2.3 Linear Estimation

2.3.1 Signal Estimation Problem

2.3.2 Optimum Estimation of Signals

2.3.3 Linear MMSE Estimation of Signals

2.4 FIR Wiener Filter

2.5 Least Squares Optimal Filtering

2.5.1 LS Method

2.5.2 LS Optimal Filtering

2.5.3 Least Squares Orthogonality

2.6 Introduction to Adaptive filtering

2.7 Adaptive Wiener Filter

2.8 The LMS Algorithm

2.9 Convergence Property of LMS Algorithm

2.10 Simplified LMS Algorithm

2.11 Applications

2.12 RLS Adaptive Filters

2.13 Adaptive Transversal Filters Using Least Squares Method

2.3 Linear Estimation

2.3.1 Signal Estimation Problem

- The problem of estimating one signal from another is one of the most important in signal processing.
- In many applications , the desired signal is not available or observable directly. Instead, the desired signal is a degraded or distorted version of the original signal.
- The signal estimation problem is to recover , in the best way possible , the desired signal from its degraded replica.

- **Examples :**
  - **(1) The desired signal may be corrupted by strong additive noise , such as weak measured evoked brain potentials against the strong background of ongoing EEGs**
  - **(2) A signal transmitted over a communications channel can suffer phase and amplitude distortions and can be subject to additive channel noise ; the problem is to recover the transmitted signal from the distorted received signal.**

2.3.2 Optimum Estimation of Signals

- **The signal estimation problem can be stated as follows:**
  We are to estimate a random signal $x(n)$ on the basis of available observations of related signal $y(n)$.
- **The available signal $y(n)$ is to be processed by an optimal processor that produces the best estimate of $x(n)$. The resulting estimate $x(n)$ will be a functional of the observations $y(n)$.**
- **If the optimal processor is linear, such as a linear filter, then the estimate $x(n)$ will be a linear function of the observations.**

- **Several major criteria for designing such optimal processors will be discussed :**

  (1) Maximum a posteriori (MAP) criterion

  (2) Maximum likelihood (ML) criterion

  (3) Minimum mean-squared error (MMSE) criterion

  (4) Linear minimum mean squared error (LMMSE) criterion

  (5) Least square (LS) criterion

- **Let us assume that the desired signal $x(n)$ is to be estimated over a finite time interval $n_a \leqq n \leqq n_b$**

Also we may assume that the observed signal y_n is also available over the same interval.

Define the vector

$$\mathbf{x} = (x(n_a)\ \ x(n_a+1)\ \ldots\ x(n_b)\ )^T \quad (2.111)$$

$$\mathbf{y} = (y(n_a)\ \ y(n_a+1)\ \ldots\ y(n_b)\ )^T \quad (2.112)$$

For each value of n, we seek the functional dependence

$$\hat{x}(n) = \hat{x}(n \mid \mathbf{y})$$

of $x(n)$ on the given observation vector y to provide best estimate of the nth sample $x(n)$.

2.3.2.1 MAP Estimation

- The criterion for the MAP estimate is to maximize the *a posterior* conditional density of $x(n)$ given that **y** already occurred ; namely

$$p(\,x(n)\;|\;\mathbf{y}\,) = \text{maximum} \qquad (2.113)$$

In other words, the optimal estimate $\hat{x}(n)$ is that $x(n)$ which maximizes this quantity for the given vector **y** .

2.1.2.2 ML Estimation

- The ML criterion , on the other hand, selects $x(n)$ to maximize the conditional density of **y** given $x(n)$; that is

$$p\,(\mathbf{y}\;|\;x(n)\,) = \text{maximum} \qquad (2.114)$$

This criterion selects $\hat{x}(n)$ as though the already collected observations **y** were the most likely ones to occur .

2.3.2.3 MMSE Estimation

The MMSE criterion minimizes the mean-squared estimation error

$$E[e^2(n)] = \text{minimum}, \qquad (2.114)$$

where $e(n) = x(n) - \hat{x}(n \mid y)$ (2.115)

and $\hat{x}(n \mid y) = E[x(n) \mid y]$

= mean-square estimate (2.116)

2.3.2.4 LMMSE Estimate

- **The LMMSE criterion requires the estimate to be a linear function of the observations**

$$x(n) = \Sigma_{i=na}^{nb} \quad h(n; i)\ y(i) \qquad (2.117)$$

For each n, the weights $h(n;i)$ are selected so as to minimize the mean-squared estimation error

$$E[e^2(n)] = \mathrm{E}\,[\,(x(n) - x\hat{}\,(n))^2] = \text{minimum} \qquad (2.118)$$

Note :

With the exception of the LMMSE estimate , all the other estimates $x(n \mid y)$ are , in general , nonlinear in y

2.3.3 Linear MMSE Estimation

- Two common problems of determining the optimal weights $h(n,i)$ according to the mean-squared minimization criterion

  are

  (1) Optimal filtering problem

  (2) Optimal prediction problem

- In these cases , the optimal estimate of $x(n)$ at a given time instant n is given by an expression of the form

$$x(n) = \Sigma_{i=na}^{nb} h(n;i)\, y(i) \qquad (2.119)$$

as a linear combination of the available observations $y(n)$ in the interval $n_a \leqq n \leqq n_b$.

2.3.3.1 Optimal Filtering

- **The optimal filtering problem requires the linear operation**

$$\hat{x}(n) = \sum_{i=na}^{nb} h(n; i)\, y(i) \qquad (2.120)$$

to be causal ; that is, only those observations that are in the present and past of the current samples $x(n)$ must be used in making up the estimate $\hat{x}(n)$.

This requires $h(n; i) = 0$ **, for** $n < i$ **(2.121)**

and then $\hat{x}(n) = \sum_{i=na}^{n} h(n; i)\, y(i)$ **(2.122)**

- **The estimate $\hat{x}(n)$ depends on the present and all the past observations, from the fixed starting point n_a to the current time instant n . As n increases , more and more observations are taking into account in making up the estimate $\hat{x}(n)$.**

- **To make the optimum filter computationally efficient and manageable, only the current and the past M observations $y(i)$; $i = n-M, n-M+1, \ldots, n-1, n$, are taking into account. That is,**

$$\hat{x}(n) = \sum_{i=n-M}^{n} h(n;i)\, y_i$$

$$= \sum_{i=n-M}^{n} h(n;i)\, y(i) \qquad (2.123)$$

This is referred to as the finite impulse response (FIR) Wiener filter.

2.3.3.2 Linear Prediction

- **The linear prediction problem is a special case of the optimal filtering problem with the additional stipulation that observations only up to time instant *n-D* must be used in obtaining the current estimate $x(n)$; this equivalent to the problem of predicting *D* units of time into the future.**
- **If we demand that the prediction be based only on the past M samples (from the current sample) , we obtain the FIR version of the prediction problem, which can be depicted below :**

$$\hat{x}(n) = \sum_{i=n-M}^{n-1} h(n;i)\, y(i) \qquad (2.124)$$

Summary :

- **In LMMSE problem , the estimate is expressed by**

$$\hat{x}(n) = \sum_{i=na}^{nb} h(n;i)\, y(i) \qquad (2.125)$$

for given observations $y(n)$.

- **The problem is a filtering problem when $n = n_b$,**

it is a prediction problem when $n > n_b$,

- **We are to set up the general orthogonality and normal equations for the optimal weights of optimal filtering.**

2.4 FIR Wiener Filter

- **Assuming that the signals are stationary , FIR Wiener filters are relatively simple to implement, inherently stable and more practical .**
- **The FIR filter is represented by the tap-weights w_k, $k = 0,1,2, \ldots, M$**
- **Denote that $W = (w_0, w_1, \ldots, w_M)^T$**

The estimate can be expressed as

$$\hat{x}(n) = \sum_{k=0}^{M} y(n-k)\, w_k \qquad (2.126)$$

Then, the estimation error in FIR Wiener filter is given by

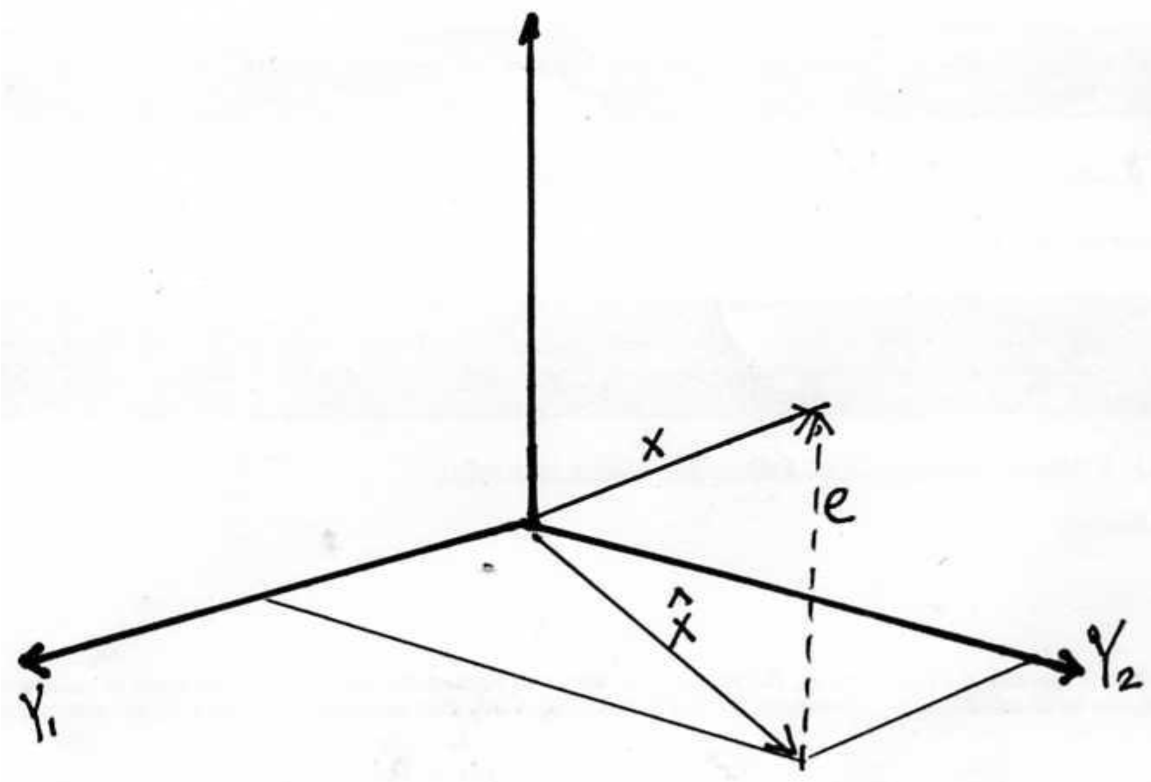
$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{k=0}^{M} y(n-k)\, w_k \qquad (2.127)$$

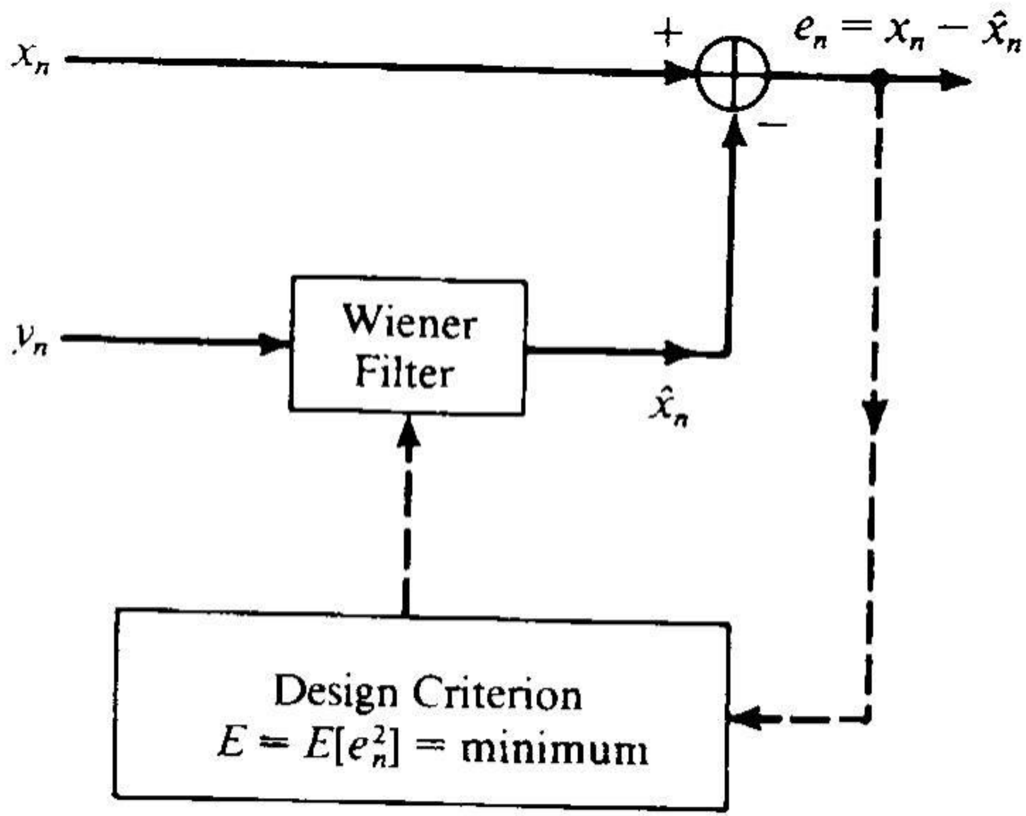
Differentiating the mean-squared estimation error with respect to each weight and setting the derivative to zero, we obtain the orthogonality equations that are enough to determine the weights :

$$\delta E[e^2(n)]/\delta w_i = 2E[e(n)(\delta e(n)/\delta w_i)]$$
$$= -2E[e(n)y(n-i)] = 0 \qquad (2.128)$$

or $R_{ey}(n) = E[e(n)y(n-i)] = 0$ for $0 \leqq i \leqq M$

$$(2.129)$$


x
e
$\hat{x}$
Y_1
Y_2



x_n
+
$e_n = x_n - \hat{x}_n$
−
y_n
Wiener Filter
$\hat{x}_n$
Design Criterion
$E = E[e_n^2] = \text{minimum}$

Thus , the estimation error is orthogonal (uncorrelated) to each observation y_i used in making up the estimate x_n . The orthogonality equations provided exactly as many equations as there are unknown weights.

- **Inserting (2.127) for e_n , the orthogonality equations may be written in an equivalent form , known as normal equations or Wiener-Hopf equations.**

$$E\left[\left\{ x(n) - \sum_{k=0}^{M} w_k y(n-k) \right\} y(n-i) \right] = 0 \qquad (2.130)$$

or

$$E[x(n)\, y(n-i)] = \sum_{k=0}^{M} w_k\, E[y(n-k)\, y(n-i)]$$

$$\text{for } 0 \leqq i \leqq M \qquad (2.131)$$

These Wiener-Hopf equations determine the optimal weights at the current time instant n .

We can write Eq.(2.131) in vector notation as

$$\mathbf{p} = E[\, x(n)\mathbf{y} \,]$$

and $\hat{x}(n) = \mathbf{W}^T \mathbf{y}$

where $\mathbf{W} = (w_0, w_1, \ldots, w_M)^T$ **is the optimum weight- vector ,**

$$\mathbf{y} = (\, y(n) \quad y(n-1) \ \ldots \ y(n-M) \,)^T$$

is the vector of observations up to the current time n **,**

- **The optimal** ***weights*** $\mathbf{W}^o$ **and the estimate are then given by**

$$\mathbf{W}^o = E\,[x(n)\mathbf{y}\,]\, E[\mathbf{y}\mathbf{y}^T]^{-1}$$

$$= \mathbf{R}_Y^{-1}\, \mathbf{p} \qquad (2.132)$$

This is identical to the correlation canceller discussed before.

Note that $\mathbf{p} = (\, p(0) \ p(1) \ \ldots \ p(M) \,)^T$

where $p(j) = E[x(n)\, y(n-j)\,]$ **(2.133)**

The *M+1* optimal filter weights $w_0, w_1, .., w_M$ are obtained by the *(M+1)* x *(M+1)* matrix inversion of the Wiener-Hopf equations (also known as normal equation):

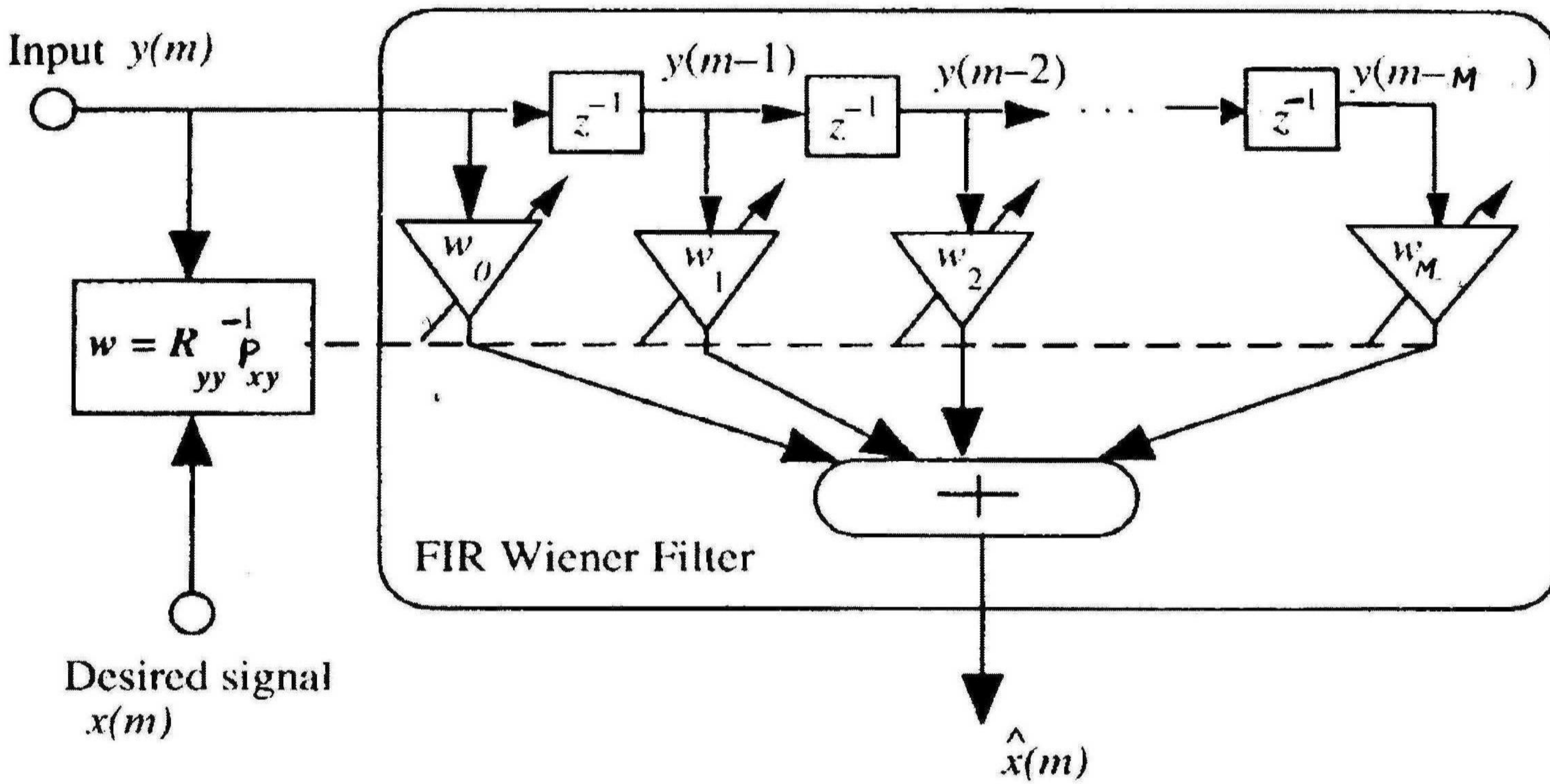


Illustration of a Wiener filter structure.

$$
\begin{matrix}
R_{yy}(0) & R_{yy}(1) & R_{yy}(2) & \dots & R_{yy}(M) & w_0 \\
R_{yy}(1) & R_{yy}(0) & R_{yy}(1) & \dots & R_{yy}(M-1) & w_1 \\
R_{yy}(2) & R_{yy}(1) & R_{yy}(0) & \dots & R_{yy}(M-2) & w_2 \\
 & & & & & \\
R_{yy}(M) & R_{yy}(M-1) & R_{yy}(M-2) & \dots & R_{yy}(0) & w_M
\end{matrix}
$$

$$
= \begin{matrix} p(0) \\ p(1) \\ p(2) \\ . \\ p(M) \end{matrix}
$$

(2.134)

- **The minimized estimation error at time instant *n* is computed by**

$$J(n) = E[e^2(n)] = E[e(n)\,e(n)] = E[e(n)\,x(n)]$$

$$= E[\{x(n) - \Sigma_{k=0}^{M} y(n-k)\,w_k\}\,x(n)]$$

$$= E[x^2(n)] - E[\Sigma_{k=0}^{M} y(n-k)\,w_k\,x(n)]$$

$$= E[x^2(n)] - E[x(n)\,y^T]\,E[yy^T]^{-1}\,E[y\,x(n)] \qquad (2.135)$$

Exploiting the Toeplitz property of the matrix R_{yy}, the above matrix equation (2.134) can be solved efficiently using Levinson's algorithm.

Example :

(1) For M=2

Equ.(2.55) becomes

$$\begin{bmatrix} R_{yy}(0) & R_{yy}(1) & R_{yy}(2) \\ R_{yy}(1) & R_{yy}(0) & R_{yy}(1) \\ R_{yy}(2) & R_{yy}(1) & R_{yy}(0) \end{bmatrix} \begin{bmatrix} w(0) \\ w(1) \\ w(2) \end{bmatrix} = \begin{bmatrix} r_{xy}(0) \\ r_{xy}(1) \\ r_{xy}(2) \end{bmatrix}$$

If observation $y(n) = x(n) + v(n)$

where $R_{xx}(k) = 2\,(0.8)^{|k|}$

$R_{vv}(k) = 2\,\delta(k)$

Thus $p(k) = r_{xy}(k) = E[\,x(n)\,\{x(n-k) + v(n-k)\}\,] = R_{xx}(k)$

$= 2\,(0.8)^{k}$

$$R_{yy}(k) = E[\,\{x(n) + v(n)\}\,\{x(n-l) + v(n-l)\}\,]$$

$$= R_{xx}(k) + R_{vv}(k) = 2\,(0.8)^{k} + 2\,\delta(k)$$

- **The normal equation becomes**

$$\begin{bmatrix} 4.00 & 1.60 & 1.28 \\ 1.60 & 4.00 & 1.60 \\ 1.28 & 1.60 & 4.00 \end{bmatrix} \begin{bmatrix} w(0) \\ w(1) \\ w(2) \end{bmatrix} = \begin{bmatrix} 2.00 \\ 1.60 \\ 1.28 \end{bmatrix}$$

Solving the above equations, we obtain

$w(0) = 0.3824$

$w(1) = 0.2000$

$w(2) = 0.1176$

Appendix : Wiener filter for complex –valued signals

- $\mathbf{W} = (w_0^*, w_1^*, \ldots, w_M^*)^T = (w_0, w_1, \ldots, w_M)^H$

  where H denotes the complex-conjugate or Hermitian .

  $\mathbf{p} = E[\, x^*(n)\, \mathbf{y}\,]$

  $\mathbf{R_{YY}} = E[\, \mathbf{y}\mathbf{y}^H\,]$

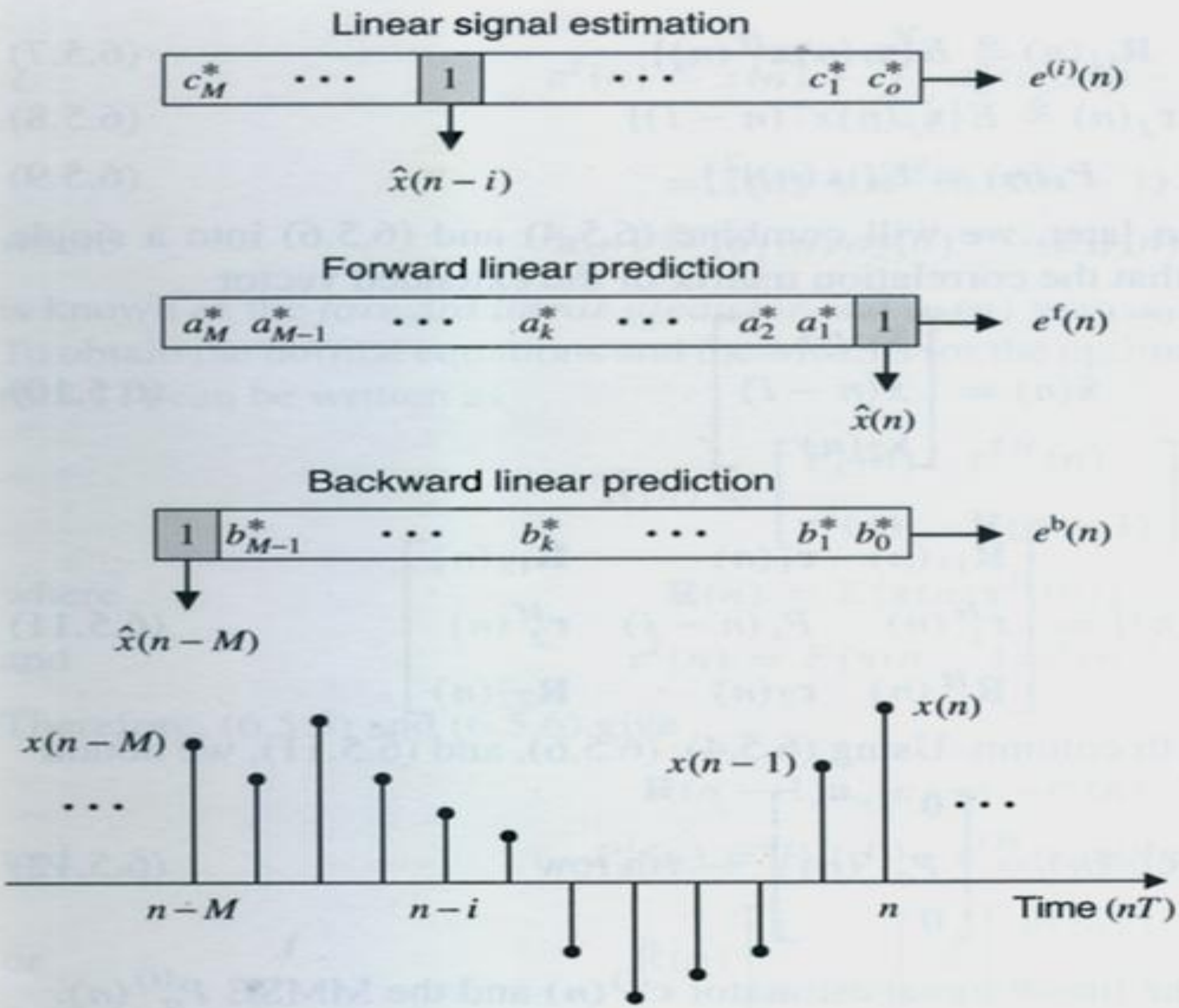
  The optimal *weights* $\mathbf{W^o}$ and the estimate are then given by

  $\mathbf{W^o} = E[\,\mathbf{yy}^H\,]^{-1}\, E\,[x^*(n)\,\mathbf{y}\,]$

  $= \mathbf{R_Y}^{-1}\,\mathbf{p}$

2.5 Linear Prediction

- **The linear prediction problem is a special case of the optimal filtering problem with the additional stipulation that observations only up to time instant $n\text{-}M$ must be used in obtaining the current estimate $x(n)$ *to* predict *one* units of time into the future. This is a one-step forward prediction.**
- **By using similar concept of prediction, we may define a backward predictor, that predicts a sample $y(n\text{-}M)$ from future samples , $y(n\text{-}M+1)$, …, $y(n)$.**



Linear signal estimation
c_M^* ··· 1 ··· c_1^* c_o^*
$e^{(i)}(n)$
$\hat{x}(n-i)$
Forward linear prediction
a_M^* a_{M-1}^* ··· a_k^* ··· a_2^* a_1^* 1
$e^{f}(n)$
$\hat{x}(n)$
Backward linear prediction
1 b_{M-1}^* ··· b_k^* ··· b_1^* b_0^*
$e^{b}(n)$
$\hat{x}(n-M)$
$x(n-M)$
$x(n-1)$
$x(n)$
$n-M$
$n-i$
n
Time (nT)

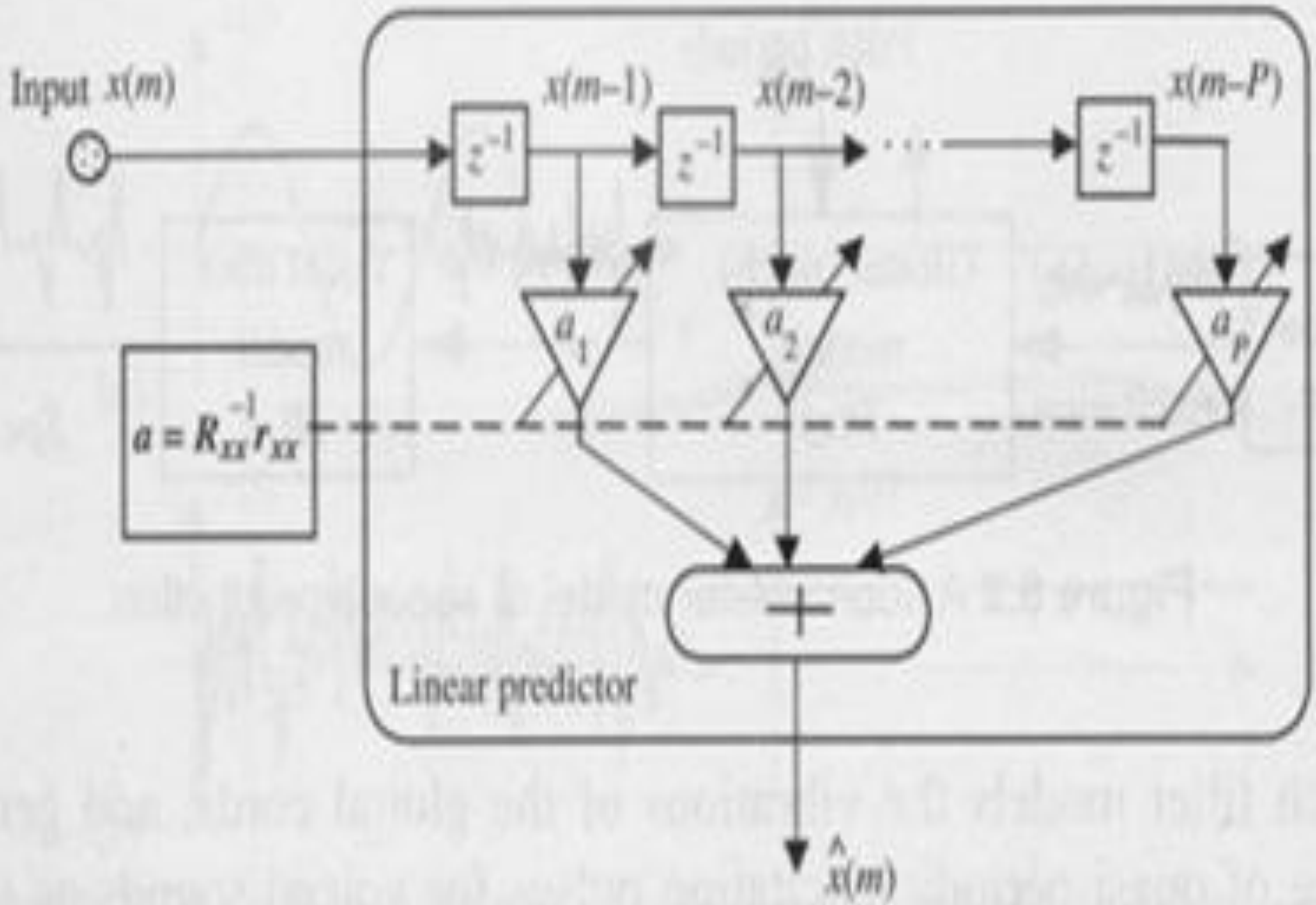


Figure 8.3 Block-diagram illustration of a linear predictor.

2.5.1 Forward Prediction

- **Assuming that the signals are stationary , the prediction process can be implemented by FIR filters which are relatively simple to realized, inherently stable and more practical .**
- **The prediction filter is represented by the tap-weights a_k, $k = 1,2, \ldots,M$**
- **Denote that $a = (a_1, \ldots, a_M)$**

The prediction can be expressed as

$$\hat{y}(n) = \sum_{k=1}^{M} a_k y(n-k) \qquad (2.136)$$

Then, the estimation error in FIR prediction is given by

$$e(n) = y(n) - \hat{y}(n) = y(n) - \sum_{k=1}^{M} a_k y(n-k) \qquad (2.137)$$

Let Y_{n-1} denote the M-dimensional linear space spanned by $y(n-1)$, $y(n-2)$,…, $y(n-M)$, and use $\hat{y}(n \mid Y_{n-1})$ to denote the predicted value of y (n) given this set of samples.
Then, $e_f(n) = y(n) - \hat{y}(n \mid Y_{n-1})$

- Differentiating the mean-squared estimation error with respect to each weight and setting the derivative to zero , we obtain the orthogonality equations that are enough to determine the weights :

$$\delta E[e^2(n)] / \delta a_i = 2 E[e(n) (\delta e(n) / \delta a_i)]$$
$$= -2 E[e(n) y(n-i)]$$
$$= 0 \quad \text{for } 0 \leq i , n \leq M$$

(2.138)

- **Inserting (2.137) for e_n, the orthogonality equations may be written in an equivalent form, known as normal equations**

$$E [\{ y(n) - \Sigma_{k=0}^{M} a_k y(n-k) \} \; y(n-i)] = 0$$

or
$$E[y(n)\, y(n-i)] = \Sigma_{k=0}^{M} a_k E [y(n-k)\, y(n-i)]$$

for $0 \leq n \leq M$

(2.139)

These Wiener-Hopf equations determine the optimal weights at the current time instant *n* .

We can write Eq.(2.138) in vector notation as

$$\mathbf{p} = E[\, y(n)\, \mathbf{Y}_{n-1} \,]$$

and $y(n) = \mathbf{a}^T \mathbf{Y}_{n-1}$

where $\mathbf{a} = (a_1, \ldots, a_M)^T$ is the optimum weight- vector ,

$$\mathbf{Y}_{n-1} = (y(n-1) \ldots y(n-M))^T$$

is the vector of observations up to time $n-1$,

- The optimal *weights* $\mathbf{a}^o$ and the prediction are then given by

$$\mathbf{a}^o = E\,[x(n)\mathbf{y}\,]\, E[\mathbf{y}\mathbf{y}^T]^{-1}$$

$$= R_{Y\, n-1}^{-1}\, p \qquad (2.140)\ .$$

Note that $\mathbf{p} = (p(1) \ldots p(M))^T$

where $p(j) = E[y(n)\, y(n-j)\,]$ (2.141)

$$
\begin{bmatrix}
R_{yy}(0) & R_{yy}(1) & R_{yy}(2) & \dots & R_{yy}(M-1) \\
R_{yy}(1) & R_{yy}(0) & R_{yy}(1) & \dots & R_{yy}(M-2) \\
R_{yy}(2) & R_{yy}(1) & R_{yy}(0) & \dots & R_{yy}(M-3) \\
\\
R_{yy}(M-1) & R_{yy}(M-2) & R_{yy}(M-3) & \dots & R_{yy}(0)
\end{bmatrix}
\begin{bmatrix}
a_1 \\ a_2 \\ a_3 \\ \\ a_M
\end{bmatrix}
=
\begin{bmatrix}
p(1) \\ p(2) \\ p(3) \\ \\ p(M)
\end{bmatrix}
\quad (2.142\ \ )
$$

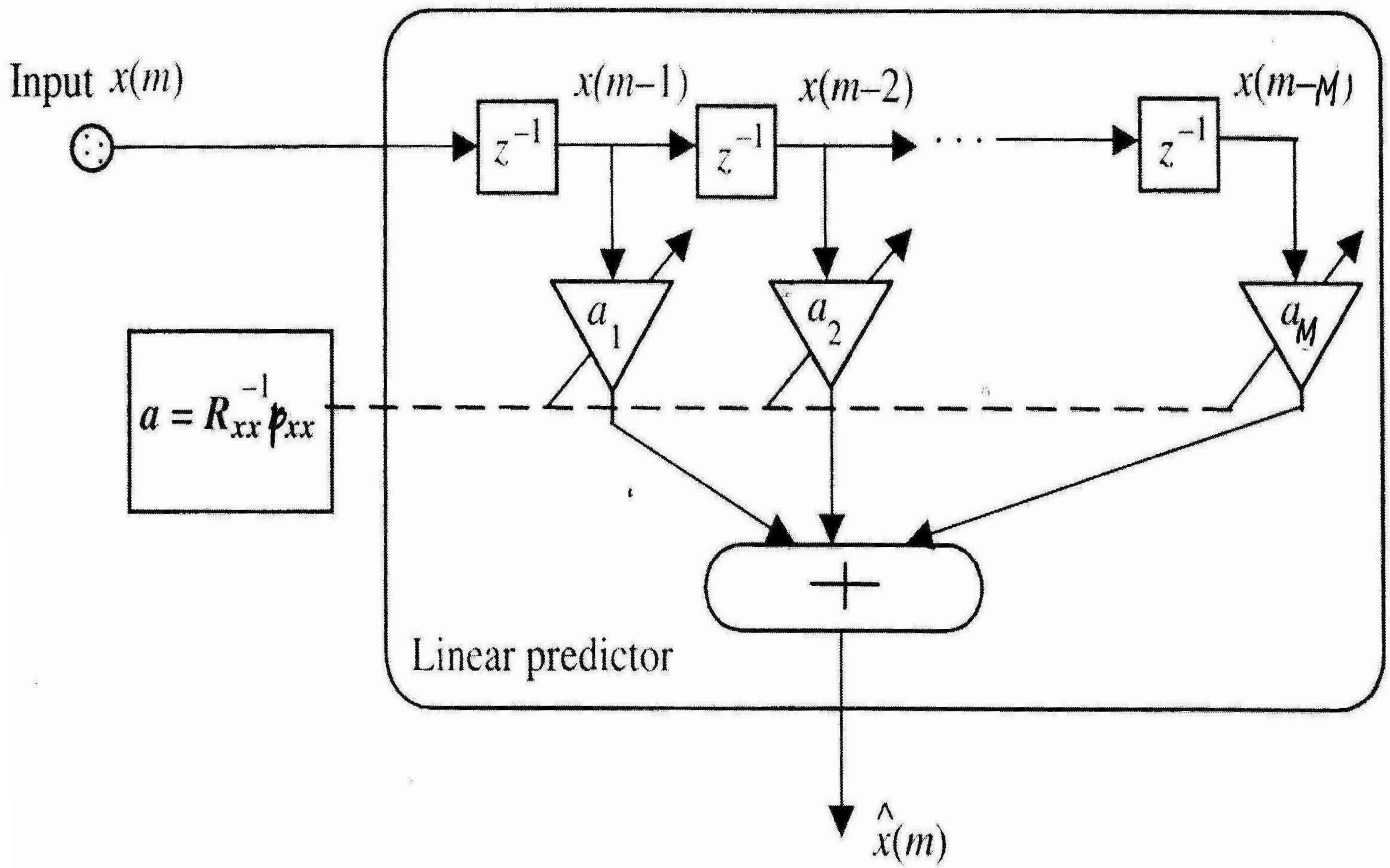


Figure 8.3 Block-diagram illustration of a linear predictor.

2.5.2 Backward Prediction

- **A backward prediction filter predicts a signal sample y(n-M) from M future samples { *y(n-1)* , *y(n-2)* ,…., *y(n-M+1*) = Y_{n-1} .**

The predicted value can be expressed as

$$\hat{y}(n-M) = \sum_{k=1}^{M} b_k\, y(n-k+1) \qquad (2.143)$$

The backward prediction error

$$e_b(n) = y(n-M) - \sum_{k=1}^{M} b_k\, y(n-k+1)$$

$$= y(n-M) - \hat{y}(n-M) \mid Y_{n-1}) \qquad (2.144)$$

- **The optimal coefficients can be obtained by he normal equation**

$$\mathbf{b} = \mathbf{R}_{Y\,n-1}^{-1}\,\mathbf{p}^{B} \qquad (2.145)$$

where $\mathbf{p}^{B} = (p(M)\;\; p(M-1)\; \ldots\; p(2)\;\;\; p(1))^{T}$

$$\mathbf{b} = (b_1\;\; b_2\; \ldots\;\; b_{M-1}\; b_M\;\;) \qquad (2.146)$$

$$
\begin{bmatrix}
R_{yy}(0) & R_{yy}(1) & R_{yy}(2) & \dots & R_{yy}(M-1) \\
R_{yy}(1) & R_{yy}(0) & R_{yy}(1) & \dots & R_{yy}(M-2) \\
R_{yy}(2) & R_{yy}(1) & R_{yy}(0) & \dots & R_{yy}(M-3) \\
 & & & & \\
R_{yy}(M-1) & R_{yy}(M-2) & R_{yy}(M-3) & \dots & R_{yy}(0)
\end{bmatrix}
\begin{bmatrix}
b_1 \\ b_2 \\ b_3 \\ \\ b_M
\end{bmatrix}
=
\begin{bmatrix}
p(M) \\ p(M-1) \\ p(M-2) \\ \\ p(1)
\end{bmatrix}
$$

(2.147)

2.6 Least Squares Optimal Filtering

(C.W. Therrien, pp.519-524)

- **In the minimum mean-squared estimation , the sequences (or vectors) y and x were regarded as random processes with known (or previously estimated) second-order (moment) statistics.**
- **Here , we are to consider the optimal filtering problem from a slightly different approach , the least square (LS) method.**

2.6.1 LS Method

Historical Notes :

The principle of least squares was introduced by the German mathematician Carl F. Gauss , who used it to determine the orbit of the asteroid Ceres in 1821 by formulating the estimation problem as an optimization problem.

Carl Gauss (1777 – 1855) was born in Braunschweig , Germany.

- **There is no presumed knowledge of the statistical properties of random vectors x and y beforehand . It is assumed that a typical data sequence of both x and y has been measured and recorded and that these sequences can be used to design the filter .**
- **We are to estimate** $\mathbf{x} = [\, x(0), x(1), \ldots, x(M) \,]^T$

  given the observation $\mathbf{y} = [\, y(0), y(1), \ldots, y(M) \,]^T$

- **If a causal FIR filter of length *M* is used , then the estimate for the given data sequence is**

$$\hat{x}(n) = \sum_{i=0}^{M} h(i)\, y(n-i) \qquad (2.148)$$

and the estimation error can be defined as

$$\varepsilon(n) = x(n) - \hat{x}(n) \qquad (2.149)$$

The approach here is to design the filter to minimize the sum of squared errors

$$S = \sum_{n=nI}^{nF} |\varepsilon(n)|^2 \qquad (2.150)$$

where n_1 and n_F are some initial and final values of n that define the interval over which to perform the minimization. Note that no probabilistic statements have been made in defining this problem.

- **The criterion (2.150) is called a least squares criterion.**
- **In matrix form, we can express (2.148) as**

$$\mathbf{x}^\wedge = \mathbf{Y}\,\mathbf{h} \qquad (2.151)$$

where $\mathbf{x}^\wedge = [\, x^\wedge(n_I)\ \ x^\wedge(n_I+1) \ldots x^\wedge(n_F)\,]^T$

$$\mathbf{Y} = \begin{bmatrix} y(n_I) & y(n_I-1) & \ldots & y(n_I-M) \\ y(n_I+1) & y(n_I) & \ldots & y(n_I-M+1) \\ & & & \\ y(n_F) & y(n_F-1) & \ldots & y(n_F-M) \end{bmatrix} \qquad (2.152)$$

and $\mathbf{h} = [\, h(0)\ \ h(1)\ \ldots\ h(M)\,]^T \qquad (2.153)$

- **The matrix Y is called the data matrix and has dimension $K \times M$ where $K = n_f - n_I + 1$.**
  It will be assumed that $K >> M$.
- **Define the error vector as**

$$\varepsilon = \mathbf{x} - \mathbf{x}^{\wedge} \qquad (2.154)$$

$\mathbf{x}$ and ε are K-dimensional vectors.
Then the problem is to minimize

$$S = \| \varepsilon \|^2 = \varepsilon^{*T} \varepsilon \qquad (2.155)$$

2.6.2 LS Optimal Filtering

- **A direct approach to this problem would be as follows . Substitute (2.148) and (2.149) into (2.150) and expand the result to obtain**

$$S = (\mathbf{x}-\mathbf{Yh})^{*T}(\mathbf{x}-\mathbf{Yh})$$
$$= \mathbf{x}^{*T}\mathbf{x} - \mathbf{h}^{*T}\mathbf{Y}^{*T}\mathbf{x} - \mathbf{x}^{*T}\mathbf{Yh} + \mathbf{h}^{*T}\mathbf{Y}^{*T}\mathbf{Y}\mathbf{h} \qquad (2.156)$$

- **Then by formal methods of differentiation , a necessary condition for the minimum can be found to be**

$$(\mathbf{Y}^{*T}\mathbf{Y})\mathbf{h} = \mathbf{Y}^{*T}\mathbf{x} \qquad (2.157)$$

This is the least squares Wiener –Hopf equation.

- **If Y has independent columns (i.e. , if it is of full rank) , then $\mathbf{Y}^{*T}\mathbf{Y}$ is also of full rank and (2.157) has the solution**

$$\mathbf{h} = (\mathbf{Y}^{*T}\mathbf{Y})^{-1}\mathbf{Y}^{*T}\mathbf{x} \qquad (2.158)$$

- The sum of the squared errors for the optimal filter can be found by returning to (2.156) and writing

$$S = (x-Yh)^{*T}(x-Yh)$$
$$= x^{*T}(x-Yh) - (Yh)^{*T}(x-Yh)$$
$$= x^{*T}x - x^{*T}Yh - h^{*T}(Y^{*T}x - Y^{*T}Yh)$$
$$= x^{*T}x - x^{*T}Yh \qquad (2.159)$$

- Denote that $Y^{+} = (Y^{*T}Y)^{-1}Y^{*T}$ (2.160)

Then we can write the optimal solution (2.158) of **h** as

$$h = Y^{+}x \qquad (2.161)$$

The matrix Y^{+} is known as Moore-Penrose pseudoinverse .

2.6.3 Least Squares Orthogonality

- **Theorem (Least Squares Orthogonality):**

  Let $\mathbf{x}^{\wedge} = \mathbf{Y}\mathbf{h}$ and $\varepsilon = \mathbf{x} - \mathbf{x}^{\wedge}$

  Then h minimizes the sum of squared errors

$$S = \| \varepsilon \|^2 \qquad (2.162)$$

  if h is chosen such that $\mathbf{Y}^{*T} \varepsilon = 0$

  Further , the sum of squared errors is given by

$$S = \mathbf{x}^{*T} \varepsilon \qquad (2.163)$$

Proof :

Let h be any vector of filter coefficients and $h_{\perp}$ be the vector that results in orthogonality.

Further , let ε be the error vector corresponding to h and $\varepsilon_\perp$ be the error vector corresponding to $h_\perp$. Then it follows that

$$\varepsilon = x - Yh = (x - Y h_\perp) + Y(h_\perp - h)$$
$$= \varepsilon_\perp + Y(h_\perp - h)$$

so that

$$\varepsilon^{*T} \varepsilon = [\varepsilon_\perp + Y(h_\perp - h)]^{*T} [\varepsilon_\perp + Y(h_\perp - h)]$$
$$= \varepsilon_\perp^{*T} \varepsilon_\perp + (h_\perp - h)^{*T} Y^{*T} \varepsilon_\perp$$
$$+ \varepsilon_\perp^{*T} Y (h_\perp - h) + (h_\perp - h)^{*T} Y^{*T} Y (h_\perp - h)$$

Since $\varepsilon_\perp^{*T} Y$ and $Y^{*T} \varepsilon_\perp$ are both zero by assumption , this leads to

$$S = \varepsilon^{*T} \varepsilon = \varepsilon_\perp^{*T} \varepsilon_\perp + (h_\perp - h)^{*T} Y^{*T} Y (h_\perp - h)$$

which is clearly minimized when $h = h_\perp$.

The minimum sum of squared errors is then

$$S = \varepsilon^{*T} \varepsilon = (x - Y h)^{*T} \varepsilon = x^{*T} \varepsilon$$

The last step follows because $Y^{*T} \varepsilon = 0$.

This proves the theorem.

The results of the above Theorem can now be applied to solve the optimal LS filtering problem. The Theorem requires that

$$Y^{*T}\varepsilon = Y^{*T}(x - Yh) = 0$$

which leads to the Wiener-Hopf equation (2.131)

$$Y^{*T}x = Y^{*T}Yh \qquad (2.164)$$

- **Further , from the theorem , the minimum sum of squared errors is given by**

$$S = x^{*T}\varepsilon = x^{*T}(x - Yh) = x^{*T}x - x^{*T}Yh \qquad (2.165)$$

as before (2.135) .

Example

The sequence {x(n); 0≦n ≦4 } = { 1,-1,1, -1,1 } is to be estimated from the observation sequence { y(n)} = { 1,-2,3, -4 ,5}

using an FIR filter of length *P*=2 . Choosing $n_I = 1$, $n_F = 4$.

We obtain the following least-squares problem :

The pseudo inverse of the data matrix is

The filter is the given by

Example #2 (Estimating Expected Values from Data)

- Assuming that $x^{(1)}$ $x^{(2)}$ … $x^{(K)}$ are samples of a random vector x.

  The expectation of a function φ(x) can be approximated as

  $$[\varphi(x)] = (1/K)\ \Sigma\ \varphi(x^K\ )$$

- When estimating for expectation of a quantity involving two random vectors from samples $x^{(1)}$ $x^{(2)}$ … $x^{(K)}$ and $y^{(1)}$ $y^{(2)}$ … $y^{(K)}$.

  The estimate for the expectation takes the form

  $$[\varphi(x,y)] = (1/K)\ \Sigma\ \varphi(x^K, y^K\ )$$

- **For complex-valued random vectors , define the data matrix X by**

$$\mathbf{X} = (\mathbf{x}^{(1)*} \ \mathbf{x}^{(2)*} \ \ldots \ \mathbf{x}^{(K)*})^T$$

Then since $\mathbf{X}^{*T}\mathbf{X} = \Sigma_{k=1}^{K} \mathbf{x}^{(k)*} \ \mathbf{x}^{(k)*T}$

The correlation matrix can be written as

$$\mathbf{R}_{xx} = (1/K) \mathbf{X}^{*T} \mathbf{X}$$

Then a typical element $r_{kl} = (1/K) x_k^{*T} x_l$

Toeplitz Matrix

- **A Toeplitz matrix** or **diagonal-constant matrix**, **named after Otto Toeplitz, is a matrix in which each descending diagonal from left to right is constant. For instance, the following matrix is a Toeplitz matrix:**

a b c d e
f a b c d
e f a b c
d e f a b
c d e f a

Toeplitz systems of form *Ax = b* can be solved by the Levinson- Durbin recursion in Θ(n^2) time.
Note that Levinson-Durbin recursion is a procedure in linear algebra to recursively calculate the solution to an equation involving a Toeplitz matrix.

Appendix

Norbert Wiener (1894-1964)

- Norbert Wiener was born in Columbia, Missouri.

  He entered Tufts College at age 11.

  Harvard awarded Wiener a Ph.D. in 1912, when he was a mere 18, for a dissertation on mathematical logic.

  In 1914, Wiener traveled to Europe, to study under Bertrand Russell and G.H.Hardy at Cambridge University, and under David Hilbert and Edmund Landau at the University of Gottingen .

- Wiener ‘s position in the Mathematics Department at MIT began in 1919. He was promoted to Professorship in 1932.

  He was a pioneer in the study of stochastic and noise processes, contributing work relevant to electronic engineering, communications and control systems.

- Wiener is perhaps best known as the founder of cybernetics, a field that formalizes the notion of feedback and has implications for engineering , systems control , computer science , biology , philosophy, and the organization of society.

- **In 1942 ,Wiener developed theory of signal transmission in the presence of a perturbative noise, in a classified monograph ,nicknamed "the yellow peril“ and then in Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications (1949).**

  In this book Wiener applies generalized harmonic analysis to stationary random signals and solves the problem of optimal elimination of the perturbative noise and of optimal prediction of the signal itself, with the help of a filtering operator.

  Quite independently, A.N. Kolmogoroff had announced results in the same domain at a time (1941) when scientific communications were interrupted.

Norman Levinson (, 1912 –1975)

was an American mathematician. He worked closely with Norbert Wiener in his early career. He joined the faculty of the MIT in 1937.

He received both his BS degree and his master degree in electrical engineering from MIT in 1934, where he had studied under Wiener and took almost all of the graduate-level courses in mathematics. He received the MIT Redfield Protor Traveling Fellowship to study at Cambridge University , with the assurance that MIT would reward him with a Ph. D. upon his return regardless of whatever he produced at Cambridge. Within the first four months in Cambridge, he had already produced two papers. In 1935, MIT awarded him with the Ph. D. in mathematics.

- **Norbert Wiener and Eberhard Hopf**

In retrospect, it seems an unlikely collaboration: American born Norbert Wiener (1894–1964) and Austrian born Eberhard Hopf (1902–1983).

- **The former was of European Jewish descent1 whilst the latter was educated in Berlin**

To the general public Norbert Wiener is widely recognised as the founder of modern cybernetics.

To mathematicians, however, he is primarily known for his highly innovative and fundamental work in what is now termed stochastic processes. His interest in randomness began in the early 1920s with studies of Brownian motion. This led him to harmonic analysis, Tauberian theorems and eventually to Paley–Wiener theory which was subsequently used to study problems involving more general stochastic processes.

- **Eberhard Hopf, on the other hand, is known primarily for his work in ergodic theory and partial differential equations—his bifurcation theory is a particular *tour de force* that is still used repeatedly today as a central element of stability analysis and dynamical systems theory.**
- **The two men differed greatly in personality. Wiener is widely acknowledged as having been absent-minded and his papers were hard to read: sometimes difficult results appeared with scarcely a proof and at other times he would present a lengthy proof of a triviality! It is also said that Wiener's lectures were difficult and often without structure.**
- **In contrast, Hopf was an excellent communicator: he had the ability to illuminate the most complex subjects and render them palatable to his colleagues and even to non-specialists.**

- **The Wiener–Hopf technique**

In 1930, having completed his Habilitation in Mathematical Astronomy at the University of Berlin, Hopf received a fellowship from the Rockefeller Foundation to study classical mechanics with Birkhoff (1884–1944) at Harvard College Observatory. A year later, and with the help of Norbert Wiener (who was already established at MIT), he joined the Department of Mathematics at the Massachusetts Institute of Technology on a temporary contract.

The collaboration between Wiener and Hopf was initiated by their mutual interest in the differential equations governing the radiation equilibrium of stars.

- It was at the end of his contract with MIT that Hopf took up a full professorship at the University of Leipzig.
  On the matter of Hopf’s return to Germany, Wiener was uncritical. He knew that, particularly when set against the United State’s economic depression, the post offered to Hopf was both lucrative and offered social prestige beyond that then available at MIT.
  He acknowledged that Hopf’s views were not strongly pro-Nazi and felt that the position was better filled by a man of moderate views. Wiener feared, however, that Hopf’s acceptance would severely damage his standing in the academic community.
  Indeed that seems to have been the case: in the years following the end of the second World War, Hopf suffered a substantial decrease in popularity which led to the neglect of his work and even to it being attributed to other mathematicians.
  It is, for example, suggested that Hopf’s name was dropped from the discrete version of the Wiener–Hopf equation, which is now referred to as the ‘Wiener filter’.

- **his standing in the academic community. Indeed that seems to have been the case: in the years following the end of**
- **the second World War, Hopf suffered a substantial decrease in popularity which led to the neglect of his work and**
- **even to it being attributed to other mathematicians.**
- **It is, for example, suggested [8] that Hopf's name was dropped**
- **from the discrete version of the Wiener–Hopf equation, which is now referred to as the 'Wiener filter'.**

- Eberhard Hopf was born in Salzburg, Austria, but his scientific career was divided between Germany and the United States. He received his Ph.D. in Mathematics in 1926 and his Habilitation in Mathematical Astronomy from the University of Berlin in 1929.
- In 1930 he received a fellowship from the Rockefeller Foundation to study classical mechanics with George Birkhoff at Harvard, but his appointment was at the Harvard College Observatory. In late 1931, with the help of Norbert Wiener, Hopf joined the Department of Mathematics of the Massachusetts Institute of Technology, accepting the position of Assistant Professor. While at MIT, Hopf did much of his work on ergodic theory.
- In Cambridge Hopf worked on many mathematical and astronomical subjects. His paper *On time average theorem in dynamics*, which appeared in the Proceedings of the National Academy of Sciences, is considered by many to be the first readable paper in modern ergodic theory. His book *Mathematical problems of radiative equilibrium* first appeared in 1934 and was reprinted in 1964. Another important contribution from this period is the theory of Wiener-Hopf equations, which he developed in collaboration with Norbert Wiener. By 1960, a discrete version of these equations was being extensively used in electrical engineering and geophysics, their use continuing until the present day. During this time, Hopf gained a reputation for his ability of illuminating the most complex subjects for his colleagues and even for non-specialists. Because of this talent, many discoveries and proofs of other mathematicians became easier to understand after they had been described by Hopf.
- In 1936 Hopf received and accepted an offer of a full professorship from the University of Leipzig. Hopf, with his wife Ilse and their infant daughter Barbara, returned to Germany, which by this time was under the control of the Nazi Party.
- The book *Ergodentheorie*, most of which was written when Hopf was still at the Massachusetts Institute of Technology, was published in 1937. In that book, containing only 81 pages, Hopf presented a precise and elegant summary of ergodic theory. In 1939 Hopf established ergodicity of the geodesic flow on compact manifolds of constant negative curvature. In 1940 Hopf was on the list of the invited lecturers to the International Congress of Mathematicians to be held in Cambridge, Massachusetts. Because of the start of World War II, however, the Congress was cancelled.
- In 1942 Hopf was drafted to work in the German Aeronautical Institute. In 1944, one year before the end of World War II, Hopf was appointed to a professorship at the University of Munich. In 1947, at the behest of Richard Courant he returned to the United States, where he presented the definitive solution of Hurewicz's problem.[*citation needed*]
- On 22 February 1949 Hopf became a US citizen and joined Indiana University at Bloomington as a Professor of Mathematics. In 1962 he was made Research Professor of Mathematics, staying in that position until his death.
- Hopf was never forgiven by many people for his moving to Germany in 1936, where the Nazi party was in power. As a result, most of his work in ergodic theory and topology was neglected or even attributed to others in the years following the end of World War II. An example of this was the expulsion of Hopf's name from the discrete version of the Wiener–Hopf equations, which were frequently referred to as "Wiener filter".