Adaptive Filters

V. John Mathews
Scott C. Douglas

Contents

Chapter 2

Linear Estimation Theory

This chapter explores the fundamentals of optimal estimation using linear system models. We develop a common framework for posing and solving linear estimation problems for several different measures of quality. This common framework is based on the notion of vector spaces, and it enables us to visualize many estimation problems using geometrical analogies. The intuitive ideas developed through this approach will prove to be useful in the following chapters.

2.1 The Linear Estimation Problem

As described in Chapter 1, an adaptive filter approximates or estimates one or more signals as a function of one or more other signals. The objective of *linear estimation* problems is to estimate one or more signals as a linear combination of several signals. We assume in our discussion that only one signal needs to be estimated from several others, although the general multiple-signal estimation problem can be handled similarly. In such problems, we estimate a *desired response signal* $d(n)$ as a weighted sum of L *input signals* $x_1(n), x_2(n) \cdots x_L(n)$. Let

$$\hat{d}(n) = \sum_{i=1}^{L} w_i x_i(n) \tag{2.1}$$

denote an estimate of $d(n)$. Our objective is to select the coefficients $w_1, w_2, \cdots, w_L$ such that the estimate $\hat{d}(n)$ is as close to $d(n)$ as possible in some sense. What we mean by "as close ... as possible" is something that we will specify shortly.

2.1.1 Examples of Linear Estimation

The most common class of problems that we will encounter in this book are *single channel linear estimation* problems. In such situations, the objective is to estimate the desired response signal $d(n)$ as a linear combination of certain samples of the input signal. Quite

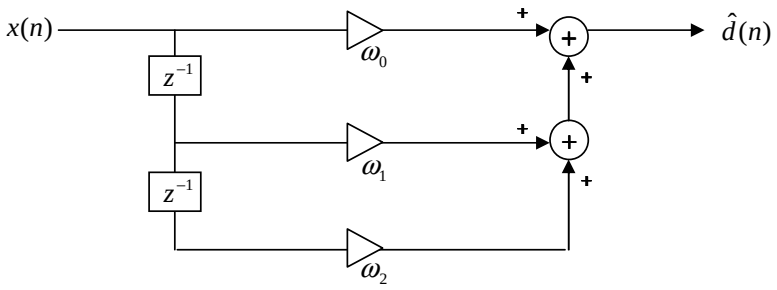

Figure 2.1: A linear, finite-memory system.

commonly, the estimate has the form

$$\hat{d}(n) = \sum_{i=L_1}^{L_2} w_i x(n-i), \tag{2.2}$$

where L_1 and L_2 are integers and $L_2 \geq L_1$. That is, the various signals $x_i(n)$ in (2.1) are formed by samples of the same signal $x(n)$. The estimator is said to be *causal* if $L_1 \geq 0$ since $\hat{d}(n)$ depends only on current and past values of the input signal. Otherwise, the estimator is *non-causal*.

Now, consider a causal estimator of the form

$$\hat{d}(n) = \sum_{i=0}^{L} w_i x(n-i). \tag{2.3}$$

The estimator is said to have *finite memory* if L is a finite number. Otherwise, the estimator has *infinite memory*. In either case, the estimator employs a *linear system model*[1]. Infinite memory estimators do not necessarily require infinite amount of computation at each time instant. They can often be realized using *recursive* structures. A recursive, linear estimator has the form

$$\hat{d}(n) = \sum_{i=0}^{M} b_i x(n-i) + \sum_{i=1}^{L} a_i \hat{d}(n-i). \tag{2.4}$$

By definition, the system of (2.3) is *non-recursive*. Figures 2.1 and 2.2 show block diagrams of finite-memory and recursive linear estimators.

[1]This definition assumes that the coefficients of the filter do not depend on its input signal.

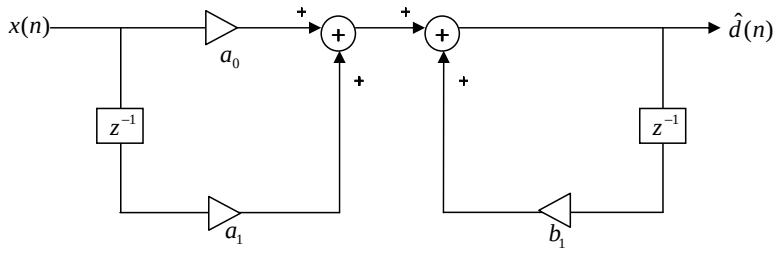

Figure 2.2: A recursive, linear system.

A special case of non-recursive estimation that is of particular interest is when $d(n) = x(n)$ and $L_1 > 0$. The problem of estimating a signal $x(n)$ using its previous samples is known as *prediction*. The objective of one-step linear prediction is to estimate $x(n)$ as

$$\hat{x}(n) = \sum_{i=1}^{L} a_i x(n-i) \tag{2.5}$$

using the most recent L samples of the signal. Figure 2.3 depicts such a system. To distinguish the general estimation problem from the prediction problem, we refer to the former as *joint-process estimation*.

Finally, in *multichannel linear estimation* problems, we attempt to estimate a desired signal $d(n)$ using samples belonging to K input signals $x_1(n), x_2(n), \cdots, x_K(n)$. A non-recursive and causal K-channel linear estimate has the form

$$\hat{d}(n) = \sum_{i=1}^{K} \sum_{j=0}^{L_i} w_{ij} x_i(n-j), \tag{2.6}$$

where $\{w_{ij}, 1 \leq i \leq K, 0 \leq j \leq L_i\}$ denote the parameters of the model.

2.1.2 A Pictorial Introduction to Estimation Theory

Consider the problem described in Figure 2.4. $\mathbf{X_1}$ and $\mathbf{D}$ are two vectors defined on the two-dimensional plane as shown. Suppose that we are interested in finding another vector $\hat{\mathbf{D}}$ which lies in the same direction as $\mathbf{X_1}$ and at the same time is the closest to $\mathbf{D}$. Our notion of what is "closest" is driven by intuition. We desire to have the difference vector $\mathbf{D} - \hat{\mathbf{D}}$ to be of the shortest possible length. We can see from the figure that $\mathbf{D} - \hat{\mathbf{D}}$ has the

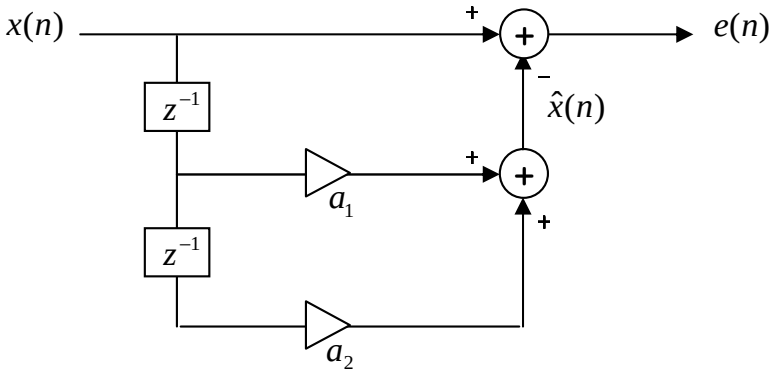

Figure 2.3: A linear predictor.

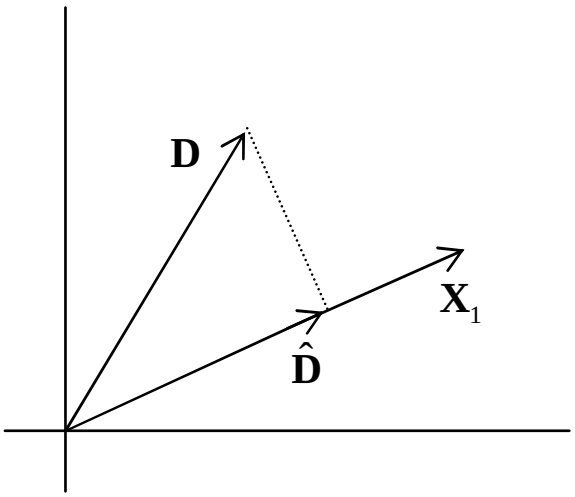

Figure 2.4: The closest vector to $\mathbf{D}$ in the direction of $\mathbf{X_1}$ can be obtained by dropping a perpendicular from $\mathbf{D}$ to $\mathbf{X_1}$.

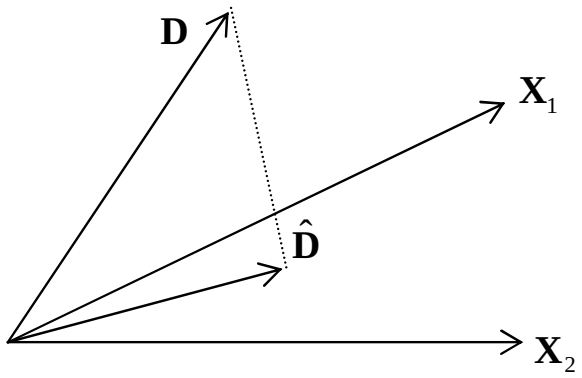

Figure 2.5: The closest vector to $\mathbf{D}$ among all possible linear combinations of $\mathbf{X_1}$ and $\mathbf{X_2}$ can be obtained by dropping a perpendicular from $\mathbf{D}$ to the plane containing $\mathbf{X_1}$ and $\mathbf{X_2}$.

shortest possible length if it is perpendicular to $\mathbf{X_1}$. To prove this formally, consider any other vector $\mathbf{D'}$ that lies in the same direction as $\mathbf{X_1}$. Since the difference vector $\mathbf{D} - \mathbf{D'}$ forms the hypotenuse of a right-angled triangle whose other sides are $\mathbf{D} - \hat{\mathbf{D}}$ and $\mathbf{D'} - \hat{\mathbf{D}}$, $\mathbf{D} - \mathbf{D'}$ is longer than $\mathbf{D} - \hat{\mathbf{D}}$.

Now look at the slightly more complex problem depicted in Figure 2.5. Here, we are interested in finding the vector that lies on a plane defined by two other vectors $\mathbf{X_1}$ and $\mathbf{X_2}$ and is closest to $\mathbf{D}$. This plane contains all vectors in the form $a\mathbf{X_1} + b\mathbf{X_2}$, *i.e.*, all linear combinations of $\mathbf{X_1}$ and $\mathbf{X_2}$. Once again, the closest vector $\hat{\mathbf{D}}$ in the plane defined by the vectors $\mathbf{X_1}$ and $\mathbf{X_2}$ is defined by the intersection of the plane and a perpendicular to the plane dropped from the end point of $\mathbf{D}$. Obviously, the difference vector $\mathbf{D} - \hat{\mathbf{D}}$ is perpendicular to the $(a\mathbf{X_1} + b\mathbf{X_2})$ plane.

Extension of this idea to spatial dimensions of four or more are straightforward, but more difficult to visualize. In such cases the task is to find a vector in the space[2] defined by the set of L vectors $\mathbf{X}_1,\ \mathbf{X}_2,\ \cdots,\ \mathbf{X}_L$ that is closest to the vector $\mathbf{D}$ to be estimated. In other words, we want to estimate $\mathbf{D}$ as

$$\hat{\mathbf{D}} = \sum_{i=1}^{L} w_i \mathbf{X}_i, \tag{2.7}$$

[2]The space defined by all possible vectors that are linear combinations of $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_L$ *i.e.*, all vectors that can be expressed in the form $\sum_{i=1}^{L} a_i \mathbf{X}_i$, is called the *linear span* of $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_L$.

and we choose $w_1, w_2, \cdots, w_L$ so that the error vector $\mathbf{D} - \hat{\mathbf{D}}$ has the shortest possible length. The estimate $\hat{\mathbf{D}}$ is defined by the intersection of the perpendicular line drawn from $\mathbf{D}$ to the space spanned by $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_L$. The error vector lies on this line and is thus perpendicular to this space.

2.1.3 Analytical Solution to the Estimation Problem

Let us return to the problem described in Figure 2.4. Let $\| \mathbf{X_1} \|$ and $\| \mathbf{D} \|$ denote the lengths of the vectors $\mathbf{X_1}$ and $\mathbf{D}$, respectively. Also, let θ be the angle between the two vectors. The best estimate of $\mathbf{D}$ in the direction of $\mathbf{X_1}$ must have the form

$$\hat{\mathbf{D}} = \| \hat{\mathbf{D}} \| \frac{\mathbf{X_1}}{\| \mathbf{X_1} \|}, \tag{2.8}$$

where $\| \hat{\mathbf{D}} \|$ is the length of $\hat{\mathbf{D}}$ and $\frac{\mathbf{X_1}}{\|\mathbf{X_1}\|}$ defines a unit vector (a vector with unit length) in the direction of $\mathbf{X_1}$. Equation (2.8) explicitly states that the direction of the estimate $\hat{\mathbf{D}}$ is determined by the vector $\mathbf{X_1}$**.** The length of the estimate $\hat{\mathbf{D}}$ is yet to be determined. We can show from Figure 2.4 that[3]

$$\| \hat{\mathbf{D}} \| = \| \mathbf{D} \| \cos\theta. \tag{2.9}$$

Substituting (2.9) in (2.8), we get the following expression for $\hat{\mathbf{D}}$:

$$\hat{\mathbf{D}} = \left(\frac{\| \mathbf{D} \|}{\| \mathbf{X_1} \|} \cos\theta \right) \mathbf{X_1}. \tag{2.10}$$

Thus, if we define $\hat{\mathbf{D}} = w\mathbf{X_1}$, the optimal coefficient w for this case is given by

$$w = \frac{\| \mathbf{D} \|}{\| \mathbf{X_1} \|} \cos\theta = \frac{\| \mathbf{D} \| \ \| \mathbf{X_1} \|}{\| \mathbf{X_1} \|^2} \cos\theta. \tag{2.11}$$

We can see from the above discussion that the problem of approximating one vector with a scaled version of another vector can be solved if the lengths of the vectors and the angle between them are known. For problems involving two or more input vectors, this result can be extended in a straightforward manner. In fact, almost all of the estimation problems that we discuss in this book can be viewed from the geometrical perspective illustrated in the problems above.

2.2 Vector Spaces

We formalize the ideas described above by developing the notion of vector spaces. For this, we first define vector spaces and then show that many estimation problems can be viewed as minimization of appropriate functions defined for particular vector spaces.

[3]Note that $\cos\theta$ can be negative. In this case, $\hat{\mathbf{D}}$ is in the opposite direction as $\mathbf{X_1}$.

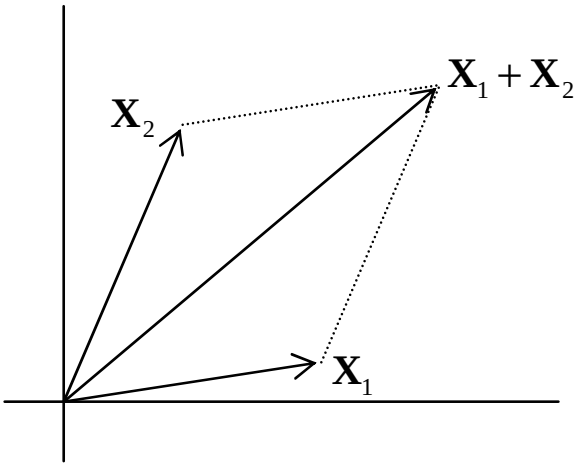

Figure 2.6: Addition of two vectors in a Euclidean space.

The Euclidean Space

The most common example of a vector space (also called a *linear space*) is the Euclidean space. In the Euclidean space, a vector is a point in an L-dimensional space and is uniquely specified by its coordinates. The vectors in this space are represented as

$$\mathbf{X} = [x_1 \; x_2 \; \cdots \; x_L]^T, \tag{2.12}$$

where $x_1, x_2, \cdots, x_L$ are the L coordinates of the vector $\mathbf{X}$. Two operations that can be performed on the vectors in the Euclidean space are addition of vectors denoted by the $+$ sign and scaling of vectors (multiplication of a vector with a scalar constant) usually denoted by the $\cdot$ sign. Figures 2.6 and 2.7 demonstrate these two operations in a two-dimensional space.

2.2.1 Definition of a Vector Space

A vector space is uniquely defined by a set of rules that governs the two operations of addition and scalar multiplication. Analogous to the notation employed for Euclidean spaces, we use $\mathbf{X} + \mathbf{Y}$ to denote the addition of the vectors $\mathbf{X}$ and $\mathbf{Y}$. Similarly, $\gamma \cdot \mathbf{X}$ represents the scalar multiple of the vector $\mathbf{X}$ with γ. We now describe the rules that govern these operations.

Rules of Addition in a Vector Space

(i)

$$\mathbf{X} + \mathbf{Y} = \mathbf{Y} + \mathbf{X} \text{ (commutative law)} \tag{2.13}$$

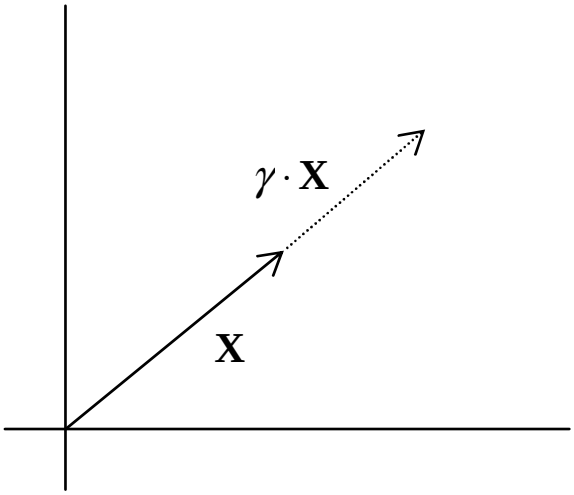

Figure 2.7: Scaling of a vector in a Euclidean space.

(ii)
$$\mathbf{X} + (\mathbf{Y} + \mathbf{Z}) = (\mathbf{X} + \mathbf{Y}) + \mathbf{Z} \quad \text{(associative law)} \tag{2.14}$$

(iii) There exists a zero vector denoted by $\mathbf{0}$ with the property that

$$\mathbf{X} + \mathbf{0} = \mathbf{X}. \tag{2.15}$$

(iv) For every vector $\mathbf{X}$ in the vector space, there exists another vector $(-\mathbf{X})$ such that

$$\mathbf{X} + (-\mathbf{X}) = \mathbf{0} \tag{2.16}$$

Rules of Scalar Multiplication

(i) For each scalar α and vector $\mathbf{X}$, $\alpha \cdot \mathbf{X}$ is a vector such that

$$\beta \cdot (\alpha \cdot \mathbf{X}) = (\alpha\beta) \cdot \mathbf{X} \text{ (associative law)} \tag{2.17}$$

for every scalar β.

(ii)
$$1 \cdot \mathbf{X} = \mathbf{X} \tag{2.18}$$

(iii)
$$0 \cdot \mathbf{X} = \mathbf{0} \tag{2.19}$$

Rules Satisfied by Addition and Multiplication

The following distributive laws must be satisfied by the two operations:

(i)
$$\alpha \cdot (\mathbf{X} + \mathbf{Y}) = \alpha \cdot \mathbf{X} + \alpha \cdot \mathbf{Y} \tag{2.20}$$

(ii)
$$(\alpha + \beta) \cdot \mathbf{X} = \alpha \cdot \mathbf{X} + \beta \cdot \mathbf{X}. \tag{2.21}$$

We leave it as an exercise for the reader to show that addition and multiplication in the Euclidean space satisfy all the above rules. In this case, the properties of the vector space also apply directly to the component elements of the vectors involved.

2.2.2 Inner Products

A special class of vector spaces called *inner product spaces* are of particular interest in estimation problems. In such spaces we define $\langle \mathbf{X}, \mathbf{Y} \rangle$ as a scalar number representing the inner product of the vectors $\mathbf{X}$ and $\mathbf{Y}$. The inner product satisfies the following rules:

(i)
$$\langle \mathbf{X}, \mathbf{Y} + \mathbf{Z} \rangle \;=\; \langle \mathbf{X}, \mathbf{Y} \rangle + \langle \mathbf{X}, \mathbf{Z} \rangle \tag{2.22}$$

(ii)
$$\langle \alpha \cdot \mathbf{X}, \mathbf{Y} \rangle \;=\; \alpha \langle \mathbf{X}, \mathbf{Y} \rangle \tag{2.23}$$

(iii)
$$\langle \mathbf{X}, \mathbf{Y} \rangle \;=\; \langle \mathbf{Y}, \mathbf{X} \rangle^*, \tag{2.24}$$

where * denotes the complex conjugate operation.

(iv) The quantity $\langle \mathbf{X}, \mathbf{X} \rangle$ is a real number and

$$\langle \mathbf{X}, \mathbf{X} \rangle \;\geq\; 0 \tag{2.25}$$

with equality if and only if $\mathbf{X} = \mathbf{0}$.

We now consider several examples of inner product spaces.

Example 2.1: Finite-Dimensional Vectors

Let $\mathbf{X}$ and $\mathbf{Y}$ belong to a complex, L-dimensional vector space such that

$$\mathbf{X} = [x_1 \; x_2 \; \cdots \; x_L]^T$$

and

$$\mathbf{Y} = [y_1 \; y_2 \; \cdots \; y_L]^T \, ,$$

respectively. Then, it can be shown that an inner product defined as

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i=1}^{N} x_i y_i^*$$

satisfies all the rules associated with inner products. Consequently, the N-dimensional vector space with the inner product as defined above is an inner product space.

Example 2.2: Finite-Dimensional Vectors

Let $\mathbf{X}$ and $\mathbf{Y}$ belong to the Euclidean vector space of real N-dimensional vectors and let $\mathbf{W}$ be a symmetric positive definite $N \times N$ matrix. Then,

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \mathbf{X}^T \mathbf{W} \mathbf{Y}$$

is also a properly defined inner product for the Euclidean space.

Example 2.3: Finite-Energy Signals

Let $x(n)$ and $y(n)$ be discrete-time (real or complex) signals of possibly infinite duration, but of finite energy.[4] We can express these signals as infinitely-long vectors of the form

$$\mathbf{X} = \begin{bmatrix} \vdots \\ x(n-1) \\ x(n) \\ x(n+1) \\ \vdots \end{bmatrix}$$

and

$$\mathbf{Y} = \begin{bmatrix} \vdots \\ y(n-1) \\ y(n) \\ y(n+1) \\ \vdots \end{bmatrix},$$

[4] A signal $x(n)$ possesses finite energy if $\sum_{n=-\infty}^{\infty} |x(n)|^2 < \infty$.

respectively. Note that vector addition and scalar multiplication of vectors as we know satisfy all the rules of addition and multiplication, and therefore we have a properly defined vector space. Furthermore,

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{n=-\infty}^{\infty} x(n)y^*(n)$$

is a well-defined inner product for this space.

Example 2.4: Wide-Sense Stationary Random Processes

Let $x(n)$ and $y(n)$ belong to the class of jointly wide-sense stationary and discrete-time random processes with finite covariances. By defining infinite-dimensional random vectors $\mathbf{X}$ and $\mathbf{Y}$ of the form in Example 2.3, it can be shown that

$$\langle \mathbf{X}, \mathbf{Y} \rangle = E\{x(n)y^*(n)\}$$

satisfies all the rules of a properly defined inner product.

The Length of a Vector

The concept of an inner product gives rise to the notion of the length of a vector. The length[5] $\| \mathbf{X} \|$ of a vector $\mathbf{X}$ is defined as

$$\| \mathbf{X} \| = \sqrt{\langle \mathbf{X}, \mathbf{X} \rangle}. \tag{2.26}$$

This quantity is also known as the *norm* of $\mathbf{X}$. In the usual Euclidean N-dimensional vector space of Example 2.1, the norm is defined as

$$\| \mathbf{X} \| = \sqrt{\sum_{i=1}^{L} |x_i|^2}, \tag{2.27}$$

which is also the definition of the length of the vector $\mathbf{X}$. Analogous to this, we interpret the norm $\| \mathbf{X} \|$ as the length of the vector $\mathbf{X}$ in any properly defined inner product space. The concept of the norm of a vector can be used to define a distance measure between two vectors. The distance between two vectors $\mathbf{X}$ and $\mathbf{Y}$ is defined as the length of the difference between the two vectors as

$$\mathcal{D}(\mathbf{X}, \mathbf{Y}) = \| \mathbf{X} - \mathbf{Y} \| . \tag{2.28}$$

Example 2.5: Distance Measures in the Euclidean Space and the Space of Stationary Random Processes

[5]This definition of length is different from the MATLAB command `length`, where it simply denotes the number of elements in the input vector.

The distance measure associated with the norm defined as

$$\langle \mathbf{X}, \mathbf{Y} \rangle \;=\; \sum_{i=1}^{L} x_i y_i$$

for a Euclidean space containing real-valued, L-dimensional vectors is

$$\mathcal{D}(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^{L} |x_i - y_i|^2}.$$

The above distance measure is the same as the familiar Euclidean distance measure. The corresponding distance measure for a space of random vectors in which the inner product of two vectors is defined using the cross-correlation of the random variables is

$$\mathcal{D}(\mathbf{X}, \mathbf{Y}) = \sqrt{E\{|x(n) - y(n)|^2\}}.$$

Unless otherwise stated, we will assume in what follows that all vectors are real-valued.

The Angle Between Two Vectors

Consider a normalized inner product of two vectors $\mathbf{X}$ and $\mathbf{Y}$, defined as

$$\rho = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\| \mathbf{X} \| \;\; \| \mathbf{Y} \|}. \tag{2.29}$$

This quantity can be related to the concept of the angle between two vectors. We first show that $|\rho| \leq 1$ for all properly defined inner products. To see this, consider

$$\begin{aligned} \| \mathbf{X} - \alpha \cdot \mathbf{Y} \|^2 &= \| \mathbf{X} \|^2 + \alpha^2 \| \mathbf{Y} \|^2 - \alpha \langle \mathbf{X}, \mathbf{Y} \rangle - \alpha \langle \mathbf{Y}, \mathbf{X} \rangle \\ &= \| \mathbf{X} \|^2 + \alpha^2 \| \mathbf{Y} \|^2 - 2\alpha \langle \mathbf{X}, \mathbf{Y} \rangle, \end{aligned} \tag{2.30}$$

where α is an arbitrary, real constant and we have used the relation

$$\langle \mathbf{X}, \mathbf{Y} \rangle \;=\; \langle \mathbf{Y}, \mathbf{X} \rangle \tag{2.31}$$

(see (2.24)) for real vectors. Since the norm of a vector is always nonnegative, we have from (2.30) that

$$\| \mathbf{X} \|^2 + \alpha^2 \| \mathbf{Y} \|^2 - 2\alpha \langle \mathbf{X}, \mathbf{Y} \rangle \;\geq\; 0. \tag{2.32}$$

In particular, the inequality in (2.32) holds for the minimum value of $\| \mathbf{X} - \alpha \cdot \mathbf{Y} \|^2$. The reader can easily verify that

$$\alpha = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\| \mathbf{Y} \|^2} \tag{2.33}$$

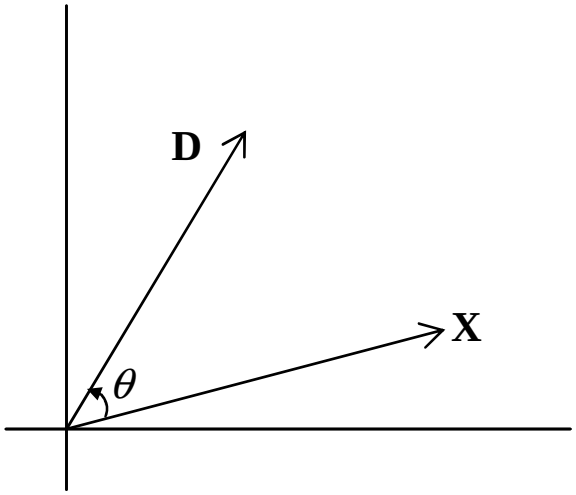

Figure 2.8: Angle between **X** and **D** in a Euclidean space matches the standard definition of the angle between vectors.

minimizes (2.30). Substituting this value in (2.30) gives

$$\| \mathbf{X} \|^2 - \frac{\langle \mathbf{X}, \mathbf{Y} \rangle^2}{\| \mathbf{Y} \|^2} \geq 0, \tag{2.34}$$

which implies that

$$\rho^2 = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle^2}{\| \mathbf{X} \|^2 \ \| \mathbf{Y} \|^2} \leq 1, \tag{2.35}$$

proving our result.

Since $|\rho| \leq 1$, it is traditional to use a geometric interpretation for ρ by assigning $\rho = \cos \theta$. Obviously, θ can be interpreted as the angle between the two vectors. In the Euclidean space, $\cos^{-1}\rho$ indeed defines the angle between the two vectors **X** and **Y**. The inequality in (2.35) is a form of the *Cauchy-Schwartz inequality*, which is usually expressed as

$$\langle \mathbf{X}, \mathbf{Y} \rangle^2 \leq \| \mathbf{X} \|^2 \| \mathbf{Y} \|^2 . \tag{2.36}$$

Example 2.6: Angle Measurement in the Euclidean Space

Consider a real-valued, N-dimensional vector space with inner product defined as in Example 2.1. The angle between the two vectors **X** and **D** shown in Figure 2.8 is defined by

$$\cos\theta = \frac{\sum_{i=1}^{L} x_i d_i}{\sqrt{\sum_{i=1}^{L} x_i^2} \sqrt{\sum_{i=1}^{L} d_i^2}}.$$

The above definition results in the usual interpretation of the angle between two vectors in the Euclidean space.

Example 2.7: Angle Measurement in the Space of Stationary Random Variables

Let X and D belong to the space of real-valued and stationary random variables with finite variance. Also, let the inner product be defined as

$$\langle X, D \rangle = E\{XD\}.$$

Then, the angle between the two processes is given by the correlation coefficients of the two random variables defined as

$$\rho = \frac{E\{XD\}}{\sqrt{E\{X^2\}}\sqrt{E\{D^2\}}}.$$

Geometric Interpretation of the Inner Product

We can show that the inner product of two vectors is a measure of the similarity between the two vectors. Substituting $\rho = \cos\theta$ in (2.29) and cross-multiplying, we see that

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \| \mathbf{X} \| \ \ \| \mathbf{Y} \| \cos\theta. \tag{2.37}$$

Referring to Figure 2.9, we notice that $\| \mathbf{Y} \| \cos\theta$ is actually the length of the component of $\mathbf{Y}$ in the direction of $\mathbf{X}$. Thus, the inner product is simply the product of the length of one of the vectors and the length of the projection of the other vector onto the first one. Note that the inner product is maximum when the two vectors are aligned in the same direction, *i.e.*, when the angle between the two vectors is zero. Similarly, the inner product is zero when the two vectors are perpendicular to each other.

2.2.3 Orthogonal Vectors

Two vectors $\mathbf{X}$ and $\mathbf{Y}$ in a given inner product space are termed *orthogonal* or *perpendicular* to each other if $\langle \mathbf{X}, \mathbf{Y} \rangle = 0$. Note that if $\langle \mathbf{X}, \mathbf{Y} \rangle = 0$ then it follows from the definition of ρ in equation (2.29) that the angle between $\mathbf{X}$ and $\mathbf{Y}$ is $\pm 90^o$. Thus the notion of perpendicularity of such vectors is appropriate in our geometric interpretation. The concept of orthogonality is crucial in estimation theory.

Example 2.8: Orthogonal Vectors in a Three-Dimensional Euclidean Space

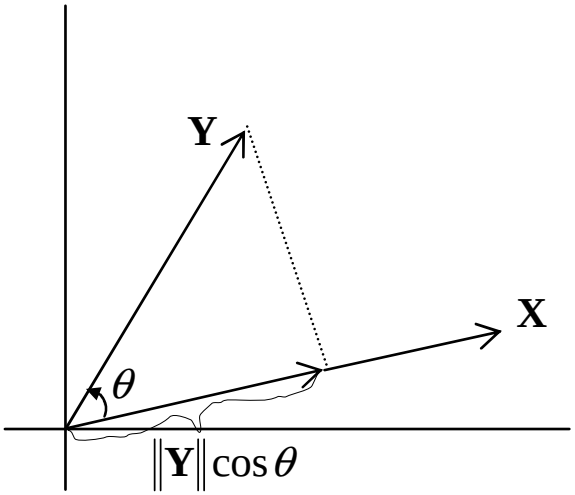

Figure 2.9: Illustration of the geometric interpretation of inner products.

Consider a Euclidean space of real-valued, three-dimensional vectors with the inner product defined as

$$\langle \mathbf{X}, \mathbf{Y} \rangle \; = \; \sum_{i=1}^{3} x_i y_i.$$

Then, the vectors

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \text{ and } \quad \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

are mutually orthogonal since

$$\langle \mathbf{e}_1, \mathbf{e}_2 \rangle \; = \; \langle \mathbf{e}_1, \mathbf{e}_3 \rangle \; = \; \langle \mathbf{e}_2, \mathbf{e}_3 \rangle \; = 0.$$

These three vectors are displayed in Figure 2.10. It is evident from the figure that there are many sets of vectors that are orthogonal to each other. Another example of three mutually orthogonal vectors is

$$\mathbf{X}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \mathbf{X}_2 = \begin{bmatrix} -1 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} \text{ and } \quad \mathbf{X}_3 = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}.$$

These vectors are shown in Figure 2.11. Note that in both cases, the angle between any two vectors in the set is 90° even though the latter set of vectors do not lie on the principal axes of the three-dimensional plane.

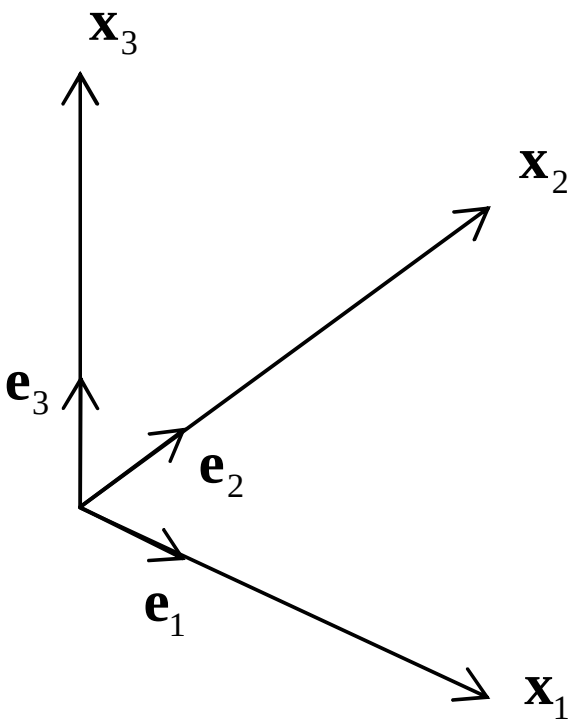

Figure 2.10: The vectors $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ are mutually orthogonal.

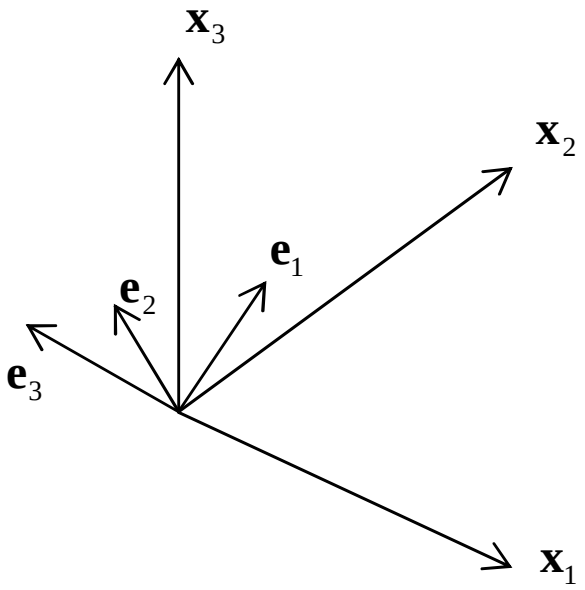

Figure 2.11: Three mutually orthogonal vectors that do not lie on the principal axes.

Example 2.9: Orthogonality in the Space of Stationary Random Variables

In a vector space of zero-mean, wide-sense stationary and real-valued random processes with inner products defined as

$$\langle \mathbf{X}, \mathbf{Y} \rangle \;=\; E\{x(n)y(n)\},$$

two vectors are orthogonal to each other if and only if the random processes corresponding to them are uncorrelated with each other at all times n.

2.3 Linear Estimation in Inner Product Spaces

We now pose and solve the linear estimation problem in a general framework using the vector space concepts and then discuss the two important cases of linear minimum mean-square estimation and least-squares estimation. Our objective is to demonstrate that all linear estimation problems can be solved using a general framework based on vector space concepts. These results can also be interpreted from a geometric point of view.

In the most general sense, a linear estimator approximates a vector $\mathbf{D}$ as a linear combination of L other vectors $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_L$ such that the squared norm of the estimation error vector is minimized. The norm must be defined with respect to some arbitrary but well-defined inner product. Let

$$\hat{\mathbf{D}} = \sum_{i=1}^{L} w_i \mathbf{X}_i \tag{2.38}$$

denote the estimate of $\mathbf{D}$. Our objective is to select the coefficients $w_1, w_2, \cdots, w_L$ so that

$$\mathcal{D}^2(\mathbf{D}, \hat{\mathbf{D}}) \;= \| \mathbf{D} - \sum_{i=1}^{L} w_i \mathbf{X}_i \|^2 \tag{2.39}$$

is minimized.

2.3.1 The Orthogonality Principle

The key to solving the estimation problem posed above is a strong relationship that exists between the optimal estimation error vector and the input vectors. We now state and prove this important result in estimation theory, known as the *orthogonality principle*.

Theorem: Consider the estimation problem described above. Then, the optimal estimation error vector is orthogonal to $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_L$, *i.e.*, for all $1 \leq i \leq L$,

$$\langle \mathbf{D} - \hat{\mathbf{D}}, \mathbf{X}_i \rangle = 0. \tag{2.40}$$

REMARK 2.1: The orthogonality principle states that the optimal minimum squared-norm error estimate of $\mathbf{D}$ can be found by determining the error vector $\mathbf{D}-\hat{\mathbf{D}}$ that is perpendicular to all of $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_L$. The above solution is identical to the geometrical solution obtained in Section 2.1. Consequently, linear estimation problems can be solved by finding a vector that passes through $\mathbf{D}$ and is perpendicular to the linear span $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_L$. The intersection of this perpendicular vector with the span of $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_L$ defines the estimate of $\mathbf{D}$.

We now prove the orthogonality principle formally. The proof assumes that all the vectors involved are real-valued. The proof for complex vectors is similar and is left as an exercise.

Proof of the Orthogonality Principle

We begin by expanding the right hand side of (2.39) using the definition of the norm in (2.26). This operation results in

$$\mathcal{D}^2(\mathbf{D}, \hat{\mathbf{D}}) = \| \mathbf{D} \|^2 + \sum_{i=1}^{L}\sum_{j=1}^{L} w_i w_j \langle \mathbf{X}_i, \mathbf{X}_j \rangle - 2\sum_{i=1}^{L} w_i \langle \mathbf{D}, \mathbf{X}_i \rangle. \tag{2.41}$$

Clearly, the function $\mathcal{D}^2(\mathbf{D}, \hat{\mathbf{D}})$ is quadratic in the coefficients $w_1, w_2, \cdots, w_L$. Furthermore, since the squared-distance measure $\mathcal{D}^2(\cdot, \cdot)$ is non-negative, the function has a unique minimum value. The optimal values of the coefficients corresponding to this minimum can be found by differentiating (2.41) with respect to each of the coefficients and setting the resulting equations to zero. This operation gives

$$2\sum_{i=1}^{L} w_{opt,i} \langle \mathbf{X}_i, \mathbf{X}_j \rangle - \ 2\langle \mathbf{D}, \mathbf{X}_j \rangle \ = 0; \ j = 1, 2, \cdots, L \tag{2.42}$$

We have used $w_{opt,i}$ in the above equation to indicate that the coefficients satisfying (2.42) correspond to the optimal solution. Rearranging the left-side of (2.42) results in

$$\langle \left(\mathbf{D} - \sum_{i=1}^{L} w_{opt,i} \mathbf{X}_i \right), \mathbf{X}_j \rangle \ = 0; \ j = 1, 2, \cdots, L, \tag{2.43}$$

which is the desired result since the second vector within the inner product is indeed the optimal estimation error vector. Note that this result is valid for all inner product spaces and therefore can be applied to all estimation problems that can be formulated as minimization of squared-distance measures in appropriate inner product spaces.

2.3.2 The Optimal Linear Estimator

Examining (2.43), we note that the optimal coefficient values for the problem satisfy a set of L linear equations. Let

$$\mathbf{W}_{opt} = [w_{opt,1}\, w_{opt,2}\, \ldots\, w_{opt,L}]^T \tag{2.44}$$

denote the optimal coefficient vector for the linear estimation problem. Let us also define an $L \times L$ matrix $\mathbf{R_{XX}}$ as

$$\mathbf{R_{XX}} = \begin{bmatrix} \langle \mathbf{X}_1, \mathbf{X}_1 \rangle & \langle \mathbf{X}_1, \mathbf{X}_2 \rangle & \cdots & \langle \mathbf{X}_1, \mathbf{X}_L \rangle \\ \langle \mathbf{X}_2, \mathbf{X}_1 \rangle & \langle \mathbf{X}_2, \mathbf{X}_2 \rangle & \cdots & \langle \mathbf{X}_2, \mathbf{X}_L \rangle \\ \vdots & \vdots & & \vdots \\ \langle \mathbf{X}_L, \mathbf{X}_1 \rangle & \langle \mathbf{X}_L, \mathbf{X}_2 \rangle & \cdots & \langle \mathbf{X}_L, \mathbf{X}_L \rangle \end{bmatrix} \tag{2.45}$$

and an $L \times 1$ vector $\mathbf{P_{DX}}$ as

$$\mathbf{P_{DX}} = \left[\langle \mathbf{D}, \mathbf{X}_1 \rangle \, \langle \mathbf{D}, \mathbf{X}_2 \rangle \cdots \langle \mathbf{D}, \mathbf{X}_L \rangle \right]^T. \tag{2.46}$$

For reasons that will become obvious in the next section, we refer to $\mathbf{R_{XX}}$ as the autocorrelation matrix of the input vectors and $\mathbf{P_{DX}}$ as the cross-correlation vector of $\mathbf{D}$ and the input vectors.

We can rewrite (2.42) using (2.45) and (2.46) as

$$\mathbf{R_{XX}} \mathbf{W}_{opt} = \mathbf{P_{DX}}. \tag{2.47}$$

The set of equations in (2.47) is known as the *normal equations* for the optimization problem defined by the minimization of (2.39). Assuming that the matrix $\mathbf{R_{XX}}$ can be inverted, the optimal solution is

$$\mathbf{W}_{opt} = \mathbf{R_{XX}^{-1}} \mathbf{P_{DX}}. \tag{2.48}$$

The Minimum Squared-Norm of the Estimation Error

The squared-norm of the estimation error vector $\mathbf{D} - \hat{\mathbf{D}}$ can be found by substituting (2.48) in (2.41). Equation (2.41) can be written using matrix notation as

$$\mathcal{D}^2(\mathbf{D}, \hat{\mathbf{D}}) = \| \mathbf{D} \|^2 + \mathbf{W}_{opt}^T \, \mathbf{R_{XX}} \, \mathbf{W}_{opt} - 2\mathbf{W}_{opt}^T \, \mathbf{P_{DX}}. \tag{2.49}$$

Substituting $\mathbf{P_{DX}}$ for $\mathbf{R_{XX}} \mathbf{W}_{opt}$ and denoting the minimum squared-norm value with $\mathcal{D}_{min}(\mathbf{D}, \hat{\mathbf{D}})$, we get

$$\begin{aligned} \mathcal{D}_{min}^2(\mathbf{D}, \hat{\mathbf{D}}) &= \| \mathbf{D} \|^2 - \mathbf{W}_{opt}^T \mathbf{P_{DX}} \\ &= \| \mathbf{D} \|^2 - \mathbf{P}_{\mathbf{DX}}^T \mathbf{R}_{\mathbf{XX}}^{-1} \mathbf{P_{DX}}. \end{aligned} \tag{2.50}$$

2.4 Some Special Cases of Linear Estimation

In this section we consider some specific examples of linear estimation. In each case, the general theory can be applied in a straightforward manner.

2.4.1 Linear, Minimum Mean-Squared Error (MMSE) Estimation of Random Variables

Consider the task of estimating a real and scalar random variable D as a linear combination of L other random variables $X_1, X_2, X_3, \cdots, X_L$ such that the mean-squared value of the estimation error is minimized. In other words, we seek the coefficients $w_1, w_2, \cdots, w_L$ such that

$$J = E\left\{\left(D - \sum_{i=1}^{L} w_i X_i\right)^2\right\} \tag{2.51}$$

is minimized among all possible choices of the coefficients $\{w_i\}$.

Vector Space Formulation

Each of the random variables $D, X_1, X_2, \cdots, X_L$ can be considered as a one-dimensional vector. It is easy to show that $E\{X_i X_j\}$ is a well-defined inner product of X_i and X_j in this space. Consequently, from (2.48), the optimal coefficient vector $\mathbf{W}_{opt}$is given by

$$\mathbf{W}_{opt} = \mathbf{R}_{\mathbf{XX}}^{-1}\mathbf{P}_{\mathbf{DX}}, \tag{2.52}$$

where

$$\mathbf{R}_{\mathbf{XX}} = E\left\{[X_1\, X_2\, \cdots\, X_L]^T\,[X_1\, X_2\, \cdots\, X_L]\right\} \tag{2.53}$$

is the autocorrelation matrix of the random variables $X_1, X_2, \ldots, X_L$ and

$$\mathbf{P}_{\mathbf{DX}} = E\left\{D\,[X_1\, X_2\, \cdots\, X_L]^T\right\} \tag{2.54}$$

is the cross-correlation vector of D and the random variables $X_1, X_2, \cdots, X_L$.

Example 2.10: Orthogonality in the Space of Stationary Random Variables

Let X, Y and D be jointly Gaussian-distributed random variables with zero mean values. The joint probability density function of X, Y and D is given by

$$f_{X,Y,D}(x,y,d) = \frac{1}{(2\pi)^{\frac{3}{2}}\det^{\frac{1}{2}}(\mathbf{C})}\ \exp\left(-\frac{1}{2}[x\,y\,d]\mathbf{C}^{-1}[x\,y\,d]^T\right)$$

where $\mathbf{C}$ is the covariance matrix given by

$$\mathbf{C} = \begin{bmatrix} c_{xx} & c_{xy} & c_{xd} \\ c_{yx} & c_{yy} & c_{yd} \\ c_{dx} & c_{dy} & c_{dd} \end{bmatrix}$$

with c_{uv} defined as

$$c_{uv} = E\{(U - E\{U\})(V - E\{V\})\}$$

for arbitrary random variables U and V. Note that since all the random variables are zero-mean quantities, their autocorrelation and covariance matrices are identical. In this example, let

$$\mathbf{C} = \begin{bmatrix} 1 & 0.5 & 0.2 \\ 0.5 & 1 & 0.5 \\ 0.2 & 0.5 & 1 \end{bmatrix}.$$

Find w_1 and w_2 such that

$$\hat{D} = w_1 X + w_2 Y$$

is the linear MMSE estimate of D as a function of X and Y. In addition, find the value of the minimum mean-squared estimation error.

Solution: We can use (2.48) and (2.50) directly here. Using (2.48), we get

$$\begin{aligned} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} &= \begin{bmatrix} c_{xx} & c_{xy} \\ c_{yx} & c_{yy} \end{bmatrix}^{-1} \begin{bmatrix} c_{dx} \\ c_{dy} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0.2 \\ 0.5 \end{bmatrix} = \begin{bmatrix} \frac{-1}{15} \\ \frac{8}{15} \end{bmatrix}. \end{aligned}$$

Let ξ_{min} denote the minimum mean-squared estimation error for this problem. Substituting the above numerical values for the parameters and the covariance matrix in (2.50) gives

$$\begin{aligned} \xi_{min} &= E\{D^2\} - [w_1 \; w_2] \begin{bmatrix} c_{dx} \\ c_{dy} \end{bmatrix} \\ &= 1 - \begin{bmatrix} \frac{-1}{15} & \frac{8}{15} \end{bmatrix} \begin{bmatrix} .2 \\ .5 \end{bmatrix} = \frac{56}{75}. \end{aligned}$$

2.4.2 Linear, MMSE Estimation of Random Processes

Let $x(n)$ and $d(n)$ be real-valued, jointly wide sense stationary random processes. Consider the problem of estimating $d(n)$ as a linear combination of the most recent L samples of $x(n)$, *i.e.*, we want to find the coefficients $w_0, w_1, w_2, \cdots, w_{L-1}$ such that

$$\hat{d}(n) = \sum_{i=0}^{L-1} w_i x(n-i) \tag{2.55}$$

is the closest to $d(n)$ in the mean-squared error sense.

Vector Space Formulation

Define infinite dimensional vectors $\mathbf{X}_0$, $\mathbf{X}_1$, $\mathbf{X}_2$,, $\cdots$, $\mathbf{X}_{L-1}$ and $\mathbf{D}_0$ as

$$\mathbf{X}_0 = \begin{bmatrix} \vdots \\ x(n-1) \\ x(n) \\ x(n+1) \\ \vdots \end{bmatrix}, \quad \mathbf{X}_k = \begin{bmatrix} \vdots \\ x(n-1-k) \\ x(n-k) \\ x(n+1-k) \\ \vdots \end{bmatrix}; \quad k = 0, 1, \cdots, L-1$$

and

$$\mathbf{D}_0 = \begin{bmatrix} \vdots \\ d(n-1) \\ d(n) \\ d(n+1) \\ \vdots \end{bmatrix}, \tag{2.56}$$

respectively. Note that $\mathbf{X}_k$ corresponds to a vector whose elements are shifted by k samples with respect to those of the vector $\mathbf{X}_0$. As we saw in Example 2.4,

$$\langle \mathbf{X}_i, \mathbf{X}_j \rangle \ = E\{x(n-i)x(n-j)\} \tag{2.57}$$

is a well-defined inner product for this vector space. We can now reformulate the problem as one of obtaining the coefficients $w_0, w_1, w_2, \cdots, w_{L-1}$ of the estimate

$$\hat{\mathbf{D}_0} = \sum_{i=0}^{L-1} w_i \mathbf{X}_i \tag{2.58}$$

so that the squared-norm of the error given by

$$\| \mathbf{D}_0 - \hat{\mathbf{D}_0} \|^2 = E\left\{ (d(n) - \hat{d}(n))^2 \right\} \tag{2.59}$$

is minimized.

Optimal Solution to the Problem

As before, the optimal coefficient vector can be found by substituting the appropriate inner product values in equation (2.48). This operation results in

$$\mathbf{W}_{opt} = \mathbf{R}_{\mathbf{XX}}^{-1} \mathbf{P}_{\mathbf{DX}}, \tag{2.60}$$

where

$$\mathbf{R_{xx}} = \begin{bmatrix} E\{x^2(n)\} & E\{x(n)x(n-1)\} & \cdots & E\{x(n)x(n-L+1)\} \\ E\{x(n-1)x(n)\} & E\{x^2(n-1)\} & \cdots & E\{x(n-1)x(n-L+1)\} \\ \vdots & \vdots & \vdots & \vdots \\ E\{x(n-L+1)x(n)\} & E\{x(n-L+1)x(n-1)\} & \cdots & E\{x^2(n-L+1)\} \end{bmatrix} \tag{2.61}$$

and

$$\mathbf{P_{DX}} = \begin{bmatrix} E\{d(n)x(n)\} \\ E\{d(n)x(n-1)\} \\ \vdots \\ E\{d(n)x(n-L+1)\} \end{bmatrix}. \tag{2.62}$$

Recall that $x(n)$ and $d(n)$ are jointly wide sense stationary random processes. As defined in Appendix B, denote the autocorrelation of $x(n)$ and cross-correlation of $x(n)$ and $d(n)$ by

$$E\{x(n)x(n-k)\} = r_{xx}(k) \tag{2.63}$$

and

$$E\{d(n)x(n-k)\} = r_{dx}(k), \tag{2.64}$$

respectively. Then, the solution is

$$\mathbf{W}_{opt} = \begin{bmatrix} r_{xx}(0) & r_{xx}(1) & r_{xx}(2) & \cdots & r_{xx}(L-1) \\ r_{xx}(1) & r_{xx}(0) & r_{xx}(1) & \cdots & r_{xx}(L-2) \\ r_{xx}(2) & r_{xx}(1) & r_{xx}(0) & \cdots & r_{xx}(L-3) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{xx}(L-1) & r_{xx}(L-2) & r_{xx}(L-3) & \cdots & r_{xx}(0) \end{bmatrix}^{-1} \begin{bmatrix} r_{dx}(0) \\ r_{dx}(1) \\ r_{dx}(2) \\ \vdots \\ r_{dx}(L-1) \end{bmatrix}. \tag{2.65}$$

One very important property of the above solution is that $\mathbf{R_{xx}}$, which is the $L \times L$ autocorrelation matrix of $X(n)$, is a Toeplitz matrix. Because of this fact we can derive efficient algorithms to compute the optimal solution. This will be discussed in Chapter 3. The set of equations in (2.65) is known as the *Wiener-Hopf* equations.

2.4.3 Applications of MMSE Estimation

Before we discuss some properties of MMSE estimators, we consider two applications.

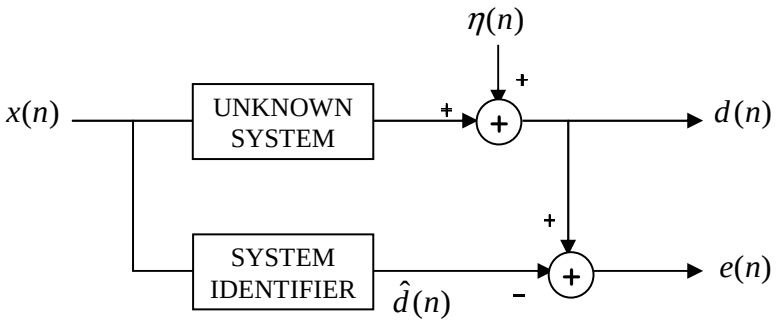

Figure 2.12: The system identification problem.

System Identification

The objective of system identification is to estimate the input-output relationship of an unknown system from knowledge of the statistics of its input signal $x(n)$ and its output signal $d(n)$ as depicted in Figure 2.12. The statistics of the input and output signals are usually unknown, and therefore must be estimated from measurements of the input and output signals. We will address this important issue later in the context of least-squares estimation. In practice, it is impossible to measure the output of the unknown system with perfect accuracy. This is partly due to the inherent noise present in the measurement instruments and other interferences in the environment. There may also be discrepancies between the unknown system and the system model. For simplicity, we model such distortions and measurement noises present in the output signal by an additive random noise sequence $\eta(n)$. For our discussion, we assume that the unknown system is linear and time-invariant. We also assume that the unknown system can be adequately represented by a causal, finite impulse response filter with L coefficients. We will discuss ways of estimating the order of the system model shortly. Our approach to the identification of the unknown system is to estimate $d(n)$ as a linear combination of the most recent L samples of $x(n)$ so that the mean-squared estimation error $E\{e^2(n)\}$, given by

$$E\{e^2(n)\} = E\{(d(n) - \hat{d}(n))^2\} \tag{2.66}$$

is minimized. The coefficients of the estimator are identical to those of the unknown system if i) the measurement noise $\eta(n)$ is uncorrelated with the output $y(n)$ of the unknown system, ii) the joint second-order statistics of $x(n)$ and $d(n)$ are exactly known, and iii) the model order of the estimator is equal to or greater than that of the unknown system. It is left as an exercise to the reader to show that the minimum mean-squared estimation error for the

system in Figure 2.12 is exactly the mean-squared value of the measurement noise $E\{\eta^2(n)\}$ when all of the above three conditions are satisfied. Since the problem is formulated exactly as in standard MMSE estimation, its solution is also given by equation (2.60).

Example 2.11: Orthogonality in the Space of Stationary Random Variables

Let the impulse response function of the unknown system in Figure 2.12 be

$$h(n) = \begin{cases} 1 & ; \quad n = 0 \\ -1 & ; \quad n = 1 \\ 0.5 & ; \quad n = 2 \\ 0 & ; \quad \text{otherwise} \end{cases}$$

We wish to identify this system using its input-output signal statistics. The input signal has zero mean value and its autocorrelation function is given by

$$r_{xx}(k) = 0.8^{|k|}.$$

The measurement noise sequence $\eta(n)$ is an i.i.d. sequence with zero mean and variance $\sigma_\eta^2 = 0.1$ and is independent of the input signal sequence. We use a model order $L = 3$ so that the estimator structure exactly matches that of the unknown system.

The 3×3-element autocorrelation matrix of the input signal is given by

$$\mathbf{R_{xx}} = \begin{bmatrix} 1.0 & 0.8 & 0.64 \\ 0.8 & 1.0 & 0.8 \\ 0.64 & 0.8 & 1.0 \end{bmatrix}.$$

Now,

$$d(n) = x(n) - x(n-1) + 0.5x(n-2) + \eta(n)$$

The cross-correlation of $d(n)$ and $x(n)$ is given by

$$\begin{aligned} r_{dx}(k) &= E\{d(n)x(n-k)\} \\ &= E\{[x(n) - x(n-1) + 0.5x(n-2) + \eta(n)]\, x(n-k)\} \\ &= r_{xx}(k) - r_{xx}(k-1) + 0.5r_{xx}(k-2). \end{aligned}$$

We made use of the fact that $\eta(n)$ and $x(n)$ are independent processes in deriving the above result. We can evaluate $r_{dx}(k)$ by substituting the numerical values for the autocorrelation function in the expression for the cross-correlation function. This operation results in

$$r_{dx}(k) = \begin{cases} 0.52 & ; \quad k = 0 \\ 0.2 & ; \quad k = 1 \\ 0.34 & ; \quad k = 2 \end{cases}$$

Substituting the relevant values in (2.60) we get the optimal solution as

$$\mathbf{W}_{opt} = \begin{bmatrix} 1.0 & 0.8 & 0.64 \\ 0.8 & 1.0 & 0.8 \\ 0.64 & 0.8 & 1.0 \end{bmatrix}^{-1} \begin{bmatrix} 0.52 \\ 0.2 \\ 0.34 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 0.5 \end{bmatrix}$$

Since the coefficients of the identified system and the unknown system are identical, we can see that the MMSE error sequence $e(n)$ and the measurement noise $\eta(n)$ are identical. Consequently, the MMSE value is given by

$$\xi_{min} = E\{\eta^2(n)\} = 0.1.$$

We leave it to the reader to verify this by using (2.50).

Model Order Selection

The problem of selecting the model order in estimation problems is in general difficult and problem-dependent. It is typical to employ prior knowledge of the task, including the physical characteristics of the system that produces the signals, to aid in the selection of the system model as well as the model order. Often, a trial-and-error method is used, in which measured input and output signals are employed to determine a suitable model order that produces an acceptably low level of estimation error. The basic idea of such procedures is that the mean-squared estimation error (or, in general, the appropriate squared-norm of the estimation error vector) decreases monotonically as the model order increases.

Example 2.11 (continued): Model Order Selection

Figure 2.13 displays the mean-squared estimation error for the estimation problem in Example 2.11 as a function of the number of coefficients. We can see that the mean-square error decreases up to a model order of three and then stays constant for higher model orders. This implies that a model order of three is appropriate in this example.

The model order selection in the above example was a relatively easy task. However, in many situations, the unknown system may not be identical to the system model employed. In such situations, the mean-square error often tends to decrease monotonically with increasing system orders without reaching some steady-state value. Therefore, we need to modify the above procedure for estimating the model order. The approach that is typically used is to recognize that increasing the model order arbitrarily may not bring about a correspondingly large reduction in the error. In order to select a model order that corresponds to a reasonable compromise between the complexity and the performance of the estimator, we attempt to minimize a cost function given by

$$J(N) = \xi_{min}(N) + \delta N, \tag{2.67}$$

where $\xi_{min}(N)$ is the MMSE when N coefficients are employed and δ is a small positive constant. Note that the first term on the right-hand side is a monotone non-increasing function of N and that the second term increases linearly with N. The second term may be thought of as a penalty term for increasing the model order. The model order is then selected as the value of N for which $J(N)$ achieves a minimum value.

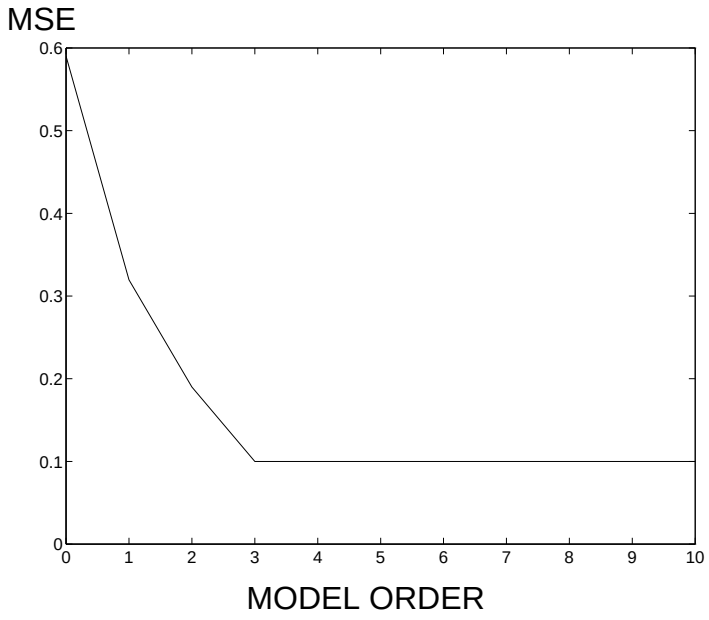

Figure 2.13: Mean-square estimation error for Example 2.11 as a function of the model order.

Example 2.12: IIR System Identification

Consider an IIR filter with input-output relationship

$$y(n) = 0.7y(n-1) + x(n).$$

We wish to identify this system using an FIR system model. The input signal $x(n)$ employed for the identification task is an i.i.d., zero-mean process with unit variance. The output signal $y(n)$ is measured in the presence of additive measurement noise that is also a zero-mean and i.i.d. process with unit variance. Our task in this example is to estimate the model order L of the FIR system model from the statistics of the input signal $x(n)$ and the measured output signal $d(n)$. Increasing the number of coefficients in this example decreases the mean-squared estimation error monotonically. In order to estimate the number of coefficients of the system model, we evaluate the cost function $J(N)$ from the actual correlation statistics of the signals for different values of N. It is straightforward to show that the relevant statistics are given by

$$r_{dd}(0) = E\{d^2(n)\} = 2.96$$

and

$$r_{dx}(k) = \begin{cases} 0.7^k; & n \geq 0 \\ 0; & \text{otherwise.} \end{cases}$$

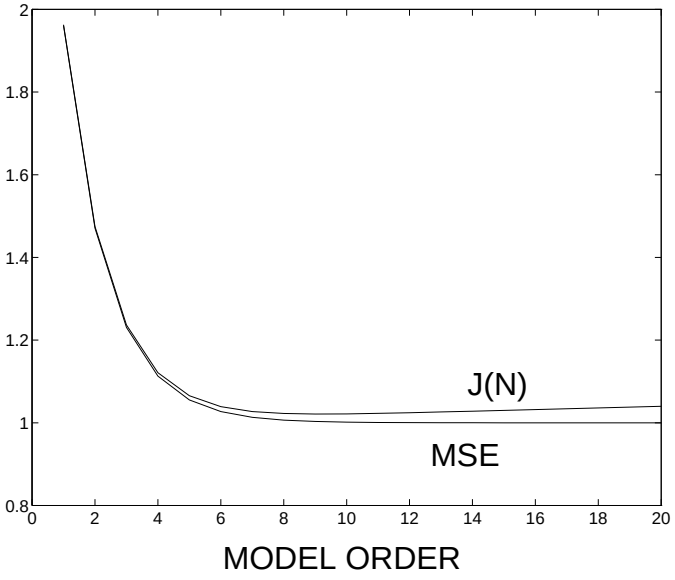

Figure 2.14: Objective function for determining the model order and the mean-squared estimation error in Example 2.12.

Since the autocorrelation matrix of the input signal is an identity matrix for all choices of the model order L, the coefficients of the estimator are given by

$$w_i = \begin{cases} 0.7^i; & 0 \leq i \leq L-1 \\ 0; & \text{otherwise.} \end{cases}$$

The cost function $J(N)$ can be evaluated using (2.67) and (2.50). Figure 2.14 displays $J(N)$ against N for $\delta = 0.002$ along with the mean-squared estimation error. We select the model order to be 9 since $J(N)$ achieves its minimum value of 1.0212 for N =9. The excess mean-square error over the minimum possible value of the MSE of one is only 0.0032 in this case, indicating that our choice of the model order is a reasonable one.

REMARK 2.2: The cost function $J(N)$ as defined in (2.67) may not have a unique minimum. However, in most practical applications, $J(N)$ exhibits monotone increasing behavior for large values of N, and therefore, it is fairly easy to identify the global minimum of $J(N)$.

REMARK 2.3: Even though model order selection as well as model selection are important problems, a thorough discussion of these topics is beyond the scope of this book. Consequently, we will not stress these issues very much from now on.

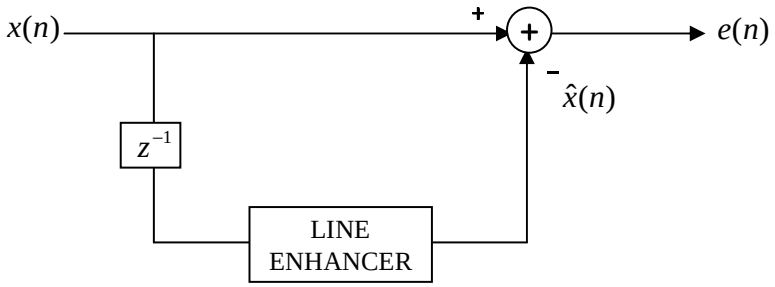

Figure 2.15: Line enhancement using linear prediction.

Line Enhancement

There are many applications in communications and sensor array processing in which broadband noise corrupting a signal consisting of one or more sinusoids must be removed. We use the term *line enhancement* to describe this problem. The name arises from the fact that the spectra of sinusoids consist of impulses or *lines* at the frequencies of the sinusoids. By removing the broadband noise from a corrupted sum-of-sinusoids signal, we *enhance* the signal. We consider a simple case in which the input signal has the form

$$X(n) = A\cos(\omega_0 n + \theta) + \eta(n), \tag{2.68}$$

where A is an arbitrary constant amplitude, θ is the initial phase of the sinusoid, and $\eta(n)$ is the additive noise component within the signal that corrupts our measurement of the sinusoid. The amplitude A and the phase θ are generally unknown for this problem although in practice we may have some *a priori* knowledge about their approximate values or distributions. For our discussion, we assume that θ is uniformly distributed in the range $[-\pi, \pi)$. We assume that $\eta(n)$ is white and uncorrelated with θ. In practice, the noise component is a broadband signal. However, the ideas discussed here can be easily extended to the more general case involving broadband signals.

The Principle of Line Enhancement. A signal that can be decomposed into M real sinusoids can be predicted exactly using $2M$ past samples of the signal. Exercise 2.12 guides the reader through the proof of this statement. White signals, on the other hand, cannot be linearly predicted. Now, consider the estimator structure depicted in Figure 2.15. If we design an estimator to predict $x(n)$ using a delayed version of $x(n)$ (say $x(n-\Delta)$), only the sinusoidal component of $x(n)$ is correlated with $x(n-\Delta)$, as long as $\Delta \geq 1$. Consequently,

the optimal prediction $\hat{x}(n)$ is more an estimate of the sinusoidal component than that of the noisy sinusoid. Thus $\hat{x}(n)$ is an enhanced version of the input signal.

Optimal Solution. It can be shown that the autocorrelation function of $x(n-\Delta)$ for any value of Δ is given by

$$\begin{aligned} r_{xx}(k) &= E\{x(n-\Delta)x(n-\Delta-k)\} \\ &= \frac{A^2}{2}\cos(\omega_o k) + \sigma_\eta^2 \delta(k), \end{aligned} \tag{2.69}$$

where σ_η^2 is the variance of the noise component $\eta(n)$. Since we are trying to predict $x(n)$ using its past values, the relevant cross-correlation values are simply the appropriate samples of the autocorrelation function, *i.e.*,

$$E\{x(n)x(n-\Delta-k)\} = r_{xx}(k+\Delta). \tag{2.70}$$

The optimal Lth order predictor coefficients are obtained by substituting (2.69) and (2.70) in (2.65), which yields

$$\mathbf{W}_{opt} = \begin{bmatrix} r_{xx}(0) & r_{xx}(1) & \cdots & r_{xx}(L-1) \\ r_{xx}(1) & r_{xx}(0) & \cdots & r_{xx}(L-2) \\ \vdots & & & \\ r_{xx}(L-1) & r_{xx}(L-2) & \cdots & r_{xx}(0) \end{bmatrix}^{-1} \begin{bmatrix} r_{xx}(\Delta) \\ r_{xx}(\Delta+1) \\ \vdots \\ r_{xx}(\Delta+L-1) \end{bmatrix}. \tag{2.71}$$

The corresponding minimum mean-squared error value is given by

$$\xi_{min}(L) = r_{xx}(0) - \mathbf{W}_{opt}^T \mathbf{P}_{x\mathbf{x}}, \tag{2.72}$$

where

$$\mathbf{P}_{x\mathbf{x}} = [r_{xx}(\Delta)\ r_{xx}(\Delta+1)\ \cdots\ r_{xx}(\Delta+L-1)]^T. \tag{2.73}$$

Given the above solution, we now ask: *1) How do we select an appropriate prediction order L and delay Δ? 2)How effective is this method in enhancing the sinusoidal components?* We address each of these issues here.

The correlation statistics of the noise $\eta(n)$ are important in the choice of both the delay Δ and the order of the estimator L. If the noise component is white, we can choose Δ to be one sampling time. If the noise component is broadband, we must choose Δ to be large enough so that $\eta(n)$ and $\eta(n-\Delta)$ are effectively uncorrelated.

The choice of predictor order is somewhat more complicated since the input signal is corrupted by additive noise. The prediction using noisy samples of the input signal is no longer exact. Moreover, we can expect that the larger the order of prediction, the better the prediction is, since the predictor tends to reduce the effect of the noise when it uses several input samples weighted and averaged together to estimate the sinusoids. It is the designer's

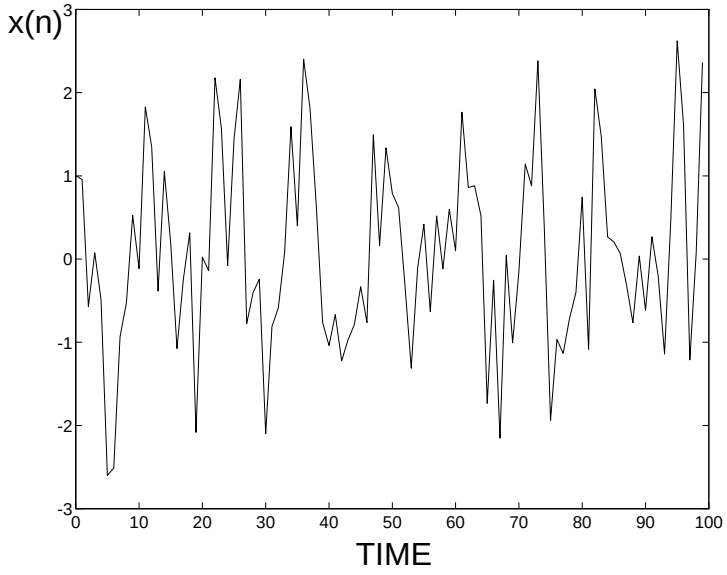

Figure 2.16: The input signal in Example 2.13.

task to pick an order estimator L such that the cost of increasing the prediction order is no longer worth the additional improvement in the signal quality. The cost associated with implementing the predictor can be hardware costs, memory requirements, etc. Methods similar to the one described earlier for model order selection are commonly employed to determine the number of predictor coefficients.

REMARK 2.4: The problem of linear prediction arises in a large number of situations in this book. Examples other than line enhancement include autoregressive spectrum estimation discussed in Section 2.4.5, orthogonalization of signals using lattice predictors considered in Section 3.2.2 and fast recursive least-squares adaptive filters described in Chapter 11. The set of equations in (2.71) is known as the *Yule-Walker equations* for the special case when $\Delta = 1$

In order to get an intuitive feel for the properties of the MMSE line enhancer, we now present a simulation example.

Example 2.13: Line Enhancement

We consider the model of (2.68) with amplitude $A = 1$ and $\omega_0 = \frac{\pi}{6}$ radians/sample. The noise $\eta(n)$ is chosen to be a i.i.d. zero-mean, Gaussian process with unit variance. In this case, the signal-to-noise ratio is -3 dB. Figure 2.16 displays one hundred samples of one realization of

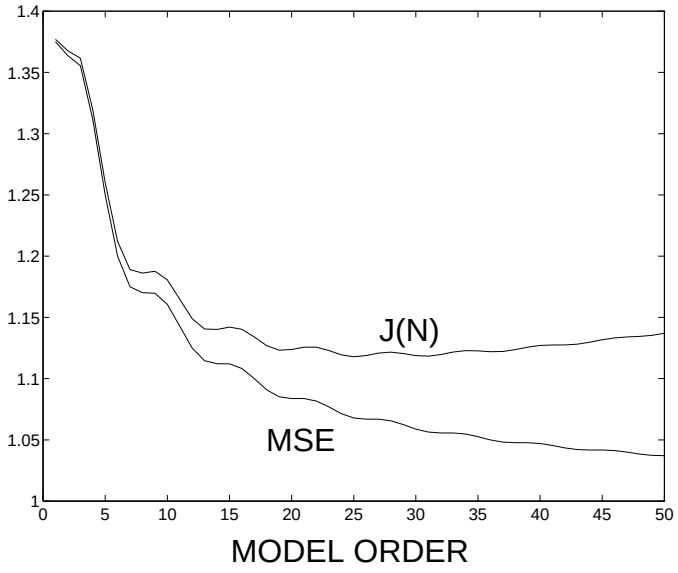

Figure 2.17: The mean-squared prediction error and the cost function to determine the model order in Example 2.13 as a function of the number of coefficients.

the input process. Notice that we can barely identify the presence of a sinusoidal component in the measured signal. Figure 2.17 demonstrates how the mean-squared prediction error behaves when $\Delta = 1$ for various prediction orders from $L = 1$ to $L = 50$. The same plot also displays the cost function in (2.67) for $\delta = 0.002$. This cost function achieves its minimum value for a model order of 25 coefficients. The predicted signal $\hat{x}(n)$ for $L = 25$ corresponding to the input shown in Figure 2.16 is plotted in Figure 2.18. We can see that a considerable amount of noise has been removed from the signal. The same result is demonstrated in the frequency domain in Figures 2.19 and 2.20. These plots contain the estimated spectrum of the input and output signals, respectively, obtained by averaging the magnitude-squared values of the discrete Fourier transform of the signals of duration 1,000 samples each over one hundred independent sets. We can see from these plots that the noise level in the predicted signals has been reduced by approximately 20 dB when compared with the corresponding plots for the input signal spectrum.

2.4.4 Linear Estimation Using Measured Signals

We have assumed in all of our discussions up to this point that the statistics of the signals involved in the linear estimation tasks are known. However, this assumption is rarely true in practice. We typically only have measurements of the signals involved. Consequently, we have to seek means of estimating the statistics from measurements of the signal. Once these

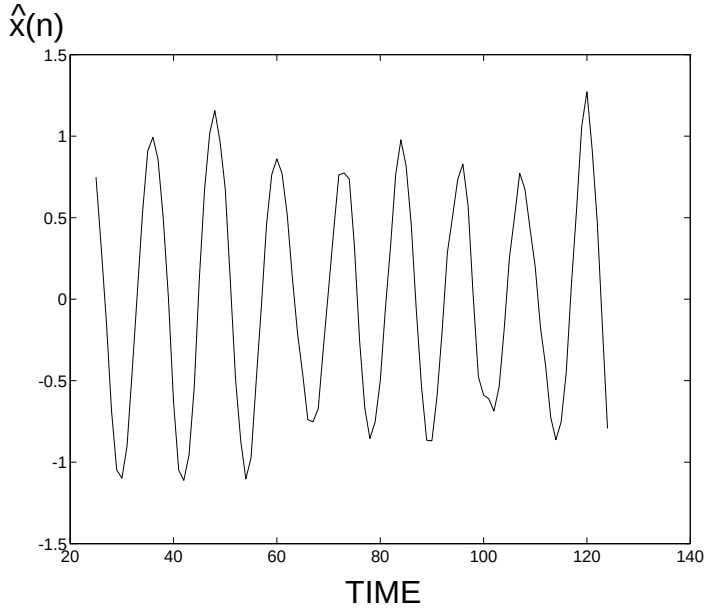

Figure 2.18: The enhanced signal in Example 2.13 for $L = 25$.

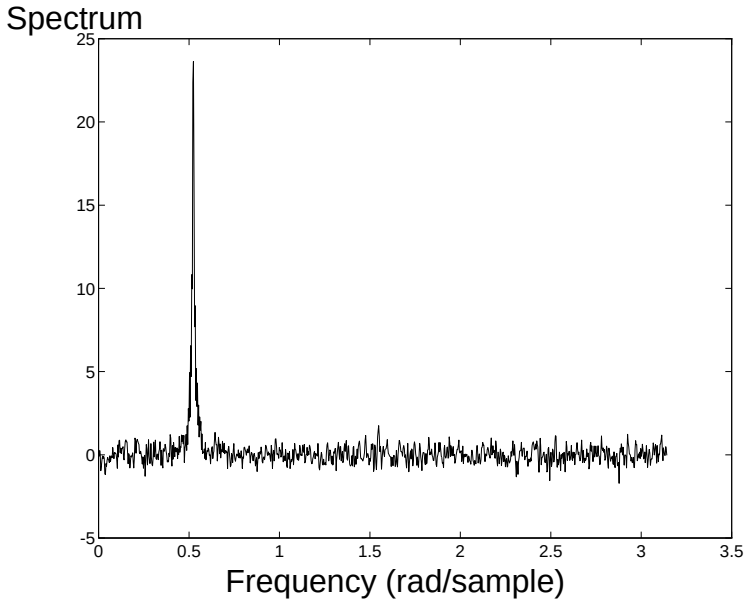

Figure 2.19: Spectrum of input signal in Example 2.13.

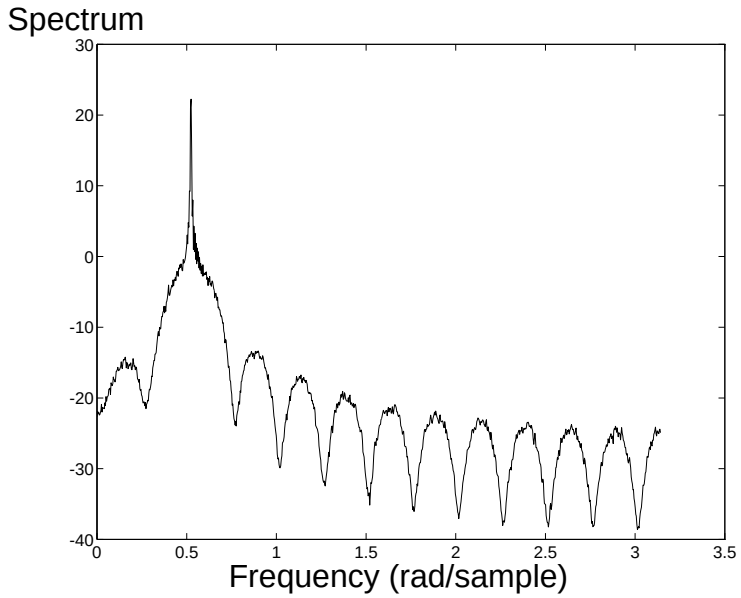

Figure 2.20: Spectrum of enhanced signal in Example 2.13.

estimates are made, we can find approximations for the optimal estimator in (2.60). If the estimates are accurate, the values of the coefficients found this way will be nearly the same as the optimal coefficients if the problem is numerically well-conditioned, and the performance of the estimator will be almost as good as that of the optimal estimator.

The statistical expectations required to solve (2.60) are ensemble averages. In practice, we usually have one limited-duration measurement of a single realization of each of the processes involved. If we assume that all of the processes are jointly ergodic, we can replace the ensemble averages by time averages. Ergodicity and stationarity are rarely satisfied in practice; however, in many situations these properties are satisfied on a short-term or *local* basis. In these situations we can estimate the statistics of the signals based on time-domain measurements of the signals. We can then use these estimates in the Wiener-Hopf equations to solve for the optimum coefficients.

Estimation of Autocorrelation and Cross-correlation Functions

Suppose that we have P samples (say, for $n = 0, 1, \cdots, P-1$) each of a single realization of two ergodic processes $x(n)$ and $d(n)$. We can estimate the autocorrelation and cross-correlation functions as

$$\hat{r}_{xx}(k) = \frac{1}{P} \sum_{n=0}^{P-1} x(n)x(n-k) \tag{2.74}$$

and

$$\hat{r}_{dx}(k) = \frac{1}{P}\sum_{n=0}^{P-1} d(n)x(n-k), \tag{2.75}$$

respectively. When the number of samples involved is much larger than the lag value k, we assume that $x(n-k) = 0$ whenever $n-k < 0$ or $n-k \geq P$. The estimates obtained from this approximation are slightly biased, but usually are adequate in most applications. We leave it to the reader to show that

$$E\{\hat{r}_{xx}(k)\} = \frac{P-k}{P} r_{xx}(k). \tag{2.76}$$

If this bias cannot be tolerated in our application, we should instead use

$$\hat{r}_{xx}(k) = \frac{1}{P-k}\sum_{n=k}^{P-1} x(n)x(n-k) \tag{2.77}$$

and

$$\hat{r}_{dx}(k) = \frac{1}{P-k}\sum_{n=k}^{P-1} d(n)x(n-k) \tag{2.78}$$

when k is positive. When k is negative, the appropriate estimates are

$$\hat{r}_{xx}(k) = \frac{1}{P+k}\sum_{n=0}^{P-1+k} x(n)x(n-k) \tag{2.79}$$

and

$$\hat{r}_{dx}(k) = \frac{1}{P+k}\sum_{n=0}^{P-1+k} d(n)x(n-k). \tag{2.80}$$

Now, the two-stage process that we have proposed – estimating the statistics of the signals and then using (2.65) to determine the estimator coefficients – provides only an approximation to the optimal MMSE solution. A natural question we can ask at this point is: *Is there an estimation problem formulated based on the measurements themselves and solved exactly to give an accurate estimate $\hat{d}(n)$ in some sense?* We explore one such problem and its solution next.

2.4.5 Linear Least-Squares Estimation

As before, we are interested in estimating the process $d(n)$ using the most recent L samples of the process $x(n)$. However, for our current discussion, we base the computation of the coefficients on P samples each of single realizations of $x(n)$ and $d(n)$[6]. Let

$$\hat{d}(n) = \sum_{i=0}^{L-1} w_i x(n-i) \tag{2.81}$$

[6]Recall that we are using the same notation for random processes as well as their realizations. The differences will be obvious from the context in most situations.

represent the desired estimate of $d(n)$. The objective of linear least-squares estimation is to choose the coefficients $w_0, w_1, w_2, \cdots, w_{L-1}$ so that

$$J(P) = \frac{1}{P}\sum_{n=0}^{P-1}\Big(d(n) - \hat{d}(n)\Big)^2 \tag{2.82}$$

is the minimum among all possible choices of the coefficients. As usual, we formulate and solve this problem using vector space concepts.

Vector Space Formulation

Let us define $(P + L - 1)$-dimensional vectors

$$\mathbf{X}_0 = \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ \vdots \\ x(P-1) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ L-1 \quad \text{zeroes} \end{matrix} \quad ; \quad \mathbf{D}_0 = \begin{bmatrix} d(0) \\ d(1) \\ d(2) \\ \vdots \\ d(P-1) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2.83}$$

and

$$\mathbf{X}_k = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ x(0) \\ x(1) \\ \vdots \\ x(P-1) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ k \ \text{zeroes} \\ \\ \\ \\ \\ \\ \\ L-1-k \ \text{zeroes} \\ \\ \end{matrix} \quad ; k = 0, 1, \cdots, L-1. \tag{2.84}$$

An inner product defined as

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \frac{1}{P}\mathbf{X}^T\mathbf{Y} \tag{2.85}$$

is a well-defined inner product for this space. With these definitions, the least-squares estimation problem can be formulated as follows: Find the coefficients $w_0, w_1, w_2, \cdots, w_{L-1}$

such that

$$J(P) = \| \mathbf{D}_0 - \sum_{i=0}^{L-1} w_i \mathbf{X}_i \|^2 \tag{2.86}$$

has the minimum value among all possible choices of the coefficients.

The Optimal Least-Squares Solution

Given the above formulation of the problem, the solution is exactly the same as that given by (2.48). The optimal coefficient vector $\mathbf{W}_{opt}$ is given by

$$\mathbf{W}_{opt} = \hat{\mathbf{R}}_{\mathbf{xx}}^{-1} \hat{\mathbf{P}}_{\mathbf{dx}}, \tag{2.87}$$

where the (i,j)th element of $\hat{\mathbf{R}}_{\mathbf{xx}}$ is given by

$$\hat{r}_{xx}(i,j) = \frac{1}{P} \mathbf{X}_i^T \mathbf{X}_j \tag{2.88}$$

and the ith element of $\hat{\mathbf{P}}_{\mathbf{dx}}$ is given by

$$\hat{r}_{dx}(i) = \frac{1}{P} \mathbf{D}_0^T \mathbf{X}_i. \tag{2.89}$$

It is left as an exercise for the reader to show that $\hat{\mathbf{R}}_{\mathbf{xx}}$ is a symmetric, Toeplitz matrix. This fact, combined with (2.88) implies that $\hat{\mathbf{R}_{\mathbf{xx}}}$ is an estimate of the autocorrelation matrix of the process $x(n)$. Our solution in (2.87) is exactly the same as that in (2.60) for the MMSE formulation, with the exception that (2.60) uses statistical averages whereas (2.87) uses the corresponding estimates obtained by data averaging. For these reasons, this approach of least-squares estimation is known as the *autocorrelation method*.

REMARK 2.5: The method described above implicitly assumed that the input signal $x(n)$ is zero for all values of time outside the window of interest given by $0 \leq n \leq P-1$. In general, the above assumption is not valid. Furthermore, as discussed earlier, *windowing* of the data typically results in biased estimates. Consequently, variations of the autocorrelation method for least squares estimates are often employed in practice. If the formulation of the estimation problem assumes that $x(n) = 0$ for $n < 0$, the procedure is known as *pre-windowing*. If, on the other hand, we assume that $x(n) = 0$ for $n \geq P$, the procedure is known as *post-windowing*. Post and pre-windowing together result in the autocorrelation method. If no post or pre-windowing is done, the estimation procedure is known as the *covariance* method. Depending on the task at hand, these other formulations may provide more accurate estimates of the optimum coefficients.

Example 2.14: Least-Squares System Identification.

Table 2.1: Statistics of the estimates in Example 2.14.

	True coeff. values	$P = 10$		$P = 100$		$P = 1000$	
		Mean	MSD	Mean	MSD	Mean	MSD
$\sigma_\eta^2 = 0.01$	1.0	0.918	0.015	0.989	0.0004	1.000	0.162×10^{-4}
	-1.0	-0.814	0.067	-0.981	0.0013	-0.999	0.231×10^{-4}
	0.5	0.345	0.047	0.485	0.0013	0.499	0.145×10^{-4}
$\sigma_\eta^2 = 0.1$	1.0	0.914	0.0354	0.992	0.0018	1.000	0.144×10^{-3}
	-1.0	-0.816	0.0974	-0.986	0.0025	-1.001	0.191×10^{-3}
	0.5	0.336	0.0596	0.482	0.0024	0.501	0.174×10^{-3}
$\sigma_\eta^2 = 1.0$	1.0	0.876	0.245	0.991	0.0158	1.001	0.0014
	-1.0	-0.782	0.446	-0.989	0.0182	-1.006	0.0019
	0.5	0.283	0.264	0.472	0.0172	0.505	0.0017

Consider the identification of the system in Example 2.11 from measurements of its input and output signals. This example compares the performance of the least-squares estimator for different sample sizes with that of the MMSE estimator. The input signal to the unknown system was generated as the output of an FIR filter with input-output relationship given by

$$x(n) = 0.6\xi(n) + 0.8\xi(n-1),$$

where the input signal $\xi(n)$ belonged to a Gaussian process with zero-mean and unit variance. The autocorrelation matrix of the input signal is identical to the $\mathbf{R_{xx}}$ matrix in Example 2.11. The measurement noise at the output signal belonged to an i.i.d. pseudo-Gaussian sequence with zero mean value and was independent of the input process to the unknown system.

Table 2.1 displays the mean values of the three coefficients for sample sizes of $P = 10,\ 100$ and 1,000 samples and measurement noise variances corresponding to $\sigma_\eta^2 = 0.01,\ 0.1$ and 1.0, when the autocorrelation method was employed. These results were obtained by averaging the parameter estimates from one hundred experiments performed with different, independent realizations of the input and output processes. This table also contains the mean-squared values of the difference (MSD) of each coefficient estimate from its true value, computed over the ensemble of the one hundred experiments.

We can observe several things from the tabulated statistics. The mean values of the estimates approach the true values for large sample sizes. The bias in the estimates decreases with increasing sample sizes. This is a direct consequence of the fact that the least-squares estimates of the cross-correlation and autocorrelation values show larger biases for smaller sample sizes.

The mean-square deviation of the estimates from their true values is a measure that combines the bias and the variance of the estimates. As we would expect, the results of our experiments indicate that the MSD reduces with increasing numbers of samples. Similarly, when the measurement error variance is small, the performance of the estimator improves.

Table 2.2: Statistics of the estimates in Example 2.15.

	True value	Mean	MSD
a_1	1.7900	1.7722	0.0012
a_2	-1.9425	-1.9113	0.0039
a_3	1.2700	1.2426	0.0039
a_4	-0.5000	-0.4895	0.0010
σ_ξ^2	0.1600	0.1636	1.012×10^{-4}

Example 2.15: Autoregressive Spectrum Estimation

In this example, we consider the problem of estimating the power spectral density of a signal generated as the output of a linear, time-invariant system with input-output relationship

$$x(n) = -\sum_{k=1}^{L} a_k x(n-i) + \xi(n)$$

when its input $\xi(n)$ is an i.i.d., zero-mean signal. The above model for signal generation is known as the *autoregressive* (AR) model. We wish to estimate the spectrum of $x(n)$ from measurement of a single realization of the process and knowledge of the signal model.

It is relatively easy to show that the true spectrum is given by

$$S_{\mathbf{xx}}(\omega) = \frac{\sigma_\xi^2}{\mid 1 + \sum_{k=1}^{L} a_k e^{-j\omega k} \mid^2},$$

where σ_ξ^2 is the variance of $\xi(n)$. It should be clear from the above expression that the spectrum can be estimated from knowledge of the parameters $\{a_k; k = 1, 2, \cdots, L\}$ and σ_ξ^2. It is left as an exercise to show that the optimal coefficients of the minimum mean-squared error linear predictor for $x(n)$ using the most recent L samples of $x(n)$, *i.e.*, $x(n-1), x(n-2), \cdots, x(n-L)$ are given by $a_1, a_2, \cdots, a_L$. Furthermore, the MMSE value for this problem is σ_ξ^2. Consequently, we can formulate a least-squares prediction problem of order L to estimate the parameters of interest.

Table 2.2 displays the mean values of the parameters estimated using the autocorrelation method calculated over one hundred independent estimates obtained using Gaussian random sequences of length 1000 samples each. The above sequences were generated using a fourth-order AR model with parameters as shown in the table. This table also shows the variances of the estimates. We can see from the table that the least-squares estimator performs reasonably well in this example. Figure 2.21 shows the plot of the average of the estimated spectrum over the one hundred estimates. The variability of the estimate from one run to the next can be seen from the overlaid plots of the one hundred estimates as shown in Figure 2.22.

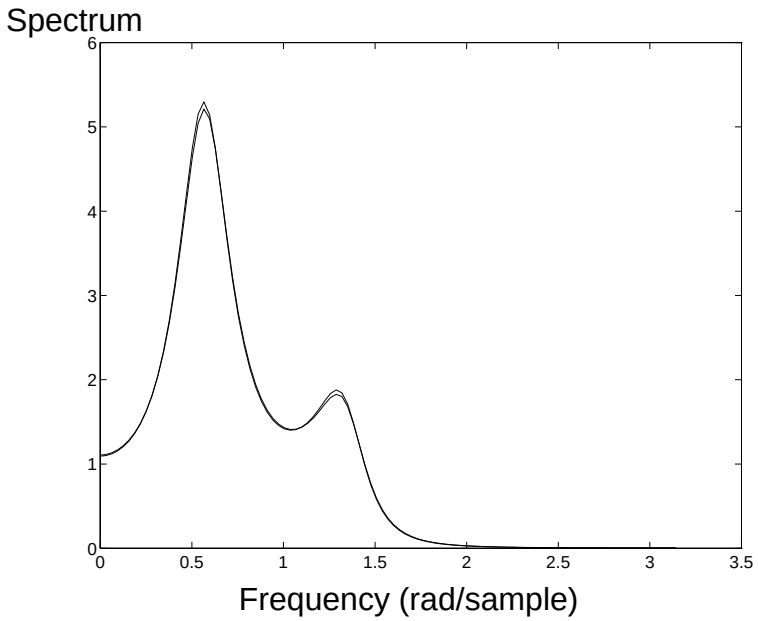

Figure 2.21: The true spectrum and the average of one hundred independent estimates of the spectrum in Example 2.15.

Example 2.16: Identification of Recursive Linear Systems.

We now consider the identification of a linear, time-invariant system with an input-output relationship given by

$$y(n) = \sum_{i=0}^{L} b_i x(n-i) + \sum_{i=1}^{N} a_i y(n-i)$$

from measurements of its input signal and a noisy version of its output signal given by

$$d(n) = y(n) + \eta(n),$$

where $\eta(n)$ is an i.i.d. zero-mean measurement noise sequence with variance $\sigma_\eta^2 = 1$ and is statistically independent of $x(n)$.

If we assume that the measurement noise is relatively small, we can attempt to estimate $d(n)$ as a linear combination of past and present samples of $x(n)$ and past samples of $d(n)$ as

$$\hat{d}(n) = \sum_{i=0}^{L} \hat{b}_i x(n-i) + \sum_{i=1}^{N} \hat{a}_i d(n-i).$$

If $\eta(n) = 0$ for all n, $d(n) = y(n)$, and therefore, the above estimate would be unbiased. However, as the measurement noise becomes large, this approach results in biased parameter estimates.

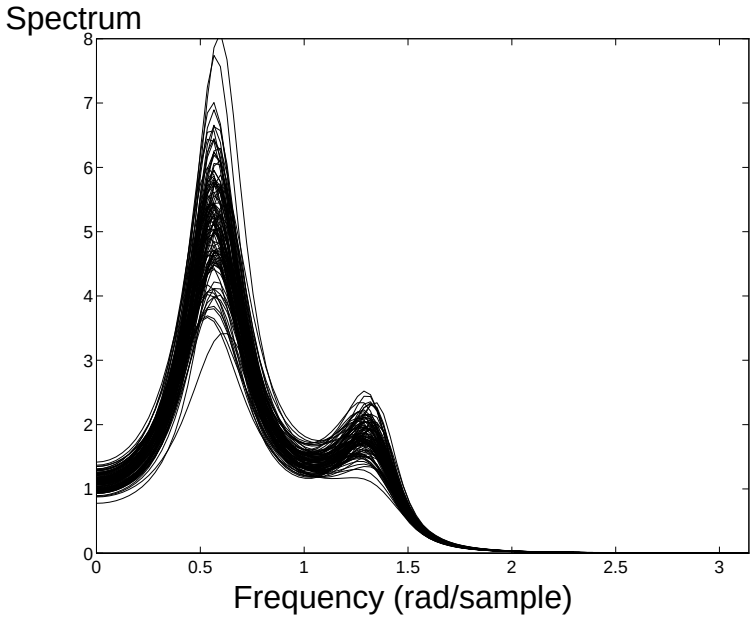

Figure 2.22: Overlaid plots of one hundred independent estimates of the spectrum in Example 2.15.

Table 2.3: Statistics of the estimates in Example 2.16.

True value	$P = 1,000$						$P = 10,000$			
	$\sigma_\eta^2 = 0$		$\sigma_\eta^2 = 0.01$		$\sigma_\eta^2 = 0.1$		$\sigma_\eta^2 = 0.01$		$\sigma_\eta^2 = 0.1$	
	Mean	MSD	Mean	MSD	Mean	MSD	Mean	MSD	Mean	MSD
1.0000	1.0000		1.000	0.000	1.000	0.001	1.000	0.000	1.000	0.000
0.5000	0.5000		0.633	0.018	1.193	0.482	0.632	0.017	1.189	0.475
-1.0000	-1.0000		-0.898	0.011	-0.446	0.310	-0.902	0.010	-0.456	0.296
0.2500	0.2500	$< 10^{-23}$	0.106	0.021	-0.448	0.489	0.108	0.020	-0.447	0.486
1.7900	1.7900		1.656	0.018	1.096	0.484	1.658	0.017	1.102	0.473
-1.9425	-1.9425		-1.736	0.043	-0.905	1.080	-1.739	0.041	-0.911	1.064
1.2700	1.2700		1.096	0.031	0.394	0.769	1.098	0.029	0.400	0.757
-0.5000	-0.5000		-0.416	0.007	-0.088	0.171	-0.417	0.007	-0.090	0.168

This property of the estimator is evident in Table 2.3. This table displays the result of least-squares estimation of the parameters using the approach described above and the autocorrelation method. In our experiments we chose $L = 3$ and $N = 4$ for both the unknown system as well as the estimator. The mean values were obtained from one hundred independent estimates obtained using Gaussian sequences of length 1,000 or 10,000 samples each. Note that as the variance of the measurement noise increases, the deviation of the mean values from the true parameter values increases. This deviation of the mean values does not change even when the number of samples used in the estimation procedure is increased. Note also that the mean-squared deviation of the coefficient estimates does not decrease significantly when the number of samples employed by the estimator is increased. This is a consequence of the fact that the bias in the estimates contributes to a large portion of the MSD value of the estimates.

We will delay our discussion of obtaining unbiased estimates of the parameters of recursive system models to Chapter 13.

Example 2.17: Identification of a Nonlinear System

We have so far concentrated on linear system models. However, the basic ideas described in this chapter can also be applied to a large number of nonlinear estimation problems. A nonlinear system model that is commonly employed in many practical applications is the *truncated Volterra* system model. The input-output relationship of a truncated Volterra system with pth order nonlinearity and L-sample memory is given by

$$\begin{aligned} y(n) &= h_0 + \sum_{i_1=0}^{L-1} h_1(i_1)x(n-i_1) + \sum_{i_1=0}^{L-1}\sum_{i_2=i_1}^{L-1} h_2(i_1,i_2)x(n-i_1)x(n-i_2) \\ &\quad + \cdots + \sum_{i_1=0}^{L-1}\sum_{i_2=i_1}^{L-1}\cdots\sum_{i_p=i_{p-1}}^{L-1} h_p(i_1,i_2,\cdots i_p)x(n-i_1)\cdots x(n-i_p) \end{aligned}$$

where $h_r(i_1, i_2, \cdots, i_r)$ is known as the rth order Volterra kernel of the system. A special case of such systems is the *homogeneous quadratic* system which contains only second-order nonlinearities. The input-output relationship of such systems is given by

$$y(n) = \sum_{i_1=0}^{L-1}\sum_{i_2=i_1}^{L-1} h_2(i_1,i_2)x(n-i_1)x(n-i_2).$$

The choice of the range of summation over i_2 avoids redundant terms in the expansion that would occur if i_2 were to range from 0 to $L-1$. It is relatively easy to see that a least-squares approach, similar to that derived for linear FIR system models, can be used to identify truncated Volterra systems, since $y(n)$ is a linear combination of nonlinear transformations of the input signal.

In this example, we consider the identification of a quadratic system with three-sample memory and coefficients as shown in Table 2.4 from measurements of 1000 consecutive samples of the input and output signals using the autocorrelation method. The input signal to the system was identical to that employed in Example 2.14. The measurement noise signal belonged to an i.i.d. Gaussian

Table 2.4: Statistics of the estimates in Example 2.17.

	True value	Mean	MSD
$h_2(0,0)$	1.00	0.993	0.0009
$h_2(0,1)$	0.30	0.305	0.0050
$h_2(0,2)$	0.10	0.106	0.0013
$h_2(1,1)$	0.50	0.506	0.0022
$h_2(1,2)$	0.15	0.142	0.0048
$h_2(2,2)$	0.20	0.203	0.0011

sequence with zero mean value and variance 0.01. As in the previous examples, we have tabulated the mean values as well as the mean-square values of the coefficient errors for the estimates computed over one hundred independent estimates. We can see in this example also that the least-squares method is capable of estimating the parameters of the system model using noisy measurements of the input and output signals even when the system model is nonlinear.

2.5 Main Points of This Chapter

- Almost all estimation problems can be formulated using vector space concepts. If $\mathbf{R_{xx}}$ is invertible, the vector
$$\mathbf{W}_{opt} = \mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{P}_{\mathbf{dx}}$$
denotes the optimal coefficient vector that minimizes the squared norm of the estimation error vector in the appropriate inner product space.

- The optimal estimation error vector is orthogonal to all the vectors used for estimating the desired vector. Furthermore, the error vector is orthogonal to the space spanned by the input vectors.

- The quantity
$$\mathcal{D}_{min}^2(\mathbf{D}, \hat{\mathbf{D}}) = \| \mathbf{D} \|^2 - \mathbf{W}_{opt}^T \mathbf{P}_{\mathbf{dx}}$$
represents the minimum value of the squared norm of the optimal estimation error vector.

- Vector space concepts can be applied to both minimum mean-square-error estimation and least-squares estimation problems. MMSE methods estimate the parameters by finding the minimum point on an error surface defined using the joint statistics of the input signals. The least-squares techniques estimate the parameters by determining the minimum point of an error surface defined deterministically using measured signals.

- Examples of linear estimation considered in this chapter include line enhancement, autoregressive spectrum estimation, and linear system identification. Linear prediction is a key component of many concepts developed in this book.

- The principles of linear estimation developed in this chapter can be extended to many nonlinear filtering problems.

2.6 Bibliographical Notes

Early Work. A very extensive survey that traces the early development of linear estimation theory can be found in [Kailath 1974]. Another historical survey that describes the contributions of several early researchers in this area is [Seal 1967]. While early Babylonians were known to practice rudiments of estimation theory [Neugebauen 1957], Kailath attributes the beginnings of the theory of estimation which attempts to minimize various functions of the estimation errors to Galileo in 1632.

Gauss is generally regarded as the first to practice least-squares estimation in 1795 [Gauss 1873]. However, the first to publish results on least-squares estimation was Legendre [Legendre 1805]. Grewal [Grewal 1993] states that the German-Swiss physicist Johann Heinrich Lambert discovered and used least-squares techniques before Gauss was born. The technique was independently discovered by Adrian in the United States [Adrian 1808]. According to [Plackett 1949], Gauss was the first to justify the use of least-squares techniques on the basis of their ability to produce unbiased linear estimates with minimum error variance. Interestingly, Gauss first used the least-squares techniques for a nonlinear estimation problem in mathematical astronomy [Grewal 1993].

Least-squares techniques were applied to the problem of predicting discrete-time random processes by Kolmogorov [Kolmogorov 1939, Kolmogorov 1941]. Krein extended the results to continuous-time random signals [Krein 1945a, Krein 1945b]. Wiener was the first to develop explicit formulae for estimators of continuous signals and use them in engineering applications [Wiener 1941].

According to [Kailath 1974], Frecht first suggested the idea of regarding random processes as elements in a metric space with the distance between elements being the variance of their difference [Frecht 1937]. Yule was the first to apply autoregressive models for spectrum estimation. His work involved fitting the AR models to sunspot numbers [Yule 1927]. Yule-Walker equations refer to the early works of Yule and Walker, another pioneer in this area [Walker 1931].

General References. This chapter provides only the basic concepts of linear estimation theory that are required to understand the rest of the book. More detailed, and relatively easy-to-understand discussions of estimation theory can be found in [Mendel 1995, Therrien 1992, Shanmugan 1988]. A good source of additional information on inner product spaces is

[Halmos 1957]. For another description of the development of concepts in estimation theory using inner product spaces, see [Honig 1985].

In all the estimation problems considered in this chapter, the error surface was convex and had a unique minimum. In many nonlinear estimation problems, the error surface may be non-convex and may have multiple local minima. Minimization of such cost functions with multiple minima are not considered in this book. A good source for minimization of complex performance surfaces is [Gill 1981].

The book [Söderström1989] is an excellent reference for adaptive and non-adaptive system identification methods. Parametric spectrum estimation techniques, including autoregressive modeling, are discussed in [Marple 1987, Kay 1988]. References on Volterra systems include [Schetzen 1989, Rugh 1981].

2.7 Exercises

2.1. *Examples of Inner Product Spaces:* Show that the definitions of inner products in Examples 2.1, 2.2, 2.3 and 2.4 satisfy all the properties that inner products must satisfy.

2.2. Consider a space of real-valued N-dimensional vectors with vector addition and scalar multiplication as defined for the Euclidean vector space. Determine if the following definitions are that of valid inner products.

a.

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \max_{i \in [1,N]} \{x_i y_i\}$$

b.

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i=1}^{N} \frac{1}{2} \left(x_i \text{sign}\{y_i\} + y_i \text{sign}\{x_i\} \right)$$

In each case that is not a valid inner product, identify the properties of inner products that are not satisfied by the definition.

2.3. *Triangle Inequality:* Show that the triangle inequality must be satisfied in all valid inner product spaces i.e.,

$$\| \mathbf{X} - \mathbf{Y} \|^2 \leq \| \mathbf{X} - \mathbf{Z} \|^2 + \| \mathbf{Y} - \mathbf{Z} \|^2$$

for any set of three vectors $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$. For simplicity, you may assume that the vectors are real-valued, even though the inequality holds for real-valued as well as complex-valued vectors.

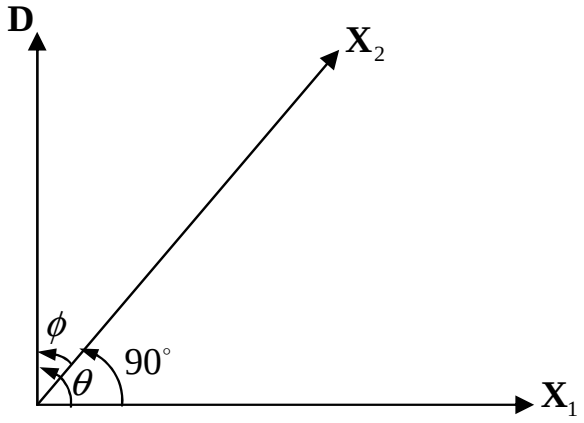

Figure 2.23: Relationship between the vectors in Exercise 2.5.

2.4. *Angle and Distance Calculation:* Determine the angle and the distance between the two vectors given in each part below.

a. $\mathbf{X}_1 = [1 \quad -.5 \quad .5]^T$, $\mathbf{X}_2 = [-1 \quad 0 \quad 1]^T$. The vectors belong to a space of real, three-dimensional vectors and the inner product is defined as in Example 2.1.

b. $\mathbf{X}_1 = \cos\theta$ and $\mathbf{X}_2 = \cos(\theta + \frac{\pi}{4})$, where θ is a uniformly distributed random variable in the range $[-\pi, \pi)$, the vector space under consideration is the space of random variables with zero mean value and finite variances, and the inner product between two vectors $\mathbf{X}$ and $\mathbf{Y}$ is defined as $E\{\mathbf{XY}\}$.

2.5. *Estimation in a Three-Dimensional Space:* Consider the three-dimensional plot of the three vectors $\mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{D}$ shown in Figure 2.23 The angles between the vectors as well as the lengths of the vector are labeled in the figure.

a. Find an expression for the optimum linear estimate of $\mathbf{D}$ using $\mathbf{X}_1$ and $\mathbf{X}_2$ as a function of the angles θ and ϕ.

b. Suppose now that the angle between $\mathbf{X}_1$ and $\mathbf{X}_2$ is no longer 90°. How will you find the optimum linear estimate of $\mathbf{D}$ using $\mathbf{X}_1$ and $\mathbf{X}_2$ in this case?

2.6. *An FIR System Identification Problem:* Let $x(n)$ be a stationary random process generated as the output of a linear system with input-output relationship

$$x(n) = 1.3x(n-1) + 0.4x(n-2) + \zeta(n),$$

where $\zeta(n)$ is a real, i.i.d. process with zero mean value and unit variance. This signal is then processed with a three-coefficient FIR filter with coefficients

$$h(k) = \begin{cases} 1 & ; \quad k = 0 \\ 0.5 & ; \quad k = 1 \\ 0.25 & ; \quad k = 2 \\ 0 & ; \quad \text{otherwise.} \end{cases}$$

Let the above output signal be $y(n)$. Our objective is to identify the above system using the MMSE criterion with the help of the statistics of the input signal $x(n)$ and the desired response signal $d(n)$, obtained by corrupting $y(n)$ with an additive noise sequence uncorrelated with $x(n)$, with zero mean value, and variance $\sigma_\eta^2 = 0.01$.

a. Show that $\zeta(n)$ is uncorrelated with $x(k)$ for $k < n$. Using this result, show that the autocorrelation of the input signal satisfies the relationship

$$r_{xx}(k) = 1.3r_{xx}(k-1) + 0.4r_{xx}(k-2) + \delta(k) \;\; ; k \geq 0,$$

where $\delta(k)$ denotes the discrete-time unit impulse function.

b. Set up the normal equations for this problem when the system model employs three coefficients. Verify by directly evaluating the relevant cross-correlations that the MMSE error sequence is uncorrelated with $x(n)$, $x(n-1)$ and $x(n-2)$.

2.7. *Computing Assignment in System Identification:* In this exercise, we will investigate several characteristics of least-squares system identification techniques.

Generate a zero-mean, Gaussian sequence $x(n)$ of length P samples and variance $\sigma_x^2 = 1$ using the `randn` command in MATLAB. Also generate a zero-mean, Gaussian noise sequence $\eta(n)$ with the same length and variance σ_η^2. The actual values of the parameters P and σ_η^2 will change from experiment to experiment. Finally, generate the noisy version of the output of the system to be identified as

$$d(n) = \sum_{i=0}^{9} x(n-i)w_{true,i} + \eta(n)$$

where $w_{true,i}$ represents the coefficients of the unknown system with numerical values given by [0.1 0.3 0.5 0.7 0.9 0.9 0.7 0.5 0.3 0.1]. For each part of the exercise repeat the experiment fifty times using independent signal sets and evaluate the desired statistics by ensemble averaging over the fifty estimates. Estimate the unknown system coefficients using the least-squares technique and the direct form system modeling. Graphically display the following information obtained from the experiments.

a. MMSE value as a function of the model order L varying from 0 to 15 when $P = 1000$ and $\sigma_\eta^2 = 0.1$. Evaluate the MMSE value as time average over the one thousand samples and then ensemble average over the fifty experiments.

b. Observe the ensemble averages of the coefficient values for several data lengths and measurement noise variances. In addition, plot the sum of the mean-square deviations of the coefficients over the fifty experiments from their ensemble mean values as a function of the data length P for different values of the measurement noise variance σ_η^2. Use $P = 100$, 200, 500, 1000, 2000, 5000, and 10,000 and $\sigma_\eta^2 = 0$, 00.1, 0.01, 0.1 and 1 in the experiments. Use $L = 10$ in all the experiments for this part. Attempt to derive a functional relationship between the performance measure and P when all the other parameters are kept constant. Similarly, develop a functional relationship between the performance measure and σ_η^2 when all the other parameters are held constant. An example of a functional relationship is: the performance measure is proportional to σ_η^2.

2.8. *Computing Assignment: Identification of a Nonlinear System.* Generate a Gaussian random sequence $x(n)$ with zero mean value, variance $\sigma_x^2 = 0.5$ and length $P = 1000$ samples using the MATLAB function `randn`. Generate another 1000-sample long sequence that is uniformly distributed in the range [0.8,1] and uncorrelated with $x(n)$ using the MATLAB function `rand`. Now create a new signal $d(n)$ using the relationship

$$d(n) = e^{\{ax(n)\}}\eta(n),$$

where the parameter $a = -0.5$ for this experiment. Our task is to estimate the parameter a from the measurements of $d(n)$ and $x(n)$ that were generated above. We can transform the problem into another one involving a linear model by taking the logarithm of $d(n)$ to get

$$y(n) = \ln\{d(n)\} = ax(n) + \ln\{\eta(n)\}.$$

Estimate a using the least-square criterion and the model $\hat{y}(n) = \hat{a}x(n)$. Repeat the experiment fifty times with an independent signal set for each experiment and evaluate the ensemble mean and variance of the estimated parameter. Why is the estimate biased? Devise a model that would result in an unbiased estimate of the parameter. Repeat the experiment using this model and verify that the estimate is unbiased.

2.9. *Exponentially-Weighted Least-Squares Estimation:* In most adaptive filtering problems, it is desirable to weight the recent samples of the signals involved as more important than signals that occurred in the distant past. The rationale is that the statistics of the input signals may have changed over time, and in order to design an estimator that measures the current relationship between the desired response signal and the input signal, the recent samples should be weighted higher than the older samples. One way of achieving this objective is to use the exponentially-weighted least-squares criterion which involves the minimization of the cost function

$$J(P) = \sum_{k=1}^{P} \lambda^{P-k} e^2(k),$$

where

$$e(k) = d(k) - \sum_{i=0}^{L-1} w_i x(k-i)$$

a. Formulate the above estimation problem in an appropriate inner product space.

b. Find a closed-form expression for the optimal coefficient vector.

2.10. *Bias in Least-Squares Estimation:* Show that

$$E\{\hat{r}_{xx}(k)\} = \frac{P-k}{P} r_{xx}(k),$$

where $\hat{r}_{xx}(k)$ is as given in (2.74).

2.11. *A Constrained Least-Squares Estimation Problem:* Suppose that we are interested in estimating a signal $d(n)$ as a linear combination of the most recent L samples of $x(n)$, and at the same time limit the magnitude of the estimator coefficients. One way to accomplish this is by modifying the cost function as follows:

$$J = E\left\{d(n) - \sum_{i=0}^{L-1} w_i x(n-i)\right\}^2 + \delta \parallel \mathbf{W} \parallel^2$$

where δ is a positive constant.

a. Derive the optimum coefficient vector that minimizes the above cost function.

b. Derive a least-squares algorithm that achieves the same objective.

2.12. *Prediction and Frequency Estimation of Sinusoidal Signals:* Let

$$x(n) = \sum_{i=1}^{N} A_i e^{j(\omega_i n + \theta_i)},$$

where A_i's are positive, real amplitudes and θ_i's are the phase values distributed in the range $[-\pi, \pi)$.

a. Show that the output of the system with transfer function

$$H(z) = \prod_{i=1}^{N} (1 - e^{j\omega_i} z^{-1})$$

is zero when its input is $x(n)$.

b. Use the above information to design a perfect N-point predictor for $x(n)$. By "perfect" we mean that the estimation error is zero. Let

$$A(z) = \sum_{i=1}^{N} a_i z^{-i}$$

denote the transfer function of this predictor. Show that

$$H(z) = 1 - A(z)$$

c. How would you estimate the frequencies of the sinusoids from the coefficients of the predictor? Modify your method if it is known that $x(n)$ contains K real sinusoids.

d. *Computing Assignment:* Generate one thousand samples of the signal

$$x(n) = \sin(0.10n + \theta_1) + \sin(0.25n + \theta_2),$$

where θ_1 and θ_2 are independent random varaibles that are uniformly distributed in the range $[-\pi, \pi)$. Estimate the two frequencies with the help of a fourth-order least-squares predictor.

The rest of the assignment involves evaluating the performance of your system when $x(n)$ is corrupted by additive Gaussian noise with zero mean value and variance σ_x^2. For this, create a zero-mean Gaussian noise with unit variance sequence $\eta(n)$ of length $P = 1000$ samples using the `randn` command in MATLAB. Now create a corrupted version of $x(n)$ as given by

$$y(n) = x(n) + \sqrt{\alpha}\eta(n),$$

where α may be chosen as $1, 0.1, 0.01, 0.001$ and 0.0001 for different experiments. For each choice of α, find the optimal fourth-order least-squares predictor for $y(n)$. Estimate the two frequencies after finding the roots of the polynomial $1 - A(z)$. Repeat each experiment using fifty independent sets of signals and tabulate the ensemble mean and variance of the parameter estimates. Describe the difficulties caused by noise in the measurements.

2.13. *Computing Assignment: Sinusoidal Interference Cancellation.* One significant problem that occurs in diagnostic equipments such as electro-cardiographs (ECG) and electro-encephalographs (EEG) is the inability to completely isolate the devices from line voltages. Since the measurements made by these machines typically range in microvolts, even a small leakage of the line voltage can completely obscure the desired measurements. Fortunately, the source of interference is known in this case and we can use this information to cancel the interference adaptively. A block diagram of the

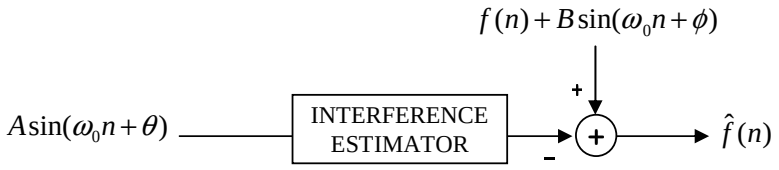

Figure 2.24: Block diagram of a sinusoidal interference canceller.

system one would employ for this application is shown in Figure 2.24. The desired response signal contains the signal $f(n)$ that we want extracted. The interference signal is different from the input signal by an unknown initial phase and an unknown amplitude value as shown in the figure. Assuming that $f(n)$ is uncorrelated with the source of interference $x(n)$, we can argue that the estimate of $d(n)$ using $x(n)$ will estimate only the interference, and therefore, the estimation error signal is a cleaner version of the signal $f(n)$.

a. To simulate an ECG signal, generate a triangular waveform $f(n)$ with period twenty samples and a peak value of 0.1 volt. Also generate a sinusoidal signal $x(n)$ with amplitude 1 volt and frequency 60 Hz and sampled at a rate of 200 samples/second. Generate 2000 samples of each signal. You can simulate the corrupted signal using the model

$$d(n) = f(n) + 0.5\sin\left(\frac{120\pi}{200}(n-0.25)\right).$$

b. From your understanding about the predictability of sinusoids, what can you say about the number of coefficients required for the estimator? Plot the enhanced version of $f(n)$ obtained as the error in estimating $d(n)$ using your choice for the number of coefficients and the least-squares error criterion. Comment on the performance of the interference canceller you developed.

2.14. *The Cost Function of (2.67) May Have Multiple Local Minima:* The cost function $J(N)$ given in (2.67) is not guaranteed to have a unique minimum. Demonstrate this fact using an example.