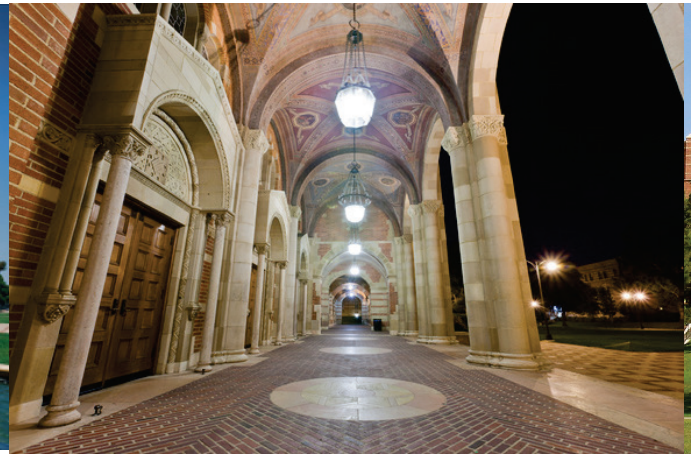




EE210A: Adaptation and Learning

Professor Ali H. Sayed

MAIN RESULT FROM LAST LECTURE

Theorem 2.1 (Optimal estimation in the vector case) The least-mean-squares estimator of a (possibly complex-valued) vector x given another (possibly complex-valued) vector y is the conditional expectation of x given y, i.e., $\hat{x} = \mathsf{E}(x|y)$. This estimator solves

$$\min_{\hat{x}} \mathsf{Tr}(R_{\tilde{x}})$$

where $R_{\tilde{x}} = \mathsf{E}\,\tilde{x}\tilde{x}^*$ and $\tilde{x} = x - \hat{x}$.

Lemma 2.2 (Cost function) The conditional expectation of x given y is optimal relative to either cost

$$\min_{\hat{x}} \mathsf{Tr}(R_{\tilde{x}}) \quad \text{or} \quad \min_{\hat{x}} R_{\tilde{x}}$$

where $R_{\tilde{x}} = \mathsf{E}\,\tilde{x}\tilde{x}^*$ and $\tilde{x} = x - \hat{x}$.

MAIN RESULT FROM LAST LECTURE

Lemma 2.1 (Circular Gaussian variables) If x and y are two circular and jointly Gaussian random variables with means $\{\bar{x}, \bar{y}\}$ and covariance matrices $\{R_x, R_y, R_{xy}\}$, then the least-mean-squares estimator of x given y is

$$\hat{\boldsymbol{x}} = \bar{x} + R_{xy} R_y^{-1} (\boldsymbol{y} - \bar{y})$$

and the resulting minimum cost is m.m.s.e. $= R_x - R_{xy} R_y^{-1} R_{yx}$.

$$\begin{aligned} R_x &= \mathsf{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{x} - \bar{x})^* \\ R_y &= \mathsf{E}\,(\boldsymbol{y} - \bar{y})(\boldsymbol{y} - \bar{y})^* \\ R_{xy} &= \mathsf{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{y} - \bar{y})^* = R_{yx}^* \end{aligned}$$

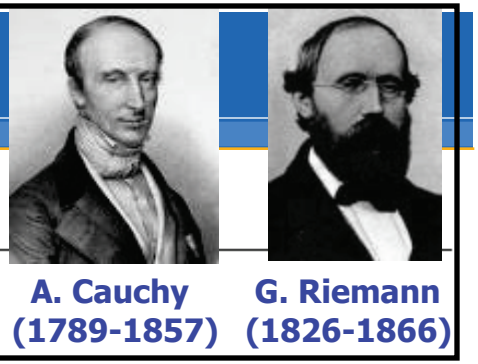
LECTURE #03

LINEAR MEAN-SQUARE ERROR ESTIMATION

Sections in order: App. C, 3.1, 3.2, 3.3, 3.4, 4.2

COMPLEX GRADIENTS

C.1 CAUCHY-RIEMANN CONDITIONS



A. Cauchy (1789-1857) G. Riemann (1826-1866)

We start with a scalar argument $z = x + jy$, where $j = \sqrt{-1}$. In this case, we can regard $g(z)$ as a function of the two real scalar variables, x and y, say

$$g(z) = u(x, y) \; + \; jv(x, y) \tag{C.1}$$

with $u(\cdot, \cdot)$ denoting its real part and $v(\cdot, \cdot)$ denoting its imaginary part. Now, from complex function theory, the derivative of $g(z)$ at a point $z_o = x_o + jy_o$ is defined as

$$\frac{dg}{dz} \triangleq \lim_{\Delta z \to 0} \frac{g(x_o + \Delta x, y_o + \Delta y) \; - \; g(x_o, y_o)}{\Delta x + j\Delta y}$$

where $\Delta z = \Delta x + j\Delta y$. For $g(z)$ to be differentiable at z_o, in which case it is also said to be *analytic* at z_o, the above limit should exist regardless of the direction from which z approaches z_o. In particular, if we assume $\Delta y = 0$ and $\Delta x \to 0$, then the above definition gives

$$\frac{dg}{dz} = \frac{\partial u}{\partial x} + j\frac{\partial v}{\partial x} \tag{C.2}$$

CAUCHY RIEMANN CONDITIONS

If, on the other hand, we assume that $\Delta x = 0$ and $\Delta y \to 0$ so that $\Delta z = j\Delta y$, then the definition gives

$$\frac{dg}{dz} = \frac{\partial v}{\partial y} - j\frac{\partial u}{\partial y} \tag{C.3}$$

The expressions (C.2) and (C.3) should coincide. Therefore, by adding them we get

$$\frac{dg}{dz} = \frac{1}{2}\left(\frac{\partial u}{\partial x} + j\frac{\partial v}{\partial x} + \frac{\partial v}{\partial y} - j\frac{\partial u}{\partial y}\right)$$

or, more compactly,

$$\boxed{\frac{dg}{dz} \triangleq \frac{1}{2}\left\{\frac{\partial g}{\partial x} - j\,\frac{\partial g}{\partial y}\right\}} \tag{C.4}$$

Observe that the equality of expressions (C.2) and (C.3) implies that the real and imaginary parts of $g(\cdot)$ should satisfy the conditions

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

SCALAR EXAMPLES

C.2 SCALAR ARGUMENTS

More generally, if g is a function of both z and z^*, we define its partial derivatives with respect to z and z^* as follows:

$$\boxed{\frac{\partial g}{\partial z} = \frac{1}{2}\left\{\frac{\partial g}{\partial x} - j\,\frac{\partial g}{\partial y}\right\}, \qquad \frac{\partial g}{\partial z^*} = \frac{1}{2}\left\{\frac{\partial g}{\partial x} + j\,\frac{\partial g}{\partial y}\right\}} \tag{C.5}$$

Note in particular, from the Cauchy-Riemann conditions, that if $g(z)$ is an analytic function of z then it must necessarily hold that $\partial g/\partial z^* = 0$.

Examples

We illustrate the definitions (C.4)–(C.5) considering several examples.

1. Let $g(z) = z = x + jy$. Then

$$\frac{\partial g}{\partial z} = (1 - j^2)/2 = 1, \qquad \frac{\partial g}{\partial z^*} = (1 + j^2)/2 = 0$$

SCALAR EXAMPLES

2. Let $g(z) = z^2 = (x+jy)(x+jy) = (x^2 - y^2) + j2xy$. Then

$$\frac{\partial g}{\partial z} = 2(x+jy) = 2z, \qquad \frac{\partial g}{\partial z^*} = 0$$

In examples 1. and 2., since g is a function of z alone, it holds that $\partial g/\partial z = dg/dz$.

3. Let $g(z) = |z|^2 = zz^* = (x+jy)(x-jy) = x^2 + y^2$. Then

$$\frac{\partial g}{\partial z} = (x-jy) = z^*, \qquad \frac{\partial g}{\partial z^*} = (x+jy) = z$$

4. Let $g(z) = \lambda + \alpha z + \beta z^* + \gamma z z^*$, where $(\lambda, \alpha, \beta, \gamma)$ are complex constants. That is,

$$g(z) = \left[\lambda + \alpha x + \beta x + \gamma(x^2 + y^2)\right] + j\left[\alpha y - \beta y\right]$$

Then

$$\frac{\partial g}{\partial z} = \alpha + \gamma z^*, \qquad \frac{\partial g}{\partial z^*} = \beta + \gamma z$$

VECTOR EXAMPLES

C.3 VECTOR ARGUMENTS

Now assume that z is a *column* vector, say

$$z = \text{col}\{z_1, z_2, \ldots, z_n\}, \quad z_i = x_i + jy_i$$

The *complex gradient* of g with respect to z is denoted by $\nabla_z g(z)$, or simply $\nabla_z g$, and is defined as the *row* vector

$$\boxed{\nabla_z g \triangleq \begin{bmatrix} \partial g/\partial z_1 & \partial g/\partial z_2 & \ldots & \partial g/\partial z_n \end{bmatrix}} \quad \left\{ \begin{array}{l} z \text{ is a column} \\ \nabla_z g \text{ is a row} \end{array} \right.$$

Likewise, the complex gradient of g with respect to z^* is defined as the *column* vector

$$\boxed{\nabla_{z^*} g \triangleq \begin{bmatrix} \partial g/\partial z_1^* \\ \partial g/\partial z_2^* \\ \vdots \\ \partial g/\partial z_n^* \end{bmatrix}} \quad \left\{ \begin{array}{l} z^* \text{ is a row} \\ \nabla_z g \text{ is a column} \end{array} \right.$$

VECTOR EXAMPLES

The reason why we choose to define $\nabla_z g$ as a row vector and $\nabla_{z^*} g$ as a column vector is because the subsequent differentiation results will be consistent with what we are used to from the standard differentiation of functions of real-valued arguments. Let us again consider a few examples:

1. Let $g(z) = \alpha^* z$, where $\{\alpha, z\}$ are column vectors. Then $\nabla_z g = \alpha^*$ and $\nabla_{z^*} g = 0$.

2. Let $g(z) = z^* \beta$, where $\{\beta, z\}$ are column vectors. Then $\nabla_z g = 0$ and $\nabla_{z^*} g = \beta$.

3. Let $g(z) = z^* z$, where z is a column vector. Then $\nabla_z g = z^*$ and $\nabla_{z^*} g = z$.

4. Let $g(z) = \lambda + \alpha^* z + z^* \beta + z^* \Gamma z$, where λ is a scalar, $\{\alpha, \beta\}$ are column vectors and Γ is a matrix. Then $\nabla_z g = \alpha^* + z^* \Gamma$ and $\nabla_{z^*} g = \beta + \Gamma z$.

OPTIMAL MSE ESTIMATOR

In Secs. 1.2 and 2.1 we studied the problem of determining the optimal function $h(\cdot)$ that minimizes the mean-square error of estimating a random variable $\boldsymbol{x}$ from another random variable $\boldsymbol{y}$. Specifically, we solved

$$\min_{h(\cdot)} \ \mathsf{E}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^* \tag{3.1}$$

over all functions $h(\cdot)$ of $\boldsymbol{y}$. The optimal solution was found to be the conditional expectation of $\boldsymbol{x}$ given $\boldsymbol{y}$, i.e.,

$$\hat{\boldsymbol{x}} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$$

Such conditional expectations are generally hard to evaluate in closed-form, except in some special cases.

AFFINE ESTIMATOR

Due to the difficulty in evaluating $\mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$ in general, it is common practice to restrict the choice of $h(\cdot)$ to the subclass of *affine* functions of $\boldsymbol{y}$, i.e., to functions of the form

$$h(\boldsymbol{y}) \;=\; K\boldsymbol{y} + b \tag{3.2}$$

for some matrix K and for some vector b to be determined. For general vector-valued

Affine functions of the form (3.2) are easier to implement than most nonlinear functions. For example, when K is a row vector, the product $K\boldsymbol{y}$ amounts to an inner product. Moreover, by deliberately restricting $h(\cdot)$ to be an affine function of $\boldsymbol{y}$, the evaluation of the optimal $h(\cdot)$ is greatly simplified. In particular, it will be seen that all we need to know in order to determine the optimal parameters $\{K, b\}$ are the first and second-order moments of $\{\boldsymbol{x}, \boldsymbol{y}\}$, namely, $\mathsf{E}\,\boldsymbol{x}$, $\mathsf{E}\,\boldsymbol{y}$, $\mathsf{E}\,\boldsymbol{x}\boldsymbol{x}^*$, $\mathsf{E}\,\boldsymbol{y}\boldsymbol{y}^*$, and $\mathsf{E}\,\boldsymbol{x}\boldsymbol{y}^*$. No other moments are needed. In contrast, evaluation of the optimal estimator, $\hat{\boldsymbol{x}} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$, requires full knowledge of the conditional pdf $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$.

3.1 MEAN-SQUARE ERROR CRITERION

Thus consider two zero-mean vector-valued random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ and let

$$\bar{x} \triangleq \mathsf{E}\,\boldsymbol{x} = 0, \quad \bar{y} \triangleq \mathsf{E}\,\boldsymbol{y} = 0, \quad R_x \triangleq \mathsf{E}\,\boldsymbol{x}\boldsymbol{x}^*, \quad R_y \triangleq \mathsf{E}\,\boldsymbol{y}\boldsymbol{y}^*, \quad R_{xy} \triangleq \mathsf{E}\,\boldsymbol{x}\boldsymbol{y}^*$$

The dimensions of $\boldsymbol{x}$ and $\boldsymbol{y}$ need not be identical, say $\boldsymbol{x}$ is $p \times 1$ and $\boldsymbol{y}$ is $q \times 1$. In this case, $\{R_x, R_y, R_{xy}\}$ are $\{p \times p, q \times q, p \times q\}$ matrices, respectively.

We now seek an affine estimator for $\boldsymbol{x}$, namely, one of the form

$$\hat{\boldsymbol{x}} = K\boldsymbol{y} + b$$

for some constants $\{K, b\}$ to be determined, where K is $p \times q$ and b is $p \times 1$. The determination of $\{K, b\}$ is based on two considerations.

UNBIASED ESTIMATOR

The determination of $\{K, b\}$ is based on two considerations. First, the estimator should be unbiased, which means that it should satisfy

$$\boxed{\mathsf{E}\,\hat{\boldsymbol{x}} = 0}$$

But since

$$\mathsf{E}\,\hat{\boldsymbol{x}} = K\mathsf{E}\,\boldsymbol{y} + b \;=\; 0 + b = b$$

we find that we must have $b = 0$. This means that the estimator that we are seeking is effectively a linear estimator, i.e., it is one of the form $\hat{\boldsymbol{x}} = K\boldsymbol{y}$. Second, the coefficient matrix K should be chosen optimally so as to minimize the error covariance matrix (or its trace), as we now explain.

MINIMIZING MSE

Let $\{k_i^*\}$ denote the individual rows of K. Then the estimator for each entry of $\boldsymbol{x}$, say $\boldsymbol{x}(i)$, is given by the inner product $k_i^*\boldsymbol{y}$,

$$\underbrace{\begin{bmatrix} \hat{\boldsymbol{x}}(0) \\ \hat{\boldsymbol{x}}(1) \\ \vdots \\ \hat{\boldsymbol{x}}(p-1) \end{bmatrix}}_{=\hat{\boldsymbol{x}}} = \begin{bmatrix} k_0^*\boldsymbol{y} \\ k_1^*\boldsymbol{y} \\ \vdots \\ k_{p-1}^*\boldsymbol{y} \end{bmatrix} = \underbrace{\begin{bmatrix} \text{---------} \\ \text{---------} \\ \vdots \\ \text{---------} \end{bmatrix}}_{=K} \boldsymbol{y}$$

The optimal choices for the column vectors $\{k_i\}$ are determined by solving

$$\boxed{\min_{k_i} \quad \mathsf{E}\,|\tilde{\boldsymbol{x}}(i)|^2} \qquad \text{for each} \quad i = 0, 1, \ldots, p-1 \tag{3.3}$$

where

$$\tilde{\boldsymbol{x}}(i) = \boldsymbol{x}(i) - \hat{\boldsymbol{x}}(i)$$

MINIMIZING MSE

The optimization problems (3.3) can be grouped together and stated equivalently as the problem of determining the matrix K by solving

$$\boxed{\min_{K} \quad \mathsf{E}\,\tilde{\boldsymbol{x}}^{*}\tilde{\boldsymbol{x}}} \tag{3.4}$$

where $\tilde{\boldsymbol{x}} = \boldsymbol{x} - \hat{\boldsymbol{x}}$. This is because the scalar quantity $\mathsf{E}\,\tilde{\boldsymbol{x}}^{*}\tilde{\boldsymbol{x}}$ in (3.4) is simply the sum of the individual error variances that appear in (3.3),

$$\mathsf{E}\,\tilde{\boldsymbol{x}}^{*}\tilde{\boldsymbol{x}} \;=\; \mathsf{E}\,|\tilde{\boldsymbol{x}}(0)|^{2} \;+\; \mathsf{E}\,|\tilde{\boldsymbol{x}}(1)|^{2} \;+\; \ldots \;+\; \mathsf{E}\,|\tilde{\boldsymbol{x}}(p-1)|^{2} \tag{3.5}$$

and each term $\mathsf{E}\,|\tilde{\boldsymbol{x}}(i)|^{2}$ depends on the corresponding k_i alone. In this way, minimizing $\mathsf{E}\,\tilde{\boldsymbol{x}}^{*}\tilde{\boldsymbol{x}}$ over K is equivalent to minimizing each term $\mathsf{E}\,|\tilde{\boldsymbol{x}}(i)|^{2}$ over its k_i, so that problems (3.3) and (3.4) are equivalent.

MINIMIZING MSE

Therefore, continuing with (3.3), we expand the cost function to obtain a *quadratic* expression in the unknown column vector k_i,

$$\begin{aligned}\mathsf{E}\,|\tilde{\boldsymbol{x}}(i)|^2 &\triangleq \mathsf{E}\,|\boldsymbol{x}(i)-k_i^*\boldsymbol{y}|^2\\ &= \mathsf{E}\,[\boldsymbol{x}(i)-k_i^*\boldsymbol{y}][\boldsymbol{x}(i)-k_i^*\boldsymbol{y}]^*\\ &= \mathsf{E}\,|\boldsymbol{x}(i)|^2-[\mathsf{E}\,\boldsymbol{x}(i)\boldsymbol{y}^*]k_i-k_i^*[\mathsf{E}\,\boldsymbol{y}\boldsymbol{x}^*(i)]+k_i^*R_yk_i\end{aligned}$$

This is a *scalar-valued* cost function of a possibly *complex-valued vector* quantity, k_i. We denote it by

$$J(k_i) \triangleq \mathsf{E}\,|\boldsymbol{x}(i)|^2-[\mathsf{E}\,\boldsymbol{x}(i)\boldsymbol{y}^*]k_i-k_i^*[\mathsf{E}\,\boldsymbol{y}\boldsymbol{x}^*(i)]+k_i^*R_yk_i \tag{3.6}$$

The quantity $\mathsf{E}\,|\boldsymbol{x}(i)|^2$ that appears in the expression for $J(k_i)$ is the variance of $\boldsymbol{x}(i)$ and is therefore equal to the i-th diagonal entry of R_x. We denote it by

$$\sigma_{x,i}^2=\mathsf{E}\,|\boldsymbol{x}(i)|^2$$

COST FUNCTION

Likewise, the quantity $\mathsf{E}\,\boldsymbol{x}(i)\boldsymbol{y}^*$ is the i-th row of the cross-covariance matrix R_{xy}. We denote it by $R_{xy,i} = \mathsf{E}\,\boldsymbol{x}(i)\boldsymbol{y}^*$. In this way, we can rewrite $J(k_i)$ as

$$\boxed{J(k_i) \;\triangleq\; \sigma_{x,i}^2 - R_{xy,i}k_i - k_i^* R_{yx,i} + k_i^* R_y k_i} \tag{3.7}$$

where $\{\sigma_{x,i}^2, R_{xy,i}, R_y\}$ are known quantities and k_i is the unknown column vector; $\sigma_{x,i}^2$ is a scalar, $R_{xy,i}$ is a row vector, and R_y is a nonnegative-definite matrix. Moreover, $R_{yx,i} = R_{xy,i}^*$. Our objective is to minimize $J(k_i)$ over k_i.

MSE MINIMIZATION

3.2 MINIMIZATION BY DIFFERENTIATION

$$J(k_i) \triangleq \sigma_{x,i}^2 - R_{xy,i}k_i - k_i^* R_{yx,i} + k_i^* R_y k_i \tag{3.7}$$

the complex gradient vector of $J(k_i)$ with respect to k_i is given by

$$\nabla_{k_i} J(k_i) = -R_{xy,i} + k_i^* R_y$$

Observe that this result is consistent with what we would expect from the standard rules of differentiation for functions of real variables, with the vectors k_i and k_i^* treated as different quantities. Also, if all data were real-valued, in which case

$$J(k_i) = \mathsf{E}\,\boldsymbol{x}^2(i) - [\mathsf{E}\,\boldsymbol{x}(i)\boldsymbol{y}^{\mathsf{T}}]k_i - k_i^{\mathsf{T}}[\mathsf{E}\,\boldsymbol{y}\boldsymbol{x}(i)] + k_i^{\mathsf{T}} R_y k_i$$

then we would have obtained instead $\nabla_{k_i} J(k_i) = -R_{xy,i} + 2k_i^{\mathsf{T}} R_y$, with an additional factor of 2 — see App. C.

DIFFERENTIATION ARGUMENT

By setting the complex gradient equal to zero at the optimal choice $k_i = k_i^o$, we find that k_i^o should satisfy the linear equations

$$k_i^{o*} R_y = R_{xy,i}, \qquad i = 0, 1, \ldots, p-1 \tag{3.8}$$

If we collect the row vectors $\{k_i^{o*}\}$ from (3.8) into a matrix K_o we find that this desired solution matrix should satisfy

$$\boxed{K_o R_y = R_{xy}} \tag{3.9}$$

These equations are called the normal equations, for reasons explained later in Remark 4.1 in the next chapter.

COMPLETION OF SQUARES ARGUMENT

3.3 MINIMIZATION BY COMPLETION OF SQUARES

$$J(k_i) = \begin{bmatrix} 1 & k_i^* \end{bmatrix} \begin{bmatrix} \sigma_{x,i}^2 & -R_{xy,i} \\ -R_{yx,i} & R_y \end{bmatrix} \begin{bmatrix} 1 \\ k_i \end{bmatrix} \tag{3.10}$$

with a Hermitian center matrix and with the unknown vector k_i, and its conjugate transpose, multiplying from both sides.

Now given any Hermitian matrix of the form

$$M = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix}$$

with $A = A^*$, $C = C^*$, and C invertible, it can be verified by direct calculation that M can be factored into a product of a block upper-triangular, block-diagonal, and block lower-triangular matrices as:

COMPLETION OF SQUARES

$$\begin{bmatrix} A & B \\ B^* & C \end{bmatrix} = \begin{bmatrix} \mathrm{I} & BC^{-1} \\ 0 & \mathrm{I} \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & C \end{bmatrix} \begin{bmatrix} \mathrm{I} & 0 \\ C^{-1}B^* & \mathrm{I} \end{bmatrix} \tag{3.11}$$

where

$$\Sigma \triangleq A - BC^{-1}B^*$$

there is a generalization of (3.11) for block matrices M with possibly singular matrices C. Indeed, it is easy to verify, also by direct calculation, that we can alternatively factor any such M as

$$\begin{bmatrix} A & B \\ B^* & C \end{bmatrix} = \begin{bmatrix} \mathrm{I} & D \\ 0 & \mathrm{I} \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & C \end{bmatrix} \begin{bmatrix} \mathrm{I} & 0 \\ D^* & \mathrm{I} \end{bmatrix} \tag{3.12}$$

where D is defined as *any* solution to the linear system of equations

$$DC = B \tag{3.13}$$

and

$$\Sigma = A - BD^*$$

COMPLETION OF SQUARES

Applying (3.12) to the center matrix in (3.10) we can write

$$\begin{bmatrix} \sigma_{x,i}^2 & -R_{xy,i} \\ -R_{yx,i} & R_y \end{bmatrix} = \begin{bmatrix} 1 & -k_i^{o*} \\ 0 & \mathrm{I} \end{bmatrix} \begin{bmatrix} \sigma_{x,i}^2 - R_{xy,i}k_i^o & 0 \\ 0 & R_y \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -k_i^o & \mathrm{I} \end{bmatrix} \tag{3.14}$$

where k_i^o is any solution to the linear system of equations

$$k_i^{o*} R_y = R_{xy,i} \tag{3.15}$$

Substituting the factorization (3.14) into expression (3.10) for $J(k_i)$ and expanding the right-hand side, we find that $J(k_i)$ can be expressed in the equivalent form

$$\boxed{J(k_i) = (\sigma_{x,i}^2 - R_{xy,i}k_i^o) \; + \; (k_i - k_i^o)^* R_y (k_i - k_i^o)} \tag{3.16}$$

$$\boxed{J(k_i^o) = \sigma_{x,i}^2 - R_{xy,i}k_i^o \; = \; \sigma_{x,i}^2 - k_i^{o*} R_y k_i^o \; = \; \sigma_{x,i}^2 - k_i^{o*} R_{yx,i}} \tag{3.17}$$

$$\boxed{\mathsf{E}\,\tilde{x}^*\tilde{x} \; = \; \mathsf{Tr}\,(R_x - K_o R_y K_o^*)} \tag{3.18}$$

3.4 MINIMIZATION OF THE ERROR COVARIANCE MATRIX

The same completion-of-squares argument can be applied directly to the solution of (3.4) rather than the solution of the individual problems (3.3). To see this, let $J(K)$ denote the error-covariance matrix,

$$\boxed{J(K) \triangleq \mathsf{E}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^* = \mathsf{E}\,(\boldsymbol{x} - K\boldsymbol{y})(\boldsymbol{x} - K\boldsymbol{y})^*}$$

so that, as in (3.10),

$$J(K) = \begin{bmatrix} \mathrm{I} & K \end{bmatrix} \begin{bmatrix} R_x & -R_{xy} \\ -R_{yx} & R_y \end{bmatrix} \begin{bmatrix} \mathrm{I} \\ K^* \end{bmatrix}$$

ERROR COVARIANCE MATRIX

Following the same arguments as in the previous section, we can factor the center matrix as

$$\begin{bmatrix} R_x & -R_{xy} \\ -R_{yx} & R_y \end{bmatrix} = \begin{bmatrix} \mathrm{I} & -K_o \\ 0 & \mathrm{I} \end{bmatrix} \begin{bmatrix} R_x - R_{xy}K_o^* & 0 \\ 0 & R_y \end{bmatrix} \begin{bmatrix} \mathrm{I} & 0 \\ -K_o^* & \mathrm{I} \end{bmatrix}$$

where K_o is any solution to $K_o R_y = R_{xy}$. It then follows that

$$J(K) = (R_x - R_{xy}K_o^*) + (K - K_o)R_y(K - K_o)^* \tag{3.19}$$

where the last term is nonnegative definite for any K since $R_y \geq 0$. Now the criterion in (3.4) is related to $J(K)$ via $\mathsf{E}\,\tilde{\boldsymbol{x}}^*\tilde{\boldsymbol{x}} = \mathsf{Tr}[J(K)]$ so that, from (3.19),

$$\mathsf{E}\,\tilde{\boldsymbol{x}}^*\tilde{\boldsymbol{x}} \geq \mathsf{Tr}(R_x - R_{xy}K_o^*) \qquad \text{for any } K \tag{3.20}$$

This is because the trace of the nonnegative-definite matrix $(K - K_o)R_y(K - K_o)^*$ is nonnegative.

ERROR COVARIANCE MATRIX

This derivation reveals another aspect of the solution K_o. It not only minimizes the trace of the error covariance matrix, as in (3.4), but it also minimizes the error-covariance matrix itself since we also get $J(K) \geq J(K_o)$ for any K. We can therefore interpret K_o as also the solution to the problem

$$\min_{K} \quad \mathsf{E}\, \tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^* \tag{3.21}$$

which is in terms of the error-covariance matrix, rather than its trace. The resulting minimum value is

$$J(K_o) = R_x - R_{xy}K_o^* \tag{3.22}$$

The optimization problem (3.21) is interesting for two reasons. First, the cost function $J(K) = \mathsf{E}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^*$ is matrix-valued. That is, it assumes *matrix values* for each choice of K. Second, the unknown argument, K, is a *matrix* itself. In this way, problem (3.21) involves minimizing a matrix-valued cost function over a matrix-valued argument.

OPTIMAL LINEAR ESTIMATOR

Theorem 3.1 (Optimal linear estimator) Given zero-mean random variables x and y, the linear least-mean-squares estimator (l.l.m.s.e.) of x given y is

$$\hat{x} = K_o y$$

where K_o is any solution to the linear system of equations $K_o R_y = R_{xy}$. This estimator minimizes the following two error measures:

$$\min_K \mathsf{E}\,\tilde{x}^*\tilde{x} \qquad \text{and} \qquad \min_K \mathsf{E}\,\tilde{x}\tilde{x}^*$$

The scalar cost on the left is the trace of the matrix cost on the right. The resulting minimum mean-square errors, as defined by (3.4) and (3.22), are given by

$$\begin{aligned} \min_K \mathsf{E}\,\tilde{x}^*\tilde{x} &= \mathsf{Tr}\,(R_x - K_o R_y K_o^*) \\ \min_K \mathsf{E}\,\tilde{x}\tilde{x}^* &= R_x - K_o R_y K_o^* \end{aligned}$$

4.2 ORTHOGONALITY CONDITION

The linear least-mean-squares estimator admits an important geometric interpretation in the form of an orthogonality condition. This can be seen by rewriting the normal equations (3.9) as $K_o \mathsf{E}\, \boldsymbol{y}\boldsymbol{y}^* = \mathsf{E}\, \boldsymbol{x}\boldsymbol{y}^*$ or, equivalently,

$$\mathsf{E}\,(\boldsymbol{x} - K_o\boldsymbol{y})\boldsymbol{y}^* \;=\; 0 \tag{4.3}$$

The difference $\boldsymbol{x} - K_o\boldsymbol{y}$ is the estimation error, $\tilde{\boldsymbol{x}}$. Therefore, equality (4.3) states that the error is orthogonal to (or uncorrelated with) the observation vector $\boldsymbol{y}$, namely,

$$\boxed{\mathsf{E}\,\tilde{\boldsymbol{x}}\boldsymbol{y}^* = 0}$$

which we also write as (see Fig. 4.4):

$$\boxed{\tilde{\boldsymbol{x}} \;\perp\; \boldsymbol{y}} \tag{4.4}$$

ORTHOGONALITY CONDITION

We thus conclude that for linear least-mean-squares estimation, the estimation error is orthogonal to the data and, in fact, to any *linear* transformation of the data, say $A\boldsymbol{y}$ for any matrix A. This fact means that no further linear transformation of $\boldsymbol{y}$ can extract additional information about $\boldsymbol{x}$ in order to further reduce the error covariance matrix. Moreover, since the estimator $\hat{\boldsymbol{x}}$ is itself a linear function of $\boldsymbol{y}$, we obtain, as a special case, that

$$\boxed{\tilde{\boldsymbol{x}} \perp \hat{\boldsymbol{x}}} \tag{4.5}$$

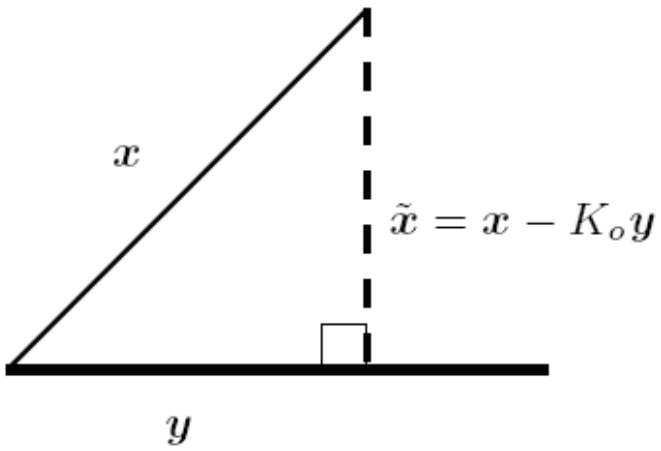


FIGURE 4.4 The orthogonality condition for linear estimation: $\tilde{\boldsymbol{x}} \perp \boldsymbol{y}$.

DEFINING PROPERTY

Theorem 4.1 (Orthogonality principle) Given two zero-mean random variables x and y, a linear estimator $\hat{x} = K_o y$ is optimal in the least-mean-squares sense (3.3) if, and only if, it satisfies $x - \hat{x} \perp y$, i.e., $\mathsf{E}\,(x - \hat{x})y^* = 0$.

Proof: One direction was argued prior to the statement of the theorem. Specifically, if $\hat{x}$ is the optimal linear estimator, then we know from (4.4) that $\tilde{x} \perp y$. Conversely, assume $\hat{x}$ is a linear estimator for x that satisfies $x - \hat{x} \perp y$, and let $\hat{x} = Ky$ for some K. It then follows from $x - \hat{x} \perp y$ that $\mathsf{E}\,(x - Ky)y^* = 0$, so that K satisfies the normal equations $KR_y = R_{xy}$ and, hence, from Thm. 3.1 we conclude that $\hat{x}$ should be the optimal linear estimator.

$\diamond$

EXAMPLE

Example 4.5 (Signal with exponential auto-correlation)

Consider a scalar zero-mean stationary random process $\{z(t)\}$ with auto-correlation function:

$$R_z(\tau) \triangleq \mathsf{E}\, z(t)z^*(t-\tau) = e^{-\alpha|\tau|} \tag{4.6}$$

That is, the samples of $z(t)$ become less correlated as the time gap between them increases. A so-called *random telegraph* signal has this property — see Prob. II.23. It is claimed that the linear least-mean squares estimator of $z(T_3)$ given $z(T_1)$ and $z(T_2)$ (assuming $T_1 < T_2 < T_3$) is

$$\hat{z}(T_3) = e^{-\alpha(T_3-T_2)}z(T_2)$$

That is, the estimator of a future value depends only on the most recent observation, $z(T_2)$. We can verify the validity of this claim by checking whether the above estimator satisfies the orthogonality condition.

EXAMPLE

So define the observation vector $\boldsymbol{y} = \text{col}\{\boldsymbol{z}(T_1), \boldsymbol{z}(T_2)\}$ and let $\boldsymbol{x} = \boldsymbol{z}(T_3)$. We can now evaluate the cross-correlation vector

$$\mathsf{E}\,(\boldsymbol{x} - \hat{\boldsymbol{x}})\boldsymbol{y}^* = \mathsf{E}\,[\boldsymbol{z}(T_3) - e^{-\alpha(T_3-T_2)}\boldsymbol{z}(T_2)]\boldsymbol{y}^*$$

If the answer is zero then the orthogonality condition is satisfied and the estimator is optimal in the linear least-mean-squares sense, as claimed. Otherwise, the estimator is not optimal. Using the given auto-correlation function (4.6), it is easy to verify that

$$\begin{aligned}
\mathsf{E}\left[\boldsymbol{z}(T_3) - e^{-\alpha(T_3-T_2)}\boldsymbol{z}(T_2)\right]\boldsymbol{y}^* &= \left[\begin{array}{cc} e^{-\alpha(T_3-T_1)} - e^{-\alpha(T_3-T_1)} & e^{-\alpha(T_3-T_2)} - e^{-\alpha(T_3-T_2)} \end{array}\right] \\
&= \left[\begin{array}{cc} 0 & 0 \end{array}\right]
\end{aligned}$$

so that the estimator is optimal.

◇

1st DESIGN EXAMPLE

Example 4.1 (Noisy measurements of a binary signal)

We reconsider Ex. 1.2 of a BPSK signal x that assumes the values ± 1 with probability $1/2$. The measurement y is $y = x + v$, where x and the disturbance v are independent of each other, with v being zero-mean Gaussian of unit variance.

Both x and y have zero means so that, according to Thm. 3.1, the optimal linear estimator of x is $\hat{x} = k_o y$, where the (now scalar) coefficient k_o is obtained from solving $k_o \sigma_y^2 = \sigma_{xy}$. We therefore need to determine the quantities $\{\sigma_y^2, \sigma_{xy}\}$. Now since $\{x, v\}$ are independent we have

$$\sigma_y^2 = \sigma_x^2 + \sigma_v^2 = 1 + 1 = 2$$

Moreover,

$$\sigma_{xy} = \mathsf{E}\, xy = \mathsf{E}\, x(x + v) = \mathsf{E}\, x^2 + 0 = \mathsf{E}\, x^2 = 1$$

so that $k_o = 1/2$, and the optimal linear estimator is $\hat{x} = y/2$. That is, we simply scale the received signal by $1/2$. In contrast, the optimal estimator was found in Ex. 1.2 to be given by the nonlinear transformation $\tanh(y)$. In addition, observe that the form of the linear estimator, $\hat{x} = y/2$, is valid regardless of whether the noise v is Gaussian or not (i.e., the Gaussian assumption on v is not needed to arrive at $k_o = 1/2$). The form of the optimal estimator, $\hat{x} = \tanh(y)$, on the other hand, is very much tied to the Gaussian assumption on v.

1st DESIGN EXAMPLE

Let us now reconsider Ex. 2.1, where we collect two noisy measurements $\boldsymbol{y}(0)$ and $\boldsymbol{y}(1)$ of $\boldsymbol{x}$, say

$$\boldsymbol{y}(0) = \boldsymbol{x} + \boldsymbol{v}(0) \quad \text{and} \quad \boldsymbol{y}(1) = \boldsymbol{x} + \boldsymbol{v}(1)$$

where $\{\boldsymbol{v}(0), \boldsymbol{v}(1)\}$ are zero-mean unit-variance Gaussian random variables that are independent of each other and of $\boldsymbol{x}$. The value of $\boldsymbol{x}$ is the same in both measurements (i.e., if it is $+1$ in the measurement $\boldsymbol{y}(0)$, it is also $+1$ in the measurement $\boldsymbol{y}(1)$, and similarly for -1) — recall Fig. 2.1. Introduce the column vector $\boldsymbol{y} = \text{col}\{\boldsymbol{y}(0), \boldsymbol{y}(1)\}$. Then, according to Thm. 3.1, the optimal linear estimator of $\boldsymbol{x}$ given $\boldsymbol{y}$ is $\hat{\boldsymbol{x}} = k_o^* \boldsymbol{y}$, where k_o^* is now 1×2 and is obtained from the solution of the normal equations $k_o^* R_y = R_{xy}$. To determine $\{R_y, R_{xy}\}$ we proceed as follows. Since $\{\boldsymbol{x}, \boldsymbol{v}(0), \boldsymbol{v}(1)\}$ are independent we get

$$R_y = \mathsf{E}\, \boldsymbol{y}\boldsymbol{y}^* = \begin{bmatrix} \mathsf{E}\,|\boldsymbol{y}(0)|^2 & \mathsf{E}\,\boldsymbol{y}(0)\boldsymbol{y}^*(1) \\ \mathsf{E}\,\boldsymbol{y}(1)\boldsymbol{y}^*(0) & \mathsf{E}\,|\boldsymbol{y}(1)|^2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

1ST DESIGN EXAMPLE

where we used the fact that

$$\mathsf{E}\, \boldsymbol{y}(0)\boldsymbol{y}^*(1) = \mathsf{E}\,(\boldsymbol{x}+\boldsymbol{v}(0))(\boldsymbol{x}+\boldsymbol{v}(1))^* = \mathsf{E}\,|\boldsymbol{x}|^2 = 1$$

Likewise,

$$R_{xy} = \mathsf{E}\,\boldsymbol{x}\boldsymbol{y}^* = \begin{bmatrix} \mathsf{E}\,\boldsymbol{x}\boldsymbol{y}^*(0) & \mathsf{E}\,\boldsymbol{x}\boldsymbol{y}^*(1) \end{bmatrix} = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

so that

$$k_o^* = R_{xy}R_y^{-1} = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} = \frac{1}{3}\begin{bmatrix} 1 & 1 \end{bmatrix}$$

That is,

$$\hat{\boldsymbol{x}} = \frac{1}{3}\,[\boldsymbol{y}(0)+\boldsymbol{y}(1)]$$

Again, this expression for the linear least-mean-squares estimator of $\boldsymbol{x}$ given $\{\boldsymbol{y}(0), \boldsymbol{y}(1)\}$ holds regardless of whether the noises $\{\boldsymbol{v}(0), \boldsymbol{v}(1)\}$ are Gaussian or not. Only the first and second moments of $\{\boldsymbol{v}(0), \boldsymbol{v}(1)\}$, namely, their means and variances, are needed to determine k_o^*. In the context of the two-antenna example of Fig. 2.1, the above result leads to the optimal *linear* receiver structure shown in Fig. 4.1.

1ST DESIGN EXAMPLE

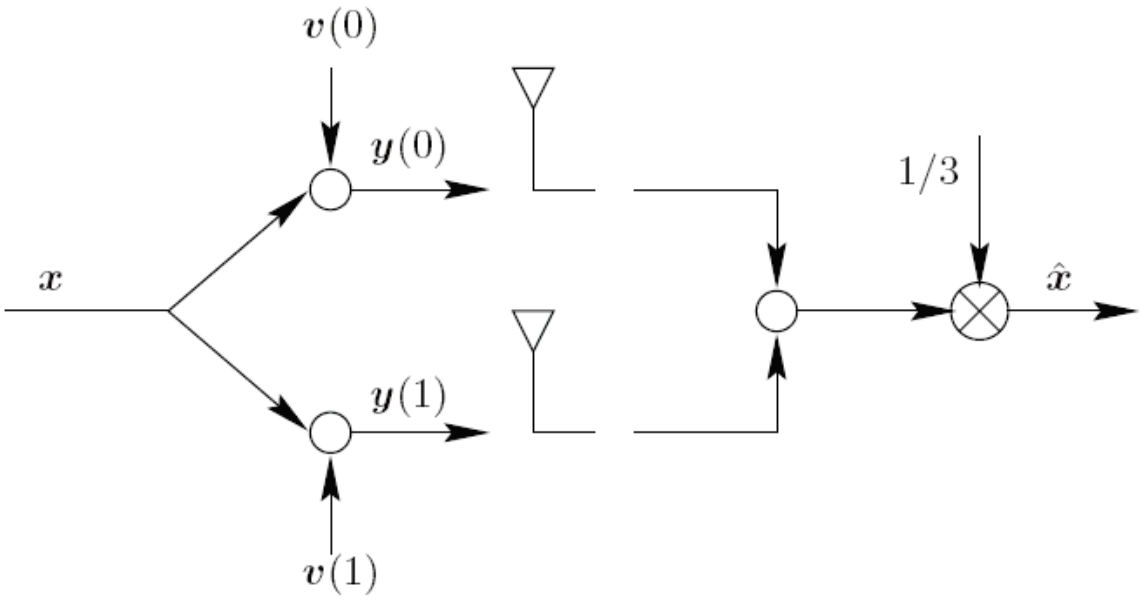


FIGURE 4.1 An optimal linear receiver for recovering a BPSK transmission from two measurements in the presence of additive unit-variance uncorrelated noises.

2ND DESIGN EXAMPLE

Example 4.2 (Multiple measurements of a binary signal)

Continuing with Ex. 4.1, let us examine what happens if we increase the number of available measurements from 2 to N, say

$$\boldsymbol{y}(i) = \boldsymbol{x} + \boldsymbol{v}(i), \qquad i = 0, 1, \ldots, N-1$$

Introduce the observation vector $\boldsymbol{y} = \text{col}\{\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(N-1)\}$. Then, say for $N = 5$,

$$R_{xy} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \qquad R_y = \begin{bmatrix} 2 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 \end{bmatrix}$$

2ND DESIGN EXAMPLE

so that

$$\hat{\boldsymbol{x}} = R_{xy}R_y^{-1}\boldsymbol{y} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 \end{bmatrix}^{-1} \right) \cdot \boldsymbol{y}$$

We need to evaluate R_y^{-1}. Due to the special structure of R_y, its inverse can be evaluated in closed form for any N, as we explain below. Later, in Sec. 5.5, when we reconsider this problem, we shall show how to evaluate $\hat{x}$ via a more direct route.

Observe that, for any N, the matrix R_y can be expressed as $R_y = \mathrm{I} + aa^{\mathsf{T}}$, where a is the $N \times 1$ column vector $a = \mathrm{col}\{1, 1, 1, \ldots, 1\}$. In other words, R_y is a rank-one modification of the identity matrix. This is a useful observation since the inverse of every such matrix has a similar form (see Prob. II.1). Specifically,

2ND DESIGN EXAMPLE

$$\left(\mathrm{I}+aa^{\mathsf{T}}\right)^{-1} = \mathrm{I}-\frac{aa^{\mathsf{T}}}{1+\|a\|^2} = \mathrm{I}-\frac{1}{N+1}aa^{\mathsf{T}}$$

where $\|a\|^2$ denotes the squared Euclidean norm of a, $\|a\|^2 = a^{\mathsf{T}}a$. Using this result we find that

$$R_{xy}R_y^{-1} = a^{\mathsf{T}}\left(\mathrm{I}-\frac{aa^{\mathsf{T}}}{N+1}\right) = a^{\mathsf{T}}-\frac{N}{N+1}a^{\mathsf{T}} = \frac{a^{\mathsf{T}}}{N+1}$$

so that

$$\hat{\boldsymbol{x}} = R_{xy}R_y^{-1}\boldsymbol{y} = \frac{a^{\mathsf{T}}}{N+1}\boldsymbol{y} = \frac{1}{N+1}\sum_{k=0}^{N-1}\boldsymbol{y}(k)$$

◇