Institut EURECOM
2229, route des Crêtes
B.P. 193
06904 Sophia Antipolis Cedex
FRANCE

Advanced Signal Processing (TdS2)

Dirk T.M. Slock

Spring 2000

Telephone: +33 (0)4 93 00 26 26
Fax: +33 (0)4 93 00 26 27
Dirk T.M. Slock: +33 (0)4 93 00 26 06

E-mail: slock@eurecom.fr
WWW: http://www.eurecom.fr

Contents

1 Parameter Estimation **1**

1.1 The Multivariate Gaussian Distribution 1

1.1.1 Gaussian Random Variables and Linear Models 8

1.2 The Parameter Estimation Problem 9

1.3 Bayes Estimation 12

1.3.1 The MMSE criterion 13

1.3.2 The absolute value cost function 14

1.3.3 The uniform cost function 15

1.3.4 Equivalences of Bayes estimators 20

1.3.5 Vector Parameters 22

1.3.6 The MMSE criterion: Vector Parameters 23

1.3.7 The Uniform Cost Function and MAP Estimation: Vector Parameters . 25

1.3.8 The Cramer-Rao Lower Bound for Stochastic Parameters 26

1.3.9 Linear MMSE Estimation 34

1.3.10 The Bayesian Linear Model 40

1.3.11 Recap: Bayesian parameter estimation 42

1.4 Deterministic Parameters 44

1.4.1 Problem setting 44

1.4.2 Some optimality properties 46

1.4.3 Finding UMVUE using complete sufficient statistics♠ 47

1.4.4 Maximum Likelihood estimation 47

1.4.5 Cramer-Rao Bound 52

1.4.6 A Non-Linear Signal Model♠ 55

1.4.7 The Deterministic Linear Model 57

1.4.8 Asymptotic (Large Sample) Properties 58

1.4.9 Recap: Properties of Estimators and Estimation Techniques 61

1.4.10 The Method of Moments 62

1.4.11 The Best Linear Unbiased Estimator (BLUE) 65

1.4.12 Least-Squares Estimation 68

1.5 Choice of an Estimator 73

1.6 Further Reading 74

1.7 Problems 75

2 Optimal Filtering **1**
2.1 Noncausal Wiener Filtering 1
2.1.1 Signal in Noise 4
2.1.2 Channel Equalization 7
2.2 Causal Wiener Filtering♠ 12
2.3 Order-Recursive FIR Wiener Filtering♠ 13
2.4 Fixed-Order FIR Wiener Filtering 16
2.5 Problems 19

3 Adaptive Filtering **1**
3.1 The Steepest-Descent Algorithm applied to FIR Wiener Filtering 2
3.2 The Least Mean Square (LMS) Algorithm 8
3.2.1 A Model for the Purpose of Analysis 9
3.2.2 Averaging Theorem 12
3.2.3 The LMS Learning Curve with Constant Stepsize 13
3.2.4 Conditions on the Stepsize for Exact Convergence 15
3.2.5 The Normalized LMS Algorithm 16
3.3 The Recursive Least-Squares (RLS) Algorithm 18
3.3.1 Derivation of the RLS algorithm 19
3.3.2 Performance analysis 19
3.3.3 A Bayesian Context - A Priori Information 20
3.3.4 Exponential Weighting 20
3.4 Adaptive Filtering Applications 21
3.5 Problems 21

Chapter 1

Parameter Estimation

In this chapter, we cover some elements from estimation theory. The techniques we shall develop here will prove an essential building block in all chapters that follow. The statistical description will be an essential ingredient in our characterization of signals. In order to illustrate how one random quantity reveals some information about another random quantity that is correlated with the first one, we shall first investigate in some detail the nature of the joint distibution of Gaussian random variables. This will shed some light on the nature of estimation. It will furthermore allow us to illustrate the concept of conditional distribution, which is fundamental to estimation theory.

1.1 The Multivariate Gaussian Distribution

We shall begin by introducing the multivariate Gaussian distribution in a constructive fashion, using the Gaussian distribution of a single variable and two postulates. Let $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_m]^T$ be a real random vector. The superscript ^{T} denotes transposition. In this course, we shall concentrate on real quantities. Extensions to complex variables are possible and are convenient for the treatment of modulated signals (see TCOM course). For clarity, we distinguish here explicitly between the random variables $\mathbf{x}_i$ (functions on a probablity space) which are denoted in boldface, and the values x_i that these random variables may take on. Later on we shall omit this distinction. For the construction of the multivariate gaussian distribution, the only moments that are required are the moments of first and second order. So we consider available the mean $m_X = E\mathbf{X} = [m_{x_1} \cdots m_{x_m}]^T$ and the covariance matrix

$$\begin{aligned} C_{XX} &= E\left(\mathbf{X}-m_X\right)\left(\mathbf{X}-m_X\right)^T = \left[E\left(\mathbf{x}_i-m_{x_i}\right)\left(\mathbf{x}_j-m_{x_j}\right)\right]_{i,j=1}^{m} \\ &= E\left(\mathbf{X}\mathbf{X}^T-\mathbf{X}m_X^T-m_X\mathbf{X}^T+m_Xm_X^T\right) \\ &= \left(E\mathbf{X}\mathbf{X}^T\right)-\left(E\mathbf{X}\right)m_X^T-m_X\left(E\mathbf{X}\right)^T+m_Xm_X^T \\ &= R_{XX}-m_Xm_X^T-m_Xm_X^T+m_Xm_X^T = R_{XX}-m_Xm_X^T \end{aligned} \tag{1.1}$$

where R_{XX} is the correlation matrix. In this derivation we exploited the linearity of the expectation operator E, which allows us to interchange it with other linear operations. By definition of expectation

$$C_{XX} = E\left(\mathbf{X}-m_X\right)\left(\mathbf{X}-m_X\right)^T = \int dx_1 \cdots \int dx_m f_{\mathbf{X}}(X)\left(X-m_X\right)\left(X-m_X\right)^T . \tag{1.2}$$

This shows that $C_{XX} = C_{XX}^T$ is symmetric. The eigenvalues λ_i and corresponding eigenvectors V_i of C_{XX} satisfy $C_{XX}V_i = \lambda_i V_i$. Symmetry of the matrix implies that its eigenvalues are real and that there exists a complete set of orthonormal eigenvectors: $V_i^T V_j = \delta_{ij}$ where δ_{ij} is the Kronecker delta (equal to 1 if $i = j$ and zero otherwise). The matrix $V = [V_1 \cdots V_m]$ the columns of which are the eigenvectors of C_{XX} is orthogonal. Indeed, $\left[V^T V\right]_{ij} = V_i^T V_j = \delta_{ij}$, hence $V^T V = I_m$. Since both the matrix C_{XX} itself and its eigenvalues are real, this also implies that the eigenvectors are real. Now $I_m = V^T V$ implies that $1 = \det\left(V^T V\right) = (\det V)^2 \Rightarrow \det V = \pm 1$. We can choose the signs of the V_i such that $\det V = 1$.

From (1.2) it follows that C_{XX} is an infinite sum (integral) of positive semidefinite (of rank one) matrices $f_{\mathbf{X}}(X)\,(X - m_X)\,(X - m_X)^T$. This implies that C_{XX} itself is positive semidefinite, which we denote as $C_{XX} \geq 0$. This means that $\forall U \in \mathcal{R}^m : U^T C_{XX} U = E\left(U^T(\mathbf{X} - m_X)\right)^2 \geq 0$ (positive definite would be $\forall U \in \mathcal{R}^m \setminus \{0\} : U^T C_{XX} U > 0$). By choosing $U = V_i$ we find $U^T C_{XX} U = V_i^T C_{XX} V_i = \lambda_i V_i^T V_i = \lambda_i \geq 0$. We can order the eigenvalues in non-increasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$. If $\lambda_m = 0$, C_{XX} is singular. In this case $V_m^T C_{XX} V_m = E\left(V_m^T(\mathbf{X} - m_X)\right)^2 = 0$ or hence the auto-correlation of the random variable $V_m^T(\mathbf{X} - m_X)$ is zero. This implies that its mean and variance are zero, which in turn implies that the random variable $V_m^T(\mathbf{X} - m_X)$ is zero with probability one (w.p. 1). This means that at least one variable $\mathbf{x}_i$ is a linear combination of the other variables and 1. We shall in general exclude this possibility and hence exclude the case of singular covariance matrices. In other words, we shall assume in general that covariance matrices are positive definite: $C_{XX} > 0$, $\lambda_i > 0$, $i = 1, \ldots, m$. Note that the vectors $C_{XX}V_i - V_i\lambda_i = 0$ are the columns of the matrix $C_{XX}V - V\Lambda = 0$ where $\Lambda = \mathrm{diag}\,\{\lambda_1, \ldots, \lambda_m\}$. Using $V^{-1} = V^T$, we find

$$C_{XX} \;=\; V\Lambda V^T \;=\; [V_1 \cdots V_M]\,\mathrm{diag}\,\{\lambda_1, \ldots, \lambda_m\}\,[V_1 \cdots V_M]^T \;=\; \sum_{i=1}^{m} \lambda_i V_i V_i^T \,. \tag{1.3}$$

Consider now a linear transformation of variables: $\mathbf{Z} = V^T(\mathbf{X} - m_X)$. Then we find for the first two moments

$$\begin{aligned} m_Z &= E\,V^T(\mathbf{X} - m_X) = V^T(E\mathbf{X} - m_X) = V^T(m_X - m_X) = 0 \\ C_{ZZ} &= E\,(\mathbf{Z} - m_Z)\,(\mathbf{Z} - m_Z)^T \;=\; E\mathbf{Z}\mathbf{Z}^T \;=\; EV^T\,(\mathbf{X} - m_X)\,(\mathbf{X} - m_X)^T\,V \\ &= V^T\left(E\,(\mathbf{X} - m_X)\,(\mathbf{X} - m_X)^T\right)V \;=\; V^T C_{XX} V \;=\; V^T V\,\Lambda\,V^T V = \Lambda \end{aligned} \tag{1.4}$$

Or hence $E\mathbf{z}_i\mathbf{z}_j = \lambda_i\delta_{ij}$: the variables $\mathbf{z}_i$ are zero mean and uncorrelated. At this point we have only specified the first two moments of the variables $\mathbf{z}_i$. We will now specify the rest of their distribution by stating that the $\mathbf{z}_i$ are jointly Gaussian. We furthermore postulate that zero mean uncorrelated Gaussian random variables are independent (note that this postulate is backwards compatible since in general independent zero mean random variables are uncorrelated; on the other hand, uncorrelated zero mean random variables are not independent in general). Hence the joint distribution of the independent Gaussian random variables $\mathbf{z}_i$ is

$$f_{\mathbf{Z}}(Z) = \prod_{i=1}^{m} f_{\mathbf{z}_i}(z_i) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\,\lambda_i}} \exp\left(-\frac{z_i^2}{2\lambda_i}\right) = (2\pi)^{-\frac{m}{2}}\,(\det\Lambda)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}Z^T\Lambda^{-1}Z\right) \tag{1.5}$$

At this point we furthermore postulate that a linear transformation of jointly Gaussian random variables produces again jointly Gaussian random variables. Since the Jacobian ($\det V^T$) of the linear transformation between $\mathbf{X}$ and $\mathbf{Z}$ equals one, we get for the joint distribution of the variables $\mathbf{x}_i$

$$\begin{aligned} f_{\mathbf{X}}(X) &= f_{\mathbf{Z}}\left(V^T(X-m_X)\right) \\ &= (2\pi)^{-\frac{m}{2}}\left(\det(V^T C_{XX} V)\right)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}\left[V^T(X-m_X)\right]^T \Lambda^{-1}\left[V^T(X-m_X)\right]\right) \\ &= (2\pi)^{-\frac{m}{2}}\left((\det V^T)(\det C_{XX})(\det V)\right)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}\left[X-m_X\right]^T V\Lambda^{-1}V^T\left[X-m_X\right]\right) \\ &= (2\pi)^{-\frac{m}{2}}(\det C_{XX})^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}\left[X-m_X\right]^T C_{XX}^{-1}\left[X-m_X\right]\right) \end{aligned} \tag{1.6}$$

This is the general expression for a multivariate Gaussian probability density function (pdf). The fact that the variables $\mathbf{x}_i$ are jointly Gaussianly distributed will be denoted as $\mathbf{X} \sim \mathcal{N}(m_X, C_{XX})$, which emphasizes the fact that the Gaussian distribution is completely specified in terms of the first and second-order moments. Because of the two postulates we introduced, it should not be surprising that one can show that zero mean uncorrelated Gaussian random variables are independent and that a linear transformation (and in particular linear filtering) preserves the Gaussian distribution.

Now let $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_m]^T$ and $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_n]^T$ be $m+n$ random variables that are jointly Gaussianly distributed. So we have $m+n$ variables which are arbitrarily split into two groups, $\mathbf{X}$ and $\mathbf{Y}$. The joint Gaussian (or normal) distribution of these $m+n$ variables can be stated as

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_X \\ m_Y \end{bmatrix}, \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix}\right) \tag{1.7}$$

where the mean is defined as

$$\begin{bmatrix} m_X \\ m_Y \end{bmatrix} = E\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \tag{1.8}$$

and the covariance matrix is defined as

$$C = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix} = E\begin{bmatrix} \mathbf{X}-m_X \\ \mathbf{Y}-m_Y \end{bmatrix}\begin{bmatrix} \mathbf{X}-m_X \\ \mathbf{Y}-m_Y \end{bmatrix}^T . \tag{1.9}$$

The joint Gaussian pdf can be written as

$$f_{\mathbf{X},\mathbf{Y}}(X,Y) = (2\pi)^{-\frac{m+n}{2}}(\det C)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\begin{bmatrix} X-m_X \\ Y-m_Y \end{bmatrix}^T \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix}^{-1} \begin{bmatrix} X-m_X \\ Y-m_Y \end{bmatrix}\right). \tag{1.10}$$

We recall the meaning of the joint pdf, which is

$$\Pr\left[\mathbf{X} \in (X, X+dX), \mathbf{Y} \in (Y, Y+dY)\right] = f_{\mathbf{X},\mathbf{Y}}(X,Y)dXdY \tag{1.11}$$

where with some abuse of notation dX once means $dX = [dx_1 \cdots dx_m]^T$ and the second time it stands for $dX = dx_1\, dx_2 \cdots dx_m$ (and similarly for dY). We are interested in finding the conditional pdf $f_{\mathbf{X}|\mathbf{Y}}(X|Y)$ of which the meaning is

$$\Pr\left[\mathbf{X} \in (X, X+dX) | \mathbf{Y} = Y\right] = f_{\mathbf{X}|\mathbf{Y}}(X|Y)dX \ . \tag{1.12}$$

The conditional distribution is defined by Bayes' rule:

$$f_{\mathbf{X}|\mathbf{Y}}(X|Y) \;=\; \frac{f_{\mathbf{X},\mathbf{Y}}(X,Y)}{f_{\mathbf{Y}}(Y)} \tag{1.13}$$

where the marginal distribution $f_{\mathbf{Y}}(Y)$ is obtained from the joint distribution by integrating out X:

$$\begin{aligned} f_{\mathbf{Y}}(Y) &= \int f_{\mathbf{X},\mathbf{Y}}(X,Y)dX \\ &= \int \cdots \int f_{\mathbf{x}_1,\ldots,\mathbf{x}_m,\mathbf{y}_1,\ldots,\mathbf{y}_n}(x_1,\ldots,x_m,y_1,\ldots,y_n)dx_1\cdots dx_m \\ &= (2\pi)^{-\frac{n}{2}}\,(\det C_{YY})^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\,[Y-m_Y]^T\,C_{YY}^{-1}\,[Y-m_Y]\right). \end{aligned} \tag{1.14}$$

This last expression can be written down immediately since we know the mean and the covariance matrix of the Gaussian variables $\mathbf{Y}$. In order to carry out the division in (1.13), we shall try to factor out $f_{\mathbf{Y}}(Y)$ from $f_{\mathbf{X},\mathbf{Y}}(X,Y)$. For this we need to consider the block Upper Diagonal Lower (UDL) triangular factorization of the covariance matrix C:

$$C \;=\; \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix} \;=\; \begin{bmatrix} I & K \\ 0 & I \end{bmatrix}\begin{bmatrix} P & 0 \\ 0 & C_{YY} \end{bmatrix}\begin{bmatrix} I & 0 \\ K^T & I \end{bmatrix} \tag{1.15}$$

with

$$K \;=\; C_{XY}C_{YY}^{-1}\;,\quad P \;=\; C_{XX} - C_{XY}C_{YY}^{-1}C_{YX}\;. \tag{1.16}$$

To verify this factorization, it suffices to check that the product of the three factors on the RHS (right hand side) of (1.15) gives C. From the UDL factorization of C, we can obtain the LDU factorization of C^{-1} by taking the inverse of both sides of (1.15), viz.

$$\begin{aligned} C^{-1} &= \begin{bmatrix} I & 0 \\ K^T & I \end{bmatrix}^{-1}\begin{bmatrix} P & 0 \\ 0 & C_{YY} \end{bmatrix}^{-1}\begin{bmatrix} I & K \\ 0 & I \end{bmatrix}^{-1} \\ &= \begin{bmatrix} I & 0 \\ -K^T & I \end{bmatrix}\begin{bmatrix} P^{-1} & 0 \\ 0 & C_{YY}^{-1} \end{bmatrix}\begin{bmatrix} I & -K \\ 0 & I \end{bmatrix} \end{aligned} \tag{1.17}$$

where we used the property that $(AB)^{-1} = B^{-1}A^{-1}$ and the expression given for the inverse of a 2×2 block triangular matrix can be easily verified. By taking determinants of both sides of (1.15), we also obtain

$$\begin{aligned} \det C &= \det\begin{bmatrix} I & K \\ 0 & I \end{bmatrix}\,\det\begin{bmatrix} P & 0 \\ 0 & C_{YY} \end{bmatrix}\,\det\begin{bmatrix} I & 0 \\ K^T & I \end{bmatrix} \\ &= 1\;.\;\det\begin{bmatrix} P & 0 \\ 0 & C_{YY} \end{bmatrix}\;.\;1 \;=\; \det P\;\;\det C_{YY} \end{aligned} \tag{1.18}$$

where we used the properties $\det(AB) = \det A\;\det B$, the determinant of a triangular matrix is the product of the diagonal elements and more generally, the determinant of a block diagonal

(or triangular) matrix is the product of the determinants of the diagonal blocks. Using (1.17), we can rewrite the exponent of the joint distribution as

$$\begin{aligned}
&\begin{bmatrix} X-m_X \\ Y-m_Y \end{bmatrix}^T \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix}^{-1} \begin{bmatrix} X-m_X \\ Y-m_Y \end{bmatrix} \\
&= \begin{bmatrix} X-m_X \\ Y-m_Y \end{bmatrix}^T \begin{bmatrix} I & 0 \\ -K^T & I \end{bmatrix} \begin{bmatrix} P^{-1} & 0 \\ 0 & C_{YY}^{-1} \end{bmatrix} \begin{bmatrix} I & -K \\ 0 & I \end{bmatrix} \begin{bmatrix} X-m_X \\ Y-m_Y \end{bmatrix} \\
&= \begin{bmatrix} X-KY-(m_X-Km_Y) \\ Y-m_Y \end{bmatrix}^T \begin{bmatrix} P^{-1} & 0 \\ 0 & C_{YY}^{-1} \end{bmatrix} \begin{bmatrix} X-KY-(m_X-Km_Y) \\ Y-m_Y \end{bmatrix} \qquad (1.19) \\
&= [X-KY-(m_X-Km_Y)]^T P^{-1} [X-KY-(m_X-Km_Y)] + [Y-m_Y]^T C_{YY}^{-1} [Y-m_Y] .
\end{aligned}$$

Using (1.10),(1.14),(1.18) and (1.19), we can carry out the division in (1.13) and prove the following theorem:

Theorem 1.1 (Gauss-Markov) *If* $\mathbf{X}$ *and* $\mathbf{Y}$ *have the joint Gaussian distribution indicated in (1.7),(1.10), then the conditional distribution is*

$$\begin{aligned}
&f_{\mathbf{X}|\mathbf{Y}}(X|Y) = \\
&(2\pi)^{-\frac{m}{2}} (\det P)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} [X-KY-(m_X-Km_Y)]^T P^{-1} [X-KY-(m_X-Km_Y)]\right) .
\end{aligned} \qquad (1.20)$$

The conditional distribution is again Gaussian with conditional mean

$$E_{\mathbf{X}|\mathbf{Y}}\mathbf{X} = E[\mathbf{X}|\mathbf{Y}=Y] = m_X + C_{XY}C_{YY}^{-1}(Y-m_Y) \qquad (1.21)$$

and conditional covariance matrix

$$E_{\mathbf{X}|\mathbf{Y}}\left[\mathbf{X}-E_{\mathbf{X}|\mathbf{Y}}\mathbf{X}\right]\left[\mathbf{X}-E_{\mathbf{X}|\mathbf{Y}}\mathbf{X}\right]^T = P = C_{XX} - C_{XY}C_{YY}^{-1}C_{YX} . \qquad (1.22)$$

Example 1.1 Two correlated Gaussian random variables

Consider the scalar case $m = n = 1$ and zero means $m_X = m_Y = 0$. We can rename $C_{XX} = \sigma_1^2$ and $C_{YY} = \sigma_2^2$. Using the Cauchy-Schwarz inequality ($|\langle \mathbf{x}, \mathbf{y} \rangle|^2 \leq \langle \mathbf{x}, \mathbf{x} \rangle \langle \mathbf{y}, \mathbf{y} \rangle$) on the inner product $\langle \mathbf{x}, \mathbf{y} \rangle = E\mathbf{x}\mathbf{y}$, one can show that $C_{XY}^2 \leq C_{XX}C_{YY}$. This motivates us to introduce the following normalized correlation coefficient

$$\rho = \frac{C_{XY}}{\sqrt{C_{XX}C_{YY}}} \in [-1, 1] . \qquad (1.23)$$

We can rewrite C_{XY} as $C_{XY} = \rho\sigma_1\sigma_2$. The joint Gaussian distribution of $\mathbf{x}$ and $\mathbf{y}$ is given by (1.10) and can now be rewritten as

$$f_{\mathbf{x},\mathbf{y}}(x, y) \leftrightarrow \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right) . \qquad (1.24)$$

The conditional pdf of $\mathbf{x}$ given $\mathbf{y}$ is given in (1.20) and can be rewritten as

$$f_{\mathbf{x}|\mathbf{y}}(x|y) \leftrightarrow \mathcal{N}\left(\rho\frac{\sigma_1}{\sigma_2}y, \sigma_1^2(1-\rho^2)\right) \qquad (1.25)$$

and is illustrated in Fig. 1.1 for various values of y.

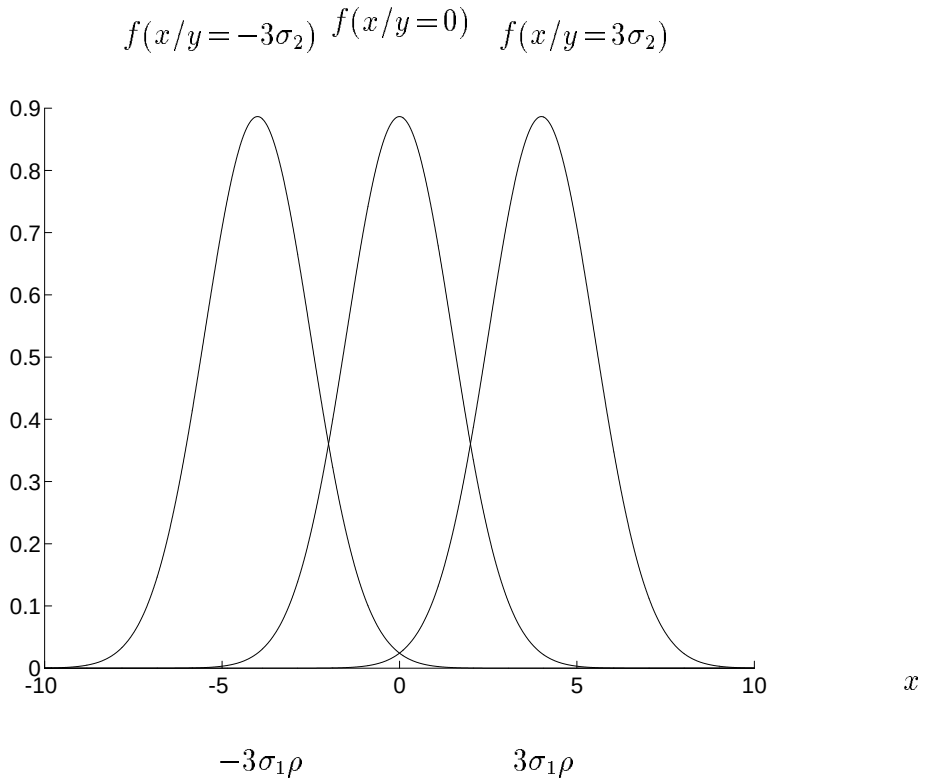


Figure 1.1: The conditional pdf $f_{\mathbf{x}|\mathbf{y}}(x|y)$ for $y = 0, \pm 3\sigma_2$.

If $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated ($\rho = 0$), then $f_{\mathbf{x}|\mathbf{y}}(x|y) = f_{\mathbf{x}}(x)$ which shows that $\mathbf{x}$ and $\mathbf{y}$ are independent in that case. As $|\rho| \to 1$, the conditional mean of $\mathbf{x}$ given $\mathbf{y} = y$ differs more and more from zero (assuming $y \neq 0$) and the conditional variance goes to zero. This means that when $\mathbf{y} = y$ is known, it gives us some information about $\mathbf{x}$ and the residual randomness in $\mathbf{x}$ decreases as $|\rho| \to 1$. We can consider the extreme cases:

$$\begin{array}{rcll} \rho = 1 & : & \mathbf{x} = \frac{\sigma_1}{\sigma_2}\mathbf{y} & \mathbf{x} \text{ and } \mathbf{y} \text{ are perfectly correlated,} \\ \rho = -1 & : & \mathbf{x} = -\frac{\sigma_1}{\sigma_2}\mathbf{y} & \mathbf{x} \text{ and } \mathbf{y} \text{ are perfectly anticorrelated.} \end{array} \tag{1.26}$$

We shall pursue the following geometric interpretation of correlation. We define a *concentration ellipse* as the set of points in the (x, y) plane for which the probability density function takes on a given value:

$$\begin{aligned} f_{\mathbf{x},\mathbf{y}}(x,y) \;=\; c' \;&\Longleftrightarrow\; \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x \\ y \end{bmatrix} \;=\; c \\ &\Longleftrightarrow\; \frac{\left(x - \rho\frac{\sigma_1}{\sigma_2}y\right)^2}{\sigma_1^2(1-\rho^2)} + \frac{y^2}{\sigma_2^2} = c \end{aligned} \tag{1.27}$$

This ellipse with $c = 1$ is illustrated in Fig. 1.2.

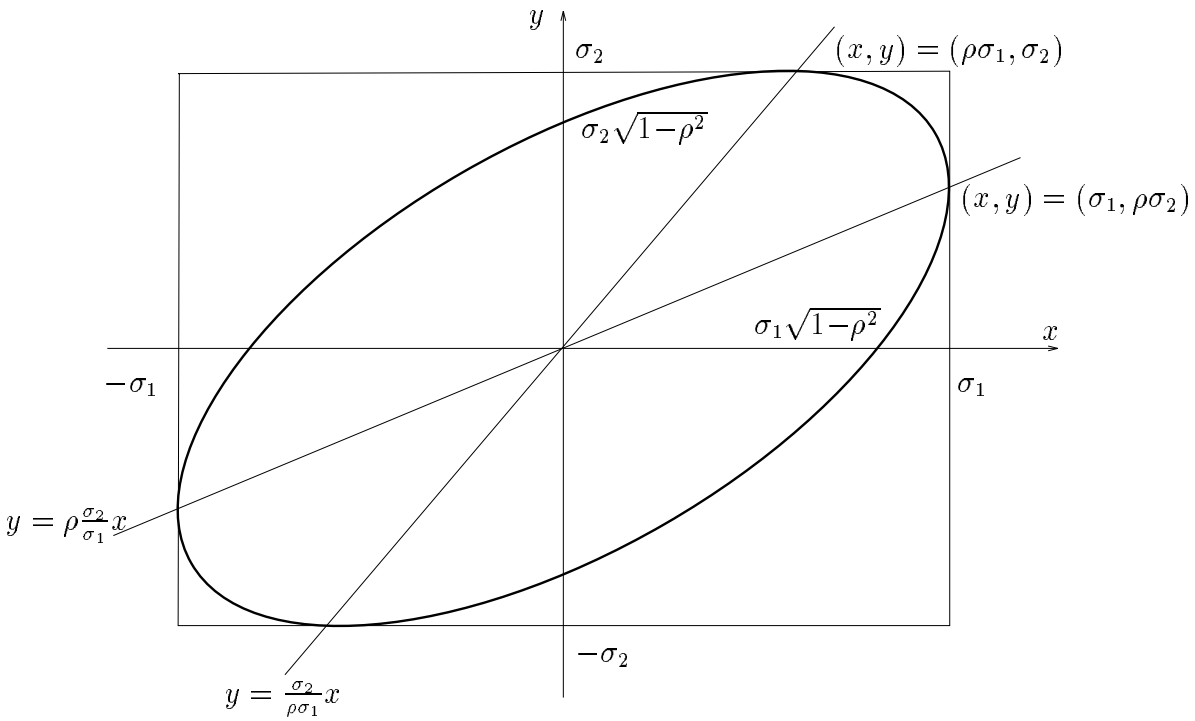


Figure 1.2: The concentration ellipse for two correlated Gaussian random variables.

In particular for $c = 1$, the points inside the ellipse, which satisfy

$$\frac{\left(x - \rho\frac{\sigma_1}{\sigma_2}y\right)^2}{\sigma_1^2(1-\rho^2)} + \frac{y^2}{\sigma_2^2} \leq 1 \tag{1.28}$$

contain a significant portion of the probability mass. The area (volume in higher dimensions) contained in the ellipse can be shown to be

$$V \;=\; \pi\,(\det C)^{\frac{1}{2}} \;=\; \pi\sigma_1\sigma_2\sqrt{1-\rho^2} \;\in\; [0, \pi\sigma_1\sigma_2]\,. \tag{1.29}$$

This means that for strongly correlated variables, the pair $(\mathbf{x}, \mathbf{y})$ takes with high probability values in a fairly small area. Hence, in some sense the randomness (or entropy) of the pair $(\mathbf{x}, \mathbf{y})$ decreases as $|\rho| \to 1$. We shall indeed see in a later chapter that for Gaussian random variables, the entropy can be measured by $\det C$ (related to the volume of the concentration ellipse). In Fig. 1.3 the concentration ellipse is illustrated for various values of ρ. When $\rho = \pm 1$, the concentration ellipse reduces to a line segment. $\diamond$

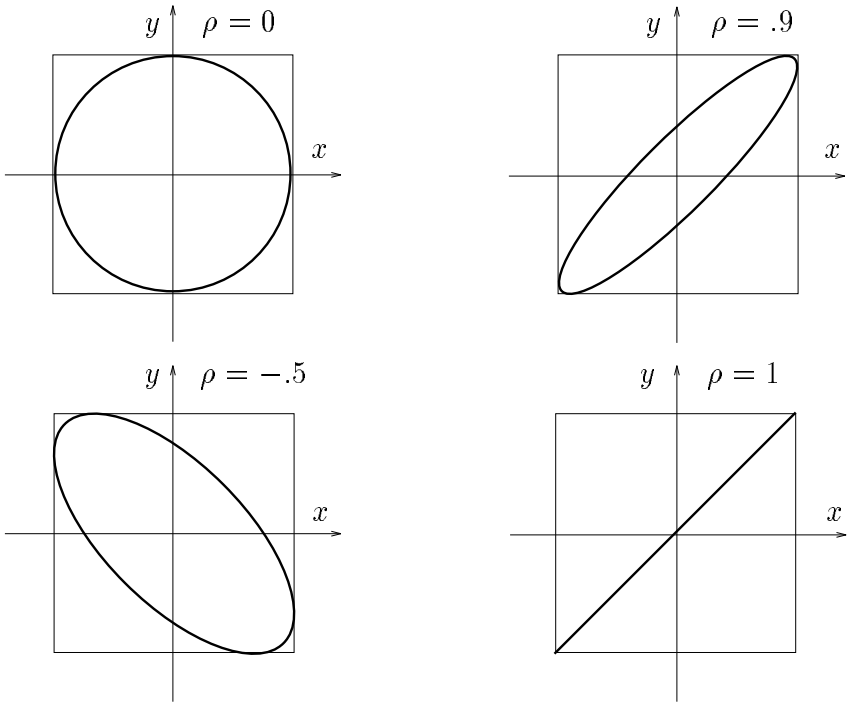


Figure 1.3: The concentration ellipse as a function of correlation: $\rho = 0, 0.9, -0.5, 1$.

1.1.1 Gaussian Random Variables and Linear Models

The conditional pdf for $\mathbf{Y}$ given $\mathbf{X}$ can be found from (1.20) by interchanging X and Y, viz.

$$f_{\mathbf{Y}|\mathbf{X}}(Y|X) \leftrightarrow \mathcal{N}\left(m_Y + C_{YX}C_{XX}^{-1}(X - m_X),\, C_{YY} - C_{YX}C_{XX}^{-1}C_{XY}\right) \tag{1.30}$$

and the corresponding LDU triangular factorization of C is

$$C = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix} = \begin{bmatrix} I & 0 \\ C_{YX}C_{XX}^{-1} & I \end{bmatrix} \begin{bmatrix} C_{XX} & 0 \\ 0 & C_{YY} - C_{YX}C_{XX}^{-1}C_{XY} \end{bmatrix} \begin{bmatrix} I & C_{XX}^{-1}C_{XY} \\ 0 & I \end{bmatrix} \tag{1.31}$$

If we now introduce the Gaussian random vector $\mathbf{V}$ of dimension n also, with distribution

$$f_{\mathbf{V}}(V) \leftrightarrow \mathcal{N}\left(m_Y - C_{YX}C_{XX}^{-1}m_X,\, C_{YY} - C_{YX}C_{XX}^{-1}C_{XY}\right) \text{ and } \mathbf{V} \text{ independent of } \mathbf{X}, \tag{1.32}$$

then $\mathbf{X}$ and $\mathbf{Y}$ generated as follows

$$\begin{bmatrix} \mathbf{X} - m_X \\ \mathbf{Y} - m_Y \end{bmatrix} = \begin{bmatrix} I & 0 \\ C_{YX}C_{XX}^{-1} & I \end{bmatrix} \begin{bmatrix} \mathbf{X} - m_X \\ \mathbf{V} - m_V \end{bmatrix} \tag{1.33}$$

are jointly Gaussian and have the correct mean and variance, hence have the correct pdf $f_{\mathbf{X},\mathbf{Y}}(X, Y)$. The second relation in (1.33) can be explicitly rewritten as

$$\mathbf{Y} = C_{YX}C_{XX}^{-1}\mathbf{X} + \mathbf{V}\,. \tag{1.34}$$

This means that for two sets of correlated Gaussian random variables, one ($\mathbf{Y}$) can be thought of as being generated from the other ($\mathbf{X}$) through a linear model ($C_{YX}C_{XX}^{-1}$) and being corrupted by independent Gaussian "measurement" noise ($\mathbf{V}$). Going back to the scalar example ($m = n = 1$), the variance of the linear model part ($C_{YX}C_{XX}^{-1}\mathbf{X}$) is $\rho^2\sigma_2^2$ while the variance of the measurement noise ($\mathbf{V}$) is $(1-\rho^2)\sigma_2^2$. The two are complementary parts of the total variance σ_2^2 of $\mathbf{Y}$. We could define a signal to noise ratio (SNR) as the ratio of the two parts which is $\frac{\rho^2}{1-\rho^2}$. The SNR increases from 0 to ∞ as $|\rho|$ increases from 0 to 1, as the correlation between $\mathbf{X}$ and $\mathbf{Y}$ increases.

1.2 The Parameter Estimation Problem

In many problems in telecommunications, we encounter signals or stochastic processes of which the description is known up to the values of a number of parameters. We will need to estimate the value of these parameters based on measurements of the signals. Some examples of such estimation problems are:

- tone detector: a push button telephone emits sinusoids of which the frequencies are distinct for different buttons. The unknown parameter of interest is the sinusoid frequency. This frequency determination problem may in fact be better approached as a detection problem as we shall see shortly
- the carrier phase and timing instants in linear digital modulation schemes
- the impulse response of the transmission channel. We may take a parameterized model (e.g. FIR model) for this impulse response and then the channel identification problem becomes a problem of estimating the parameters in the channel model
- similar impulse response identification problems occur in the problems of the cancellation of electrical echos caused by hybrids connecting 2-wire and 4-wire sections of telephone line, and acoustic echos in handsfree telephony (e.g. teleconferencing) systems
- the average rate in a Poisson process occurring in queuing theory and network performance analysis
- the pixel value in the noisy capture of an image by a satellite
- the position and orientation of an object in an image

In a first instance, we shall consider the *Bayesian* framework. In the Bayesian approach, we consider the parameters to be random variables. The formal chain of parameter generation, measurements and estimation is depicted in Fig. 1.4. The parameters have some *a priori* distribution $f_{\boldsymbol{\theta}}(\theta)$. This distribution is called a priori because it summarizes the knowledge we have about θ before making any measurement. Next we'll carry out one experiment in which we shall make some measurements, collected in Y. This experiment correponds to one particular realization θ of $\boldsymbol{\theta}$ and one particular realization Y of $\mathbf{Y}$. Using the measurements Y, it is this unique value θ of the unknown parameters that we shall estimate. The measurement is not a deterministic function of the unknown parameters. The random aspect of this relation

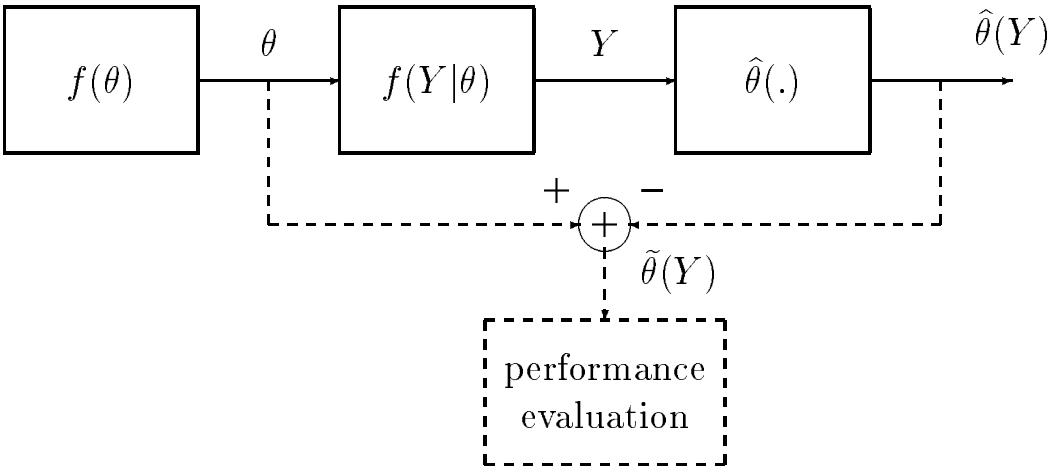


Figure 1.4: The relation between parameters, measurements and estimates.

is captured by the conditional distribution $f_{\mathbf{Y}|\boldsymbol{\theta}}(Y|\theta)$. For the vector of measurements, there are two points of view. Often, we'll collect a set of scalar measurements y_i that are i.i.d. (independent identically distributed) given θ. This implies that each y_i is distributed according to $f_{\mathbf{y}_1|\boldsymbol{\theta}}(y_i|\theta)$. These y_i can be collected in a vector $Y = [y_1 \cdots y_n]^T$. Or we can measure a vector Y of quantities. Whether we obtain the measurements y_i sequentially or simultaneously, we can always collect them into a vector Y of measurements corresponding to the same parameter value θ.

Example 1.2 Additive independent measurement noise

An example of the stochastic relation between θ and Y is illustrated in Fig. 1.5. We assume here θ, v and y to be scalars. The measurement noise v is assumed to be independent of the parameter θ. It is this measurement noise that introduces a random aspect in the relation between θ and y.

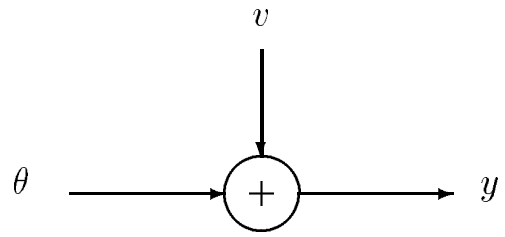


Figure 1.5: Parameter measurement with additive independent measurement noise.

The conditional distribution $f_{\mathbf{y}|\boldsymbol{\theta}}(y|\theta)$ for this example can be obtained as follows:

$$f_{\mathbf{y}|\boldsymbol{\theta}}(y|\theta) = \frac{f_{\mathbf{y},\boldsymbol{\theta}}(y,\theta)}{f_{\boldsymbol{\theta}}(\theta)} = \frac{f_{\mathbf{v},\boldsymbol{\theta}}(y-\theta,\theta)}{f_{\boldsymbol{\theta}}(\theta)} = \frac{f_{\mathbf{v}}(y-\theta)\, f_{\boldsymbol{\theta}}(\theta)}{f_{\boldsymbol{\theta}}(\theta)} = f_{\mathbf{v}}(y-\theta) \qquad (1.35)$$

where we used Bayes' rule in the first identity, the independence of v and θ in the third identity, and to obtain the second identity, we considered the following transformation of

variables

$$\begin{bmatrix} v \\ \theta \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y \\ \theta \end{bmatrix} . \tag{1.36}$$

Since this transformation is linear, the Jacobian is simply the determinant of the transformation matrix, which equals 1. The calculation of the conditional distribution in this example will prove very useful because this type of situation occurs quite frequently. ◇

The final element in the estimation chain is the estimator itself. We should make a distinction here between the *estimator* and the *estimate*. The estimator is a function $\widehat{\theta}(.)$ to be applied to the measurement $\mathbf{Y}$. The estimate $\widehat{\theta}(Y)$ is the estimator $\widehat{\theta}(\mathbf{Y})$ evaluated at $\mathbf{Y} = Y$. The estimator we consider here is a *point estimator*, i.e. it delivers one value $\widehat{\theta} = \widehat{\theta}(Y) = t(y_1, \ldots, y_n)$ that we hope to be close to θ. The function $\widehat{\theta}(\mathbf{Y}) = t(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ of the measurement data is called a *statistic*. Another type of estimator would be an *interval estimator*: two statistics $t_1(y_1, \ldots, y_n)$ and $t_2(y_1, \ldots, y_n)$ are used to define an interval such that

$$\Pr\{\theta \in (t_1(y_1, \ldots, y_n),\, t_2(y_1, \ldots, y_n))\} \;=\; \gamma \tag{1.37}$$

can be determined. γ is called the *confidence coefficient*.

Example 1.3 Point and Interval estimators for the mean of Gaussian variables

Let the $\mathbf{y}_i \sim \mathcal{N}(\theta, 5)$ be i.i.d. (independent and identically distributed) normal random variables with unknown mean θ and variance equal to 5. The arithmetic mean is a point estimator of θ :

$$\widehat{\theta} \;=\; \overline{y} \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i \, . \tag{1.38}$$

Now, since $\overline{y} \sim \mathcal{N}(\theta, \frac{5}{n})$, we can also state

$$\Pr\left\{\theta \in \left(\overline{y} - 2\sqrt{\frac{5}{n}}, \overline{y} + 2\sqrt{\frac{5}{n}}\right)\right\} \;=\; 0.95 \tag{1.39}$$

so that $\left(\overline{y} - 2\sqrt{\frac{5}{n}}, \overline{y} + 2\sqrt{\frac{5}{n}}\right)$ is an interval estimator for θ with confidence coefficient 0.95. ◇

The design of interval estimators is the result of the following compromise. On the one hand, one wants a small interval so that θ can be known fairly accurately. On the other hand, a small interval necessarily leads to a small confidence coefficient since in general we can only state with a low probability level that θ is contained in a small interval. But we would like the confidence coefficient to be high so that we would be able to say with high probability where θ is located. This again would lead to a large interval etc. We will restrict the further discussion to point estimators. In general, the estimators we shall consider will asymptotically (for many measurements, large n) have a Gaussian distribution so that we can easily construct an interval estimator from the point estimator as in the example above.

As a final remark, we shall assume that the parameters (and their estimators) can take on a continuous range of values so that their distribution is described by a probability density function. In case the parameters can take on only a discrete set of values and the task is to decide which one of the values the parameters actually take, then the estimation problem is called a *detection* or *decision* problem. Such problems will be pursued in the TCOM course.

1.3 Bayes Estimation

We shall first concentrate on a single parameter θ. The Bayes procedure for determining an estimator is to introduce a nonnegative cost function $\mathcal{C}(\theta, \widehat{\theta})$ that depends on the parameter θ and its estimate $\widehat{\theta}(Y)$. Since θ and $\widehat{\theta}(Y)$ are random variables, we can never hope to make this cost function small for every outcome of θ and Y. All we can hope for is to minimize the expected value of the cost function, also called the risk

$$\mathcal{R}(\widehat{\theta}(.)) \;=\; E\,\mathcal{C}(\theta, \widehat{\theta}(Y)) \;=\; E_{\theta,Y}\mathcal{C}(\theta, \widehat{\theta}(Y))\;. \tag{1.40}$$

It is useful at this point to elaborate a bit on the Bayesian experiment depicted in Fig. 1.4. In one particular experiment, one particular value of θ is realized and that is the value we are trying to estimate. So even if $f(Y|\theta) = \prod_{k=1}^{n} f(y_k|\theta)$ in which case the vector measurement Y consists of n independent measurements y_k, those n measurements all correspond to the same value of θ. Hence in one experiment, we estimate one and only one realization of the parameters θ using one correponding value Y of the measurements $\mathbf{Y}$. However, in order to evaluate the performance of an estimator, we shall assume that the experiment may get repeated such that in different experiments different values of θ will be drawn from an a priori distribution $f(\theta)$ and different measurement values Y will be obtained corresponding to the different values of θ according to $f(Y|\theta)$. For each experiment separately, we would construct an estimate of the value θ of $\boldsymbol{\theta}$ that is involved in that experiment, using the measurements Y in that experiment. We shall be interested in the average performance of an estimator over all possible experiments.

So the estimator is that function $\widehat{\theta}(.)$ that minimizes the Bayes risk

$$\widehat{\theta}(.) \;=\; \arg\min_{\widehat{\theta}(.)} \mathcal{R}(\widehat{\theta}(.)) \quad . \tag{1.41}$$

It is common practice to limit the choice of the cost function to a nonnegative cost function of the parameter estimation error

$$\widetilde{\theta} \;=\; \theta - \widehat{\theta}(Y) \tag{1.42}$$

only, in which case the estimation problem becomes

$$\widehat{\theta}(.) \;=\; \arg\min_{\widehat{\theta}(.)} E\,\mathcal{C}(\widetilde{\theta}) \quad . \tag{1.43}$$

Three common choices for $\mathcal{C}(\widetilde{\theta})$ are shown in Fig. 1.6.

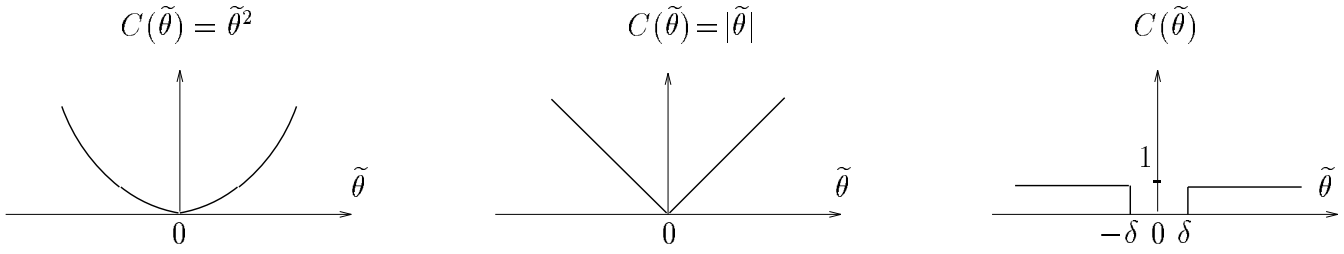


Figure 1.6: Three popular Bayes cost functions: squared parameter deviation, absolute parameter deviation, and the uniform cost function.

All three choices assign no cost when there is no error. The Mean Squared Error (MSE) risk assigns a cost that increases quadratically with $|\tilde{\theta}|$, the cost for the absolute value criterion is simply $|\tilde{\theta}|$ itself, while the uniform cost function is constant for every parameter error, except for a region around zero with arbitrarily small radius δ. Before deriving the estimators corresponding to these three cost functions, we shall manipulate the Bayes optimization criterion a bit. We can rewrite (1.43) as

$$\begin{aligned} \min_{\widehat{\theta}(.)} \mathcal{R}(\widehat{\theta}(.)) \;=\; \min_{\widehat{\theta}(.)} E\,\mathcal{C}(\tilde{\theta}) \;&=\; \min_{\widehat{\theta}(.)} \int\int f(Y,\theta)\,\mathcal{C}(\theta - \widehat{\theta}(Y))\,dY\,d\theta \\ &=\; \min_{\widehat{\theta}(.)} \int f(Y)dY \underbrace{\int f(\theta|Y)\,\mathcal{C}(\theta-\widehat{\theta}(Y))d\theta}_{\mathcal{R}(\widehat{\theta}(.)|Y)} \,. \end{aligned} \tag{1.44}$$

The function $\mathcal{R}(\widehat{\theta}(.)|Y) = \mathcal{R}(\widehat{\theta}(Y)|Y)$ is a function of Y. Since the contributions of the conditional risk $\mathcal{R}(\widehat{\theta}(Y)|Y)$ for every Y are combined via the nonnegative weighting factor $f(Y)$ to obtain the global risk $\mathcal{R}(\widehat{\theta}(.))$, we can minimize $\mathcal{R}(\widehat{\theta}(.))$ by minimizing $\mathcal{R}(\widehat{\theta}(Y)|Y)$ for every particular Y. While the minimization of the global risk $\mathcal{R}(\widehat{\theta}(.))$ is w.r.t. the estimator function $\widehat{\theta}(.)$, the minimization of the conditional risk $\mathcal{R}(\widehat{\theta}(Y)|Y)$ is w.r.t. the particular estimate $\widehat{\theta}(Y)$, which is simply a number (the estimator function $\widehat{\theta}(.)$ evaluated at $\mathbf{Y} = Y$). This last minimization problem requires the *a posteriori* distribution $f(\theta|Y)$ of θ given the measurement Y. The a posteriori distribution describes the randomness left in θ after we have made a measurement of (the related quantity) Y. The a posteriori distribution $f(\theta|Y)$ can be determined from the conditional distribution $f(Y|\theta)$ and the a priori distribution $f(\theta)$, which are normally given in the problem description, as follows (using Bayes' rule):

$$f(\theta|Y) \;=\; \frac{f(Y,\theta)}{f(Y)} \;=\; \frac{f(Y|\theta)f(\theta)}{\int_\Theta f(Y,\theta)d\theta} \;=\; \frac{f(Y|\theta)f(\theta)}{\int_\Theta f(Y|\theta)f(\theta)d\theta} \tag{1.45}$$

where Θ is the region of support for θ ($f(\theta) \geq 0,\ \theta \in \Theta,\ \ f(\theta) \equiv 0,\ \theta \notin \Theta$).

1.3.1 The MMSE criterion

For the Minimum MSE (MMSE) criterion, we have the quadratic cost function $\mathcal{C}_{MMSE}(\tilde{\theta}) = |\tilde{\theta}|^2$. The minimization problem of the conditional Bayes risk becomes

$$\min_{\widehat{\theta}(Y)} \mathcal{R}_{MMSE}(\widehat{\theta}(Y)|Y) \;=\; \min_{\widehat{\theta}(Y)} \int_{-\infty}^{\infty} f(\theta|Y)\,(\theta-\widehat{\theta}(Y))^2 d\theta \,. \tag{1.46}$$

To be able to compute the minimum, we want to take the derivative w.r.t. $\widehat{\theta}$. In order to do this, it is useful to recall Leibnitz's rule:

$$\frac{\partial}{\partial u}\int_{g_1(u)}^{g_2(u)} h(u,v)\,dv \;=\; \int_{g_1(u)}^{g_2(u)} \frac{\partial\, h(u,v)}{\partial u}\,dv + \frac{dg_2(u)}{du}h(u,g_2(u)) - \frac{dg_1(u)}{du}h(u,g_1(u)) \,. \tag{1.47}$$

Using Leibnitz's rule, we obtain

$$\frac{\partial}{\partial\widehat{\theta}}\mathcal{R}_{MMSE}(\widehat{\theta}|Y) \;=\; 2\int_{-\infty}^{\infty} f(\theta|Y)\,(\widehat{\theta}-\theta)d\theta \;\;=0 \tag{1.48}$$

where we put the derivative equal to zero in order to find an extremum. (1.48) can be rewritten as

$$\widehat{\theta}(Y)\;\underbrace{\int_{-\infty}^{\infty} f(\theta|Y)\,d\theta}_{=1} \;=\; \int_{-\infty}^{\infty}\theta\, f(\theta|Y)\,d\theta \tag{1.49}$$

or hence

$$\widehat{\theta}_{MMSE}(Y) \;=\; E(\theta|Y) \tag{1.50}$$

which is the *mean* of the a posteriori distribution of θ given Y. To know whether this extremum is a local minimum, we can apply Leibnitz's rule a second time to obtain

$$\frac{\partial^2}{\partial\widehat{\theta}^2}\mathcal{R}_{MMSE}(\widehat{\theta}|Y) \;=\; 2\int_{-\infty}^{\infty} f(\theta|Y)\,d\theta \;=\; 2 \;>0\,. \tag{1.51}$$

Hence the extremum at $\widehat{\theta}(Y) \;=\; E(\theta|Y)$ is a local minimum and it is furthermore the global minimum since it is the unique local minimum. Note that $E(\theta|Y)$ is indeed a function of Y.

1.3.2 The absolute value cost function

For the absolute value cost function $\mathcal{C}_{ABS}(\widetilde{\theta}) = |\widetilde{\theta}|$, the minimization problem of the conditional Bayes risk becomes

$$\begin{aligned}\min_{\widehat{\theta}(Y)}\mathcal{R}_{ABS}(\widehat{\theta}(Y)|Y) \;&=\; \min_{\widehat{\theta}(Y)}\int_{-\infty}^{\infty} f(\theta|Y)\,|\theta-\widehat{\theta}(Y)|d\theta \\ &=\; \min_{\widehat{\theta}(Y)}\left[\int_{-\infty}^{\widehat{\theta}} f(\theta|Y)\,(\widehat{\theta}-\theta)d\theta \;+\; \int_{\widehat{\theta}}^{\infty} f(\theta|Y)\,(\theta-\widehat{\theta})d\theta\right]\,.\end{aligned} \tag{1.52}$$

We shall again use Leibnitz's rule to take the derivative. For the first integral, we let $h(\widehat{\theta},\theta) = f(\theta|Y)\,(\widehat{\theta}-\theta)$ and we get

$$h(\widehat{\theta},g_2(\widehat{\theta})) = h(\widehat{\theta},\widehat{\theta}) = f(\theta|Y)\,(\widehat{\theta}-\widehat{\theta}) = 0 \tag{1.53}$$

and $\frac{d\,g_1(\widehat{\theta})}{d\widehat{\theta}} = 0$ since the lower limit does not depend on $\widehat{\theta}$. The corresponding terms for the second integral are similarly equal to zero. So the only terms in the derivative remaining are those obtained by differentiating the integrands directly:

$$\frac{\partial}{\partial\widehat{\theta}}\mathcal{R}_{ABS}(\widehat{\theta}|Y) \;=\; \int_{-\infty}^{\widehat{\theta}} f(\theta|Y)\,d\theta \;-\; \int_{\widehat{\theta}}^{\infty} f(\theta|Y)\,d\theta \;=\; 0 \tag{1.54}$$

where again we put the derivative equal to zero in order to find an extremum. So the condition for an extremum becomes

$$F(\widehat{\theta}|Y) \;=\; \int_{-\infty}^{\widehat{\theta}} f(\theta|Y)\,d\theta \;=\; \int_{\widehat{\theta}}^{\infty} f(\theta|Y)\,d\theta \;=\; 1-F(\widehat{\theta}|Y) \tag{1.55}$$

where $F(\theta|Y)$ is the a posteriori cumulative distribution function (cdf) of θ given Y. We can rewrite (1.55) also as

$$F(\widehat{\theta}_{ABS}(Y)|Y) \;=\; \frac{1}{2} \tag{1.56}$$

which means that $\widehat{\theta}_{ABS}(Y)$ is the *median* of the a posteriori distribution of θ given Y. Again, to know whether this extremum is a local minimum, we can apply Leibnitz's rule a second time to obtain

$$\frac{\partial^2}{\partial \widehat{\theta}^2} \mathcal{R}_{ABS}(\widehat{\theta}|Y)\bigg|_{\widehat{\theta}=\widehat{\theta}_{ABS}} \;=\; 2\, f(\widehat{\theta}_{ABS}|Y) \;\geq\; 0\;. \tag{1.57}$$

If $f(\widehat{\theta}_{ABS}|Y) = 0$ then, since $f(\theta|Y) \geq 0$, the first nonzero derivative w.r.t. θ of $f(\theta|Y)$ will be of even order and positive (this reasoning can be extended to the case where $f(\theta|Y)$ would not be sufficiently differentiable). Hence the extremum at $\widehat{\theta}_{ABS}(Y)$ is a local minimum and it is furthermore the global minimum since it is the unique local minimum. Note again that $\widehat{\theta}_{ABS}(Y)$ is indeed a function of Y.

1.3.3 The uniform cost function

For the uniform cost function the minimization problem of the conditional Bayes risk becomes

$$\begin{aligned} \min_{\widehat{\theta}(Y)} \mathcal{R}_{UNIF}(\widehat{\theta}(Y)|Y) \;&=\; \min_{\widehat{\theta}(Y)} \left[\left(\int_{-\infty}^{\widehat{\theta}-\delta} + \int_{\widehat{\theta}+\delta}^{\infty} \right) f(\theta|Y)\, d\theta \right] \\ &=\; \min_{\widehat{\theta}(Y)} \Bigg[\underbrace{\int_{-\infty}^{\infty} f(\theta|Y)\, d\theta}_{=\,1} \;-\; \int_{\widehat{\theta}-\delta}^{\widehat{\theta}+\delta} f(\theta|Y)\, d\theta \Bigg]\;. \end{aligned} \tag{1.58}$$

Since we take δ to be arbitrarily small, the optimization problem becomes

$$\max_{\widehat{\theta}(Y)} \int_{\widehat{\theta}-\delta}^{\widehat{\theta}+\delta} f(\theta|Y)\, d\theta \;\approx\; \max_{\widehat{\theta}(Y)} 2\delta\, f(\widehat{\theta}|Y) \;=\; 2\delta \max_{\widehat{\theta}(Y)} f(\widehat{\theta}|Y)\;. \tag{1.59}$$

Hence, for δ arbitrarily small, $\mathcal{R}_{UNIF}(\widehat{\theta}(Y)|Y)$ is minimized by choosing $\widehat{\theta}$ to be the *mode* (location of the maximum) of the a posteriori distribution of θ given Y. For this reason, the estimator corresponding to a uniform cost function is normally called the *Maximum A Posteriori* (MAP) estimator:

$$\widehat{\theta}_{MAP}(Y) \;=\; \arg\max_{\theta} f(\theta|Y) \tag{1.60}$$

which is again a function of Y.

Remarks:

(i) We may note that the same estimator is obtained by choosing the cost function $\mathcal{C}(\tilde{\theta}) = 1 - \delta(\tilde{\theta})$. While this cost function is cleaner in that it involves no limiting operation with δ, it might be considered non-intuitive since it is not nonnegative. However, one should keep in mind that adding or subtracting a constant to a cost function does not influence its minimizing argument.

(ii) Instead of maximizing $f(\theta|Y)$, any strictly increasing function of it may be maximized. Since $f(\theta|Y)$ is often given in factored form and often contains exponential distributions, a convenient choice is to maximize

$$\ln f(\theta|Y) \;=\; \ln f(Y|\theta) + \ln f(\theta) - \ln f(Y)\;. \tag{1.61}$$

(iii) Often $f(\theta|Y)$ satisfies certain regularity conditions so that $\widehat{\theta}_{MAP}$ is a solution of

$$\frac{\partial}{\partial\theta}\ln f(\theta|Y) \;=\; 0 \;=\; \frac{\partial}{\partial\theta}\ln f(Y|\theta) + \frac{\partial}{\partial\theta}\ln f(\theta)\;\;. \tag{1.62}$$

Note that $\frac{\partial \ln f(\theta|Y)}{\partial\theta} = \frac{1}{f(\theta|Y)}\frac{\partial f(\theta|Y)}{\partial\theta}$. The conditions for a maximum (rather than another form of extremum) need to be verified of course. Equation (1.62) indicates that, starting from the description of the problem which includes $f(Y|\theta)$ and $f(\theta)$, the calculation of $\widehat{\theta}_{MAP}$ should be relatively straightforward.

(iv) The MAP estimator is given by the *global* maximum of $f(\theta|Y)$. If there are several local maxima, all of them need to be examined and compared to find the global maximum.

(v) Even if $f(\theta|Y)$ satisfies regularity conditions, the maximum may occur at the boundary of the parameter space Θ (which may not necessarily be $(-\infty,\infty)$). In that case, the maximum is not a local extremum.

In summary, the estimators that minimize the Bayes risk for the three cost functions we have considered are the mean, the median, and the mode of the posterior pdf. In general, these three estimators may differ from one another as is illustrated in Fig. 1.7.

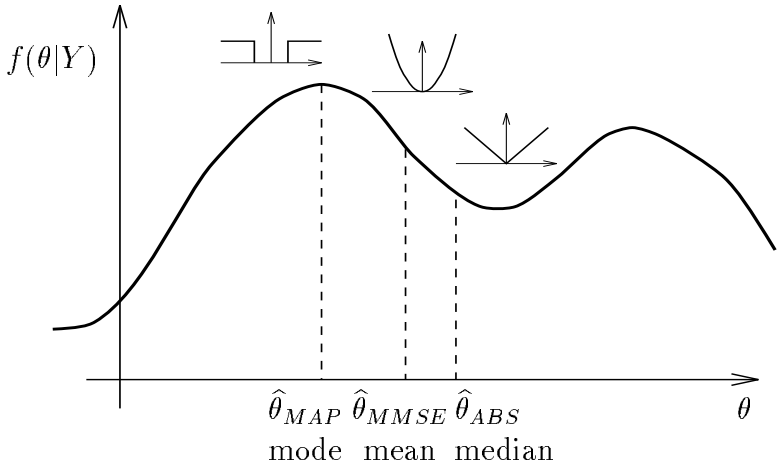


Figure 1.7: An example of the relative position of three Bayes estimators.

Example 1.4 Gaussian mean in Gaussian noise

Consider the problem of estimating an unknown dc level in additive Gaussian noise with zero mean. For instance, the technology used in a satellite to capture an image may lead to a noisy image, due to constraints of weight and power consumption. A digital image in sampled is the spatial directions and the samples are called pixels. Each pixel has a

certain grey value, of which we obtain a noisy measurement. However, the image a satellite is looking at normally remains stationary for quite some time. So if the satellite captures images of a certain scene at a rate of say 30 images per second, in a few seconds we get quite a number of noisy measurements of the same grey value at any given pixel. Since the scene the satellite camera is looking at may nevertheless change very slowly, we may organize the consecutive images in groups and say that e.g. 100 consecutive images are noisy measurements of the same image, while the next 100 shots are measurements of a possibly different image. For any given pixel, the measurement problem can be modeled as

$$y_i \;=\; \theta + v_i \;,\quad i = 1, \ldots, n \tag{1.63}$$

where θ is the true grey value of the pixel, y_i are the consecutive noisy measured grey values, the $v_i \sim \mathcal{N}(0, \sigma_v^2)$ are i.i.d.: the measurement noise has zero mean, a constant variance that depends on the equipment, is independent from one shot to another, and may be the cumulative effect of a number of influences, justifying a Gaussian approximation based on the Central Limit Theorem. In our example, $n = 100$. Even though the images vary very slowly, the image of one group of 100 shots will not differ very much from the image of the previous 100 shots. Therefore, we can consider the image estimate (with value m_θ at the pixel considered) from the previous group to be prior information for the problem of estimating the image of the current group (with value θ at the same pixel). This prior information is not perfect due to the estimation variance associated with the processing of the previous group and also the variation from one image (group) to the next. Therefore we model the prior information as $\theta \sim \mathcal{N}(m_\theta, \sigma_\theta^2)$ where σ_θ^2 reflects the joint effect of estimation variability and variablity due to change in time. Also, θ is assumed to be independent of the v_i.

We can write the measurements in vector form

$$Y \;=\; \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \;=\; \theta \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \;=\; \theta\,\mathbf{1} + V \tag{1.64}$$

With $X = \theta$ and equation (1.64), we can consider the following invertible transformation

$$\begin{bmatrix} X \\ Y \end{bmatrix} \;=\; \begin{bmatrix} 1 & 0 \\ \mathbf{1} & I_n \end{bmatrix} \begin{bmatrix} \theta \\ V \end{bmatrix} \;. \tag{1.65}$$

Now, since θ and V are independent and both Gaussian, they are jointly Gaussian. In problem 1.1, it is shown that the linear transformation of jointly Gaussian random variables leads again to Gaussian random variables. Hence, $X = \theta$ and Y are jointly Gaussian. In order to determine a Bayes estimator, we need the a posteriori pdf $f(\theta|Y) = f(X|Y)$. Since X and Y are jointly Gaussian, the Gauss-Markov theorem in section 1.1 tells us that $f(X|Y)$ is also Gaussian and it also tells us how to compute it from the first and second-order moments of X and Y. So we shall compute these moments. The first-order moments are

$$\begin{aligned} m_X \;&=\; E\,X \;=\; m_\theta \\ m_Y \;&=\; E\,Y \;=\; m_\theta \mathbf{1} \;. \end{aligned} \tag{1.66}$$

The joint covariance matrix is

$$\begin{aligned} C &= E\left[\begin{array}{c} X-m_X \\ Y-m_Y \end{array}\right]\left[\begin{array}{c} X-m_X \\ Y-m_Y \end{array}\right]^T = E\left[\begin{array}{c} \theta-m_\theta \\ (\theta-m_\theta)\,\mathbf{1}+V \end{array}\right]\left[\begin{array}{c} \theta-m_\theta \\ (\theta-m_\theta)\,\mathbf{1}+V \end{array}\right]^T \\ &= \left[\begin{array}{cc} \sigma_\theta^2 & \sigma_\theta^2\mathbf{1}^T \\ \sigma_\theta^2\mathbf{1} & \sigma_\theta^2\mathbf{1}\mathbf{1}^T+\sigma_v^2 I \end{array}\right] = \left[\begin{array}{cc} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{array}\right] \end{aligned} \tag{1.67}$$

where we used $E\{(\theta-m_\theta)V\} = E\{(\theta-m_\theta)\}\ E\{V\} = 0.0 = 0$. A key quantity that appears in the expression for $f(X|Y)$ is $C_{XY}C_{YY}^{-1}$. From (1.67), we get

$$\begin{aligned} C_{XY}C_{YY}^{-1} &= \sigma_\theta^2\mathbf{1}^T\left[\sigma_v^2 I+\sigma_\theta^2\,\mathbf{1}\,\mathbf{1}^T\right]^{-1} \\ &= \tfrac{\sigma_\theta^2}{\sigma_v^2}\mathbf{1}^T\left[I+\tfrac{\sigma_\theta^2}{\sigma_v^2}\,\mathbf{1}\,\mathbf{1}^T\right]^{-1}. \end{aligned} \tag{1.68}$$

In order to compute the inverse of the matrix in brackets, we shall use the following identity.

Lemma 1.1 (Matrix Inversion Lemma) *If A and C are respectively $n\times n$ and $m\times m$ invertible matrices and B and D are respectively $n\times m$ and $m\times n$ matrices, then*

$$[A+BCD]^{-1} = A^{-1}-A^{-1}B\left[DA^{-1}B+C^{-1}\right]^{-1}DA^{-1} \tag{1.69}$$

if the first inverse exists.

This lemma can be shown by straightforward verification: the product of $A+BCD$ with the RHS of the equation gives the identity matrix. This lemma is very useful when we have to compute $[A+BCD]^{-1}$, we know A^{-1}, and m is (much) smaller than n. We shall apply this lemma with $A=I$, $B=\mathbf{1}$, $C=\frac{\sigma_\theta^2}{\sigma_v^2}$ and $D=\mathbf{1}^T$. Hence $m=1$. We obtain

$$\begin{aligned} C_{XY}C_{YY}^{-1} &= \tfrac{\sigma_\theta^2}{\sigma_v^2}\mathbf{1}^T[I-\mathbf{1}(\underbrace{\mathbf{1}^T\mathbf{1}+\frac{\sigma_v^2}{\sigma_\theta^2}}_{=n+\frac{\sigma_v^2}{\sigma_\theta^2}})^{-1}\mathbf{1}^T] \\ &= \tfrac{\sigma_\theta^2}{\sigma_v^2}\left(1-\tfrac{n}{n+\frac{\sigma_v^2}{\sigma_\theta^2}}\right)\mathbf{1}^T = \tfrac{1}{n+\frac{\sigma_v^2}{\sigma_\theta^2}}\mathbf{1}^T. \end{aligned} \tag{1.70}$$

From the Gauss-Markov theorem, we can now compute the a posteriori mean

$$\begin{aligned} E[\theta|Y] &= m_\theta + C_{XY}C_{YY}^{-1}(Y-m_Y) \\ &= m_\theta + \tfrac{1}{n+\frac{\sigma_v^2}{\sigma_\theta^2}}\mathbf{1}^T(Y-m_\theta\mathbf{1}) \\ &= \left(\tfrac{1}{\sigma_\theta^2}m_\theta+\tfrac{n}{\sigma_v^2}\overline{y}\right)/\left(\tfrac{1}{\sigma_\theta^2}+\tfrac{n}{\sigma_v^2}\right) \end{aligned} \tag{1.71}$$

where we introduced the sample mean $\overline{y} = \frac{1}{n}\mathbf{1}^T Y = \frac{1}{n}\sum_{i=1}^{n} y_i$. Note that $E[\theta|Y]$ is of the form $\alpha\, m_\theta + (1-\alpha)\overline{y}$, which is a convex combination between the a priori mean and the sample mean obtained from the data. The respective weighting factors are proportional to the inverse of the variance associated with each of the two components. We shall call the inverse of the variance the amount of information available:

- $\frac{1}{\sigma_\theta^2}$ = information in the prior distribution $f(\theta)$,
- $\frac{n}{\sigma_v^2}$ = information in the n independent measurements (conditional distribution $f(y_i|\theta) = f_{\mathbf{v}_i}(y_i - \theta)$).

We can consider two extreme cases:

- $\sigma_\theta^2 \ll \frac{\sigma_v^2}{n}$: $E[\theta|Y] \approx m_\theta$. In this case, the measurements are so noisy that the reliable a priori information dominates.
- $\sigma_\theta^2 \gg \frac{\sigma_v^2}{n}$: $E[\theta|Y] \approx \overline{y}$. In this case, the accurate measurements dominate the unreliable a priori information. Note that this second case will eventually occur when $n \to \infty$.

From the Gauss-Markov theorem, we can also determine the a posteriori variance, which we shall denote by $\sigma_{\widehat{\theta}}^2$:

$$\begin{aligned} \sigma_{\widehat{\theta}}^2 \;=\; Var[\theta|Y] \;&=\; C_{XX} - C_{XY}C_{YY}^{-1}C_{YX} \\ &=\; \sigma_\theta^2 - \frac{1}{n+\frac{\sigma_v^2}{\sigma_\theta^2}}\mathbf{1}^T\mathbf{1}\sigma_\theta^2 \;=\; \cdots \;=\; \left(\frac{1}{\sigma_\theta^2} + \frac{n}{\sigma_v^2}\right)^{-1} \end{aligned} \tag{1.72}$$

or hence

$$\frac{1}{\sigma_{\widehat{\theta}}^2} \;=\; \frac{1}{\sigma_\theta^2} + \frac{n}{\sigma_v^2} \tag{1.73}$$

which means that the a priori information and the information in the n independent measurements add up to form the total information in the posterior distribution. (1.73) allows us to rewrite the a posteriori mean in (1.71) as

$$\frac{1}{\sigma_{\widehat{\theta}}^2}\, E[\theta|Y] \;=\; \frac{1}{\sigma_\theta^2} m_\theta + \frac{n}{\sigma_v^2}\overline{y}\,. \tag{1.74}$$

Using the Gauss-Markov theorem, we are finally ready to write the a posteriori distribution as

$$f(\theta|Y) \;\leftrightarrow\; \mathcal{N}\left(E[\theta|Y],\; \sigma_{\widehat{\theta}}^2\right) \tag{1.75}$$

where $E[\theta|Y]$ is given by (1.71) or (1.74) and $\sigma_{\widehat{\theta}}^2$ is given in (1.73). Since the posterior distribution is Gaussian, its mean, mode and median coincide. Hence we get

$$\widehat{\theta}_{MMSE} \;=\; \widehat{\theta}_{MAP} \;=\; \widehat{\theta}_{ABS} \;=\; E[\theta|Y]\;. \tag{1.76}$$

Note that the posterior distribution $f(\theta|Y)$ depends on Y only through $\overline{y}$. In this case we say that the statistic $\overline{y}$ (a function of the data Y) is a *sufficient statistic*, meaning that, for the purpose of estimating θ, the only thing we need to know about $y_1, \ldots, y_n$ is $\overline{y}$. $\diamond$

1.3.4 Equivalences of Bayes estimators

One characteristic of Bayes estimation is that a Bayes estimator depends on the cost function it is associated with. In the following, we shall examine three cases in which the Bayes estimator coincides with $\widehat{\theta}_{MMSE}$ meaning that the estimator minimizing the risk based on some non-quadratic cost function coincides with $\widehat{\theta}_{MMSE}$, the estimator minimizing a quadratic cost function.

1. Consider a cost function $C(\widetilde{\theta})$ with the following properties:

   - symmetric: $C(\widetilde{\theta}) = C(-\widetilde{\theta})$
   - convex : $C(\alpha\widetilde{\theta}_1+(1-\alpha)\widetilde{\theta}_2) \leq \alpha\, C(\widetilde{\theta}_1)+(1-\alpha)\, C(\widetilde{\theta}_2)\,,\ \alpha \in [0,1]$

   and let $f(\theta|Y)$ be symmetric about $E[\theta|Y]$. Then $\widehat{\theta}_C = \widehat{\theta}_{MMSE}$.

2. Consider a cost function $C(\widetilde{\theta})$ with the following properties:

   - symmetric: $C(\widetilde{\theta}) \ = \ C'(|\widetilde{\theta}|)$
   - $C'(|\widetilde{\theta}|)$ is a non-decreasing function of $|\widetilde{\theta}|$ (this condition is weaker than $C(\widetilde{\theta})$ being convex)

   and let $f(\theta|Y)$ have the following properties

   - symmetric about $E[\theta|Y]$
   - unimodal (only one local maximum)

   and finally let $C(.)$ and $f(.|Y)$ be such that $\lim\limits_{\theta\to\infty} C(\theta)\, f(\theta|Y) \ = \ 0\,,\ \forall Y$, then again $\widehat{\theta}_C = \widehat{\theta}_{MMSE}$.

3. If $f(\theta|Y)$ is Gaussian, then $\widehat{\theta}_{MMSE} = \widehat{\theta}_{MAP} = \widehat{\theta}_{ABS}$.

These equivalences show the relative importance of $\widehat{\theta}_{MMSE}$. In general however, $\widehat{\theta}_{MAP}$ is the easiest to compute, as is illustrated in the following example.

Example 1.5 Poisson process

Consider a communication network node where messages are passing. Let the observation $\mathbf{N}$ be the number of messages that pass during a certain observation period of duration τ. $\mathbf{N}$ has a Poisson distribution

$$\Pr\left[\mathbf{N} = N|\theta\right] \ = \ (\theta\tau)^N \, \frac{e^{-\theta\tau}}{N!}\,, \quad N = 0,1,2,\ldots \tag{1.77}$$

where the parameter θ represents the average number of messages passing per second by the node we are considering. This average frequency has itself an a priori distribution (over the different nodes in the network) which we take to be exponential with average value $m_\theta = 1/\lambda$:

$$f(\theta) \ = \ \begin{cases} \lambda\, e^{-\lambda\,\theta} & ,\ \theta > 0 \\ 0 & ,\ \theta \leq 0\,. \end{cases} \tag{1.78}$$

In order to find $\widehat{\theta}_{MMSE}$, we shall compute the a posteriori distribution of θ given the measurement N. Using Bayes' rule we get

$$\begin{aligned} f(\theta|N) &= \frac{\Pr[\mathbf{N}=N|\theta]\, f(\theta)}{\Pr[\mathbf{N}=N]} \\ &= \frac{1}{\Pr[\mathbf{N}=N]} \lambda \frac{\tau^N}{N!} \theta^N e^{-\theta(\tau+\lambda)} , \quad \theta > 0 \\ &= k(N) \theta^N e^{-\theta(\tau+\lambda)} , \quad \theta > 0 \end{aligned} \tag{1.79}$$

where $k(N)$ depends on N but not on θ. Note that $f(\theta|N) = 0, \ \theta \leq 0$. In order to determine $k(N)$, we use the constraint

$$\int_0^\infty f(\theta|N) d\theta \ = \ 1 \ = \ k(N) \underbrace{\int_0^\infty \theta^N e^{-\theta(\tau+\lambda)} d\theta}_{g(N)} . \tag{1.80}$$

Hence $k(N) = 1/g(N)$. We shall calculate $g(N)$ by partial integration:

$$g(N) \ = \ \left[\frac{\theta^N e^{-\theta(\tau+\lambda)}}{-(\tau+\lambda)} \right]_0^\infty + \frac{N}{\tau+\lambda} \int_0^\infty \theta^{N-1} e^{-\theta(\tau+\lambda)} d\theta . \tag{1.81}$$

Hence

$$g(N) \ = \ \frac{N}{\tau+\lambda} g(N-1) \ = \cdots = \ \frac{N!}{(\tau+\lambda)^N} g(0) \tag{1.82}$$

where

$$g(0) \ = \ \int_0^\infty e^{-\theta(\tau+\lambda)} d\theta \ = \ \left[\frac{e^{-\theta(\tau+\lambda)}}{-(\tau+\lambda)} \right]_0^\infty \ = \ \frac{1}{\tau+\lambda} \tag{1.83}$$

which finally leads to

$$g(N) \ = \ \frac{N!}{(\tau+\lambda)^{N+1}} , k(N) \ = \ 1/g(N) \ = \ \frac{(\tau+\lambda)^{N+1}}{N!} . \tag{1.84}$$

Now we can calculate

$$\begin{aligned} \widehat{\theta}_{MMSE} &= E\left[\theta|N\right] \ = \ \textstyle\int_0^\infty \theta\, f(\theta|N)\, d\theta \\ &= \frac{1}{g(N)} \int_0^\infty \theta^{N+1} e^{-\theta(\tau+\lambda)} d\theta \ = \ \frac{g(N+1)}{g(N)} \ = \ \frac{N+1}{\tau+\lambda} . \end{aligned} \tag{1.85}$$

To determine $\widehat{\theta}_{MAP}$, consider the logarithm of the posterior pdf

$$\ln f(\theta|N) \ = \ -\ln g(N) + N \ln \theta - \theta(\tau+\lambda) . \tag{1.86}$$

Differentiation w.r.t. θ yields

$$\frac{\partial}{\partial \theta} \ln f(\theta|Y) \ = \ 0 \ = \ \frac{N}{\theta} - (\tau+\lambda) . \tag{1.87}$$

So we obtain

$$\widehat{\theta}_{MAP} \ = \ \frac{N}{\tau+\lambda} \ \neq \ \frac{N+1}{\tau+\lambda} \ = \ \widehat{\theta}_{MMSE} . \tag{1.88}$$

Note however that if $\tau \gg \lambda$ (observation duration much longer than the a priori expected time between observations), then with high probability $N \gg 1$ and hence $\widehat{\theta}_{MAP} \approx \widehat{\theta}_{MMSE} \approx \frac{N}{\tau}$, which is simply the sample average of the number of messages per second.

$\diamond$

1.3.5 Vector Parameters

In many estimation problems, more than one, say m, parameters are unknown. We shall still denote the vector parameter as

$$\theta \;=\; \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix} . \tag{1.89}$$

Most of the notions from the scalar case carry through to the vector case. In particular, we have a prior distribution $f_{\boldsymbol{\theta}}(\theta)$ for the random parameters and a conditional distribution $f_{\mathbf{Y}|\boldsymbol{\theta}}(Y|\theta)$ describing the uncertain relation that links the parameters θ to the observed variables Y. Bayes' rule allows us to obtain the joint distribution $f_{\mathbf{Y},\boldsymbol{\theta}}(Y,\theta) = f_{\mathbf{Y}|\boldsymbol{\theta}}(Y|\theta) f_{\boldsymbol{\theta}}(\theta)$. We again introduce an estimator for θ as a statistic, a nonlinear function of the data Y,

$$\widehat{\theta}(Y) \;=\; \begin{bmatrix} \widehat{\theta}_1(Y) \\ \vdots \\ \widehat{\theta}_m(Y) \end{bmatrix} , \quad \widetilde{\theta} \;=\; \widetilde{\theta}(\theta, Y) \;=\; \theta - \widehat{\theta}(Y) . \tag{1.90}$$

We shall again associate a cost $\mathcal{C}(\theta, \widehat{\theta}(Y))$ which will often be directly a funtion of the estimation error $\widetilde{\theta}$ and in fact often a function of the length of the estimation error $\left\|\widetilde{\theta}\right\|$. We obtain the estimator function $\widehat{\theta}(.)$ by minimizing the risk, which is the expected value of the cost:

$$\begin{aligned} \min_{\widehat{\theta}(.)} \mathcal{R}(\widehat{\theta}(.)) \;&=\; \min_{\widehat{\theta}(.)} E\,\mathcal{C}(\theta, \widehat{\theta}(Y)) \;=\; \min_{\widehat{\theta}(.)} E_{\mathbf{Y},\boldsymbol{\theta}} \mathcal{C}(\theta, \widehat{\theta}(Y)) \\ &=\; \min_{\widehat{\theta}(.)} E_{\mathbf{Y}} E_{\boldsymbol{\theta}|\mathbf{Y}} \mathcal{C}(\theta, \widehat{\theta}(Y)) \;=\; E_{\mathbf{Y}} \left[\min_{\widehat{\theta}(Y)} E_{\boldsymbol{\theta}|\mathbf{Y}} \mathcal{C}(\theta, \widehat{\theta}(Y))\right] \\ &=\; E_{\mathbf{Y}} \left[\min_{\widehat{\theta}(Y)} \mathcal{R}(\widehat{\theta}(Y)|Y)\right] . \end{aligned} \tag{1.91}$$

$\mathcal{R}(\widehat{\theta}(.))$ is a weighted average of $\mathcal{R}(\widehat{\theta}(Y)|Y)$, weighted by the nonnegative weighting function $f_{\mathbf{Y}}(Y)$. The minimum of $\mathcal{R}(\widehat{\theta}(.))$ w.r.t. $\widehat{\theta}(.)$ will hence be obtained by minimizing $\mathcal{R}(\widehat{\theta}(Y)|Y)$ w.r.t. $\widehat{\theta}(Y)$ for every Y. Note that $\mathcal{R}(\widehat{\theta}(Y)|Y)$ depends on the posterior distribution $f_{\boldsymbol{\theta}|\mathbf{Y}}(\theta|Y)$.

For the minimization of cost functions involving a vector parameter, we need to consider the gradient w.r.t. the vector parameter. If $g(\theta)$ is a $1 \times l$ row vector function then the gradient w.r.t. θ is defined as

$$\frac{\partial\, g(\theta)}{\partial \theta} \;=\; \begin{bmatrix} \frac{\partial\, g(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial\, g(\theta)}{\partial \theta_m} \end{bmatrix} \tag{1.92}$$

which is a $m \times l$ matrix. If $g(\theta)$ is a scalar ($l = 1$), then $\frac{\partial\, g(\theta)}{\partial \theta}$ is a column vector of the same dimensions as θ. The gradient operator commutes with linear operations. Some often-used formulas are the following.

$$\frac{\partial\, \theta^T}{\partial \theta} \;=\; \left[\frac{\partial\, \theta_j}{\partial \theta_i}\right] \;=\; I_m . \tag{1.93}$$

Let X be a $m \times 1$ column vector.

$$\frac{\partial}{\partial \theta}\left(\theta^T X\right) \;=\; \left(\frac{\partial\, \theta^T}{\partial \theta}\right) X \;=\; I_m\, X \;=\; X . \tag{1.94}$$

Since a scalar equals its transpose, we get

$$\frac{\partial}{\partial\theta}\left(X^T\theta\right) \;=\; \frac{\partial}{\partial\theta}\left(\theta^T X\right) \;=\; X\;. \tag{1.95}$$

If A is a $m \times l$ matrix, then we get

$$\frac{\partial}{\partial\theta}\left(\theta^T A\right) \;=\; \left(\frac{\partial\,\theta^T}{\partial\theta}\right) A \;=\; I_m\, A \;=\; A\;. \tag{1.96}$$

In the following, we use an extension of the well-known rule for the scalar case, $(uv)' = u'\,v + u\,v'$, to the vector case. Let $g(\theta)$ and $h(\theta)$ both be $l \times 1$ column vector functions of θ. Since $g^T(\theta)h(\theta) = \left(g^T(\theta)h(\theta)\right)^T = h^T(\theta)g(\theta)$, we get

$$\frac{\partial}{\partial\theta}\left(g^T(\theta)h(\theta)\right) \;=\; \left(\frac{\partial g^T(\theta)}{\partial\theta}\right) h(\theta) + \left(\frac{\partial h^T(\theta)}{\partial\theta}\right) g(\theta)\;. \tag{1.97}$$

A particular application of this rule with $g(\theta) = \theta$ and $h(\theta) = A\theta$ leads to

$$\frac{\partial}{\partial\theta}\left(\theta^T A\theta\right) \;=\; \left(\frac{\partial\,\theta^T}{\partial\theta}\right) A\theta + \left(\frac{\partial\,\theta^T}{\partial\theta}\right) A^T\theta \;=\; \left(A + A^T\right)\theta\;. \tag{1.98}$$

When A is symmetric, this gradient reduces to $2\,A\theta$.

1.3.6 The MMSE criterion: Vector Parameters

For the Minimum Mean Squared Error criterion, we have the quadratic cost function $\mathcal{C}_{MMSE}(\theta,\widehat{\theta}) = \left\|\widetilde{\theta}\right\|_2^2 = \widetilde{\theta}^T\widetilde{\theta}$. The minimization problem of the conditional Bayes risk becomes

$$\min_{\widehat{\theta}(Y)} \mathcal{R}_{MMSE}(\widehat{\theta}(Y)|Y) \;=\; \min_{\widehat{\theta}(Y)} \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\theta|Y)\,\left(\theta-\widehat{\theta}\right)^T\left(\theta-\widehat{\theta}\right) d\theta_1\cdots d\theta_m\;. \tag{1.99}$$

In order to find the minimum, we investigate the extrema. Using Leibnitz's rule and (1.97) with $g(\theta) = h(\theta) = \theta - \widehat{\theta}$, we obtain (using the simplified notation $d\theta \;=\; d\theta_1\cdots d\theta_m$):

$$\begin{array}{l} \frac{\partial}{\partial\widehat{\theta}}\mathcal{R}_{MMSE}(\widehat{\theta}|Y) \;=\; \frac{\partial}{\partial\widehat{\theta}}\int f(\theta|Y)\,\left(\theta-\widehat{\theta}\right)^T\left(\theta-\widehat{\theta}\right)\,d\theta \\ =\; \int f(\theta|Y)\,\left(\frac{\partial}{\partial\widehat{\theta}}\left(\theta-\widehat{\theta}\right)^T\left(\theta-\widehat{\theta}\right)\right)\,d\theta \;=\; 2\int f(\theta|Y)\,(\widehat{\theta}-\theta)\,d\theta \;\;=0 \end{array} \tag{1.100}$$

where we put the derivative equal to zero in order to find an extremum. (1.100) can be rewritten as

$$\widehat{\theta}(Y)\;\underbrace{\int f(\theta|Y)\,d\theta}_{=\,1} \;=\; \int \theta\, f(\theta|Y)\,d\theta \tag{1.101}$$

or hence

$$\widehat{\theta}_{MMSE}(Y) \;=\; E(\theta|Y) \tag{1.102}$$

which is again the *mean* of the a posteriori distribution of θ given Y. To know whether this extremum is a minimum, we investigate the second-order derivatives

$$\begin{aligned} Hessian &= \left[\frac{\partial^2}{\partial\widehat{\theta}_i\,\partial\widehat{\theta}_j}\mathcal{R}_{MMSE}(\widehat{\theta}|Y)\right] = \frac{\partial}{\partial\widehat{\theta}}\left(\frac{\partial}{\partial\widehat{\theta}}\mathcal{R}_{MMSE}(\widehat{\theta}|Y)\right)^T \\ &= 2\int f(\theta|Y)\left[\frac{\partial\widehat{\theta}^T}{\partial\widehat{\theta}} - \frac{\partial\theta^T}{\partial\widehat{\theta}}\right] d\theta = 2I\int f(\theta|Y)\,d\theta = 2I > 0\,. \end{aligned} \tag{1.103}$$

Hence the extremum at $\widehat{\theta}(Y) = E(\theta|Y)$ is a local minimum and it is furthermore the global minimum since it is the unique local minimum. Note that $E(\theta|Y)$ is indeed a function of Y.

One property is that MMSE estimation commutes over linear transformations. Indeed, let $\phi = A\theta$ $(l \times 1)$ be another set of parameters obtained by linear transformation of θ. Then

$$\widehat{\phi}_{MMSE} = E(\phi|Y) = E(A\theta|Y) = A\,E(\theta|Y) = A\,\widehat{\theta}_{MMSE}\,. \tag{1.104}$$

Orthogonality property of MMSE estimators♠

The following orthogonality property of the conditional mean shows that there are two equivalent ways of defining the conditional mean. Let $\boldsymbol{\theta}$ and $\mathbf{Y}$ be random vectors ($\mathbf{Y}$ could even be a discrete-time or continuous-time stochastic process, observed over an infinite time interval) so that we can define

$$E(\boldsymbol{\theta}|\mathbf{Y}) = \int \theta\, f_{\boldsymbol{\theta}|\mathbf{Y}}(\theta|Y)\,d\theta\,. \tag{1.105}$$

Then the following orthogonality property holds

$$\widehat{\theta}(\mathbf{Y}) = E(\boldsymbol{\theta}|\mathbf{Y}) \quad \text{iff} \quad E((\boldsymbol{\theta} - \widehat{\theta}(\mathbf{Y}))\,g(\mathbf{Y})) = 0\,, \;\; \forall\, g(.) \tag{1.106}$$

where $g(.)$ is a scalar function. The property can equivalently be written as

$$E(\widehat{\theta}(\mathbf{Y})\,g(\mathbf{Y})) = E(\boldsymbol{\theta}\,g(\mathbf{Y}))\,, \;\; \forall\, g(.)\,. \tag{1.107}$$

(1.105) represents an alternative way of defining $E(\boldsymbol{\theta}|\mathbf{Y})$.
Proof: Only if: Assume $\widehat{\theta}(\mathbf{Y}) = E(\boldsymbol{\theta}|\mathbf{Y})$. Then for any $g(.)$

$$\begin{aligned} E(\widehat{\theta}(\mathbf{Y})\,g(\mathbf{Y})) &= \int\int d\theta\,dY\, f_{\boldsymbol{\theta},\mathbf{Y}}(\theta,Y)\,g(Y)\int du\,u\,f_{\boldsymbol{\theta}|\mathbf{Y}}(u|Y) \\ &= \int dY\, f_{\mathbf{Y}}(Y)\,g(Y)\int du\,u\,f_{\boldsymbol{\theta}|\mathbf{Y}}(u|Y)\underbrace{\int d\theta\, f_{\boldsymbol{\theta}|\mathbf{Y}}(\theta|Y)}_{=1} \\ &= \int\int du\,dY\, f_{\boldsymbol{\theta},\mathbf{Y}}(u,Y)\,u\,g(Y) = E(\boldsymbol{\theta}\,g(\mathbf{Y}))\,. \end{aligned} \tag{1.108}$$

If: Suppose now that $\widehat{\theta}(.)$ is a function of Y that satisfies for all $g(.)$

$$\begin{aligned} E(\widehat{\theta}(\mathbf{Y})\,g(\mathbf{Y})) &= E(\boldsymbol{\theta}\,g(\mathbf{Y})) \\ \Leftrightarrow \int\int d\theta\,dY\,\widehat{\theta}(Y)g(Y)f_{\boldsymbol{\theta},\mathbf{Y}}(\theta,Y) &= \int\int d\theta\,dY\,\theta g(Y)f_{\boldsymbol{\theta},\mathbf{Y}}(\theta,Y) \\ \Leftrightarrow \int dY\,\widehat{\theta}(Y)g(Y)f_{\mathbf{Y}}(Y)\underbrace{\int d\theta f_{\boldsymbol{\theta}|\mathbf{Y}}(\theta|Y)}_{=1} &= \int dY\,g(Y)f_{\mathbf{Y}}(Y)\int d\theta\,\theta\,f_{\boldsymbol{\theta}|\mathbf{Y}}(\theta|Y) \end{aligned} \tag{1.109}$$

Take in particular $g(Y) = \delta(Y-V)$ $(= \delta(Y_1-V_1)\cdots\delta(Y_n-V_n)$ if Y is a vector with n components). Then we get

$$\widehat{\theta}(V)f_{\mathbf{Y}}(V) \;=\; f_{\mathbf{Y}}(V)\int d\theta\;\theta\, f_{\boldsymbol{\theta}|\mathbf{Y}}(\theta|V) \quad \text{or} \quad f_{\mathbf{Y}}(Y)\left(\widehat{\theta}(Y) - \int d\theta\;\theta\, f_{\boldsymbol{\theta}|\mathbf{Y}}(\theta|Y)\right) = 0\;. \tag{1.110}$$

Hence

$$\widehat{\theta}(Y) \;=\; \int d\theta\;\theta\, f_{\boldsymbol{\theta}|\mathbf{Y}}(\theta|Y) \tag{1.111}$$

whenever $f_{\mathbf{Y}}(Y) > 0$ (and we do not care about the value of $\widehat{\theta}(Y)$ when $f_{\mathbf{Y}}(Y) = 0$ because when $f_{\mathbf{Y}}(Y) = 0$, the value of $\widehat{\theta}(Y)$ does not influence the integral in the orthogonality property (1.106). Note that for the minimization of the Bayes risk in (1.91), the value of $\widehat{\theta}(Y)$ did not matter either whenever $f_{\mathbf{Y}}(Y) = 0$). □

The orthogonality property provides us with an elegant way of showing that the MMSE Bayes risk is indeed minimized for $\widehat{\theta}_{MMSE}(Y) = E(\theta|Y)$. Indeed, let $\widehat{\theta}(Y)$ be any $m \times 1$ nonlinear function of Y. Then

$$\begin{aligned} E\left\|\theta - \widehat{\theta}(Y)\right\|_2^2 \;&=\; E\left\|\theta - E(\theta|Y) + E(\theta|Y) - \widehat{\theta}(Y)\right\|_2^2 \\ &=\; E\left\|\theta - E(\theta|Y)\right\|_2^2 + \underbrace{E\left\|E(\theta|Y) - \widehat{\theta}(Y)\right\|_2^2}_{\geq 0} + 2\,\underbrace{E\left((\theta - E(\theta|Y))^T(E(\theta|Y) - \widehat{\theta}(Y))\right)}_{=0} \\ &\geq\; E\left\|\theta - E(\theta|Y)\right\|_2^2 \end{aligned} \tag{1.112}$$

where the crossterms are zero because of the orthogonality property of $E(\theta|Y)$ (indeed, $E(\theta|Y) - \widehat{\theta}(Y)$ is a function of Y). This shows that no $\widehat{\theta}(Y)$ leads to a smaller MMSE risk than the one obtained with $\widehat{\theta}_{MMSE}(Y) = E(\theta|Y)$. In a similar way, it can be shown that the MMSE estimator minimizes the correlation matrix of the estimation errors, viz.

$$E\,(\theta - E(\theta|Y))(\theta - E(\theta|Y))^T \;\leq\; E\,(\theta - \widehat{\theta})(\theta - \widehat{\theta})^T\;. \tag{1.113}$$

In fact, inequality (1.112) follows from inequality (1.113) by taking the trace (= sum of the diagonal elements).

1.3.7 The Uniform Cost Function and MAP Estimation: Vector Parameters

Let us first introduce the notion of a ball centered around θ_o with radius δ as the set

$$\mathcal{B}_\delta(\theta_o) \;=\; \{\theta \in \Theta : \|\theta - \theta_0\|_2 \leq \delta\}\;. \tag{1.114}$$

Then the natural extension to the vector case of the uniform cost function is

$$\mathcal{C}_{UNIF}(\theta, \widehat{\theta}) \;=\; \begin{cases} 0 & ,\; \theta \in \mathcal{B}_\delta(\widehat{\theta}) \\ 1 & ,\; \theta \in \Theta \setminus \mathcal{B}_\delta(\widehat{\theta}) \end{cases} \tag{1.115}$$

The conditional Bayes risk becomes

$$\begin{aligned} \mathcal{R}_{UNIF}(\widehat{\theta}(Y)|Y) &= \int_{\Theta} f(\theta|Y)\, \mathcal{C}_{UNIF}(\theta, \widehat{\theta}) d\theta = \int_{\Theta \setminus \mathcal{B}_\delta(\widehat{\theta})} f(\theta|Y)\, d\theta \\ &= \int_{\Theta} f(\theta|Y)\, d\theta - \int_{\mathcal{B}_\delta(\widehat{\theta})} f(\theta|Y)\, d\theta = 1 - \int_{\mathcal{B}_\delta(\widehat{\theta})} f(\theta|Y)\, d\theta \,. \end{aligned} \tag{1.116}$$

The optimization problem $\min\limits_{\widehat{\theta}(Y)} \mathcal{R}_{UNIF}(\widehat{\theta}(Y)|Y)$ hence leads to

$$\max_{\widehat{\theta}(Y)} \int_{\mathcal{B}_\delta(\widehat{\theta})} f(\theta|Y)\, d\theta \;\approx\; \mathrm{Vol}\,(\mathcal{B}_\delta(0)) \;\max_{\widehat{\theta}(Y)} f(\widehat{\theta}|Y) \tag{1.117}$$

where the approximation becomes arbitrarily accurate as δ becomes arbitrarily small. This leads us to the Maximum A Posteriori (probability) estimator

$$\widehat{\theta}_{MAP}(Y) = \arg\max_{\theta \in \Theta} f(\theta|Y) \,. \tag{1.118}$$

The same remarks as in the scalar case hold here also. In particular, one may equivalently obtain $\widehat{\theta}_{MAP}(Y)$ from the optimization problem $\widehat{\theta}_{MAP}(Y) = \arg\max\limits_{\theta \in \Theta} \ln f(\theta|Y)$. Under certain regularity conditions, $\widehat{\theta}_{MAP}(Y)$ can be found as a solution to

$$\frac{\partial}{\partial \theta} \ln f(\theta|Y) = 0 = \frac{\partial}{\partial \theta} \ln f(Y|\theta) + \frac{\partial}{\partial \theta} \ln f(\theta) \;. \tag{1.119}$$

Also MAP estimation commutes over linear transformations. Let again $\phi = A\theta$ $(m \times 1)$ be another set of parameters obtained by linear transformation of θ. Then

$$\begin{aligned} \widehat{\phi}_{MAP}(Y) &= \arg\max_{\phi} f_{\boldsymbol{\phi}|\mathbf{Y}}(\phi|Y) = A \arg\max_{\theta} f_{\boldsymbol{\phi}|\mathbf{Y}}(A\theta|Y) \\ &= A \arg\max_{\theta} \frac{1}{|\det A|} f_{\boldsymbol{\theta}|\mathbf{Y}}(\theta|Y) = A\, \widehat{\theta}_{MAP}(Y) \,. \end{aligned} \tag{1.120}$$

This argument can be extended to the case in which ϕ and θ do not have the same dimensions.

1.3.8 The Cramer-Rao Lower Bound for Stochastic Parameters

There exists a lower bound on the correlation matrix of the estimator errors. The lower bound to be discussed is independent of the Bayes estimator used, and depends only on the posterior distribution. The lower bound is specified in terms of the *information matrix*.

The Fisher Information Matrix

The idea behind the information matrix is to express in quantitive terms the information carried by the posterior distribution about the parameters θ. For such an information measure, the following properties are desirable:

- The information should increase as the *sensitivity* of $f(\theta|Y)$ to changes in θ increases. Hence, the information should be an increasing function of $\frac{\partial\, f(\theta|Y)}{\partial \theta}$ or of $\frac{\partial\, \ln f(\theta|Y)}{\partial \theta}$.

- The information should be *additive* in the sense that it should be the sum of the informations from the prior distribution ($f(\theta)$) and from the data ($f(Y|\theta)$). Furthermore if, given θ, Y_1 and Y_2 are independent ($f(Y_1, Y_2|\theta) = f(Y_1|\theta)f(Y_2|\theta)$), then the informations in Y_1 and Y_2 should add up.

- The information should be positive and should be insensitive to a change of sign of θ.

- The information should be a *deterministic* quantity, it should not depend on a particular realization of the random variables involved.

The information matrix is defined as

$$J \;=\; E\left(\frac{\partial\ \ln\ f(\theta|Y)}{\partial\theta}\right)\left(\frac{\partial\ \ln\ f(\theta|Y)}{\partial\theta}\right)^T\ . \tag{1.121}$$

It can be shown to satisfy all the properties mentioned above. We first develop an alternative expression in terms of second-order derivatives. Note that

$$\frac{\partial\ \ln\ f(\theta|Y)}{\partial\theta} \;=\; \frac{1}{f(\theta|Y)}\frac{\partial\ f(\theta|Y)}{\partial\theta}\ . \tag{1.122}$$

This allows us to write the Hessian of $\ln\ f(\theta|Y)$ as

$$\begin{aligned}\tfrac{\partial}{\partial\theta}\left(\tfrac{\partial\ \ln\ f(\theta|Y)}{\partial\theta}\right)^T &= \tfrac{1}{f^2(\theta|Y)}\left[f(\theta|Y)\,\tfrac{\partial}{\partial\theta}\left(\tfrac{\partial\ f(\theta|Y)}{\partial\theta}\right)^T - \left(\tfrac{\partial\ f(\theta|Y)}{\partial\theta}\right)\left(\tfrac{\partial\ f(\theta|Y)}{\partial\theta}\right)^T\right] \\ &= \tfrac{1}{f(\theta|Y)}\,\tfrac{\partial}{\partial\theta}\left(\tfrac{\partial\ f(\theta|Y)}{\partial\theta}\right)^T - \left(\tfrac{\partial\ \ln\ f(\theta|Y)}{\partial\theta}\right)\left(\tfrac{\partial\ \ln\ f(\theta|Y)}{\partial\theta}\right)^T\ .\end{aligned} \tag{1.123}$$

For the expectation of the first term, we get

$$\begin{aligned}E\,\tfrac{1}{f(\theta|Y)}\,\tfrac{\partial}{\partial\theta}\left(\tfrac{\partial\ f(\theta|Y)}{\partial\theta}\right)^T &= \int d\theta \int dY\ f(Y)\,\tfrac{\partial}{\partial\theta}\left(\tfrac{\partial\ f(\theta|Y)}{\partial\theta}\right)^T \\ &= \tfrac{\partial}{\partial\theta}\left(\tfrac{\partial}{\partial\theta}\int d\theta\int dY\ f(Y)f(\theta|Y)\right)^T \;=\; \tfrac{\partial}{\partial\theta}\left(\tfrac{\partial}{\partial\theta}1\right)^T \;=\; 0\ .\end{aligned} \tag{1.124}$$

From (1.121), (1.123) and (1.124), it follows that

$$J \;=\; -E\,\frac{\partial}{\partial\theta}\left(\frac{\partial\ \ln\ f(\theta|Y)}{\partial\theta}\right)^T\ . \tag{1.125}$$

This expression will often allow us to obtain J more easily than via (1.121). Note also that

$$\frac{\partial\ \ln\ f(Y,\theta)}{\partial\theta} \;=\; \frac{\partial\ \ln\ f(\theta|Y)}{\partial\theta} + \frac{\partial\ \ln\ f(Y)}{\partial\theta} \;=\; \frac{\partial\ \ln\ f(\theta|Y)}{\partial\theta} \tag{1.126}$$

so that as long as derivatives are taken, we can interchange $f(Y,\theta)$ and $f(\theta|Y)$.

Conditions on the Estimator Bias

The (conditional) *bias* of an estimator $\widehat{\theta}(Y)$ of θ is defined as

$$b_{\widehat{\theta}}(\theta) \;=\; -E_{\mathbf{Y}|\boldsymbol{\theta}}\widetilde{\theta} \;=\; E_{\mathbf{Y}|\boldsymbol{\theta}}\left(\widehat{\theta}(Y)-\theta\right) \;=\; E_{\mathbf{Y}|\boldsymbol{\theta}}\widehat{\theta}(Y) \;-\; \theta\;. \tag{1.127}$$

In order to show the Cramer-Rao lower bound, we shall assume that one of the following two conditions holds. Either

$$E_{\boldsymbol{\theta}}\, b_{\widehat{\theta}}(\theta) \;=\; 0 \tag{1.128}$$

which, stated in words, means that the unconditional or average bias is zero, or

$$\lim_{\theta\to\partial\Theta} f(\theta) b_{\widehat{\theta}}(\theta) \;=\; 0 \tag{1.129}$$

where Θ is the domain for θ and $\partial\Theta$ is its boundary. (1.129) means that the product $f(\theta)b_{\widehat{\theta}}(\theta)$ disappears as θ approaches any point on the boundary of Θ. Often $\Theta=\mathcal{R}^m$ in which case the boundaries are at infinity. In that case, (1.129) will be satisfied if the prior density $f(\theta)$ decays to zero faster than the bias $b_{\widehat{\theta}}(\theta)$ explodes as $|\theta_i|\to\infty$ for at least one $i\in\{1,\ldots,m\}$. This does not represent a stringent requirement on the estimator, especially if the decay of $f(\theta)$ is exponential.

Lemma 1.2 (Unit Cross Correlation) *If either condition (1.128) or condition (1.129) is satisfied, then*

$$E\,\frac{\partial\,\ln f(Y,\theta)}{\partial\theta}\left(\widehat{\theta}-\theta\right)^T \;=\; I\;. \tag{1.130}$$

In words, the cross correlation matrix between $\frac{\partial\,\ln f(Y,\theta)}{\partial\theta}$ and the estimation error of any estimator satisfying (1.128) or (1.129) is the identity matrix.

Proof: In the case that condition (1.128) holds, we have

$$\int d\theta \int dY\; f(Y,\theta)\left(\widehat{\theta}-\theta\right)^T \;=\; 0\;. \tag{1.131}$$

By taking the gradient w.r.t. θ, we get

$$\int d\theta \int dY\,\frac{\partial\, f(Y,\theta)}{\partial\theta}\left(\widehat{\theta}-\theta\right)^T - I\,\underbrace{\int d\theta\int dY\; f(Y,\theta)}_{=1} \;=\; 0\;. \tag{1.132}$$

Hence

$$\begin{aligned} I \;&=\; \textstyle\int d\theta\int dY\,\frac{\partial\, f(Y,\theta)}{\partial\theta}\left(\widehat{\theta}-\theta\right)^T \;=\; \int d\theta\int dY\; f(Y,\theta)\frac{\partial\,\ln f(Y,\theta)}{\partial\theta}\left(\widehat{\theta}-\theta\right)^T \\ &=\; E\,\frac{\partial\,\ln f(Y,\theta)}{\partial\theta}\left(\widehat{\theta}-\theta\right)^T\;. \end{aligned} \tag{1.133}$$

In the case condition (1.129) holds, consider

$$\begin{aligned} \frac{\partial}{\partial\theta}\left(f(\theta)b^T_{\widehat{\theta}}(\theta)\right) \;&=\; \frac{\partial}{\partial\theta}\left(f(\theta)\; E_{\mathbf{Y}|\boldsymbol{\theta}}\left(\widehat{\theta}(Y)-\theta\right)^T\right) \\ &=\; -\,I\int dY\; f(Y,\theta) \;+\; \int dY\,\frac{\partial f(Y,\theta)}{\partial\theta}\left(\widehat{\theta}(Y)-\theta\right)^T\;. \end{aligned} \tag{1.134}$$

By integrating both sides over θ, we find

$$0 \;=\; \int_\Theta d\theta \,\frac{\partial}{\partial\theta}\left(f(\theta)b_{\hat\theta}^T(\theta)\right) \;=\; -I \;+\; E\,\frac{\partial\,\ln f(Y,\theta)}{\partial\theta}\left(\hat\theta-\theta\right)^T \tag{1.135}$$

where the first equality follows from (1.129). □

Schur Complements are Positive Semidefinite♠

We shall formulate the following property in its full generality. It is in fact a generalization of the Cauchy-Schwarz inequality to matrix-valued inner products. Before considering matrix-valued inner products, we shall recall the defining properties of a classical inner product. An inner product $<.,.>$ associates a real number $<x,y>\in\mathcal{R}$ with two vectors x and y of the vector space $\mathcal{V}$ we are considering, and it has the following properties:

1. linearity: $\forall\alpha_1,\alpha_2,\beta_1,\beta_2\in\mathcal{R},\quad \forall x,x_1,x_2,y,y_1,y_2\in\mathcal{V}$:

$$\begin{aligned} <\alpha_1x_1+\alpha_2x_2,y> &= \alpha_1<x_1,y>+\alpha_2<x_2,y> \\ <x,\beta_1y_1+\beta_2y_2> &= <x,y_1>\beta_1+<x,y_2>\beta_2 \end{aligned} \tag{1.136}$$

2. symmetry: $<x,y>=<y,x>$

3. non-degeneracy (of the norm induced by the inner product): $<x,x>=\|x\|^2\geq 0$. If $\|x\|=0$, then $x=0$.

One particular example is a space of random variables with the correlation as inner product: $<x,y>=E\,xy$. The non-degeneracy requirement holds also in this case modulo the following subtlety:

$$E\,x^2 \;=\; 0 \quad\Rightarrow\quad x=0 \text{ a.s. or a.e. or w.p. 1} \tag{1.137}$$

where the three equivalent notions are "almost surely" or "almost everywhere" or "with probability 1". Indeed, $Ex^2 \;=\; m_x^2+\sigma_x^2 \;=\; 0$ implies that both the mean m_x and the variance σ_x^2 of x are zero. From the Tchebycheff inequality

$$\Pr\left(|x-m_x|\geq\epsilon\right)\leq\frac{\sigma_x^2}{\epsilon^2}\,,\quad\forall\epsilon>0\,, \tag{1.138}$$

we get that $\Pr(x=0)=1$. The meaning of $x=0$ w.p. 1 is illustrated in Fig. 1.8. A real random variable $\mathbf{x}$ is a function $\mathbf{x}(\omega)$ from the sample space Ω to the real line. Assume that Ω' is a subset of Ω such that all probability mass is concentrated in Ω': $\Pr(\Omega')=1$. Then $\mathbf{x}=0$ w.p. 1 if

$$\mathbf{x}(\omega) \;=\; \begin{cases} 0 & ,\,\omega\in\Omega' \\ \text{anything} & ,\,\omega\in\Omega\setminus\Omega' \end{cases} \tag{1.139}$$

This means that $\mathbf{x}=0$ in all realizations that can occur.

We now consider a vector space in which the vectors have multiple components such that the inner product is a real matrix. For example, consider a vector space of random vectors with inner product $<X,Y>=E\,XY^T$ where X and Y are column vectors of random variables (not

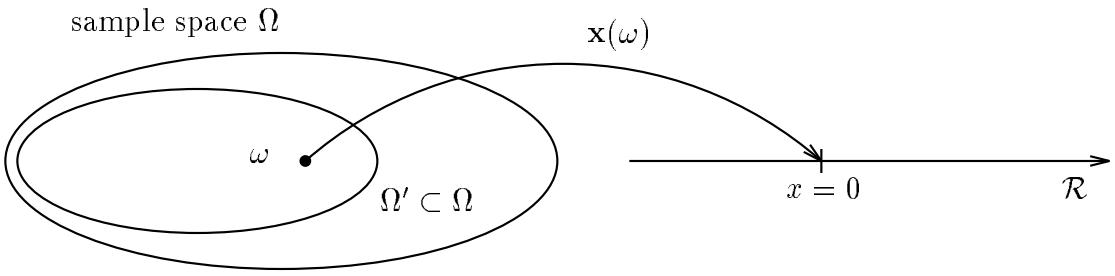


Figure 1.8: An illustration of $\mathbf{x} = 0$ w.p. 1.

necessarily with the same number of rows). Or consider a vector space in which the "vectors" are real matrices that all have the same number of columns but possibly different numbers of rows. The inner product of two such matrices X and Y could then be $< X, Y >= XY^T$. Matrix valued inner products satisfy the following properties, which are natural generalizations of the scalar case.

1. linearity: let $X, X_1, X_2 \in \mathcal{V}$ have m rows and $Y, Y_1, Y_2 \in \mathcal{V}$ have n rows. Then $\forall \alpha_1, \alpha_2 \in \mathcal{R}^{k\times m}$, $\forall \beta_1, \beta_2 \in \mathcal{R}^{l\times n}$, for any k and l,
$$\begin{aligned} < \alpha_1 X_1 + \alpha_2 X_2, Y > &= \alpha_1 < X_1, Y > + \alpha_2 < X_2, Y > \\ < X, \beta_1 Y_1 + \beta_2 Y_2 > &= < X, Y_1 > \beta_1^T + < X, Y_2 > \beta_2^T \end{aligned} \tag{1.140}$$
2. symmetry: $< X, Y > = < Y, X >^T$
3. non-degeneracy: $< X, X >= \|X\|^2 \geq 0$. If $\|X\|^2 = 0$, then $X = 0$.

These properties can be verified for the two examples we cited above. Again, in the example $< X, Y >= E\, XY^T$, $\|X\|^2 = 0$ implies $X = 0$ w.p. 1. This follows from the fact that every diagonal element of $\|X\|^2$ is the norm squared of the corresponding element of X and hence the non-degeneracy of every element of X implies the non-degeneracy of X as a whole. We are now ready to show the following lemma.

Lemma 1.3 (Schur Complements) *Let X_1 and X_2 be vectors in a certain vector space with a certain inner product and denote $R_{ij} =< X_i, X_j >$, $i, j = 1, 2$ so that $R_{ij} = R_{ji}^T$. Assume that R_{11} is nonsingular. Then, because of property 3 of the inner product (non-degeneracy), we have*
$$\begin{aligned} \|X_2 - R_{21}R_{11}^{-1}X_1\|^2 &= < X_2 - R_{21}R_{11}^{-1}X_1, X_2 - R_{21}R_{11}^{-1}X_1 > \\ &= R_{22} - 2R_{21}R_{11}^{-1}R_{12} + R_{21}R_{11}^{-1}R_{11}R_{11}^{-1}R_{12} \\ &= R_{22} - R_{21}R_{11}^{-1}R_{12} \;\geq\; 0 \end{aligned} \tag{1.141}$$
with equality iff $X_2 = R_{21}R_{11}^{-1}X_1$.

The name for this lemma stems from the following congruence relation
$$\begin{aligned} \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} &= \begin{bmatrix} I & O \\ R_{21}R_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} R_{11} & 0 \\ 0 & R_{22} - R_{21}R_{11}^{-1}R_{12} \end{bmatrix} \begin{bmatrix} I & R_{11}^{-1}R_{12} \\ O & I \end{bmatrix} \\ &= \begin{bmatrix} I \\ R_{21}R_{11}^{-1} \end{bmatrix} R_{11} \begin{bmatrix} I \\ R_{21}R_{11}^{-1} \end{bmatrix}^T + \begin{bmatrix} 0 & 0 \\ 0 & R_{22} - R_{21}R_{11}^{-1}R_{12} \end{bmatrix}. \end{aligned} \tag{1.142}$$

Note that this is the first step of a procedure that can be repeated to obtain the Lower-Diagonal-Upper (LDU) triangular factorization of the matrix on the LHS . The matrix $R_{22} - R_{21}R_{11}^{-1}R_{12}$ is called the Schur complement of R_{11} within the big matrix on the LHS .

We are now ready to show the following theorem.

Theorem 1.2 (Cramer-Rao Bound (Stochastic Parameters)) *If the bias $b_{\widehat{\theta}}(\theta)$ of the estimator $\widehat{\theta}(Y)$ of θ satisfies one of the conditions (1.128) or (1.129), then the correlation matrix of the parameter estimation errors $\widetilde{\theta}$ is bounded below by the inverse of the information matrix (assuming it exists):*

$$R_{\widetilde{\theta}\widetilde{\theta}} \;=\; E\,(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \geq J^{-1} \tag{1.143}$$

with equality iff

$$\widehat{\theta}(Y)-\theta \;=\; J^{-1}\frac{\partial \ln f(\theta|Y)}{\partial\,\theta} \quad a.s. \tag{1.144}$$

An estimator that achieves the lower bound is called *efficient.*

Proof: We shall apply the lemma on Schur complements to a vector space of random vectors with inner product $< X, Y >= E\,XY^T$. In particular, we choose $X_1 = \frac{\partial \ln f(\theta|Y)}{\partial\,\theta}$ and $X_2 = \widehat{\theta}-\theta$. Since one of the conditions (1.128) or (1.129) is assumed to be satisfied, the Unit Cross Correlation lemma applies and we find

$$E\left[\begin{array}{c}\frac{\partial \ln f(\theta|Y)}{\partial\,\theta}\\ \widehat{\theta}-\theta\end{array}\right]\left[\begin{array}{c}\frac{\partial \ln f(\theta|Y)}{\partial\,\theta}\\ \widehat{\theta}-\theta\end{array}\right]^T \;=\; \left[\begin{array}{cc} J & I\\ I & R_{\widetilde{\theta}\widetilde{\theta}}\end{array}\right] \;=\; \left[\begin{array}{cc} R_{11} & R_{12}\\ R_{21} & R_{22}\end{array}\right]. \tag{1.145}$$

Then from (1.141), we have $R_{22} \geq R_{21}R_{11}^{-1}R_{12}$ which is (1.143), with equality iff $X_2 = R_{21}R_{11}^{-1}X_1$ which is (1.144). □

Remarks:

(i) In general, $R_{\widetilde{\theta}\widetilde{\theta}} = C_{\widetilde{\theta}\widetilde{\theta}} + (E\widetilde{\theta})(E\widetilde{\theta})^T$. If the bias satisfies condition (1.128) ($E\widetilde{\theta} = -E_{\boldsymbol{\theta}}b_{\widehat{\theta}}(\theta) = 0$: estimator *unbiased on the average*), then $R_{\widetilde{\theta}\widetilde{\theta}} = C_{\widetilde{\theta}\widetilde{\theta}}$, the estimation error covariance matrix.

(ii) We can investigate what the condition for equality (the condition for efficiency) implies for the posterior distribution. (1.144) can be written as

$$\frac{\partial \ln f(\theta|Y)}{\partial\,\theta} \;=\; J\widehat{\theta}(Y) - J\theta\,. \tag{1.146}$$

Integration over θ yields

$$\ln f(\theta|Y) \;=\; c(Y) + \widehat{\theta}^T(Y)\,J\,\theta - \frac{1}{2}\theta^T J\theta \;=\; c'(Y) - \frac{1}{2}(\theta-\widehat{\theta})^T J\,(\theta-\widehat{\theta}) \tag{1.147}$$

where $c(Y)$ and $c'(Y)$ are scalar functions of Y. (1.147) implies that $f(\theta|Y)$ is Gaussian. Using the constraint $\int_\Theta f(\theta|Y)\,d\theta = 1$, we can determine the proper integration constant and we get

$$f(\theta|Y) \;=\; \sqrt{\frac{\det J}{(2\pi)^m}}\,\exp\left(-\frac{1}{2}(\theta-\hat{\theta}(Y))^T J\,(\theta-\hat{\theta}(Y))\right) \tag{1.148}$$

or in other words $f(\theta|Y) \leftrightarrow \mathcal{N}(\hat{\theta}(Y), J^{-1})$. So the posterior distribution should be Gaussian with constant covariance matrix. In that case, the posterior mean (which may depend on the data) is an efficient estimator. Note that neither the prior distribution nor the conditional distribution $f(Y|\theta)$ need to be Gaussian for the posterior distribution to be Gaussian. So efficiency may occur in other than just the plain Gaussian case.

(iii) One should not confuse the estimation error correlation matrix and the covariance matrix of the posterior density:

$$\begin{aligned} R_{\widetilde{\theta}\widetilde{\theta}} &= E\,(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T \\ C_{\theta|Y} &= E_{\boldsymbol{\theta}|\mathbf{Y}}\;(\theta - E[\theta|Y])\,(\theta-E[\theta|Y])^T \end{aligned} \tag{1.149}$$

If $\hat{\theta} = \hat{\theta}_{MMSE} = E[\theta|Y]$, then we have

$$R_{\widetilde{\theta}\widetilde{\theta}} \;=\; E_{\mathbf{Y}}\,C_{\theta|Y}\;. \tag{1.150}$$

If $C_{\theta|Y}$ does not depend on Y as is the case when the posterior density is Gaussian, then we have furthermore $R_{\widetilde{\theta}\widetilde{\theta}} = E_{\mathbf{Y}}\,C_{\theta|Y} = C_{\theta|Y}$.

(iv) *Concentration ellipsoids.*
A geometrical interpretation of the Cramer-Rao lower bound exists. It involves the concentration ellipsoids that we introduced in the discussion of multivariate Gaussian distributions. For a Gaussian random vector, a concentration ellipsoid corresponds to the volume in which the random vector occurs which a certain probability. For a non-Gaussian random vector, this is not strictly true. Nevertheless, the concentration ellipsoid can still be interpreted as a measure of the spread of the random vector in space.

We shall show that the Cramer-Rao bound implies that the concentration ellipsoid for the estimation errors of any unbiased estimator lies outside or on the concentration ellipsoid of an efficient estimator. This means that the estimation errors are the most concentrated in space (around the origin) for an efficient estimator. The concentration ellipsoid for the estimation errors of any estimator is the set of points $\widetilde{\theta}$ that satisfy

$$\widetilde{\theta}^T R_{\widetilde{\theta}\widetilde{\theta}}^{-1}\widetilde{\theta} \;=\; 1 \tag{1.151}$$

while the concentration ellipsoid for an efficient estimator is

$$\widetilde{\theta}^T J\,\widetilde{\theta} \;=\; 1\;. \tag{1.152}$$

For the estimation errors in the direction of an arbitrary unit vector U we have: $\widetilde{\theta} = \|\widetilde{\theta}\|\, U$, $\widetilde{\theta}_{eff} = \|\widetilde{\theta}_{eff}\|\, U$. Using (1.151) and (1.152), this implies

$$\|\widetilde{\theta}\|^2\; U^T R_{\widetilde{\theta}\widetilde{\theta}}^{-1} U = \|\widetilde{\theta}_{eff}\|^2\; U^T J\, U = 1 \quad \Rightarrow \quad \frac{\|\widetilde{\theta}\|^2}{\|\widetilde{\theta}_{eff}\|^2} = \frac{U^T J\, U}{U^T R_{\widetilde{\theta}\widetilde{\theta}}^{-1} U} \geq 1\,. \tag{1.153}$$

This last inequality follows from the CRB, i.e.

$$R_{\widetilde{\theta}\widetilde{\theta}} \geq J^{-1} > 0 \quad \Rightarrow \quad 0 < R_{\widetilde{\theta}\widetilde{\theta}}^{-1} \leq J \quad \Rightarrow \quad 0 < U^T R_{\widetilde{\theta}\widetilde{\theta}}^{-1} U \leq U^T J\, U \tag{1.154}$$

where we assumed that the information matrix J is non-singular. $\|\widetilde{\theta}_{eff}\| \leq \|\widetilde{\theta}\|$ means that one ellipsoid is circumscribed by the other one. If $R_{\widetilde{\theta}\widetilde{\theta}} > J$ then the efficient ellipsoid is strictly inside the other one, as shown in Fig. 1.9 for the case $m = 2$ (in which case the ellipsoids are ellipses). If on the other hand $\mathrm{rank}(R_{\widetilde{\theta}\widetilde{\theta}} - J^{-1}) = r < m$, then $\mathrm{rank}(J - R_{\widetilde{\theta}\widetilde{\theta}}^{-1}) = r$, which implies that $U^T \left(J - R_{\widetilde{\theta}\widetilde{\theta}}^{-1}\right) U = 0$ for U in some $(m-r)$-dimensional subspace. This means that the two m-dimensional ellipsoids coincide over an $(m-r)$-dimensional ellipsoid.

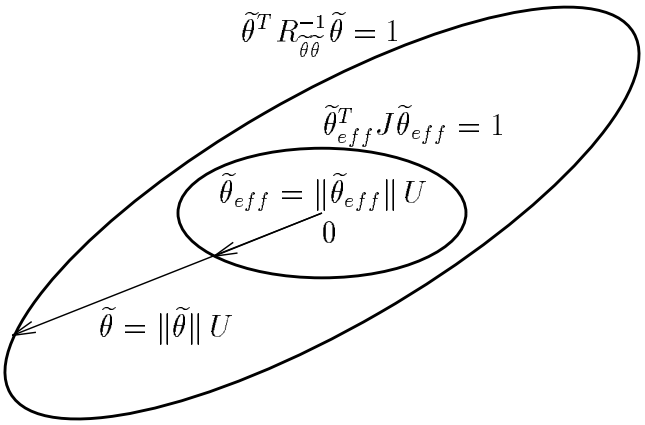


Figure 1.9: The Cramer-Rao bound in terms of concentration ellipses.

(v) *Additivity of the information matrix.*
Using Bayes' rule, we can write

$$\begin{aligned} J &= -E\,\tfrac{\partial}{\partial\theta}\left(\tfrac{\partial\, \ln f(\theta|Y)}{\partial\theta}\right)^T \\ &= -E\,\tfrac{\partial}{\partial\theta}\left(\tfrac{\partial\, \ln f(\theta)}{\partial\theta}\right)^T - E\,\tfrac{\partial}{\partial\theta}\left(\tfrac{\partial\, \ln f(Y|\theta)}{\partial\theta}\right)^T = J_{prior} + J_Y\,. \end{aligned} \tag{1.155}$$

Hence the total information J is the sum of information J_{prior} obtained from the prior distribution and information J_Y obtained from the data. If the distribution $f(\theta)$ is Gaussian then the corresponding information matrix is the inverse of the covariance matrix. If the different data y_i are independent given θ, then

$$f(Y|\theta) = \prod_{i=1}^{n} f(y_i|\theta) \quad \Rightarrow \quad J_Y = \sum_{i=1}^{n} J_{y_i}\,. \tag{1.156}$$

If furthermore the data are i.i.d. given θ, then $J_Y = n J_{y_1}$.

(vi) The importance of the Cramer-Rao bound stems primarily from the fact that it reveals how the correlation matrix of the estimation errors is bounded below in terms of the information we get from the prior distribution and the data. Another potential use of the Cramer-Rao bound goes as follows. The Cramer-Rao bound evaluates the MSE criterion and the MSE is indeed probably the most important performance criterion. Since the MMSE estimator minimizes this criterion, it is often the preferred estimator. However, the MMSE estimator is the posterior mean which may often be difficult to obtain. The MAP estimator on the other hand can be obtained more easily. In particular, it does not require the computation of the posterior distribution $f(\theta|Y)$ and its derivation can proceed directly from the given prior and conditional distributions $f(\theta)$ and $f(Y|\theta)$. It is not much extra work to verify if the equality condition (1.144) for efficiency is satisfied. If it is found that the MAP estimator achieves efficiency, then it equals the MMSE estimator since the MMSE estimator minimizes the MSE criterion but the minimum value of the MSE criterion cannot be lower than J^{-1}. So this approach allows one to verify if the MMSE estimate can be obtained in this indirect but simpler way.

Example 1.6 Gaussian mean in Gaussian noise - Example 1.4 Continued

In example 1.4, the posterior distribution was Gaussian with constant variance (see (1.75)). We had $\widehat{\theta}_{MMSE} = \widehat{\theta}_{MAP} = \widehat{\theta}_{ABS}$ which is hence an efficient estimator. We found indeed for the estimation error correlation

$$\frac{1}{\sigma^2_{\widehat{\theta}}} \;=\; J \;=\; J_{prior} + J_Y \;=\; \frac{1}{\sigma^2_\theta} + \frac{n}{\sigma^2_v} \tag{1.157}$$

which decomposes indeed into the prior information and n times the information in one measurement (all distributions involved are Gaussian). $\diamond$

1.3.9 Linear MMSE Estimation

So the MMSE criterion is a desirable criterion but the resulting Bayes estimator $E[\theta|Y]$, the mean of the posterior distribution $f(\theta|Y)$, may be complicated to derive. In practice, one often resorts to suboptimal estimators. Perhaps the most natural choice for a suboptimal estimator is to restrict the estimator to be a linear function of the data. Before proceeding, remark that the MMSE criterion for a vector parameter decomposes:

$$\min_{\widehat{\theta}} \; E\,(\widehat{\theta}-\theta)^T(\widehat{\theta}-\theta) \;=\; \min_{\widehat{\theta}} \; E\sum_{i=1}^{m}(\widehat{\theta}_i-\theta_i)^2 \;=\; \sum_{i=1}^{m}\min_{\widehat{\theta}_i} \; E\,(\widehat{\theta}_i-\theta_i)^2\,. \tag{1.158}$$

Hence it suffices to concentrate on the estimator $\widehat{\theta}_i$ of a scalar component θ_i of θ and then stack the scalar estimators in a vector to obtain $\widehat{\theta} = \left[\widehat{\theta}_1 \cdots \widehat{\theta}_m\right]^T$.

Linear MMSE Estimators

So we shall constrain $\widehat{\theta}_i(Y)$ to be linear:

$$\widehat{\theta}_i(Y) \;=\; H_i^T Y \tag{1.159}$$

where H_i is a $n \times 1$ vector of parameters (of the same dimension as the data vector Y). The MSE risk function is

$$\mathcal{R}_{LMMSE}(H_i) \;=\; E\,(\theta_i - \widehat{\theta}_i)^2 \;=\; E\,(\theta_i - H_i^T Y)^2 \tag{1.160}$$

and we shall obtain H_i as $H_i = \arg\min_{H_i} \mathcal{R}_{LMMSE}(H_i)$. Setting the gradient equal to zero leads to

$$\frac{\partial}{\partial\, H_i}\mathcal{R}_{LMMSE}(H_i) \;=\; -2E\,(\theta_i - H_i^T Y)Y \;=\; 0 \quad\Rightarrow\quad H_i^T = (E\,\theta_i Y)(E\,YY)^{-1} = R_{\theta_i Y} R_{YY}^{-1}\;. \tag{1.161}$$

The Hessian can be verified to be

$$\frac{\partial}{\partial\, H_i}\left(\frac{\partial}{\partial\, H_i}\mathcal{R}_{LMMSE}(H_i)\right)^T \;=\; 2\,R_{YY} \;>\; 0\;. \tag{1.162}$$

Hence, the unique extremum we found in (1.161) is indeed the global minimum. This can be done for each component θ_i of θ independently to yield the minimizer of the overall MSE criterion (1.158). We get

$$\widehat{\theta} \;=\; \begin{bmatrix} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_m \end{bmatrix} \;=\; \begin{bmatrix} H_1^T Y \\ \vdots \\ H_m^T Y \end{bmatrix} \;=\; H\,Y \;=\; R_{\theta Y} R_{YY}^{-1} Y \;=\; \begin{bmatrix} R_{\theta_1 Y} R_{YY}^{-1} Y \\ \vdots \\ R_{\theta_m Y} R_{YY}^{-1} Y \end{bmatrix} \tag{1.163}$$

where $H = [H_1 \cdots H_m]^T$. We can investigate the correlation matrix of the parameter estimation errors:

$$\begin{aligned} R_{\widetilde{\theta}\widetilde{\theta}}(H) \;&=\; E\,(\theta - \widehat{\theta})(\theta - \widehat{\theta})^T \;=\; E\,(\theta - HY)(\theta^T - Y^T H^T) \\ &=\; R_{\theta\theta} - R_{\theta Y} H^T - H\,R_{Y\theta} + H\,R_{YY} H^T\;. \end{aligned} \tag{1.164}$$

Evaluated at the minimum, this gives

$$R_{\widetilde{\theta}\widetilde{\theta}}^{LMMSE} \;=\; R_{\widetilde{\theta}\widetilde{\theta}}(R_{\theta Y} R_{YY}^{-1}) \;=\; R_{\theta\theta} - R_{\theta Y} R_{YY}^{-1} R_{Y\theta}\;. \tag{1.165}$$

Note that the MSE criterion is just the trace of this correlation matrix, hence

$$\min_H E\,(\theta - HY)^T(\theta - HY) \;=\; \mathrm{tr}\left\{R_{\widetilde{\theta}\widetilde{\theta}}^{LMMSE}\right\} \;=\; \mathrm{tr}\left\{R_{\theta\theta} - R_{\theta Y} R_{YY}^{-1} R_{Y\theta}\right\}\;. \tag{1.166}$$

Affine MMSE Estimators

If the random variables involved have non-zero mean, it may be advantageous to include a constant term in the estimator. So consider a $\widehat{\theta}_i(Y)$ that is affine:

$$\widehat{\theta}_i(Y) \;=\; H_i^T Y + g_i \tag{1.167}$$

where g_i is a scalar. The MSE risk function is now

$$\mathcal{R}_{AMMSE}(H_i, g_i) \;=\; E\,(\theta_i - \widehat{\theta}_i)^2 \;=\; E\,(\theta_i - H_i^T Y - g_i)^2 \tag{1.168}$$

and we shall obtain H_i and g_i from the optimization problem $\min_{H_i,g_i} \mathcal{R}_{AMMSE}(H_i,g_i)$. Setting the gradients equal to zero leads to

$$\begin{array}{rcccl|l} \frac{\partial}{\partial H_i}\mathcal{R}_{AMMSE}(H_i,g_i) & = & 0 & = & -2E\,(\theta_i - H_i^T Y - g_i)Y & 1 \\ \frac{\partial}{\partial g_i}\mathcal{R}_{AMMSE}(H_i,g_i) & = & 0 & = & -2E\,(\theta_i - H_i^T Y - g_i) & -m_Y \end{array} \tag{1.169}$$

From the second equation, we get

$$g_i \;=\; m_{\theta_i} - H_i^T m_Y \;. \tag{1.170}$$

By forming a linear combination of both equations, as indicated in (1.169), we get furthermore

$$\begin{aligned} 0 \;&=\; E\,(\theta_i - H_i^T Y - g_i)(Y - m_Y) \;=\; E\,(\theta_i - m_{\theta_i} - H_i^T(Y - m_Y))(Y - m_Y) \\ &=\; E\,(Y - m_Y)(\theta_i - m_{\theta_i} - (Y - m_Y)^T H_i) \\ &=\; E\,\{(Y - m_Y)(\theta_i - m_{\theta_i})\} \;+\; E\left\{(Y - m_Y)(Y - m_Y)^T\right\} H_i \;=\; C_{Y\theta_i} - C_{YY} H_i \end{aligned} \tag{1.171}$$

which leads to

$$\widehat{\theta}_i(Y) \;=\; H_i^T Y + g_i \;=\; m_{\theta_i} + C_{\theta_i Y} C_{YY}^{-1}(Y - m_Y) \;. \tag{1.172}$$

The Hessian can be verified to be

$$\begin{aligned} &\left[\begin{array}{cc} \frac{\partial}{\partial H_i}\left(\frac{\partial}{\partial H_i}\mathcal{R}_{AMMSE}\right)^T & \frac{\partial}{\partial H_i}\left(\frac{\partial}{\partial g_i}\mathcal{R}_{AMMSE}\right)^T \\ \frac{\partial}{\partial g_i}\left(\frac{\partial}{\partial H_i}\mathcal{R}_{AMMSE}\right)^T & \frac{\partial}{\partial g_i}\left(\frac{\partial}{\partial g_i}\mathcal{R}_{AMMSE}\right)^T \end{array}\right]\Bigg|_{\left\{\begin{array}{l} H_i = C_{\theta_i Y} C_{YY}^{-1} \\ g_i = m_{\theta_i} - C_{\theta_i Y} C_{YY}^{-1} m_Y \end{array}\right.} \\ &=\; 2\left[\begin{array}{cc} R_{YY} & m_Y \\ m_Y^T & 1 \end{array}\right] \;=\; 2\left[\begin{array}{cc} C_{YY} & 0 \\ 0 & 0 \end{array}\right] + 2\left[\begin{array}{c} m_Y \\ 1 \end{array}\right]\left[\begin{array}{c} m_Y \\ 1 \end{array}\right]^T \\ &=\; 2\left[\begin{array}{cc} I & m_Y \\ 0 & 1 \end{array}\right]\left[\begin{array}{cc} C_{YY} & 0 \\ 0 & 1 \end{array}\right]\left[\begin{array}{cc} I & m_Y \\ 0 & 1 \end{array}\right]^T \;>\; 0 \end{aligned} \tag{1.173}$$

since the matrix in the middle is positive definite, a triangular matrix with unit diagonal is nonsingular, and a congruence transformation of a positive definite matrix with a nonsingular matrix preserves positive definiteness. Hence, the unique extremum we found from (1.169) is indeed the global minimum. We also used the property that for X and Y random vectors

$$R_{XY} \;=\; C_{XY} + m_X\, m_Y^T \;. \tag{1.174}$$

This can again be done for each component θ_i of θ independently to yield the minimizer of the overall MSE criterion (1.158). We get

$$\widehat{\theta} \;=\; \left[\begin{array}{c} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_m \end{array}\right] \;=\; \left[\begin{array}{c} H_1^T Y + g_1 \\ \vdots \\ H_m^T Y + g_m \end{array}\right] \;=\; H\,Y \;+\; g \;=\; m_\theta + C_{\theta Y} C_{YY}^{-1}(Y - m_Y) \tag{1.175}$$

where $H = [H_1 \cdots H_M]^T$ and $g = [g_1 \cdots g_m]^T$. The affine estimator is unbiased: $E\,\widehat{\theta} \;=\; m_\theta$ or $E\,\widetilde{\theta} = 0$. We can again investigate the correlation matrix of the parameter estimation

errors. Evaluated for the optimal estimator (1.175), this yields

$$\begin{aligned} R_{\widetilde{\theta}\widetilde{\theta}}^{AMMSE} &= E\,(\theta-\widehat{\theta})(\theta-\widehat{\theta})^T \\ &= E\,[\theta - m_\theta - C_{\theta Y}C_{YY}^{-1}(Y-m_Y)][\theta - m_\theta - C_{\theta Y}C_{YY}^{-1}(Y-m_Y)]^T \\ &= C_{\theta\theta} - C_{\theta Y}C_{YY}^{-1}C_{Y\theta} \;=\; C_{\widetilde{\theta}\widetilde{\theta}}^{AMMSE}\,. \end{aligned} \tag{1.176}$$

Note that the MSE criterion is just the trace of this correlation matrix, hence

$$\min_{H,g} E\,(\theta - HY - g)^T(\theta - HY - g) \;=\; \mathrm{tr}\left\{C_{\widetilde{\theta}\widetilde{\theta}}^{AMMSE}\right\} \;=\; \mathrm{tr}\left\{C_{\theta\theta} - C_{\theta Y}C_{YY}^{-1}C_{Y\theta}\right\}\,. \tag{1.177}$$

Remarks:

(i) From section 1.1, we know that when θ and Y are jointly Gaussian, then

$$E\;[\theta|Y] \;=\; m_\theta + C_{\theta Y}C_{YY}^{-1}(Y-m_Y)\,. \tag{1.178}$$

Hence, in the case of Gaussian variables, the Affine MMSE estimator is optimal!

(ii) Whereas general Bayes estimators require the knowledge of the complete joint distribution $f(Y,\theta)$, the Linear and Affine MMSE estimators only require the joint first and second order moments of θ and Y.

(iii) If θ and Y have non-zero means, it is advantageous to use an affine estimator. Indeed, using $R_{XY} = C_{XY} + m_X m_Y^T$ and the Matrix Inversion Lemma (lemma 1.1) on $R_{YY}^{-1} = (C_{YY} + m_Y m_Y^T)^{-1}$, one can show from (1.165) and (1.176) that

$$\begin{aligned} R_{\widetilde{\theta}\widetilde{\theta}}^{LMMSE} &= R_{\widetilde{\theta}\widetilde{\theta}}^{AMMSE} + \underbrace{(m_\theta - C_{\theta Y}C_{YY}^{-1}m_Y)(m_Y^TC_{YY}^{-1}m_Y+1)^{-1}(m_\theta - C_{\theta Y}C_{YY}^{-1}m_Y)^T}_{\geq 0} \\ &\geq R_{\widetilde{\theta}\widetilde{\theta}}^{AMMSE} \;=\; C_{\widetilde{\theta}\widetilde{\theta}}^{AMMSE}\,. \end{aligned} \tag{1.179}$$

By taking the trace of these expressions, one finds that the affine estimator leads to lower MSE than the linear estimator, in the case of non-zero means.

(iv) Since the form of the Linear MMSE estimator is simpler than the form of the Affine MMSE estimator, we shall seek to reduce the case of Affine MMSE estimation to the case of Linear MMSE estimation. One way to achieve this is to introduce a linear estimator for an augmented problem as

$$\theta' = \theta\,,\; Y' = \begin{bmatrix} Y \\ 1 \end{bmatrix}\,,\quad \widehat{\theta}_L' \;=\; H'Y' \;=\; [H\;g]\begin{bmatrix} Y \\ 1 \end{bmatrix} \;=\; HY + g \;=\; \widehat{\theta}_A\,. \tag{1.180}$$

In problem 1.7, it is asked to show that the Linear MMSE estimator for the augmented problem $\{\theta, Y'\}$

$$\widehat{\theta}_{LMMSE}' \;=\; R_{\theta Y'}R_{Y'Y'}^{-1}Y' \;=\; m_\theta + C_{\theta Y}C_{YY}^{-1}(Y-m_Y) \;=\; \widehat{\theta}_{AMMSE} \tag{1.181}$$

is in fact the Affine MMSE estimator for the original problem $\{\theta, Y\}$.

(v) Another, and more popular, way to reduce the affine case to the linear case goes as follows. Note that in the case of zero means, $m_\theta = 0$ and $m_Y = 0$, the affine estimator reduces to the linear estimator. This encourages us to centralize the variables before further treatment so that effectively we work with

$$\left\{ \begin{array}{rcl} \theta' & = & \theta - m_\theta \\ Y' & = & Y - m_Y \ . \end{array} \right. \tag{1.182}$$

Then the linear and the affine MMSE estimators coincide

$$\widehat{\theta'} \ = \ R_{\theta' Y'} R_{Y'Y'}^{-1} Y' \ = \ C_{\theta' Y'} C_{Y'Y'}^{-1} Y' \ = \ C_{\theta Y} C_{YY}^{-1} Y' \ . \tag{1.183}$$

Unless the contrary is explicitly stated, we shall henceforth assume that the random variables are zero mean. This will allow us to use the expressions for the Linear MMSE estimator to handle the affine case at the same time.

(vi) Except in the jointly Gaussian case, the MSE increases by imposing the constraint of linearity on the estimator. This increase can be displayed by decomposing the MSE associated with a linear estimator as follows:

$$\begin{aligned} & E(\theta - HY)^T(\theta - HY) \\ & = \ E(\theta - E[\theta|Y] + E[\theta|Y] - HY)^T(\theta - E[\theta|Y] + E[\theta|Y] - HY) \\ & = \ E(\theta - E[\theta|Y])^T(\theta - E[\theta|Y]) + E(E[\theta|Y] - HY)^T(E[\theta|Y] - HY) \\ & \quad +2 \underbrace{E(\theta - E[\theta|Y])^T(E[\theta|Y] - HY)}_{=0} \\ & = \ E(\theta - E[\theta|Y])^T(\theta - E[\theta|Y]) + \underbrace{E(E[\theta|Y] - HY)^T(E[\theta|Y] - HY)}_{\geq 0} \\ & \geq \ E(\theta - E[\theta|Y])^T(\theta - E[\theta|Y]) \end{aligned} \tag{1.184}$$

where the cross-terms disappear due to the orthogonality of $\theta - E[\theta|Y]$ on any nonlinear function of Y such as $E[\theta|Y] - HY$ in this case. This shows that the difference between the LMMSE and the MMSE is the MSE in approximating the conditional mean $E[\theta|Y]$ by a linear function HY. In fact, the best linear approximation of $E[\theta|Y]$ is also the best linear approximation of θ since (1.184) implies

$$R_{\theta Y} R_{YY}^{-1} \ = \ \arg\min_H E(\theta - HY)^T(\theta - HY) \ = \ \arg\min_H E_Y(E[\theta|Y] - HY)^T(E[\theta|Y] - HY) \ . \tag{1.185}$$

Geometric Interpretations and the Orthogonality Property of LMMSE

Assume for a moment that θ and y are vectors in $\mathcal{R}^2$, or that θ, y_1 and y_2 are vectors in $\mathcal{R}^3$ (see Fig. 1.10). Then we know that the best approximation of θ (yielding the shortest error vector, $\|\widetilde{\theta}\|$ smallest) in terms of y in 2D (y_1 and y_2 in 3D) is obtained by taking the orthogonal projection of θ onto the line passing through y in 2D (plane spanned by y_1 and y_2

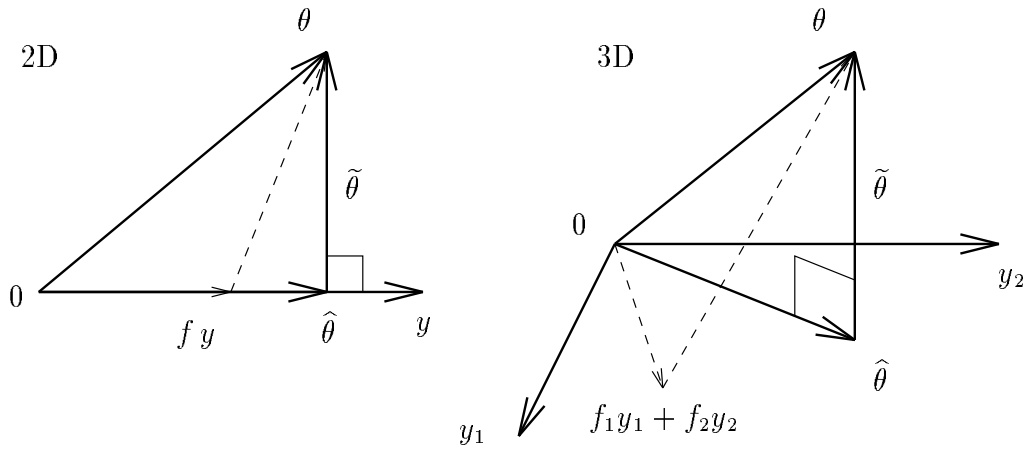


Figure 1.10: The orthogonality property of LMMSE in 2D and 3D Euclidean space.

in 3D). The resulting error vector $\tilde{\theta} = \theta - \hat{\theta}$ satisfies $< \tilde{\theta}, y >= 0$ in 2D ($< \tilde{\theta}, y_i >= 0,\ i = 1, 2$ in 3D).

This geometrical result from Euclidean vector spaces carries over to our estimation problem in a vector space of random variables. Let $\theta, y_1, \ldots, y_n$ be random variables that span a $(n+1)$-dimensional vector space. We take the inner product in this vector space to be $< x, y >= E\, xy$. From our previous discussion on inner products, we know this is a valid inner product. We shall form a linear estimate (approximation) of θ in terms of $y_1, \ldots, y_n$:

$$\hat{\theta} \;=\; h^T Y \;=\; \sum_{i=1}^{n} h_i\, y_i \;. \tag{1.186}$$

We shall determine the combination coefficients h_i by minimizing the MSE

$$\min_{h_i} E\,(\theta - \hat{\theta})^2 \;=\; \min_{h_i} E\,(\theta - \sum_{i=1}^{n} h_i\, y_i)^2 \;=\; \min_{h_i} \left\| \theta - \hat{\theta} \right\|^2 \;. \tag{1.187}$$

We know from before that the minimum is obtained by setting the gradient equal to zero. So the optimal coefficients h_i satisfy

$$\frac{\partial}{\partial\, h_i} E\,(\theta - \hat{\theta})^2 \;=\; -2\, E\,(\theta - \hat{\theta}) y_i \;=\; 0 \quad \Rightarrow \quad < \theta - \hat{\theta},\, y_i >= 0\;, \quad i = 1, \ldots, n\;. \tag{1.188}$$

Hence the LMMSE estimate also satisfies the orthogonality conditions.

When $\theta = [\theta_1 \cdots \theta_m]^T$ is a random vector, then we can again consider the space spanned by the variables $\theta_1, \ldots, \theta_m, y_1, \ldots, y_n$. We shall now more generally consider vectors of random variables and the matrix inner product between them. If X and Y are random vectors, then we take their inner product to be $< X, Y >= E\, XY^T$. We have discussed before the validity of this matrix valued inner product. A linear estimator for θ in terms of Y is now

$$\hat{\theta} \;=\; H\, Y \tag{1.189}$$

where H is a $m \times n$ matrix of combination coefficients. We cannot follow directly the same path as for the case of a scalar θ since we have not seen how to take gradients w.r.t. a matrix

(and even less how to consider the corresponding Hessian). However, we shall rely on our geometrical intuition to claim that the optimal H is the one that satisfies the orthogonality condition:

$$0 = <\theta - \hat{\theta}, Y> = <\theta - HY, Y> = <\theta, Y> - H <Y, Y> \tag{1.190}$$

or hence

$$H = <\theta, Y><Y, Y>^{-1} = R_{\theta Y} R_{YY}^{-1} \tag{1.191}$$

which is indeed the result we found before using a different route. We can now show that this H which satisfies the orthogonality condition (1.190) minimizes the correlation matrix $R_{\tilde{\theta}\tilde{\theta}}$ of the estimation errors. Indeed, let KY be any other linear estimator of θ. Then

$$\begin{aligned} R_{\tilde{\theta}\tilde{\theta}}(K) &= \|\theta - KY\|^2 = <\theta - KY, \theta - KY> \\ &= <\theta - HY + HY - KY, \theta - HY + HY - KY> \\ &= \|\theta - HY\|^2 + \|(H-K)Y\|^2 \\ &\quad + \underbrace{<\theta - HY, Y>}_{=0}(H-K)^T + (H-K)\underbrace{<Y, \theta - HY>}_{=0} \\ &= \|\theta - HY\|^2 + \underbrace{(H-K)\|Y\|^2(H-K)^T}_{\geq 0 \ (=0 \text{ iff } H=K \ (\|Y\|^2 = R_{YY} > 0))} \\ &\geq \|\theta - HY\|^2 = R_{\tilde{\theta}\tilde{\theta}}(H) \, . \end{aligned} \tag{1.192}$$

Since the MSE is the trace of $R_{\tilde{\theta}\tilde{\theta}}$, this shows again that $\hat{\theta} = R_{\theta Y} R_{YY}^{-1} Y$ is the LMMSE estimator, with a demonstration based on the orthogonality property.

1.3.10 The Bayesian Linear Model

We have seen in section 1.1.1 that when θ and Y are jointly Gaussian, then it is always possible to interpret the stochastic relation between them in the form of a linear model as in (1.34). Note that the estimation paradigm we have considered so far, as illustrated in Fig. 1.4, is one in which the parameters are the main quantities of interest. We first have the parameters θ, though they are unaccessible to us, and then we measure the data Y whose only purpose is to reveal information about θ. The observation of jointly Gaussian θ and Y leading to a linear model fits in this paradigm. However, in the linear model discussion, the paradigm is mainly one of signal estimation. In this case, the observations $\{y_k\}$ are the main quantities of interest. They are considered to be noisy samples of the actual signal of interest. In the linear model case, this signal of interest is assumed to be a linear combination of known basis functions. So the measurements are modeled as

$$y_k = s_k + v_k = \sum_{i=1}^{m} \theta_i \, h_i(k) + v_k = c_k^T \theta + v_k \tag{1.193}$$

where $c_k^T = [h_1(k) \cdots h_m(k)]$. The linear combination coefficients θ_i are the parameters and the basis functions $h_i(k)$ are known. A signal model that is used very often is the solution of

a homogenous difference equation with constant coefficients. In the general case, in which the roots of the characteristic equation may be repeated, the solution of such a difference equation can be written as

$$\begin{aligned} s_k &= \sum_{i=1}^{m_0} \Big(\sum_{j=1}^{m_i} \alpha_{ij}\, k^{j-1} \Big)\, \lambda_i^k \;=\; c_k^T \theta \\ c_k^T &= \Big[k^0 \lambda_1^k \cdots k^{m_1-1} \lambda_1^k \quad k^0 \lambda_2^k \cdots k^{m_{m_0}-1} \lambda_{m_0}^k \Big] \\ \theta^T &= \Big[\alpha_{1\,1} \cdots \alpha_{1\,m_1} \quad \alpha_{2\,1} \cdots \alpha_{m_0\,m_{m_0}} \Big] \end{aligned} \tag{1.194}$$

in which λ_i is a root that occurs with multiplicity m_i and there are m_0 distinct roots (if the signal is to be real, the complex roots occur in complex conjugate pairs and the corresponding coefficients α also). The signal s_k in (1.194) is the solution of the following difference equation

$$\prod_{i=1}^{m_0} (1 - \lambda_i\, q^{-1})^{m_i}\, s_k \;=\; 0 \tag{1.195}$$

where q^{-1} is the delay operator: $q^{-1} s_k = s_{k-1}$ (q^{-1} transforms to a multiplication by z^{-1} when taking the z-transform). The total order of the difference equation is $m = \sum_{i=1}^{m_0} m_i$.

One important particular case occurs when $m_0 = 1$ and $\lambda_1 = 1$. In this case, s_k is a polynomial function of k. In particular, when furthermore $m_1 = 1$, then $s_k = \alpha_{11}$ is a constant. We have treated this case in example 1.4. Another important case is when m_0 is even, the λ_i are on the unit circle ($\lambda_i = e^{j\omega_i}$) and occur in complex conjugate pairs, and $m_i = 1$, $i \geq 1$. In this case, it is useful to do the following reparameterization:

$$s_k \;=\; \sum_{i=1}^{m_0/2} \left(\alpha_i\, e^{j\omega_i k} + \alpha_i^*\, e^{-j\omega_i k} \right) \;=\; \sum_{i=1}^{m_0/2} \left(a_i\, \cos(\omega_i k) + b_i\, \sin(\omega_i k) \right) \;. \tag{1.196}$$

The case of one sinusoid ($m_0/2 = 1$) is treated in problem 1.3.

We can stack the data from (1.193) into a vector $Y = [y_1 \cdots y_n]^T$ to obtain

$$Y \;=\; H\,\theta + V \tag{1.197}$$

where $H = [c_1 \cdots c_n]^T$ is a known $n \times m$ matrix that should not be confused with the H matrix we used in the expression for LMMSE estimators. The Bayesian linear model assumes, apart from a linear model structure, also Gaussian distributions of the form

$$\theta = [\theta_1 \cdots \theta_m]^T \sim \mathcal{N}(m_\theta, C_{\theta\theta}) \;\;, \quad V = [v_1 \cdots v_n]^T \sim \mathcal{N}(0, C_{VV}) \;, \quad \theta \text{ and } V \text{ are independent.} \tag{1.198}$$

Hence $f(\theta, V) = f(\theta) f(V)$, which shows that θ and V are jointly Gaussian. Using (1.197), we can write

$$\begin{bmatrix} \theta \\ Y \end{bmatrix} \;=\; \begin{bmatrix} I & 0 \\ H & I \end{bmatrix} \begin{bmatrix} \theta \\ V \end{bmatrix} \;. \tag{1.199}$$

Since θ and V are jointly Gaussian and $[\theta^T \; Y^T]^T$ is obtained as a linear transformation of $[\theta^T \; V^T]^T$, also θ and Y are jointly Gaussian (see problem 1.1). By taking the expected value of both sides of (1.197), we find $m_Y = H\, m_\theta$, and by subtracting this from (1.197), we obtain

$$Y - m_Y \;=\; H\,(\theta - m_\theta) + V \;. \tag{1.200}$$

From (1.200), we obtain immediately

$$\begin{aligned} C_{YY} &= H\,C_{\theta\theta}\,H^T + C_{VV} \\ C_{Y\theta} &= H\,C_{\theta\theta}\ . \end{aligned} \tag{1.201}$$

Using the Gauss-Markov theorem (theorem 1.1), we find

$$\begin{aligned} f(\theta|Y) &\leftrightarrow \mathcal{N}\left(m_{\theta|Y},\, C_{\theta|Y}\right) \\ m_{\theta|Y} &= m_\theta + C_{\theta\theta}H^T\left(H\,C_{\theta\theta}\,H^T + C_{VV}\right)^{-1}(Y - H\,m_\theta) \\ C_{\theta|Y} &= C_{\theta\theta} - C_{\theta\theta}H^T\left(H\,C_{\theta\theta}\,H^T + C_{VV}\right)^{-1}H\,C_{\theta\theta} \end{aligned} \tag{1.202}$$

Since θ and Y are jointly Gaussian, the MMSE estimator and the AMMSE estimator are equal, and they equal $\widehat{\theta}(Y) = m_{\theta|Y}$. Using the Matrix Inversion Lemma (lemma 1.1) on the expression for C_{YY}, one can show (parallel to the development in example 1.4) that $m_{\theta|Y}$ can be written also as

$$m_{\theta|Y} \;=\; m_\theta + \left(C_{\theta\theta}^{-1} + H^T\,C_{VV}^{-1}\,H\right)^{-1}H^T\,C_{VV}^{-1}\,(Y - H\,m_\theta)\ . \tag{1.203}$$

Using furthermore the Matrix Inversion Lemma on the expression for $C_{\theta|Y}$, one can also show

$$C_{\theta|Y}^{-1} \;=\; C_{\theta\theta}^{-1} \;+\; H^T\,C_{VV}^{-1}\,H\ . \tag{1.204}$$

Since $f(\theta|Y)$ is Gaussian, the (A)MMSE estimator is efficient and $R_{\widetilde{\theta\theta}} = C_{\theta|Y} = J^{-1}$. (1.204) is in fact nothing else but the decomposition of the information in prior information and information coming from the data: $J = J_{prior} + J_Y$. If furthermore the noise samples are uncorrelated so that C_{VV} is diagonal: $C_{VV} = \mathrm{diag}\{\sigma_{v_1}^2 \cdots \sigma_{v_n}^2\}$, then the information from the data also decomposes

$$J_Y \;=\; H^T\,C_{VV}^{-1}\,H \;=\; \sum_{k=1}^{n} c_k\,\sigma_{v_k}^{-2}\,c_k^T\ . \tag{1.205}$$

Note that for $n < m$ J_Y has rank (at most) equal to n and hence is singular. However, due to $J_{prior} > 0$, J becomes positive definite and hence non-singular. This illustrates the importance of prior information when only little measurement data is available. On the other hand, as $n \to \infty$, J_{prior} becomes of negligible importance compared to J_Y. So in general, the influence of the prior information disappears asymptotically as the number of measurements becomes large.

1.3.11 Recap: Bayesian parameter estimation

In the Bayesian context, an estimator $\widehat{\theta}(.)$ is obtained by minimizing a risk function, an average cost function:

$$\min_{\widehat{\theta}(.)} \mathcal{R}(\widehat{\theta}(.)) = E_{\mathbf{Y}}\left[\min_{\widehat{\theta}(Y)} E_{\boldsymbol{\theta}|\mathbf{Y}}\mathcal{C}(\theta, \widehat{\theta}(Y))\right] = E_{\mathbf{Y}}\left[\min_{\widehat{\theta}(Y)} \mathcal{R}(\widehat{\theta}(Y)|Y)\right]\ . \tag{1.206}$$

The minimization of the conditional risk $\mathcal{R}(\widehat{\theta}(Y)|Y)$ requires the posterior distribution $f_{\boldsymbol{\theta}|\mathbf{Y}}(\theta|Y) = f_{\mathbf{Y}|\boldsymbol{\theta}}(Y|\theta)f_{\boldsymbol{\theta}}(\theta)/f_{\mathbf{Y}}(Y)$ which using Bayes' identity can be expressed in terms of the prior distribution and the conditional distribution of the measurements. Two important special cases:

- quadratic cost function (risk=MSE): $\widehat{\theta}_{MMSE}(Y) = E(\theta|Y)$
- uniform cost function : $\widehat{\theta}_{MAP}(Y) = \arg\max_{\theta} f(\theta|Y) = \arg\max_{\theta} f(Y|\theta)f(\theta)$

In practice, the most popular performance measure of the parameter estimation error $\widetilde{\theta} = \theta - \widehat{\theta}$ is the

$$\text{MSE } = E_{\boldsymbol{\theta},\mathbf{Y}}\|\widetilde{\theta}\|^2 = E_{\boldsymbol{\theta},\mathbf{Y}}\widetilde{\theta}^T\widetilde{\theta} \; . \tag{1.207}$$

A more complete description of the estimation errors is their correlation matrix

$$R_{\widetilde{\theta}\widetilde{\theta}} = E_{\boldsymbol{\theta},\mathbf{Y}}\widetilde{\theta}\widetilde{\theta}^T = C_{\widetilde{\theta}\widetilde{\theta}} + (E_{\boldsymbol{\theta},\mathbf{Y}}\widetilde{\theta})(E_{\boldsymbol{\theta},\mathbf{Y}}\widetilde{\theta})^T = C_{\widetilde{\theta}\widetilde{\theta}} + (E_{\boldsymbol{\theta}}\, b_{\widehat{\theta}}(\theta))(E_{\boldsymbol{\theta}}\, b_{\widehat{\theta}}(\theta))^T \; . \tag{1.208}$$

where $C_{\widetilde{\theta}\widetilde{\theta}}$ is the error covariance matrix and $E_{\boldsymbol{\theta}}\, b_{\widehat{\theta}}(\theta)$ is the average bias. The (conditional) bias is $b_{\widehat{\theta}}(\theta) = -E_{\mathbf{Y}|\boldsymbol{\theta}}\widetilde{\theta} = E_{\mathbf{Y}|\boldsymbol{\theta}}\widehat{\theta} \; - \theta$. The MSE summarizes the error correlation matrix in one number:

$$\text{MSE } = \text{tr}\left\{R_{\widetilde{\theta}\widetilde{\theta}}\right\} \; . \tag{1.209}$$

The MMSE estimator minimizes both the MSE and $R_{\widetilde{\theta}\widetilde{\theta}}$. An estimator-independent lower bound for $R_{\widetilde{\theta}\widetilde{\theta}}$ exists. To this end, consider the Fisher information matrix

$$\begin{aligned} J &= E\left(\tfrac{\partial \ln f(\theta|Y)}{\partial\theta}\right)\left(\tfrac{\partial \ln f(\theta|Y)}{\partial\theta}\right)^T = -E\,\tfrac{\partial}{\partial\theta}\left(\tfrac{\partial \ln f(\theta|Y)}{\partial\theta}\right)^T \\ &= -E\,\tfrac{\partial}{\partial\theta}\left(\tfrac{\partial \ln f(\theta)}{\partial\theta}\right)^T - E\,\tfrac{\partial}{\partial\theta}\left(\tfrac{\partial \ln f(Y|\theta)}{\partial\theta}\right)^T = J_{prior} + J_Y \; . \end{aligned} \tag{1.210}$$

In the Bayesian context, an estimator is called unbiased if

- either $E_{\boldsymbol{\theta}}\, b_{\widehat{\theta}}(\theta) = 0$ (average bias)
- or $\lim_{\theta\to\partial\Theta} f(\theta) b_{\widehat{\theta}}(\theta) = 0$

where $\partial\Theta$ is the border of Θ, the domain of θ. The Cramer-Rao lower bound states that for any unbiased estimator:

$$R_{\widetilde{\theta}\widetilde{\theta}} \geq J^{-1} = \text{ CRB} \tag{1.211}$$

with equality (*efficiency*) iff

$$\widehat{\theta}(Y) - \theta \; = \; J^{-1}\frac{\partial \ln f(\theta|Y)}{\partial\,\theta} \quad \Rightarrow \quad f(\theta|Y) \text{ Gaussian.} \tag{1.212}$$

In general:

$$R_{\widetilde{\theta}_{MAP}\widetilde{\theta}_{MAP}} \geq R_{\widetilde{\theta}_{MMSE}\widetilde{\theta}_{MMSE}} \geq J^{-1} \tag{1.213}$$

In general $\widehat{\theta}_{MAP}$ is more easily obtainable than $\widehat{\theta}_{MMSE}$ (the second one requires the computation of $f(\theta|Y)$ whereas the first one doesn't). An estimator that still uses the MSE criterion but that is more easily derived than $\widehat{\theta}_{MMSE}$ is obtained by imposing the estimator to be a linear function of the measurements:

$$\widehat{\theta}_{LMMSE} = R_{\theta Y}R_{YY}^{-1}Y \; . \tag{1.214}$$

In general $R_{\widetilde{\theta}_{LMMSE}\widetilde{\theta}_{LMMSE}} \geq R_{\widetilde{\theta}_{MMSE}\widetilde{\theta}_{MMSE}}$. In the Gaussian case ($\theta$ and Y jointly Gaussian):

$$\widehat{\theta}_{LMMSE} = \widehat{\theta}_{MMSE} = \widehat{\theta}_{MAP} \; . \tag{1.215}$$

An important special Gaussian case is the linear model:

$$Y = H\,\theta + V \tag{1.216}$$

in which θ and V are each Gaussian and independent.

1.4 Deterministic Parameters

1.4.1 Problem setting

There are two points of view leading to the deterministic parameter estimation problem formulation:

- the parameters θ are unknown deterministic quantities
- the parameters θ are stochastic, but their prior distribution $f(\theta)$ is unknown

For some people, the only way to formulate a parameter estimation problem is according to the first point of view. They find the notion of prior knowledge artificial. One has to admit that in certain applications some creativity is required to come up with prior information. However, in other applications such as in many optimal filtering applications that we shall see later, prior information is readily available and can/should not be ignored. Some other people adhere to the Bayesian point of view and they consider θ to be random (their point of view will in fact allow us to make some useful connections between Bayesian and deterministic parameter estimation). However, in certain applications we are confronted with the problem that no meaningful (or easily describable) prior information can be found (or that so many/good measurements are available that the inclusion of prior information would complicate things unnecessarily). Whatever the point of view, in deterministic parameter estimation, the problem can be described as in Fig. 1.11.

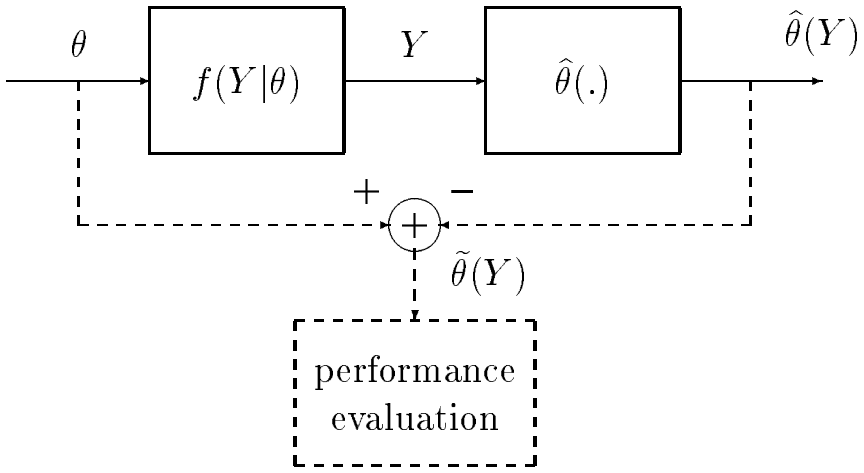


Figure 1.11: The deterministic parameter estimation problem.

There is only one experiment with one particular unknown value θ. The only stochastic description available is the conditional density $f(Y|\theta)$ describing the stochastic relation between the unknown parameters θ and the observed measurements Y. Since θ is not necessarily a random vector but just a set of parameters on which the distribution of Y depends, we often find the notations

$$f(Y|\theta) \;=\; f(Y;\theta) \;=\; f_\theta(Y) \tag{1.217}$$

but we shall continue to use $f(Y|\theta)$ which takes on the meaning of conditional distribution of Y given θ the moment we would consider θ to be random. Since this distribution describes the

only randomness there is in the deterministic context, expectation now means $E = E_{\mathbf{Y}|\boldsymbol{\theta}}$. An estimator $\widehat{\theta}(Y)$ of θ is again a function of Y (a statistic), with estimation error $\widetilde{\theta} = \theta - \widehat{\theta}(Y)$. To evaluate the quality of an estimator, we shall again introduce a *risk* function as the average value of a *cost* function

$$\mathcal{R}(\widehat{\theta}(.)|\theta) \;=\; E_{\mathbf{Y}|\boldsymbol{\theta}}\, \mathcal{C}(\widetilde{\theta}) \;=\; \int f(Y|\theta)\mathcal{C}(\theta - \widehat{\theta}(Y))dY \tag{1.218}$$

Note that $\mathcal{R}(\widehat{\theta}(.)|\theta)$ is a function of θ in general (the Bayesian risk, corresponding to $E_{\boldsymbol{\theta}}\mathcal{R}(\widehat{\theta}(.)|\theta)$, was not a function of θ). The most popular risk function is the MSE $= E_{\mathbf{Y}|\boldsymbol{\theta}}\, \|\widetilde{\theta}\|^2$. Now, the minimization of the risk function leads to $\widehat{\theta} = \theta$ (and $\mathcal{R} = 0$): this is not an acceptable estimator since the resulting $\widehat{\theta}$ depends on the unknown θ (and is not a function of Y). So, as in the Bayesian case, the quality of an estimator is evaluated using a risk. However, unlike in the Bayesian case, in the deterministic case we cannot obtain estimators by minimizing the risk w.r.t. the estimator function.

Ideally, we would like $\widehat{\theta}(.)$ to be such that $\mathcal{R}(\widehat{\theta}(.)|\theta)$ is minimized $\forall \theta \in \Theta$. However, this is impossible! Consider $\widehat{\theta}(Y) = \theta_0 \in \Theta$: this estimator ignores Y and simply chooses a certain possible value of θ. Now, $\mathcal{R}(\widehat{\theta}(.)|\theta_0) = 0$: if the true value of θ happens to be θ_0, then this estimator produces zero error!. However, it produces a potentially large error if the true value is not θ_0. Given two estimators, one is usually not uniformly better (lower risk) than the other one as illustrated in Fig. 1.12: $\widehat{\theta}_1(Y)$ is better than $\widehat{\theta}_2(Y)$ for certain values of $\theta \in \Theta$, while the situation is reversed for other values of θ. In general, a uniformly minimum risk estimator does not exist (lower (or equal) risk than all other estimators, for every possible value of θ). Hence, we are forced to consider some other desirable properties.

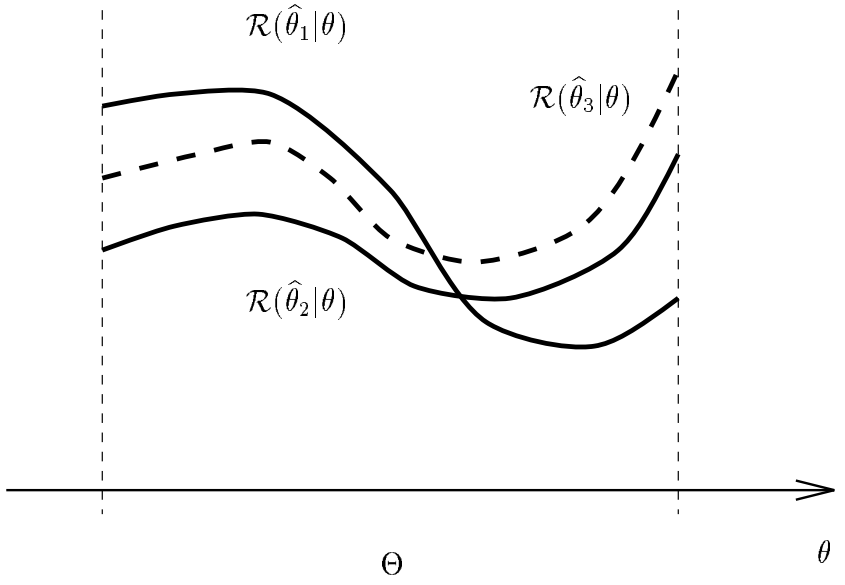


Figure 1.12: Risk functions of different estimators, as a function of $\theta \in \Theta$.

1.4.2 Some optimality properties

A first characterization of the estimation error concerns its first order moment. We introduce again the estimator *bias* $b_{\widehat{\theta}}(\theta)$ as the average deviation from the true parameter

$$b_{\widehat{\theta}}(\theta) \;=\; -E_{\mathbf{Y}|\boldsymbol{\theta}}\widetilde{\theta} \;=\; E_{\mathbf{Y}|\boldsymbol{\theta}}\left(\widehat{\theta}(Y)-\theta\right) \;=\; E_{\mathbf{Y}|\boldsymbol{\theta}}\widehat{\theta}(Y) \;-\; \theta\;. \tag{1.219}$$

An estimator will be called *unbiased* if

$$b_{\widehat{\theta}}(\theta) = 0\;,\; \forall\theta\in\Theta\;. \tag{1.220}$$

where it is important to emphasize that the bias should be zero everywhere in Θ. Now, unbiasedness is a weak property: an estimator can be correct on the average, but with large deviations. And on the other hand, good estimators exist that are biased.

Example 1.7 Mean of Gaussian i.i.d. variables

Consider the (conditional on θ) i.i.d. measurements $y_i \sim \mathcal{N}(\theta,1)\;,\quad i=1,\ldots,n$. We take as estimator of the mean $\widehat{\theta}(Y)=\overline{y}=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}y_i$, the sample mean. We have $E_{\mathbf{Y}|\boldsymbol{\theta}}\widehat{\theta} = E_{\mathbf{Y}|\boldsymbol{\theta}}\overline{y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}E_{\mathbf{Y}|\boldsymbol{\theta}}y_i = \dfrac{1}{n}\sum_{i=1}^{n}E_{\mathbf{Y}|\boldsymbol{\theta}}\theta = \theta$: the sample mean is an umbiased estimator of the mean. ◇

Example 1.8 A trivial biased estimator

The estimator $\widehat{\theta}(Y)\equiv\theta_0\in\Theta$ which always produces its preferred value θ_0 is biased: $b_{\widehat{\theta}}(\theta)=\theta_0-\theta$. The bias is only zero for $\theta=\theta_0$ and not everywhere in Θ. ◇

Another desirable property is admissibility. An estimator $\widehat{\theta}(.)$ is called *inadmissible* if another estimator $\widehat{\theta}'(.)$ has uniformly lower risk:

$$\forall\theta\in\Theta:\mathcal{R}(\widehat{\theta}'|\theta)\le\mathcal{R}(\widehat{\theta}|\theta)\;,\quad \exists\theta_0\in\Theta:\mathcal{R}(\widehat{\theta}'|\theta_0)<\mathcal{R}(\widehat{\theta}|\theta_0)\;. \tag{1.221}$$

$\widehat{\theta}$ is called *admissible* if no such $\widehat{\theta}'$ exists. For example, $\widehat{\theta}_3$ in Fig. 1.12 is inadmissible since $\widehat{\theta}_2$ has lower risk for all $\theta\in\Theta$. Admissibility characterizes a large class of estimators but does not lead to a specific optimal estimator.

An optimal estimator may be obtained by considering the following criterion. $\widehat{\theta}(Y)$ is said to be *minimax* if it satisfies

$$\sup_{\theta\in\Theta}\mathcal{R}(\widehat{\theta}|\theta) \;=\; \inf_{\widehat{\theta}'}\sup_{\theta\in\Theta}\mathcal{R}(\widehat{\theta}'|\theta)\;. \tag{1.222}$$

The notions infimum/supremum over a set that is not closed can be defined as the minimum/maximum over the corresponding closed set (including its borders). A minimax estimator minimizes (compared to other estimators) the maximum risk over Θ. Minimax estimators are difficult to obtain in general.

For biased estimators, the following relationship exists between the error correlation and covariance matrices

$$R_{\widetilde{\theta}\widetilde{\theta}} = E_{\mathbf{Y}|\boldsymbol{\theta}}\widetilde{\theta}\widetilde{\theta}^T = C_{\widetilde{\theta}\widetilde{\theta}} + (E_{\mathbf{Y}|\boldsymbol{\theta}}\widetilde{\theta})(E_{\mathbf{Y}|\boldsymbol{\theta}}\widetilde{\theta})^T = C_{\widetilde{\theta}\widetilde{\theta}} + b_{\widehat{\theta}}(\theta)b_{\widehat{\theta}}(\theta)^T \ . \tag{1.223}$$

As before, the minimum mean squared error summarizes the error correlation matrix in one number

$$\text{MSE} \ = \text{tr}\left\{R_{\widetilde{\theta}\widetilde{\theta}}\right\} \ . \tag{1.224}$$

Note however that in the deterministic context $R_{\widetilde{\theta}\widetilde{\theta}}$, $C_{\widetilde{\theta}\widetilde{\theta}}$ and the MSE are functions of θ.

Uniformly minimum risk estimators may be found if we restrict the class of estimators to the unbiased estimators. $\widehat{\theta}$ is called a *uniformly minimum variance unbiased estimator* (UMVUE) if it is unbiased and if for any other unbiased estimator $\widehat{\theta}'$

$$C_{\widetilde{\theta}\widetilde{\theta}} = R_{\widetilde{\theta}\widetilde{\theta}} = E_{\mathbf{Y}|\boldsymbol{\theta}}(\widehat{\theta}(Y)-\theta)(\widehat{\theta}(Y)-\theta)^T \ \leq \ E_{\mathbf{Y}|\boldsymbol{\theta}}(\widehat{\theta}'(Y)-\theta)(\widehat{\theta}'(Y)-\theta)^T = R_{\widetilde{\theta}'\widetilde{\theta}'} = C_{\widetilde{\theta}'\widetilde{\theta}'} \quad \forall\theta\in\Theta \ . \tag{1.225}$$

The term uniformly refers to the fact that this inequality holds everywhere in Θ. The inequality also holds for the traces of the correlation matrices which is the MSE ($\text{tr}\{R_{\widetilde{\theta}'\widetilde{\theta}'}\}-\text{tr}\{R_{\widetilde{\theta}\widetilde{\theta}}\} = \text{tr}\{R_{\widetilde{\theta}'\widetilde{\theta}'} - R_{\widetilde{\theta}\widetilde{\theta}}\} \geq 0$: the trace of a positive semidefinite matrix is nonnegative). In the unbiased case, the MSE is also the variance $E_{\mathbf{Y}|\boldsymbol{\theta}}\|\widetilde{\theta}-E_{\mathbf{Y}|\boldsymbol{\theta}}\widetilde{\theta}\|^2$, hence the name minimum variance. Since they are in fact unbiased MMSE estimators, UMVUE estimators are highly desirable. However, they may not exist or they may be difficult to compute. They can be computed if a *complete sufficient statistic* can be found.

Estimates of functions of the parameters♠

Let $\widehat{\phi}(Y)$ be an estimate of $\phi(\theta)$, a function of θ. $\widehat{\phi}(Y)$ is said to be unbiased if

$$E_{\mathbf{Y}|\boldsymbol{\theta}}\left(\widehat{\phi}(Y)-\phi(\theta)\right) \ = \ 0 \ , \quad \forall\theta\in\Theta \ . \tag{1.226}$$

The function $\phi(\theta)$ is said to be *estimable* if an unbiased estimator $\widehat{\phi}(Y)$ for it exists.

1.4.3 Finding UMVUE using complete sufficient statistics♠

1.4.4 Maximum Likelihood estimation

We introduce here what is probably the most popular technique for the estimation of deterministic parameters. The maximum likelihood (ML) estimation philosophy consists of choosing that value of the parameters that renders the observation Y the most likely realization of $\mathbf{Y}$:

$$\widehat{\theta}_{ML}(Y) \ = \ \arg\max_{\theta\in\Theta} \ f(Y|\theta) \ . \tag{1.227}$$

$f(Y|\theta)$ as conditional distribution of Y given θ is a function of Y and for the rest depends on one value of θ, the true value. However, for the purpose of ML estimation, things get reversed. We observe one realization (one value) Y of $\mathbf{Y}$ which now fixed. For the rest, $f(Y|\theta)$ is now a function of θ, the value of which we don't know and hence is variable. In order to emphasize

the dependence of $f(Y|\theta)$ on θ and the fact that the observation Y is fixed, it is often denoted as

$$L(\theta;Y) \;=\; f(Y|\theta)\;. \tag{1.228}$$

$L(\theta;Y)$ $(=\; f(Y|\theta))$ is called the *likelihood function*. It is the likelihood that Y would be observed for the value θ. It is a function of θ that depends on the realization Y.

Example 1.9 Estimation of the mean of triangular noise

We observe an unknown constant θ in additive noise v with a triangular density function

$$y = \theta + v\;,\; f_{\mathbf{v}}(v) = \begin{cases} 1-|v| & ,\; |v| \le 1 \\ 0 & ,\; |v| > 1 \end{cases} \qquad f(y|\theta) = f_{\mathbf{v}}(y-\theta) = L(\theta;y)\,. \tag{1.229}$$

In the left part of Fig. 1.13, we find $f(y|\theta)$. θ is the true value and the height of the dashed line indicates the probability density with which the realization y has occured. In the right part of the figure, the value y is the value that we have observed (the realization of $\mathbf{y}$). The likelihood function indicates for each possible value of θ the likelihood (probability density) that y would have occured if θ were the true value. In particular for the value $\widehat{\theta}$, the height of the dashed line indicates the likelihood that y would have occured if $\widehat{\theta}$ were the true value of θ. The ML estimate chooses the value for θ which makes the observation y the most likely realization of $\mathbf{y}$ and hence

$$\widehat{\theta}_{ML}(y) \;=\; \arg\max_{\theta\in\Theta} L(\theta;y) \;=\; y \tag{1.230}$$

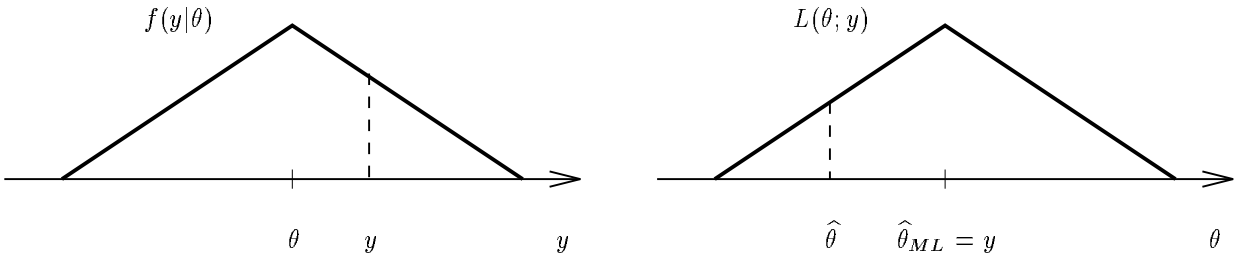


Figure 1.13: Probability density and likelihood functions for the mean of triangular noise.

$\diamond$

Remarks:

(i) Since the logarithmic function is strictly monotone, the maximum point of $L(\theta;Y)$ corresponds with the maximum point of $\ln L(\theta;Y)$, called the *log likelihood function*.

(ii) Often $f(\theta|Y)$ satisfies certain regularity conditions so that $\widehat{\theta}_{ML}$ is a solution of

$$\frac{\partial}{\partial\theta} \ln f(Y|\theta) \;=\; 0\;. \tag{1.231}$$

The conditions for a maximum (rather than another form of extremum) need to be verified of course.

(iii) The ML estimator is given by the *global* maximum of $L(\theta;Y)$. If there are several local maxima, all of them need to be examined and compared to find the global maximum.

(iv) Even if $f(Y|\theta)$ satisfies regularity conditions, the maximum may occur at the boundary of the parameter space Θ (which may not necessarily be $(-\infty,\infty)$ for every θ_i). In that case, the maximum is not a local extremum.

(v) Assume now that we consider $\boldsymbol{\theta}$ to be random but that the only reason why we perform deterministic parameter estimation is that we are not sure about the prior distribution $f_{\boldsymbol{\theta}}(\theta)$. In this case, the ML estimator can be seen as a limiting case of the MAP estimator when the prior distribution $f(\theta)$ becomes uninformative (uniform distribution). Indeed

$$\begin{aligned}\widehat{\theta}_{MAP}(Y) &= \arg\max_{\theta\in\Theta} f(\theta|Y) = \arg\max_{\theta\in\Theta} \frac{f(Y|\theta)f(\theta)}{f(Y)} \\ &= \arg\max_{\theta\in\Theta} f(Y|\theta)f(\theta) \overset{f(\theta)=c^t}{=} \arg\max_{\theta\in\Theta} f(Y|\theta) = \widehat{\theta}_{ML}(Y)\ .\end{aligned} \tag{1.232}$$

This means that the ML estimate can be seen to be a special case of the MAP estimate, when the prior distribution is uniform. This provides another justification for considering the ML estimate. For a constant prior density function, we get $\frac{\partial \ln f(\theta)}{\partial\theta} = 0,\ \forall\theta\in\Theta$ and hence

$$J_{prior} = -E\,\frac{\partial}{\partial\theta}\left(\frac{\partial \ln f(\theta)}{\partial\theta}\right)^T = -E\,0 = 0 \tag{1.233}$$

A uniform prior distribution contains no information about $\boldsymbol{\theta}$. For those components θ_i of θ for which the support is unbounded (e.g. a Gaussian distribution), this means that $\sigma^2_{\theta_i}\to\infty$ (information $\to$ 0).

Example 1.10 ML estimation of mean and variance of Gaussian i.i.d. variables

We observe $y_i = \mu + \sigma v_i,\ v_i \sim \mathcal{N}(0,1)$ i.i.d. or alternatively $y_i \sim \mathcal{N}(\mu,\sigma^2)$ i.i.d. for $i = 1,\ldots,n$. The vector of unknown parameters is $\theta = [\mu\ \ \sigma^2]^T$. We shall derive $\widehat{\theta}_{ML}$ and investigate its bias. Since the measurements are i.i.d. given θ, we get for the joint conditional distribution

$$f(Y|\mu,\sigma^2) = \prod_{i=1}^n f(y_i|\mu,\sigma^2) = (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}\exp[-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\mu)^2]\ . \tag{1.234}$$

The loglikelihood function is

$$\ln L(\theta;Y) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\mu)^2\ . \tag{1.235}$$

Searching for extrema, we consider the gradient condition

$$\begin{cases} \dfrac{\partial}{\partial\mu}\ln L(\theta;Y) = 0 = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^n(y_i-\mu) & (1) \\[2ex] \dfrac{\partial}{\partial\sigma^2}\ln L(\theta;Y) = 0 = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^n(y_i-\mu)^2 & (2) \end{cases} \tag{1.236}$$

which leads us to the following extrema

$$\begin{cases} (1) \Rightarrow \ \widehat{\mu}_{ML} = \dfrac{1}{n}\sum_{i=1}^{n} y_i = \overline{y} \quad \text{sample mean} \\ \\ (2) \Rightarrow \ \widehat{\sigma^2}_{ML} = \dfrac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \overline{(y-\overline{y})^2} \quad \text{sample variance} \end{cases} \tag{1.237}$$

There is only one local extremum. By investigating the Hessian, we find it is a local maximum. Checking the boundary conditions, we find that $\ln L(\theta; Y) \to -\infty$ as $\mu \to \pm\infty$ or $\sigma^2 \to 0$ or $+\infty$. Hence this local maximum is indeed the global maximum.

To investigate the bias of $\widehat{\mu}_{ML}$, consider

$$E[\widehat{\mu}_{ML}|\mu,\sigma^2] = E[\overline{y}|\mu,\sigma^2] = \frac{1}{n}\sum_{i=1}^{n} E[y_i|\mu,\sigma^2] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \ , \ \forall \mu \ . \tag{1.238}$$

Hence $\widehat{\mu}_{ML}$ is unbiased! For the calculation of the mean of $\widehat{\sigma^2}_{ML}$, consider the following manipulation: with $\overline{y} = \frac{1}{n}\mathbf{1}^T Y$, we get

$$\begin{aligned} n\,\widehat{\sigma^2}_{ML} \ &= \sum_{i=1}^{n}(y_i - \overline{y})^2 = (Y - \mu\mathbf{1} + \mu\mathbf{1} - \overline{y}\mathbf{1})^T (Y - \mu\mathbf{1} + \mu\mathbf{1} - \overline{y}\mathbf{1}) = (Y-\mu\mathbf{1})^T(Y-\mu\mathbf{1}) \\ &+ (\overline{y}-\mu)^2 \underbrace{\mathbf{1}^T\mathbf{1}}_{=n} - 2(\overline{y}-\mu)\underbrace{\mathbf{1}^T(Y-\mu\mathbf{1})}_{=n(\overline{y}-\mu)} = (Y-\mu\mathbf{1})^T(Y-\mu\mathbf{1}) - \tfrac{1}{n}(Y-\mu\mathbf{1})^T\mathbf{1}\mathbf{1}^T(Y-\mu\mathbf{1}) \end{aligned} \tag{1.239}$$

Hence

$$\begin{aligned} E[\widehat{\sigma^2}_{ML}|\mu,\sigma^2] &= \tfrac{1}{n}\,E_{Y|\mu,\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 - \frac{1}{n^2}\,\mathrm{tr}\left\{\mathbf{1}\mathbf{1}^T\,E_{Y|\mu,\sigma^2}(Y-\mu\mathbf{1})(Y-\mu\mathbf{1})^T\right\} \\ &= \sigma^2 - \tfrac{1}{n^2}\,\mathrm{tr}\{\mathbf{1}\mathbf{1}^T\,\sigma^2\,I_n\} = \sigma^2 - \tfrac{1}{n^2}\,\sigma^2\underbrace{\mathbf{1}^T I_n \mathbf{1}}_{=n} = (1-\tfrac{1}{n})\sigma^2 = \tfrac{n-1}{n}\sigma^2 \neq \sigma^2 \end{aligned} \tag{1.240}$$

Hence $\widehat{\sigma^2}_{ML}$ is biased! Remark that we could introduce an unbiased estimate of the variance:

$$\widehat{\sigma^2}_{ub} = \frac{n}{n-1}\widehat{\sigma^2}_{ML} = \frac{1}{n-1}\sum_{i-1}^{n}(y_i - \overline{y})^2 \ . \tag{1.241}$$

However, it can be shown that $MSE_{\widehat{\sigma^2}_{ub}} > MSE_{\widehat{\sigma^2}_{ML}}$. So the ML estimate is better even though it is biased. $\diamond$

Example 1.11 ML estimation of the mean of uniform i.i.d. random variables

We observe n i.i.d. uniformly distributed random variables $y_i \sim \mathcal{U}[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$ with unknown mean θ. In other words,

$$f(y_i|\theta) = \begin{cases} 1 & , \ y_i \in [\theta - \frac{1}{2}, \theta + \frac{1}{2}] \\ 0 & , \ \text{elsewhere} \end{cases} \tag{1.242}$$

To find $\widehat{\theta}_{ML}$, it will be useful to introduce the indicator function

$$f(y_i|\theta) = I_{[\theta-\frac{1}{2},\theta+\frac{1}{2}]}(y_i) = 1 \quad \text{if} \quad \theta - \frac{1}{2} \leq y_i \leq \theta + \frac{1}{2} \;\Leftrightarrow\; y_i - \frac{1}{2} \leq \theta \leq y_i + \frac{1}{2}\,. \tag{1.243}$$

We can now write for the joint distribution and hence the likelihood function

$$\begin{aligned} f(Y|\theta) &= \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} I_{[\theta-\frac{1}{2},\theta+\frac{1}{2}]}(y_i) = \prod_{i=1}^{n} I_{[y_i-\frac{1}{2},y_i+\frac{1}{2}]}(\theta) \\ &= I_{\bigcap_{i=1}^{n}[y_i-\frac{1}{2},y_i+\frac{1}{2}]}(\theta) = I_{[y_{max}-\frac{1}{2},y_{min}+\frac{1}{2}]}(\theta) = L(\theta;Y) \end{aligned} \tag{1.244}$$

where $y_{max/min} = \max / \min\{y_1, \ldots, y_n\}$. Hence any $\widehat{\theta}_{ML} \in [y_{max}-\frac{1}{2}, y_{min}+\frac{1}{2}]$ maximizes the likelihood function. So we get a whole interval for the ML estimates! In practice, to make the ML estimator a point estimator, one may be inclined to choose the midpoint of the interval, i.e. $\widehat{\theta}_{ML} = \dfrac{y_{min} + y_{max}}{2}$.

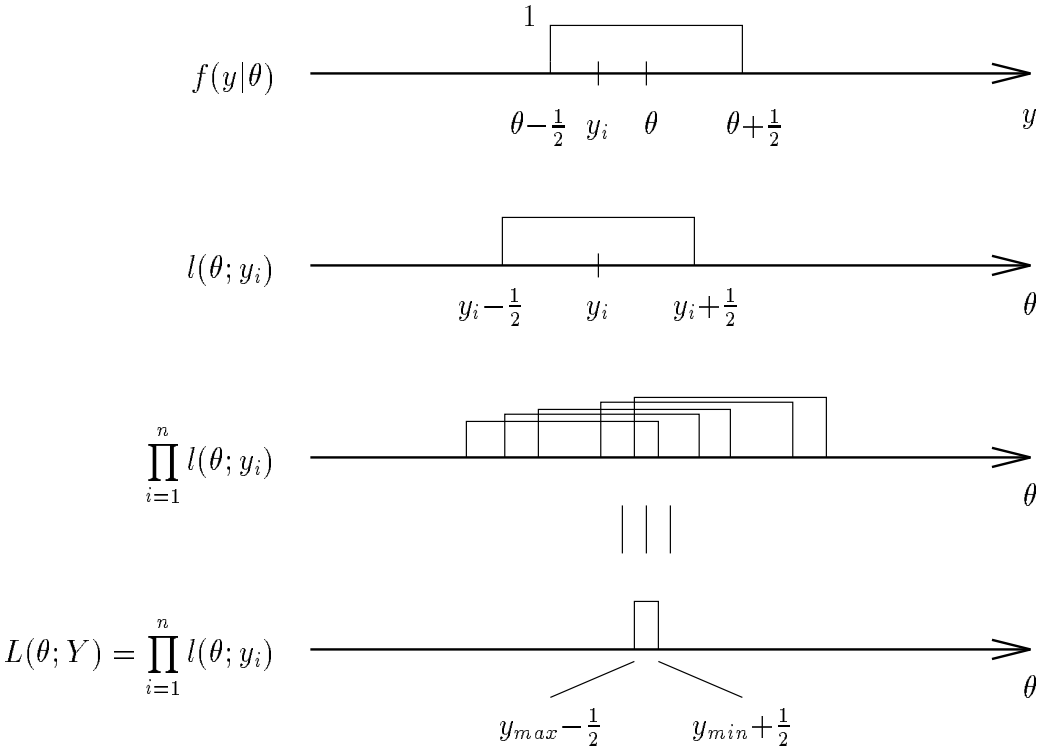


Figure 1.14: ML estimation of the mean of i.i.d. uniform variables.

◇

ML Invariance under Transformations♠

Assume that the data Y are still parameterized in terms of θ but that we want to estimate a set of transformed parameters. So consider the function $g(\theta) : \Theta \to \Phi \subset \mathcal{R}^p$, $p \leq m$ and we

want to estimate $\phi = g(\theta)$. For any $\phi \in \Phi$, $g^{-1}(\phi) \subseteq \Theta$ denotes the inverse image of ϕ. We shall define the *induced likelihood function* $L(\phi; Y)$ of $\phi = g(\theta)$ as

$$L(\phi; Y) \;=\; \sup_{\theta \in g^{-1}(\phi)} f(Y|\theta)\;. \tag{1.245}$$

If $g(.)$ is an invertible transformation, then $L(\phi; Y) = f(Y|\phi)$. A natural estimate for ϕ is $\widehat{\phi} = g(\widehat{\theta}_{ML})$. The following theorem will justify calling $g(\widehat{\theta}_{ML}) = \widehat{\phi}_{ML}$, the ML estimate of ϕ.

Theorem 1.3 (ML Estimates of Transformed Parameters) *If $\widehat{\theta}$ is the ML estimate of θ, then the global maximum of the induced likelihood $L(\phi; Y)$ is attained at $\widehat{\phi} = g(\widehat{\theta})$.*

Proof:

$$\begin{aligned} f(Y|\widehat{\theta}) \;&\leq\; \sup\nolimits_{\theta \in g^{-1}(g(\widehat{\theta}))} f(Y|\theta) = \sup\nolimits_{\theta \in g^{-1}(\widehat{\phi})} f(Y|\theta) = L(\widehat{\phi}; Y) \\ &\leq\; \sup_{\phi \in \Phi} L(\phi; Y) = \sup_{\theta \in \Theta} f(Y|\theta) = f(Y|\widehat{\theta}) \end{aligned} \tag{1.246}$$

Hence $L(\widehat{\phi}; Y) = \sup\limits_{\phi \in \Phi} L(\phi; Y)$. □

Example 1.12 ML estimation of the mean squared value of a Gaussian variable

Consider example 1.10 again with $y_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. for $i = 1, \ldots, n$ and $\theta = [\mu \;\; \sigma^2]^T$. Assume we want to estimate now the men squared value $\phi = Ey^2 = \mu^2 + \sigma^2 = g(\theta)$. Then according to the previous theorem

$$\widehat{\phi}_{ML} \;=\; (\widehat{\mu}_{ML})^2 + \widehat{\sigma^2}_{ML}\;. \tag{1.247}$$

◇

1.4.5 Cramer-Rao Bound

In the Bayesian case, the MMSE estimator minimizes the parameter estimation error correlation matrix. In the deterministic case, the UMVUE does the same thing within the class of the unbiased estimators. However, the UMVUE is hardly used in practice. The much more popular ML estimator is not based on the minimization of the parameter estimation error correlation matrix (some simpler estimators that we shall see later are based on the minimization of the MSE, but subject to certain structural constraints on the estimator). Therefore, for the evaluation of the performance of the ML and other estimators, it will be important to be able to compare the performance to a bound.

As in the Bayesian case, this performance bound is based on the Fisher Information Matrix (FIM). The FIM for deterministic parameters is defined by the first expression below

$$J(\theta) \;=\; E_{Y|\theta}\left(\frac{\partial \ln f(Y|\theta)}{\partial \theta}\right)\left(\frac{\partial \ln f(Y|\theta)}{\partial \theta}\right)^T \;=\; -E_{Y|\theta}\,\frac{\partial}{\partial \theta}\left(\frac{\partial \ln f(Y|\theta)}{\partial \theta}\right)^T \tag{1.248}$$

where the equality to the second expression can be shown as in the Bayesian case. The FIM can again be shown to satisfy all the properties we specified for an information matrix. Note that in contrast to the Bayesian case, $J(\theta)$ now depends on the true parameter value θ.

Lemma 1.4 (Unit Cross Correlation) *For any unbiased estimator* $\widehat{\theta}(Y)$

$$E_{Y|\theta} \frac{\partial \ln f(Y|\theta)}{\partial \theta} \left(\widehat{\theta} - \theta\right)^T = I \,. \tag{1.249}$$

In words, the cross correlation matrix between $\frac{\partial \ln f(Y|\theta)}{\partial \theta}$ and the estimation error of any unbiased estimator is the identity matrix. The proof is similar to the Bayesian case. We are now ready to formulate the Cramer-Rao Bound (CRB).

Theorem 1.4 (CRB for Deterministic Parameters) *If the estimator* $\widehat{\theta}(Y)$ *of* θ *is unbiased, then the covariance matrix of the parameter estimation errors* $\widetilde{\theta}$ *is bounded below by the inverse of the information matrix:*

$$C_{\widetilde{\theta}\widetilde{\theta}} = R_{\widetilde{\theta}\widetilde{\theta}} = E_{Y|\theta} (\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T \geq J^{-1}(\theta) \tag{1.250}$$

with equality iff

$$\widehat{\theta}(Y) - \theta = J^{-1}(\theta) \frac{\partial \ln f(Y|\theta)}{\partial \theta} \quad a.e.\ (\theta) \,. \tag{1.251}$$

An estimator that achieves the lower bound $(\forall \theta \in \Theta)$ *is called efficient.*

Remarks:

(i) It will be interesting to investigate what the efficiency condition implies for the conditional distribution $f(Y|\theta)$. When the equality (1.251) holds, we can integrate to get

$$f(Y|\theta) = h(Y) \exp[c_1^T(\theta)\widehat{\theta}(Y) - c_0(\theta)] \tag{1.252}$$

where $c_1(\theta)$ and $c_0(\theta)$ are vector and scalar functions of θ resp. such that $\frac{\partial c_1^T(\theta)}{\partial \theta} = J(\theta)$ and $\frac{\partial c_0(\theta)}{\partial \theta} = J(\theta)\theta$. Hence the functions $f(Y|\theta)$ that satisfy (1.252) form an exponential family and $\widehat{\theta}(Y)$ is a sufficient statistic. Remark that in the deterministic case, the class of functions $f(Y|\theta)$ (the exponential family) that lead to efficiency is much larger than in the Bayesian case, in which case $f(\theta|Y)$ was required to be Gaussian, a specific exponential distribution.

(ii) The CRB $J^{-1}(\theta)$ only depends on $f(Y|\theta)$, not on $\widehat{\theta}(Y)$.

(iii) The CRB has two uses:

(1) evaluate the performance of unbiased estimators: $\widehat{\theta}$ with $b_{\widehat{\theta}}(\theta) \equiv 0$:
if $C_{\widetilde{\theta}\widetilde{\theta}} - J^{-1}(\theta)$ is small enough, then $\widehat{\theta}$ is good enough.

(2) find UMVUE: $\min_{\widehat{\theta}: b_{\widehat{\theta}} \equiv 0} C_{\widetilde{\theta}\widetilde{\theta}} \geq J^{-1}(\theta)$.

If $\widehat{\theta}$ is efficient: $\forall \theta \in \Theta$, $C_{\widetilde{\theta}\widetilde{\theta}} = J^{-1}(\theta)$, then $\widehat{\theta}$ is UMVUE since no unbiased estimator can have a smaller error correlation matrix than the CRB!

(iv) **Theorem 1.5 (Efficient ML estimators)** *Suppose $\widehat{\theta}_{ML}$ is obtained by $\frac{\partial}{\partial\theta} f(Y|\theta)\Big|_{\theta=\widehat{\theta}_{ML}} = 0$. Then if an efficient estimator exists, it is $\widehat{\theta}_{ML}$.*

Proof: $\widehat{\theta}_{eff}$ satisfies

$$\frac{\partial \ln f(Y|\theta)}{\partial \theta} = \underbrace{J(\theta)}_{>0} [\widehat{\theta}_{eff} - \theta] . \tag{1.253}$$

For $\theta = \widehat{\theta}_{ML}$, the LHS equals 0. Hence the RHS equals 0 which implies: $\widehat{\theta}_{eff} = \widehat{\theta}_{ML}$. □

Example 1.13 ML estimation and CRB for the mean of Gaussian variables

Consider n i.i.d. measurements $y_i \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 known, the only unknown parameter is $\theta = \mu$. The likelihood function is

$$f(Y|\mu) = \prod_{i=1}^{n} f(y_i|\mu) = (2\pi\sigma^2)^{-n/2} \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2] . \tag{1.254}$$

The first two derivatives of the loglikelihood function are

$$\frac{\partial \ln f(Y|\mu)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) , \quad \frac{\partial^2 \ln f(Y|\mu)}{\partial \mu^2} = -\frac{n}{\sigma^2} . \tag{1.255}$$

Before considering an estimator, we can investigate the FIM and the CRB

$$J = -E_{Y|\mu}\frac{\partial^2 \ln f(Y|\mu)}{\partial \mu^2} = \frac{n}{\sigma^2} , \quad C_{\widetilde{\mu}\widetilde{\mu}} = E_{Y|\mu}(\widehat{\mu} - \mu)^2 \geq J^{-1} = \frac{\sigma^2}{n} . \tag{1.256}$$

Note that in this particular example the FIM J does not depend on $\theta = \mu$. An intuitive estimator for the mean is the sample mean

$$\widehat{\mu} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i , \quad E_{Y|\mu}\widehat{\mu} = \mu : \text{ unbiased} \tag{1.257}$$

Investigating the estimation error variance, we find

$$C_{\widetilde{\mu}\widetilde{\mu}} = E_{Y|\mu}(\widehat{\mu} - \mu)^2 = E_{Y|\mu}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)\right)^2 = \frac{1}{n^2} n\,\sigma^2 = \frac{\sigma^2}{n} = J^{-1} . \tag{1.258}$$

The sample mean is an efficient estimator since it reaches the CRB! We find in this example the ML estimator by setting the gradient in (1.255) equal to zero. We find $\widehat{\mu}_{ML} = \overline{y}$: according to theorem 1.5, $\widehat{\mu}_{ML}$ had indeed to be equal to the sample mean, an efficient estimator, since $\widehat{\mu}_{ML}$ is obtained by setting the gradient of the likelihood function equal to zero. Another way of checking that $\widehat{\mu}_{ML}$ is efficient is to check that the efficiency condition (1.251) is satisfied:

$$\frac{\partial \ln f(Y|\mu)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = \frac{n}{\sigma^2}(\overline{y} - \mu) = J^{-1}(\widehat{\mu}_{ML} - \mu) . \tag{1.259}$$

◇

1.4.6 A Non-Linear Signal Model♠

We shall elaborate here on the FIM calculation for a particular signal model. We shall assume that the measurements contain a signal of interest that depends on parameters in a non-linear but known way. However, the value of those parameters is unknown. The signal of interest is measured in additive noise. So we get for the measurements

$$y_k = s_k(\theta) + v_k \ , k = 1, \ldots, n \quad \Rightarrow \quad Y = S(\theta) + V \ . \tag{1.260}$$

$S(\theta)$ may not depend on all the parameters θ_i, some θ_j may be used to characterize V (e.g. σ_v^2). We get the conditional distribution by extending the reasoning in example 1.2

$$f_{\mathbf{Y}|\boldsymbol{\theta}}(Y|\theta) = f_{\mathbf{V}+\mathbf{S}(\boldsymbol{\theta})|\boldsymbol{\theta}}(Y|\theta) = f_{\mathbf{V}|\boldsymbol{\theta}}(Y - S(\theta)|\theta) \ . \tag{1.261}$$

The last expression does not simplify to $f_{\mathbf{V}}(Y - S(\theta))$ since the distribution of $\mathbf{V}$ may depend on some of the θ_j. We get for the caculation of the FIM

$$\begin{aligned}
J(\theta) &= -E_{\mathbf{Y}|\boldsymbol{\theta}} \frac{\partial}{\partial \theta} \left(\frac{\partial \ \ln f_{\mathbf{Y}|\boldsymbol{\theta}}(Y|\theta)}{\partial \theta} \right)^T = - \int dY \, f_{\mathbf{Y}|\boldsymbol{\theta}}(Y|\theta) \frac{\partial}{\partial \theta} \left(\frac{\partial \ \ln f_{\mathbf{Y}|\boldsymbol{\theta}}(Y|\theta)}{\partial \theta} \right)^T \\
&= - \int dY \, f_{\mathbf{V}|\boldsymbol{\theta}}(Y - S(\theta)|\theta) \frac{\partial}{\partial \theta} \left(\frac{\partial \ \ln f_{\mathbf{V}|\boldsymbol{\theta}}(Y - S(\theta)|\theta)}{\partial \theta} \right)^T \\
&= - \int dV \, f_{\mathbf{V}|\boldsymbol{\theta}}(V|\theta) \ \frac{\partial}{\partial \theta} \left(\frac{\partial \ \ln f_{\mathbf{V}|\boldsymbol{\theta}}(Y - S(\theta)|\theta)}{\partial \theta} \right)^T \Bigg|_{Y = V + S(\theta)} \\
&= -E_{\mathbf{V}|\boldsymbol{\theta}} \ \frac{\partial}{\partial \theta} \left(\frac{\partial \ \ln f_{\mathbf{V}|\boldsymbol{\theta}}(Y - S(\theta)|\theta)}{\partial \theta} \right)^T \Bigg|_{Y = V + S(\theta)}
\end{aligned} \tag{1.262}$$

Hence the FIM calculation corresponds to the sequence of the following three operations:

- calculate $-\frac{\partial}{\partial \theta} \left(\frac{\partial \ \ln f_{\mathbf{V}|\boldsymbol{\theta}}(Y - S(\theta)|\theta)}{\partial \theta} \right)^T$
- substitute Y by $V + S(\theta)$
- average over $V|\theta$

A more specific case is: $Y = S(\theta) + V$ but now $V \sim \mathcal{N}(0, C_{VV}(\theta))$, then

$$\ln f(Y|\theta) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln(\det C_{VV}(\theta)) - \frac{1}{2} (Y - S(\theta))^T C_{VV}^{-1}(\theta) (Y - S(\theta)) \ . \tag{1.263}$$

We shall consider a simple example of this specific case in which

$$C_{VV}(\theta) = \sigma_v^2 I_n, \ \theta = \begin{bmatrix} \sigma_v^2 \\ \overline{\theta} \end{bmatrix} \ , \ \ S = S(\overline{\theta}) \ , \ \ f_{\mathbf{Y}|\boldsymbol{\theta}}(Y|\theta) = f_{\mathbf{V}|\boldsymbol{\sigma_v^2}}(Y - S(\overline{\theta})|\sigma_v^2) \ . \tag{1.264}$$

In this case

$$\frac{\partial \ln f(Y|\theta)}{\partial \theta} = \begin{bmatrix} -\frac{n}{2} \frac{1}{\sigma_v^2} + \frac{1}{2\sigma_v^4} (Y - S(\overline{\theta}))^T (Y - S(\overline{\theta})) \\ \frac{1}{\sigma_v^2} \frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}} (Y - S(\overline{\theta})) \end{bmatrix} \ . \tag{1.265}$$

For the purpose of ML estimation, we set the gradient in (1.265) equal to zero after having substituted Y by its specific measurements. This leads to the ML estimates

$$\widehat{\sigma_v^2}_{ML} = \frac{1}{n}(Y-S(\widehat{\overline{\theta}}_{ML}))^T(Y-S(\widehat{\overline{\theta}}_{ML}))\,, \quad \left.\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}}\right|_{\overline{\theta}=\widehat{\overline{\theta}}_{ML}} (Y-S(\widehat{\overline{\theta}}_{ML})) = 0 \;\Rightarrow\; \widehat{\overline{\theta}}_{ML} \tag{1.266}$$

where the equation for $\widehat{\overline{\theta}}_{ML}$ is implicit. For the purpose of the calculation of the FIM however, Y is a random vector. For the computation of the second order derivatives, it will be useful to rewrite the first order derivative (1.265) as

$$\left(\frac{\partial \ \ln f_{\mathbf{V}|\boldsymbol{\sigma_v^2}}(Y - S(\overline{\theta})|\sigma_v^2)}{\partial \theta}\right)^T =$$

$$[\, -\tfrac{n}{2}\tfrac{1}{\sigma_v^2} + \tfrac{1}{2\sigma_v^4}(Y-S(\overline{\theta}))^T(Y-S(\overline{\theta})) \qquad \underbrace{\frac{1}{\sigma_v^2}(Y-S(\overline{\theta}))^T\left(\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}}\right)^T}_{=\frac{1}{\sigma_v^2}\sum_{k=1}^{n}(y_k-s_k(\overline{\theta}))\left(\frac{\partial s_k(\overline{\theta})}{\partial \overline{\theta}}\right)^T} \,] \tag{1.267}$$

We get for the Hessian

$$\frac{\partial}{\partial \theta}\left(\frac{\partial \ \ln f_{\mathbf{V}|\boldsymbol{\sigma_v^2}}(Y - S(\overline{\theta})|\sigma_v^2)}{\partial \theta}\right)^T =$$

$$\left[\begin{array}{cc} \frac{n}{2\sigma_v^4} - \frac{1}{\sigma_v^6}(Y-S)^T(Y-S) & -\frac{1}{\sigma_v^4}(Y-S(\overline{\theta}))^T\left(\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}}\right)^T \\ -\frac{1}{\sigma_v^4}\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}}(Y-S(\overline{\theta})) & \frac{1}{\sigma_v^2}\left\{-\left(\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}}\right)\left(\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}}\right)^T + \sum_{k=1}^{n}(y_k-s_k(\overline{\theta}))\frac{\partial}{\partial \overline{\theta}}\left(\frac{\partial s_k(\overline{\theta})}{\partial \overline{\theta}}\right)^T\right\} \end{array}\right] \tag{1.268}$$

We now substitute Y by $V + S(\theta)$ which leads to

$$\frac{\partial}{\partial \theta}\left(\frac{\partial \ \ln f_{\mathbf{V}|\boldsymbol{\sigma_v^2}}(Y - S(\overline{\theta})|\sigma_v^2)}{\partial \theta}\right)^T\Bigg|_{Y=V+S(\overline{\theta})} =$$

$$\left[\begin{array}{cc} \frac{n}{2\sigma_v^4} - \frac{1}{\sigma_v^6}V^TV & -\frac{1}{\sigma_v^4}V^T\left(\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}}\right)^T \\ -\frac{1}{\sigma_v^4}\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}}V & \frac{1}{\sigma_v^2}\left\{-\left(\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}}\right)\left(\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}}\right)^T + \sum_{k=1}^{n} v_k\frac{\partial}{\partial \overline{\theta}}\left(\frac{\partial s_k(\overline{\theta})}{\partial \overline{\theta}}\right)^T\right\} \end{array}\right] \tag{1.269}$$

Averaging over $V|\sigma_v^2$ gives

$$J(\theta) = -E_{\mathbf{V}|\boldsymbol{\sigma_v^2}}\,\frac{\partial}{\partial \theta}\left(\frac{\partial \ \ln f_{\mathbf{V}|\boldsymbol{\sigma_v^2}}(Y - S(\overline{\theta})|\sigma_v^2)}{\partial \theta}\right)^T\Bigg|_{Y=V+S(\overline{\theta})} = \left[\begin{array}{cc} \frac{n}{2\sigma_v^4} & 0 \\ 0 & \frac{1}{\sigma_v^2}\left(\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}}\right)\left(\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}}\right)^T \end{array}\right] \tag{1.270}$$

which shows that J depends indeed on θ. We notice that on the level of the FIM, the estimation of σ_v^2 and $\overline{\theta}$ is decoupled (J is block diagonal). The CRB $= J^{-1}$ is also block diagonal:

$$J^{-1}(\theta) = \begin{bmatrix} \frac{2\sigma_v^4}{n} & 0 \\ 0 & \sigma_v^2 \left[\left(\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}} \right) \left(\frac{\partial S^T(\overline{\theta})}{\partial \overline{\theta}} \right)^T \right]^{-1} \end{bmatrix} . \tag{1.271}$$

This means that the CRB for the estimation of one of the two subsets of parameters, σ_v^2 and $\overline{\theta}$, does not change whether we know the other parameters or not. Indeed, if we would have calculated the CRB for σ_v^2 or for $\overline{\theta}$ separately, assuming the other one was known, then we would have found the same values.

1.4.7 The Deterministic Linear Model

The linear model is a special case of the non-linear model discussed in the previous subsection with $\theta = \overline{\theta}$, $S(\theta) = H\theta$. We get: $Y = H\theta + V$, with $V \sim \mathcal{N}(0, C_{VV})$ where C_{VV} is constant. We refer to the Bayesian linear model for examples of linear models. The deterministic linear model does not differ from the Bayesian linear model, except that the parameters θ are now considered to be deterministic. We get for the conditional distribution

$$f_{\mathbf{Y}|\boldsymbol{\theta}}(Y|\theta) = f_{\mathbf{V}}(Y - H\theta) \tag{1.272}$$

and for the gradient and the ML estimate

$$\frac{\partial \ \ln \ f_{\mathbf{V}}(Y - H\theta)}{\partial \theta} = H^T C_{VV}^{-1}(Y - H\theta) = 0 \quad \Rightarrow \quad \widehat{\theta}_{ML} = (H^T C_{VV}^{-1} H)^{-1} \, H^T C_{VV}^{-1} Y \, . \tag{1.273}$$

The Hessian (which becomes deterministic here) is

$$\frac{\partial}{\partial \theta} \left(\frac{\partial \ \ln \ f_{\mathbf{V}}(Y - H\theta)}{\partial \theta} \right)^T = \underbrace{- H^T \underbrace{C_{VV}^{-1}}_{>0} H}_{>0} = -J < 0 \tag{1.274}$$

hence we have indeed a maximum. We should remark though that $J > 0$ requires that $n \geq m$: the number of measurements is at least equal to the number of unknowns. If $n < m$, then J is singular which means that certain linear combinations of the parameters can be estimated, but others cannot: the inverse of zero leads to ∞ in the CRB. If $n < m$ or if n does not exceed m by much, the Bayesian formulation (with the introduction of prior information to make J non-singular) is strongly recommended, if not mandatory. We shall assume here that J is nonsingular: $C_{VV} > 0$ and H of rank m. We get for the estimation error

$$\widetilde{\theta} = \theta - \widehat{\theta} = -(H^T C_{VV}^{-1} H)^{-1} \, H^T C_{VV}^{-1} V \ , \ \ E_{Y|\theta} \widetilde{\theta} = E_V \widetilde{\theta} = 0 \tag{1.275}$$

hence $\widehat{\theta}_{ML}$ is unbiased! The estimation error correlation matrix becomes

$$C_{\widetilde{\theta}\widetilde{\theta}} = R_{\widetilde{\theta}\widetilde{\theta}} = E_{Y|\theta} \widetilde{\theta}\widetilde{\theta}^T = E_V \widetilde{\theta}\widetilde{\theta}^T = (H^T C_{VV}^{-1} H)^{-1} = J^{-1} \tag{1.276}$$

which means that $\widehat{\theta}_{ML}$ is efficient since it reaches the Cramer-Rao lower bound. We can independently verify efficiency by checking (1.251):

$$\frac{\partial \ \ln \ f_{\mathbf{V}}(Y - H\theta)}{\partial \theta} = H^T C_{VV}^{-1} Y - H^T C_{VV}^{-1} H\theta = J(\widehat{\theta} - \theta) \tag{1.277}$$

hence once again $\widehat{\theta}_{ML}$ is efficient.

1.4.8 Asymptotic (Large Sample) Properties

The sample size is n. By asymptotic or large sample, we mean $n \to \infty$. For a finite sample size n, unbiasedness and efficiency are properties that don't occur easily. However, when the sample size becomes large, the picture looks more optimistic for some estimators, in particular for the ML estimator as we shall see.

A first asymptotic property is that of *asymptotical unbiasedness* :

$$\lim_{n\to\infty} b_n(\theta) = 0 \ , \quad \forall \theta \in \Theta \tag{1.278}$$

where we explicitly denote the dependence of the bias on the sample size n. An estimator may be biased but asymptotically unbiased as shown in the following example.

Example 1.14 Mean and Variance of Gaussian i.i.d. variables - Example 1.10 cont'd

In example 1.10, we found that the ML estimator for the variance of Gaussian i.i.d. variables with unknown mean also was the sample variance. The expected value for this estimator was

$$E[\widehat{\sigma^2}_{ML}|\mu, \sigma^2] \ = \ \frac{n-1}{n}\sigma^2 \ \neq \ \sigma^2 \tag{1.279}$$

hence the estimator was biased. However, when we investigate the bias as a function of sample size, we find

$$b_n \ = \ \frac{n-1}{n}\sigma^2 - \sigma^2 \ = \ -\frac{\sigma^2}{n} \stackrel{n\to\infty}{\longrightarrow} 0 \tag{1.280}$$

so that the sample variance $\widehat{\sigma^2}_{ML}$ is a biased, but asymptotically unbiased estimator of the variance. $\diamond$

As we have more and more measurements available, one may wonder if the estimate gets close to the true value. One may indeed hope that the availability of an infinity of data should be able to overcome the randomness in the relation $f(Y|\theta)$ between Y and θ. *Consistency* is about the convergence of $\widehat{\theta}_n = \widehat{\theta}(Y_n) \ \to \theta$. Since the $\widehat{\theta}_n$ form a series of random vectors, in order to investigate the convergence of the parameter estimates to the true deterministic value, we have to investigate the convergence of a series of random vectors. The convergence of a series of random quantities can occur in a number of ways:

- convergence in probability
- convergence in mean square

- convergence with probability one
- convergence in distribution

We shall first consider the first three forms of convergence. To each type of convergence corresponds a form of consistency of the parameter estimates. The sequence of estimates $\widehat{\theta}(Y_n)$ is said to be

- *simply* or *weakly consistent* if

$$\lim_{n\to\infty} \Pr_{Y_n|\theta}\left\{\|\widehat{\theta}(Y_n) - \theta\| < \epsilon\right\} = 1\,, \quad \forall \epsilon > 0,\ \forall \theta \in \Theta \tag{1.281}$$

  The limit of the probability that $\widehat{\theta}_n$ does not deviate from θ is one.

- *mean-square consistent* if

$$\lim_{n\to\infty} E_{Y_n|\theta}\|\widehat{\theta}(Y_n) - \theta\|^2 = 0\,, \quad \forall \theta \in \Theta \tag{1.282}$$

  The limit of the MSE is zero.

- *strongly consistent* if

$$\Pr_{Y_n|\theta}\{\lim_{n\to\infty} \widehat{\theta}(Y_n) = \theta\} = 1\,, \quad \forall \theta \in \Theta \tag{1.283}$$

  The probability that the limit of $\widehat{\theta}_n$ equals θ is one.

Any of these 3 types of consistency implies asymptotic unbiasedness: the parameter estimate cannot converge to the true value if the estimation error does not become zero on the average. E.g. for mean-square consistency:

$$\begin{aligned} E_{Y_n|\theta}\|\widehat{\theta}(Y_n) - \theta\|^2 = \|E_{Y_n|\theta}\widehat{\theta}(Y_n) - \theta\|^2 + E_{Y_n|\theta}\|\widehat{\theta}(Y_n) - E_{Y_n|\theta}\widehat{\theta}(Y_n)\|^2 \to 0 \\ \Rightarrow\ \lim_{n\to\infty} E_{Y_n|\theta}\widehat{\theta}(Y_n) = \theta \end{aligned} \tag{1.284}$$

The MSE decomposes into the square of the bias plus the variance. Both terms are nonnegative. Hence if their sum goes to zero, each term goes to zero.

Strong and mean-square consistency do not imply each other in general. Either of them implies weak consistency (e.g. use the Chebyshev inequality to show that mean-square consistency implies weak consistency), but not conversely. Except when Θ is bounded: then weak consistency implies mean-square consistency.

Example 1.15 Mean of Gaussian i.i.d. variables - Example 1.13 cont'd

We had i.i.d. $y_i \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = \mu$, σ^2 known. $\widehat{\mu}_{ML} = \overline{y}$. We have $E\,\widehat{\mu}_{ML} = \mu$: unbiased. And

$$\text{MSE } = Var(\widehat{\mu}_{ML}) = \frac{\sigma^2}{n} \overset{n\to\infty}{\longrightarrow} 0 \tag{1.285}$$

which implies mean-square consistency. ◇

Example 1.16 Mean of uniform i.i.d. variables - Example 1.11 cont'd

We had i.i.d. $y_i \sim U[\theta-\frac{1}{2},\theta+\frac{1}{2}]$, $\widehat{\theta}_{ML} = \frac{y_{min}+y_{max}}{2}$ It can be shown that

$$\begin{cases} y_{min} & \to & \theta-\frac{1}{2} & \text{in probability} \\ y_{max} & \to & \theta+\frac{1}{2} & \text{in probability} \end{cases} \tag{1.286}$$

$$\widehat{\theta}_{ML} \to \theta \quad \text{in probability}$$

which implies weak consistency. Mean-square consistency can also be shown. ◇

If $\widehat{\theta}_n$ is consistent, then $\widetilde{\theta} \to 0$ according to the corresponding mode of convergence. Now, different estimators can all have their estimation errors converge to zero, but some may nevertheless be better than others (converge to zero faster than others). In order to investigate the shrinking estimation error, we shall introduce a magnifying glass: we shall consider $d_n(\widehat{\theta}_n - \theta)$ where $0 < d_{n-1} \leq d_n \to \infty$ is a sequence of magnification factors that increases to infinity. Usually $d_n = \sqrt{n}$. For the magnified estimation error, we shall consider the fourth type of convergence, *convergence in distribution*. This type of convergence is weaker than the three other forms of convergence of sequences of random vectors mentioned before. We have convergence in distribution if the distribution of $d_n(\widehat{\theta}_n - \theta)$ converges to the distribution of some random vector ξ, and we shall write $d_n(\widehat{\theta}_n - \theta) \overset{in\ dist.}{\longrightarrow} \xi$. Convergence in distribution is a weak form of convergence in the sense that the random vector $d_n(\widehat{\theta}_n - \theta)$ itself does not converge to any specific random quantity, only its distribution converges to something. Nevertheless, when we have $d_n(\widehat{\theta}_n - \theta) \overset{in\ dist.}{\longrightarrow} \xi$, then the distribution of ξ will be a useful characterization of the limiting behavior of $\widehat{\theta}_n$.

We shall say that $\widehat{\theta}_n$ is *consistent asymptotically normal* (CAN) if $\widehat{\theta}_n$ is simply consistent and $d_n(\widehat{\theta}_n - \theta) \overset{in\ dist.}{\longrightarrow} \mathcal{N}(0, \Xi(\theta))$. This means that $\widehat{\theta}_n$ is asymptotically unbiased, and Ξ is the asymptotic normalized covariance of $\widetilde{\theta}_n$.

Strictly speaking, it is necessary to distinguish $\Xi(\theta)$ from $V(\theta) = \lim_{n\to\infty} d_n^2\, C_{\widetilde{\theta}\widetilde{\theta}}(\theta)$ which may not even exist for a CAN estimate (if $\widehat{\theta}_n$ is simply consistent but not mean-square consistent). $V(\theta)$ exists for a mean-square consistent $\widehat{\theta}_n$, but is in that case not necessarily equal to $\Xi(\theta)$. Nevertheless, for mean-square consistent estimates, we will often have that $\Xi(\theta) = V(\theta)$.

The *normalized asymptotic information matrix* is defined as $J_0(\theta) = \lim_{n\to\infty} \frac{1}{d_n^2} J_n(\theta)$ if the limit exists. An estimator is called *best asymptotically normal* (BAN) if it is CAN and $\Xi(\theta) = J_0^{-1}(\theta)$. A BAN estimator is also called *asymptotically efficient* since it reaches the normalized CRB asymptotically.

Under some regularity conditions (maximum of the likelihood function unique, y_i given θ i.i.d., or y_i and y_j i.i.d. as $|i-j| \to \infty$, ...) the ML estimate is strongly consistent and BAN with $d_n = \sqrt{n}$. In particular, the ML estimate is

- asymptotically unbiased
- asymptotically efficient (if i.i.d.: $J_n = nJ_1 \;\Rightarrow\; J_0 = J_1$)
- asymptotically normal

This means that usually in the large sample case, the ML estimator is a UMVUE and hence is as good as an estimator can get!

Example 1.17 Mean of Gaussian i.i.d. variables - Example 1.13 cont'd

We had i.i.d. $y_i \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = \mu$, σ^2 known. $\widehat{\mu}_{ML} = \overline{y}$.

$$\widehat{\mu}_{ML} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \quad \longrightarrow \quad \sqrt{n}(\widehat{\mu}_{ML} - \mu) \sim \mathcal{N}(0, \sigma^2) \ , \quad J_n = \frac{n}{\sigma^2} \Rightarrow J_0^{-1} = \sigma^2 \ . \tag{1.287}$$

In this example, the normalized estimation error has for any sample size a Gaussian distribution with zero mean, and a variance equal to the inverse of the normalized information matrix. Hence these properties continue to hold in the limit as $n \to \infty$. $\diamond$

1.4.9 Recap: Properties of Estimators and Estimation Techniques

Finite sample properties (finite n):

- *bias*: $b_{\widehat{\theta}}(\theta) = E_{Y|\theta}\widehat{\theta}(Y) - \theta \quad (= 0, \ \forall \theta \in \Theta$: unbiased)
- *error correlation*: $R_{\widetilde{\theta}\widetilde{\theta}} = E_{Y|\theta}\left(\widehat{\theta}(Y) - \theta\right)\left(\widehat{\theta}(Y) - \theta\right)^T$

  Cramer-Rao Bound : $\widehat{\theta}$ unbiased: $R_{\widetilde{\theta}\widetilde{\theta}} = C_{\widetilde{\theta}\widetilde{\theta}} = C_{\widehat{\theta}\widehat{\theta}}$

  $$C_{\widetilde{\theta}\widetilde{\theta}} \geq J^{-1}(\theta) \ , \quad J(\theta) = -E_{Y|\theta}\frac{\partial}{\partial \theta}\left(\frac{\partial \ln f(Y|\theta)}{\partial \theta}\right)^T \quad \text{information matrix} \tag{1.288}$$

  efficient: $C_{\widetilde{\theta}\widetilde{\theta}} = J^{-1}(\theta) \ , \ \forall \theta \in \Theta \quad \Rightarrow \quad \widehat{\theta}(Y)$ is UMVUE

Large sample (asymptotic, $n \to \infty$) properties:

- *asymptotically unbiased*: $\lim_{n\to\infty} b_{\widehat{\theta}}(\theta) = 0, \ \forall \theta \in \Theta$
- *consistency* (weak, in mean square, strong): $\Rightarrow$ asymptotically unbiased
- *asymptotic normality*:

$$\text{BAN} \left\{ \begin{array}{l} \diamond \text{ weakly consistent} \\ \diamond \text{ asymptotically normal} \\ \diamond \text{ asymptotically efficient} \end{array} \right\} \text{CAN}$$

Estimation Techniques:

- *Uniformly Minimum Variance Unbiased Estimator* (UMVUE): complicated (via "complete sufficient statistics")
- *Maximum likelihood* (ML): $\widehat{\theta}_{ML} = \arg\max_{\theta} f(Y|\theta)$

  Qualities:

- ◇ if $\exists$ efficient $\widehat{\theta} = \widehat{\theta}_{eff}$ and $\widehat{\theta}_{ML}$ is obtained from $\dfrac{\partial \ln f(Y|\theta)}{\partial \theta} = 0$
  $\Rightarrow\ \widehat{\theta}_{eff} = \widehat{\theta}_{ML} = \widehat{\theta}_{UMVUE}$
- ◇ $\widehat{\theta}_{ML}$ = BAN

Problems:

- ◇ what if $f(Y|\theta)$ is unknown?
- ◇ if $f(Y|\theta)$ is not concave (local maxima)

- simplified estimators:
  - ◇ *Method of Moments*
  - ◇ *Best Linear Unbiased Estimator* (BLUE) $\rightarrow$ linear model
  - ◇ *Least-Squares* (LS) $\rightarrow$ linear model

If $f(Y|\theta)$ is not known, then we are forced to consider an estimation strategy different from ML. We can use one of the three simpler estimators, depending on the amount of information we have. Another problem with ML that is often encountered in practice is that often we will not be able to determine the extrema analytically by setting the gradient of the likelihood function equal to zero. That means that we will have to employ a general optimization strategy to find the global maximum. The way these optimization algorithms work is that starting from an initial point in parameter space, they will lead to one of the local maxima close by the initialization. Indeed, the likelihood function is often not concave to such an extent that there are several local maxima. One can find the global maximum by exhaustively and repeatedly starting the optimization algorithm at a large number of points spread out over the parameter space. If we want to run the optimization algorithm only once, we should have an initialization that is reasonably close to the global maximum (and hence presumably close to the true parameter value) in such a way that the global maximum is the local maximum that is closest to the initialization in some sense. Such an initialization, which is in fact another parameter estimate, can be obtained by using one of the three simpler estimators described below. For one of these estimators to provide good initialization points for the optimization of the likelihood function, it is desirable (and sufficient in the large sample case) that the estimator is consistent.

1.4.10 The Method of Moments

The principle of the method of moments (MM) goes as follows. We have m unknown parameters $\theta = [\theta_1 \cdots \theta_m]^T$. The distribution of Y, $f(Y|\theta)$, depends on θ. Hence the moments (mean, second-order, higher-order) of Y also depend on θ. Take m moments

$$\mu = g(\theta) = \begin{bmatrix} g_1(\theta) \\ \vdots \\ g_m(\theta) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_m \end{bmatrix} \tag{1.289}$$

such that the function $g(.)$ is invertible, i.e. $\theta = g^{-1}(\mu)$: we can determine θ from μ. Estimate the moments: $\widehat{\mu}$, using any estimator (e.g. the sample moments). Then the estimate according to the method of moments is

$$\widehat{\theta}_{MM} = g^{-1}(\widehat{\mu}) \ . \tag{1.290}$$

The characteristic function of $Y|\theta$ is the (n-dimensional) Fourier transform of the probability density function $f(Y|theta)$. By developing the exponential into a Taylor series, the characteristic function can be developed into a series, the coefficients of which are the moments of increasing order. Hence, the specification of $f(Y|\theta)$ is equivalent to the specification of an infinite number of moments. The method of moments reduces this amount of information by concentrating on just a finite number (m) of moments.

Example 1.18 Estimation of the mixture parameter of a Gaussian mixture

We observe $y_i,\ i = 1,\ldots,n$ i.i.d., which are distributed according to $f(y|\theta)$, a mixture distribution with mixture parameter θ:

$$f(y|\theta) = (1-\theta)\phi_1(y) + \theta\phi_2(y) \ , \quad \phi_k(y) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{y^2}{2\sigma_k^2}}, k = 1,2 \tag{1.291}$$

We get for the estimation of any function $h(y)$ of y

$$E_f h(y) = \int h(y) f(y|\theta) dy = (1-\theta) \int h(y)\phi_1(y)dy + \theta \int h(y)\phi_2(y)dy = (1-\theta)\, E_{\phi_1} h(y) + \theta\, E_{\phi_2} h(y) \ . \tag{1.292}$$

We shall consider the following moment of y

$$\mu = E\left(y^2|\theta\right) = (1-\theta)\,\sigma_1^2 + \theta\,\sigma_2^2 = g(\theta) \ \Rightarrow \ \theta = g^{-1}(\mu) = \frac{\mu - \sigma_1^2}{\sigma_2^2 - \sigma_1^2} \tag{1.293}$$

where we used (1.291) with $h(y) = y^2$. The method of moments suggests the following estimate

$$\widehat{\theta}_{MM} = g^{-1}(\widehat{\mu}) = \frac{\widehat{\mu} - \sigma_1^2}{\sigma_2^2 - \sigma_1^2} \ , \quad \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 \tag{1.294}$$

where $\widehat{\mu}$ is the sample mean square value of y. To investigate the bias, consider:

$$E\widehat{\theta}_{MM} = \frac{1}{\sigma_2^2 - \sigma_1^2} E\widehat{\mu} - \frac{\sigma_1^2}{\sigma_2^2 - \sigma_1^2} = \frac{1}{\sigma_2^2 - \sigma_1^2}\mu - \frac{\sigma_1^2}{\sigma_2^2 - \sigma_1^2} = \theta \tag{1.295}$$

hence $\widehat{\theta}_{MM}$ is unbiased. In general, if $\widehat{\mu}$ is unbiased and $g(.)$ is linear, then $\widehat{\theta}_{MM}$ will be unbiased. To investigate the variance of $\widehat{\theta}_{MM}$, we get

$$Var\left(\widehat{\theta}_{MM}\right) = Var\left(\tfrac{1}{\sigma_2^2-\sigma_1^2}\widehat{\mu} - \tfrac{\sigma_1^2}{\sigma_2^2-\sigma_1^2}\right) = \tfrac{1}{\left(\sigma_2^2-\sigma_1^2\right)^2} Var\left(\widehat{\mu}\right) = \tfrac{1}{\left(\sigma_2^2-\sigma_1^2\right)^2} Var\left(\tfrac{1}{n}\sum_{i=1}^{n} y_i^2\right)$$

$$= \tfrac{1}{\left(\sigma_2^2-\sigma_1^2\right)^2} \sum_{i=1}^{n} Var\left(\frac{1}{n} y_i^2\right) = \frac{1}{\left(\sigma_2^2 - \sigma_1^2\right)^2} \sum_{i=1}^{n} \frac{1}{n^2} Var\left(y_i^2\right) = \frac{1}{\left(\sigma_2^2 - \sigma_1^2\right)^2} \frac{1}{n}\, Var\left(y^2\right) \tag{1.296}$$

Using (1.292), we get

$$Var(y^2) = Ey^4 - (Ey^2)^2 \ , \quad E(y^4|\theta) = (1-\theta)\, 3\sigma_1^4 + \theta\, 3\sigma_2^4 \tag{1.297}$$

where we used the fact that $Ey^4 = 3\sigma^4$ if $y \sim \mathcal{N}(0, \sigma^2)$. Combining with (1.296), we get

$$Var(\widehat{\theta}_{MM}) = \frac{3(1-\theta)\sigma_1^4 + 3\theta\sigma_2^4 - \left[(1-\theta)\sigma_1^2 + \theta\sigma_2^2\right]^2}{n\,(\sigma_1^2 - \sigma_2^2)^2} \stackrel{n\to\infty}{\longrightarrow} 0 \ . \tag{1.298}$$

Hence $\widehat{\theta}_{MM}$ is mean-square consistent. $\diamond$

Example 1.19 Estimation of the parameters of a sinusoid in white noise

We observe $y_k = s_k + v_k = A\cos(\omega k + \phi) + v_k \,, \quad k = 1, \ldots, n$. In what follows, only the first and second-order moments of V are needed: $EV = 0$, $EVV^T = \sigma_v^2 I_n$ in which σ_v^2 is an unknown parameter. We consider the amplitude A, the angular frequency ω and the noise variance σ_v^2 as deterministic, and the phase ϕ as random. We are interested in estimating the following parameters

$$\theta = \begin{bmatrix} A \\ \omega \\ \sigma_v^2 \end{bmatrix} \ , \quad \Theta : A > 0, \omega \in [0, \pi], \sigma_v^2 > 0 \ . \tag{1.299}$$

We consider ϕ as a random nuisance parameter (we don't know its value, but we are not interested in it) with distribution $\phi \sim \mathcal{U}[0, 2\pi]$ independent of V, θ. The randomness in this problem is completely described by

$$f(Y, \phi|\theta) = f(\phi|\theta)\, f(Y|\theta, \phi) = f(\phi)\, f_{\mathbf{V}|\boldsymbol{\sigma_v^2}}(Y - S(A, \omega, \phi)|\sigma_v^2) \tag{1.300}$$

and expectations are taken according to this $f(Y, \phi|\theta)$. Note that the uniformly distributed phase turns the signal $A\cos(\omega k + \phi)$ and hence y_k into a stationary sequence. We get in particular for the mean:

$$E_{Y,\phi|\theta}\, y_k = AE\cos(\omega k + \phi) + Ev_k = 0 \ . \tag{1.301}$$

The covariance sequence of y_k is

$$\begin{aligned} r_{yy}(i) &= Ey_k y_{k+i} = A^2 E\cos(\omega k + \phi)\cos(\omega k + \phi + \omega i) \\ &\quad + AE\cos(\omega k + \phi)Ev_{k+i} + AE\cos(\omega k + \phi + \omega i)Ev_k + Ev_k v_{k+i} \\ &= \frac{A^2}{2} E\cos(2\omega k + 2\phi + \omega i) + \frac{A^2}{2} E\cos(\omega i) + \sigma_v^2 \delta_{i0} \\ &= \frac{A^2}{2}\cos(\omega i) + \sigma_v^2 \delta_{i0} \end{aligned} \tag{1.302}$$

Note that for sample estimates of the mean and the covariance sequence to converge to the statistical averages, we need in principle to dispose of realizations of y_k that correspond to different realizations of ϕ. However, it turns out that if we have a large enough number

n of samples y_k corresponding to one realization of ϕ, then the influence of the value of ϕ disappears asymptotically as $n \to \infty$.

The particular moments we shall consider here are

$$\mu = \begin{bmatrix} r_{yy}(0) \\ r_{yy}(1) \\ r_{yy}(2) \end{bmatrix} = \begin{bmatrix} \frac{A^2}{2} + \sigma_v^2 \\ \frac{A^2}{2}\cos(\omega) \\ \frac{A^2}{2}\cos(2\omega) \end{bmatrix} = g(\theta) \; . \tag{1.303}$$

It turns out that these moments constitute an invertible function of our parameters θ; the inverse function $\theta = g^{-1}(\mu)$ can be specified as

$$\omega = \begin{cases} \arccos\left(\dfrac{r_{yy}(2) + \sqrt{r_{yy}^2(2) + 8r_{yy}^2(1)}}{4\, r_{yy}(1)}\right) & , \, r_{yy}(1) \neq 0 \\ \dfrac{\pi}{2} & , \, r_{yy}(1) = 0 \end{cases}$$

$$A = \begin{cases} \sqrt{\dfrac{2\, r_{yy}(1)}{\cos(\omega)}} & , \, r_{yy}(1) \neq 0 \\ \sqrt{-2\, r_{yy}(2)} & , \, r_{yy}(1) = 0 \end{cases} \quad , \quad \sigma_v^2 = r_{yy}(0) - \frac{A^2}{2} \tag{1.304}$$

We can use the sample moments as $\widehat{\mu}$:

$$\widehat{r}_{yy}(i) = \frac{1}{n}\sum_{k=1}^{n-i} y_k y_{k+i} \, , \; i = 0, 1, 2 \tag{1.305}$$

leading to $\widehat{\theta}_{MM} = g^{-1}(\widehat{\mu})$ via (1.304). $\diamond$

Since $\widehat{\mu}$ is easy to compute, $\widehat{\theta}_{MM} = g^{-1}(\widehat{\mu})$ is straightforward to compute if μ (and hence $g(.)$) is chosen well, hence $\widehat{\theta}_{MM}$ is easy to determine in general and easy to implement. The estimate is furthermore unique (no problem with local maxima as in ML). The MM estimate has no optimality properties in general but is usually consistent (since $\widehat{\mu}$ is consistent). If the performance of $\widehat{\theta}_{MM}$ is not satisfactory, one can use $\widehat{\theta}_{MM}$ as initialization in an iterative optimization procedure that finds $\widehat{\theta}_{ML}$.

1.4.11 The Best Linear Unbiased Estimator (BLUE)

Just like the UMVUE is the deterministic analog of the MMSE estimator in the Bayesian case, the BLUE is the deterministic analog of the LMMSE estimator. The three characteristics of the BLUE are

- *linear*: $\widehat{\theta}(Y) = AY \quad (A : m \times n)$

- *unbiased*: $E_{Y|\theta}\widehat{\theta} = A\,E(Y|\theta) = \theta\,, \quad \forall\theta \in \Theta$
- *best = minimum variance*: $\min C_{\widetilde{\theta}\widetilde{\theta}}$

Remarks:

(i) The BLUE is inferior to the UMVUE unless the UMVUE is linear.

(ii) Generalizations of the formulation are possible to allow non-linear functions of the data: if $X = g(Y)$, then $\widehat{\theta}(Y) = A\,X = A\,g(Y)$ is linear in X, which is a non-linear transformation of the data. For instance, an estimator that is linear in Y is inappropriate if $\theta \neq 0$ and $E(Y|\theta) = 0$ since in that case there is no linear unbiased estimator.

Example 1.20 Example of a non-linear data transformation.

We observe $y_i \sim \mathcal{N}(0,\sigma^2)$ i.i.d., $\theta = \sigma^2$, $Y = [y_1 \cdots y_n]^T$. If we take a linear estimator $\widehat{\sigma^2} = AY$, then $E_{Y|\sigma^2}\widehat{\sigma^2} = AE(Y|\sigma^2) = 0 \neq \sigma^2$: no linear unbiased estimator $\widehat{\sigma^2}$ exists. However, if we let $x_i = y_i^2$, $X = [y_1^2 \cdots y_n^2]^T$, then we get for an estimator linear in X

$$\widehat{\sigma^2} = A\,X \;\Rightarrow\; E_{Y|\sigma^2}\widehat{\sigma^2} = AE\left(X|\sigma^2\right) = \sigma^2\,A\,\mathbf{1} = \sigma^2 \;\Rightarrow\; A\,\mathbf{1} = 1\,. \tag{1.306}$$

Hence all estimators linear in X that satisfy $A\,\mathbf{1} = 1$ are unbiased. It turns out that for this problem: $\widehat{\sigma^2}_{UMVUE} = \frac{1}{n}\mathbf{1}^T X = \widehat{\sigma^2}_{BLUE}$ $(A = \frac{1}{n}\mathbf{1}^T)$. $\diamond$

Based on the above specification of the BLUE problem, it is not possible to derive the BLUE. In order to be able to derive the BLUE, it is necessary to introduce two assumptions, arising from considerations of the moments of order one and two. Consider the unbiasedness requirement: $AE(Y|\theta) = \theta\,, \quad \forall\theta \in \Theta$. This implies that $AE(Y|\theta)$ is a linear function of θ. If we want the unbiasedness requirement to hold for a large class of linear unbiased estimators (many A satisfying $AE(Y|\theta) = \theta$, $\forall\theta \in \Theta$), then the following assumption arises naturally:

Assumption 1 : $E(Y|\theta) = H\,\theta\,, \quad (H : n \times m)$

meaning that $E(Y|\theta)$ should be a linear function of θ. The unbiasedness requirement then implies that $A\,H = I_m$ $(\Rightarrow\; n \geq m)$. In order to be able to find the linear unbiased estimator with the lowest variance, we will have to introduce an assumption concerning the covariance matrix of $Y|\theta$. The covariance matrix of the estimation errors is

$$\begin{aligned} C_{\widetilde{\theta}\widetilde{\theta}} = C_{\widehat{\theta}\widehat{\theta}} &= E_{Y|\theta}\left(\widehat{\theta} - E_{Y|\theta}\widehat{\theta}\right)\left(\widehat{\theta} - E_{Y|\theta}\widehat{\theta}\right)^T \\ &= E_{Y|\theta}\left(AY - AE(Y|\theta)\right)\left(AY - AE(Y|\theta)\right)^T \\ &= A\,E_{Y|\theta}\left(Y - E(Y|\theta)\right)\left(Y - E(Y|\theta)\right)^T A^T = A\,C_{YY}(\theta)\,A^T\,. \end{aligned} \tag{1.307}$$

In order to be able to minimize $C_{\widetilde{\theta}\widetilde{\theta}}$ w.r.t. A in such a way that the solution for A does not depend on the true parameter value θ, we have to introduce the following assumption:

Assumption 2 : $C_{YY}(\theta) = c(\theta)\,C$

where $c(\theta)$ $(> 0, \forall\theta)$ is a scalar function of θ, and the matrix C is constant w.r.t. θ.

We are now able to formulate the BLUE optimization problem

$$\min_{\widehat{\theta}: E_{Y|\theta}\widehat{\theta}(Y)=\theta} C_{\widetilde{\theta}\widetilde{\theta}} \tag{1.308}$$

as

$$\min_{A:\, AH=I} A\, C\, A^T \ . \tag{1.309}$$

In order to solve this problem, it will be convenient to introduce the notion of a matrix square-root. The matrix $B\,(n \times n)$ is called a matrix square root of the matrix $C = C^T > 0\,(n \times n)$ if $C = B\,B^T$. We shall introduce the following notation: $B = C^{1/2}$, $C^{T/2} = (C^{1/2})^T$. This allows us to write $C = C^{1/2}C^{T/2}$, $C^{-1} = C^{-T/2}C^{-1/2}$. A possible choice for the matrix square root is a triangular matrix; the matrix square-root is called Cholesky factor in that case.

Consider now a vector space of $m \times n$ matrices with matrix inner product $\langle X_1, X_2 \rangle = X_1 X_2^T$. Take $X_1 = H^T C^{-T/2}$, $X_2 = A\, C^{1/2}$. With $AH = I$, we get:

$$\left\langle \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \right\rangle = \begin{bmatrix} H^T C^{-T/2} \\ A\, C^{1/2} \end{bmatrix} \begin{bmatrix} H^T C^{-T/2} \\ A\, C^{1/2} \end{bmatrix}^T = \begin{bmatrix} H^T C^{-1} H & I \\ I & ACA^T \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \tag{1.310}$$

From the Schur Complements Lemma, we know that $R_{22} \geq R_{21} R_{11}^{-1} R_{12}$ with equality iff $X_2 = R_{21} R_{11}^{-1} X_1$. Hence

$$\min_{A:\, AH=I} A\, C^T A^T = \left(H^T C^{-1} H\right)^{-1} \quad \text{for} \quad A = \left(H^T C^{-1} H\right)^{-1} H^T C^{-1} \ . \tag{1.311}$$

Or also, since $\widehat{\theta} = A\, Y$,

$$\widehat{\theta}_{BLUE} = \left(H^T C^{-1} H\right)^{-1} H^T C^{-1} Y \quad \text{with} \quad C_{\widetilde{\theta}\widetilde{\theta}} = c(\theta) \left(H^T C^{-1} H\right)^{-1} = \left(H^T C_{YY}^{-1} H\right)^{-1} \ . \tag{1.312}$$

Example 1.21 Example 1.20 cont'd.

We had $y_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d., $\theta = \sigma^2$, $x_i = y_i^2$, $\widehat{\sigma^2} = A\, X$. So the role of Y is here played by X. We can verify that the BLUE assumptions are satisfied:

$$E\left(X|\sigma^2\right) = \mathbf{1}\, \sigma^2 = H\, \theta\,, \quad C_{XX} = 2\sigma^4\, I = c(\theta)\, C \tag{1.313}$$

with $c(\theta) = 2\sigma^4$ and $C = I$. We find

$$\widehat{\sigma^2}_{BLUE} = \left(H^T C^{-1} H\right)^{-1} H^T C^{-1} X = \frac{1}{n} \mathbf{1}^T X \text{ with } C_{\widehat{\sigma^2}\widehat{\sigma^2}}(\sigma^2) = \left(H^T C_{XX}^{-1} H\right)^{-1} = \frac{2\sigma^4}{n} \,. \tag{1.314}$$

as we announced before. ◇

We recapitulate the BLUE assumptions:

$$\begin{cases} (1)\ E(Y|\theta) = H\, \theta \\ (2)\ C_{YY}(\theta) = c(\theta)\, C \end{cases}$$

We only need to know the first two moments of $f(Y|\theta)$, and they need to satisfy these assumptions. The higher-order moments of $f(Y|\theta)$ don't need to be known and can be arbitrary functions of θ.

The BLUE Linear Model

We get for the specification of the linear model:

$$Y = H\,\theta + V\ ,\ \ EV = 0\,,\ EVV^T = C_{VV} \tag{1.315}$$

where EV and C_{VV} do not depend on θ. We note that only the first two moments of V need to be specified. We can verify that the BLUE assumptions are satisfied:

$$\left\{\begin{array}{l} E\,(Y|\theta) \ = \ H\,\theta \\ C_{YY}(\theta) = E_{\mathbf{Y}|\boldsymbol{\theta}}\,(Y - E\,(Y|\theta))\,(Y - E\,(Y|\theta))^T = E_{\mathbf{V}} VV^T = C_{VV} = C\ \ (c(\theta) \equiv 1) \end{array}\right. \tag{1.316}$$

We find

$$\widehat{\theta}_{BLUE} = \left(H^T C_{VV}^{-1} H\right)^{-1} H^T C_{VV}^{-1} Y \text{ with } C_{\widetilde{\theta}\widetilde{\theta}} = \left(H^T C_{VV}^{-1} H\right)^{-1}\ . \tag{1.317}$$

If $V \sim \mathcal{N}(0, C_{VV})$, then $\widehat{\theta}_{BLUE} = \widehat{\theta}_{ML}$ and the estimator is efficient, hence also $\widehat{\theta}_{BLUE} = \widehat{\theta}_{UMVUE}$ for the linear model.

1.4.12 Least-Squares Estimation

In the Method of Moments, the information in the complete description of the distribution $f(Y|\theta)$ is reduced to a number of moments. In the BLUE approach, the required information gets further reduced to only the first- and second-order moments, which furthermore need to depend on θ in a specified way. The third approach, the least-squares method, can be specified without any statistical knowledge at all, as far as the derivation of the estimator is concerned. For its performance evaluation however, a statistical context will of course be required.

To apply the least-squares method, we shall assume that we dispose of a model that describes the relations between the unknown parameters θ and the measurements Y. In general, we can assume a nonlinear model that is described by a set of n (implicit) model functions:

$$g_k(\theta, Y) = 0\ ,\quad k = 1, \ldots, n\ . \tag{1.318}$$

Example 1.22 A sinusoidal signal

Consider the following sinusoid

$$y_k = s_k = A\,\cos(\omega k + \phi) \tag{1.319}$$

where the unknown parameter that we are interested in this time is the frequency: $\theta = \omega$. It can be shown that this sinusoid satisfies the following homogeneous (RHS=0) difference equation

$$y_k - 2\,\cos\omega\, y_{k-1} + y_{k-2} = g_k(\theta, Y) = 0\ . \tag{1.320}$$

Indeed, taking the z-transform of both sides of the difference equation (which holds $\forall k$), we get

$$\mathrm{Y}(z) - 2\,\cos\omega\, z^{-1}\mathrm{Y}(z) + z^{-2}\mathrm{Y}(z) = (z^2 - 2\,\cos\omega\, z + 1)z^{-2}\mathrm{Y}(z) = 0\ . \tag{1.321}$$

This leads us to the characteristic equation:

$$z^2 - 2\cos\omega\, z + 1 = 0 \quad \Rightarrow \quad z_{1,2} = e^{\pm j\omega} \tag{1.322}$$

with roots that occur as a pair of complex numbers since the equation has real coefficients. The solution of the difference equation is of the form

$$y_k = \alpha\, z_1^k + \alpha^*\, z_2^k = \frac{Ae^{j\phi}}{2} e^{j\omega k} + \frac{Ae^{-j\phi}}{2} e^{-j\omega k} \tag{1.323}$$

where the combination coefficients also occur as a pair of complex conjugate numbers so that y_k is a real signal. Any complex number can be written as $\alpha = \frac{Ae^{j\phi}}{2}$ for some amplitude A and some phase ϕ. The second expression in (1.323) can now be seen to be the sinusoid appearing in (1.319). ◇

One important application of the method of least-squares is an extension of the Method of Moments to the case where we want to use more moments then there are parameters to be estimated. The measurements in that case are the moments μ_k and the model equations are $\mu_k - g_k(\theta) = 0$.

Now assume that the model functions are in fact verified for clean measurement signals, but that in fact we can only make the measurements with a limited precision, up to some error, so that if we plug the actual measurements Y into the model functions, the model equations (1.318) are actually not perfectly satisfied:

$$g_k(\theta, Y) \;=\; e_k \;\neq\; 0\,. \tag{1.324}$$

The error term e_k is called the equation error, it is the error that prevents the model equation from being satisfied. For instance, referring to the previous example, if we measure the sinusoidal signal up to some error v_k: $y_k = s_k + v_k$, then the equation error is $e_k = v_k - 2\cos\omega\, v_{k-1} + v_{k-2}$. Or in the extension of the method of moments, we will not actually use the true values of the moments but estimates $\widehat{\mu}_k$ thereof (the sample moments), so that the model functions $\widehat{\mu}_k - g_k(\theta) \neq 0$. Finally, the equation error can also come about because our model functions $g_k(\theta, Y)$ are possibly only approximately true.

Let us collect all the (scalar valued) model functions and equation errors into vectors

$$G(\theta, Y) = \begin{bmatrix} g_1(\theta, Y) \\ \vdots \\ g_n(\theta, Y) \end{bmatrix}, \quad E = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}. \tag{1.325}$$

The least-squares method now chooses that value of the parameters that makes the sum of the squares of the equation errors, $\sum_{k=1}^{n} e_k^2 = E^T E$, as small as possible. Of course, in order to have a criterion that can be optimized w.r.t. θ, we need to express the equation errors in terms of the model functions. So the Least-Squares (LS) estimate is obtained as

$$\widehat{\theta}_{LS} \;=\; \arg\min_{\theta} G^T(\theta, Y)\, G(\theta, Y)\,. \tag{1.326}$$

This cost function can be generalized by introducing a positive definite symmetric weighting matrix $W = W^T > 0$. This leads to the Weighted Least-Squares (WLS) estimate

$$\widehat{\theta}_{WLS} \;=\; \arg\min_{\theta} G^T(\theta, Y)\, W\, G(\theta, Y) \;. \tag{1.327}$$

For a general nonlinear model, the WLS criterion may be a complicated function of θ and its minimization may require general optimization tools. The solution may be easily found explicitly in the following case.

Model Linear in the Parameters

Consider the case in which the model functions depend linearly on the parameters θ:

$$g_k(\theta, Y) \;=\; f_k(Y) - C_k(Y)\,\theta \;, \quad f_k(Y) : \; 1 \times 1, \; C_k(Y) : \; 1 \times m \;. \tag{1.328}$$

In the previous example, the model function in (1.320) depends on ω in a nonlinear way. However, if we reparameterize the problem and take $\theta = 2\cos\omega$ then the model functions in (1.320) can be seen to be of the form in (1.328) with

$$\left\{ \begin{array}{rcl} f_k(Y) & = & y_k + y_{k-2} \\ C_k(Y) & = & y_{k-1} \end{array} \right. \tag{1.329}$$

To collect all the equation errors into a vector, consider

$$F(Y) = \left[\begin{array}{c} f_1(Y) \\ \vdots \\ f_n(Y) \end{array} \right] \;:\; n \times 1 \,, \quad H(Y) = \left[\begin{array}{c} C_1(Y) \\ \vdots \\ C_n(Y) \end{array} \right] \;:\; n \times m \;\; . \tag{1.330}$$

The WLS estimate can now be found as

$$\widehat{\theta}_{WLS} \;=\; \arg\min_{\theta} \left[F(Y) - H(Y)\,\theta\right]^T W \; \left[F(Y) - H(Y)\,\theta\right] \;. \tag{1.331}$$

An analysis of different situations is in order now. More explicitly, the vector of errors can be written as

$$F(Y) - H(Y)\,\theta = \left[\begin{array}{c} f_1 \\ \vdots \\ f_n \end{array} \right] - \left[\begin{array}{c} C_1 \\ \vdots \\ C_n \end{array} \right] \left[\begin{array}{c} \theta_1 \\ \vdots \\ \theta_m \end{array} \right] = \left[\begin{array}{c} f_1 \\ \vdots \\ f_n \end{array} \right] - [H_1 \cdots H_m] \left[\begin{array}{c} \theta_1 \\ \vdots \\ \theta_m \end{array} \right] = \left[\begin{array}{c} e_1 \\ \vdots \\ e_n \end{array} \right] = E \tag{1.332}$$

where we denoted column i of $H(Y)$ as H_i. We have n equations in m unknowns. In general, the following three cases can occur:

- $n > m$: *overdetermined* case

  In this case, an exact fit of the equations is impossible in general. That is why we introduced the least-squares fit. We shall assume that H has full rank (= full column rank) so that there exists a unique solution.

- $n = m$: *exactly determined* case

  If H has full rank ($\Rightarrow$ H^{-1} exists), then $\widehat{\theta} = H^{-1}F$ is the unique solution of the model equations ($E = 0$). No averaging of the equation errors occurs though.

- $n < m$: *underdetermined* case

  In this case, ∞^{m-n} solutions exist (their differences lie in the nullspace of H). We can make the solution unique by requiring it to have minimum norm $\left\|\widehat{\theta}\right\|$.

We shall henceforth assume that we are dealing with a genuine LS problem: $n > m$, and we assume also that rank$(H) = m$.

The WLS problem is of the form $\min_{\theta} \xi(\theta)$ with

$$\xi(\theta) = [F(Y) - H(Y)\,\theta]^T\, W\, [F(Y) - H(Y)\,\theta] \ . \tag{1.333}$$

We find an extremum by setting the gradient equal to zero:

$$\begin{aligned} &\frac{\partial \xi}{\partial \theta} = -2H^T(Y)\, W\, [F(Y) - H(Y)\,\theta] = 0 \\ &\Rightarrow \quad \widehat{\theta}_{WLS} = \left(H^T(Y)\, W\, H(Y)\right)^{-1} H^T(Y)\, W\, F(Y) \ . \end{aligned} \tag{1.334}$$

We check the Hessian $= \dfrac{\partial}{\partial \theta}\left(\dfrac{\partial \xi}{\partial \theta}\right)^T = 2H^T(Y)\, W\, H(Y) > 0$ since $W > 0$ and $H(Y)$ has full rank. Hence the extremum is a local minimum and it is the only one, hence it is the global minimum.

The Linear Model

Further simplifications of the model in (1.328) can be obtained by taking $F(Y) = Y$ and $H(Y) = H$, a matrix that does not depend on Y. We get the familiar linear model

$$\begin{cases} y_k = C_k\,\theta + v_k \ , \ k = 1, \ldots, n \quad v_k = e_k = \text{ error} \\ Y = H\,\theta + V \quad \begin{cases} H\,\theta = \text{ signal component} \\ V = \text{ noise} \end{cases} \end{cases} \tag{1.335}$$

In the linear model, $H\theta$ is considered as the clean signal component of the measurements Y and V as the associated measurement noise. The WLS estimate is immediately obtained from (1.334) as

$$\widehat{\theta}_{WLS} = \left(H^T W\, H\right)^{-1} H^T W\, Y \ . \tag{1.336}$$

We have seen several examples of the linear model already. Here are some more explicit examples.

Example 1.23 Amplitude and phase estimation of a noisy sinusoid

The measurements are

$$\begin{aligned} y_k &= A\cos(\omega k + \phi) + v_k \\ &= A\cos\phi\,\cos(\omega k) - a\sin\phi\,\sin(\omega k) + v_k \\ &= \underbrace{[\cos(\omega k) \quad \sin(\omega k)]}_{C_k} \underbrace{\begin{bmatrix} A\cos\phi \\ -A\sin\phi \end{bmatrix}}_{\theta} + v_k \end{aligned} \tag{1.337}$$

If we consider the frequency to be known, then the measurements are linear in the parameters $A\cos\phi$ and $-A\sin\phi$. $\diamond$

Example 1.24 Line fitting

In this classical example of LS fitting, we want to fit a straight line through a collection of points (x_k, y_k) in the plane. We can write

$$y_k = a\,x_k + b + v_k = \underbrace{[x_k \quad 1]}_{C_k} \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{\theta} + v_k\ . \tag{1.338}$$

We want to determine the parameters a and b of the line by minimizing the sum of the squares of the errors v_k between the actual ordinates y_k and their model value $a\,x_k + b$ if the point would have laid on the line $y = a\,x + b$. $\diamond$

No statistical information (about V) is needed to derive $\hat{\theta}_{WLS}$. However, in order to evaluate its performance, we need to introduce a stochastic context: we assume V random with $E\,V = 0$, $E\,VV^T = C_{VV}$. (The performance analysis below is in fact valid for general $F(Y)$ and $H(Y)$, but we shall keep the notation of the linear model). The parameter estimation error can be written as

$$\hat{\theta}_{WLS} - \theta = \left(H^T W\,H\right)^{-1} H^T W\ (H\,\theta + V) - \theta = \left(H^T W\,H\right)^{-1} H^T W\,V\ . \tag{1.339}$$

Taking expected value, we find

$$E\,\hat{\theta}_{WLS} - \theta = \left(H^T W\,H\right)^{-1} H^T W E\,V = 0 \tag{1.340}$$

which means that under the above assumptions, the WLS estimator is unbiased. We can compute the estimation error covariance matrix and we find

$$C_{\tilde{\theta}\tilde{\theta}}(W) = C_{\tilde{\theta}\tilde{\theta}}(W) = \left(H^T W\,H\right)^{-1} H^T W\,C_{VV}\,W\,H\,\left(H^T W\,H\right)^{-1} \tag{1.341}$$

where we have stressed the fact that $C_{\tilde{\theta}\tilde{\theta}}$ depends on the choice of the weighting matrix W. A natural question to ask now is what is the optimal choice for W. The answer is

$$W = C_{VV}^{-1}\ :\ \ C_{\tilde{\theta}\tilde{\theta}}(W) \geq C_{\tilde{\theta}\tilde{\theta}}(C_{VV}^{-1}) = \left(H^T C_{VV}^{-1} H\right)^{-1} \tag{1.342}$$

which is not surprising: we find the BLUE since we are satsifying the BLUE assumptions.

To show the optimality of $W = C_{VV}^{-1}$, consider again a vector space of $m \times n$ matrices with matrix inner product $\langle X_1, X_2 \rangle = X_1 X_2^T$. Take $X_1 = H^T C_{VV}^{-T/2}$, $X_2 = H^T W C_{VV}^{1/2}$. We get

$$\begin{aligned} \left\langle \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \right\rangle &= \begin{bmatrix} H^T C_{VV}^{-T/2} \\ H^T W C_{VV}^{1/2} \end{bmatrix} \begin{bmatrix} H^T C_{VV}^{-T/2} \\ H^T W C_{VV}^{1/2} \end{bmatrix}^T \\ &= \begin{bmatrix} H^T C_{VV}^{-1} H & H^T W H \\ H^T W H & H^T W C_{VV} W H \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \end{aligned} \tag{1.343}$$

From the Schur Complements Lemma, $R_{21}^{-1} R_{22} R_{12}^{-1} \geq R_{11}^{-1}$ with equality iff $X_2 = R_{21} R_{11}^{-1} X_1$. Hence,

$$\min_W \left(H^T W H \right)^{-1} H^T W C_{VV} W H \left(H^T W H \right)^{-1} = \left(H^T C_{VV}^{-1} H \right)^{-1} \tag{1.344}$$

with equality iff

$$\left(H^T W H \right)^{-1} H^T W = \left(H^T C_{VV}^{-1} H \right)^{-1} H^T C_{VV}^{-1} \ . \tag{1.345}$$

$W = C_{VV}^{-1}$ is one optimal choice (in general: $W = C_{VV}^{-1} + P_H^{\perp} Q P_H^{\perp}$, any Q).

If we introduce further statistical knowledge, we get more optimality properties:

$$\widehat{\theta}_{WLS} = \widehat{\theta}_{ML} \ \text{ if } \ V \sim \mathcal{N}(0, W^{-1}) \text{ and independent of } \theta \ . \tag{1.346}$$

Due to efficiency, furthermore $\widehat{\theta}_{WLS} = \widehat{\theta}_{UMVUE}$ in this case.

1.5 Choice of an Estimator

If we review the definitions for the information matrix in the Bayesian and the deterministic case, then we can relate the stochastic (Bayesian) information matrix to the deterministic information matrix as follows:

$$J_{stoch} \ = \ J_{prior} + J_Y \ = \ J_{prior} + E_{\boldsymbol{\theta}} \, J_{det}(\theta) \tag{1.347}$$

where $E_{\boldsymbol{\theta}}$ is over the prior distribution of θ. As J_{det} is roughly (or exactly if the $y_i|\theta$ are i.i.d.) proportional to n, J_{det} dominates as $n \gg 1$. Hence if lots of data are available, then (uncertain) prior information is of little relevance (asymptotically irrelevant). Hence, the deterministic approach is sufficient in that case. If on the other hand little data is available, and certainly if fewer measurements are available than the number of parameters to be estimated, then a prior distribution (even if invented) is necessary to regularize the problem, to avoid the singularity of J_{det}, by (averaging it over θ and) adding J_{prior} to it.
For Bayesian estimation:

- $\widehat{\theta}_{MMSE}$ preferable
- $\widehat{\theta}_{MAP}$ easier to calculate
- $\widehat{\theta}_{LMMSE}$ simple, acceptable if everything $\approx$ Gaussian (model $\approx$ linear)

For deterministic (classical) estimation:

- Maximum Likelihood (ML) if possible
- if ML too complex or if a good initialization is required for an iterative optimization of ML or if $f(Y|\theta)$ is not available: Method of Moments or BLUE or Least-Squares
- Linear Gaussian model: all good estimators identical

1.6 Further Reading

The classical reference on the subject of parameter estimation is [1]. Excellent recent references are [2] and [3]. A very thorough book, that only treats deterministic parameter estimation however, is [4]. This reference gives for instance a very good introduction to the notion of information matrix. The multivariate Gaussian distribution and the Bayesian linear model are discussed in e.g. [2] and [3]. The unusual unbiasedness condition in the Bayesian case was found in [5]. The orthogonality condition for (nonlinear) MMSE estimation appears in [6]. The UMVUE and its computation via complete sufficient statistics are discussed in [2] and [3]. An extensive discussion of performance bounds, properties of the ML estimator and of the Method of Moments can be found in [4]. The BLUE is discussed in [7] and [8]. The LS method is treated in detail in [2]. Another reference on these topics is [9]

1.7 Problems

1.1. *Gaussian Random Variables.* Let $X \sim \mathcal{N}(m_X, C_{XX})$, i.e. $X = [x_1 \dots x_m]^T$ is a random vector that is normally distributed with mean m_X and covariance matrix C_{XX}. Let Y be a linear transformation of X,

$$Y = AX + b\,. \tag{1.348}$$

The dimension n of Y can be different from m, the dimension of X. So $A \in \mathcal{R}^{n\times m}, b \in \mathcal{R}^n$. Derive the distribution of Y. Give its mean and covariance matrix.

1.2. *Soft Decisions.* Let $y = \theta + v$ where θ and v are independent random variables. v is Gaussian with zero mean and unit variance and θ takes on the values ± 1 with equal probability. Show that the MMSE estimate is

$$\widehat{\theta}_{MMSE} \;=\; \tanh y \;=\; \frac{e^y - e^{-y}}{e^y + e^{-y}}\;. \tag{1.349}$$

This problem models a very simple communication problem in which we have binary communication over an additive white Gaussian noise channel. In detection (decision) theory, one would try, on the basis of the observation y, to determine whether θ is $+1$ or -1. Here, we provide a continuously valued estimate for θ. For $y \gg 1$, $\widehat{\theta} \to +1$: the chance for v to take on the value $y-1$ is small, but still much higher than the chance for v to take on the value $y+1$. Similarly, $\widehat{\theta} \to -1$ as $y \ll -1$. When y is around 0, we are really not sure about the value of θ and we prefer to give a very cautious estimate rather than running the risk of making a wrong decision (in which case $|\widetilde{\theta}| = 2$). In any case, $\widehat{\theta} \in [-1, 1]$ and $\widehat{\theta}$ can be considered as a convex combination of the two possible values ± 1. Though for detection purposes, we will have to make up our mind about $\theta = \pm 1$, for the purpose of blindly adapting a channel equalizer, it might be wiser to base the adaptation on the estimated symbols discussed here, rather than on the detected symbols (hard decisions). See the next chapter.

1.3. *Bayesian Fourier Analysis.* Many signals exhibit cyclical behavior. It is common practice to determine the presence of strong cyclical components by employing Fourier analysis. Large Fourier coefficients are indicative of strong components. Here we show how Fourier analysis can result from a well-founded estimation approach. Assume we have the following signal model

$$y_k \;=\; a\,\cos(2\pi f_0 k) + b\,\sin(2\pi f_0 k) + v_k\;,\quad k = 1, \dots, n \tag{1.350}$$

where f_0 is assumed to be an integer multiple of $\frac{1}{n}$ (as in the DFT), excepting 0 or $\frac{1}{2}$ (for which $\sin(2\pi f_0 k)$ would be identically zero), and v_k is WGN (white Gaussian noise): $V = [v_1 \cdots v_n]^T \sim \mathcal{N}(0, \sigma_v^2 I_n)$. It is desired to estimate $\theta = [a\ b]^T$. We assume that a and b are random variables with prior pdf $f(\theta) \leftrightarrow \mathcal{N}(0, \sigma_\theta^2 I_2)$ and θ is assumed independent of V. This type of model is referred to as a *Rayleigh fading sinusoid* because the sinusoid amplitude $\sqrt{a^2 + b^2}$ has a Rayleigh distribution (the phase is uniform in $[0, 2\pi)$). This model is frequently used to model a sinusoid that has propagated through a dispersive medium.

Find the MMSE estimator $\widehat{\theta}_{MMSE}$ and its associated covariance matrix.

1.4. *Imperfect Geiger Counter.* During a certain time period, a radioactive source emits n radioactive particles, and an imperfect Geiger counter only counts $k \leq n$ of them. The problem is to estimate the parameter n from the measurement k. To proceed, we assume that n is drawn from an a priori distribution that is a Poisson distribution with mean λ (average number of particles emitted during the given time period):

$$P(n) \;=\; e^{-\lambda}\frac{\lambda^n}{n!}\;,\;\; n \geq 0\;. \tag{1.351}$$

This a priori information may be based on the physics of the problem and λ depends on the specific radioactive material we are considering. The imperfection of the counter, which counts only k particles out of n emitted and hence misses $n-k$ of them, is given by the following conditional distribution

$$P(k|n) \;=\; \binom{n}{k} p^k (1-p)^{n-k}\;,\;\; 0 \leq k \leq n \tag{1.352}$$

which is a binomial distribution.

Find the MMSE estimator $\widehat{n}(k)$ and its associated variance.

1.5. *Rayleigh variable plus Gaussian noise.* A certain signal level θ is observed in additive noise:

$$y_i \;=\; \theta + v_i\;,\;\; i = 1, \ldots, n\;. \tag{1.353}$$

The a priori distribution of θ is a Rayleigh distribution:

$$f(\theta) \;=\; \begin{cases} \frac{\theta}{\sigma_\theta^2} e^{-\frac{\theta^2}{2\sigma_\theta^2}} & ,\;\; \theta \geq 0 \\ 0 & ,\;\; \theta < 0\;. \end{cases} \tag{1.354}$$

The additive noise samples v_i are i.i.d., Gaussian: $v_i \sim \mathcal{N}(0, \sigma_v^2)$, and independent of θ. Find the Maximum A Posteriori estimator $\widehat{\theta}_{MAP}$. Show that for $n \gg 1$, $\widehat{\theta}_{MAP} \approx \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

1.6. *Poisson process.* Consider a network node where messages are passing and let the single observation y be the time between the passage of two consecutive messages. y has an exponential distribution

$$f(y|\theta) \;=\; \begin{cases} \theta\, e^{-\theta\, y} & ,\;\; y \geq 0 \\ 0 & ,\;\; y < 0 \end{cases} \tag{1.355}$$

where the parameter θ represents the average frequency of messages passing by the node we are considering. This average frequency has itself an a priori distribution (over the different nodes in the network) which we take to be exponential with average value $1/\lambda$:

$$f(\theta) \;=\; \begin{cases} \lambda\, e^{-\lambda\, \theta} & ,\;\; \theta > 0 \\ 0 & ,\;\; \theta \leq 0\;. \end{cases} \tag{1.356}$$

Find the a posteriori distribution of θ given the measurement y and then find the estimators $\widehat{\theta}_{MMSE}$, $\widehat{\theta}_{MAP}$, $\widehat{\theta}_{ABS}$ which correspond to the mean, the mode and the median

of this distribution (to find the last one, the root of a transcendental equation will have to be found (by e.g. Matlab)). Observe how these three Bayes estimators are similar, but not identical.

1.7. *LMMSE Estimation for Non-Centered Data.* The Linear Minimum Mean Square Error Estimate is given either without constant term

$$\widehat{\theta}_{LMMSE} = H\,Y = R_{\theta Y} R_{YY}^{-1} Y \tag{1.357}$$

or with constant term

$$\widehat{\theta}_{AMMSE} = H\,Y + g = m_\theta + C_{\theta Y} C_{YY}^{-1}\,(Y - m_Y) \tag{1.358}$$

where $R_{XY} = E\,XY^T$ and $C_{XY} = E(X - m_X)(Y - m_Y)^T$. As we see, (1.358) seems more complicated than (1.357). We can nevertheless include (1.358) into (1.357) by considering the augmented problem

$$\widehat{\theta}_{AMMSE} \;=\; \widehat{\theta}_{LMMSE'} \;=\; H'\,Y' \tag{1.359}$$

where $H' = [H \;\; g]$ and $Y' = [Y^T \;\; 1]^T$. Show that for this augmented problem, the estimator of the form (1.357) becomes the estimator in (1.358).

1.8. *LMMSE Estimation in the Singular Case.* If the correlation matrix of the data is singular (i.e. R_{YY} has no inverse), there is no unique optimal filter. In this case, all solutions of $H\,R_{YY} = R_{\theta Y}$ will have the form:

$$H = H_0 + C \tag{1.360}$$

where C is any matrix such that $C\,R_{YY} = 0$ and H_0 is a particular solution. Show that no matter which such H we use, we obtain the same estimate $\widehat{\theta} = HY$ and the same mean square error.

1.9. *Other Minimization Criteria related to MMSE.* Show that the LMMSE estimate

$$\widehat{\theta} = R_{\theta Y} R_{YY}^{-1} Y \tag{1.361}$$

which minimizes the criterion

$$E\,\left(\theta - \widehat{\theta}\right)\left(\theta - \widehat{\theta}\right)^T \tag{1.362}$$

also minimizes

(i) $\quad \mathrm{tr}\left\{E\,\left(\theta - \widehat{\theta}\right)\left(\theta - \widehat{\theta}\right)^T\right\} \;=\; E\,\left(\theta - \widehat{\theta}\right)^T\left(\theta - \widehat{\theta}\right)$

(ii) $\quad \det\left\{E\,\left(\theta - \widehat{\theta}\right)\left(\theta - \widehat{\theta}\right)^T\right\}$

(iii) $\quad E\,\left(\theta - \widehat{\theta}\right)^T W\left(\theta - \widehat{\theta}\right) \quad , \qquad W = W^T > 0\;.$

Hint: For (ii) note that $A - B \geq 0$ implies $\det(A) \geq \det(B)\;.$

1.10. *Sinusoid in White Noise (ML).* An important estimation problem in signal processing is the estimation of amplitude, phase and frequency of a sinusoidal signal model. Let the following measurements be taken:

$$y_k = s_k + v_k = A\cos(\omega k + \phi) + v_k \quad , \quad k = 0, \ldots, n-1 \quad . \tag{1.363}$$

The signal parameters ($A > 0, \phi \in [0, 2\pi), \omega > 0$) are unknown. Let the noise term $v_k \sim \mathcal{N}(0, \sigma^2)$ be a sequence of i.i.d. random variables. The noise variance $\sigma^2 > 0$ is also unknown. So the parameter vector is

$$\theta = \begin{bmatrix} \sigma^2 \\ A \\ \phi \\ \omega \end{bmatrix} \quad . \tag{1.364}$$

Given θ, the measurements are a sequence of independent $\mathcal{N}(s_k, \sigma^2)$ random variables. The problem is to find the maximum likelihood estimate $\widehat{\theta}_{ML}$ and its covariance matrix. We shall do this in several steps.
(i) Find the log likelihood function $\ln f(Y|\theta)$ where $Y = [y_0 \cdots y_{n-1}]^T$.
(ii) Given A, ϕ, and ω maximize $\ln f(Y|\theta)$ w.r.t. σ^2 to find $\widehat{\sigma^2} = \widehat{\sigma^2}(Y, A, \phi, \omega)$.
(iii) Show that maximizing $\ln f(Y|\widehat{\sigma^2}(Y, A, \phi, \omega), A, \phi, \omega)$ with respect to A, ϕ, ω amounts to the same thing as minimizing $\widehat{\sigma^2}(Y, A, \phi, \omega)$ w.r.t. A, ϕ and ω.
(iv) In developing the terms in $\widehat{\sigma^2}(Y, A, \phi, \omega)$, neglect the term $\sum_{k=0}^{n-1} \cos(2\omega k + 2\phi)$ (because it is normally not proportional to n).
(v) Minimize $\widehat{\sigma^2}(Y, A, \phi, \omega)$ w.r.t. ϕ to obtain $\widehat{\phi}(Y, \omega)$. It will be useful to introduce $\mathcal{Y}(\omega) = \sum_{k=0}^{n-1} y_k e^{-j\omega k} = |\mathcal{Y}(\omega)| e^{j \arg \mathcal{Y}(\omega)}$, the Fourier transform of the sequence y_k .
(vi) Minimize $\widehat{\sigma^2}(Y, A, \widehat{\phi}(Y, \omega), \omega)$ with respect to A to obtain $\widehat{A}(Y, \omega)$.
(vii) Minimize $\widehat{\sigma^2}(Y, \widehat{A}(Y, \omega), \widehat{\phi}(Y, \omega), \omega)$ with respect to ω to obtain $\widehat{\omega}_{ML}(Y)$.
Then the other ML estimates are

$$\begin{aligned} \widehat{A}_{ML}(Y) &= \widehat{A}(Y, \widehat{\omega}_{ML}(Y)) \\ \widehat{\phi}_{ML}(Y) &= \widehat{\phi}(Y, \widehat{\omega}_{ML}(Y)) \\ \widehat{\sigma^2}_{ML}(Y) &= \widehat{\sigma^2}(Y, \widehat{A}_{ML}(Y), \widehat{\phi}_{ML}(Y), \widehat{\omega}_{ML}(Y)) \quad . \end{aligned} \tag{1.365}$$

The Fisher information matrix for deterministic parameters is

$$J(\theta) = E_{\mathbf{Y}|\boldsymbol{\theta}} \left(\frac{\partial}{\partial \theta} \ln f(Y|\theta) \right) \left(\frac{\partial}{\partial \theta} \ln f(Y|\theta) \right)^T = -E_{\mathbf{Y}|\boldsymbol{\theta}} \frac{\partial}{\partial \theta} \left(\frac{\partial}{\partial \theta} \ln f(Y|\theta) \right)^T \quad . \tag{1.366}$$

The Cramer-Rao bound (CRB) specifies a lower bound on the covariance matrix of an unbiased estimator $\widehat{\theta}$ as

$$R_{\widehat{\theta}\widehat{\theta}} = E_{\mathbf{Y}|\boldsymbol{\theta}} (\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^T \geq CRB_{\widehat{\theta}\widehat{\theta}} = J^{-1}(\theta) \quad . \tag{1.367}$$

The ML estimator is asymptotically (as $n \to \infty$) unbiased and efficient ($R_{\widetilde{\theta\theta}} = J^{-1}$). Note that the information matrix $J(\theta)$ is evaluated at the true parameters θ. At the true paramater values, $y_k - s_k = v_k$, and $E_{\mathbf{Y}|\boldsymbol{\theta}} = E_{\mathbf{V}}$.

(viii) Show that the information matrix equals (hint: use the second expression in (1.366))

$$J(\theta) = \begin{bmatrix} \frac{n}{2\sigma^4} & 0 & 0 & 0 \\ 0 & \frac{1}{\sigma^2}\sum_{k=0}^{n-1}\cos^2(\omega k+\phi) & \frac{-A}{2\sigma^2}\sum_{k=0}^{n-1}\sin(2\omega k+2\phi) & \frac{-A}{2\sigma^2}\sum_{k=0}^{n-1}k\sin(2\omega k+2\phi) \\ 0 & \frac{-A}{2\sigma^2}\sum_{k=0}^{n-1}\sin(2\omega k+2\phi) & \frac{A^2}{\sigma^2}\sum_{k=0}^{n-1}\sin^2(\omega k+\phi) & \frac{A^2}{\sigma^2}\sum_{k=0}^{n-1}k\sin^2(\omega k+\phi) \\ 0 & \frac{-A}{2\sigma^2}\sum_{k=0}^{n-1}k\sin(2\omega k+2\phi) & \frac{A^2}{\sigma^2}\sum_{k=0}^{n-1}k\sin^2(\omega k+\phi) & \frac{A^2}{\sigma^2}\sum_{k=0}^{n-1}k^2\sin^2(\omega k+\phi) \end{bmatrix} \tag{1.368}$$

When $n \gg \frac{1}{\sin\omega}$, the following approximation can be shown to be valid. It is obtained by keeping the most significant term (as a function of n) in the diagonal elements. In the off-diagonal element J_{ij} the term of the same order as $\sqrt{J_{ii}J_{jj}}$ is kept.

$$J \approx \begin{bmatrix} \frac{n}{2\sigma^4} & 0 & 0 & 0 \\ 0 & \frac{n}{2\sigma^2} & 0 & 0 \\ 0 & 0 & \frac{nA^2}{2\sigma^2} & \frac{n^2A^2}{4\sigma^2} \\ 0 & 0 & \frac{n^2A^2}{4\sigma^2} & \frac{n^3A^2}{6\sigma^2} \end{bmatrix} . \tag{1.369}$$

(ix) Using the above approximation for J, compute $CRB_{\widetilde{\theta\theta}}$.

(x) Since $E\left(\widehat{\theta}_i - \theta_i\right)^2 = \left(R_{\widetilde{\theta\theta}}\right)_{ii} \geq \left(CRB_{\widetilde{\theta\theta}}\right)_{ii}$, $i = 1,2,3,4$, how do the CRBs of the four parameters compare (when considered as a function of n) ?

Assume we divide the set of parameters into two groups $\theta = \begin{bmatrix} a \\ b \end{bmatrix}$, $J = J_{\theta\theta} = \begin{bmatrix} J_{aa} & J_{ab} \\ J_{ba} & J_{bb} \end{bmatrix}$. If the parameters a would be known, then the information matrix for the parameters b is still J_{bb} and hence $CRB_{\tilde{b}\tilde{b}} = J_{bb}^{-1}$ when a is known whereas $CRB_{\tilde{b}\tilde{b}} = \left(J_{\theta\theta}^{-1}\right)_{bb}$ when a is also to be estimated.

(xi) For each of the four parameters, comment on how its CRB decreases if some or all of the other three parameters would be known.

(xii) The estimated sinusoid is $\widehat{s}_k = \widehat{A}\cos(\widehat{\omega}k + \widehat{\phi})$. Using a first-order expansion, we can write

$$\widehat{s}_k - s_k = [\cos(\omega k+\phi) \quad -A\sin(\omega k+\phi) \quad -Ak\sin(\omega k+\phi)] \begin{bmatrix} \widehat{A}-A \\ \widehat{\phi}-\phi \\ \widehat{\omega}-\omega \end{bmatrix} . \tag{1.370}$$

Show that the a posteriori variance on the sinusoid, $E\,(\widehat{s}_k - s_k)^2$, is bounded below as follows (using the result from **(ix)**, assuming an unbiased estimator $\widehat{\theta}$)

$$E(\widehat{s}_k - s_k)^2 \geq \frac{2\sigma^2}{n} + \frac{6\sigma^2}{n}\sin^2(\omega k + \phi)\left(1 - \frac{2k}{n}\right)^2 \quad . \tag{1.371}$$

Hence the variance is the smallest ($\approx \frac{2\sigma^2}{n}$) in the middle of the sequence ($k = \frac{n}{2}$) and the largest near both ends (fluctuating between $\frac{2\sigma^2}{n}$ and $\frac{8\sigma^2}{n}$).

1.11. *Amplitude Estimation.* Amplitude modulation consists essentially in multiplying a known signal $\{s_k\}$ with an amplitude θ which contains the information. Due to perturbations in the transmission medium, the received signal y_k can be written as

$$y_k = \theta s_k + v_k \quad , \tag{1.372}$$

where v_k is some zero-mean noise, and s_k is a deterministic signal. The task of the optimal receiver consists in determining the amplitude θ from a number of observations $\{y_1,\ y_2, \ldots, y_n\}$. We can write (1.372) for $k = 1, \cdots, n$ in vector form as

$$Y \ = \ S\,\theta + V \quad , \tag{1.373}$$

where $Y = [y_1\ y_2 \cdots y_n]^T$ and similarly for S and V.
(i) Give the Weighted Least-Squares estimate $\widehat{\theta}_{WLS}$ of the amplitude for some symmetric positive definite weighting matrix W.
(ii) Let $C = E\,VV^T$ be the covariance matrix of the noise vector ($E\,V = 0$). Give the variance of the amplitude estimator $\widehat{\theta}_{WLS}$. For which W is it minimal?
(iii) We assume that $V \sim \mathcal{N}(0, C)$. Give the maximum likelihood estimate $\widehat{\theta}_{ML}$ of θ. Is the Cramer-Rao bound attained by this estimator?
(iv) We shall now consider the amplitude θ to be a random variable with mean m_θ and variance σ_θ^2 and θ is uncorrelated with V which is a zero-mean noise with covariance C. S is still deterministic. Compute the affine MMSE estimator $\widehat{\theta}_{AMMSE}$ of θ from Y and the corresponding MSE.
(v) Give the MMSE estimate $\widehat{\theta}_{MMSE}$ of θ and the corresponding MSE when assuming that $\theta \sim \mathcal{N}(m_\theta, \sigma_\theta^2)$, $V \sim \mathcal{N}(0, C)$ and independent of θ. Compare to the result from **(vi)**. What happens when $\sigma_\theta^{-2} \to 0$?

Chapter 2

Optimal Filtering

In the first chapter, we have seen the estimation of a finite number of (random) parameters given (stochastically) related measurements. In the previous chapter, we let the number of parameters go to infinity when we estimated the autocorrelation function. In this chapter, we again let the number of parameters go to infinity and we consider the estimation of a random process given a related random process. Such an estimation operation can be called a filtering operation. Classical filter design is mostly based on distortion criteria. Here, we shall consider the Mean Squared Error (MSE) criterion. A filter that is designed to minimize the MSE is called an optimal filter.

Optimal filtering is a subject area that was addressed by prof. Norbert Wiener from MIT in the 1930s. The problem posed to him came (as often) from a military context. The problem was to point an anti-aircraft machine gun to a plane (with trajectory x_k) on the basis of inaccurate measurements (y_k) of the plane trajectory. The issue was to filter the noisy position measurements with a causal filter to produce a reasonable estimate of the trajectory. Furthermore, since the gunload takes a while to reach its destination, the trajectory estimate had to predicted ahead into the future. Prof. Wiener solved the causal filtering problem in the early 1930s. We shall first consider the simpler optimal filtering problem without causality constraints.

2.1 Noncausal Wiener Filtering

Consider two stochastic processes $\{x_k,\, k \in \mathcal{Z}\}$ and $\{y_k,\, k \in \mathcal{Z}\}$ that are correlated. We observe the process $\{y_k,\, k \in \mathcal{Z}\}$ but we are interested in the process $\{x_k,\, k \in \mathcal{Z}\}$ that we cannot observe. We shall consider estimating $\{x_k,\, k \in \mathcal{Z}\}$ from $\{y_k,\, k \in \mathcal{Z}\}$. Since $\{x_k,\, k \in \mathcal{Z}\}$ is a discrete-time random process, each of its samples can be considered as a random variable. Therefore, the estimation problem becomes again one of parameter estimation where x_k at any time instant k can be considered a random parameter:

$$\{y_n,\, n \in \mathcal{Z}\} \quad \rightarrow \quad \widehat{x}_k \ . \tag{2.1}$$

The estimation criterion we consider is the Minimum Mean-Squared Error criterion and hence the MMSE estimator is the conditional mean: $\widehat{x}_k = E[x_k | y_n,\, n \in \mathcal{Z}]$. Since this estimator is in general a nonlinear operator on the data $\{y_n,\, n \in \mathcal{Z}\}$, optimal estimation is often called

nonlinear filtering in this context. However, we shall assume the processes to be zero mean and we shall restrict the estimator to be a linear functional of the data. This allows us to depict the estimation problem as in Fig. 2.1. Indeed, a linear combination of the $\{y_n,\, n \in \mathcal{Z}\}$ can be considered a filtering operation in which the combination coefficients constitute the filter impulse response.

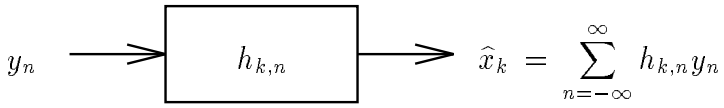


Figure 2.1: The linear MMSE (Wiener) filter.

The filter coefficients have two time indices. Indeed, the set of optimal filter coefficients may depend on the time instant k at which we are estimating the sample x_k. Therefore, the first index of $h_{k,n}$ is k. The second index n indicates which sample y_n we are combining. So we shall determine the filter coefficients $h_{k,n}$ from the following LMMSE estimation problem

$$\min_{h_{k,n}} E(x_k - \widehat{x}_k)^2 \;=\; \min_{h_{k,n}} E(x_k - \sum_{n=-\infty}^{\infty} h_{k,n} y_n)^2 \;. \tag{2.2}$$

We shall avoid carrying out the optimization explicitly (which would involve checking the positive definiteness of an infinite dimensional Hessian) by invoking the orthogonality property of LMMSE estimation: the optimal $h_{k,n}$ satisfy the orthogonality conditions

$$E\,(x_k - \widehat{x}_k) y_m \;=\; 0\;,\;\; \forall m \in \mathcal{Z} \tag{2.3}$$

or hence

$$E\,\widehat{x}_k y_m \;=\; \sum_{n=-\infty}^{\infty} h_{k,n}\, E\, y_n y_m \;=\; E\, x_k y_m\;,\;\; \forall m \in \mathcal{Z}\;. \tag{2.4}$$

At this point, we shall assume that the processes $\{x_k,\, k \in \mathcal{Z}\}$ and $\{y_k,\, k \in \mathcal{Z}\}$ are jointly stationary. The orthogonality conditions (2.4) become

$$\sum_{n=-\infty}^{\infty} h_{k,n}\, r_{yy}(n-m) \;=\; r_{xy}(k-m)\;,\;\; \forall m \in \mathcal{Z}\;. \tag{2.5}$$

This set of equations is again often called the *normal equations*. Performing the substitution $k-m \;\rightarrow\; m$ leads to

$$\sum_{n=-\infty}^{\infty} h_{k,n}\, r_{yy}(n-k+m) \;=\; r_{xy}(m)\;,\;\; \forall m \in \mathcal{Z} \tag{2.6}$$

while the substitution $n-k \;\rightarrow\; -n$ leads to

$$\sum_{n=-\infty}^{\infty} h_{k,k-n}\, r_{yy}(m-n) \;=\; r_{xy}(m)\;,\;\; \forall m \in \mathcal{Z}\;. \tag{2.7}$$

At this point, we notice that in the equations (2.7) the coefficients do not depend on k. Only the unknowns $h_{k,k-n}$ depend on k. This means that the solution for the $h_{k,k-n}$ is the same for any k. Hence,

$$h_{k,k-n} \;=\; h_{0,-n} \;=\; h_n \tag{2.8}$$

(since $h_{0,-n}$ only depends on n, we can call it h_n). This means that when the processes are (wide-sense) stationary, the optimal linear filter is time-invariant and satisfies

$$\sum_{n=-\infty}^{\infty} h_n \, r_{yy}(m-n) \;=\; r_{xy}(m) \;, \quad \forall m \in \mathcal{Z} \;. \tag{2.9}$$

This is an infinite set of equations in an infinite number of unknowns h_n. However, we can recognize the LHS of equation (2.9) to be a convolution. We can transform this convolution into a simple product by taking z-transforms. Let

$$\mathrm{S}_{xy}(z) \;=\; \sum_{m=-\infty}^{\infty} r_{xy}(m) z^{-m} \;\;, \quad \mathrm{H}(z) \;=\; \sum_{m=-\infty}^{\infty} h_m z^{-m} \;\;. \tag{2.10}$$

z-transforming both sides of equation (2.9) leads to $\mathrm{H}(z)\mathrm{S}_{yy}(z) = \mathrm{S}_{xy}(z)$ and hence

$$\mathrm{H}(z) \;=\; \frac{\mathrm{S}_{xy}(z)}{\mathrm{S}_{yy}(z)} \tag{2.11}$$

The Fourier transform of $\widehat{x}_k = \sum_{m=-\infty}^{\infty} h_m y_{k-m}$ (the Fourier transform is the z-transform evaluated at $z = e^{j2\pi f}$) is

$$\widehat{X}(f) \;=\; H(f) Y(f) \;\;, \quad H(f) \;=\; \frac{S_{xy}(f)}{S_{yy}(f)} \tag{2.12}$$

where $Y(f) = \mathrm{Y}(e^{j2\pi f}) = \sum_{m=-\infty}^{\infty} y_m e^{-j2\pi f m}$ etc. The form of (2.12) reminds of the LMMSE problem in the scalar case: let x and y be scalar random variables, then the LMMSE estimator is

$$\widehat{x} \;=\; h\, y \;\;, \quad h \;=\; \frac{R_{xy}}{R_{yy}} \;. \tag{2.13}$$

At any particular frequency f, the estimation problem in (2.12) is indeed a scalar estimation problem of estimating $X(f)$ from $Y(f)$ and one can show the following result (using the Fourier Transform Correlation Theorem (Chapter 2, Signal Modeling and Coding)):

$$\frac{R_{X(f)Y(f)}}{R_{Y(f)Y(f)}} \;=\; \frac{S_{xy}(f)}{S_{yy}(f)} \tag{2.14}$$

(the numerators on both sides are not equal and neither are the denominators, but the ratios are equal). So the Wiener filtering problem of one random process from another is like a scalar LMMSE estimation of the Fourier transforms of those processes at every frequency.

The orthogonality property of the LMMSE estimator implies

$$E\,(x_k - \widehat{x}_k)\widehat{x}_k \;=\; \sum_{n=-\infty}^{\infty} h_n \underbrace{E\,(x_k - \widehat{x}_k) y_{k-n}}_{=\,0} \;=\; 0 \;\;\Rightarrow\;\; E\,x_k\widehat{x}_k \;=\; E\,\widehat{x}_k^2 \;. \tag{2.15}$$

This allows us to demonstrate the following Pythagorean property

$$\begin{aligned} \text{MMSE} \;=\; E\,\widetilde{x}_k^2 \;&=\; E\,(x_k-\widehat{x}_k)^2 \;=\; E\,(x_k-\widehat{x}_k)x_k - \underbrace{E\,(x_k-\widehat{x}_k)\widehat{x}_k}_{=\,0} \\ &=\; E\,x_k^2 - E\,x_k\widehat{x}_k \;=\; E\,x_k^2 - \underbrace{E\,\widehat{x}_k^2}_{\geq\,0} \;\leq\; E\,x_k^2 . \end{aligned} \tag{2.16}$$

$Ex_k^2 = r_{xx}(0)$ is the MSE if we have no observations. In that case, $\widehat{x}_k = Ex_k = 0$ is our best estimator, leading indeed to Ex_k^2 as MSE. $E\widehat{x}_k^2 = r_{\widehat{x}\widehat{x}}(0) \geq 0$ is the reduction in MSE by estimating x_k using the Wiener filter on the data $\{y_n,\, n \in \mathcal{Z}\}$.

Further insight can be obtained by analyzing the MMSE in the frequency domain. Indeed

$$\text{MMSE} \;=\; r_{xx}(0) - r_{\widehat{x}\widehat{x}}(0) \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx}(f)df - \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{\widehat{x}\widehat{x}}(f)df\ . \tag{2.17}$$

Since $\widehat{x}_k = h_k * y_k$, we have for the power spectral density functions

$$S_{\widehat{x}\widehat{x}}(f) \;=\; |H(f)|^2 S_{yy}(f) \;=\; |S_{xy}(f)|^2/S_{yy}(f)\ . \tag{2.18}$$

Let us introduce the *cross-power spectral density coefficient*

$$\rho_{xy}(f) \;=\; \frac{S_{xy}(f)}{\sqrt{S_{xx}(f)S_{yy}(f)}} \tag{2.19}$$

which is defined as zero whenever $S_{xx}(f) = 0$ or $S_{yy}(f) = 0$. From (2.17),(2.18),(2.19), we get

$$\text{MMSE} \;=\; r_{\widetilde{x}\widetilde{x}}(0) \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{\widetilde{x}\widetilde{x}}(f)df \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} S_{xx}(f)\left[1-|\rho_{xy}(f)|^2\right]df\ . \tag{2.20}$$

In particular we find that

$$1-|\rho_{xy}(f)|^2 \;=\; \frac{S_{\widetilde{x}\widetilde{x}}(f)}{S_{xx}(f)} \geq 0 \tag{2.21}$$

which shows that $|\rho_{xy}(f)| \leq 1$. Hence $\rho_{xy}(f)$ is normalized and it can in fact be interpreted as the correlation coefficient between $X(f)$ and $Y(f)$. Since now $1-|\rho_{xy}(f)|^2 \in [0,1]$, (2.20) shows how the power spectral density of x_k gets attenuated as a function of frequency to obtain the power spectral density of the estimation error $\widetilde{x}_k$. At frequencies where $X(f)$ and $Y(f)$ are strongly correlated, $S_{\widetilde{x}\widetilde{x}}(f)$ will be significantly reduced w.r.t. $S_{xx}(f)$ whereas this will not be the case at frequencies where $X(f)$ and $Y(f)$ are hardly correlated.

2.1.1 Signal in Noise

More insight can be gained by considering the specific case of estimating a signal x_k by filtering a noisy measurement $y_k = x_k + v_k$ of it, see Fig. 2.2. The additive noise v_k is considered to have zero mean and to be uncorrelated with the signal of interest x_k: $E\,x_n v_k = 0,\ \forall n,k$. The quantities that determine the Wiener filter are

$$\begin{aligned} r_{xy}(n) &= r_{xx}(n) + \underbrace{r_{xv}(n)}_{=0} = r_{xx}(n) && ,\ \mathrm{S}_{xy}(z) = \mathrm{S}_{xx}(z) \\ r_{yy}(n) &= r_{xx}(n) + \underbrace{r_{xv}(n)}_{=0} + \underbrace{r_{vx}(n)}_{=0} + r_{vv}(n) \\ &= r_{xx}(n) + r_{vv}(n) && ,\ \mathrm{S}_{yy}(z) = \mathrm{S}_{xx}(z) + \mathrm{S}_{vv}(z) \end{aligned} \tag{2.22}$$

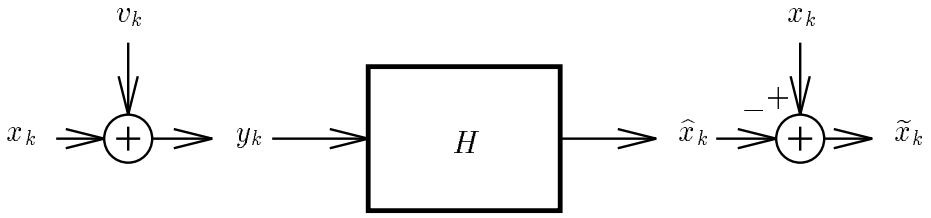


Figure 2.2: Wiener filtering for a signal in noise.

From (2.11) we can now write the Wiener filter as

$$H(f) \;=\; \frac{S_{xx}(f)}{S_{xx}(f)+S_{vv}(f)} \;\in\; [0,1]\;. \tag{2.23}$$

The optimum filter can indeed be interpreted as a frequency dependent weighting function that varies as the signal-to-noise ratio varies with frequency. It is possible to show that

$$\frac{1}{S_{\widetilde{x}\widetilde{x}}(f)} \;=\; \frac{1}{S_{xx}(f)}+\frac{1}{S_{vv}(f)} \;\geq\; \frac{1}{S_{xx}(f)} \tag{2.24}$$

which also leads to the following expression for the MMSE

$$\text{MMSE} \;=\; E\,\widetilde{x}_k^2 \;=\; \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{S_{xx}(f)\,S_{vv}(f)}{S_{xx}(f)+S_{vv}(f)}df\;. \tag{2.25}$$

The MMSE can be seen to be smaller than both $E\,x_k^2$ and $E\,v_k^2$.

Example 2.1 Bandlimited Random Process in White Noise

Consider a signal x_k with a triangular power spectral density function as indicated in Fig. 2.3. More precisely,

$$S_{xx}(f) \;=\; \begin{cases} A(1-\frac{|f|}{f_c}) & ,\; |f|\leq f_c \\ 0 & ,\; f_c\leq|f|\leq\frac{1}{2} \end{cases} \tag{2.26}$$

which is furthermore periodic with period one. What we receive is the noisy measurement $y_k = x_k + v_k$. The additive noise v_k is white with variance σ_v^2: $S_{vv}(f)=\sigma_v^2$. The task is to filter y_k so that the filter output $\widehat{x}_k$ approximates x_k well.

Classical filter design is based on the notion of distortion. So in the classical, non-statistical approach, we would design a filter that passes x_k without distortion but for the rest cuts out the noise as much as possible. Since x_k is bandlimited, we can choose an ideal low-pass filter $I(f)$ matched to the bandwidth f_c of x_k:

$$I(f) \;=\; \begin{cases} 1 & ,\; |f|\leq f_c \\ 0 & ,\; f_c<|f|\leq\frac{1}{2} \end{cases} \tag{2.27}$$

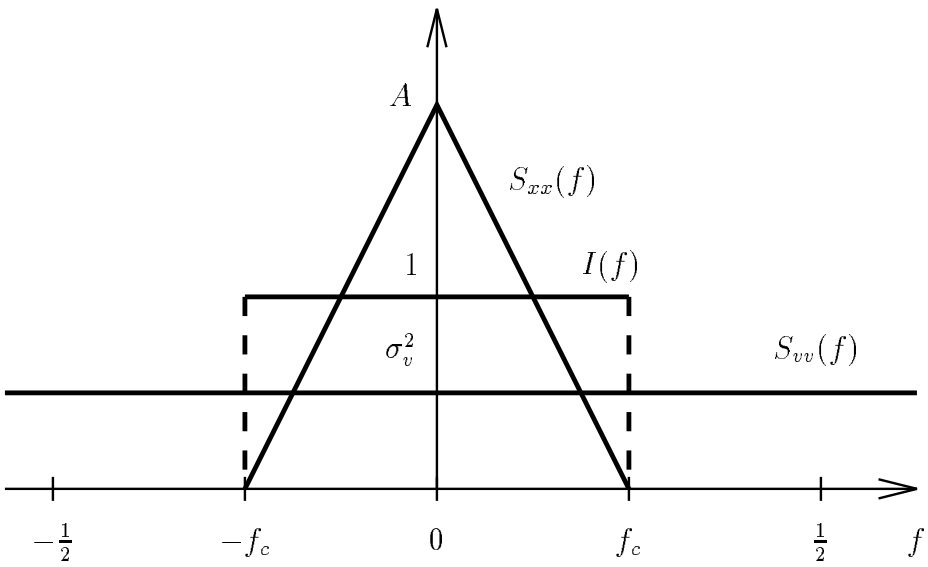


Figure 2.3: Wiener filtering and classical filtering illustrated for a triangular psdf.

The output $\widehat{x}_k(I)$ of the filter $I(f)$ will be equal to x_k minus an error $\widetilde{x}_k(I)$ which is a low-pass filtered version of v_k. Hence the variance of the error is

$$E\,\widetilde{x}_k^2(I) \;=\; \int_{-f_c}^{f_c} S_{vv}(f)df \;=\; 2\sigma_v^2 f_c\;. \tag{2.28}$$

Since $f_c \leq 0.5$, $E\,\widetilde{x}_k^2(I) \leq E\,v_k^2$. Hence, the filtering operation with $I(f)$ has reduced the noise level w.r.t. the measurement y_k while leaving the signal component x_k undistorted.

The Wiener approach is going to trade some signal distortion for a further reduction in overall error variance. Thus we find for the optimum filter

$$H(f) \;=\; \frac{S_{xx}(f)}{S_{xx}(f)+S_{vv}(f)} \;=\; \begin{cases} \dfrac{A}{\sigma_v^2}\,\dfrac{1-\frac{|f|}{f_c}}{1+\frac{A}{\sigma_v^2}(1-\frac{|f|}{f_c})} & ,\ |f|\leq f_c \\ 0 & ,\ f_c<|f|\leq \frac{1}{2} \end{cases} \tag{2.29}$$

We can analyze the nature of the optimal filter at high or low signal-to-noise ratio (SNR):

$$\begin{aligned} &\text{low SNR: } \frac{A}{\sigma_v^2}\to 0 \;:\quad H(f)\to \frac{S_{xx}(f)}{S_{vv}(f)} = \begin{cases} \dfrac{A}{\sigma_v^2}(1-\dfrac{|f|}{f_c}) & ,\ |f|\leq f_c \\ 0 & ,\ f_c<|f|\leq \frac{1}{2} \end{cases} \\ &\text{high SNR: } \frac{A}{\sigma_v^2}\to \infty \;:\quad H(f)\to I(f) \end{aligned} \tag{2.30}$$

So for low SNR, the filter $H(f)$ becomes proportional to the ratio of the psdf's of the signal of interest and the noise. For high SNR, the filter $H(f)$ approaches the classical distortion-based design, passing perfectly all frequencies where the signal of interest is present. We

find for the MMSE:

$$\text{MMSE} \;=\; E\,\tilde{x}_k^2(H) \;=\; [1 - \frac{\sigma_v^2}{A}\,\ln(1 + \frac{A}{\sigma_v^2})]\,E\,\tilde{x}_k^2(I)\;. \tag{2.31}$$

We can again analyze the limiting behavior for high or low SNR:

$$\begin{aligned} \text{low SNR: } \frac{A}{\sigma_v^2} \to 0 \;:\quad & E\,\tilde{x}_k^2(H) \to E\,x_k^2 = A\,f_c \\ \text{high SNR: } \frac{A}{\sigma_v^2} \to \infty \;:\quad & E\,\tilde{x}_k^2(H) \to\; E\,\tilde{x}_k^2(I) = 2 f_c \sigma_v^2 \end{aligned} \tag{2.32}$$

The ratio $E\,\tilde{x}_k^2(H)/E\,\tilde{x}_k^2(I)$ is sketched in Fig. 2.4. For high SNR, the optimal filter behaves like the classical distortion criterion based design. The variance of the error is the variance of the noise in the signal band. For low SNR, the optimal filter works much better than the classical one. Indeed, the variance of the error becomes equal to the signal variance, even though the noise level is much higher! It is true though that at low SNR, the performance of even the optimal filter is not very good in this example.

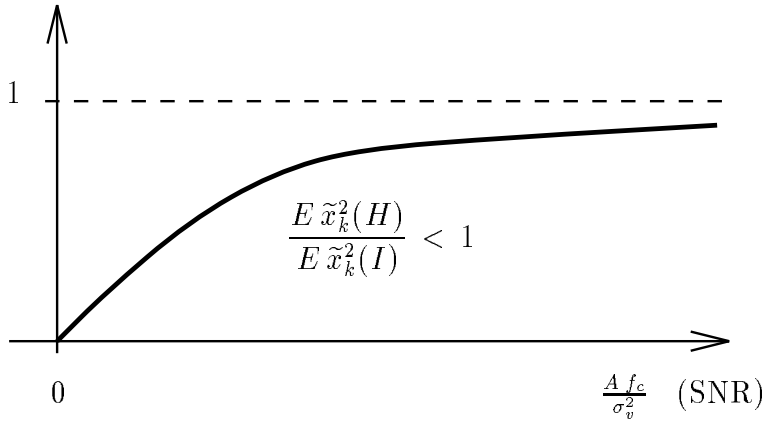


Figure 2.4: The ratio of the Mean Squared Error for the Wiener filter and the classical filter as a function of SNR.

◇

2.1.2 Channel Equalization

A variation on the signal in noise theme is the case in which the signal x_k has passed a filter $\mathrm{C}(z) = \sum c_k z^{-k}$ before being measured with additive noise v_k. This paradigm arises in digital communications. In that case the x_k represent a sequence of symbols that are transmitted over a linear time-invariant channel using a linear modulation scheme. If the received signal gets sampled at the symbol rate (the rate at which the symbols x_k are sent), then we get a discrete time received signal y_k that contains a noisy version of the x_k filtered by the cascade of the transmission (pulse shaping) filter, the actual channel and the receiving filter (ideally

a matched filter). We shall loosely call the sampled version of this cascade the channel $C(z)$. So we get the following discrete-time system

$$y_k \;=\; \mathrm{C}(q)\, x_k + v_k \;. \tag{2.33}$$

In what follows, we shall assume that x_k and v_k are independent white stationary sequences with zero mean and variances σ_x^2 and σ_v^2 respectively. We also assume here that all signals involved are real and scalar. The x_k have a discrete distribution and take on values in a finite alphabet. The problem of deciding on the basis of the y_k which discrete values x_k have been sent is called the *detection* problem. If the channel impulse response c_k has only one non-zero sample, then the optimal detection can be done instantaneously since the noise samples v_k are independent. This means that if w.l.o.g. we consider c_0 to be the non-zero sample, x_k can be detected from the sample

$$y_k = c_0\, x_k + v_k \;. \tag{2.34}$$

If the channel c_k has more than one non-zero samples, then the symbols appear superimposed in the received signal y_k and this phenomenon is called *intersymbol interference* (ISI). The problem of detecting the symbols in the presence of ISI is called *equalization*. Optimal (maximum-likelihood) equalization can be done for FIR channels using the so-called Viterbi algorithm (which corresponds to dynamic programming). The performance of the Viterbi algorithm is bound by the so-called Matched Filter Bound (MFB). The MFB expresses the maximum SNR achievable for the detection of a certain symbol x_k assuming that all other symbols have been correctly detected. This means that all other symbols are known so that their contribution can be subtracted from the received signal y_k. If we assume w.l.o.g. that the symbol we want to detect is x_0, then the resulting received signal with the contributions of the $x_k, k \neq 0$, removed is

$$y_k \;=\; c_k\, x_0 + v_k \;. \tag{2.35}$$

If the white noise v_k is Gaussian, then it turns out that the optimal detection of x_0 can be done on the basis of the unconstrained ML estimate

$$\widehat{x}_0^{ML} = \frac{\sum_k c_k\, y_k}{\sum_k c_k^2} \;. \tag{2.36}$$

Another interpretation of this estimate goes as follows. Consider filtering the y_k from (2.35) with a filter $\mathrm{H}(z)$ to obtain the signal $\widehat{x}_k$ and consider in particular the output at time 0, $\widehat{x}_0$. The part of $\widehat{x}_0$ due to x_0 is called the signal part while the part due to the v_k is called the noise part. It turns out that the filter $\mathrm{H}(z)$ that maximizes the SNR in $\widehat{x}_0$ is the *matched filter* $\mathrm{H}(z) = \mathrm{C}^\dagger(z) = \mathrm{C}(1/z)$, matched to the channel $\mathrm{C}(z)$. The SNR at the output of the matched filter is the MFB

$$\mathrm{MFB} \;= \frac{\sigma_x^2}{\sigma_v^2} \frac{1}{2\pi j} \oint \frac{dz}{z} \mathrm{C}^\dagger(z)\mathrm{C}(z) = \frac{\sigma_x^2}{\sigma_v^2} \int_{-\frac{1}{2}}^{\frac{1}{2}} df\, C^*(f) C(f) \tag{2.37}$$

The MFB is proportional to the energy in the channel response. (The term MFB is also often used to denote the corresponding probability of error in the detection of $\widehat{x}_0^{ML}$).

The Viterbi equalizer can be fairly complex however. A class of simple suboptimal equalizers can be obtained as the cascade of a simple linear estimator followed by an instantaneous

detector (decision element). The class of so-called *linear equalizers* (LEs) (see Fig. 2.5) performs linear estimation on the basis of the signal y_k alone. The more sophisticated class of *decision-feedback equalizers* (DFEs) performs linear estimation on the basis of all y_k plus also the previously detected x_k (hence, feedback of previous decisions). We shall consider here the linear equalizers in detail.

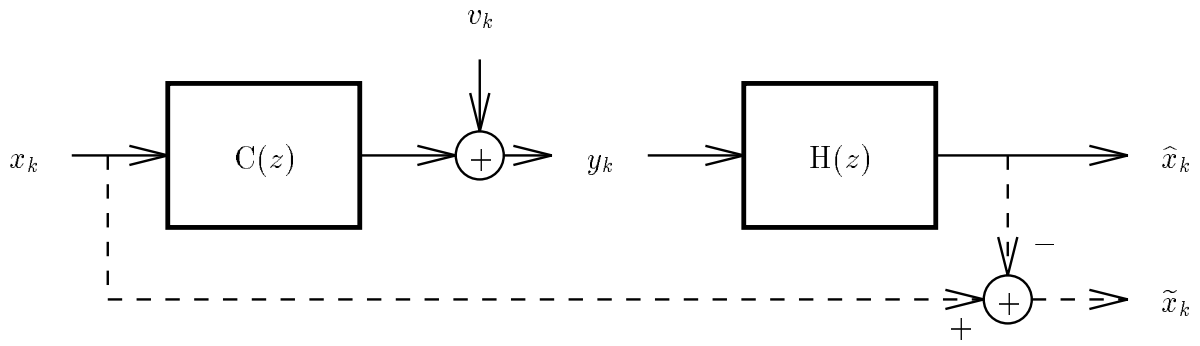


Figure 2.5: The linear equalizer set-up.

Zero-Forcing Linear Equalizers

Linear equalizers are simply linear filters $\mathrm{H}(z)$ filtering the signal y_k and their output $\widehat{x}_k = \mathrm{H}(q)y_k$ gets processed by a decision element. The objective in the classical point of view of filtering a signal in noise would be to achieve zero distortion for the signal part. The signal part in $\widehat{x}_k$ is $\mathrm{H}(q)\mathrm{C}(q)x_k$. To obtain zero distortion (signal part of $\widehat{x}_k$ equal to x_k) would mean that

$$\mathrm{H}(z)\mathrm{C}(z) = 1 \quad \Rightarrow \quad \mathrm{H}_{ZF}(z) = \frac{1}{\mathrm{C}(z)} = \frac{1}{\mathrm{C}_{min}(z)} \frac{1}{\mathrm{C}_{max}(z)} \tag{2.38}$$

Since this equalizer forces the resulting ISI to zero, this solution is called the *zero-forcing* (ZF) equalizer. The factors $\mathrm{C}_{min}(z)$ and $\mathrm{C}_{max}(z)$ are the minimum-phase and maximum-phase factors of $\mathrm{C}(z)$ (which will in general not be minimum-phase nor maximum-phase; $\mathrm{C}(z)$ is assumed rational for this discussion). The inverses of the minimum-phase and maximum-phase factors are causal and anti-causal respectively. This means that in general $\mathrm{H}_{ZF}(z)$ will have an impulse response that extends from $-\infty$ to $+\infty$ (even if $\mathrm{C}(z)$ is FIR). In practice, $\mathrm{H}(z)$ will normally be approximated with an FIR filter. This FIR approximation will have to be non-causal in order to get a good approximation. Such finite non-causality can be dealt with by introducing a corresponding delay. The FIR approximation will have to be longer and longer as the zeros of $\mathrm{C}(z)$ approach the unit circle. The ZF equalizer does not exist if in the limit the zeros of $\mathrm{C}(z)$ are on the unit circle.

The error signal $\widetilde{x}_k = x_k - \widehat{x}_k = \mathrm{H}(q)v_k = \frac{1}{\mathrm{C}(q)}v_k$ only contains noise, due to the ZF condition. The resulting MSE is

$$MSE_{ZF-LE} = E\,\widetilde{x}_k^2 = \frac{\sigma_v^2}{2\pi j} \oint \frac{dz}{z} \frac{1}{\mathrm{C}^\dagger(z)\mathrm{C}(z)} \tag{2.39}$$

The MSE can get very big as some zeros of $\mathrm{C}(z)$ approach the unit circle. This phenomenon is called *noise enhancement*. The perfect cancellation of the ISI is obtained at the cost of enhancing the noise. The SNR of the linear equalizer is defined as

$$SNR_{ZF-LE} = \frac{\sigma_x^2}{MSE_{ZF-LE}} \,. \tag{2.40}$$

Using the Cauchy-Schwarz inequality, one can show that

$$SNR_{ZF-LE} \leq MFB \,. \tag{2.41}$$

MMSE Linear Equalizers

The optimal filtering point of view simply takes the MSE as optimality criterion. Hence we get the Wiener filter

$$\mathrm{H}_{MMSE}(z) = \mathrm{S}_{xy}(z)\mathrm{S}_{yy}^{-1}(z) = \frac{\sigma_x^2 \mathrm{C}^\dagger(z)}{\sigma_x^2 \mathrm{C}^\dagger(z)\mathrm{C}(z) + \sigma_v^2} \,. \tag{2.42}$$

Remark that the factor $\mathrm{C}^\dagger(z)$ in the numerator represents the matched filter. We find for the MSE

$$\begin{aligned} MSE_{MMSE-LE} &= E\,(x_k - \widehat{x}_k)x_k \,=\, \sigma_x^2 \left(1 - \frac{\sigma_x^2}{2\pi j} \oint \frac{dz}{z}\, \mathrm{C}^\dagger(z)\mathrm{S}_{yy}^{-1}(z)\mathrm{C}(z)\right) \\ &= \frac{\sigma_x^2 \sigma_v^2}{2\pi j} \oint \frac{dz}{z}\, \mathrm{S}_{yy}^{-1}(z) \,=\, \frac{\sigma_v^2}{2\pi j} \oint \frac{dz}{z}\, \frac{1}{\mathrm{C}^\dagger(z)\mathrm{C}(z) + \frac{\sigma_v^2}{\sigma_x^2}} \,. \end{aligned} \tag{2.43}$$

From the previous expression, it is clear that

$$MSE_{MMSE-LE} \leq \min\left\{\sigma_x^2, MSE_{ZF-LE}\right\} \,. \tag{2.44}$$

This means that the MMSE-LE always does at least as well as the ZF-LE, at least in terms of MSE. Using the Cauchy-Schwarz inequality, one can show that

$$SNR_{MMSE-LE} \leq MFB + 1 \,. \tag{2.45}$$

Remark that the MMSE equalizer converges to the ZF equalizer as $\sigma_v^2 \to 0$, i.e. the two are identical in the noiseless case.

Unbiased MMSE Linear Equalizers

We remark from (2.45) that $SNR_{MMSE-LE}$ can be larger than the MFB. This is due to the fact that the MMSE LE gives a biased estimate of x_k. In other words, the coefficient of x_k appearing in $\widehat{x}_k$ is not equal to 1. Although the unconstrained MMSE LE gives the lowest MSE, the bias in $\widehat{x}_k$ will increase the probability of error in the decision process. The decision element expects indeed to see x_k plus some random deviations at its input, whereas a bias (in the form of $\alpha\, x_k$) is not a random perturbation w.r.t. x_k. Random perturbations are perturbations due to the $x_i, i \neq k$, and the v_i. The unbiased MMSE (UMMSE) LE minimizes

the MSE subject to the constraint that the estimator be unbiased, i.e. $E\,[\widehat{x}_k|x_k] = x_k$, or $\frac{1}{2\pi j}\oint \frac{dz}{z}\mathrm{H}(z)\mathrm{C}(z) = \sum_k h_k c_{-k} = 1$. The MMSE equalizer takes the Bayesian viewpoint in which all the x_k and the v_k are considered random. The UMMSE equalizer takes the more deterministic viewpoint in which the symbol to be estimated x_k is considered deterministic, but the other $x_i, i \neq k$, and the v_i are still random (this the viewpoint of the BLUE estimator, but considered here for an infinite number of measurements y_k). The UMMSE equalizer design problem is hence

$$\min_{h_i:\ \sum_i h_i c_{-i} = 1} E\,(x_k - \sum_i h_i y_{k-i})^2\ . \tag{2.46}$$

We can turn this constrained optimization problem into an unconstrained optimization problem by introducing a Lagrange multiplier λ. The unconstrained problem becomes

$$\min_{h_i,\lambda} f(\mathrm{H}(.),\lambda) \ = \ \min_{h_i,\lambda}\left\{E\,(x_k - \sum_i h_i y_{k-i})^2 \ + \lambda(\sum_i h_i c_{-i} - 1)\right\}\ . \tag{2.47}$$

By setting derivatives equal to zero, we find

$$\begin{array}{rcl} \dfrac{\partial f}{\partial h_i} & = & 2\,E\left(x_k - \sum_n h_n y_{k-n}\right)(-y_{k-i}) + \lambda c_{-i} = 0 \\ \dfrac{\partial f}{\partial \lambda} & = & \sum_i h_i c_{-i} - 1 = 0 \end{array} \tag{2.48}$$

The first equation leads to

$$\begin{array}{l} -r_{xy}(i) + \sum_n h_n r_{yy}(i-n) + \dfrac{\lambda}{2}c_{-i} = 0 \\ \Rightarrow\ -\mathrm{S}_{xy}(z) + \mathrm{H}(z)\mathrm{S}_{yy}(z) + \dfrac{\lambda}{2}\mathrm{C}^\dagger(z) = 0\ \Rightarrow\ \mathrm{H}(z) = (\sigma_x^2 - \dfrac{\lambda}{2})\mathrm{C}^\dagger(z)\mathrm{S}_{yy}^{-1}(z)\ . \end{array} \tag{2.49}$$

From the constraint, we find

$$\frac{1}{2\pi j}\oint \frac{dz}{z}\mathrm{H}(z)\mathrm{C}(z) = 1 = \frac{\sigma_x^2 - \dfrac{\lambda}{2}}{2\pi j}\oint \frac{dz}{z}\mathrm{C}^\dagger(z)\mathrm{S}_{yy}^{-1}(z)\mathrm{C}(z)\ . \tag{2.50}$$

Hence

$$\mathrm{H}_{UMMSE}(z) = \left(\frac{1}{2\pi j}\oint \frac{dz}{z}\mathrm{C}^\dagger(z)\mathrm{S}_{yy}^{-1}(z)\mathrm{C}(z)\right)^{-1}\mathrm{C}^\dagger(z)\mathrm{S}_{yy}^{-1}(z) \tag{2.51}$$

Hence, the UMMSE equalizer is simply proportional to the MMSE equalizer, with the proportionality factor adjusted for unbiasedness. We find for the MSE

$$\begin{array}{rcl} MSE_{UMMSE-LE} & = & E\,(x_k - \widehat{x}_k)^2 = \frac{1}{2\pi j}\oint \frac{dz}{z}(\mathrm{S}_{xx}(z) - \mathrm{S}_{x\widehat{x}}(z) - \mathrm{S}_{\widehat{x}x}(z) + \mathrm{S}_{\widehat{x}\widehat{x}}(z)) \\ & = & \frac{1}{2\pi j}\oint \frac{dz}{z}(\mathrm{S}_{xx}(z) - \mathrm{S}_{xx}(z) - \mathrm{S}_{xx}(z) + \mathrm{S}_{\widehat{x}\widehat{x}}(z)) \\ & = & \frac{1}{2\pi j}\oint \frac{dz}{z}\mathrm{S}_{\widehat{x}\widehat{x}}(z) \ - \ \frac{1}{2\pi j}\oint \frac{dz}{z}\mathrm{S}_{xx}(z) \ = \ \left(\frac{1}{2\pi j}\oint \frac{dz}{z}\mathrm{C}^\dagger(z)\mathrm{S}_{yy}^{-1}(z)\mathrm{C}(z)\right)^{-1} - \sigma_x^2 \end{array} \tag{2.52}$$

from which we can find the SNR

$$SNR_{UMMSE-LE} = \frac{\sigma_x^2}{MSE_{UMMSE-LE}} = \frac{\frac{\sigma_x^2}{2\pi j}\oint \frac{dz}{z}\mathrm{C}^\dagger(z)\mathrm{S}_{yy}^{-1}(z)\mathrm{C}(z)}{1-\frac{\sigma_x^2}{2\pi j}\oint \frac{dz}{z}\mathrm{C}^\dagger(z)\mathrm{S}_{yy}^{-1}(z)\mathrm{C}(z)} = \frac{1}{\frac{\sigma_v^2}{2\pi j}\oint \frac{dz}{z}\mathrm{S}_{yy}^{-1}(z)} - 1 \quad . \tag{2.53}$$

From (2.53) and (2.43), we find that

$$SNR_{MMSE} \;=\; SNR_{UMMSE} + 1 \quad ! \tag{2.54}$$

Even though the UMMSE LE has a lower SNR than the MMSE LE, it can be shown that its probability of error is lower (using the Gaussian assumption on the interfering symbols - Central Limit Theorem). Since the UMMSE has the highest SNR of all unbiased LEs, it has the lowest probability of error of all LEs. We also have

$$SNR_{ZF-LE} \;\leq\; SNR_{UMMSE-LE} \;\leq\; MFB = \frac{1}{\sigma_v^2\, 2\pi j}\oint \frac{dz}{z}\mathrm{S}_{yy}(z) - 1 \tag{2.55}$$

where the second inequality (compare to (2.53)) is due to the fact that (Cauchy-Schwarz inequality) $\left(\frac{1}{2\pi j}\oint \frac{dz}{z}\mathrm{S}_{yy}^{-1}(z)\right)^{-1} \leq \frac{1}{2\pi j}\oint \frac{dz}{z}\mathrm{S}_{yy}(z)$ in which equality holds iff $S_{yy}(f)$ and hence $|C(f)|$ is constant as a function of frequency. In this last case, also $SNR_{ZF-LE} = MFB$.

2.2 Causal Wiener Filtering♠

The optimal filter obtained in the previous section is non-causal in general. This may not pose a problem in cases where the independent variable is not time but space (e.g. image processing). However, in the context of temporal processing, we shall need to constrain the filter to be causal in order to arrive at a design that can be implemented in real time. So apart from the linearity constraint, we shall impose the estimate to be a causal filtering operation on the measurements y_k. Since now the data $\{y_n,\, n \leq k\}$ are localized in time due to the causality constraint, it makes a difference whether we are going to use these data to estimate x_k or x_{k+10} or x_{k-12} (in the absence of causality as in the previous section, this was not an issue). So consider using the data up to time k to estimate $x_{k+\lambda},\ \lambda \in \mathcal{Z}$. We distinguish the following three cases:

- $\lambda = 0$: *filtering*
- $\lambda > 0$: *prediction*
- $\lambda < 0$: *smoothing*

The filter coefficients $\{h_{k,n},\ n \leq k\}$ are determined from the following LMMSE estimation problem

$$\min_{h_{k,n}}\; E(x_{k+\lambda} - \widehat{x}_{k+\lambda})^2 \;\;=\;\; \min_{h_{k,n}}\; E(x_{k+\lambda} - \sum_{n=-\infty}^{k} h_{k,n}y_n)^2 \quad . \tag{2.56}$$

As before, the result can be most easily expressed in the z-transform domain. In summary, we get

$$\mathrm{H}(z) \;=\; \frac{1}{\mathrm{S}^+_{yy}(z)} \left\{ \frac{\mathrm{S}_{xy}(z)\, z^{\lambda}}{\mathrm{S}^+_{yy}(z^{-1})} \right\}_+ \tag{2.57}$$

where we have used the following notation

- $\{.\}_+$: "take the causal part of".
- $\mathrm{S}_{yy}(z) = \mathrm{S}^+_{yy}(z)\, \mathrm{S}^+_{yy}(z^{-1})$: *spectral factorization.* Subject to certain conditions, a psdf can be factored into its causal minimum-phase factor $\mathrm{S}^+_{yy}(z)$ and its anti-causal maximum-phase counterpart $\mathrm{S}^+_{yy}(z^{-1})$.

If we drop the causality constraint and replace the $\{.\}_+$ operator by an identity operator, then we find the non-casual Wiener filter from the previous section.

The filtering operation indicated in (2.57) can be simplified by considering the following two-step procedure. Instead of working with the signal y_k directly, consider its whitened version $f_{\infty,k}$ which is obtained by filtering y_k with $\mathrm{A}_\infty(z) = \mathrm{S}^+_{yy}(\infty)/\mathrm{S}^+_{yy}(z)$, a causal and causally invertible filter. The infinite-order forward prediction errors form a white noise with variance $\sigma^2_{f,\infty}$. They are also called the *innovations* process since $f_{\infty,k}$ represents the part of y_k that could not be predicted from the previous y's, hence $f_{\infty,k}$ is the innovation of y_k w.r.t. its past. Estimating $x_{k+\lambda}$ in terms of the causally transformed measurements $f_{\infty,k}$ leads to the optimal filter

$$\mathrm{H}^f(z) \;=\; \frac{1}{\sigma_{f,\infty}} \left\{ \frac{\mathrm{S}_{x\, f_\infty}(z)\, z^{\lambda}}{\sigma_{f,\infty}} \right\}_+ \;=\; \frac{1}{\sigma^2_{f,\infty}} \left\{ \mathrm{S}_{x\, f_\infty}(z)\, z^{\lambda} \right\}_+ \;. \tag{2.58}$$

This result follows immediately from (2.57) by substituting y by f_∞. However, (2.58) can be determined very straightforwardly from the normal equations resulting from (2.56) (with y_k substituted by $f_{\infty,k}$): this infinite set of normal equations has $\sigma^2_{f,\infty} I$ as matrix of coefficients and hence $h^f_{k,n} = \dfrac{r_{x\, f_\infty}(\lambda + k - n)}{\sigma^2_{f,\infty}} = h^f_{0,n-k} = h^f_{k-n},\ n \le k$. From $\mathrm{H}^f(z)$ it is now easy to determine $\mathrm{H}(z)$ as

$$\mathrm{H}(z) \;=\; \mathrm{H}^f(z)\, \mathrm{A}_\infty(z) = \frac{1}{\sigma^2_{f,\infty}} \left\{ \mathrm{S}_{x\, f_\infty}(z)\, z^{\lambda} \right\}_+ \frac{\sigma_{f,\infty}}{\mathrm{S}^+_{yy}(z)} = \frac{1}{\mathrm{S}^+_{yy}(z)} \left\{ \frac{\mathrm{S}_{xy}(z)\, z^{\lambda}}{\mathrm{S}^+_{yy}(z^{-1})} \right\}_+ \;. \tag{2.59}$$

So the innovations approach can be used to find the result (2.57) or to actually carry out the filtering operation in two steps.

2.3 Order-Recursive FIR Wiener Filtering♠

So far we have assumed that x_k and y_k are jointly wide-sense stationary processes. The approach followed in this section is applicable to non-stationary processes. As a result, the optimal filter will be time-varying and transform-domain techniques will no longer be applicable. We shall therefore follow a time-domain approach. In this time-domain approach, we

shall not only consider causality of the filter, but also causality of the signals. In other words, we consider starting the filtering operation at some point in time. Without loss of generality, we can take time zero. In this way, the time-domain approach will lead to causal FIR Wiener filtering problems of which the filtering order increases as time increases. This will allow us, if so desired, to converge to the infinite length causal Wiener filter of the stationary case considered in the previous section.

The non-stationarity considered here can be due to a variety of reasons:

- transients: signals that start at time zero are inherently nonstationary, but if they are obtained by prewindowing stationary signals, then the nonstationary transients will decay as time grows,
- genuine nonstationarity.

So assume data start coming in at time zero. Then at time k we can form the linear causal estimate

$$\widehat{x}_k \;=\; \sum_{i=0}^{k} h_{k,i} y_i \;. \tag{2.60}$$

The filter coefficients $h_{k,i}$ are obtained from the following LMMSE estimation problem

$$\min_{h_{k,i}} E\,(x_k - \widehat{x}_k)^2 \;. \tag{2.61}$$

The optimal $h_{k,i}$ are again found by considering the orthogonality conditions:

$$E\,(x_k - \widehat{x}_k)\, y_m = 0 \;,\quad m = 0, 1, \ldots, k \;. \tag{2.62}$$

Or hence

$$E\,\widehat{x}_k y_m \;=\; \sum_{i=0}^{k} h_{k,i}\, E\, y_m y_i \;=\; \sum_{i=0}^{k} h_{k,i}\, r_{yy}(m,i) \;=\; r_{xy}(k,m) \;=\; E\, x_k y_m \;,\quad m = 0, 1, \ldots, k \;. \tag{2.63}$$

These are the normal equations which can be spelled out as

$$\begin{bmatrix} r_{yy}(0,0) & r_{yy}(0,1) & \cdots & r_{yy}(0,k) \\ r_{yy}(1,0) & r_{yy}(1,1) & \cdots & r_{yy}(1,k) \\ \vdots & \vdots & \ddots & \vdots \\ r_{yy}(k,0) & r_{yy}(k,1) & \cdots & r_{yy}(k,k) \end{bmatrix} \begin{bmatrix} h_{k,0} \\ h_{k,1} \\ \vdots \\ h_{k,k} \end{bmatrix} \;=\; \begin{bmatrix} r_{xy}(k,0) \\ r_{xy}(k,1) \\ \vdots \\ r_{xy}(k,k) \end{bmatrix} \tag{2.64}$$

and can be denoted as

$$R_k\, H_k \;=\; P_k \;. \tag{2.65}$$

The normal equations are a set of $k{+}1$ equations in $k{+}1$ unknowns and hence take $\mathcal{O}((k{+}1)^3)$ operations to be solved.

One way of solving them involves the *lower-diagonal-upper (LDU) triangular factorization* of the covariance matrix R_k. In fact this factorization comes again about by working in terms of the innovations process. Consider *Gram-Schmidt orthogonalization* of the random variables

$\{y_0, y_1, \ldots, y_k\}$, leading to the innovations $\{f_{0,0}, f_{1,1}, \ldots, f_{k,k}\}$. This orthogonalization procedure can be seen to correspond to consecutive forward prediction problems of increasing order at consecutive time instants. Indeed, we get

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ A_{1,1} & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ A_{k,k} & \cdots & A_{k,1} & 1 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} f_{0,0} \\ f_{1,1} \\ \vdots \\ f_{k,k} \end{bmatrix} \quad \rightarrow \quad A_{0:k}\, y_{0:k} \;=\; f_{0:k} \tag{2.66}$$

where the last equation is a short-hand notation for the first one. Note that $A_{0:k}$ is a lower-triangular matrix and corresponds to a causal filtering operation. By multiplying both sides of (2.66) with its respective transpose and taking expectation, we get

$$A_{0:k} R_k A_{0:k}^T \;=\; D_k \;=\; E\, f_{0:k} f_{0:k}^T = \; \mathrm{diag}\{\sigma_{f,0}^2, \sigma_{f,1}^2, \ldots, \sigma_{f,k}^2\} \quad \Rightarrow \quad R_k = L_k D_k L_k^T \tag{2.67}$$

with $L_k = A_{0:k}^{-1}$ being a lower-triangular matrix. This is the LDU decomposition of R_k. A related decomposition is the *Cholesky* decomposition which is

$$R_k \;=\; \overline{L}_k\, \overline{L}_k^T \quad \text{where} \quad \overline{L}_k \;=\; L_k D_k^{\frac{1}{2}} \; . \tag{2.68}$$

In fact, we have the following correspondences between time-domain and frequency-domain expressions

$$\begin{array}{rcl} A_{0:k} R_k A_{0:k}^T \;=\; D_k & \leftrightarrow & \mathrm{A}_\infty(z) S_{yy}(z) \mathrm{A}_\infty(z^{-1}) \;=\; \sigma_{f,\infty}^2 \\ A_{0:k} \;=\; D_k^{\frac{1}{2}}\, \overline{L}_k^{-1} & \leftrightarrow & \mathrm{A}_\infty(z) \;=\; \dfrac{\sigma_{f,\infty}}{\mathrm{S}_{yy}^+(z)} \\ R_k \;=\; \overline{L}_k\, \overline{L}_k^T & \leftrightarrow & \mathrm{S}_{yy}(z) \;=\; \mathrm{S}_{yy}^+(z)\, \mathrm{S}_{yy}^+(z^{-1}) \end{array} \tag{2.69}$$

We can also estimate x_k in terms of the innovations, the transformed measurements,

$$\begin{array}{c} \widehat{x}_k = H_k^T y_{0:k} = H_k^T L_k f_{0:k} = (L_k^T H_k)^T f_{0:k} = H_k^{f\,T} f_{0:k} \\ \Rightarrow \quad H_k^f \;=\; L_k^T H_k \; . \end{array} \tag{2.70}$$

The optimal filter H_k^f can now be determined from the normal equations (2.64) in which we substitute $y_{0:k}$ by $f_{0:k}$:

$$D_k\, H_k^f \;=\; P_k^f = \; E\, f_{0:k} x_k = L_k^{-1}\, E\, y_{0:k} x_k \;=\; L_k^{-1} P_k \; . \tag{2.71}$$

We can now find the optimal filter for the original data by a sequence of two operations

$$\begin{array}{rcl} L_k D_k\, H_k^f & = & P_k \\ L_k^T\, H_k & = & H_k^f \end{array} \tag{2.72}$$

which is a well-known technique for solving a set of normal equations $R_k H_k = L_k D_k\, L_k^T H_k = P_k$ using the LDU decomposition of R_k and the two backsubstitution operations required for solving the two triangular sets of equations in (2.72). When comparing to (2.59), the causal and anti-causal operations may not seem to correspond. However, they do correspond when

one realizes that the entries in the vectors P and H appear in reversed order (consider the stationary, time-invariant case).

Again, the innovations technique allows to solve for the filter to be applied to the original data. Or we can first do the orthogonalization and then work with the innovations. Given the innovations, the filtering problem is extremely easy and can be solved in a recursive manner. Indeed, the diagonal set of equations (2.71) for H_k^f implies

$$H_k^f \;=\; \begin{bmatrix} H_{k-1}^f \\ h_{k,k}^f \end{bmatrix} . \tag{2.73}$$

Hence we obtain

$$\widehat{x}_k \;=\; \widehat{x}_{k-1} + h_{k,k}^f \, f_{k,k} \quad , \quad h_{k,k}^f \;=\; r_{xf}(k,k)/\sigma_{f,k}^2 . \tag{2.74}$$

Hence, the main task in the innovations approach is the generation of the innovations themselves. Once the innovations are available, the filtering task is easy, due to the uncorrelatedness of the innovations. To obtain the innovations, we need to find the forward prediction filters which are themselves determined by normal equations. In general, it is about as difficult to solve the normal equations for the prediction filters as for a general filtering problem. Simplifications arise in the following two cases:

- *Stationarity*: if $r_{yy}(i,j) = r_{yy}(|i-j|)$, then R_k is Toeplitz and the Levinson algorithm of the previous chapter can be used to generate the first k innovations in $\mathcal{O}(k^2)$ operations. In this case, the filtering operations in equation (2.74) correspond to a ladder part that gets added to the lattice filter for the prediction part (in which the roles of the forward and backward prediction errors get interchanged).

- If y_k is the output of a n-dimensional state-space model (even time-varying), then the first k innovations can be generated in $\mathcal{O}(kn^3)$ operations. This is accomplished by the *Kalman filter*.

2.4 Fixed-Order FIR Wiener Filtering

Now fix the filter order of the FIR Wiener filter to be N. Using the causality and FIR conditions, we construct a linear estimator (FIR filter) for the (unavailable) signal of interest x_k (desired signal) from the N most recent samples of a correlated available signal y_k :

$$\widehat{x}_k \;=\; \sum_{i=0}^{N-1} h_i y_{k-i} \;=\; Y_k^T H \;=\; H^T Y_k \tag{2.75}$$

where

$$Y_k \;=\; \begin{bmatrix} y_k \\ y_{k-1} \\ \vdots \\ y_{k-N+1} \end{bmatrix} \quad , \quad H \;=\; \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_{N-1} \end{bmatrix} . \tag{2.76}$$

As criterion to determine the filter coefficients H, we consider the mean squared error in estimating x_k with $\widehat{x}_k$:

$$\begin{aligned} \text{MSE} \;=\; \xi(H) \;&=\; E\,(x_k - \widehat{x}_k)^2 \;=\; E\left(x_k - H^T Y_k\right)^2 \\ &=\; E\left(x_k^2 \;-\; 2H^T Y_k x_k \;+\; H^T Y_k Y_k^T H\right) \\ &=\; \sigma_x^2 \;-\; 2H^T E\,(Y_k x_k) \;+\; H^T E\left(Y_k Y_k^T\right) H \\ &=\; \sigma_x^2 \;-\; 2H^T R_{Yx} \;+\; H^T R_{YY} H \end{aligned} \tag{2.77}$$

where

$$\begin{aligned} R_{YY} \;=\; E Y_k Y_k^T \;&=\; \begin{bmatrix} r_{yy}(k,k) & \cdots & r_{yy}(k,k-N+1) \\ \vdots & & \vdots \\ r_{yy}(k-N+1,k) & \cdots & r_{yy}(k-N+1,k-N+1) \end{bmatrix} \\ R_{Yx} \;=\; E Y_k x_k \;&=\; \begin{bmatrix} r_{yx}(k,k) \\ \vdots \\ r_{yx}(k-N+1,k) \end{bmatrix} . \end{aligned} \tag{2.78}$$

The MSE $\xi(H)$ in (2.77) clearly is a quadratic cost function in H. Minimization of the MSE ξ requires setting the gradient to zero. The gradient is

$$\frac{\partial\,\xi}{\partial\,H} \;\triangleq\; \begin{bmatrix} \frac{\partial\xi}{\partial h_0} \\ \vdots \\ \frac{\partial\xi}{\partial h_{N-1}} \end{bmatrix} = \begin{bmatrix} -2r_{yx}(k,k) + 2\sum_{i=0}^{N-1} h_i r_{yy}(k,k-i) \\ \vdots \\ -2r_{yx}(k-N+1,k) + 2\sum_{i=0}^{N-1} h_i r_{yy}(k-N+1,k-i) \end{bmatrix} = 2R_{YY}H - 2R_{Yx}\,. \tag{2.79}$$

These expressions can be obtained by taking partial derivatives of $H^T R_{Yx} = \sum_{i=0}^{N-1} h_i r_{yx}(k-i,k)$ and $H^T R_{YY} H = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} h_i h_j r_{yy}(k-i,k-j)$. We find that the filter setting H^o that extremizes the MSE can be found from

$$\frac{\partial\,\xi}{\partial\,H} \;=\; 2\,(R_{YY}H^o - R_{Yx}) \;=\; 0 \quad\Rightarrow\quad H^o \;=\; R_{YY}^{-1} R_{Yx}\,. \tag{2.80}$$

The extremum thus found is indeed a minimum because the (constant) Hessian of the criterion $\xi(H)$ is

$$\left[\frac{\partial^2\xi}{\partial h_i\,\partial h_j}\right]_{i,j=0}^{N-1} \;=\; 2R_{YY} \;>\; 0 \tag{2.81}$$

which is a covariance matrix and hence is symmetric and positive semidefinite in general, but we shall here exclude degenerate cases and assume positive definiteness. The minimum value of the MSE (MMSE) can be found to be

$$\begin{aligned} \xi^o \;=\; \xi(H^o) \;&=\; \sigma_x^2 - 2R_{Yx}^T H^o + H^{o\,T} R_{YY} H^o \\ &=\; \cdots =\; \sigma_x^2 - R_{xY}\,R_{YY}^{-1} R_{Yx} \;=\; \sigma_x^2 - R_{xY} H^o \end{aligned} \tag{2.82}$$

where $R_{xY} = R_{Yx}^T$. One can check that we can alternatively write the criterion ξ as

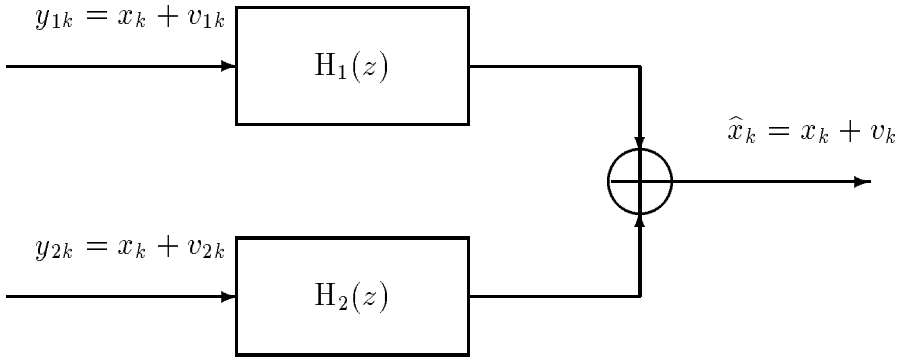
$$\xi(H) \;=\; \xi^o + (H - H^o)^T \, R_{YY} \, (H - H^o) \tag{2.83}$$

which brings out clearly its parabolic character.

2.5 Problems

2.1. *Wiener Filtering.* Let x_k be a zero-mean stationary random signal. By using this signal as the input to two linear filters with transfer functions $H_1(z)$ and $H_2(z)$ resp., we obtain two different outputs signals y_{1k} and y_{2k}. Compute the frequency response of the non-causal filter giving the LMMSE of y_{2k} in terms of y_{1k} and calculate the corresponding MSE.

2.2. *Constrained Wiener Filtering.* Consider the problem of combining two independent noisy measurements of the same signal as depicted in the figure below. Often a spectral



description of x_k is not available, or it may not be properly modeled as a random process. Furthermore, one may require that the signal be filtered without delay or distortion. We consider x_k to be deterministic in this case, and constrain the two filters to satisfy

$$\mathrm{H}_1(z) + \mathrm{H}_2(z) = 1, \quad \forall z \ , \tag{2.84}$$

so that x_k passes undistortedly.
Show how to choose $\mathrm{H}_1(z)$ so as to minimize the variance of the output noise $v_k = -\widetilde{x}_k$.

Chapter 3

Adaptive Filtering

So far in FIR Wiener Filtering, we have assumed that we have the second-order moments R_{YY} and R_{Yx} available to us. In practice, that may not be the case. But instead, we may have (part of) a realization $\{x_k, y_k, k \geq 0\}$ available. From these samples, we could of course try to estimate R_{YY} and R_{Yx}, and below we shall see that the *Recursive Least-Squares* (RLS) algorithm effectively does just that. However, we may also find that such an approach becomes too complicated, too computationally demanding. The *Least Mean Square* (LMS) algorithm takes an alternative approach in which it simply applies the steepest-descent algorithm (the simplest optimization algorithm) to the instantaneous squared error signal. To better comprehend the behavior of the LMS algorithm, we shall first apply the steepest-descent technique to the mean squared error to find the FIR Wiener filter iteratively.

The reader may note that at this point, we are deviating from the original optimal filtering starting point. Indeed, we have motivated the optimal filtering problem as follows: we are interested in a signal x_k that we cannot measure, but we can measure a related signal y_k. We also assumed the joint second-order statistics of the signals x_k and y_k to be available. In adaptive filtering, we do not assume any statistics to be given but we assume a realization of both signals to be available. The point is still to try to estimate the signal x_k from the signal y_k. Not because x_k would not be available, but in order to decompose x_k into its part that can be predicted from the signal y_k and its complementary part, which cannot be predicted from y_k. In practice, we are interested in one of these two complementary parts of x_k, $\widetilde{x}_k$ or $\widehat{x}_k$, or H, as can be seen in the applications.

In some applications such as the channel equalization problem, we may dispose of both x_k (the transmitted symbols) and y_k (the received signal) for a limited time, namely for the training sequence, which is a sequence of transmitted symbols that is known to the receiver. The role of adaptive filtering in that case is to find the FIR Wiener (MMSE) equalizer during the training sequence period, so that the equalizer can be applied afterwards, when actual unknown symbols x_k are transmitted, to produce symbol estimates $\widehat{x}_k$. In a more advanced category of techniques, known as *blind* equalization, the equalizer gets adapted without training sequence.

3.1 The Steepest-Descent Algorithm applied to FIR Wiener Filtering

We see from (2.80) that the N coefficients of the optimal filter H^o are found by solving a set of N equations, the so-called *normal equations*:

$$R_{YY}\, H^o \;=\; R_{Yx}\;. \tag{3.1}$$

This requires $O(N^3)$ operations (multiplications/additions) in general. We may want to solve this set of equations in an iterative fashion, where each iteration takes significantly less work than the amount of work required in solving a set of equations. The following iterative procedure requires a matrix-vector multiplication at each iteration, which takes $O(N^2)$ operations in general. If the matrix R_{YY} would be sparse (for instance, banded), the amount of computations could be even significantly further reduced. An iterative procedure allows us in each iteration to correct for mistakes (due to e.g. round-off errors) made in the previous iterations. We shall see further that this observation is the key idea behind the development of a corresponding adaptive algorithm.

There exist iterative methods other than the one described below, which approximate R_{YY} by a circulant matrix so that FFT techniques can be used, or which approximate R_{YY}^{-1} by a banded matrix. Iterations are then introduced to compensate for the approximation. The amount of work required per iteration is significantly less than $O(N^3)$.

One of the simplest methods of optimization is the *steepest-descent* algorithm. The idea is to start at an initial guess for the solution and to iterate the following process. The current approximate solution is improved upon by taken a certain step in the direction of the negative of the gradient, evaluated at the current approximate solution. Indeed, at any given point in H space, the gradient $\frac{\partial \xi}{\partial H}$ points in the direction of steepest ascent, the direction in which $\xi(H)$ increases the most. Hence $-\frac{\partial \xi}{\partial H}$ points in the direction of steepest descent, the direction in which $\xi(H)$ decreases the most (at least locally), see Fig. 3.1.

Let μ be the stepsize parameter. It determines how far we shall go in the direction of steepest descent. If H_k represents the approximate solution at iteration step k, then we get the following recursion:

$$H_{k+1} \;=\; H_k - \frac{\mu}{2}\nabla_k \tag{3.2}$$

where

$$\nabla_k \;=\; \left.\frac{\partial \xi}{\partial H}\right|_{H=H_k} \;=\; 2R_{YY}H_k - 2R_{Yx} \;=\; 2R_{YY}\,(H_k - H^o) \;=\; -2R_{YY}\widetilde{H}_k \tag{3.3}$$

where we have introduced the filter approximation error

$$\widetilde{H}_k \;=\; H^o - H_k\;. \tag{3.4}$$

So the iteration (3.2) now becomes

$$H_{k+1} \;=\; H_k - \mu\,(R_{YY}H_k - R_{Yx}) \;=\; (I - \mu R_{YY})\,H_k + \mu R_{Yx}\;. \tag{3.5}$$

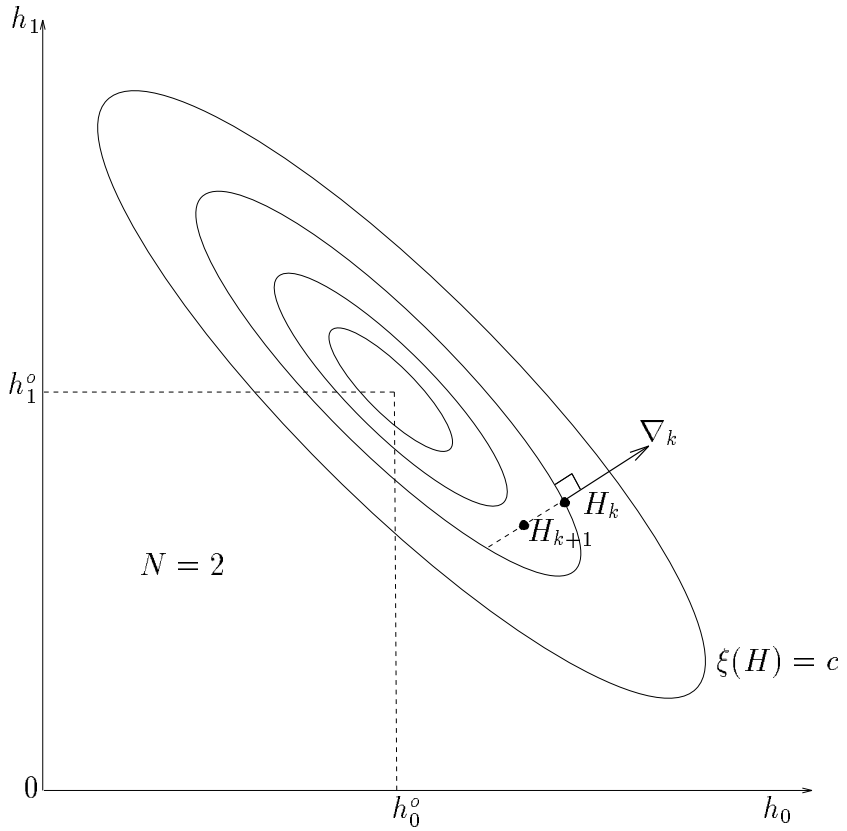


Figure 3.1: Orthogonality of the gradient on the elliptic isocost lines (loci of constant cost function).

Interpreting this iteration as an iterative equation solver, we see that $\frac{1}{2}\nabla_k = R_{YY}\,H_k - R_{Yx}$, the error in the satisfaction of the system of normal equations of which we are seeking the solution, is used to adjust the approximation to the solution.

Equation (3.4) describes the steepest descent algorithm. Its complexity is indeed dominated by that of forming the product $R_{YY}\,H_k$. Using R_{YY}, R_{Yx} and an initial guess H_0, (3.4) allows us to generate a sequence of approximations H_k. The question is whether this sequence will converge, and if so whether it will converge to H^o. For the analysis of the convergence process, we shall assume knowledge of $H^o = R_{YY}^{-1}\,R_{Yx}$ but we shall see that the conditions for convergence do not depend on H^o. The conditions for convergence are conditions on the step-size μ. From the derivation of the algorithm, it appears intuitively clear that the algorithm should work for a small positive μ since in that case we can be sure to make descending steps. If the steps we take are too big however, it is clear that we pass the valley and climb up the hill on the other side of the valley. So for large μ we do not expect convergence. Let us see what the analysis gives.

By introducing a translation of coordinates and referring the approximate solution with respect to the optimal solution, we get a set of homogeneous equations

$$\widetilde{H}_{k+1} \;=\; (I - \mu R_{YY})\,\widetilde{H}_k \;. \tag{3.6}$$

This is a set of coupled equations since R_{YY} is not a diagonal matrix in general. Consider the

eigen decomposition of the covariance matrix R:

$$R_{YY} \;=\; EY_kY_k^T \;=\; V\Lambda V^T \;=\; \sum_{i=1}^{N} \lambda_i V_i V_i^T \tag{3.7}$$

where

$$\Lambda \;=\; \text{diag}\;\{\lambda_1 \cdots \lambda_N\}\;, \quad V \;=\; [V_1\; V_2 \cdots V_N] \tag{3.8}$$

are matrices containing the eigenvalues and eigenvectors respectively. Note that V is an orthogonal matrix : $V^TV = I = VV^T$ or in other words, the eigenvectors are orthonormal: $V_i^TV_j = \delta_{ij}$. We also assume that the eigenvalues are ordered: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N > 0$ (the last inequality holds if the covariance matrix is positive definite, which we assume to be the case).

By rotating the principal axes using V, we get for the transformed error system (from (3.6))

$$\begin{aligned} V^T\widetilde{H}_{k+1} &= V^T\,(I - \mu R_{YY})\,\widetilde{H}_k \\ &= \left(V^T - \mu\Lambda V^T\right)\widetilde{H}_k \\ &= (I - \mu\Lambda)\,V^T\widetilde{H}_k \end{aligned} \tag{3.9}$$

which is now a system of decoupled equations ($I-\mu\Lambda$ is a diagonal matrix). If we introduce the error components $v_i(k)$ as

$$\left(V^T\widetilde{H}_k\right) \;\triangleq\; [v_1(k)\;\; v_2(k)\cdots v_N(k)]^T \tag{3.10}$$

then we find the following decoupled dynamics for the ith *natural mode*:

$$v_i(k+1) \;=\; (1 - \mu\lambda_i)\,v_i(k)\;, \quad i = 1, \ldots, N\;. \tag{3.11}$$

The solution of these homogeneous difference equations of first order is readily obtained as

$$v_i(k) \;=\; (1 - \mu\lambda_i)^k\,v_i(o)\;, \quad i = 1, \ldots, N\;. \tag{3.12}$$

The numbers $v_i(k)$ represent a *geometric series* with a geometric ratio equal to $1-\mu\lambda_i$. For *stability* or *convergence* of the steepest-descent algorithm, the magnitude of this geometric ratio must be less than 1 for all i, or

$$-1 < 1-\mu\lambda_i < 1\;, \quad i = 1, \ldots, N\;, \tag{3.13}$$

which will guarantee

$$\lim_{k\to\infty} (1-\mu\lambda_i)^k \;=\; 0\;, \quad i = 1, \ldots, N\;, \tag{3.14}$$

Then, as the number of iterations, k, approaches infinity, all the natural modes of the steepest-descent algorithm die out, irrespective of the initial conditions. This is equivalent to saying that the FIR filter coefficient vector H_k approaches the optimum solution H^o as k approaches infinity.

With the eigenvalues of R_{YY} being real and positive, and ordered as mentioned before, the necessary and sufficient condition (3.13) for the convergence of the algorithm is that the stepsize parameter μ satisfy the following condition:

$$0 \;<\; \mu \;<\; \frac{2}{\lambda_1}\;. \tag{3.15}$$

In absence of knowledge of the eigenvalues of R_{YY}, the following reasoning leads to a safe operating region:

$$\lambda_1 \;<\; \sum_{i=1}^{N} \lambda_i \;=\; \mathrm{tr}\, R_{YY} \;=\; N\sigma_y^2 \,, \tag{3.16}$$

assuming that y_k is stationary so that all diagonal elements of R_{YY} are σ_y^2, from which we obtain

$$0 \;<\; \mu \;<\; \frac{2}{N\sigma_y^2} \;<\; \frac{2}{\lambda_1} \,. \tag{3.17}$$

In practice, it will be much easier to determine σ_y^2 than λ_1.

By taking the unit of time to be the duration of one iteration cycle, we can associate a time constant τ_i with the ith natural mode (for what follows, we assume $1-\mu\lambda_i > 0$):

$$1 - \mu\lambda_i \;=\; e^{-\frac{1}{\tau_i}} \quad\Rightarrow\quad \tau_i \;=\; \frac{-1}{ln\,(1-\mu\lambda_i)} \tag{3.18}$$

which can be approximated, for small values of μ, as

$$\tau_i \;\approx\; \frac{1}{\mu\lambda_i} \,, \quad \mu \ll 1 \,. \tag{3.19}$$

This shows that the smaller the stepsize parameter μ, the slower will be the convergence of the steepest-descent algorithm.

To see how the natural modes determine the evolution of the filter estimate H_k in the original coordinate system, consider

$$H \;=\; H^o - \widetilde{H} \;=\; H^o - V\left(V^T\widetilde{H}_k\right) \;=\; H^o - \sum_{i=1}^{N} V_i v_i(k) \,. \tag{3.20}$$

This coordinate transformation represents a translation followed by a generalized rotation (a combination of rotations, reflections and permutations of the basis vectors). Indeed, after a translation over H^o and a reflection w.r.t. the origin, H becomes $\widetilde{H}$. $\widetilde{H}$ is represented w.r.t. to a standard coordinate system:

$$\widetilde{H} \;=\; I\,\widetilde{H} \;=\; \sum_{i=1}^{N} \begin{bmatrix} 0_{(i-1)\times 1} \\ 1 \\ 0_{(N-i)\times 1} \end{bmatrix} \tilde{h}_{i-1} \;=\; V\left(V^T\widetilde{H}\right) \;=\; [V_1 \cdots V_N] \begin{bmatrix} v_1 \\ \vdots \\ v_N \end{bmatrix} \;=\; \sum_{i=1}^{N} V_i v_i \,. \tag{3.21}$$

So, when represented in the basis composed of the eigen vectors, the coordinates of $\widetilde{H}$ are $[v_1 \cdots v_N]$. The evolution of the filter estimate H_k in terms of the natural modes becomes

$$\begin{aligned} H_k \;&=\; H^o - \widetilde{H}_k \;=\; H^o - V\left(V^T\widetilde{H}_k\right) \\ &=\; H^o - \sum_{i=1}^{N} V_i v_i(k) \\ &=\; H^o - \sum_{i=1}^{N} V_i v_i(0)\,(1-\mu\lambda_i)^k \,. \end{aligned} \tag{3.22}$$

Further insight into the operation of the steepest-descent algorithm may be obtained by considering the evolution of the MSE. From (2.83), we get

$$\begin{aligned}\xi(H) &= \xi^o + \widetilde{H}^T R_{YY} \widetilde{H} = \xi^o + \widetilde{H}^T V \Lambda V^T \widetilde{H} \\ &= \xi^o + [v_1 \cdots v_N] \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_N \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_N \end{bmatrix} = \xi^o + \sum_{i=1}^{N} \lambda_i v_i^2 \, . \end{aligned} \tag{3.23}$$

Hence, the set of points H of constant cost function, $\xi(H) = c'$, is given by

$$\sum_{i=1}^{N} \frac{v_i^2}{1/\lambda_i} = c = c' - \xi^o \, . \tag{3.24}$$

This is an ellipsoid (ellips when $N = 2$) centered at H^o and with principal axes coinciding with the eigenvectors of R_{YY}, see Fig. 3.2 for $N = 2$. In Fig. 3.2, the intersections with the principal axes are indicated for the ellips corresponding to $c = 1$. The convergence of the H_k is traced for a case in which $1-\mu\lambda_1 < 0$, $1-\mu\lambda_2 > 0$ and $|1-\mu\lambda_1| \ll |1-\mu\lambda_2|$. So for the component $v_1(k)$ along the V_1-axis, we get alternating signs and a fairly fast convergence. For the component $v_2(k)$ along the V_2-axis, we get a more slowly decaying exponential of constant sign.

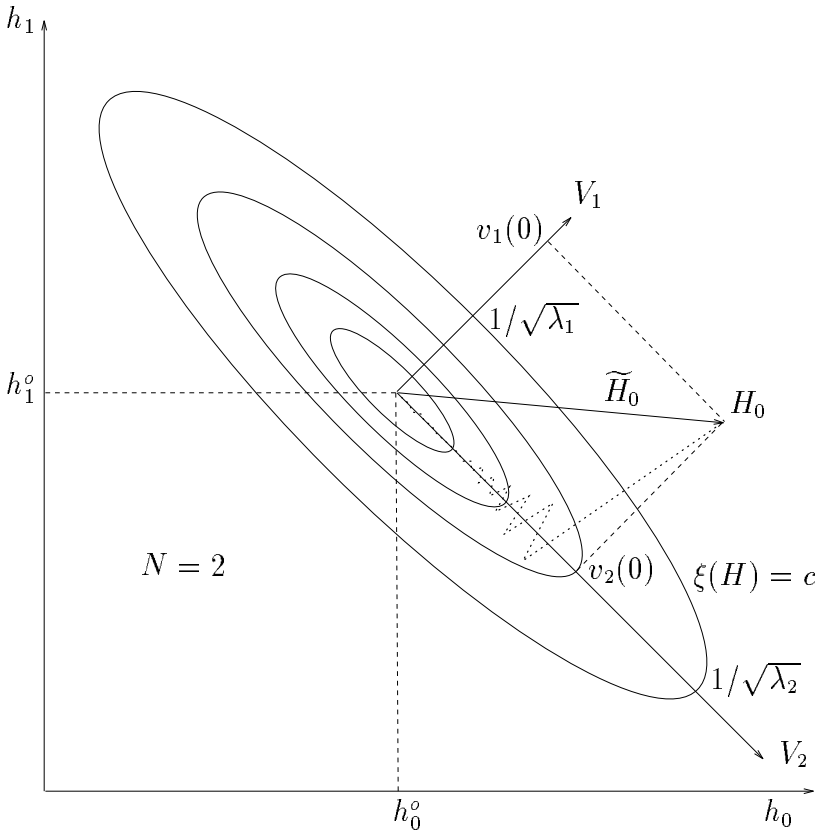


Figure 3.2: Convergence of H_k decomposed.

The evolution of the MSE can now be written as

$$\begin{aligned} \xi_k \;=\; \xi(H_k) \;&=\; \xi^o + \widetilde{H}_k^T R_{YY}\, \widetilde{H}_k \;=\; \xi^o + \widetilde{H}_k^T V \Lambda V^T \widetilde{H}_k \;=\; \xi^o + \sum_{i=1}^{N} \lambda_i v_i^2(k) \\ &=\; \xi^o + \sum_{i=1}^{N} \lambda_i \left(1-\mu\lambda_i\right)^{2k} v_i^2(0) \;\triangleq\; \xi^o + \xi_k^e \;\geq\; \xi^o \,. \end{aligned} \tag{3.25}$$

The MSE is the sum of the minimum MSE (MMSE) ξ^o and what is called the Excess MSE (EMSE) ξ^e. The curve obtained by plotting the MSE ξ_k as a function of iteration number k is called the *learning curve*. So in general, the learning curve of the steepest-descent algorithm consists of a sum of N exponentials, each one of which corresponds to a natural mode of the algorithm. Again, when the stepsize satisfies the convergence condition (3.15), we get that irrespective of the initial conditions

$$\lim_{k\to\infty} \xi_k \;=\; \xi^o \;,\;\; \lim_{k\to\infty} \xi_k^e \;=\; 0 \,. \tag{3.26}$$

Note that the time constants for the MSE are half of the corresponding time constants for the natural modes in the convergence of the filter estimate. As shown in Fig. 3.3, the learning curve is dominated by different modes at different stages (if the eigenvalues are well separated). The (approximate) piecewise linear form of the learning curve indicated in Fig. 3.3 holds when $\lambda_1 \gg \lambda_2 \gg \cdots \gg \lambda_N$ and $1-\mu\lambda_1 > 0$. In that case $\tau_1 \ll \tau_2 \ll \cdots \ll \tau_N$. In Fig. 3.3, the MSE is plotted in dB. Hence an exponential decay of ξ_k corresponds to a linear decay of $10\log_{10}\xi_k$ in the figure. In the beginning, the learning curve follows the decay of the fastest mode, associated with λ_1. After about $2\tau_1 = 4\frac{\tau_1}{2}$ (the time constants for ξ_k are roughly half those for H_k), the first mode has died out while the other modes have hardly decreased in this short time span. Now the decay associated with mode two dominates the learning curve until about $2\tau_2$ etc.

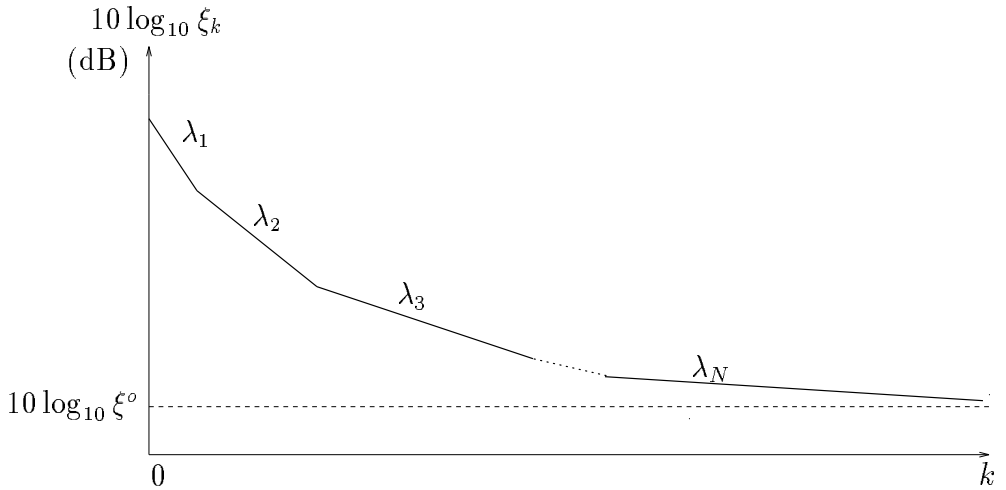


Figure 3.3: The learning curve.

If $1-\mu\lambda_1 < 0$, then a similar reasoning still applies. But then the time constants are not ordered according to the λ_i but according to the magnitudes $|1-\mu\lambda_i|$.

The value of the stepsize μ which leads to fastest convergence can be obtained by considering the following optimization problem

$$\min_{\mu} \max_{i} |1-\mu\lambda_i| \tag{3.27}$$

which leads to the following condition

$$1-\mu^o\lambda_1 = -(1-\mu^o\lambda_N) \tag{3.28}$$

with solution

$$\mu^o = \frac{2}{\lambda_1+\lambda_N} < \frac{2}{\lambda_1} \tag{3.29}$$

which is the inverse of the arithmetic average of largest and smallest eigenvalue. We also get for the slowest mode (but corresponding to the optimal stepsize)

$$\min_{\mu} \max_{i} |1-\mu\lambda_i| = \frac{\lambda_1-\lambda_N}{\lambda_1+\lambda_N} . \tag{3.30}$$

This shows that in general the convergence of the steepest-descent algorithm slows down as the eigenvalue spread (the ratio λ_1/λ_N) increases, as illustrated in Fig. 3.4. We say "in general", because it would be possible that e.g. $v_N(0) = 0$. In that case, it is the spread λ_1/λ_{N-1} that counts. Equation (3.30) shows that the convergence is the fastest when all eigenvalues are equal. In that case

$$\min_{\mu} \max_{i} |1-\mu\lambda_i| = 0 \tag{3.31}$$

for $\mu^o = 1/\lambda_1$. This means that convergence occurs in one iteration! Indeed, when all eigenvalues are equal, then the ellipsoids in (3.24) become spheres. In this case, all negative gradients (at any point in H space) point towards H^o. So it suffices to take the right stepsize to end up at H^o in one step.

3.2 The Least Mean Square (LMS) Algorithm

Let's just drop the mathematical expectation operator E in the MSE criterion, and instead consider the instantaneous squared error. So the new cost function is simply

$$J_k(H) = \epsilon_k^2 = (x_k - H^T Y_k)^2 . \tag{3.32}$$

So we have dropped the statistical averaging operation, but we hope to replace it by some amount of time domain averaging inherent in the low-pass filtering nature of the adaptation process. Just as the criterion (3.32) can be considered an instantaneous estimate of the MSE, the true criterion of interest, the gradient of J_k can be considered to be an estimate of the true gradient. We get

$$\widehat{\nabla}_k = \left.\frac{\partial J_k}{\partial H}\right|_{H=H_{k-1}} = \left[\begin{array}{c} \frac{\partial \epsilon_k^2}{\partial h_0} \\ \frac{\partial \epsilon_k^2}{\partial h_1} \\ \vdots \\ \frac{\partial \epsilon_k^2}{\partial h_{N-1}} \end{array}\right]\Bigg|_{H=H_{k-1}} = 2\epsilon_k^p \left[\begin{array}{c} \frac{\partial \epsilon_k}{\partial h_0} \\ \frac{\partial \epsilon_k}{\partial h_1} \\ \vdots \\ \frac{\partial \epsilon_k}{\partial h_{N-1}} \end{array}\right]\Bigg|_{H=H_{k-1}} = -2\epsilon_k^p Y_k \tag{3.33}$$

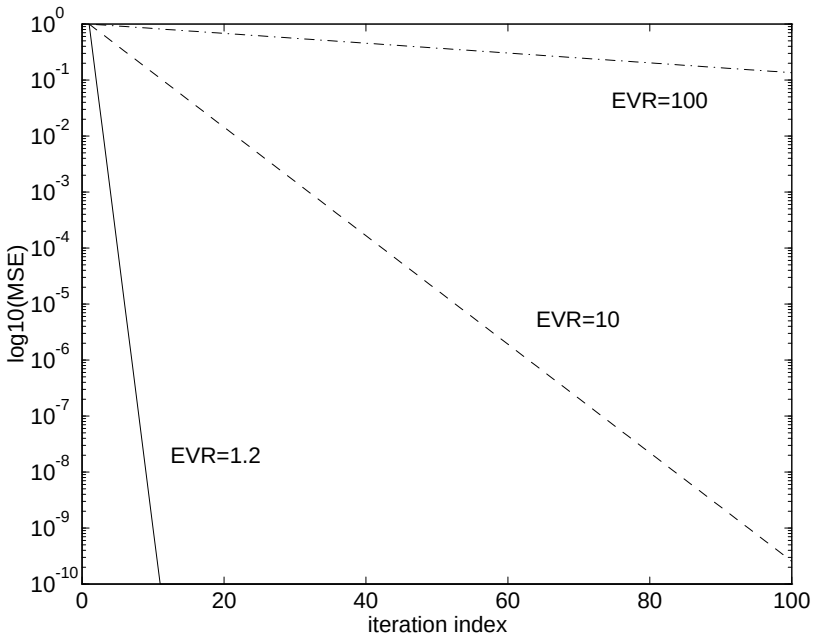


Figure 3.4: The learning curve for N=2, optimal stepsize μ^o, in the case of three eigenvalue ratios EVR=λ_1/λ_N.

where

$$\epsilon_k^p \triangleq x_k - H_{k-1}^T Y_k \tag{3.34}$$

is the *predicted* or *a priori* error signal, obtained as the difference of the desired response and the filter output at time k, but using the filter estimate at time $k-1$.

With this simple estimate of the gradient, we can specify a steepest-descent type of adaptive algorithm. We get

$$\begin{aligned} H_k &= H_{k-1} - \tfrac{\mu}{2}\widehat{\nabla}_k \\ &= H_{k-1} + \mu\epsilon_k^p Y_k \end{aligned} \tag{3.35}$$

Equations (3.34) and (3.35) specify the LMS algorithm. Whereas the index k in the steepest-descent algorithm was just an iteration number, here k is really the discrete time index, which proceeds as new samples enter the adaptation process. In fact, the LMS algorithm does one iteration per sample period. The LMS algorithm is a specific type of stochastic gradient algorithm, so called because the gradient is stochastic (is determined by stochastic processes). Since the filter adjustments at each sample instant are based on imperfect gradient estimates, we would expect the adaptive process to be noisy; that is, it would not follow the true line of steepest descent on the performance surface $\xi(H)$. Note that the LMS algorithm can be implemented in a practical system without squaring, averaging, or differentiation and is elegant in its simplicity and computational efficiency ($2N$ operations per sample).

The main question now is, does H_k converge to the Wiener solution H^o in any sense?

3.2.1 A Model for the Purpose of Analysis

We now proceed to discuss some of the properties of the LMS algorithm. The convergence analysis of stochastic gradient algorithms requires very sophisticated tools to be carried out

properly. We shall restrict the discussion to a simplified treatment that can be made rigourous though, by filling in the omitted mathematical details.

Consider the output $\widehat{x}_k = Y_k^T H^o$ produced by the optimal Wiener filter $H^o = R_{YY}^{-1} R_{Yx}$. Then we can write the filtering error as $\widetilde{x}_k \stackrel{\triangle}{=} x_k - \widehat{x}_k$. Using the filtering error, we can decompose the signal x_k as

$$x_k \;=\; \widehat{x}_k + \widetilde{x}_k \;=\; Y_k^T H^o + \widetilde{x}_k \tag{3.36}$$

which is a decomposition into the part of x_k that can be linearly predicted from Y_k and the remainder, the part that is orthogonal to Y_k.

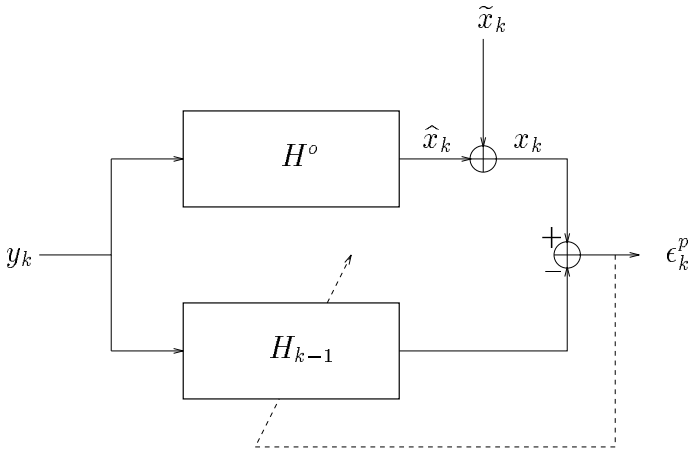


Figure 3.5: The system identification set-up.

With H^o being the optimal Wiener filter, the zero-mean noise $\widetilde{x}_k$ in the model (3.36) for the signal x_k satisfies the following conditions:

(i) $E\widetilde{x}_k y_i = 0,\ i = k, k-1, \ldots, k-N+1$: since H^o satisfies the orthogonality condition, $\widetilde{x}_k$ is uncorrelated with Y_k.

(ii) $E\widetilde{x}_k^2 = \xi^o = \sigma_x^2 - R_{xY} R_{YY}^{-1} R_{Yx} =$ constant if $\{x_k\}$ and $\{y_k\}$ are jointly stationary.

In order to simplify the analysis of the LMS algorithm, we shall augment these conditions which follow from the optimality of the FIR Wiener filter with some assumptions. Indeed, we shall assume that in fact the signal x_k is generated by passing the signal y_k through some filter H^o and adding some noise $\widetilde{x}_k$ to the filter output (see equation (3.36)). A complete specification of y_k and $\widetilde{x}_k$ jointly then determines the joint description of y_k and x_k. We shall assume the following about the stationary zero-mean processes y_k and $\widetilde{x}_k$:

(i') $\widetilde{x}_k$ is independent of $\{y_n\}$ (and not only uncorrelated with Y_k as in (i)),

(ii') $\{\widetilde{x}_k\}$ is i.i.d. (independent identically distributed, white noise) (and not only of constant variance as in (ii)).

Since we have only augmented the conditions (i),(ii) to obtain (i'),(ii'), the conditions (i),(ii) are still satisfied, which implies that the linear FIR filter H^o that is used to generate x_k is the Wiener filter. Indeed, H^o satisfies the orthogonality conditions

$$E(x_k - Y_k^T H^o)Y_k \;=\; E\widetilde{x}_k Y_k \;=\; 0\,. \tag{3.37}$$

If we now consider estimating x_k from Y_k by some arbitrary FIR filter H, then the associated MSE becomes

$$\begin{aligned}
\xi(H) &= E(x_k - H^T Y_k)^2 \;=\; E(\widetilde{x}_k + H^{oT} Y_k - H^T Y_k)^2 \\
&= E(\widetilde{H}^T Y_k + \widetilde{x}_k)^2 \;\;,\;\; \widetilde{H} = H^o - H \\
&= E\widetilde{x}_k^2 + 2\widetilde{H}^T E Y_k \widetilde{x}_k + \widetilde{H}^T (E Y_k Y_k^T)\widetilde{H} \\
&= \xi^o + \widetilde{H}^T R_{YY} \widetilde{H}
\end{aligned} \tag{3.38}$$

where we used $EY_k\widetilde{x}_k = (EY_k)(E\widetilde{x}_k) = 0$. This shows among other things that $H = H^o$ is indeed the Wiener solution. It also allows us to evaluate the MSE associated with an arbitrary filter H and in particular it shows how the MSE augments from the MMSE ξ^o for $H = H^o$. We could use expression (3.38) to evaluate the MSE associated with the estimate H_k provided by the LMS algorithm. Since H_k is in this case a stochastic quantity, we would still need to average (3.38) over H_k though. But we should also remark that the evaluation of $\xi(H)$ in (3.38) requires the availability of ξ^o, R_{YY} and H^o (to form $\widetilde{H}$). We shall show now that $E\xi(H_{k-1})$ can be approximated by the variance of the a priori filtering error ϵ_k^p, which can be more easily estimated in practice. The approximation involved comes from the so-called *independence assumption* that we shall introduce. This assumption treats H_{k-1} and Y_k as if they were independent. Though this assumption is hardly ever true in practice, the approximation that follows from applying this assumption gives acceptable results. So consider now

$$\begin{aligned}
E(\epsilon_k^p)^2 &= E(x_k - H_{k-1}^T Y_k)^2 \;=\; E(\widetilde{x}_k + H^{oT} Y_k - H_{k-1}^T Y_k)^2 \\
&= E(\widetilde{H}_{k-1}^T Y_k + \widetilde{x}_k)^2 \;\;,\;\; \widetilde{H}_k = H^o - H_k \\
&= E\widetilde{x}_k^2 + 2E\widetilde{H}_{k-1}^T Y_k \widetilde{x}_k + E\widetilde{H}_{k-1}^T (Y_k Y_k^T)\widetilde{H}_{k-1} \\
&= E\widetilde{x}_k^2 + 2(E\widetilde{H}_{k-1}^T Y_k)(E\widetilde{x}_k) + \mathrm{tr}\left(E(Y_k Y_k^T)(\widetilde{H}_{k-1}\widetilde{H}_{k-1}^T)\right) \\
&= E\widetilde{x}_k^2 + 2(E\widetilde{H}_{k-1}^T Y_k)0 + \mathrm{tr}\left((EY_k Y_k^T)(E\widetilde{H}_{k-1}\widetilde{H}_{k-1}^T)\right) \\
&= \xi^o + \mathrm{tr}(R_{YY} C_{k-1}) \;=\; E\xi(H_{k-1})
\end{aligned} \tag{3.39}$$

where we used the independence assumption and we introduced $C_k \triangleq E\widetilde{H}_k\widetilde{H}_k^T$. If we have convergence of the correlation matrix C_k to zero: $C_k \to 0$ as $k \to \infty$, then we have convergence of the filter estimate:

$$\widetilde{H}_k \to 0 \quad \text{or} \quad H_k \to H^o \quad \text{in mean square.} \tag{3.40}$$

Note that convergence of the correlation matrix implies convergence of the mean and covariance of the estimation error:

$$C_k = E\widetilde{H}_k\widetilde{H}_k^T = (E\widetilde{H}_k)(E\widetilde{H}_k)^T + E(\widetilde{H}_k - E\widetilde{H}_k)(\widetilde{H}_k - E\widetilde{H}_k)^T\,. \tag{3.41}$$

In summary, in order to investigate the learning curve (MSE) of the LMS algorithm (the curve of $E\xi(H_k)$), we need to investigate how C_k evolves.

From the LMS algorithm we get

$$\begin{aligned} \widetilde{H}_k \;=\; H^o - H_k \;&=\; H^o - H_{k-1} - \mu\epsilon_k^p Y_k \\ &=\; \widetilde{H}_{k-1} - \mu(Y_k^T \widetilde{H}_{k-1} + \widetilde{x}_k)Y_k \\ \widetilde{H}_k \;&=\; \left(I - \mu Y_k Y_k^T\right)\widetilde{H}_{k-1} - \mu\widetilde{x}_k Y_k \end{aligned} \tag{3.42}$$

By applying the independence assumption, the analysis of the first order moments of $\widetilde{H}_k$ becomes straightforward. Taking expected value of both sides of (3.42) yields

$$E\widetilde{H}_k \;=\; (I - \mu\, R_{YY})\; E\widetilde{H}_{k-1} \tag{3.43}$$

where we used the independence of Y_k and $\widetilde{H}_{k-1}$ (independence assumption), and the independence of $\widetilde{x}_k$ and Y_k who both have zero mean. So the mean filter error $E\widetilde{H}_k$ in the LMS algorithm converges in the same fashion as the filter error in the steepest-descent algorithm. However, it is not because the mean of the filter error converges to zero that the filter error converges to zero. To see how $E(\epsilon_k^p)^2$ converges, we have to analyze the second-order moments of $\widetilde{H}_k$.

Compared to the steepest-descent algorithm, the system (3.42) for $\widetilde{H}_k$ now has a driving term (input to the system). Also, in the system transition matrix, R_{YY} got replaced by $Y_k Y_k^T$. The system matrix $I-\mu Y_k Y_k^T$ is now stochastic. This makes the system (3.42) very hard to analyze. However, it is possible to introduce a simplification when the stepsize is small.

3.2.2 Averaging Theorem

Consider a stochastic dynamic system

$$\zeta_k \;=\; (I - \mu F_k)\zeta_{k-1} + W_k \tag{3.44}$$

where ζ_k and W_k are vectors of length N and F_k is a matrix valued stochastic process of size $N \times N$ with bounded moments. Assume that the stochastic processes involved are stationary. Now consider a different system in which the stochastic dynamics $I-\mu F_k$ are replaced by the average dynamics $I-\mu EF_k$, while the input W_k remains unchanged:

$$\overline{\zeta}_k \;=\; (I - \mu EF_k)\overline{\zeta}_{k-1} + W_k\ . \tag{3.45}$$

To compare the difference between the two systems, consider

$$\zeta_k \;=\; (I - \mu EF_k + \mu(EF_k - F_k))\zeta_{k-1} + W_k \tag{3.46}$$

and hence by subtracting both sides of (3.45) from both sides of (3.46), we get

$$(\zeta_k - \overline{\zeta}_k) \;=\; (I - \mu EF_k)(\zeta_{k-1} - \overline{\zeta}_{k-1}) + \mu(EF_k - F_k))\zeta_{k-1} \tag{3.47}$$

and hence

$$\zeta_k - \overline{\zeta}_k \;\sim\; \mu \tag{3.48}$$

The averaging theorem consists of the following weak convergence result:

$$\textit{distribution of } \{\zeta_k\} \;\;\rightarrow\;\; \textit{distribution of } \left\{\overline{\zeta}_k\right\} \quad \textit{as } \mu \rightarrow 0 \tag{3.49}$$

if the averaged system is exponentially stable. So if we want to study first and second order moments of ζ_k, we may as well study those of $\overline{\zeta}_k$ which is generated by the simpler averaged system. The results become correct in the limit as the stepsize becomes very small.

3.2.3 The LMS Learning Curve with Constant Stepsize

Applying the averaging theorem to the LMS algorithm, we replace the system (3.42) by the system with averaged dynamics

$$\widetilde{H}_k = (I - \mu R_{YY})\,\widetilde{H}_{k-1} - \mu \tilde{x}_k Y_k \tag{3.50}$$

and we shall keep the same notation for the output $\widetilde{H}_k$ of this averaged system. This averaged system is the same as the system for the steepest-descent algorithm, except that the input is nonzero now. Due to this stochatic input, the $\widetilde{H}_k$ form a stochastic process. Let us first investigate the mean of this process. Taking expected value of both sides of (3.50), we find

$$E\widetilde{H}_k = (I - \mu R_{YY})\,E\widetilde{H}_{k-1} \tag{3.51}$$

where we used the independence and zero mean of $\tilde{x}_k$ and Y_k to put $E\tilde{x}_k Y_k = 0$. We find that the mean of the filter estimated by the LMS algorithm behaves exactly as the filter values appearing in the iterations of the steepest-descent algorithm. So we can refer to the analysis of the steepest-descent algorithm for further results on the behavior of $E\widetilde{H}_k$.

We are mostly interested in the behavior of the learning curve (MSE of the filter output) though. We have seen that the MSE depends on the second-order moments of the filter estimation error $\widetilde{H}_k$. From (3.39), we have

$$\begin{aligned} \xi_k \triangleq \xi(H_k) = E(\epsilon^p_{k+1})^2 &= \xi^o + \mathrm{tr}(R_{YY}C_k) \\ &= \xi^o + \mathrm{tr}(V\Lambda V^T C_k) = \xi^o + \mathrm{tr}(\Lambda V^T C_k V) \\ &= \xi^o + \textstyle\sum_{i=1}^N \lambda_i (V^T C_k V)_{ii} \end{aligned} \tag{3.52}$$

Now, we have that

$$V^T C_k V = V^T E\widetilde{H}_k \widetilde{H}_k^T V = E\left(V^T\widetilde{H}_k\right)\left(V^T\widetilde{H}_k\right)^T = E\begin{bmatrix} v_1(k) \\ \vdots \\ v_N(k) \end{bmatrix}\begin{bmatrix} v_1(k) \\ \vdots \\ v_N(k) \end{bmatrix}^T \tag{3.53}$$

and hence

$$(V^T C_k V)_{ii} = E v_i^2(k) \tag{3.54}$$

which allows us to rewrite (3.52) as

$$\xi_k = \xi^o + \sum_{i=1}^N \lambda_i E v_i^2(k)\,. \tag{3.55}$$

Representing (3.50) in transformed coordinates with transformation matrix V, we get

$$V^T\widetilde{H}_k = (I - \mu\Lambda)\,V^T\widetilde{H}_{k-1} - \mu\tilde{x}_k V^T Y_k\,. \tag{3.56}$$

By putting

$$\begin{aligned} V^T\widetilde{H}_k &= \begin{bmatrix} v_1(k) \\ \vdots \\ v_N(k) \end{bmatrix}, \quad V^T Y_k = \begin{bmatrix} w_1(k) \\ \vdots \\ w_N(k) \end{bmatrix}, \\ E\begin{bmatrix} w_1(k) \\ \vdots \\ w_N(k) \end{bmatrix}&\begin{bmatrix} w_1(k) \\ \vdots \\ w_N(k) \end{bmatrix}^T = EV^T Y_k Y_k^T V = V^T R_{YY} V = \Lambda\,, \end{aligned} \tag{3.57}$$

we can rewrite (3.56) for component i as

$$v_i(k) = (1 - \mu\lambda_i)\, v_i(k-1) - \mu\widetilde{x}_k w_i(k)\ . \tag{3.58}$$

By taking the expected value of the square of both sides, we get

$$Ev_i^2(k) = (1-\mu\lambda_i)^2\, Ev_i^2(k-1) + \mu^2 E\widetilde{x}_k^2 w_i^2(k) \;\; = \;\; (1-\mu\lambda_i)^2\, Ev_i^2(k-1) + \mu^2\xi^o\lambda_i \tag{3.59}$$

where the crossterms disappear due to the independence of $\widetilde{x}_k$ and $w_i(k)$, $v_i(k-1)$. Since we have the same dynamics as in the steepest-descent algorithm, we shall again have stability (convergence) when $0 < \mu < \frac{2}{\lambda_1}$. However, the results here are based on the averaging theorem which assumes that μ is small. If based on such a theory, we derive a condition for the maximum possible value for μ to have convergence, then we can expect that result to be erroneous. Indeed, the range of stable stepsize values for convergence of the second-order moments in practice turns out to be quite a bit more conservative than $\frac{2}{\lambda_1}$. In any case, for μ within the stable range, the scalar systems in (3.59) converge exponentially fast and $Ev_i^2(k)$ will converge to some value $Ev_i^2(\infty)$. This value can be obtained from (3.59) by taking $k = \infty$ and one gets

$$Ev_i^2(\infty) \;\; = \;\; \frac{\mu}{2-\mu\lambda_i}\xi^o\ . \tag{3.60}$$

The learning curve converges to the value

$$\xi_\infty \;\; = \;\; \xi^o + \sum_{i=1}^{N}\lambda_i Ev_i^2(\infty) \;\; = \;\; \xi^o + \xi^o\mu\sum_{i=1}^{N}\frac{\lambda_i}{2-\mu\lambda_i} = \xi^o(1+M) \tag{3.61}$$

where

$$M \;\; = \;\; \mu\sum_{i=1}^{N}\frac{\lambda_i}{2-\mu\lambda_i} \;\; \approx \;\; \frac{\mu}{2}\sum_{i=1}^{N}\lambda_i = \frac{\mu N\sigma_y^2}{2} \tag{3.62}$$

is called the *misadjustment* factor and the indicated approximation holds for small μ. Although $EH_\infty = H^o$ (the estimate is asymptotically unbiased), the estimate H_k of H^o is not *consistent* because its variance does not decrease to zero. Due to the use of a finite nonzero stepsize, the LMS algorithm will, even in steady-state, continue taking steps in the noisy gradient direction (even though the true gradient would be zero at convergence). This residual amount of variance is measured by the misadjustment factor M or by the Excess MSE $\xi_\infty^e = \xi^o M$. The design of a constant stepsize μ results from a compromise: small values for μ lead to low misadjustment but slow convergence dynamics, large values lead to the opposite. To elaborate on this compromise, we can find ξ_k explicitly from (3.55) and (3.59) and we get

$$\xi_k \;\; = \;\; \xi^o + \sum_{i=1}^{N}\lambda_i(1-\mu\lambda_i)^{2k}\, v_i^2(0) + \xi^o\,\mu\sum_{i=1}^{N}\frac{1-(1-\mu\lambda_i)^{2k}}{2-\mu\lambda_i}\lambda_i\ , \quad k \geq 1\,. \tag{3.63}$$

The first term is the MMSE. The second term is due to the mean of $\widetilde{H}_k$. This term is initially dominating, but converges exponentially to zero. The third term is due to the variance of $\widetilde{H}_k$. It is initially zero, but it builds up to the excess MSE. To make H_k converge fast, we have to decrease $E\widetilde{H}_k$ as fast a possible with a large step size μ so that H_k gets in the neighborhood of H^o. However, once the mean $E\widetilde{H}_k$ has converged, H_k continues to jitter around H^o. To

limit the variance of this jitter (excess MSE), we have to limit the value of the step size μ. We have illustrated this compromise in the design of μ in Fig. 3.6 for a white noise input with $N = 10$, initial MSE $\xi_0 = 1$ (0 dB), MMSE $\xi^o = 0.01$ (-20 dB) and two values of the step size μ: $\mu = 1/N\sigma_y^2$ and $\mu = 0.1/N\sigma_y^2$. The large step size allows the algorithm to converge in about 40 samples, but leads to a misadjustment factor of almost 2 dB. The smaller step size leads to a much slower convergence in 400 samples, but to a negligible misadjustment of 0.2 dB.

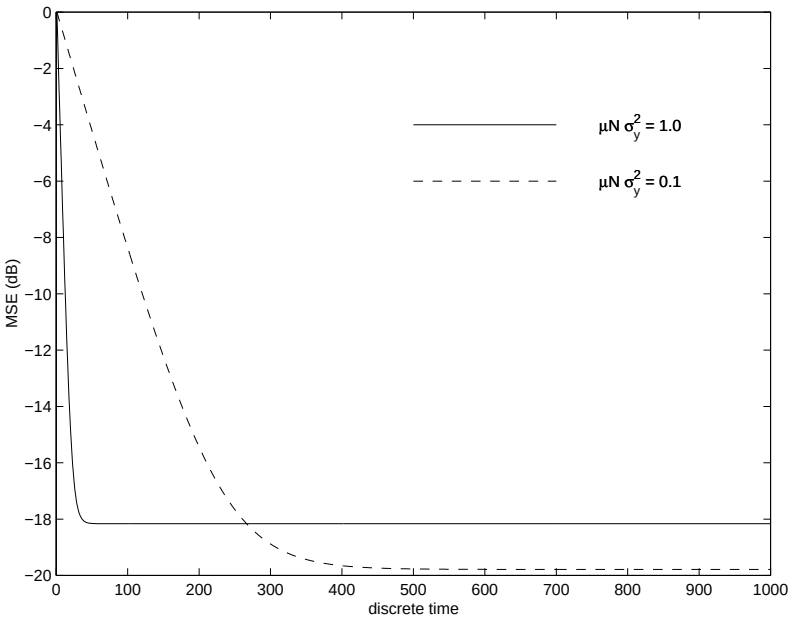


Figure 3.6: LMS learning curves $10 \log \xi_k$ (in dB) for N=10, $\xi_0 = 1$ (0 dB), $\xi^o = 0.01$ (-20 dB) and two values of the step size μ.

3.2.4 Conditions on the Stepsize for Exact Convergence

What if we require $M = 0$. Then we need $\mu = 0$ from our previous steady-state considerations. However, $\mu = 0$ will clearly not allow any adaptation. So what we need to do is to vary μ with k. The time-varying μ_k will have some finite value initially and decrease to zero as $k \to \infty$. One can show that under the following conditions

$$\begin{array}{rcl} \mu_k & \geq & 0 \\ \sum_{k=1}^{\infty} \mu_k & = & \infty \\ \sum_{k=1}^{\infty} \mu_k^2 & < & \infty \end{array} \tag{3.64}$$

the misadjustment converges to zero and $H_k \to H^o$, the Wiener solution, in mean square. A typical stepsize sequence that satisfies these conditions is

$$\mu_k = \frac{c}{k} \tag{3.65}$$

where c is some constant.

3.2.5 The Normalized LMS Algorithm

The LMS algorithm applies the steepest descent approach to the instantaneous squared filtering error $\epsilon_k^2(H) = \left(x_k - H^T Y_k\right)^2$, and can be written as

$$\begin{cases} \epsilon_k^p &= x_k - H_{k-1}^T Y_k \\ H_k &= H_{k-1} + \mu \epsilon_k^p Y_k \quad , \end{cases} \tag{3.66}$$

where $\epsilon_k^p = \epsilon_k(H_{k-1})$ is called the *a priori* (or *predicted*) filtering error. From our previous analysis, it is clear that the stable range for the stepsize is inversely proportional to the variance of the input signal: $\mu \sim 1/\sigma_y^2$. Therefore, in order to be robust w.r.t. possible variations in the level of the input signal y_k, it is desirable to normalize the stepsize, to divide the stepsize by a quantity that behaves roughly as the variance σ_y^2. Therefore, the so-called Normalized LMS (NLMS) algorithm normalizes the stepsize by $\|Y_k\|^2 = Y_k^T Y_k$,

$$\mu_k = \frac{\overline{\mu}}{\|Y_k\|^2} , \tag{3.67}$$

and hence is given by

$$\begin{cases} \epsilon_k^p &= x_k - H_{k-1}^T Y_k \\ H_k &= H_{k-1} + \dfrac{\overline{\mu}}{\|Y_k\|^2} \epsilon_k^p Y_k \quad . \end{cases} \tag{3.68}$$

Note that $\mu_k = \dfrac{\overline{\mu}}{\|Y_k\|^2}$ is indeed roughly inversely proportional to σ_y^2 since $E\|Y_k\|^2 = N\sigma_y^2$ (when $\{y_k\}$ is stationary). However, the consequences of the normalization of the stepsize by $\|Y_k\|^2$ rather than by (a multiple of) σ_y^2 reach much farther than can be imagined from the considerations put forward so far.

Indeed, consider the *a posteriori* filtering error, which is the filtering error evaluated at the updated coefficients H_k , $\epsilon_k = \epsilon_k(H_k)$. One can show that for

$$\begin{cases} \text{LMS}: \quad \epsilon_k &= (1 - \mu\|Y_k\|^2)\, \epsilon_k^p \\ \text{NLMS}: \quad \epsilon_k &= (1 - \overline{\mu})\, \epsilon_k^p \quad . \end{cases} \tag{3.69}$$

While the NLMS algorithm allows a precise control of the magnitude of the *a posteriori* filtering error ϵ_k , and in particular the choice $\overline{\mu} = 1$ makes $\epsilon_k \equiv 0$, the LMS algorithm only allows to make ϵ_k small on the average, with possibly large values occuring at certain time instants, depending on the variations of $\|Y_k\|^2$ with time.

With the model for x_k, $x_k = H^{o\,T} Y_k + \tilde{x}_k$, that we have used for the convergence analysis (with H^o being the optimal Wiener filter coefficients), we can write for the filter deviation $\widetilde{H}_k = H^o - H_k$,

$$\begin{cases} \epsilon_k^p &= \widetilde{H}_{k-1}^T Y_k + \tilde{x}_k \\ \text{LMS}: \quad \widetilde{H}_k &= \Phi_k^{LMS} \widetilde{H}_{k-1} - \mu \tilde{x}_k Y_k \quad , & \Phi_k^{LMS} = I - \mu Y_k Y_k^T \\ \text{NLMS}: \quad \widetilde{H}_k &= \Phi_k^{NLMS} \widetilde{H}_{k-1} - \dfrac{\overline{\mu}}{\|Y_k\|^2} \tilde{x}_k Y_k \quad , & \Phi_k^{NLMS} = I - \overline{\mu} \dfrac{Y_k Y_k^T}{Y_k^T Y_k} . \end{cases} \tag{3.70}$$

The state transition matrix Φ_k determines the dynamics and in particular the stability of the error system $\widetilde{H}_k$. To investigate the long-run dynamics, we have used the averaging theorem, leading to the approximations

$$\begin{cases} \Phi_k^{LMS} & \approx & I - \mu R_{YY} \\ \Phi_k^{NLMS} & \approx & I - \dfrac{\overline{\mu}}{\mathrm{tr} R_{YY}} R_{YY} \quad , \end{cases} \tag{3.71}$$

where we have used for NLMS an additional approximation that is valid for large N. The analysis based on the averaging theorem has revealed the dependence of the dynamics on the eigenvalue spread λ_1/λ_N of R_{YY}. However, it is also interesting to look at the instantaneous dynamics. Φ_k can be shown to have the following eigenvalues

$$\begin{cases} \nu_1(k) & = & \nu_2 = \cdots = \nu_{N-1}(k) = 1 \quad \text{for LMS and NLMS} \\ \nu_N^{LMS}(k) & = & 1 - \mu \|Y_k\|^2 \\ \nu_N^{NLMS}(k) & = & 1 - \overline{\mu} \quad , \end{cases} \tag{3.72}$$

and the eigenvector $W_N(k)$ corresponding to $\nu_N(k)$ is proportional to Y_k (what are the eigenvectors corresponding to $\nu_1(k), \nu_2(k), \cdots, \nu_{N-1}(k)$?). Remark that we can write for both algorithms

$$\epsilon_k = \nu_N(k) \epsilon_k^p \quad . \tag{3.73}$$

Let $W_i(k)$ be the eigenvector associated with eigenvalue $\nu_i(k)$. The eigenvectors form an orthonormal basis for $\mathcal{R}^N$ and we can always write $\widetilde{H}_k = WW^T \widetilde{H}_k = \sum_{i=1}^{N} W_i \left(W_i^T \widetilde{H}_k \right)$. We can write for both algorithms

$$\begin{cases} W_i^T \widetilde{H}_k & = & W_i^T \widetilde{H}_{k-1} \quad , \; i = 1, 2, \cdots, N-1 \\ W_N^T \widetilde{H}_k & = & \nu_N(k) W_N^T \widetilde{H}_{k-1} - \mu_k \widetilde{x}_k W_N^T Y_k \end{cases} \tag{3.74}$$

where

$$\mu_k = \begin{cases} \mu & \text{for LMS,} \\ \dfrac{\overline{\mu}}{\|Y_k\|^2} & \text{for NLMS.} \end{cases} \tag{3.75}$$

Hence, for the components of $\widetilde{H}_{k-1}$ along directions orthogonal to Y_k, nothing changes. The component of $\widetilde{H}_{k-1}$ along the direction of Y_k however gets multiplied by $\nu_N(k)$. For the NLMS algorithm, $\nu_N^{NLMS}(k)$ can be controlled very precisely by $\overline{\mu}$. The choice for $\overline{\mu}$ results from a trade-off between fast dynamics = small $|1-\overline{\mu}|$ = big $\overline{\mu}$, and a small noise amplification factor $\overline{\mu}$ (second (driving) term). In particular, $\overline{\mu} = 1$ results in complete elimination of the accumulated error component of $\widetilde{H}_{k-1}$ in the direction of Y_k (also, $\epsilon_k = 0$ for $\overline{\mu} = 1$). However, with $\widetilde{x}_k \neq 0$, some noise gets injected in the component of $\widetilde{H}_k$ along Y_k. For the LMS algorithm, if we want to make $\left| \nu_N^{LMS}(k) \right|$ small on the average , we need to choose μ fairly big. However, due to the statistical fluctuations of $\|Y_k\|$, $\left| \nu_N^{LMS}(k) \right|$ may get quite a bit bigger than 1 at certain times k, if μ is big. That means that even though LMS is converging on the average, it may actually take diverging steps at certain isolated time instants. This

happens especially when μ is relatively large, when we want the average convergence to be as fast as possible. This problem of instantaneous diverging steps does not occur in the NLMS algorithm. Therefore, NLMS always converges faster than LMS when both algorithms have a stepsize that is optimized for fastest convergence speed.

3.3 The Recursive Least-Squares (RLS) Algorithm

The optimal FIR Wiener filter H^o minimizes

$$\xi(H) \;=\; E\left(x_k - H^T Y_k\right)^2 \;. \tag{3.76}$$

As mentioned before, this requires the knowledge of the joint second-order statistics of x_k and Y_k. Often, this statistical information is not available in practice. In the LMS approach, this lack of information has led us to drop the statistical averaging operation. We have seen that this averaging operation gets replaced by some amount of low-pass filtering (time-domain averaging) due to the iterative nature of stochastic gradient algorithms.

An alternative approach consists of replacing the statistical averaging operation by a time-domain (sample) averaging operation. In the LS approach, we use the data up to time k and obtain H_k by minimizing the following criterion

$$\xi_k\left(H\right) = \sum_{i=1}^{k}\left(x_i - H^T Y_i\right)^2 + \left(H - H_0\right)^T R_0 \left(H - H_0\right) \;\;, \tag{3.77}$$

where the second term with $R_0 = R_0^T > 0$ allows for a proper initialization of the algorithm (the first term alone has a singular Hessian $(= 2\sum_{i=1}^{k} Y_i Y_i^T)$ for $k < N$). We can rewrite (3.77) as

$$\begin{aligned}
\xi_k\left(H\right) &= H^T\left(\sum_{i=1}^{k} Y_i Y_i^T\right) H - 2H^T\left(\sum_{i=1}^{k} Y_i x_i\right) + \sum_{i=1}^{k} x_i^2 + \left(H - H_0\right)^T R_0 \left(H - H_0\right) \\
&= H^T\left(R_0 + \sum_{i=1}^{k} Y_i Y_i^T\right) H - 2H^T\left(R_0 H_0 + \sum_{i=1}^{k} Y_i x_i\right) + \sum_{i=1}^{k} x_i^2 + H_0^T R_0 H_0 \\
&= H^T R_k H - 2H^T P_k + \sum_{i=1}^{k} x_i^2 + H_0^T R_0 H_0
\end{aligned} \tag{3.78}$$

where

$$\begin{aligned}
R_k &= R_0 + \sum_{i=1}^{k} Y_i Y_i^T \;\;\;\; = \;\; R_{k-1} + Y_k Y_k^T \\
P_k &= R_0\, H_0 + \sum_{i=1}^{k} Y_i\, x_i \;\; = \;\; P_{k-1} + Y_k x_k \;\;.
\end{aligned} \tag{3.79}$$

By putting the gradient of $\xi_k(H)$ equal to zero and noting that the Hessian $2R_k > 0$, we find that the LS filter H_k that minimizes (3.77) solves the following normal equations

$$R_k\, H_k = P_k \;\;. \tag{3.80}$$

To solve this set of equations at each time instant k would take $\mathcal{O}(N^3)$ operations at each time instant. In what follows, we shall derive the Recursive LS algorithm, which allows us, using information obtained at time $k-1$, to obtain H_k with only $\mathcal{O}(N^2)$ operations.

3.3.1 Derivation of the RLS algorithm

By using the recursive relations for R_k and P_k indicated in (3.79), we can find a recursion for H_k satisfying (3.80). Using (3.80), we can rewrite $P_k = P_{k-1} + Y_k x_k$ as

$$\begin{aligned} R_k H_k &= R_{k-1} H_{k-1} + Y_k x_k \\ &= \left(R_k - Y_k Y_k^T\right) H_{k-1} + Y_k x_k \\ &= R_k H_{k-1} + Y_k \epsilon_k^p \end{aligned} \tag{3.81}$$

where $\epsilon_k^p = x_k - H_{k-1}^T Y_k$ as in the LMS algorithm. This leads immediately to

$$H_k = H_{k-1} + R_k^{-1} Y_k \epsilon_k^p \tag{3.82}$$

where $R_k^{-1} Y_k$ is called the Kalman gain (the RLS algorithm is a special case of the so-called Kalman filter). Clearly, the RLS algorithm requires the recursive update of R_k^{-1}. This can be obtained using the Matrix Inversion Lemma:

$$\begin{aligned} R_k^{-1} &= \left(R_{k-1} + Y_k Y_k^T\right)^{-1} \\ &= R_{k-1}^{-1} - R_{k-1}^{-1} Y_k \left(1 + Y_k^T R_{k-1}^{-1} Y_k\right)^{-1} Y_k^T R_{k-1}^{-1} \quad . \end{aligned} \tag{3.83}$$

This equation allows us to obtain R_k^{-1} from R_{k-1}^{-1} and Y_k using $\mathcal{O}(N^2)$ operations. When multiplying both sides of (3.83) with Y_k to the right, we obtain

$$R_k^{-1} Y_k = R_{k-1}^{-1} Y_k \left(1 + Y_k^T R_{k-1}^{-1} Y_k\right)^{-1} \quad . \tag{3.84}$$

Using (3.82) and then (3.84), we can write for the *a posteriori* error

$$\epsilon_k = x_k - H_k^T Y_k = \left(1 - Y_k^T R_k^{-1} Y_k\right) \epsilon_k^p = \left(1 - Y_k^T R_{k-1}^{-1} Y_k\right)^{-1} \epsilon_k^p \; . \tag{3.85}$$

All this can be formulated as the RLS algorithm:

$$\left\{ \begin{array}{rcl} \epsilon_k^p &=& x_k - H_{k-1}^T Y_k \\ \epsilon_k &=& \epsilon_k^p \left(1 + Y_k^T R_{k-1}^{-1} Y_k\right)^{-1} \\ H_k &=& H_{k-1} + R_{k-1}^{-1} Y_k \epsilon_k \\ R_k^{-1} &=& R_{k-1}^{-1} - R_{k-1}^{-1} Y_k \left(1 + Y_k^T R_{k-1}^{-1} Y_k\right)^{-1} Y_k^T R_{k-1}^{-1} \quad . \end{array} \right. \tag{3.86}$$

The initial values for R_k^{-1} and H_k are R_0^{-1} and H_0 which appear in (3.77). Compared to the LMS algorithm, the scalar stepsize μ gets replaced by a matrix stepsize R_k^{-1}. The RLS algorithm takes $\mathcal{O}(N^2)$ operations while the LMS algorithm takes only $2N$ operations. However, it converges much faster.

3.3.2 Performance analysis

Assume now our usual model for x_k,

$$x_k = H^{oT} Y_k + \widetilde{x}_k \quad , \tag{3.87}$$

where the $\widetilde{x}_k$ are iid with zero mean and variance $\sigma^2_{\widetilde{x}}$. Consider here $\{y_k\}$ as a deterministic signal, so the only randomness comes from the $\{\widetilde{x}_k\}$. The H^o are the unknown parameters governing the model. The analysis of the RLS algorithm is much simpler than that of the LMS algorithm since now we have a closed-form expression, $H_k = R_k^{-1} P_k$, for the estimate H_k. Assume now $k \geq N, H_0 = 0, R_0 = 0$, and assume that R_k is nonsingular. We can write $P_k = \sum_{i=1}^{k} Y_i x_i = R_k H^o + \sum_{i=1}^{k} Y_i \widetilde{x}_i$. Hence with $\widetilde{H}_k = H^o - H_k$, we obtain

$$\widetilde{H}_k = -R_k^{-1} \sum_{i=1}^{k} Y_i \widetilde{x}_i \quad . \tag{3.88}$$

From this, we obtain

$$C_k \stackrel{\triangle}{=} E\, \widetilde{H}_k \widetilde{H}_k^T = \sigma^2_{\widetilde{x}} R_k^{-1} \quad . \tag{3.89}$$

Since R_k^{-1} behaves as $1/k$, we see that C_k converges to zero as $1/k$. If the $\widetilde{x}_k$ are furthermore Gaussian, then we see that H_k is the Maximum Likelihood estimate of H^o. The Maximum Likelihood estimate attains in this case the Cramer-Rao lower bound, which is given in (3.89).

3.3.3 A Bayesian Context - A Priori Information

Instead of treating the filter coefficients H^o as unknown constants that we are trying to estimate, we could also consider H^o as a stochastic vector about which we have some prior information, possibly from previous adaptive filtering experience. Assume now that, prior to obtaining the measurements $y_1, y_2, \cdots$, we know that H^o has a distribution with mean $E\, H^o = H_0$ and covariance $E\, (H^o - H_0)\, (H^o - H_0)^T = C_0$. So now the randomness comes from the $\widetilde{x}_k$ in (3.87) and H^o.

Since the problem formulation can now be recognized to be one of a Bayesian Linear Model, it is pretty straightforward to determine the Affine MMSE estimator. The AMMSE estimator can be shown to be the filter estimate resulting from (3.77),(3.80) with $R_0 = \sigma^2_{\widetilde{x}} C_0^{-1}$. With again $\widetilde{H}_k = H^o - H_k$, the correlation matrix of the estimate H_k, $C_k = E\, \widetilde{H}_k \widetilde{H}_k^T$ satisfies

$$C_k^{-1} = \sigma_{\widetilde{x}}^{-2} R_k = \sigma_{\widetilde{x}}^{-2} \sum_{i=1}^{k} Y_i\, Y_i^T + C_0^{-1} \quad . \tag{3.90}$$

Note that C_k^{-1} is an increasing function of C_0^{-1} and hence C_k is a decreasing function of C_0^{-1}. So we see that H_0 and R_0 in the LS cost function (3.77) have the interpretation of the prior mean and the inverse of the prior covariance of H^o. We'll choose R_0 small if we don't have a lot of confidence in our prior guess H_0 (C_0 big). In practice, R_0 is often chosen as $R_0 = \eta I$, a multiple of the identity matrix.

3.3.4 Exponential Weighting

In order to be able to track a possibly time-varying $H^o = H_k^o$, one introduces an exponential forgetting factor $\lambda \in (0, 1)$ into the cost function (3.77) to obtain

$$\xi_k\,(H) = \sum_{i=1}^{k} \lambda^{k-i} \left(x_i - H^T Y_i\right)^2 + \lambda^k\, (H - H_0)^T\, R_0\, (H - H_0) \ . \tag{3.91}$$

This implies that the past (and in particular the initial conditions H_0, R_0) is forgotten exponentially fast with a window with time constant $\frac{1}{1-\lambda}$. Note that (3.77) is a special case of (3.91) with $\lambda = 1$. It is possible to obtain results analogous to the ones above, but now with $\lambda < 1$.

3.4 Adaptive Filtering Applications

See Chapter 1 of the course Signal Modeling and Coding (applications of the Least-Squares technique).

3.5 Problems

3.1. *Steepest Descent on an AR(1) process.* Consider the steepest-descent algorithm for iterating towards the Wiener FIR filter for the case of two filter coefficients ($N = 2$) and the filter input y_k having the covariance sequence $r_{yy}(k) = \rho^{|k|}$. Determine both the maximum stepsize μ for convergence and the stepsize that gives the fastest convergence as a function of ρ.

3.2. *LMS as an Optimization Problem.* Show that the LMS algorithm can be obtained exactly from the following optimization problem

$$H_k = \arg\min_H \left\{ \left(x_k - Y_k^T H\right)^2 + (\frac{1}{\mu} - Y_k^T Y_k) \left(H - H_{k-1}\right)^T \left(H - H_{k-1}\right) \right\} . \qquad (3.92)$$

This shows the LMS update as the result of a compromise between taking into account the new data at time k and adhering to the previous estimate H_{k-1}. However, the weighting factor of the second term is not guaranteed to be positive and if it is not, the Hessian of the optimization problem in (3.92) will not be positive definite, meaning that the extremum at H_k will not be a minimum. This illustrates again the tricky stability issue of the LMS algorithm (unless μ is very small). Note that $\mu(\frac{1}{\mu} - Y_k^T Y_k) = 1 - \mu Y_k^T Y_k$ is the one eigenvalue different from one of the transition matrix $I - \mu Y_k Y_k^T$.

3.3. Consider the LMS algorithm in the system identification set-up with fixed optimal parameters. Using the independence assumption, but not the averaging theorem, the correlation matrix of the parameter estimation error vector $\widetilde{H}_k$ is found to satisfy the recursion

$$\begin{aligned} C_k &= E\left(\left[I - \mu Y_k Y_k^T\right] \widetilde{H}_{k-1} \widetilde{H}_{k-1}^T \left[I - \mu Y_k Y_k^T\right]\right) + \mu^2 \xi^o R_{YY} \\ &= E_{Y_k}\left(\left[I - \mu Y_k Y_k^T\right] C_{k-1} \left[I - \mu Y_k Y_k^T\right]\right) + \mu^2 \xi^o R_{YY} \end{aligned} . \qquad (3.93)$$

The remaining expectation operator E_{Y_k} is over the elements of the input vector $Y_k = [y_k \; y_{k-1} \cdots y_{k-N+1}]^T$. Assume now that the input signal y_k is Gaussian.

(a) Use the following property for Gaussian random variables

$$E\{y_1y_2y_3y_4\} = E\{y_1y_2\}E\{y_3y_4\} + E\{y_1y_3\}E\{y_2y_4\} + E\{y_1y_4\}E\{y_2y_3\}$$

to show that (3.93) becomes

$$\begin{aligned} C_k &= C_{k-1} - \mu C_{k-1}R_{YY} - \mu R_{YY}C_{k-1} + 2\mu^2 R_{YY}C_{k-1}R_{YY} \\ &\quad + \mu^2\left[\xi^o + \mathrm{tr}\{R_{YY}C_{k-1}\}\right]R_{YY} \ . \end{aligned} \tag{3.94}$$

(b) With the usual eigen decomposition $R_{YY} = V\Lambda V^T = \sum_{i=1}^{N}\lambda_i V_i V_i^T$, let $V_i^T C_k V_i = E\,v_i^2(k)$ as usual. Introduce the vectors $D_k = [E\,v_1^2(k)\cdots E\,v_N^2(k)]^T$, $\lambda = [\lambda_1\cdots\lambda_N]^T$. From (3.94), find a recursion for the vector D_k (hint: the system matrix will be diagonal, except for a rank one term).

(c) Assuming convergence, find the steady-state values D_∞ and ξ_∞^e.

(d) What is the maximal stepsize for convergence? (hint: to have convergence, ξ_∞^e should be finite). Observe that the averaging approach led to a very optimistic value for the maximal stepsize.

3.4. *Mean-square Convergence of the LMS Algorithm.*
Consider Venter's Theorem: If

$$\beta_k \ \le\ (1-\alpha_k)\beta_{k-1} + \delta_k$$

and

(i) $\sum_{k=0}^{\infty}\alpha_k = \infty$

(ii) $\alpha_k \to 0$ as $k \to \infty$

(iii) α_k, δ_k and $\beta_k \ge 0$

(iv) $\sum_{k=0}^{\infty}\delta_k < \infty$

then

$$\beta_k \to 0 \ \text{ as } \ k \to \infty \ .$$

All quantities above are scalars. By taking the stepsize μ_k in the LMS algorithm to be time-varying, and using Venter's Theorem, show that the following conditions must be satisfied

(i) $\sum_{k=0}^{\infty}\mu_k = \infty$

(ii) $\mu_k \to 0$ as $k \to \infty$

(iii) $\mu_k \ge 0$

(iv) $\sum_{k=0}^{\infty} \mu_k^2 < \infty$

in order for the LMS algorithm to converge in mean-square: $C_k \to 0$ as $k \to \infty$. Give an example of a stepsize sequence μ_k which satisfies these conditions.

3.5. *Sign-error LMS Algorithm.* Apply the derivation technique for the LMS algorithm to the following criterion: $\min_H E\left|x_k - H^T Y_k\right|$. Comment on how this leads to an adaptive algorithm that can be considered as a computational simplification of the LMS algorithm. Knowing that the MSE criterion is well adjusted to the case where the measurement noise $\widetilde{x}_k$ is Gaussian, to what kind of distribution of $\widetilde{x}_k$ would the criterion considered here be adapted?

3.6. *The Leaky LMS Algorithm.* Apply the derivation technique of the LMS algorithm to the following criterion:

$$(x_k - H^T Y_k)^2 + \rho \|H\|^2 \ .$$

(i) In the usual system identification set-up with a fixed optimal filter H^o, does the mean of the resulting parameter estimation error $\widetilde{H}_k$ converge to zero?

(ii) Using the averaging theorem, analyze the convergence behavior of the correlation matrix C_k and of the learning curve ξ_k. In particular, investigate the condition on ρ and μ for convergence and find the steady-state values C_∞ and ξ_∞.

(iii) Explain why this algorithm may have beneficial properties when the input covariance matrix R_{YY} is singular. For this, take into account that for the actual implementation on a processor with finite precision, the filter coefficients of the LMS algorithm satisfy the following recursion

$$H_k \ = \ H_{k-1} + \mu Y_k \epsilon_k^p + W_k \tag{3.95}$$

where $W_k \in \mathcal{R}^N$ is the vector of instantaneous round-off errors made during the update at time k, that is to say, the errors involved in computing H_k from H_{k-1}. You may assume $E\, W_k = 0$ and $E\, W_k W_k^T = Q$ (constant since H_k becomes stationary in steady-state). Investigate the effect of the round-off errors W_k on ξ_∞ (using the averaging theorem) when R_{YY} is singular (e.g. $\lambda_N = 0$) or nearly singular (λ_N very small). Show how the introduction of the leakage factor ρ reduces the influence of the round-off errors in such cases.

3.7. *The Instrumental Variable (IV) LMS Algorithm.* The LMS algorithm is a stochastic gradient approach to minimizing the MSE. Due to the orthogonality property of least-squares, the LMS algorithm can alternatively be viewed as an algorithm that tries to enforce the orthogonality between the error signal and the input data vector:

$$\begin{aligned} \epsilon_k^p &= x_k - Y_k^T H_{k-1} \\ H_k &= H_{k-1} + \mu Y_k \epsilon_k^p \ . \end{aligned} \tag{3.96}$$

Indeed, the posterior error signal is $\epsilon_k = x_k - Y_k^T H_k = (1-\mu Y_k^T Y_k)\,\epsilon_k^p$. Hence, after the update, we get

$$\|Y_k \epsilon_k\|^2 = (\epsilon_k)^2 \|Y_k\|^2 = (1-\mu \|Y_k\|^2)\,(\epsilon_k^p)^2 \|Y_k\|^2 = (1-\mu \|Y_k\|^2)\,\|Y_k \epsilon_k^p\|^2 < \|Y_k \epsilon_k^p\|^2 \tag{3.97}$$

assuming μ sufficiently small.

An *instrumental variable* is any signal z_k such that with $Z_k = [z_k \cdots z_{k-N+1}]^T$, R_{ZY} is positive definite and $R_{Z\widetilde{x}} = 0$. The IV LMS algorithm tries to enforce the orthogonality between the error signal and Z_k:

$$\begin{aligned} \epsilon_k^p &= x_k - Y_k^T H_{k-1} \\ H_k &= H_{k-1} + \mu Z_k \epsilon_k^p \,. \end{aligned} \tag{3.98}$$

(i) Using the independence assumption (H_{k-1} independent of Y_k or Z_k), investigate the convergence of $E\,H_k$. With the system identification set-up ($x_k = \widetilde{x}_k + Y_k^T H^o$), show that $E\,H_\infty = H^o$ if convergence occurs.

(ii) Consider the particular choice in which z_k are the innovations of the signal y_k (whitened version). This implies that R_{ZY} is a lower triangular Toeplitz matrix with diagonal elements equal to σ_y^2. Using the averaging theorem, show that the convergence dynamics of this IV LMS algorithm do not depend on the eigenvalue distribution of R_{YY}.

3.8. *Sign-data LMS Algorithm.* As an alternative way to eliminate multiplication operations in the LMS algorithm (alternative to sign-error LMS), consider the sign-data LMS algorithm

$$\begin{aligned} \epsilon_k^p &= x_k - Y_k^T H_{k-1} \\ H_k &= H_{k-1} + \mu \,\mathrm{sign}(Y_k)\epsilon_k^p \,. \end{aligned} \tag{3.99}$$

This is a particular case of the IV LMS algorithm with $z_k = \mathrm{sign}(y_k)$. Analyze the convergence of the learning curve using the averaging theorem for the case when y_k is a Gaussian signal.

3.9. *Computational Complexity of NLMS.* The computational complexity of the LMS algorithm is about $2N$ multiplications/additions. In a straightforward implementation, the computational complexity of the NLMS algorithm would be $3N$. Show how to reorganize the computations in the NLMS algorithm so that the number of multiplications required to do one update is $2N$ plus a constant.

3.10. *NLMS with optimal time-varying stepsize.* Consider the NLMS algorithm in which the step-size $\overline{\mu}_k$ is made time-varying:

$$\begin{cases} \epsilon_k^p &= x_k - Y_k^T H_{k-1} \\ H_k &= H_{k-1} + \dfrac{\overline{\mu}_k}{Y_k^T Y_k} Y_k \,\epsilon_k^p \,. \end{cases} \tag{3.100}$$

Let the input signal y_k be white noise ($R_{YY} = \sigma_y^2\, I_N$). Using the independence assumption and the approximation $\frac{1}{Y_k^T Y_k} = \frac{1}{\mathrm{tr}\, R_{YY}}$, show that the excess MSE, ξ_k^e (defined by

$\xi_k = \xi^o + \xi_k^e$), satisfies the recursion

$$\xi_{k+1}^e \;=\; \left(1 + \frac{-2\overline{\mu}_k + \overline{\mu}_k^2}{N}\right) \xi_k^e + \frac{\overline{\mu}_k^2}{N}\xi^o \quad . \tag{3.101}$$

Find a recursion for the optimal stepsize sequence $\overline{\mu}_k^o$ which minimizes ξ_{k+1}^e at all time instants. Show that $\overline{\mu}_k^o$ behaves asymptotically (for large k) as $\frac{c}{k}$ and find the proportionality factor c. Is this asymptotic behavior expected? Why?

3.11. *RLS with Exponential Weighting.* Derive the exponentially weighted Recursive Least-Squares algorithm, which minimizes the following criterion recursively:

$$\xi_k\,(H) = \sum_{i=1}^{k} \lambda^{k-i}\,\left(x_i - H^T Y_i\right)^2 + \lambda^k\,\left(H - H_0\right)^T R_0\,(H - H_0)\;.$$

3.12. *RLS with Exponential Weighting Continued.* Derive a recursion for the minimum value of the cost function $\xi_k(H_k)$ of the RLS algorithm with exponential window.

3.13. *Stochastic Newton algorithm.* Consider the filter design approach in which the filter coefficients H_k are obtained by minimizing the criterion

$$\xi_k^{SN}(H) \;=\; \left(x_k - Y_k^T H\right)^2 + (H - H_{k-1})^T\,A_{k-1}\,(H - H_{k-1}) \tag{3.102}$$

where $A_{k-1} = A_{k-1}^T > 0$ is a symmetric positive definite matrix.

(i) Consider on the other hand the Bayes estimation problem in which we are estimating the random parameter vector H. We have prior information in the sense that we know the prior distribution of H to be $H \sim \mathcal{N}(H_{k-1}, A_{k-1}^{-1})$. We furthermore make the measurement $x_k = Y_k^T H + v_k$ where Y_k is deterministic (known) and the measurement noise v_k is Gaussian, $v_k \sim \mathcal{N}(0,1)$, and independent of H. Show that the MMSE Bayes estimate of H based on the measurement x_k and the prior information is found by minimizing the criterion in (3.102).

(ii) Find $H_k = \arg\min_H \xi_k^{SN}(H)$ in terms of H_{k-1}.

(iii) Put $A_{k-1} \;=\; \left(\frac{1}{\mu} - Y_k^T A^{-1} Y_k\right) A$ in (3.102) where $A = A^T > 0$. Find again H_k as in (ii). Using the averaging theorem, analyze the convergence of the learning curve associated with the adaptive algorithm thus obtained. Apply an appropriate change of coordinates to $\widetilde{H}$ that will render the system decoupled.

(iv) Show how the NLMS algorithm and the RLS algorithm with exponential weighting fall out as special cases of the approach in (3.102).

(v) In the Bayes set-up above, what is the posterior distribution of H_k? In the RLS algorithm, this posterior distribution becomes the prior distribution when we make the next measurement. Hence the RLS algorithm determines at every time instant k the mean and the covariance matrix of the posterior distribution after taking into account the measurement at time k.

(vi) Consider the RLS algorithm with exponential weighting. Due to the exponential weighting, the effect of the initial conditions will have died out after a few time constants $\frac{1}{1-\lambda}$. After this initial transient, we can approximate the sample covariance matrix with exponential weighting as $R_k \approx E\, R_k = \frac{1}{1-\lambda} R_{YY}$. And for large enough N (but not too large so that $(1-\lambda)N \ll 1$), $Y_k^T R_{YY}^{-1} Y_k \approx E\, Y_k^T R_{YY}^{-1} Y_k = N$. Using your results from (iii), show how the convergence of the RLS algorithm, after the initial transient, does not depend on the eigenvalue spread of R_{YY}. In fact, the algorithm in (ii) with A being a multiple of R_{YY}^{-1} is called a stochastic Newton algorithm, whereas with A a multiple of I we have a stochastic gradient algorithm.

3.14. *RLS with Sliding Rectangular Window.* Derive a Recursive Least-Squares algorithm for the following criterion:

$$\xi_k(H) = \sum_{i=k-L+1}^{k} \left(x_i - H^T Y_i\right)^2 . \tag{3.103}$$

This is algorithm offers an alternative approach for tracking time-varying optimal parameters $H^o = H_k^o$. Only the data in a finite rectangular window of length L into the past are kept.

Bibliography

[1] H.L. Van Trees. *Detection, Estimation and Modulation Theory*, volume 1. Wiley, New York, 1968.

[2] L. Scharf. *Statistical Signal Processing.* Addison-Wesley, Reading, MA, 1991.

[3] S.M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory.* Prenctice Hall, 1993.

[4] B. Porat. *Digital Processing of Random Signals: Theory and Methods.* Prentice Hall, 1994.

[5] C.W. Therrien. *Discrete Random Signals and Statistical Signal Processing.* Prentice Hall, 1992.

[6] T. Kailath. *Lectures on Wiener and Kalman Filtering.* Springer-Verlag, Wien – New York, 1981.

[7] Lennart Ljung. *System Identification: Theory for the User.* Prentice-Hall, Englewood Cliffs, NJ, 1987.

[8] T. Söderström and P.G. Stoica. *System Identification.* Prentice Hall, 1989.

[9] J.M. Mendel. *Lessons in Estimation Theory for Signal Processing, Communications and Control.* Prentice Hall, 1995.