

# CRB and MVU Estimation

## Outline:

- Cramér-Rao bound (CRB).
- Constructing minimum-variance unbiased (MVU) estimators.

**Reading:** Kay-I, Chapter 3.

# CRB

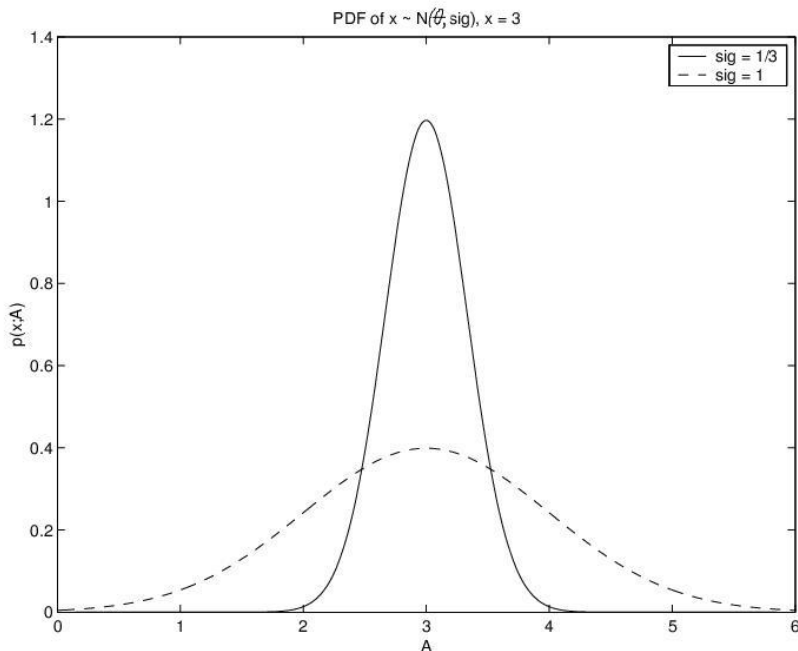
How accurately we can estimate a parameter  $\theta$  depends on the pdf or pmf of the observation  $X$  given  $\theta$  (i.e. on the likelihood function of  $\theta$ ).

## Example (Kay-I, Chapter 3):

$$X = \theta + W$$

where  $\theta$  is the unknown parameter and

$$W \sim \mathcal{N}(0, \sigma^2).$$



Intuitively, sharpness of the pdf (pmf)  $f_{X|\theta}(x|\theta)$  ( $p_{X|\theta}(x|\theta)$ ) determines how accurately we can estimate  $\theta$ .

## Cramér-Rao Bound: Regularity Conditions

We make two assumptions on  $f_{X|\theta}(x|\theta)$ :

(i) The support set of  $f_{X|\theta}(x|\theta)$ :

$$A = \{x \mid f_{X|\theta}(x|\theta) > 0\}$$

*does not depend on  $\theta$* . For all  $x \in A$  and  $\theta$  in the parameter space  $\Theta$ ,  $\partial/\partial\theta[\ln f_{X|\theta}(x|\theta)]$  exists and is finite.

(ii) If  $T(x)$  is any statistic satisfying  $E_{X|\theta}(|T(X)||\theta) < \infty$  for all  $\theta \in \Theta$ , then integration over  $x$  and differentiation by  $\theta$  can be interchanged in  $\int T(x) p(x; \theta) dx$ , i.e.

$$\frac{\partial}{\partial\theta} \left[ \int T(x) f_{X|\theta}(x|\theta) dx \right] = \int T(x) \frac{\partial}{\partial\theta} f_{X|\theta}(x|\theta) dx \quad (1)$$

whenever the right-hand side is finite. In particular, (1) should hold for  $T(x) = 1$ : we will use this special case in Lemma 1.

**Note:** Checking assumption (ii) is not very practical. We need simple sufficient conditions on  $f_{X|\theta}(x|\theta)$  so that (ii) holds. The assumption (ii) is coupled with (i): if (i) does not hold, it does not make sense to talk about changing the order of integration and differentiation with respect to  $\theta$ . The following

proposition describes a family of distributions that satisfy (i) and (ii).

**Proposition 1. (One-parameter exponential family of distributions)** *If*

$$f_{X|\theta}(x|\theta) = h(x) \exp[\eta(\theta) T(x) - B(\theta)] \quad (2)$$

and  $\eta(\theta)$  has a nonvanishing continuous derivative on the parameter space  $\Theta$ , then (i) and (ii) hold.

If (i) holds, it is possible to define an important characteristic of  $f_{X|\theta}(x|\theta)$ , the *Fisher information*  $\mathcal{I}(\theta)$ :

$$\begin{aligned} \mathcal{I}(\theta) &= \mathbb{E}_{X|\theta} \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_{X|\theta}(X|\theta) \right]^2 \middle| \theta \right\} \\ &= \int \left[ \frac{\partial}{\partial \theta} \ln f_{X|\theta}(x|\theta) \right]^2 f_{X|\theta}(x|\theta) dx. \end{aligned} \quad (3)$$

Note that  $0 \leq \mathcal{I}(\theta) \leq \infty$ .

**Terminology:**

$$\frac{\partial}{\partial \theta} \ln f_{X|\theta}(x|\theta)$$

is known as the *score function* for the parameter  $\theta$  and data  $x$ .

**Lemma 1.** *Suppose that (i) and (ii) hold and that*

$$\mathbb{E}_X \left| \frac{\partial}{\partial \theta} \ln f_{X|\theta}(X|\theta) \right| < \infty.$$

Then

$$\mathbb{E}_{X|\theta} \left( \frac{\partial}{\partial \theta} \ln f_{X|\theta}(X|\theta) \mid \theta \right) = 0$$

and, thus,

$$\mathcal{I}(\theta) = \text{var}_{X|\theta} \left( \frac{\partial}{\partial \theta} \ln f_{X|\theta}(X|\theta) \mid \theta \right). \quad (4)$$

**Proof.**

$$\begin{aligned} & \mathbb{E}_{X|\theta} \left( \frac{\partial}{\partial \theta} \ln f_{X|\theta}(X|\theta) \mid \theta \right) \\ &= \int \left\{ \left[ \frac{\partial}{\partial \theta} f_{X|\theta}(x|\theta) \right] / f_{X|\theta}(x|\theta) \right\} f_{X|\theta}(x|\theta) dx \\ &= \int \frac{\partial}{\partial \theta} f_{X|\theta}(x|\theta) dx = \frac{\partial}{\partial \theta} \int f_{X|\theta}(x|\theta) dx = 0. \end{aligned}$$

Here, we have utilized the chain rule of differentiation:

$$\frac{df(p(z))}{dz} = \frac{\partial f(w)}{\partial w} \Big|_{w=p(z)} \cdot \frac{dp(z)}{dz}$$

with  $f(\cdot) = \ln(\cdot)$ .  $\square$

**Comments:**

- We have just shown that the score function has mean zero and variance equal to the Fisher information  $\mathcal{I}(\theta)$ .

- The score function is zero when we equate  $\theta$  in it to the ML estimator of  $\theta$ .

**Example.** Suppose  $X[0], X[1], \dots, X[N - 1]$  are i.i.d. measurements from Poisson( $\lambda$ ) distribution:

$$p_{X|\lambda}(x[n]|\lambda) = \frac{\lambda^{x[n]} e^{-\lambda}}{x[n]!}$$

see the distribution table. Then

$$p_{\mathbf{X}|\lambda}(\mathbf{x}|\lambda) = \prod_{n=0}^{N-1} \left( \frac{\lambda^{x[n]} e^{-\lambda}}{x[n]!} \right) = \frac{\lambda^{\sum_n x[n]} e^{-N\lambda}}{\prod_n x[n]!}$$

where  $\mathbf{x} = [x[0], x[1], \dots, x[N - 1]]^T$  and

$$\frac{\partial}{\partial \lambda} [\ln p_{\mathbf{X}|\lambda}(\mathbf{x}|\lambda)] = \frac{\sum_{n=0}^{N-1} x[n]}{\lambda} - N \quad (5)$$

$$\begin{aligned} \mathcal{I}(\lambda) &\stackrel{\text{see (4)}}{=} \text{var}_{\mathbf{X}|\lambda} \left( \frac{\sum_{n=0}^{N-1} X[n]}{\lambda} \mid \lambda \right) \\ &= \frac{1}{\lambda^2} \text{var}_{\mathbf{X}|\lambda} \left( \sum_{n=0}^{N-1} X[n] \mid \lambda \right) \\ &\stackrel{X[n] \text{ i.i.d.}}{=} \frac{1}{\lambda^2} \cdot N \lambda = \frac{N}{\lambda}. \end{aligned} \quad (6)$$

Here, we have used the fact that, for  $\{X[n] | \lambda\} \sim \text{Poisson}(\lambda)$ ,

$$\text{var}_{X | \lambda}(X[n] | \lambda) = \lambda \quad (7)$$

see the distribution table.

**Theorem 1. (Information Inequality)** Consider statistics  $T(X)$  that satisfy

$$\text{var}_{X | \theta}[T(X) | \theta] < \infty$$

for all  $\theta$ . Suppose that (i) and (ii) hold and  $0 < \mathcal{I}(\theta) < \infty$ , where  $\mathcal{I}(\theta)$  is the Fisher information for  $\theta$ , defined in (3). Define

$$\mathbb{E}_{X | \theta}[T(X) | \theta] = \psi(\theta).$$

Then, for all  $\theta$ ,

$$\text{var}_{X | \theta}[T(X) | \theta] \geq \frac{|\psi'(\theta)|^2}{\mathcal{I}(\theta)} \quad (8)$$

where

$$\psi'(\theta) = \frac{d\psi(\theta)}{d\theta}.$$

**Proof.** Using (i) and (ii) leads to

$$\begin{aligned}\psi'(\theta) &= \frac{\partial}{\partial \theta} \int T(x) f_{X|\theta}(x|\theta) dx \\ &= \int T(x) \frac{\partial f_{X|\theta}(x|\theta)}{\partial \theta} dx \\ &= \int T(x) \frac{\partial \ln f_{X|\theta}(x|\theta)}{\partial \theta} f_{X|\theta}(x|\theta) dx \\ &= \mathbb{E}_{X|\theta} \left( T(X) \cdot \frac{\partial \ln f_{X|\theta}(X|\theta)}{\partial \theta} \right)\end{aligned}$$

and, therefore,

$$\psi'(\theta) = \text{cov}_{X|\theta} \left( \frac{\partial \ln f_{X|\theta}(X|\theta)}{\partial \theta}, T(X) \right).$$

(Recall that

$$\begin{aligned}\text{cov}_{P,Q}(P, Q) &\triangleq \mathbb{E}_{P,Q}[(P - \mathbb{E}_P[P]) (Q - \mathbb{E}_Q[Q])] \\ &\stackrel{\text{if } \mathbb{E}_P[P] = 0 \text{ or } \mathbb{E}_Q[Q] = 0}{=} \mathbb{E}_{P,Q}[P Q].\end{aligned}$$

Apply the Cauchy-Schwartz inequality

$$[\text{cov}_{P,Q}(P, Q)]^2 \leq \text{var}_P(P) \cdot \text{var}_Q(Q)$$

to the random variables  $\overbrace{\partial \ln f_{X|\theta}(X|\theta)/\partial\theta}^P$  and  $\overbrace{T(X)}^Q$ :

$$|\psi'(\theta)|^2 \leq \text{var}_{X|\theta}[T(X)|\theta] \cdot \text{var}_{X|\theta}\left(\frac{\partial \ln f_{X|\theta}(X|\theta)}{\partial\theta} \middle| \theta\right).$$

Theorem 1 follows by using Lemma 1.  $\square$

## Digression

It is instructive to derive the Cauchy-Schwartz inequality. First, remember that any covariance matrix needs to be positive semidefinite. Therefore,

$$\begin{array}{ccc}
 \downarrow & \text{COV}_{P,Q} \left( \begin{array}{c} P \\ Q \end{array} \right) & \downarrow \\
 \text{determinant} & \text{covariance matrix} & \text{determinant} \\
 = & \left| \begin{array}{cc} \text{var}_P(P) & \text{COV}_{P,Q}(P, Q) \\ \text{COV}_{P,Q}(P, Q) & \text{var}_Q(Q) \end{array} \right| & \geq 0
 \end{array}$$

and the Cauchy-Schwartz inequality follows.

But, why does a covariance matrix of  $\begin{bmatrix} P \\ Q \end{bmatrix}$  need to be positive semidefinite? Because the following holds for arbitrary  $a$  and  $b$ :

$$\text{var}_{P,Q}(aP + bQ) \geq 0$$

which can be rewritten as

$$[a, b] \text{COV}_{P,Q} \left( \begin{array}{c} P \\ Q \end{array} \right) \begin{array}{c} a \\ b \end{array} \geq 0 \quad \forall a, b$$

which, by the definition of positive semidefiniteness, implies that  $\text{COV}_{P,Q} \left( \begin{array}{c} P \\ Q \end{array} \right)$  is a positive semidefinite matrix.

## (Back to the Main Track) Comments

- If we view  $T(X)$  as a (generally biased) estimator of  $\theta$ , then

$$E_{X|\theta}[T(X) | \theta] = \psi(\theta) = \theta + \underbrace{b(\theta)}_{\text{bias}}$$

and (8) is a lower bound on the variance of  $T(X)$ :

$$\text{var}_{X|\theta}[T(X) | \theta] \geq \frac{|1 + b'(\theta)|^2}{\mathcal{I}(\theta)} \quad (9)$$

We can bound the MSE of  $T(X)$  as follows:

$$\begin{aligned} \text{MSE}\{T(X)\} &= \text{var}_{X|\theta}[T(X) | \theta] + b^2(\theta) \\ &\geq \frac{|1 + b'(\theta)|^2}{\mathcal{I}(\theta)} + b^2(\theta). \end{aligned} \quad (10)$$

In practice, this result may not be very useful, since it is typically hard to analytically compute bias.

- Since

$$E_{X|\theta}[T(X) | \theta] = \psi(\theta) \triangleq \psi$$

we can view  $T(X)$  as an unbiased estimator of  $\psi$ ; then (8) gives a lower bound on the variance of  $T(X)$ , expressed in terms of the Fisher information  $\mathcal{I}(\theta)$  for  $\theta$ .

The lower bound depends on  $T(X)$  through its expectation  $\psi(\theta)$ . If we consider *only the class of unbiased estimators  $T(X)$  of  $\theta$* , i.e.

$$\psi(\theta) = \theta$$

we obtain a *universal lower bound* on variance of such estimators, given by the following.

**Corollary 1.** *Suppose the conditions of Theorem 1 hold and that  $T(X)$  is an unbiased estimator of  $\theta$ , i.e.*

$$\mathbb{E}_{X|\theta}[T(X) | \theta] = \theta.$$

Then

$$\text{var}_{X|\theta}[T(X)] \geq \frac{1}{\mathcal{I}(\theta)}.$$

Here,

$$\text{CRB}(\theta) = \frac{1}{\mathcal{I}(\theta)} \quad (11)$$

is often referred to as the *Cramér-Rao bound (CRB)* on the variance of an unbiased estimator of  $\theta$ .

For one-to-one mappings  $\psi = \psi(\theta)$  and  $\theta = \theta(\psi)$ , the general result in (8) can be interpreted as a change-of-variables formula:

$$\text{CRB}(\psi) = |\psi'(\theta)|^2 \text{CRB}(\theta) \Big|_{\theta=\theta(\psi)}. \quad (12)$$

**Proposition 2.** *Suppose that  $f_{X|\theta}(x|\theta)$  satisfies, in addition to (i) and (ii), the following condition:*

(iii)  $f_{X|\theta}(x|\theta)$  is twice differentiable and it is permitted to interchange integration (with respect to  $x$ ) and differentiation (with respect to  $\theta$ ).

Then

$$\mathcal{I}(\theta) = -\mathbb{E}_{X|\theta} \left[ \frac{\partial^2}{\partial \theta^2} \ln f_{X|\theta}(X|\theta) \right]. \quad (13)$$

**Proof.**

$$\begin{aligned} \frac{\partial^2}{\partial \theta^2} \ln f_{X|\theta}(x|\theta) &= \frac{1}{f_{X|\theta}(x|\theta)} \cdot \frac{\partial^2}{\partial \theta^2} f_{X|\theta}(x|\theta) \\ &\quad - \left[ \frac{\partial}{\partial \theta} \ln f_{X|\theta}(x|\theta) \right]^2 \end{aligned}$$

and apply expectation with respect to  $X$  given  $\theta$  to both sides, i.e. multiply by  $f_{X|\theta}(x|\theta)$  and integrate.  $\square$

The above results provides another way of computing the Fisher information, which may be more convenient than taking the expectation of the squares score function. The exponential family of distributions satisfies (iii).

**Example.** Back to the Poisson example and apply Proposition

2:

$$\begin{aligned}\mathcal{I}(\lambda) &= \mathbb{E}_{\mathbf{X}|\lambda} \left[ -\frac{\partial^2}{\partial \lambda^2} \ln f_{\mathbf{X}|\lambda}(\mathbf{X}|\lambda) \mid \lambda \right] \\ &\stackrel{\text{diff. (5)}}{=} \frac{1}{\lambda^2} \mathbb{E}_{\mathbf{X}|\lambda} \left( \sum_{n=0}^{N-1} X[n] \mid \lambda \right) = \frac{N}{\lambda}\end{aligned}$$

which is easier than the derivation on p. 6. Here, we have used the fact that, for  $\{X[n] \mid \lambda\} \sim \text{Poisson}(\lambda)$ ,

$$\mathbb{E}_{X|\lambda}(X[n] \mid \lambda) = \lambda$$

see the distribution table.

In this example,

$$\bar{X} = \frac{1}{N} \sum_{n=0}^{N-1} X[n] \quad (\text{sample mean})$$

is the ML estimator of  $\lambda$  and it is unbiased. In the Poisson case, we have

$$\text{var}_{X|\lambda}(X[n] \mid \lambda) = \lambda$$

see also (7). Then,

$$\begin{aligned}
 \text{var}_{\mathbf{X} | \lambda}(\bar{X} | \lambda) &= \text{var}_{\mathbf{X} | \lambda} \left( \frac{\sum_{n=0}^{N-1} X[n]}{N} \mid \lambda \right) \\
 &= \frac{1}{N^2} \text{var}_{\mathbf{X} | \lambda} \left( \sum_{n=0}^{N-1} X[n] \mid \lambda \right) \\
 &\stackrel{\text{X}[n] \text{ i.i.d.}}{=} \frac{1}{N^2} \cdot N \lambda = \frac{\lambda}{N} \\
 &\stackrel{\text{see (6)}}{=} \frac{1}{\mathcal{I}(\lambda)} = \text{CRB}(\lambda)
 \end{aligned}$$

and, by Corollary 1,  $\bar{X}$  is minimum-variance unbiased (MVU) estimator of  $\lambda$ .

**Example:** Continue with the same Poisson example, but consider unbiased estimators  $T(\mathbf{X})$  of  $\psi = \lambda^2$ . Here,

$$\psi(\lambda) = \lambda^2 \quad \text{and, therefore,} \quad \psi'(\lambda) = 2\lambda$$

and

$$\begin{aligned}
 \text{var}_{\mathbf{X} | \lambda}[T(\mathbf{X}) | \lambda] &\stackrel{\text{see (8)}}{\geq} \frac{|\psi'(\lambda)|^2}{\mathcal{I}(\lambda)} = \frac{4\lambda^2}{N/\lambda} = \frac{4\lambda^3}{N} \\
 \text{CRB}(\psi) &\stackrel{\text{see (12)}}{=} |\psi'(\lambda)|^2 \text{CRB}(\lambda) = \frac{4\lambda^3}{N} = \frac{4\psi^{3/2}}{N}.
 \end{aligned}$$

# One-parameter Canonical Exponential Family

On p. 34 of handout # 1, we introduced the one-parameter canonical exponential family:

$$f_{\mathbf{X}|\eta}(\mathbf{x}|\eta) = h(\mathbf{x}) \exp [\eta T(\mathbf{x}) - A(\eta)] \quad (14)$$

and stated its property:

$$\mathbb{E}_{\mathbf{X}|\eta}[T(\mathbf{X})|\eta] = \frac{dA(\eta)}{d\eta}, \quad \text{var}_{\mathbf{X}|\eta}[T(\mathbf{X})|\eta] = \frac{d^2 A(\eta)}{d\eta^2} \quad (15)$$

see (14) in handout # 1. Now

$$\begin{aligned} \frac{\partial \ln f_{\mathbf{X}|\eta}(\mathbf{x}|\eta)}{\partial \eta} &= T(\mathbf{x}) - \frac{dA(\eta)}{d\eta} \\ -\frac{\partial^2 \ln f_{\mathbf{X}|\eta}(\mathbf{x}|\eta)}{\partial \eta^2} &= \frac{d^2 A(\eta)}{d\eta^2} \end{aligned}$$

and apply Proposition 2, which we can do since the exponential family (14) satisfies (i)–(iii):

$$\mathcal{I}(\eta) = \frac{d^2 A(\eta)}{d\eta^2} \quad (16)$$

see (13). Apply (15) and combine with (16):

$$\text{var}_{\mathbf{X}|\eta}[T(\mathbf{X})|\eta] = \frac{d^2 A(\eta)}{d\eta^2} = \mathcal{I}(\eta). \quad (17)$$

We now apply the change-of-variables formula (12) with  $\psi$  and  $\eta$  in place of  $\psi$  and  $\theta$ , where

$$\psi = \frac{dA(\eta)}{d\eta}$$

and, therefore,

$$\text{CRB}(\psi) = |\psi'(\eta)|^2 \text{CRB}(\eta) = \frac{|\psi'(\eta)|^2}{\mathcal{I}(\eta)}$$

and

$$\mathbb{E}_{\mathbf{X} | \psi}[T(\mathbf{X}) | \psi] = \psi.$$

Now,

$$\begin{aligned} \text{CRB}(\psi) &= \left| \frac{d^2 A(\eta)}{d\eta^2} \right|^2 / \mathcal{I}(\eta) \\ &= \left| \frac{d^2 A(\eta)}{d\eta^2} \right|^2 / \frac{d^2 A(\eta)}{d\eta^2} = \frac{d^2 A(\eta)}{d\eta^2} \\ &\stackrel{\text{see (17)}}{=} \text{var}_{\mathbf{X} | \eta}[T(\mathbf{X}) | \eta] = \text{var}_{\mathbf{X} | \psi}[T(\mathbf{X}) | \psi] \end{aligned}$$

implying that  $T(\mathbf{X})$  is an efficient estimator of its expectation

$$\mathbb{E}_{\mathbf{X} | \psi}[T(\mathbf{X}) | \psi] = \psi = \mathbb{E}_{\mathbf{X} | \eta}[T(\mathbf{X}) | \eta] = \frac{dA(\eta)}{d\eta}$$

*and*

$$\text{CRB}(\psi) = \text{var}_{\mathbf{X} | \psi}[T(\mathbf{X}) | \psi] = \frac{d^2 A(\eta)}{d\eta^2} \Big|_{\eta=\eta(\psi)}.$$

# Fisher Information for I.I.D. Measurements

**Corollary 2.** *Suppose that the elements of  $\mathbf{X} = [X[0], X[1], \dots, X[N - 1]]^T$  are i.i.d. with density  $f_{X|\theta}(x|\theta)$  and that the conditions (i) and (ii) hold. Define the contribution of a single measurement ( $X[0]$ , say) to the Fisher information:*

$$\mathcal{I}_1(\theta) = \mathbb{E}_{X|\theta} \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_{X|\theta}(X[0]|\theta) \right]^2 \right\}.$$

*(Here, we arbitrarily pick  $X[0]$  as our single measurement, its contribution to the Fisher information is equal to that of  $X[1]$  etc.) Then, the Fisher information  $\mathcal{I}(\theta)$  for  $\theta$  based on all observations  $\mathbf{x}$  is easy to compute using  $\mathcal{I}_1(\theta)$ :*

$$\mathcal{I}(\theta) = N \mathcal{I}_1(\theta) \quad \text{and} \quad \text{var}_{\mathbf{X}|\theta}[T(\mathbf{X})] \geq \frac{|\psi'(\theta)|^2}{N \mathcal{I}_1(\theta)}.$$

## Proof.

$$\begin{aligned}\mathcal{I}(\theta) &= \text{var}_{\mathbf{X}|\theta} \left[ \frac{\partial}{\partial \theta} \ln f_{\mathbf{X}|\theta}(\mathbf{X}|\theta) \right] \\ &= \text{var}_{X|\theta} \left[ \sum_{n=0}^{N-1} \frac{\partial}{\partial \theta} \ln f_{X|\theta}(X[n]|\theta) \right] \\ &\stackrel{X[n] \text{ i.i.d.}}{=} \sum_{n=0}^{N-1} \text{var}_{X|\theta} \left[ \frac{\partial}{\partial \theta} \ln f_{X|\theta}(X[n]|\theta) \right] = N \mathcal{I}_1(\theta).\end{aligned}$$

□

**Example:** Suppose that  $X[0], X[1], \dots, X[N-1]$  are i.i.d. observations from  $\mathcal{N}(\mu, \sigma^2)$ , conditional on  $\mu$ . Here,  $\mu$  is the unknown parameter and  $\sigma^2$  is a known constant. Note that the conditions (i) and (ii) hold, since  $\mathcal{N}(\mu, \sigma^2)$  with parameter  $\mu$  is a member of the exponential family of distributions. Then

$$\begin{aligned}\ln f_{X|\mu}(x[0]|\mu) &= -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (x[0] - \mu)^2 \\ \frac{\partial \ln f_{X|\mu}(x[0]|\mu)}{\partial \mu} &= \frac{x[0] - \mu}{\sigma^2}\end{aligned}$$

and

$$\begin{aligned}\mathcal{I}_1(\mu) &= \mathbb{E}_{X|\mu} \left[ \left( \frac{\partial \ln f_{X|\mu}(X[0]|\mu)}{\partial \mu} \right)^2 \right] \\ &= \mathbb{E}_{X|\mu} \left[ \left( \frac{X[0] - \mu}{\sigma^2} \right)^2 \right] = \frac{1}{\sigma^2}.\end{aligned}$$

Since the exponential family of distributions satisfies (iii) as well, we could differentiate the score function with respect to  $\mu$ :

$$\frac{\partial^2 \ln f_{X|\mu}(x[0]|\mu)}{\partial \mu^2} = -\frac{1}{\sigma^2}$$

and apply (13) in Proposition 2:

$$\mathcal{I}_1(\mu) = -\mathbb{E}_{X|\mu} \left[ \frac{\partial^2}{\partial \mu^2} \ln f_{X|\mu}(X[0]|\mu) \right] = \frac{1}{\sigma^2}.$$

By Corollary 2, the Fisher information for  $\mu$  based on  $X[0], X[1], \dots, X[N-1]$  is

$$\mathcal{I}(\mu) = N \mathcal{I}_1(\mu) = \frac{N}{\sigma^2}. \quad (18)$$

Is the sample mean

$$\bar{X} = \frac{1}{N} \sum_{n=0}^{N-1} X[n]$$

an MVU estimator of  $\mu$ ? Observe that

$$\begin{aligned} \mathbb{E}_{\mathbf{X}|\mu}(\bar{X}|\mu) &= \mu \\ \text{var}_{\mathbf{X}|\mu}(\bar{X}|\mu) &= \frac{1}{N^2} \text{var}_{\mathbf{X}|\mu} \left( \sum_{n=0}^{N-1} X[n] | \mu \right) \\ &\stackrel{\text{X}[n] \text{ i.i.d.}}{=} \frac{1}{N^2} N \text{var}_{X|\mu}(X[n]|\mu) = \frac{\sigma^2}{N} \\ &\stackrel{\text{see (18)}}{=} \frac{1}{\mathcal{I}(\mu)} = \text{CRB}(\mu) \end{aligned}$$

and, therefore,  $\bar{X}$  is an MVU estimator of  $\mu$ . Since  $\bar{X}$  does not depend on  $\sigma^2$ , it is MVU for any  $\sigma^2$ . Hence,  $\bar{X}$  is an MVU estimator of  $\mu$  even if  $\sigma^2$  is unknown.

# Efficiency

**Definition.** *An unbiased estimator of  $\theta$  that attains CRB for  $\theta$  for all  $\theta$  in the parameter space  $\Theta$  is said to be **efficient**.*

**Note:** Efficiency implies MVU. However, MVU does not imply efficiency, because CRB is not always attainable by MVU estimators (at least not for finite samples, i.e. finite  $N$ ).

Under certain regularity conditions, maximum-likelihood (ML) estimators attain CRB asymptotically (for large  $N$ ); hence they are **asymptotically efficient**, which is one of the main reasons for their popularity, see handout # 3.

**Proof that efficiency implies MVU.** Recall Corollary 1: for **any unbiased** estimator, its variance must be greater than or equal to the CRB. If there exists an unbiased estimator whose variance is equal to CRB for all  $\theta$  in the parameter space  $\Theta$ , then this estimator must be MVU.

In the following theorem, we give necessary and sufficient conditions for CRB to be attainable. This theorem formalizes the results on pp. 16–18.

**Theorem 2.** *Suppose that assumptions (i) and (ii) hold and there exists an unbiased estimate  $T$  of  $\psi(\theta)$  that achieves the*

lower bound of the information inequality theorem (Theorem 1) for every  $\theta$ . Then  $f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$  is a one-parameter exponential family with pdf/pmf of the form

$$f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) = h(\mathbf{x}) \exp[\eta(\theta) T(\mathbf{x}) - B(\theta)] \quad (19)$$

see also (2). Conversely, if  $f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)$  belongs to the one-parameter exponential family of the above form and  $\eta(\theta)$  has a continuous nonvanishing derivative on the parameter space  $\Theta$ , then  $T(\mathbf{X})$  achieves the CRB and is the MVU estimator of  $\mathbb{E}_{\mathbf{X}|\theta}[T(\mathbf{X})|\theta]$ . Hence,  $T(\mathbf{X})$  is an efficient estimator of  $\mathbb{E}_{\mathbf{X}|\theta}[T(\mathbf{X})|\theta] = \psi(\theta)$ .

**Proof.** See Bickel & Doksum, Theorem 3.4.2. Sufficiency (i.e. that having exp. family implies that the natural suff. stat is efficient) was proved on pp. 16–18.  $\square$

For further reading, see also Property 4 in Section 4.4.4 of Hero's notes.

Theorem 2 gives both the necessary and sufficient conditions for an efficient estimator.

## Comments:

- (19) can be used to show efficiency of  $S(\mathbf{X})$  if there is an affine relationship between  $S(\mathbf{X})$  and the natural sufficient

statistic  $T(\mathbf{X})$ :

$$S(\mathbf{X}) = \underbrace{a}_{\neq 0} T(\mathbf{X}) + b \quad (20)$$

where  $T(\mathbf{X})$  is the natural sufficient statistic of the exponential family in (19).

**Note:** if  $T(\mathbf{X})$  is an efficient estimator of its expectation  $E_{\mathbf{X}|\theta}[T(\mathbf{X})|\theta]$ , this *does not* imply that a non-affine function of  $T(\mathbf{X})$  is an efficient estimator of its expectation. For example, suppose that  $T(\mathbf{X})$  is an efficient estimator of its expectation; then  $S(\mathbf{X}) = 1/T(\mathbf{X})$  will not be an efficient estimator of  $E_{\mathbf{X}|\theta}[1/T(\mathbf{X})]$  in general.

**Example.**  $X[0], X[1], \dots, X[N-1]$  are i.i.d. Poisson( $\lambda$ ):

$$\begin{aligned} p_{\mathbf{X}|\lambda}(\mathbf{x}|\lambda) &= \frac{\lambda^{\sum_{n=0}^{N-1} x[n]}}{\prod_{n=0}^{N-1} x[n]!} \exp(-N\lambda) \\ &= \frac{1}{\prod_{n=0}^{N-1} x[n]!} \exp\left(\underbrace{\sum_{n=0}^{N-1} x[n]}_{T(\mathbf{x})} \underbrace{\ln \lambda}_{\eta} - \underbrace{N\lambda}_{A(\eta)}\right) \quad (21) \end{aligned}$$

where  $\mathbf{X} = [X[0], X[1], \dots, X[N-1]]^T$  and  $\mathbf{x} = [x[0], x[1], \dots, x[N-1]]^T$ . Note that  $p_{\mathbf{X}|\lambda}(\mathbf{x}|\lambda)$  in (21) belongs to the exponential family of distributions, with

$\lambda = \exp(\eta)$ , and

$$A(\eta) = N \lambda = N \exp(\eta).$$

Now, apply Theorem 2: the natural sufficient statistic

$$T(\mathbf{X}) = \sum_{n=0}^{N-1} X[n]$$

is an efficient estimator of

$$\mathbb{E}_{\mathbf{X} | \eta}[T(\mathbf{X}) | \eta] = \frac{dA(\eta)}{d\eta} = N \exp(\eta) = N \lambda$$

and

$$\text{var}_{\mathbf{X} | \eta}[T(\mathbf{X})] = \frac{d^2 A(\eta)}{d\eta^2} = N \exp(\eta) = N \lambda.$$

We apply the affine transform (20) with  $a = 1/N$  and  $b = 0$  and conclude that

$$S(\mathbf{X}) = \frac{T(\mathbf{X})}{N} = \frac{1}{N} \cdot \sum_{n=0}^{N-1} X[n] = \bar{X}$$

is an efficient estimator of

$$\lambda = \mathbb{E}_{\mathbf{X} | \eta}(\bar{X} | \eta) = \frac{\mathbb{E}_{\mathbf{X} | \eta}[T(\mathbf{X}) | \eta]}{N} = \exp(\eta)$$

and

$$\text{var}_{\mathbf{X} | \eta}(\bar{X} | \eta) = \frac{1}{N^2} \text{var}_{\mathbf{X} | \eta}(T(\mathbf{X}) | \eta) = \frac{1}{N^2} \cdot N \exp(\eta) = \frac{\lambda}{N}.$$

By efficiency,

$$\text{var}_{\mathbf{X} | \lambda}(\bar{X} | \lambda) = \frac{\lambda}{N} = \text{CRB}(\lambda)$$

which is consistent with the CRB result for the Poisson mean parameter  $\lambda$  in (6).

## Cramér-Rao Bound: Example

Consider a sinusoid of *unknown frequency* but known amplitude and phase:

$$\begin{aligned} s[n; f] &= A \cos(2\pi f n + \phi) \quad 0 < f < 0.5 \\ X[n] &= s[n; f] + W[n] \quad 0 \leq n \leq N - 1. \end{aligned}$$

Assume that  $W[n]$  is additive white Gaussian noise (AWGN) with known variance  $\sigma^2$ . (Recall, the AWGN assumption on  $W[n]$  is that  $W[n]$  are i.i.d. zero-mean Gaussian with constant variance  $\sigma^2$ ). Then

$$f_{X|f}(x[n] | f) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{1}{2\sigma^2} \cdot (x[n] - s[n; f])^2\right].$$

Since the observations are independent, the likelihood function of  $f$  for the data  $\mathbf{x}$  is

$$f_{\mathbf{X}|f}(\mathbf{x} | f) = \prod_{n=0}^{N-1} f_{X|f}(x[n] | f)$$

where  $\mathbf{x} = [x[0], x[1], \dots, x[N-1]]^T$ . Taking the logarithm

yields

$$\begin{aligned}\ln f_{\mathbf{X} | f}(\mathbf{x} | f) &= \sum_{n=0}^{N-1} \ln f_{X | f}(x[n] | f) \\ &= -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - s[n; f])^2 + \underbrace{\text{const}}_{\text{not a function of } f}\end{aligned}$$

Differentiate with respect to  $f$ :

$$\frac{\partial \ln f_{\mathbf{X} | f}(\mathbf{x} | f)}{\partial f} = \frac{1}{\sigma^2} \cdot \sum_{n=0}^{N-1} \frac{\partial s[n; f]}{\partial f} \cdot (x[n] - s[n; f])$$

and once more:

$$\begin{aligned}\frac{\partial^2 \ln f_{\mathbf{X} | f}(\mathbf{x} | f)}{\partial f^2} &= \frac{1}{\sigma^2} \cdot \sum_{n=0}^{N-1} \frac{\partial^2 s[n; f]}{\partial f^2} \cdot (x[n] - s[n; f]) \\ &\quad - \frac{1}{\sigma^2} \cdot \sum_{n=0}^{N-1} \left( \frac{\partial s[n; f]}{\partial f} \right)^2.\end{aligned}$$

The negative expected value of this expression is the Fisher

information:

$$\begin{aligned}\mathcal{I}(f) &= -\mathbb{E}_{\mathbf{x}|f} \left[ \frac{\partial \ln \ln f_{\mathbf{X}|f}(\mathbf{x}|f)}{\partial f} \right] = \frac{1}{\sigma^2} \cdot \sum_{n=0}^{N-1} \left( \frac{\partial s[n; f]}{\partial f} \right)^2 \\ &= \text{SNR} \cdot \sum_{n=0}^{N-1} [2\pi N \cdot \sin(2\pi f n + \phi)]^2\end{aligned}$$

where  $\text{SNR} = A^2/\sigma^2$  is the signal-to-noise ratio. The CRB is

$$1/\mathcal{I}(f) \leq \text{var}_{\mathbf{x}|f}(\hat{f})$$

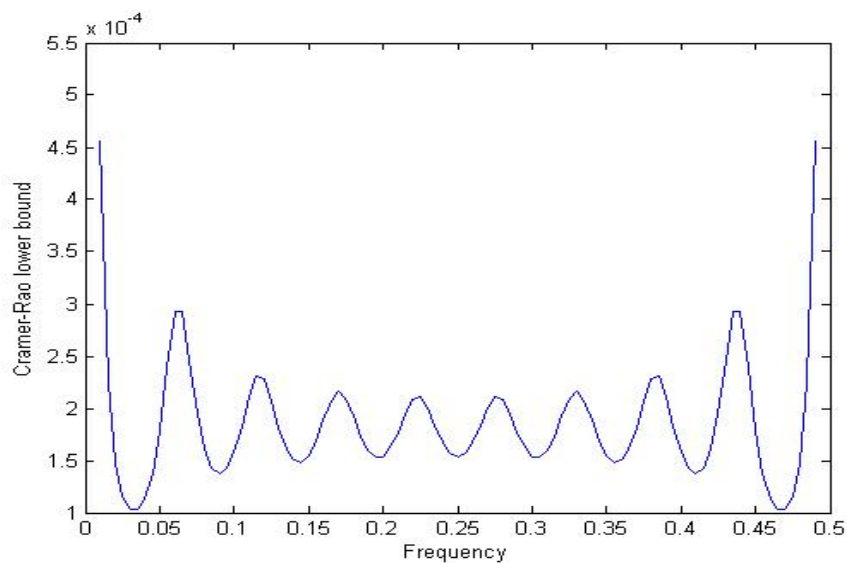
for unbiased frequency estimators  $\hat{f}$ .

## Cramér-Rao bound – Example (cont.)

Consider the case where  $\text{SNR} = 1$ ,  $N = 10$ , and  $\phi = 0$ . Then

$$s[n; f] = A \cos(2 \pi f n).$$

Recall that  $N$ ,  $A$ ,  $\phi$ , and  $\sigma^2$  are assumed *known*.

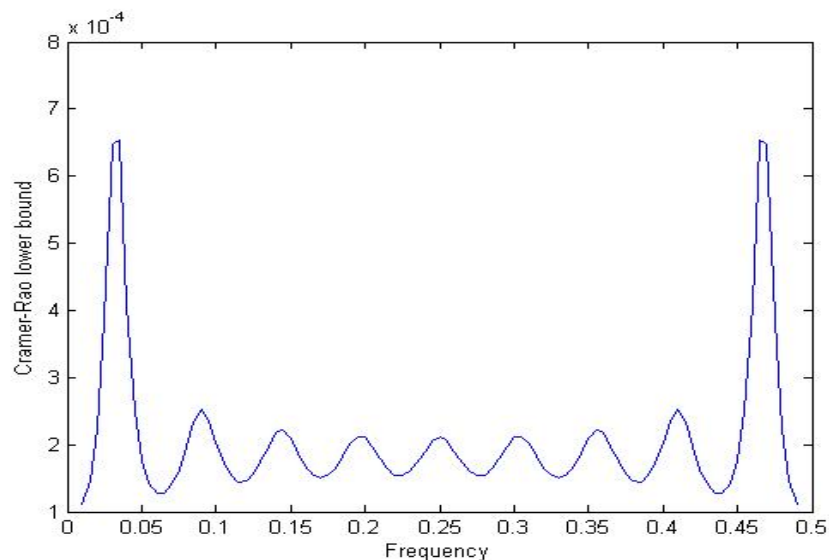


CRB for  $f$  as a function of  $f$ , for  $\text{SNR} = 1$ ,  $N = 10$ , and  $\phi = 0$ .

There are preferred frequencies.

Consider now the case where  $\text{SNR} = 1$ ,  $N = 10$ , and  $\phi = -\pi/2$ . Then

$$s[n; f] = A \sin(2 \pi f n).$$



CRB for  $f$  as a function of  $f$ , for  $\text{SNR} = 1$ ,  $N = 10$ , and  $\phi = -\pi/2$ .

Here,  $f \searrow 0$  is good for frequency estimation because we can easily differentiate between the case of no signal at all (which happens at  $f = 0$ ) and a sinusoid with amplitude  $A$ .

CRB results can be used to design a good frequency-estimation system.

In general, CRB is used as a

- measure of the potential performance attainable from the system,
- benchmark for assessing algorithm performance,
- measure for system design.

# Multiparameter CRB

We extend the CRB to the case of several parameters, i.e. a parameter vector

$$\boldsymbol{\theta} = [\theta_1, \dots, \theta_d]^T.$$

**Proposition.** Assume that the parameter space  $\Theta$  is an open subset of  $\mathbf{R}^d$  and that  $f_{X|\boldsymbol{\theta}}(x|\boldsymbol{\theta})$  satisfies conditions (i) and (ii) when differentiation is with respect to  $\theta_i$   $i = 1, 2, \dots, d$ . We define the  $d \times d$  Fisher information matrix (FIM) for the parameter vector  $\boldsymbol{\theta}$  as

$$\mathcal{I}(\boldsymbol{\theta}) = (\mathcal{I}_{i,k}(\boldsymbol{\theta})) \quad i, k \in \{1, 2, \dots, d\}$$

where

$$\mathcal{I}_{i,k}(\boldsymbol{\theta}) = \mathbb{E}_{X|\boldsymbol{\theta}} \left[ \frac{\partial}{\partial \theta_i} \ln f_{X|\boldsymbol{\theta}}(X|\boldsymbol{\theta}) \frac{\partial}{\partial \theta_k} \ln f_{X|\boldsymbol{\theta}}(X|\boldsymbol{\theta}) \mid \boldsymbol{\theta} \right].$$

Under the above conditions, the following hold:

(a)

$$\mathbb{E}_{X|\boldsymbol{\theta}} \left[ \frac{\partial}{\partial \theta_i} \ln f_{X|\boldsymbol{\theta}}(X|\boldsymbol{\theta}) \mid \boldsymbol{\theta} \right] = 0 \quad i = 1, 2, \dots, d$$

$$\mathcal{I}_{i,k} = \text{COV}_{X|\boldsymbol{\theta}} \left[ \frac{\partial}{\partial \theta_i} \ln f_{X|\boldsymbol{\theta}}(X|\boldsymbol{\theta}), \frac{\partial}{\partial \theta_k} \ln f_{X|\boldsymbol{\theta}}(X|\boldsymbol{\theta}) \mid \boldsymbol{\theta} \right]$$

for  $i, k \in \{1, 2, \dots, d\}$ . Using matrix notation, we rewrite these results as

$$\mathbb{E}_{X|\theta} \left[ \frac{\partial}{\partial \theta} \ln f_{X|\theta}(X|\theta) \mid \theta \right] = \underbrace{\mathbf{0}}_{d \times 1 \text{ vector of zeros}}$$

and

$$\mathcal{I}(\theta) = \text{COV}_{X|\theta} \left[ \frac{\partial}{\partial \theta} \ln f_{X|\theta}(X|\theta) \mid \theta \right].$$

**(b)** For i.i.d. measurements  $X[0], X[1], \dots, X[N-1]$  and  $\mathbf{X} = [X[0], X[1], \dots, X[N-1]]^T$ , FIM for  $\theta$  is

$$N \mathcal{I}_1(\theta)$$

where  $\mathcal{I}_1(\theta)$  is FIM for  $\theta$  and a single measurement  $X[0]$ , say (or  $X[1]$  etc).

**(c)** If, in addition to the above regularity conditions,  $f_{X|\theta}(x|\theta)$  is twice differentiable and integration (with respect to  $X$ ) and double differentiation (with respect to  $\theta$ ) under the integral sign can be interchanged, then

$$[\mathcal{I}(\theta)]_{i,k} = -\mathbb{E}_{X|\theta} \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_k} \ln f_{X|\theta}(X|\theta) \right] \quad i, k \in \{1, 2, \dots, d\}.$$

**Example.** Suppose

$$X | \boldsymbol{\theta} \sim \mathcal{N}(\mu, \sigma^2)$$

and

$$\boldsymbol{\theta} = [\mu, \sigma^2]^T.$$

Then

$$\ln f_{X|\boldsymbol{\theta}}(x|\boldsymbol{\theta}) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (x - \mu)^2$$

$$\begin{aligned} \mathcal{I}_{11}(\boldsymbol{\theta}) &= -\mathbb{E}_{X|\boldsymbol{\theta}} \left\{ \frac{\partial^2}{\partial \mu^2} \ln[f_{X|\boldsymbol{\theta}}(X|\boldsymbol{\theta})] \right\} \\ &= \mathbb{E}_{X|\boldsymbol{\theta}}(\sigma^{-2}) = \sigma^{-2} \end{aligned}$$

$$\begin{aligned} \mathcal{I}_{12}(\boldsymbol{\theta}) &= -\mathbb{E}_{X|\boldsymbol{\theta}} \left\{ \frac{\partial}{\partial \sigma^2} \frac{\partial}{\partial \mu} \ln[f_{X|\boldsymbol{\theta}}(X|\boldsymbol{\theta})] \right\} \\ &= -\sigma^{-4} \mathbb{E}_{X|\boldsymbol{\theta}}(X - \mu) = 0 = \mathcal{I}_{21}(\boldsymbol{\theta}) \end{aligned}$$

$$\mathcal{I}_{22}(\boldsymbol{\theta}) = -\mathbb{E}_{X|\boldsymbol{\theta}} \left[ \frac{\partial^2}{\partial (\sigma^2)^2} \ln[f_{X|\boldsymbol{\theta}}(X|\boldsymbol{\theta})] \right] = \sigma^{-4}/2.$$

Therefore

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & \sigma^{-4}/2 \end{bmatrix}. \quad (22)$$

**Multiple I.I.D. Observations:**

$$X[n] | \boldsymbol{\theta} \sim \mathcal{N}(\mu, \sigma^2) \quad n = 0, 1, \dots, N-1$$

with

$$\boldsymbol{\theta} = [\mu, \sigma^2]^T.$$

Then, (22) implies that

$$\mathcal{I}_1(\boldsymbol{\theta}) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & \sigma^{-4}/2 \end{bmatrix}$$

and, consequently,

$$\mathcal{I}(\boldsymbol{\theta}) = N \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & \sigma^{-4}/2 \end{bmatrix}. \quad (23)$$

**Decoupling:** FIM in this example is diagonal. Therefore, CRB for  $\mu$  remains the same whether or not  $\sigma^2$  is known. Similarly, CRB for  $\sigma^2$  is the same regardless of whether or not  $\mu$  is known.

In general, the more parameters<sup>1</sup>, the larger (or equal) the CRB; the CRBs are equal in the case of decoupling. See problems 3.11 and 3.12 in Kay-I.

**Theorem 3.** Assume that the regularity conditions from p. 3 hold and suppose that FIM  $\mathcal{I}(\boldsymbol{\theta})$  is positive definite (hence nonsingular). Then, for

$$\mathbb{E}_{X|\boldsymbol{\theta}}[T(X) | \boldsymbol{\theta}] = \boldsymbol{\psi}(\boldsymbol{\theta})$$

---

<sup>1</sup>We have to compare nested models; otherwise, we would be comparing apples and oranges.

the following holds:

$$\text{var}_{\mathbf{X}|\boldsymbol{\theta}}[\mathbf{T}(X) | \boldsymbol{\theta}] \geq \frac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T} \mathcal{I}(\boldsymbol{\theta})^{-1} \frac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}. \quad (24)$$

More generally, for a  $d$ -dimensional statistic  $\mathbf{T}(X) = [T_1(X), \dots, T_d(X)]^T$  and

$$\boldsymbol{\psi}(\boldsymbol{\theta}) \triangleq \mathbb{E}_{\mathbf{X}|\boldsymbol{\theta}}[\mathbf{T}(X) | \boldsymbol{\theta}] = [\psi_1(\boldsymbol{\theta}), \dots, \psi_d(\boldsymbol{\theta})]^T \quad (25)$$

we have

$$\text{cov}_{\mathbf{X}|\boldsymbol{\theta}}[\mathbf{T}(X) | \boldsymbol{\theta}] \geq \frac{\partial\boldsymbol{\psi}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T} \mathcal{I}(\boldsymbol{\theta})^{-1} \frac{\partial\boldsymbol{\psi}(\boldsymbol{\theta})^T}{\partial\boldsymbol{\theta}}.$$

## Comments:

- $$\text{cov}_{\mathbf{X}|\boldsymbol{\theta}}[\mathbf{T}(X) | \boldsymbol{\theta}] \geq \frac{\partial\boldsymbol{\psi}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T} \mathcal{I}(\boldsymbol{\theta})^{-1} \frac{\partial\boldsymbol{\psi}(\boldsymbol{\theta})^T}{\partial\boldsymbol{\theta}}.$$

means that

$$\text{cov}_{\mathbf{X}|\boldsymbol{\theta}}[\mathbf{T}(X) | \boldsymbol{\theta}] - \frac{\partial\boldsymbol{\psi}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T} \mathcal{I}(\boldsymbol{\theta})^{-1} \frac{\partial\boldsymbol{\psi}(\boldsymbol{\theta})^T}{\partial\boldsymbol{\theta}} \geq 0$$

i.e. the matrix on the left is positive semidefinite. Recall: a matrix  $A$  is positive semidefinite if

$$\mathbf{q}^T A \mathbf{q} \geq 0 \quad \forall \mathbf{q}. \quad (26)$$

- If  $\mathbf{T}(X)$  is an unbiased estimator of  $\boldsymbol{\theta}$ , i.e.

$$E_{\mathbf{X} | \boldsymbol{\theta}}[\mathbf{T}(\mathbf{X}) | \boldsymbol{\theta}] = \boldsymbol{\psi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$$

then

$$\text{COV}_{\mathbf{X} | \boldsymbol{\theta}}[\mathbf{T}(X) | \boldsymbol{\theta}] \geq \mathcal{I}^{-1}(\boldsymbol{\theta}).$$

- Suppose now that

$$\boldsymbol{\psi}(\boldsymbol{\theta}) = \theta_i$$

corresponding to  $T_i(X)$ , where  $T_i(X)$  is the  $i$ th element of  $\mathbf{T}(X)$  (and  $\mathbf{T}(X)$  is an unbiased estimator of  $\boldsymbol{\theta}$ ). Now,

$$\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = [0, 0, \dots, 0, \underbrace{1}_{i\text{th place}}, 0, \dots, 0]^T$$

and, consequently,

$$\text{var}_{\mathbf{X} | \boldsymbol{\theta}}[T_i(\mathbf{X}) | \boldsymbol{\theta}] \geq [\mathcal{I}(\boldsymbol{\theta})^{-1}]_{i,i} = \underbrace{\text{CRB}_{i,i}(\boldsymbol{\theta})}_{(i,i)\text{th element of CRB matrix for } \boldsymbol{\theta}}.$$

- **Notation:** If

$$\mathbf{a}(\boldsymbol{\theta}) = \begin{bmatrix} a_1(\boldsymbol{\theta}) \\ a_2(\boldsymbol{\theta}) \\ \vdots \\ a_m(\boldsymbol{\theta}) \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix}$$

then

$$\frac{\partial \mathbf{a}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \begin{bmatrix} \partial a_1(\boldsymbol{\theta})/\partial \theta_1 & \partial a_1(\boldsymbol{\theta})/\partial \theta_2 & \cdots & \partial a_1(\boldsymbol{\theta})/\partial \theta_d \\ \partial a_2(\boldsymbol{\theta})/\partial \theta_1 & \partial a_2(\boldsymbol{\theta})/\partial \theta_2 & \cdots & \partial a_2(\boldsymbol{\theta})/\partial \theta_d \\ \vdots & \vdots & \cdots & \vdots \\ \partial a_m(\boldsymbol{\theta})/\partial \theta_1 & \partial a_m(\boldsymbol{\theta})/\partial \theta_2 & \cdots & \partial a_m(\boldsymbol{\theta})/\partial \theta_d \end{bmatrix}$$

and

$$\frac{\partial \mathbf{a}(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} = \left( \frac{\partial \mathbf{a}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right)^T.$$

# Multiparameter Exponential Family and Efficiency

Consider the canonical  $d$ -parameter exponential family:

$$f_{X|\boldsymbol{\eta}}(\mathbf{x}|\boldsymbol{\eta}) = \exp \left[ \underbrace{\sum_{i=1}^d T_i(\mathbf{x}) \eta_i}_{\mathbf{T}^T(\mathbf{x}) \boldsymbol{\eta}} - A(\boldsymbol{\eta}) \right] h(\mathbf{x})$$

and assume that the parameter space of  $\boldsymbol{\eta}$  is an open subset of  $\mathbb{R}^d$ . Then

$$\frac{\partial \ln f_{X|\boldsymbol{\eta}}(\mathbf{x}|\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \mathbf{T}(\mathbf{x}) - \frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}. \quad (27)$$

Hence, the Fisher information matrix is

$$\mathcal{I}(\boldsymbol{\eta}) = \text{cov}_{X|\boldsymbol{\eta}}[\mathbf{T}(X)|\boldsymbol{\eta}] = \underbrace{\frac{\partial^2 A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T}}_{d \times d \text{ matrix}}. \quad (28)$$

**Theorem 4.** Each  $T_i(X)$  is an MVU estimator of  $E_{X|\boldsymbol{\eta}}[T_i(X)|\boldsymbol{\eta}]$ .

**Comment:** The claim in Theorem 4 is different from stating that  $T_i(X)$  is an MVU estimator for estimating

$$E_{X|\eta}[T_i(X) | \eta] = \frac{\partial A(\eta)}{\partial \eta_i}$$

if  $\eta_k, k \neq i$  are *known*, which follows directly from Theorem 2. How do we show this new claim?

**Proof. (Theorem 4)** Without loss of generality, we focus on  $i = 1$ . Note that (15) implies

$$\begin{aligned} \text{var}_{X|\eta}[T_1(X) | \eta] &= \frac{\partial^2 A(\eta)}{\partial \eta_1^2} \\ E_{X|\eta}[T_1(X) | \eta] &= \frac{\partial A(\eta)}{\partial \eta_1} \triangleq \psi(\eta). \end{aligned}$$

Therefore

$$\frac{\partial \psi(\eta)}{\partial \eta^T} = \frac{\partial A(\eta)}{\partial \eta_1 \partial \eta^T}$$

is the first row of

$$\mathcal{I}(\eta) = \frac{\partial^2 A(\eta)}{\partial \eta \partial \eta^T}$$

implying that

$$\frac{\partial \psi(\eta)}{\partial \eta^T} \mathcal{I}(\eta)^{-1} = \underbrace{[1, 0, \dots, 0]}_{\text{first row of } \mathcal{I}(\eta)} \mathcal{I}(\eta)^{-1} = [1, 0, \dots, 0]$$

and, finally,

$$\frac{\partial \psi(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T} \mathcal{I}(\boldsymbol{\eta})^{-1} \frac{\partial \psi(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \frac{\partial \psi(\boldsymbol{\eta})}{\partial \eta_1} = \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \eta_1^2} = \text{var}_{X|\boldsymbol{\eta}}[T_1(X)]$$

i.e. (24) in Theorem 3 is satisfied *with equality* and  $T_1(X)$  is MVU for  $\mathbb{E}_{X|\boldsymbol{\eta}}[T_1(X) | \boldsymbol{\eta}]$ .  $\square$

**Example:** If  $X[0], X[1], \dots, X[N-1]$  given  $\mu$  and  $\sigma^2$  are i.i.d.  $\mathcal{N}(\mu, \sigma^2)$ , then

$$\bar{X} = \frac{1}{N} \sum_{n=0}^{N-1} X[n]$$

is the MVU estimator of  $\mu$  and

$$\frac{1}{N} \sum_{n=0}^{N-1} X^2[n]$$

is the MVU estimator of  $\mu^2 + \sigma^2$ . This result follows by noting

that

$$\begin{aligned}
 f_{\mathbf{X} | \boldsymbol{\theta}}(\mathbf{x} | \boldsymbol{\theta}) &= (2 \pi \sigma^2)^{-N/2} \exp \left\{ -\frac{1}{2 \sigma^2} \sum_{n=0}^{N-1} (x[n] - \mu)^2 \right\} \\
 &= (2 \pi \sigma^2)^{-N/2} \exp \left( -\frac{N \mu^2}{2 \sigma^2} \right) \\
 &\quad \cdot \exp \left[ -\frac{1}{2 \sigma^2} \left( \underbrace{N \frac{1}{N} \sum_{n=0}^{N-1} x^2[n]}_{T_2(\mathbf{x})} - 2 \mu N \cdot \underbrace{\bar{x}}_{T_1(\mathbf{x})} \right) \right]
 \end{aligned}$$

belongs to the two-parameter exponential family of distributions and that  $T_1(\mathbf{X})$  and  $T_2(\mathbf{X})$  are natural sufficient statistics. But, it *does not* follow that

$$\frac{1}{N-1} \cdot \sum_{n=0}^{N-1} (X[n] - \bar{X})^2$$

is the MVU estimator of  $\sigma^2$ .

## Gaussian CRB

**Theorem 5.** Suppose that  $\mathbf{X}$  has an  $N$ -variate Gaussian distribution,

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{C}(\boldsymbol{\theta}))$$

i.e.

$$f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^N |\mathbf{C}|}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right].$$

Then, the  $(i, k)$ th element of FIM for  $\boldsymbol{\theta}$  is given by

$$\mathcal{I}_{i,k} = \frac{\partial \boldsymbol{\mu}^T}{\partial \theta_i} \mathbf{C}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \theta_k} + \frac{1}{2} \text{tr} \left( \mathbf{C}^{-1} \frac{\partial \mathbf{C}}{\partial \theta_i} \mathbf{C}^{-1} \frac{\partial \mathbf{C}}{\partial \theta_k} \right). \quad (29)$$

(29) is a convenient general formula for analysis.

**Proof.** See Kay-I, App. 3C.  $\square$

**Example:** Consider the following signal-plus-noise model:

$$X[n] = s[n; \theta] + W[n] \quad n = 0, 1, \dots, N-1$$

where  $\theta$  is the unknown parameter and  $W[n]$  is AWGN with known variance  $\sigma^2$ . Then, we can write this model specification in a vector form as follows:

$$\{\mathbf{X}|\boldsymbol{\theta}\} = \boldsymbol{\mu}(\boldsymbol{\theta}) + \mathbf{W} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \underbrace{\sigma^2 \overbrace{\mathbf{I}}^{N \times N \text{ identity matrix}}}_{\mathbf{C}}).$$

where

$$\boldsymbol{\mu}(\theta) = \begin{bmatrix} s[1; \theta] \\ s[2; \theta] \\ \vdots \\ s[N-1; \theta] \end{bmatrix}$$

and  $C$  does not depend on  $\theta$  (and, furthermore, is completely known).

$$\mathcal{I}(\theta) = \frac{1}{\sigma^2} \frac{\partial \boldsymbol{\mu}^T}{\partial \theta} \frac{\partial \boldsymbol{\mu}}{\partial \theta} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} \left( \frac{\partial s[n; \theta]}{\partial \theta} \right)^2$$

which is the familiar expression that we derived earlier, see p. 19. What if we have a vector of parameters  $\boldsymbol{\theta}$ ? In this case,

$$\mathcal{I}_{i,k} = \frac{1}{\sigma^2} \frac{\partial \boldsymbol{\mu}^T}{\partial \theta_i} \frac{\partial \boldsymbol{\mu}}{\partial \theta_k} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} \frac{\partial s[n; \theta]}{\partial \theta_i} \frac{\partial s[n; \theta]}{\partial \theta_k}.$$

**Example:**  $X[n]$   $n = 0, 1, \dots, N-1$  is AWGN with variance  $\sigma^2$ , i.e.

$$\{\mathbf{X} \mid \sigma^2\} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_N).$$

If  $\sigma^2$  is the unknown parameter, then

$$\mathcal{I}(\sigma^2) = \frac{1}{2} \text{tr} \left( C^{-1} \frac{\partial C}{\partial \sigma^2} C^{-1} \frac{\partial C}{\partial \sigma^2} \right) = \frac{1}{2 (\sigma^2)^2} \text{tr}(I_N) = \frac{N}{2 (\sigma^2)^2} \quad (30)$$

and, therefore,

$$\text{CRB}(\sigma^2) = [\mathcal{I}(\sigma^2)]^{-1} = \frac{2(\sigma^2)^2}{N}. \quad (31)$$

Say we wish to compute CRB for  $\sigma$ :

$$\text{CRB}(\sigma) = [\mathcal{I}(\sigma)]^{-1} = \left[ \frac{1}{2} \cdot \text{tr} \left( C^{-1} \frac{\partial C}{\partial \sigma} C^{-1} \frac{\partial C}{\partial \sigma} \right) \right]^{-1} = \frac{\sigma^2}{2N}$$

which can also be computed using (8):  $\psi(\sigma^2) = (\sigma^2)^{1/2}$ ,  $\psi'(\sigma^2) = 1/2 \cdot (\sigma^2)^{-1/2}$ , and

$$\frac{|\psi'(\sigma^2)|^2}{\mathcal{I}(\sigma^2)} = \frac{(1/4) \cdot \sigma^{-2}}{N/(2\sigma^4)} = \frac{\sigma^2}{2N}.$$

Here, we consider the same measurement model as Example 2 in handout # 1. There, we studied the following family of estimators of  $\sigma^2$ :

$$\hat{\sigma}^2 = c \frac{1}{N} \sum_{n=0}^{N-1} X^2[n]$$

and found that

$$c_{\text{OPT}} = \frac{N}{N+2}$$

yields an estimator

$$\hat{\sigma}_*^2 = c_{\text{OPT}} \cdot \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] = \frac{1}{N+2} \sum_{n=0}^{N-1} x^2[n]$$

whose MSE is the smallest within the family:

$$\text{MSE}_{\text{MIN}} = \frac{2(\sigma^2)^2}{N+2} < \frac{2(\sigma^2)^2}{N} = \text{CRB}(\sigma^2)$$

see (31). Note that  $\hat{\sigma}_*^2$  is a *biased* estimator of  $\sigma^2$  and that CRB is a lower bound on variance of *unbiased* estimators only. Applying (10) leads to

$$\text{MSE}\{\hat{\sigma}^2\} = \text{var}_{X|\sigma^2}\{\hat{\sigma}^2\} + b^2(\sigma^2) \geq \frac{2c^2\sigma^4}{N} + (c-1)^2(\sigma^2)^2$$

since  $b(\sigma^2) = \psi(\sigma^2) - \sigma^2 = (c-1)\sigma^2$ . Interestingly, the above inequality becomes equality for

$$c = c_{\text{OPT}} = \frac{N}{N+2}.$$

The MSE bound (10) is not always attainable; we just happen to be lucky in this case.