EE5137 : Stochastic Processes (Spring 2021)
Estimation Theory

Vincent Y. F. Tan

April 16, 2021

We will cover this material in Lectures 12 and 13.

1 Bayesian Parameter Estimation

The problem of hypothesis testing, which we have covered can be viewed as one of making decisions about the value of a binary variable taking values in the alphabet $\mathcal{H} = \{H_0, H_1\}$. When the unknown to be estimated is not binary, but a general discrete- or continuous-valued vector $\mathbf{x}$ taking on values in an alphabet $\mathcal{X}$, this is called *estimation theory*. As before, our observations $\mathbf{Y}$ are random and take values in some alphabet $\mathcal{Y}$.

In the Bayesian world, there is an a priori distribution $p_{\mathbf{X}}(\cdot)$ for the unknown parameter $\mathbf{X}$. This represents our belief about $\mathbf{X}$ prior to any observation of the measurement $\mathbf{Y}$. In addition, the observation model takes the form of a conditional distribution $f_{\mathbf{Y}|\mathbf{X}}(\cdot|\mathbf{x})$, which fully specifies the way in which $\mathbf{Y}$ contains information about $\mathbf{X}$.

Example 1. Suppose that $\mathbf{Y}$ is a noise-corrupted measurement of some function of $\mathbf{X}$, viz.

$$\mathbf{Y} = \mathbf{h}(\mathbf{X}) + \mathbf{W} \tag{1}$$

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

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = f_{\mathbf{W}}(\mathbf{Y} - \mathbf{h}(\mathbf{X})). \tag{2}$$

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

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \mathbf{A}\mathbf{x}, \mathbf{\Lambda}). \tag{3}$$

The observation model $f_{\mathbf{Y}|\mathbf{X}}(\cdot|\mathbf{x})$ and prior distribution $f_{\mathbf{X}}(\mathbf{x})$ together constitute a full statistical characterization of $\mathbf{X}$ and $\mathbf{Y}$. In particular, the joint distribution is given by their product, i.e.,

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}). \tag{4}$$

As in hypothesis testing, for a given observation $\mathbf{Y}$, a complete characterization of our knowledge of the parameter $\mathbf{X}$ is given by the *posterior* distribution for $\mathbf{X}$, i.e.,

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{\int_{\mathcal{X}} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}') f_{\mathbf{X}}(\mathbf{x}') \, \mathrm{d}\mathbf{x}'} \tag{5}$$

In general, to find good estimator for $\mathbf{X}$, we need some measure of goodness of candidate estimators. In other words, we need a suitable performance criterion with respect to which we optimize our choice of estimator. In the Bayesian formulation, 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.,

$$\hat{\mathbf{x}}(\cdot) := \arg\min_{\mathbf{g}(\cdot)} \mathbb{E}[C(\mathbf{X}, \mathbf{g}(\mathbf{Y}))]. \tag{6}$$

Note that the expectation in (6) 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.

As in Bayesian hypothesis testing, solving for the optimum function $\hat{\mathbf{x}}(\cdot)$ in (6) can, in fact, be accomplished on a *pointwise* basis. To show this, write the objective function in (6) as

$$\mathbb{E}[C(\mathbf{X}, \mathbf{g}(\mathbf{Y}))] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} C(\mathbf{x}, \mathbf{g}(\mathbf{y})) f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \tag{7}$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} C(\mathbf{x}, \mathbf{g}(\mathbf{y})) f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x} \right] f_{\mathbf{Y}}(\mathbf{y}) \, \mathrm{d}\mathbf{y} \tag{8}$$

Since $f_{\mathbf{Y}}(\mathbf{y}) \geq 0$, we clearly minimize (8) if we choose $\hat{\mathbf{x}}(\mathbf{y})$ to minimize the term in brackets for each individual value of $\mathbf{y}$, i.e.,

$$\hat{\mathbf{x}}(\mathbf{y}) = \arg\min_{\mathbf{a}} \int_{-\infty}^{\infty} C(\mathbf{x}, \mathbf{a}) f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x}. \tag{9}$$

In the following, we will specialize the above rule to various cost functions.

1.1 Minimum Absolute-Error Estimation

Now we consider the case in which the parameter that we want to estimate is scalar-valued. 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

$$C(a, \hat{a}) = |a - \hat{a}|. \tag{10}$$

Theorem 1. *The MAE estimate is the median of the belief $f_{X|\mathbf{Y}}(x|\mathbf{y})$.*

Proof. Substituting (10) into (9), we obtain

$$\hat{x}_{\mathrm{MAE}}(\mathbf{y}) = \arg\min_{a} \int_{-\infty}^{\infty} |x - a| f_{X|\mathbf{Y}}(x|\mathbf{y}) \, \mathrm{d}x \tag{11}$$

$$= \arg\min_{a} \left\{ \int_{-\infty}^{a} (a - x) f_{X|\mathbf{Y}}(x|\mathbf{y}) \, \mathrm{d}x + \int_{a}^{\infty} (x - a) f_{X|\mathbf{Y}}(x|\mathbf{y}) \, \mathrm{d}x \right\}. \tag{12}$$

Differentiating the quantity inside braces in (12) with respect to a gives, via Leibnitz' rule,[1] the condition

$$\left[\int_{-\infty}^{a} f_{X|\mathbf{Y}}(x|\mathbf{y}) \, \mathrm{d}x - \int_{a}^{\infty} f_{X|\mathbf{Y}}(x|\mathbf{y}) \, \mathrm{d}x \right]_{a=\hat{x}_{\mathrm{MAE}}(\mathbf{y})} = 0. \tag{13}$$

Rewriting (13), we obtain

$$\int_{-\infty}^{\hat{x}_{\mathrm{MAE}}(\mathbf{y})} f_{X|\mathbf{Y}}(x|\mathbf{y}) \, \mathrm{d}x = \int_{\hat{x}_{\mathrm{MAE}}(\mathbf{y})}^{\infty} f_{X|\mathbf{Y}}(x|\mathbf{y}) \, \mathrm{d}x = \frac{1}{2}. \tag{14}$$

From (14) we see that $\hat{x}_{\mathrm{MAE}}(\mathbf{y})$ is the threshold in x of the belief $f_{X|\mathbf{Y}}(x|\mathbf{y})$ for which half the probability is located above the threshold and, hence, half is also below the threshold. Hence, the MAE estimator for X given $\mathbf{Y} = \mathbf{y}$ is the median of the belief. □

[1] Leibnitz' rule states that

$$\frac{\mathrm{d}}{\mathrm{d}a} \int_{b(a)}^{c(a)} g(a, x) \, \mathrm{d}x = g(a, c(a)) \frac{\mathrm{d}}{\mathrm{d}a} c(a) - g(a, b(a)) \frac{\mathrm{d}}{\mathrm{d}a} b(a) + \int_{b(a)}^{c(a)} \frac{\partial}{\partial a} g(a, x) \, \mathrm{d}x.$$

Example 2. Suppose we have the posterior density

$$f_{X|Y}(x|y) = \begin{cases} 1/(3y) & 0 < x < y \\ 2/(3y) & y < x < 2y \\ 0 & \text{otherwise} \end{cases}. \tag{15}$$

Then

$$\hat{x}_{\text{MAE}}(y) = (1 + \Delta)y \tag{16}$$

for an appropriate choice of $\Delta > 0$. To solve for Δ, we use (14) to obtain

$$\frac{1}{3y} \cdot y + \frac{2}{3y} \cdot y\Delta = \frac{1}{2}, \tag{17}$$

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

We also note that the median of a density is not necessarily unique, as the following example illustrates.

Example 3. Let $y > 0$. Suppose

$$f_{X|Y}(x|y) = \begin{cases} 1/(2y) & 0 < x < y \text{ and } 2y < x < 3y \\ 0 & \text{otherwise} \end{cases}. \tag{18}$$

Then the median of (18) is any number between y and $2y$; hence, the MAE estimators for X given $Y = y$ are all of the form

$$\hat{x}_{\text{MAE}}(y) = \alpha \tag{19}$$

where α is any constant satisfying $y \leq \alpha \leq 2y$.

1.2 Maximum A Posterior Estimation

As an alternative to that considered in the previous section, consider the minimum uniform cost (MUC) cost criterion, whereby

$$C(a, \hat{a}) = \begin{cases} 1 & |a - \hat{a}| > \epsilon \\ 0 & \text{otherwise} \end{cases}. \tag{20}$$

which uniformly penalizes all estimation errors with magnitude bigger than ϵ.

Theorem 2. *In the limit $\epsilon \to 0$, the MUC estimate is the mode of the belief $f_{X|\mathbf{Y}}(x|\mathbf{y})$ i.e., it is the maximum a posteriori (MAP) estimate*

$$\hat{x}_{\text{MAP}}(\mathbf{y}) = \arg\max_a f_{X|\mathbf{Y}}(a|\mathbf{y}). \tag{21}$$

From this claim, we see that the MAP estimator can be viewed as resulting from a Bayes' cost formulation in which all errors are, in an appropriate sense, equally bad.

Proof. Substituting (20) into (9), we see that

$$\hat{x}_{\text{MUC}}(\mathbf{y}) = \arg\min_a \left[1 - \int_{a-\epsilon}^{a+\epsilon} f_{X|\mathbf{Y}}(x|\mathbf{y})\, \mathrm{d}x\right] \tag{22}$$

$$= \arg\max_a \int_{a-\epsilon}^{a+\epsilon} f_{X|\mathbf{Y}}(x|\mathbf{y})\, \mathrm{d}x. \tag{23}$$

Note from (23) that the minimum uniform cost estimator $\hat{x}_{\text{MUC}}(\mathbf{y})$ corresponds to the value of a that makes the probability $\Pr(|X - \hat{x}_{\text{MUC}}(\mathbf{Y})| < \epsilon \mid \mathbf{Y} = \mathbf{y})$ as large as possible. This means finding the interval of length 2ϵ where the posterior density $f_{X|\mathbf{Y}}(x|\mathbf{y})$ is most concentrated.

If we carry this perspective a little further, we see that if we let ϵ get sufficiently small then the $\hat{x}_{\text{MUC}}(\mathbf{y})$ approaches the point corresponding to the peak of the posterior density, i.e., the MAP estimate (21)

$$\lim_{\epsilon \to 0} \hat{x}_{\text{MUC}}(\mathbf{y}) = \arg\max_a f_{X|\mathbf{Y}}(a|\mathbf{y}). \tag{24}$$

□

1.3 Bayes' Least-Squares Estimation

Perhaps the most popular popular Bayesian estimator is based on a quadratic cost criterion, which we now develop. Specifically, we consider the mean-square error (MSE) cost criterion

$$C(\mathbf{a}, \hat{\mathbf{a}}) = \|\mathbf{a} - \hat{\mathbf{a}}\|^2 = (\mathbf{a} - \hat{\mathbf{a}})^T(\mathbf{a} - \hat{\mathbf{a}}) = \sum_{i=1}^{N}(a_i - \hat{a}_i)^2, \tag{25}$$

from which we obtain what is termed 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 $\hat{x}_{\text{MMSE}}(\cdot)$.

Theorem 3. *The BLS estimate is the mean of the belief $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$.*

Proof. Substituting (25) into (9) yields

$$\hat{x}_{\text{BLS}}(\mathbf{y}) = \arg\min_{\mathbf{a}} \int_{-\infty}^{\infty} (\mathbf{x} - \mathbf{a})^T(\mathbf{x} - \mathbf{a}) f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})\, \mathrm{d}\mathbf{x}. \tag{26}$$

Let us begin with the simpler case of scalar estimation, for which (26) becomes

$$\hat{x}_{\text{BLS}}(\mathbf{y}) = \arg\min_{a} \int_{-\infty}^{\infty} (x - a)^2 f_{X|\mathbf{Y}}(x|\mathbf{y})\, \mathrm{d}x. \tag{27}$$

As we did in the case of MAE estimation, we can perform the minimization in (27) by differentiating with respect to a and setting the result to zero to find the local extrema. Differentiating the integral in (27) we obtain

$$\frac{\partial}{\partial a}\left[\int_{-\infty}^{\infty} (x - a)^2 f_{X|\mathbf{Y}}(x|\mathbf{y})\, \mathrm{d}x\right] = \int_{-\infty}^{\infty} \frac{\partial}{\partial a}(x - a)^2 f_{X|\mathbf{Y}}(x|\mathbf{y})\, \mathrm{d}x = -2\int_{-\infty}^{\infty} (x - a) f_{X|\mathbf{Y}}(x|\mathbf{y})\, \mathrm{d}x. \tag{28}$$

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

$$\left[\int_{-\infty}^{\infty} (x - a) f_{X|\mathbf{Y}}(x|\mathbf{y})\, \mathrm{d}x\right]_{a=\hat{x}_{\text{BLS}}(\mathbf{y})} = \int_{-\infty}^{\infty} x f_{X|\mathbf{Y}}(x|\mathbf{y})\, \mathrm{d}x - \int_{-\infty}^{\infty} \hat{x}_{\text{BLS}}(\mathbf{y}) f_{X|\mathbf{Y}}(x|\mathbf{y})\, \mathrm{d}x \tag{29}$$

$$= \mathbb{E}[X|\mathbf{Y} = \mathbf{y}] - \hat{x}_{\text{BLS}}(\mathbf{y}) = 0. \tag{30}$$

Hence,

$$\hat{x}_{\text{BLS}}(\mathbf{y}) = \mathbb{E}[X|\mathbf{Y} = \mathbf{y}]. \tag{31}$$

That is, the BLS or MMSE estimate of X given $\mathbf{Y} = \mathbf{y}$ is the mean of the belief $f_{X|\mathbf{Y}}(x|\mathbf{y})$.

When $\mathbf{x}$ is a vector, it suffices to note that since the cost criterion (25) is additive, the minimum is achieved by minimizing the mean-square estimation error in each scalar component. Hence, we obtain

$$\hat{\mathbf{x}}_{\text{BLS}}(\mathbf{Y}) = \mathbb{E}[\mathbf{X}|\mathbf{Y}]. \tag{32}$$

from which we see that the BLS estimate of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$ is in general the mean of the belief $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$. □

We will focus on the scalar case for simplicity, though there is a lot of theory for the vector case.

Example 4. Suppose X and W are independent random variables that are both uniformly distributed over the range $[-1, 1]$, and let

$$Y = \text{sgn}(X) + W. \tag{33}$$

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

$$f_{Y|X}(y|x) = \begin{cases} 1/2 & 0 < y < 2 \\ 0 & \text{otherwise} \end{cases}. \tag{34}$$

while for $x < 0$, we have

$$f_{Y|X}(y|x) = \begin{cases} 1/2 & -2 < y < 0 \\ 0 & \text{otherwise} \end{cases}. \tag{35}$$

Hence, the joint density is

$$f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x) = \begin{cases} 1/4 & 0 < x < 1 \text{ and } 0 < y < 2 \\ 1/4 & -1 < x < 0 \text{ and } -2 < y < 0 \\ 0 & \text{otherwise} \end{cases} \tag{36}$$

and so for $y > 0$ we have the posterior

$$f_{X|Y}(x|y) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

and for $y < 0$, we have

$$f_{X|Y}(x|y) = \begin{cases} 1 & -1 < x < 0 \\ 0 & \text{otherwise} \end{cases} \tag{38}$$

Thus, from (37) and (37), we have

$$\hat{x}_{\text{BLS}}(y) = \mathbb{E}[X|Y=y] = \frac{1}{2}\text{sgn}(y) = \begin{cases} 1/2 & y > 0 \\ -1/2 & y < 0 \end{cases}. \tag{39}$$

Also note that

$$\mathsf{Var}(X|Y=y) = \frac{1}{12}. \tag{40}$$

2 NonBayesian Parameter Estimation

In many estimation problems, it may be unnatural to assign a prior distribution to the latent variable of interest, and thus the Bayesian framework cannot be used. In such cases, an alternative is to treat the variable as deterministic, but unknown. In such cases, the observation model is not a distribution for $\mathbf{Y}$ conditioned on the latent variable, but rather a distribution for $\mathbf{Y}$ that is *parameterized* by this variable.

As an example, suppose we have a sequence of independent identically distributed Gaussian random variables $Y_1, \ldots, Y_n$, where the mean μ and variance σ^2 that parameterize the density are unknown. In this section we describe some classical approaches to the problem of developing good estimators for such nonrandom parameters.

In our treatment, we use $\mathbf{x}$ to denote the vector of parameters we seek to estimate, and write the density for the vector of observations $\mathbf{Y}$ as $f_{\mathbf{Y}}(\mathbf{y};\mathbf{x})$ so as to make the parameterization explicit. In addition, we will use $\mu_{\mathbf{Y}}(\mathbf{x})$ and $\Lambda_{\mathbf{Y}}(\mathbf{x})$ to denote, respectively, the mean vector and covariance matrix of $\mathbf{Y}$, again to make the parameterization on $\mathbf{x}$ explicit, i.e.,

$$\mu_{\mathbf{Y}}(\mathbf{x}) = \mathbb{E}_{\mathbf{x}}[\mathbf{Y}] = \int_{-\infty}^{\infty} \mathbf{y} f_{\mathbf{Y}}(\mathbf{y};\mathbf{x})\, \mathrm{d}\mathbf{y}, \quad \text{and} \tag{41}$$

$$\Lambda_{\mathbf{Y}}(\mathbf{x}) = \mathsf{Cov}_{\mathbf{x}}(\mathbf{Y}) = \int_{-\infty}^{\infty} (\mathbf{y} - \mu_{\mathbf{Y}}(\mathbf{x}))(\mathbf{y} - \mu_{\mathbf{Y}}(\mathbf{x}))^T f_{\mathbf{Y}}(\mathbf{y};\mathbf{x})\, \mathrm{d}\mathbf{y}. \tag{42}$$

Throughout this section, we focus on mean-square estimation error as the performance measure of interest. We begin by noting that the Bayesian (least-squares) framework developed earlier cannot be immediately adapted to handle nonrandom parameter estimation. To see this, consider the case of a scalar parameter x. If we attempt to construct a minimum mean-square error estimate $\hat{x}(\mathbf{Y})$ via

$$\hat{x}(\cdot) = \arg\min_{g(\cdot)} \mathbb{E}[(x - g(\mathbf{Y}))^2] \tag{43}$$

we encounter a difficulty. In particular, since the expectation in (43) is over $\mathbf{Y}$ alone (since x is deterministic), we immediately obtain that the right-hand side of (43) is minimized by choosing $\hat{x}(\mathbf{Y}) = x$, and hence the optimum estimator according to (43) depends on the very parameter we're trying to estimate!

This observation reveals an important insight on nonrandom parameter estimation: in any meaningful formulation of such problems, we need to explicitly restrict our search to estimators that don't depend explicitly on the parameters we're trying to estimate.

Definition 1. *An estimator is* valid *if it does not depend explicitly on the parameters being estimated.*

In the sequel, we describe some traditional approaches to finding valid estimators that yield good mean-square error performance.

2.1 Bias and Error Covariance

Given an estimator $\hat{\mathbf{x}}(\cdot)$, we can define two important performance metrics as follows. Using

$$\mathbf{E} = \mathbf{e}(\mathbf{Y}) := \hat{\mathbf{x}}(\mathbf{Y}) - \mathbf{x} = \hat{\mathbf{X}} - \mathbf{x} \tag{44}$$

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

$$\mathbf{b}_{\hat{\mathbf{X}}}(\mathbf{x}) = \mathbb{E}[\mathbf{e}(\mathbf{Y})] = \mathbb{E}[\hat{\mathbf{x}}(\mathbf{Y}) - \mathbf{x}] = \left[\int_{-\infty}^{\infty} \hat{\mathbf{x}}(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}; \mathbf{x})\, \mathrm{d}\mathbf{y}\right] - \mathbf{x}. \tag{45}$$

Likewise, we express the error covariance as

$$\mathbf{\Lambda}_{\mathbf{E}}(\mathbf{x}) = \mathbb{E}\left[(\mathbf{e}(\mathbf{Y}) - \mathbf{b}_{\hat{\mathbf{X}}}(\mathbf{x}))(\mathbf{e}(\mathbf{Y}) - \mathbf{b}_{\hat{\mathbf{X}}}(\mathbf{x}))^T\right], \tag{46}$$

where, again, the expectation is with respect to $\mathbf{Y}$. Not surprisingly, both the bias (45) and error covariance (46) are, in general, functions of the parameter $\mathbf{x}$.

Definition 2. *An estimator $\hat{\mathbf{x}}(\cdot)$ for a nonrandom parameter $\mathbf{x}$ is* unbiased *if $\mathbf{b}_{\hat{\mathbf{X}}}(\mathbf{x}) = \mathbf{0}$ for all possible values of $\mathbf{x}$.*

This is the notion underlying well-known *minimum-variance unbiased (MVU) estimators*, which we discuss next.

As one final comment before proceeding, 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.,

$$\mathbf{\Lambda}_{\mathbf{E}}(\mathbf{x}) = \mathbf{\Lambda}_{\hat{\mathbf{X}}}(\mathbf{x}) = \mathbb{E}\left[(\hat{\mathbf{x}}(\mathbf{Y}) - \mathbb{E}[\hat{\mathbf{x}}(\mathbf{Y})])(\hat{\mathbf{x}}(\mathbf{Y}) - \mathbb{E}[\hat{\mathbf{x}}(\mathbf{Y})])^T\right]. \tag{47}$$

To see this, note that

$$\begin{aligned}
\mathbf{\Lambda}_{\mathbf{E}}(\mathbf{x}) &= \mathbb{E}\left[(\mathbf{e}(\mathbf{Y}) - \mathbf{b}_{\hat{\mathbf{X}}}(\mathbf{x}))(\mathbf{e}(\mathbf{Y}) - \mathbf{b}_{\hat{\mathbf{X}}}(\mathbf{x}))^T\right] && (48)\\
&= \mathbb{E}\left[\big((\hat{\mathbf{x}}(\mathbf{Y}) - \mathbf{x}) - (\mathbb{E}[\hat{\mathbf{x}}(\mathbf{Y})] - \mathbf{x})\big)\big((\hat{\mathbf{x}}(\mathbf{Y}) - \mathbf{x}) - (\mathbb{E}[\hat{\mathbf{x}}(\mathbf{Y})] - \mathbf{x})\big)^T\right] && (49)\\
&= \mathbb{E}\left[(\hat{\mathbf{x}}(\mathbf{Y}) - \mathbb{E}[\hat{\mathbf{x}}(\mathbf{Y})])(\hat{\mathbf{x}}(\mathbf{Y}) - \mathbb{E}[\hat{\mathbf{x}}(\mathbf{Y})])^T\right] && (50)\\
&= \mathbf{\Lambda}_{\hat{\mathbf{X}}}(\mathbf{x}). && (51)
\end{aligned}$$

In the remainder of our discussion, we restrict our attention to the case where the parameter to be estimated is a scalar x. As in the case of Bayesian estimation, vector parameter extensions with the mean-square error criterion can be constructed in a component-wise manner.

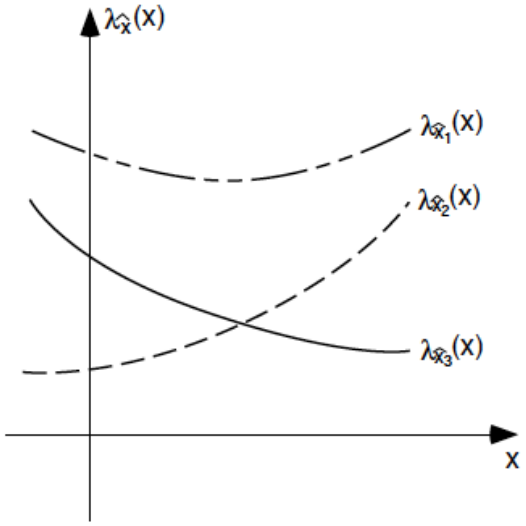

Figure 1: The variances of three hypothetical unbiased estimators.

2.2 Minimum-Variance Unbiased Estimators

We begin by defining the admissible set of estimators.

Definition 3. *An* admissible *estimator is one that is both valid (i.e., does not depend on x) and unbiased. We use*

$$\mathcal{A} := \{\hat{x}(\cdot) : \hat{x}(\cdot) \text{ is valid and } b_{\hat{X}}(x) = 0, \ \forall\, x\} \tag{52}$$

to denote the set of admissible estimators.

In turn, when it exists, a minimum-variance unbiased (MVU) estimator for x is defined to be the admissible estimator with the smallest variance, i.e.,

$$\hat{x}_{\mathrm{MVU}}(\cdot) := \arg\min_{\hat{X} \in \mathcal{A}} \lambda_{\hat{X}}(x), \quad \forall\, x. \tag{53}$$

Several observations regarding (53) are worth emphasizing. The first is that $\hat{x}_{\mathrm{MVU}}(\cdot)$ 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 uniformly smaller variance than all the others, i.e., for all values of the parameter x. Suppose for example that $\mathcal{A}$ consists of three hypothetical 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. 1. 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 $\hat{x}_{\mathrm{MVU}}(\cdot)$ 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, there are cases in which such estimators can be computed, as we'll discuss. To this end, it is sometimes useful to exploit a bound on $\hat{x}_{\mathrm{MVU}}(\cdot)$ in the pursuit of MVU estimators. A celebrated bound for this purpose is the Cramér-Rao bound, as we develop next.

2.3 The Cramér-Rao Bound

When it exists, the Cramér-Rao bound gives a lower bound on the variance of any admissible estimator $\hat{x}(\cdot)$ for x. In particular, we have the following.

Theorem 4. *Provided $f_{\mathbf{Y}}(\mathbf{y};x)$ satisfies the regularity condition*

$$\mathbb{E}\left[\frac{\partial}{\partial x}\log f_{\mathbf{Y}}(\mathbf{y};x)\right]=0,\quad \forall\, x \tag{54}$$

we have for any $\hat{x}(\cdot)$ satisfying Definition 3,

$$\lambda_{\hat{x}}(x)=\mathsf{Var}_x(\hat{x}(\mathbf{Y}))\geq\frac{1}{J_{\mathbf{Y}}(x)} \tag{55}$$

where

$$J_{\mathbf{Y}}(x)=\mathbb{E}\left[\left(\frac{\partial}{\partial x}\log f_{\mathbf{Y}}(\mathbf{y};x)\right)^2\right] \tag{56}$$

is referred to as the Fisher information *in* $\mathbf{Y}$ *about* x*.*

Proof. To derive the Cramér-Rao bound (55), we begin by recalling that for unbiased estimators the error

$$e(\mathbf{y})=\hat{x}(\mathbf{y})-x \tag{57}$$

has zero mean, i.e.,

$$\mathbb{E}[e(\mathbf{Y})]=0, \tag{58}$$

and variance

$$\mathsf{Var}(e(\mathbf{Y}))=\mathbb{E}[e^2(\mathbf{Y})]=\lambda_{\hat{x}}(x). \tag{59}$$

Next we define what is known as the *score function*

$$h(\mathbf{y})=\frac{\partial}{\partial x}\log f_{\mathbf{Y}}(\mathbf{y};x) \tag{60}$$

and note that using the identity

$$\frac{\partial}{\partial x}\log f_{\mathbf{Y}}(\mathbf{y};x)=\frac{1}{f_{\mathbf{Y}}(\mathbf{y};x)}\cdot\frac{\partial}{\partial x}f_{\mathbf{Y}}(\mathbf{y};x), \tag{61}$$

we get that $h(\mathbf{Y})$ has zero mean:

$$\begin{aligned}\mathbb{E}[h(\mathbf{Y})]&=\mathbb{E}\left[\frac{1}{f_{\mathbf{Y}}(\mathbf{Y};x)}\cdot\frac{\partial}{\partial x}f_{\mathbf{Y}}(\mathbf{Y};x)\right] &(62)\\ &=\int_{-\infty}^{\infty}\frac{\partial}{\partial x}f_{\mathbf{Y}}(\mathbf{y};x)\,\mathrm{d}\mathbf{y} &(63)\\ &=\frac{\partial}{\partial x}\int_{-\infty}^{\infty}f_{\mathbf{Y}}(\mathbf{y};x)\,\mathrm{d}\mathbf{y} &(64)\\ &=\frac{\partial}{\partial x}1=0, &(65)\end{aligned}$$

and in turn the variance

$$\mathsf{Var}(h(\mathbf{Y}))=\mathbb{E}[h^2(\mathbf{Y})]=J_{\mathbf{Y}}(x). \tag{66}$$

Finally by using the identity in (61), the covariance between $e(\mathbf{Y})$ and $h(\mathbf{Y})$ is given

$$\begin{aligned}\mathsf{Cov}(e(\mathbf{Y}),h(\mathbf{Y}))&=\mathbb{E}[e(\mathbf{Y})h(\mathbf{Y})] &(67)\\ &=\int_{-\infty}^{\infty}(\hat{x}(\mathbf{y})-x)\frac{\partial}{\partial x}f_{\mathbf{Y}}(\mathbf{y};x)\,\mathrm{d}\mathbf{y} &(68)\\ &=\left[\frac{\partial}{\partial x}\int_{-\infty}^{\infty}\hat{x}(\mathbf{y})f_{\mathbf{Y}}(\mathbf{y};x)\,\mathrm{d}\mathbf{y}\right]-\left[x\frac{\partial}{\partial x}\int_{-\infty}^{\infty}f_{\mathbf{Y}}(\mathbf{y};x)\,\mathrm{d}\mathbf{y}\right] &(69)\\ &=1-0=1, &(70)\end{aligned}$$

where the last step follows from the fact that $\hat{x}(\mathbf{Y})$ is unbiased and so $\mathbb{E}[\hat{x}(\mathbf{Y})] = x$. Now recall that correlation coefficients have at most unit magnitude; this is the Cauchy-Schwarz inequality for random variables. Indeed, if Z and W are zero-mean random variables and we define $K := (Z - \alpha W)^2$, we have

$$0 \leq \mathbb{E}[K] = \mathbb{E}[(Z - \alpha W)^2] = \mathbb{E}[Z^2] - 2\alpha\mathbb{E}[ZW] + \alpha^2\mathbb{E}[W^2]. \tag{71}$$

Now, we let $g(\alpha) = \mathbb{E}[Z^2] - 2\alpha\mathbb{E}[ZW] + \alpha^2\mathbb{E}[W^2]$. Then we know that $g(\alpha) \geq 0$ for all $\alpha \in \mathbb{R}$. Moreover if $g(\alpha) = 0$ for some α, then we have $\mathbb{E}[K] = \mathbb{E}[(Z - \alpha W)^2] = 0$, which essentially means that $Z = \alpha W$ with probability one. To prove the Cauchy-Schwarz inequality, let $\alpha = \mathbb{E}[ZW]/\mathbb{E}[W^2]$, then,

$$0 \leq \mathbb{E}[Z^2] - 2\frac{\mathbb{E}[ZW]}{\mathbb{E}[W^2]}\mathbb{E}[ZW] + \left(\frac{\mathbb{E}[ZW]}{\mathbb{E}[W^2]}\right)^2 \mathbb{E}[W^2] = \mathbb{E}[Z^2] - \frac{(\mathbb{E}[ZW])^2}{\mathbb{E}[W^2]}. \tag{72}$$

This means that

$$\mathbb{E}[ZW] \leq \sqrt{\mathbb{E}[Z^2]\mathbb{E}[W^2]}, \tag{73}$$

completing the proof of the Cauchy-Schwarz inequality for random variables. We note that equality holds if and only if $Z = \alpha W$, which means that one random variable must be a scalar multiple of the other.

As a result,

$$\rho^2_{e,h} = \frac{\mathsf{Cov}(e(\mathbf{Y}), h(\mathbf{Y}))^2}{\mathsf{Var}(e(\mathbf{Y}))\,\mathsf{Var}(h(\mathbf{Y}))} \leq 1. \tag{74}$$

Finally, substituting (59), (66) and (70) into (74), we obtain the Cramér-Rao bound. □

In addition, we note that Fisher information is often convenient to express as follows.

Corollary 5. *The Fisher information* (56) *can be equivalently expressed in the form*

$$J_{\mathbf{Y}}(\mathbf{x}) = -\mathbb{E}\left[\frac{\partial^2}{\partial x^2}\log f_{\mathbf{Y}}(\mathbf{y}; x)\right]. \tag{75}$$

Proof. To verify (75), we begin by observing

$$\int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y}; x)\,\mathrm{d}\mathbf{y} = 1. \tag{76}$$

Differentiating this with respect to x and using the identity in (61) yields

$$\int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y}; x)\frac{\partial}{\partial x}\log f_{\mathbf{Y}}(\mathbf{y}; x)\,\mathrm{d}\mathbf{y} = 0. \tag{77}$$

Finally, differentiating this once more and using (61) yields

$$\int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y}; x)\left[\frac{\partial^2}{\partial x^2}\log f_{\mathbf{Y}}(\mathbf{y}; x)\right]\,\mathrm{d}\mathbf{y} + \int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y}; x)\left[\frac{\partial}{\partial x}\log f_{\mathbf{Y}}(\mathbf{y}; x)\right]^2\,\mathrm{d}\mathbf{y} = 0. \tag{78}$$

□

A few remarks on the Cramér-Rao bound are in order.

1. We emphasize that the Fisher information cannot be computed in all problems, i.e., the regularity condition may not be satisfied, in which case no Cramér-Rao bound exists. For example, for densities such as

$$f_Y(y; x) = \begin{cases} 1 & x < y < x + 1 \\ 0 & \text{otherwise} \end{cases} \tag{79}$$

which are not strictly positive for all x and y, the logarithm in (56) doesn't exist and hence $J_Y(x)$ cannot be calculated.

2. The Fisher information (56) can be interpreted as a measure of curvature: it measures, on average, how "peaky" $\log f_{\mathbf{Y}}(\mathbf{y}; x)$ is as a function of x. As such, the larger $J_{\mathbf{Y}}(x)$, 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.

3. The regularity condition is a statement of the interchangeability of the order of integration and differentiation in (54).

4. Any estimator that satisfies the Cramér-Rao bound with equality must be an 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!

We illustrate the results with some examples.

Example 5. Consider the scalar Gaussian problem

$$Y = x + W, \tag{80}$$

where $W \sim \mathcal{N}(0, \sigma^2)$. To determine the Cramér-Rao bound, we first calculate

$$\log f_Y(y; x) = -\frac{1}{2\sigma^2}(x - Y)^2 - \frac{1}{2}\log(2\pi\sigma^2), \tag{81}$$

from which we obtain

$$\frac{\partial}{\partial x}\log f_Y(y; x) = -\frac{1}{\sigma^2}(x - Y) = \frac{1}{\sigma^2}W. \tag{82}$$

Hence, the Fisher information is

$$J_Y(x) = \frac{1}{\sigma^4}\mathbb{E}[W^2] = \frac{1}{\sigma^2}, \tag{83}$$

and, thus, the variance of any unbiased estimator $\hat{x}$ satisfies

$$\lambda_{\hat{x}}(x) \geq \sigma^2. \tag{84}$$

Moreover, we see that the smaller the variance σ^2, the sharper the peak of (82) is as a function of x.

2.4 Maximum Likelihood Estimation

Perhaps the most widely used estimators in practice are maximum likelihood estimators, defined as follows.

Definition 4. *The maximum likelihood*[2] *estimate* $\hat{x}_{\mathrm{ML}}(\mathbf{y})$ *for parameter* x *based on observations* $\mathbf{y}$ *is defined via*

$$\hat{x}_{\mathrm{ML}}(\mathbf{y}) = \arg\max_{x \in \mathcal{X}} f_{\mathbf{Y}}(\mathbf{y}; x). \tag{85}$$

Maximum likelihood (ML) estimators are particularly pleasing. In particular, if an estimator achieves the Cramér-Rao bound, it must be the ML estimator.

Definition 5. *An unbiased estimator is* efficient *if it satisfies the Cramér-Rao bound* (55) *with equality.*

From our derivation of the Cramér-Rao bound and in particular (74), we see that equality holds if and only if there exists some constant $k(x) > 0$ (that can only depend on x) such that

$$e(\mathbf{y}) = k(x)h(\mathbf{y}) \qquad \text{for all } \mathbf{y}. \tag{86}$$

Thus efficient estimators must have this property. Rearranging and noticing that $e(\mathbf{y}) = \hat{x}(\mathbf{y}) - x$, we see that

$$\hat{x}(\mathbf{y}) = x + k(x)\frac{\partial}{\partial x}\log f_{\mathbf{Y}}(\mathbf{y}; x). \tag{87}$$

[2]The function $f_{\mathbf{Y}}(\cdot;\cdot)$ can be viewed two different ones. In particular, while we refer to $f_{\mathbf{Y}}(\cdot; x)$ as a model for our data, we refer to $f_{\mathbf{Y}}(\mathbf{y};\cdot)$ as the likelihood function for the possible parameters.

Hence, an efficient estimator must be one for which the above estimator is valid, i.e., the right-hand-side does not depend on x. However, $k(x)$ cannot be arbitrary. To see this, let us suppose that an efficient estimator exists, so that $\lambda_{\hat{x}}(x) = 1/J_{\mathbf{Y}}(x)$. In this case,

$$\mathbb{E}[e^2(\mathbf{Y})] = \mathbb{E}[e(\mathbf{Y})k(x)h(\mathbf{Y})] = k(x)\mathbb{E}[e(\mathbf{Y})h(\mathbf{Y})] = k(x) \tag{88}$$

since the Cauchy-Schwartz inequality must hold with equality. We conclude that

$$k(x) = \frac{1}{J_{\mathbf{Y}}(x)}. \tag{89}$$

We conclude that $\hat{x}(\cdot)$ is efficient if and only if

$$\hat{x}(\mathbf{y}) = x + \frac{1}{J_{\mathbf{Y}}(x)} \frac{\partial}{\partial x} \log f_{\mathbf{Y}}(\mathbf{y}; x). \tag{90}$$

where the right-hand side must be independent of x for the estimator to be valid. Note from this expression that an efficient estimator is automatically unbiased because

$$\mathbb{E}[\hat{x}(\mathbf{Y})] = x + \frac{1}{J_{\mathbf{Y}}(x)} \mathbb{E}\left[\frac{\partial}{\partial x} \log f_{\mathbf{Y}}(\mathbf{y}; x)\right] = x, \tag{91}$$

where the final equality follows from the regularity assumption in (54).

The following claim says that when it *when it exists*, the (unique) efficient estimator is equivalent to the ML estimator for the problem.

Theorem 6. *When an efficient estimator $\hat{x}_{\text{eff}}(\cdot)$ exists, it is the ML estimator, i.e.,*

$$\hat{x}_{\text{eff}}(\cdot) = \hat{x}_{\text{ML}}(\cdot). \tag{92}$$

Proof. Suppose an efficient estimator $\hat{x}_{\text{eff}}(\cdot)$ exists. Then it can be written as

$$\hat{x}_{\text{eff}}(\mathbf{y}) = x + \frac{1}{J_{\mathbf{Y}}(x)} \frac{\partial}{\partial x} \log f_{\mathbf{Y}}(\mathbf{y}; x). \tag{93}$$

for any x and we can compute it directly. Now since the right-hand side is independent of the value of x, we are free to choose any value of x in this expression, so let us judiciously choose x to be the number

$$\hat{x}_{\text{ML}}(\mathbf{y}) = \arg\max_x f_{\mathbf{Y}}(\mathbf{y}; x), \tag{94}$$

which is the maximum likelihood estimator of x given $\mathbf{y}$. Provided the likelihood function is strictly positive and differentiable, the ML estimator satisfies

$$\left[\frac{\partial}{\partial x} \log f_{\mathbf{Y}}(\mathbf{y}; x)\right]\Bigg|_{x=\hat{x}_{\text{ML}}(\mathbf{y})} = 0. \tag{95}$$

Then since $J_{\mathbf{Y}}(x) > 0$, for all x except in the trivial case, (93) becomes

$$\hat{x}_{\text{eff}}(\mathbf{y}) = \hat{x}_{\text{ML}}(\mathbf{y}). \tag{96}$$

From this we can conclude that *when it exists*, the (unique) efficient estimator is equivalent to the ML estimator for the problem. □

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 often have desirable asymptotic properties, in the limit of a large number of (independent) observations, that make them attractive. We will not cover these properties. For the interested reader, please see the book by Van Trees [Van68] or Poor [Poo98].

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

Example 6. Suppose that the random variable Y is exponentially-distributed with unknown mean $x > 0$, i.e.,

$$f_Y(y;x) = \frac{1}{x} e^{-y/x}, \quad y \geq 0. \tag{97}$$

We obtain the ML estimate as

$$\frac{\partial}{\partial x} \log f_Y(y;x) = \frac{\partial}{\partial x} \left[-\log x - \frac{y}{x} \right] = -\frac{1}{x} + \frac{y}{x^2} = 0. \tag{98}$$

Thus,

$$\hat{x}_{\mathrm{ML}}(y) = y. \tag{99}$$

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

$$\lambda_{\mathrm{ML}}(x) = \mathsf{Var}(\hat{x}_{\mathrm{ML}}(Y)) = \mathsf{Var}(Y) = x^2, \tag{100}$$

we obtain

$$J_Y(x) = \mathbb{E}\left[\left(\frac{\partial}{\partial x} \log f_Y(y;x) \right)^2 \right] = \mathbb{E}\left[\frac{(Y-x)^2}{x^4} \right] = \frac{1}{x^4} \cdot x^2 = \frac{1}{x^2} = \frac{1}{\lambda_{\mathrm{ML}}(x)}. \tag{101}$$

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 7. Suppose we observe a vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_M)^T$ of independent Poisson random variables with unknown mean x, i.e., for $i = 1, 2, \ldots, M$,

$$p_{Y_i}(y_i;x) = \Pr(Y_i = y_i; x) = \frac{x^{y_i} e^{-x}}{y_i!}, \quad y_i \in \mathbb{N} \cup \{0\}. \tag{102}$$

In this case,

$$\log p_{\mathbf{Y}}(\mathbf{Y};x) = \sum_{i=1}^{M} \log p_{Y_i}(y_i;x) = \sum_{i=1}^{M} (y_i \log x - x) - \sum_{i=1}^{M} \log(y_i!), \tag{103}$$

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

$$\frac{\partial}{\partial x} \log p_{\mathbf{Y}}(\mathbf{Y};x) = \sum_{i=1}^{M} \left(\frac{Y_i}{x} - 1 \right) = 0. \tag{104}$$

In particular, we obtain that

$$\hat{x}_{\mathrm{ML}}(\mathbf{y}) = \frac{1}{M} \sum_{i=1}^{M} Y_i, \tag{105}$$

which again is then unbiased. Since the variance of a Poisson random variable equals its mean, we have

$$\lambda_{\mathrm{ML}} = \frac{1}{M^2} \sum_{i=1}^{M} x = \frac{x}{M}. \tag{106}$$

Using the alternative formula for the Fisher information in (75), we obtain that the Fisher information is

$$J_{\mathbf{Y}}(x) = -\mathbb{E}\left[\frac{\partial^2}{\partial x^2} \log p_{\mathbf{Y}}(\mathbf{Y};x) \right] = \frac{1}{x^2} \mathbb{E}\left[\sum_{i=1}^{M} Y_i \right] = \frac{M}{x}, \tag{107}$$

so comparing (106) and (107) shows that the ML estimate is efficient.

References

[Poo98] H. V. Poor. *An Introduction to Signal Detection and Estimation.* Springer, 2nd, 1998.

[Van68] H. Van Trees. *Detection, Estimation and Modulation Theory: Part I.* New York, Wiley, 1968.