*Estimation with Applications to Tracking and Navigation: Theory Algorithms and Software*
Yaakov Bar-Shalom, X. Rong Li, Thiagalingam Kirubarajan

ISBNs: 0-471-41655-X (Hardback) 0-471-22127-9 (Electronic)

# Chapter 2

# BASIC CONCEPTS IN ESTIMATION

## 2.1 INTRODUCTION

### 2.1.1 Outline

This chapter introduces some of the basic techniques of estimation that provide the foundation for state estimation and its applications like tracking, navigation, etc.

The problem of parameter estimation is defined in Section 2.2, where the two most commonly used models for unknown parameters (nonrandom and random) are also described. Section 2.3 deals with the maximum likelihood (ML) and the maximum a posteriori (MAP) estimators. The least squares (LS) and the minimum mean square error (MMSE) estimators are presented in Section 2.4.

The remaining sections deal with various "measures of quality" of estimators. Section 2.5 discusses unbiasedness and Section 2.6 discusses the variances of estimators. The consistency of estimators is discussed in Section 2.7, together with "information limit" results: the Cramer-Rao lower bound, the Fisher information, and estimator efficiency.

### 2.1.2 Basic Concepts – Summary of Objectives

Distinguish between

- Random parameters
- Nonrandom parameters

Define the following estimates

- Maximum likelihood
- Maximum a posteriori
- Least squares

- Minimum mean square error

Present "measures of quality" of estimators and "information limit" results

- Unbiasedness
- Variance
- Consistency
- The Cramer-Rao lower bound, Fisher information
- Efficiency

## 2.2 THE PROBLEM OF PARAMETER ESTIMATION

### 2.2.1 Definitions

The term ***parameter*** is used to designate a quantity (scalar or vector valued) that is assumed to be *time invariant.* If it does change with time, it can be designated (with a slight abuse of language) as a "time-varying parameter," but its time variation must be "slow" compared to the state variables of a system. *State estimation*, which is for *dynamic systems*, is covered starting with Chapter 5.

The problem of estimating a (time invariant) parameter $x$ is the following. Given the measurements

$$z(j) = h[j, x, w(j)] \qquad j = 1, \ldots, k \tag{2.2.1-1}$$

made in the presence of the disturbances (noises) $w(j)$, find a function of the $k$ observations

$$\boxed{\hat{x}(k) \triangleq \hat{x}[k, Z^k]} \tag{2.2.1-2}$$

where these observations are denoted compactly as

$$Z^k \triangleq \{z(j)\}_{j=1}^{k} \tag{2.2.1-3}$$

that estimates the value of $x$ in some sense.

The function (2.2.1-2) is called the ***estimator***. The value of this function is the ***estimate***. These terms, while not the same, will be used (sometimes) interchangeably.

The ***estimation error*** corresponding to the estimate $\hat{x}$ is

$$\boxed{\tilde{x} \triangleq x - \hat{x}} \tag{2.2.1-4}$$

An alternate notation instead of (2.2.1-2) that will be used when $k$ is fixed (and, therefore, can be omitted) is

$$\hat{x}(Z) \triangleq \hat{x}[k, Z^k] \tag{2.2.1-5}$$

where $Z$ is the set of observations.

**Remark**

Parameter estimation can be viewed as a ***static estimation problem***, while state estimation can be viewed as a ***dynamic estimation problem***.

### 2.2.2 Models for Estimation of a Parameter

There are two models one can use in the estimation of a (time invariant) parameter:

1. Nonrandom ("unknown constant"): There is an unknown true value $x_0$. This is also called the ***non-Bayesian*** or ***Fisher approach***.
2. Random: The parameter is a random variable with a *prior* (or *a priori*) pdf $p(x)$ — a *realization* (see Subsection 1.4.2) of $x$ according to $p(x)$ is assumed to have occurred; this value then stays constant during the measurement process. This is also called the ***Bayesian approach***.

**The Bayesian Approach**

In the ***Bayesian approach***, one starts with the ***prior*** pdf of the parameter from which one can obtain its ***posterior*** pdf (or ***a posteriori*** pdf) using ***Bayes' formula***:

$$\boxed{p(x|Z) = \frac{p(Z|x)p(x)}{p(Z)} = \frac{1}{c}p(Z|x)p(x)} \tag{2.2.2-1}$$

where $c$ is the ***normalization constant***, which does not depend on $x$.

The posterior pdf can be used in several ways to estimate $x$.

**The Non-Bayesian (Likelihood Function) Approach**

In contrast to the above, in the ***non-Bayesian approach*** there is no prior pdf associated with the parameter and thus one cannot define a posterior pdf for it.

In this case, one has the *pdf of the measurements conditioned on the parameter*, called the ***likelihood function (LF)*** of the parameter

$$\boxed{\Lambda_Z(x) \triangleq p(Z|x)} \tag{2.2.2-2}$$

or

$$\Lambda_k(x) \triangleq p(Z^k|x) \tag{2.2.2-3}$$

as a measure of how "likely" a parameter value is given the obtained observations. The likelihood function serves as a measure of the ***evidence from the data***.

The use of the LF (2.2.2-2) and a similar usage of the posterior pdf (2.2.2-1) are discussed in the next section.

## 2.3 MAXIMUM LIKELIHOOD AND MAXIMUM A POSTERIORI ESTIMATORS

### 2.3.1 Definitions of ML and MAP Estimators

#### Maximum Likelihood Estimator

A common method of estimating nonrandom parameters is the ***maximum likelihood method*** that maximizes the likelihood function (2.2.2-2). This yields the ***maximum likelihood estimator (MLE)***

$$\boxed{\hat{x}^{\mathrm{ML}}(Z) = \arg\max_x \Lambda_Z(x) = \arg\max_x p(Z|x)} \tag{2.3.1-1}$$

Note that, while $x$ is an unknown constant, $\hat{x}^{\mathrm{ML}}(Z)$, being a function of the set of random observations $Z$, is a random variable.

The MLE is the solution of the ***likelihood equation***

$$\frac{d\Lambda_Z(x)}{dx} = \frac{dp(Z|x)}{dx} = 0 \tag{2.3.1-2}$$

#### Maximum A Posteriori Estimator

The corresponding estimate for a random parameter is the ***maximum a posteriori (MAP)*** estimator, which follows from the maximization of the posterior pdf (2.2.2-1):

$$\boxed{\hat{x}^{\mathrm{MAP}}(Z) = \arg\max_x p(x|Z) = \arg\max_x [p(Z|x)p(x)]} \tag{2.3.1-3}$$

The last equality above follows from the fact that, when using Bayes' formula (2.2.2-1), the normalization constant is irrelevant for the maximization.

The MAP estimate, which depends on the observations $Z$, and through them on the realization of $x$ is, obviously, a random variable.

### 2.3.2 MLE vs. MAP Estimator with Gaussian Prior

Consider the single measurement

$$z = x + w \tag{2.3.2-1}$$

of the unknown parameter $x$ in the presence of the additive measurement noise $w$, assumed to be a normally (Gaussian) distributed random variable with mean zero and variance $\sigma^2$, that is,

$$w \sim \mathcal{N}(0, \sigma^2) \tag{2.3.2-2}$$

First assume that $x$ is an unknown constant (no prior information about it is available). The likelihood function of $x$ (denoted here without a subscript, for simplicity) is

$$\Lambda(x) = p(z|x) = \mathcal{N}(z; x, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-x)^2}{2\sigma^2}} \tag{2.3.2-3}$$

Then

$$\hat{x}^{\text{ML}} = \arg\max_x \Lambda(x) = z \tag{2.3.2-4}$$

since the peak or ***mode*** of (2.3.2-3) occurs at $x = z$.

Next assume that the prior information about the parameter is that $x$ is Gaussian with mean $\bar{x}$ and variance $\sigma_0^2$, that is,

$$p(x) = \mathcal{N}(x; \bar{x}, \sigma_0^2) \tag{2.3.2-5}$$

It is also assumed that $x$ is independent of $w$.

Then the posterior pdf of $x$ conditioned on the observation $z$ is

$$p(x|z) = \frac{p(z|x)p(x)}{p(z)} = \frac{1}{c} e^{-\frac{(z-x)^2}{2\sigma^2} - \frac{(x-\bar{x})^2}{2\sigma_0^2}} \tag{2.3.2-6}$$

where

$$c = 2\pi\sigma\sigma_0 p(z) \tag{2.3.2-7}$$

is the normalization constant independent of $x$. This normalization constant, which guarantees that the pdf integrates to unity, is given explicitly next, after rearranging the exponent in (2.3.2-6).

After rearranging the exponent in the above by completing the squares in $x$, it can be easily shown in a manner similar to the one used in Subsection 1.4.14 that the posterior pdf of $x$ is

$$p(x|z) = \mathcal{N}[x; \xi(z), \sigma_1^2] = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{[x-\xi(z)]^2}{2\sigma_1^2}} \tag{2.3.2-8}$$

i.e., Gaussian, where

$$\xi(z) \triangleq \frac{\sigma^2}{\sigma_0^2 + \sigma^2}\bar{x} + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2} z = \bar{x} + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}(z - \bar{x}) \tag{2.3.2-9}$$

and (the "parallel resistors formula")

$$\sigma_1^2 \triangleq \frac{\sigma_0^2 \sigma^2}{\sigma_0^2 + \sigma^2} \tag{2.3.2-10}$$

The maximization of (2.3.2-8) with respect to $x$ yields immediately

$$\hat{x}^{\text{MAP}} = \xi(z) \tag{2.3.2-11}$$

that is, $\xi(z)$ given by (2.3.2-9) is the *maximum a posteriori estimator* for the random parameter $x$ with the prior pdf (2.3.2-5).

Note that the MAP estimator (2.3.2-9) for this (purely Gaussian) problem is a weighted combination of

1. $z$, the MLE, which is the peak (or mode) of the likelihood function;
2. $\bar{x}$, which is the peak of the prior pdf of the parameter to be estimated.

Equation (2.3.2-9) can be rewritten as follows:

$$\begin{aligned} \hat{x}^{\text{MAP}} &= (\sigma_0^{-2} + \sigma^{-2})^{-1}\sigma_0^{-2}\bar{x} + (\sigma_0^{-2} + \sigma^{-2})^{-1}\sigma^{-2}z \\ &= (\sigma_0^{-2} + \sigma^{-2})^{-1}\left[\frac{\bar{x}}{\sigma_0^2} + \frac{z}{\sigma^2}\right] \end{aligned} \tag{2.3.2-12}$$

which indicates that the weightings of the prior mean and the measurement are *inversely proportional to their variances.*

Similarly, (2.3.2-10) can be rewritten as follows:

$$\sigma_1^{-2} = \sigma_0^{-2} + \sigma^{-2} \tag{2.3.2-13}$$

which shows that the *inverse variances* (also called ***information*** — this will be discussed in more detail in Subsection 3.4.2) are *additive.* This additivity property of information holds in general when the information sources are *independent.* (See also problem 2-8.)

### 2.3.3 MAP Estimator with One-Sided Exponential Prior

Consider the same problem as before except that the prior pdf of $x$ is a ***one-sided exponential pdf***

$$p(x) = ae^{-ax} \qquad x \geq 0 \tag{2.3.3-1}$$

This can model, for instance, the arrival time in a stochastic process where the number of arrivals is Poisson distributed.

The ML estimate is the same as before in (2.3.2-4), that is,

$$\hat{x}^{\text{ML}} = z \tag{2.3.3-2}$$

The posterior pdf of $x$ is now

$$p(x|z) = c(z)e^{-\frac{(z-x)^2}{2\sigma^2} - ax} \qquad x \geq 0 \tag{2.3.3-3}$$

Since the exponent is quadratic in $x$, the above posterior pdf is Gaussian but truncated due to the fact that $x$ cannot be negative as modeled by the prior given in (2.3.3-1).

In view of the fact that it cannot be negative, the maximizing argument of (2.3.3-3) is given by

$$\hat{x}^{\text{MAP}} = \max(z - \sigma^2 a, 0) \tag{2.3.3-4}$$

Note that the MAP estimate (2.3.3-4) in this case will always be smaller than the MLE (2.3.3-2) as long as the latter is not negative because the prior (2.3.3-1) attaches higher probability to smaller values of $x$.

### 2.3.4 MAP Estimator with Diffuse Prior

While $\hat{x}^{\text{ML}}$ is based on a non-Bayesian approach and $\hat{x}^{\text{MAP}}$ is based on the Bayesian approach, the latter will coincide with the former for a certain prior pdf, called a ***diffuse pdf.***

This can be seen by rewriting the denominator in Bayes' formula

$$p(x|Z) = \frac{p(Z|x)p(x)}{p(Z)} \tag{2.3.4-1}$$

with the total probability theorem

$$p(Z) = \int_{-\infty}^{\infty} p(Z|x)p(x)\,dx \tag{2.3.4-2}$$

and assuming a ***diffuse uniform prior pdf*** for the parameter

$$p(x) = \epsilon \qquad \text{for } |x| < \frac{1}{2\epsilon} \tag{2.3.4-3}$$

over a "sufficiently large" region of length $1/\epsilon$ where $\epsilon > 0$ but small. Using (2.3.4-3) in (2.3.4-2) yields

$$p(Z) = \epsilon \int_{-\infty}^{\infty} p(Z|x)dx = \epsilon g(Z) \tag{2.3.4-4}$$

where $g$ does not depend on $x$.

Then, inserting (2.3.4-4) into Bayes' formula (2.3.4-1) yields

$$p(x|Z) = \frac{p(Z|x)\epsilon}{\epsilon g(Z)} = \frac{p(Z|x)}{g(Z)} = \frac{1}{c}p(Z|x) \tag{2.3.4-5}$$

since $\epsilon \neq 0$.

This diffuse pdf is also called ***improper pdf*** because as $\epsilon \to 0$, it does not integrate to unity as a ***proper pdf*** does. Another name for it is ***noninformative pdf*** because it carries no information about the parameter: uniform distribution over an infinite interval at the limit.

A diffuse prior causes the posterior pdf of $x$ to be proportional to its likelihood function and, thus, the MAP estimate to coincide with the MLE.

#### Bayesian vs. Non-Bayesian Philosophies

The non-Bayesian MLE is, in view of the above discussion, nothing but the Bayesian MAP estimate with complete prior ignorance, reflected by the diffuse prior (2.3.4-3). *This provides a philosophically unifying view of the Bayesian and non-Bayesian approaches to estimation.*[1]

[1]Thus one can say that statisticians can be divided into two categories: Bayesians and closet Bayesians. This is a particular case of the following general taxonomy theorem: People can be divided into two categories. (*Proof*: The categories are those who believe that people can be divided into two categories and those who do not.)

### Remark

In spite of the fact that the diffuse prior (2.3.4-3) does not integrate to unity and has no moments (i.e., it is *not a proper pdf*), the posterior pdf of $x$ will be, in general, proper.

### Example

Consider the problem of Subsection 2.3.2 where the prior (2.3.2-5) is made diffuse by making $\sigma_0 \to \infty$. A Gaussian pdf with very large variance becomes flat and at the limit looks like a uniform pdf over the whole real line.

When $\sigma_0 \to \infty$ it can be seen from (2.3.2-9) that

$$\lim_{\sigma_0 \to \infty} \xi(z) = z \tag{2.3.4-6}$$

that is, $\hat{x}^{\text{MAP}}$ coincides with $\hat{x}^{\text{ML}}$. This occurs regardless of the value of $\bar{x}$, which becomes irrelevant when $\sigma_0 \to \infty$ (i.e., this is a *noninformative prior*). Thus the non-Bayesian approach can be seen as a degenerate case of the Bayesian approach.

### The Philosophical Meaning of the Prior

The prior pdf assumed in a problem is in many cases the subjective assessment of phenomena. The uniform prior assumes Nature as "indifferent." In game theory, Nature is assumed to be opposed to our interests. While neither of these two extreme points of view is correct, it is useful to keep in mind the well-known ***principle of perversity of inanimate objects.***[2]

## 2.3.5 The Sufficient Statistic and the Likelihood Equation

If the likelihood function of a parameter can be decomposed as follows

$$\Lambda(x) \triangleq p(Z|x) = f_1[g(Z), x] f_2(Z) \tag{2.3.5-1}$$

then it is clear that the maximum likelihood estimate of $x$ depends only on the function $g(Z)$, called the ***sufficient statistic***, rather than on the entire data set $Z$.

The sufficient statistic *summarizes the information about $x$ contained in the entire data.*

[2] As mentioned by Richard Bellman [Bellman61], this has been established by a number of experiments. The most conclusive of these involved dropping a piece of buttered toast on a rug. In 79.3% of the trials the toast fell buttered side down; for a mathematical proof of this principle, which does not hold in the land of the Brobdingnagians, see problem 1-17. For an extensive discussion on priors, see [Raiffa72].

**Example**

Consider the scalar measurements

$$z(j) = x + w(j) \qquad j = 1, \ldots, k \tag{2.3.5-2}$$

If the noise components $w(j)$, $j = 1, \ldots, k$, are independent and identically distributed zero-mean Gaussian random variables with variance $\sigma^2$, that is,

$$w(j) \sim \mathcal{N}(0, \sigma^2) \tag{2.3.5-3}$$

then

$$z(j) \sim \mathcal{N}(x, \sigma^2) \tag{2.3.5-4}$$

and, conditioned on $x$, the observations $z(j)$ are mutually independent.

Thus, the ***likelihood function*** of $x$ in terms of

$$Z^k \triangleq \{z(j), j = 1, \ldots, k\} \tag{2.3.5-5}$$

is then

$$\begin{aligned} \Lambda_k(x) &\triangleq p(Z^k|x) \triangleq p[z(1), \ldots, z(k)|x] \\ &= \prod_{j=1}^{k} \mathcal{N}[z(j); x, \sigma^2] = ce^{-\frac{1}{2\sigma^2}\sum_{j=1}^{k}[z(j)-x]^2} \end{aligned} \tag{2.3.5-6}$$

The likelihood function (2.3.5-6) can be rewritten into the product of two functions as in (2.3.5-1) as follows:

$$\begin{aligned} \Lambda_k(x) &= c\, e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{k} z(j)^2 + \frac{1}{2\sigma^2} 2\sum_{j=1}^{k} z(j)x - \frac{1}{2\sigma^2}kx^2} \\ &= c\, e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{k} z(j)^2} e^{-\frac{1}{2\sigma^2}kx\left[x - \frac{2}{k}\sum_{j=1}^{k} z(j)\right]} \\ &\triangleq f_2(Z) f_1[g(Z), x] \end{aligned} \tag{2.3.5-7}$$

where

$$f_2(Z) \triangleq ce^{-\frac{1}{2\sigma^2}\sum_{j=1}^{k} z(j)^2} \tag{2.3.5-8}$$

$$f_1[g(Z), x] \triangleq e^{-\frac{1}{2\sigma^2}kx[x-2\bar{z}]} \tag{2.3.5-9}$$

$$g(Z) \triangleq \frac{1}{k}\sum_{j=1}^{k} z(j) \triangleq \bar{z} \tag{2.3.5-10}$$

Thus, according to the definition (2.3.5-1), $\bar{z}$ is the *sufficient statistic* for estimating $x$.

### The Likelihood Equation

To maximize the likelihood function (2.3.5-7), one sets its derivative with respect to $x$ to zero. Since $f_2$ is independent of $x$, the ***likelihood equation*** is

$$\frac{d\Lambda_k(x)}{dx} = 0 \qquad \Longleftrightarrow \qquad \frac{df_1[g(Z), x]}{dx} = 0 \tag{2.3.5-11}$$

Since the logarithm is a monotonic transformation, equivalently one can use the derivative of the ***log-likelihood function***

$$\frac{d \ln \Lambda_k(x)}{dx} = \frac{d \ln f_1[g(Z), x]}{dx} = -\frac{k}{2\sigma^2} 2(x - \bar{z}) = 0 \tag{2.3.5-12}$$

which yields

$$\hat{x}^{\mathrm{ML}} = \bar{z} \tag{2.3.5-13}$$

The concept of sufficient statistic carries over to the MAP procedure in a completely analogous manner.

## 2.4 LEAST SQUARES AND MINIMUM MEAN SQUARE ERROR ESTIMATION

### 2.4.1 Definitions of LS and MMSE Estimators

### The LS Estimator

Another common estimation procedure for nonrandom parameters is the ***least squares (LS) method***. Given the (scalar and nonlinear) measurements

$$z(j) = h(j, x) + w(j) \qquad j = 1, \ldots, k \tag{2.4.1-1}$$

the ***least squares estimator (LSE)*** of $x$ is, with notation (2.2.1-2),

$$\boxed{\hat{x}^{\mathrm{LS}}(k) = \arg\min_x \left\{ \sum_{j=1}^{k} [z(j) - h(j, x)]^2 \right\}} \tag{2.4.1-2}$$

This is the ***nonlinear LS problem*** — if the function $h$ is linear in $x$, then one has the ***linear LS problem***. The linear LS problem is considered in more detail for the vector case in Section 3.4.

The criterion in (2.4.1-2) makes no assumptions about the "measurement errors" or "noises" $w(j)$. If these are independent and identically distributed zero-mean Gaussian random variables, that is,

$$w(j) \sim \mathcal{N}(0, \sigma^2) \tag{2.4.1-3}$$

then the LSE (2.4.1-2) coincides with the MLE under these assumptions. In this case,

$$z(j) \sim \mathcal{N}[h(j,x), \sigma^2] \qquad j = 1, \ldots, k \tag{2.4.1-4}$$

The likelihood function of $x$ is then

$$\begin{aligned} \Lambda_k(x) &\triangleq p(Z^k|x) \triangleq p[z(1), \ldots, z(k)|x] \\ &= \prod_{j=1}^{k} \mathcal{N}[z(j); h(j,x), \sigma^2] = ce^{-\frac{1}{2\sigma^2}\sum_{j=1}^{k}[z(j)-h(j,x)]^2} \end{aligned} \tag{2.4.1-5}$$

and the minimization (2.4.1-2) is equivalent to the maximization of (2.4.1-5); that is, *the LS method is a "disguised" ML approach.*

### The MMSE Estimator

For random parameters, the counterpart of the above is the ***minimum mean square error (MMSE) estimator***

$$\hat{x}^{\text{MMSE}}(Z) = \arg\min_{\hat{x}} E[(\hat{x} - x)^2|Z] \tag{2.4.1-6}$$

The solution to (2.4.1-6) is the ***conditional mean*** of $x$

$$\boxed{\hat{x}^{\text{MMSE}}(Z) = E[x|Z] \triangleq \int_{-\infty}^{\infty} xp(x|Z)\,dx} \tag{2.4.1-7}$$

where the expectation is with respect to the conditional pdf (2.2.2-1).

The above follows by setting the derivative of (2.4.1-6) with respect to $\hat{x}$ to zero:

$$\frac{d}{d\hat{x}} E[(\hat{x} - x)^2|Z] = E[2(\hat{x} - x)|Z] = 2(\hat{x} - E[x|Z]) = 0 \tag{2.4.1-8}$$

For vector random variables, (2.4.1-7) is obtained similarly by setting the gradient of the mean of the squared norm of the error to zero, that is

$$\nabla_{\hat{x}} E[(\hat{x} - x)'(\hat{x} - x)|Z] = 2(\hat{x} - E[x|Z]) = 0 \tag{2.4.1-9}$$

from which (2.4.1-7) follows immediately.

### Remarks

1. With $x$ being an unknown constant (nonrandom) and the noises in (2.4.1-1) modeled as random (not necessarily Gaussian), the LSE is a random variable.
2. The MMSE estimate (2.4.1-7) is a random variable that depends on the observations $Z$ and, through them, on (the realization of) $x$. Also, for a *given* $Z$, $x$ is a random variable with a conditional pdf (2.2.2-1).
3. The MMSE estimation problem (2.4.1-6) is a particular case of Bayesian estimation where the expected value of a (positive definite) ***cost function*** $C(\hat{x} - x)$ is to be minimized. The MMSE cost function is a quadratic. The widespread use of the quadratic criterion is due primarily to the (relative) ease of obtaining the solution. (See also problem 2-3.)

### 2.4.2 Some LS Estimators

#### LS Estimator from a Single Measurement

For the problem of a single measurement of the unknown parameter $x$,

$$z = x + w \tag{2.4.2-1}$$

the least squares criterion leads to

$$\hat{x}^{\text{LS}} = \arg\min_x[(z-x)^2] = z \tag{2.4.2-2}$$

which is the same result as $\hat{x}^{\text{ML}}$ if $w$ is zero-mean Gaussian. This is due to the fact that maximizing the likelihood function, which is a Gaussian pdf, is equivalent to minimizing the square in its exponent.

#### LS Estimator from Several Measurements

Assume now that $k$ measurements are made

$$z(j) = x + w(j) \qquad j = 1, \ldots, k \tag{2.4.2-3}$$

where $w(j)$ are independent, identically distributed, normal, zero mean, and with common variance $\sigma^2$.

The likelihood function is, as in (2.4.1-5),

$$\Lambda_k(x) = ce^{-\frac{1}{2\sigma^2}\sum_{j=1}^{k}[z(j)-x]^2} \tag{2.4.2-4}$$

As before, the ML and LS estimates coincide and it can be easily shown that they are given by the following expression:

$$\hat{x}^{\text{ML}}(k) = \hat{x}^{\text{LS}}(k) = \frac{1}{k}\sum_{j=1}^{k} z(j) = \bar{z} \tag{2.4.2-5}$$

This estimate is known as the ***sample mean*** or ***sample average***, since it estimates the unknown mean $x$ of the $k$ random variables from (2.4.2-3).

### 2.4.3 MMSE vs. MAP Estimator in Gaussian Noise

In the single measurement example with a prior pdf on the parameter to be estimated, discussed in Subsection 2.3.2, the posterior pdf of $x$ was obtained in (2.3.2-8) as

$$p(x|z) = \frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{[x-\xi(z)]^2}{2\sigma_1^2}} \tag{2.4.3-1}$$

It is apparent by inspection that the mean of this Gaussian pdf is $\xi(z)$, which is also the ***mode*** (peak) of this pdf.

Thus

$$\hat{x}^{\text{MMSE}} = E[x|z] = \xi(z) = \hat{x}^{\text{MAP}} \tag{2.4.3-2}$$

i.e., the MMSE estimator (the conditional mean) *coincides* with the MAP estimator.

This is due to the fact that the *mean* and the *mode* of a Gaussian pdf, which is symmetric and ***unimodal***, coincide.

Note that, in view of (2.4.3-2), equation (2.4.3-1) can be also written as

$$p(x|z) = \mathcal{N}(x; \hat{x}^{\text{MMSE}}, \sigma_1^2) = \mathcal{N}(x; \hat{x}^{\text{MAP}}, \sigma_1^2) \tag{2.4.3-3}$$

## 2.5 UNBIASED ESTIMATORS

### 2.5.1 Definition

#### Non-Bayesian Case

For a nonrandom parameter, an estimator is said to be ***unbiased*** if

$$E[\hat{x}(k, Z^k)] = x_0 \tag{2.5.1-1}$$

where $x_0$ is the true value of the parameter. The expectation in (2.5.1-1) is over the estimate, which is a random variable since it is a function of the measurements (2.2.1-3), and is taken with respect to the conditional pdf $p(Z^k|x = x_0)$.

#### Bayesian Case

If $x$ is a random variable with a prior pdf $p(x)$, then the unbiasedness property is written as

$$E[\hat{x}(k, Z^k)] = E[x] \tag{2.5.1-2}$$

where the expectation on the left-hand side above is with respect to the joint pdf $p(Z^k, x)$ and the one on the right-hand side is with respect to $p(x)$.

#### General Definition

The above unbiasedness requirements can be unified by requiring that the ***estimation error***

$$\tilde{x} \triangleq x - \hat{x} \tag{2.5.1-3}$$

be zero mean, that is,

$$\boxed{E[\tilde{x}] = 0} \tag{2.5.1-4}$$

Equation (2.5.1-4) covers both cases, with the expectation being taken over $Z^k$ in the first case and over $Z^k$ *and* $x$ in the second case.

An estimator is unbiased if (2.5.1-4) holds for all $k$ and is ***asymptotically unbiased*** if it holds in the limit as $k \to \infty$.

### 2.5.2 Unbiasedness of an ML and a MAP Estimator

Consider the ML estimator (2.3.2-4) of the parameter $x$ with true value $x_0$

$$\hat{x}^{\text{ML}} = z \tag{2.5.2-1}$$

from the single measurement

$$z = x + w \tag{2.5.2-2}$$

Its mean is

$$E[\hat{x}^{\text{ML}}] = E[z] = E[x_0 + w] = x_0 + E[w] = x_0 \tag{2.5.2-3}$$

since the mean of the Gaussian random variable $w$ is zero.

For the MAP estimate (2.3.2-11) of $x$ modeled as a Gaussian random variable with prior mean $\bar{x}$, prior variance $\sigma^2$, and independent of $w$,

$$\hat{x}^{\text{MAP}} = \xi(z) \triangleq \frac{\sigma^2}{\sigma_0^2 + \sigma^2}\bar{x} + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}z \tag{2.5.2-4}$$

one has

$$\begin{aligned} E[\hat{x}^{\text{MAP}}] &= E[\xi(z)] = \frac{\sigma^2}{\sigma_0^2 + \sigma^2}\bar{x} + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}E[z] \\ &= \frac{\sigma^2}{\sigma_0^2 + \sigma^2}\bar{x} + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}[\bar{x} + E(w)] = \bar{x} = E[x] \end{aligned} \tag{2.5.2-5}$$

Thus, both of these estimates are unbiased.

### 2.5.3 Bias in the ML Estimation of Two Parameters

Consider the problem of estimating the unknown mean $x$ of a set of $k$ measurements as in (2.4.2-3), with the additional parameter to be estimated being the variance $\sigma^2$, now also assumed to be unknown.

The likelihood function for the unknown parameters $x$ and $\sigma$ is

$$\Lambda_k(x, \sigma) = p[z(1), \ldots, z(k)|x, \sigma] = \frac{1}{(2\pi)^{k/2}\sigma^k} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{k}[z(j)-x]^2} \tag{2.5.3-1}$$

To maximize the above, one writes the ***likelihood equation*** by setting to zero the derivatives of $\Lambda$ or, more conveniently, of $\ln \Lambda$, with respect to $x$ and $\sigma$

$$\frac{\partial \ln \Lambda_k}{\partial x} = \frac{1}{\sigma^2}\sum_{j=1}^{k}[z(j) - x] = 0 \tag{2.5.3-2}$$

$$\frac{\partial \ln \Lambda_k}{\partial \sigma} = -\frac{k}{\sigma} + \frac{1}{\sigma^3}\sum_{j=1}^{k}[z(j) - x]^2 = 0 \tag{2.5.3-3}$$

### The Solution of the Likelihood Equation

The first equation yields $\hat{x}^{\mathrm{ML}}$ as before in (2.4.2-5), that is, the sample mean — in this problem the estimate of $x$ is not affected at all by the fact that $\sigma$ is also unknown. Substituting this into the second equation yields

$$\frac{\partial \ln \Lambda_k}{\partial \sigma} = -\frac{k}{\sigma} + \frac{1}{\sigma^3}\sum_{j=1}^{k}[z(j) - \hat{x}^{\mathrm{ML}}]^2 = 0 \tag{2.5.3-4}$$

The resulting estimate, known as the ***sample variance*** based on $k$ observations, is

$$[\hat{\sigma}^{\mathrm{ML}}(k)]^2 = \frac{1}{k}\sum_{j=1}^{k}[z(j) - \hat{x}^{\mathrm{ML}}]^2 = \frac{1}{k}\sum_{i=1}^{k}\left[z(j) - \frac{1}{k}\sum_{i=1}^{k} z(i)\right]^2 \tag{2.5.3-5}$$

### The Means of the Sample Mean and Sample Variance

Denote the true values of the parameters by $x_0$ and $\sigma_0$. The expected value of the sample mean is

$$E[\hat{x}^{\mathrm{ML}}(k)] = E\left[\frac{1}{k}\sum_{j=1}^{k} z(j)\right] = x_0 \tag{2.5.3-6}$$

that is, the sample mean estimator (2.4.2-5) is unbiased.

The expected value of the sample variance (2.5.3-5) is

$$\begin{aligned} E\left\{[\hat{\sigma}^{\mathrm{ML}}(k)]^2\right\} &= E\left\{\frac{1}{k}\sum_{j=1}^{k}\left[z(j) - \frac{1}{k}\sum_{i=1}^{k} z(i)\right]^2\right\} \\ &= \frac{1}{k}\sum_{j=1}^{k} E\left\{\left[w(j) - \frac{1}{k}\sum_{i=1}^{k} w(i)\right]^2\right\} \\ &= \frac{1}{k^3}\sum_{j=1}^{k} E\left\{\left[(k-1)w(j) - \sum_{\substack{i=1 \\ i\neq j}}^{k} w(i)\right]^2\right\} \\ &= \frac{1}{k^2}[(k-1)^2 + k - 1]\sigma_0^2 \\ &= \frac{k-1}{k}\sigma_0^2 \end{aligned} \tag{2.5.3-7}$$

Thus the sample variance (2.5.3-5) is *biased*, even though it becomes unbiased as $k \to \infty$, i.e., it is *asymptotically unbiased*. In order to be unbiased, the denominator in (2.5.3-5) should be $k-1$ rather than $k$:

$$[\hat{\sigma}(k)]^2 = \frac{1}{k-1}\sum_{j=1}^{k}\left[z(j) - \frac{1}{k}\sum_{i=1}^{k} z(i)\right]^2 \tag{2.5.3-8}$$

Expression (2.5.3-8) is the more common sample variance used. However, for reasonably large $k$ this is not going to make a significant difference. (See also problem 2-4.)

## 2.6 THE VARIANCE AND MSE OF AN ESTIMATOR

### 2.6.1 Definitions of Estimator Variances

#### Non-Bayesian Case

For a non-Bayesian estimator, $\hat{x}(Z)$, (LS or ML) the ***variance of the estimator*** is

$$\text{var}[\hat{x}(Z)] \triangleq E\left[\{\hat{x}(Z) - E[\hat{x}(Z)]\}^2\right] \tag{2.6.1-1}$$

where the averaging is over the observation set $Z$.

If this estimator is *unbiased*, that is,

$$E[\hat{x}(Z)] = x_0 \tag{2.6.1-2}$$

where $x_0$ is the true value, then

$$\text{var}[\hat{x}(Z)] = E[[\hat{x}(Z) - x_0]^2] \tag{2.6.1-3}$$

If this estimator is *biased*, then (2.6.1-3) is its ***mean square error (MSE)***[3]

$$\text{MSE}[\hat{x}(Z)] = E[[\hat{x}(Z) - x_0]^2] \tag{2.6.1-4}$$

#### Bayesian Case

For a Bayesian estimator, the ***unconditional MSE*** is[4]

$$\text{MSE}[\hat{x}(Z)] \triangleq E[[\hat{x}(Z) - x]^2] \tag{2.6.1-5}$$

where the averaging is with respect to the joint pdf of the observations $Z$ and the random parameter $x$. The above can be rewritten, using the smoothing property of expectations (see Subsection 1.4.12), as follows:

$$\text{MSE}[\hat{x}(Z)] = E\left[E\{[\hat{x}(Z) - x]^2|Z\}\right] = E[\text{MSE}[\hat{x}(Z)|Z]] \tag{2.6.1-6}$$

where the last expression inside the braces is the ***conditional MSE***, i.e., for a given realization (or value) of the observations $Z$.

For the MMSE estimator, the conditional MSE is

$$\begin{aligned} E[[\hat{x}^{\text{MMSE}}(Z) - x]^2|Z] &= E[[x - E(x|Z)]^2|Z] \\ &= \text{var}(x|Z) \end{aligned} \tag{2.6.1-7}$$

that is, the ***conditional variance*** of $x$ *given* $Z$. Note that the expectations in (2.6.1-7) are with respect to $p(x|Z)$.

Averaging over $Z$ yields

$$E[\text{var}(x|Z)] = E[[x - E(x|Z)]^2] \tag{2.6.1-8}$$

which is the unconditional MSE (2.6.1-5) of the estimate $\hat{x}^{\text{MMSE}}$. This is the "average squared error over all the possible observations."

---

[3]Mean square is a personality type from the official list compiled by psycho-statisticians.

[4]The variance, since it would be about the mean $E[x]$ (if unbiased), has no real meaning.

### General Definition

With the definition of the estimation error

$$\tilde{x} \triangleq x - \hat{x} \tag{2.6.1-9}$$

one can say in a unified manner that the expected value of the square of the estimation error is the estimator's variance or MSE:

$$E[\tilde{x}^2] = \begin{cases} \text{var}(\hat{x}) & \text{if } \hat{x} \text{ is unbiased and } x \text{ is nonrandom} \\ \text{MSE}(\hat{x}) & \text{in all cases} \end{cases} \tag{2.6.1-10}$$

where the expectations are to be taken according to the discussion above.

The square root of the variance (or MSE) of an estimator

$$\sigma_{\hat{x}} \triangleq \sqrt{\text{var}(\hat{x})} \tag{2.6.1-11}$$

is its ***standard error***, also called the ***standard deviation associated with the estimator*** or the ***standard deviation of the estimation error.***

The standard error provides a measure of the accuracy of the estimator: assuming the estimation error to be Gaussian, the difference between the estimate and the true value will be up to 2 standard errors with 95% probability.

### 2.6.2 Comparison of Variances of an ML and a MAP Estimator

The "qualities" of the ML and MAP estimators discussed in Subsection 2.3.2 (from a single observation), as measured by their variances, will be compared next.

For the MLE given by (2.3.2-4) one has

$$\text{var}(\hat{x}^{\text{ML}}) = E[(\hat{x}^{\text{ML}} - x_0)^2] = E[(z - x_0)^2] \triangleq \sigma^2 \tag{2.6.2-1}$$

For the MAP estimate given by (2.3.2-11), which has a Gaussian prior in this case, one has[5]

$$\begin{aligned} \text{var}(\hat{x}^{\text{MAP}}) &= E[(\hat{x}^{\text{MAP}} - x)^2] \\ &= E\left\{\left[\frac{\sigma^2}{\sigma_0^2 + \sigma^2}\bar{x} + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}(x + w) - x\right]^2\right\} \\ &= E\left[\left[\frac{\sigma^2}{\sigma_0^2 + \sigma^2}(\bar{x} - x) + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}w\right]^2\right] \\ &= \frac{\sigma_0^2\sigma^2}{\sigma_0^2 + \sigma^2} < \sigma^2 = \text{var}(\hat{x}^{\text{ML}}) \end{aligned} \tag{2.6.2-2}$$

[5]With abuse of notation because this is really the MSE; such abuses of notation are very common in the literature.

Thus it can be seen that the variance of the MAP estimator (given by the "parallel resistors formula") is *smaller* than that of the MLE — this is due to the availability of *prior information.*

Note that in (2.6.2-1) the averaging is only over $z$ (or, equivalently, $w$) while in (2.6.2-2) the averaging is over $w$ *and* $x$, which is assumed random. (See also problem 2-1.)

### 2.6.3 The Variances of the Sample Mean and Sample Variance

The ***variance of the sample mean*** (2.4.2-5) — the square of its ***standard error*** — is obtained as

$$E\left[[\hat{x}^{\mathrm{ML}}(k) - x_0]^2\right] = E\left\{\left[\frac{1}{k}\sum_{j=1}^{k}[z(j) - x_0]\right]^2\right\} = \frac{\sigma^2}{k} \tag{2.6.3-1}$$

which, as $k \to \infty$, converges to zero; that is, this estimator is consistent (the definition is given in the next subsection).

The ***variance of the sample variance*** (2.5.3-5) is computed next. For simplicity, it is assumed that the mean is zero and known. The estimator of the variance in this case is

$$(\hat{\sigma}^{\mathrm{ML}})^2 = \frac{1}{k}\sum_{j=1}^{k} z(j)^2 \tag{2.6.3-2}$$

and it can be easily shown that it is unbiased.

The variance of this estimator is, with the true value denoted by $\sigma^2$,

$$\begin{aligned} E\left[[(\hat{\sigma}^{\mathrm{ML}})^2 - \sigma^2]^2\right] &= E\left\{\left[\frac{1}{k}\sum_{j=1}^{k} z(j)^2 - \sigma^2\right]^2\right\} \\ &= \frac{1}{k^2}\sum_{j=1}^{k}\sum_{i=1}^{k} E[w(j)^2 w(i)^2] - 2\sigma^2\frac{1}{k}\sum_{j=1}^{k} E[w(j)^2] + \sigma^4 \\ &= \frac{1}{k^2}[k(k-1)\sigma^4 + k3\sigma^4] - \frac{2}{k}k\sigma^4 + \sigma^4 = \frac{2\sigma^4}{k} \end{aligned} \tag{2.6.3-3}$$

which also converges to zero as $k \to \infty$.

The following relationship has been used in (2.6.3-3)

$$E[w(i)^2 w(j)^2] = \begin{cases} \sigma^4 & \text{if } i \neq j \\ 3\sigma^4 & \text{if } i = j \end{cases} \tag{2.6.3-4}$$

The fourth moment of $w$ is needed here and, assuming it to be Gaussian, use was made of (1.4.15-7).

The ***standard error of the sample variance*** from $k$ samples is therefore, from (2.6.3-3), given by

$$\sigma_{(\hat{\sigma}^{\mathrm{ML}})^2} = \sigma^2\sqrt{2/k} \tag{2.6.3-5}$$

In the above the notation, $\sigma_\xi$ has been used to denote the ***standard error of the estimate*** $\xi$.

**Application — The Number of Samples Needed to Estimate a Variance with a Given Accuracy**

Based on this result, the number of samples needed to obtain the sample variance *within 10% of the true value with probability of 95%* can be obtained as follows.

Assuming for convenience that the sample variance is normally distributed [6] about its mean (equal to the true variance) with standard error as above, one has the 95% confidence region

$$P\left\{|(\hat{\sigma}^{\mathrm{ML}})^2 - \sigma^2| \leq 1.96\sigma^2\sqrt{2/k}\right\} = 0.95 \tag{2.6.3-6}$$

The requirement of at most 10% error in the variance estimate, that is,

$$\frac{|(\hat{\sigma}^{\mathrm{ML}})^2 - \sigma^2|}{\sigma^2} = 0.1 \tag{2.6.3-7}$$

leads to setting

$$1.96\sqrt{2/k} = 0.1 \tag{2.6.3-8}$$

which yields

$$k \approx 800 \tag{2.6.3-9}$$

The resulting very large number of samples necessary for the required accuracy justifies the use of the CLT in (2.6.3-6). (See also problem 1-2.)

### 2.6.4 Estimation of the Probability of an Event

The ***estimation of the probability of an event*** can be done as follows. If in $N$ independent identical experiments an event occurs $N_0$ times, then one can use the following estimate of the probability $p$ of this event:

$$\hat{p} = \frac{N_0}{N} \tag{2.6.4-1}$$

It can be easily seen that the above is an unbiased estimate since

$$E\hat{p} = E\frac{N_0}{N} = \frac{Np}{N} = p \tag{2.6.4-2}$$

Now, since $N_0$ is a Bernoulli random variable (the sum of $N$ i.i.d. binary random variables), its distribution is given exactly by the binomial distribution. However, one can gain more insight into this by assuming $N$ large enough, in which case one can use, in view of the CLT, the Gaussian approximation of the pdf of $\hat{p}$. ICBES that

$$\mathrm{var}[\hat{p}] = \frac{p(1-p)}{N} \tag{2.6.4-3}$$

[6] In view of the CLT, see Subsection 1.4.23.

Thus, using the 95% ($2\sigma$) confidence region, one can say that

$$P\left\{|\hat{p}-p| \leq 2\sqrt{\frac{p(1-p)}{N}}\right\} = 0.95 \tag{2.6.4-4}$$

Since we do not know $p$, one can use $\hat{p}$ instead of $p$ in the expression of the variance above.

As an example, if $N = 1000$ and $\hat{p} = 0.1$, the standard deviation of the estimate is $\sqrt{0.1 \cdot 0.9/1000} \approx 0.01$. Therefore, the above confidence region becomes [0.08, 0.12].

Such techniques are commonly used in public opinion polls, but in a more conservative way: The worst case of $p = 0.5$ is assumed for the variance. For $N = 1000$ the standard deviation of the estimate is then $\sqrt{0.5 \cdot 0.5/1000} \approx 0.016$ with the confidence region being $[\hat{p} - 0.03, \hat{p} + 0.03]$. This is how to interpret the results from a poll, which state a ***margin of error*** of $\pm 3\%$.

## 2.7 CONSISTENCY AND EFFICIENCY OF ESTIMATORS

### 2.7.1 Consistency

An estimator of a *nonrandom parameter* is said to be a ***consistent estimator*** if the estimate (which is a random variable) converges to the true value in some stochastic sense.[7] Using the ***convergence in mean square*** criterion, then

$$\lim_{k\to\infty} E\left[[\hat{x}(k, Z^k) - x_0]^2\right] = 0 \tag{2.7.1-1}$$

is the condition for ***consistency in the mean square sense***. The expectation is taken over $Z^k$, as in (2.5.1-1).

For a *random parameter*, convergence of its estimator in the mean square sense requires

$$\lim_{k\to\infty} E\left[[\hat{x}(k, Z^k) - x]^2\right] = 0 \tag{2.7.1-2}$$

where the expectation is over $Z^k$ *and* $x$, as in (2.5.1-2).

Similarly to the unbiasedness case, consistency can be expressed as the requirement that the estimation error converge to zero, that is,

$$\lim_{k\to\infty} \tilde{x}(k, Z^k) = 0 \tag{2.7.1-3}$$

in some stochastic (e.g., mean square) sense.

[7] Another definition of consistency is "the last refuge of the unimaginative" (Oscar Wilde).

### Remark

The consistency defined above is an *asymptotic* property; that is, it is defined for the case when the sample size $k$ tends to infinity and the object of estimation is a *fixed parameter*. Later, in the context of state estimation, where the object of estimation is an *evolving state*, there will be another definition of consistency as a *finite sample size* property.

## 2.7.2 The Cramer-Rao Lower Bound and the Fisher Information Matrix

According to the ***Cramer-Rao lower bound (CRLB)***, the mean square error corresponding to the estimator of a parameter *cannot be smaller* than a certain quantity related to the likelihood function.

### Scalar Case

In the estimation of a scalar *nonrandom* parameter $x$ with an *unbiased* estimator $\hat{x}(Z)$, the variance is bounded from below as follows:

$$\boxed{E\left[[\hat{x}(Z) - x_0]^2\right] \geq J^{-1}} \tag{2.7.2-1}$$

where

$$J \triangleq -E\left[\frac{\partial^2 \ln \Lambda(x)}{\partial x^2}\right]\Bigg|_{x=x_0} = E\left\{\left[\frac{\partial \ln \Lambda(x)}{\partial x}\right]^2\right\}\Bigg|_{x=x_0} \tag{2.7.2-2}$$

is the ***Fisher information***, $\Lambda(x) = p(Z|x)$ is the likelihood function (2.2.2-2) denoted for simplicity without subscript, and $x_0$ is the true value of the unknown constant $x$.

For a scalar *random* parameter $x$ estimated by an unbiased estimator $\hat{x}(Z)$, the variance is bounded from below by a similar expression, namely,

$$E\left[[\hat{x}(Z) - x]^2\right] \geq J^{-1} \tag{2.7.2-3}$$

where

$$J \triangleq -E\left[\frac{\partial^2 \ln p(Z,x)}{\partial x^2}\right] = E\left\{\left[\frac{\partial \ln p(Z,x)}{\partial x}\right]^2\right\} \tag{2.7.2-4}$$

The expectations in (2.7.2-2) and (2.7.2-4) are taken as in (2.7.1-1) and (2.7.1-2), respectively.

Note that the Fisher information has two forms in (2.7.2-2) as well as in (2.7.2-4). The proof of (2.7.2-1) and the equivalence of the two forms in (2.7.2-2) — one with first partial derivatives and the other with second partials — is given later.

If an estimator's variance is equal to the CRLB, then such an estimator is called ***efficient***.

### Multidimensional Case

For *nonrandom vector parameters*, the CRLB states that the covariance matrix of an unbiased estimator is bounded from below as follows:

$$\boxed{E[[\hat{x}(Z) - x_0][\hat{x}(Z) - x_0]'] \geq J^{-1}} \tag{2.7.2-5}$$

where the ***Fisher information matrix (FIM)*** is

$$\boxed{J \triangleq -E[\nabla_x \nabla_x' \ln \Lambda(x)] \,|_{x=x_0} = E[[\nabla_x \ln \Lambda(x)][\nabla_x \ln \Lambda(x)]']|_{x=x_0}} \tag{2.7.2-6}$$

and $x_0$ is the true value of the vector parameter $x$.

As in the scalar case, note the two forms of the FIM: one with the Hessian of the log-likelihood function and the other with the dyad of its gradient.

The matrix inequality in (2.7.2-5) is to be interpreted as follows:

$$A \geq B \qquad \Longleftrightarrow \qquad C \triangleq A - B \geq 0 \tag{2.7.2-7}$$

that is, the difference $C$ of the two matrices is positive semidefinite.

A similar expression holds for the case of a multidimensional random parameter.

### Remarks

The FIM can be seen as a quantification of the (maximum) ***existing information*** in the data about a parameter. Efficiency amounts to the ***extracted information*** being equal to the existing one, i.e., all the information has been extracted.

A necessary condition for an estimator to be consistent in the mean square sense is that there must be an increasing amount of information (in the sense of Fisher) about the parameter in the measurements — the Fisher information has to tend to infinity as $k \to \infty$. Then the CRLB converges to zero as $k \to \infty$ and thus the variance can also converge to zero.

### Note

For estimators that are *biased*, there is a modified version of the CRLB (e.g., [Van Trees68]).

## 2.7.3 Proof of the Cramer-Rao Lower Bound

Let $\hat{x}(z)$ be an unbiased estimate of the nonrandom real-valued parameter $x$ based on the observation (or set of observations) denoted now as $z$. The likelihood function of $x$ is

$$\Lambda(x) = p(z|x) \tag{2.7.3-1}$$

It will be assumed that the first and second derivatives of (2.7.3-1) with respect to $x$ exist and are absolutely integrable.

From the unbiasedness condition on the estimate $\hat{x}(z)$, one has (the true value is denoted now also as $x$)

$$E[\hat{x}(z) - x] = \int_{-\infty}^{\infty} [\hat{x}(z) - x] p(z|x)\, dz = 0 \tag{2.7.3-2}$$

The derivative of the above with respect to $x$ is

$$\begin{aligned} \frac{d}{dx} \int_{-\infty}^{\infty} [\hat{x}(z) - x] p(z|x)\, dz &= \int_{-\infty}^{\infty} \frac{\partial}{\partial x} \Big\{ [\hat{x}(z) - x] p(z|x) \Big\}\, dz \\ &= -\int_{-\infty}^{\infty} p(z|x)\, dz + \int_{-\infty}^{\infty} [\hat{x}(z) - x] \frac{\partial p(z|x)}{\partial x}\, dz \\ &= 0 \end{aligned} \tag{2.7.3-3}$$

Using the fact that the first integral in the last line above is equal to unity and the identity

$$\frac{\partial p(z|x)}{\partial x} = \frac{\partial \ln p(z|x)}{\partial x} p(z|x) \tag{2.7.3-4}$$

yields from (2.7.3-3)

$$\int_{-\infty}^{\infty} [\hat{x}(z) - x] \frac{\partial \ln p(z|x)}{\partial x} p(z|x)\, dz = 1 \tag{2.7.3-5}$$

Equation (2.7.3-5) can be rewritten as

$$\int_{-\infty}^{\infty} \Big\{ [\hat{x}(z) - x] \sqrt{p(z|x)} \Big\} \left\{ \frac{\partial \ln p(z|x)}{\partial x} \sqrt{p(z|x)} \right\} dz = 1 \tag{2.7.3-6}$$

The ***Schwarz inequality*** for real-valued functions, which is a generalized version of (1.3.4-2), is

$$|\langle f_1, f_2 \rangle| \triangleq \int_{-\infty}^{\infty} f_1(z) f_2(z)\, dz \leq \|f_1\|\, \|f_2\| \tag{2.7.3-7}$$

where

$$\|f_i\| \triangleq \{\langle f_i, f_i \rangle\}^{1/2} = \left\{ \int_{-\infty}^{\infty} f_i(z)^2\, dz \right\}^{1/2} \tag{2.7.3-8}$$

The equality in (2.7.3-7) holds if and only if

$$f_1(z) = c f_2(z) \qquad \forall z \tag{2.7.3-9}$$

Note that the left-hand side of (2.7.3-6) is an inner product of two functions as in (2.7.3-7). Using (2.7.3-7) to majorize the left-hand side of (2.7.3-6) yields

$$\left\{ \int_{-\infty}^{\infty} [\hat{x}(z) - x]^2 p(z|x)\, dz \right\}^{1/2} \left\{ \int_{-\infty}^{\infty} \left[ \frac{\partial \ln p(z|x)}{\partial x} \right]^2 p(z|x)\, dz \right\}^{1/2} \geq 1 \tag{2.7.3-10}$$

which can be rewritten as

$$E\{[\hat{x}(z) - x]^2\} \geq \left\{ E\left[ \frac{\partial \ln p(z|x)}{\partial x} \right]^2 \right\}^{-1} \tag{2.7.3-11}$$

with equality holding if and only if

$$\frac{\partial \ln p(z|x)}{\partial x} = c(x)[\hat{x}(z) - x] \qquad \forall z \tag{2.7.3-12}$$

Equation (2.7.3-11) is equivalent to (2.7.2-1), which completes the proof of the CRLB for a nonrandom scalar parameter. In view of (2.7.3-2), which holds at the true value of the parameter, all the partial derivatives are to be evaluated at the true value of the parameter, which is indicated explicitly only in (2.7.2-2).

### Equivalence of the Two Forms of the Fisher Information

To prove the equivalence of the two forms of the Fisher information in (2.7.2-2), consider the identity

$$\int_{-\infty}^{\infty} p(z|x)\, dz = 1 \tag{2.7.3-13}$$

Taking the derivative of the above with respect to $x$ yields

$$\int_{-\infty}^{\infty} \frac{\partial p(z|x)}{\partial x}\, dz = 0 \tag{2.7.3-14}$$

Using identity (2.7.3-4), the above can be rewritten as

$$\int_{-\infty}^{\infty} \frac{\partial \ln p(z|x)}{\partial x} p(z|x)\, dz = 0 \tag{2.7.3-15}$$

Taking now the derivative of (2.7.3-15) with respect to $x$ leads to

$$\int_{-\infty}^{\infty} \frac{\partial^2 \ln p(z|x)}{\partial x^2} p(z|x)\, dz + \int_{-\infty}^{\infty} \left[ \frac{\partial \ln p(z|x)}{\partial x} \right]^2 p(z|x)\, dz = 0 \tag{2.7.3-16}$$

which proves the equivalence of the expression of the Fisher information with the second partial derivative of the log-likelihood function with the one that has the square of the first partial derivative, as in (2.7.2-2).

## 2.7.4 An Example of Efficient Estimator

Consider the likelihood function (2.4.2-4) for the estimation of the mean $x$ from a set of $k$ independent and identically distributed measurements with Gaussian noises.

The Fisher information in this case is the scalar quantity

$$J = -E\left[\frac{\partial^2 \ln \Lambda_k(x)}{\partial x^2}\right]\Bigg|_{x=x_0} = \frac{k}{\sigma^2} \tag{2.7.4-1}$$

Thus

$$E\left[[\hat{x}^{\mathrm{ML}}(k) - x_0]^2\right] \geq J^{-1} = \frac{\sigma^2}{k} \tag{2.7.4-2}$$

Comparing the above to (2.6.3-1), it is seen that the CRLB is met; that is, the ML estimator (which, in this case, is the sample mean) is efficient. This is because condition (2.7.3-12) is satisfied.

Since the variance (2.7.4-2) converges to zero as $k \to \infty$, this estimator is also consistent. (See also problem 2-2.)

### Evaluation of the CRLB

In the simple case considered above, the expression of $J$ is independent of $x$. In general this is not true and there is need to evaluate $J$ at the *true value* of $x$.

If the true value of the parameter is not available, then the ***evaluation of the CRLB***, which amounts to a linearization, is done *at the estimate.* Caution has to be exercised in this case, since the unavoidable estimation errors can lead to a *possibly incorrect value* of the resulting Fisher information $J$, which, in general, is a matrix.

## 2.7.5 Large Sample Properties of the ML Estimator

The following are the ***large-sample properties of the ML estimator***:

1. It is ***asymptotically unbiased.***
2. It is ***asymptotically efficient.***

Thus, if there is "enough information" in the measurements, in which case the CRLB will tend to zero, the variance of the ML estimate will also converge to zero. Therefore, the ML estimate will converge to the true value — it will be consistent.

Another property of the ML estimator is the following:

3. It is ***asymptotically Gaussian.***

Combining all the above, *the ML estimate is asymptotically Gaussian with the mean equal to the true value of the parameter to be estimated and variance given by the CRLB.*

This can be summarized for a vector parameter as

$$\hat{x}^{\mathrm{ML}}(k) \sim \mathcal{N}(x, J^{-1}) \qquad \text{for large } k \tag{2.7.5-1}$$

where $J$ is the Fisher information matrix.

Comparing the (non-Bayesian) MLE (2.7.5-1) with the (Bayesian) MMSE estimate, the conditional mean

$$\hat{x}^{\text{MMSE}} = E[x|z] \tag{2.7.5-2}$$

points out the contrast between these two philosophies:

1. In (2.7.5-1), given $x$, the estimate $\hat{x}^{\text{ML}}$ is a random variable, function of $z$;
2. In (2.7.5-2), given $z$, the true value $x$ is a random variable.

## 2.8 SUMMARY

### 2.8.1 Summary of Estimators

***Estimator of a parameter*** — a function of the measurements that yields a "best approximation" for the value of a parameter.

***Estimate of a parameter*** — the value taken by the estimator for the given values (realizations) of the measurements.

Models for the parameter to be estimated:

1. *Unknown constant (nonrandom).*
2. *Random*: a (single) realization of a random variable according to a certain prior pdf.

Model 2 yields the ***Bayesian approach***, whereas model 1 leads to what is called the ***non-Bayesian approach***.

***Likelihood function*** of a (nonrandom) parameter — pdf of the measurements conditioned on the parameter.

***Bayes' formula*** — given a prior pdf of a (random) parameter, this formula yields its posterior pdf conditioned on the measurements.

***ML estimate*** (of a nonrandom parameter) — the value of the parameter that maximizes its likelihood function.

***MAP estimate*** (of a random parameter) — the value of the parameter that maximizes its posterior pdf.

The MAP estimate of a parameter with a *diffuse (noninformative)* prior pdf coincides with its MLE.

***LS*** estimate (of a nonrandom parameter) — minimizes the sum of the squares of the errors between the measurements and the observed function of the parameter.

***MMSE*** estimate (of a random parameter) — minimizes the expected value (mean) of the square of the parameter estimation error conditioned on the measurements. This estimate is the *conditional mean* of the parameter given the measurements.

If in a given set of measurements the errors are additive, zero mean, *Gaussian*, and independent, then the *LS* estimate coincides with the *ML* estimate.

The *MAP* estimate of a *Gaussian* random variable coincides with its *MMSE* estimate (conditional mean).

### 2.8.2 Summary of Estimator Properties

***Unbiased estimator*** — if the mean of the corresponding error is zero.

***Variance/MSE of an estimator*** — the expected value of the square of the estimation error of an unbiased/biased estimator. The variance of the estimator of a parameter modeled as random (with some prior) is *smaller* than when it is modeled as an unknown constant.

***Consistent estimator*** — if the corresponding error converges to zero in some stochastic sense (most common: in mean square).

***CRLB*** — lower bound on the achievable variance in the estimation of a parameter. For an unbiased estimator, it is given by the inverse of the ***Fisher information matrix (FIM)***.

***FIM*** — quantifies the existing total information about the parameter of interest in the observations.

***Efficient estimator*** — if its variance meets the CRLB, that is, if all the existing information has been extracted.

### On the Terminology

In (most of) the literature there is little or no distinction between the terms LS and MMSE estimation. The MMSE estimation is sometimes called LS, which is incorrect according to our definition, or ***least mean square (LMS)***, which is a valid alternate designation. Another term used is ***minimum variance (MV)***.

## 2.9 NOTES AND PROBLEMS

### 2.9.1 Bibliographical Notes

The basic concepts in estimation are discussed, for example, in [Van Trees68, Sage71, Melsa78]. The proof of the CRLB for vector-valued parameters can be found in [Van Trees68] and [Ljung87, p. 206].

Another model of uncertainty in parameter estimation is the "unknown but bounded" approach discussed in [Schweppe73].

## 2.9.2 Problems

**2-1 Estimators for a discrete-valued parameter.** A discrete-valued parameter with the prior pdf

$$p(x) = \sum_{i=1}^{2} p_i \delta(x - i)$$

is measured with the additive noise $w \sim \mathcal{N}(0, \sigma^2)$

$$z = x + w$$

1. Find the posterior pdf of the parameter.
2. Find its MAP estimate and the associated MSE conditioned on $z$.
3. Find its MMSE estimate and the associated variance.
4. Evaluate these estimates and MSE for

| Case | $p_1$ | $\sigma$ | $z$ |
|---|---|---|---|
| A | 0.5 | 1 | 1.5 |
| B | 0.5 | 1 | 3 |
| C | 0.3 | 1 | 1.5 |
| D | 0.5 | 0.1 | 1.8 |

5. Comment on the meaningfulness of the two estimates in the above four cases.

**2-2 ML Estimation with correlated noises.** A parameter $x$ is measured with correlated rather than independent additive Gaussian noises

$$z_k = x + w_k \qquad k = 1, \ldots, n$$

with

$$E[w_k] = 0 \qquad E[w_k w_j] = \begin{cases} 1 & k = j \\ \rho & |k - j| = 1 \\ 0 & |k - j| > 1 \end{cases}$$

For $n = 2$:

1. Write the likelihood function of the parameter $x$.
2. Find the MLE of $x$. What happens if $\rho = 1$? What happens if $\rho = -1$?
3. Find the CRLB for the estimation of $x$. Show the effect of $\rho > 0$ versus $\rho < 0$. Explain what happens at $\rho = -1$.
4. Is the MLE efficient? Can one have a perfect (zero-variance) estimate?

(The remaining items are more challenging.) For general $n$, let

$$z \triangleq [z_1 \ \cdots \ z_n]' \qquad \mathbf{1} \triangleq [1 \ \cdots \ 1]' \qquad w \triangleq [w_1 \ \cdots \ w_n]' \qquad P \triangleq E[ww']$$

5. Using the above notations, write the likelihood function of $x$.
6. Find the MLE of $x$.

**2-3 Estimation criteria that lead to the conditional mean.** Show that, in estimating a random vector $x$ with the following criteria

1. $\min_{\hat{x}} E[(x-\hat{x})'A(x-\hat{x})|z], \quad \forall A > 0$ (positive definite)
2. $\min_{\hat{x}} \text{tr}[P]$ with $P \triangleq E[(x-\hat{x})(x-\hat{x})'|z]$
3. $\min_{\hat{x}} \text{tr}[AP]$ with $A$ and $P$ as above

all yield the same result $\hat{x} = E[x|z]$.

**2-4 Estimate of the variance with the smallest MSE.** Consider the problem of estimating the mean and the variance of a set of independent and identically distributed Gaussian random variables $z(j)$, $j = 1, \ldots, k$, as in Subsection 2.5.3, with the true mean $x_0$ and true variance $\sigma_0^2$.

1. Show that the value of $n$ in

$$[\hat{\sigma}(k,n)]^2 = \frac{1}{n}\sum_{j=1}^{k}\left[z(j) - \frac{1}{k}\sum_{i=1}^{k} z(i)\right]^2$$

that minimizes the MSE of the above (defined according to (2.6.1-4) with respect to $\sigma_0^2$) is $n = k+1$.

2. Can a biased estimator have a smaller MSE than an unbiased one?

**2-5 MAP estimate with two-sided exponential (Laplacian) prior pdf.** Consider the same problem as in Subsection 2.3.2 but with a two-sided exponential prior

$$p(x) = \frac{a}{2}\, e^{-a|x|}$$

1. Write the posterior pdf of $x$.
2. Find $\hat{x}^{\text{MAP}}$.

**2-6 Two-sided exponential prior made diffuse.**

1. Specify the limiting process that will make the prior from problem 2-5 into a diffuse one.
2. Show that the resulting MAP estimate coincides with the MLE.

**2-7 Minimum magnitude error estimate.** Given $p(x,z)$, show that the Bayesian estimation that minimizes the expected value of the cost function

$$C(x-\hat{x}) \triangleq |x-\hat{x}|$$

yields $\hat{x} = x_m$, the ***median*** of $x$, defined as

$$\int_{-\infty}^{x_m} p(x|z)\, dx = \frac{1}{2}$$

**2-8 MAP with Gaussian prior — vector version.** Given $z = x + w$, where all the variables are $n$-vectors, with

$$w \sim \mathcal{N}(0, P) \qquad\qquad x \sim \mathcal{N}(\bar{x}, P_0)$$

and $x$ independent of $w$. Find the MAP estimator of $x$ in terms of $z$ and the covariance of this estimator.

**2-9 Conditional variance versus unconditional variance.** Let

$$\bar{x} \triangleq E[x] \qquad\qquad \text{var}(x) \triangleq E[(x - \bar{x})^2]$$

$$\hat{x} \triangleq E[x|z] \qquad\qquad \text{var}(x|z) \triangleq E[(x - \hat{x})^2|z]$$

Prove that

$$\text{var}(x) \geq E[\text{var}(x|z)]$$

**2-10 MMSE with exponential prior.** Given the prior pdf of $x$ as $p(x) = e^{-x}$, $x \geq 0$ and the observation $z = x + w$ where $w \sim \mathcal{N}(0, 1)$ and independent of $x$, find the following:

1. $p(z|x)$
2. $p(x|z)$
3. $E[x|z]$
4. $\text{var}[x|z]$

**2-11 Altitude estimation from slant range.** A sensor is located at $(0, 0)$. It is desired to estimate the height (altitude) $y$ of a point (an aircraft) located at $(d, y)$, where $d$ (the horizontal range) is known, based on the "slant range" measurement $z = r + w$ where $r = h(y) = \sqrt{d^2 + y^2}$ and $w \sim \mathcal{N}(0, \sigma^2)$.

1. Write the likelihood function of $y$.
2. Find the CRLB for estimating $y$.
3. Evaluate the standard deviation of the estimate according to the CRLB for $d = 10^5$, $\sigma = 10^2$, and assumed true value $y = 10^3$. How useful would such an estimate be?
4. Find the expression of the MLE of $y$ in terms of $z$ and $d$.

**2-12 Superefficiency?** A scalar parameter is estimated in $N$ Monte Carlo runs. The CRLB for this problem is $\sigma^2_{CRLB} = 10$. The sample variance obtained from $N = 100$ runs is $\hat{\sigma}^2$=7.811. Your best friend is concerned and tells you that you must have made a mistake somewhere. What is your answer to this? Give a quantitative justification for it — you can make any reasonable assumptions.

**2-13 MLE from correlated measurements.** Given the three estimates of the scalar $x$

$$z_i \triangleq \hat{x}_i = x + \tilde{x}_i \qquad i = 1, 2, 3$$

with the estimation errors $\tilde{x}_i$ jointly Gaussian, zero-mean, with

$$E[\tilde{x}_i \tilde{x}_j] = P_{ij} \qquad i, j = 1, 2, 3$$

find

1. The MLE $\hat{x}(z_1, z_2, z_3)$
2. The variance $\sigma^2$ of the above MLE

**2-14 Measurement error variance.** We have a measuring device for which we are to ascertain that its measurement error variance is less than the borderline acceptable value of 100. You can assume that the errors are zero-mean.

We carry out $N$ independent trials and the result is that the estimate of the variance is $\hat{\sigma}^2 = 80$.

For what $N$ are you willing to risk your job by saying that there is a 5% (or less) probability to get such an estimate while the true value is $\sigma^2 = 100$ (or higher)?

*Hint*: You can assume a Gaussian distribution for the error in the estimate of the variance and use a probability region about the borderline value.

**2-15 Estimation with correlated measurements.** Given the scalar random variables $x_i$, $i = 1, \ldots, N+1$, with $Ex_i = \mu$ and $\text{cov}[x_i, x_j] = \sigma^2 \rho^{|i-j|}$, where $|\rho < 1$.

1. Find the mean and variance of the following estimate of $\mu$

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

You can assume $N \gg 1$.

2. Is this a consistent estimator? Justify the answer.

3. To how many i.i.d. random variables with the same first two moments are the above equivalent (i.e., they yield the same variance) for $\rho = 0.5$?

**2-16 Public opinion polls.**

1. Find the margin of error of a public opinion poll with $N = 625$ subjects.
2. How many subjects are needed for a margin of error of 1%?