CHAPTER 1

Scalar-Valued Data

In this first part of the book we focus on the basic, yet fundamental, problem of estimating an unobservable quantity from a collection of measurements in the least-mean-squares sense. The estimation task is made more or less difficult depending on how much information the measured data convey about the unobservable quantity. We shall study this estimation problem with increasing degrees of complexity, starting from a simple scenario and building up to more sophisticated cases.

The material is developed initially at a slow pace. This is done deliberately in order to familiarize readers (and especially students) with the basic concepts of estimation theory for both real- *and* complex-valued random variables, as well as for scalar- *and* vector-valued random variables. We hope that, by the end of our exposition, the reader will be convinced that these different scenarios (of real vs. complex and scalar vs. vector) can be masked by adopting a uniform vector and complex-conjugation notation. The notation is introduced gradually in the two initial chapters and will be used throughout the book thereafter.

Before plunging into a study of least-mean-squares estimation theory, and the reasons for its widespread use, the reader is advised to consult the review material in Secs. A.1–A.4. These sections provide an intuitive explanation for what the variance of a random variable means. The sections also introduce several useful concepts such as complex- and vector-valued random variables and the notions of independence and uncorrelatedness between two random variables. The explanations will help the reader appreciate the value of the least-mean-squares criterion, which is used extensively in later sections and chapters.

1.1 ESTIMATION WITHOUT OBSERVATIONS

We initiate our discussions of estimation theory by posing and solving a simple (almost trivial) estimation problem. Thus, suppose that all we know about a real-valued random variable $\boldsymbol{x}$ is its mean $\bar{x}$ and its variance σ_x^2, and that we wish to estimate the value that $\boldsymbol{x}$ will assume in a given experiment. We shall denote the *estimate* of $\boldsymbol{x}$ by $\hat{x}$; it is a deterministic quantity (i.e., a number). But how do we come up with a value for $\hat{x}$? And how do we decide whether this value is optimal or not? And if optimal, in what sense? These inquiries are at the heart of every estimation problem.

To answer these questions, we first need to choose a cost function to penalize the estimation error. The resulting estimate $\hat{x}$ will be optimal only in the sense that it leads to the smallest cost value. Different choices for the cost function will generally lead to different choices for $\hat{x}$, each of which will be optimal in its own way.

The design criterion we shall adopt is the *mean-square-error* criterion. It is based on introducing the error signal

$$\tilde{\boldsymbol{x}} \overset{\Delta}{=} \boldsymbol{x} - \hat{x}$$

Adaptive Filters, by Ali H. Sayed

and then determining $\hat{x}$ by minimizing the mean-square-error (m.s.e.), which is defined as the expected value of $\tilde{\boldsymbol{x}}^2$, i.e.,

$$\min_{\hat{x}} \ \mathsf{E}\ \tilde{\boldsymbol{x}}^2 \tag{1.1}$$

The error $\tilde{\boldsymbol{x}}$ is a random variable since $\boldsymbol{x}$ is random. The resulting estimate, $\hat{x}$, will be called the *least-mean-squares estimate* of $\boldsymbol{x}$. The following result is immediate (and, in fact, intuitively obvious as we explain below).

Lemma 1.1 (Lack of observations) The least-mean-squares estimate of $\boldsymbol{x}$ given knowledge of $(\bar{x}, \sigma_x^2)$ is $\hat{x} = \bar{x}$. The resulting minimum cost is $\mathsf{E}\,\tilde{\boldsymbol{x}}^2 = \sigma_x^2$.

Proof: Expand the mean-square error by subtracting and adding $\bar{x}$ as follows:

$$\mathsf{E}\,\tilde{\boldsymbol{x}}^2 = \mathsf{E}\,(\boldsymbol{x}-\hat{x})^2 = \mathsf{E}\,[(\boldsymbol{x}-\bar{x}) + (\bar{x}-\hat{x})]^2 = \sigma_x^2 + (\bar{x}-\hat{x})^2$$

The choice of $\hat{x}$ that minimizes the m.s.e. is now evident. Only the term $(\bar{x}-\hat{x})^2$ is dependent on $\hat{x}$ and this term can be annihilated by choosing $\hat{x} = \bar{x}$. The resulting minimum mean-square error (m.m.s.e.) is then

$$\text{m.m.s.e.} \ \stackrel{\Delta}{=} \ \mathsf{E}\,\tilde{\boldsymbol{x}}^2 \ = \ \sigma_x^2$$

An alternative derivation would be to expand the cost function as

$$\mathsf{E}\,(\boldsymbol{x}-\hat{x})^2 = \mathsf{E}\,\boldsymbol{x}^2 - 2\bar{x}\hat{x} + \hat{x}^2$$

and to differentiate it with respect to $\hat{x}$. By setting the derivative equal to zero we arrive at the same conclusion, namely, $\hat{x} = \bar{x}$.

◇

There are several good reasons for choosing the mean-square-error criterion (1.1). The simplest one perhaps is that the criterion is amenable to mathematical manipulations, more so than any other criterion. In addition, the criterion is essentially attempting to force the estimation error to assume values close to its mean, which happens to be zero. This is because

$$\mathsf{E}\,\tilde{\boldsymbol{x}} = \mathsf{E}\,(\boldsymbol{x}-\hat{x}) = \mathsf{E}\,(\boldsymbol{x}-\bar{x}) = \bar{x}-\bar{x} = 0$$

and, by minimizing $\mathsf{E}\,\tilde{\boldsymbol{x}}^2$, we are in effect minimizing the variance of the error, $\tilde{\boldsymbol{x}}$. In view of the discussion in Sec. A.1 regarding the interpretation of the variance of a random variable, we find that the mean-square-error criterion is therefore attempting to increase the likelihood of small errors.

The effectiveness of the estimation procedure (1.1) can be measured by examining the value of the minimum cost, which is the variance of the resulting estimation error. The above lemma tells us that the minimum cost is equal to σ_x^2. That is,

$$\sigma_{\tilde{x}}^2 \ = \ \sigma_x^2$$

so that the estimate $\hat{x} = \bar{x}$ does not reduce our initial uncertainty about $\boldsymbol{x}$ since the error variable still has the same variance as $\boldsymbol{x}$ itself! We thus find that the performance of the mean-square-error design procedure is limited in this case. Clearly, we are more interested in estimation procedures that result in error variances that are smaller than the original signal variance. We shall discuss one such procedure in the next section.

The reason for the poor performance of the estimate $\hat{x} = \bar{x}$ lies in the lack of more sophisticated prior information about $\boldsymbol{x}$. Note that Lemma 1.1 simply tells us that the best we can do, in the absence of any other information about a random variable $\boldsymbol{x}$, other than its mean and variance, is to use the mean value of $\boldsymbol{x}$ as our estimate. This statement is, in a sense, intuitive. After all, the mean value of a random variable is, by definition, an indication of the value that we would expect to occur on average in repeated experiments. Hence, in answer to the question: what is the best guess for $\boldsymbol{x}$?, the analysis tells us that the best guess is what we would expect for $\boldsymbol{x}$ on average! This is a circular answer, but one that is at least consistent with intuition.

Example 1.1 (Binary signal)

Assume $\boldsymbol{x}$ represents a BPSK (binary phase-shift keying) signal that is equal to ± 1 with probability $1/2$ each. Then

$$\bar{x} = \frac{1}{2} \cdot (1) \; + \; \frac{1}{2} \cdot (-1) \; = \; 0$$

and

$$\sigma_x^2 = \mathsf{E}\, \boldsymbol{x}^2 = 1$$

Now given knowledge of $\{\bar{x}, \sigma_x^2\}$ alone, the best estimate of $\boldsymbol{x}$ in the least-mean-squares sense is $\hat{x} = \bar{x} = 0$. This example shows that the least-mean-squares (and, hence, optimal) estimate does not always lead to a meaningful solution! In this case, $\hat{x} = 0$ is not useful in guessing whether $\boldsymbol{x}$ is 1 or -1 in a given realization. If we could incorporate into the design of the estimator the knowledge that $\boldsymbol{x}$ is a BPSK signal, or some other related information, then we could perhaps come up with a better estimate for $\boldsymbol{x}$.

◇

1.2 ESTIMATION GIVEN DEPENDENT OBSERVATIONS

So let us examine the case in which more is known about a random variable $\boldsymbol{x}$, other than its mean and variance. Specifically, let us assume that we have access to an observation of a second random variable $\boldsymbol{y}$ that is related to $\boldsymbol{x}$ in some way. For example, $\boldsymbol{y}$ could be a noisy measurement of $\boldsymbol{x}$, say, $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v}$, where $\boldsymbol{v}$ denotes the disturbance, or $\boldsymbol{y}$ could be the sign of $\boldsymbol{x}$, or dependent on $\boldsymbol{x}$ in some other manner.

Given two dependent random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$, we therefore pose the problem of determining the least-mean-squares *estimator* of $\boldsymbol{x}$ given $\boldsymbol{y}$. Observe that we are now employing the terminology *estimator* of $\boldsymbol{x}$ as opposed to *estimate* of $\boldsymbol{x}$. In order to highlight this distinction, we denote the estimator of $\boldsymbol{x}$ by the boldface notation $\hat{\boldsymbol{x}}$; it is a random variable that is defined as a function of $\boldsymbol{y}$, say,

$$\hat{\boldsymbol{x}} = h(\boldsymbol{y})$$

for some function $h(\cdot)$ to be determined. Once the function $h(\cdot)$ has been determined, evaluating it at a particular occurrence of $\boldsymbol{y}$, say, for $\boldsymbol{y} = y$, will result in an estimate for $\boldsymbol{x}$, i.e.,

$$\hat{x} = \left. h(\boldsymbol{y}) \right|_{\boldsymbol{y}=y} = h(y)$$

Different occurrences for $\boldsymbol{y}$ lead to different estimates $\hat{x}$. In Sec. 1.1 we did not need to make this distinction between an estimator $\hat{\boldsymbol{x}}$ and an estimate $\hat{x}$. There we sought directly an estimate $\hat{x}$ for $\boldsymbol{x}$ since we did not have access to a random variable $\boldsymbol{y}$; we only had access to the deterministic quantities $\{\bar{x}, \sigma_x^2\}$.

The criterion we shall use to determine the estimator $\hat{\boldsymbol{x}}$ is still the mean-square-error criterion. We define the error signal

$$\boxed{\tilde{\boldsymbol{x}} \ \overset{\Delta}{=} \ \boldsymbol{x} - \hat{\boldsymbol{x}}} \tag{1.2}$$

and then determine $\hat{\boldsymbol{x}}$ by minimizing the mean-square-error over all possible functions $h(\cdot)$:

$$\boxed{\min_{h(\cdot)} \quad \mathsf{E}\ \tilde{\boldsymbol{x}}^2} \tag{1.3}$$

The solution is given by the following statement.

Theorem 1.1 (Optimal mean-square-error estimator) The least-mean-squares estimator (l.m.s.e.) of $\boldsymbol{x}$ given $\boldsymbol{y}$ is the conditional expectation of $\boldsymbol{x}$ given $\boldsymbol{y}$, i.e., $\hat{\boldsymbol{x}} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$. The resulting estimate is

$$\hat{x} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y} = y) \ = \ \int_{\mathcal{S}_x} x f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)\mathrm{d}x$$

where $\mathcal{S}_x$ denotes the support (or domain) of the random variable $\boldsymbol{x}$. Moreover, the estimator is unbiased, i.e., $\mathsf{E}\,\hat{\boldsymbol{x}} = \bar{x}$, and the resulting minimum cost is $\mathsf{E}\,\tilde{\boldsymbol{x}}^2 \ = \ \sigma_x^2 \ - \ \sigma_{\hat{x}}^2$.

Proof: There are several ways to establish the result. Our argument is based on recalling that for any two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, it holds that (see Prob. I.4):

$$\boxed{\mathsf{E}\,\boldsymbol{x} \ = \ \mathsf{E}\,[\mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})]} \tag{1.4}$$

where the outermost expectation on the right-hand side is with respect to $\boldsymbol{y}$, while the innermost expectation is with respect to $\boldsymbol{x}$. We shall indicate these facts explicitly by showing the variables with respect to which the expectations are performed, so that (1.4) is rewritten as

$$\mathsf{E}\,\boldsymbol{x} \ = \ \mathsf{E}_{\,y}[\mathsf{E}_{\,x}(\boldsymbol{x}|\boldsymbol{y})]$$

It now follows that, for any function of $\boldsymbol{y}$, say, $g(\boldsymbol{y})$, it holds that

$$\mathsf{E}_{\,x,y}\ \boldsymbol{x}g(\boldsymbol{y}) = \mathsf{E}_{\,y}\left[\mathsf{E}_{\,x}\big(\boldsymbol{x}g(\boldsymbol{y})|\boldsymbol{y}\big)\right] = \mathsf{E}_{\,y}\left[\mathsf{E}_{\,x}\big(\boldsymbol{x}|\boldsymbol{y}\big)g(\boldsymbol{y})\right] = \mathsf{E}_{\,x,y}\left[\mathsf{E}_{\,x}\big(\boldsymbol{x}|\boldsymbol{y}\big)\right]g(\boldsymbol{y})$$

This means that, for any $g(\boldsymbol{y})$, it holds that $\mathsf{E}_{\,x,y}\left[\boldsymbol{x} - \mathsf{E}_{\,x}\big(\boldsymbol{x}|\boldsymbol{y}\big)\right]g(\boldsymbol{y}) \ = \ 0$, which we write more compactly as

$$\boxed{\mathsf{E}\ [\boldsymbol{x} - \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})]\,g(\boldsymbol{y}) = 0} \tag{1.5}$$

Expression (1.5) states that the random variable $\boldsymbol{x} - \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$ is uncorrelated with any function $g(\cdot)$ of $\boldsymbol{y}$. Indeed, as mentioned before in Sec. A.2, two random variables $\boldsymbol{a}$ and $\boldsymbol{b}$ are uncorrelated if, and only if, their cross-correlation is zero, i.e., $\mathsf{E}\,(\boldsymbol{a} - \bar{a})(\boldsymbol{b} - \bar{b}) \ = \ 0$. On the other hand, the random variables are said to be *orthogonal* if, and only if, $\mathsf{E}\,\boldsymbol{a}\boldsymbol{b} \ = \ 0$. It is easy to verify that the concepts of orthogonality and uncorrelatedness coincide if at least one of the random variables is zero mean. From equation (1.5) we conclude that the variables $\boldsymbol{x} - \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$ and $g(\boldsymbol{y})$ are orthogonal. However, since $\boldsymbol{x} - \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$ is zero mean, then we can also say that they are uncorrelated.

Using this intermediate result, we return to the cost function (1.3), add and subtract $\mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$ to its argument, and express it as

$$\mathsf{E}\,(\boldsymbol{x}-\hat{\boldsymbol{x}})^2 = \mathsf{E}\,[\boldsymbol{x}-\mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})+\mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})-\hat{\boldsymbol{x}}]^2$$

The term $\mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})-\hat{\boldsymbol{x}}$ is a function of $\boldsymbol{y}$. Therefore, if we choose $g(\boldsymbol{y}) = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})-\hat{\boldsymbol{x}}$, then from the orthogonality property (1.5) we conclude that

$$\mathsf{E}\,(\boldsymbol{x}-\hat{\boldsymbol{x}})^2 = \mathsf{E}\,[\boldsymbol{x}-\mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})]^2 \;+\; \mathsf{E}\,[\mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})-\hat{\boldsymbol{x}}]^2$$

Now only the second term on the right-hand side is dependent on $\hat{\boldsymbol{x}}$ and the m.s.e. is minimized by choosing $\hat{\boldsymbol{x}} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$. To evaluate the resulting m.m.s.e. we first note that the optimal estimator is unbiased since

$$\mathsf{E}\,\hat{\boldsymbol{x}} \;=\; \mathsf{E}\,[\mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})] \;=\; \mathsf{E}\,\boldsymbol{x} \;=\; \bar{x}$$

and its variance is therefore given by $\sigma_{\hat{x}}^2 = \mathsf{E}\,\hat{\boldsymbol{x}}^2 - \bar{x}^2$. Moreover, in view of the orthogonality property (1.5), and in view of the fact that $\hat{\boldsymbol{x}} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$ is itself a function of $\boldsymbol{y}$, we have

$$\boxed{\mathsf{E}\,(\boldsymbol{x}-\hat{\boldsymbol{x}})\hat{\boldsymbol{x}} = 0} \tag{1.6}$$

In other words, the estimation error, $\tilde{\boldsymbol{x}}$, is uncorrelated with the optimal estimator. Using this result, we can evaluate the m.m.s.e. as follows:

$$\begin{aligned}
\mathsf{E}\,\tilde{\boldsymbol{x}}^2 &= \mathsf{E}\,[\boldsymbol{x}-\hat{\boldsymbol{x}}][\boldsymbol{x}-\hat{\boldsymbol{x}}] \;=\; \mathsf{E}\,[\boldsymbol{x}-\hat{\boldsymbol{x}}]\boldsymbol{x} \qquad \text{(because of (1.6))}\\
&= \mathsf{E}\,\boldsymbol{x}^2 \;-\; \mathsf{E}\,\hat{\boldsymbol{x}}[\tilde{\boldsymbol{x}}+\hat{\boldsymbol{x}}]\\
&= \mathsf{E}\,\boldsymbol{x}^2 \;-\; \mathsf{E}\,\hat{\boldsymbol{x}}^2 \qquad \text{(because of (1.6))}\\
&= \left(\mathsf{E}\,\boldsymbol{x}^2 \;-\; \bar{x}^2\right) \;+\; \left(\bar{x}^2 \;-\; \mathsf{E}\,\hat{\boldsymbol{x}}^2\right) \;=\; \sigma_x^2 - \sigma_{\hat{x}}^2
\end{aligned}$$

◇

Theorem 1.1 tells us that the least-mean-squares estimator of $\boldsymbol{x}$ is its conditional expectation given $\boldsymbol{y}$. This result is again intuitive. In answer to the question: what is the best guess for $\boldsymbol{x}$ given that we observed $\boldsymbol{y}$?, the analysis tells us that the best guess is what we would expect for $\boldsymbol{x}$ given the occurrence of $\boldsymbol{y}$!

Example 1.2 (Noisy measurement of a binary signal)

Let us return to Ex. 1.1, where $\boldsymbol{x}$ is a BPSK signal that assumes the values ± 1 with probability $1/2$. Assume now that in addition to the mean and variance of $\boldsymbol{x}$, we also have access to a noisy observation of $\boldsymbol{x}$, say,

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v}$$

Assume further that the signal $\boldsymbol{x}$ and the disturbance $\boldsymbol{v}$ are independent, with $\boldsymbol{v}$ being a zero-mean Gaussian random variable of unit variance, i.e., its pdf is given by

$$f_{\boldsymbol{v}}(v) = \frac{1}{\sqrt{2\pi}}\, e^{-v^2/2}$$

Our intuition tells us that we should be able to do better here than in Ex. 1.1. But beware, even here, we shall make some interesting observations.

According to Thm. 1.1, the optimal estimate of $\boldsymbol{x}$ given an observation of $\boldsymbol{y}$ is

$$\hat{x} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y}=y) \;=\; \int_{-\infty}^{\infty} x f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)\mathsf{d}x \tag{1.7}$$

We therefore need to determine the conditional pdf, $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$, and evaluate the integral (1.7). For this purpose, we start by noting, from probability theory, that the pdf of the sum of two independent random variables, namely, $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v}$, is equal to the convolution of their individual pdfs, i.e.,

$$f_{\boldsymbol{y}}(y) = \int_{-\infty}^{\infty} f_{\boldsymbol{x}}(x) f_{\boldsymbol{v}}(y-x)\mathsf{d}x$$

In this example, we have

$$f_{\boldsymbol{x}}(x) = \frac{1}{2}\delta(x-1) \; + \; \frac{1}{2}\delta(x+1)$$

where $\delta(\cdot)$ is the Dirac-delta function, so that $f_{\boldsymbol{y}}(y)$ is given by

$$f_{\boldsymbol{y}}(y) = \frac{1}{2} f_{\boldsymbol{v}}(y+1) + \frac{1}{2} f_{\boldsymbol{v}}(y-1) \tag{1.8}$$

Moreover, the joint pdf of $\{\boldsymbol{x}, \boldsymbol{y}\}$ is given by

$$\begin{aligned} f_{\boldsymbol{x},\boldsymbol{y}}(x,y) &= f_{\boldsymbol{x}}(x) \cdot f_{\boldsymbol{y}|\boldsymbol{x}}(y|x) \\ &= \left[\frac{1}{2}\delta(x-1) \; + \; \frac{1}{2}\delta(x+1)\right] \cdot f_{\boldsymbol{v}}(y-x) \\ &= \frac{1}{2} f_{\boldsymbol{v}}(y-1)\delta(x-1) + \frac{1}{2} f_{\boldsymbol{v}}(y+1)\delta(x+1) \end{aligned}$$

Using (A.6) we get

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \;=\; \frac{f_{\boldsymbol{x},\boldsymbol{y}}(x,y)}{f_{\boldsymbol{y}}(y)} \;=\; \frac{f_{\boldsymbol{v}}(y-1)\delta(x-1)}{f_{\boldsymbol{v}}(y+1) + f_{\boldsymbol{v}}(y-1)} + \frac{f_{\boldsymbol{v}}(y+1)\delta(x+1)}{f_{\boldsymbol{v}}(y+1) + f_{\boldsymbol{v}}(y-1)}$$

Substituting into expression (1.7) for $\hat{x}$ and integrating we obtain

$$\begin{aligned} \hat{x} &= \frac{f_{\boldsymbol{v}}(y-1)}{f_{\boldsymbol{v}}(y+1) + f_{\boldsymbol{v}}(y-1)} - \frac{f_{\boldsymbol{v}}(y+1)}{f_{\boldsymbol{v}}(y+1) + f_{\boldsymbol{v}}(y-1)} \\ &= \frac{1}{\left(\dfrac{e^{-(y+1)^2/2}}{e^{-(y-1)^2/2}}\right) + 1} - \frac{1}{\left(\dfrac{e^{-(y-1)^2/2}}{e^{-(y+1)^2/2}}\right) + 1} = \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}} \;\stackrel{\Delta}{=}\; \mathsf{tanh}\, y \end{aligned}$$

In other words, the least-mean-squares estimator of $\boldsymbol{x}$ is the hyperbolic tangent function,

$$\boxed{\hat{\boldsymbol{x}} = \mathsf{tanh}(\boldsymbol{y})} \tag{1.9}$$

The result is represented schematically in Fig. 1.1.

Figure 1.2 plots the function $\mathsf{tanh}(y)$. We see that it tends to ± 1 as $y \longrightarrow \pm\infty$. For other values of y, the function assumes real values that are distinct from ± 1. This is a bit puzzling from the designer's perspective. The designer is interested in knowing whether the symbol $\boldsymbol{x}$ is $+1$ or -1 based on the observed value of $\boldsymbol{y}$. The above construction tells the designer to estimate $\boldsymbol{x}$ by computing $\mathsf{tanh}(y)$. But this value will never be exactly $+1$ or -1; it will be a real number inside

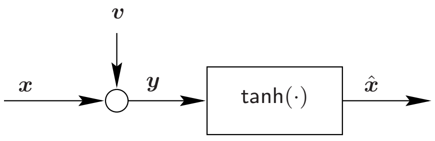

FIGURE 1.1 Optimal estimation of a BPSK signal embedded in unit-variance additive Gaussian noise.

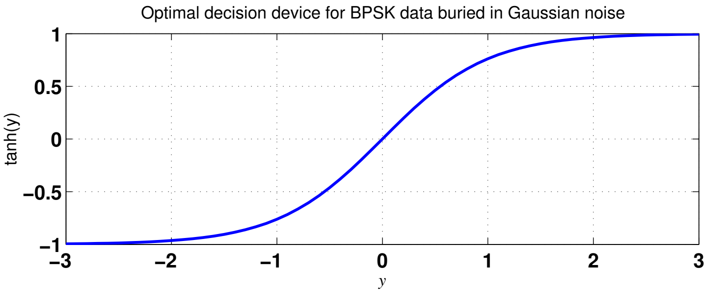

FIGURE 1.2 A plot of the function $\mathsf{tanh}(y)$.

the interval $(-1, 1)$. The designer will then be induced to make a hard decision of the form:

$$\text{decide in favor of } \begin{cases} +1 & \text{if } \hat{x} \text{ is nonnegative} \\ -1 & \text{if } \hat{x} \text{ is negative} \end{cases}$$

In effect, the designer ends up implementing the alternative estimator:

$$\hat{\boldsymbol{x}} \;=\; \mathsf{sign}[\mathsf{tanh}(\boldsymbol{y})] \tag{1.10}$$

where $\mathsf{sign}(\cdot)$ denotes the sign of its argument; it is equal to $+1$ if the argument is nonnegative and -1 otherwise.

We therefore have a situation where the optimal estimator, although known in closed form, does not solve the original problem of recovering the symbols ± 1 directly. Instead, the designer is forced to implement a suboptimal solution; it is suboptimal from a least-mean-squares point of view. Even more puzzling, the designer could consider implementing the alternative (and simpler) suboptimal estimator:

$$\hat{\boldsymbol{x}} = \mathsf{sign}(\boldsymbol{y}) \tag{1.11}$$

where the $\mathsf{sign}(\cdot)$ function operates directly on $\boldsymbol{y}$ rather than on $\mathsf{tanh}(\boldsymbol{y})$ — see Fig. 1.3. Both suboptimal implementations (1.10) and (1.11) lead to the same result since, as is evident from Fig. 1.2, $\mathsf{sign}[\mathsf{tanh}(y)] = \mathsf{sign}(y)$. In the computer project at the end of this part we shall compare the performance of the optimal and suboptimal estimators (1.9)–(1.11).

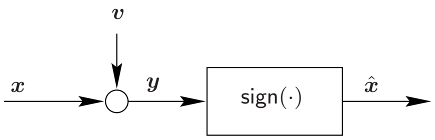

FIGURE 1.3 Suboptimal estimation of a BPSK signal embedded in unit-variance additive Gaussian noise.

We may mention that in the digital communications literature, especially in studies on equalization methods, an implementation using (1.11) is usually said to be based on *hard decisions*, while an implementation using (1.9) is said to be based on *soft decisions*.

◇

Remark 1.1 (Complexity of optimal estimation) Example 1.2 highlights one of the inconveniences of working with the optimal estimator of Thm. 1.1. Although the form of the optimal solution is given explicitly by $\hat{\boldsymbol{x}} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$, in general it is not an easy task to find a closed-form expression for the conditional expectation of two random variables (especially for other choices of probability density functions). Moreover, even when a closed-form expression can be found, one is usually led to a nonlinear estimator whose implementation may not be practical or may even be costly. For this reason, from Part II (*Linear Estimation*) onwards, we shall restrict the class of estimators to *linear* estimators, and study the capabilities of these estimators.

◇

The purpose of Exs. 1.1 and 1.2 is not to confuse the reader, but rather to stress the fact that an optimal estimator is optimal only in the sense that it satisfies a certain optimality criterion. One should not confuse an optimal guess with a perfect guess. One should also not confuse an optimal guess with a practical one; an optimal guess need not be perfect or even practical, though it can suggest good practical solutions.

1.3 ORTHOGONALITY PRINCIPLE

There are two important conclusions that follow from the proof of Thm. 1.1, namely, the orthogonality properties (1.5) and (1.6). The first one states that the difference

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

is orthogonal to any function $g(\cdot)$ of $\boldsymbol{y}$. Now since we already know that the conditional expectation, $\mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$, is the optimal least-mean-squares estimator of $\boldsymbol{x}$, we can restate this result by saying that the estimation error $\tilde{\boldsymbol{x}}$ is orthogonal to any function of $\boldsymbol{y}$,

$$\boxed{\mathsf{E}\,\tilde{\boldsymbol{x}}\; g(\boldsymbol{y}) = 0} \tag{1.12}$$

We shall sometimes use a geometric notation to refer to this result and write instead

$$\boxed{\tilde{\boldsymbol{x}} \;\perp\; g(\boldsymbol{y})} \tag{1.13}$$

where the symbol $\perp$ is used to signify that the two random variables are orthogonal; a schematic representation of this orthogonality property is shown in Fig. 1.4.
Relation (1.13) admits the following interpretation. It states that the optimal estimator $\hat{\boldsymbol{x}} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$ is such that the resulting error, $\tilde{\boldsymbol{x}}$, is orthogonal to (and, in fact, also uncorrelated with) any transformation of the data $\boldsymbol{y}$. In other words, the optimal estimator is such that no matter how we modify the data $\boldsymbol{y}$, there is no way we can extract additional information from the data in order to reduce the variance of $\tilde{\boldsymbol{x}}$ any further. This is because any additional processing of $\boldsymbol{y}$ will remain uncorrelated with $\tilde{\boldsymbol{x}}$.

The second orthogonality property (1.6) is a special case of (1.13). It states that

$$\boxed{\tilde{\boldsymbol{x}} \;\perp\; \hat{\boldsymbol{x}}}$$

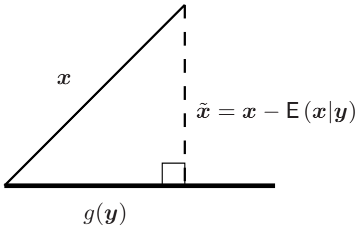
$\boldsymbol{x}$
$\tilde{\boldsymbol{x}} = \boldsymbol{x} - \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$
$g(\boldsymbol{y})$

FIGURE 1.4 The orthogonality condition: $\tilde{\boldsymbol{x}} \perp g(\boldsymbol{y})$.

That is, the estimation error is orthogonal to (or uncorrelated with) the estimator itself. This is a special case of (1.13) since $\hat{\boldsymbol{x}}$ is a function of $\boldsymbol{y}$ by virtue of the result $\hat{\boldsymbol{x}} = \mathsf{E}(\boldsymbol{x}|\boldsymbol{y})$.

In summary, the optimal least-mean-squares estimator is such that the estimation error is orthogonal to the estimator and, more generally, to any function of the observation. It turns out that the converse statement is also true so that the orthogonality condition (1.13) is in fact a *defining* property of optimality in the least-mean-squares sense.

Theorem 1.2 (Orthogonality condition) Given two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, an estimator $\hat{\boldsymbol{x}} = h(\boldsymbol{y})$ is optimal in the least-mean-squares sense (1.3) if, and only if, $\hat{\boldsymbol{x}}$ is unbiased (i.e., $\mathsf{E}\,\hat{\boldsymbol{x}} = \bar{x}$) and $\boldsymbol{x} - \hat{\boldsymbol{x}} \perp g(\boldsymbol{y})$ for any function $g(\cdot)$.

Proof: One direction has already been proven prior to the statement of the theorem, namely, if $\hat{\boldsymbol{x}}$ is the optimal estimator and hence, $\hat{\boldsymbol{x}} = \mathsf{E}(\boldsymbol{x}|\boldsymbol{y})$, then we already know from (1.13) that $\tilde{\boldsymbol{x}} \perp g(\boldsymbol{y})$, for any $g(\cdot)$. Moreover, we know from Thm. 1.1 that this estimator is unbiased.

Conversely, assume $\hat{\boldsymbol{x}}$ is some unbiased estimator for $\boldsymbol{x}$ and that it satisfies $\boldsymbol{x} - \hat{\boldsymbol{x}} \perp g(\boldsymbol{y})$, for any $g(\cdot)$. Define the random variable $\boldsymbol{z} = \hat{\boldsymbol{x}} - \mathsf{E}(\boldsymbol{x}|\boldsymbol{y})$ and let us show that it is the zero variable with probability one. For this purpose, we note first that $\boldsymbol{z}$ is zero mean since

$$\mathsf{E}\,\boldsymbol{z} = \mathsf{E}\,\hat{\boldsymbol{x}} \;-\; \mathsf{E}\left(\mathsf{E}(\boldsymbol{x}|\boldsymbol{y})\right) \;=\; \bar{x} - \bar{x} \;=\; 0$$

Moreover, from (1.5) we have $\boldsymbol{x} - \mathsf{E}(\boldsymbol{x}|\boldsymbol{y}) \perp g(\boldsymbol{y})$ and, by assumption, we have $\boldsymbol{x} - \hat{\boldsymbol{x}} \perp g(\boldsymbol{y})$ for any $g(\cdot)$. Subtracting these two conditions we conclude that $\boldsymbol{z} \;\perp g(\boldsymbol{y})$, which is the same as $\mathsf{E}\,\boldsymbol{z}g(\boldsymbol{y}) = 0$. Now since the variable $\boldsymbol{z}$ itself is a function of $\boldsymbol{y}$, we may choose $g(\boldsymbol{y}) = \boldsymbol{z}$ to get $\mathsf{E}\,\boldsymbol{z}^2 = 0$. We thus find that $\boldsymbol{z}$ is zero mean and has zero variance, so that, from Remark A.1, we conclude that $\boldsymbol{z} = 0$, or equivalently, $\hat{\boldsymbol{x}} = \mathsf{E}(\boldsymbol{x}|\boldsymbol{y})$, with probability one.

◇

Example 1.3 (Suboptimal estimator for a binary signal)

Consider again Ex. 1.2, where $\boldsymbol{x}$ is a BPSK signal that assumes the values ± 1 with probability $1/2$. Let us verify that the estimator $\hat{\boldsymbol{x}} = \mathsf{sign}(\boldsymbol{y})$ is not optimal in the least-mean squares sense. We already know that this is the case because we found in Ex. 1.2 that the optimal estimator is $\mathsf{tanh}(\boldsymbol{y})$. Here we wish to verify the sub-optimality of $\mathsf{sign}(\boldsymbol{y})$ without assuming prior knowledge of the optimal estimator, and by relying solely on the orthogonality condition (1.12).

According to Thm. 1.2, we need to verify that the estimator $\mathsf{sign}(\boldsymbol{y})$ fails the orthogonality test. In particular, we shall exhibit a function $g(\boldsymbol{y})$ such that the difference $\boldsymbol{x} - \mathsf{sign}(\boldsymbol{y})$ is correlated with it. Actually, we shall choose $g(\boldsymbol{y}) = \mathsf{sign}(\boldsymbol{y})$ and verify that

$$\mathsf{E}\left[\boldsymbol{x} - \mathsf{sign}(\boldsymbol{y})\right]\mathsf{sign}(\boldsymbol{y}) \;\neq\; 0 \tag{1.14}$$

Let us first check whether the estimator $\hat{\boldsymbol{x}} = \mathsf{sign}(\boldsymbol{y})$ is biased or not. For this purpose we recall that $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v}$ and that

$$\mathsf{sign}(\boldsymbol{x} + \boldsymbol{v}) \;=\; \begin{cases} +1 & \text{if } \boldsymbol{x} + \boldsymbol{v} \geq 0 \\ -1 & \text{if } \boldsymbol{x} + \boldsymbol{v} < 0 \end{cases}$$

We therefore need to evaluate the probability of the events $\boldsymbol{x} + \boldsymbol{v} \geq 0$ and $\boldsymbol{x} + \boldsymbol{v} < 0$. For the first case we have

$$\boldsymbol{x} + \boldsymbol{v} \geq 0 \Longleftrightarrow \; (\boldsymbol{x} = +1 \text{ and } \boldsymbol{v} \geq -1) \;\text{ or }\; (\boldsymbol{x} = -1 \text{ and } \boldsymbol{v} \geq 1)$$

Now recall that $\boldsymbol{x}$ and $\boldsymbol{v}$ are independent and that $\boldsymbol{v}$ is a zero-mean unit-variance Gaussian random variable. Thus, let

$$P(\boldsymbol{v} \geq 1) \ \stackrel{\Delta}{=}\ \alpha \tag{1.15}$$

Then

$$P(\boldsymbol{v} \geq -1) \ = \ 1 - P(\boldsymbol{v} \leq -1) \ = \ 1 - P(\boldsymbol{v} \geq 1) \ = \ 1 - \alpha$$

and we obtain

$$P(\boldsymbol{x} + \boldsymbol{v} \geq 0) = (1-\alpha)/2 + \alpha/2 \ = \ 1/2$$

Consequently, $P(\boldsymbol{x} + \boldsymbol{v} < 0) = 1/2$ and $\mathsf{E}\,\mathsf{sign}(\boldsymbol{x}+\boldsymbol{v}) = 0$. This means that the estimator $\hat{\boldsymbol{x}} = \mathsf{sign}(\boldsymbol{y})$ is unbiased. We now return to (1.14) and note that

$$\mathsf{E}\,[\boldsymbol{x} - \mathsf{sign}(\boldsymbol{y})]\mathsf{sign}(\boldsymbol{y}) \ = \ \mathsf{E}\ \boldsymbol{x}\mathsf{sign}(\boldsymbol{y}) \ - \ 1$$

Therefore, all we need to do in order to verify that (1.14) holds is to check that $\mathsf{E}\,\boldsymbol{x}\mathsf{sign}(\boldsymbol{y})$ does not evaluate to one. To do this, we introduce the random variable $\boldsymbol{z} = \boldsymbol{x}\mathsf{sign}(\boldsymbol{y})$ and proceed to evaluate its mean. It is clear from the definition of $\boldsymbol{z}$ that

$$\boldsymbol{z} = \begin{cases} +1 & \text{if } (\boldsymbol{x} = +1 \text{ and } \boldsymbol{v} \geq -1) \ \text{ or } \ (\boldsymbol{x} = -1 \text{ and } \boldsymbol{v} < 1) \\ -1 & \text{if } (\boldsymbol{x} = +1 \text{ and } \boldsymbol{v} < -1) \ \text{ or } \ (\boldsymbol{x} = -1 \text{ and } \boldsymbol{v} \geq 1) \end{cases}$$

The events

$$(\boldsymbol{x} = +1 \text{ and } \boldsymbol{v} \geq -1) \quad \text{or} \quad (\boldsymbol{x} = -1 \text{ and } \boldsymbol{v} < 1)$$

each has probability $0.5(1-\alpha)$. Likewise, the events

$$(\boldsymbol{x} = +1 \text{ and } \boldsymbol{v} < -1) \quad \text{or} \quad (\boldsymbol{x} = -1 \text{ and } \boldsymbol{v} \geq 1)$$

each has probability 0.5α. It then follows that $\mathsf{E}\,\boldsymbol{z} = 1 - 2\alpha \neq 1$, so that $\boldsymbol{x} - \mathsf{sign}(\boldsymbol{y})$ is correlated with $\mathsf{sign}(\boldsymbol{y})$. Hence, the estimator $\mathsf{sign}(\boldsymbol{y})$ does not satisfy the orthogonality condition (1.12) and, therefore, it cannot be the optimal least-mean-squares estimator.

◇

1.4 GAUSSIAN RANDOM VARIABLES

We mentioned earlier in Remark 1.1 that it is not always possible to determine a closed form expression for the optimal estimator $\mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$. Only in some special cases this calculation can be carried out to completion (as we did in Ex. 1.2 and as we shall do in another example below). This difficulty will motivate us to limit ourselves in Part II (*Linear Estimation*) to the subclass of *linear (or affine) estimators*, namely, to choices of $h(\cdot)$ in (1.3) that are *affine* functions of the observation, say, $h(\boldsymbol{y}) = k\boldsymbol{y} + b$ for some constants k and b to be determined. Despite its apparent narrowness, this class of estimators performs reasonably well in many applications.

There is an important special case for which the *optimal* estimator of Thm. 1.1 turns out to be affine in $\boldsymbol{y}$. This scenario happens when the random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly Gaussian. To see this, let us introduce the matrix

$$R \ \stackrel{\Delta}{=}\ \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$$

where $\{\sigma_x^2, \sigma_y^2, \sigma_{xy}\}$ denote the variances and cross-correlation of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively,

$$\sigma_x^2 = \mathsf{E}\,(\boldsymbol{x} - \bar{x})^2, \qquad \sigma_y^2 = \mathsf{E}\,(\boldsymbol{y} - \bar{y})^2, \qquad \sigma_{xy} = \mathsf{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{y} - \bar{y})$$

The matrix R can be regarded as the *covariance* matrix of the column vector $\mathsf{col}\{\boldsymbol{x}, \boldsymbol{y}\}$, namely,

$$R = \mathsf{E}\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} - \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}\right)\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} - \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}\right)^{\mathsf{T}} \tag{1.16}$$

where the symbol T denotes vector transposition, and the notation $\text{col}\{\alpha, \beta\}$ denotes a column vector whose entries are α and β. As explained in Sec. A.4, every such covariance matrix is necessarily symmetric, $R = R^{\mathsf{T}}$. Moreover, R is also nonnegative-definite, written as $R \geq 0$. To proceed with the analysis, we are going to assume that the covariance matrix R is positive-definite and, hence, invertible — see Prob. I.6.

Now the joint pdf of two jointly Gaussian random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ is given by (see Sec. A.5 for a review of Gaussian random variables and their probability density functions):

$$\boxed{f_{\boldsymbol{x},\boldsymbol{y}}(x, y) = \frac{1}{2\pi}\frac{1}{\sqrt{\det R}}\ \exp\left\{-\tfrac{1}{2}\begin{bmatrix} x-\bar{x} & y-\bar{y} \end{bmatrix} R^{-1}\begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix}\right\}} \tag{1.17}$$

Also, the individual probability density functions of $\boldsymbol{x}$ and $\boldsymbol{y}$ are given by

$$\begin{aligned} f_{\boldsymbol{x}}(x) &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_x}\ \exp\left\{-(x-\bar{x})^2/2\sigma_x^2\right\} \\ f_{\boldsymbol{y}}(y) &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_y}\ \exp\left\{-(y-\bar{y})^2/2\sigma_y^2\right\} \end{aligned}$$

According to Thm. 1.1, the least-mean-squares estimator of $\boldsymbol{x}$ given $\boldsymbol{y}$ is $\hat{\boldsymbol{x}} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$, which requires that we determine the conditional pdf $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$. This pdf can be obtained from the calculation:

$$\begin{aligned} f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) &= \frac{f_{\boldsymbol{x},\boldsymbol{y}}(x, y)}{f_{\boldsymbol{y}}(y)} \\ &= \frac{\dfrac{1}{2\pi}\dfrac{1}{\sqrt{\det R}}\ \exp\left\{-\tfrac{1}{2}\begin{bmatrix} x-\bar{x} & y-\bar{y} \end{bmatrix} R^{-1}\begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix}\right\}}{\dfrac{1}{\sqrt{2\pi}}\dfrac{1}{\sigma_y}\ \exp\left\{-(y-\bar{y})^2/2\sigma_y^2\right\}} \end{aligned} \tag{1.18}$$

In order to simplify the above ratio, we shall use the fact that R can be factored into a product of an upper-triangular, diagonal, and lower-triangular matrices, as follows (this can be checked by straightforward algebra):

$$\boxed{R = \begin{bmatrix} 1 & \sigma_{xy}/\sigma_y^2 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ \sigma_{xy}/\sigma_y^2 & 1 \end{bmatrix}} \tag{1.19}$$

where we introduced the scalar

$$\sigma^2 \ \stackrel{\Delta}{=}\ \sigma_x^2 - \sigma_{xy}^2/\sigma_y^2$$

which is called the *Schur complement* of σ_y^2 in R; it is guaranteed to be positive in view of the assumed positive-definiteness of R itself. Indeed, and more generally, let

$$R = \begin{bmatrix} A & B \\ B^{\mathsf{T}} & C \end{bmatrix}$$

be any symmetric matrix with possibly matrix-valued entries $\{A, B, C\}$ satisfying $A = A^{\mathsf{T}}$ and $C = C^{\mathsf{T}}$. Assume further that C is invertible. Then it is easy to verify by direct calculation that every such matrix can be factored in the form

$$\left[\begin{array}{cc} A & B \\ B^{\mathsf{T}} & C \end{array}\right] = \left[\begin{array}{cc} \mathrm{I} & BC^{-1} \\ 0 & \mathrm{I} \end{array}\right] \left[\begin{array}{cc} \Sigma & 0 \\ 0 & C \end{array}\right] \left[\begin{array}{cc} \mathrm{I} & 0 \\ C^{-1}B^{\mathsf{T}} & \mathrm{I} \end{array}\right]$$

where

$$\Sigma = A - BC^{-1}B^{\mathsf{T}}$$

is called the Schur complement of R with respect to C. The factorization (1.19) is a special case of this result where the entries $\{A, B, C\}$ are scalars: $A = \sigma_x^2$, $B = \sigma_{xy}$, and $C = \sigma_y^2$. Moreover, the determinant of a positive-definite matrix is always positive — see Sec. B.1. We see from (1.19) that $\det R = \sigma^2\sigma_y^2$, so that σ^2 is necessarily positive since $\det R > 0$.

Now, by inverting both sides of (1.19), we find that the inverse of R can be factored as

$$R^{-1} = \left[\begin{array}{cc} 1 & 0 \\ -\sigma_{xy}/\sigma_y^2 & 1 \end{array}\right] \left[\begin{array}{cc} 1/\sigma^2 & 0 \\ 0 & 1/\sigma_y^2 \end{array}\right] \left[\begin{array}{cc} 1 & -\sigma_{xy}/\sigma_y^2 \\ 0 & 1 \end{array}\right] \tag{1.20}$$

where we used the simple fact that for any scalar a,

$$\left[\begin{array}{cc} 1 & 0 \\ a & 1 \end{array}\right]^{-1} = \left[\begin{array}{cc} 1 & 0 \\ -a & 1 \end{array}\right], \qquad \left[\begin{array}{cc} 1 & a \\ 0 & 1 \end{array}\right]^{-1} = \left[\begin{array}{cc} 1 & -a \\ 0 & 1 \end{array}\right]$$

Then

$$\left[\begin{array}{cc} x - \bar{x} & y - \bar{y} \end{array}\right] R^{-1} \left[\begin{array}{c} x - \bar{x} \\ y - \bar{y} \end{array}\right] = \frac{[(x-\bar{x}) - \sigma_{xy}\sigma_y^{-2}(y-\bar{y})]^2}{\sigma^2} + \frac{(y-\bar{y})^2}{\sigma_y^2}$$

where the right-hand side is expressed as the sum of two quadratic terms. It follows that

$$\exp\left\{-\frac{1}{2}\left[\begin{array}{cc} x - \bar{x} & y - \bar{y} \end{array}\right] R^{-1} \left[\begin{array}{c} x - \bar{x} \\ y - \bar{y} \end{array}\right]\right\} =$$
$$\exp\left\{-[(x-\bar{x}) - \sigma_{xy}\sigma_y^{-2}(y-\bar{y})]^2/2\sigma^2\right\}\ \exp\left\{-(y-\bar{y})^2/2\sigma_y^2\right\}$$

This equality, along with $\det R = \sigma^2\sigma_y^2$, allows us to simplify expression (1.18) for $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ to

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\sigma^2}}\ \exp\left\{-[(x-\bar{x}) - \sigma_{xy}\sigma_y^{-2}(y-\bar{y})]^2/2\sigma^2\right\}$$

This expression has the form of the pdf of a Gaussian random variable with variance σ^2 and mean value $\bar{x} + \sigma_{xy}\sigma_y^{-2}(y - \bar{y})$. Consequently, the optimal estimator is given by the *affine* relation:

$$\boxed{\hat{\boldsymbol{x}} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y}) = \bar{x} + \frac{\sigma_{xy}}{\sigma_y^2}(\boldsymbol{y} - \bar{y})} \tag{1.21}$$

Moreover, the resulting $\mathsf{m.m.s.e.}$, which is the variance of $\tilde{\boldsymbol{x}} = \boldsymbol{x} - \hat{\boldsymbol{x}}$, is given by

$$\boxed{\mathsf{m.m.s.e.} \ \stackrel{\Delta}{=}\ \sigma_{\tilde{x}}^2 \ = \ \sigma_x^2 - \sigma_{\hat{x}}^2 \ = \ \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2} \ = \ \sigma^2} \tag{1.22}$$

Observe that, in this Gaussian case, the m.m.s.e. is completely specified by the second-order statistics of the random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ (namely, σ_x^2, σ_y^2, and σ_{xy}). Note also that the m.m.s.e. is smaller than σ_x^2.

Example 1.4 (Correlation coefficient)

A measure of the correlation between two random variables is their correlation coefficient, defined by

$$\rho_{xy} \;\triangleq\; \sigma_{xy}/\sigma_x\sigma_y$$

It is shown in Prob. I.6 that ρ_{xy} always lies in the interval $[-1, 1]$. As ρ_{xy} moves closer to zero, the variables $\boldsymbol{x}$ and $\boldsymbol{y}$ become more uncorrelated (in the Gaussian case, this also means that the variables become less dependent). We see from (1.22) that the m.m.s.e. in the Gaussian case can be rewritten in the form

$$\text{m.m.s.e.} \;=\; \sigma_x^2(1-\rho_{xy}^2)$$

This shows that when $\rho_{xy} = 0$, which occurs when $\sigma_{xy} = 0$, the resulting m.m.s.e. is σ_x^2. Also, from (1.21), the estimator collapses to $\hat{\boldsymbol{x}} = \bar{x}$. That is, we are reduced to the simple estimator studied in Sec. 1.1. This is expected since in the Gaussian case, a zero cross-correlation means that the random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent so there is no additional information available that we can use to estimate $\boldsymbol{x}$, besides its mean and variance.

◇

Example 1.5 (Gaussian noise)

Let $\boldsymbol{x}$ denote a Gaussian random variable with mean $\bar{x} = 1$ and variance $\sigma_x^2 = 2$. Similarly, let $\boldsymbol{v}$ denote a Gaussian random variable independent of $\boldsymbol{x}$, with mean $\bar{v} = 2$ and variance σ_v^2. Now consider the noisy measurement

$$\boldsymbol{y} = 2\boldsymbol{x} + \boldsymbol{v}$$

and let us estimate $\boldsymbol{x}$ from $\boldsymbol{y}$. According to (1.21), we need to determine the quantities $\{\bar{y}, \sigma_{xy}, \sigma_y^2\}$. From the above equation we find that

$$\bar{y} = 2\bar{x} + \bar{v} = 4$$

The independence of $\boldsymbol{x}$ and $\boldsymbol{v}$ implies that

$$\sigma_y^2 = 4\sigma_x^2 + \sigma_v^2 = 8 + \sigma_v^2$$

Finally, the cross-correlation σ_{xy} is given by

$$\sigma_{xy} = \mathsf{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y}) = \mathsf{E}\,(\boldsymbol{x}-1)(2\boldsymbol{x}+\boldsymbol{v}-4) \;= 4$$

where we used

$$\mathsf{E}\,\boldsymbol{x}^2 = \sigma_x^2 + \bar{x}^2 = 3$$

and

$$\mathsf{E}\,\boldsymbol{x}\boldsymbol{v} = \mathsf{E}\,\boldsymbol{x}\mathsf{E}\,\boldsymbol{v} = 2$$

Using (1.21) and (1.22) we obtain

$$\hat{\boldsymbol{x}} = 1 + \frac{4}{8+\sigma_v^2}(\boldsymbol{y}-4) \quad \text{and} \quad \sigma_{\tilde{x}}^2 = 2 - \frac{16}{8+\sigma_v^2} \;=\; \frac{2\sigma_v^2}{8+\sigma_v^2}$$

Moreover, since $\sigma_{\tilde{x}}^2 \;=\; \sigma_x^2 - \sigma_{\hat{x}}^2$, we also find that

$$\sigma_{\hat{x}}^2 = \frac{16}{8+\sigma_v^2}$$

◇