

EE6617: 01 Minimum Unbiased Estimator

In this lecture note, we will study fundamentals of minimum unbiased estimator. The concept establishes here is the foundations of classical estimation theory.

Major reference: Fundamentals of Statistical Processing, Volume I: Estimation Theory, 1/E, by Steven M. Kay, Prentice Hall, 1993

In particular, our goals in this lecture are stated as follows:

- Want to design a good estimator for **unknown deterministic parameter** $\theta \in \mathbb{R}$ or parameters $\boldsymbol{\theta} \in \mathbb{R}^{m \times 1}$
- Want to design an **estimator** $\hat{\theta}$ which on the average yields the **true parameter value** θ
- Find the one that exhibits the **least (smallest) variance**

To fulfill the goals, we need first the concept of unbiased estimator.

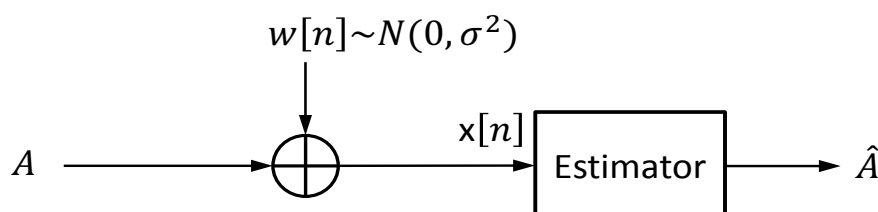
1 Unbiased Estimator

For the parameter θ , the estimator $\hat{\theta}$ that yields θ on the average is called the unbiased estimator. Mathematically, the estimator is **unbiased** if

$$E[\hat{\theta}] = \theta.$$

This concept can be checked from the following example.

Example 1: Unbiased Estimator for DC Level in white Gaussian Noise



Consider the observations:

$$x[n] = A + w[n], \quad n = 0, 1, \dots, N-1, \quad w[n] \sim \mathcal{N}(0, \sigma^2)$$

where $w[n]$ is independent and identically distributed (i.i.d.) according to $\mathcal{N}(0, \sigma^2)$. We want to estimate the parameter A given the observation $\{x[n]\}_{n=1}^{N-1}$. An **intuitive estimator** is the

sample mean

$$\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] : \text{sample mean.}$$

Now, check whether or not this estimator is biased. Notice that from the above figure the **estimator $\hat{A}$ is a function of $x[n]$** , $n = 0, \dots, N-1$. We evaluate $E[\hat{A}]$,

$$E[\hat{A}] = \frac{1}{N} \sum_{n=0}^{N-1} E[x[n]] = \frac{1}{N} NA = A.$$

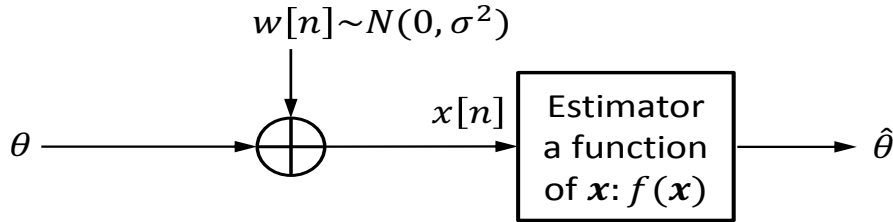
This is because $E[x[n]] = A + E[w[n]] = A$. Since $E[\hat{A}] = A$, it is **unbiased**.

The **variance of the estimator** could also be used to measure the estimator performance,

$$\begin{aligned} \text{var}[\hat{A}] &= E\left[\left(\hat{A} - A\right)^2\right] = E\left[\left(\frac{1}{N} \sum_{n=0}^{N-1} x[n] - A\right)^2\right] \\ &= E\left[\left(\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - A)\right)^2\right] = E\left[\left(\frac{1}{N} \sum_{n=0}^{N-1} w[n]\right)^2\right] = \frac{\sigma^2}{N} \end{aligned} \quad (1)$$

Thus, $\hat{A}$ is the unbiased estimator with $\text{var}[\hat{A}] = \frac{\sigma^2}{N} \xrightarrow{N \rightarrow \infty} 0$. ■

In general, $\hat{\theta}$ is a function of the received signal $\mathbf{x} = [x[1] \ x[2] \ \dots \ x[N-1]]^T \in \mathbb{R}^{N \times 1}$,



$$\mathbf{x} \xrightarrow{f} \hat{\theta}, \quad \hat{\theta} = f(\mathbf{x})$$

In this general setting, the unbiasedness of $\hat{\theta}$ implies

$$E[\hat{\theta}] = \int f(\mathbf{x}) p(\mathbf{x}; \theta) d\mathbf{x} = \theta.$$

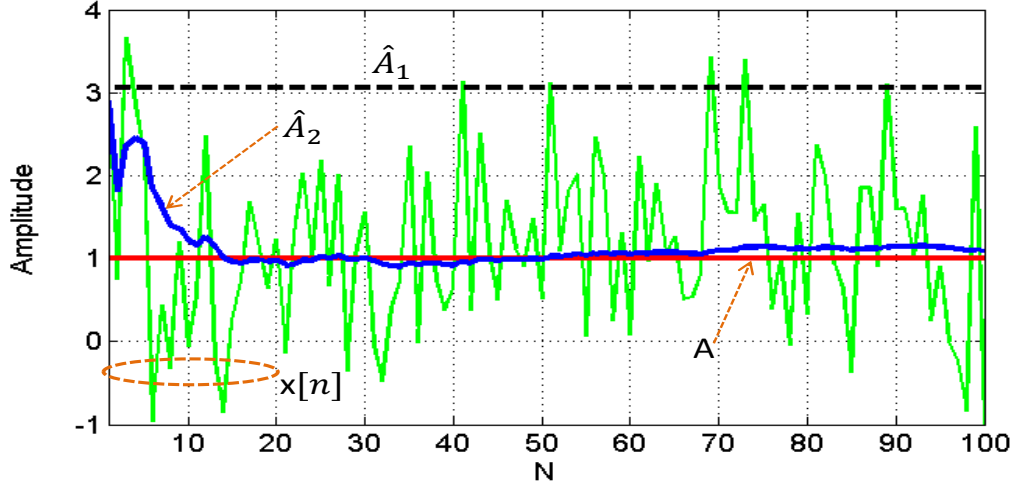
Question: Why unbiased estimator? The unbiased estimator is a good estimator?

We attempt to answer this question by looking at some examples. Suppose two estimators

$$\hat{A}_1 = x[0] \text{ and } \hat{A}_2 = \frac{19}{20N} \sum_{n=0}^{N-1} x[n],$$

1. $\hat{A}_1 = x[0], E[\hat{A}_1] = E[x[0]] = A$: unbiased
2. $\hat{A}_2 = \frac{19}{20N} \sum_{n=0}^{N-1} x[n], E[\hat{A}_2] = \frac{19}{20}A$: $\begin{cases} \text{unbiased} & \text{if } A = 0 \\ \text{biased} & \text{if } A \neq 0 \end{cases}$

Q: Which estimator is better?



A: From the figure, we can see $\hat{A}_2$ is better than $\hat{A}_1$ (obviously, this answer is solely based on the observation). This result, at least, elucidates that we might get better estimation performance by having an unbiased estimator.

Then, why do we care about the unbiased estimator? This will be clearer in the next section.

We define the **bias of the estimator $\hat{\theta}$ as $b(\theta)$** . We shall write $E[\hat{\theta}] = \theta + b(\theta)$ and if $b(\theta) \neq 0$, the estimator $\hat{\theta}$ is said to be **biased** and

$$b(\theta) = E[\hat{\theta}] - \theta \neq 0$$

2 Estimator Performance Criterion

From the previous section, we have seen the unbiased estimator is not necessarily a good estimator and biased estimator could yield better estimation performance. Then, **how can we tell whether an estimator is good or not?**

In determining **an optimal estimator**, we need to adopt some **optimality criterion**. **Mean square error (MSE)** of an estimator has been popularly used as the optimality criterion

$$\text{mse}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right].$$

This measures the **average mean square deviation of the estimator from the true value**.

The $\text{mse}(\hat{\theta})$ can be expressed as

$$\text{mse}(\hat{\theta}) = \text{var}(\hat{\theta}) + b^2(\theta).$$

How? To see this,

$$\begin{aligned} \text{mse}(\hat{\theta}) &= E \left[(\hat{\theta} - \theta)^2 \right] = E \left[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2 \right] \\ &= \underbrace{E \left[(\hat{\theta} - E[\hat{\theta}])^2 \right]}_{=\text{var}(\hat{\theta})} + \underbrace{E \left[(E[\hat{\theta}] - \theta)^2 \right]}_{b^2(\theta)} + \underbrace{2 E \left[(\hat{\theta} - E[\hat{\theta}]) (E[\hat{\theta}] - \theta) \right]}_{\text{cross term}}, \end{aligned}$$

and the cross term is

$$E \left[(\hat{\theta} - E[\hat{\theta}]) (E[\hat{\theta}] - \theta) \right] = E[\hat{\theta}]^2 - E[\hat{\theta}] \theta - E[\hat{\theta}]^2 + E[\hat{\theta}] \theta = 0,$$

implying

$$\text{mse}(\hat{\theta}) = \text{var}(\hat{\theta}) + b^2(\theta).$$

$\text{mse}(\hat{\theta})$ depends on both:

- **Variance of $\hat{\theta}$, $\text{var}(\hat{\theta})$**
- **Bias of $\hat{\theta}$, $b(\theta)$**

Hence, designing an estimator by minimizing MSE should be sought as a design criterion. However, the dependency to the bias term $b(\theta)$, in most of cases, makes the **MSE criterion unrealizable** to design the estimator. This can be checked by looking at the following example.

Example 2: Modified Estimator for DC Level in White Gaussian Noise

Same as Example 1, we want to design the estimator of A from the observation:

$$\begin{aligned} x[n] &= A + w[n], \quad n = 0, \dots, N-1, \\ \mathbf{w} &= [w[0] \ w[1] \ \dots \ w[N-1]]^T \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad E[\mathbf{w}\mathbf{w}^T] = \sigma^2 \mathbf{I} \end{aligned}$$

In this example, we consider a **modified estimator**,

$$\bar{A} = a \frac{1}{N} \sum_{n=0}^{N-1} x[n] = a\hat{A}$$

where $a \in \mathbb{R}$ is an optimization parameter that should be selected to minimize the MSE of $\bar{A}$.

Let us first check whether or not it is biased

$$\begin{aligned} E[\bar{A}] &= aA \\ b(A) &= E[\bar{A}] - A = (a - 1)A : \begin{cases} \text{unbiased} & \text{if } A = 0 \text{ or } a = 1 \\ \text{biased} & \text{otherwise} \end{cases} \end{aligned}$$

Now, compute the variance,

$$\text{var}(\bar{A}) = a^2 \text{var}(\hat{A}) = a^2 \frac{\sigma^2}{N}.$$

The MSE is formulated

$$\text{mse}(\bar{A}) = \frac{a^2 \sigma^2}{N} + (a - 1)^2 A^2$$

Optimize a to minimize $\text{mse}(\bar{A})$, i.e.,

$$a_{opt} = \underset{a}{\text{argmin}} \text{mse}(\bar{A})$$

Try the first and second derivatives of $\text{mse}(\bar{A})$ with respect to a

$$\begin{aligned} \text{mse}(\bar{A}) &= \frac{a^2 \sigma^2}{N} + (a - 1)^2 A^2 \\ \frac{\partial \text{mse}(\bar{A})}{\partial a} &= 2 \frac{a \sigma^2}{N} + 2(a - 1)A^2 \\ \frac{\partial^2 \text{mse}(\bar{A})}{\partial^2 a} &= 2 \frac{\sigma^2}{N} + 2A^2 \geq 0 \Rightarrow \text{mse}(\bar{A}) \text{ is convex in } a \end{aligned}$$

The optimal a that minimize MSE is determined by

$$\begin{aligned} \frac{\partial \text{mse}(\bar{A})}{\partial a} &= 0 \Rightarrow a_{opt} = \frac{A^2}{A^2 + \sigma^2/N} \\ \bar{A}_{opt} &= \left(\frac{A^2}{A^2 + \sigma^2/N} \right) \frac{1}{N} \sum_{n=0}^{N-1} x[n] \end{aligned} \quad (2)$$

This is the optimal estimator that minimize the MSE.

Notice that in order to realize $\bar{A}_{opt}$, A^2 or A should be known. However, if we know A , the optimal estimator is $\bar{A}_{opt} = A$ which is the perfect estimator of A . ■

So far, we have observed from Example 2 that

- In general, the MSE estimator is **unrealizable**. This is because the bias term $b(A) = (a - 1)A$ is always a function of the unknown, i.e., the estimation parameter A .

- Any criterion which depends on the bias will lead to an unrealizable estimator.

⇒ The MSE criterion is often not a good criterion for getting a practical estimator for an unknown deterministic parameter θ .

The practical motivation leads us to focus on unbiased estimators, while minimizing the MSE. For the unbiased estimator, minimizing the MSE is equivalent to minimizing $\text{var}(\hat{\theta})$ and we call a such estimator as minimum variance unbiased (MVU) estimator.

MSE criterion vs. MVU criterion

- MSE criterion:

$$\hat{\theta}_{mse} = \underset{\hat{\theta}}{\text{argmin}} \text{var}(\hat{\theta}) + b^2(\theta)$$

- MVU criterion:

$$\hat{\theta}_{mvu} = \underset{\hat{\theta}}{\text{argmin}} \text{var}(\hat{\theta}) \quad \text{subject to (s.t.) } b(\theta) = 0$$

Q: What is the MSE of MVU estimator?

Q: Let $\hat{\theta}_{mse} \in \mathcal{Q}_{mse}$ and $\hat{\theta}_{mvu} \in \mathcal{Q}_{mvu}$. Is $\mathcal{Q}_{mse} \subset \mathcal{Q}_{mvu}$?

One popular approach to design the MVU estimator is to use Cramer-Rao lower bound (CRLB).

3 Crammer-Rao Lower Bound (CRLB)

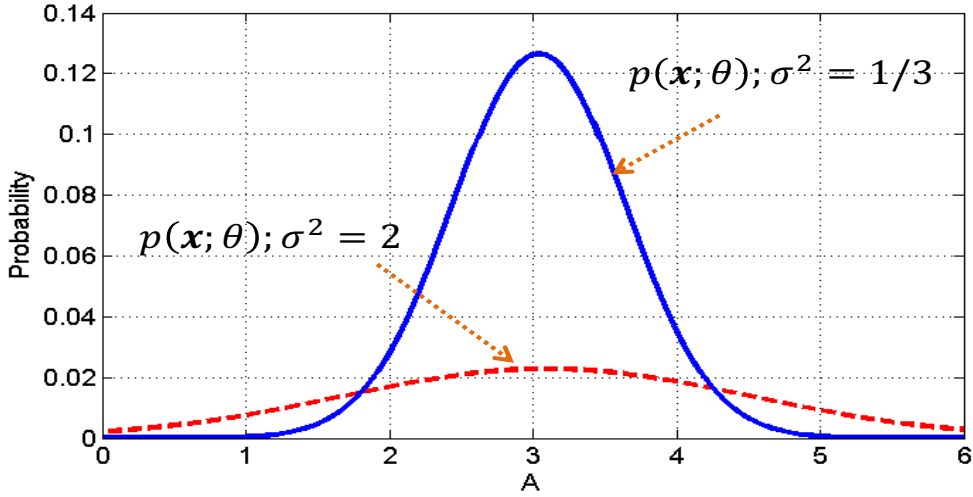
When the PDF $p(\mathbf{x}; \theta)$ is viewed as a function of the parameter θ (with $\mathbf{x}$ fixed, i.e., with the observation $\mathbf{x}$), the $p(\mathbf{x}; \theta)$ is termed as the **likelihood function**. What does $p(\mathbf{x}; \theta)$ tell us about the difficulty of the estimation problem?

3.1 Example 3: Likelihood Function and its Implication

Consider the observation: $x[n] = A + w[n]$, $n = 0, \dots, N-1$, $\mathbf{w} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. The likelihood function is

$$p(\mathbf{x}; \theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2\right).$$

Let us plot $p(\mathbf{x}; \theta)$ for a realization $\mathbf{x} = \mathbf{x}_0$, $\sigma^2 = 2$ and $\sigma^2 = 1/3$.



The two figures exhibit different **concentration of $\mathbf{x}$ around $A = 3$** and it is obvious to see that the case with $\sigma^2 = 1/3$ (smaller variance) reveals **sharper concentration** than the other case. The sharpness of $p(\mathbf{x}; \theta)$ determines how accurately we can estimate the parameter A . ■

Then, the next question is **how to measure the sharpness of $p(\mathbf{x}; \theta)$** . From calculus, we know that the sharpness of $p(\mathbf{x}; \theta)$ can be effectively measured by the **negative of the second derivative of the $\ln(p(\mathbf{x}; \theta))$** , i.e.,

$$\begin{aligned}
 p(\mathbf{x}; \theta) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2\right) \\
 -\ln(p(\mathbf{x}; \theta)) &= \frac{N}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \\
 -\frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial A} &= -\frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A) \\
 -\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 A} &= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} 1 = \frac{N}{\sigma^2}.
 \end{aligned} \tag{3}$$

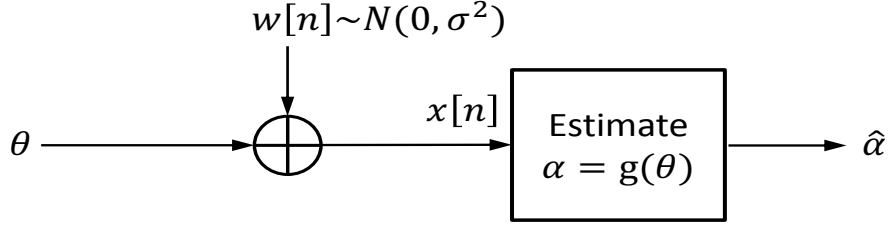
The $\frac{N}{\sigma^2}$ captures the **easiness (or sharpness)** of the the estimation problem. Consider the sample mean estimator $\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$. Given that $\text{var}(\hat{A}) = \frac{\sigma^2}{N}$ in (1), we can equate from (15) that

$$\text{var}(\hat{A}) = \frac{1}{-\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 A}}. \tag{4}$$

$\Rightarrow$ **variance of $\hat{A}$ decreases as the curvature increases**. This is the basic idea of Cramer-Rao Lower Bound (CRLB).

3.2 Cramer-Rao Lower Bound (CRLB) - Scalar θ

We want to estimate $\alpha = g(\theta)$, where $g(\cdot)$ is a known function.



The CRLB tells us two important aspects of a estimator:

- **The performance of the estimator $\hat{\alpha}$** in terms of $\text{var}(\hat{\alpha})$.
- **How to construct the MVU estimator $\hat{\alpha}$.**

The Cramer-Rao Lower Bound (CRLB) is summarized in the following theorem.

Theorem 1 Suppose that the likelihood function satisfies the *regularity condition*

$$E_{\mathbf{x}} \left[\frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} \right] = 0.$$

Then, the *variance of any unbiased estimator $\hat{\alpha}$* must satisfy

$$\text{var}(\hat{\alpha}) \geq \frac{\left(\frac{\partial g(\theta)}{\partial \theta} \right)^2}{-E_{\mathbf{x}} \left[\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} \right]} = \frac{\left(\frac{\partial g(\theta)}{\partial \theta} \right)^2}{I(\theta)}$$

where $I(\theta) = -E_{\mathbf{x}} \left[\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} \right]$ is called *Fisher information*. An unbiased estimator that achieves this bound can be found if and only if

$$\frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} = \frac{1}{c(\theta)} (\hat{\alpha}_{\text{opt}} - \alpha) \quad (5)$$

and *MVU estimator is $\hat{\alpha}_{\text{opt}}$.*

Proof : See appendix in Section 4.1. ■

If $\alpha = g(\theta) = \theta$,

$$\text{var}(\hat{\theta}) \geq -\frac{1}{E_{\mathbf{x}} \left[\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} \right]}.$$

This bounds is achieved as an equality, from (5), if and only if the unbiased estimator $\hat{\theta}$ meets

$$\frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} = \frac{1}{c(\theta)} (\hat{\theta} - \theta).$$

We can check this by taking the derivative with respect to θ and the expectation, i.e.,

$$\begin{aligned}\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} &= \frac{\partial \frac{1}{c(\theta)}}{\partial \theta} (\hat{\theta} - \theta) - \frac{1}{c(\theta)} \\ \Rightarrow E_{\mathbf{x}} \left[\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} \right] &= -\frac{1}{c(\theta)}.\end{aligned}$$

This is because $E[\hat{\theta}] = \theta$. Then,

$$\begin{aligned}c(\theta) &= \frac{1}{-E_{\mathbf{x}} \left[\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} \right]} = \frac{1}{I(\theta)} \\ \Rightarrow \text{var}(\hat{\theta}) &= \frac{1}{I(\theta)}.\end{aligned}$$

3.3 Example 4: Unbiased Estimator for DC level in White Gaussian noise

Design an unbiased estimator that estimates the parameter A from the observation $x[n] = A + w[n]$, $n = 0, \dots, N-1$, $\mathbf{w} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ by using the CRLB theorem. The likelihood function is

$$p(\mathbf{x}; A) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right).$$

- Check regularity condition $E_{\mathbf{x}} \left[\frac{\partial \ln(p(\mathbf{x}; A))}{\partial A} \right] = 0$

$$E \left[-\frac{\partial \ln(p(\mathbf{x}; A))}{\partial A} \right] = E \left[-\frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A) \right] = -\frac{1}{\sigma^2} \sum_{n=0}^{N-1} (E[x[n]] - A) = 0$$

- Evaluate CRLB

$$E \left[-\frac{\partial^2 \ln(p(\mathbf{x}; A))}{\partial^2 A} \right] = E \left[\frac{N}{\sigma^2} \right] = \frac{N}{\sigma^2} \Rightarrow \text{var}(\hat{A}) \geq \frac{\sigma^2}{N} \quad (6)$$

- $I(A) = \frac{N}{\sigma^2}$

- Apply iff condition (5), i.e., $\frac{\partial \ln(p(\mathbf{x}; A))}{\partial A} = I(A) (\hat{A}_{opt} - A)$, to find $\hat{A}_{opt}$

$$\begin{aligned}\frac{\partial \ln(p(\mathbf{x}; A))}{\partial A} &= I(A) (\hat{A}_{opt} - A) = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A) \\ &= \frac{N}{\sigma^2} \frac{1}{N} \sum_{n=1}^{N-1} (x[n] - A) \\ &= \frac{N}{\sigma^2} \left(\frac{1}{N} \sum_{n=1}^{N-1} x[n] - A \right)\end{aligned}$$

- $\hat{A}_{opt} = \frac{1}{N} \sum_{n=1}^{N-1} x[n]$ is the **optimal estimator** in the sense that

- It is **unbiased**
- It achieves the **minimum variance** $\text{var}(\hat{A}) = \frac{\sigma^2}{N}$. Why? ■

Definition 1 We call an estimator is **efficient**, if it is **unbiased** and **attains the CRLB**.

3.4 Fisher Information $I(\theta)$

The $I(\theta)$ is called **Fisher Information**

$$I(\theta) = E \left[-\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} \right].$$

As we have already known, it is the measure of **average sharpness or concentration of information** about θ contained in $\mathbf{x}$.

Estimation Theory: Design the estimator of θ to maximize Fisher information $I(\theta)$.

Information theory: Design the coding scheme θ to maximize Mutual information $\tilde{I}(\theta) = E \left[\ln \frac{p(\mathbf{x}; \theta)}{p(\mathbf{x})p(\theta)} \right]$, a value for the amount of information about θ contained in $\mathbf{x}$.

See? There is fundamental similarity between Estimation Theory and Information Theory

3.5 Affine or Linear Mapping

Suppose that we have already constructed the **efficient** estimator $\hat{\theta}_{opt}$, i.e., the MVU estimator that achieves the CRLB,

$$E[\hat{\theta}_{opt}] = \theta \quad \text{and} \quad \text{var}(\hat{\theta}_{opt}) = \frac{1}{-E \left[\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} \right]}.$$

We want to estimate $\alpha = g(\theta) = a\theta + b$. Try CRLB

$$\text{var}(\hat{\alpha}) \geq \frac{\left(\frac{\partial g(\theta)}{\partial \theta} \right)^2}{-E \left[\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} \right]} = a^2 \text{var}(\hat{\theta}_{opt}). \quad (7)$$

Consider a linear mapping of $\hat{\theta}_{opt}$ to $\hat{\alpha} = a\hat{\theta}_{opt} + b$. Then,

$$\begin{aligned} E[\hat{\alpha}] &= aE[\hat{\theta}_{opt}] + b = a\theta + b = \alpha \Rightarrow \text{unbiased} \\ \text{var}(\hat{\alpha}) &= a^2 \text{var}(\hat{\theta}_{opt}) = a^2 \frac{1}{-E \left[\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} \right]}. \end{aligned} \quad (8)$$

From (7) and (8), we can conclude that **estimating linear function of parameters, e.g., $g(\theta) = a\theta + b$** is equivalent to **estimating the parameters, e.g., $\hat{\theta}_{opt}$ first and doing a linear mapping $g(\hat{\theta}_{opt}) = a\hat{\theta}_{opt} + b$.**

This does not hold for non-linear mapping in general.

3.6 CRLB-Vector Parameter θ

We now extend the CRLB-scalar case to estimate the vector parameter $\theta = [\theta_1, \dots, \theta_p] \in \mathbb{R}^{p \times 1}$. Again, we consider the **unbiased estimators**

$$b(\theta) = E[\hat{\theta}] - \theta = \mathbf{0}.$$

Theorem 2 Suppose the likelihood function $p(\mathbf{x}; \theta)$ satisfies the **regularity conditions**

$$E_{\mathbf{x}} \left[\frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} \right] = \mathbf{0}.$$

Then, the **covariance matrix of any unbiased estimator $\hat{\theta}$** satisfies

$$\mathbf{C}_{\hat{\theta}} - \mathbf{I}^{-1}(\theta) \geq \mathbf{0} \quad (9)$$

where the bound means $\mathbf{C}_{\hat{\theta}} - \mathbf{I}^{-1}(\theta)$ is positive definite. The Fisher information matrix $\mathbf{I}(\theta)$ is given by

$$[\mathbf{I}(\theta)]_{ij} = -E_{\mathbf{x}} \left[\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial \theta_i \partial \theta_j} \right] \quad (10)$$

An unbiased estimator may be found that achieves the bound (9) if and only if

$$\frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} = \mathbf{I}(\theta) (\hat{\alpha}(\mathbf{x}) - \theta). \quad (11)$$

The **MVU estimator is $\hat{\theta} = \hat{\alpha}(\mathbf{x})$** and its covariance matrix is $\mathbf{C}_{\hat{\theta}} = \mathbf{I}^{-1}(\theta)$.

A few comments concerning Theorem 2 are in order.

- The differentiation of a multivariate function $f(\mathbf{a})$ is $\frac{\partial f(\mathbf{a})}{\partial \mathbf{a}} = \left[\frac{\partial f(\mathbf{a})}{\partial a_1} \quad \frac{\partial f(\mathbf{a})}{\partial a_2} \quad \dots \quad \frac{\partial f(\mathbf{a})}{\partial a_p} \right]^T \in \mathbb{R}^{p \times 1}$.
For example, given $f(\mathbf{a}) = a_1^2 + a_2^2$ with $\mathbf{a} = [a_1 \ a_2]^T$,

$$\frac{\partial f(\mathbf{a})}{\partial \mathbf{a}} = \left[\begin{array}{c} \frac{\partial f(\mathbf{a})}{\partial a_1} \\ \frac{\partial f(\mathbf{a})}{\partial a_2} \end{array} \right] = \left[\begin{array}{c} 2a_1 \\ 2a_2 \end{array} \right].$$

- The **Fisher information matrix** is defined by

$$\mathbf{I}(\boldsymbol{\theta}) = -E \begin{bmatrix} \frac{\partial^2 \ln(p(\mathbf{x};\boldsymbol{\theta}))}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 \ln(p(\mathbf{x};\boldsymbol{\theta}))}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 \ln(p(\mathbf{x};\boldsymbol{\theta}))}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial^2 \ln(p(\mathbf{x};\boldsymbol{\theta}))}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln(p(\mathbf{x};\boldsymbol{\theta}))}{\partial \theta_2 \partial \theta_2} & \dots & \frac{\partial^2 \ln(p(\mathbf{x};\boldsymbol{\theta}))}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ln(p(\mathbf{x};\boldsymbol{\theta}))}{\partial \theta_p \partial \theta_1} & \dots & \frac{\partial^2 \ln(p(\mathbf{x};\boldsymbol{\theta}))}{\partial \theta_p \partial \theta_{p-1}} & \frac{\partial^2 \ln(p(\mathbf{x};\boldsymbol{\theta}))}{\partial \theta_p \partial \theta_p} \end{bmatrix} \quad (12)$$

- The i th diagonal element of $\mathbf{C}_{\hat{\boldsymbol{\theta}}}$ corresponds to the variance of the i th estimator, i.e., $[\mathbf{C}_{\hat{\boldsymbol{\theta}}}]_{ii} = \text{var}(\hat{\theta}_i)$, and the bound (9) gives us the bound on the estimation performance of $\hat{\theta}_i$

$$\text{var}(\hat{\theta}_i) \geq [\mathbf{I}^{-1}(\boldsymbol{\theta})]_{ii}$$

- The proof of the theorem is in Section 4.2.

3.7 CRLB-Multivariate Function of $\boldsymbol{\theta}$, $\mathbf{g}(\boldsymbol{\theta})$

We often need to estimate a **multivariate function** $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta}) \in \mathbb{R}^{r \times 1}$, where $\mathbf{g} : \mathbb{R}^{p \times 1} \rightarrow \mathbb{R}^{r \times 1}$ is a known function.

$$\boldsymbol{\theta} \in \mathbb{R}^{p \times 1} \xrightarrow{\mathbf{g}} \boldsymbol{\alpha} \in \mathbb{R}^{r \times 1}.$$

When the regularity condition, $E \left[\frac{\partial \ln(p(\mathbf{x};\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right] = \mathbf{0}$ satisfies, the **CRLB for the multivariate function $\mathbf{g}(\boldsymbol{\theta})$ is**

$$\mathbf{C}_{\hat{\boldsymbol{\alpha}}} - \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \mathbf{I}^{-1}(\boldsymbol{\theta}) \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \geq \mathbf{0} \quad (13)$$

where

$$\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} (\mathbf{g}(\boldsymbol{\theta}))^T = \begin{bmatrix} \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial g_2(\boldsymbol{\theta})}{\partial \theta_1} & \dots & \frac{\partial g_r(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_2} & \frac{\partial g_2(\boldsymbol{\theta})}{\partial \theta_2} & \dots & \frac{\partial g_r(\boldsymbol{\theta})}{\partial \theta_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_p} & \frac{\partial g_2(\boldsymbol{\theta})}{\partial \theta_p} & \dots & \frac{\partial g_r(\boldsymbol{\theta})}{\partial \theta_p} \end{bmatrix} \in \mathbb{R}^{p \times r}.$$

The proof of (13) is in Section 4.2.

3.8 Multivariate Linear or Affine Mapping

Similar to Section 3.5, the **efficiency of the multivariate estimator does not change under linear or affine mapping**. Suppose we are given the estimator $\hat{\boldsymbol{\theta}}$, and want to estimate

$$\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta} + \mathbf{b}$$

where $\mathbf{A} \in \mathbb{R}^{r \times p}$ and $\mathbf{b} \in \mathbb{R}^{r \times 1}$.

If $\hat{\boldsymbol{\theta}}$ is **efficient**, i.e., $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \mathbf{I}^{-1}(\boldsymbol{\theta})$ and $E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$, its linear mapping $\hat{\boldsymbol{\alpha}} = \mathbf{A}\hat{\boldsymbol{\theta}} + \mathbf{b}$ is also **efficient**.

- It is easy to check its unbiasedness by $E[\hat{\alpha}] = E[\mathbf{A}\hat{\theta} + \mathbf{B}] = \mathbf{A}E[\hat{\theta}] + \mathbf{B} = \mathbf{A}\theta + \mathbf{B} = \alpha$
- The covariance of $\hat{\alpha}$ is calculated by $\mathbf{C}_{\hat{\alpha}} = E[(\alpha - \hat{\alpha})(\alpha - \hat{\alpha})^T] = \mathbf{A}\mathbf{C}_{\hat{\theta}}\mathbf{A}^T$. Given $\mathbf{C}_{\hat{\alpha}} = \mathbf{A}\mathbf{C}_{\hat{\theta}}\mathbf{A}^T$ it is straightforward to show $\mathbf{C}_{\hat{\alpha}}$ achieves the CRLB in (13) by evaluating $\left(\frac{\partial \mathbf{g}(\theta)}{\partial \theta}\right)^T \mathbf{I}^{-1}(\theta) \left(\frac{\partial \mathbf{g}(\theta)}{\partial \theta}\right)$.

4 Appendices

4.1 Proof of Theorem 1

Let $\alpha = g(\theta)$ and $\hat{\alpha}$ be the unbiased estimator of α such that $E[\hat{\alpha}] = \alpha = g(\theta)$ or

$$\int \hat{\alpha} p(\mathbf{x}; \alpha) d\mathbf{x} = g(\theta). \quad (14)$$

We first claim the regularity condition $E_{\mathbf{x}} \left[\frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} \right] = 0$ holds.

$$\begin{aligned} \int \frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} p(\mathbf{x}; \theta) d\mathbf{x} &= \int \frac{\partial p(\mathbf{x}; \theta)}{\partial \theta} d\mathbf{x}; \text{ why? } \frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} = \frac{1}{p(\mathbf{x}; \theta)} \frac{\partial p(\mathbf{x}; \theta)}{\partial \theta} \\ &= \int \lim_{h \rightarrow 0} \frac{p(\mathbf{x}; \theta + h) - p(\mathbf{x}; \theta)}{h} d\mathbf{x} \\ &= \lim_{h \rightarrow 0} \frac{\int p(\mathbf{x}; \theta + h) d\mathbf{x} - \int p(\mathbf{x}; \theta) d\mathbf{x}}{h} \\ &= \frac{\partial}{\partial \theta} \int p(\mathbf{x}; \theta) d\mathbf{x} = 0. \end{aligned} \quad (15)$$

In general, we can interchange integral and differentiation, when the function is measurable and bounded. Taking the derivative both sides of (14) with respect to θ ,

$$\begin{aligned} \int \hat{\alpha} \frac{\partial p(\mathbf{x}; \theta)}{\partial \theta} d\mathbf{x} &= \frac{\partial g(\theta)}{\partial \theta} \\ \int \hat{\alpha} \frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} p(\mathbf{x}; \theta) d\mathbf{x} - \alpha \int \frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} p(\mathbf{x}; \theta) d\mathbf{x} &= \frac{\partial g(\theta)}{\partial \theta}; \text{ follows from (15)} \\ \int (\hat{\alpha} - \alpha) \frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} p(\mathbf{x}; \theta) d\mathbf{x} &= \frac{\partial g(\theta)}{\partial \theta} \end{aligned} \quad (16)$$

By Cauchy-Schwartz inequality,

$$\left(\int w(\mathbf{x}) f(\mathbf{x}) h(\mathbf{x}) d\mathbf{x} \right)^2 \leq \int w(\mathbf{x}) f^2(\mathbf{x}) d\mathbf{x} \int w(\mathbf{x}) h^2(\mathbf{x}) d\mathbf{x} \quad (17)$$

with equality if and only if $f(\mathbf{x}) = ch(\mathbf{x})$ for any $c \in \mathbb{R}$, which is not a function of $\mathbf{x}$, we have from (16)

$$\begin{aligned} \left(\frac{\partial g(\theta)}{\partial \theta} \right)^2 &\leq \int p(\mathbf{x}; \theta) \left(\frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} \right)^2 d\mathbf{x} \int p(\mathbf{x}; \theta) (\hat{\alpha} - \alpha)^2 d\mathbf{x} \\ &= \text{var}(\hat{\alpha}) E_{\mathbf{x}} \left[\left(\frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} \right)^2 \right] \\ \text{var}(\hat{\alpha}) &\geq \frac{\left(\frac{\partial g(\theta)}{\partial \theta} \right)^2}{E_{\mathbf{x}} \left[\left(\frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} \right)^2 \right]}. \end{aligned}$$

On the other hand, from (15)

$$\begin{aligned}
0 &= \int \frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} p(\mathbf{x}; \theta) d\mathbf{x} = \int \left(\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} p(\mathbf{x}; \theta) + \frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} \frac{\partial p(\mathbf{x}; \theta)}{\partial \theta} \right) d\mathbf{x} \\
&= \int \left(\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} p(\mathbf{x}; \theta) + \frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} \frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} p(\mathbf{x}; \theta) \right) d\mathbf{x} \\
&\Rightarrow -E_{\mathbf{x}} \left[\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} \right] = E_{\mathbf{x}} \left[\left(\frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} \right)^2 \right]
\end{aligned} \tag{18}$$

$$\text{var}(\hat{\alpha}) \geq \frac{\left(\frac{\partial g(\theta)}{\partial \theta} \right)^2}{-E_{\mathbf{x}} \left[\frac{\partial^2 \ln(p(\mathbf{x}; \theta))}{\partial^2 \theta} \right]}. \tag{19}$$

If $c \frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} = (\hat{\alpha} - \alpha)$, the bound (19) is achieved.

$$\frac{\partial \ln(p(\mathbf{x}; \theta))}{\partial \theta} = \frac{1}{c(\theta)} (\hat{\alpha} - \alpha) \tag{20}$$

which equals to (5) and $\hat{\alpha}_{opt} = \hat{\alpha}$. ■

4.2 Proof of Theorem 2 or CRLB in (13)

Let $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta})$ such that $E[\hat{\alpha}_i] = \alpha_i = [\mathbf{g}(\boldsymbol{\theta})]_i$, $i = 1, 2, \dots, r$. The Theorem 2 is the special case with $\mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}$.

The regularity condition $E_{\mathbf{x}} \left[\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right] = \mathbf{0}$ is now easy to show by looking at $E_{\mathbf{x}} \left[\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right]_i = E_{\mathbf{x}} \left[\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \theta_i} \right]$ and applying (15).

Similarly, from (16), we have

$$\int (\hat{\alpha}_i - \alpha_i) \frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \theta_i} p(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} = \frac{\partial [\mathbf{g}(\boldsymbol{\theta})]_i}{\partial \theta_i}.$$

Now consider for $j \neq i$ and

$$\int (\hat{\alpha}_i - \alpha_i) \frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \theta_j} p(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} = \int \hat{\alpha}_i \frac{\partial p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_j} d\mathbf{x} - \int \alpha_i \frac{\partial p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_j} d\mathbf{x}. \tag{21}$$

Notice that $\hat{\alpha}_i$ is the function of $\mathbf{x}$, while α_i is not a function of $\mathbf{x}$. Thus, we have $\int \alpha_i \frac{\partial p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_j} d\mathbf{x} = \alpha_i \int \frac{\partial p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_j} d\mathbf{x} = 0$ from (15). This simplifies (21) as

$$\begin{aligned}
\int (\hat{\alpha}_i - \alpha_i) \frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \theta_j} p(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} &= \int \hat{\alpha}_i \frac{\partial p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_j} d\mathbf{x} = \frac{\partial}{\partial \theta_j} \int \hat{\alpha}_i p(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} \\
&= \frac{\partial \alpha_i}{\partial \theta_j} = \frac{\partial [\mathbf{g}(\boldsymbol{\theta})]_i}{\partial \theta_j}.
\end{aligned} \tag{22}$$

Collecting (22) for all i, j in a vector form yields

$$\int (\hat{\alpha} - \alpha) \left(\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right)^T p(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} = \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T. \quad (23)$$

Similar to the scalar case, we want to find a bound of (23) by applying Cauchy-Schwarz inequality. Notice that Cauchy-Schwarz inequality is defined for scalar operands. In order to apply the bound, we employ arbitrary vectors $\mathbf{a} \in \mathbb{R}^{r \times 1}$, $\mathbf{b} \in \mathbb{R}^{p \times 1}$ so that

$$\int \underbrace{\mathbf{a}^T (\hat{\alpha} - \alpha)}_{f(\mathbf{x})} \underbrace{\left(\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right)^T}_{h(\mathbf{x})} \underbrace{\mathbf{b} p(\mathbf{x}; \boldsymbol{\theta})}_{w(\mathbf{x})} d\mathbf{x} = \mathbf{a}^T \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \mathbf{b}$$

and applying the Cauchy-Schwarz inequality in (17) yields

$$\begin{aligned} \left(\mathbf{a}^T \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \mathbf{b} \right)^2 &\leq \int \mathbf{a}^T (\hat{\alpha} - \alpha) (\hat{\alpha} - \alpha)^T \mathbf{a} p(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} \\ &\quad \times \int \mathbf{b}^T \left(\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right)^T \mathbf{b} p(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} \\ &= \mathbf{a}^T \mathbf{C}_{\hat{\alpha}} \mathbf{a} \mathbf{b}^T \mathbf{I}(\boldsymbol{\theta}) \mathbf{b}. \end{aligned}$$

The last equality follows from (18) that specifies

$$E_{\mathbf{x}} \left[\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \theta_i} \frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \theta_j} \right] = -E \left[\frac{\partial^2 \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \theta_i \partial \theta_j} \right] = [\mathbf{I}(\boldsymbol{\theta})]_{ij}. \quad (24)$$

Since $\mathbf{a}$ and $\mathbf{b}$ are arbitrary, we set $\mathbf{b} = \mathbf{I}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{a}$. Then

$$\begin{aligned} \left(\mathbf{a}^T \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \mathbf{I}^{-1}(\boldsymbol{\theta}) \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \mathbf{a} \right)^2 &\leq \mathbf{a}^T \mathbf{C}_{\hat{\alpha}} \mathbf{a} \left(\mathbf{a}^T \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \mathbf{I}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{a} \right) \quad (25) \\ \Rightarrow \mathbf{a}^T \left(\mathbf{C}_{\hat{\alpha}} - \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \mathbf{I}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \mathbf{a} &\geq 0 \\ \Rightarrow \mathbf{C}_{\hat{\alpha}} - \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \mathbf{I}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &\geq 0 \end{aligned}$$

implying $\mathbf{C}_{\hat{\alpha}} - \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \mathbf{I}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ must be positive semidefinite.

By the condition for the equality

$$\begin{aligned} \mathbf{a}^T (\hat{\alpha} - \alpha) &= c \left(\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right)^T \mathbf{b} \\ &= c \left(\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right)^T \mathbf{I}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{a}. \end{aligned}$$

Since $\mathbf{a}$ is arbitrary

$$\begin{aligned} \frac{1}{c}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) &= \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \mathbf{I}^{-1}(\boldsymbol{\theta}) \frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \\ \Rightarrow \frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} &= \frac{1}{c(\boldsymbol{\theta})} \left(\left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \mathbf{I}^{-1}(\boldsymbol{\theta}) \right)^{-1} (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}). \end{aligned}$$

Notice that $\left(\left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \mathbf{I}^{-1}(\boldsymbol{\theta}) \right)^{-1}$ exists, when $\left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \mathbf{I}^{-1}(\boldsymbol{\theta})$ is invertible.

If $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, then $\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{I}$,

$$\mathbf{C}_{\hat{\boldsymbol{\alpha}}} - \mathbf{I}^{-1}(\boldsymbol{\theta}) \geq 0.$$

and the if and only if condition to achieve the CRLB is

$$\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \frac{1}{c(\boldsymbol{\theta})} \mathbf{I}(\boldsymbol{\theta}) (\hat{\boldsymbol{\alpha}} - \boldsymbol{\theta}). \quad (26)$$

Note that $c(\boldsymbol{\theta})$ may depend on $\boldsymbol{\theta}$ or not. To be more rigorous, we focus on the i th element of $\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}$

$$\begin{aligned} \frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \theta_i} &= \sum_{k=1}^p \frac{[\mathbf{I}(\boldsymbol{\theta})]_{ik}}{c(\boldsymbol{\theta})} (\hat{\theta}_k - \theta_k) \\ \frac{\partial^2 \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \theta_i \partial \theta_j} &= \sum_{k=1}^p \left(\frac{[\mathbf{I}(\boldsymbol{\theta})]_{ik}}{c(\boldsymbol{\theta})} (-\delta_{jk}) + \frac{\partial \left(\frac{[\mathbf{I}(\boldsymbol{\theta})]_{ik}}{c(\boldsymbol{\theta})} \right)}{\partial \theta_j} (\hat{\theta}_k - \theta_k) \right). \end{aligned}$$

From (24) and the above, we have

$$-E \left[\frac{\partial^2 \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \theta_i \partial \theta_j} \right] = [\mathbf{I}(\boldsymbol{\theta})]_{ij} = \frac{[\mathbf{I}(\boldsymbol{\theta})]_{ij}}{c(\boldsymbol{\theta})}$$

and thus $c(\boldsymbol{\theta}) = 1$. When $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, the if and only if condition for the achievability is

$$\frac{\partial \ln(p(\mathbf{x}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta}) (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}).$$