STOCHASTIC PROCESSES, DETECTION AND ESTIMATION

6.432 Course Notes

Alan S. Willsky, Gregory W. Wornell, and Jeffrey H. Shapiro

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139

Fall 2003

3

Estimation Theory

3.1 INTRODUCTION

This chapter of the notes provides a fairly self-contained introduction to the fundamental concepts and results in estimation theory. The prototype problem we will consider is that of estimating the value of a vector $\mathbf{x}$ based on observations of a related vector $\mathbf{y}$. As an example, $\mathbf{x}$ might be a vector of the position and velocity of an aircraft, and $\mathbf{y}$ might be a vector of radar return measurements from several sensors.

As in our treatment of hypothesis testing and detection theory, there are two fundamentally rather different approaches to these kinds of estimation problems. In the first, we view the quantity to be estimated as a *random* vector $\mathbf{x}$. In this case, the conditional density $p_{\mathbf{y}|\mathbf{x}}(\mathrm{y}|\mathrm{x})$ fully characterizes the relationship between $\mathbf{x}$ and the observation $\mathbf{y}$. In the second case, we view the quantity as a *nonrandom* but unknown quantity x. In this case, we express the relationship between x and the observed data $\mathbf{y}$ by writing x as a *parameter* of the density for $\mathbf{y}$, i.e., $p_{\mathbf{y}}(\mathrm{y};\mathrm{x})$. We emphasize that for nonrandom parameter estimation, a probability density is *not* defined for x, and as a consequence, we will not need to distinguish between $\mathbf{x}$ and x in this case. Note, however, that in both the random and nonrandom parameter cases the observations $\mathbf{y}$ have some inherent randomness, and hence $\mathbf{y}$ is always specified probabilistically.

Before we begin, it is worth commenting that the hypothesis testing problems we considered in the last chapter of the notes can, at least in principle, be viewed as a special case of the more general estimation problem. In particular, we can view the M-ary hypothesis testing problem as one of estimating the value of a

quantity $\mathbf{x}$ that takes on one of M distinct values, each of which corresponds to one of the hypotheses $H_0, H_1, \ldots, H_{M-1}$. From this perspective, at least conceptually we can view the problem of estimation of a vector $\mathbf{x}$ as one of making a decision among a *continuum* of candidate hypotheses. However, in practice this perspective turns out to be a better way to interpret our estimation theory results than to first derive them. As a result, we will develop estimation theory independently.

3.2 ESTIMATION OF RANDOM VECTORS: A BAYESIAN FORMULATION

A natural framework for the estimation of random vectors arises out of what is referred to as "Bayesian estimation theory." This Bayesian framework will be the subject of this section. As will become apparent, there is a close connection between the Bayesian estimation problem we consider here and the Bayesian hypothesis testing problem we discussed in Chapter 2.

In the Bayesian framework, we refer to the density $p_{\mathbf{x}}(\mathbf{x})$ for the vector $\mathbf{x} \in \mathbb{R}^n$ of quantities to be estimated as the *prior density*. This is because this density fully specifies our knowledge about $\mathbf{x}$ prior to any observation of the measurement $\mathbf{y}$.

The conditional density $p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$, which fully specifies the way in which $\mathbf{y}$ contains information about $\mathbf{x}$, is often not specified directly but is inferred from a *measurement model*.

Example 3.1

Suppose that $\mathbf{y}$ is a noise-corrupted measurement of some function of $\mathbf{x}$, viz.,

$$\mathbf{y} = \mathbf{h}(\mathbf{x}) + \mathbf{w} \tag{3.1}$$

where $\mathbf{w}$ is a random noise vector that is independent of $\mathbf{x}$ and has density $p_{\mathbf{w}}(\mathbf{w})$. Then

$$p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = p_{\mathbf{w}}(\mathbf{y} - \mathbf{h}(\mathbf{x})). \tag{3.2}$$

Suppose in addition, $\mathbf{h}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ and $\mathbf{w} \sim N(\mathbf{0}, \mathbf{\Lambda})$ where the matrix $\mathbf{A}$ and covariance matrix $\mathbf{\Lambda}$ are arbitrary. Then

$$p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = N(\mathbf{y}; \mathbf{A}\mathbf{x}, \mathbf{\Lambda}).$$

Note that the measurement model $p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ and prior density $p_{\mathbf{x}}(\mathbf{x})$ together constitute a fully statistical characterization of $\mathbf{x}$ and $\mathbf{y}$. In particular, the joint density is given by their product, i.e.,

$$p_{\mathbf{y},\mathbf{x}}(\mathbf{y}, \mathbf{x}) = p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{x}}(\mathbf{x}) \tag{3.3}$$

from which we can get all other statistical information. As an example, we can get the marginal density $p_{\mathbf{y}}(\mathbf{y})$ for the observed data via

$$p_{\mathbf{y}}(\mathbf{y}) = \int_{-\infty}^{+\infty} p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x}.$$

In turn, we can also get the *posterior* density for $\mathbf{x}$, i.e., the density for $\mathbf{x}$ given that $\mathbf{y} = \mathbf{y}$ has been observed, via

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{x}}(\mathbf{x})}{p_{\mathbf{y}}(\mathbf{y})}. \tag{3.4}$$

In our treatment, we will use $\hat{\mathbf{x}}(\mathbf{y})$ to denote our estimate of $\mathbf{x}$ based on observing that the measurement $\mathbf{y} = \mathbf{y}$. Note that what we are estimating is actually an entire vector *function* $\hat{\mathbf{x}}(\cdot)$, not just an individual vector. In particular, for each possible observed value $\mathbf{y}$, the quantity $\hat{\mathbf{x}}(\mathbf{y})$ represents the estimate of the corresponding value of $\mathbf{x}$. We call this function the "estimator."

In the Bayesian framework, we choose the estimator to optimize a suitable performance criterion. In particular, we begin by choosing a deterministic scalar-valued function $C(\mathbf{a}, \hat{\mathbf{a}})$ that specifies the cost of estimating an arbitrary vector $\mathbf{a}$ as $\hat{\mathbf{a}}$. Then, we choose our estimator $\hat{\mathbf{x}}(\cdot)$ as that function which minimizes the average cost, i.e.,

$$\hat{\mathbf{x}}(\cdot) = \arg\min_{\mathbf{f}(\cdot)} E\left[C(\mathbf{x}, \mathbf{f}(\mathbf{y}))\right]. \tag{3.5}$$

Note that the expectation in (3.5) is over $\mathbf{x}$ and $\mathbf{y}$ jointly, and hence $\hat{\mathbf{x}}(\cdot)$ is that function which minimizes the cost averaged over all possible $(\mathbf{x}, \mathbf{y})$ pairs.

Solving for the optimum function $\hat{\mathbf{x}}(\cdot)$ in (3.5) can, in fact, be accomplished on a *pointwise basis*, i.e., for each particular value $\mathbf{y}$ that is observed, we find the best possible choice (in the sense of (3.5)) for the corresponding estimate $\hat{\mathbf{x}}(\mathbf{y})$. To see this, using (3.3) we first rewrite our objective function in (3.5) in the form

$$\begin{aligned} E\left[C(\mathbf{x}, \mathbf{f}(\mathbf{y}))\right] &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} C(\mathbf{x}, \mathbf{f}(\mathbf{y}))\, p_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y})\, d\mathbf{x}\, d\mathbf{y} \\ &= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} C(\mathbf{x}, \mathbf{f}(\mathbf{y}))\, p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, d\mathbf{x} \right] p_{\mathbf{y}}(\mathbf{y})\, d\mathbf{y}. \end{aligned} \tag{3.6}$$

Then, since $p_{\mathbf{y}}(\mathbf{y}) \geq 0$, we clearly will minimize (3.6) if we choose $\hat{\mathbf{x}}(\mathbf{y})$ to minimize the term in brackets for each individual value of $\mathbf{y}$, i.e.,

$$\hat{\mathbf{x}}(\mathbf{y}) = \arg\min_{\mathbf{a}} \int_{-\infty}^{+\infty} C(\mathbf{x}, \mathbf{a})\, p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, d\mathbf{x}. \tag{3.7}$$

As (3.7) indicates, the posterior density $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ summarizes everything we need to know about the $\mathbf{x}$ and $\mathbf{y}$ to construct the optimal Bayesian estimators for *any* given cost criterion. From this perspective, we see that the posterior density plays a role analogous to that played by the likelihood ratio in hypothesis testing problems. As discussed earlier, computation of the posterior density is generally accomplished via (3.4). However, since the denominator is simply a normalization factor (independent of $\mathbf{x}$), it is worth emphasizing that we can rewrite (3.7) more directly in terms of the measurement model and prior density as

$$\hat{\mathbf{x}}(\mathbf{y}) = \arg\min_{\mathbf{a}} \int_{-\infty}^{+\infty} C(\mathbf{x}, \mathbf{a})\, p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x}. \tag{3.8}$$

As an additional remark, we note that the result (3.7) is actually a direct extension of the corresponding M-ary Bayesian hypothesis testing result developed in the preceding chapter of the course notes. Specifically, if $\mathbf{x}$ takes on one of only M values—which, for convenience, we label $\mathbf{H}_0$, $\mathbf{H}_1$, ..., $\mathbf{H}_{M-1}$—then the pdf for $\mathbf{x}$ consists of M impulses and the integral in (3.7) becomes a summation, i.e.,

$$\hat{\mathbf{x}}(\mathbf{y}) = \hat{\mathbf{H}}(\mathbf{y}) = \underset{\mathbf{a}\in\{\mathbf{H}_0,\mathbf{H}_1,\ldots,\mathbf{H}_{M-1}\}}{\arg\min} \sum_{i=0}^{M-1} C(\mathbf{H}_i, \mathbf{a}) \Pr\left[\mathbf{x} = \mathbf{H}_i \mid \mathbf{y} = \mathbf{y}\right]. \tag{3.9}$$

Choosing a suitable cost criterion for a particular problem depends on a variety of factors. For example, the cost criterion should reflect the relative importance of various kinds of errors in the application of interest. However, from a practical standpoint, if we choose extremely complicated cost criteria, solving for the optimal estimator may be intractable. As a result, selecting a good cost criterion involves a tradeoff between capturing the aspects of interest in the error behavior and obtaining a framework that lends itself to analysis.

In the remainder of this section, we focus on some examples of practical cost criteria. For simplicity, we will generally restrict our attention to the case of estimating scalar variables x from vector observations $\mathbf{y}$. Keep in mind, however, that the more general case of estimating vector variables $\mathbf{x}$ can be handled in a component-wise manner.

3.2.1 Minimum Absolute-Error Estimation

One possible choice for the cost function is based on a minimum absolute-error (MAE) criterion. The cost function of interest in this case is

$$C(a, \hat{a}) = |a - \hat{a}|. \tag{3.10}$$

Substituting (3.10) into (3.7) we obtain

$$\begin{aligned}\hat{x}_{\mathrm{MAE}}(\mathbf{y}) &= \arg\min_a \int_{-\infty}^{+\infty} |x - a|\, p_{x|\mathbf{y}}(x|\mathbf{y})\, dx \\ &= \arg\min_a \left\{ \int_{-\infty}^{a} (a - x) p_{x|\mathbf{y}}(x|\mathbf{y})\, dx + \int_{a}^{+\infty} (x - a) p_{x|\mathbf{y}}(x|\mathbf{y})\, dx \right\}.\end{aligned} \tag{3.11}$$

Differentiating the quantity inside braces in (3.11) with respect to a gives, via Leibnitz' rule, the condition

$$\left[\int_{-\infty}^{a} p_{x|\mathbf{y}}(x|\mathbf{y})\, dx - \int_{a}^{+\infty} p_{x|\mathbf{y}}(x|\mathbf{y})\, dx \right]\Bigg|_{a=\hat{x}_{\mathrm{MAE}}(\mathbf{y})} = 0. \tag{3.12}$$

Rewriting (3.12) we obtain

$$\int_{-\infty}^{\hat{x}_{\mathrm{MAE}}(\mathbf{y})} p_{x|\mathbf{y}}(x|\mathbf{y})\, dx = \int_{\hat{x}_{\mathrm{MAE}}(\mathbf{y})}^{+\infty} p_{x|\mathbf{y}}(x|\mathbf{y})\, dx = \frac{1}{2}. \tag{3.13}$$

From (3.13) we see that the $\hat{x}_{\mathrm{MAE}}(\mathbf{y})$ is the threshold in x of the posterior density $p_{x|\mathbf{y}}(x|\mathbf{y})$ for which half the probability is located above the threshold and, hence, half is also below the threshold. This quantity is more generally known as the *median* of a probability density. Hence, the MAE estimator for x given $\mathbf{y} = \mathbf{y}$ is the median of the posterior density.

Note that, in general, there is no explicit formula for the median of a density, but rather it is specified implicitly as a solution to (3.13). As a result, the median is often calculated through an iterative, numerical optimization procedure.

Example 3.2

Suppose we have the posterior density

$$p_{x|y}(x|y) = \begin{cases} 1/(3y) & 0 < x < y \\ 2/(3y) & y < x < 2y \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\hat{x}_{\mathrm{MAE}}(y) = (1 + \Delta)y$$

for an appropriate choice of $\Delta > 0$. To solve for Δ, we use (3.13) to obtain

$$\frac{1}{3y} \cdot y + \frac{2}{3y} \cdot y\Delta = 1/2$$

from which we deduce that $\Delta = 1/4$.

We also note that the median of a density is not necessarily unique.

Example 3.3

Suppose

$$p_{x|y}(x|y) = \begin{cases} 1/2y & 0 < x < y \text{ and } 2y < x < 3y \\ 0 & \text{otherwise} \end{cases} \tag{3.14}$$

Then the median of (3.14) is any number between y and $2y$; hence, the MAE estimators for x given $y = y$ are all of the form

$$\hat{x}_{\mathrm{MAE}}(y) = \alpha$$

where α is any constant satisfying $y \leq \alpha \leq 2y$ (assuming $y \geq 0$) or $2y \leq \alpha \leq y$ (assuming $y < 0$).

3.2.2 Maximum A Posteriori Estimation

As an alternative to that considered in the previous section, consider the cost function

$$C(a, \hat{a}) = \begin{cases} 1 & |a - \hat{a}| > \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

which uniformly penalizes all estimation errors with magnitude bigger than ϵ. This time, substituting (3.15) into (3.7) we obtain that the minimum uniform cost (MUC) estimator satisfies

$$\begin{aligned}\hat{x}_{\mathrm{MUC}}(\mathbf{y}) &= \arg\min_a \left[1 - \int_{a-\epsilon}^{a+\epsilon} p_{x|\mathbf{y}}(x|\mathbf{y})\, dx\right] \\ &= \arg\max_a \int_{a-\epsilon}^{a+\epsilon} p_{x|\mathbf{y}}(x|\mathbf{y})\, dx. \end{aligned} \tag{3.16}$$

Note that via (3.16) we see that $\hat{x}_{\mathrm{MUC}}(\mathbf{y})$ corresponds to the value of a that makes $\Pr\left[|x - \hat{x}_{\mathrm{MUC}}(\mathbf{y})| < \epsilon \mid \mathbf{y} = \mathbf{y}\right]$ as large as possible. This means finding the interval of length 2ϵ where the posterior density $p_{x|\mathbf{y}}(x|\mathbf{y})$ is most concentrated.

If we carry this perspective a little further, we see that if we let ϵ get sufficiently small then the $\hat{x}_{\mathrm{MUC}}(\mathbf{y})$ approaches the point corresponding to the peak of the posterior density. For this reason, this limiting case estimator is referred to as the "maximum *a posteriori*" (MAP) estimator, which we denote using

$$\hat{x}_{\mathrm{MAP}}(\mathbf{y}) = \arg\max_a p_{x|\mathbf{y}}(a|\mathbf{y}) = \lim_{\epsilon \to 0} \hat{x}_{\mathrm{MUC}}(\mathbf{y}). \tag{3.17}$$

The peak value of a density is referred to as its *mode*. Hence, we see that the MAP estimate of x based on observing $\mathbf{y} = \mathbf{y}$ is the mode of the posterior density $p_{x|\mathbf{y}}(x|\mathbf{y})$. From our limiting argument, we see that the MAP estimator can be viewed as resulting from a Bayes' cost formulation in which all errors are, in the appropriate sense, equally bad.

As a final remark, we note that the vector form of the MAP is a straightforward generalization of (3.17); specifically,

$$\hat{\mathbf{x}}_{\mathrm{MAP}}(\mathbf{y}) = \arg\max_{\mathbf{a}} p_{\mathbf{x}|\mathbf{y}}(\mathbf{a}|\mathbf{y}). \tag{3.18}$$

We conclude our development of MAP estimation with a brief discussion of some computational issues. For this discussion, let us restrict our attention to the fairly typical case in which the posterior density is differentiable in $\mathbf{x}$. In this case, we first look for the MAP estimate among the stationary points of the posterior density—the values of $\mathbf{x}$ for which the Jacobian is zero, i.e.,

$$\frac{\partial}{\partial \mathbf{x}} p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = \mathbf{0}. \tag{3.19}$$

Recall that in this most general vector case, (3.19) is a *set* of equations resulting from differentiation with respect to each component of $\mathbf{x}$. A solution of (3.19) is a local maximum of the posterior density if the corresponding Hessian matrix satisfies

$$\frac{\partial^2}{\partial \mathbf{x}^2} p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) < \mathbf{0} \tag{3.20}$$

where, as discussed in Appendix 1.A, the matrix inequality in (3.20) is to be interpreted in the sense of negative definiteness. If there are several local maxima,

the relative sizes of the posterior density $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ at each of these must be determined. In addition, if $\mathbf{x}$ takes on values only in some restricted range, then the MAP estimate may be on the boundary of this set even if (3.19) is not satisfied at such a point. Consequently, solving for the MAP estimator in general involves finding all values of $\mathbf{x}$ corresponding to local maxima of $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ as well as all boundary points corresponding to the range of $\mathbf{x}$, and taking as $\hat{\mathbf{x}}_{\mathrm{MAP}}(\mathbf{y})$ the value that maximizes $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ over all these points.

As a final remark, it is worth pointing out that in many problems it is more convenient to maximize other monotonic functions of the posterior density. For example, maximizing $\ln p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ with respect to $\mathbf{x}$ is sometimes easier than maximizing the posterior density directly. In this example, using

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})p_{\mathbf{x}}(\mathbf{x})}{p_{\mathbf{y}}(\mathbf{y})},$$

taking logarithms, and then differentiating with respect to $\mathbf{x}$ we obtain the MAP equations

$$\frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) + \frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{x}}(\mathbf{x}) = \mathbf{0}. \tag{3.21}$$

Solutions of (3.21) that also satisfy (3.20) are again the local maxima of the posterior density $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$.

3.2.3 Bias and Variance

Let us briefly discuss some general and useful measures of performance for estimators $\hat{\mathbf{x}}(\cdot)$ regardless of the cost criterion we choose. One very important quantity is the estimate *bias*. Specifically, if we define the estimation *error* via

$$\mathbf{e}(\mathbf{x}, \mathbf{y}) = \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}, \tag{3.22}$$

then the bias is the average value of this error, i.e.,

$$\mathbf{b} = E\left[\mathbf{e}(\mathbf{x}, \mathbf{y})\right] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left[\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}\right] p_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y})\, d\mathbf{x}\, d\mathbf{y}. \tag{3.23}$$

The second performance measure is the error covariance

$$\mathbf{\Lambda}_{\mathbf{e}} = E\left[(\mathbf{e}(\mathbf{x}, \mathbf{y}) - \mathbf{b})(\mathbf{e}(\mathbf{x}, \mathbf{y}) - \mathbf{b})^{\mathrm{T}}\right]. \tag{3.24}$$

Note that using (3.24) the associated error correlation matrix $E\left[\mathbf{e}\mathbf{e}^{\mathrm{T}}\right]$ can be expressed in the form

$$E\left[\mathbf{e}\mathbf{e}^{\mathrm{T}}\right] = \mathbf{\Lambda}_{\mathbf{e}} + \mathbf{b}\mathbf{b}^{\mathrm{T}}, \tag{3.25}$$

so that both the bias and covariance contribute to the error correlation and, in turn, mean-square estimation error.

Since $\mathbf{b}$ is a deterministic vector it is, in principle, straightforward to correct for the bias: take as the estimate $\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{b}$. From this perspective, we see that adding

the constraint that our estimator be unbiased need not be a serious restriction. It should be pointed out, however, that in some problems, it may be difficult to compute $\mathbf{b}$ and therefore compensate for it. In such cases, there may be a tradeoff between choosing an estimator with a small covariance or one with a small bias.

As a final remark, if error covariance is the performance metric of primary interest for our estimator, then the natural cost criterion is in fact the least-squares one, which we'll now develop in detail.

3.2.4 Bayes' Least-Squares Estimation

In this section we consider the mean-square error (MSE) cost criterion

$$C(\mathbf{a}, \hat{\mathbf{a}}) = \|\mathbf{a} - \hat{\mathbf{a}}\|^2 = (\mathbf{a} - \hat{\mathbf{a}})^{\mathrm{T}}(\mathbf{a} - \hat{\mathbf{a}}) = \sum_{i=1}^{N} (a_i - \hat{a}_i)^2 \tag{3.26}$$

In this case, substituting (3.26) into (3.7) yields

$$\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) = \arg\min_{\mathbf{a}} \int_{-\infty}^{+\infty} (\mathbf{x} - \mathbf{a})^{\mathrm{T}} (\mathbf{x} - \mathbf{a})\, p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, d\mathbf{x} \tag{3.27}$$

where we have used $\hat{\mathbf{x}}_{\mathrm{BLS}}(\cdot)$ to specifically denote the Bayes least-squares (BLS) estimator. Since this estimator minimizes the mean-square estimation error, it is often alternatively referred to as the minimum mean-square error (MMSE) estimator and denoted using $\hat{\mathbf{x}}_{\mathrm{MMSE}}(\cdot)$.

Let us begin with the simpler case of scalar estimation, for which (3.27) becomes

$$\hat{x}_{\mathrm{BLS}}(\mathbf{y}) = \arg\min_{a} \int_{-\infty}^{+\infty} (x - a)^2\, p_{x|\mathbf{y}}(x|\mathbf{y})\, dx. \tag{3.28}$$

As we did in the case of MAE estimation, we can perform the minimization in (3.28) by differentiating with respect to a and setting the result to zero to find the local extrema. Differentiating the integral in (3.28) we obtain

$$\begin{aligned} \frac{\partial}{\partial a}\left[\int_{-\infty}^{+\infty} (x-a)^2 p_{x|\mathbf{y}}(x|\mathbf{y})\, dx\right] &= \int_{-\infty}^{+\infty} \frac{\partial}{\partial a}(x-a)^2 p_{x|\mathbf{y}}(x|\mathbf{y})\, dx \\ &= -2\int_{-\infty}^{+\infty} (x-a) p_{x|\mathbf{y}}(x|\mathbf{y})\, dx. \end{aligned} \tag{3.29}$$

Setting (3.29) to zero at $a = \hat{x}_{\mathrm{BLS}}(\mathbf{y})$ we see that

$$\begin{aligned}
&\left[\int_{-\infty}^{+\infty} (x-a) p_{x|\mathbf{y}}(x|\mathbf{y})\, dx\right]\Big|_{a=\hat{x}_{\mathrm{BLS}}(\mathbf{y})} \\
&= \int_{-\infty}^{+\infty} x\, p_{x|\mathbf{y}}(x|\mathbf{y})\, dx - \int_{-\infty}^{+\infty} \hat{x}_{\mathrm{BLS}}(\mathbf{y})\, p_{x|\mathbf{y}}(x|\mathbf{y})\, dx \\
&= E\left[x|\mathbf{y}=\mathbf{y}\right] - \hat{x}_{\mathrm{BLS}}(\mathbf{y}) \int_{-\infty}^{+\infty} p_{x|\mathbf{y}}(x|\mathbf{y})\, dx \\
&= E\left[x|\mathbf{y}=\mathbf{y}\right] - \hat{x}_{\mathrm{BLS}}(\mathbf{y}) = 0. \qquad (3.30)
\end{aligned}$$

A simple rearrangement of (3.30) then yields our final result

$$\hat{x}_{\mathrm{BLS}}(\mathbf{y}) = E\left[x|\mathbf{y}=\mathbf{y}\right], \qquad (3.31)$$

i.e., that the BLS or MMSE estimate of x given $\mathbf{y} = \mathbf{y}$ is the *mean* of the posterior density $p_{x|\mathbf{y}}(x|\mathbf{y})$.

Note that since the quantity being minimized in (3.28) is nonnegative, and since our derivation above concluded that there exists a single local extremum, this extremum—and hence our estimate (3.31)—must correspond to a global minimum.[1]

The preceding derivation generalizes rather easily when $\mathbf{x}$ is a vector. Specifically, since the cost criterion (3.26) is a sum of individual squared estimation errors for the components of $\mathbf{x}$, the minimum is achieved by minimizing the mean-square estimation error in each scalar component. Hence, we obtain

$$\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) = E\left[\mathbf{x}|\mathbf{y}\right], \qquad (3.33)$$

from which we see that in the vector case as well the BLS estimate of $\mathbf{x}$ given $\mathbf{y} = \mathbf{y}$ is the mean of the posterior density $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$.

Let us next turn our attention to the performance characteristics of the BLS estimator, in particular its bias and error covariance. First, we note that the BLS estimator is always unbiased: using (3.23) we have

$$\mathbf{b}_{\mathrm{BLS}} = E\left[\mathbf{e}(\mathbf{x},\mathbf{y})\right] = E\left[\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) - \mathbf{x}\right] = E\left[E\left[\mathbf{x}|\mathbf{y}\right]\right] - E\left[\mathbf{x}\right] = \mathbf{0}, \qquad (3.34)$$

where the last equality follows from a simple application of the law of iterated expectation.

Next, using (3.22), (3.24), and (3.34) we obtain that the associated error covariance is given by

$$\mathbf{\Lambda}_{\mathrm{BLS}} \triangleq \mathbf{\Lambda}_{\mathbf{e}} = E\left[\mathbf{e}\mathbf{e}^{\mathrm{T}}\right] = E\left[\left(\mathbf{x} - E\left[\mathbf{x}|\mathbf{y}\right]\right)\left(\mathbf{x} - E\left[\mathbf{x}|\mathbf{y}\right]\right)^{\mathrm{T}}\right] \qquad (3.35)$$

[1] We can also verify this independently by taking a second derivative of the integral in (3.28), i.e.,

$$\frac{\partial^2}{\partial a^2}\left[\int_{-\infty}^{+\infty} (x-a)^2 p_{x|\mathbf{y}}(x|\mathbf{y})\, dx\right] = 2\int_{-\infty}^{+\infty} p_{x|\mathbf{y}}(x|\mathbf{y})\, dx = 2 > 0, \qquad (3.32)$$

which also establishes that the objective function is convex.

where we emphasize that the notation $\mathbf{\Lambda}_{\mathrm{BLS}}$ is used to refer to the error covariance of the BLS estimator. Applying iterated expectation to (3.35) we see that the error covariance can be written as

$$\mathbf{\Lambda}_{\mathrm{BLS}} = E\left[E\left[(\mathbf{x} - E\left[\mathbf{x}|\mathbf{y}\right])\left(\mathbf{x} - E\left[\mathbf{x}|\mathbf{y}\right]\right)^{\mathrm{T}} \mid \mathbf{y}\right]\right]. \tag{3.36}$$

However, the inner expectation in (3.36) is simply the covariance of the posterior density, i.e., $\mathbf{\Lambda}_{\mathbf{x}|\mathbf{y}}$, which in general depends on $\mathbf{y}$.[2] Hence, the error covariance of the BLS estimator is simply the average of the covariance of the posterior density, where this averaging is over all possible values of $\mathbf{y}$, i.e.,

$$\mathbf{\Lambda}_{\mathrm{BLS}} = E\left[\mathbf{\Lambda}_{\mathbf{x}|\mathbf{y}}(\mathbf{y})\right]. \tag{3.37}$$

As a final remark before we proceed to an example, note that using the identity (1.204) from Appendix 1.A of Chapter 1, we have that at its minimum value the expected cost objective function in (3.27) can be expressed as

$$\begin{aligned} E\left[C(\mathbf{x}, \hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}))\right] &= E\left[\left(E\left[\mathbf{x}|\mathbf{y}\right] - \mathbf{x}\right)^{\mathrm{T}}\left(E\left[\mathbf{x}|\mathbf{y}\right] - \mathbf{x}\right)\right] \\ &= E\left[\mathrm{tr}\left\{\left(E\left[\mathbf{x}|\mathbf{y}\right] - \mathbf{x}\right)\left(E\left[\mathbf{x}|\mathbf{y}\right] - \mathbf{x}\right)^{\mathrm{T}}\right\}\right] \\ &= \mathrm{tr}\left(E\left[\left(E\left[\mathbf{x}|\mathbf{y}\right] - \mathbf{x}\right)\left(E\left[\mathbf{x}|\mathbf{y}\right] - \mathbf{x}\right)^{\mathrm{T}}\right]\right) \\ &= \mathrm{tr}\left(\mathbf{\Lambda}_{\mathrm{BLS}}\right). \end{aligned} \tag{3.38}$$

Example 3.4

Suppose x and w are independent random variables that are both uniformly distributed over the range $[-1, 1]$, and let

$$y = \mathrm{sgn}\, x + w.$$

Let's determine the BLS estimate of x given y. First we construct the joint density. Note that for $x > 0$, we have

$$p_{y|x}(y|x) = \begin{cases} 1/2 & 0 < y < 2 \\ 0 & \text{otherwise} \end{cases}$$

while for $x < 0$, we have

$$p_{y|x}(y|x) = \begin{cases} 1/2 & -2 < y < 0 \\ 0 & \text{otherwise} \end{cases}.$$

Hence, the joint density is

$$p_{x,y}(x,y) = p_{y|x}(y|x)\, p_x(x) = \begin{cases} 1/4 & 0 < x < 1 \text{ and } 0 < y < 2 \\ 1/4 & -1 < x < 0 \text{ and } -2 < y < 0 \\ 0 & \text{otherwise} \end{cases}$$

[2] Note that given an observed value of $\mathbf{y}$, this posterior covariance $\mathbf{\Lambda}_{\mathbf{x}|\mathbf{y}=\mathbf{y}}$ is in general a function of $\mathbf{y}$. To emphasize this dependence, and for future convenience, we'll frequently use the alternative notation $\mathbf{\Lambda}_{\mathbf{x}|\mathbf{y}}(\mathbf{y})$ for this covariance.

and so for $y > 0$ we have

$$p_{x|y}(x|y) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.39a}$$

and for $y < 0$ we have

$$p_{x|y}(x|y) = \begin{cases} 1 & -1 < x < 0 \\ 0 & \text{otherwise} \end{cases}. \tag{3.39b}$$

Thus from (3.39) we conclude that

$$\hat{x}_{\text{BLS}}(y) = E\left[x|y = y\right] = \frac{1}{2}\operatorname{sgn} y = \begin{cases} 1/2 & y > 0 \\ -1/2 & y < 0 \end{cases}. \tag{3.40}$$

Since

$$\lambda_{x|y}(y) = 1/12$$

is independent of y in this example, we have that the corresponding error variance is simply

$$\lambda_{\text{BLS}} = E\left[\lambda_{x|y}(y)\right] = 1/12. \tag{3.41}$$

Additional Properties of BLS Estimators

Let us briefly consider some additional important properties and an alternate characterization of the Bayes' least-squares estimate $\hat{\mathbf{x}}_{\text{BLS}}(\mathbf{y})$. Recall that we have already shown (see (3.34)) that the Bayes' least-squares estimate is unbiased.

Next we show that Bayes' least-squares estimates are unique in having an important *orthogonality* property. Specifically, we have the following theorem.

Theorem 3.1 *An estimator $\hat{\mathbf{x}}(\cdot)$ is the Bayes' least-squares estimator, i.e., $\hat{\mathbf{x}}(\cdot) = \hat{\mathbf{x}}_{\text{BLS}}(\cdot)$, if and only if the associated estimation error $\mathbf{e}(\mathbf{x}, \mathbf{y}) = \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}$ is orthogonal to* any *(vector-valued) function $\mathbf{g}(\cdot)$ of the data, i.e.,*

$$E\left[[\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}]\,\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] = \mathbf{0}. \tag{3.42}$$

In establishing this result, it will be convenient to first rewrite the condition (3.42) as

$$E\left[\mathbf{x}\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] = E\left[\hat{\mathbf{x}}(\mathbf{y})\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right]. \tag{3.43}$$

and note, using the law of iterated expectation, that the left-hand side of (3.43) can in turn be expressed in the form

$$E\left[\mathbf{x}\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] = E\left[E\left[\mathbf{x}\mathbf{g}^{\mathrm{T}}(\mathbf{y}) \mid \mathbf{y}\right]\right] = E\left[E\left[\mathbf{x}|\mathbf{y}\right]\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right]. \tag{3.44}$$

Then, to prove the "only if" statement, simply let $\hat{\mathbf{x}}(\cdot) = \hat{\mathbf{x}}_{\text{BLS}}(\cdot)$ in (3.43), and note that in this case the right-hand expressions in both (3.44) and (3.43) are identical, verifying (3.42).

To prove the converse, let us rewrite (3.42) using (3.43) and (3.44) as

$$\begin{aligned}\mathbf{0} &= E\left[\mathbf{x}\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] - E\left[\hat{\mathbf{x}}(\mathbf{y})\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] \\ &= E\left[E\left[\mathbf{x}|\mathbf{y}\right]\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] - E\left[\hat{\mathbf{x}}(\mathbf{y})\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] \\ &= E\left[\left[E\left[\mathbf{x}|\mathbf{y}\right] - \hat{\mathbf{x}}(\mathbf{y})\right]\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right]. \end{aligned} \tag{3.45}$$

Then, since (3.45) must hold for all $\mathbf{g}(\cdot)$, let us choose $\mathbf{g}(\mathbf{y}) = E\left[\mathbf{x}|\mathbf{y}\right] - \hat{\mathbf{x}}(\mathbf{y})$ where $\hat{\mathbf{x}}(\cdot)$ is our estimator. In this case (3.45) becomes

$$E\left[\left[E\left[\mathbf{x}|\mathbf{y}\right] - \hat{\mathbf{x}}(\mathbf{y})\right]\left[E\left[\mathbf{x}|\mathbf{y}\right] - \hat{\mathbf{x}}(\mathbf{y})\right]^{\mathrm{T}}\right] = \mathbf{0},$$

from which we can immediately conclude that $\hat{\mathbf{x}}(\mathbf{y}) = E\left[\mathbf{x}|\mathbf{y}\right]$.[3]

It is worth emphasizing that Theorem 3.1 ensures what we would expect of an estimator that yields the minimum mean-square error: that since the error $\mathbf{e}(\mathbf{x},\mathbf{y}) = \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}$ is uncorrelated with *any* function of the data we might construct, there is no further processing that can be done on the data to further reduce the error covariance in the estimate.

One final property of the BLS estimator is given in terms of the following matrix inequality (again to be interpreted in the sense of positive semidefiniteness as discussed in Appendix 1.A). Let $\mathbf{\Lambda_e}$ be the error covariance of *any* estimator $\hat{\mathbf{x}}(\cdot)$. Then the error covariance of the BLS estimator, i.e., $\mathbf{\Lambda}_{\mathrm{BLS}}$, satisfies

$$\mathbf{\Lambda}_{\mathrm{BLS}} \leq \mathbf{\Lambda_e} \tag{3.46}$$

with equality if and only if

$$\hat{\mathbf{x}}(\mathbf{y}) - E\left[\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}\right] = \hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) = E\left[\mathbf{x}|\mathbf{y}\right]. \tag{3.47}$$

In essence, this states that Bayes' least-squares estimator is guaranteed to yield less uncertainty in the value of $\mathbf{x}$ (as measured by the covariance) than any other estimator—biased or unbiased.

Before proving this result, we point out that a useful corollary results from the special case corresponding to choosing $\hat{\mathbf{x}}(\mathbf{y}) = \mathbf{m_x}$. In this case the corresponding error covariance is $\mathbf{\Lambda_e} = \mathbf{\Lambda_x}$, so we have that

$$\mathbf{\Lambda}_{\mathrm{BLS}} \leq \mathbf{\Lambda_x} \tag{3.48}$$

with equality if and only if

$$\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) = E\left[\mathbf{x}|\mathbf{y}\right] = E\left[\mathbf{x}\right]. \tag{3.49}$$

We stress that as discussed in Chapter 1, (3.49) is not equivalent to $\mathbf{x}$ and $\mathbf{y}$ being either statistically independent or uncorrelated. While $\mathbf{x}$ and $\mathbf{y}$ being uncorrelated is a necessary condition for (3.49) to hold, it is not sufficient one. On the other hand, for (3.49) to hold it is sufficient but not necessary that $\mathbf{x}$ and $\mathbf{y}$ be independent.

---

[3]Here we are using a straightforward consequence of the Chebyshev inequality—that if $E\left[\mathbf{z}\mathbf{z}^{\mathrm{T}}\right] = 0$ then $\mathbf{z} = 0$, or more precisely, $\Pr\left[\mathbf{z} = 0\right] = 1$.

To prove (3.46), let $\mathbf{b}$ denote the bias in our estimator $\hat{\mathbf{x}}(\cdot)$, let

$$\mathbf{g}(\mathbf{y}) = \hat{\mathbf{x}}(\mathbf{y}) - \hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) - \mathbf{b}, \tag{3.50}$$

and let us begin by noting that

$$\begin{aligned}
\mathbf{\Lambda_e} &= E\left[(\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x} - \mathbf{b})\,(\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x} - \mathbf{b})^{\mathrm{T}}\right] \\
&= E\left[[\mathbf{g}(\mathbf{y}) + (\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) - \mathbf{x})]\,[\mathbf{g}(\mathbf{y}) + (\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) - \mathbf{x})]^{\mathrm{T}}\right] \\
&= E\left[\mathbf{g}(\mathbf{y})\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] + E\left[(\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) - \mathbf{x})\,(\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) - \mathbf{x})^{\mathrm{T}}\right] \\
&\quad + E\left[(\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) - \mathbf{x})\,\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right] + E\left[(\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) - \mathbf{x})\,\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right]^{\mathrm{T}}.
\end{aligned} \tag{3.51}$$

From Theorem 3.1 we get that the last two terms in (3.51) are zero. Using this together with the definition of $\mathbf{\Lambda}_{\mathrm{BLS}}$ we get

$$\mathbf{\Lambda_e} - \mathbf{\Lambda}_{\mathrm{BLS}} = E\left[\mathbf{g}(\mathbf{y})\mathbf{g}^{\mathrm{T}}(\mathbf{y})\right]. \tag{3.52}$$

The right-hand side of (3.52) is in general positive semidefinite, which verifies (3.46), and equal to zero if and only if $\mathbf{g}(\mathbf{y}) = \mathbf{0}$, which using (3.50) yields (3.47).

Weighted Least-Squares Estimators

As a final comment on Bayes' least-squares estimation, in this section we show that the posterior or conditional mean $E[\mathbf{x}|\mathbf{y}]$ is also the optimal estimator for a more general *weighted* least-squares cost criterion.

In particular, consider the weighted least-squares cost criterion

$$C(\mathbf{a}, \hat{\mathbf{a}}) = (\mathbf{a} - \hat{\mathbf{a}})^{\mathrm{T}}\mathbf{M}(\mathbf{a} - \hat{\mathbf{a}}), \tag{3.53}$$

where $\mathbf{M}$ is an arbitrary positive definite matrix. In this case (3.7) becomes

$$\hat{\mathbf{x}}_{\mathrm{WLS}}(\mathbf{y}) = \arg\min_{\mathbf{a}} \int_{-\infty}^{+\infty} (\mathbf{x} - \mathbf{a})^{\mathrm{T}}\mathbf{M}(\mathbf{x} - \mathbf{a}) p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, d\mathbf{x}. \tag{3.54}$$

The objective function in (3.54) is a function of the vector of variables comprising $\mathbf{a}$. To find the minimum of this function we set the Jacobian of the function to zero and look for local extrema. Applying this procedure to (3.54) we obtain

$$\begin{aligned}
\mathbf{0} &= \left[\frac{\partial}{\partial \mathbf{a}} \int_{-\infty}^{+\infty} (\mathbf{x} - \mathbf{a})^{\mathrm{T}}\mathbf{M}(\mathbf{x} - \mathbf{a})\, p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, d\mathbf{x}\right]\Bigg|_{\mathbf{a}=\hat{\mathbf{x}}_{\mathrm{WLS}}(\mathbf{y})} \\
&= \left[\int_{-\infty}^{+\infty} \frac{\partial}{\partial \mathbf{a}} \left[(\mathbf{x} - \mathbf{a})^{\mathrm{T}}\mathbf{M}(\mathbf{x} - \mathbf{a})\right]\, p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, d\mathbf{x}\right]\Bigg|_{\mathbf{a}=\hat{\mathbf{x}}_{\mathrm{WLS}}(\mathbf{y})} \\
&= \left[2\int_{-\infty}^{+\infty} (\mathbf{x} - \mathbf{a})^{\mathrm{T}}\mathbf{M}\, p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, d\mathbf{x}\right]\Bigg|_{\mathbf{a}=\hat{\mathbf{x}}_{\mathrm{WLS}}(\mathbf{y})}.
\end{aligned} \tag{3.55}$$

Minor rearrangement of (3.55) then yields

$$\mathbf{M}E\left[\mathbf{x}|\mathbf{y}\right] = \mathbf{M}\hat{\mathbf{x}}_{\mathrm{WLS}}(\mathbf{y}), \tag{3.56}$$

which, since $\mathbf{M}$ is invertible, implies that (3.56) has a unique solution, and hence the objective function in (3.54) has a unique local extremum. But since $\mathbf{M}$ is positive definite, the cost function $C(\cdot,\cdot)$ in (3.53) is non-negative for every value of $\mathbf{a}$. Hence, the unique local extremum must be a global minimum;[4] hence from (3.56) we obtain

$$\hat{\mathbf{x}}_{\mathrm{WLS}}(\mathbf{y}) = E\left[\mathbf{x}|\mathbf{y}\right] = \hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}). \tag{3.57}$$

3.2.5 Linear Least-Squares Estimation

Two important observations about the Bayes' least-squares estimator $\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) = E\left[\mathbf{x}|\mathbf{y}\right]$ should be made. First, this estimator is in general a nonlinear (and often highly nonlinear) function of the data $\mathbf{y}$. Second, computing this estimator requires that we have access to a *complete* statistical characterization of the relationship between $\mathbf{x}$ and $\mathbf{y}$. In particular, we need full knowledge of $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ or, equivalently, $p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ and $p_{\mathbf{x}}(\mathbf{x})$.

However, there are many application scenarios when even though the least-squares cost criterion is appropriate, the resulting Bayes' least-squares estimator is not practicable either because implementing the nonlinear estimator is computationally too expensive, or because a complete statistical characterization of the relationship between $\mathbf{x}$ and $\mathbf{y}$ is not available from which to compute the estimator.

In these situations, we often must settle for a suboptimal estimator. One way to obtain such an estimator is to add a constraint on the form of the estimator. As an important example, in this section of the notes we'll develop in detail estimators that minimize the average Bayes' least-squares cost (3.26), but subject to the additional constraint that the estimator be a linear[5] function of the data. Specifically, we let $\hat{\mathbf{x}}_{\mathrm{LLS}}(\cdot)$ denote this linear least-squares (LLS) estimator, and define it as

$$\hat{\mathbf{x}}_{\mathrm{LLS}}(\cdot) = \underset{\mathbf{f}(\cdot)\in\mathcal{B}}{\arg\min}\, E\left[\|\mathbf{x}-\mathbf{f}(\mathbf{y})\|^2\right] \tag{3.58a}$$

where

$$\mathcal{B} = \left\{\mathbf{f}(\cdot) \mid \mathbf{f}(\mathbf{y}) = \mathbf{A}\mathbf{y}+\mathbf{d} \text{ for some } \mathbf{A} \text{ and } \mathbf{d}\right\}. \tag{3.58b}$$

---

[4] In fact, the objective function not only has a unique global minimum, but is convex as well:

$$\frac{\partial^2}{\partial \mathbf{a}^2}\int_{-\infty}^{+\infty}(\mathbf{x}-\mathbf{a})^{\mathrm{T}}\mathbf{M}(\mathbf{x}-\mathbf{a})\,p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\,d\mathbf{x} = \mathbf{M}\int_{-\infty}^{+\infty}p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\,d\mathbf{x} = \mathbf{M} > 0.$$

[5] Throughout this course, we'll use the term "linear" to refer to estimators of the form $\mathbf{A}\mathbf{y}+\mathbf{d}$ since this has become standard practice in estimation theory. More precise terminology would have us refer to such estimators as "affine" and estimators of the form $\mathbf{A}\mathbf{y}$ as "linear."

As we'll see, this estimator is not only particularly efficient to implement, but we'll need access to only the joint second-order statistics of $\mathbf{x}$ and $\mathbf{y}$ in order to compute it.

Although there are a variety of ways to derive the optimum estimator, we'll follow a powerful approach based on the abstract vector space concepts we developed in Section 1.7. With this formulation, several important perspectives and properties of linear estimators will become apparent.

The key to exploiting vector space concepts in this problem lies in reinterpreting (3.58) as a problem in linear *approximation*. For simplicity, we'll begin by examining the case in which we wish to estimate a scalar x based on observation of a vector

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_M \end{bmatrix}^{\mathrm{T}}. \tag{3.59}$$

In particular, let $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote the inner-product and associated norm, respectively, of the inner product space $\mathcal{V} = L^2(\Omega)$ of finite mean-square random variables with

$$\langle x, y \rangle = E[xy] \tag{3.60}$$

and thus

$$\|x\|^2 = E\left[x^2\right]. \tag{3.61}$$

Then we can rewrite (3.58) as

$$\hat{x}_{\mathrm{LLS}}(\mathbf{y}) = \arg\min_{w \in \mathcal{Y}} \|w - x\|^2 \tag{3.62a}$$

where

$$\begin{aligned} \mathcal{Y} &= \mathrm{span}(1, y_1, y_2, \ldots, y_M) \\ &= \left\{ w \in \mathcal{V} \mid w = d + \sum_{i=1}^{M} a_i y_i \text{ for some } a_1, a_2, \ldots, a_M \right\}. \end{aligned} \tag{3.62b}$$

Note that by rephrasing our problem in terms of (3.62), we have abstracted our estimation problem into one of approximation in an arbitrary inner product space. That is, we can now develop a solution to (3.62) without having to take into account the particular inner product that is relevant to this problem. As a result, we'll obtain a solution that can simultaneously solve a rich class of related approximation problems, one of which is the particular one we are interested in at present.

The problem that (3.62) poses in abstract vector space is the following. Given a particular element x in a vector space, how do we optimally approximate this element with an arbitrary linear combination of other elements $y_1, y_2, \ldots, y_M$ in this space. Note that all linear combinations of these approximating elements constitute a subspace of the vector space, which can be thought of as a hyperplane. Thus, our problem is to choose the element in this hyperplane that best approximates our element x, where x itself generally does not lie in this hyperplane.

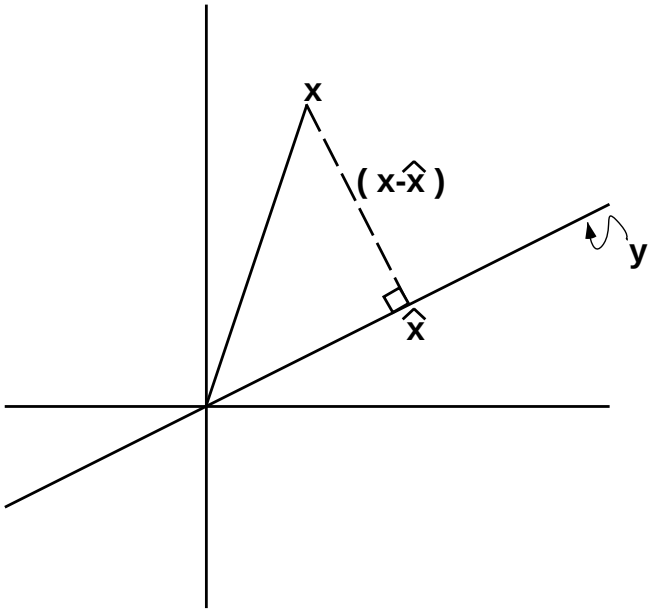

Figure 3.1. Approximation by projection in $\mathbb{R}^2$.

Based on our intuition about such approximations in $\mathbb{R}^N$, we would expect that the optimum approximation $\hat{x}$ would correspond to a *projection* of x onto the hyperplane, which would make the approximation error orthogonal to the hyperplane. Fig. 3.1 depicts this result in $\mathbb{R}^2$ when there is $M = 1$ approximating element.

In fact, this intuition and reasoning can be used to construct optimum approximations in *arbitrary* abstract vector spaces. The generalization and optimality of this result is given in terms of the celebrated *Orthogonal Projection Theorem*, which we now develop.

Theorem 3.2 (Orthogonal Projection) *Let $\mathcal{Y}$ be a subspace of a (complete) inner product space $\mathcal{V}$, and let $x \in \mathcal{V}$ be an arbitrary element. Then*

$$\hat{x} = \arg\min_{w \in \mathcal{Y}} \|w - x\|, \tag{3.63}$$

i.e., the approximation $\hat{x} \in \mathcal{Y}$ minimizes $\|\hat{x} - x\|$, if

$$e = \hat{x} - x \in \mathcal{Y}^{\perp},$$

i.e., if

$$(\hat{x} - x) \perp w \qquad \textit{for all } w \in \mathcal{Y}. \tag{3.64}$$

In order to simplify our proof of Theorem 3.2, let us first establish *Pythagoras' Theorem:* two elements v_1 and v_2 in an inner product space $\mathcal{V}$ are orthogonal, i.e., $v_1 \perp v_2$, if and only if

$$\|v_1 + v_2\|^2 = \|v_1\|^2 + \|v_2\|^2. \tag{3.65}$$

To derive this intermediate result, it suffices to note

$$\begin{aligned}\|v_1 + v_2\|^2 &= \langle v_1 + v_2, v_1 + v_2\rangle \\ &= \langle v_1, v_1\rangle + \langle v_1, v_2\rangle + \langle v_2, v_1\rangle + \langle v_2, v_2\rangle \\ &= \|v_1\|^2 + \|v_2\|^2 + 2\langle v_1, v_2\rangle . \end{aligned} \tag{3.66}$$

Clearly, the right-hand sides of (3.66) and (3.65) are equal if and only if v_1 and v_2 are orthogonal.

Returning now to our proof of Theorem 3.2, since $\hat{x} - x$ is orthogonal to every element of $\mathcal{Y}$, it is orthogonal to the particular element $w - \hat{x}$ for $w \in \mathcal{Y}$, i.e., $(\hat{x} - x) \perp (w - \hat{x})$. Hence, by Pythagoras' Theorem we have

$$\|x - w\|^2 = \|(\hat{x} - x) + (w - \hat{x})\|^2 = \|\hat{x} - x\|^2 + \|w - \hat{x}\|^2$$

from which we can conclude

$$\|x - w\|^2 \geq \|\hat{x} - x\|^2$$

with equality if and only if $w = \hat{x}$. Thus $\hat{x}$ is the solution to (3.63).

Note that the theorem yields a remarkably general result. Not only does it not depend on the choice of inner product, but it doesn't depend on the dimension of $\mathcal{Y}$ either. Indeed, $\mathcal{Y}$ could have infinite dimension, and in fact the infinite dimensional case will be an important focus later in the course. However, in the estimation problem we consider in this section, i.e., (3.62), we'll assume the corresponding subspace has finite dimension, i.e., $M < \infty$.

In the finite dimensional case, the Orthogonal Projection Theorem can be used to obtain a matrix representation for the optimum approximation $\hat{x}$. To see this, when $\mathcal{Y}$ is M-dimensional and spanned by the elements

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_M \end{bmatrix}^{\mathrm{T}},$$

any element of $\mathcal{Y}$ can be expressed as a linear combination of the y_i's, and in particular, since $\hat{x} \in \mathcal{Y}$, we can write

$$\hat{x} = \hat{\mathbf{a}}^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{M} \hat{a}_i y_i \tag{3.67}$$

for an appropriate choice of weights

$$\hat{\mathbf{a}} = \begin{bmatrix} \hat{a}_1 & \hat{a}_2 & \cdots & \hat{a}_M \end{bmatrix}^{\mathrm{T}}.$$

Let's now use the orthogonality condition (3.64) to determine what these weights must be. In particular, since $\hat{x}$ must satisfy

$$\langle x - \hat{x}, y_j\rangle = 0, \qquad j = 1, 2, \ldots, M, \tag{3.68}$$

we can substitute (3.67) into (3.68) to obtain

$$\langle x, y_j\rangle = \left\langle \left(\sum_{i=1}^{M} \hat{a}_i y_i\right), y_j \right\rangle = \sum_{i=1}^{M} \hat{a}_i \langle y_i, y_j\rangle \qquad j = 1, 2, \ldots, M. \tag{3.69}$$

Collecting the M equations of (3.69) into vector form we obtain the *normal equations*

$$\mathbf{R}_{x\mathbf{y}}^{\mathrm{T}} = \mathbf{R}_{\mathbf{yy}}\hat{\mathbf{a}}, \tag{3.70}$$

where

$$\hat{\mathbf{a}} = \begin{bmatrix}\hat{a}_1 & \hat{a}_2 & \cdots & \hat{a}_M\end{bmatrix}^{\mathrm{T}} \tag{3.71}$$

$$\mathbf{R}_{x\mathbf{y}} = \begin{bmatrix}\langle x, y_1\rangle & \langle x, y_2\rangle & \cdots & \langle x, y_M\rangle\end{bmatrix} \tag{3.72}$$

$$\mathbf{R}_{\mathbf{yy}} = \begin{bmatrix}\langle y_1, y_1\rangle & \langle y_2, y_1\rangle & \cdots & \langle y_M, y_1\rangle \\ \langle y_1, y_2\rangle & \langle y_2, y_2\rangle & \cdots & \langle y_M, y_2\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle y_1, y_M\rangle & \langle y_2, y_M\rangle & \cdots & \langle y_M, y_M\rangle\end{bmatrix}. \tag{3.73}$$

The matrix $\mathbf{R}_{\mathbf{yy}}$ in (3.73) is referred to a the *Grammian* matrix associated with the approximation problem. It is straightforward to verify that the Grammian is a positive semidefinite matrix.[6] Furthermore, if the $y_1, y_2, \ldots, y_M$ are a basis for $\mathcal{Y}$ (i.e., are a linearly independent set), then the Grammian is strictly positive definite and hence invertible. In this case, the optimal weights are given by

$$\hat{\mathbf{a}} = \mathbf{R}_{\mathbf{yy}}^{-1}\mathbf{R}_{x\mathbf{y}}^{\mathrm{T}}.$$

When the y_i's are not linearly independent, there is still a solution, but optimal weights $\hat{\mathbf{a}}$ are no longer unique. In particular, any solution of (3.70) will be optimal.

Again, we stress that this framework is remarkably general and can be used to solve a host of approximation problems. For example, when we let $\mathcal{V} = L^2(\mathbb{R})$ or $\mathcal{V} = \ell^2(\mathbb{Z})$, this framework solves the continuous- or discrete-time deterministic linear least-squares approximation problem.[7] Estimation problems can be viewed as a specific class of approximation problems involving random variables. In particular, when $\mathcal{V} = L^2(\Omega)$, the framework solves the linear least-squares estimation problem that is of primary interest in this section, which we now develop.

---

[6] Let $z = \mathbf{a}^{\mathrm{T}}\mathbf{y}$ where

$$\mathbf{a} = \begin{bmatrix}a_1 & a_2 & \cdots & a_M\end{bmatrix}^{\mathrm{T}}$$

and note that

$$0 \le \|z\|^2 = \langle z, z\rangle = \mathbf{a}^{\mathrm{T}}\mathbf{R}_{\mathbf{yy}}\mathbf{a}.$$

[7] A prototypical deterministic least-squares problem involves approximating some deterministic known but arbitrary function $x(t)$ as a linear combination of other deterministic known but fixed functions $y_1(t), y_2(t), \ldots, y_M(t)$, i.e.,

$$\hat{x}(t) = \sum_{i=1}^{M} a_i y_i(t)$$

so as to minimize the energy in the approximation error, i.e.,

$$\int_{-\infty}^{+\infty} [\hat{x}(t) - x(t)]^2 \, dt.$$

As an example, the approximating functions could be polynomials, e.g., $y_i(t) = t^i$.

In this case, we choose the particular inner product and associated norm given by (3.60) and (3.61), respectively, and let our approximating elements be the random variables (3.59). Now any affine estimator for x based on $\mathbf{y}$ can, without loss of generality, be expressed in the form

$$\hat{x}(\mathbf{y}) = d + \mathbf{a}^{\mathrm{T}}(\mathbf{y} - \mathbf{m}_{\mathbf{y}}). \tag{3.74}$$

In order to accommodate the additional constant term, we need to augment our observed data with one additional (deterministic) random variable that is the constant 1, so our data is now

$$\tilde{\mathbf{y}}_{+} = \begin{bmatrix} 1 \\ \tilde{\mathbf{y}} \end{bmatrix} \tag{3.75}$$

with

$$\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{m}_{\mathbf{y}}. \tag{3.76}$$

With this notation our estimator (3.74) takes the form

$$\hat{x}(\mathbf{y}) = \mathbf{a}_{+}^{\mathrm{T}}\tilde{\mathbf{y}}_{+}, \tag{3.77}$$

where

$$\mathbf{a}_{+} = \begin{bmatrix} d \\ \mathbf{a} \end{bmatrix}. \tag{3.78}$$

Proceeding, the corresponding normal equations (3.70), i.e.,

$$E\left[x\tilde{\mathbf{y}}_{+}\right] = E\left[\tilde{\mathbf{y}}_{+}\tilde{\mathbf{y}}_{+}^{\mathrm{T}}\right]\hat{\mathbf{a}}_{+},$$

become

$$\begin{bmatrix} m_x \\ \mathbf{\Lambda}_{x\mathbf{y}}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{\mathbf{y}} \end{bmatrix} \begin{bmatrix} \hat{d} \\ \hat{\mathbf{a}} \end{bmatrix}$$

from which we get the two equations

$$\hat{d} = m_x \tag{3.79}$$

and

$$\mathbf{\Lambda}_{x\mathbf{y}}^{\mathrm{T}} = \mathbf{\Lambda}_{\mathbf{y}}\hat{\mathbf{a}}. \tag{3.80}$$

Substituting (3.79) and (3.80) into (3.74) we obtain, when $\mathbf{\Lambda}_{\mathbf{y}}$ is nonsingular,

$$\hat{x}_{\mathrm{LLS}}(\mathbf{y}) = m_x + \hat{\mathbf{a}}^{\mathrm{T}}\tilde{\mathbf{y}} = m_x + \mathbf{\Lambda}_{x\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}\tilde{\mathbf{y}}. \tag{3.81}$$

Finally, substituting (3.76) into (3.81) gives

$$\hat{x}_{\mathrm{LLS}}(\mathbf{y}) = m_x + \mathbf{\Lambda}_{x\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{y}}). \tag{3.82}$$

Turning next to the performance of the resulting LLS estimator, we first note from (3.82) that the estimator is unbiased. This follows immediately from the fact that the estimation error must be orthogonal to the constant 1, which was an element of the data vector (3.75), i.e.,

$$E\left[(\hat{x}_{\mathrm{LLS}}(\mathbf{y}) - x) \cdot 1\right] = E\left[\hat{x}_{\mathrm{LLS}}(\mathbf{y}) - x\right] = 0, \tag{3.83}$$

which in turn means of course that the mean-square estimation error is the same as the error variance.

Second, we note that the error variance associated with this estimator can also be obtained in a particularly straightforward manner using the orthogonality condition. In particular, via Pythagoras' Theorem we have that

$$\begin{aligned}\lambda_{\mathrm{LLS}} = \lambda_e &= E\left[(x - \hat{x}_{\mathrm{LLS}}(\mathbf{y}))^2\right] \\ &= \|x - \hat{x}\|^2 = \|x\|^2 - \|\hat{x}\|^2 \\ &= E\left[x^2\right] - \|m_x + \hat{\mathbf{a}}^{\mathrm{T}}\tilde{\mathbf{y}}\|^2 \\ &= \lambda_x - \hat{\mathbf{a}}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{y}}\hat{\mathbf{a}} \\ &= \lambda_x - \mathbf{\Lambda}_{x\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}\mathbf{\Lambda}_{x\mathbf{y}}^{\mathrm{T}}, \end{aligned} \tag{3.84}$$

where to obtain the last equality in (3.84) we have used (3.80). This completes our derivation of the scalar LLS estimator.

Our results are easily extended to the case in which we want to construct the linear least-squares estimate of an N-vector $\mathbf{x}$ based on observations $\mathbf{y}$. To do this, we solve the problem in a component-wise manner, constructing the optimum linear least-squares solution for the estimation of each component of $\mathbf{x}$. The result is a set of estimates

$$\hat{x}_{i,\mathrm{LLS}}(\mathbf{y}) = m_{x_i} + \mathbf{\Lambda}_{x_i\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{y}}) \qquad i = 1, 2, \ldots, N, \tag{3.85}$$

which we collect into vector form as

$$\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) = \mathbf{m}_{\mathbf{x}} + \mathbf{\Lambda}_{\mathbf{xy}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{y}}). \tag{3.86}$$

Exploiting that this estimator is unbiased, the error covariance can be readily calculated. In particular, with the error written as

$$\mathbf{e} = \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x} = \mathbf{\Lambda}_{\mathbf{xy}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}\tilde{\mathbf{y}} - \tilde{\mathbf{x}}$$

where $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{m}_{\mathbf{x}}$, we obtain

$$\begin{aligned}\mathbf{\Lambda}_{\mathrm{LLS}} &= E\left[\mathbf{e}\mathbf{e}^{\mathrm{T}}\right] \\ &= E\left[\left(\mathbf{\Lambda}_{\mathbf{xy}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}\tilde{\mathbf{y}} - \tilde{\mathbf{x}}\right)\left(\mathbf{\Lambda}_{\mathbf{xy}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}\tilde{\mathbf{y}} - \tilde{\mathbf{x}}\right)^{\mathrm{T}}\right] \\ &= \mathbf{\Lambda}_{\mathbf{x}} - \mathbf{\Lambda}_{\mathbf{xy}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}\mathbf{\Lambda}_{\mathbf{xy}}^{\mathrm{T}}. \end{aligned} \tag{3.87}$$

As a final comment before we proceed to an example, note that as we anticipated both the LLS estimator (3.86) and its performance (3.87) depend only on the joint second-order statistics of $\mathbf{x}$ and $\mathbf{y}$. As a consequence, this means that we do not need to know the complete statistical characterization of $\mathbf{x}$, $\mathbf{y}$ and their relationship in order to construct and evaluate this estimator. On the other hand, we must also keep in mind that this implies the LLS estimator cannot exploit the higher-order statistical dependencies among $\mathbf{x}$ and $\mathbf{y}$ to better estimate $\mathbf{x}$. As a result, our LLS estimators can never give us a lower mean-square estimation error than our BLS estimators, which do fully exploit this additional information.

Example 3.5

Let's consider the random variables x and y from Example 3.4 again, but now find the LLS estimator for x based on y. First we note that by symmetry

$$m_x = m_y = 0. \tag{3.88a}$$

Furthermore, since x and w are independent we have

$$\lambda_{xy} = E[xy] = E[x(\operatorname{sgn} x + w)] = E[|x|] = 1/2, \tag{3.88b}$$

and

$$\lambda_y = E\left[y^2\right] = E\left[(\operatorname{sgn} x + w)^2\right] = E\left[\operatorname{sgn}^2 x\right] + E\left[w^2\right] = 1 + 1/3 = 4/3. \tag{3.88c}$$

Substituting (3.88) into (3.86) and (3.87) we obtain

$$\hat{x}_{\mathrm{LLS}}(y) = \frac{\lambda_{xy}}{\lambda_y} y = \frac{3}{8} y \tag{3.89}$$

and

$$\lambda_{\mathrm{LLS}} = \lambda_x - \frac{\lambda_{xy}^2}{\lambda_y} = \frac{1}{3} - \frac{(1/2)^2}{(4/3)} = 7/48. \tag{3.90}$$

Comparing (3.90) with (3.41) we see that, as expected, constraining our estimator to be linear leads to a larger mean-square estimation error.

Additional Properties of LLS Estimators

We've already established some important properties of LLS estimators. For instance we've shown that the LLS estimator is always unbiased [see (3.83)]. In this section, we develop several additional special properties of LLS estimators.

In Section 3.2.4 we showed in Theorem 3.1 that an estimator $\hat{\mathbf{x}}(\mathbf{y})$ is the Bayes' least-squares estimator if and only if the corresponding estimation error $\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}$ is orthogonal to *any* function of the observed data $\mathbf{y}$. From the orthogonality principle we used to derive the linear least-squares estimator, we immediately obtain the counterpart to Theorem 3.1 for linear estimators.

Theorem 3.3 *A linear estimator $\hat{\mathbf{x}}_{\mathrm{L}}(\cdot)$ is the linear least-squares estimator, i.e., $\hat{\mathbf{x}}_{\mathrm{L}}(\cdot) = \hat{\mathbf{x}}_{\mathrm{LLS}}(\cdot)$, if and only if the associated estimation error $\mathbf{e}(\mathbf{x}, \mathbf{y}) = \hat{\mathbf{x}}_{\mathrm{L}}(\mathbf{y}) - \mathbf{x}$ is orthogonal to* any *vector-valued linear (i.e., affine) function of the data, i.e.,*

$$E\left[[\hat{\mathbf{x}}_{\mathrm{L}}(\mathbf{y}) - \mathbf{x}]\,[\mathbf{F}\mathbf{y} + \mathbf{g}]^{\mathrm{T}}\right] = \mathbf{0} \tag{3.91}$$

for any constant matrix $\mathbf{F}$ and any constant vector $\mathbf{g}$.

The proof follows immediately from the fact that the estimation error is orthogonal to any of the elements of (3.75), and thus any linear combination of these elements as well.

As another property of LLS estimators, we have that the covariance of the BLS and LLS estimators are related according to the matrix inequalities

$$0 \leq \mathbf{\Lambda}_{\mathrm{BLS}} \leq \mathbf{\Lambda}_{\mathrm{LLS}} \leq \mathbf{\Lambda}_{\mathrm{L}} \tag{3.92}$$

where $\mathbf{\Lambda}_{\mathrm{L}}$ is the error covariance of *any* linear (i.e., affine) estimator $\hat{\mathbf{x}}_{\mathrm{L}}(\cdot)$, and where the rightmost inequality is satisfied with equality if and only if

$$\hat{\mathbf{x}}_{\mathrm{L}}(\mathbf{y}) - E\left[\hat{\mathbf{x}}_{\mathrm{L}}(\mathbf{y}) - \mathbf{x}\right] = \hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}). \tag{3.93}$$

Before deriving (3.92), we make several observations. First, the leftmost inequality in (3.92) is merely a restatement of the fact that covariance matrices are positive semidefinite, and equality holds when $\mathbf{x}$ can be determined with certainty from the data—this is what is called the "singular estimation" scenario.

Second, the middle inequality in (3.92) is an immediate consequence of (3.46) which holds for any estimator $\hat{\mathbf{x}}(\cdot)$ and therefore any linear estimator $\hat{\mathbf{x}}_{\mathrm{L}}(\cdot)$. Furthermore, as we will see shortly, this middle inequality in (3.92) is satisfied with equality when $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian. However, the converse is not true: there *do* exist non-Gaussian examples where the BLS estimator turns out to be a linear estimator.

Finally, since $\hat{\mathbf{x}}_{\mathrm{L}}(\mathbf{y}) = \mathbf{m}_{\mathbf{x}}$ is a valid linear estimator, and has an associated error covariance of $\mathbf{\Lambda}_{\mathrm{L}} = \mathbf{\Lambda}_{\mathbf{x}}$, we have as a special case of (3.92) the statement

$$\mathbf{\Lambda}_{\mathrm{LLS}} \leq \mathbf{\Lambda}_{\mathbf{x}}. \tag{3.94}$$

From (3.86), we see that (3.94) is satisfied with equality if and only if $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated.

We derive the rightmost inequality in (3.92) by following an approach analogous to that used to derive the corresponding result for BLS estimators, i.e., (3.46). In particular, let $\mathbf{b}_{\mathrm{L}}$ denote the bias in our estimator $\hat{\mathbf{x}}_{\mathrm{L}}(\cdot)$, let

$$\mathbf{h}(\mathbf{y}) = \hat{\mathbf{x}}_{\mathrm{L}}(\mathbf{y}) - \hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) - \mathbf{b}_{\mathrm{L}}, \tag{3.95}$$

and let us begin by noting that

$$\begin{aligned}
\mathbf{\Lambda}_{\mathrm{L}} &= E\left[\left(\hat{\mathbf{x}}_{\mathrm{L}}(\mathbf{y}) - \mathbf{x} - \mathbf{b}_{\mathrm{L}}\right)\left(\hat{\mathbf{x}}_{\mathrm{L}}(\mathbf{y}) - \mathbf{x} - \mathbf{b}_{\mathrm{L}}\right)^{\mathrm{T}}\right] \\
&= E\left[\left[\mathbf{h}(\mathbf{y}) + \left(\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) - \mathbf{x}\right)\right]\left[\mathbf{h}(\mathbf{y}) + \left(\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) - \mathbf{x}\right)\right]^{\mathrm{T}}\right] \\
&= E\left[\mathbf{h}(\mathbf{y})\mathbf{h}^{\mathrm{T}}(\mathbf{y})\right] + E\left[\left(\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) - \mathbf{x}\right)\left(\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) - \mathbf{x}\right)^{\mathrm{T}}\right] \\
&\quad + E\left[\left(\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) - \mathbf{x}\right)\mathbf{h}^{\mathrm{T}}(\mathbf{y})\right] + E\left[\left(\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) - \mathbf{x}\right)\mathbf{h}^{\mathrm{T}}(\mathbf{y})\right]^{\mathrm{T}}.
\end{aligned} \tag{3.96}$$

Since $\mathbf{h}(\mathbf{y})$ is a linear (i.e., affine) function of $\mathbf{y}$, from Theorem 3.3 we get that the last two terms in (3.96) are zero. Using this together with the definition of $\mathbf{\Lambda}_{\mathrm{LLS}}$ we get

$$\mathbf{\Lambda}_{\mathrm{L}} - \mathbf{\Lambda}_{\mathrm{LLS}} = E\left[\mathbf{h}(\mathbf{y})\mathbf{h}^{\mathrm{T}}(\mathbf{y})\right]. \tag{3.97}$$

The right-hand side of (3.97) is in general positive semidefinite, which verifies the rightmost inequality in (3.92), and equal to zero if and only if $\mathbf{h}(\mathbf{y}) = \mathbf{0}$, which, using (3.95), yields (3.93).

We finish this section with two examples of linear least-squares estimation.

Example 3.6

Consider the scalar problem of estimating a random variable x whose mean is m_x and whose variance is σ_x^2 based on observations of the form

$$y = hx + w \tag{3.98}$$

where h is a known deterministic constant, and where w has zero mean, variance σ_w^2, and is independent of x. In this case, to construct the LLS estimator (3.86) we need only determine the appropriate statistics. In particular, we have

$$\lambda_{xy} = h\sigma_x^2 \tag{3.99}$$

$$\sigma_y^2 = h^2\sigma_x^2 + \sigma_w^2 \tag{3.100}$$

$$m_y = hm_x \tag{3.101}$$

so that the LLS estimator is

$$\begin{aligned}\hat{x}_{\mathrm{LLS}}(y) &= m_x + \frac{h\sigma_x^2}{h^2\sigma_x^2 + \sigma_w^2}(y - hm_x) \\ &= \left[\frac{\sigma_w^2}{h^2\sigma_x^2 + \sigma_w^2}\right] m_x + \left[\frac{h^2\sigma_x^2}{h^2\sigma_x^2 + \sigma_w^2}\right] \frac{y}{h}\end{aligned} \tag{3.102}$$

and the associated error covariance is

$$\lambda_{\mathrm{LLS}} = \sigma_x^2 - \frac{h^2\sigma_x^4}{h^2\sigma_x^2 + \sigma_w^2} = \frac{\sigma_x^2\sigma_w^2}{h^2\sigma_x^2 + \sigma_w^2}. \tag{3.103}$$

Example 3.7

Next let's consider the vector generalization of Example 3.6, which arises in a host of practical problems. Specifically, suppose that $\mathbf{x}$ has mean $\mathbf{m_x}$ and covariance $\mathbf{\Lambda_x}$ and that our observations are noisy measurements of linear functions of $\mathbf{x}$, i.e.,

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w} \tag{3.104}$$

where $\mathbf{H}$ is a given matrix and $\mathbf{w}$ has, for convenience, zero-mean, covariance $\mathbf{\Lambda_w}$, and is uncorrelated with $\mathbf{x}$.[8]

[8]There is no loss of generality in assuming that $\mathbf{w}$ has zero-mean; if the measurement noise $\mathbf{w}$ were nonzero mean, we could simply subtract this mean from $\mathbf{y}$ to obtain an equivalent problem with zero-mean. Also, it is certainly reasonable to consider a scenario in which $\mathbf{w}$ is correlated with $\mathbf{x}$, though slightly more complicated expressions result in this case.

To construct the linear least-squares estimator for this problem simply requires that we determine the appropriate statistics in (3.86). In particular, we have

$$\begin{aligned}
\mathbf{m_y} &= E[\mathbf{y}] = \mathbf{H}E[\mathbf{x}] + E[\mathbf{w}] = \mathbf{H}\mathbf{m_x} && (3.105)\\
\mathbf{\Lambda_{xy}} &= E\left[(\mathbf{x}-\mathbf{m_x})(\mathbf{y}-\mathbf{m_y})^{\mathrm{T}}\right] \\
&= E\left[(\mathbf{x}-\mathbf{m_x})(\mathbf{H}(\mathbf{x}-\mathbf{m_x})+\mathbf{w})^{\mathrm{T}}\right] \\
&= E\left[(\mathbf{x}-\mathbf{m_x})(\mathbf{x}-\mathbf{m_x})^{\mathrm{T}}\right]\mathbf{H}^{\mathrm{T}} + E\left[(\mathbf{x}-\mathbf{m_x})\mathbf{w}^{\mathrm{T}}\right] \\
&= \mathbf{\Lambda_x}\mathbf{H}^{\mathrm{T}}, && (3.106)\\
\mathbf{\Lambda_y} &= E\left[(\mathbf{y}-\mathbf{m_y})(\mathbf{y}-\mathbf{m_y})^{\mathrm{T}}\right] \\
&= E\left[[\mathbf{H}(\mathbf{x}-\mathbf{m_x})+\mathbf{w}]\left[(\mathbf{x}-\mathbf{m_x})^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}+\mathbf{w}^{\mathrm{T}}\right]\right] \\
&= \mathbf{H}\mathbf{\Lambda_x}\mathbf{H}^{\mathrm{T}} + \mathbf{\Lambda_w}. && (3.107)
\end{aligned}$$

Then, from (3.86), (3.87) we obtain that

$$\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) = \mathbf{m_x} + \mathbf{K}(\mathbf{y}-\mathbf{H}\mathbf{m_x}) \quad (3.108)$$

$$\mathbf{\Lambda}_{\mathrm{LLS}} = \mathbf{\Lambda_x} - \mathbf{K}\left(\mathbf{H}\mathbf{\Lambda_x}\mathbf{H}^{\mathrm{T}} + \mathbf{\Lambda_w}\right)\mathbf{K}^{\mathrm{T}}, \quad (3.109)$$

where $\mathbf{K}$ is a gain matrix defined as

$$\mathbf{K} = \mathbf{\Lambda_x}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{\Lambda_x}\mathbf{H}^{\mathrm{T}} + \mathbf{\Lambda_w})^{-1}. \quad (3.110)$$

Note the intuitively appealing structure of (3.108)—our posterior estimate $\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y})$ equals our prior estimate $\mathbf{m_x}$ plus a correction term that is proportional to the difference between the observation $\mathbf{y}$ and our best prediction of the observation based on the prior information, i.e., $\mathbf{m_y} = \mathbf{H}\mathbf{m_x}$.

Note that the gain matrix $\mathbf{K}$ controls the relative weight placed on our prior information versus the observation. In particular, as $\mathbf{\Lambda_x}$ increases (e.g., in the sense of its trace), our prior information degrades in quality, and one would therefore want to place more weight on $\mathbf{y}$. If $\mathbf{\Lambda_w}$ increases (again in the sense of its trace), the quality of the measurement decreases and we would want less weight placed on the measurement. The gain matrix makes the tradeoff in a statistically optimal manner.

The gain matrix $\mathbf{K}$ also optimally captures interdependencies among the components of the vector to be estimated. This is especially important in applications where it is only possible to obtain measurements of *some* of the variables of interest. For example, consider the estimation of vehicle position x_1 and velocity x_2 in a tracking problem, and let

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Furthermore, suppose we are only able to obtain measurements of the position, so that

$$y = x_1 + w.$$

In this case, the only way in which y helps us to estimate the velocity x_2 is through its correlation with position x_1. This dependency is exploited as efficiently as possible via the gain matrix $\mathbf{K}$.

As a final comment, an alternate expression for $\mathbf{\Lambda}_{\mathrm{LLS}}$ in (3.109) that is derived in Appendix 3.A is (see (3.343))

$$\mathbf{\Lambda}_{\mathrm{LLS}}^{-1} = \mathbf{\Lambda_x}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{\Lambda_w}^{-1}\mathbf{H}. \quad (3.111)$$

Which form is more useful for practical computations of $\mathbf{\Lambda}_{\mathrm{LLS}}$ depends on a number of factors. Note for example, that (3.111) involves inversions of a matrix of size N, the dimension of $\mathbf{x}$, while (3.109) involves inversion of a matrix of size M, the dimensions of $\mathbf{y}$. Depending on the relative sizes of M and N, one form may be preferable to the other. Other considerations that influence the choice involve numerical stability issues, which we won't develop here.

In any case, the form (3.111) provides us with some valuable intuition. As we'll see shortly, the inverse of a covariance has a useful interpretation as a measure of information. What (3.111) states is that the information $\mathbf{\Lambda}_{\mathrm{LLS}}^{-1}$ about $\mathbf{x}$ after the measurement equals the prior information $\mathbf{\Lambda}_{\mathbf{x}}^{-1}$ plus the information $\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H}$ contained in the measurement.

Estimation in the Jointly Gaussian Case

In this section, we establish yet another very special property of jointly Gaussian random variables. In particular we have the remarkable result that if $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian random vectors, then

$$\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) = \hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}). \tag{3.112}$$

To see this, let

$$\mathbf{e}_{\mathrm{LLS}} = \hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) - \mathbf{x}, \tag{3.113}$$

and note that by Theorem 3.3 we have that $\mathbf{e}_{\mathrm{LLS}}$ must be orthogonal to every linear function of $\mathbf{y}$ and hence $\mathbf{y}$ itself. But since $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian, this means that $\mathbf{e}$ is actually statistically independent of $\mathbf{y}$. This implies, for example, that

$$E\left[\mathbf{e}_{\mathrm{LLS}}|\mathbf{y}\right] = E\left[\mathbf{e}_{\mathrm{LLS}}\right] = 0 \tag{3.114}$$

where the last equality follows from the fact that the LLS estimate is unbiased. But we also have directly from (3.113) and from (3.33) that

$$E\left[\mathbf{e}_{\mathrm{LLS}}|\mathbf{y}\right] = E\left[\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y})|\mathbf{y}\right] - E\left[\mathbf{x}|\mathbf{y}\right] = \hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) - \hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}). \tag{3.115}$$

Comparing (3.114) and (3.115) completes our derivation.

Note that this result provides a convenient derivation of the mean and covariance associated with the posterior density $p_{\mathbf{x}|\mathbf{y}}(\mathrm{x}|\mathrm{y})$ in the jointly Gaussian case, i.e., (1.149) and (1.150), respectively, from Chapter 1. To establish the mean expression (1.149) it suffices to combine (3.112) with (3.33) and (3.86). To establish the covariance expression (1.150) we first note, using (3.112) and the fact that $\mathbf{e}_{\mathrm{LLS}}$ and $\mathbf{y}$ are jointly Gaussian, that

$$\begin{aligned}
\mathbf{\Lambda}_{\mathbf{x}|\mathbf{y}} &= E\left[\left(\mathbf{x} - E\left[\mathbf{x}|\mathbf{y}\right]\right)\left(\mathbf{x} - E\left[\mathbf{x}|\mathbf{y}\right]\right)^{\mathrm{T}} \mid \mathbf{y}\right] \\
&= E\left[\mathbf{e}_{\mathrm{LLS}}\mathbf{e}_{\mathrm{LLS}}^{\mathrm{T}} \mid \mathbf{y}\right] \\
&= E\left[\mathbf{e}_{\mathrm{LLS}}\mathbf{e}_{\mathrm{LLS}}^{\mathrm{T}}\right] = \mathbf{\Lambda}_{\mathrm{LLS}}.
\end{aligned} \tag{3.116}$$

Combining (3.116) with (3.87) we get our desired posterior covariance expression. Again we emphasize that the resulting posterior covariance (1.150) is not a function of $\mathbf{y}$ in this jointly Gaussian case.

Finally, to verify that the posterior density for $\mathbf{x}$ is actually Gaussian we need only recognize that since

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{m_x}\\ \mathbf{m_y}\end{bmatrix}, \begin{bmatrix}\mathbf{\Lambda_x} & \mathbf{\Lambda_{xy}}\\ \mathbf{\Lambda_{xy}}^{\mathrm{T}} & \mathbf{\Lambda_y}\end{bmatrix}\right) \tag{3.117}$$

and since

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y})}{p_{\mathbf{y}}(\mathbf{y})},$$

we have that when viewed as a function of $\mathbf{x}$ alone (with $\mathbf{y}$ fixed), the posterior density satisfies

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\begin{bmatrix}\mathbf{x}-\mathbf{m_x}\\ \mathbf{y}-\mathbf{m_y}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{\Lambda_x} & \mathbf{\Lambda_{xy}}\\ \mathbf{\Lambda_{xy}}^{\mathrm{T}} & \mathbf{\Lambda_y}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{x}-\mathbf{m_x}\\ \mathbf{y}-\mathbf{m_y}\end{bmatrix}\right) \tag{3.118}$$

which is an exponentiated quadratic in $\mathbf{x}$. But the exponentiated quadratic family of densities are precisely the Gaussian densities, and thus the posterior density is Gaussian with the mean and covariance determined above. Obviously, results for scalar x and y correspond to a special case of what we've obtained above.

Note that as a consequence of this result we have that although the BLS estimator is in general a nonlinear function of the data, in the jointly Gaussian case it turns out to be a linear function of the data. Furthermore, as we mentioned earlier would be the case, from (3.112) we get immediately that the middle inequality in (3.92), i.e.,

$$\mathbf{\Lambda}_{\mathrm{BLS}} \leq \mathbf{\Lambda}_{\mathrm{LLS}},$$

is satisfied with equality in the jointly Gaussian case. In turn this implies that there are no statistical dependencies among $\mathbf{x}$ and $\mathbf{y}$ beyond second-order ones that the BLS can exploit in the jointly Gaussian case.

In summary then we have that when $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian,

$$\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) = \mathbf{m_x} + \mathbf{\Lambda_{xy}}\mathbf{\Lambda_y}^{-1}(\mathbf{y}-\mathbf{m_y}) \tag{3.119}$$

$$\mathbf{\Lambda}_{\mathrm{BLS}} = E\left[\mathbf{\Lambda_{x|y}}\right] = \mathbf{\Lambda_{x|y}} = \mathbf{\Lambda_x} - \mathbf{\Lambda_{xy}}\mathbf{\Lambda_y}^{-1}\mathbf{\Lambda_{xy}}^{\mathrm{T}}. \tag{3.120}$$

And we note that when $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated ($\mathbf{\Lambda_{xy}} = \mathbf{0}$), they are also independent in this case, and the BLS estimator and its performance degenerate to the prior statistics on $\mathbf{x}$ since $\mathbf{y}$ provides no information about $\mathbf{x}$.

There are other consequences of the fact that the posterior density for $\mathbf{x}$ is Gaussian. For example, since the Gaussian density is symmetric and unimodal, its mean is also its mode. Since the posterior mean is the BLS estimator and the posterior mode is the MAP estimator, we have that when $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian the two estimators coincide, i.e.,

$$\hat{\mathbf{x}}_{\mathrm{BLS}}(\mathbf{y}) = \hat{\mathbf{x}}_{\mathrm{MAP}}(\mathbf{y}). \tag{3.121}$$

As a result, in the jointly Gaussian case, the MAP estimator inherits a variety of useful properties. For example, in general MAP estimates are biased. However,

in the jointly Gaussian case they are unbiased. Note, however, that $\mathbf{x}$ and $\mathbf{y}$ need not be jointly Gaussian for (3.121) to hold. Indeed, any $p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y})$ such that the corresponding posterior density $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ is symmetric and unimodal, for instance, will have this property.

We finish this section with a scalar example.

Example 3.8

Suppose that x and y are scalar, jointly Gaussian random variables with

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} m_x \\ m_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \lambda_{xy} \\ \lambda_{xy} & \sigma_y^2 \end{bmatrix} \right). \tag{3.122}$$

Then

$$p_{x|y}(x|y) = N\left(x; \hat{x}_{\mathrm{BLS}}(y), \lambda_{\mathrm{BLS}}\right), \tag{3.123}$$

where

$$\hat{x}_{\mathrm{BLS}}(y) = m_x + \frac{\lambda_{xy}}{\sigma_y^2}(y - m_y) = m_x + \rho_{xy}\left(\frac{\sigma_x}{\sigma_y}\right)(y - m_y). \tag{3.124}$$

Furthermore,

$$\lambda_{\mathrm{BLS}} = \lambda_{x|y} = \sigma_x^2 - \frac{\lambda_{xy}^2}{\sigma_y^2} = \sigma_x^2(1 - \rho_{xy}^2), \tag{3.125}$$

where

$$\rho_{xy} = \frac{\lambda_{xy}}{\sigma_x \sigma_y}$$

is the correlation coefficient.

Several observations should be re-emphasized. First, in the jointly Gaussian case we were able to express the posterior variance as λ_{BLS} which does not depend on y since $\lambda_{x|y}(y)$ does not depend on the actual observed value of y, i.e.,

$$\lambda_{\mathrm{BLS}} = E\left[\lambda_{x|y}\right] = \lambda_{x|y}. \tag{3.126}$$

Second, note that because x and y are jointly Gaussian, when $\rho_{xy} = 0$ they are also independent. In this case, y contains no information about x, which is reflected in the fact that (3.124) and (3.125) reduce to the prior statistics on x. Conversely, larger values of $|\rho_{xy}|$ result in more weight being placed on information from the measurement, and a reduction in the posterior uncertainty—i.e., the uncertainty in x after incorporating knowledge of y.

Again we emphasize that because x and y are jointly Gaussian the resulting estimator (3.124) is a linear (or more precisely affine) function of the data y, and that in non-Gaussian cases BLS estimators are generally *nonlinear* functions of the data y.

Finally, from our earlier comments, we note that the MAP estimator and BLS estimators are identical in the jointly Gaussian case, so we immediately obtain

$$\hat{x}_{\mathrm{MAP}}(y) = \hat{x}_{\mathrm{BLS}}(y).$$

Note too that since the posterior density is symmetric and unimodal, its mean is also its median. Since the median of the posterior density is the minimum absolute-error estimator, we have as well

$$\hat{x}_{\mathrm{MAE}}(y) = \hat{x}_{\mathrm{BLS}}(y).$$

BLS Estimation with only Second-Moment Information

At the outset of Section 3.2.5, we showed that the construction of the BLS estimator of a random variable x from a random vector $\mathbf{y}$ subject to the constraint that the estimator be linear requires only knowledge of the joint second-moment properties of $(x, \mathbf{y})$.

This observation raises an interesting related question. Suppose we only have knowledge of the joint second-moment properties of a pair $(x, \mathbf{y})$, then what is the best possible estimator $\hat{x}(\mathbf{y})$ (in a BLS sense) we can construct, and how does it perform? In particular, we might reasonably ask whether the LLS estimator is also the solution to this problem.

To answer this question requires posing the problem as a game between two adversaries: the system designer tries to find the best estimator, and nature tries to find the model that makes the performance of the chosen estimator as bad as possible subject to the constraint that the $(x, \mathbf{y})$ statistics match the prescribed second-moment information.

For this game, we now determine the best estimator choice for the system designer, the worst joint distribution we can encounter, and the resulting mean-square estimator error. With the given moments being m_x, $\mathbf{m_y}$, σ_x^2, σ_y^2, and λ_{xy}, we seek to evaluate

$$\max_{p_{x\mathbf{y}} \in \mathcal{M}} \min_{f(\cdot)} E\left[(x - f(\mathbf{y}))^2\right], \tag{3.127}$$

where

$$\mathcal{M} = \left\{ p_{x\mathbf{y}} : \quad E\left[\begin{bmatrix} x \\ \mathbf{y} \end{bmatrix}\right] = \begin{bmatrix} m_x \\ \mathbf{m_y} \end{bmatrix}, \quad \operatorname{cov}\left(\begin{bmatrix} x \\ \mathbf{y} \end{bmatrix}, \begin{bmatrix} x \\ \mathbf{y} \end{bmatrix}\right) = \begin{bmatrix} \sigma_x^2 & \mathbf{\Lambda}_{x\mathbf{y}} \\ \mathbf{\Lambda}_{x\mathbf{y}}^{\mathrm{T}} & \mathbf{\Lambda}_{\mathbf{yy}} \end{bmatrix} \right\}. \tag{3.128}$$

We first note that, in general, a max-min can be upper bounded by a min-max. For our problem, this means that

$$\max_{p_{x\mathbf{y}} \in \mathcal{M}} \min_{f(\cdot)} E\left[(x - f(\mathbf{y}))^2\right] \leq \min_{f(\cdot)} \max_{p_{x\mathbf{y}} \in \mathcal{M}} E\left[(x - f(\mathbf{y}))^2\right]. \tag{3.129}$$

Intuitively, the minimizer is more powerful on the left side of (3.129) while the maximizer is more powerful on the right side. Indeed, the minimizer on the left side of (3.129) gets to choose an estimating function f that depends on the distribution chosen by the maximizer, while the opposite is true for the right side.

We can further upper bound the right side of (3.129) by substituting any function f_0. That is,

$$\min_{f(\cdot)} \max_{p_{x\mathbf{y}} \in \mathcal{M}} E\left[(x - f(\mathbf{y}))^2\right] \leq \max_{p_{x\mathbf{y}} \in \mathcal{M}} E\left[(x - f_0(\mathbf{y}))^2\right]. \tag{3.130}$$

For any linear function $f_0(y) = d + \mathbf{a}^{\mathrm{T}}(\mathbf{y} - \mathbf{m_y})$ the right side of (3.130) only depends on the second-moment statistics, which are fixed. Let us further choose $d = m_x$ and $\mathbf{a}^{\mathrm{T}} = \mathbf{\Lambda}_{x\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}$, which results in

$$E\left[(x - f_0(\mathbf{y}))^2\right] = E\left[\left(x - m_x - \mathbf{\Lambda}_{x\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}(\mathbf{y} - \mathbf{m_y})\right)^2\right] = \sigma_x^2 - \mathbf{\Lambda}_{x\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}\mathbf{\Lambda}_{x\mathbf{y}}^{\mathrm{T}} \tag{3.131}$$

for any $p_{x\mathbf{y}} \in \mathcal{M}$. Thus, the right side of (3.131) is an upper bound to (3.127).

We now proceed to show that the right side of (3.131) is also a lower bound to (3.127). Similarly to our upper bound in (3.130), we can lower bound (3.127) by choosing any distribution on x and y, i.e.,

$$\max_{p_{x\mathbf{y}} \in \mathcal{M}} \min_{f(\cdot)} E\left[(x - f(\mathbf{y}))^2\right] \geq \min_{f(\cdot)} E\left[(x - f(\mathbf{y}))^2\right] \tag{3.132}$$

where the expectation on the righthand side of (3.132) is with respect to an arbitrary distribution $p^*_{x\mathbf{y}} \in \mathcal{M}$. Let us choose $p^*_{x\mathbf{y}}$ as that corresponding to x and y being jointly Gaussian with the specified second-moment statistics. For jointly Gaussian random variables, the BLS estimate is the linear estimate $\hat{x} = m_x + \mathbf{\Lambda}_{x\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{y}})$. The resulting mean square error is that given on the right side of (3.131).

Since (3.127) is both upper and lower bounded by the right side of (3.131), we conclude that 1) the system designer should choose the LLS estimator, 2) nature should choose the jointly Gaussian model matching the second-moment statistics, and 3) the resulting mean-square error performance will be $\sigma_x^2 - \mathbf{\Lambda}_{x\mathbf{y}}\mathbf{\Lambda}_{\mathbf{y}}^{-1}\mathbf{\Lambda}_{x\mathbf{y}}^{\mathrm{T}}$.

Gram-Schmidt Orthogonalization

In many approximation problems in vector space, it turns out to be much more convenient to work with approximating elements that are orthogonal to one another, and normalized. Indeed, when the approximating elements form an orthonormal set, the normal equations take a particularly simple form since the Grammian is then just the identity matrix!

Now any set of elements can be replaced with a corresponding set of orthonormal elements that span the same space or subspace. Furthermore, this new orthonormal set can be generated from the original set of elements in a efficient recursive manner. This is the essence of the Gram-Schmidt orthogonalization process. You've probably seen this procedure used a linear algebra course for regular vectors in $\mathbb{R}^N$. However, with our broader view of vector space in this course, we now see that the appropriate abstraction of this procedure can be used with arbitrary vector spaces. In particular, if we apply this to some subspace of zero-mean random variables, then the Gram-Schmidt procedure generates a set of uncorrelated random variables from a set of correlated ones. This notion will form the basis for some very efficient and powerful estimation algorithms we'll discuss later in the course.

To begin our discussion of the Gram-Schmidt process for arbitrary vector spaces, we consider the general problem of approximating an element $x \in \mathcal{V}$ as a linear combination of $y_1, y_2, \ldots, y_M \in \mathcal{V}$, i.e., as in (3.67). As developed earlier, the optimal values of the $\hat{a}_i$ are obtained from the normal equations (3.70) by inverting the Grammian, which is a computationally intensive task in general. In addition, suppose that we add one more vector, y_{M+1}, to the set on which we will base our approximation. In general, the optimum values of the coefficients $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_M$

of the other y_i's *will all change*—i.e., the procedure is not a simple recursive one. Suppose, however, the y_i are orthogonal, i.e., $\langle y_i, y_j \rangle = 0$ if $i \neq j$. In this case, the solution to (3.73) yields

$$\hat{a}_i = \frac{\langle x, y_i \rangle}{\langle y_i, y_i \rangle}, \qquad i = 1, 2, \ldots, M \tag{3.133}$$

which implies that each $\hat{a}_i$ can be calculated individually, in essence representing the best approximation of x using that single element y_i.

The preceding remarks suggest a powerful strategy for solving the normal equations in general which is of particular importance in *recursive* approximation in which the y_i are received *sequentially*. The basic idea here is that if the y_i's are not orthogonal, we'll transform them so that the resulting elements *are* orthogonal. This procedure for accomplishing this, which we now describe, is referred to as *Gram-Schmidt* orthogonalization.

First note that if we let $\hat{y}[i|i-1]$ denote the best linear approximation of y_i based on the preceding y's, i.e., based on $y_1, y_2, \ldots, y_{i-1}$, then the sequence $z_1, z_2, \ldots$ obtained from the sequence $y_1, y_2, \ldots$ according to

$$z_1 = y_1 \tag{3.134a}$$

$$z_i = y_i - \hat{y}[i|i-1] \qquad i \geq 2 \tag{3.134b}$$

has two important properties.

First, the z's are orthogonal, i.e.,

$$\langle z_i, z_j \rangle = 0 \qquad i \neq j \tag{3.135}$$

To see this, assume, without loss of generality, that $i > j$. Then note that from its definition (3.134b) as an *approximation error* and the Orthogonal Projection Theorem, z_i is orthogonal to $y_1, y_2, \ldots, y_{i-1}$. However, z_j is a linear combination of $y_1, y_2, \ldots, y_j$ with $j \leq i-1$. Therefore, (3.135) holds.

The second important property is

$$\text{span}(y_1, y_2, \ldots, y_k) = \text{span}(z_1, z_2, \ldots, z_k), \qquad k = 1, 2, \ldots \tag{3.136}$$

This implies that the best approximation that can be obtained in terms of the y's is the same as that which can be obtained in terms of the z's. Since the latter are orthogonal, the approximation in terms of the z's is easier to compute.

To verify (3.136), we begin by noting that since z_i is *defined* as a linear combination of $y_1, y_2, \ldots, y_i$, we need only show the reverse, i.e., that each y_i is also a linear combination of $z_1, z_2, \ldots, z_i$. We'll show this by mathematical induction. First, from (3.134a) we see that (3.136) is trivially true for $k = 1$. Next, we assume that (3.136) is true for $k \leq i-1$, and proceed to show that this implies it must be true for $k = i$. In particular, using (3.134b) we see that

$$y_i = z_i + \hat{y}[i|i-1] \tag{3.137}$$

but $\hat{y}[i|i-1]$ is a linear combination of $y_1, y_2, \ldots, y_{i-1}$, which by the induction hypothesis can also be written as a linear combination of $z_1, z_2, \ldots, z_{i-1}$. This completes our proof.

The result just described suggests the following recursive approximation procedure. Let $\hat{x}[i]$ denote the best approximation to x based on $y_1, y_2, \ldots, y_i$ or equivalently, $z_1, z_2, \ldots, z_i$. We then begin by computing

$$\hat{x}[1] = K_1 z_1, \qquad K_1 = \frac{\langle x, z_1 \rangle}{\langle z_1, z_1 \rangle} \tag{3.138}$$

At the next step, we receive y_2 and first compute

$$z_2 = y_2 - \gamma_{21} z_1, \qquad \gamma_{21} = \frac{\langle y_2, z_1 \rangle}{\langle z_1, z_1 \rangle} \tag{3.139}$$

and then

$$\hat{x}[2] = K_1 z_1 + K_2 z_2, \qquad K_2 = \frac{\langle x, z_2 \rangle}{\langle z_2, z_2 \rangle}. \tag{3.140}$$

We stress that once we have computed z_2, (3.140) tells us that it is very simple to compute $\hat{x}[2]$ by *updating* our previous approximation, i.e., (3.140) states that

$$\hat{x}[2] = \hat{x}[1] + K_2 z_2 \tag{3.141}$$

More generally, at step i we have

$$\hat{x}[i] = \hat{x}[i-1] + K_i z_i, \qquad K_i = \frac{\langle x, z_i \rangle}{\langle z_i, z_i \rangle} \tag{3.142}$$

In general, of course, z_i must be computed via

$$z_i = y_i - \gamma_{i,1} z_1 - \gamma_{i,2} z_2 - \cdots - \gamma_{i,i-1} z_{i-1} \tag{3.143a}$$

with

$$\gamma_{ij} = \frac{\langle y_i, z_j \rangle}{\langle z_j, z_j \rangle}. \tag{3.143b}$$

The preceding algorithm has a rather natural interpretation when our vector space is the subspace of $L^2(\Omega)$ corresponding to zero-mean random variables. In this case x and the $y_1, y_2, \ldots$ are all zero-mean random variables, and $\hat{y}[i|i-1]$ is the best *estimate* of y_i given previous measurements of $y_1, y_2, \ldots, y_{i-1}$. Moreover, z_i is the corresponding estimation error, i.e., we can think of z_i as the *one-step prediction error* in estimating y_i. We can also think of z_i as the *new* information in y_i that is not predictable based on observations of $y_1, y_2, \ldots, y_{i-1}$. For this reason the sequence $z_1, z_2, \ldots$ is often referred to as the *innovations* sequence. In the terminology of stochastic processes to be developed in the next chapter of the notes, this innovations sequence is an example of a *white* process, in that the sequence consists of uncorrelated random variables, and the Gram-Schmidt procedure in this context can be viewed as a *whitening filter*,—it takes the correlated sequence $y_1, y_2, \ldots$ as input and produces the white sequence $z_1, z_2, \ldots$ as output.

We can also interpret the Gram-Schmidt procedure applied to random variables as another strategy for diagonalizing a covariance matrix. To see this, we begin by observing that transformation from y's to z's—and hence, via (3.136), from z's to y's—is linear. We can express this in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \mathbf{\Gamma} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_M \end{bmatrix}. \tag{3.144}$$

for some matrix $\mathbf{\Gamma}$. Note however, that because of the recursive nature of the algorithm, the associated matrix $\mathbf{\Gamma}$ is *lower triangular*, and its coefficients in the lower triangle are precisely the γ_{ij}'s we constructed via projections in (3.143b), i.e.,

$$[\mathbf{\Gamma}]_{ij} = \begin{cases} 1 & i = j \\ \gamma_{ij} & i > j \\ 0 & \text{otherwise} \end{cases}$$

Note that by substituting recursively for the z's on the right-hand side of (3.143b), we can also readily compute $\mathbf{\Gamma}^{-1}$, which we see is also lower triangular. Specifically,

$$\begin{aligned} z_1 &= y_1 \\ z_2 &= y_2 - \gamma_{21} z_1 = y_2 - \gamma_{21} y_1 \\ z_3 &= y_3 - \gamma_{31} z_1 - \gamma_{32} z_2 \\ &= y_3 - \gamma_{31} y_1 - \gamma_{32}(y_2 - \gamma_{21} y_1) \\ &\vdots \end{aligned} \tag{3.145}$$

To complete our interpretation of Gram-Schmidt as a diagonalization, note that using (3.144) we get that

$$\mathbf{\Lambda_y} = E\left[\mathbf{y}\mathbf{y}^{\mathrm{T}}\right] = \mathbf{\Gamma}\mathbf{\Lambda_z}\mathbf{\Gamma}^{\mathrm{T}}, \tag{3.146}$$

and, in turn,

$$\mathbf{\Lambda_y}^{-1} = \mathbf{\Gamma}^{-\mathrm{T}}\mathbf{\Lambda_z}^{-1}\mathbf{\Gamma}^{-1}. \tag{3.147}$$

Since the z's are uncorrelated, this means that $\mathbf{\Lambda_z}$ is indeed *diagonal*. The factorization (3.146) is generally referred to as an LDU (lower-triangular–diagonal–upper-triangular) decomposition of a matrix, and as (3.147) indicates, once we have this decomposition, it is particularly easy to compute $\mathbf{\Lambda_y}^{-1}$. This interpretation provides additional perspective on the the Gram-Schmidt procedure.

It is worth commenting that this is now the second method for diagonalizing a covariance matrix that we've encountered. In Chapter 1 we developed a diagonalization based on the eigenvalue decomposition of the covariance matrix. There are, in fact, others as well that are useful. What you should keep in mind is that

which diagonalization is most useful in practice depends greatly on the type of problem being addressed.

As a final comment, we note that in many problems of interest, the computation of the z's can be extremely simple. For example, in a large class of estimation problems the observations y_i are given explicitly as functions of the variable x to be estimated. In such a case $\hat{y}[i|i-1]$ can be computed directly from $\hat{x}[i-1]$. We illustrate this through the following example.

Example 3.9

Suppose that

$$y_i = h_i x + w_i, \qquad i = 1, 2, \ldots \tag{3.148}$$

where the h_i are known numbers, x has zero mean and variance σ_x^2, and the w_i are uncorrelated, zero-mean random variables with variances σ_i^2 and are also uncorrelated with x.

We again let $\hat{x}[i]$ denote our optimum estimate of x based on observing $y_1, y_2, \ldots, y_i$, and denote the corresponding mean-square error in these estimates by $\sigma_x^2[i]$. With this notation, we use $\sigma_x^2[0]$ to denote the variance of x, i.e., the mean-square error before any observation is made; hence $\sigma_x^2[0] = \sigma_x^2$.

To initiate the recursion, we consider the estimation of x based on the first measurement y_1. This is just the scalar version of our LLS estimator problem, so from (3.102) and (3.103) we obtain

$$\hat{x}[1] = K_1 y_1 \tag{3.149a}$$

$$K_1 = \frac{\sigma_x^2[0] h_1}{h_1^2 \sigma_x^2[0] + \sigma_1^2}, \tag{3.149b}$$

and the variance of the estimation error $\hat{x}[1] - x$ is

$$\sigma_x^2[1] = \sigma_x^2[0] - \frac{h_1^2 \lambda_x^2[0]}{h_1^2 \sigma_x^2[0] + \sigma_1^2} = \frac{\sigma_1^2 \sigma_x^2[0]}{h_1^2 \sigma_x^2[0] + \sigma_1^2}. \tag{3.150}$$

At the next step, we first need to compute the best estimate of $y_2 = h_2 x + w_2$ based on y_1. However, since w_2 is uncorrelated with y_1, we have that

$$\hat{y}[2|1] = h_2 \hat{x}[1], \tag{3.151}$$

so

$$z_2 = y_2 - h_2 \hat{x}[1]. \tag{3.152}$$

Note that no new estimates are needed to generate z_2. More generally,

$$z_i = y_i - h_i \hat{x}[i-1], \tag{3.153}$$

and furthermore, from (3.142) we have that

$$\begin{aligned} \hat{x}[i] &= \hat{x}[i-1] + K_i z_i \\ &= \hat{x}[i-1] + K_i [y_i - h_i \hat{x}[i-1]], \end{aligned} \tag{3.154}$$

where

$$K_i = \frac{\lambda_{x z_i}}{\lambda_{z_i}}. \tag{3.155}$$

To calculate the statistics in (3.155), we obtain z_i from (3.148) and (3.153) as

$$z_i = h_i(x - \hat{x}[i-1]) + w_i. \tag{3.156}$$

Then, since w_i is uncorrelated with both x and $y_1, y_2, \ldots$ we see that

$$\lambda_{z_i} = h_i^2 \sigma_x^2[i-1] + \sigma_i^2. \tag{3.157}$$

Similarly, we obtain that

$$\begin{aligned}\lambda_{xz_i} &= h_i E\left[x(x - \hat{x}[i-1])\right] + E\left[x w_i\right] \\ &= h_i E\left[(x - \hat{x}[i-1])^2\right] + h_i E\left[\hat{x}[i-1](x - \hat{x}[i-1])\right] \\ &= h_i \sigma_x^2[i-1],\end{aligned} \tag{3.158}$$

where we have used the fact that x and w_i are uncorrelated, and the fact that the error $\hat{x}[i-1] - x$ is uncorrelated with any linear combination of $y_1, y_2, \ldots, y_{i-1}$ and thus in particular with $\hat{x}[i-1]$. Substituting (3.157) and (3.158) into (3.155) and combining the result with (3.154) we obtain the recursion for our estimator, viz.,

$$\hat{x}[i] = \hat{x}[i-1] + K_i\left(y_i - h_i \hat{x}[i-1]\right) \tag{3.159a}$$

$$K_i = \frac{h_i \sigma_x^2[i-1]}{h_i^2 \sigma_x^2[i-1] + \sigma_i^2}. \tag{3.159b}$$

Since the computation (3.159) involves the sequence of mean-square errors $\sigma_x^2[1], \sigma_x^2[2], \ldots$ we must also calculate these recursively. We obtain this recursion by direct computation. In particular, using (3.148) and (3.159) we obtain

$$\begin{aligned}\sigma_x^2[i] &= E\left[(x - \hat{x}[i])^2\right] \\ &= E\left[\left[(1 - K_i h_i)(x - \hat{x}[i-1]) - K_i w_i\right]^2\right] \\ &= (1 - K_i h_i)^2 \sigma_x^2[i-1] + K_i^2 \sigma_i^2 \\ &= \sigma_x^2[i-1] - \frac{h_i^2 \sigma_x^2[i-1]^2}{h_i^2 \sigma_x^2[i-1] + \sigma_i^2} \\ &= \frac{\sigma_i^2 \sigma_x^2[i-1]}{h_i^2 \sigma_x^2[i-1] + \sigma_i^2}.\end{aligned} \tag{3.160}$$

What we have just derived is a simple example of a *Kalman filter*. The general development and study of this and related topics forms the focus of a major portion of the advanced graduate subject in recursive estimation, 6.433.

3.3 NONRANDOM PARAMETER ESTIMATION

In many types of applications, we are interested in estimating certain parameters of the observed data. For example, given noisy measurements of a sinusoid, we might be interested in estimating the frequency of the sinusoid. In such cases it is often unnatural or inappropriate to view these parameters as random. Rather, it makes more sense to view them as deterministic quantities, but quantities that are nevertheless unknown.

As an example, suppose we have a sequence of independent identically distributed Gaussian random variables $y_1, y_2, \ldots, y_N$, where the mean m and variance σ^2 that parameterize the density are unknown. We'll explore some approaches to the problem of developing good estimators for these kinds of nonrandom parameters, and in doing so, we'll explore the kinds of performance criteria that are typically used in evaluating such estimators.

We stress at the outset that our Bayesian framework can't be adapted in any straightforward way to handle nonrandom parameter estimators. To see this, consider the scalar parameter case with a least-squares cost criterion. If we attempt to construct an estimate $\hat{x}(\mathbf{y})$ via

$$\hat{x}(\cdot) = \arg\min_{f(\cdot)} E\left[(x - f(\mathbf{y}))^2\right] \tag{3.161}$$

we see that we obtain a degenerate solution. In particular, noting that the expectation in (3.161) is over $\mathbf{y}$ alone (since x is deterministic), we immediately obtain that the right-hand side of (3.161) is minimized by choosing $\hat{x}(\mathbf{y}) = x$, and hence the optimum estimator according to (3.161) depends on the very parameter we're trying to estimate!

Obviously, regardless of the performance criterion we choose we'll want to restrict our search to "valid" estimators, i.e., estimators that don't depend explicitly on the parameters we're trying to estimate. In this portion of the notes, we'll develop some ways of thinking about and approaching the problem of finding valid estimators that yield good performance.

In our treatment, we will use $\mathbf{x}$ to denote the vector of parameters we wish to estimate, and write the density for the vector of observations $\mathbf{y}$ as $p_{\mathbf{y}}(\mathbf{y}; \mathbf{x})$ so as to make the parameterization explicit. In addition, we will use $\mathbf{m}_{\mathbf{y}}(\mathbf{x})$ and $\mathbf{\Lambda}_{\mathbf{y}}(\mathbf{x})$ to denote, respectively, the mean vector and covariance matrix of $\mathbf{y}$, again to make the parameterization explicit.

3.3.1 Bias and Error Covariance

As in the case of random parameters, two important measures of the performance of an estimator for nonrandom parameters are the bias and error covariance. However, there are some important distinctions between these quantities in the nonrandom case, which will become apparent in this section.

Using

$$\mathbf{e}(\mathbf{y}) = \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x} = \hat{\mathbf{x}} - \mathbf{x} \tag{3.162}$$

as our notation for the error, we define the bias in an estimator $\hat{\mathbf{x}}(\cdot)$ as

$$\begin{aligned}\mathbf{b}_{\hat{\mathbf{x}}}(\mathbf{x}) &= E\left[\mathbf{e}(\mathbf{y})\right] = E\left[\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}\right] \\ &= \int_{-\infty}^{+\infty} \left[\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}\right] p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\, d\mathbf{y} \\ &= \left[\int_{-\infty}^{+\infty} \hat{\mathbf{x}}(\mathbf{y})\, p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\, d\mathbf{y}\right] - \mathbf{x} \end{aligned} \tag{3.163}$$

Likewise, we express the error covariance as

$$\mathbf{\Lambda}_{\mathbf{e}}(\mathbf{x}) = E\left[\left[\mathbf{e}(\mathbf{y}) - \mathbf{b}_{\hat{\mathbf{x}}}(\mathbf{x})\right]\left[\mathbf{e}(\mathbf{y}) - \mathbf{b}_{\hat{\mathbf{x}}}(\mathbf{x})\right]^{\mathrm{T}}\right], \tag{3.164}$$

where, again, the expectation is with respect to $\mathbf{y}$.

We stress that both the bias (3.163) and error covariance (3.164) are, in general, functions of the parameter $\mathbf{x}$. Moreover, since the parameter $\mathbf{x}$ is unknown, removing a bias $\mathbf{b}_{\hat{\mathbf{x}}}(\mathbf{x})$ as we discussed in the case of random parameter estimators is generally not feasible in the nonrandom case—what bias to subtract would be a function of the very quantity we wish to estimate. In other words, modifying such an estimator by subtracting the bias will render it an "invalid" estimator.

In general the error correlation (and, in turn, its trace—the mean-square estimation error) depends on both bias and error covariance; specifically

$$E\left[\mathbf{e}(\mathbf{y})\mathbf{e}^{\mathrm{T}}(\mathbf{y})\right] = \mathbf{\Lambda}_{\mathbf{e}}(\mathbf{x}) + \mathbf{b}_{\hat{\mathbf{x}}}(\mathbf{x})\mathbf{b}_{\hat{\mathbf{x}}}^{\mathrm{T}}(\mathbf{x}) \tag{3.165}$$

From this expression we see that we may not want simply to minimize $\mathbf{\Lambda}_{\mathbf{e}}(\mathbf{x})$ if this leads to a large bias. To illustrate this point, suppose, for example, we take as our estimate $\hat{\mathbf{x}}(\mathbf{y})$ a constant vector independent of $\mathbf{y}$. In this case, $\mathbf{\Lambda}_{\mathbf{e}}(\mathbf{x}) = \mathbf{0}$, but the bias could be arbitrarily large.

For these reasons, a reasonable approach to developing good estimators for a nonrandom parameter is to explicitly restrict our search for estimators to those that are valid and unbiased,[9] and among this class, choose the one having the smallest variance. This is the notion underlying minimum-variance unbiased estimators, which we discuss next.

As one final comment before we explore this topic, note that in contrast to the case of random parameters, for nonrandom parameter estimators we have that the error covariance is the same as the covariance of the estimator itself, i.e.,

$$\mathbf{\Lambda}_{\mathbf{e}}(\mathbf{x}) = \mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x}) = E\left[\left(\hat{\mathbf{x}}(\mathbf{y}) - E\left[\hat{\mathbf{x}}(\mathbf{y})\right]\right)\left(\hat{\mathbf{x}}(\mathbf{y}) - E\left[\hat{\mathbf{x}}(\mathbf{y})\right]\right)^{\mathrm{T}}\right].$$

To see this in the scalar parameter case, simply note that using (3.162), we have

$$\begin{aligned}\lambda_{\mathsf{e}}(x) = E\left[e^2(\mathbf{y})\right] &= E\left[\left(\left(\hat{x}(\mathbf{y}) - x\right) - \left(E\left[\hat{x}(\mathbf{y})\right] - x\right)\right)^2\right] \\ &= E\left[\left(\hat{x}(\mathbf{y}) - E\left[\hat{x}(\mathbf{y})\right]\right)^2\right] \\ &= \lambda_{\hat{x}}(x).\end{aligned}$$

[9] We say an estimator $\hat{x}(\cdot)$ for a nonrandom parameter x is unbiased if $b_{\hat{x}}(x) = 0$ for *all* possible values of x.

We begin by considering the case where the parameter to be estimated is a scalar x, which simplifies our exposition much as it did for the case of random parameter estimation. We again note in advance that what we develop will readily generalize to the vector case, since the estimation of a vector of parameters can be accomplished in a component-wise manner.

3.3.2 Minimum-Variance Unbiased Estimators

To begin, let $\mathcal{A}$ denote the set of all estimators that are valid (i.e., don't depend on x) and unbiased, i.e.,

$$\mathcal{A} = \{\hat{x}(\cdot) \mid \hat{x}(\cdot) \text{ is valid and } b_{\hat{x}}(x) = 0\}$$

Then, when it exists, a minimum-variance unbiased (MVU) estimator for x is defined to be the estimator in $\mathcal{A}$ with the smallest variance, i.e.,

$$\hat{x}_{\text{MVU}}(\cdot) = \arg\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}(x) \qquad \text{for all } x \tag{3.166}$$

Several observations regarding (3.166) are worth emphasizing. The first is that $\hat{x}_{\text{MVU}}(\cdot)$ may not exist! For example, for some problems the set $\mathcal{A}$ is empty—there are no valid unbiased estimators. In other cases, $\mathcal{A}$ is not empty, but no estimator in $\mathcal{A}$ has a smaller variance than all the others *for all values of the parameter* x. Suppose for example that $\mathcal{A}$ consists of three estimators $\hat{x}_1(\cdot)$, $\hat{x}_2(\cdot)$, and $\hat{x}_3(\cdot)$, whose variances are plotted as a function of the unknown parameter x in Fig. 3.2. In this case, there is no estimator having a smaller variance than all the others for all values of x.

It should also be emphasized that even when $\hat{x}_{\text{MVU}}(\cdot)$ does exist, it may be difficult to find. In fact in general there is no systematic procedure for either determining whether an MVU estimator exists, or for computing it when it is does exist. However, fortunately there are cases in which such estimators can be computed, as we'll discuss later. As an example, we can determine the MVU estimator in the linear/Gaussian case. Also, when we further restrict our attention to *linear* MVU estimators, we'll see that these can often be computed as well—in particular, whenever the mean of the observations is a linear (affine) function of the parameter.

Sometimes, it is useful to exploit a bound on $\lambda_{\hat{x}}(x)$ in our quest for MVU estimators. A particular useful bound for this purpose is the Cramér-Rao bound, which we explore next.

3.3.3 The Cramér-Rao Bound

Again we consider first the estimation of an unknown scalar parameter x given a measurement vector $\mathbf{y}$ with density $p_{\mathbf{y}}(\mathbf{y}; x)$. When it exists, the Cramér-Rao bound

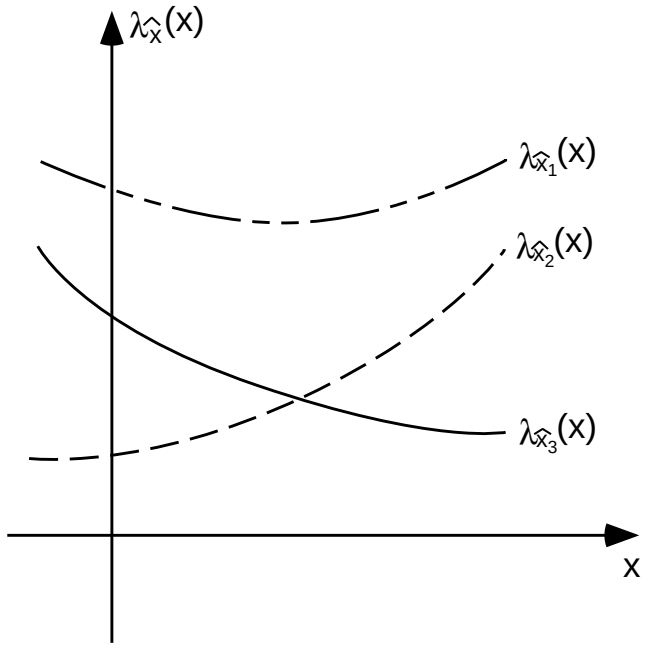

Figure 3.2. The variances of three unbiased estimators.

gives a lower bound on the variance of *any* valid unbiased estimator $\hat{x}(\cdot)$ for x. In particular, the Cramér-Rao bound for any $\hat{x}(\cdot) \in \mathcal{A}$ is

$$\lambda_{\hat{x}}(x) \geq \frac{1}{I_{\mathbf{y}}(x)}, \tag{3.167}$$

where the nonnegative quantity $I_{\mathbf{y}}(x)$ is referred to as the *Fisher information* in $\mathbf{y}$ about x, which is defined by

$$I_{\mathbf{y}}(x) = E\left[\left(\frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y};x)\right)^2\right]. \tag{3.168}$$

Some preliminary remarks are worth making. First, we stress that the Fisher information cannot be computed in all problems, in which case no Cramér-Rao bound exists. For example, for densities such as

$$p_y(y;x) = \begin{cases} 1 & x < y < x+1 \\ 0 & \text{otherwise} \end{cases},$$

which are not strictly positive for all x and y, the logarithm in (3.168) doesn't exist and hence $I_{\mathbf{y}}(x)$ can't be calculated.

Second, the notion of referring to (3.168) as an information measure comes from the fact that $I_{\mathbf{y}}(x)$ is both nonnegative and additive, i.e., whenever

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_M \end{bmatrix}^{\mathrm{T}}$$

consists of mutually independent components we have

$$I_{\mathbf{y}}(x) = \sum_{i=1}^{N} I_{y_i}(x).$$

Third, the Fisher information (3.168) can be interpreted as a measure of curvature: it measures, on average, how "peaky" $\ln p_{\mathbf{y}}(\mathbf{y};x)$ is as a function of x. As such, the larger $I_{\mathbf{y}}(x)$, the better we expect to be able to resolve the value of x from the observations, and hence the smaller we expect $\lambda_{\hat{x}}(x)$ to be. We'll develop this interpretation further in Section 3.4.

Example 3.10

Consider the scalar Gaussian problem

$$y = x + w,$$

where $w \sim N(0,\sigma^2)$. Then

$$\ln p_y(y;x) = -\frac{1}{2\sigma^2}(x-y)^2 - \frac{1}{2}\ln(2\pi\sigma^2) \tag{3.169}$$

Here the Fisher information is

$$I_y(x) = \frac{1}{\sigma^2},$$

so the smaller the variance σ^2 the sharper the peak of (3.169) is as a function of x.

To derive the Cramér-Rao bound (3.167), we begin by recalling that for unbiased estimators the error

$$e(\mathbf{y}) = \hat{x}(\mathbf{y}) - x \tag{3.170}$$

has zero mean, i.e.,

$$E\left[e(\mathbf{y})\right] = 0, \tag{3.171}$$

and variance

$$\operatorname{var} e(\mathbf{y}) = E\left[e^2(\mathbf{y})\right] = \lambda_{\hat{x}}(x). \tag{3.172}$$

Next we define

$$f(\mathbf{y}) = \frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y};x) \tag{3.173}$$

and note that using the identity

$$\frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y};x) = \frac{1}{p_{\mathbf{y}}(\mathbf{y};x)}\frac{\partial}{\partial x}p_{\mathbf{y}}(\mathbf{y};x), \tag{3.174}$$

we get that $f(\mathbf{y})$ has zero mean:

$$\begin{aligned} E\left[f(\mathbf{y})\right] &= E\left[\frac{1}{p_{\mathbf{y}}(\mathbf{y};x)}\frac{\partial}{\partial x}p_{\mathbf{y}}(\mathbf{y};x)\right] \\ &= \int_{-\infty}^{+\infty}\frac{\partial}{\partial x}p_{\mathbf{y}}(\mathbf{y};x)\,d\mathbf{y} \\ &= \frac{\partial}{\partial x}\int_{-\infty}^{+\infty}p_{\mathbf{y}}(\mathbf{y};x)\,d\mathbf{y} = \frac{\partial}{\partial x}1 = 0, \end{aligned} \tag{3.175}$$

and, in turn, variance

$$\operatorname{var} f(\mathbf{y}) = E\left[f^2(\mathbf{y})\right] = I_{\mathbf{y}}(x). \tag{3.176}$$

Finally, again using the identity (3.174), the covariance between $e(\mathbf{y})$ and $f(\mathbf{y})$ is given by

$$\begin{aligned}\operatorname{cov}\left(e(\mathbf{y}), f(\mathbf{y})\right) &= E\left[e(\mathbf{y})f(\mathbf{y})\right] \\ &= \int_{-\infty}^{+\infty} (\hat{x}(\mathbf{y}) - x)\frac{\partial}{\partial x} p_{\mathbf{y}}(\mathbf{y}; x)\, d\mathbf{y} \\ &= \left[\frac{\partial}{\partial x}\int_{-\infty}^{+\infty} \hat{x}(\mathbf{y})\, p_{\mathbf{y}}(\mathbf{y}; x)\, d\mathbf{y}\right] - \left[x\frac{\partial}{\partial x}\int_{-\infty}^{+\infty} p_y(\mathbf{y}; x)\, d\mathbf{y}\right] \\ &= 1 - 0 = +1. \end{aligned} \tag{3.177}$$

Now recall from from Chapter 1 that the correlation coefficient associated with $e(\mathbf{y})$ and $f(\mathbf{y})$ satisfies, via the Cauchy-Schwarz inequality, the bound

$$\rho_{ef}^2 = \frac{\left[\operatorname{cov}\left(e(\mathbf{y}), f(\mathbf{y})\right)\right]^2}{\operatorname{var} e(\mathbf{y}) \operatorname{var} f(\mathbf{y})} \leq 1. \tag{3.178}$$

Finally, substituting (3.172), (3.176) and (3.177) into (3.178) we get the Cramér-Rao bound (3.167).

Several additional comments regarding the bound provide important insights. First, the bound (3.167) in general depends on x, which we don't know. However, by plotting it as a function of x, we get a sense for the relative difficulty of estimating x as a function of its true value. In addition, we can extract best- and worst-case scenarios.

Second, any estimator that satisfies the Cramér-Rao bound with equality must be a MVU estimator. Note however, that the converse is not true: the Cramér-Rao bound may not be tight. Sometimes no estimator can meet the bound for all x, or even for any x! An estimator which achieves the bound, i.e., satisfies (3.167) with equality is referred to as an *efficient* estimator. Hence, efficient estimators are MVU estimators, but the converse need not be true. We'll return to a discussion of efficiency shortly.

Third, the Cramér-Rao bound can be generalized in a variety of ways. For example, a Cramér-Rao bound can be constructed for *biased* estimates. However, in practice this bound is not particularly useful. Likewise, there is an analogous bound for random parameters (see, e.g., Van Trees). However, this bound is not widely used, primarily because we always have a tight bound on the error variance of random parameter estimates, viz.,

$$\operatorname{var} e(x, \mathbf{y}) = \operatorname{var}\left[\hat{x}(\mathbf{y}) - x\right] \geq E\left[\lambda_{x|\mathbf{y}}(\mathbf{y})\right]$$

with equality if and only if $\hat{x}(\mathbf{y}) = \hat{x}_{\mathrm{BLS}}(\mathbf{y}) = E\left[x|\mathbf{y}\right]$.

In practice, the Fisher information (3.168) is frequently more useful when re-expressed in the following form

$$I_{\mathbf{y}}(x) = -E\left[\frac{\partial^2}{\partial x^2}\ln p_{\mathbf{y}}(\mathbf{y}; x)\right]. \tag{3.179}$$

To verify (3.179), we begin by observing

$$\int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};x)\,d\mathbf{y} = 1. \tag{3.180}$$

Differentiating (3.180) with respect to x and using the identity (3.174) yields

$$\int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};x)\frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y};x)\,d\mathbf{y} = 0. \tag{3.181}$$

Finally, differentiating (3.181) once more with respect to x and again using (3.174) we obtain

$$\int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};x)\left[\frac{\partial^2}{\partial x^2}\ln p_{\mathbf{y}}(\mathbf{y};x)\right]\,d\mathbf{y} + \int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};x)\left[\frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y};x)\right]^2\,d\mathbf{y} = 0, \tag{3.182}$$

which verifies that (3.168) and (3.179) are consistent.

Efficiency and Consistency

From our derivation of the Cramér-Rao bound (3.167) and in particular from (3.178), we note that the Cramér-Rao bound is satisfied with equality if and only if the functions $e(\mathbf{y})$ and $f(\mathbf{y})$ defined in (3.170) and (3.173), respectively, are perfectly positively correlated, i.e., if and only if there exists some constant $k(x) > 0$ (i.e., that can only depend on x) such that

$$e(\mathbf{y}) = k(x)f(\mathbf{y}) \qquad \text{for all } \mathbf{y}. \tag{3.183}$$

As we mentioned earlier, we refer to estimators that satisfy the Cramér-Rao bound with equality as efficient estimators. Rearranging (3.183) using (3.170) and (3.173), we obtain that an efficient estimator $\hat{x}(\cdot)$ must take the form

$$\hat{x}(\mathbf{y}) = x + k(x)\,\frac{\partial}{\partial x}\ln p_{\mathbf{y}}(\mathbf{y};x). \tag{3.184}$$

Hence, an efficient estimator exists if and only if (3.184) is a valid estimator, i.e., if and only if the right-hand side of (3.184) is independent of x for some $k(x)$.

However, $k(x)$ cannot, in fact, be arbitrary. To see this, let us suppose that an efficient estimator exists, so that (3.167) is satisfied with equality. Then, via (3.172) we must have

$$E\left[e^2(\mathbf{y})\right] = \lambda_{\tilde{x}}(x) = \frac{1}{I_{\mathbf{y}}(x)}. \tag{3.185}$$

Next note that using (3.183), (3.176), and (3.177) we obtain

$$E\left[e^2(\mathbf{y})\right] = E\left[e(\mathbf{y})\cdot k(x)f(\mathbf{y})\right] = k(x)E\left[e(\mathbf{y})f(\mathbf{y})\right] = k(x) \tag{3.186}$$

Comparing (3.185) and (3.186), we can then conclude that

$$k(x) = \frac{1}{I_{\mathbf{y}}(x)}. \tag{3.187}$$

Thus, substituting (3.187) into (3.184), we obtain the following characterization for efficient estimators: an estimator $\hat{x}(\cdot)$ is efficient if and only if it can be expressed in the form

$$\hat{x}(\mathbf{y}) = x + \frac{1}{I_{\mathbf{y}}(x)} \frac{\partial}{\partial x} \ln p_{\mathbf{y}}(\mathbf{y}; x) \tag{3.188}$$

where the right-hand side must be independent of x for the estimator to be valid.

Three final remarks are important. First, note that an efficient estimator, i.e., a valid estimator satisfying (3.188) is guaranteed to be unbiased: taking the expectation of (3.188) we get, using (3.175),

$$E\left[\hat{x}(\mathbf{y})\right] = x + \frac{1}{I_{\mathbf{y}}(x)} E\left[f(\mathbf{y})\right] = x.$$

Second, we note that (3.188) implies that when it exists, an efficient estimator is also *unique*—clearly no two estimators could satisfy (3.188) and be distinct. Finally, since it meets a lower bound on the estimator variance, when it exists, an efficient estimator must be the unique MVU estimator for a problem.

Let us now consider another desirable property of estimators. Suppose we have an estimator for x based on a sequence of observations $y_1, y_2, \ldots$, and let us specifically denote by $\hat{x}_M$ the estimate of x based on $y_1, y_2, \ldots, y_M$, i.e.,

$$\hat{x}_M = \hat{x}(y_1, y_2, \ldots, y_M).$$

Then we will say the estimator $\hat{x}_M$ is a *consistent* estimator for x if

$$\hat{x}_M \to x \qquad \text{as } M \to \infty. \tag{3.189}$$

Note that since $\hat{x}_M$ is a random variable for each M we have to say what we mean by the convergence condition (3.189), so let's briefly discuss the four main notions of convergence that are associated with sequences of random variables. To make the discussion as general as possible, let z be an arbitrary random variable and let $z_1, z_2, \ldots$ denote a sequence of related random variables. We note in advance that z being a deterministic constant corresponds to a special case of our discussion; z then has the degenerate density consisting of a single impulse at its actual value.

The weakest form of convergence is termed "convergence in distribution." The sequence $z_1, z_2, \ldots$ is said to converge in distribution to z if, when $P_z(\cdot)$ denotes the distribution function for z and $P_{z_n}(\cdot)$ denotes that for z_n, we have

$$\lim_{n\to\infty} P_{z_n}(z) = P_z(z)$$

for all z at which $P_z(z)$ is continuous. The notation

$$z_n \xrightarrow{\text{d}} z$$

is often used to specifically denote convergence in distribution. Note that this does not say that the values of the random variables z_n are getting close to the

value of z for large n, merely that their statistics are. This is, of course, the kind of convergence that the Central Limit Theorem we discussed in Chapter 1 involves. We also emphasize that convergence in distribution does not ensure convergence in density, as was apparent in our discussion of the Central Limit Theorem in particular.

A second form of convergence is termed "convergence in probability" or "p-convergence." We say that $z_1, z_2, \ldots$ converges in probability to z if for every fixed $\epsilon > 0$ we have

$$\lim_{n\to\infty} \Pr\left[|z_n - z| > \epsilon\right] = 0.$$

The notation

$$z_n \xrightarrow{\text{p}} z$$

is sometimes used to denote convergence in probability. This kind of convergence is much stronger than convergence in distribution, and says something about the actual values of the z_n's converging to z. Convergence in probability implies convergence in distribution, then, but of course the converse is not true. As an example, the weak law of large numbers is a statement about the convergence in probability of certain averages.

A still stronger notion of convergence is termed "mean-square convergence" or "convergence in the mean." We say $z_1, z_2, \ldots$ converges in mean-square (or "in the mean") to z if

$$\lim_{n\to\infty} E\left[(z_n - z)^2\right] = 0.$$

This kind of convergence is usually denoted using

$$z_n \xrightarrow{\text{m.s.}} z$$

or sometimes

$$\operatorname*{l.i.m.}_{n\to\infty} z_n = z.$$

Using the Chebyshev inequality we discussed in Chapter 1, we can readily establish that convergence in mean-square implies convergence in probability. However, it is important to note (although it may not be obvious at first glance) that the converse isn't true.

Another very strong notion of convergence is termed "almost-sure" or "probability-1" convergence. We say $z_1, z_2, \ldots$ converges almost-surely (or with probability-1) to z if

$$\Pr\left[\lim_{n\to\infty} z_n(\omega) = z(\omega), \quad \text{for all } \omega \in \mathcal{A} \subset \Omega, \Pr[\mathcal{A}] = 1\right] = 1.$$

This kind of convergence is often denoted using

$$z_n \xrightarrow{\text{a.s.}} z,$$

and, while a technically somewhat difficult definition to digest, effectively requires that (almost) every realization $z_1, z_2, \ldots$ of the sequence of random variables $z_1, z_2, \ldots$ converges to the corresponding realization z of z. It is this kind

of convergence that is involved in the strong law of large numbers. Almost-sure convergence also implies convergence in probability, but again the converse is not true.

Almost-sure convergence is the most desirable form of convergence in many problems. However, it is often difficult to establish. By contrast, mean-square convergence is often comparatively easier to work with, and is well-suited to engineering problems. For this reason, we'll generally restrict our attention to the latter form of convergence. Keep in mind, however, that mean-square convergence is not a weaker form of convergence than almost-sure convergence. In particular, mean-square convergence neither implies nor is implied by almost-sure convergence. A full discussion of these issues and counterexamples is beyond the scope of our treatment. However, good discussions can be found in a variety of advanced probability texts.

Let us now return to our development of the notion of consistency of an estimator. Based on our discussion above, we'll restrict our attention to consistency in the sense of mean-square convergence of (3.189), so $\hat{x}_M$ will be a consistent estimator when

$$E\left[(\hat{x}_M - x)^2\right] \to 0 \qquad \text{as } M \to \infty. \tag{3.190}$$

As a final comment before we explore a couple of examples, note that in general there is no relationship between efficiency and consistency: an efficient estimator need not be consistent, and a consistent estimator need not be efficient.

Example 3.11

Let's continue with the linear Gaussian problem we began in Example 3.10, i.e.,

$$y = x + w, \tag{3.191}$$

where $w \sim N(0, \sigma^2)$. In this case

$$\ln p_y(y; x) = -\ln(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y - x)^2, \tag{3.192}$$

so that

$$\frac{\partial^2}{\partial x^2} \ln p_y(y; x) = -\frac{1}{\sigma^2}, \tag{3.193}$$

and from (3.179)

$$I_y(x) = \frac{1}{\sigma^2}. \tag{3.194}$$

From the Cramér-Rao bound (3.167), we get that variance of any unbiased estimator satisfies

$$\lambda_{\hat{x}}(x) \geq \sigma^2. \tag{3.195}$$

Constructing the right-hand side of (3.188) using (3.192) and (3.194) we obtain

$$\hat{x}(y) = y \tag{3.196}$$

which we note is not a function of x and is therefore valid. Hence, we can immediately conclude that $\hat{x} = \hat{x}(y)$ defined via (3.196) is unbiased and has a variance equal to the Cramér-Rao bound, i.e.,

$$\lambda_{\hat{x}}(x) = \sigma^2. \tag{3.197}$$

Hence, we can conclude that (3.196) is an efficient estimator, and hence the unique MVU estimator for the problem.

Example 3.12

Let's consider a generalization of Example 3.11. In particular, suppose that we now have a set of observations of x of the form

$$y_i = x + w_i \qquad i = 1, 2, \ldots, M \tag{3.198}$$

where the w_i are independent identically-distributed random variables with densities $N(0, \sigma^2)$. In this case,

$$\ln p_{\mathbf{y}}(\mathbf{y}; x) = \frac{-M}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{M} (y_i - x)^2, \tag{3.199}$$

and hence

$$\frac{\partial}{\partial x} \ln p_{\mathbf{y}}(\mathbf{y}; x) = \frac{1}{\sigma^2} \sum_{i=1}^{M} (y_i - x). \tag{3.200}$$

From (3.200) and (3.168) we then obtain

$$I_{\mathbf{y}}(x) = \frac{M}{\sigma^2}. \tag{3.201}$$

If we again construct an estimator from the right-hand side of (3.188) using (3.200) and (3.201), we obtain

$$\hat{x}(\mathbf{y}) = \frac{1}{M} \sum_{i=1}^{M} y_i \tag{3.202}$$

which a valid estimator. Hence, (3.202) is unbiased and also an efficient estimator for the problem, so its variance is

$$\lambda_{\hat{x}} = 1/I_{\mathbf{y}}(x) = \frac{\sigma^2}{M}. \tag{3.203}$$

Note too that our estimator (3.202) also happens to be consistent in this example, i.e., from (3.203) we have

$$\lambda_{\hat{x}} = \frac{\sigma^2}{M} \to 0 \qquad \text{as } M \to \infty.$$

3.3.4 Maximum Likelihood Estimation

To develop the topic of maximum likelihood estimators, we begin with the following observation regarding efficient estimators. Specifically, suppose an efficient estimator exists for a particular problem of interest, and let $\hat{x}_{\text{eff}}(\cdot)$ denote this estimator. Hence, for any particular value of the data $\mathbf{y}$ we have, rewriting (3.188),

$$\hat{x}_{\text{eff}}(\mathbf{y}) = x + \frac{1}{I_{\mathbf{y}}(x)} \frac{\partial}{\partial x} \ln p_{\mathbf{y}}(\mathbf{y}; x). \tag{3.204}$$

which we can compute directly. Now since the right-hand side of (3.204) is independent of the value of x, we are free to choose *any* value of x in this expression,[10] so let us judiciously choose x to be the number

$$\hat{x}_{\mathrm{ML}}(\mathbf{y}) = \arg\max_x p_{\mathbf{y}}(\mathbf{y}; x). \tag{3.205}$$

Since $p_{\mathbf{y}}(\mathbf{y}; x)$ is typically referred to as the *likelihood function* of the data $\mathbf{y}$, (3.205) is referred to as the *maximum likelihood* (ML) estimator for x based on $\mathbf{y}$.

From (3.205) we see that provided the likelihood function is strictly positive and differentiable, the ML estimator satisfies

$$\left[\frac{\partial}{\partial x} \ln p_{\mathbf{y}}(\mathbf{y}; x)\right]\Bigg|_{x=\hat{x}_{\mathrm{ML}}(\mathbf{y})} = 0. \tag{3.206}$$

Thus, since $I_{\mathbf{y}}(x) > 0$ for all x except in the trivial case, (3.204) becomes

$$\hat{x}_{\mathrm{eff}}(\mathbf{y}) = \hat{x}_{\mathrm{ML}}(\mathbf{y}). \tag{3.207}$$

From this we can conclude that *when it exists*, the (unique) efficient estimator is equivalent to the ML estimator for the problem. For future convenience, we'll use $\lambda_{\mathrm{ML}}(x)$ to denote the variance (and hence error variance) of the estimator (3.205).

However, several points should be stressed. This does not mean the ML estimators are always efficient! When an efficient estimator doesn't exist for a problem, then the ML estimator need not have any special properties. This means, for example, that when an efficient estimator does not exist, the ML estimator may not have good variance properties or even be unbiased.

Nevertheless, ML estimators are highly practical—in particular, there exists a systematic procedure for obtaining them from data. In problems where the likelihood, for a particular observed value of the data $\mathbf{y}$, is a sufficiently tractable and differentiable function of the parameter x, we may compute the ML estimate for that $\mathbf{y}$ as follows. First, we analytically determine local maxima of the likelihood function, i.e., solutions to

$$\frac{\partial}{\partial x} p_{\mathbf{y}}(\mathbf{y}; x) = 0, \tag{3.208}$$

for which

$$\frac{\partial^2}{\partial x^2} p_{\mathbf{y}}(\mathbf{y}; x) < 0.$$

Then, we search over these local maxima and any boundary values for the largest value of the likelihood function. In some problems, it often turns out to be easier to maximize some monotonic function of the likelihood rather than the likelihood itself. For example, in a variety of problems maximizing the log-likelihood function $\ln p_{\mathbf{y}}(\mathbf{y}; x)$ simplifies computations significantly.

It is worth pointing out, however, that the number of problems for which solutions to (3.208) can be obtained as closed-form expressions is relatively small.

[10] In particular, we need not choose x to be its true value.

More typically, iterative numerical techniques such as gradient searches (Newton-Raphson) are used to find the local maxima of the likelihood function, from which the global maximum is selected. In addition to general-purpose iterative ascent algorithms, there also exist iterative ascent algorithms that are specifically tailored to the special characteristics of likelihood functions. One class of these algorithms are the so-called Estimate-Maximize (EM) algorithms, which have proven useful in a wide range of practical estimation problems.

There are additional reasons why ML estimators have proven popular in many applications even when they aren't efficient estimators. For example, in many cases these estimators have good *asymptotic* properties, i.e., when the size of the vector $\mathbf{y}$ gets sufficiently large.

To develop the necessary concepts, as we did in Section 3.3.3, we again consider the estimation of a parameter x based on a sequence of related observations $y_1, y_2, \ldots$, and we let $\hat{x}_M$ denote the estimate based on the first M observations, i.e., on $y_1, y_2, \ldots, y_M$. We then say $\hat{x}_M$ is an *asymptotically unbiased* estimator if

$$E[\hat{x}_M] \to x \qquad \text{as } M \to \infty. \tag{3.209}$$

In addition, we say that an at least asymptotically unbiased estimator $\hat{x}_M$ is *weakly asymptotically efficient* if

$$\lambda_{\hat{x}_M} - I^{-1}_{y_1,y_2,\ldots,y_M}(x) \to 0 \qquad \text{as } M \to \infty. \tag{3.210}$$

Furthermore, we say that an estimator is *strongly asymptotically efficient* if it is weakly asymptotically efficient and

$$\frac{I^{-1}_{y_1,y_2,\ldots,y_M}(x) - I^{-1}_{y_1,y_2,\ldots,y_\infty}(x)}{\lambda_{\hat{x}_M} - \lambda_{\hat{x}_\infty}} \to 1 \qquad \text{as } M \to \infty. \tag{3.211}$$

Also, note that in general the concepts of consistency and asymptotic efficiency need not be related. However, if for all x

$$I_{y_1,y_2,\ldots,y_M}(x) \to \infty \qquad \text{as } M \to \infty,$$

then an estimator that is even weakly asymptotically efficient is also consistent.

In many problems, the ML estimator is not efficient but has the property that it is asymptotically efficient and often consistent. Moreover, in a substantial subset of such problems it is not only asymptotically efficient but also asymptotically Gaussian; specifically

$$\hat{x}_M \sim N\left(x, 1/I_{y_1,y_2,\ldots,y_M}(x)\right) \qquad \text{as } M \to \infty.$$

While these "folk theorems" are often used casually to justify the optimality of ML estimators, it should be emphasized that while they are often true, it is easy to construct counterexamples for which the ML estimator is neither asymptotically efficient nor even asymptotically unbiased. For example, as will become apparent later in this chapter, for observations of the form

$$y_1 = h(x) + w_1 \tag{3.212}$$

$$y_i = w_i \qquad i \geq 2 \tag{3.213}$$

where $h(\cdot)$ is an invertible nonlinear function and where the w_i are independent identically-distributed $N(0, \sigma^2)$ random variables, the ML estimator for x based on $y_1, y_2, \ldots, y_M$ for *any* M is

$$\hat{x}_{\mathrm{ML}}(\mathbf{y}) = h^{-1}(y_1) \tag{3.214}$$

However, for almost any choice of $h(\cdot)$ the ML estimator (3.214) is neither efficient nor unbiased. Thus, since (3.214) is also independent of M, it is neither asymptotically efficient nor unbiased either.

One class of problems for which ML estimators are always efficient and therefore MVU estimators are the linear/Gaussian problems. We consider the canonical scalar version of this problem in the following example.

Example 3.13

Consider the scalar linear/Gaussian problem

$$y = hx + w \tag{3.215}$$

where $w \sim N(0, \sigma_w^2)$. Note that

$$p_y(y; x) = N(y; hx, \sigma_w^2) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left[-\frac{1}{2\sigma_w^2}(y - hx)^2\right] \tag{3.216}$$

so that the ML estimator simply inverts h and ignores the noise, i.e.,

$$\hat{x}_{\mathrm{ML}}(y) = \frac{y}{h}. \tag{3.217}$$

It is straightforward to verify this estimator is unbiased, i.e.,

$$E\left[\hat{x}_{\mathrm{ML}}(y) - x\right] = E\left[\frac{hx + w}{h} - x\right] = \frac{1}{h}E\left[w\right] = 0 \tag{3.218}$$

and that its variance is

$$\lambda_{\mathrm{ML}}(x) = E\left[\frac{w^2}{h^2}\right] = \frac{\sigma_w^2}{h^2} \tag{3.219}$$

Note that in this case the estimator variance turns out to be independent of x. Furthermore, the estimator variance is equal to the reciprocal of the Fisher information for the problem, i.e.,

$$\lambda_{\mathrm{ML}}(x) = \frac{\sigma_w^2}{h^2} = 1/I_y(x),$$

and therefore the ML estimator is efficient.

It is interesting to compare the ML estimator in this example to the LLS estimator for the closely related problem developed in Example 3.6. In both examples, the measurement models (3.215) and (3.98) are identical, but in this example x is a nonrandom parameter while in Example 3.6 we have a random parameter x with zero-mean and variance σ_x^2.

If we add the Gaussian assumptions to Example 3.6, we can conclude that the resulting LLS estimator (3.102) is also the BLS estimator, the MAP estimator, and the MAE estimator for the problem. For this reason, we'll simply use $\hat{x}_{\mathrm{B}}(y)$ to denote this estimator for the remainder of this example. Furthermore, since our ML

estimate is efficient, it is the MVU estimator for the nonrandom parameter estimation problem, so we'll use $\hat{x}_{\mathrm{MVU}}(y)$ to denote this estimator for the remainder of this example.

First, comparing (3.102) and (3.217) we see that

$$\lim_{\sigma_x^2 \to \infty} \hat{x}_{\mathrm{B}}(y) = \hat{x}_{\mathrm{MVU}}(y) \tag{3.220}$$

which indicates that as our prior knowledge about x in the random parameter case deteriorates (so that $p_x(x)$ becomes increasingly flat) the Bayesian estimate approaches the MVU estimate. In fact, from (3.102) we can see that the Bayesian estimate is a linear combination of the best prior estimate m_x and the MVU estimate y/h, where the weights are determined by the relative quality of the prior information and the measurement. Indeed, if we define a signal-to-noise ratio (SNR) of the form

$$\mathrm{SNR} = \frac{\text{mean-square contribution of "signal" portion of } y}{\text{mean-square contribution of noise in } y} = \frac{h^2\sigma_x^2}{\sigma_w^2} \tag{3.221}$$

we have that

$$\hat{x}_{\mathrm{B}}(y) = \left[\frac{1}{1+\mathrm{SNR}}\right] m_x + \left[\frac{\mathrm{SNR}}{1+\mathrm{SNR}}\right] \hat{x}_{\mathrm{MVU}}(y) \tag{3.222}$$

Similarly, we can relate the performance of these estimators according to

$$\frac{1}{\lambda_{\mathrm{B}}} = \frac{1}{\lambda_{\mathrm{MVU}}} + \frac{1}{\sigma_x^2}, \tag{3.223}$$

to which we can attach the interpretation that the information after the measurement equals the sum of the information in the measurement plus the prior information.

Let's consider a couple of other examples of ML estimators that happen to be efficient.

Example 3.14

Suppose that the random variable y is exponentially-distributed with unknown mean $x \geq 0$, i.e.,

$$p_y(y;x) = \frac{1}{x} e^{-y/x}\, u(y). \tag{3.224}$$

Since $p_y(y;x)$ and $\ln p_y(y;x)$ have the same maximum, we obtain the ML estimate as the solution of

$$\begin{aligned} \frac{\partial}{\partial x} \ln p_y(y;x) &= \frac{\partial}{\partial x}\left[-\ln x - \frac{y}{x}\right] \\ &= -\frac{1}{x} + \frac{y}{x^2} = 0. \end{aligned} \tag{3.225}$$

In particular, from (3.225) we get

$$\hat{x}_{\mathrm{ML}}(y) = y. \tag{3.226}$$

Since the mean of y is x, this estimate is unbiased. Furthermore, using the fact that

$$\lambda_{\mathrm{ML}}(x) = \mathrm{var}\, y = x^2$$

we obtain

$$I_y(x) = E\left[\left(\frac{\partial}{\partial x}\ln p_y(y;x)\right)\right] = E\left[\frac{(y-x)^2}{x^4}\right]$$
$$= \frac{1}{x^4}x^2 = \frac{1}{x^2} = \frac{1}{\lambda_{\mathrm{ML}}(x)}. \tag{3.227}$$

Hence, the Cramér-Rao lower bound is tight and the ML estimate is efficient. Note that in this case the variance of the estimator and thus the Cramér-Rao bound are functions of x.

All of our results on nonrandom parameter estimation apply equally well to the case in which $\mathbf{y}$ is discrete-valued, as we illustrate with the following example.

Example 3.15

Suppose we observe a vector

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_M \end{bmatrix}^{\mathrm{T}}$$

of independent Poisson random variables with unknown mean x, i.e., for $i = 1, 2, \ldots, M$ we have

$$p_{y_i}[y_i;x] = \Pr[y_i = y_i;x] = \frac{x^{y_i}e^{-x}}{y_i!}. \tag{3.228}$$

In this case,

$$\ln p_{\mathbf{y}}[\mathbf{y};x] = \sum_{i=1}^{M}\ln p_{y_i}[y_i;x] = \sum_{i=1}^{M}(y_i\ln x - x) - \sum_{i=1}^{M}\ln(y_i!) \tag{3.229}$$

so that $\hat{x}_{\mathrm{ML}}(\mathbf{y})$ is the unique solution to

$$\frac{\partial \ln p_{\mathbf{y}}[\mathbf{y};x]}{\partial x} = \sum_{i=1}^{M}\left(\frac{y_i}{x} - 1\right) = 0. \tag{3.230}$$

In particular, we obtain

$$\hat{x}_{\mathrm{ML}}(\mathbf{y}) = \frac{1}{M}\sum_{i=1}^{M} y_i, \tag{3.231}$$

which again is then unbiased.

Since the variance of a Poisson random variable equals its mean, we have

$$\lambda_{\mathrm{ML}} = \frac{1}{M^2}\sum_{i=1}^{M} x = \frac{x}{M}. \tag{3.232}$$

Using (3.179) with (3.229) we get that the Fisher information is

$$I_y(x) = \frac{1}{x^2}E\left[\sum_{i=1}^{M} y_i\right] = \frac{M}{x}, \tag{3.233}$$

so comparing (3.233) with (3.232) we get that the ML estimate is efficient. Furthermore, since $\lambda_{\mathrm{ML}} \to 0$ as $M \to \infty$, we see that the ML estimate is also consistent.

3.3.5 Estimation of Nonrandom Vectors

In this section, we explore some extensions of the preceding results to the problem of estimating a vector of nonrandom parameters $\mathbf{x}$. To begin, let's briefly discuss the extension of the Cramér-Rao bound to this case. In particular, we have that the covariance matrix $\mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x})$ of any unbiased estimator satisfies the matrix inequality

$$\mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x}) \geq \mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x}), \tag{3.234}$$

where $\mathbf{I}_{\mathbf{y}}(\mathbf{x})$ is now the Fisher Information *matrix*

$$\begin{aligned}\mathbf{I}_{\mathbf{y}}(\mathbf{x}) &= E\left[\left[\frac{\partial \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})}{\partial \mathbf{x}}\right]^{\mathrm{T}}\left[\frac{\partial \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})}{\partial \mathbf{x}}\right]\right] \\ &= -E\left[\frac{\partial^2 \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})}{\partial \mathbf{x}^2}\right]\end{aligned} \tag{3.235}$$

Note that from the diagonal elements of (3.234) we obtain a set of scalar Cramér-Rao bounds on the variances of individual components of $\mathbf{x}$. Also an unbiased efficient estimate $\hat{\mathbf{x}}(\mathbf{y})$ exists if and only if

$$\hat{\mathbf{x}}(\mathbf{y}) = \mathbf{x} + \mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\left[\frac{\partial \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})}{\partial \mathbf{x}}\right]^{\mathrm{T}} \tag{3.236}$$

is a valid estimator, i.e., if and only if the right-hand side of (3.236) does not depend on $\mathbf{x}$. Also, if an efficient unbiased estimate exists, it is the ML estimate.

To derive the matrix Cramér-Rao bound (3.234), we follow an approach analogous to that used to obtain (3.167), but which requires some additional steps. In particular, we begin by recalling that for unbiased estimators the error

$$\mathbf{e}(\mathbf{y}) = \hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x} \tag{3.237}$$

has zero mean, i.e.,

$$E\left[\mathbf{e}(\mathbf{y})\right] = 0 \tag{3.238}$$

and covariance

$$E\left[\mathbf{e}(\mathbf{y})\mathbf{e}^{\mathrm{T}}(\mathbf{y})\right] = \mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x}). \tag{3.239}$$

Next we define

$$\mathbf{f}^{\mathrm{T}}(\mathbf{y}) = \frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) \tag{3.240}$$

and note that using the identity

$$\frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) = \frac{1}{p_{\mathbf{y}}(\mathbf{y};\mathbf{x})}\frac{\partial}{\partial \mathbf{x}} p_{\mathbf{y}}(\mathbf{y};\mathbf{x}), \tag{3.241}$$

we get that $\mathbf{f}(\mathbf{y})$ has zero mean:

$$\begin{aligned} E\left[\mathbf{f}^{\mathrm{T}}(\mathbf{y})\right] &= E\left[\frac{1}{p_{\mathbf{y}}(\mathbf{y};\mathbf{x})}\frac{\partial}{\partial \mathbf{x}}p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\right] \\ &= \int_{-\infty}^{+\infty}\frac{\partial}{\partial \mathbf{x}}p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\,d\mathbf{y} \\ &= \frac{\partial}{\partial \mathbf{x}}\int_{-\infty}^{+\infty}p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\,d\mathbf{y} = \frac{\partial}{\partial \mathbf{x}}1 = \mathbf{0}, \end{aligned} \tag{3.242}$$

and, in turn, covariance

$$\mathbf{\Lambda}_{\mathbf{f}}(\mathbf{x}) = E\left[\mathbf{f}(\mathbf{y})\mathbf{f}^{\mathrm{T}}(\mathbf{y})\right] = \mathbf{I}_{\mathbf{y}}(\mathbf{x}). \tag{3.243}$$

Finally, again using the identity (3.241), the covariance between $\mathbf{e}(\mathbf{y})$ and $\mathbf{f}(\mathbf{y})$ is given by

$$\begin{aligned} \mathrm{cov}\left(\mathbf{e}(\mathbf{y}),\mathbf{f}(\mathbf{y})\right) &= E\left[\mathbf{e}(\mathbf{y})\mathbf{f}^{\mathrm{T}}(\mathbf{y})\right] \\ &= \int_{-\infty}^{+\infty}(\hat{\mathbf{x}}(\mathbf{y})-\mathbf{x})\frac{\partial}{\partial \mathbf{x}}p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\,d\mathbf{y} \\ &= \left[\frac{\partial}{\partial \mathbf{x}}\int_{-\infty}^{+\infty}\hat{\mathbf{x}}(\mathbf{y})\,p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\,d\mathbf{y}\right] - \left[\mathbf{x}\frac{\partial}{\partial \mathbf{x}}\int_{-\infty}^{+\infty}p_{y}(\mathbf{y};\mathbf{x})\,d\mathbf{y}\right] \\ &= \mathbf{I}-\mathbf{0}=\mathbf{I}. \end{aligned} \tag{3.244}$$

Next, for an arbitrary choice of $\mathbf{c}$ we let

$$\tilde{e}(\mathbf{y}) = \mathbf{c}^{\mathrm{T}}\mathbf{e}(\mathbf{y}) \tag{3.245}$$

and

$$\tilde{f}(\mathbf{y}) = \mathbf{c}^{\mathrm{T}}\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\mathbf{f}(\mathbf{y}) = \mathbf{f}^{\mathrm{T}}(\mathbf{y})\,\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\,\mathbf{c}. \tag{3.246}$$

Then both $\tilde{e}(\mathbf{y})$ and $\tilde{f}(\mathbf{y})$ have zero-mean and, using (3.239), (3.243) and (3.244), we have

$$\mathrm{var}\,\tilde{e}(\mathbf{y}) = \mathbf{c}^{\mathrm{T}}\mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x})\mathbf{c} \tag{3.247a}$$

$$\mathrm{var}\,\tilde{f}(\mathbf{y}) = \mathbf{c}^{\mathrm{T}}\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\mathbf{c} \tag{3.247b}$$

$$\mathrm{cov}\left(\tilde{e}(\mathbf{y}),\tilde{f}(\mathbf{y})\right) = \mathbf{c}^{\mathrm{T}}\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\mathbf{c}. \tag{3.247c}$$

Now since the covariance between $\tilde{e}(\mathbf{y})$ and $\tilde{f}(\mathbf{y})$ satisfies the bound

$$\left[\mathrm{cov}\left(\tilde{e}(\mathbf{y}),\tilde{f}(\mathbf{y})\right)\right]^2 \le \mathrm{var}\,\tilde{e}(\mathbf{y})\,\mathrm{var}\,\tilde{f}(\mathbf{y}) \tag{3.248}$$

we can substitute (3.247) into (3.248) to obtain, after some simple manipulation,

$$\mathbf{c}^{\mathrm{T}}\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\mathbf{c}\left[\mathbf{c}^{\mathrm{T}}\mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x})\mathbf{c} - \mathbf{c}^{\mathrm{T}}\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\mathbf{c}\right] \ge 0. \tag{3.249}$$

However, since $\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})$ is positive semidefinite, the term to the left of the brackets in (3.249) is non-negative. Hence, the term in brackets must be non-negative. But then since $\mathbf{c}$ is arbitrary this means $\mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x}) - \mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})$ must be positive semidefinite, which establishes (3.234) as desired.

Finally, equality is satisfied in (3.248) (and therefore (3.249)) if and only if $\tilde{e}(\mathbf{y}) = k(\mathbf{x})\tilde{f}(\mathbf{y})$ for some function $k(\mathbf{x})$ that doesn't depend on $\mathbf{y}$, i.e., if and only if,

$$\mathbf{c}^{\mathrm{T}}\mathbf{e}(\mathbf{y}) = \mathbf{c}^{\mathrm{T}}k(\mathbf{x})\,\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\,\mathbf{f}(\mathbf{y}). \tag{3.250}$$

However, since (3.250) holds for any choice of $\mathbf{c}$ we must have

$$\mathbf{e}(\mathbf{y}) = k(\mathbf{x})\,\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\,\mathbf{f}(\mathbf{y}). \tag{3.251}$$

Again $k(\mathbf{x})$ can't be arbitrary. In particular, when the bound (3.234) is satisfied with equality we have

$$E\left[\mathbf{e}(\mathbf{y})\,\mathbf{e}^{\mathrm{T}}(\mathbf{y})\right] = \mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x}) = \mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x}). \tag{3.252}$$

However, using (3.251), (3.243), and (3.244) we have

$$E\left[\mathbf{e}(\mathbf{y})\,\mathbf{e}^{\mathrm{T}}(\mathbf{y})\right] = E\left[\mathbf{e}(\mathbf{y})\,\mathbf{f}^{\mathrm{T}}(\mathbf{y})\,\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\,k(\mathbf{x})\right] = \mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\,k(\mathbf{x}). \tag{3.253}$$

Comparing (3.252) with (3.253) we obtain

$$k(\mathbf{x}) = 1, \tag{3.254}$$

which when substituted into (3.251) yields the following: $\hat{\mathbf{x}}(\mathbf{y})$ is an efficient estimator, i.e., satisfies the bound (3.234) with equality if an only if it can be expressed in the form (3.236) where the right-hand side must be independent of $\mathbf{x}$ for the estimator to be valid.

We can also readily verify that the second form of the Fisher information in (3.235) is equivalent to the first. Analogous to our approach in the scalar case, we begin by observing

$$\int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\,d\mathbf{y} = 1. \tag{3.255}$$

Computing the Jacobian of (3.255) with respect to $\mathbf{x}$ and using the identity (3.241) yields

$$\int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\left[\frac{\partial}{\partial\mathbf{x}}\ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\right]^{\mathrm{T}} d\mathbf{y} = \mathbf{0}. \tag{3.256}$$

Finally, computing the Hessian of (3.255) with respect to $\mathbf{x}$ and again using (3.241) we obtain

$$\begin{aligned}\int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\left[\frac{\partial^2}{\partial\mathbf{x}^2}\ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\right] d\mathbf{y} \\ + \int_{-\infty}^{+\infty} p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\left[\frac{\partial}{\partial\mathbf{x}}\ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\right]^{\mathrm{T}}\left[\frac{\partial}{\partial\mathbf{x}}\ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\right] d\mathbf{y} = \mathbf{0}.\end{aligned} \tag{3.257}$$

which verifies that the two expressions in (3.235) are consistent.

It is also straightforward to verify that when an efficient estimator $\hat{\mathbf{x}}_{\text{eff}}(\mathbf{y})$ exists, it must be the ML estimator. Again we follow an approach analogous to the scalar case. Since (3.236) must not be a function of $\mathbf{x}$ when an efficient estimator exists, we can then freely choose any value of $\mathbf{x}$ in this expression without effect. If we choose the value $\mathbf{x} = \hat{\mathbf{x}}_{\text{eff}}(\mathbf{y})$, we obtain

$$\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x}) \left[\frac{\partial}{\partial \mathbf{x}} \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x})\right]^{\mathrm{T}}\Bigg|_{\mathbf{x}=\hat{\mathbf{x}}_{\text{eff}}(\mathbf{y})} = \mathbf{0}. \tag{3.258}$$

But since $\mathbf{I}_{\mathbf{y}}(\mathbf{x})$ is nonsingular except in the trivial case, we have that the term in brackets in (3.258) must be zero, i.e.,

$$\hat{\mathbf{x}}_{\text{eff}}(\mathbf{y}) = \hat{\mathbf{x}}_{\text{ML}}(\mathbf{y}) = \arg\max_{\mathbf{x}} p_{\mathbf{y}}(\mathbf{y};\mathbf{x}). \tag{3.259}$$

Again we stress that one should not infer from these results that the ML estimator is always efficient. When no efficient estimator exists, the ML estimate can still be computed; however it need not have any special properties. As in the scalar case, though, even when an efficient estimator doesn't exist, the ML estimator often has good asymptotic properties in several problems. One class of problems in which the ML estimator is always efficient are the linear/Gaussian problems. We conclude this section with the canonical example.

Example 3.16

Suppose we that our observed data $\mathbf{y}$ depends on our parameter vector $\mathbf{x}$ through the linear model

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}, \tag{3.260}$$

where $\mathbf{w} \sim N(\mathbf{0}, \mathbf{\Lambda}_{\mathbf{w}})$. In this case

$$p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) = N(\mathbf{y};\mathbf{H}\mathbf{x},\mathbf{\Lambda}_{\mathbf{w}}) \propto \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\right] \tag{3.261}$$

so that maximizing $p_{\mathbf{y}}(\mathbf{y};\mathbf{x})$ with respect to $\mathbf{x}$ is equivalent to *minimizing*

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}) \tag{3.262}$$

with respect to $\mathbf{x}$. Since (3.262) is a non-negative function, its unique stationary point, which we obtain by setting the Jacobian of (3.262) to zero, is its global minimum and thus gives the ML estimate

$$\hat{\mathbf{x}}_{\text{ML}}(\mathbf{y}) = (\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{y} \tag{3.263}$$

This estimate is unbiased, since

$$E\left[\hat{\mathbf{x}}_{\text{ML}}(\mathbf{y})\right] = (\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}(\mathbf{H}\mathbf{x} + E\left[\mathbf{w}\right]) = \mathbf{x} \tag{3.264}$$

and its error covariance is

$$\begin{aligned}\mathbf{\Lambda}_{\text{ML}} &= E\left[\left[(\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{w}\right]\left[(\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{w}\right]^{\mathrm{T}}\right] \\ &= (\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{\Lambda}_{\mathbf{w}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H}(\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1} \\ &= (\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H})^{-1}.\end{aligned} \tag{3.265}$$

Note that for this estimate to make sense, $\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H}$ must be invertible, and this in turn requires that the dimension of $\mathbf{y}$ (or, more precisely, the rank of $\mathbf{\Lambda}_{\mathbf{w}}$) be at least as large as the dimension of $\mathbf{x}$. Phrased differently, the number of degrees of freedom in the measurements must equal or exceed the number of parameters to be estimated.

The Fisher information matrix for this problem is obtained using the second form of (3.235) and yields

$$I_{\mathbf{y}}(\mathbf{x}) = -\frac{d^2}{d\mathbf{x}^2}J(\mathbf{x}) = \mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H} \tag{3.266}$$

which by comparison to (3.265) allows us to conclude that the ML estimate is, in fact, efficient. Note as well that the estimator covariance (and thus the Fisher matrix) is independent of $\mathbf{x}$ in this example.

As in the scalar case, it is again interesting to compare the ML estimator in this example to the LLS estimator for the closely related problem developed in Example 3.7. In both examples, the measurement models (3.260) and (3.104) are identical, but in this example $\mathbf{x}$ is a nonrandom parameter vector while in Example 3.7 we have a random parameter $\mathbf{x}$ with zero-mean and covariance $\mathbf{\Lambda}_{\mathbf{x}}$.

If we added the Gaussian assumptions to Example 3.7, we can again conclude that the resulting LLS estimator (3.102) is also the BLS estimator and the MAP estimator for the problem. For this reason, we'll simply use $\hat{\mathbf{x}}_{\mathrm{B}}(y)$ to denote this estimator and $\mathbf{\Lambda}_{\mathrm{B}}$ to denote its error covariance for the remainder of this example. Furthermore, since our ML estimate is efficient, it is the MVU estimator for the nonrandom parameter estimation problem,[11] so we'll use $\hat{\mathbf{x}}_{\mathrm{MVU}}(\mathbf{y})$ to denote this estimator and $\mathbf{\Lambda}_{\mathrm{MVU}}$ to denote its covariance for the remainder of this example.

In this case we have, using the alternative matrix forms developed in Appendix 3.A,

$$\hat{\mathbf{x}}_{\mathrm{B}}(\mathbf{y}) = \mathbf{m}_{\mathbf{x}} + \mathbf{\Lambda}_{\mathrm{B}}\mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}(\mathbf{y} - \mathbf{H}\mathbf{m}_{\mathbf{x}}) \tag{3.267}$$

$$\mathbf{\Lambda}_{\mathrm{B}}^{-1} = \mathbf{\Lambda}_{\mathbf{x}}^{-1} + \mathbf{\Lambda}_{\mathrm{MVU}}^{-1} \tag{3.268}$$

$$\tag{3.269}$$

From these expressions we see that as $\mathbf{\Lambda}_{\mathbf{x}} \to \infty$ (again in the sense of its trace),

$$\mathbf{\Lambda}_{\mathrm{B}} \to \mathbf{\Lambda}_{\mathrm{MVU}} \tag{3.270}$$

and, in turn,

$$\hat{\mathbf{x}}_{\mathrm{B}}(\mathbf{y}) \to \hat{\mathbf{x}}_{\mathrm{MVU}}(\mathbf{y}). \tag{3.271}$$

3.4 NONLINEAR ESTIMATION

Quite frequently in practice our observations $\mathbf{y}$ correspond to some noisy nonlinear function of the parameters $\mathbf{x}$. Let us explore this general problem in the context of the nonrandom parameter estimation theory of Section 3.3. There are lots of important examples of problems that fall into this category, and we'll explore

[11] This result is referred to as the Gauss-Markov theorem.

in detail one involving estimation of the parameters of a sinusoid in Section 3.4.1. Before we do that, however, let us begin with some preliminary observations.

Example 3.17

Consider the following nonlinear measurement

$$y = h(x) + w, \tag{3.272}$$

where $w \sim N(0, \sigma^2)$. In this case,

$$p_y(y; x) = N(y; h(x), \sigma^2), \tag{3.273}$$

so that

$$\frac{\partial \ln p_y(y; x)}{\partial x} = \left(\frac{y - h(x)}{\sigma^2} \right) \frac{dh(x)}{dx}. \tag{3.274}$$

Let's compute the Cramér-Rao bound on the performance of arbitrary unbiased estimates $\hat{x}(\cdot)$ for x. Using (3.274) we obtain that

$$I_y(x) = E\left[\left[\left(\frac{y - h(x)}{\sigma^2} \right) \frac{dh(x)}{dx} \right]^2 \right] = \left[\frac{dh(x)}{dx} \right]^2 E\left[\left(\frac{w}{\sigma^2} \right)^2 \right] = \frac{1}{\sigma^2} \left[\frac{dh(x)}{dx} \right]^2, \tag{3.275}$$

so that

$$\lambda_{\hat{x}}(x) \geq \frac{\sigma^2}{(dh(x)/dx)^2} \tag{3.276}$$

for any unbiased estimate. Now an efficient estimate exists if and only if (3.188) is a valid estimator, i.e., if and only if

$$x + \frac{1}{I_y(x)} \frac{\partial}{\partial x} \ln p_y(y; x) = \left(x - \frac{h(x)}{dh(x)/dx} \right) + \frac{y}{dh(x)/dx} \tag{3.277}$$

is a function only of y. However, since the right-hand term in (3.277) is the only one that depends on y and since y can be arbitrary, we can conclude that no efficient estimate can exist unless $dh(x)/dx$ does not depend on x. However, this will only be the case when $h(\cdot)$ is a linear (affine) function. Hence, efficient estimates fail to exist in the strictly nonlinear case.

Consider, for example, $h(x) = x^3$. In this case, (3.277) becomes

$$x + \frac{y - x^3}{3x^2} = \frac{2}{3}x + \frac{1}{3}\frac{y}{x^2}, \tag{3.278}$$

from which we see that there is no efficient estimate.

Since an efficient estimate generally doesn't exist, the ML estimate, which we'll now compute, needn't have any special properties in the nonlinear case. When $h(\cdot)$ is invertible, as we'll assume in this example, we get immediately from (3.274) that

$$\hat{x}_{\mathrm{ML}}(y) = h^{-1}(y), \tag{3.279}$$

where $h^{-1}(\cdot)$ is the inverse function of $h(\cdot)$, i.e., $h^{-1}(h(x)) = x$. Calculating the bias $b_{\mathrm{ML}}(x)$ and variance $\lambda_{\mathrm{ML}}(x)$ of this estimate is difficult in general, though in general it will be biased. And when biased, this means we cannot even conclude that its variance $\lambda_{\mathrm{ML}}(x)$ satisfies (3.276) for even one value of x.

While ML estimates in nonlinear problems needn't have good variance characteristics, they do have some attractive features. For example, suppose that a parameter θ is related to x via

$$\theta = g(x),$$

where $g(\cdot)$ is a nonlinear but invertible transformation. Then it is a straightforward exercise to show that the ML estimates are also related by

$$\hat{\theta}_{\mathrm{ML}}(\mathbf{y}) = g(\hat{x}_{\mathrm{ML}}(\mathbf{y})), \tag{3.280}$$

i.e., ML estimates commute under nonlinear transformations. This is an extremely convenient property, and one that is not shared by most of the other estimators we've explored in this chapter of the notes. For example, Bayesian estimators almost never commute with nonlinear transformations, i.e., if x is a random parameter, then

$$\hat{\theta}_{\mathrm{B}}(\mathbf{y}) \neq g(\hat{x}_{\mathrm{B}}(\mathbf{y})),$$

for almost any nontrivial cost criterion. We also remark that although (3.280) doesn't apply when $g(\cdot)$ is not invertible, straightforward extensions of this result can be developed to handle the non-invertible case.

We remark, however, that even if $\hat{x}_{\mathrm{ML}}(\mathbf{y})$ has nice properties, these properties are generally not preserved under the transformation (3.280). For example, since typically

$$E\left[\hat{\theta}_{\mathrm{ML}}(\mathbf{y})\right] = E\left[g(\hat{x}_{\mathrm{ML}}(\mathbf{y}))\right] \neq g\left(E\left[\hat{x}_{\mathrm{ML}}(\mathbf{y})\right]\right)$$

we wouldn't expect $\hat{\theta}_{\mathrm{ML}}(\mathbf{y})$ to be unbiased even if $\hat{x}_{\mathrm{ML}}(\mathbf{y})$ were.

Nonlinear estimation problems are distinguished in other important ways from inherently linear estimation problems. In the remainder of the chapter, we explore such distinguishing characteristics. To illustrate the main ideas before we develop them in detail, we first explore as a case study a particular nonlinear estimation problem that arises in an extraordinarily wide range of practical applications.

3.4.1 Sinusoid Estimation

In this section we explore a basic problem involving sinusoid estimation. In particular, given noisy observations of the form

$$y[n] = A\cos(\omega_0 n + \Theta) + w[n], \qquad n = 0, 1, \ldots, N-1 \tag{3.281}$$

we wish to estimate one or more of the nonrandom parameters A, ω_0, or Θ, where $A > 0$ and $0 \leq \omega_0 < \pi$. In (3.281), we'll assume that the noise samples $w[n]$ are independent, identically distributed $N(0, \sigma^2)$ random variables; this will facilitate our analysis, and is often a good model in applications.

Among the enormous number of applications in which this model arises are analog communications, Doppler radar, noise cancellation, interference suppression, radio astronomy, and sonar direction-finding.

Analog Communication

Amplitude Modulation (AM) In AM systems, the frequency ω_0 is known, but the amplitude varies with time and carries the information. In such systems we can generally approximate the amplitude as constant over the block of N samples, and consider the problem of recovering the amplitude for each block as an estimation problem. In such problems, the phase Θ may be known, but more typically is an unknown parameter that, while not of interest, must be simultaneously estimated. Note that if ω_0 and Θ are both known, then the resulting estimation problem is inherently linear, and will exhibit certain associated characteristic behavior.

Phase Modulation (PM) In PM systems, it is the phase Θ that carries the information while the frequency remains essentially fixed (and known). In such problems, the amplitude is generally distorted by the channel and while similarly not of interest, must be jointly estimated as well.

Frequency Modulation (FM) In FM systems, the frequency ω_0 carries the information and varies with time accordingly. In such systems, we then wish to estimate ω_0, which is modeled as essentially constant over the block of length N. Typically, the communication channel distorts both the amplitude and phase, so these quantities must be simultaneously (i.e., jointly) estimated.

Noise and Interference Cancellation A wide variety of noise and interference encountered in practice is inherently sinusoidal in nature. Examples include 60 Hz (line-frequency) interference in systems due to AC power supplies, noise from rotating machinery, propeller noise in aircraft and on ships, and narrowband jamming—hostile or inadvertent—in wireless communication systems. In such cases, the sinusoidal term in (3.281) may be the unwanted interference and $w[n]$ may represent the (broadband) signal of interest. For these scenarios, an effective interference suppression strategy involves estimating the parameters of the sinusoidal interferer, then subtracting it out from the observations to recover the signal of interest.

Doppler Radar In radar systems, (3.281) can be used to model the radar return, where the deviation of ω_0 from some nominal value is a Doppler shift used to measure the velocity of the target.

Radio Astronomy In radio astronomy applications, one is often interested in detecting and locating spectral lines corresponding to emissions from distant sources of radiation, and even most experimental apparatus designed for searching for extraterrestrials uses (3.281) as the basic model for the signal being sent by ET!

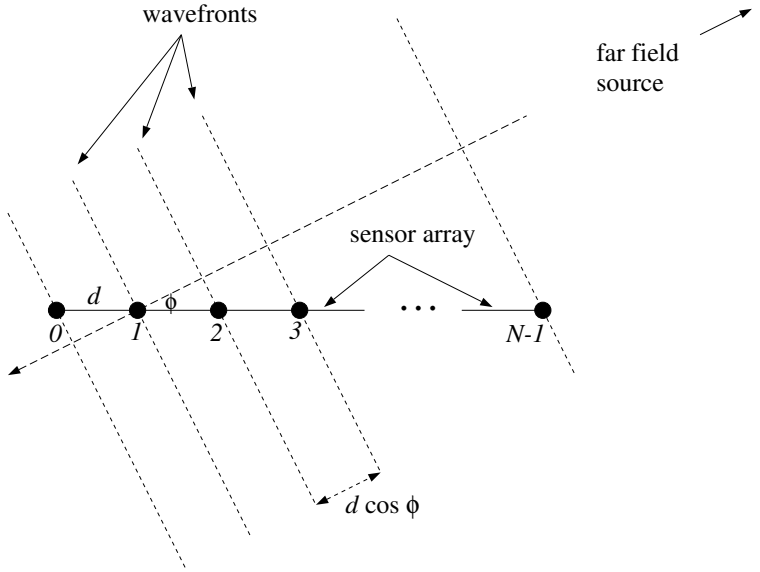

Figure 3.3. Estimating the direction of arrival of a far field source using an N-element linear array of sensors.

Sonar Direction-Finding Sonar systems are often used to locate the direction from which an acoustic source is propagating. To illustrate this, suppose the source is emitting a pure tone (sinusoid) of the form

$$x(t) = A\cos(\Omega_0 t), \tag{3.282}$$

and that this signal is being picked up at a linear, horizontal array of N sensors (hydrophones), as depicted in Fig. 3.3. Let d denote the distance between sensors, and let us assume that the source is sufficiently distant to allow a so-called "far-field" approximation: the signal arrives at the array as a plane-wave, with the wavefronts consisting of straight lines (rather than circles) as Fig. 3.3 reflects. Let ϕ denote the angle at which the plane wave impinges on the array.

In this system, the propagation time to the nth sensor is

$$t_n = t_0 - n\frac{d}{c}\cos\phi, \qquad n = 0, 1, \ldots, N-1 \tag{3.283}$$

where c is the propagation speed (i.e., phase velocity), so that the signal observed at the nth sensor is, for some Θ',

$$y_n(t) = A\cos(\Omega_0(t - t_n) + \Theta') + w_n(t). \tag{3.284}$$

If we take a snapshot of all the sensors at a particular time instant $t = t_*$, we obtain the vector of samples

$$\begin{aligned} y[n] &= y_n(t_*) \\ &= A\cos\left[\left(\Omega_0 \frac{d}{c}\cos\phi\right) n + \Theta\right] + w_n(t_*) \\ &= A\cos(\omega_0 n + \Theta) + w[n], \end{aligned} \tag{3.285}$$

where $\Theta = \Theta' + \Omega_0(t_* - t_0)$, $\omega_0 = (\Omega_0 d/c)\cos\phi$, and $w[n] = w_n(t_*)$. Hence, by estimating the spatial frequency ω_0, we can indirectly obtain an estimate of the direction-of-arrival ϕ.

Performance Issues and Cramér-Rao Bounds

Let us begin by considering the most general problem, wherein the parameters A, ω_0, and Θ in the model (3.281) are all unknown, and explore the form of the associated Cramér-Rao bounds. These bounds will give us some insight into how we can expect estimator performance to vary with the signal-to-noise ratio

$$\gamma = \frac{1}{2}A^2/\sigma^2, \tag{3.286}$$

the data length N, and the actual values of the parameters A, ω_0, and Θ.

When we collect the unknown parameters into a vector $\mathbf{x}$, i.e.,

$$\mathbf{x} = \begin{bmatrix} A \\ \omega_0 \\ \Theta \end{bmatrix}, \tag{3.287}$$

and do the same for the data, i.e.,

$$\mathbf{y} = \begin{bmatrix} y[0] \\ y[1] \\ \vdots \\ y[N-1] \end{bmatrix}, \tag{3.288}$$

the elements of the Fisher information matrix then take the form

$$[\mathbf{I}_{\mathbf{y}}(\mathbf{x})]_{ij} = -E\left[\frac{\partial^2}{\partial x_i \partial x_j}\ell(\mathbf{y};\mathbf{x})\right], \tag{3.289}$$

where

$$\ell(\mathbf{y};\mathbf{x}) = \ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\big[y[n] - A\cos(\omega_0 n + \Theta)\big]^2. \tag{3.290}$$

The calculation of the quantities (3.289) is straightforward but somewhat lengthy; the details are provided in Appendix 3.B. For arbitrary N, the results

are also somewhat cumbersome. However, as the Appendix shows, in the large N regime, corresponding to at least moderately sized data sets, the Fisher information can be expressed using order notation[12] in the following comparatively simple form

$$\mathbf{I}_{\mathbf{y}}(\mathbf{x}) = \frac{1}{\sigma^2}\begin{bmatrix} N/2 + o(N) & o(N^2) & o(N) \\ o(N^2) & A^2T_N/2 + o(N^3) & A^2S_N/2 + o(N^2) \\ o(N) & A^2S_N/2 + o(N^2) & A^2N/2 + o(N) \end{bmatrix}, \tag{3.291}$$

where

$$S_N = \sum_{n=0}^{N-1} n = \frac{1}{2}N(N-1) \tag{3.292}$$

$$T_N = \sum_{n=0}^{N-1} n^2 = \frac{1}{6}N(N-1)(2N-1). \tag{3.293}$$

Computing the inverse of (3.291) we then obtain our Cramér-Rao bound on unbiased estimates $\hat{A}(\mathbf{y})$, $\hat{\omega}_0(\mathbf{y})$, and $\hat{\Theta}(\mathbf{y})$ of the parameters in the large N regime. In particular, we obtain, using (3.286),

$$\operatorname{var}\left(\frac{\hat{A}(\mathbf{y})}{A}\right) \geq \frac{1}{A^2}\left[\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\right]_{11} \approx \frac{2\sigma^2}{A^2N} = \frac{1}{\gamma N} \sim O\left(\frac{1}{\gamma N}\right) \tag{3.294a}$$

$$\operatorname{var}\hat{\omega}_0(\mathbf{y}) \geq \left[\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\right]_{22} \approx \frac{12}{\gamma N(N^2-1)} \sim O\left(\frac{1}{\gamma N^3}\right) \tag{3.294b}$$

$$\operatorname{var}\hat{\Theta}(\mathbf{y}) \geq \left[\mathbf{I}_{\mathbf{y}}^{-1}(\mathbf{x})\right]_{33} \approx \frac{2(2N-1)}{\gamma N(N+1)} \sim O\left(\frac{1}{\gamma N}\right) \tag{3.294c}$$

It should be emphasized that the size of N necessary for the approximations in (3.294) to be valid depends on the true value of ω_0—in general, the closer ω_0 is to 0 or π, the larger N must be. More specifically, as the development in Appendix 3.B reveals, the approximation is valid provided

$$\frac{\pi}{N} \ll \omega_0 \ll \pi\left(1 - \frac{1}{N}\right). \tag{3.295}$$

---

[12]For functions $f(\cdot)$ and $g(\cdot)$ we use the notation $f(N) \sim o(g(N))$ to indicate that $f(N)$ grows strictly slower than $g(N)$, i.e.,

$$\lim_{N\to\infty} \frac{f(N)}{g(N)} = 0,$$

As related order notation, we write $f(N) \sim O(g(N))$ if $f(N)$ grows no faster than $g(N)$, i.e.,

$$\lim_{N\to\infty} \frac{f(N)}{g(N)} < \infty.$$

As examples, $f(N) \sim o(N)$ means that $f(N)$ grows slower than linearly with N, while $f(N) \sim O(N)$ means that $f(N)$ grows no faster than linearly with N.

The asymptotic Cramér-Rao lower bounds (3.294) reveal some key characteristics of the estimation problem. As we would expect, all the bounds decrease inversely with the SNR γ and the data length N. However, data length has the most profound impact on the bound for the frequency estimate. This suggests that it may be possible to estimate this parameter with very high accuracy at moderate data lengths. Whether this is possible depends, of course, on whether estimators can be developed whose performance comes close to the bound. We explore this issue, among others, in the context of developing asymptotic ML estimates for the parameters in the next section.

As a final remark, it is worth emphasizing that the Fisher information (3.291) contains all the information necessary to asymptotically bound the performance of related sinusoid estimation problems. In particular, when some of the parameters A, ω_0, Θ are known, the associated Cramér-Rao bounds for the remaining parameters are obtained by inverting a submatrix of (3.291) formed by discarding the rows and columns corresponding to the known parameters. Using this approach, it can be readily verified that, for example, when the frequency ω_0 is known, the asymptotic Cramér-Rao bounds for $\hat{A}$ and $\hat{\Theta}$ are still $O(1/\gamma N)$ as in (3.294a) and (3.294c), respectively. Likewise, when both the frequency ω_0 and phase Θ are known, the Cramér-Rao bound on $\hat{A}$ remains $O(1/\gamma N)$ as in (3.294a).

Maximum Likelihood Estimates

In this section, we obtain ML estimates for the sinusoid estimation problem, develop their properties, and relate the performance of resulting estimators to the corresponding Cramér-Rao bounds.

To begin, the ML parameter estimates

$$\hat{x}(\mathbf{y}) = \arg\max_{\mathbf{x}} \ell(\mathbf{y}; \mathbf{x})$$

are the solutions to the nonlinear least-squares problem

$$(\hat{A}, \hat{\omega}_0, \hat{\Theta}) = \arg\min_{(A,\omega_0,\Theta)} J(A, \omega_0, \Theta) \tag{3.296a}$$

where, via (3.290),

$$J(A, \omega_0, \Theta) = \sum_{n=0}^{N-1} \big[y[n] - A\cos(\omega_0 n + \Theta)\big]^2. \tag{3.296b}$$

In principle, the optimization (3.296) can be performed numerically without exploiting any of the special structure in the problem. However, in the large N regime, more direct expressions are possible, which have intuitively satisfying interpretations and lead to efficient implementations. These expressions are in terms of the normalized, length-N discrete-time Fourier transform of the data segment

comprising $\mathbf{y}$, i.e.,[13]

$$Y_N(e^{j\omega}) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} y[n] e^{-j\omega n}. \tag{3.299}$$

The magnitude-squared of (3.299), i.e., $|Y_N(e^{j\omega})|^2$, is referred to as the *periodogram* of the data.

A natural periodogram-based estimator for the sinusoid estimation problem is defined as follows.

Definition 3.1 (Periodogram-Based Estimator) *The periodogram-based frequency estimate $\hat{\omega}_0$ is the location of the peak of the periodogram, i.e.,*

$$\hat{\omega}_0 = \arg\max_{\omega} \left| Y_N(e^{j\omega}) \right|^2. \tag{3.300}$$

In turn, the magnitude of this peak yields the associated amplitude estimate, i.e.,

$$\hat{A}^2 = \frac{4}{N} \left| Y_N(e^{j\hat{\omega}_0}) \right|^2, \tag{3.301}$$

and the associated phase estimate corresponds to the (negated) phase of $Y_N(e^{j\omega})$ at the location of the peak, i.e.,

$$\hat{\Theta} = -\measuredangle Y_N(e^{j\hat{\omega}_0}) = -\tan^{-1}\left(\frac{\mathrm{Im}\left\{ Y_N(e^{j\hat{\omega}_0}) \right\}}{\mathrm{Re}\left\{ Y_N(e^{j\hat{\omega}_0}) \right\}} \right) \tag{3.302}$$

Note that the estimator in Definition 3.1 is both intuitively appealing and highly practical. Indeed, to identify a sinusoid it is rather natural to compute the Fourier transform of the noisy data segment and locate the amplitude, frequency, and phase of its peak. Moreover, these estimators can be implemented very efficiently in practice. In particular, the frequency estimate can be computed by taking a sufficiently large discrete Fourier transform (DFT) of the data—i.e., with sufficient zero-padding of the data—and searching for index of the largest DFT coefficient. The computation of these DFT's can be conveniently carried out using an efficient fast Fourier transform (FFT) algorithm, which has $O(N \log N)$ complexity.

The estimator of Definition 3.1 also has some important optimality properties, and in fact is closely related to the ML estimator for the problem. In particular,

---

[13] It is often convenient to view (3.299) as the Fourier transform of a windowed version of the sequence $y[n]$, i.e.,

$$Y_N(e^{j\omega}) = \mathcal{F}\left\{ g_N[n]\, y[n] \right\} \tag{3.297}$$

where $g[n]$ is the unit-energy window

$$g_N[n] = \begin{cases} 1/\sqrt{N} & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases}. \tag{3.298}$$

the periodogram-based and ML estimators are effectively equivalent when N is large in the sense of (3.295); from this perspective, we can view the periodogram-based estimator as the asymptotic ML estimator. A relatively straightforward derivation of this result is developed in Appendix 3.C.

In order to analyze the performance of the periodogram-based estimator, it is useful to decompose the periodogram into signal and noise components. In particular, we write $\mathsf{Y}_N(e^{j\omega})$ in the form

$$\mathsf{Y}_N(e^{j\omega}) = X_N(e^{j\omega}) + \mathsf{W}_N(e^{j\omega}) \tag{3.303}$$

where

$$\mathsf{W}_N(e^{j\omega}) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} \mathsf{w}[n]\, e^{-j\omega n} \tag{3.304}$$

and

$$X_N(e^{j\omega}) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n]\, e^{-j\omega n}, \tag{3.305}$$

with

$$x[n] = A\cos(\omega_0 n + \Theta). \tag{3.306}$$

The noise component $\mathsf{W}_N(e^{j\omega})$ has mean

$$E\left[\mathsf{W}_N(e^{j\omega})\right] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} E\left[\mathsf{w}[n]\right]\, e^{-j\omega n} = 0 \tag{3.307}$$

and variance

$$E\left[\left|\mathsf{W}_N(e^{j\omega})\right|^2\right] = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} E\left[\mathsf{w}[n]\, \mathsf{w}[m]\right] e^{j\omega(n-m)} = \sigma^2. \tag{3.308}$$

With this decomposition, the "signal" and "noise" components of the periodogram are naturally defined as, respectively,

$$\left|E\left[\mathsf{Y}_N(e^{j\omega})\right]\right|^2 = \left|X_N(e^{j\omega})\right|^2 \tag{3.309}$$

and

$$\text{var}\, \mathsf{Y}_N(e^{j\omega}) = E\left[\left|\mathsf{W}_N(e^{j\omega})\right|^2\right] = \sigma^2. \tag{3.310}$$

These two components are depicted in Fig. 3.4. In turn, we define the SNR at a particular frequency as

$$\gamma(\omega) = \frac{\left|E\left[\mathsf{Y}_N(e^{j\omega})\right]\right|^2}{\text{var}\, \mathsf{Y}_N(e^{j\omega})} = \frac{\left|X_N(e^{j\omega})\right|^2}{\sigma^2}. \tag{3.311}$$

In the large N regime (3.295) we have

$$\left|X_N(e^{j\omega})\right|^2 \approx \frac{A^2}{4N} \left\{ \left[\frac{\sin(\omega-\omega_0)N/2}{\sin(\omega-\omega_0)/2}\right]^2 + \left[\frac{\sin(\omega+\omega_0)N/2}{\sin(\omega+\omega_0)/2}\right]^2 \right\}. \tag{3.312}$$

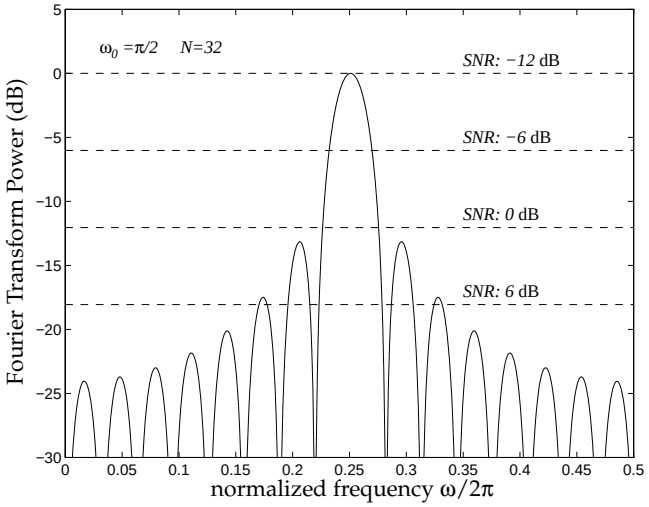

Figure 3.4. Signal and noise components of the periodogram, when $\omega_0 = \pi/2$ and $N = 32$. The solid curve depicts the signal component, $10\log_{10}(|X_N(e^{j\omega})|^2)$, while the dash lines depict the noise components $\sigma^2 = 10\log_{10}(E\left[|W_N(e^{j\omega})|^2\right])$ corresponding to various values of SNR.

and thus the SNR (3.311) effectively attains its peak at $\omega = \omega_0$. Since

$$\left|X_N(e^{j\omega_0})\right|^2 \approx \frac{A^2 N}{4},$$

the peak SNR is, using (3.286),

$$\gamma(\omega_0) = \max_{\omega} \gamma(\omega) = \frac{A^2 N}{4\sigma^2} = \frac{1}{2}\gamma N. \tag{3.313}$$

Note that (3.313) is a factor of $N/2$ larger than γ, the SNR for the original data.

The performance characteristics of the periodogram-based parameter estimators have some special features. To illustrate this, the variance in the amplitude and frequency estimates are plotted as a function of SNR in Figs. 3.5 and 3.6, respectively, along with the associated Cramér-Rao bounds. These figures reveal a distinct threshold phenomenon: for a given data length, there exists a SNR threshold above which the estimator variance closely tracks the Cramér-Rao bound, and below which the estimator variance diverges sharply from the bound. The phenomenon is particularly pronounced for the frequency estimator, but arises with the amplitude estimator as well.

This threshold behavior, which is also referred to as the "capture" effect, can be understood as follows. When the SNR is high enough that the peak in the periodogram at the true frequency protrudes prominently above the noise, the peak can be located quite accurately and the parameter estimation errors are due to slight, noise-induced distortion of the true peak. This is the regime in which the estimator performance tracks the Cramér-Rao bound. On the other hand, when the SNR is low enough that the correct peak lies below the noise and is obscured by other peaks, catastrophic estimation errors due to the estimator selecting the wrong peak entirely, leading to anomalous parameter estimates. This is the regime in which the estimator performance diverges from the Cramér-Rao bound. Sample periodograms corresponding to the different regimes are depicted in Fig. 3.7.

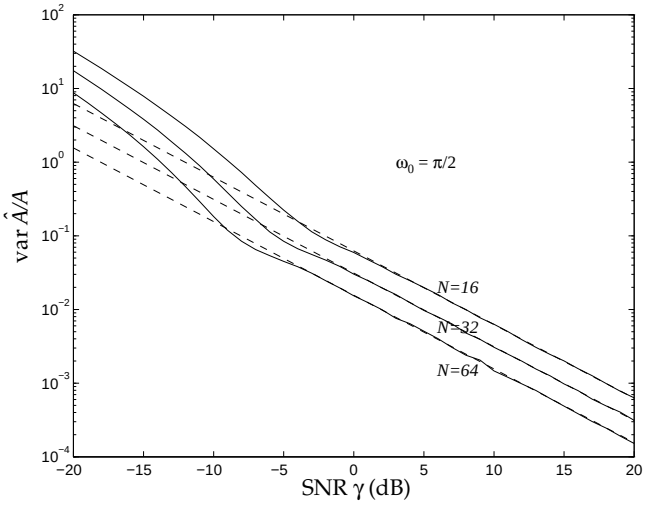

Figure 3.5. Variance of the ML estimate of the amplitude of an unknown sinusoid in white Gaussian noise as a function of SNR γ for various data lengths N. The dashed lines are the associated Cramér-Rao bounds.

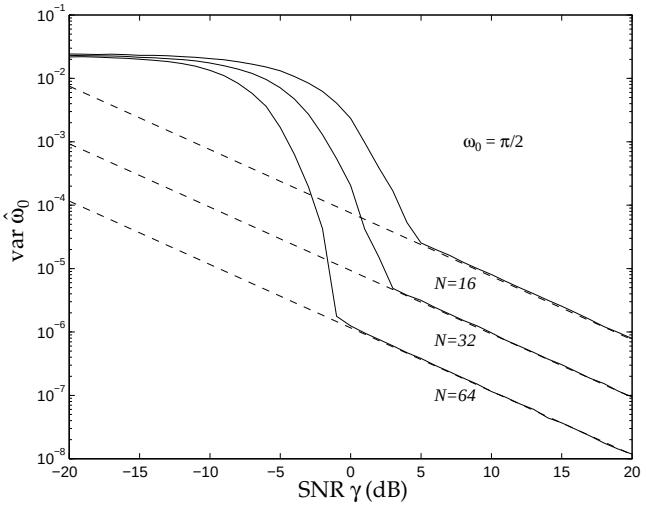

Figure 3.6. Variance of the ML estimate of the frequency of a unknown sinusoid in white Gaussian noise as a function of SNR γ for various data lengths N. The dashed lines are the associated Cramér-Rao bounds.

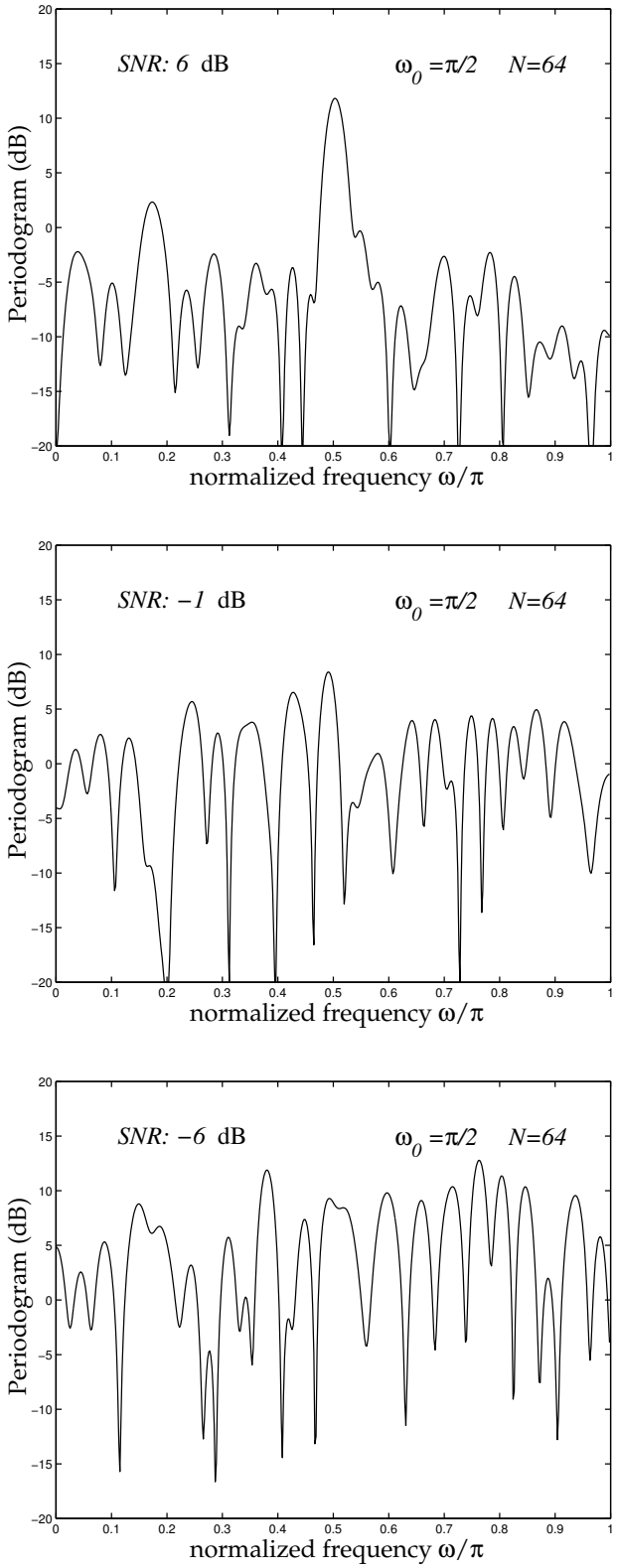

Figure 3.7. Sample periodograms for a sinusoid of frequency $\omega_0 = \pi/2$ and length $N = 64$. The top figure corresponds to SNR of 6 dB, which is above threshold. The middle figure corresponds to the approximate threshold SNR of -1 dB. The bottom figure corresponds to a sub-threshold SNR of -6 dB.

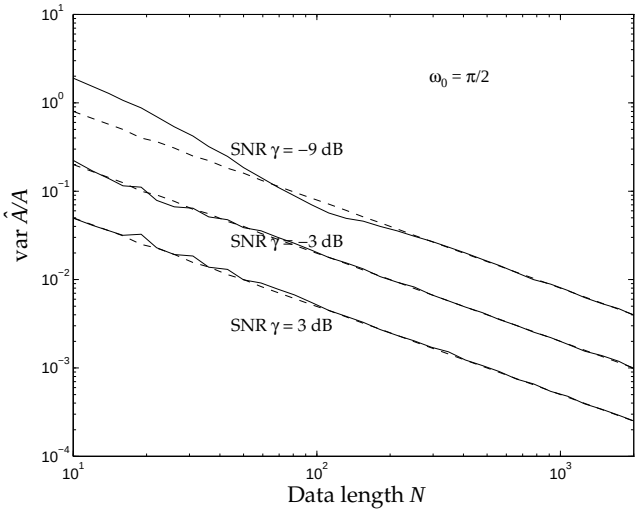

Figure 3.8. Variance of the ML estimate of the amplitude of an unknown sinusoid in white Gaussian noise as a function of data length N for various SNRs γ. The dashed lines are the associated Cramér-Rao bounds.

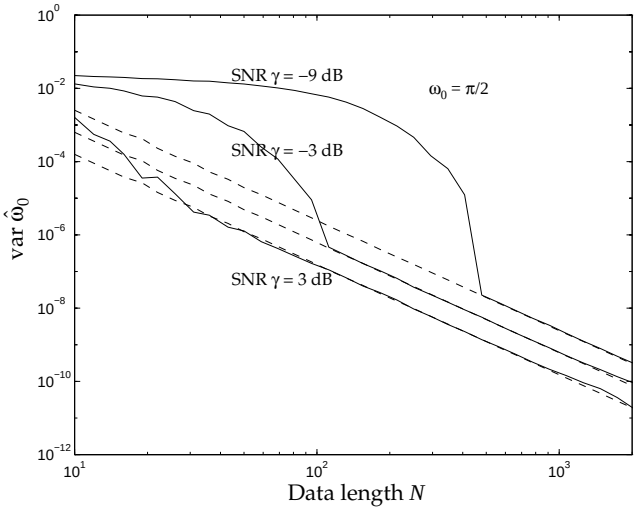

Figure 3.9. Variance of the ML estimate of the frequency of a unknown sinusoid in white Gaussian noise as a function of data length N for various SNRs γ. The dashed lines are the associated Cramér-Rao bounds.

Among other features revealed by Figs. 3.5 and 3.6, we see that since the peak SNR (3.313) is proportional to data length N, the threshold SNR decreases as N increases. Also, the slope of the bounds in the two figures are the same, reflecting the same inverse dependence on SNR γ [cf. (3.294a) and (3.294b)], though the offsets are quite different due to the different nature of the dependence on N. Performance variations with block length N are more fully apparent in Figs. 3.8 and 3.9. These figures show the variance in the estimates of A and ω_0, respectively, plotted as a function of N for several values of the SNR γ. Note that the slope of the Cramér-Rao bounds is greater by a factor of 3 (on the log-log scale) for the frequency estimate. This is because of the $1/N^3$ vs. $1/N$ dependence in the bounds apparent in comparisons of (3.294b) and (3.294a).

It is important to emphasize that, by contrast, linear estimation problems do not exhibit the kind of threshold behavior observed above. In fact, for linear estimation problems involving Gaussian data, we established that ML estimates are

efficient, so the associated Cramér-Rao bounds are accurate predictors of the performance attainable in practice. This is the case in sinusoid estimation problems where only the amplitude is unknown.

These distinctions underlie the familiar differences in the way signal quality varies in, e.g., AM and FM radio reception. AM reception has the characteristic that the quality degrades steadily with increasing distance from the source of the transmission. On the other hand, FM systems have the characteristic that within a certain radius of the source the quality of the reception is higher than corresponding AM systems, but that that outside this service area reception deteriorates sharply as the SNR drops below threshold.

More generally, the capture effect is a dominant feature of systems in many applications where there are inherent nonlinearities. In the next section, we discuss how the effect arises in this more general setting, and view the sinusoid estimation problem as a special instance of the phenomenon.

3.4.2 Threshold Behavior and the Capture Phenomenon

In this section, let's consider a vector generalization of Example 3.17 in which the measurements $\mathbf{y}$ depend on the parameter vector $\mathbf{x}$ via

$$\mathbf{y} = \mathbf{h}(\mathbf{x}) + \mathbf{w} \tag{3.314}$$

with $\mathbf{w} \sim N(\mathbf{0}, \mathbf{\Lambda_w})$, so that the measurements take the form of Gaussian random vector.

Let us first determine the associated Cramér-Rao bound for the problem. To begin, first note that, provided $\mathbf{\Lambda_w} > 0$ and $\mathbf{h}(\cdot)$ is differentiable,

$$\ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) = -\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{\Lambda_w}| - \frac{1}{2}(\mathbf{y}-\mathbf{h}(\mathbf{x}))^{\mathrm{T}}\mathbf{\Lambda_w}^{-1}(\mathbf{y}-\mathbf{h}(\mathbf{x})), \tag{3.315}$$

so

$$\frac{\partial}{\partial \mathbf{x}}\ln p_{\mathbf{y}}(\mathbf{y};\mathbf{x}) = (\mathbf{y}-\mathbf{h}(\mathbf{x}))^{\mathrm{T}}\mathbf{\Lambda_w}^{-1}\frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}}. \tag{3.316}$$

In turn, using (3.316) in (3.235) we obtain the Fisher matrix

$$\begin{aligned}\mathbf{I_y}(\mathbf{x}) &= E\left[\frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}}^{\mathrm{T}}\mathbf{\Lambda_w}^{-1}(\mathbf{y}-\mathbf{h}(\mathbf{x}))(\mathbf{y}-\mathbf{h}(\mathbf{x}))^{\mathrm{T}}\mathbf{\Lambda_w}^{-1}\frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}}\right] \\ &= \frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}}^{\mathrm{T}}\mathbf{\Lambda_w}^{-1}E\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\mathbf{\Lambda_w}^{-1}\frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}} \\ &= \frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}}^{\mathrm{T}}\mathbf{\Lambda_w}^{-1}\frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}}.\end{aligned} \tag{3.317}$$

From (3.317) we see that $\mathbf{I_y}(\mathbf{x}) > 0$ if and only if the Jacobian matrix[14] $d\mathbf{h}(\mathbf{x})/d\mathbf{x}$ is nonsingular. In this case, the Cramér-Rao bound on the covariance of unbiased

[14]This matrix was defined in Appendix 1.B of Chapter 1.

estimates $\hat{\mathbf{x}}(\mathbf{y})$ of $\mathbf{x}$ is

$$\mathbf{\Lambda}_{\hat{\mathbf{x}}}(\mathbf{x}) \geq \left[\frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}}^{\mathrm{T}} \mathbf{\Lambda}_{\mathbf{w}}^{-1} \frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}} \right]^{-1}. \tag{3.318}$$

As in Example 3.17, it is straightforward to show that an efficient estimate fails to exist when $\mathbf{h}(\cdot)$ is strictly nonlinear, i.e., unless $\mathbf{h}(\cdot)$ is a linear (affine) function. Nevertheless, the Cramér-Rao bound (3.329) does have the useful interpretation as the performance of a closely related linear system. To see this, consider a linearization of $\mathbf{h}(\cdot)$ about a particular value $\mathbf{x}_*$ that is near $\mathbf{x}$, and let us use $\tilde{\mathbf{x}}$ to denote the deviation, i.e.,

$$\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}_*. \tag{3.319}$$

This linearization is obtained from a Taylor series expansion of $\mathbf{h}(\mathbf{x})$ about $\mathbf{x}_*$, which, as described in Appendix 1.B of Chapter 1, takes the form

$$\mathbf{h}(\mathbf{x}) = \mathbf{h}(\mathbf{x}_*) + \left[\frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}} \right] \Bigg|_{\mathbf{x}=\mathbf{x}_*} \tilde{\mathbf{x}} + \cdots \tag{3.320}$$

In particular, when the higher-order terms $(\cdots)$ are neglected, the measurement deviations

$$\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{h}(\mathbf{x}_*) \tag{3.321}$$

are related to the parameter deviation $\tilde{\mathbf{x}}$ according to the *linear* model

$$\tilde{\mathbf{y}} \approx \left[\frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}} \right] \Bigg|_{\mathbf{x}=\mathbf{x}_*} \tilde{\mathbf{x}} + \mathbf{w}. \tag{3.322}$$

To verify this it suffices to substitute (3.320) into (3.321). This linearization is depicted in Fig. 3.10 for the case in which both $\mathbf{x}$ and $\mathbf{y}$ are scalars.

When (3.322) holds with equality, the minimum variance unbiased estimate of $\tilde{\mathbf{x}}$ based on $\tilde{\mathbf{y}}$ is the ML estimate, as is that for $\mathbf{x}$ based on $\mathbf{y}$. In particular, via the Gauss-Markov theorem (Example 3.16) we have

$$\hat{\mathbf{x}}_{\mathrm{MVU}}(\mathbf{y}) = \mathbf{x}_* + \Phi^{-1}(\mathbf{x}_*) \left. \frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}}^{\mathrm{T}} \right|_{\mathbf{x}=\mathbf{x}_*} \mathbf{\Lambda}_{\mathbf{w}}^{-1} \left(\mathbf{y} - \mathbf{h}(\mathbf{x}_*) \right) \tag{3.323}$$

where

$$\Phi(\mathbf{x}) = \frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}}^{\mathrm{T}} \mathbf{\Lambda}_{\mathbf{w}}^{-1} \frac{d\mathbf{h}(\mathbf{x})}{d\mathbf{x}}. \tag{3.324}$$

Moreover, via the Gauss-Markov theorem we also have that the covariance of this estimate is

$$\mathbf{\Lambda}_{\mathrm{MVU}} = \Phi^{-1}(\mathbf{x}_*), \tag{3.325}$$

which when $\mathbf{x}_* = \mathbf{x}$ corresponds to the Cramér-Rao bound for the problem [cf. (3.318)].

This analysis implies that if we know *a priori* that $\mathbf{x}$ lies in a neighborhood of $\mathbf{x}_*$, and that the neighborhood is small enough that the Jacobian matrix $d\mathbf{h}(\mathbf{x})/d\mathbf{x}$

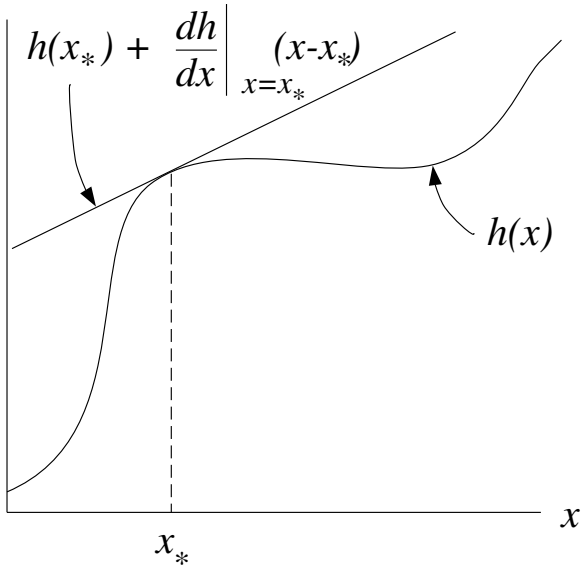

Figure 3.10. Linearization of the measurement function $\mathbf{h}(\cdot)$ in the Gaussian nonlinear estimation problem.

can be well approximated as essentially constant in that neighborhood, then the Cramér-Rao bound will yield an accurate measure of achievable performance. For this reason, the Cramér-Rao bound is frequently referred to as a *local* bound.

When, as is the case more generally, such *a priori* localization information is not available, then ML estimates $\hat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y})$ are generally not efficient. However, these estimates are often asymptotically efficient at high SNR, where estimation errors are small. To understand this property, we explore the ML estimate and its relationship to the Cramér-Rao bound in more depth.

To begin, the ML estimate is obtained by maximizing (3.315) or, equivalently, as the solution to the following nonlinear least-squares minimization problem

$$\hat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y}) = \arg\min_{\mathbf{a}} J(\mathbf{a}) \tag{3.326a}$$

where

$$J(\mathbf{a}) = (\mathbf{y} - \mathbf{h}(\mathbf{a}))^{\mathrm{T}} \mathbf{\Lambda}_{\mathbf{w}}^{-1} (\mathbf{y} - \mathbf{h}(\mathbf{a})). \tag{3.326b}$$

Because the resulting estimate will generally be biased, the bound (3.318) won't apply, but again may asymptotically.

Although the ML estimate may lack specific optimality properties, it is an at least intuitively reasonable estimate, as the form of (3.326b) reveals. In particular, the maximum likelihood estimate minimizes a weighted sum of squared errors, where the weighting is determined by $\mathbf{\Lambda}_{\mathbf{w}}^{-1}$ so that more accurate observations are weighted more heavily. For this reason, this type of "least-squares" estimate is often used without any justification in terms of ML estimation.

In the sequel, it will be convenient to express the ML estimate in a different form. In particular, expanding out the quadratic form (3.326b) and discarding the quadratic term in $\mathbf{y}$ (since it doesn't depend on $\mathbf{a}$ and hence won't affect our

optimization over $\mathbf{a}$) we see that the ML estimate can be rewritten as

$$\hat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y}) = \arg\max_{\mathbf{a}} r(\mathbf{y};\mathbf{a}), \tag{3.327a}$$

where

$$r(\mathbf{y};\mathbf{a}) = \mathbf{h}^{\mathrm{T}}(\mathbf{a})\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{y} - \frac{1}{2}\mathbf{h}^{\mathrm{T}}(\mathbf{a})\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{h}(\mathbf{a}). \tag{3.327b}$$

To simplify our exposition, in the sequel let us restrict our attention to the case in which $\mathbf{x}$ is an unknown scalar x. In this case, the measurement model (3.314) and Cramér-Rao bound (3.318) specialize to, respectively,

$$\mathbf{y} = \mathbf{h}(x) + \mathbf{w} \tag{3.328}$$

and

$$\mathbf{\Lambda}_{\hat{x}}(x) \geq \left[\frac{d\mathbf{h}^{\mathrm{T}}(x)}{dx}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\frac{d\mathbf{h}(x)}{dx}\right]^{-1}. \tag{3.329}$$

Also, we rewrite (3.327b) in this case as

$$r(\mathbf{y};a) = \mathbf{h}^{\mathrm{T}}(a)\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{y} - \frac{1}{2}\mathbf{h}^{\mathrm{T}}(a)\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{h}(a) \tag{3.330}$$

and view the first term on the right-hand side of (3.330) as a (weighted) inner product of the observed value $\mathbf{y}$ and its candidate values $\mathbf{h}(a)$. The second term in (3.330) is an energy (norm) term, i.e., it is a measure of the corresponding SNR we expect to see. As we'll see, in the linear case the *dominant* effect a has is on signal energy; this is the case, for example, when we are estimating the amplitude of a sinusoid of known frequency (the AM problem). On the other hand, in many nonlinear problems a has little or no effect on SNR; this is the case, for example, when we are estimating the frequency of a sinusoid (the FM problem). It's these fundamental differences that generally leads to the Cramér-Rao bound being overly optimistic in the nonlinear problem.

The statistics of the objective function (3.330), and hence overall system performance, can be expressed completely in terms of "nonlinear inner products" of the form

$$C(x_1, x_2) = \mathbf{h}(x_1)^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{h}(x_2) \tag{3.331}$$

which in essence measure how similar the observations $\mathbf{y}$ will be on average if a is x_1 versus x_2. In particular, since $r(\mathbf{y};a)$ is a linear function of $\mathbf{y}$, it is a Gaussian random variable, and thus is fully described by its mean

$$m_r(a) = E\left[r(\mathbf{y};a)\right] = C(a,x) - \frac{1}{2}C(a,a) \tag{3.332}$$

and variance

$$\lambda_r(a) = \mathrm{var}\, r(\mathbf{y};a) = C(a,a) = \mathbf{h}^{\mathrm{T}}(a)\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{h}(a). \tag{3.333}$$

Fig. 3.11 illustrates a typical example of what these statistics look like as a function of a. This example corresponds to a special case of the sinusoid estimation problem

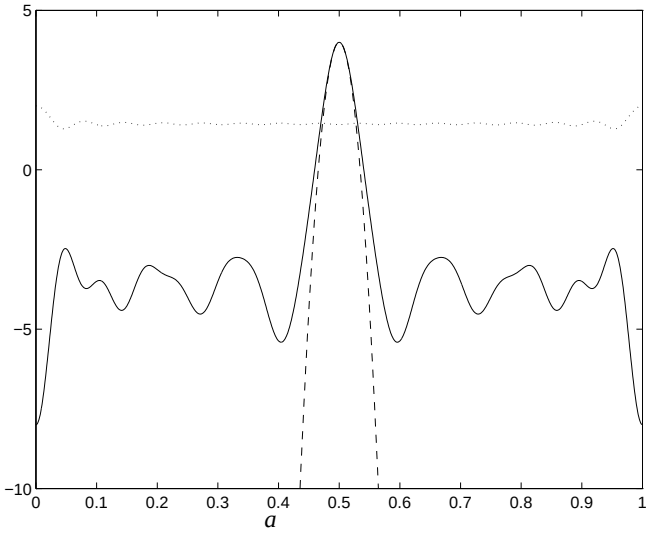

Figure 3.11. Statistics of an objective function $r(\mathbf{y}; a)$. This example is generated from the sinusoid estimation problem involving $N = 16$ data samples corrupted by white Gaussian noise, with the unknown parameter x being the normalized frequency of the sinusoid ($1/2$ in this case). The solid curve depicts the mean $m_r(a)$, while the dashed curve depicts a quadratic function with the same curvature at its peak. The dotted curve depicts $\sqrt{\lambda_r(a)}$ corresponding to an SNR of 12 dB.

of Section 3.4.1 in which only the frequency parameter is unknown (and is denoted using x). As this figure reflects, on average the peak of the objective function lies at the true parameter value. Noise in $\mathbf{y}$ perturbs the values of $r(\mathbf{y}; x)$ away from the solid curve in the figure, which in turn leads to the peak value shifting and, hence, estimation error. As the figure also reflects, the standard deviation of the associated perturbations does not vary strongly with the independent variable a.

A quadratic with the same curvature as $m_r(a)$ at its peak is also depicted in Fig. 3.11. The curvature of $m_r(a)$ at its peak (i.e., $a = x$) is intimately related to the Cramér-Rao bound for the problem. To see this, note that using (3.315) and (3.327b) we can express $r(\mathbf{y}; a)$ as

$$r(\mathbf{y}; a) = \ln p_{\mathbf{y}}(\mathbf{y}; a) + f(\mathbf{y}) \tag{3.334}$$

where $f(\mathbf{y})$ does not depend on a. Thus, the curvature is given by

$$\begin{aligned} \frac{d^2}{da^2} m_r(a)\Big|_{a=x} &= E\left[\frac{\partial^2}{\partial a^2} r(\mathbf{y}; a)\right]\Bigg|_{a=x} \\ &= E\left[\frac{\partial^2}{\partial a^2} \ln p_{\mathbf{y}}(\mathbf{y}; a)\right]\Bigg|_{a=x} \\ &= -I_{\mathbf{y}}(x). \end{aligned} \tag{3.335}$$

More generally, for the case of vector parameters $\mathbf{x}$, it is straightforward to verify that the Hessian matrix (as defined in Appendix 1.B) for $m_r(\mathbf{a})$ at its peak $\mathbf{a} = \mathbf{x}$ is the negative of the Fisher information, i.e.,

$$\frac{d^2}{d\mathbf{a}^2} m_r(\mathbf{a})\Big|_{\mathbf{a}=\mathbf{x}} = -\mathbf{I}_{\mathbf{y}}(\mathbf{x}). \tag{3.336}$$

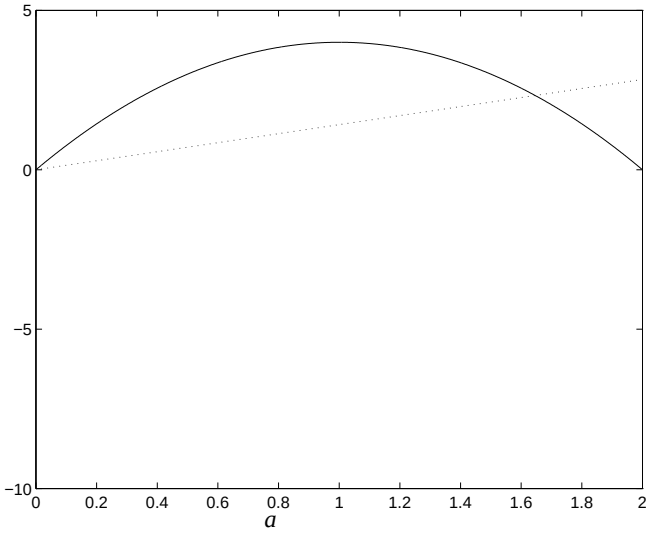

Figure 3.12. Statistics of an objective function $r(\mathbf{y}; a)$. This example is generated from the sinusoid estimation problem involving $N = 16$ data samples corrupted by white Gaussian noise, with the unknown parameter x being the amplitude of the sinusoid (unity in this case). The solid curve depicts the mean $m_r(a)$, while the dotted curve depicts $\lambda_r(a)$ corresponding to an SNR of 12 dB.

Note that when $\mathbf{h}(\cdot)$ is a linear function, e.g., $\mathbf{h}(x) = \mathbf{c}x$, then $m_r(a)$ is quadratic in a:

$$m_r(a) = \left(\mathbf{c}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{c}\right)\, a\, x - \frac{1}{2}\left(\mathbf{c}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{c}\right)\, a^2. \tag{3.337}$$

This is illustrated in Fig. 3.12. Since the Cramér-Rao bound is tight in this linear case, the curvature of $m_r(a)$ at its peak fully characterizes the performance of the ML estimator.

From this perspective, in the more general nonlinear case the Cramér-Rao bound corresponds to is fitting a quadratic at the peak of $m_r(a)$ as depicted in Fig. 3.11, and using the curvature as a measure of the performance. However, the dashed line in Fig. 3.11 falls off sharply as a function of x, consistent with the fact that in the linear case signal energy is a strong function of x. For this reason, making very large errors in a linear problem is extremely unlikely because of the enormous noise energy needed to cause such an error. However, in nonlinear problems where signal energy is at most a weak function of x, behavior like that depicted in Fig. 3.11 is more typical. In this case much smaller noise values are needed to push a value of $r(\mathbf{y}; a)$ located far from $a = x$ above the value of $r(\mathbf{y}; x)$, and therefore the Cramér-Rao bound tends to grossly underestimate the probability of large estimation errors. Whether this is ultimately significant or not depends upon the size of the noise variance.

This type of behavior manifests itself as the capture phenomenon we saw with the sinusoid estimation problem explored in Section 3.4.1. For small noise variances, the Cramér-Rao bound is accurate, since large errors occur with negligible frequency. As the noise increases, however, a threshold effect occurs at some value of the noise variance beyond which large errors become significant. At this point achievable performance becomes considerably worse than the optimistic Cramér-Rao bound prediction.

This interpretation provide additional insight into the qualitative differences in behavior between AM and FM radio reception discussed at the end of Section 3.4.1. For FM systems, where the parameter of interest is frequency, the associated $m_r(a)$ is as depicted in Fig. 3.11. For AM systems, where the parameter of interest is amplitude and appears linearly in the observations, the associated $m_r(a)$ is a quadratic as depicted in Fig. 3.12. As a comparison between these figures reflects, the peak is much broader in the AM case than the FM case. Since the breadth of the peak is determined by the curvature, this implies that at reasonably high SNR, errors in FM demodulation are much smaller than those in AM demodulation. As the SNR degrades, however, the AM reception degrades in a rather uniform manner with SNR, while FM reception degrades abruptly when the SNR reaches the threshold value.

3.4.3 Computation of ML Estimates

We close this chapter with a discussion of issues associated with the computation of the ML estimate. As (3.327) reflects, $\hat{x}_{\mathrm{ML}}$ is in principle determined by evaluating $r(\mathbf{y}; a)$ for all values of a and choosing that for which the maximum is attained. Obviously this isn't viable in practice. One practical approach that can be used is the following successive-linearization strategy:

1. Make an initial guess of the estimate, i.e., let $i = 0$ and let $\hat{x}^{(i)} = x_*$ for some x_*.
2. Linearize $\mathbf{h}(x)$ about $x = \hat{x}^{(i)}$, i.e., assume $x = \hat{x}^{(i)} + \Delta_i$ and solve the linearized estimation problem for $\hat{\Delta}_i$.
3. Generate a new estimate via $\hat{x}^{(i+1)} = \hat{x}^{(i)} + \hat{\Delta}_i$, and increment i.
4. Go to step 2.

This iterative algorithm typically converges to a local maximum of the objective function $r(\mathbf{y}; x)$. Alternative methods for finding such local maxima can also be used.

To find the global maximum, a two-stage procedure can be used in principle. First a coarse search is performed essentially as an M-ary hypothesis test. In particular, we choose a discrete set of values $x_1 < x_2 < \cdots < x_M$ and let

$$\hat{x}^{(0)} = x_{\hat{i}} \quad \text{where} \quad \hat{i} = \arg\max_i r(\mathbf{y}; x_i). \tag{3.338}$$

The estimate (3.338) can then be used as the initial estimate for a local search based on, e.g., successive-linearization.

Note that for the coarse search to be useful, the grid points $x_1, x_2, \ldots, x_M$ need to be chosen so that $r(\mathbf{y}; x_i)$ is likely to be larger than $r(\mathbf{y}; x_j)$ for $j \neq i$ if the actual value of x is closest to x_i. For example, in a scenario like that depicted

in Fig. 3.11, we might partition the a-axis up into small intervals of width on the order of the width of the main lobe of the solid curve and take as the x_i the centers of these intervals.

Finally, the two-stage algorithm leads to one more convenient interpretation of the capture effect in a nonlinear estimation problem. Specifically, when we choose the spacing between the x_i to be sufficiently small that the $E[r(\mathbf{y}; x_i)]$ are roughly quadratic near x_i the Cramér-Rao bound provides an accurate measure of estimation error provided the correct x_i is chosen in the first, coarse estimation stage. As the noise variance increases, however, there is an increasing probability of making an error in this first stage, and it is this behavior that leads to the threshold phenomenon.

3.A ALTERNATE FORMULAS FOR LINEAR LEAST-SQUARES ESTIMATION

As mentioned in Section 3.2.5 there are a number of alternate expressions for the quantities involved in linear-least squares estimation. Recall that the problem is that of estimating $\mathbf{x}$ given the measurements

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w} \tag{3.339}$$

For this problem, the LLS estimator (which also corresponds to the BLS estimator when $\mathbf{x}$ and $\mathbf{w}$ are independent Gaussian random vectors) and its performance are given by, respectively, (3.108) and (3.109) with (3.110), which we repeat here for convenience:

$$\hat{\mathbf{x}}_{\mathrm{LLS}}(\mathbf{y}) = \mathbf{m}_{\mathbf{x}} + \mathbf{K}\left(\mathbf{y} - \mathbf{H}\mathbf{m}_{\mathbf{x}}\right) \tag{3.340}$$

$$\mathbf{\Lambda}_{\mathrm{LLS}} = \mathbf{\Lambda}_{\mathbf{x}} - \mathbf{K}\left(\mathbf{H}\mathbf{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}} + \mathbf{\Lambda}_{\mathbf{w}}\right)\mathbf{K}^{\mathrm{T}}, \tag{3.341}$$

where

$$\mathbf{K} = \mathbf{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}} + \mathbf{\Lambda}_{\mathbf{w}})^{-1}. \tag{3.342}$$

Let us first derive the following alternative form for the error covariance

$$\mathbf{\Lambda}_{\mathrm{LLS}} = \left[\mathbf{\Lambda}_{\mathbf{x}}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H}\right]^{-1}. \tag{3.343}$$

Showing (3.343), i.e.,

$$\mathbf{\Lambda}_{\mathrm{LLS}}\left[\mathbf{\Lambda}_{\mathbf{x}}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H}\right] = \mathbf{I}$$

is, using (3.341) with (3.342), equivalent to showing that

$$\left[\mathbf{\Lambda}_{\mathbf{x}} - \mathbf{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}\left[\mathbf{H}\mathbf{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}} + \mathbf{\Lambda}_{\mathbf{w}}\right]^{-1}\mathbf{H}\mathbf{\Lambda}_{\mathbf{x}}\right]\left[\mathbf{\Lambda}_{\mathbf{x}}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H}\right] - \mathbf{I} = \mathbf{0}. \tag{3.344}$$

Eq. (3.344) can in fact be obtained from the expressions (1.231)–(1.235) in Appendix 1.A to Chapter 1 for inverting block matrices. Here, however, we verify

this more directly. Expanding the expressions in (3.344) and rearranging terms we find that the left-hand side of (3.344) is equivalent to

$$\begin{aligned}
&\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}\left(-\left[\mathbf{H}\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}+\boldsymbol{\Lambda}_{\mathbf{w}}\right]^{-1}\boldsymbol{\Lambda}_{\mathbf{w}}+\mathbf{I}-\left[\mathbf{H}\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}+\boldsymbol{\Lambda}_{\mathbf{w}}\right]^{-1}\mathbf{H}\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}\right)\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H} \\
&\quad=\ \boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{I}-\left[\mathbf{H}\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}+\boldsymbol{\Lambda}_{\mathbf{w}}\right]^{-1}\left[\mathbf{H}\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}+\boldsymbol{\Lambda}_{\mathbf{w}}\right]\right)\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}\mathbf{H} \\
&\quad=\ \mathbf{0}
\end{aligned} \tag{3.345}$$

so that (3.343) is verified.

Let us consider one additional expression for the error covariance. In practice, (3.343) is typically not well-suited for actual numerical computation of the error covariance $\boldsymbol{\Lambda}_{\mathrm{LLS}}$. Neither, however, is (3.341). Specifically, as a covariance matrix $\boldsymbol{\Lambda}_{\mathrm{LLS}}$ is at least positive semidefinite. Eq. (3.341) expresses $\boldsymbol{\Lambda}_{\mathrm{LLS}}$ as the difference between two positive semidefinite matrices, and, in cases in which this difference involves subtracting large numbers, it is possible that numerical errors can lead to the computed value of $\boldsymbol{\Lambda}_{\mathrm{LLS}}$ losing its definiteness. A detailed investigation of numerical computation issues is beyond the scope of this course. However, we point out that rewriting the error covariance in the form

$$\boldsymbol{\Lambda}_{\mathrm{LLS}}=\left[\mathbf{I}-\mathbf{K}\mathbf{H}\right]\boldsymbol{\Lambda}_{\mathbf{x}}\left[\mathbf{I}-\mathbf{K}\mathbf{H}\right]^{\mathrm{T}}+\mathbf{K}\boldsymbol{\Lambda}_{\mathbf{w}}\mathbf{K}^{\mathrm{T}}, \tag{3.346}$$

which involves the *sum* of positive definite matrices, is much preferred for numerical computation. To derive (3.346), we use (3.339), (3.340) to write

$$\mathbf{e}(\mathbf{x},\mathbf{y})=\mathbf{x}-\hat{\mathbf{x}}(\mathbf{y})=\left[\mathbf{I}-\mathbf{K}\mathbf{H}\right](\mathbf{x}-\mathbf{m}_{\mathbf{x}})-\mathbf{K}\mathbf{w} \tag{3.347}$$

Then, since $\mathbf{x}$ and $\mathbf{w}$ are uncorrelated, we immediately obtain the expression (3.346) for $\boldsymbol{\Lambda}_{\mathrm{LLS}}$.

As a final comment, we note that the gain (3.342) can also be written in an alternative form, viz.,

$$\mathbf{K}=\boldsymbol{\Lambda}_{\mathrm{LLS}}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}. \tag{3.348}$$

To verify this, we use (3.341) to write

$$\begin{aligned}
\boldsymbol{\Lambda}_{\mathrm{LLS}}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1} &= \left[\boldsymbol{\Lambda}_{\mathbf{x}}-\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}\left[\mathbf{H}\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}+\boldsymbol{\Lambda}_{\mathbf{w}}\right]^{-1}\mathbf{H}\boldsymbol{\Lambda}_{\mathbf{x}}\right]\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1} \\
&= \boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}\left[\mathbf{H}\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}+\boldsymbol{\Lambda}_{\mathbf{w}}\right]^{-1}\left[\mathbf{H}\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}+\boldsymbol{\Lambda}_{\mathbf{w}}-\mathbf{H}\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}\right]\boldsymbol{\Lambda}_{\mathbf{w}}^{-1} \\
&= \boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}\left[\mathbf{H}\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}+\boldsymbol{\Lambda}_{\mathbf{w}}\right]^{-1}
\end{aligned} \tag{3.349}$$

which agrees with (3.342).

3.B FISHER INFORMATION CALCULATIONS FOR SINUSOID ESTIMATION

We compute the Fisher matrix entries one at a time. First,

$$\begin{aligned}
[\mathbf{I}_{\mathbf{y}}(\mathbf{x})]_{11} &= -E\left[\frac{\partial^2}{\partial A^2}\ell(\mathbf{y};\mathbf{x})\right] \\
&= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\cos^2(\omega_0 n+\Theta) \\
&= \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}[1+\cos(2\omega_0 n+2\Theta)] \\
&= \frac{N}{2\sigma^2}+\frac{N}{2\sigma^2}\left[\frac{1}{N}\sum_{n=0}^{N-1}\cos(2\omega_0 n+2\Theta)\right] \\
&= \frac{N}{2\sigma^2}+\frac{N}{2\sigma^2}\operatorname{Re}\{\xi(\omega_0)\}, \qquad (3.350)
\end{aligned}$$

where we introduce the function $\xi(\cdot)$ defined via

$$\xi(\omega_0)=\frac{1}{N}\sum_{n=0}^{N-1}e^{j(2\omega_0 n+2\Theta)}. \qquad (3.351)$$

As we'll see, this function and its first and second derivatives, respectively

$$\xi'(\omega_0)=\frac{2}{N}\sum_{n=0}^{N-1}n\,e^{j(2\omega_0 n+2\Theta)} \qquad (3.352)$$

and

$$\xi''(\omega_0)=\frac{4}{N}\sum_{n=0}^{N-1}n^2\,e^{j(2\omega_0 n+2\Theta)}, \qquad (3.353)$$

play a central role in the Fisher information for the problem.

Next, using (3.352),

$$\begin{aligned}
[\mathbf{I}_{\mathbf{y}}(\mathbf{x})]_{12} &= -E\left[\frac{\partial^2}{\partial A\partial\omega_0}\ell(\mathbf{y};\mathbf{x})\right] \\
&= -\frac{1}{\sigma^2}\sum_{n=0}^{N-1}An\cos(\omega_0 n+\Theta)\sin(\omega_0 n+\Theta) \\
&= -\frac{AN}{2\sigma^2}\left[\frac{1}{N}\sum_{n=0}^{N-1}n\sin(2\omega_0 n+2\Theta)\right] \\
&= -\frac{AN}{2\sigma^2}\frac{1}{2}\operatorname{Im}\{\xi'(\omega_0)\}, \qquad (3.354)
\end{aligned}$$

and, using (3.351),

$$\begin{aligned}
[\mathbf{I}_{\mathbf{y}}(\mathbf{x})]_{13} &= -E\left[\frac{\partial^2}{\partial A \partial \Theta}\ell(\mathbf{y};\mathbf{x})\right] \\
&= -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} A\cos(\omega_0 n+\Theta)\sin(\omega_0 n+\Theta) \\
&= -\frac{AN}{2\sigma^2}\left[\frac{1}{N}\sum_{n=0}^{N-1}\sin(2\omega_0 n+2\Theta)\right] \\
&= -\frac{AN}{2\sigma^2}\operatorname{Im}\{\xi(\omega_0)\}\,. \qquad (3.355)
\end{aligned}$$

Proceeding, using (3.293) and (3.353),

$$\begin{aligned}
[\mathbf{I}_{\mathbf{y}}(\mathbf{x})]_{22} &= -E\left[\frac{\partial^2}{\partial {\omega_0}^2}\ell(\mathbf{y};\mathbf{x})\right] \\
&= -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} A^2 n^2\sin^2(\omega_0 n+\Theta) \\
&= \frac{A^2}{2\sigma^2}\sum_{n=0}^{N-1} n^2\left[1-\cos(2\omega_0 n+2\Theta)\right] \\
&= \frac{A^2}{2\sigma^2}\sum_{n=0}^{N-1} n^2\left[1-\cos(2\omega_0 n+2\Theta)\right] \\
&= \frac{A^2}{2\sigma^2}\left[\sum_{n=0}^{N-1} n^2\right] - \frac{A^2N}{2\sigma^2}\left[\frac{1}{N}\sum_{n=0}^{N-1} n^2\cos(2\omega_0 n+2\Theta)\right] \\
&= \frac{A^2}{2\sigma^2}T_N - \frac{A^2N}{2\sigma^2}\frac{1}{4}\operatorname{Re}\{\xi''(\omega_0)\}\,, \qquad (3.356)
\end{aligned}$$

and, using (3.292) and (3.352),

$$\begin{aligned}
[\mathbf{I}_{\mathbf{y}}(\mathbf{x})]_{23} &= -E\left[\frac{\partial^2}{\partial \omega_0 \partial \Theta}\ell(\mathbf{y};\mathbf{x})\right] \\
&= -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} A^2 n\sin^2(\omega_0 n+\Theta) \\
&= -\frac{A^2}{2\sigma^2}\sum_{n=0}^{N-1} n\left[1-\cos(2\omega_0 n+2\Theta)\right] \\
&= \frac{A^2}{2\sigma^2}\left[\sum_{n=0}^{N-1} n\right] - \frac{A^2N}{2\sigma^2}\left[\sum_{n=0}^{N-1} n\cos(2\omega_0 n+2\Theta)\right] \\
&= \frac{A^2}{2\sigma^2}S_N - \frac{A^2N}{2\sigma^2}\frac{1}{2}\operatorname{Re}\{\xi'(\omega_0)\}\,. \qquad (3.357)
\end{aligned}$$

Finally, using (3.351),

$$\begin{aligned}
[\mathbf{I}_{\mathbf{y}}(\mathbf{x})]_{33} &= -E\left[\frac{\partial^2}{\partial^2\Theta}\ell(\mathbf{y};\mathbf{x})\right] \\
&= -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} A^2\sin^2(\omega_0 n+\Theta) \\
&= \frac{A^2}{2\sigma^2}\sum_{n=0}^{N-1}[1-\cos(2\omega_0+2\Theta)] \\
&= \frac{A^2N}{2\sigma^2} - \frac{A^2N}{2\sigma^2}\left[\frac{1}{N}\sum_{n=0}^{N-1}\cos(2\omega_0 n+2\Theta)\right] \\
&= \frac{A^2N}{2\sigma^2} - \frac{A^2N}{2\sigma^2}\,\mathrm{Re}\left\{\xi(\omega_0)\right\}. \qquad (3.358)
\end{aligned}$$

To develop the asymptotic (large N) behavior of the Fisher information, we explore the corresponding behavior of the function $\xi(\omega_0)$ for $0 \le \omega_0 < \pi$. To begin, since (3.351) is a finite geometric sum, we readily obtain

$$|\xi(\omega_0)| = \left|\frac{1}{N}e^{j2\Theta}\cdot\frac{1-e^{j2\omega_0 N}}{1-e^{j2\omega_0}}\right| = \frac{1}{N}\left|\frac{\sin\omega_0 N}{\sin\omega_0}\right|, \qquad (3.359)$$

which is depicted in Fig. 3.13. As this figure reflects, and consistent with (3.359), the function $\xi(\omega_0)$ has a mainlobe at $\omega_0 = 0$ of unit height and a series of sidelobes spaced apart by π/N; i.e.,

$$\xi\left(\frac{n\pi}{N}\right) = \begin{cases} 1 & n = 0 \\ 0 & n = 1,2,\ldots,N-1 \end{cases}. \qquad (3.360)$$

The first sidelobe has its peak near $\omega_0 = 3\pi/2N$, and successive sidelobes get progressively smaller. In particular, when N is large, this first sidelobe has height

$$\left|\xi\left(\frac{3\pi}{2N}\right)\right| = \frac{1}{N}\left|\frac{\sin(3\pi/2)}{\sin(3\pi/2N)}\right| \approx \frac{2}{3\pi}, \qquad (3.361)$$

while the height of the smallest sidelobe is no larger than $1/N$. Indeed, using (3.359), we see

$$\left|\xi\left(\frac{\pi}{2}\right)\right| = \frac{1}{N}\left|\sin\left(\frac{\pi N}{2}\right)\right| \le \frac{1}{N}. \qquad (3.362)$$

More generally, in the large N regime (3.295) we have[15]

$$|\xi(\omega_0)| \ll 1, \qquad \frac{\xi'(\omega_0)}{N} \ll 1, \qquad \frac{\xi''(\omega_0)}{N^2} \ll 1, \qquad (3.363)$$

[15]More precisely, for any fixed ω_0 such that $0 < \omega_0 < \pi$ and given an $\varepsilon > 0$ that can be arbitrarily small, there exists an $N = N(\omega_0)$ such that

$$|\xi(\omega_0)| < \varepsilon, \qquad \frac{\xi'(\omega_0)}{N} < \varepsilon, \qquad \frac{\xi''(\omega_0)}{N^2} < \varepsilon.$$

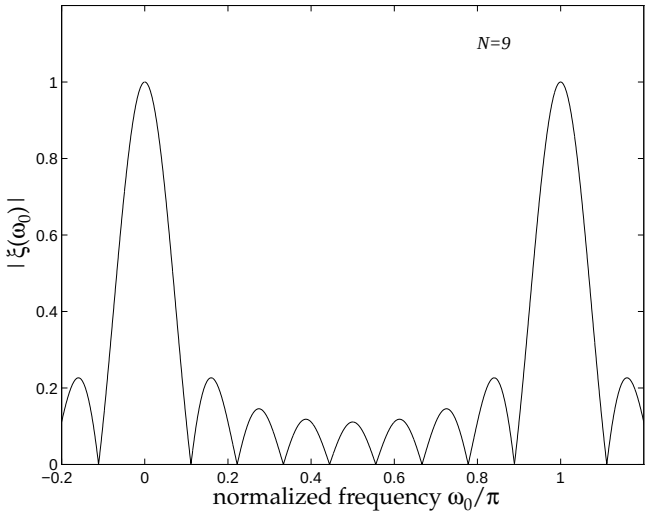

Figure 3.13. Plot of the magnitude of the function $\xi(\omega_0)$ defined in (3.351) when $N = 9$.

i.e.,

$$\xi(\omega_0) \sim o(1), \qquad \xi'(\omega_0) \sim o(N), \qquad \xi''(\omega_0) \sim o(N^2). \tag{3.364}$$

Using (3.364) in (3.350)–(3.358) yields the desired (3.291).

3.C MAXIMUM LIKELIHOOD SINUSOID ESTIMATOR DERIVATION

To obtain our solution, we perform an invertible transformation of the parameter set of the form

$$(A, \omega_0, \Theta) \longrightarrow (\alpha_1, \alpha_2, \omega_0)$$

where

$$\alpha_1 = A\cos\Theta, \qquad \alpha_2 = A\sin\Theta \tag{3.365}$$

and develop ML estimates $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\omega}_0)$ of the new parameters. Then via the invariance property of ML estimation, the corresponding ML estimates $(\hat{A}, \hat{\omega}_0, \hat{\Theta})$ follow immediately as

$$\hat{A} = \sqrt{\hat{\alpha}_1^2 + \hat{\alpha}_2^2} \qquad \hat{\Theta} = -\tan^{-1}\left(\frac{\hat{\alpha}_2}{\hat{\alpha}_1}\right). \tag{3.366}$$

A straightforward solution in terms of the new parameters involves casting the problem as a nested optimization: we first find the optimum $\hat{\alpha}_1$ and $\hat{\alpha}_2$ in terms of ω_0, then solve for the optimum $\hat{\omega}_0$, i.e.,

$$\begin{aligned}\min_{\alpha_1,\alpha_2,\omega_0} J(\alpha_1, \alpha_2, \omega_0) &= \min_{\omega_0}\left\{\min_{\alpha_1,\alpha_2} J(\alpha_1, \alpha_2, \omega_0)\right\} \\ &= \min_{\omega_0} J(\hat{\alpha}_1(\omega_0), \hat{\alpha}_2(\omega_0), \omega_0) \\ &= J(\hat{\alpha}_1(\hat{\omega}_0), \hat{\alpha}_2(\hat{\omega}_0), \hat{\omega}_0).\end{aligned} \tag{3.367}$$

The first minimization is straightforward because α_1 and α_2 appear linearly in the objective function. In particular, using

$$\mathbf{c}(\omega_0) = \begin{bmatrix} 1 \\ \cos\omega_0 \\ \vdots \\ \cos\omega_0(N-1) \end{bmatrix}, \qquad \mathbf{s}(\omega_0) = \begin{bmatrix} 0 \\ \sin\omega_0 \\ \vdots \\ \sin\omega_0(N-1) \end{bmatrix} \tag{3.368}$$

we can write

$$\begin{aligned} \mathbf{y} &= \alpha_1\mathbf{c}(\omega_0) + \alpha_2\mathbf{s}(\omega_0) + \mathbf{w} \\ &= \mathbf{H}(\omega_0)\boldsymbol{\alpha} + \mathbf{w} \end{aligned} \tag{3.369}$$

where

$$\mathbf{H}(\omega_0) = \begin{bmatrix} \mathbf{c}(\omega_0) & \mathbf{s}(\omega_0) \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}. \tag{3.370}$$

Hence, for a given ω_0, solving for $\boldsymbol{\alpha}$ is a standard linear-Gaussian ML estimation problem. In particular, we obtain,

$$\begin{aligned} \hat{\boldsymbol{\alpha}}(\omega_0) &= \arg\min_{\alpha_1,\alpha_2} J(\alpha_1, \alpha_2, \omega_0) \\ &= \min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{H}(\omega_0)\boldsymbol{\alpha}\|^2 \\ &= \left[\mathbf{H}(\omega_0)^{\mathrm{T}}\mathbf{H}(\omega_0)\right]^{-1}\mathbf{H}(\omega_0)^{\mathrm{T}}\mathbf{y}. \end{aligned} \tag{3.371}$$

For the second stage of optimization, we then have

$$\hat{\omega}_0 = \arg\min_{\omega_0} J_1(\omega_0) \tag{3.372}$$

where

$$\begin{aligned} J_1(\omega_0) &= J(\hat{\alpha}_1(\omega_0), \hat{\alpha}_2(\omega_0), \omega_0) \\ &= \|\mathbf{y} - \mathbf{H}(\omega_0)\hat{\boldsymbol{\alpha}}(\omega_0)\|^2 \\ &= \mathbf{y}^{\mathrm{T}}\left[\mathbf{I} - \mathbf{H}(\omega_0)\left[\mathbf{H}(\omega_0)^{\mathrm{T}}\mathbf{H}(\omega_0)\right]^{-1}\mathbf{H}(\omega_0)^{\mathrm{T}}\right]\mathbf{y}. \end{aligned} \tag{3.373}$$

Thus, from (3.373), we obtain

$$\hat{\omega}_0 = \arg\min_{\omega_0} J_1(\omega_0) = \arg\max_{\omega_0} J_2(\omega_0) \tag{3.374}$$

where

$$J_2(\omega_0) = \mathbf{y}^{\mathrm{T}}\mathbf{H}(\omega_0)\left[\mathbf{H}(\omega_0)^{\mathrm{T}}\mathbf{H}(\omega_0)\right]^{-1}\mathbf{H}(\omega_0)^{\mathrm{T}}\mathbf{y}. \tag{3.375}$$

Now

$$\mathbf{H}(\omega_0)^{\mathrm{T}}\mathbf{H}(\omega_0) = \begin{bmatrix} \mathbf{c}(\omega_0)^{\mathrm{T}}\mathbf{c}(\omega_0) & \mathbf{c}(\omega_0)^{\mathrm{T}}\mathbf{s}(\omega_0) \\ \mathbf{c}(\omega_0)^{\mathrm{T}}\mathbf{s}(\omega_0) & \mathbf{s}(\omega_0)^{\mathrm{T}}\mathbf{s}(\omega_0) \end{bmatrix} \tag{3.376}$$

and

$$\mathbf{H}(\omega_0)^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} \mathbf{c}(\omega_0)^{\mathrm{T}}\mathbf{y} \\ \mathbf{s}(\omega_0)^{\mathrm{T}}\mathbf{y} \end{bmatrix}. \tag{3.377}$$

But in the large N regime (3.295) we have, again using order notation,

$$\mathbf{c}(\omega_0)^{\mathrm{T}}\mathbf{c}(\omega_0) = \sum_{n=0}^{N-1} \cos^2 \omega_0 n = \frac{N}{2} + o(N) \tag{3.378a}$$

$$\mathbf{c}(\omega_0)^{\mathrm{T}}\mathbf{s}(\omega_0) = \sum_{n=0}^{N-1} \cos \omega_0 n \, \sin \omega_0 n = o(N) \tag{3.378b}$$

$$\mathbf{s}(\omega_0)^{\mathrm{T}}\mathbf{s}(\omega_0) = \sum_{n=0}^{N-1} \sin^2 \omega_0 n = \frac{N}{2} + o(N). \tag{3.378c}$$

Hence, using (3.377) and (3.376) with (3.378) in (3.375) we obtain

$$\begin{aligned} J_2(\omega_0) &= \begin{bmatrix} \mathbf{c}(\omega_0)^{\mathrm{T}}\mathbf{y} & \mathbf{s}(\omega_0)^{\mathrm{T}}\mathbf{y} \end{bmatrix} \begin{bmatrix} N/2 + o(N) & o(N) \\ o(N) & N/2 + o(N) \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{c}(\omega_0)^{\mathrm{T}}\mathbf{y} \\ \mathbf{s}(\omega_0)^{\mathrm{T}}\mathbf{y} \end{bmatrix} \\ &\approx \frac{2}{N} \left[(\mathbf{c}(\omega_0)^{\mathrm{T}}\mathbf{y})^2 + (\mathbf{s}(\omega_0)^{\mathrm{T}}\mathbf{y})^2 \right] \\ &= \frac{2}{N} \left[\left(\sum_{n=0}^{N-1} y[n] \cos \omega_0 n \right)^2 + \left(\sum_{n=0}^{N-1} y[n] \sin(\omega_0 n) \right)^2 \right] \\ &= 2 \, |Y_N(e^{j\omega_0})|^2 \end{aligned} \tag{3.379}$$

where $Y_N(e^{j\omega})$ is as given in (3.299). Thus, using (3.379) in (3.374), we conclude that for N in the regime (3.295), the ML estimate $\hat{\omega}_0$ effectively corresponds to the location of the peak of the periodogram of the data, i.e., (3.300).

Given $\hat{\omega}_0$, the remaining parameters follow almost immediately. In particular, since

$$\begin{aligned} \hat{\boldsymbol{\alpha}}(\hat{\omega}_0) &= (\mathbf{H}(\hat{\omega}_0)^{\mathrm{T}}\mathbf{H}(\hat{\omega}_0))^{-1}\mathbf{H}(\hat{\omega}_0)^{\mathrm{T}}\mathbf{y} \\ &\approx \frac{2}{N} \begin{bmatrix} \mathbf{c}(\hat{\omega}_0)^{\mathrm{T}}\mathbf{y} \\ \mathbf{s}(\hat{\omega}_0)^{\mathrm{T}}\mathbf{y} \end{bmatrix} \\ &= \frac{2}{\sqrt{N}} \begin{bmatrix} \mathrm{Re}\left\{ Y_N(e^{j\hat{\omega}_0}) \right\} \\ \mathrm{Im}\left\{ Y_N(e^{j\hat{\omega}_0}) \right\} \end{bmatrix}, \end{aligned} \tag{3.380}$$

we have

$$\hat{A}^2 = \hat{\alpha}_1^2(\hat{\omega}_0) + \hat{\alpha}_2^2(\hat{\omega}_0) = \hat{\boldsymbol{\alpha}}(\hat{\omega}_0)^{\mathrm{T}}\hat{\boldsymbol{\alpha}}(\hat{\omega}_0) = \frac{4}{N} \left| Y_N(e^{j\hat{\omega}_0}) \right|^2 \tag{3.381}$$

corresponding to (3.301), and

$$\hat{\Theta} = -\tan^{-1}\left(\frac{\hat{\alpha}_2(\hat{\omega}_0)}{\hat{\alpha}_1(\hat{\omega}_0)} \right) = -\tan^{-1}\left(\frac{\mathrm{Im}\left\{ Y_N(e^{j\hat{\omega}_0}) \right\}}{\mathrm{Re}\left\{ Y_N(e^{j\hat{\omega}_0}) \right\}} \right) \tag{3.382}$$

corresponding to (3.302).