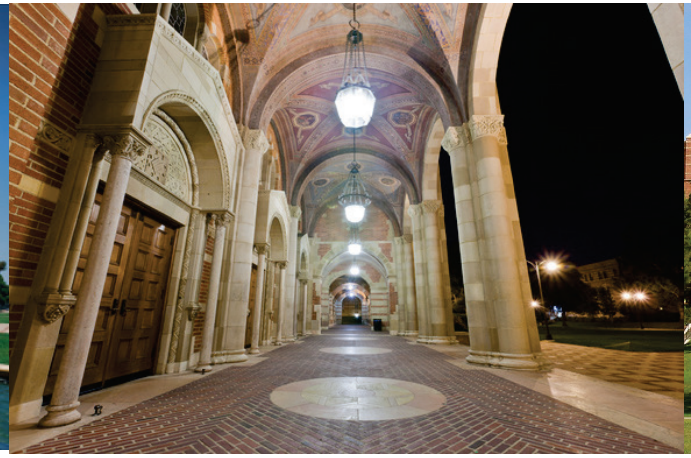




EE210A: Adaptation and Learning

Professor Ali H. Sayed

MOTIVATION: WHAT IS ADAPTATION?

- Adaptation endows systems and devices with **learning** abilities.

- The human being is a fantastic example of an adaptive system; a continuous learning experience from childhood to adulthood.

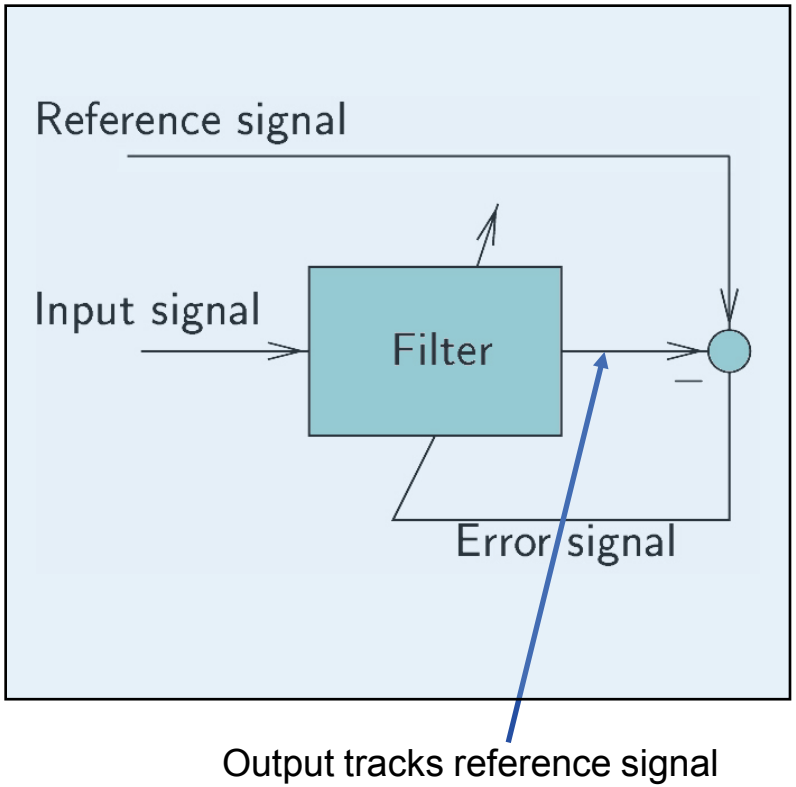
- Broadly, adaptive systems **interact** with their environment; **learn** from the interaction; and adjust their behavior for optimal **performance**.

WHAT IS AN ADAPTIVE FILTER?

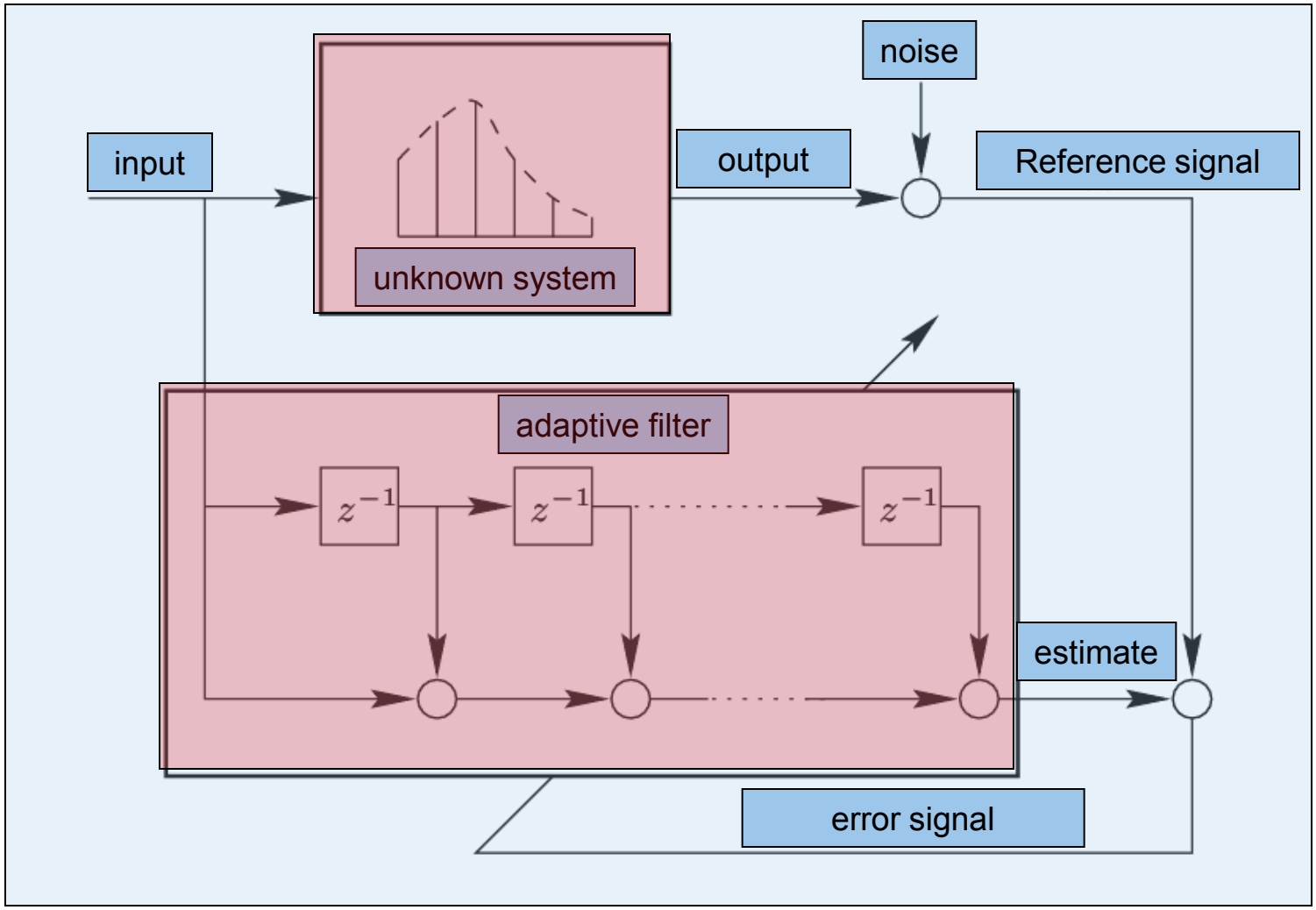
- It is generally a digital filter whose coefficients vary in time according to certain **rules**.

- **Objective:** The filter output should **track** a reference signal in a certain optimal manner.

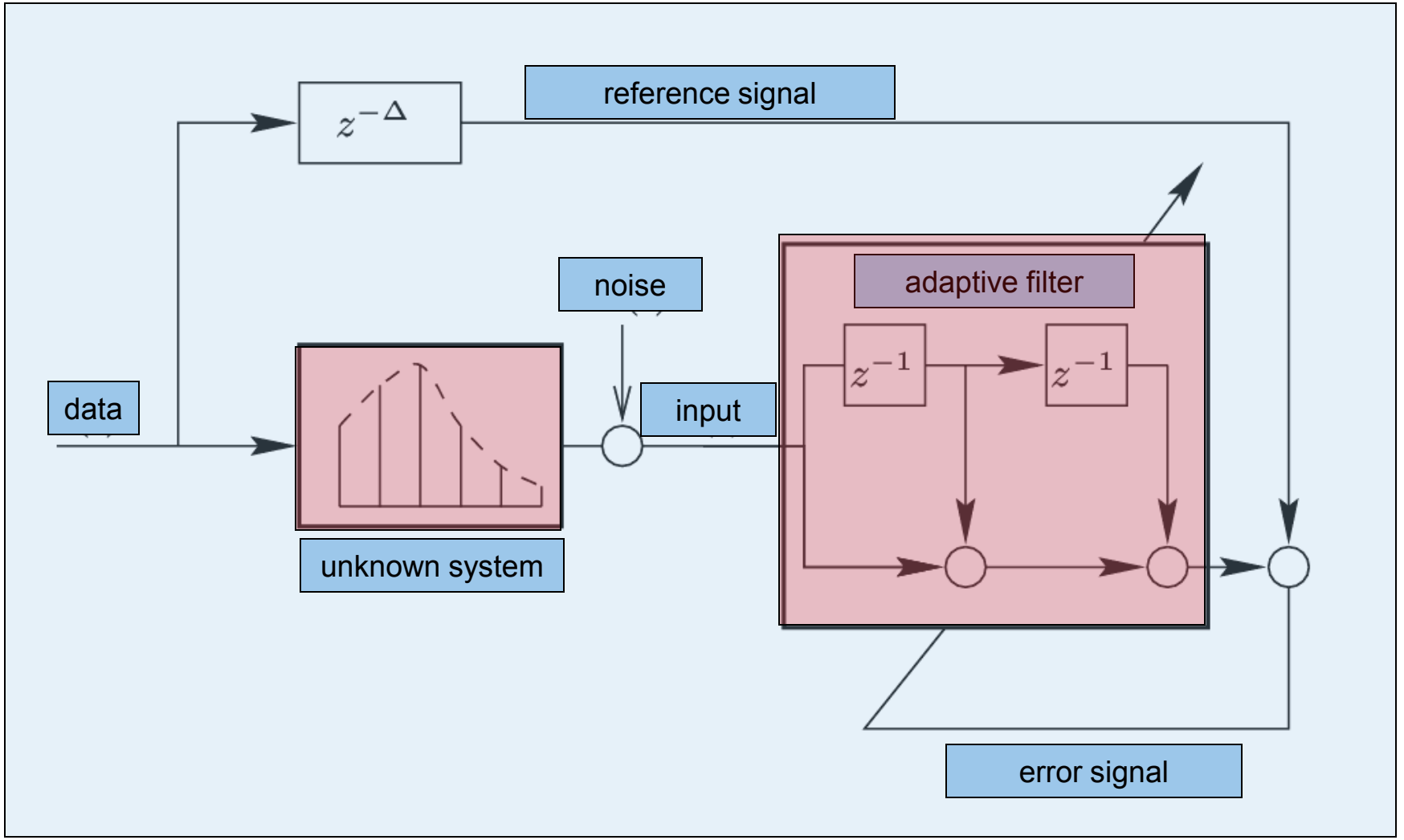
- **Property:** The filter is able to **respond** to variations in the statistical properties of its signals.



APPLICATION: CHANNEL ESTIMATION



APPLICATION: CHANNEL EQUALIZATION



RICH HISTORY

- Early results can be traced back to the 1950s; maybe 1940s with the contributions of Kolmogorov, Wiener, Krein, and Levinson in related areas:

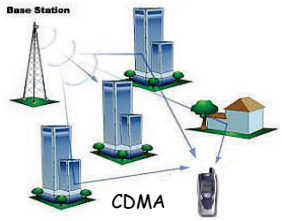
  - ✓ Plackett (1950): "modern" RLS or recursive least-squares.
  - ✓ **Robbins and Monroe** (1951): Stochastic approximation.
  - ✓ **Widrow and Hoff** (1960): LMS filter.
  - ✓ Kalman (1960): Kalman filter.

- **Gauss** (1795, at the age of 18) was the forerunner in formulating the powerful least squares criterion and its recursive version.

Carl F. Gauss
(1777-1855)

MODERN TECHNOLOGIES

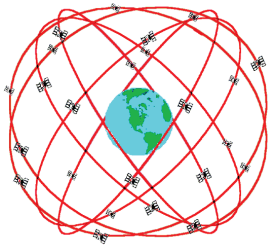
- Interest in communications, bioengineering & information technologies have motivated **heightened research** on adaptive systems in order to:
  - ✓ Understand their limits of performance.
  - ✓ Develop variants that meet more stringent specifications.

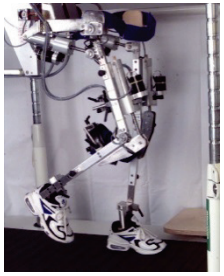


- **Applications**:
  - ✓ Wireless networks.
  - ✓ Wireless communications.
  - ✓ Wearable computing.
  - ✓ Biometrics.
  - ✓ Voice over IP.
  - ✓ DSL.
  - ✓ Home phone networking.
  - ✓ HDTV.

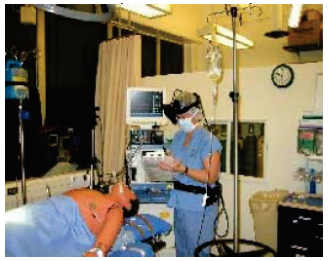

WLAN


HRL wearable system


GPS

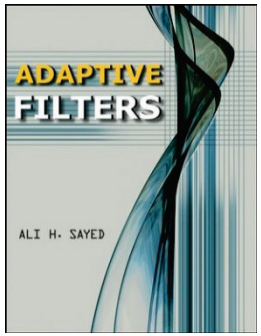

Lokomat ETH


E911 eLocate UCLA


NOMAD wearable system

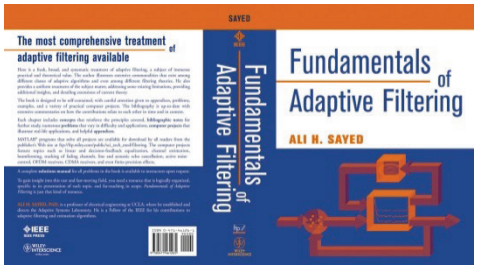
REFERENCES

- A. H. Sayed, *Adaptive Filters*, Wiley, NJ, 2008. (textbook used for this course)



- A. H. Sayed, *Adaptation, Learning, and Optimization over Networks,* NOW Publishers, 2014.

- A. H. Sayed, *Fundamentals of Adaptive Filtering*, Wiley, NJ, 2003.



- S. Haykin, *Adaptive Filter Theory*, Prentice Hall, NJ, 2000.

- B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice Hall, NJ, 1985.

LECTURE #01

SCALAR AND REAL-VALUED MEAN-SQUARE ERROR ESTIMATION

Sections in order: A.1, 1.1, A.2, 1.2, 1.3

A.1 VARIANCE OF A RANDOM VARIABLE

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

$$\boxed{\bar{x} \triangleq \mathsf{E}\,\boldsymbol{x}, \qquad \sigma_x^2 \triangleq \mathsf{E}\,(\boldsymbol{x}-\bar{x})^2 = \mathsf{E}\,\boldsymbol{x}^2 - \bar{x}^2} \tag{A.1}$$

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 $\sigma_x^2 = \mathsf{E}\,\boldsymbol{x}^2$. 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 values to expect. These two situations are illustrated in Figs. A.1 and A.2 for two different probability density functions.

GAUSSIAN DISTRIBUTION

Figure A.1 plots the probability density function (pdf) of a Gaussian random variable 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

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

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

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

i.e., the probability of $\boldsymbol{x}$ assuming values inside the interval $[a, b]$. From Fig. A.1 we observe that the smaller the variance of $\boldsymbol{x}$, the more concentrated its pdf is around its mean.

GAUSSIAN DISTRIBUTION

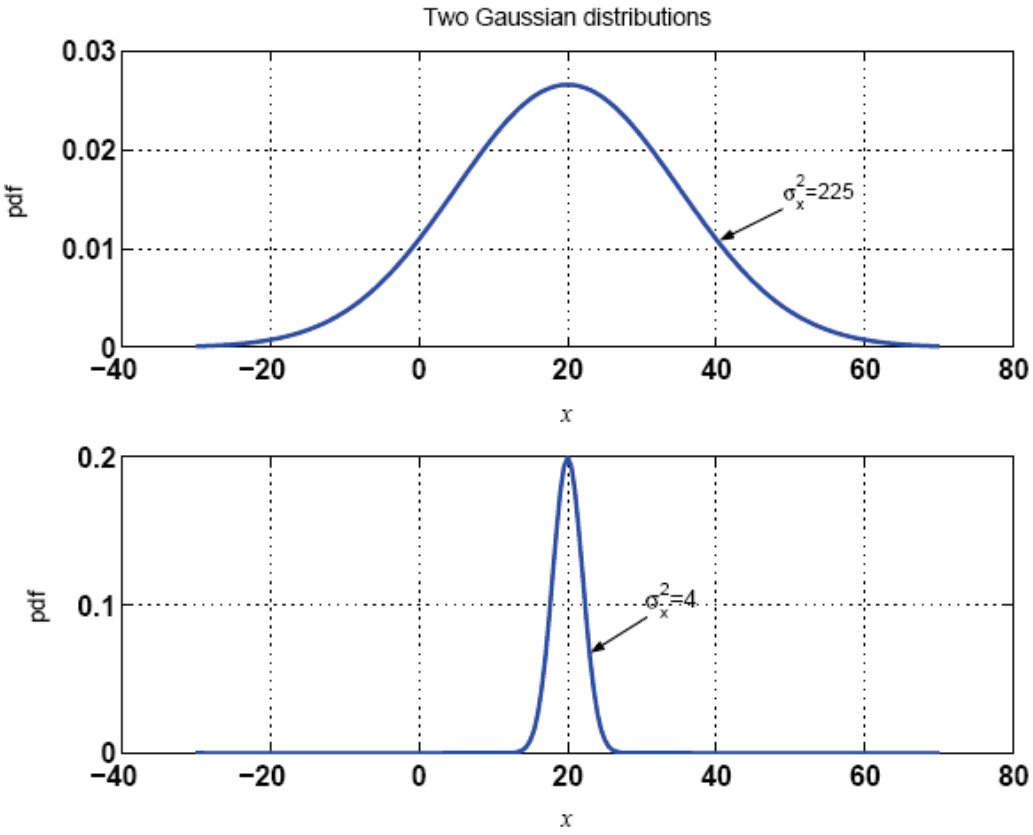


FIGURE A.1 The figure shows the plots of the probability density functions of a Gaussian random variable x with mean $\bar{x} = 20$, variance $\sigma_x^2 = 225$ in the top plot, and variance $\sigma_x^2 = 4$ in the bottom plot.

RAYLEIGH DISTRIBUTION

Figure A.2 provides similar plots for a random variable x with a Rayleigh distribution, namely, with a pdf given by

$$\boxed{f_{\boldsymbol{x}}(x) = \frac{x}{\alpha^2}\, e^{-\frac{x^2}{2\alpha^2}}, \qquad x \geq 0, \quad \alpha > 0} \tag{A.3}$$

where α is a positive parameter that determines the mean and variance of x according to the expressions (see Prob. I.1):

$$\bar{x} = \alpha \sqrt{\frac{\pi}{2}}, \qquad \sigma_x^2 = \left(2 - \frac{\pi}{2}\right) \alpha^2 \tag{A.4}$$

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. A.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 $\sigma_x^2 = 2.4592$.

Lord Rayleigh
(1842-1919)

RAYLEIGH DISTRIBUTION

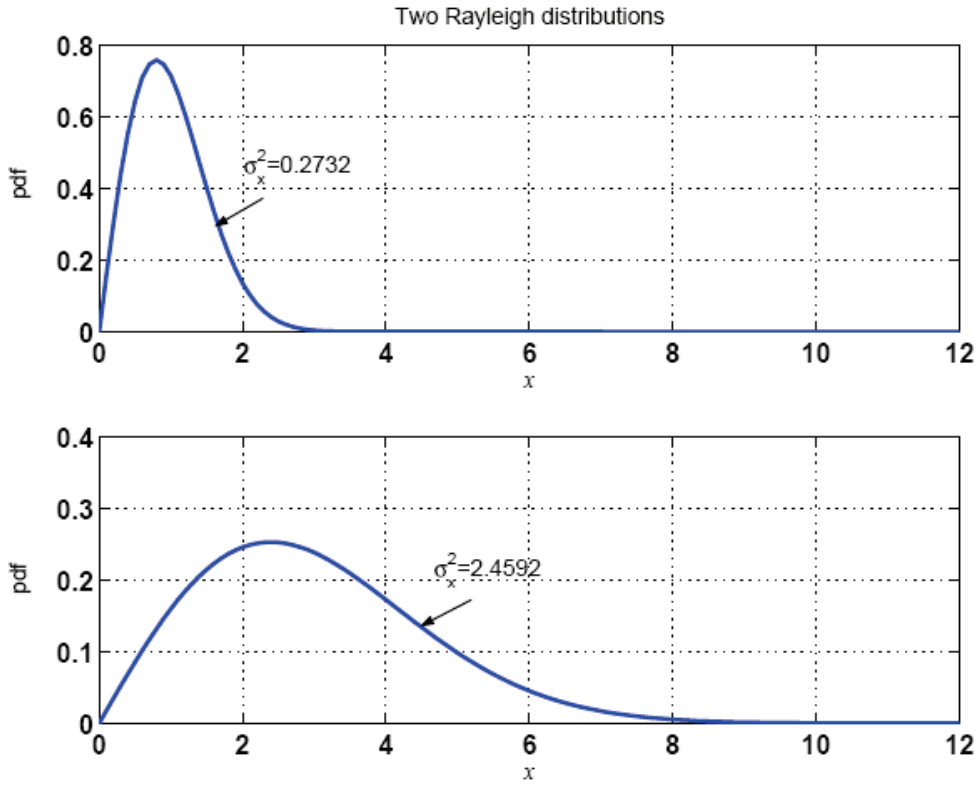


FIGURE A.2 The figure shows the plots of the probability density functions of a Rayleigh random variable x with mean $\bar{x} = 1$ and variance $\sigma_x^2 = 0.2732$ in the top plot, and mean $\bar{x} = 3$ and variance $\sigma_x^2 = 2.4592$ in the bottom plot.

CHEBYSHEV'S INEQUALITY

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. I.2 and I.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

$$\boxed{P(|\boldsymbol{x} - \bar{x}| \geq \delta) \leq \sigma_x^2/\delta^2} \tag{A.5}$$

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 (A.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 $(\bar{x} - 5\sigma_x, \bar{x} + 5\sigma_x)$.

CHEBYSHEV'S INEQUALITY

Actually, the bound that is provided by Chebyshev's inequality is generally not tight. Consider, for example, a zero-mean Gaussian random variable x with variance σ_x^2 and choose $\delta = 2\sigma_x$. Then, from Chebyshev's inequality (A.5) we would obtain

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

whereas direct evaluation of the integral

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

yields

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

P. L. Chebyshev (1821-1894)

ZERO-VARIANCE RANDOM VARIABLE

Remark A.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,

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

This is because, for any small $\delta > 0$, we obtain from (A.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. 1.2).

$\diamond$

INITIAL ESTIMATION PROBLEM

1.1 ESTIMATION WITHOUT OBSERVATIONS

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

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

MEAN-SQUARE ERROR

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

$$\tilde{\boldsymbol{x}} \triangleq \boldsymbol{x} - \hat{x}$$

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.,

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

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

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

ARGUMENT

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

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

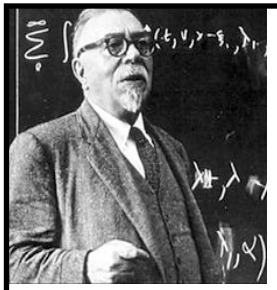
The choice of $\hat{x}$ that minimizes the m.s.e. is now evident. Only the term $(\bar{x}-\hat{x})^2$ is dependent on $\hat{x}$ and this term can be annihilated by choosing $\hat{x}=\bar{x}$. The resulting minimum mean-square error (m.m.s.e.) is then

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

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, $\hat{x}=\bar{x}$.



N. Wiener (1894-1964)

A. Kolmogorov (1903-1987)

WHY MSE?

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

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

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

MMSE PERFORMANCE

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

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

so that the estimate $\hat{x} = \bar{x}$ does not reduce our initial uncertainty about 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 limited in this case. Clearly, we are more interested in estimation procedures that result in error variances that are smaller than the original signal variance. We shall discuss one such procedure in the next section.

INTUITION

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 1.1 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

Example 1.1 (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 of 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.

$\diamond$

1.2 ESTIMATION GIVEN DEPENDENT OBSERVATIONS

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

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

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

for some function $h(\cdot)$ to be determined.

MSE CRITERION

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 $\boldsymbol{x}$,

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

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

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

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

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

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

A.2 DEPENDENT RANDOM VARIABLES

The dependency between two real-valued random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ is characterized by their *joint* probability density function (pdf). Thus let $f_{\boldsymbol{x},\boldsymbol{y}}(x, y)$ denote the joint (continuous) 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

$$\boxed{f_{\boldsymbol{x},\boldsymbol{y}}(x, y) = f_{\boldsymbol{y}}(y)\, f_{\boldsymbol{x}|\boldsymbol{y}}(x|y) = f_{\boldsymbol{x}}(x)\, f_{\boldsymbol{y}|\boldsymbol{x}}(y|x)} \tag{A.6}$$

INDEPENDENT RANDOM VARIABLES

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 $\mathsf{E}\,\boldsymbol{xy} = \mathsf{E}\,\boldsymbol{x}\mathsf{E}\,\boldsymbol{y}$. 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:

$$\begin{aligned}
\sigma_{xy} &\triangleq \mathsf{E}\,(\boldsymbol{x} - \bar{x})(\boldsymbol{y} - \bar{y}) \\
&= \mathsf{E}\,\boldsymbol{xy} - \bar{x}\bar{y} \\
&= \mathsf{E}\,\boldsymbol{x}\;\mathsf{E}\,\boldsymbol{y} - \bar{x}\bar{y} \\
&= 0
\end{aligned}$$

The converse statement is not true: uncorrelated random variables can be dependent. Consider the following example. Let $\boldsymbol{\theta}$ be a random variable that is uniformly distributed over the interval $[0, 2\pi]$. Define the zero-mean random variables $\boldsymbol{x} = \cos\boldsymbol{\theta}$ and $\boldsymbol{y} = \sin\boldsymbol{\theta}$. Then $\boldsymbol{x}^2 + \boldsymbol{y}^2 = 1$ so that $\boldsymbol{x}$ and $\boldsymbol{y}$ are dependent. However, $\mathsf{E}\,\boldsymbol{xy} = \mathsf{E}\,\cos\boldsymbol{\theta}\sin\boldsymbol{\theta} = 0.5\mathsf{E}\,\sin 2\boldsymbol{\theta} = 0$, so that $\boldsymbol{x}$ and $\boldsymbol{y}$ are uncorrelated.

OPTIMAL MSE ESTIMATOR

Theorem 1.1 (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 $\mathsf{E}\,\tilde{x}^2 \;=\; \sigma_x^2 \;-\; \sigma_{\hat{x}}^2$.

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. I.4):

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

ARGUMENT

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

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

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

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

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

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

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

ARGUMENT

Using this intermediate result, we return to the cost function (1.3), 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 orthogonality property (1.5) we conclude that

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

Now 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

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

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

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

ARGUMENT AND INTUITION

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

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

◇

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

EXAMPLE

Example 1.2 (Noisy measurement of a binary signal)

Let us return to Ex. 1.1, 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

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

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

According to Thm. 1.1, the optimal estimate of x given an observation of y is

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

EXAMPLE

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

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

In this example, we have

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

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

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

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

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

EXAMPLE

Using (A.6) we get

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

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

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

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

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

The result is represented schematically in Fig. 1.1.

EXAMPLE

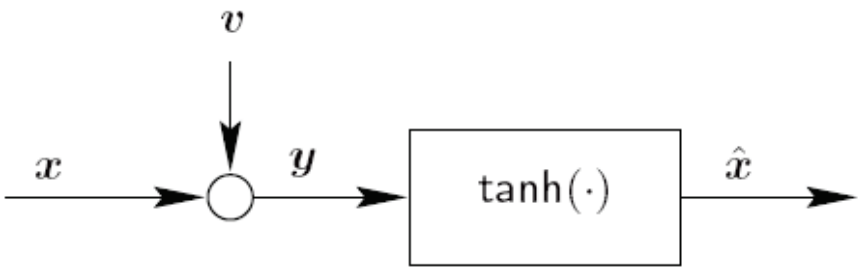


FIGURE 1.1 Optimal estimation of a BPSK signal embedded in unit-variance additive Gaussian noise.

Figure 1.2 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 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:

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

EXAMPLE

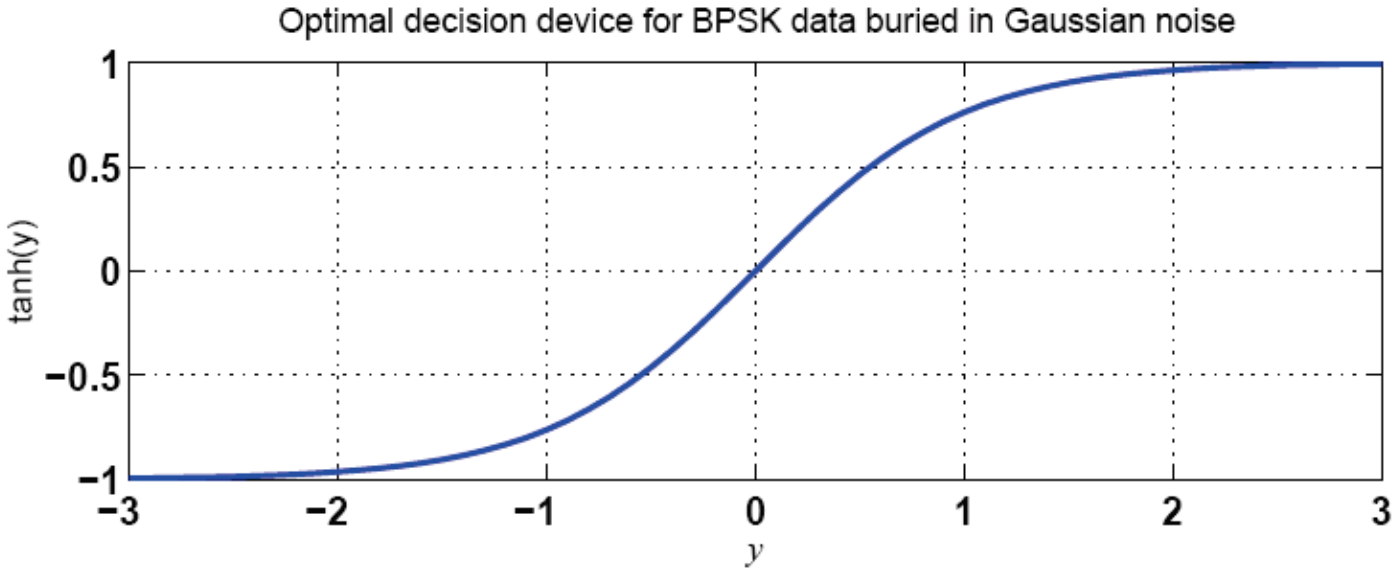


FIGURE 1.2 A plot of the function $\tanh(y)$.

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

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

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

EXAMPLE

We therefore have a situation where the optimal estimator, although known in closed form, does not solve the original problem of recovering the symbols $\pm 1's$ 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:

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

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

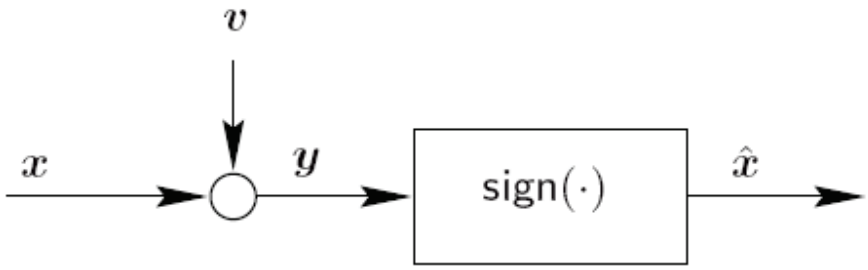


FIGURE 1.3 Sub-optimal estimation of a BPSK signal embedded in unit-variance additive Gaussian noise.

1.3 ORTHOGONALITY PRINCIPLE

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

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

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

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

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

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

where the symbol $\perp$ is used to signify that the two random variables are orthogonal; a schematic representation of this orthogonality property is shown in Fig. 1.4.

ORTHOGONALITY PRINCIPLE

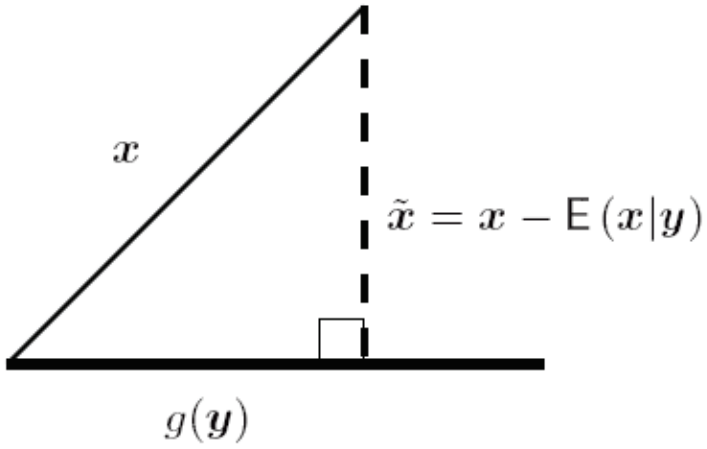


FIGURE 1.4 The orthogonality condition: $\tilde{x} \perp g(y)$.

Relation (1.13) admits the following interpretation. It states that the optimal estimator $\hat{x} = \mathsf{E}\,(x|y)$ is such that the resulting error, $\tilde{x}$, is orthogonal to (and, in fact, also uncorrelated with) any transformation of the data y. In other words, the optimal estimator is such that no matter how we modify the data y, there is no way we can extract additional information from the data in order to reduce the variance of $\tilde{x}$ any further. This is because any additional processing of y will remain uncorrelated with $\tilde{x}$.

OROTHOGONALITY PRINCIPLE

The second orthogonality property (1.6) is a special case of (1.13). 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 (1.13) since $\hat{\boldsymbol{x}}$ is a function of $\boldsymbol{y}$ by virtue of the result $\hat{\boldsymbol{x}} = \mathsf{E}\,(\boldsymbol{x}|\boldsymbol{y})$.

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

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

ARGUMENT

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

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

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

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

$\diamond$