}}_{t|t} = E\{\underline{X}_t | \underline{Y}_0^t\}$) that the filtering and prediction errors generated by the Kalman filter are orthogonal to the data. I.e., show that

$$E\{(\underline{X}_t - \hat{\underline{X}}_{t|t})\underline{Y}_k^T\} = \mathbf{0}, \quad 0 \leq k \leq t \tag{V.E.1}$$

and

$$E\{(\underline{X}_t - \hat{\underline{X}}_{t|t-1})\underline{Y}_k^T\} = \mathbf{0}, \quad 0 \leq k \leq t-1. \tag{V.E.2}$$

where $\mathbf{0}$ denotes a matrix of all zeros.

2. Suppose the state equation in the Kalman-Bucy model is modified as follows:

$$\underline{X}_{k+1} = F_k\underline{X}_k + G_k\underline{U}_k + \Gamma_k\underline{s}_k$$

where $\{\underline{s}_k\}_{k=0}^{\infty}$ is a known sequence of vectors and $\{\Gamma_k\}_{k=0}^{\infty}$ is a known sequence of matrices of appropriate dimension. (Note, e.g., that $\{\underline{s}_k\}_{k=0}^{\infty}$ could be a sequence of controls.) Find the appropriate modification of the Kalman-Bucy recursions.

3. Repeat Exercise 2 for the situation in which each $\underline{s}_k$ is allowed to be a function of the past measurement; i.e., $\underline{s}_k$ can be a function of $\underline{Y}_0^k$. (So, for example, $\{\underline{s}_k\}_{k=0}^{\infty}$ could be a sequence of feedback controls.)

4. Suppose we return to the original Kalman-Bucy model, but allow for correlation between the state and measurement noises; i.e., assume everything as before except

$$\mathrm{Cov}(\underline{U}_k, \underline{V}_l) = \begin{cases} \mathbf{C}_k, & k = l \\ \mathbf{0}, & k \neq l \end{cases}$$

where $\mathbf{C}_k$ is a matrix of appropriate dimension. Show that the Kalman predictor is given by

$$\hat{\underline{X}}_{t+1|t} = \mathbf{F}_t\hat{\underline{X}}_{t|t-1} + \mathbf{K}_t(\underline{Y}_t - \mathbf{H}_t\hat{\underline{X}}_{t|t-1})$$

with

$$\hat{\underline{X}}_{0|-1} = \underline{m}_0,$$

where

$$\mathbf{K}_t = (\mathbf{F}_t\mathbf{\Sigma}_{t|t-1}\mathbf{H}_t^T + \mathbf{G}_t\mathbf{C}_t)(\mathbf{H}_t\mathbf{\Sigma}_{t|t-1}\mathbf{H}_t^T + \mathbf{R}_t)^{-1}$$

and

$$\begin{aligned} \mathbf{\Sigma}_{t+1|t} &= \mathbf{F}_t\mathbf{\Sigma}_{t|t-1}\mathbf{F}_t^T \\ &\quad - \mathbf{K}_t(\mathbf{F}_t\mathbf{\Sigma}_{t|t-1}\mathbf{H}_t^T + \mathbf{G}_t\mathbf{C}_t) + \mathbf{G}_t\mathbf{Q}_t\mathbf{G}_t^T \end{aligned}$$

with

$$\mathbf{\Sigma}_{0|-1} = \mathrm{Cov}(\underline{X}_0).$$

5. Consider the model (X_k's are scalars)

$$X_{k+1} = \tfrac{1}{2}X_k + U_k, \quad k = 0, 1, \ldots,$$
$$Y_k = \Theta X_k + V_k, \quad k = 0, 1, \ldots,$$

where $U_0, U_1, \ldots$ are i.i.d. $\mathcal{N}(0, q)$ random variables, $V_0, V_1, \ldots,$ are i.i.d. $\mathcal{N}(0, r)$ random variables, X_0 is a $\mathcal{N}(0, \Sigma_0)$ random variable, and all U_k's, V_k's, and X_0 are independent of one another.

(a) Suppose $\Theta \equiv 1$ and $r = q = 1$. Find the initial state variance Σ_0 such that the optimal prediction filter (i.e., $\hat{X}_{t+1|t}$) is time-invariant. Write the recursion for $\hat{X}_{t+1|t}$ in this case. What is the mean-squared prediction error for this case?

(b) Suppose $r = q = 1$ and Σ_0 is the answer from part (a). What is the structure of the optimum decision rules for deciding $\Theta = 0$ versus $\Theta = 1$ based on observations for $k = 0, 1, \ldots, n$?

6. Consider the standard Kalman-Bucy model with states $\underline{X}_k$ and observations $\underline{Y}_k$. Suppose $0 \le j \le t$ and we wish to estimate $\underline{X}_j$ from $\underline{Y}_0^t$. Consider the estimator defined recursively (in t) by

$$\hat{\underline{X}}_{j|t} = \hat{\underline{X}}_{j|t-1} + \mathbf{K}_t^a(\underline{Y}_t - \mathbf{H}_t\hat{\underline{X}}_{t|t-1})$$

where

$$\mathbf{K}_t^a = \mathbf{\Sigma}_{t|t-1}^a \mathbf{H}_t^T[\mathbf{H}_t\mathbf{\Sigma}_{t|t-1}^a\mathbf{H}_t^T + \mathbf{R}_t]^{-1}$$

and

$$\mathbf{\Sigma}_{t+1|t}^a = \mathbf{\Sigma}_{t|t-1}^a[\mathbf{F}_t - \mathbf{K}_t\mathbf{H}_t]^T$$

with

$$\mathbf{\Sigma}_{j|j-1}^a = \mathbf{\Sigma}_{j|j-1}$$

where $\mathbf{H}_t, \hat{\underline{X}}_{t|t-1}, \mathbf{\Sigma}_{t|t-1}, \mathbf{R}_t, \mathbf{F}_t,$ and $\mathbf{K}_t$ are as in the one-step prediction problem.

(a) Show that $\mathbf{\Sigma}_{t|t-1}^a = E\{(\underline{X}_j - \hat{\underline{X}}_{j|t-1})\underline{X}_t^T\}$.

(b) Show that $\hat{\underline{X}}_{j|t} = E\{\underline{X}_j|\underline{Y}_0^t\}$.

7. Suppose that $\underline{\Theta}, \underline{Z}_1, \underline{Z}_2, \ldots, \underline{Z}_n$, are jointly Gaussian random vectors, and that $\underline{Z}_1, \underline{Z}_2, \ldots, \underline{Z}_n$, are mutually independent and have zero means. Show that the minimum-mean-squared error estimate of $\underline{\Theta}$ given $\underline{Z}_1, \underline{Z}_2, \ldots, \underline{Z}_n$, is given by

$$\hat{\underline{\theta}} = \sum_{k=1}^{n} E\{\underline{\Theta}|\underline{Z}_k\} - (n-1)E\{\underline{\Theta}\}.$$

8. Use the result of Exercise 7 to derive the Kalman-Bucy measurement update formula (V.B.14a).

9. Consider the observation model:

$$Y_k = N_k + \Theta s_k, \quad k = 1, 2, \ldots,$$

where $N_1, N_2, \ldots$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables, $\Theta \sim \mathcal{N}(\mu, \nu^2)$ is independent of $N_1, N_2, \ldots,$ and $s_1, s_2, \ldots$ is a known sequence. Let $\hat{\theta}_n$ denote the MMSE estimate of Θ given $Y_1, \ldots, Y_n$. Find recursions for $\hat{\theta}_n$ and for the minimum mean-squared error, $E\{(\hat{\theta}_n - \Theta)^2\}$, by recasting this problem as a Kalman filtering problem.

10. Consider a sequence $\{X_k\}_{k=0}^\infty$ of binary random variables, each taking on the values 0 and 1. Suppose that the probabilistic structure of this sequence is such that

$$\begin{aligned} P(X_k &= x_k|X_0 = x_0, \ldots, X_{k-1} = x_{k-1}) \\ &= P(X_k = x_k|X_{k-1} = x_{k-1}) \triangleq p_{x_k, x_{k-1}}, \end{aligned}$$

for all integers $k \ge 1$, and for all binary sequences $\{x_k\}_{k=0}^\infty$. Consider the observation model

$$Y_k = X_k + N_k, \quad k = 0, 1, 2, \ldots,$$

where $\{N_k\}_{k=0}^\infty$ is a sequence of independent and identically distributed random variables, independent of $\{X_k\}_{k=0}^\infty$ and having marginal probability density function f. For each integer $t \ge 0$, let $\hat{X}_{t|t}$ denote the minimum-mean-squared-error (MMSE) estimate of X_t, given measurements $Y_0, Y_1, \ldots, Y_t$, and let $\hat{X}_{t|t-1}$ denote the MMSE estimate of X_t, given measurements $Y_0, Y_1, \ldots, Y_{t-1}$, with $\hat{X}_{0|-1} \equiv E\{X_0\}$. Show that $\hat{X}_{t|t}$ and $\hat{X}_{t|t-1}$ satisfy the joint recursion

$$\hat{X}_{t|t} = \frac{\hat{X}_{t|t-1}f(y_t - 1)}{\hat{X}_{t|t-1}f(y_t - 1) + \left(1 - \hat{X}_{t|t-1}\right)f(y_t)}, \quad t \ge 0,$$

and

$$\hat{X}_{t+1|t} = p_{1,1}\hat{X}_{t|t} + p_{1,0}\left(1 - \hat{X}_{t|t}\right), \quad t \ge 0,$$

with initial condition $\hat{X}_{0|-1} = P(X_0 = 1)$. Can you generalize this result to nonbinary sequences satisfying the above property?

11. Verify that the Levinson algorithm given in (V.C.26)-(V.C.29) solves the Yule-Walker equations (V.C.25).

12. Consider the model of Example V.D.1 with $\lambda = 0$. Consider an estimate of S_t given by

$$\hat{S}_t = \sum_{n=-\infty}^{\infty} h_{t-n} Y_n.$$

Show that

$$E\{(\hat{S}_{t+\lambda} - S_{t+\lambda})^2\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |1 - H(\omega)|^2 \phi_S(\omega) d\omega + \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(\omega)|^2 \phi_N(\omega) d\omega.$$

where H is the transfer function corresponding to $\{h_n\}_{n=-\infty}^{\infty}$. Interpret the two terms in this error formula.

13. Show directly that the filter with transfer function

$$H(\omega) = \frac{\phi_S(\omega)}{\phi_S(\omega) + \phi_N(\omega)}$$

minimizes the error expression given in Exercise 12.

VI

Signal Detection in Continuous Time

VI.A Introduction

In the preceding chapters we have presented the basic principles of signal detection and estimation, assuming throughout that our observation set is either a set of vectors or is a discrete set. Throughout this analysis a key role was played by a family of densities $\{p_\theta; \theta \in \Lambda\}$ on the observation space, either through the likelihood ratio in hypothesis testing, through the computation of an *a posteriori* parameter distribution in Bayesian estimation, or through the study of MVUEs and MLEs in nonBayesian parameter estimation. This necessity of specifying a family of densities on the observation space is the primary reason for restricting our observation sets in the way that we have done. In particular, all the problems considered thus far have been treated using the ordinary probability calculus of probability density functions and probability mass functions.[1]

Although the observation sets treated thus far are of considerable interest in practice, there are many applications in which our observations are best modeled as a continuous-time random process. That is, our overall observation Y is a collection of random variables $\{Y_t; t \in [0,T]\}$ indexed by a continuous parameter t, where for convenience we have chosen our observation interval to be $[0,T]$ for some $T > 0$. In this chapter and the following one, we consider signal detection and estimation problems with this type of observation. Signal detection is treated in this chapter, with signal estimation being treated in Chapter VII.

In continuous-time problems, the observation set Γ becomes a set each of whose elements is a continuous-time waveform. Such a set is called a *function space*. In order to model signal detection and estimation problems in this setting, we need to construct families of densities on such sets. Since a density is a function that can be integrated (or summed) to give probabilities, the notion of a density in continuous time requires a method of

[1] An exception is the linear estimation problem treated in Section V.D. Since we needed only a second-order statistical description for this problem, we were in fact able to extend our observation set to include sets of infinite sequences.