Chapter 1

OPTIMAL ESTIMATION

In this first chapter 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 presentation in the chapter relies on some basic concepts from probability theory and random variables. For the benefit of the reader, we shall motivate these concepts whenever needed, as well as highlight their relevance in the estimation context. In this way, readers will be introduced to the necessary concepts in a gradual and motivated manner, and they will come to appreciate their significance away from unnecessary abstractions.

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 chapter and will be used throughout the book thereafter.

1.1 VARIANCE OF A RANDOM VARIABLE

Before plunging into a discussion of least-mean-squares estimation theory, and before giving some reasons for its widespread use, we find it useful to provide an intuitive explanation for what the variance of a random variable means. The explanation will help the reader appreciate the value of the least-mean-squares criterion, which is used extensively in later sections and chapters.

Consider a scalar *real-valued* random variable $\boldsymbol{x}$ with mean value $\bar{x}$ and variance σ_x^2, i.e.,

$$\boxed{\bar{x} \triangleq \mathsf{E}\,\boldsymbol{x}, \qquad \sigma_x^2 \triangleq \mathsf{E}\,(\boldsymbol{x}-\bar{x})^2 = \mathsf{E}\,\boldsymbol{x}^2 - \bar{x}^2} \tag{1.1.1}$$

where the symbol E denotes the expectation operator. Observe that we are using boldface letters to denote random variables, which will be our convention in this book. When $\boldsymbol{x}$ has zero mean, its variance is simply given by $\sigma_x^2 = \mathsf{E}\,\boldsymbol{x}^2$. Intuitively, the variance of $\boldsymbol{x}$ defines an interval on the real axis around $\bar{x}$ where the values of $\boldsymbol{x}$ are most likely to occur:

1. A small σ_x^2 indicates that $\boldsymbol{x}$ is more likely to assume values that are close to its mean, $\bar{x}$.

2. A large σ_x^2 indicates that $\boldsymbol{x}$ can assume values over a wider interval around its mean.

For this reason, it is customary to regard the variance of a random variable as a measure of *uncertainty* about the value it can assume in a given experiment. A small variance indicates that we are more certain about what values to expect for $\boldsymbol{x}$ (namely, values that are close to its mean), while a large variance indicates that we are less certain about what to expect. These two situations are illustrated in Figs. 1.1 and 1.2 for two different probability density functions.

Figure 1.1 plots the probability density function (pdf) of a Gaussian random variable $\boldsymbol{x}$ for two different variances. In both cases, the mean of the random variable is fixed at $\bar{x} = 20$ while the variance is $\sigma_x^2 = 225$ in one case and $\sigma_x^2 = 4$ in the other. Recall that the pdf of a Gaussian random variable is defined in terms of $(\bar{x}, \sigma_x^2)$ by the expression

$$\boxed{f_{\boldsymbol{x}}(x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma_x}\, e^{-\frac{(x-\bar{x})^2}{2\sigma_x^2}}, \quad x \in (-\infty,\infty)} \tag{1.1.2}$$

where σ_x is called the *standard deviation* of $\boldsymbol{x}$. Recall further that the pdf of a random variable is useful in several respects. In particular, it allows us to evaluate probabilities of events of the form

$$P(a \le \boldsymbol{x} \le b) \;=\; \int_a^b f_{\boldsymbol{x}}(x)\mathrm{d}x$$

i.e., the probability of $\boldsymbol{x}$ assuming values inside the interval $[a,b]$. From Fig. 1.1 we find that the smaller the variance of $\boldsymbol{x}$, the more concentrated its pdf is around its mean.

Figure 1.2 provides similar plots for a random variable $\boldsymbol{x}$ with a Rayleigh distribution, namely, with a pdf given by

$$\boxed{f_{\boldsymbol{x}}(x) \;=\; \frac{x}{\alpha^2}\, e^{-\frac{x^2}{2\alpha^2}}, \qquad x \ge 0, \;\; \alpha > 0} \tag{1.1.3}$$

where α is a positive parameter that determines the mean and the variance of $\boldsymbol{x}$ according to the expressions (see Prob. 1.1):

$$\bar{x} = \alpha\sqrt{\frac{\pi}{2}}, \qquad \sigma_x^2 = \left(2 - \frac{\pi}{2}\right)\alpha^2 \tag{1.1.4}$$

Observe in particular, and in contrast to the Gaussian case, that the mean and variance of a Rayleigh-distributed random variable cannot be chosen independently of each other since they are linked through the parameter α. In Fig. 1.2, the top plot corresponds to $\bar{x} = 1$ and $\sigma_x^2 = 0.2732$, while the bottom plot corresponds to $\bar{x} = 3$ and $\sigma_x^2 = 2.4592$.

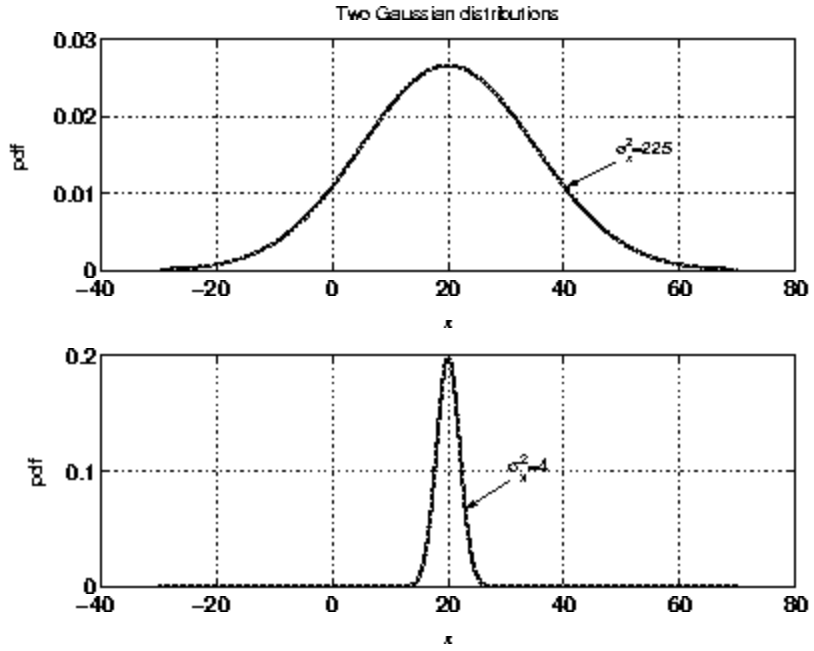

Figure 1.1. The figure shows the plots of the probability density functions of a Gaussian random variable $\boldsymbol{x}$ with mean $\bar{x} = 20$, variance $\sigma_x^2 = 225$ in the top plot, and variance $\sigma_x^2 = 4$ in the bottom plot.

These remarks on the variance of a random variable can be further qualified by invoking a well-known result from probability theory known as Chebyshev's inequality — see Probs. 1.2 and 1.3. The result states that for a random variable $\boldsymbol{x}$ with mean $\bar{x}$ and variance σ_x^2, and for any given scalar $\delta > 0$, it holds that

$$\boxed{P(|\boldsymbol{x} - \bar{x}| \geq \delta) \;\leq\; \sigma_x^2/\delta^2} \tag{1.1.5}$$

That is, the probability that $\boldsymbol{x}$ assumes values outside the interval $(\bar{x} - \delta, \bar{x} + \delta)$ does not exceed σ_x^2/δ^2, with the bound being proportional to the variance of $\boldsymbol{x}$. Hence, for a fixed δ, the smaller the variance of $\boldsymbol{x}$ the smaller the probability that $\boldsymbol{x}$ will assume values outside the interval $(\bar{x} - \delta, \bar{x} + \delta)$. Choose, for instance, $\delta = 5\sigma_x$. Then (1.1.5) gives

$$P(|\boldsymbol{x} - \bar{x}| \geq 5\sigma_x) \;\leq\; 1/25 = 4\%$$

In other words, there is at most 4% chance that $\boldsymbol{x}$ will assume values outside the interval $(\bar{x} - 5\sigma_x, \bar{x} + 5\sigma_x)$.

Actually, the bound that is provided by Chebyshev's inequality is generally not tight. Consider, for example, a zero-mean Gaussian random variable $\boldsymbol{x}$ with variance σ_x^2 and choose $\delta = 2\sigma_x$. Then, from Chebyshev's inequality (1.1.5) we would obtain

$$P(|\boldsymbol{x}| \geq 2\sigma_x) \;\leq\; 1/4 = 25\%$$

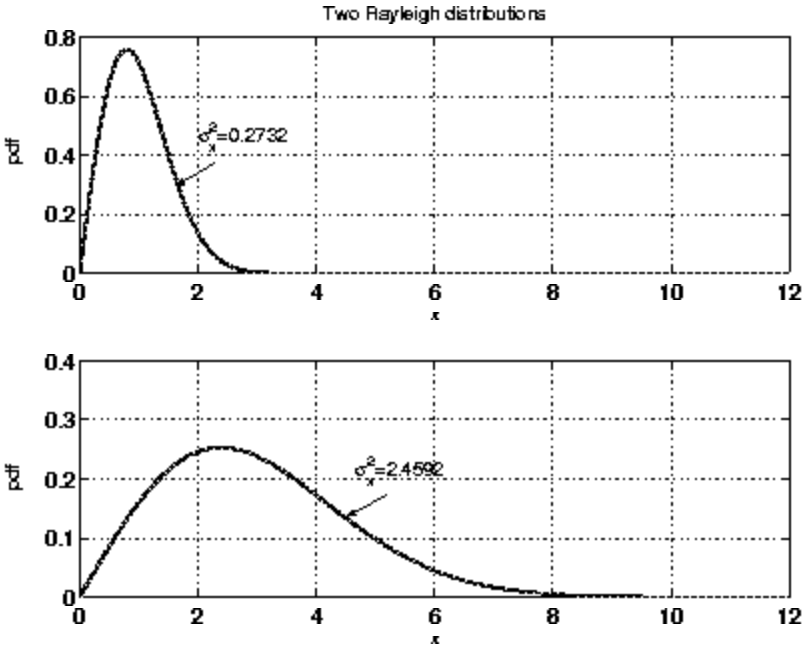

Figure 1.2. The figure shows the plots of the probability density functions of a Rayleigh random variable x with mean $\bar{x} = 1$ and variance $\sigma_x^2 = 0.2732$ in the top plot, and mean $\bar{x} = 3$ and variance $\sigma_x^2 = 2.4592$ in the bottom plot.

whereas direct evaluation of the integral[1]

$$P(|\boldsymbol{x}| \geq 2\sigma_x) \;\stackrel{\Delta}{=}\; 1 - 2\left(\frac{1}{\sqrt{2\pi}\,\sigma_x}\int_0^{2\sigma_x} e^{-\frac{x^2}{2\sigma_x^2}}\,\mathrm{d}x\right)$$

yields

$$P(|\boldsymbol{x}| \geq 2\sigma_x) \;\approx\; 4.56\%$$

Remark 1 [Zero-variance random variables] One useful consequence of Chebyshev's inequality is the following. It allows us to interpret a zero-variance random variable as one that is equal to its mean with probability one. That is,

$$\boxed{\sigma_x^2 = 0 \;\;\Longrightarrow\;\; \boldsymbol{x} = \bar{x} \quad \text{with probability one}}$$

This is because, for any small $\delta > 0$, we obtain from (1.1.5) that

$$P(|\boldsymbol{x} - \bar{x}| \geq \delta) \;\leq\; 0$$

But since the probability of any event is necessarily a nonnegative number, we conclude that $P(|\boldsymbol{x} - \bar{x}| \geq \delta) = 0$, for any $\delta > 0$, so that $\boldsymbol{x} = \bar{x}$ with probability one. We shall call upon this result on several occasions (see, e.g., the proof of Thm. 1.3.2).

$\diamondsuit$

[1] Many books on statistics, and also on digital communications theory, contain tables with the values of such integral expressions in the Gaussian case for different values of δ. In the communications context, such integrals are useful in quantifying the probability of erroneous decisions.

1.2 ESTIMATION GIVEN NO OBSERVATIONS

We now 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 in general 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 so-called *mean-square-error* criterion. It is based on introducing the error signal

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

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.,

$$\min_{\hat{x}} \mathsf{E}\, \tilde{\boldsymbol{x}}^2 \tag{1.2.1}$$

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 1.2.1 (Lack of observations) The least-mean-squares estimate of $\boldsymbol{x}$ given knowledge of only $(\bar{x}, \sigma_x^2)$ 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 (1.2.1). The simplest one perhaps is that the criterion is amenable to mathematical manipulations, more so than any other criterion. In addition, the criterion is in effect attempting to force the estimation error to assume values close to its mean, which happens to be zero since

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

Therefore, by minimizing $\mathsf{E}\,\tilde{x}^2$ we are in effect minimizing the variance of the error. And in view of the discussion in Sec. 1.1 regarding the interpretation of the variance of a random variable, we see that the mean-square-error criterion tries to increase the likelihood of small errors.

The effectiveness of this estimation procedure can be measured by examining the value of the resulting 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 x since the error variable still has the same variance as x itself! We thus find that the performance of the mean-square-error design procedure is rather limited in this case. Of course, 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 x. Note that Lemma 1.2.1 simply tells us that the best we can do, in the absence of any other information about a random variable x, other than its mean and variance, is to use the mean value of 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 x?, the analysis tells us that the best guess is what we would expect for x on average! This is a circular answer, but one that is at least consistent with intuition.

Example 1.2.1 (Binary signal) Assume 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

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

Now given knowledge of $\{\bar{x}, \sigma_x^2\}$ alone, the best estimate for 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 x is 1 or -1 in a given realization. If we could incorporate into the design of the estimator the knowledge that x is a BPSK signal, or some other related information, then we could perhaps come up with a better estimate for x.

$\diamond$

1.3 ESTIMATION GIVEN DEPENDENT OBSERVATIONS

So let us now examine the case in which more is known about a random variable x, other than its mean and variance. Specifically, let us assume that we have access to an observation of a second random variable y that is related to x in some way. For example, y could be a noisy measurement of x, say

$$y = x + v$$

where v denotes the disturbance, or y could be the sign of x, or dependent on x in some other way.

The dependency between two real-valued random variables $\{x, y\}$ is usually described in terms of their *joint* probability density function (pdf). Thus let $f_{x,y}(x, y)$ denote the joint

pdf of $\boldsymbol{x}$ and $\boldsymbol{y}$, this function allows us to evaluate probabilities of events of the form:

$$P(a \leq \boldsymbol{x} \leq b, c \leq \boldsymbol{y} \leq d) = \int_c^d \int_a^b f_{\boldsymbol{x},\boldsymbol{y}}(x,y)\mathrm{d}x\mathrm{d}y$$

namely, the probability that $\boldsymbol{x}$ and $\boldsymbol{y}$ assume values inside the intervals $[a,b]$ and $[c,d]$, respectively. Let also $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$ denote the *conditional* pdf of $\boldsymbol{x}$ given $\boldsymbol{y}$; this function allows us to evaluate probabilities of events of the form

$$P(a \leq \boldsymbol{x} \leq b \mid \boldsymbol{y} = y) = \int_a^b f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)\mathrm{d}x$$

namely, the probability that $\boldsymbol{x}$ assumes values inside the interval $[a,b]$ given that $\boldsymbol{y}$ is fixed at the value y. It is known that the joint and conditional pdfs of two random variables are related via Bayes' rule, which states that

$$\boxed{f_{\boldsymbol{x},\boldsymbol{y}}(x,y) \;=\; f_{\boldsymbol{y}}(y)\; f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \;=\; f_{\boldsymbol{x}}(x)\; f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)} \tag{1.3.1}$$

in terms of the probability density functions of the individual random variables $\boldsymbol{x}$ and $\boldsymbol{y}$.

The variables $\{\boldsymbol{x},\boldsymbol{y}\}$ are said to be *independent* if

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) = f_{\boldsymbol{x}}(x) \quad \text{and} \quad f_{\boldsymbol{y}|\boldsymbol{x}}(y|x) = f_{\boldsymbol{y}}(y)$$

in which case the pdfs of $\boldsymbol{x}$ and $\boldsymbol{y}$ are not modified by conditioning on $\boldsymbol{y}$ and $\boldsymbol{x}$, respectively. Otherwise, the variables are said to be *dependent*. In particular, when the variables are independent, it follows that $\mathsf{E}\,\boldsymbol{x}\boldsymbol{y} = \mathsf{E}\,\boldsymbol{x}\mathsf{E}\,\boldsymbol{y}$. It also follows that independent random variables are uncorrelated, meaning that their cross-correlation is zero as can be verified from the definition of cross-correlation:

$$\sigma_{\boldsymbol{x}\boldsymbol{y}} \;\overset{\Delta}{=}\; \mathsf{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y}) = \mathsf{E}\,\boldsymbol{x}\boldsymbol{y} - \bar{x}\bar{y} = \mathsf{E}\,\boldsymbol{x}\,\mathsf{E}\,\boldsymbol{y} - \bar{x}\bar{y} = 0$$

The converse statement is not true: uncorrelated random variables can be dependent.[2]

Now given two dependent random variables $\{\boldsymbol{x},\boldsymbol{y}\}$, we can 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 $\hat{\boldsymbol{x}}$; 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.,

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

Different occurrences $\boldsymbol{y} = y$ lead to different estimates $\hat{x}$. In Sec. 1.2 we did not need to make this distinction between an estimator and an estimate. 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 $\{\bar{x},\sigma_x^2\}$ and we could only come up with an estimate for $\boldsymbol{x}$.

[2] Consider the following example. Let θ be a random variable that is uniformly distributed over the interval $[0, 2\pi]$. Define the zero-mean random variables $\boldsymbol{x} = \cos\theta$ and $\boldsymbol{y} = \sin\theta$. Then $\boldsymbol{x}^2 + \boldsymbol{y}^2 = 1$ so that $\boldsymbol{x}$ and $\boldsymbol{y}$ are dependent. However, $\mathsf{E}\,\boldsymbol{x}\boldsymbol{y} = \mathsf{E}\cos\theta\sin\theta = 0.5\mathsf{E}\sin 2\theta = 0$, so that $\boldsymbol{x}$ and $\boldsymbol{y}$ are uncorrelated.

1.3.1 Mean-Square-Error Criterion

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

$$\boxed{\tilde{x} \triangleq x - \hat{x}} \tag{1.3.2}$$

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

$$\boxed{\min_{h(\cdot)} \ \mathsf{E}\,\tilde{x}^2} \tag{1.3.3}$$

The solution is given by the following statement.

Theorem 1.3.1 (Optimal mean-square-error estimator) The least-mean-squares estimator (l.m.s.e.) of x given y is the conditional expectation of x given y, i.e., $\hat{x} = \mathsf{E}(x|y)$. The resulting estimate is

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

where $\mathcal{S}_x$ denotes the support (or domain) of the random variable x. Moreover, the estimator is unbiased, i.e., $\mathsf{E}\,\hat{x} = \bar{x}$, and the resulting minimum cost is given by either expression

$$\mathsf{E}\,\tilde{x}^2 = \mathsf{E}\,x^2 \ - \ \mathsf{E}\,\hat{x}^2 \ = \sigma_x^2 \ - \ \sigma_{\hat{x}}^2$$

Proof: There are several ways to establish the result. Our argument is based on recalling that for any two random variables x and y, it holds that (see Prob. 1.4):

$$\boxed{\mathsf{E}\,x \ = \ \mathsf{E}[\mathsf{E}(x|y)]} \tag{1.3.4}$$

where the outermost expectation on the right-hand side is with respect to y, while the innermost expectation is with respect to x. We shall indicate these facts explicitly by showing the variables with respect to which the expectations are performed, so that

$$\mathsf{E}\,x \ = \ \mathsf{E}_y[\mathsf{E}_x(x|y)]$$

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

$$\mathsf{E}_{x,y}\, xg(y) = \mathsf{E}_y\left[\mathsf{E}_x(xg(y)|y)\right] = \mathsf{E}_y\left[\mathsf{E}_x(x|y)g(y)\right] = \mathsf{E}_{x,y}\left[\mathsf{E}_x(x|y)g(y)\right]$$

This means that, for any $g(y)$,

$$\mathsf{E}_{x,y}\left[x - \mathsf{E}_x(x|y)\right] g(y) = 0$$

which we write more compactly as

$$\boxed{\mathsf{E}\left[x - \mathsf{E}(x|y)\right] g(y) = 0} \tag{1.3.5}$$

Expression (1.3.5) states that the random variable $x - \mathsf{E}(x|y)$ is uncorrelated with any function $g(\cdot)$ of y.[3]

[3] As mentioned before, two random variables x and y are uncorrelated if, and only if, their cross-correlation is zero, i.e., $\mathsf{E}(x - \bar{x})(y - \bar{y}) = 0$. On the other hand, the random variables are said to be *orthogonal* if, and only if, $\mathsf{E}\,xy = 0$. 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 (1.3.5) we conclude that the variables $x - \mathsf{E}(x|y)$ and $g(y)$ are orthogonal. However, since $x - \mathsf{E}(x|y)$ is zero mean, then we can also say that they are uncorrelated.

Using this intermediate result, we return to the cost function (1.3.3), add and subtract $\mathsf{E}(x|y)$ to its argument, and express it as

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

The term $\mathsf{E}(x|y)-\hat{x}$ is a function of y. Therefore, if we choose $g(y)=\mathsf{E}(x|y)-\hat{x}$, then from the uncorrelatedness property (1.3.5) we conclude that

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

Only the second term on the right-hand side is dependent on $\hat{x}$ and the m.s.e. is minimized by choosing $\hat{x}=\mathsf{E}(x|y)$.

To evaluate the resulting m.m.s.e. we first note that the optimal estimator is unbiased since

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

so that its variance is given by

$$\sigma_{\hat{x}}^2 = \mathsf{E}\hat{x}^2 \;-\; \bar{x}^2$$

Moreover, in view of the uncorrelatedness property (1.3.5), and in view of the fact that the optimal estimator $\hat{x}=\mathsf{E}(x|y)$ is itself a function of y, we have

$$\boxed{\mathsf{E}(x-\hat{x})\hat{x}=0} \tag{1.3.6}$$

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

$$\begin{aligned}
\mathsf{E}\tilde{x}^2 &= \mathsf{E}[x-\hat{x}][x-\hat{x}] \\
&= \mathsf{E}[x-\hat{x}]x \qquad \text{(because of (1.3.6))} \\
&= \mathsf{E}x^2 \;-\; \mathsf{E}\hat{x}x \\
&= \mathsf{E}x^2 \;-\; \mathsf{E}\hat{x}[\tilde{x}+\hat{x}] \\
&= \mathsf{E}x^2 \;-\; \mathsf{E}\hat{x}^2 \qquad \text{(because of (1.3.6))} \\
&= (\mathsf{E}x^2 \;-\; \bar{x}^2) \;+\; (\bar{x}^2 \;-\; \mathsf{E}\hat{x}^2) \\
&= \sigma_x^2 - \sigma_{\hat{x}}^2
\end{aligned}$$

$\Diamond$

Theorem 1.3.1 tells us that the least-mean-squares estimator of x is its conditional expectation given y. This result is again intuitive. In answer to the question: what is the best guess for x given that we observed y?, the analysis tells us that the best guess is what we would expect for x given the occurrence of y!

Example 1.3.1 (Noisy measurement of a binary signal) Let us return to Ex. 1.2.1, where 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 x, we also have access to a noisy observation of x, say

$$y = x + v$$

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

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

Our intuition tells us that we should be able to do better here than in Ex. 1.2.1. But beware, even here, we shall be able to make some interesting observations.

According to Thm. 1.3.1, the optimal estimate of x given an observation of y is

$$\hat{x} = \mathsf{E}(x|y=y) \;=\; \int_{-\infty}^{\infty} x f_{x|y}(x|y)\mathrm{d}x \tag{1.3.7}$$

We therefore need to determine the conditional pdf, $f_{x|y}(x|y)$, and evaluate the integral (1.3.7). For this purpose, we start by noting that since $y = x + v$, and since x and v are independent, the pdf of y is given by[1]

$$f_y(y) = \frac{1}{2} f_v(y+1) + \frac{1}{2} f_v(y-1) \tag{1.3.8}$$

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

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

Using (1.3.1) we get

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

Substituting into expression (1.3.7) for $\hat{x}$ and integrating we obtain

$$\begin{aligned} \hat{x} \;&=\; \frac{f_v(y-1)}{f_v(y+1)+f_v(y-1)} - \frac{f_v(y+1)}{f_v(y+1)+f_v(y-1)} \\ &=\; \frac{1}{\left(\frac{e^{-(y+1)^2/2}}{e^{-(y-1)^2/2}}\right)+1} - \frac{1}{\left(\frac{e^{-(y-1)^2/2}}{e^{-(y+1)^2/2}}\right)+1} \;=\; \frac{e^{y}-e^{-y}}{e^{y}+e^{-y}} \;\triangleq\; \tanh y \end{aligned}$$

In other words, the least-mean-squares estimator of x is the hyperbolic tangent function,

$$\boxed{\hat{x} = \tanh(y)} \tag{1.3.9}$$

The result is represented schematically in Fig. 1.3.

Figure 1.3. Optimal estimation of a BPSK signal embedded in unit-variance additive Gaussian noise.

Figure 1.4 plots the function $\tanh(y)$. 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

[1] From probability theory, it is known that the pdf of the sum of two independent random variables is equal to the convolution of the individual pdfs, i.e.,

$$f_y(y) = \int_{-\infty}^{\infty} f_x(x) f_v(y-x)\mathrm{d}x$$

In this example,

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

where $\delta(\cdot)$ is the dirac-delta function, so that $f_y(y)$ is given by (1.3.8).

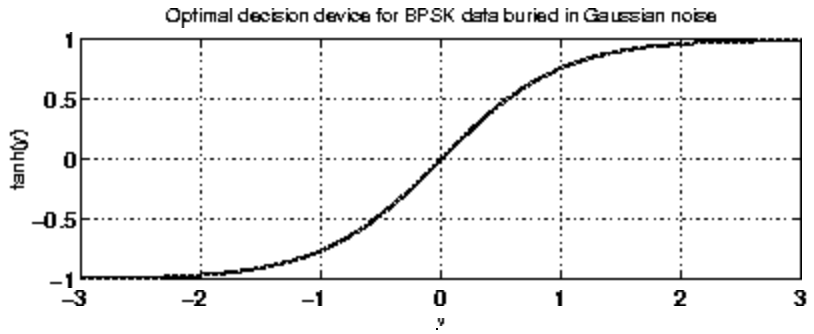

Figure 1.4. A plot of the function $\tanh(y)$.

the designer's perspective. The designer is interested in knowing whether the symbol x is $+1$ or -1 based on the observed value of y. The above construction tells the designer to estimate x by computing $\tanh(y)$. But this value will never be exactly $+1$ or -1; it will be a real number inside 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:

$$\hat{x} \;=\; \text{sign}[\tanh(y)] \tag{1.3.10}$$

where $\text{sign}(\cdot)$ 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's 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:

$$\hat{x} = \text{sign}(y) \tag{1.3.11}$$

where the $\text{sign}(\cdot)$ function operates directly on y rather than on $\tanh(y)$ — see Fig. 1.5. Both suboptimal implementations (1.3.10) and (1.3.11) lead to the same result since, as is evident from Fig. 1.4, $\text{sign}[\tanh(y)] = \text{sign}(y)$. In the computer project at the end of the chapter we shall compare the performance of the optimal and suboptimal estimators (1.3.9)–(1.3.11).[5]

Figure 1.5. Sub-optimal estimation of a BPSK signal embedded in unit-variance additive Gaussian noise.

[5] The purpose of Exs. 1.2.1 and 1.3.1 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 does not need to be perfect or even practical, though it can suggest good practical solutions.

We may mention that in the digital communications literature, especially in studies on equalization methods, an implementation using (1.3.11) is usually said to be based on *hard decisions*, while an implementation using (1.3.9) is said to be based on *soft decisions*.

$\diamondsuit$

Remark 2 [Complexity of optimal estimation] Example 1.3.1 highlights one of the inconveniences of working with the optimal estimator of Thm. 1.3.1. Although the form of the optimal solution is known, 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 Chapter 2 onwards, we shall restrict the class of estimators to *linear* estimators, and study the capabilities of these estimators.

$\diamondsuit$

1.3.2 Orthogonality Principle

There are two important conclusions that follow from the proof of Thm. 1.3.1, namely, the orthogonality properties (1.3.5) and (1.3.6). 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 re-state this result by saying that the estimation error $\tilde{\boldsymbol{x}}$ is orthogonal to any function of $\boldsymbol{y}$,

$$\boxed{\mathsf{E}\,\tilde{\boldsymbol{x}}\,g(\boldsymbol{y}) = 0} \tag{1.3.12}$$

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

$$\boxed{\tilde{\boldsymbol{x}} \;\perp\; g(\boldsymbol{y})} \tag{1.3.13}$$

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. 1.6.

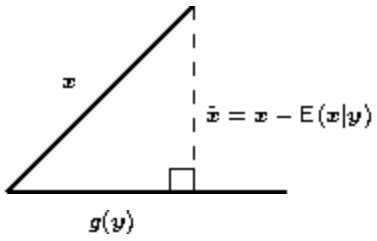

Figure 1.6. The orthogonality condition: $\tilde{\boldsymbol{x}} \perp g(\boldsymbol{y})$.

Relation (1.3.13) 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 in order to reduce the variance of $\tilde{\boldsymbol{x}}$ any further other than the information that has already been extracted by $\hat{\boldsymbol{x}}$. This is because any additional processing of $\boldsymbol{y}$ will remain uncorrelated with $\tilde{\boldsymbol{x}}$.

The second orthogonality property (1.3.6) is a special case of (1.3.13). It states that

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

That is, the estimation error is orthogonal to (or uncorrelated with) the estimator itself. This is a special case of (1.3.13) 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 (1.3.13) is in fact a *defining* property of optimality in the least-mean-squares sense.

Theorem 1.3.2 (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 (1.3.3) 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 (1.3.13) that $\tilde{\boldsymbol{x}} \perp g(\boldsymbol{y})$, for any $g(\cdot)$. Moreover, we know from Thm. 1.3.1 that this estimator is unbiased.

Conversely, assume $\hat{\boldsymbol{x}}$ is an 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 $z = \hat{\boldsymbol{x}} - \mathsf{E}(\boldsymbol{x}|\boldsymbol{y})$ and let us show that it is the zero variable with probability one. For this purpose, we note first that z is zero mean since

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

Moreover, from (1.3.5) 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

$$z \perp g(\boldsymbol{y})$$

which is the same as $\mathsf{E}\,z\,g(\boldsymbol{y}) = 0$. Since the variable z itself is a function of $\boldsymbol{y}$, we choose $g(\boldsymbol{y}) = z$ to get $\mathsf{E}\,z^2 = 0$. We thus find that z is zero mean and has zero variance, so that, from Remark 1 at the end of Section 1.1, we conclude that $z = 0$, or equivalently, $\hat{\boldsymbol{x}} = \mathsf{E}(\boldsymbol{x}|\boldsymbol{y})$, with probability one. ♢

Example 1.3.2 (Suboptimal estimator for a binary signal) Consider again Ex. 1.3.1, where $\boldsymbol{x}$ is a BPSK signal that assumes the values ± 1 with probability 1/2. Let us verify that the estimator $\hat{\boldsymbol{x}} = \mathrm{sign}(\boldsymbol{y})$ is not optimal in the least-mean squares sense. We already know that this is the case because we found in Ex. 1.3.1 that the optimal estimator is $\tanh(\boldsymbol{y})$. Here we wish to verify the sub-optimality of $\mathrm{sign}(\boldsymbol{y})$ without assuming prior knowledge of the optimal estimator, and by relying solely on the orthogonality condition.

According to Thm. 1.3.2, we need to verify that the estimator $\mathrm{sign}(\boldsymbol{y})$ fails the orthogonality test. In particular, we shall exhibit a function $g(\boldsymbol{y})$ such that the difference $\boldsymbol{x} - \mathrm{sign}(\boldsymbol{y})$ is correlated with it. Actually, we shall simply choose $g(\boldsymbol{y}) = \mathrm{sign}(\boldsymbol{y})$ and verify that

$$\mathsf{E}[\boldsymbol{x} - \mathrm{sign}(\boldsymbol{y})]\mathrm{sign}(\boldsymbol{y}) \neq 0 \tag{1.3.14}$$

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

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

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

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

Now recall that x and v are independent and that v is a zero-mean unit-variance Gaussian random variable. Thus let

$$P(v \geq 1) \triangleq \alpha \tag{1.3.15}$$

Then

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

and we obtain

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

Consequently,

$$P(x + v < 0) = 1/2$$

so that

$$\mathsf{E}\ \mathrm{sign}(x + v) = 0$$

This means that the estimator $\hat{x} = \mathrm{sign}(y)$ is unbiased.

We now return to (1.3.14) and note that

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

Therefore, all we need to do in order to verify that (1.3.14) holds is to check that $\mathsf{E}\, x\mathrm{sign}(y)$ is not unity. To do this, we introduce the random variable $z = x\mathrm{sign}(y)$ and proceed to compute its mean.

It is clear from the definition of z that

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

The events

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

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

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

each has probability 0.5α. It then follows that

$$\mathsf{E}\ z = 1 - 2\alpha \neq 1$$

so that $x - \mathrm{sign}(y)$ is correlated with $\mathrm{sign}(y)$. Hence, the estimator $\mathrm{sign}(y)$ does not satisfy the orthogonality condition and, therefore, it cannot be the optimal least-mean-squares estimator.

Let us further compute the enhancement in the signal-to-noise ratio (SNR) that results from the use of this suboptimal estimator. Recall that the variances of x and v are both unity so that, for this example, the SNR prior to the estimation procedure is[6]

$$\mathsf{SNR}_{in} \triangleq 10 \log\left(\frac{\sigma_x^2}{\sigma_v^2}\right) = 0 \text{ dB}$$

After the estimation procedure, the SNR is taken as

$$\mathsf{SNR}_{out} \triangleq 10 \log\left(\frac{\sigma_x^2}{\sigma_{\tilde{x}}^2}\right)$$

where

$$\mathsf{E}\,\tilde{x}^2 = \mathsf{E}[x - \mathrm{sign}(y)]^2 = \sigma_x^2 - 2\,\mathsf{E}\, x\mathrm{sign}(y) + 1 = 2 - 2(1 - 2\alpha) = 4\alpha$$

so that

$$\mathsf{SNR}_{out} = -10 \log(4\alpha)$$

The improvement in SNR is then given by

$$\mathsf{SNR}_{out} - \mathsf{SNR}_{in} = -10 \log(4\alpha)$$

[6] We write $\log(\cdot)$ to refer to the logarithm of a positive number relative to base 10.

Recall from the definition of a in (1.3.15) that it is a number in the interval $(0,1)$ and is given by

$$a = \frac{1}{\sqrt{2\pi}} \int_1^\infty e^{-v^2/2} dv \approx 0.1587$$

The improvement in SNR is therefore approximately 1.9736 dB. This example is pursued further in Probs. 1.5 and 1.6.

$\diamond$

1.3.3 Gaussian Random Variables

We mentioned earlier in Remark 2 prior to Sec. 1.3.2 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. 1.3.1 and as we shall do in another example below). This difficulty will motivate us to limit ourselves in Chapter 2 to the subclass of *linear (or affine) estimators*, namely, to choices of $h(\cdot)$ in (1.3.3) that are *affine* functions of the observation, say $h(\boldsymbol{y}) = a\boldsymbol{y} + b$ for some constants a 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. 1.3.1 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

$$R \triangleq \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$$

where $\{\sigma_x^2, \sigma_y^2\}$ denote the variances of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively,

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

whereas σ_{xy} denotes their cross-correlation

$$\sigma_{xy} \triangleq \mathsf{E}(\boldsymbol{x} - \bar{x})(\boldsymbol{y} - \bar{y})$$

We further assume that R is nonsingular, in fact positive-definite since it can be regarded as the *covariance* matrix of the column vector $\mathrm{col}\{\boldsymbol{x}, \boldsymbol{y}\}$,[7] namely,

$$R = \mathsf{E}\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} - \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}\right)\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} - \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}\right)^{\mathsf{T}} \tag{1.3.16}$$

where the symbol T denotes vector transposition. Every such covariance matrix is necessarily symmetric, $R = R^{\mathsf{T}}$. It is also nonnegative-definite, written as $R \geq 0$ — see the argument prior to Ex. 1.4.2 further ahead.[8] Here we are requiring it to be positive-definite, $R > 0$, and, hence, invertible — see Prob. 1.7.

The joint pdf of $\{\boldsymbol{x}, \boldsymbol{y}\}$ is given by (see App. 1.B for a review of Gaussian random variables and their probability density functions):

$$\boxed{f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = \frac{1}{2\pi}\frac{1}{\sqrt{\det R}} \exp\left\{-\tfrac{1}{2}\begin{bmatrix} x - \bar{x} & y - \bar{y} \end{bmatrix} R^{-1} \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix}\right\}} \tag{1.3.17}$$

[7] The notation $\mathrm{col}\{a, b\}$ denotes a column vector whose entries are a and b.

[8] A symmetric matrix R is said to be nonnegative-definite (written as $R \geq 0$) if, and only if, $a^{\mathsf{T}} R a \geq 0$ for all column vectors a. It is said to be positive-definite (written as $R > 0$) if, and only if, $a^{\mathsf{T}} R a > 0$ for all nonzero column vectors a. Every positive-definite matrix is necessarily invertible — see App. 1.A for a brief review of Hermitian and sign-definite matrices.

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

$$f_{\boldsymbol{x}}(x) = \frac{1}{\sqrt{2\pi}}\,\frac{1}{\sigma_x}\,\exp\{-(x-\bar{x})^2/2\sigma_x^2\}, \qquad f_{\boldsymbol{y}}(y) = \frac{1}{\sqrt{2\pi}}\,\frac{1}{\sigma_y}\,\exp\{-(y-\bar{y})^2/2\sigma_y^2\}$$

According to Thm. 1.3.1, the least-mean-squares estimator of $\boldsymbol{x}$ 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:

$$\begin{aligned} f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) &= \frac{f_{\boldsymbol{x},\boldsymbol{y}}(x,y)}{f_{\boldsymbol{y}}(y)} \\ &= \frac{\dfrac{1}{2\pi}\,\dfrac{1}{\sqrt{\det R}}\,\exp\left\{-\frac{1}{2}\begin{bmatrix} x-\bar{x} & y-\bar{y}\end{bmatrix} R^{-1}\begin{bmatrix} x-\bar{x} \\ y-\bar{y}\end{bmatrix}\right\}}{\dfrac{1}{\sqrt{2\pi}}\,\dfrac{1}{\sigma_y}\,\exp\{-(y-\bar{y})^2/2\sigma_y^2\}} \end{aligned} \tag{1.3.18}$$

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):[9]

$$\boxed{R = \begin{bmatrix} 1 & \sigma_{xy}/\sigma_y^2 \\ 0 & 1\end{bmatrix}\begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma_y^2\end{bmatrix}\begin{bmatrix} 1 & 0 \\ \sigma_{xy}/\sigma_y^2 & 1\end{bmatrix}} \tag{1.3.19}$$

where we introduced the scalar

$$\sigma^2 \;\triangleq\; \sigma_x^2 - \sigma_{xy}^2/\sigma_y^2$$

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.[10]

Now, by inverting both sides of (1.3.19), we find that the inverse of R can be factored as:[11]

$$R^{-1} = \begin{bmatrix} 1 & 0 \\ -\sigma_{xy}/\sigma_y^2 & 1\end{bmatrix}\begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/\sigma_y^2\end{bmatrix}\begin{bmatrix} 1 & -\sigma_{xy}/\sigma_y^2 \\ 0 & 1\end{bmatrix} \tag{1.3.20}$$

This factorization of R^{-1} allows us to express the term

$$\begin{bmatrix} x-\bar{x} & y-\bar{y}\end{bmatrix} R^{-1}\begin{bmatrix} x-\bar{x} \\ y-\bar{y}\end{bmatrix}$$

[9] 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

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

where $\Sigma = A - BC^{-1}B^{\mathsf{T}}$ is called the Schur complement of R with respect to C. The factorization (1.3.19) is a special case of this result where the entries $\{A, B, C\}$ are scalars: $A = \sigma_x^2$, $B = \sigma_{xy}$, and $C = \sigma_y^2$.

[10] The determinant of a positive-definite matrix is positive — see App. 1.A. We see from (1.3.19) that $\det R = \sigma^2\sigma_y^2$, so that σ^2 is necessarily positive.

[11] Here we used the simple fact that for any scalar a,

$$\begin{bmatrix} 1 & 0 \\ a & 1\end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ -a & 1\end{bmatrix}, \qquad \begin{bmatrix} 1 & a \\ 0 & 1\end{bmatrix}^{-1} = \begin{bmatrix} 1 & -a \\ 0 & 1\end{bmatrix}$$

which appears in expression (1.3.18), as a separable sum of two quadratic terms. Indeed, direct calculation using (1.3.20) shows that

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

which allows us to write

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

This equality, along with $\det R = \sigma^2\sigma_y^2$, allows us to simplify expression (1.3.18) for $f_{x|y}(x|y)$ to

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

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:

$$\boxed{\hat{\boldsymbol{x}} = \mathsf{E}(\boldsymbol{x}|\boldsymbol{y}) = \bar{x} + \frac{\sigma_{xy}}{\sigma_y^2}(\boldsymbol{y}-\bar{y})} \tag{1.3.21}$$

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

$$\boxed{\text{m.m.s.e.} \triangleq \sigma_{\tilde{x}}^2 = \sigma_x^2 - \sigma_{\hat{x}}^2 = \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2} = \sigma^2} \tag{1.3.22}$$

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 1.3.3 (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. 1.7 that ρ_{xy} always lies in the interval $[-1,1]$. As ρ_{xy} moves closer to zero, the variables x and y become more uncorrelated (in the Gaussian case, this also means that the variables become less dependent). We see from (1.3.22) 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 (1.3.21), the estimator collapses to $\hat{x}=\bar{x}$. That is, we are reduced to the simple estimator studied in Sec. 1.2. This is expected since in the Gaussian case, a zero cross-correlation means that the random variables x and y are independent so there is no additional information available that we can use to estimate x, besides its mean and variance.

$\Diamond$

Example 1.3.4 (Gaussian noise) Let x denote a Gaussian random variable with mean $\bar{x} = 1$ and variance $\sigma_x^2 = 2$. Similarly, let v denote a Gaussian random variable independent of x, with mean $\bar{v} = 2$ and variance σ_v^2. Now consider the noisy measurement

$$y = 2x + v$$

and let us estimate x from y. According to (1.3.21), 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 x and 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

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

where we used $\mathsf{E}\,x^2 = \sigma_x^2 + \bar{x}^2 = 3$ and $\mathsf{E}\,xv = \mathsf{E}\,x\mathsf{E}\,v = 2$.

Using (1.3.21) we obtain

$$\hat{x} = 1 + \frac{4}{8 + \sigma_v^2}(y - 4)$$

and the resulting m.m.s.e. from (1.3.22) is

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

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

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

Figure 1.7 shows the result of 50 random simulations for two noise variances, $\sigma_v^2 = 0.5$ (top plot) and $\sigma_v^2 = 0.1$ (bottom plot). The dotted lines indicate the values of x during the experiments, while the solid lines indicate the resulting values of $\hat{x}$.

◊

1.4 ESTIMATION IN THE COMPLEX AND VECTOR CASES

We have focused so far in the chapter on *scalar real-valued* random variables. The results however can be extended in a straightforward manner, by using the convenience and power of the vector notation, to the cases of vector-valued and even complex-valued random variables.

These two situations are common in applications. For example, in channel estimation, the quantities to be estimated are the samples of the impulse response sequence of a supposedly finite-impulse-response (FIR) channel. If we group these samples into a vector x, then we are faced with the problem of estimating a vector rather than a scalar quantity. Likewise, in quadrature amplitude modulation (QAM) or in quadrature phase-shift keying (QPSK) transmissions over a communications channel, the transmitted symbols are complex-valued. The recovery of these symbols at the receiver requires that we solve an estimation problem that involves estimating complex-valued quantities.

1.4.1 Complex-Valued Random Variables

Now a complex-valued random variable is one whose real and imaginary parts are *real*-valued random variables themselves, say

$$\boxed{x = x_r + jx_i, \quad j \triangleq \sqrt{-1}}$$

where x_r and x_i denote the real and imaginary parts of x. Therefore, the pdf of a complex-valued random variable x is fully characterized in terms of the joint pdf, $f_{x_r,x_i}(\cdot,\cdot)$, of its

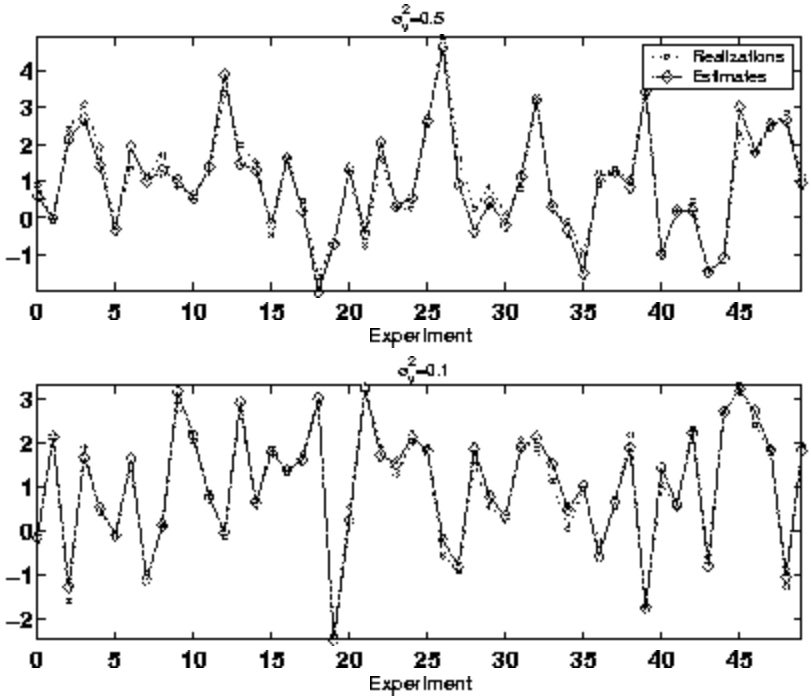

Figure 1.7. Plots of the values of $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ over 50 random experiments with $\sigma_v^2 = 0.5$ (*top figure*) and $\sigma_v^2 = 0.1$ (*bottom figure*) for Ex. 1.3.1. The *dotted* lines with circles correspond to realizations of $\boldsymbol{x}$ while the *solid* lines with diamonds correspond to realizations of the estimator $\hat{\boldsymbol{x}}$.

real and imaginary parts. This means that we can treat a complex random variable as a function of two real random variables. The mean of $\boldsymbol{x}$ will then be defined as

$$\mathsf{E}\,\boldsymbol{x} \stackrel{\Delta}{=} \mathsf{E}\,\boldsymbol{x}_{\mathrm{r}} + j\mathsf{E}\,\boldsymbol{x}_{\mathrm{i}} = \bar{x}_{\mathrm{r}} + j\bar{x}_{\mathrm{i}}$$

in terms of the means of its real and imaginary parts. Its variance, however, will be defined as

$$\boxed{\sigma_x^2 \stackrel{\Delta}{=} \mathsf{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^* = \mathsf{E}\,|\boldsymbol{x}-\bar{x}|^2} \tag{1.4.1}$$

where the symbol $*$ denotes complex conjugation. Comparing with the definition (1.1.1) in the real case, we see that the above definition is different because of the use of conjugation (in the real case, the conjugate of $(\boldsymbol{x}-\bar{x})$ is $(\boldsymbol{x}-\bar{x})$ itself and the above definition collapses to (1.1.1)). The use of the conjugate term in (1.4.1) is necessary in order to guarantee that σ_x^2 will be a nonnegative real number. In particular, it is immediate to verify from (1.4.1) that

$$\sigma_x^2 = \sigma_{x_{\mathrm{r}}}^2 + \sigma_{x_{\mathrm{i}}}^2$$

in terms of the individual variances of $\boldsymbol{x}_{\mathrm{r}}$ and $\boldsymbol{x}_{\mathrm{i}}$.

We shall say that two complex-valued random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are uncorrelated if, and only if, their cross-correlation is zero, i.e., if

$$\sigma_{xy} \stackrel{\Delta}{=} \mathsf{E}(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y})^* = 0$$

On the other hand, we shall say that they are *orthogonal* if, and only if,

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

It can be immediately verified that the concepts of orthogonality and uncorrelatedness coincide if at least one of the random variables is zero mean.

Example 1.4.1 (QPSK constellation) Consider a signal x that is chosen uniformly from a QPSK constellation, i.e., x assumes any of the values

$$\pm\frac{\sqrt{2}}{2} \pm j\frac{\sqrt{2}}{2}$$

with probability 1/4 (see Fig. 1.8). Clearly, x is a complex-valued random variable; its mean and variance are easily found to be $\bar{x} = 0$ and $\sigma_x^2 = 1$.

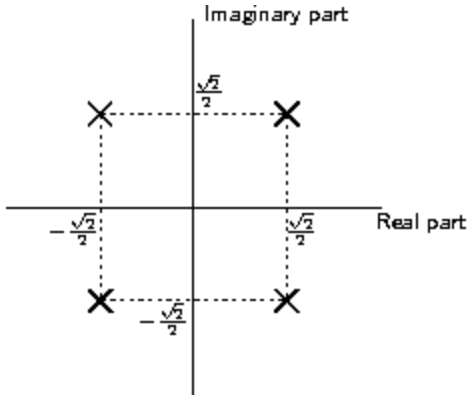

Figure 1.8. A QPSK constellation.

◊

1.4.2 Vector-Valued Random Variables

A vector-valued random variable, on the other hand, is a collection (in column or row vector forms) of random variables. The individual entries can be real or complex-valued themselves. For example, if $\boldsymbol{x} = \mathrm{col}\{\boldsymbol{x}(0), \boldsymbol{x}(1)\}$ is a random vector with entries $\{\boldsymbol{x}(0), \boldsymbol{x}(1)\}$,[12] then we shall define its *mean* as the vector of individual means,

$$\mathsf{E}\,\boldsymbol{x} = \begin{bmatrix} \bar{x}(0) \\ \bar{x}(1) \end{bmatrix}$$

and its covariance *matrix* as

$$\boxed{R_x \triangleq \mathsf{E}\,(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^*} \tag{1.4.2}$$

where the symbol $*$ now denotes complex-conjugate transposition (i.e., we transpose the vector and then replace each of its entries by the corresponding conjugate value).[13] Comparing with the definition (1.3.16) in the real-valued case, we see that the symbol $*$ replaces

[12] We use parenthesis to index the scalar entries of a vector, e.g., $\boldsymbol{x}(k)$ denotes the k–th entry of $\boldsymbol{x}$.

[13] We may mention in passing that if $\boldsymbol{x}$ were a *row* random vector, rather than a *column* random vector, then its covariance matrix would be defined as

$$R_x \triangleq \mathsf{E}\,(\boldsymbol{x}-\bar{x})^*(\boldsymbol{x}-\bar{x})$$

with the conjugate term coming first. This is because in this case, it is the product $(\boldsymbol{x}-\bar{x})^*(\boldsymbol{x}-\bar{x})$ that yields a matrix, while the product $(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^*$ would be a scalar.

the transposition symbol T. Moreover, we also see that R_x is not symmetric anymore but rather *Hermitian*. That is, R_x satisfies

$$R_x = R_x^*$$

For the above two-element vector $\boldsymbol{x}$ we obtain

$$R_x = \begin{bmatrix} \mathsf{E}\,|\boldsymbol{x}(0)-\bar{x}(0)|^2 & \mathsf{E}\,[\boldsymbol{x}(0)-\bar{x}(0)][\boldsymbol{x}(1)-\bar{x}(1)]^* \\ \mathsf{E}\,[\boldsymbol{x}(1)-\bar{x}(1)][\boldsymbol{x}(0)-\bar{x}(0)]^* & \mathsf{E}\,|\boldsymbol{x}(1)-\bar{x}(1)|^2 \end{bmatrix}$$

with the individual variances of the variables $\{\boldsymbol{x}(0), \boldsymbol{x}(1)\}$ appearing on the diagonal and the cross-correlations between them appearing on the off-diagonal entries. In the zero-mean case, the definition of R_x, and the above expression, simplify to

$$\boxed{R_x \triangleq \mathsf{E}\,\boldsymbol{x}\boldsymbol{x}^*}$$

and

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

We mentioned before, following the definition (1.3.16), that the covariance matrix of a random vector is always nonnegative-definite.[14] In order to verify this claim in the general complex and vector-valued case, we introduce the scalar-valued random variable $y = a^*(\boldsymbol{x} - \bar{x})$, where a is an arbitrary column vector. Then y has zero mean and

$$\sigma_y^2 = \mathsf{E}\,|y|^2 = a^* R_x a$$

But since the variance of any scalar-valued random variable is always nonnegative, we conclude that $a^* R_x a \geq 0$ for any a. This means that R_x is nonnegative definite, as claimed.

Example 1.4.2 (Transmissions over a noisy channel) Consider the setting of Fig. 1.9. A sequence of independent and identically distributed (i.i.d.) symbols $\{s(i)\}$ is transmitted over an initially relaxed FIR channel with transfer function $C(z) = 1 + 0.5z^{-1}$, where z^{-1} denotes the unit-time delay in the $z-$transform domain. Each symbol is either $+1$ with probability p or -1 with probability $1-p$. The output of the channel is corrupted by zero-mean additive white[15] Gaussian noise $v(i)$ of unit variance, i.e.,

$$\mathsf{E}\,v(i)v^*(j) = \delta_{ij}$$

where δ_{ij} denotes the Kronecker delta function that is equal to unity when $i = j$ and zero otherwise. The noise and the symbols are assumed independent.

The output of the channel at any specific time instant i is given by

$$y(i) = s(i) + 0.5s(i-1) + v(i)$$

Assume we collect $N+1$ measurements, $\{y(i), i = 0, 1, \ldots, N\}$, into a column vector y, and then pose the problem of recovering the transmitted symbols $\{s(i), i = 0, 1, \ldots, N\}$ over the same interval of time. If we collect the symbols $\{s(i)\}$ into a column vector $\boldsymbol{x}$ as well,

$$\boldsymbol{x} \triangleq \mathrm{col}\{s(0), s(1), \ldots, s(N)\}$$

[14] We defined in an earlier footnote what we mean by a nonnegative-definite matrix. There however we were dealing with real-valued matrices. In the general complex-valued case, we say that a Hermitian matrix R is nonnegative definite if, and only if, $a^* R a \geq 0$ for any column vector a (real or complex-valued). We say that R is positive-definite if, and only if, $a^* R a > 0$ for any nonzero column vector a.

[15] A random process $v(i)$ is white if $\mathsf{E}\,v(i)v^*(j) = 0$ for all $i \neq j$; i.e., if its terms are uncorrelated with each other.

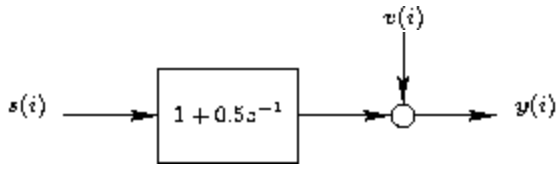

Figure 1.9. Transmission over an additive white Gaussian FIR channel.

then we are faced with the problem of estimating a vector $\boldsymbol{x}$ from a vector $\boldsymbol{y}$. Here the entries of $\boldsymbol{x}$ and $\boldsymbol{y}$ are all real-valued. If the symbols $s(i)$ were instead chosen from a QPSK constellation, then both $\boldsymbol{x}$ and $\boldsymbol{y}$ will be complex-valued. We shall return to this problem further ahead in Ex. 1.4.4 and also in Chapter 2 (see Ex. 2.2.3).

$\diamondsuit$

1.4.3 Optimal Estimator in the Vector Case

It turns out that the optimal estimator in the general vector and complex-valued case is still given by the conditional expectation of $\boldsymbol{x}$ given $\boldsymbol{y}$. To see this, let us start with a special case.

Assume $\boldsymbol{x}$ and $\boldsymbol{y}$ are both real-valued with $\boldsymbol{x}$ a *scalar* and $\boldsymbol{y}$ a vector, say

$$\boldsymbol{y} = \mathrm{col}\{\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)\}$$

As before, let $\hat{\boldsymbol{x}} = h(\boldsymbol{y})$ denote an estimator for $\boldsymbol{x}$. Since $\boldsymbol{y}$ is vector-valued, the function $h(\cdot)$ operates on the entries of $\boldsymbol{y}$ and provides a real scalar quantity as a result. More explicitly, we write

$$\hat{\boldsymbol{x}} = h[\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)]$$

The function $h(\cdot)$ is to be chosen optimally by minimizing the variance of the error $\tilde{\boldsymbol{x}} = \boldsymbol{x} - \hat{\boldsymbol{x}}$, i.e., by solving

$$\min_{h(\cdot)} \mathsf{E}\, \tilde{\boldsymbol{x}}^2$$

The same argument that we used to establish Thm. 1.3.1 can be repeated here to verify that the optimal estimator is still given by

$$\hat{\boldsymbol{x}} \;=\; \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y}) \;=\; \mathsf{E}[\boldsymbol{x}|\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)] \tag{1.4.3}$$

The only difference between this result and that of Thm. 1.3.1 is that the conditional expectation is now computed relative to a collection of random variables $\{\boldsymbol{y}(i)\}$, rather than a single random variable. Moreover, the orthogonality condition (1.3.5) extends to this case and is still given by

$$\mathsf{E}\,[\boldsymbol{x} - \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})]\, g(\boldsymbol{y}) = 0 \tag{1.4.4}$$

for any function $g(\cdot)$ of $\boldsymbol{y}$.

Example 1.4.3 (Various noisy measurements of a BPSK signal) Let us return to Ex. 1.3.1, where $\boldsymbol{x}$ is a BPSK signal that is either $+1$ or -1 with probability $1/2$ each. Assume that we collect two noisy measurements $y(0)$ and $y(1)$ of $\boldsymbol{x}$, say

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

where $\{v(0), v(1)\}$ 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 $y(0)$, it is also $+1$ in the measurement $y(1)$, and similarly for -1.) For instance, we could interpret $\{y(0), y(1)\}$ as the noisy signals measured at two antennas as a result of transmitting x over two additive Gaussian-noise channels — see Fig. 1.10.

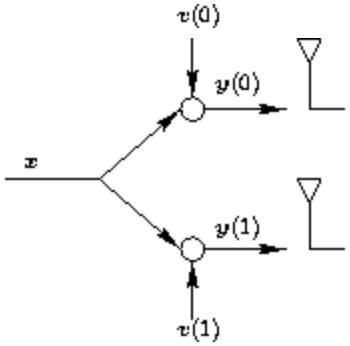

Figure 1.10. Reception by two antennas of a symbol x transmitted over two additive Gaussian-noise channels.

We can then pose the problem of estimating x given *both* measurements $\{y(0), y(1)\}$. According to (1.4.3), the solution is given by

$$\hat{x} = \mathsf{E}\,[x|y(0), y(1)]$$

The evaluation of the conditional expectation in this case is a trivial extension of the derivation given in Ex. 1.3.1, and it is left as an exercise to the reader — see Prob. 1.11, where the more general case of multiple measurements is treated. The result of that problem shows that

$$\hat{x} = \tanh[y(0) + y(1)]$$

In the context of the two-antenna example of Fig. 1.10, this result leads to the optimal receiver structure shown in Fig. 1.11.

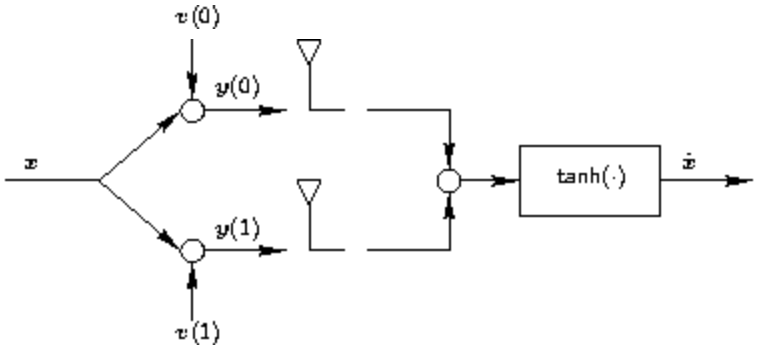

Figure 1.11. Optimal receiver structure for recovering a symbol x from two separate measurements over additive Gaussian-noise channels.

◊

Let us now study the general case and determine the form of the optimal estimator for a *vector-valued* random variable x given another vector-valued random variable y, with both variables allowed to be complex-valued as well. Thus assume that x is $p-$dimensional while y is $q-$dimensional.

Again, let $\hat{\boldsymbol{x}} = h(\boldsymbol{y})$ denote an estimator for $\boldsymbol{x}$. Since $\boldsymbol{x}$ and $\boldsymbol{y}$ are vector-valued, the function $h(\cdot)$ operates on the entries of $\boldsymbol{y}$ and provides a vector quantity as a result. More explicitly, we can write for the individual entries of $\hat{\boldsymbol{x}}$ and $\boldsymbol{y}$,

$$\begin{bmatrix} \hat{\boldsymbol{x}}(0) \\ \hat{\boldsymbol{x}}(1) \\ \hat{\boldsymbol{x}}(2) \\ \vdots \\ \hat{\boldsymbol{x}}(p-1) \end{bmatrix} = \begin{bmatrix} h_0[\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)] \\ h_1[\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)] \\ h_2[\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)] \\ \vdots \\ h_{p-1}[\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(q-1)] \end{bmatrix}$$

where the $\{h_k(\cdot)\}$ represent the individual mappings from the observation vector $\boldsymbol{y}$ to the estimators $\{\hat{\boldsymbol{x}}(k)\}$. We can then seek optimal functions $\{h_k(\cdot)\}$ that minimize the variance of the error in each component of $\boldsymbol{x}$, namely, each $h_k(\cdot)$ is determined by solving

$$\boxed{\min_{h_k(\cdot)} \quad \mathsf{E}\,|\tilde{\boldsymbol{x}}(k)|^2} \tag{1.4.5}$$

where

$$\tilde{\boldsymbol{x}}(k) \;\triangleq\; \boldsymbol{x}(k) - h_k(\boldsymbol{y})$$

This formulation is also equivalent to solving over all $\{h_k(\cdot)\}$ the following problem:

$$\boxed{\min_{\{h_k(\cdot)\}} \quad \mathsf{E}\;\tilde{\boldsymbol{x}}^*\tilde{\boldsymbol{x}}} \tag{1.4.6}$$

This is because the quantity $\mathsf{E}\,\tilde{\boldsymbol{x}}^*\tilde{\boldsymbol{x}}$ is the sum of the individual terms $\mathsf{E}\,|\tilde{\boldsymbol{x}}(k)|^2$,

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

with each term $\mathsf{E}\,|\tilde{\boldsymbol{x}}(k)|^2$ depending only on the corresponding function $h_k(\cdot)$. In this way, minimizing the sum $\mathsf{E}\,\tilde{\boldsymbol{x}}^*\tilde{\boldsymbol{x}}$ over all $\{h_k(\cdot)\}$ is equivalent to minimizing each individual term, $\mathsf{E}\,|\tilde{\boldsymbol{x}}(k)|^2$, over its $h_k(\cdot)$. Note further that

$$\mathsf{E}\,\tilde{\boldsymbol{x}}^*\tilde{\boldsymbol{x}} \;=\; \mathsf{Tr}(\mathsf{E}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^*) \;=\; \mathsf{Tr}\,(R_{\tilde{\boldsymbol{x}}})$$

That is, the scalar quantity $\mathsf{E}\,\tilde{\boldsymbol{x}}^*\tilde{\boldsymbol{x}}$ in (1.4.6) is equal to the trace of the error covariance matrix $R_{\tilde{\boldsymbol{x}}}$,[16] so that (1.4.6) is also equivalent to solving over all $\{h_k(\cdot)\}$:

$$\boxed{\min_{\{h_k(\cdot)\}} \quad \mathsf{Tr}(R_{\tilde{\boldsymbol{x}}})} \tag{1.4.7}$$

Now the solution to the general problem (1.4.5) follows from the special case discussed at the beginning of this section. Indeed, if we express $\boldsymbol{x}(k)$ and $h_k(\cdot)$ in terms of their real and imaginary parts, say

$$\boldsymbol{x}(k) \;\triangleq\; \boldsymbol{x}_{\mathrm{r}}(k) + j\;\boldsymbol{x}_{\mathrm{i}}(k), \qquad h_k \;\triangleq\; h_{\mathrm{r},k} + j\;h_{\mathrm{i},k}$$

then we can expand the error criterion as

$$\mathsf{E}\,|\boldsymbol{x}(k) - h_k(\boldsymbol{y})|^2 = \mathsf{E}\,[\boldsymbol{x}_{\mathrm{r}}(k) - h_{\mathrm{r},k}(\boldsymbol{y})]^2 \;+\; \mathsf{E}[\boldsymbol{x}_{\mathrm{i}}(k) - h_{\mathrm{i},k}(\boldsymbol{y})]^2$$

[16] The trace of a matrix is equal to the sum of its diagonal elements. Moreover, for any column vector a, it holds that $a^*a = \mathsf{Tr}(aa^*)$.

and we are reduced to minimizing the sum of two nonnegative quantities over the unknowns $\{h_{r,k}(\cdot), h_{i,k}(\cdot)\}$. This is equivalent to minimizing each term separately,

$$\min_{h_{r,k}(\cdot)} \mathsf{E}\,[\boldsymbol{x}_r(k) - h_{r,k}(\boldsymbol{y})]^2, \qquad \min_{h_{i,k}(\cdot)} \mathsf{E}\,[\boldsymbol{x}_i(k) - h_{i,k}(\boldsymbol{y})]^2$$

and the solution we already know from (1.4.3) to be given by

$$\begin{aligned}\hat{\boldsymbol{x}}_r(k) &= \mathsf{E}[\boldsymbol{x}_r(k)|\boldsymbol{y}_r(0), \boldsymbol{y}_i(0), \boldsymbol{y}_r(1), \boldsymbol{y}_i(1), \ldots, \boldsymbol{y}_r(q-1), \boldsymbol{y}_i(q-1)] \\ \hat{\boldsymbol{x}}_i(k) &= \mathsf{E}[\boldsymbol{x}_i(k)|\boldsymbol{y}_r(0), \boldsymbol{y}_i(0), \boldsymbol{y}_r(1), \boldsymbol{y}_i(1), \ldots, \boldsymbol{y}_r(q-1), \boldsymbol{y}_i(q-1))\end{aligned}$$

Therefore, the optimal choice for $h_k(\cdot)$ is

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

so that the optimal estimator that minimizes the variances of the individual errors $\{\tilde{\boldsymbol{x}}(k)\}$ is

$$\boxed{\hat{\boldsymbol{x}} = \mathsf{E}(\boldsymbol{x}|\boldsymbol{y}) \triangleq \begin{bmatrix} \mathsf{E}[\boldsymbol{x}(0)|\boldsymbol{y}] \\ \mathsf{E}[\boldsymbol{x}(1)|\boldsymbol{y}] \\ \vdots \\ \mathsf{E}[\boldsymbol{x}(p-1)|\boldsymbol{y}] \end{bmatrix}} \tag{1.4.8}$$

Likewise, using the property (1.3.5) of conditional expectations, we conclude that the orthogonality condition in this case is still given by

$$\mathsf{E}\,[\boldsymbol{x} - \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})]g(\boldsymbol{y}) = 0 \tag{1.4.9}$$

for any function $g(\cdot)$ of the observation $\boldsymbol{y}$.

Theorem 1.4.1 (Optimal estimation in the vector case) The least-mean-squares estimator of a (possibly complex-valued) vector $\boldsymbol{x}$ given another (possibly complex-valued) vector $\boldsymbol{y}$ is still the conditional expectation of $\boldsymbol{x}$ given $\boldsymbol{y}$, i.e., $\hat{\boldsymbol{x}} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$. This estimator solves

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

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

Example 1.4.4 (Estimation of transmitted symbols) Consider again the setting of Ex. 1.4.2 and assume $N = 2$, so that we are interested in estimating the vector $\boldsymbol{x} = \mathrm{col}\{s(0), s(1)\}$ from the observation $\boldsymbol{y} = \mathrm{col}\{y(0), y(1)\}$, where

$$y(0) = s(0) + v(0), \qquad y(1) = s(1) + 0.5s(0) + v(1)$$

Here we are assuming that transmissions start at time 0 so that $s(-1) = 0$. If we introduce the 2×2 matrix

$$H = \begin{bmatrix} 1 & 0 \\ 0.5 & 1 \end{bmatrix}$$

and the vector $v = \mathrm{col}\{v(0), v(1)\}$, then the above equations can be written more compactly in matrix form as follows:

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

We are therefore faced with the problem of estimating x from y, with the noise term v assumed independent of x. Now recall that the symbols $s(i)$ are either $+1$ or -1 with probabilities p or $1-p$, respectively. Hence, the vector x can assume any of four values:

$$x \in \left\{ \begin{bmatrix} +1 \\ +1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} +1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ +1 \end{bmatrix} \right\} \triangleq \{m_0, m_1, m_2, m_3\}$$

with probabilities

$$\{p^2,\ (1-p)^2,\ p(1-p),\ p(1-p)\}$$

respectively. Observe that we are denoting the four possibilities of x by $\{m_i, i = 0, \ldots, 3\}$ for compactness of notation. Let also $q = 1-p$.

Moreover, the pdf of v is Gaussian and given by

$$f_v(v) = \frac{1}{2\pi} \exp\{-v^T v/2\}$$

since the covariance matrix of v is assumed to be the identity matrix. It then follows that the pdf of y is given by

$$f_y(y) = p^2 f_v(y - Hm_0) + q^2 f_v(y - Hm_1) + pq f_v(y - Hm_2) + pq f_v(y - Hm_3)$$

Similarly, we obtain, as in Ex. 1.3.1, that

$$\begin{aligned} f_{x,y}(x,y) &= f_x(x) \cdot f_v(y - Hx) \\ &= p^2 f_v(y - Hm_0)\delta(x - m_0) + q^2 f_v(y - Hm_1)\delta(x - m_1) \\ &\quad + pq f_v(y - Hm_2)\delta(x - m_2) + pq f_v(y - Hm_3)\delta(x - m_3) \end{aligned}$$

The expressions so derived for $f_y(y)$ and $f_{x,y}(x,y)$ allow us to evaluate $f_{x|y}(x|y)$, from which we can evaluate the desired conditional expectation $\mathsf{E}(x|y)$ and, consequently, $\{\hat{s}(0), \hat{s}(1)\}$. This final computation is left as an exercise to the reader — see Prob. 1.12.

$\diamond$

1.4.4 Equivalent Optimization Criterion

A useful fact to highlight here is that the optimal estimator $\mathsf{E}(x|y)$ defined by (1.4.8), which solves problems (1.4.5)–(1.4.7), is also the optimal solution of another related *matrix-valued* error criterion (cf. (1.4.11) further ahead), as we now explain.

Thus consider the following alternative formulation. Assume that we pose the problem of estimating x from y by requiring that the functions $\{h_k(\cdot)\}$ be such that they minimize the variance of any *arbitrary* linear combination of the entries of the error vector, say $a^*(x - h(y))$ for any a. That is, assume we replace the optimization problem (1.4.5) by the alternative problem

$$\min_{\{h_k(\cdot)\}} \mathsf{E}|a^* \tilde{x}|^2, \quad \text{for any column vector } a \tag{1.4.10}$$

The error vector $\tilde{x}$ is dependent on the choice of h and, therefore, the covariance matrix $\mathsf{E}\,\tilde{x}\tilde{x}^*$ is also dependent on h. Let us indicate this fact explicitly by writing

$$R_{\tilde{x}}(h) \triangleq \mathsf{E}\,\tilde{x}\tilde{x}^*$$

Now note that

$$\mathsf{E}|a^* \tilde{x}|^2 = a^* R_{\tilde{x}}(h) a$$

so that problem (1.4.10) is in effect seeking an optimal function h^o such that, for any vector a and for any other h,

$$a^* R_{\tilde{x}}(h) a \geq a^* R_{\tilde{x}}(h^o) a$$

That is, the difference matrix $R_{\tilde{x}}(h) - R_{\tilde{x}}(h^o)$ should be nonnegative-definite for all h. For this reason, we can equivalently interpret (1.4.10) as the problem of minimizing the error covariance matrix $R_{\tilde{x}}$ itself, written as

$$\boxed{\min_{h(\cdot)} \quad \mathsf{E}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^*} \tag{1.4.11}$$

Comparing with (1.4.6) we see that we are replacing the scalar $\mathsf{E}\,\tilde{\boldsymbol{x}}^*\tilde{\boldsymbol{x}}$ by the matrix $\mathsf{E}\,\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^*$.

Let us now verify that the solution to (1.4.10), or equivalently (1.4.11), is again $h^o(\boldsymbol{y}) = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$. For this purpose, we recall that, for any $h(\boldsymbol{y})$,

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

so that the covariance matrix $R_{\tilde{x}}(h)$ is given by

$$\begin{aligned} R_{\tilde{x}}(h) &= \mathsf{E}\,[\boldsymbol{x} - h(\boldsymbol{y})][\boldsymbol{x} - h(\boldsymbol{y})]^* \\ &= \mathsf{E}\,\boldsymbol{x}\boldsymbol{x}^* - \mathsf{E}\,\boldsymbol{x}h^*(\boldsymbol{y}) - \mathsf{E}\,h(\boldsymbol{y})\boldsymbol{x}^* + \mathsf{E}\,h(\boldsymbol{y})h^*(\boldsymbol{y}) \end{aligned}$$

We now verify that

$$R_{\tilde{x}}(h) - R_{\tilde{x}}(h^o) \geq 0$$

for any choice of h. Indeed, from the orthogonality property (1.3.5), we have that $\boldsymbol{x} - h^o(\boldsymbol{y})$ is uncorrelated with any function of $\boldsymbol{y}$. Hence,

$$R_{\tilde{x}}(h^o) = \mathsf{E}\,[\boldsymbol{x} - h^o(\boldsymbol{y})][\boldsymbol{x} - h^o(\boldsymbol{y})]^* \;=\; \mathsf{E}\,[\boldsymbol{x} - h^o(\boldsymbol{y})]\boldsymbol{x}^* \;=\; \mathsf{E}\,\boldsymbol{x}\boldsymbol{x}^* - \mathsf{E}\,h^o(\boldsymbol{y})\boldsymbol{x}^*$$

Subtracting from $R_{\tilde{x}}(h)$ leads to

$$R_{\tilde{x}}(h) - R_{\tilde{x}}(h^o) \;=\; -\mathsf{E}\,\boldsymbol{x}h^*(\boldsymbol{y}) - \mathsf{E}\,h(\boldsymbol{y})\boldsymbol{x}^* + \mathsf{E}\,h(\boldsymbol{y})h^*(\boldsymbol{y}) + \mathsf{E}\,h^o(\boldsymbol{y})\boldsymbol{x}^*$$

From the orthogonality property (1.3.5) we again have that

$$\mathsf{E}\,[\boldsymbol{x} - h^o(\boldsymbol{y})]h^{o*}(\boldsymbol{y}) = 0, \quad \mathsf{E}\,[\boldsymbol{x} - h^o(\boldsymbol{y})]h^*(\boldsymbol{y}) = 0$$

so that

$$\mathsf{E}\,\boldsymbol{x}h^{o*}(\boldsymbol{y}) = \mathsf{E}\,h^o(\boldsymbol{y})h^{o*}(\boldsymbol{y}) \qquad \text{and} \qquad \mathsf{E}\,\boldsymbol{x}h^*(\boldsymbol{y}) = \mathsf{E}\,h^o(\boldsymbol{y})h^*(\boldsymbol{y})$$

These two equalities allow us to rewrite the difference $R_{\tilde{x}}(h) - R_{\tilde{x}}(h^o)$ as a perfect square:

$$R_{\tilde{x}}(h) - R_{\tilde{x}}(h^o) = \mathsf{E}\,[h^o(\boldsymbol{y}) - h(\boldsymbol{y})][h^o(\boldsymbol{y}) - h(\boldsymbol{y})]^*$$

The right-hand side is nonnegative-definite for all h, as claimed. Finally, since the cost used in (1.4.6) is simply the trace of the error covariance matrix, we conclude that minimizing the error covariance matrix is equivalent to minimizing its trace.

Lemma 1.4.1 (Cost function) The conditional expectation of $\boldsymbol{x}$ given $\boldsymbol{y}$ is optimal relative to either cost

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

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

1.4.5 Spherically Invariant Gaussian Variables

We saw earlier in Sec. 1.3.3 that for *scalar* real-valued Gaussian random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$, the optimal estimator of $\boldsymbol{x}$ given $\boldsymbol{y}$ depends in an affine manner on the observation $\boldsymbol{y}$. The same conclusion holds in the general *vector* complex-valued case.

So assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly Gaussian random *vector* variables with a nonsingular covariance matrix

$$R \triangleq \begin{bmatrix} R_x & R_{xy} \\ R_{yx} & R_y \end{bmatrix}$$

where

$$R_x = \mathsf{E}(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^*, \quad R_y = \mathsf{E}(\boldsymbol{y}-\bar{y})(\boldsymbol{y}-\bar{y})^*$$

and

$$R_{xy} = \mathsf{E}(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y})^* = R_{yx}^*$$

The variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ are assumed to be complex-valued with dimensions $p \times 1$ for $\boldsymbol{x}$ and $q \times 1$ for $\boldsymbol{y}$.

If $\boldsymbol{x}$ and $\boldsymbol{y}$ were *real-valued*, then their individual probability density functions, as well as their joint pdf, would be given by (see App. 1.B):

$$f_{\boldsymbol{x}}(x) = \frac{1}{\sqrt{(2\pi)^p}} \frac{1}{\sqrt{\det R_x}} \exp\left\{-\tfrac{1}{2}(x-\bar{x})^{\mathsf{T}} R_x^{-1}(x-\bar{x})\right\}$$

$$f_{\boldsymbol{y}}(x) = \frac{1}{\sqrt{(2\pi)^q}} \frac{1}{\sqrt{\det R_y}} \exp\left\{-\tfrac{1}{2}(y-\bar{y})^{\mathsf{T}} R_y^{-1}(y-\bar{y})\right\}$$

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = \frac{1}{\sqrt{(2\pi)^{p+q}}} \frac{1}{\sqrt{\det R}} \exp\left\{-\tfrac{1}{2}\begin{bmatrix} (x-\bar{x})^{\mathsf{T}} & (y-\bar{y})^{\mathsf{T}} \end{bmatrix} R^{-1} \begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix}\right\}$$

In particular, observe that if $\boldsymbol{x}$ and $\boldsymbol{y}$ were *uncorrelated*, i.e., if $R_{xy} = 0$, then the covariance matrix R becomes block diagonal, with entries $\{R_x, R_y\}$, and it is straightforward to verify from the above pdf expressions that in this case

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) = f_{\boldsymbol{x}}(x) \cdot f_{\boldsymbol{y}}(y)$$

In other words, uncorrelated real-valued Gaussian random variables are also independent.

When, on the other hand, $\boldsymbol{x}$ and $\boldsymbol{y}$ are *complex-valued*, they need to satisfy two conditions in order for their individual and joint pdfs to have forms similar to the above in the Gaussian case. These conditions are known as *circularity* assumptions, and the need for them is explained in App. 1.B. The conditions are as follows. Each variable is required to be *circular*, meaning that $\{\boldsymbol{x}, \boldsymbol{y}\}$ should satisfy

$$\mathsf{E}(\boldsymbol{y}-\bar{y})(\boldsymbol{y}-\bar{y})^{\mathsf{T}} = 0 \quad \text{and} \quad \mathsf{E}(\boldsymbol{x}-\bar{x})(\boldsymbol{x}-\bar{x})^{\mathsf{T}} = 0$$

with the transposition symbol T used instead of the conjugation symbol $*$. The variables are also required to be second-order circular, i.e.,

$$\mathsf{E}(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y})^{\mathsf{T}} = 0$$

These circularity assumptions are not needed when the variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ are real-valued. The circularity of $\boldsymbol{x}$ guarantees that its pdf in the Gaussian case will have the form

$$f_{\boldsymbol{x}}(x) = \frac{1}{\pi^p} \frac{1}{\det R_x} \exp\left\{-(x-\bar{x})^* R_x^{-1}(x-\bar{x})\right\}$$

Likewise, the circularity of $\boldsymbol{y}$ guarantees that its pdf will have the form

$$f_{\boldsymbol{y}}(y) = \frac{1}{\pi^q}\frac{1}{\det R_{\boldsymbol{y}}}\ \exp\{-(y-\bar{y})^* R_{\boldsymbol{y}}^{-1}(y-\bar{y})\}$$

The second-order circularity of $\boldsymbol{x}$ and $\boldsymbol{y}$ guarantees that the joint pdf of $\{\boldsymbol{x},\boldsymbol{y}\}$ will have the form

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) \ = \ \frac{1}{\pi^{p+q}}\frac{1}{\det R}\ \exp\left\{-\left[\begin{array}{cc}(x-\bar{x})^* & (y-\bar{y})^*\end{array}\right]R^{-1}\left[\begin{array}{c}x-\bar{x}\\ y-\bar{y}\end{array}\right]\right\}$$

Thus observe again that if $\boldsymbol{x}$ and $\boldsymbol{y}$ were *uncorrelated*, then the above pdf expressions lead to

$$f_{\boldsymbol{x},\boldsymbol{y}}(x,y) \ = \ f_{\boldsymbol{x}}(x)\cdot f_{\boldsymbol{y}}(y)$$

which shows that uncorrelated circular Gaussian random variables are also independent. This conclusion would not have held without the circularity assumptions in the complex case. We may add that circular Gaussian random variables are also called spherically-invariant Gaussian random variables.

Now the least-mean-squares estimator of $\boldsymbol{x}$ given $\boldsymbol{y}$ requires that we determine the conditional pdf $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$. This can be obtained from the calculation

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \ = \ \frac{f_{\boldsymbol{x},\boldsymbol{y}}(x,y)}{f_{\boldsymbol{y}}(y)} \ = \ \frac{1}{\pi^p}\frac{\det R_{\boldsymbol{y}}}{\det R}\frac{\exp\left\{-\left[\begin{array}{cc}(x-\bar{x})^* & (y-\bar{y})^*\end{array}\right]R^{-1}\left[\begin{array}{c}x-\bar{x}\\ y-\bar{y}\end{array}\right]\right\}}{\exp\{-(y-\bar{y})^* R_{\boldsymbol{y}}^{-1}(y-\bar{y})\}}$$

Following the same argument that we used earlier in Sec. 1.3.3, we can simplify the above expression by introducing the *block* upper-diagonal-lower triangular factorization (whose validity can again be verified, e.g., by direct calculation):

$$R \ \stackrel{\Delta}{=} \ \left[\begin{array}{cc}R_{\boldsymbol{x}} & R_{\boldsymbol{x}\boldsymbol{y}}\\ R_{\boldsymbol{y}\boldsymbol{x}} & R_{\boldsymbol{y}}\end{array}\right] = \left[\begin{array}{cc}\mathrm{I} & R_{\boldsymbol{x}\boldsymbol{y}}R_{\boldsymbol{y}}^{-1}\\ 0 & \mathrm{I}\end{array}\right]\left[\begin{array}{cc}\Sigma & 0\\ 0 & R_{\boldsymbol{y}}\end{array}\right]\left[\begin{array}{cc}\mathrm{I} & 0\\ R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{y}\boldsymbol{x}} & \mathrm{I}\end{array}\right]$$

where Σ is the Schur complement of $R_{\boldsymbol{y}}$ in R, namely,

$$\Sigma \ \stackrel{\Delta}{=} \ R_{\boldsymbol{x}} - R_{\boldsymbol{x}\boldsymbol{y}}R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{y}\boldsymbol{x}}$$

Inverting both sides of the above factorization for R we get

$$R^{-1} = \left[\begin{array}{cc}\mathrm{I} & 0\\ -R_{\boldsymbol{y}}^{-1}R_{\boldsymbol{y}\boldsymbol{x}} & \mathrm{I}\end{array}\right]\left[\begin{array}{cc}\Sigma^{-1} & 0\\ 0 & R_{\boldsymbol{y}}^{-1}\end{array}\right]\left[\begin{array}{cc}\mathrm{I} & -R_{\boldsymbol{x}\boldsymbol{y}}R_{\boldsymbol{y}}^{-1}\\ 0 & \mathrm{I}\end{array}\right]$$

which allows us to express the term

$$\left[\begin{array}{cc}(x-\bar{x})^* & (y-\bar{y})^*\end{array}\right]R^{-1}\left[\begin{array}{c}x-\bar{x}\\ y-\bar{y}\end{array}\right]$$

which appears in the expression for $f_{\boldsymbol{x},\boldsymbol{y}}(x,y)$, as a separable sum of two quadratic terms,

$$[(x-\bar{x}) - R_{\boldsymbol{x}\boldsymbol{y}}R_{\boldsymbol{y}}^{-1}(y-\bar{y})]^*\Sigma^{-1}[(x-\bar{x}) - R_{\boldsymbol{x}\boldsymbol{y}}R_{\boldsymbol{y}}^{-1}(y-\bar{y})] \ + \ (y-\bar{y})^* R_{\boldsymbol{y}}^{-1}(y-\bar{y})$$

Substituting this equality into the expression for $f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)$, and using $\det R = \det\Sigma\cdot\det R_{\boldsymbol{y}}$, we conclude that

$$f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \ = \ \frac{1}{\pi^p}\frac{1}{\det\Sigma}\exp\{-[(x-\bar{x}) - R_{\boldsymbol{x}\boldsymbol{y}}R_{\boldsymbol{y}}^{-1}(y-\bar{y})]^*\Sigma^{-1}[(x-\bar{x}) - R_{\boldsymbol{x}\boldsymbol{y}}R_{\boldsymbol{y}}^{-1}(y-\bar{y})]\}$$

which can be interpreted as the pdf of a circular Gaussian random variable with covariance matrix Σ and mean value $\bar{x} + R_{xy}R_y^{-1}(y - \bar{y})$. We therefore conclude that

$$\boxed{\hat{x} \triangleq \mathsf{E}(x|y) = \bar{x} + R_{xy}R_y^{-1}(y - \bar{y})}$$

and the resulting m.m.s.e. matrix is

$$\boxed{\text{m.m.s.e.} \triangleq R_{\tilde{x}} = R_x - R_{\hat{x}} = R_x - R_{xy}R_y^{-1}R_{yx}}$$

These are the extensions to the vector case of expressions (1.3.21) and (1.3.22) in the scalar case. Note further that in the zero-mean case we obtain

$$\boxed{\hat{x} = R_{xy}R_y^{-1}y}$$

with $\{R_x, R_y, R_{xy}\}$ defined accordingly,

$$R_x = \mathsf{E}\,xx^*, \quad R_y = \mathsf{E}\,yy^*, \quad R_{xy} = \mathsf{E}\,xy^*$$

Observe from the above expressions that the solution of the optimal estimation problem in the Gaussian case is completely determined by the second-order moments of the variables $\{x, y\}$ (i.e., by R_x, R_y, and R_{xy}). This means that, in the Gaussian case, the m.m.s.e. matrix can be evaluated *beforehand* by the designer (i.e., prior to the collection of the observations); a step that provides a mechanism for checking whether the least-mean-squares estimator will be an acceptable solution.

Lemma 1.4.2 (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

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

and the resulting minimum cost is

$$\text{m.m.s.e.} = R_x - R_{xy}R_y^{-1}R_{yx}$$

1.5 SUMMARY OF MAIN RESULTS

This chapter highlights several concepts and results in least-mean-squares estimation. Some of these concepts are reproduced here in a less technical language in order to reinforce their importance.

1. The variance of a random variable serves as a measure of the amount of uncertainty about the variable: the larger the variance the less certain we are about the value it may assume in an experiment.
2. The least-mean-squares error criterion is useful in that it leads to tractable mathematical solutions. The criterion is also intuitively appealing. By seeking to minimize the variance of the estimation error we are in effect attempting to force this error to assume values close to its mean and, hence, to assume small values since the mean is zero.

3. The least-mean-squares estimator of a random variable x given another random variable y is the conditional expectation estimator, namely, $\hat{x} = \mathsf{E}(x|y)$. This estimator is optimal in the sense that it minimizes the covariance matrix of the error vector (or, equivalently, its trace), i.e., it solves

$$\min_{h(\cdot)} \ \mathsf{E}\,\tilde{x}\tilde{x}^* \qquad \text{or} \qquad \min_{h(\cdot)} \ \mathsf{Tr}(R_{\tilde{x}})$$

4. A defining property of the least-mean-squares estimator is that the resulting estimation error is uncorrelated with any function of the observations, namely, it holds that

$$\mathsf{E}(x - \hat{x})g(y) = 0 \quad \text{for any function } g(\cdot) \text{ of } y$$

In particular,

$$\tilde{x} \ \perp \ \hat{x} \qquad \text{and} \qquad \tilde{x} \ \perp \ y$$

5. The evaluation of the conditional expectation, $\mathsf{E}(x|y)$, is a formidable task in most cases. However, for circular Gaussian random variables the estimator $\hat{x}$ is related to the observation y in an affine manner. Specifically, it holds that

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

where

$$R_x = \mathsf{E}(x - \bar{x})(x - \bar{x})^*, \quad R_y = \mathsf{E}(y - \bar{y})(y - \bar{y})^*$$

and

$$R_{xy} = \mathsf{E}(x - \bar{x})(y - \bar{y})^*$$

In particular, the estimator is completely determined from knowledge of the first and second-order moments of $\{x, y\}$, namely, their means, covariances and cross-covariance.

1.6 BIBLIOGRAPHIC NOTES

Probability theory. The exposition in this chapter assumes some basic knowledge of probability theory; mainly with regards to the concepts of mean, variance, probability density function, and vector-random variables. Most of these ideas were defined and introduced in the chapter from first principles. If additional help is needed, some accessible references on probability theory and basic random variable concepts are Papoulis (1991), Picinbono (1993), Leon-Garcia (1994), Stark and Woods (1994), and Durrett (1996). The textbook by Leon-Garcia (1994) is rich in examples, and is particularly directed to an engineering audience.

Mean-square-error performance. The squared-error criterion, whereby the square of the estimation error is used as a measure of performance, has a very distinguished history. It dates back to C. F. Gauss (1795), who developed a deterministic least-squares-error criterion as opposed to the stochastic least-mean-squares criterion of this chapter. Gauss' formulation was motivated by his work on celestial orbits, and we shall comment on it more fully in the concluding remarks of Chapter 11 when we study the least-squares criterion. A distinctive feature of the square-error criterion is that it penalizes large errors more than small errors. In this way, it is more sensitive to the presence of outliers in the data. This is in contrast, for example, to Laplace's proposition to use the absolute error criterion as a performance measure (see Sheynin (1977)). Gauss was very much aware of the distinction between both design criteria and this is how he commented on his squared-error criterion in relation to Laplace's absolute-error criterion:

" Laplace has also considered the problem in a similar manner, but he adopted the absolute value of the error as the measure of this loss. Now if I am not mistaken, this convention is no less arbitrary than mine. Should an error of double size be considered as tolerable as a single error twice repeated or worse? Is it better to assign only twice as much influence to a double error or more? The answers are not self-evident, and the problem cannot be resolved by mathematical proofs, but only by an arbitrary decision."

Extracted from the translation by Stewart (1995).

Besides Gauss' motivation, there are many good reasons for using the mean-square-error criterion, not the least of which is the fact that it leads to a closed-form characterization of the solution as a conditional mean. In addition, for Gaussian random variables, it can even be argued that the least-mean-squares error estimator is practically optimal for any other choice of the error cost function (quadratic or otherwise) — see, for example, Pugachev (1958) and Zakai (1964).

Statistical theory. There is extensive work on the least-mean-squares error criterion in the statistical literature. For instance, the result of Thm. 1.3.1 on the conditional mean estimator is related to the so-called Rao-Blackwell theorem from statistics (see, e.g., Caines (1988) and Scharf (1991)). However, in statistics, there is often a distinction between what is known as the classical approach to estimation and the alternative so-called Bayesian approach to estimation. In the classical approach, the unknown quantity to be estimated is modeled as a deterministic but unknown constant; we shall encounter this situation in Chapter 3 while studying the Gauss-Markov theorem. The Bayesian approach, on the other hand, models the unknown quantity as a random variable, which is the point of view we adopted in this chapter. Such Bayesian formulations allow us to incorporate prior knowledge about the unknown variable itself into the solution, such as information about its probability density function. This fact helps explain why Bayesian techniques are dominant in many successful filtering and estimation designs; still the Bayesian approach has not been immune to controversies along its history (see Box and Tiao (1973)).

Complex random variables. Complex variables, as well as complex random variables, are frequent in electrical engineering (and perhaps more so than in any other discipline). One notable example arises in digital communications whereby symbols are often selected at random from a complex constellation (or even in the complex representation of bandpass signals). Since complex random variables will play a prominent role throughout this textbook, we have chosen to motivate them from first principles in the body of the chapter. In App. 1.B we pursue their study more closely and focus, in particular, on the important class of complex-valued *Gaussian* random variables. It is explained in the appendix that a certain circularity assumption needs to be satisfied if the resulting pdf in the complex case is to be uniquely determined by the first and second-order moments of the complex random variable, as happens in the real case. The main conclusion appears in the statement of Lemma 1.B.1, which shows the form of a complex Gaussian distribution under the circularity assumption. The original derivation of this form is due to Wooding (1956) — see also Goodman (1963) and Miller (1974). It is for this reason that, in future discussions, whenever we refer to a complex Gaussian distribution we shall often attach the qualification "circular" to it and refer instead to a circular Gaussian distribution.

Linear algebra. Throughout the book, the reader will be exposed to a variety of similar concepts from linear algebra and matrix theory in a self-contained and motivated manner. In this way, after progressing sufficiently enough into the book, students will be able to master many useful concepts. If additional help is needed, some accessible references on matrix theory are the two volumes by Gantmacher (1959), the book by Bellman (1970), and the two volumes by Horn and Johnson (1987,1994). Accessible references on linear algebra are, for example, the books by Strang (1988,1993), Lay (1994), and Lax (1997).

1.7 PROBLEMS

Problem 1.1 (Rayleigh distribution) Consider a Rayleigh-distributed random variable x with pdf given by (1.1.3). Show that its mean and variance are given by (1.1.4).

Remark. Recall that for a general random variable x with pdf $f_x(x)$, the mean and variance are defined by

$$\bar{x} \triangleq \int_{-\infty}^{\infty} x f_x(x) \mathrm{d}x. \qquad \sigma_x^2 \triangleq \left(\int_{-\infty}^{\infty} x^2 f_x(x) \mathrm{d}x \right) - \bar{x}^2$$

Problem 1.2 (Markov's inequality) Suppose x is a scalar *nonnegative* real-valued random variable with probability density function $f_x(x)$. Establish the inequality $P[x \geq a] \leq \mathsf{E}\, x/a$.

Problem 1.3 (Chebyshev's inequality) Consider a scalar real-valued random variable x with mean $\bar{x}$ and variance σ_x^2. Define the nonnegative random variable $y = (x - \bar{x})^2$, whose mean is clearly σ_x^2. Use Markov's inequality to establish Chebyshev's inequality (1.1.5).

Problem 1.4 (Conditional expectation) Consider two real-valued random variables x and y. Establish that $\mathsf{E}[\mathsf{E}(x|y)] = \mathsf{E}\, x$. That is, show that

$$\int_{\mathcal{S}_x} x f_x(x) \mathrm{d}x \;=\; \int_{\mathcal{S}_y} f_y(y) \left[\int_{\mathcal{S}_x} x f_{x|y}(x|y) \mathrm{d}x \right] \mathrm{d}y$$

where $\mathcal{S}_x$ and $\mathcal{S}_y$ denote the supports of the variables x and y, respectively.

Remark. This identity states that we can split the evaluation of $\mathsf{E}\, x$ into two separate expectations: one is the conditional expectation of x given y (the result of which is a function of y), and the other is an expectation over y.

Problem 1.5 (Estimator for a binary signal) Consider the same setting of Ex. 1.3.1 but assume now that the noise v has a generic variance σ_v^2.

(a) Show that the optimal least-mean-squares estimator of x given y is $\hat{x} = \tanh(y/\sigma_v^2)$. Plot the estimate $\hat{x}$ as a function of y for the values $\sigma_v^2 = 0.5, 1, 2$.

(b) Argue that $\hat{x} = \mathrm{sign}(y)$ can be taken as a suboptimal estimator. Follow the derivation at the end of Ex. 1.3.2 to show that the improvement in SNR is given by

$$\mathsf{SNR}_{out} - \mathsf{SNR}_{in} \;=\; 10 \log(\sigma_v^2/4a)$$

where

$$a = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_v} \int_1^{\infty} e^{-\frac{v^2}{2\sigma_v^2}} \mathrm{d}v$$

(c) Plot the improvement in SNR as a function of σ_v^2.

Problem 1.6 (Biased measurements) Consider the same setting of Ex. 1.3.1 but assume now that the noise v has mean $\bar{v}$ and unit variance.

(a) Show that the optimal least-mean-squares estimator of x given y is $\hat{x} = \tanh(y - \bar{v})$. Plot the estimate $\hat{x}$ as a function of y for the values $\bar{v} = -0.5, 0, 0.5$.

(b) Argue that $\hat{x} = \mathrm{sign}(y - \bar{v})$ can be taken as a suboptimal estimator.

(c) Following the derivation at the end of Ex. 1.3.2, verify that the improvement in SNR is given by

$$\mathsf{SNR}_{out} - \mathsf{SNR}_{in} \;=\; 10 \log\left(\frac{1 + \bar{v}^2}{4a} \right)$$

where

$$a = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_v} \int_1^{\infty} e^{-\frac{(v-\bar{v})^2}{2}} \mathrm{d}v$$

Remark. Since v is not zero mean, we are measuring its power by using $\mathsf{E}\, v^2$ and not σ_v^2.

Problem 1.7 (Correlation coefficient) Consider two scalar random variables $\{x, y\}$ with means $\{\bar{x}, \bar{y}\}$, variances $\{\sigma_x^2, \sigma_y^2\}$, and cross-correlation σ_{xy}. Define the correlation coefficient $\rho_{xy} = \sigma_{xy}/(\sigma_x \sigma_y)$. Use the fact that the covariance matrix

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

is nonnegative-definite, to conclude that $|\rho_{xy}| \leq 1$.

Problem 1.8 (Fully correlated random variables) Consider two scalar real-valued random variables x and y with correlation coefficient ρ_{xy}, means $\{\bar{x}, \bar{y}\}$, and variances $\{\sigma_x^2, \sigma_y^2\}$. Show that $|\rho_{xy}| = 1$ if, and only if, the random variables are related as

$$x - \bar{x} = \pm \frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Problem 1.9 (Chi-square distribution) Let x be a real-valued random variable with pdf $f_x(x)$. Define $y = x^2$.

(a) Use the fact that for any nonnegative y, the event $\{y \leq y\}$ occurs whenever $\{-\sqrt{y} \leq x \leq \sqrt{y}\}$ to conclude that the pdf of y is given by

$$f_y(y) = \frac{1}{2}\frac{f_x(\sqrt{y})}{\sqrt{y}} + \frac{1}{2}\frac{f_x(-\sqrt{y})}{\sqrt{y}}, \quad y > 0$$

(b) Assume x is Gaussian with zero mean and unit variance. Use part (a) to conclude that

$$f_y(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2}, \quad y > 0$$

Remark. The above pdf is known as the Chi-square distribution with one degree of freedom. More generally, a Chi-square distribution with k degrees of freedom is characterized by the pdf

$$f_y(y) = \frac{1}{2^{k/2}\Gamma(k/2)}\, y^{(k-2)/2} e^{-y/2}, \quad y > 0$$

where $\Gamma(\cdot)$ is the so-called Gamma function, which is defined by the integral

$$\Gamma(z) \triangleq \int_0^\infty x^{z-1} e^{-x} dx, \qquad z > 0$$

The function $\Gamma(\cdot)$ has the following useful properties: $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(z+1) = z\Gamma(z)$ for any $z > 0$, and $\Gamma(n+1) = n!$ for any $n \geq 0$.

Problem 1.10 (Rayleigh distribution) Consider an FIR channel with two real-valued taps, $x(1)$ and $x(2)$. The taps are assumed to be independent zero-mean unit-variance Gaussian random variables.

(a) Use the result of part (b) of Prob. 1.9 to show that the random variable $w = x^2(1) + x^2(2)$ has a Chi-square distribution with two degrees of freedom, i.e.,

$$f_w(w) = \frac{1}{2} e^{-w/2}, \quad w \geq 0$$

(b) Conclude that the random variable $z = \sqrt{x^2(1) + x^2(2)}$ has a Rayleigh distribution, namely, show that

$$f_z(z) = z e^{-z^2/2}, \quad z \geq 0$$

with $\bar{z} = \sqrt{\pi/2}$ and $\sigma_z^2 = (2 - \pi/2)$.

Problem 1.11 (BPSK signal) Consider noisy observations $y(i) = x + v(i)$, where x and $v(i)$ are independent real-valued random variables, $v(i)$ is a white-noise Gaussian random process with zero mean and variance σ_v^2, and x takes the values ± 1 with equal probability. The value of x is either $+1$ or -1 for all measurements $\{y(i)\}$. The whiteness assumption on $v(i)$ means that $\mathsf{E}\, v(i)v(j) = 0$ for $i \neq j$.

(a) Show that the least-mean-squares estimate of x given N observations $\{y(0), \ldots, y(N-1)\}$ is

$$\hat{x}_N = \tanh\left(\sum_{i=0}^{N-1} y(i)/\sigma_v^2\right)$$

(b) Assume x takes the value 1 with probability p and the value -1 with probability $1-p$. Show that the least-mean-squares estimate of x given N observations $\{y(0), \ldots, y(N-1)\}$ is given by

$$\hat{x}_N = \tanh\left[\frac{1}{2}\ln\left(\frac{p}{1-p}\right) + \sum_{i=0}^{N-1} y(i)/\sigma_v^2\right]$$

in terms of the natural logarithm of $p/(1-p)$.

(c) Assume the noise is instead correlated. Specifically, define $v = \mathrm{col}\{v(0), v(1), \ldots, v(N)\}$ and let $R_v = \mathsf{E}\, vv^*$. Show that the least-mean-squares estimate of x given N observations $\{y(0), \ldots, y(N-1)\}$ is now given by

$$\hat{x}_N = \tanh\left[\frac{1}{2}\ln\left(\frac{p}{1-p}\right) + y^{\mathsf{T}} R_v^{-1} h\right]$$

where $y = \mathrm{col}\{y(0), y(1), \ldots, y(N)\}$ and $h = \mathrm{col}\{1, 1, \ldots, 1\}$.

Problem 1.12 (Optimal receiver) Let us complete the derivation of Ex. 1.4.4 and evaluate $\mathsf{E}(x|y)$.

(a) Verify that

$$\hat{x} = \frac{1}{f_y(y)}\left(\sum_{k=0}^{3} a(k) m_k f_v(y - Hm_k)\right)$$

where $a(0) = p^2$, $a(1) = q^2$ and $a(2) = a(3) = pq$.

(b) Introduce

$$a = p^2 \cdot e^{-\frac{1}{2}(-2y(0) - 3y(1) + 2)} \qquad b = q^2 \cdot e^{-\frac{1}{2}(-2y(0) + 3y(1) + 2)}$$

$$c = pq \cdot e^{-\frac{1}{2}(-2y(0) + y(1))} \qquad d = pq \cdot e^{-\frac{1}{2}(2y(0) - y(1))}$$

Show that the expression in part (a) simplifies to

$$\begin{bmatrix} \hat{s}(0) \\ \hat{s}(1) \end{bmatrix} = \frac{1}{a+b+c+d}\begin{bmatrix} a - b + c - d \\ a - b - c + d \end{bmatrix}$$

Problem 1.13 (Exponential distribution) Suppose we observe $y = x + v$ where x and v are independent real-valued random variables with exponential distributions with parameters λ_1 and λ_2 ($\lambda_1 \neq \lambda_2$). That is, the pdfs of x and v are $f_x(x) = \lambda_1 e^{-\lambda_1 x}$ for $x \geq 0$ and $f_v(v) = \lambda_2 e^{-\lambda_2 v}$ for $v \geq 0$, respectively.

(a) Using the fact that the pdf of the sum of two independent random variables is the convolution of the individual pdfs, show that

$$f_y(y) = \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} \cdot e^{-\lambda_2 y} \cdot \left[e^{(\lambda_2 - \lambda_1) y} - 1\right]$$

for $y \geq 0$.

(b) Show that the joint pdf of x and y is $f_{x,y}(x, y) = \lambda_1 \lambda_2 e^{(\lambda_2 - \lambda_1)x - \lambda_2 y}$ for $x \geq 0$ and $y \geq 0$.

(c) Show that the least-mean-squares estimate of x given $y = y$ is

$$\hat{x} = \frac{1}{\lambda_1 - \lambda_2} - \frac{e^{-\lambda_1 y}}{e^{-\lambda_2 y} - e^{-\lambda_1 y}}\, y$$

Problem 1.14 (Equivalent criteria) Show that the least-mean-squares estimator $\hat{x} = \mathsf{E}(x|y)$ also minimizes $\mathsf{E}\,\tilde{x}^* W \tilde{x}$, for any Hermitian nonnegative-definite matrix W; it does not even need to be invertible. [*Hint*: Introduce the eigen-decomposition of W (cf. App. 1.A), say $W = U\Lambda U^*$, and estimate U^*x from y.]

Problem 1.15 (Second- and fourth-order moments) Let R denote an $M \times M$ positive-definite matrix and introduce its eigen-decomposition (cf. App. 1.A),

$$R = \sum_{i=1}^{M} \lambda_i u_i u_i^*$$

where the λ_i are the eigenvalues of R (all positive), and the u_i are the eigenvectors of R. The u_i are orthonormal, i.e., $u_j^* u_i = 0$ for all $i \neq j$ and $u_i^* u_i = 1$. Let $\boldsymbol{h}$ be a random vector with probability distribution $P(\boldsymbol{h} = u_j) = \lambda_j / \mathsf{Tr}(R)$. That is, the probability that $\boldsymbol{h}$ coincides with the $j-$th eigenvector of R is proportional to the corresponding eigenvalue.

(a) Show that $\mathsf{E}\,\boldsymbol{h}\boldsymbol{h}^* = R/\mathsf{Tr}(R)$ and $\mathsf{E}\,\boldsymbol{h}\boldsymbol{h}^*\boldsymbol{h}\boldsymbol{h}^* = R/\mathsf{Tr}(R)$.

(b) Show that $\mathsf{E}\,\boldsymbol{h}^* R^{-1} \boldsymbol{h} = M/\mathsf{Tr}(R)$ and $\mathsf{E}\,\boldsymbol{h}\boldsymbol{h}^* R^{-1} \boldsymbol{h}\boldsymbol{h}^* = \mathrm{I}/\mathsf{Tr}(R)$.

(c) Show that $\mathsf{E}\,\boldsymbol{h}^*\boldsymbol{h} = 1$ and

$$\mathsf{E}\,\boldsymbol{h} = \frac{1}{\mathsf{Tr}(R)} \sum_{j=1}^{M} \lambda_j u_j$$

Problem 1.16 (Independent and Gaussian variables) Consider two independent zero-mean random variables $\{\boldsymbol{u}, \boldsymbol{w}\}$, where $\boldsymbol{u}$ is a row vector and $\boldsymbol{w}$ is a column vector; both are M-dimensional. The covariance matrices of $\boldsymbol{u}$ and $\boldsymbol{w}$ are defined by $\mathsf{E}\,\boldsymbol{u}^*\boldsymbol{u} = \sigma_u^2 \mathrm{I}$ and $\mathsf{E}\,\boldsymbol{w}\boldsymbol{w}^* = C$. In addition, $\boldsymbol{u}$ is assumed to be circular Gaussian. Define a third scalar-valued variable as $\boldsymbol{e}_a = \boldsymbol{u}\boldsymbol{w}$.

(a) Show that $\mathsf{E}\,|\boldsymbol{e}_a|^2 = \sigma_u^2 \mathsf{Tr}(C)$.

(b) Use the result of Lemma 1.B.3 to show that $\mathsf{E}\,\|\boldsymbol{u}\|^2 \cdot |\boldsymbol{e}_a|^2 = (M+1)\sigma_u^4 \mathsf{Tr}(C)$, where the notation $\|\cdot\|$ denotes the Euclidean norm of its argument, i.e., $\|\boldsymbol{u}\|^2 = \boldsymbol{u}\boldsymbol{u}^*$ (since $\boldsymbol{u}$ is a row vector).

Problem 1.17 (Fourth-moment) Assume $\boldsymbol{u}$ is a circular Gaussian random row vector with a diagonal covariance matrix Λ. Define $\boldsymbol{z} = \|\boldsymbol{u}\|^2$. What is the variance of $\boldsymbol{z}$?

Problem 1.18 (Covariance equation) Consider two column vectors $\{\boldsymbol{w}, \boldsymbol{z}\}$ that are related via

$$\boldsymbol{z} = \boldsymbol{w} + \mu \boldsymbol{u}^* (\boldsymbol{d} - \boldsymbol{u}\boldsymbol{w})$$

where $\boldsymbol{u}$ is a circular Gaussian random variable with a diagonal covariance matrix, $\mathsf{E}\,\boldsymbol{u}^*\boldsymbol{u} = \Lambda$ ($\boldsymbol{u}$ is a row vector). Moreover, μ is a positive constant and $\boldsymbol{d} = \boldsymbol{u}w^o + \boldsymbol{v}$, for some constant vector w^o and random scalar $\boldsymbol{v}$ with variance σ_v^2. The variables $\{\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{w}\}$ are independent of each other. Define $\boldsymbol{e}_a = \boldsymbol{u}(w^o - \boldsymbol{w})$, as well as the error vectors $\tilde{\boldsymbol{z}} = w^o - \boldsymbol{z}$ and $\tilde{\boldsymbol{w}} = w^o - \boldsymbol{w}$, and denote their covariances by $\{R_{\tilde{z}}, R_{\tilde{w}}\}$. Assume $\mathsf{E}\,\boldsymbol{z} = \mathsf{E}\,\boldsymbol{w} = w^o$, while all other random variables are zero-mean.

(a) Verify that $\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{w}} - \mu \boldsymbol{u}^* (\boldsymbol{e}_a + \boldsymbol{v})$.

(b) Use the result of Lemma 1.B.3 to show that

$$R_{\tilde{z}} = R_{\tilde{w}} - \mu R_{\tilde{w}} \Lambda - \mu \Lambda R_{\tilde{w}} + \mu^2 \left(\Lambda \mathsf{Tr}(R_{\tilde{w}} \Lambda) + \Lambda R_{\tilde{w}} \Lambda \right) + \mu^2 \sigma_v^2 \Lambda$$

(c) How would the result of part (b) change if $\boldsymbol{u}$ were real-valued Gaussian?

1.8 COMPUTER PROJECT

Project 1.1 (Comparing optimal and suboptimal estimators) The purpose of this project is to compare the performance of an optimal least-mean-squares estimator with three approximations for it, along the lines discussed in Ex. 1.3.1. Thus consider the setting of Prob. 1.11.

(a) Write a MATLAB program that generates a BPSK random variable x that is equal to $+1$ with probability p and to -1 with probability $1-p$.

(b) Simulate the estimator of part (b) of Prob. 1.11 for different number of observations N. For instance, generate observations $\{y(i)\}$ and plot $\hat{x}_N$ as a function of N for $1 \leq N \leq 10$, with all observations assumed generated by the same value of x — either $+1$ or -1. Plot $\hat{x}_N$ for the cases $p = 0.1, 0.3, 0.5, 0.8$. Observe how the estimate gets closer to the true value of x as the value of N increases. Do you notice any differences in behavior for the different values of p?

(c) Compare the performance of the optimal estimate $\hat{x}_N$ with the averaged estimate

$$\hat{x}_{N,av} \triangleq \frac{1}{N}\sum_{i=0}^{N-1} y(i)$$

for several values of N, say, for $1 \leq N \leq 300$, and for the same values of p in part (b). Does it take many more samples N for the averaged estimate $\hat{x}_{N,av}$ to provide a good result compared with the optimal nonlinear estimate $\hat{x}_N$?

(d) Fix $p = 1/2$, and define the nonlinear decision device:

$$\text{sign}[a] = \begin{cases} +1 & \text{if } a \geq 0 \\ -1 & \text{if } a < 0 \end{cases}$$

Consider also the alternative (sign-of-optimal) estimate

$$\hat{x}_{dec} = \text{sign}\,[\hat{x}_N]$$

It is clear that $\hat{x}_{dec}$ assumes the values ± 1, whereas the optimal estimate $\hat{x}_N$ does not. Is $\hat{x}_{dec}$ a better estimate than $\hat{x}_N$? The answer in the mean-square sense is of course negative since we already know that $\hat{x}_N$ is the best estimate. To verify this fact do the following. Fix the number of observations at $N = 10$. Then perform 1000 experiments, with each experiment i resulting in an optimal estimate $\hat{x}_{10}(i)$ and an estimate $\hat{x}_{dec}(i)$. For each estimate, the value of x is fixed at either $+1$ or -1. Compute the sample variances

$$\frac{1}{1000}\sum_{i=1}^{1000} |x - \hat{x}_{10}(i)|^2, \qquad \frac{1}{1000}\sum_{i=1}^{1000} |x - \hat{x}_{dec}(i)|^2$$

Which one is smaller? Repeat for the following (sign-of-average) estimate:

$$\hat{x}_{sign} = \text{sign}\,[\hat{x}_{N,av}]$$

That is, apply the decision device to the estimate that is obtained from averaging.

The programs that solve this problem are the following.

1. psk.m This function generates a BPSK signal x that assumes the value $+1$ with probability p and the value -1 with probability $1-p$.
2. partB.m This program generates four plots of $\hat{x}_N$, as a function of N, one for each value of p. Each plot will converge to $+1$ or -1 depending on whether the corresponding value of x is $+1$ or -1. A typical output of this program is shown in Fig. 1.12.
3. partC.m This program generates a figure with four plots. The figure shows $\{\hat{x}_N, \hat{x}_{N,av}\}$ for the four different values of p and over the entire interval $1 \leq N \leq 300$. A typical output of this program is shown in Fig. 1.13. The dotted lines correspond to the optimal estimator, while the solid lines correspond to the averaged estimator. Observe how the averaged estimator requires many more experiments for good performance.
4. partD.m This program estimates the variances of the estimators $\{\hat{x}_N, \hat{x}_{N,av}, \hat{x}_{dec}, \hat{x}_{sign}\}$. Typical values are $\{0.0031, 0.1027, 0.0040, 0.0040\}$, respectively.

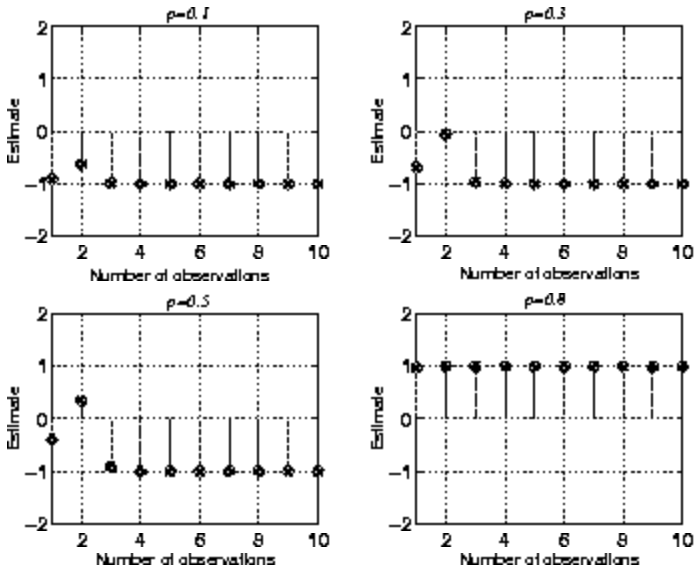

Figure 1.12. The plots show the values of the optimal estimates $\hat{x}_N$ for different choices of N (the number of observations) and for different values of p (which determines the probability distribution of x).

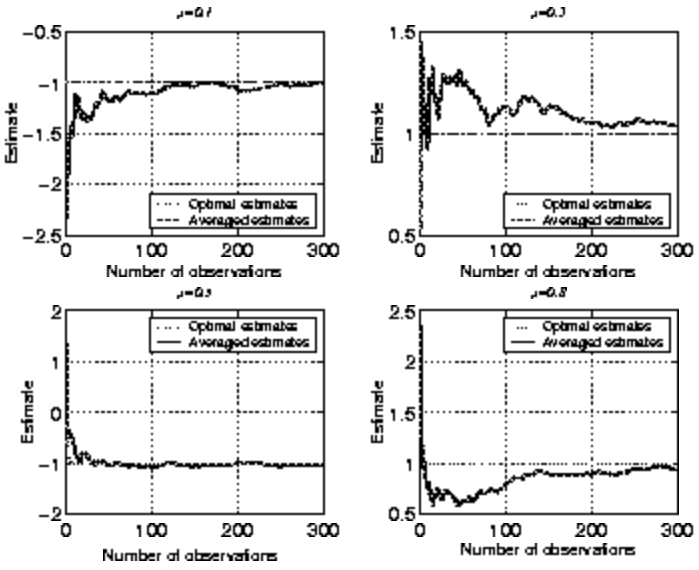

Figure 1.13. The plots show the values of the optimal estimates $\hat{x}_N$ (*dotted lines*) and the averaged estimates $\hat{x}_{N,av}$ (*solid lines*) for different choices of N (the number of observations) and for different values of p (which determines the probability distribution of x). Observe how the averaged estimates are significantly less reliable for a smaller number of observations.

1.A HERMITIAN AND POSITIVE-DEFINITE MATRICES

The Hermitian conjugate, A^*, of a matrix A is the complex conjugate of its transpose, e.g., if

$$A = \begin{bmatrix} 1 & -j \\ 2+j & 1-j \end{bmatrix}$$

then

$$A^* = \begin{bmatrix} 1 & 2-j \\ j & 1+j \end{bmatrix}$$

where $j = \sqrt{-1}$.

Hermitian matrices. A Hermitian matrix is a square matrix satisfying $A^* = A$, e.g., if

$$A = \begin{bmatrix} 1 & 1+j \\ 1-j & 1 \end{bmatrix}$$

then

$$A^* = \begin{bmatrix} 1 & 1+j \\ 1-j & 1 \end{bmatrix} = A$$

so that A is Hermitian.

Spectral decomposition. Hermitian matrices can only have *real* eigenvalues. To see this, assume u_i is an eigenvector of A corresponding to an eigenvalue λ_i, i.e., $Au_i = \lambda_i u_i$. Multiplying from the left by u_i^* we get

$$u_i^* A u_i = \lambda_i \|u_i\|^2$$

where $\|\cdot\|$ denotes the Euclidean norm of its argument. Now the scalar quantity on the left-hand side of the above equality is real since it coincides with its complex conjugate,

$$(u_i^* A u_i)^* = u_i^* A^* u_i = u_i^* A u_i$$

Therefore, λ_i must be real too.

Another important property of Hermitian matrices, whose proof requires a more involved argument, is that such matrices always have a *full* set of orthonormal eigenvectors. That is, if A is $n \times n$ Hermitian, then there will exist n orthonormal eigenvectors u_i satisfying

$$Au_i = \lambda_i u_i, \quad \|u_i\|^2 = 1, \quad u_i^* u_j = 0 \text{ for } i \neq j$$

In compact matrix notation we can write this so-called spectral (or modal or eigen-) decomposition of A as

$$\boxed{A = U \Lambda U^*}$$

where[17]

$$\Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}, \qquad U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}$$

and U satisfies

$$\boxed{UU^* = U^* U = I}$$

We say that U is a unitary matrix.

Positive-definite matrices. An $n \times n$ Hermitian matrix A is positive semi-definite (also called nonnegative definite) if it satisfies

$$x^* A x \geq 0 \quad \text{for all column vectors } x$$

It is positive definite if $x^* A x > 0$ except when $x = 0$. We denote a positive-definite matrix by writing $A > 0$ and a positive semi-definite matrix by writing $A \geq 0$. Among the several characterizations of positive-definite matrices, we note the following.

[17] The notation $\text{diag}\{a, b\}$ denotes a diagonal matrix with diagonal entries a and b.

Lemma 1.A.1 (Eigenvalues of positive-definite matrices) An $n \times n$ Hermitian matrix A is positive-definite if, and only if, all its eigenvalues are positive:

$$A > 0 \iff \{\lambda_i > 0\}$$

Proof: Let $A = U\Lambda U^*$ denote the spectral decomposition of A. Let also u_i be the $i-$th column of U with λ_i the corresponding eigenvalue,

$$Au_i = \lambda_i u_i. \quad \|u_i\|^2 = 1$$

If we multiply this equality from the left by u_i^* we get

$$u_i^* A u_i = \lambda_i \|u_i\|^2 = \lambda_i > 0$$

where the last inequality follows from the fact that $x^*Ax > 0$ for any nonzero vector x. Therefore, $A > 0$ implies $\lambda_i > 0$. Conversely, assume $\lambda_i > 0$ and multiply the equality $A = U\Lambda U^*$ by any nonzero vector x and its conjugate transpose, from right and left, to get

$$x^*Ax = x^*U\Lambda U^*x$$

Now define the matrix

$$\Lambda^{1/2} \triangleq \text{diag}\left\{\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}\right\}$$

and the vector $y = \Lambda^{1/2}U^*x$. The vector y is nonzero since U and $\Lambda^{1/2}$ are nonsingular matrices and, therefore, the product $\Lambda^{1/2}U^*$ cannot map a nonzero vector x to 0. Then the above equality becomes $x^*Ax = \|y\|^2 > 0$, which establishes that $A > 0$.

◊

In a similar vein, we can show that

$$\boxed{A \geq 0 \iff \lambda_i \geq 0}$$

Note further that since

$$\det A = (\det U)\ (\det \Lambda)\ (\det U^*)$$

and

$$\det U \det U^* = 1$$

we find that

$$\det A = \det \Lambda = \prod_{i=1}^{n} \lambda_i$$

Therefore, the determinant of a positive-definite matrix is positive,

$$\boxed{A > 0 \implies \det A > 0}$$

Rayleigh-Ritz characterization of eigenvalues. If A is an $n \times n$ Hermitian matrix, then it holds that for all vectors x

$$\boxed{\lambda_{min}\|x\|^2 \leq x^*Ax \leq \lambda_{max}\|x\|^2}$$

as well as

$$\lambda_{min} = \min_{x \neq 0}\left(\frac{x^*Ax}{x^*x}\right) = \min_{\|x\|=1} x^*Ax$$

$$\lambda_{max} = \max_{x \neq 0}\left(\frac{x^*Ax}{x^*x}\right) = \max_{\|x\|=1} x^*Ax$$

where $\{\lambda_{min}, \lambda_{max}\}$ denote the smallest and largest eigenvalues of A. The ratio x^*Ax/x^*x is called the Rayleigh-Ritz ratio.

One simple proof of these claims follows by invoking the spectral decomposition $A = U\Lambda U^*$, where U is unitary and Λ has real entries. Thus let $y = U^*x$ for any vector x. Then

$$x^*Ax = x^*U\Lambda U^*x = y^*\Lambda y = \sum_{k=1}^{n} \lambda_k |y(k)|^2$$

with the $\{y(k)\}$ denoting the individual entries of y. Now since the squared terms $\{|y(k)|^2\}$ are nonnegative, we get

$$\lambda_{\min} \sum_{k=1}^{n} |y(k)|^2 \leq \sum_{k=1}^{n} \lambda_k |y(k)|^2 \leq \lambda_{\max} \sum_{k=1}^{n} |y(k)|^2$$

or, equivalently,

$$\lambda_{\min} \|y\|^2 \leq x^*Ax \leq \lambda_{\max} \|y\|^2$$

Using the fact that U is unitary and, hence,

$$\|y\|^2 = y^*y = x\underbrace{UU^*}_{=I}x = \|x\|^2$$

we conclude that

$$\lambda_{\min} \|x\|^2 \leq x^*Ax \leq \lambda_{\max} \|x\|^2$$

The lower and upper bounds are achieved when x is chosen as the eigenvector corresponding to $\lambda_{\min}$ or to $\lambda_{\max}$, respectively.

1.B GAUSSIAN RANDOM VECTORS

Gaussian random variables play an important role in many situations, especially when we deal with the sum of a large number of random variables. In this case, a fundamental result in probability theory, known as the *central limit theorem*, states that under conditions often reasonable in applications, the probability density function (pdf) of the sum of independent random variables approaches that of a Gaussian random variable. Specifically, if $\{x(i), i = 1, 2, \ldots, N\}$ are independent real-valued Gaussian random variables with mean $\bar{x}(i)$ and variance $\sigma_x^2(i)$ each, then the pdf of the normalized variable

$$y \triangleq \frac{\sum_{i=1}^{N}[x(i) - \bar{x}(i)]}{\sqrt{\sum_{i=1}^{N}\sigma_x^2(i)}}$$

approaches that of a Gaussian distribution with zero mean and unit variance, i.e.,

$$f_y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \quad \text{as } N \longrightarrow \infty$$

or, equivalently,

$$\boxed{\lim_{N\to\infty} P(y \le a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-y^2/2} dy}$$

It is for this reason that whenever we say "Gaussian noise" in practice, the term essentially refers to the combined effect of many independent disturbances.

In this appendix we shall describe the general form of the pdf of a vector Gaussian random variable. However, as the discussion will show, we need to distinguish between two cases depending on whether the random variable is real or complex. In the complex case, the random variable will need to satisfy a certain *circularity* assumption in order for the given form of the pdf to be valid.

Real-Valued Gaussian Random Variables

We start with the real case. Thus consider a $p \times 1$ random vector x with mean $\bar{x}$ and covariance matrix

$$R_x = \mathsf{E}(x - \bar{x})(x - \bar{x})^{\mathsf{T}}$$

assumed nonsingular. We say that x has a Gaussian distribution if its pdf has the form

$$\boxed{f_x(x) = \frac{1}{\sqrt{(2\pi)^p}} \frac{1}{\sqrt{\det R_x}} \exp\left\{-\tfrac{1}{2}(x-\bar{x})^{\mathsf{T}} R_x^{-1}(x-\bar{x})\right\}} \tag{1.B.1}$$

in terms of the determinant of R_x. Of course, when $p = 1$, the above expression reduces to the pdf considered in the text in (1.1.2) with R_x replaced by σ_x^2.

Now consider a second $q \times 1$ Gaussian random vector y with mean $\bar{y}$ and covariance matrix

$$R_y = \mathsf{E}(y - \bar{y})(y - \bar{y})^{\mathsf{T}}$$

so that its pdf is given by

$$\boxed{f_y(y) = \frac{1}{\sqrt{(2\pi)^q}} \frac{1}{\sqrt{\det R_y}} \exp\left\{-\tfrac{1}{2}(y-\bar{y})^{\mathsf{T}} R_y^{-1}(y-\bar{y})\right\}}$$

Let R_{xy} denote the cross-covariance matrix between x and y, i.e.,

$$R_{xy} = \mathsf{E}(x - \bar{x})(y - \bar{y})^{\mathsf{T}}$$

We then say that the random variables $\{x, y\}$ have a joint Gaussian distribution if their joint pdf has the form

$$\boxed{f_{x,y}(x,y) = \frac{1}{\sqrt{(2\pi)^{p+q}}} \frac{1}{\sqrt{\det R}} \exp\left\{-\tfrac{1}{2}\begin{bmatrix} (x-\bar{x})^{\mathsf{T}} & (y-\bar{y})^{\mathsf{T}} \end{bmatrix} R^{-1} \begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix}\right\}} \tag{1.B.2}$$

in terms of the covariance matrix R of $\mathrm{col}\{x,y\}$, namely,

$$R = \mathsf{E}\left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}\right)\left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}\right)^{\mathsf{T}} = \begin{bmatrix} R_x & R_{xy} \\ R_{xy}^{\mathsf{T}} & R_y \end{bmatrix}$$

It can be seen from (1.B.2) that the joint pdf of $\{x,y\}$ is completely determined by the mean, covariances, and cross-covariance of $\{x,y\}$, i.e., by the first and second-order moments $\{\bar{x},\bar{y},R_x,R_y,R_{xy}\}$.

Complex-Valued Random Variables and Circularity

Let us now examine the case of complex-valued random vectors. We consider again two real random vectors $\{x,y\}$, both assumed of size $p\times 1$, with joint pdf given by (cf. (1.B.2)):

$$f_{x,y}(x,y) = \frac{1}{(2\pi)^p}\frac{1}{\sqrt{\det R}}\exp\left\{-\tfrac{1}{2}\begin{bmatrix} (x-\bar{x})^{\mathsf{T}} & (y-\bar{y})^{\mathsf{T}} \end{bmatrix} R^{-1}\begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix}\right\} \tag{1.B.3}$$

Let $z = x + jy$, where $j=\sqrt{-1}$, denote a complex-valued random variable defined in terms of $\{x,y\}$. Its mean is simply

$$\bar{z} = \mathsf{E}\,z = \bar{x} + j\bar{y}$$

while its covariance matrix is

$$R_z \;\triangleq\; \mathsf{E}(z-\bar{z})(z-\bar{z})^* = (R_x + R_y) + j(R_{yx} - R_{xy}) \tag{1.B.4}$$

which, as shown above, can be expressed in terms of the covariances and cross-covariance of $\{x,y\}$.

We shall say that the complex variable z has a Gaussian distribution if its real and imaginary parts $\{x,y\}$ are jointly Gaussian. Since z is a function of $\{x,y\}$, its pdf is characterized by the joint pdf of $\{x,y\}$ as in (1.B.3), i.e., in terms of $\{\bar{x},\bar{y},R_x,R_y,R_{xy}\}$. However, we would like to express the pdf of z in terms of its own terms first and second-order moments, i.e., in terms of $\{z,\bar{z},R_z\}$. It turns out that this is not always possible. This is because knowledge of $\{\bar{z},R_z\}$ alone is not enough recover the moments $\{\bar{x},\bar{y},R_x,R_y,R_{xy}\}$. More information is needed in the form of a circularity condition.

To see this, assume we only know $\{\bar{z},R_z\}$. Then this information is enough to recover $\{\bar{x},\bar{y}\}$ since $\bar{z} = \bar{x} + j\bar{y}$. However, the information is not enough to determine the required covariance matrices $\{R_x,R_y,R_{xy}\}$. This is because, as we see from (1.B.4), knowledge of R_z allows us to recover the values of (R_x+R_y) and $(R_{yx}-R_{xy})$ via

$$R_x + R_y \;=\; \mathsf{Re}(R_z), \qquad R_{yx} - R_{xy} \;=\; \mathsf{Im}(R_z) \tag{1.B.5}$$

This information is not sufficient to determine the individual covariances (R_x,R_y,R_{xy}).

In order to be able to uniquely recover $\{R_x,R_y,R_{xy}\}$ from R_z, it is generally assumed that the random variable z satisfies a so-called *circularity* condition. This means that z should satisfy

$$\boxed{\mathsf{E}(z-\bar{z})(z-\bar{z})^{\mathsf{T}} \;=\; 0} \qquad \text{(circularity condition)}$$

with the transposition symbol T used instead of Hermitian conjugation. Knowledge of R_z and this circularity condition are enough to recover $\{R_x,R_y,R_{xy}\}$ from R_z. Indeed, using the fact that

$$\mathsf{E}(z-\bar{z})(z-\bar{z})^{\mathsf{T}} \;=\; (R_x - R_y) + j(R_{yx} + R_{xy})$$

we find that, in view of the circularity assumption, it must hold that $R_x = R_y$ and $R_{xy} = -R_{yx}$. Consequently, combining with (1.B.5), we can solve for $\{R_x,R_y,R_{xy}\}$ to get

$$R_x \;=\; R_y \;=\; \frac{1}{2}\,\mathsf{Re}(R_z) \qquad \text{and} \qquad R_{xy} = -R_{yx} = -\frac{1}{2}\,\mathsf{Im}(R_z) \tag{1.B.6}$$

in terms of the real and imaginary parts of R_z. It follows that the covariance matrix of $\mathrm{col}\{x,y\}$ can be recovered from R_z as

$$R = \frac{1}{2}\begin{bmatrix} \mathsf{Re}(R_z) & -\mathsf{Im}(R_z) \\ \mathsf{Im}(R_z) & \mathsf{Re}(R_z) \end{bmatrix}$$

Actually, it also follows that R should have the symmetry structure

$$R = \begin{bmatrix} R_x & R_{xy} \\ -R_{xy} & R_x \end{bmatrix} \tag{1.B.7}$$

with the same matrix R_x appearing on the diagonal, and with R_{xy} and its negative appearing at the off-diagonal locations. Observe further that when z happens to be scalar-valued, then R_{xy} becomes a scalar, say σ_{xy}, and the condition $R_{xy} = -R_{yx}$ can only hold if $\sigma_{xy} = 0$. That is, the real and imaginary parts of z will need to be independent in the scalar case.

Using the result (1.B.7), we can now verify that the joint pdf of $\{x, y\}$ in (1.B.3) can be rewritten in terms of $\{\bar{z}, R_z\}$ as shown below — compare with (1.B.1) in the real case. Observe in particular that the factors of 2, as well as the square-roots, disappear from the pdf expression in the complex case.

Lemma 1.B.1 (Circular Gaussian random variables) The pdf of a complex-valued circular (or spherically invariant) Gaussian random variable z of dimension p is given by

$$f_z(z) = \frac{1}{\pi^p} \frac{1}{\det R_z} \exp\left\{-(z-\bar{z})^* R_z^{-1} (z-\bar{z})\right\} \tag{1.B.8}$$

Proof: Using (1.B.7) we get

$$\det R = \det(R_x) \cdot \det(R_x + R_{xy} R_x^{-1} R_{xy})$$

Likewise, using the expression $R_z = 2[R_x - jR_{xy}]$ we obtain

$$\begin{aligned} [\det R_z]^2 &= \det(R_z) \cdot \det(R_z^{\mathsf{T}}) \\ &= 2^{2p} \det[R_x(I - jR_x^{-1}R_{xy})] \cdot \det(R_x - jR_{xy}^{\mathsf{T}}) \end{aligned}$$

But $R_{xy}^{\mathsf{T}} = R_{yx} = -R_{xy}$ and, for matrices A and B of compatible dimensions, $\det(AB) = \det(BA)$. Hence,

$$\begin{aligned} [\det R_z]^2 &= 2^{2p} \det R_x \det[(R_x + jR_{xy})(I - jR_x^{-1}R_{xy})] \\ &= 2^{2p} \det(R_x) \cdot \det(R_x + R_{xy} R_x^{-1} R_{xy}) \end{aligned}$$

We conclude that $\det R = 2^{-2p} (\det R_z)^2$. Finally, some algebra will show that the exponents in (1.B.2) and (1.B.8) are identical.

♢

Figure 1.14 plots the pdf of a scalar zero-mean complex-valued and circular-Gaussian random variable z using

$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

i.e., $\sigma_x^2 = \sigma_y^2 = 1$ and $\sigma_{xy} = 0$ so that $\sigma_z^2 = 2$. Therefore, in this example, the real and imaginary parts of z are independent Gaussian random variables with identical variances

When (1.B.8) holds, we can check that uncorrelated jointly Gaussian random variables will also be independent; this is one of the main reasons for the assumption of circularity.

Two Fourth-Order Moment Results

We establish below two useful results concerning the evaluation of fourth-order moments of Gaussian random variables, in both cases of real and complex-valued data. Although these results will only be used later in Sec. 6.5 when we study the mean-square performance of adaptive filters, we list them here because their proofs relate to the earlier discussions on Gaussian random variables.

Lemma 1.B.2 (Fourth-moment of real Gaussian variables) Let x be a real-valued Gaussian random column vector with zero-mean and a diagonal covariance matrix, say $\mathsf{E}\, xx^{\mathsf{T}} = \Lambda$. Then for any symmetric matrix W of compatible dimensions it holds that

$$\mathsf{E}\left\{xx^{\mathsf{T}} W xx^{\mathsf{T}}\right\} = \Lambda \mathrm{Tr}(W\Lambda) + 2\Lambda W \Lambda \tag{1.B.9}$$

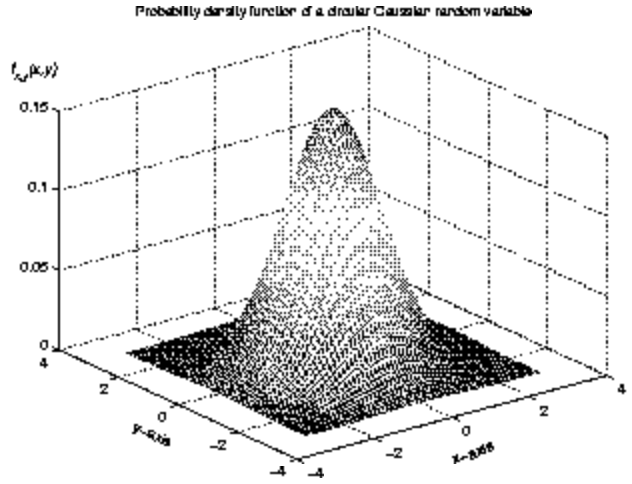

Figure 1.14. A typical plot of the probability density function of a zero-mean scalar and circular-Gaussian random variable.

Proof: The argument is based on the fact that uncorrelated Gaussian random variables are also independent, so that if $x(i)$ is the i-th element of x, then $x(i)$ is independent of $x(j)$ for $i \neq j$. Now let S denote the desired matrix, i.e.,

$$S = \mathsf{E}\left\{xx^{\mathsf{T}}Wxx^{\mathsf{T}}\right\}$$

and let S_{ij} denote its (i,j)-th element. Assume also that x is M-dimensional. Then

$$S_{ij} = \mathsf{E}\left\{x(i)x(j)\left(\sum_{m=0}^{M-1}\sum_{n=0}^{M-1} x(m)W_{mn}x(n)\right)\right\}$$

The right-hand side is nonzero only when there are two pairs of equal indices $\{i=j,\, m=n\}$ or $\{i=m,\, j=n\}$ or $\{i=n,\, j=m\}$. Assume first that $i=j$ (which corresponds to the diagonal elements of S). Then the expectation is nonzero only for $m=n$, i.e.,

$$S_{ii} = \mathsf{E}\left\{x^2(i)\sum_{m=0}^{M-1} W_{mm}x^2(m)\right\} = \sum_{m=0}^{M-1} W_{mm}\mathsf{E}\left\{x^2(i)x^2(m)\right\} = \lambda_i \mathrm{Tr}(W\Lambda) + 2W_{ii}\lambda_i^2$$

where we used the fact that for a zero-mean *real* scalar-valued Gaussian random variable a we have $\mathsf{E}\,a^4 = 3\big(\mathsf{E}\,a^2\big)^2 = 3\sigma_a^4$, where σ_a^2 denotes the variance of a, $\sigma_a^2 = \mathsf{E}\,a^2$. We are also denoting the diagonal entries of Λ by $\{\lambda_i\}$.

For the off-diagonal elements of S (i.e., for $i \neq j$), we must have either $i=n$, $j=m$, or $i=m$, $j=n$, so that

$$\begin{aligned} S_{ij} &= \mathsf{E}\left\{x(i)x(j)\big(x(i)W_{ij}x(j)\big)\right\} + \mathsf{E}\left\{x(i)x(j)\big(x(j)W_{ji}x(i)\big)\right\} \\ &= \big(W_{ij}+W_{ji}\big)\mathsf{E}\left\{x^2(i)x^2(j)\right\} = \big(W_{ij}+W_{ji}\big)\lambda_i\lambda_j \end{aligned}$$

Using the fact that W is symmetric, so that $W_{ij} = W_{ji}$, and collecting the expressions for S_{ij}, in both cases of $i=j$ and $i \neq j$, into matrix form we get the desired result (1.B.9).

♢

The equivalent result for complex-valued circular Gaussian random variables is the following. The only difference is an additional factor of 2 in (1.B.9).

Lemma 1.B.3 (Fourth-moment of complex Gaussian variables) Let z be a circular complex-valued Gaussian random column vector with zero-mean and a diagonal covariance matrix, say $\mathsf{E}\,zz^* = \Lambda$. Then for any Hermitian matrix W of compatible dimensions it holds that

$$\mathsf{E}\left\{zz^*Wzz^*\right\} = \Lambda\mathsf{Tr}(W\Lambda) + \Lambda W\Lambda \tag{1.B.10}$$

Proof: The argument is the same as in the proof of the previous lemma, with the main difference being the fact that the fourth-order moment of a zero-mean *complex* scalar-valued circular random variable a, of variance $\sigma_a^2 = \mathsf{E}\,|a|^2$, is given by $\mathsf{E}\,|a|^4 = 2\big(\mathsf{E}\,|a|^2\big)^2 = 2\sigma_a^4$. Indeed, since in this case

$$S_{ij} = \mathsf{E}\left\{z(i)z^*(j)\left(\sum_{m=0}^{M-1}\sum_{n=0}^{M-1} z^*(m)W_{mn}z(n)\right)\right\}$$

we find that for $i = j$ we have

$$S_{ii} = \mathsf{E}\left\{|z(i)|^2\sum_{m=0}^{M-1} W_{mm}|z(m)|^2\right\} = \sum_{m=0}^{M-1} W_{mm}\mathsf{E}\left\{|z(i)|^2|z(m)|^2\right\} = \lambda_i\mathsf{Tr}(W\Lambda) + W_{ii}\lambda_i^2$$

Moreover, for $i \neq j$,

$$\begin{aligned} S_{ij} &= \mathsf{E}\left\{z(i)z^*(j)\big(z^*(i)W_{ij}z(j)\big)\right\} + \mathsf{E}\left\{z(i)z^*(j)\big(z^*(j)W_{ji}z(i)\big)\right\} \\ &= \mathsf{E}\left\{|z(i)|^2W_{ij}|z(j)|^2\right\} + 0 = \lambda_iW_{ij}\lambda_j \end{aligned}$$

The zero in the second equality follows from the circularity assumption on z, namely, $\mathsf{E}\,zz^{\mathsf{T}} = 0$, which guarantees

$$\mathsf{E}\,z^2(i) = 0 \quad \text{for all } i$$

$\diamondsuit$