27 Mean-Square-Error Inference

Inference deals with the estimation of hidden parameters or random variables from observations of other related variables. In this chapter, we study the basic, yet fundamental, problem of inferring an unknown random quantity from observations of another random quantity by using the *mean-square-error* (MSE) criterion. Several other design criteria can be used for inference purposes besides MSE, such as the mean-absolute error (MAE) and the maximum a-posteriori (MAP) criteria. We will encounter these possibilities in future chapters, starting with the next chapter. We initiate our discussions of inference problems though by focusing on the MSE criterion due to its mathematical tractability and because it sheds light on several important questions that arise in the study of inference problems in general.

In our treatment of inference problems, we will encounter three broad formulations:

(a) In one instance, we will model the unknown as a *random* variable, denoted by the boldface symbol $\boldsymbol{x}$. The objective will be to predict (or infer) the value of $\boldsymbol{x}$ from observations of a related variable $\boldsymbol{y}$. The predictor or estimator for $\boldsymbol{x}$ will be denoted by $\widehat{\boldsymbol{x}}$. Different design criteria will lead to different constructions for $\widehat{\boldsymbol{x}}$. In this chapter, we discuss the popular MSE criterion.

(b) In another instance, we will continue to model the unknown $\boldsymbol{x}$ as a random variable but will limit its values to *discrete* levels, such as having $\boldsymbol{x}$ assume the values $+1$ or -1. This type of formulation is prevalent in classification problems and will be discussed at length in future chapters, starting with the next chapter, where we examine the Bayes classifier.

(c) In a third instance, we will model the unobservable as an unknown *constant*, denoted by the Greek symbol θ, rather than a random variable. This type of problem is frequent in applications requiring fitting models onto data and will be discussed in future chapters, for example, when we introduce the maximum-likelihood and expectation maximization paradigms.

27.1 INFERENCE WITHOUT OBSERVATIONS

We consider first a simple yet useful estimation problem, which relates to estimating a random variable $\boldsymbol{x} \in \mathbb{R}$ when the only information that is available about $\boldsymbol{x}$ is its mean, $\bar{x}$. Our objective is to estimate the value that $\boldsymbol{x}$ will assume in a given experiment. We denote the *estimate* for $\boldsymbol{x}$ by the notation $\widehat{x}$; it is a *deterministic* quantity (i.e., a number). But how do we come up with a value for $\widehat{x}$? 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 inference problem. To answer them, we first need to choose a cost (also called *risk*) function to penalize the estimation error. The resulting estimate $\widehat{x}$ will be optimal only in the sense that it leads to the smallest cost or risk value. Different choices for the cost function will generally lead to different choices for $\widehat{x}$, each of which will be optimal in its own way.

27.1.1 Problem Formulation

The design criterion we study first is the famed MSE criterion; several other criteria are possible and we discuss other important choices in the next chapter. The MSE criterion is based on introducing the error signal:

$$\widetilde{\boldsymbol{x}} \triangleq \boldsymbol{x} - \widehat{x} \tag{27.1}$$

and on determining $\widehat{x}$ by minimizing the MSE, which is defined as the mean of the squared error $\widetilde{\boldsymbol{x}}^2$:

$$\widehat{x} \triangleq \operatorname*{argmin}_{\widehat{x}} \mathbb{E}\,\widetilde{\boldsymbol{x}}^2 \tag{27.2}$$

The error $\widetilde{\boldsymbol{x}}$ is a random variable since $\boldsymbol{x}$ is random. The resulting estimate, $\widehat{x}$, is called the *least-mean-squares estimate* (LMSE) of $\boldsymbol{x}$. For added emphasis, we could have written $\widehat{x}_{\text{MSE}}$, with the subscript MSE, in order to highlight the fact that this is an estimate for $\boldsymbol{x}$ that is based on minimizing the MSE criterion defined by (27.2). This notation is unnecessary in this chapter because we will be discussing mainly the MSE criterion. However, in later chapters, when we introduce other design criteria, it will become necessary to use the MSE subscript to distinguish the MSE estimate, $\widehat{x}_{\text{MSE}}$, from other estimates such as $\widehat{x}_{\text{MAP}}$, $\widehat{x}_{\text{MAE}}$, or $\widehat{x}_{\text{ML}}$, where the subscripts MAP, MAE, and ML will be referring to maximum a-posteriori, mean-absolute error, and maximum-likelihood estimators.

Returning to (27.2), it is immediate to verify that the solution $\widehat{x}$ is given by

$$\boxed{\widehat{x} = \bar{x}} \tag{27.3}$$

and that the resulting minimum mean-square error (MMSE) is

$$\boxed{\text{MMSE} \;=\; \mathbb{E}\,\widetilde{\boldsymbol{x}}^2 = \sigma_x^2} \tag{27.4}$$

We therefore say that the best estimate of a random variable (best from the perspective of MSE estimation) when only its mean is known is the mean value of the random variable itself.

Proof of (27.3)–(27.4) using differentiation: We expand the MSE from (27.2) to get

$$\mathbb{E}\,\widetilde{\boldsymbol{x}}^2 \;=\; \mathbb{E}\,(\boldsymbol{x}-\widehat{x})^2 = \mathbb{E}\,\boldsymbol{x}^2 - 2\bar{x}\widehat{x} + \widehat{x}^2 \tag{27.5}$$

and then differentiate the right-hand side with respect to the unknown $\widehat{x}$. Setting the derivative to zero gives

$$\frac{\partial\,\mathbb{E}\,(\boldsymbol{x}-\widehat{x})^2}{\partial\widehat{x}} \;=\; -2\bar{x} + 2\widehat{x} = 0 \;\Longrightarrow \widehat{x} = \bar{x} \tag{27.6}$$

This choice for $\widehat{x}$ minimizes the MSE since the risk (27.5) is quadratic in $\widehat{x}$.

■

Proof of (27.3)–(27.4) using completion-of-squares: An alternative argument to arrive at the same conclusion relies on the use of a completion-of-squares step. It is immediate to verify that the risk function in (27.2) can be rewritten in the following form, by adding and subtracting $\bar{x}$:

$$\begin{aligned}\mathbb{E}\,\widetilde{\boldsymbol{x}}^2 &= \mathbb{E}\,\left((\boldsymbol{x}-\bar{x}) + (\bar{x}-\widehat{x})\right)^2 \\ &= \mathbb{E}\,(\boldsymbol{x}-\bar{x})^2 \;+\; (\bar{x}-\widehat{x})^2 \;+\; 2\,\underbrace{\mathbb{E}\,(\boldsymbol{x}-\bar{x})}_{=0}(\bar{x}-\widehat{x}) \\ &= \sigma_x^2 \;+\; (\bar{x}-\widehat{x})^2\end{aligned} \tag{27.7}$$

The above result expresses the risk as the sum of two *nonnegative* terms where only the second term depends on the unknown, $\widehat{x}$. It is clear that the choice $\widehat{x} = \bar{x}$ annihilates the second term and results in the smallest possible value for the MSE. This value is referred to as the MMSE and is equal to σ_x^2.

■

27.1.2 Interpretation

There are good reasons for using the MSE criterion (27.2). The simplest one perhaps is that the criterion is amenable to mathematical manipulations, more so than any other criterion. In addition, the criterion is attempting to force the estimation error, $\widetilde{\boldsymbol{x}}$, to assume values close to its mean, which is zero since

$$\mathbb{E}\,\widetilde{\boldsymbol{x}} = \mathbb{E}\,(\boldsymbol{x}-\widehat{x}) = \mathbb{E}\,(\boldsymbol{x}-\bar{x}) = \bar{x}-\bar{x} = 0 \tag{27.8}$$

Therefore, by minimizing $\mathbb{E}\,\widetilde{\boldsymbol{x}}^2$, we are in effect minimizing the variance of the error, $\widetilde{\boldsymbol{x}}$. Then, in view of the discussion in Section 3.2 regarding the interpretation of the variance of a random variable, we find that the MSE criterion is attempting to increase the likelihood of small errors.

The effectiveness of the estimation procedure (27.2) can be measured by examining the value of the resulting minimum risk, which is the variance of the estimation error, denoted by $\sigma_{\widetilde{x}}^2 = \mathbb{E}\,\widetilde{\boldsymbol{x}}^2$. The above discussion tells us that

$$\sigma_{\widetilde{x}}^2 \;=\; \sigma_x^2 \tag{27.9}$$

so that the estimate $\widehat{x} = \bar{x}$ *does not* reduce our initial uncertainty about $\boldsymbol{x}$: The error variable has the same variance as $\boldsymbol{x}$ itself. We therefore find that, in this initial scenario, the performance of the MSE design procedure is limited. We are interested in estimation procedures that result in error variances that are smaller than the original signal variance. We discuss one such procedure in the next section.

The reason for the poor performance of the estimate $\widehat{x} = \bar{x}$ lies in the lack of more sophisticated prior information about $\boldsymbol{x}$. Note that result (27.3) simply tells us that the best we can do, in the absence of any other information about a random variable $\boldsymbol{x}$, other than its mean, is to use the mean value of $\boldsymbol{x}$ as our estimate. This result is at least intuitive. After all, the mean value of a random variable is, by definition, an indication of the value that we expect to observe on average in repeated experiments. Hence, in answer to the question, "*What is the best guess for* $\boldsymbol{x}$*?*" the analysis tells us that the best guess is what we would expect for $\boldsymbol{x}$ on average! This is a circular answer, but one that is at least consistent with intuition.

Example 27.1 (**Guessing the class of an image**) Assume a box contains an equal number of images of cats and dogs. An image is selected at random from the box and a random variable $\boldsymbol{x}$ is associated with this experiment. The variable $\boldsymbol{x}$ assumes the value $x = +1$ if the selected image is a cat and it assumes the value $x = -1$ otherwise. We say that $\boldsymbol{x}$ represents the class (or label) variable: Its value specifies whether the selected image belongs to one class ($+1$ corresponding to cats) or the other (-1 corresponding to dogs). It is clear that $\boldsymbol{x}$ is a *binary* random variable assuming the values ± 1 with equal probability. Then,

$$\bar{x} = \frac{1}{2} \times (+1) \;+\; \frac{1}{2} \times (-1) \;=\; 0 \tag{27.10}$$

$$\sigma_x^2 = \mathbb{E}\,\boldsymbol{x}^2 = 1 \tag{27.11}$$

If a user were to predict the class of the image that will be selected from the box beforehand then, according to the MSE criterion, the best estimate for $\boldsymbol{x}$ will be $\widehat{x} = \bar{x} = 0$. This estimate value is neither $+1$ nor -1. This example shows that the LMSE estimate does not always lead to a meaningful solution! In this case, using $\widehat{x} = 0$ is not useful in guessing whether the realization for $\boldsymbol{x}$ will be $+1$ or -1. If we could incorporate into the design procedure some additional information, besides the mean of $\boldsymbol{x}$, then we could perhaps come up with a better prediction for the class of the image.

Example 27.2 (**Predicting a crime statistic**) The US Federal Bureau of Investigation (FBI) publishes statistics on the crime rates in the country on an annual basis. Figure 27.1 plots the burglary rates per 100,000 inhabitants for the period 1997–2016. Assume we model the annual burglary rate as a random variable $\boldsymbol{x}$ with some mean $\bar{x}$. By examining the plot, we find that this assumption is more or less reasonable only over the shorter range 2000–2009 during which the burglary rate remained practically flat

with fluctuations around some nominal average value. The rates are declining before 2000 and after 2010. Assume we did not know the burglary rate for the year 2010 and wanted to predict its value from the burglary rates observed in prior years. In this example, the probability distribution of $\boldsymbol{x}$ is not known to evaluate its mean $\bar{x}$. Instead, we have access to measurements for the years 1997–2015. We can use the data from the years 2000–2009 to compute a sample mean and use it to predict $x(2010)$:

$$\widehat{x}(2010) = \frac{1}{10}\sum_{n=2000}^{2009} x(n) \approx 732.6 \tag{27.12}$$

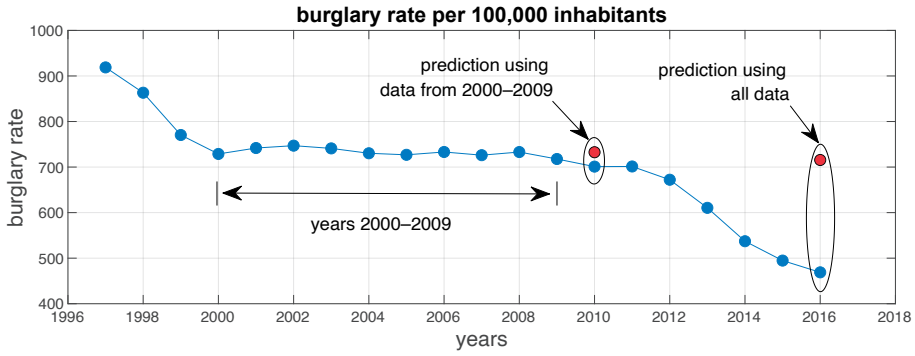

Figure 27.1 Plot of the annual burglary rates per 100,000 inhabitants in the United States during 1997–2016. Source: US Criminal Justice Information Services Division. Data found at `https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-1`.

The value 732.6 is close enough to the actual burglary rate observed for 2010, which is 701. If we were instead to predict the burglary rate for the year 2016 by using the data for the entire period 1997–2015, we would end up with

$$\widehat{x}(2016) \approx 715.5 \tag{27.13}$$

which is clearly a bad estimate since the actual value is 468.9.

This example illustrates the fact that we will often be dealing with distributions that vary (i.e., drift) over time for various reasons, such as changing environmental conditions or, in the case of this example, crime deterrence policies that may have been put in place. This possibility necessitates the development of inference techniques that are able to adapt to variations in the statistical properties of the data in an *automated* manner. In this example, the statistical properties of the data during the period 2000–2009 are clearly different from the periods before 2000 and after 2009.

27.2 INFERENCE WITH OBSERVATIONS

Let us examine next the case in which more is known about the random variable $\boldsymbol{x}$, beyond its mean. 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 way.

27.2.1 Conditional Mean Estimator

Given two dependent random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$, we pose the problem of determining the LMSE *estimator* for $\boldsymbol{x}$ from an observation of $\boldsymbol{y}$. 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 $\widehat{\boldsymbol{x}}$; it is now a *random* variable since it will be a function of $\boldsymbol{y}$, i.e.,

$$\widehat{\boldsymbol{x}} = c(\boldsymbol{y}) \tag{27.14}$$

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

$$\widehat{x} \;=\; c(\boldsymbol{y})\Big|_{\boldsymbol{y}=y} \;=\; c(y) \tag{27.15}$$

Different occurrences for $\boldsymbol{y}$ will lead to different estimates $\widehat{x}$. In Section 27.1 we did not need to make this distinction between an estimator $\widehat{\boldsymbol{x}}$ and an estimate $\widehat{x}$. There we sought directly an estimate $\widehat{x}$ for $\boldsymbol{x}$ since we did not have access to a random variable $\boldsymbol{y}$; we only had access to the deterministic quantity $\bar{x}$.

The criterion we use to determine the estimator $\widehat{\boldsymbol{x}}$ will continue to be the same MSE criterion. We define the error signal:

$$\widetilde{\boldsymbol{x}} \;\triangleq\; \boldsymbol{x} - \widehat{\boldsymbol{x}} \tag{27.16}$$

and determine $\widehat{\boldsymbol{x}}$ by minimizing the MSE over all possible choices for the function $c(\cdot)$:

$$\min_{c(\cdot)} \; \mathbb{E}\,\widetilde{\boldsymbol{x}}^2, \quad \text{subject to} \;\; \widehat{\boldsymbol{x}} = c(\boldsymbol{y}) \tag{27.17}$$

We establish in the following that the solution of (27.17) is given by the *conditional* mean estimator:

$$\boxed{\widehat{x} = \mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y} = y)} \tag{27.18}$$

That is, the optimal choice for $c(y)$ in (27.15) is

$$c^o(y) = \mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y} = y) \tag{27.19}$$

Recall from (3.100) that for continuous random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, the conditional expectation is computed via the integral expression:

$$\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y} = y) \;=\; \int_{x \in \mathcal{X}} x f_{\boldsymbol{x}|\boldsymbol{y}}(x|y)dx \tag{27.20a}$$

over the domain of $\boldsymbol{x}$, while for discrete random variables:

$$\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y}=y) \;=\; \sum_{x} x\,\mathbb{P}(\boldsymbol{x}=x|\boldsymbol{y}=y) \tag{27.20b}$$

in terms of the conditional probability values and where the sum is over the possible realizations for $\boldsymbol{x}$. We continue our presentation by using the notation for continuous random variables without loss in generality.

Returning to (27.18), this estimator is obviously unbiased since from the result of Prob. 3.25 we know that

$$\mathbb{E}\,\widehat{\boldsymbol{x}} \;=\; \mathbb{E}\,\Big(\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})\Big) \;=\; \mathbb{E}\,\boldsymbol{x} \;=\; \bar{x} \tag{27.21}$$

Moreover, the resulting minimum cost or MMSE will be given by

$$\boxed{\text{MMSE} \;\triangleq\; \mathbb{E}\,\widetilde{\boldsymbol{x}}^2 \;=\; \sigma_x^2 \;-\; \sigma_{\widehat{x}}^2} \tag{27.22}$$

which is smaller than the earlier value (27.4). Result (27.18) states 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 observe* $\boldsymbol{y}$?" the analysis tells us that the best guess is what we would expect for $\boldsymbol{x}$ given the occurrence of $\boldsymbol{y}$!

Derivation of (27.18) using differentiation: Using again the result of Prob. 3.25 we have

$$\begin{aligned}
\mathbb{E}\,(\boldsymbol{x}-\widehat{\boldsymbol{x}})^2 &= \mathbb{E}\,\Big\{\mathbb{E}\,\Big((\boldsymbol{x}-\widehat{\boldsymbol{x}})^2\,|\,\boldsymbol{y}\Big)\Big\} \\
&= \mathbb{E}\,\Big\{\mathbb{E}\,\Big(\boldsymbol{x}^2-2\boldsymbol{x}\widehat{\boldsymbol{x}}+\widehat{\boldsymbol{x}}^2\,|\,\boldsymbol{y}\Big)\Big\} \\
&= \mathbb{E}\,\Big\{\mathbb{E}\,\Big(\boldsymbol{x}^2-2\boldsymbol{x}c(\boldsymbol{y})+c^2(\boldsymbol{y})\,|\,\boldsymbol{y}\Big)\Big\},\;\;\text{since } \widehat{\boldsymbol{x}}=c(\boldsymbol{y}) \\
&= \mathbb{E}\,\Big\{\mathbb{E}\,(\boldsymbol{x}^2|\boldsymbol{y})-2c(\boldsymbol{y})\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})+c^2(\boldsymbol{y})\Big\}
\end{aligned} \tag{27.23}$$

It is sufficient to minimize the inner expectation relative to $c(y)$, for any realization $\boldsymbol{y}=y$. Differentiating and setting the derivative to zero at the optimal solution $c^o(y)$ gives $c^o(y)=\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y}=y)$.

■

Derivation of (27.18) using completion-of-squares: In this second derivation, we will establish two useful intermediate results, namely, the orthogonality conditions (27.26) and (27.29). We again refer to the result of Prob. 3.25, which we write more explicitly as

$$\mathbb{E}\,\boldsymbol{x} \;=\; \mathbb{E}_{\,\boldsymbol{y}}\Big(\mathbb{E}_{\,\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}|\boldsymbol{y})\Big) \tag{27.24}$$

where the outer expectation is relative to the distribution of $\boldsymbol{y}$, while the inner expectation is relative to the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$. It follows that, for any real-valued function of $\boldsymbol{y}$, say, $g(\boldsymbol{y})$,

$$\begin{aligned}\mathbb{E}\,\boldsymbol{x}g(\boldsymbol{y}) &= \mathbb{E}_{\boldsymbol{y}}\Big\{\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\Big(\boldsymbol{x}g(\boldsymbol{y})\,|\,\boldsymbol{y}\Big)\Big\}\\ &= \mathbb{E}_{\boldsymbol{y}}\Big\{\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}|\boldsymbol{y})g(\boldsymbol{y})\Big\}\\ &= \mathbb{E}_{\boldsymbol{y}}\Big\{\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}|\boldsymbol{y})\Big\}g(\boldsymbol{y})\end{aligned} \tag{27.25}$$

This means that, for any $g(\boldsymbol{y})$, it holds that

$$\boxed{\mathbb{E}\Big(\boldsymbol{x}-\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})\Big)g(\boldsymbol{y})=0 \iff \widetilde{\boldsymbol{x}}\perp g(\boldsymbol{y})} \quad (\textbf{orthogonality condition}) \tag{27.26}$$

Result (27.26) states that the error variable $\widetilde{\boldsymbol{x}}=\boldsymbol{x}-\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})$ is orthogonal to any function $g(\cdot)$ of $\boldsymbol{y}$; we also represent this result by the compact notation $\widetilde{\boldsymbol{x}}\perp g(\boldsymbol{y})$. However, since $\boldsymbol{x}-\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})$ is zero mean, then it also holds that $\widetilde{\boldsymbol{x}}$ is uncorrelated with $g(\boldsymbol{y})$.

Using this intermediate result, we return to the risk (27.17), add and subtract $\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})$ to its argument, and express it as

$$\mathbb{E}\,(\boldsymbol{x}-\widehat{\boldsymbol{x}})^2=\mathbb{E}\,\Big((\boldsymbol{x}-\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})+\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})-\widehat{\boldsymbol{x}}\Big)^2 \tag{27.27}$$

The term $\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})-\widehat{\boldsymbol{x}}$ is a function of $\boldsymbol{y}$. Therefore, if we choose $g(\boldsymbol{y})=\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})-\widehat{\boldsymbol{x}}$, then from the orthogonality property (27.26) we conclude that

$$\mathbb{E}\,(\boldsymbol{x}-\widehat{\boldsymbol{x}})^2=\mathbb{E}\,\Big(\boldsymbol{x}-\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})\Big)^2+\mathbb{E}\,\Big(\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})-\widehat{\boldsymbol{x}}\Big)^2 \tag{27.28}$$

Now, only the second term on the right-hand side is dependent on $\widehat{\boldsymbol{x}}$ and the MSE is minimized by choosing $\widehat{\boldsymbol{x}}=\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})$. To evaluate the resulting MMSE we first use the orthogonality property (27.26), along with the fact that $\widehat{\boldsymbol{x}}=\mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})$ is itself a function of $\boldsymbol{y}$, to conclude that

$$\boxed{\mathbb{E}\,(\boldsymbol{x}-\widehat{\boldsymbol{x}})\widehat{\boldsymbol{x}}=0 \iff \widetilde{\boldsymbol{x}}\perp\widehat{\boldsymbol{x}}} \tag{27.29}$$

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

$$\begin{aligned}\mathbb{E}\,\widetilde{\boldsymbol{x}}^2 &= \mathbb{E}\,(\boldsymbol{x}-\widehat{\boldsymbol{x}})(\boldsymbol{x}-\widehat{\boldsymbol{x}})\\ &= \mathbb{E}\,(\boldsymbol{x}-\widehat{\boldsymbol{x}})\boldsymbol{x} \qquad \text{(because of (27.29))}\\ &= \mathbb{E}\,\boldsymbol{x}^2-\mathbb{E}\,\widehat{\boldsymbol{x}}(\widetilde{\boldsymbol{x}}+\widehat{\boldsymbol{x}}) \quad (\text{since } \boldsymbol{x}=\widetilde{\boldsymbol{x}}+\widehat{\boldsymbol{x}})\\ &= \mathbb{E}\,\boldsymbol{x}^2-\mathbb{E}\,\widehat{\boldsymbol{x}}^2 \qquad \text{(because of (27.29) again)}\\ &= \big(\mathbb{E}\,\boldsymbol{x}^2-\bar{x}^2\big)+\big(\bar{x}^2-\mathbb{E}\,\widehat{\boldsymbol{x}}^2\big)\\ &= \sigma_x^2-\sigma_{\widehat{x}}^2\end{aligned} \tag{27.30}$$

■

Example 27.3 (**From soft to hard decisions**) Let us return to Example 27.1, where $\boldsymbol{x}$ is a binary signal that assumes the values ± 1 with probability $1/2$. Recall that $\boldsymbol{x}$ represents the class of the selected image (cats or dogs). In that example, we only assumed knowledge of the mean of $\boldsymbol{x}$ and concluded that the resulting MSE estimate was not meaningful because it led to $\widehat{x}=0$, which is neither $+1$ nor -1. We are going to assume now that we have access to some additional information about $\boldsymbol{x}$. Specifically,

we are going to assume that we have some noisy measurement of $\boldsymbol{x}$, denoted by $\boldsymbol{y}$, say, as

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{v} \tag{27.31}$$

where the symbol $\boldsymbol{v}$ denotes the disturbance. How do measurements $\boldsymbol{y}$ of this type arise? We are going to encounter later in this text several inference techniques that provide approximate estimates for discrete variables. Rather than detect whether the unknown $\boldsymbol{x}$ is $+1$ or -1 (which we refer to as performing *hard* decisions), these other methods will return approximate values for $\boldsymbol{x}$ such as claiming that it is 0.9 or -0.7 (which we refer to as performing *soft* decisions). The soft value is a *real* (and not discrete) number and it can be interpreted as being a perturbed version of the actual label $\boldsymbol{x}$.

We are then faced with the problem of deciding whether $\boldsymbol{x} = \pm 1$ from the soft version $\boldsymbol{y}$. Obviously, the nature of the perturbation $\boldsymbol{v}$ in (27.31) depends on the method that is used to generate the approximation $\boldsymbol{y}$. In this example, and in order to keep the analysis tractable, we will assume that $\boldsymbol{v}$ and $\boldsymbol{x}$ are independent of each other and, moreover, that $\boldsymbol{v}$ has zero mean, unit variance, and is Gaussian-distributed with probability density function (pdf):

$$f_{\boldsymbol{v}}(v) = \frac{1}{\sqrt{2\pi}}\, e^{-v^2/2} \tag{27.32}$$

Our intuition tells us that we should be able to do better here than in Example 27.1. But beware, even here, we will arrive at some interesting conclusions. According to (27.18), the optimal estimator for $\boldsymbol{x}$ given $\boldsymbol{y}$ is the conditional mean $\widehat{\boldsymbol{x}} = \mathbb{E}\,(\boldsymbol{x}|\boldsymbol{y})$, which we evaluated earlier in (3.117) and determined that:

$$\boxed{\widehat{\boldsymbol{x}} = \tanh(\boldsymbol{y}) \;\triangleq\; \frac{e^{\boldsymbol{y}} - e^{-\boldsymbol{y}}}{e^{\boldsymbol{y}} + e^{-\boldsymbol{y}}}} \tag{27.33}$$

The result is represented in Fig. 27.2. If the variance of the measurement noise $\boldsymbol{v}$ were not fixed at 1 but denoted more generically by σ_v^2, then the same argument would lead to $\widehat{\boldsymbol{x}} = \tanh(\boldsymbol{y}/\sigma_v^2)$ with $\boldsymbol{y}$ scaled by σ_v^2 – see Prob. 27.2.

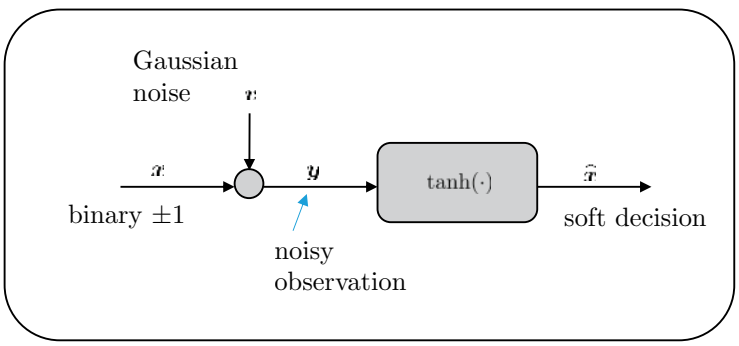

Figure 27.2 Estimation of a binary signal ± 1 observed under unit-variance additive Gaussian noise.

Figure 27.3 plots the function $\tanh(y)$. We see that it tends to ± 1 as $y \to \pm\infty$. For other values of y, the function assumes real values that are distinct from ± 1. This is a bit puzzling from the designer's perspective. The designer is interested in knowing whether the symbol $\boldsymbol{x}$ is $+1$ or -1 based on the observed value of $\boldsymbol{y}$. Construction

(27.33) tells the designer to estimate x by computing $\tanh(y)$. But, once again, this value will not be $+1$ or -1; it will be a real number somewhere inside the interval $(-1, 1)$. The designer will be induced to make a hard decision of the form:

$$\text{decide in favor of } \begin{cases} +1, & \text{if } \tanh(y) \text{ is nonnegative} \\ -1, & \text{if } \tanh(y) \text{ is negative} \end{cases} \tag{27.34}$$

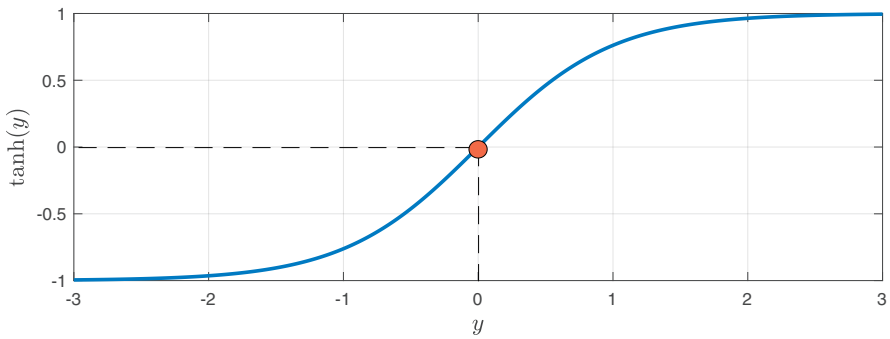

Figure 27.3 A plot of the hyperbolic tangent function, $\tanh(y)$. Observe that the curve tends to ± 1 as $y \to \pm\infty$.

In effect, the designer is implementing the alternative estimator:

$$\widehat{\boldsymbol{x}} \;=\; \text{sign}\Big(\tanh(\boldsymbol{y})\Big) \tag{27.35}$$

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

$$\text{sign}(x) \;\triangleq\; \begin{cases} +1, & \text{if } x \geq 0 \\ -1, & \text{if } x < 0 \end{cases} \tag{27.36}$$

We therefore have a situation where the optimal estimator (27.33), although known in closed form, does not solve the original problem of recovering the symbols ± 1 directly. Instead, the designer is forced to implement a *suboptimal* solution; it is suboptimal from a LMSE point of view. Actually, the designer could consider implementing the following simpler suboptimal estimator directly:

$$\widehat{\boldsymbol{x}} = \text{sign}(\boldsymbol{y}) \tag{27.37}$$

where the $\text{sign}(\cdot)$ function operates directly on $\boldsymbol{y}$ rather than on $\tanh(\boldsymbol{y})$ – see Fig. 27.4. Both suboptimal implementations (27.35) and (27.37) lead to the *same* result since, as is evident from Fig. 27.3:

$$\text{sign}\Big(\tanh(y)\Big) \;=\; \text{sign}(y) \tag{27.38}$$

We say that implementation (27.37) provides *hard decisions*, while implementation (27.33) provides *soft decisions*. We will revisit this problem later in Example 28.3 and show how the estimator (27.37) can be interpreted as being optimal relative to another design criterion; specifically, it will be the optimal Bayes classifier for the situation under study.

The purpose of Examples 27.1 and 27.3 is not to confuse the reader, but 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