Chapter 2
Wiener Filtering

Abstract Before moving to the actual adaptive filtering problem, we need to solve the optimum linear filtering problem (particularly, in the mean-square-error sense). We start by explaining the analogy between linear estimation and linear optimum filtering. We develop the principle of orthogonality, derive the Wiener–Hopf equation (whose solution lead to the optimum Wiener filter) and study the error surface. Finally, we applied the Wiener filter to the problem of linear prediction (forward and backward).

2.1 Optimal Linear Mean Square Estimation

Lets assume we have a set of samples $\{x(n)\}$ and $\{d(n)\}$ coming from a jointly wide sense stationary (WSS) process with zero mean. Suppose now we want to find a linear estimate of $d(n)$ based on the L-most recent samples of $x(n)$, i.e.,

$$\hat{d}(n) = \mathbf{w}^T \mathbf{x}(n) = \sum_{l=0}^{L-1} w_l x(n-l), \quad \mathbf{w}, \mathbf{x}(n) \in \mathbb{R}^L \quad \text{and} \quad n = 0, 1, \ldots \tag{2.1}$$

The introduction of a particular criterion to quantify how well $d(n)$ is estimated by $\hat{d}(n)$ would influence how the coefficients w_l will be computed. We propose to use the *Mean Squared Error* (MSE), which is defined by

$$J_{\text{MSE}}(\mathbf{w}) = E\left[|e(n)|^2\right] = E\left[|d(n) - \hat{d}(n)|^2\right], \tag{2.2}$$

where $E[\cdot]$ is the expectation operator and $e(n)$ is the estimation error. Then, the estimation problem can be seen as finding the vector $\mathbf{w}$ that minimizes the cost function $J_{\text{MSE}}(\mathbf{w})$. The solution to this problem is sometimes called the *stochastic*

L. Rey Vega and H. Rey, *A Rapid Introduction to Adaptive Filtering*,
SpringerBriefs in Electrical and Computer Engineering,
DOI: 10.1007/978-3-642-30299-2_2, © The Author(s) 2013

least squares solution, which is in contrast with the deterministic solution we will study in Chap. 5

If we choose the MSE cost function (2.2), the optimal solution to the linear estimation problem can be presented as:

$$\mathbf{w}_{\text{opt}} = \arg \min_{\mathbf{w} \in \mathbb{R}^L} J_{\text{MSE}}(\mathbf{w}). \tag{2.3}$$

Replacing (2.1) in (2.2), the latter can be expanded as

$$J_{\text{MSE}}(\mathbf{w}) = E\left[|d(n)|^2 - 2d(n)\mathbf{x}(n)^T\mathbf{w} + \mathbf{w}^T\mathbf{x}(n)\mathbf{x}^T(n)\mathbf{w}\right]. \tag{2.4}$$

As this is a quadratic form, the optimal solution will be at the point where the cost function has zero gradient, i.e.,

$$\nabla_{\mathbf{w}} J_{\text{MSE}}(\mathbf{w}) = \frac{\partial J_{\text{MSE}}}{\partial \mathbf{w}} = \mathbf{0}_{L\times 1}, \tag{2.5}$$

or in other words, the partial derivative of J_{MSE} with respect to each coefficient w_l should be zero.

2.2 The Principle of Orthogonality

Using (2.1) in (2.2), we can compute the gradient as

$$\frac{\partial J_{\text{MSE}}}{\partial \mathbf{w}} = 2E\left[e(n)\frac{\partial e(n)}{\partial \mathbf{w}}\right] = -2E\left[e(n)\mathbf{x}(n)\right]. \tag{2.6}$$

Then, at the minimum,[1] the condition that should hold is:

$$E\left[e_{\min}(n)\mathbf{x}(n)\right] = \mathbf{0}_{L\times 1}, \tag{2.7}$$

or equivalently

$$E\left[e_{\min}(n)x(n-l)\right] = 0, \quad l = 0, 1, \ldots, L-1. \tag{2.8}$$

This is called the *principle of orthogonality*, and it implies that the optimal condition is achieved if and only if the error $e(n)$ is decorrelated from the samples $x(n-l)$, $l = 0, 1, \ldots, L-1$. Actually, the error will also be decorrelated from the estimate $\hat{d}(n)$ since

[1] The Hessian matrix of J_{MSE} is positive definite (in general), so the gradient of J_{MSE} is nulled at its minimum.

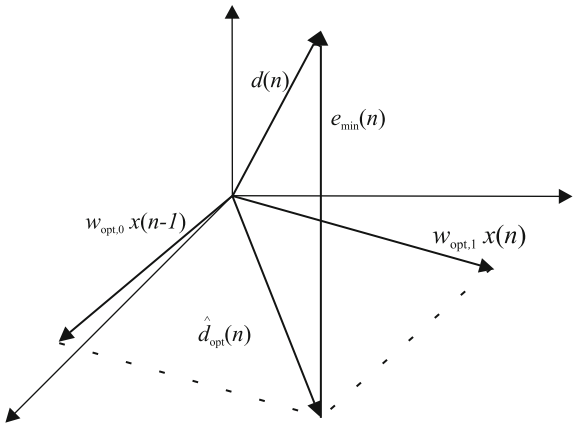

Fig. 2.1 Illustration of the principle of orthogonality for $L = 2$. The optimal error $e_{\min}(n)$ is orthogonal to the input samples $x(n)$ and $x(n-1)$, and to the optimal estimate $\hat{d}_{\text{opt}}(n)$. It should be noticed that the notion of orthogonality in this chapter is equivalent to the notion of decorrelation

$$E\left[e_{\min}(n)\hat{d}_{\text{opt}}(n)\right] = E\left[e_{\min}(n)\mathbf{w}_{\text{opt}}^T\mathbf{x}(n)\right] = \mathbf{w}_{\text{opt}}^T E\left[e_{\min}(n)\mathbf{x}(n)\right] = 0. \quad (2.9)$$

Fig. 2.1 illustrates the orthogonality principle for the case $L = 2$.

2.3 Linear Optimum Filtering

Consider a signal $x(n)$ as the input to a finite impulse response (FIR) filter of length L, $\mathbf{w}_{\text{T}} = [w_{\text{T},0}, w_{\text{T},1}, \ldots, w_{\text{T},L-1}]^T$. This filtering operation generates an output

$$y(n) = \mathbf{w}_{\text{T}}^T\mathbf{x}(n), \quad (2.10)$$

with $\mathbf{x}(n) = [x(n), x(n-1), \ldots, x(n-L+1)]^T$. As the output of the filter is observed, it can be corrupted by an additive measurement noise $v(n)$, leading to a linear regression model for the observed output

$$d(n) = \mathbf{w}_{\text{T}}^T\mathbf{x}(n) + v(n). \quad (2.11)$$

It should be noticed that this linear regression model can also be used even if the input-output relation of the given data pairs $[\mathbf{x}(n), d(n)]$ is nonlinear, with $\mathbf{w}_{\text{T}}$ being a linear approximation to the actual relation between them. In that case, in $v(n)$ there would be a component associated to the additive noise perturbations, but also another one representing, for example, modeling errors.

In the context of (2.11), we can look at $\mathbf{w}_{\text{T}}$ as the quantity to be estimated by a linear filter $\mathbf{w} \in \mathbb{R}^L$, with (2.1) giving the output of this filter. This output can still be seen as an estimate of the reference signal $d(n)$ or the system's output $y(n)$. Therefore, the problem of optimal filtering is analogous to the one of linear estimation.

When J_{MSE} is the cost function to be optimized, the orthogonality principle (2.7) holds, which can be put as:

$$E\left[e_{\min}(n)\mathbf{x}(n)\right] = E\left\{\left[d(n) - \mathbf{w}_{\text{opt}}^T\mathbf{x}(n)\right]\mathbf{x}(n)\right\} = \mathbf{0}_{L\times 1}, \tag{2.12}$$

From (2.12) we can conclude that given the signals $x(n)$ and $d(n)$, we can always assume that $d(n)$ was generated by the linear regression model (2.11). To do this, the system $\mathbf{w}_{\text{T}}$ would be equal to the optimal filter $\mathbf{w}_{\text{opt}}$, while $v(n)$ would be associated to the residual error $e_{\min}(n)$, which will be uncorrelated to the input $\mathbf{x}(n)$ [1].

It should be noticed that (2.8) is not just a condition for the cost function to reach its minimum, but also a mean for testing whether a linear filter is operating in the optimal condition. Here, the principle of orthogonality illustrated in Fig. 2.1 can be interpreted as follows: at time n the input vector $\mathbf{x}(n) = [x(n), x(n-1)]^T$ will pass through the optimal filter $\mathbf{w}_{\text{opt}} = [w_{\text{opt},0}, w_{\text{opt},1}]^T$ to generate the output $\hat{d}_{\text{opt}}(n)$. Given $d(n)$, $\hat{d}_{\text{opt}}(n)$ is the only element in the space spanned by $\mathbf{x}(n)$ that leads to an error $e(n)$ that is orthogonal to $x(n)$, $x(n-1)$, and $\hat{d}_{\text{opt}}(n)$.[2]

2.4 Wiener–Hopf Equation

Now we focus on the computation of the optimal solution. From (2.12), we have

$$E\left[\mathbf{x}(n)\mathbf{x}^T(n)\right]\mathbf{w}_{\text{opt}} = E\left[\mathbf{x}(n)d(n)\right]. \tag{2.13}$$

We introduce the following definitions

$$\mathbf{R}_{\mathbf{x}} = E\left[\mathbf{x}(n)\mathbf{x}^T(n)\right] \quad \text{and} \quad \mathbf{r}_{\mathbf{x}d} = E\left[\mathbf{x}(n)d(n)\right] \tag{2.14}$$

for the input autocorrelation matrix and the cross correlation vector, respectively. Note that as the joint process is WSS, the matrix $\mathbf{R}_{\mathbf{x}}$ is symmetric, positive definite[3] and Toeplitz [2]. Using these definitions, equation (2.13) can be put as

$$\mathbf{R}_{\mathbf{x}}\mathbf{w}_{\text{opt}} = \mathbf{r}_{\mathbf{x}d}. \tag{2.15}$$

This is the compact matrix form of a set of L equations known as *Wiener–Hopf equations* and provides a way for computing the optimal filter (in MSE sense) based on

[2] In this chapter we use the idea of orthogonality as a synonym of decorrelation. This orthogonality we are referring to is given in a certain space of random variables with finite variance and an inner product between x and y defined by $E[xy]$.

[3] The autocorrelation matrix is certainly positive semidefinite. For it not to be positive definite, some linear dependencies between the random variable conforming $\mathbf{x}(n)$ would be required. However, this is very rare in practice.

some statistical properties of the input and reference processes. Under the assumption on the positive definiteness of $\mathbf{R}_\mathbf{x}$ (so that it will be nonsingular), the solution to (2.15) is:

$$\mathbf{w}_{\text{opt}} = \mathbf{R}_\mathbf{x}^{-1}\mathbf{r}_{\mathbf{x}d}, \tag{2.16}$$

which is known as the *Wiener filter*. An alternative way to find it is the following. Using the definitions (2.14) into (2.4) results in

$$J_{\text{MSE}}(\mathbf{w}) = E\left[|d(n)|^2\right] - 2\mathbf{r}_{\mathbf{x}d}^T\mathbf{w} + \mathbf{w}^T\mathbf{R}_\mathbf{x}\mathbf{w}. \tag{2.17}$$

In addition, it can be easily shown that the following factorization holds:

$$\mathbf{w}^T\mathbf{R}_\mathbf{x}\mathbf{w} - 2\mathbf{r}_{\mathbf{x}d}^T\mathbf{w} = (\mathbf{R}_\mathbf{x}\mathbf{w} - \mathbf{r}_{\mathbf{x}d})^T\,\mathbf{R}_\mathbf{x}^{-1}\,(\mathbf{R}_\mathbf{x}\mathbf{w} - \mathbf{r}_{\mathbf{x}d}) - \mathbf{r}_{\mathbf{x}d}^T\mathbf{R}_\mathbf{x}^{-1}\mathbf{r}_{\mathbf{x}d}. \tag{2.18}$$

Replacing (2.18) in (2.17) leads to

$$J_{\text{MSE}}(\mathbf{w}) = E\left[|d(n)|^2\right] - \mathbf{r}_{\mathbf{x}d}^T\mathbf{R}_\mathbf{x}^{-1}\mathbf{r}_{\mathbf{x}d} + \left(\mathbf{w} - \mathbf{R}_\mathbf{x}^{-1}\mathbf{r}_{\mathbf{x}d}\right)^T\mathbf{R}_\mathbf{x}\left(\mathbf{w} - \mathbf{R}_\mathbf{x}^{-1}\mathbf{r}_{\mathbf{x}d}\right). \tag{2.19}$$

Using the fact that $\mathbf{R}_\mathbf{x}$ is positive definite (and therefore, so is its inverse), it turns out that the cost function reaches its minimum when the filter takes the form of (2.16), i.e., the Wiener filter. The minimum MSE value (MMSE) on the surface (2.19) is:

$$J_{\text{MMSE}} = J_{\text{MSE}}(\mathbf{w}_{\text{opt}}) = E\left[|d(n)|^2\right] - \mathbf{r}_{\mathbf{x}d}^T\mathbf{R}_\mathbf{x}^{-1}\mathbf{r}_{\mathbf{x}d} = E\left[|d(n)|^2\right] - E\left[|\hat{d}_{\text{opt}}(n)|^2\right]. \tag{2.20}$$

We could have also arrived to this result by noticing that $e_{\min}(n) = d(n) - \hat{d}_{\text{opt}}(n)$ and using the orthogonality principle as in (2.9). Therefore, the MMSE is given by the difference between the variance of the reference signal $d(n)$ and the variance of its optimal estimate $\hat{d}_{\text{opt}}(n)$.

It should be noticed that if the signals $x(n)$ and $d(n)$ are orthogonal ($\mathbf{r}_{\mathbf{x}d} = 0$), the optimal filter will be the null vector and $J_{\text{MMSE}} = E\left[|d(n)|^2\right]$. This is reasonable since nothing can be done with the filter $\mathbf{w}$ if the input signal carries no information about the reference signal (as they are orthogonal). Actually, (2.17) shows that in this case, if any of the filter coefficients is nonzero, the MSE would be increased by the term $\mathbf{w}^T\mathbf{R}_\mathbf{x}\mathbf{w}$, so it would not be optimal. On the other hand, if the reference signal is generated by passing the input signal through a system $\mathbf{w}_\text{T}$ as in (2.11), with the noise $v(n)$ being uncorrelated from the input $x(n)$, the optimal filter will be

$$\mathbf{w}_{\text{opt}} = \mathbf{R}_\mathbf{x}^{-1}\mathbf{r}_{\mathbf{x}d} = \mathbf{R}_\mathbf{x}^{-1}E\left\{\mathbf{x}(n)\left[\mathbf{x}^T(n)\mathbf{w}_\text{T} + v(n)\right]\right\} = \mathbf{w}_\text{T}. \tag{2.21}$$

This means that the Wiener solution will be able to identify the system $\mathbf{w}_\text{T}$ with a resulting error given by $v(n)$. Therefore, in this case $J_{\text{MMSE}} = E\left[|v(n)|^2\right] = \sigma_v^2$.

Finally, it should be noticed that the autocorrelation matrix admits the eigendecomposition:

$$\mathbf{R}_\mathbf{x} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T, \tag{2.22}$$

with $\boldsymbol{\Lambda}$ being a diagonal matrix determined by the eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{L-1}$ of $\mathbf{R}_\mathbf{x}$, and $\mathbf{Q}$ a (unitary) matrix that has the associated eigenvectors $\mathbf{q}_0, \mathbf{q}_1, \ldots, \mathbf{q}_{L-1}$ as its columns [2]. Lets define the *misalignment vector* (or weight error vector)

$$\tilde{\mathbf{w}} = \mathbf{w}_{\text{opt}} - \mathbf{w}, \tag{2.23}$$

and its transformed version

$$\mathbf{u} = \mathbf{Q}^T \tilde{\mathbf{w}}. \tag{2.24}$$

Using (2.20), (2.16), (2.23), (2.22), and (2.24) in (2.19), results in

$$J_{\text{MSE}}(\mathbf{w}) = J_{\text{MMSE}} + \mathbf{u}^T \boldsymbol{\Lambda} \mathbf{u}. \tag{2.25}$$

This is called the *canonical form* of the quadratic form $J_{\text{MSE}}(\mathbf{w})$ and it contains no cross-product terms. Since the eigenvalues are non-negative, it is clear that the surface describes an elliptic hyperparaboloid, with the eigenvectors being the principal axes of the hyperellipses of constant MSE value.

2.5 Example: Linear Prediction

In the filtering problem studied in this chapter, we use the L-most recent samples $x(n), x(n-1), \ldots, x(n-L+1)$ and estimate the value of the reference signal at time n. The idea behind a *forward linear prediction* is to use a certain set of samples $x(n-1), x(n-2), \ldots$ to estimate (with a linear combination) the value $x(n+k)$ for $k \geq 0$. On the other hand, in a *backward linear prediction* (also known as *smoothing*) the set of samples $x(n), x(n-1), \ldots, x(n-M+1)$ is used to linearly estimate the value $x(n-k)$ for $k \geq M$.

2.5.1 Forward Linear Prediction

Firstly, we explore the forward prediction case of estimating $x(n)$ based on the previous L samples. Since $\mathbf{x}(n-1) = [x(n-1), x(n-2), \ldots, x(n-L)]^T$, using a transversal filter $\mathbf{w}$ the forward linear prediction error can be put as

$$e_{f,L}(n) = x(n) - \sum_{j=1}^{L} w_j x(n-j) = x(n) - \mathbf{w}^T \mathbf{x}(n-1). \tag{2.26}$$

To find the optimum forward filter $\mathbf{w}_{f,L} = [w_{f,1}, w_{f,2}, \ldots, w_{f,L}]^T$ we minimize the MSE. The input correlation matrix would be

$$E\Big[\mathbf{x}(n-1)\mathbf{x}^T(n-1)\Big] = E\Big[\mathbf{x}(n)\mathbf{x}^T(n)\Big] = \mathbf{R}_\mathbf{x} = \begin{bmatrix} r_x(0) & r_x(1) & \cdots & r_x(L-1) \\ r_x(1) & r_x(0) & \cdots & r_x(L-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(L-1) & r_x(L-2) & \cdots & r_x(0) \end{bmatrix}, \tag{2.27}$$

where $r_x(k)$ is the autocorrelation function for lag k of the WSS input process. As for the cross correlation vector, the desired signal would be $x(n)$, so

$$\mathbf{r}_f = E\left[\mathbf{x}(n-1)x(n)\right] = [r_x(1), r_x(2), \ldots, r_x(L)]^T. \tag{2.28}$$

As $\mathbf{w}_{f,L}$ will be the Wiener filter, it satisfies the modified Wiener–Hopf equation

$$\mathbf{R}_\mathbf{x}\mathbf{w}_{f,L} = \mathbf{r}_f. \tag{2.29}$$

In addition, we can use (2.20) to write the forward prediction error power

$$P_{f,L} = r_x(0) - \mathbf{r}_f^T\mathbf{w}_{f,L}. \tag{2.30}$$

Actually, (2.29) and (2.30) can be put together into the *augmented Wiener–Hopf equation* as:

$$\begin{bmatrix} r_x(0) & \mathbf{r}_f^T \\ \mathbf{r}_f & \mathbf{R}_\mathbf{x} \end{bmatrix} \mathbf{a}_L = \begin{bmatrix} P_{f,L} \\ \mathbf{0}_{L\times 1} \end{bmatrix}, \tag{2.31}$$

where $\mathbf{a}_L = \begin{bmatrix} 1 & -\mathbf{w}_{f,L}^T \end{bmatrix}^T$. In fact the block matrix on the left hand side is the autocorrelation matrix of the $(L+1)\times 1$ input vector $[x(n), x(n-1), \ldots, x(n-L)]^T$. According to (2.26), when this vector passes through the filter $\mathbf{a}_L$ it produces the forward linear prediction error as its output. For this reason, $\mathbf{a}_L$ is known as the *forward prediction error filter*.

Now, in order to estimate $x(n)$ we might use only the $(L-i)$-most recent samples, leading to a prediction error

$$e_{f,L-i}(n) = x(n) - \sum_{j=1}^{L-i} w_j x(n-j). \tag{2.32}$$

But the orthogonality principle tells us that when using the optimum forward filter

$$E\left[e_{f,L}(n)\mathbf{x}(n-1)\right] = \mathbf{0}_{L\times 1}. \tag{2.33}$$

Then, we can see that for $1 \le i \le L$,

$$E\left[e_{f,L}(n)e_{f,L-i}(n-i)\right] = E\left\{e_{f,L}(n)\mathbf{a}_{L-i}^{T}\left[x(n-i),\ldots,x(n-L)\right]^{T}\right\} = 0. \tag{2.34}$$

Therefore, we see that as $L \to \infty$, $E\left[e_f(n)e_f(n-i)\right] = 0$, which means that the sequence of forward errors $e_f(n)$ is asymptotically white. This means that a sufficiently long forward prediction error filter is capable of *whitening* a stationary discrete-time stochastic process applied to its input.

2.5.2 Backward Linear Prediction

In this case we start by trying to estimate $x(n-L)$ based on the next L samples, so the backward linear prediction error can be put as

$$e_{b,L}(n) = x(n-L) - \sum_{j=1}^{L} w_j x(n-j+1) = x(n-L) - \mathbf{w}^T\mathbf{x}(n). \tag{2.35}$$

To find the optimum backward filter $\mathbf{w}_{b,L} = [w_{b,1}, w_{b,2}, \ldots, w_{b,L}]^T$ we minimize the MSE. Following a similar procedure as before to solve the Wiener filter, the augmented Wiener–Hopf equation has the form

$$\begin{bmatrix} \mathbf{R}_{\mathbf{x}} & \mathbf{r}_b \\ \mathbf{r}_b^T & r_x(0) \end{bmatrix} \mathbf{b}_L = \begin{bmatrix} \mathbf{0}_{L\times 1} \\ P_{b,L} \end{bmatrix}, \tag{2.36}$$

where $\mathbf{r}_b = E\left[\mathbf{x}(n)x(n-L)\right] = [r_x(L), r_x(L-1), \ldots, r_x(1)]^T$, $P_{b,L} = r_x(0) - \mathbf{r}_b^T\mathbf{w}_{b,L}$, and $\mathbf{b}_L = \begin{bmatrix} -\mathbf{w}_{b,L}^T & 1 \end{bmatrix}^T$ is the *backward prediction error filter*.

Consider now a stack of backward prediction error filters from order 0 to L. If we compute the errors $e_{b,i}(n)$ for $0 \le i \le L$, it leads to

$$\mathbf{e}_b(n) = \begin{bmatrix} e_{b,0}(n) \\ e_{b,1}(n) \\ e_{b,2}(n) \\ \vdots \\ e_{b,L-1}(n) \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}_{1\times(L-1)} \\ \mathbf{b}_1^T & \mathbf{0}_{1\times(L-2)} \\ \mathbf{b}_2^T & \mathbf{0}_{1\times(L-3)} \\ \vdots & \vdots \\ -\mathbf{w}_{b,L-1}^T & 1 \end{bmatrix} \mathbf{x}(n) = \mathbf{T}_b\mathbf{x}(n). \tag{2.37}$$

The $L \times L$ matrix $\mathbf{T}_b$, which is defined in terms of the backward prediction error filter coefficients, is lower triangular with 1's along its main diagonal. The transformation (2.37) is known as *Gram–Schmidt orthogonalization* [3], which defines a one-to-one correspondence between $\mathbf{e}_b(n)$ and $\mathbf{x}(n)$.[4]

In this case, the principle of orthogonality states that

[4] The Gram–Schmidt process is also used for the orthogonalization of a set of linearly independent vectors in a linear space with a defined inner product.

$$E\left[e_{b,i}(n)x(n-k)\right] = 0 \qquad 0 \leq k \leq i-1. \tag{2.38}$$

Then, it is easy to show that, at each time n, the sequence of backward prediction errors of increasing order $\{e_{b,i}(n)\}$ will be decorrelated. This means that the autocorrelation matrix of the backward prediction errors is diagonal. More precisely,

$$E\left[\mathbf{e}_b(n)\mathbf{e}_b^T(n)\right] = \text{diag}\{P_{b,i}\} \qquad 0 \leq i \leq L-1. \tag{2.39}$$

Another way to get to this result comes from using (2.37) to write

$$E\left[\mathbf{e}_b(n)\mathbf{e}_b^T(n)\right] = \mathbf{T}_b\mathbf{R}_\mathbf{x}\mathbf{T}_b^T. \tag{2.40}$$

By definition, this is a symmetric matrix. From (2.36), it is easy to show that $\mathbf{R}_\mathbf{x}\mathbf{T}_b^T$ is a lower triangular matrix with $P_{b,i}$ being the elements on its main diagonal. However, since $\mathbf{T}_b$ is also a lower triangular matrix, the product of both matrices must retain the same structure. But it has to be also symmetric, and hence, it must be diagonal.

Moreover, since the determinant of $\mathbf{T}_b$ is 1, it is a nonsingular matrix. Therefore, from (2.39) and (2.40) we can put

$$\mathbf{R}_\mathbf{x}^{-1} = \mathbf{T}_b^T\,\text{diag}\{P_{b,i}\}^{-1}\mathbf{T}_b = \left(\text{diag}\{P_{b,i}\}^{-1/2}\mathbf{T}_b\right)^T\,\text{diag}\{P_{b,i}\}^{-1/2}\mathbf{T}_b. \tag{2.41}$$

This is called the *Cholesky decomposition* of the inverse of the autocorrelation matrix. Notice that the inverse of the autocorrelation matrix is factorized into the product of an upper and lower triangular matrices that are related to each other through a transposition operation. These matrices are completely determined by the coefficients of the backward prediction error filter and the backward prediction error powers.

2.6 Final Remarks on Linear Prediction

It should be noticed that a sufficiently long (high order) forward prediction error filter transforms a (possibly) correlated signal into a white sequence of forward errors (the sequence progresses with time index n). On the other hand, the Gram–Schmidt orthogonalization transforms the input vector $\mathbf{x}(n)$ into an equivalent vector $\mathbf{e}_b(n)$, where its components (associated to the order of the backward prediction error filter) are uncorrelated.

By comparing the results shown for forward and backward predictions, it can be seen that: i) the forward and backward prediction error powers are the same. ii) the coefficients of the optimum backward filter can be obtained by reversing the ones of the optimum forward filter. Based on these relations, the *Levinson–Durbin algorithm*

[2] provides a mean of recurrently solving the linear prediction problem of order L with a complexity $O(L^2)$ instead of $O(L^3)$.[5]

2.7 Further Comments

The MSE defined in (2.2) uses the linear estimator $\hat{d}(n)$ defined in (2.1). If we relax the linear constraint on the estimator and look for a function of the input, i.e., $\hat{d}(n) = g(\mathbf{x}(n))$, the optimal estimator in mean square sense is given by the conditional expectation $E[d(n)|\mathbf{x}(n)]$ [4]. Its calculation requires knowledge of the joint distribution between $d(n)$ and $\mathbf{x}(n)$, and in general, it is a nonlinear function of $\mathbf{x}(n)$ (unless certain symmetry conditions on the joint distribution are fulfilled, as it is the case for Gaussian distributions). Moreover, once calculated it might be very hard to implement it. For all these reasons, linear estimators are usually preferred (which as we have seen, depend only on second order statistics).

On a historical note, Norbert Wiener solved a continuous-time prediction problem under causality constraints by means of an elegant technique now known as the Wiener–Hopf factorization technique. This is a much more complicated problem than the one presented in 2.3. Later, Norman Levinson formulated the Wiener filter in discrete time.

It should be noticed that the orthogonality principle used to derive the Wiener filter does not apply to FIR filters only; it can be applied to IIR (infinite impulse response) filtering, and even noncausal filtering. For the general case, the output of the noncausal filter can be put as

$$\hat{d}(n) = \sum_{i=-\infty}^{\infty} w_i x(n-i). \tag{2.42}$$

Then, minimizing the mean square error leads to the Wiener–Hopf equations

$$\sum_{i=-\infty}^{\infty} w_{\text{opt},i} r_x(k-i) = r_{xd}(k), \quad -\infty < k < \infty \tag{2.43}$$

which can be solved using Z-transform methods [5]. In addition, a general expression for the minimum mean square error is

[5] We use the Landau notation in order to quantify the computational cost of a numerical operation [3]. Assume that the numerical cost or memory requirement of an algorithm are given by a positive function $f(n)$, where n is the problem dimensionality. The notation $f(n) = O(g(n))$, where $g(n)$ is a given positive function (usually simpler than $f(n)$), means that there exists constants $M, n_0 > 0$ such that:

$$f(n) \leq Mg(n), \ \forall n \geq n_0$$

.

$$J_{\text{MMSE}} = r_d(0) - \sum_{i=-\infty}^{\infty} w_{\text{opt},i} r_{xd}(i) \tag{2.44}$$

From this general case, we can derive the FIR filter studied before (index i in the summation and lag k in (2.43) go from 0 to $L-1$) and the causal IIR filter (index i in the summation and lag k in (2.43) go from 0 to ∞).

Finally we would like to comment on the stationarity of the processes. We assume the input and reference processes were WSS. If this were not the case, the statistics would be time-dependent. However, we could still find the Wiener filter at each time n as the one that makes the estimation error orthogonal to the input, i.e., the principle of orthogonality still holds. A less costly alternative would be to recalculate the filter for every block of N signal samples. However, nearly two decades after Wiener's work, Rudolf Kalman developed the *Kalman filter*, which is the optimum mean square linear filter for nonstationary processes (evolving under a certain state space model) and stationary ones (converging in steady state to the Wiener's solution).

References

1. A.H. Sayed, *Adaptive Filters* (John Wiley & Sons, Hoboken, 2008)
2. S. Haykin, *Adaptive Filter Theory*, 4th edn. (Prentice-Hall, Upper Saddle River, 2002)
3. G.H. Golub, C.F. van Loan, *Matrix Computations* (The John Hopkins University Press, Baltimore, 1996)
4. B.D.O. Anderson, J.B. Moore, *Optimal Filtering* (Prentice-Hall, Englewood Cliffs, 1979)
5. T. Kailath, A.H. Sayed, B. Hassibi, *Linear estimation* (Prentice-Hall, Upper Saddle River, 2000)

http://www.springer.com/978-3-642-30298-5

A Rapid Introduction to Adaptive Filtering

Vega, L.R.; Rey, H.

2013, XII, 122 p. 23 illus., Softcover
ISBN: 978-3-642-30298-5