

# ECE531 Lecture 9: Information Inequality and the Cramer-Rao Lower Bound

D. Richard Brown III

Worcester Polytechnic Institute

30-March-2011

# Introduction

- ▶ In this lecture, we continue our study of estimators under the **squared error** cost function.
- ▶ Squared error: Estimator variance determines performance.
- ▶ We will develop a new procedure for finding MVU estimators:
  - ▶ Compute a lower bound on the variance.
  - ▶ Guess at a good unbiased estimator and compare its variance to the lower bound.
  - ▶ If a given unbiased estimator achieves the lower bound, it must be the MVU estimator. **“Guessing and checking” might be easier sometimes than grinding through the RBL theorem.**
- ▶ A good lower bound also provides a benchmark by which we can compare the performance of different estimators.
- ▶ We will develop a lower bound on estimator variance that can be applied to **both biased and unbiased** estimators.
- ▶ In the special case of unbiased estimators, this lower bound simplifies to the famous Cramer-Rao lower bound (CRLB).

## Intuition: When Can We Expect Low Variance?

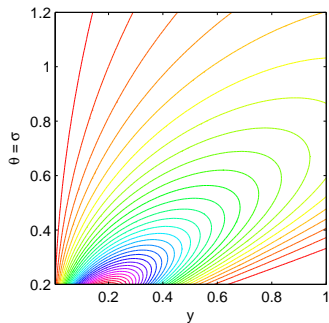
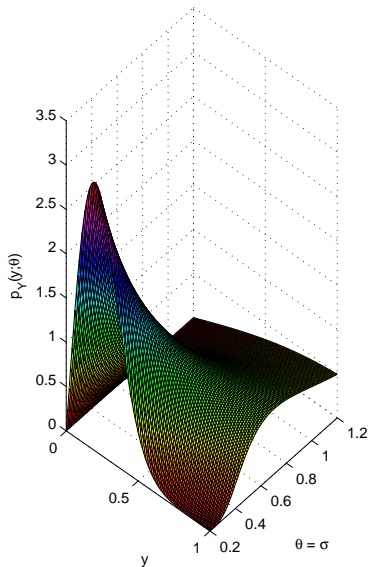
Suppose our parameter space  $\Lambda = \mathbb{R}$  and we get a scalar observation distributed as  $p_Y(y; \theta) = \mathcal{U}(0, 1)$ . What can we say about the performance of a good estimator  $\hat{\theta}(y)$  in this case?

Suppose now that we get a scalar observation distributed as  $p_Y(y; \theta) = \mathcal{U}(\theta - \epsilon, \theta + \epsilon)$  for some small value of  $\epsilon$ . What can we say about the performance of a good estimator  $\hat{\theta}(y)$  in this case?

# Intuition: When Can We Expect Low Variance?

- ▶ The minimum achievable variance of an estimator is somehow related to the **sensitivity** of the density  $p_Y(y; \theta)$  to changes in the parameter  $\theta$ .
- ▶ If the density  $p_Y(y; \theta)$  is **insensitive** to the parameter  $\theta$ , then we can't expect even the MVU estimator to do very well.
- ▶ If the density  $p_Y(y; \theta)$  is **sensitive** to changes in the parameter  $\theta$ , then the achievable performance (minimum variance) should be better.
- ▶ Our notion of sensitivity:
  - ▶ Hold  $y$  fixed.
  - ▶ How “steep” is  $p_Y(y; \theta)$  as we vary the parameter  $\theta$ ?
  - ▶ This steepness should somehow be averaged over the observations.
- ▶ Terminology: When we discuss  $p_Y(y; \theta)$  with  $y$  fixed and  $\theta$  as a variable, we call this a “likelihood function”. It is not a valid pdf in  $\theta$ .  
**Recall that  $\theta$  is not a random variable.**

Example: Rayleigh Family  $p_Y(y; \theta) = \frac{y}{\sigma^2} e^{-\frac{y^2}{\sigma^2}}$  with  $\theta = \sigma$



# Calculus Review

Some useful results:

$$\frac{\partial}{\partial \theta} \ln f(\theta) = \frac{\frac{\partial}{\partial \theta} f(\theta)}{f(\theta)}$$

$$\begin{aligned} \frac{\partial^2}{\partial \theta^2} \ln f(\theta) &= \frac{\partial}{\partial \theta} \left[ \frac{\frac{\partial}{\partial \theta} f(\theta)}{f(\theta)} \right] \\ &= \frac{\left( \frac{\partial^2}{\partial \theta^2} f(\theta) \right) f(\theta) - \left( \frac{\partial}{\partial \theta} f(\theta) \right)^2}{f^2(\theta)} \\ &= \frac{\frac{\partial^2}{\partial \theta^2} f(\theta)}{f(\theta)} - \left( \frac{\partial}{\partial \theta} \ln f(\theta) \right)^2 \end{aligned}$$

# A Definition of “Sensitivity” (Scalar Parameter – part 1)

- ▶ We require the likelihood function  $p_Y(y; \theta)$  to be differentiable with respect to  $\theta$  for each  $y \in \mathcal{Y}$ .
- ▶ Holding  $y$  fixed, the **relative steepness** of the likelihood function  $p_Y(y; \theta)$  (as a function of  $\theta$ ) can be expressed as

$$\psi(y; \theta) := \frac{\frac{\partial}{\partial \theta} p_Y(y; \theta)}{p_Y(y; \theta)} = \frac{\partial}{\partial \theta} \ln p_Y(y; \theta)$$

- ▶ Two problems:
  1. We don't care if the relative steepness is positive or negative. So we should square this result to give a non-negative measure of **squared relative steepness**.
  2. This “relative steepness”  $\psi(y; \theta)$  or “squared relative steepness”  $\psi^2(y; \theta)$  is only for a **particular observation**  $Y = y$ . We need to average this result over  $Y$  (holding  $\theta$  fixed).

# A Definition of “Sensitivity” (Scalar Parameter – part 2)

- ▶ Averaging the squared relative steepness: We compute the mean squared value of  $\psi$  as

$$\begin{aligned}
 I(\theta) &:= \mathbb{E}_\theta[\psi^2(Y; \theta)] \\
 &= \mathbb{E}_\theta \left[ \left( \frac{\partial}{\partial \theta} \ln p_Y(Y; \theta) \right)^2 \right] \\
 &= \int_{\mathcal{Y}} \left( \frac{\partial}{\partial \theta} \ln p_Y(y; \theta) \right)^2 p_Y(y; \theta) dy \\
 &= \int_{\mathcal{Y}} \left( \frac{\frac{\partial}{\partial \theta} p_Y(y; \theta)}{p_Y(y; \theta)} \right)^2 p_Y(y; \theta) dy
 \end{aligned}$$

- ▶ Terminology:  $I(\theta)$  is called the “Fisher information” that the random observation  $Y$  can tell us, on average, about the parameter  $\theta$ .
- ▶ The Fisher information  $I(\theta)$  is only a function of  $\theta$  (and other known quantities). It is not a function of  $y$ .
- ▶ Fisher information  $\neq$  mutual information (information theory).



## Example: Single Sample of Unknown Parameter in Noise

Suppose we get one sample of an unknown parameter  $\theta \in \mathbb{R}$  corrupted by zero-mean additive Gaussian noise, i.e.  $Y = \theta + W$  where  $W \sim \mathcal{N}(0, \sigma^2)$ . The likelihood function is then

$$p_Y(y; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y - \theta)^2}{2\sigma^2}\right)$$

The relative slope of  $p_Y(y; \theta)$  with respect to  $\theta$  can be easily computed

$$\psi(y; \theta) := \frac{\frac{\partial}{\partial \theta} p_Y(y; \theta)}{p_Y(y; \theta)} = \frac{\theta - y}{\sigma^2}$$

The Fisher information is then

$$\begin{aligned} I(\theta) &= \int_{-\infty}^{\infty} \left(\frac{\theta - y}{\sigma^2}\right)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y - \theta)^2}{2\sigma^2}\right) dy \\ &= \frac{1}{\sqrt{2\pi}\sigma^2} \int_{-\infty}^{\infty} t^2 \exp\left(\frac{-t^2}{2}\right) dt = \frac{1}{\sigma^2} \end{aligned}$$

## Theorem (The Information Inequality (scalar parameter))

Suppose that  $\hat{\theta}(y)$  is an estimator of the parameter  $\theta$  and that we have a family of densities  $\{p_Y(y; \theta); \theta \in \Lambda\}$ . If the following conditions hold:

1.  $\Lambda$  is an open interval
2.  $\mathcal{Y}_\theta := \{y \in \mathcal{Y} \mid p_Y(y; \theta) > 0\}$  is the same for all  $\theta \in \Lambda$  (all densities in the family share common support in  $\mathcal{Y}$ )
3.  $\frac{\partial}{\partial \theta} p_Y(y; \theta)$  exists and is finite for all  $\theta \in \Lambda$  and all  $y$  in the common support of  $\{p_Y(y; \theta); \theta \in \Lambda\}$
4.  $\frac{\partial}{\partial \theta} \int_{\mathcal{Y}} h(y) p_Y(y; \theta) dy$  exists and equals  $\int_{\mathcal{Y}} h(y) \frac{\partial}{\partial \theta} p_Y(y; \theta) dy$  for all  $\theta \in \Lambda$ , for  $h(y) = \hat{\theta}(y)$  and  $h(y) = 1$

then

$$\text{var}_\theta[\hat{\theta}(Y)] \geq \frac{\left[ \frac{\partial}{\partial \theta} \mathbf{E}_\theta \left\{ \hat{\theta}(Y) \right\} \right]^2}{I(\theta)}$$

## Example (continued)

Let's return to our example where we get a scalar observation of an unknown parameter  $\theta \in \mathbb{R}$  in zero-mean Gaussian noise:

$$p_Y(y; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y - \theta)^2}{2\sigma^2}\right)$$

We've already computed the Fisher information:

$$I(\theta) = \frac{1}{\sigma^2}$$

Suppose we restrict our attention to unbiased estimators. What can we say about  $\frac{\partial}{\partial \theta} \mathbb{E}_\theta \{ \hat{\theta}(Y) \}$ ?

Since all of the regularity conditions are satisfied (check this!), we can say

$$\text{var}_\theta[\hat{\theta}(Y)] \geq \sigma^2.$$

The estimator  $\hat{\theta}(y) = y$  is unbiased and it is easy to show that it achieves this minimum variance bound. Hence  $\hat{\theta}(y) = y$  is MVU. No need to use RBLS.

# The Information Inequality (Scalar Parameter)

A key claim in the proof of the information inequality:

$$\frac{\partial}{\partial \theta} \mathbb{E}_{\theta} \left\{ \hat{\theta}(Y) \right\} = \int_{\mathcal{Y}} \left[ \hat{\theta}(y) - \mathbb{E}_{\theta} \left\{ \hat{\theta}(Y) \right\} \right] \frac{\partial}{\partial \theta} p_Y(y; \theta) dy.$$

To show this:

$$\begin{aligned} \int_{\mathcal{Y}} \left[ \hat{\theta}(y) - \mathbb{E}_{\theta} \left\{ \hat{\theta}(Y) \right\} \right] \frac{\partial}{\partial \theta} p_Y(y; \theta) dy &= \int_{\mathcal{Y}} \hat{\theta}(y) \frac{\partial}{\partial \theta} p_Y(y; \theta) dy \\ &\quad - \mathbb{E}_{\theta} \left\{ \hat{\theta}(Y) \right\} \int_{\mathcal{Y}} \frac{\partial}{\partial \theta} p_Y(y; \theta) dy \\ &\stackrel{c4}{=} \frac{\partial}{\partial \theta} \int_{\mathcal{Y}} \hat{\theta}(y) p_Y(y; \theta) dy \\ &\quad - \mathbb{E}_{\theta} \left\{ \hat{\theta}(Y) \right\} \frac{\partial}{\partial \theta} \int_{\mathcal{Y}} p_Y(y; \theta) dy \\ &= \frac{\partial}{\partial \theta} \mathbb{E}_{\theta} \left\{ \hat{\theta}(Y) \right\} - \mathbb{E}_{\theta} \left\{ \hat{\theta}(Y) \right\} \frac{\partial}{\partial \theta} (1) \\ &= \frac{\partial}{\partial \theta} \mathbb{E}_{\theta} \left\{ \hat{\theta}(Y) \right\} \end{aligned}$$

# The Information Inequality (Scalar Parameter)

Taking the claim as a given now, we can write:

$$\begin{aligned}
 \left[ \frac{\partial}{\partial \theta} \mathbb{E}_\theta \{ \hat{\theta}(Y) \} \right]^2 &= \left[ \int_{\mathcal{Y}} \left[ \hat{\theta}(y) - \mathbb{E}_\theta \{ \hat{\theta}(Y) \} \right] \frac{\partial}{\partial \theta} p_Y(y; \theta) dy \right]^2 \\
 &= \left[ \int_{\mathcal{Y}} \left[ \hat{\theta}(y) - \mathbb{E}_\theta \{ \hat{\theta}(Y) \} \right] \frac{\frac{\partial}{\partial \theta} p_Y(y; \theta)}{p_Y(y; \theta)} p_Y(y; \theta) dy \right]^2 \\
 &= \left[ \int_{\mathcal{Y}} \left[ \hat{\theta}(y) - \mathbb{E}_\theta \{ \hat{\theta}(Y) \} \right] \left[ \frac{\partial}{\partial \theta} \ln p_Y(y; \theta) \right] p_Y(y; \theta) dy \right]^2 \\
 &= \left[ \mathbb{E}_\theta \left\{ \left[ \hat{\theta}(Y) - \mathbb{E}_\theta \{ \hat{\theta}(Y) \} \right] \left[ \frac{\partial}{\partial \theta} \ln p_Y(Y; \theta) \right] \right\} \right]^2 \\
 &\stackrel{\text{Schwarz}}{\leq} \mathbb{E}_\theta \left\{ \left[ \hat{\theta}(Y) - \mathbb{E}_\theta \{ \hat{\theta}(Y) \} \right]^2 \right\} \mathbb{E}_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln p_Y(Y; \theta) \right]^2 \right\} \\
 &= \text{var}_\theta \left[ \hat{\theta}(Y) \right] \cdot I(\theta)
 \end{aligned}$$

Hence,  $\text{var}_\theta \left[ \hat{\theta}(Y) \right] \geq \frac{\left[ \frac{\partial}{\partial \theta} \mathbb{E}_\theta \{ \hat{\theta}(Y) \} \right]^2}{I(\theta)}$ , which is the result we wanted.

# Remarks

- ▶ A key consequence of the the regularity conditions that we used in our derivation of the bound is that

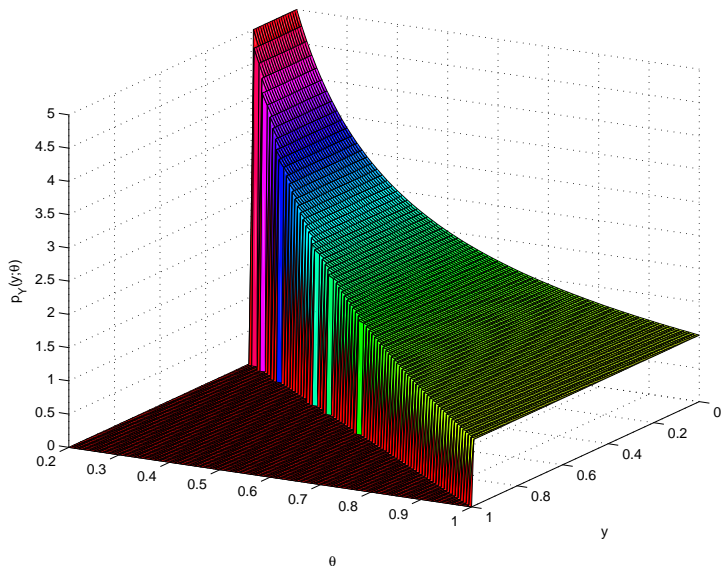
$$E_{\theta} \left[ \frac{\partial}{\partial \theta} \ln p_Y(Y; \theta) \right] = \int_{\mathcal{Y}} \frac{\partial}{\partial \theta} p_Y(y; \theta) dy = 0 \text{ for all } \theta \in \Lambda$$

- ▶ Suppose our observations were  $Y \sim \mathcal{U}(0, \theta)$  for  $\theta > 0$ .
  - ▶ Obviously, this fails to satisfy our regularity conditions, e.g. lack common support for all densities  $p_Y(y; \theta)$ .
  - ▶ But the real problem is that

$$\begin{aligned} E_{\theta} \left[ \frac{\partial}{\partial \theta} \ln p_Y(Y; \theta) \right] &= \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} p_Y(y; \theta) dy = \int_0^{\theta} \frac{\partial}{\partial \theta} \frac{1}{\theta} dy \\ &= -\frac{1}{\theta^2} \int_0^{\theta} dy = -\frac{1}{\theta} \neq 0 \end{aligned}$$

- ▶ When  $E_{\theta} \left[ \frac{\partial}{\partial \theta} \ln p_Y(Y; \theta) \right] \neq 0$ , the whole derivation of the information inequality breaks down.
- ▶ Checking the regularity conditions is important.

$$Y \sim \mathcal{U}(0, \theta) \text{ for } \theta > 0$$



# The Information Inequality (Scalar Parameter)

## Lemma

If, in addition to conditions 1-4, we also have

5.  $\frac{\partial^2}{\partial \theta^2} p_Y(y; \theta)$  exists for all  $\theta \in \Lambda$  and  $y$  in the common support of  $p_Y(y; \theta)$  and

$$\int \frac{\partial^2}{\partial \theta^2} p_Y(y; \theta) dy = \frac{\partial^2}{\partial \theta^2} \int p_Y(y; \theta) dy$$

then  $I(\theta) = -E_{\theta} \left\{ \frac{\partial^2}{\partial \theta^2} \ln p_Y(Y; \theta) \right\}$ .

**Proof:** (Lehmann TPE 1998).

We can use our calculus result derived earlier to write

$$\frac{\partial^2}{\partial \theta^2} \ln p_Y(y; \theta) = \frac{\frac{\partial^2}{\partial \theta^2} p_Y(y; \theta)}{p_Y(y; \theta)} - \left[ \frac{\frac{\partial}{\partial \theta} p_Y(y; \theta)}{p_Y(y; \theta)} \right]^2.$$

The result follows by taking the expectation of both sides and applying condition 5.  $\square$



# Unbiased Estimators: The Cramer-Rao Lower Bound

For the particular case when the estimator  $\hat{\theta}(y)$  is unbiased, we know that

$$\mathbb{E}_{\theta} \left\{ \hat{\theta}(y) \right\} = \theta$$

and, consequently

$$\frac{\partial}{\partial \theta} \mathbb{E}_{\theta} \left\{ \hat{\theta}(y) \right\} = 1.$$

Hence,

$$\boxed{\text{var}_{\theta} \left\{ \hat{\theta}(y) \right\} \geq \frac{1}{I(\theta)}}.$$

This result is known as the Cramer-Rao lower bound (originally described by Fisher in 1922 but not well-known until Rao and Cramer worked on it in 1945 and 1946, respectively).

# Example: Estimating a Constant in White Gaussian Noise

Suppose we have  $n$  observations given by

$$Y_k = \theta + W_k \quad k = 0, \dots, n-1$$

where  $W_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ . The unknown parameter  $\theta$  can take on any value on the real line and we have no prior pdf.

We know from our study of the RBLT theorem that the sample mean estimator  $\hat{\theta}(y) = \bar{y} = \frac{1}{n} \sum_{k=0}^{n-1} y_k$  is MVU. Let's compute the CRLB to see if it also attains the minimum variance bound. We have

$$p_Y(y; \theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ \frac{-\sum_{k=0}^{n-1} (y_k - \theta)^2}{2\sigma^2} \right\}$$

It is not difficult to compute

$$\frac{\partial}{\partial \theta} \ln p_Y(y; \theta) = \frac{1}{\sigma^2} \sum_{k=0}^{n-1} (y_k - \theta) = \frac{n}{\sigma^2} (\bar{y} - \theta)$$

# Example: Estimating a Constant in White Gaussian Noise

Since condition 5 holds, we can take another derivative to get:

$$\frac{\partial^2}{\partial \theta^2} \ln p_Y(y; \theta) = \frac{-n}{\sigma^2}$$

Hence

$$I(\theta) = -\mathbb{E}_\theta \left[ \frac{\partial^2}{\partial \theta^2} \ln p_Y(Y; \theta) \right] = \frac{n}{\sigma^2}$$

and we can say that

$$\text{var}_\theta \left[ \hat{\theta}(Y) \right] \geq \frac{\sigma^2}{n}$$

for any unbiased estimator. Note that the lower bound is equal to the variance of our MVU estimator  $\hat{\theta}(y) = \bar{y}$ .

# Remarks

- ▶ If an unbiased estimator attains the CRLB, it must be MVU.
- ▶ The converse is not always true. In other words, not all MVU estimators attain the CRLB.
- ▶ An estimator that is unbiased and attains the CRLB is said to be **efficient**.
- ▶ When we had one observation, the information was  $I(\theta) = \frac{1}{\sigma^2}$ . When we had  $n$  observations, the information became  $I(\theta) = \frac{n}{\sigma^2}$ . This additive information property is only true when the observations are independent.

# Additive Information from Independent Observations

## Lemma

*If  $Y_1$  and  $Y_2$  are independent random variables satisfying all of the regularity conditions with densities  $p_{Y_1}(y; \theta)$  and  $p_{Y_2}(y; \theta)$  parameterized by  $\theta$  then*

$$I(\theta) = I_{Y_1}(\theta) + I_{Y_2}(\theta)$$

*where  $I_{Y_1}(\theta)$ ,  $I_{Y_2}(\theta)$ , and  $I(\theta)$  are the information about  $\theta$  contained in  $Y_1$ ,  $Y_2$ , and  $\{Y_1, Y_2\}$ , respectively.*

## Corollary

*If  $Y_0, \dots, Y_{n-1}$  are i.i.d. satisfying all of the regularity conditions, and each has information  $I(\theta)$  about  $\theta$ , then the information in  $\{Y_0, \dots, Y_{n-1}\}$  about  $\theta$  is  $nI(\theta)$ .*

# Additive Information from Independent Observations

## Proof of Lemma.

Since  $Y_1$  and  $Y_2$  are independent, their joint pdf can be written as a product of the marginals. We can then write

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \ln p_{Y_1}(Y_1; \theta) + \frac{\partial}{\partial \theta} \ln p_{Y_2}(Y_2; \theta) \right]^2 \\ &= I_{Y_1}(\theta) + 2\mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \ln p_{Y_1}(Y_1; \theta) \right] \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \ln p_{Y_2}(Y_2; \theta) \right] + I_{Y_2}(\theta) \end{aligned}$$

The term

$$\mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta} \ln p_{Y_1}(Y_1; \theta) \right] = \int_{\mathcal{Y}_1} \frac{\frac{\partial}{\partial \theta} p_{Y_1}(x; \theta)}{p_{Y_1}(x; \theta)} p_{Y_1}(x; \theta) dx = \frac{\partial}{\partial \theta} \int_{\mathcal{Y}_1} p_{Y_1}(x; \theta) dx = 0$$

and the desired result follows immediately.  $\square$

# CRLB for Signals in Zero-Mean White Gaussian Noise

We assume the general system model

$$Y_k = s_k(\theta) + W_k \text{ for } k = 0, 1, \dots, n-1$$

where  $s_k(\theta)$  is a deterministic signal with an unknown real-valued non-random scalar parameter  $\theta$  and where  $W_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ . We only assume that  $s_k(\theta)$  doesn't violate any of our regularity conditions.

To compute the Fisher information, we can differentiate twice to get:

$$\frac{\partial^2}{\partial \theta^2} \ln p_Y(y; \theta) = \frac{1}{\sigma^2} \sum_{k=0}^{n-1} \left\{ [y_k - s_k(\theta)] \frac{\partial^2}{\partial \theta^2} s_k(\theta) - \left( \frac{\partial}{\partial \theta} s_k(\theta) \right)^2 \right\}$$

We then take the expected value (over the observations) to get

$$I(\theta) = -\mathbb{E}_\theta \left[ \frac{\partial^2}{\partial \theta^2} \ln p_Y(Y; \theta) \right] = \frac{1}{\sigma^2} \sum_{k=0}^{n-1} \left( \frac{\partial}{\partial \theta} s_k(\theta) \right)^2$$

and the CRLB follows immediately as  $1/I(\theta)$ . Note additive information.

# Example: Sinusoidal Frequency Estimation in AWGN

Consider the case where

$$Y_k = a \cos(\theta k + \phi) + W_k \text{ for } k = 0, 1, \dots, n-1$$

where  $a$  and  $\phi$  are known,  $\theta \in (0, \pi)$ , and  $W_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ . You can confirm that the regularity conditions 1-5 are all satisfied here.

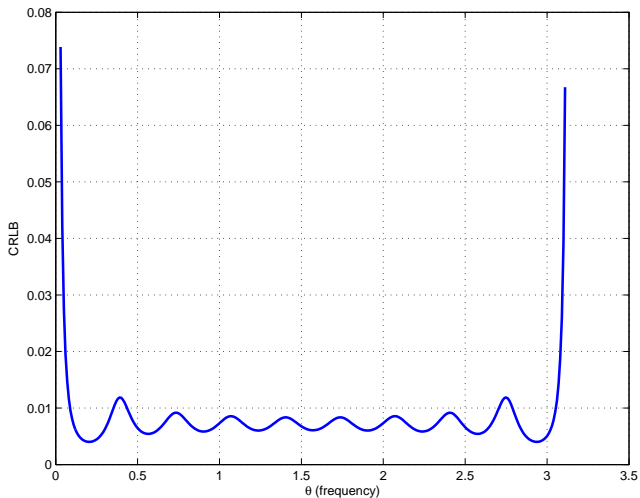
To compute the CRLB, we can apply our general result for signals in zero-mean AWGN.

$$\text{var}_\theta \left[ \hat{\theta}(Y) \right] \geq \frac{\sigma^2}{\sum_{k=0}^{n-1} \left( \frac{\partial}{\partial \theta} s_k(\theta) \right)^2} = \frac{\sigma^2}{a^2 \sum_{k=0}^{n-1} (k \sin(\theta k + \phi))^2}$$



# Example: Sinusoidal Frequency Estimation in AWGN

$n = 10$ ,  $\sigma^2/a^2 = 1$ , and  $\phi = 0$ .



# Attainability of the Information Bound

Recall that the only inequality used to derive the information inequality was the Cauchy-Schwarz inequality:

$$\left[ \int_a^b f_1(y; \theta) f_2(y; \theta) dy \right]^2 \leq \int_a^b f_1^2(y; \theta) dy \cdot \int_a^b f_2^2(y; \theta) dy$$

Under what conditions does this inequality become an equality? If and only if  $f_2(y; \theta) = k(\theta) f_1(y; \theta)$  on  $y \in (a, b)$ . In our problem, we have

$$f_1(y; \theta) = \hat{\theta}(y) - \mathbb{E}_\theta \left\{ \hat{\theta}(Y) \right\}$$

$$f_2(y; \theta) = \frac{\partial}{\partial \theta} \ln p_Y(y; \theta)$$

hence, an estimator  $\hat{\theta}(y)$  has variance equal to the information lower bound for all  $\theta \in \Lambda$  if and only if

$$\frac{\partial}{\partial \theta} \ln p_Y(Y; \theta) = k(\theta) \left[ \hat{\theta}(Y) - \mathbb{E}_\theta \left\{ \hat{\theta}(Y) \right\} \right]$$

almost surely for some  $k(\theta)$ .

# Attainability of the Information Bound

To attain the information bound, we require

$$\text{Prob} \left\{ \frac{\partial}{\partial \theta} \ln p_Y(Y; \theta) = k(\theta) \left[ \hat{\theta}(Y) - \mathbb{E}_\theta \left\{ \hat{\theta}(Y) \right\} \right] \right\} = 1$$

for some  $k(\theta)$ . What does this imply about the form of  $p_Y(Y; \theta)$  required to attain the information bound?

To answer that question, we can “undo” the derivative and the logarithm on the LHS to write

$$\begin{aligned} p_Y(y; \theta) &= h(y) \exp \left\{ \int_a^\theta k(t) \left[ \hat{\theta}(y) - \mathbb{E}_t \left\{ \hat{\theta}(Y) \right\} \right] dt \right\} \\ &= \underbrace{h(y) C(\theta) \exp \{g(\theta) T(y)\}}_{\text{one parameter exponential family}} \end{aligned}$$

for all  $y \in \mathcal{Y}$ . Note that  $h(y)$  does not depend on  $\theta$ .

# Attainability of the Information Bound

Remarks:

- ▶ The information bound is achieved by  $\hat{\theta}(y)$  if and only if  $\hat{\theta}(y) = T(y)$  in a one-parameter exponential family (the estimator is the sufficient statistic).
- ▶ It can also be shown that  $k(\theta)$  here must be equal to  $\frac{I(\theta)}{\frac{\partial}{\partial \theta} E_{\theta}[\hat{\theta}(Y)]}$ .

# Attainability of the Information Bound: Unbiased Case

When  $\hat{\theta}(y)$  is unbiased,  $E_{\theta} \left\{ \hat{\theta}(Y) \right\} = \theta$ . Hence, the necessary and sufficient attainability condition can be written as

$$\frac{\partial}{\partial \theta} \ln p_Y(Y; \theta) = k(\theta) \left[ \hat{\theta}(Y) - \theta \right]$$

almost surely for some  $k(\theta)$ . Squaring both sides and taking the expectation, we can write

$$E_{\theta} \left[ \left( \frac{\partial}{\partial \theta} \ln p_Y(Y; \theta) \right)^2 \right] = k^2(\theta) E_{\theta} \left[ \left( \hat{\theta}(Y) - \theta \right)^2 \right]$$

$$I(\theta) = k^2(\theta) \frac{1}{I(\theta)}$$

hence  $k(\theta) = \pm I(\theta)$ . The negative option can be eliminated. The necessary and sufficient attainability condition becomes

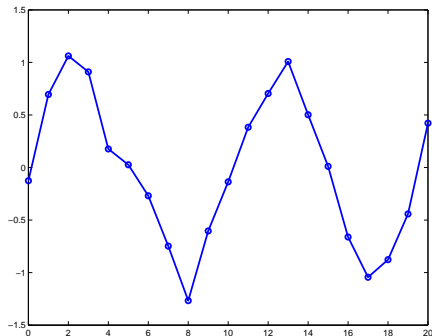
$$\frac{\partial}{\partial \theta} \ln p_Y(Y; \theta) = I(\theta) \left[ \hat{\theta}(Y) - \theta \right]$$

# Multiparameter Estimation Problems

In many problems, we have more than one parameter that we would like to estimate. For example,

$$Y_k = a \cos(\omega k + \phi) + W_k \text{ for } k = 0, 1, \dots, n - 1$$

where  $a > 0$ ,  $\phi \in (-\pi, \pi)$ , and  $\omega \in (0, \pi)$  are all non-random parameters and  $W_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ . In this problem  $\theta = [a, \phi, \omega]$ .



# Fisher Information Matrix

Recall, in the scalar parameter case, the Fisher information was motivated by a computation of the mean squared relative slope of the likelihood function:

$$I(\theta) := \mathbb{E}_\theta \left[ \left( \frac{\frac{\partial}{\partial \theta} p_Y(y; \theta)}{p_Y(y; \theta)} \right)^2 \right] = \int_{\mathcal{Y}} \left( \frac{\partial}{\partial \theta} \ln p_Y(Y; \theta) \right)^2 p_Y(y; \theta) dy$$

In multiparameter problems, we are now concerned with the relative slope of the likelihood function with respect to each of the parameters. A natural choice (assuming that all of the required derivatives exist) would be

$$I(\theta) = \mathbb{E}_\theta \left[ (\nabla_\theta \ln p_Y(Y; \theta)) (\nabla_\theta \ln p_Y(Y; \theta))^\top \right] \in \mathbb{R}^{m \times m}$$

where  $\nabla_x$  is the gradient operator defined as

$$\nabla_x f(x) := \left[ \frac{\partial}{\partial x_0} f(x), \dots, \frac{\partial}{\partial x_{m-1}} f(x) \right]^\top.$$

# Fisher Information Matrix

Let  $p := p_Y(Y; \theta)$ . The Fisher information matrix is then

$$I(\theta) = \begin{bmatrix} \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta_0} \ln p \cdot \frac{\partial}{\partial \theta_0} \ln p \right] & \dots & \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta_0} \ln p \cdot \frac{\partial}{\partial \theta_{m-1}} \ln p \right] \\ \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta_1} \ln p \cdot \frac{\partial}{\partial \theta_0} \ln p \right] & \dots & \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta_1} \ln p \cdot \frac{\partial}{\partial \theta_{m-1}} \ln p \right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta_{m-1}} \ln p \cdot \frac{\partial}{\partial \theta_0} \ln p \right] & \dots & \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta_{m-1}} \ln p \cdot \frac{\partial}{\partial \theta_{m-1}} \ln p \right] \end{bmatrix}$$

Note that the  $ij$ th element of the Fisher information matrix is given as

$$I_{ij}(\theta) = \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta_i} \ln p_Y\{Y; \theta\} \cdot \frac{\partial}{\partial \theta_j} \ln p_Y\{Y; \theta\} \right]$$

hence we can say that  $I(\theta)$  is symmetric.



# Fisher Information Matrix

When the second derivatives all exist, we can write

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln p_Y(y; \theta) = \frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j} p_Y(y; \theta)}{p_Y(y; \theta)} - \frac{\frac{\partial}{\partial\theta_i} p_Y(y; \theta)}{p_Y(y; \theta)} \frac{\frac{\partial}{\partial\theta_j} p_Y(y; \theta)}{p_Y(y; \theta)}$$

and, under the regularity assumptions, we can write

$$E_\theta \left[ \frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln p_Y(y; \theta) \right] = -E_\theta \left[ \frac{\partial}{\partial\theta_i} \ln p_Y(y; \theta) \cdot \frac{\partial}{\partial\theta_j} \ln p_Y(y; \theta) \right] = -I_{ij}(\theta).$$

Hence, we can say that

$$I_{ij}(\theta) = -E_\theta \left[ \frac{\partial^2}{\partial\theta_i\partial\theta_j} \ln p_Y(y; \theta) \right]$$

This expression is often more convenient to compute than the former expression for  $I_{ij}(\theta)$ .

# Fisher Information Matrix

Under our regularity conditions

$$\begin{aligned} \mathbb{E}_\theta \left[ \frac{\partial}{\partial \theta_i} \ln p_Y\{Y; \theta\} \right] &= \int_{\mathcal{Y}} \frac{\frac{\partial}{\partial \theta_i} p_Y(y; \theta)}{p_Y(y; \theta)} p_Y(y; \theta) dy \\ &= \frac{\partial}{\partial \theta_i} \int_{\mathcal{Y}} p_Y(y; \theta) dy = 0 \end{aligned}$$

Hence

$$I_{ij}(\theta) = \text{cov}_\theta \left\{ \frac{\partial}{\partial \theta_i} \ln p_Y\{Y; \theta\}, \frac{\partial}{\partial \theta_j} \ln p_Y\{Y; \theta\} \right\}.$$

- ▶ Since  $I(\theta)$  is a covariance matrix,  $I(\theta)$  is positive semidefinite.
- ▶ The information inequality and CRLB for scalar parameters required us to compute  $\frac{1}{I(\theta)}$ . We can expect that we might need to compute  $I^{-1}(\theta)$  when we have vector parameter.
- ▶ For  $I(\theta)$  to be invertible, we need  $I(\theta)$  to be positive definite.

# Covariance Matrices and Positive Definiteness

- ▶ Suppose  $X \in \mathbb{R}^m$  is a zero mean random vector.
- ▶ A covariance matrix  $\text{cov}(X, X) = \text{E}[X X^\top]$  fails to be positive definite only if one or more random variables  $X_i$  can be written as linear combinations of the other random variables.
- ▶ The random variables (parameterized by  $\theta$ ) in the Fisher information matrix are  $X_i(\theta) := \frac{\partial}{\partial \theta_i} \ln p_Y\{Y; \theta\}$   $i = 0, \dots, m - 1$ .
- ▶ What can we do if  $\{X_i(\theta)\}_{i=0}^{m-1}$  are linearly dependent?
- ▶ Linear dependence implies that one or more of the  $X_i$  are extraneous.
- ▶ We can excise the extraneous random variables to form a smaller set of linearly independent variables  $\{X'_i(\theta)\}_{i=0}^{p-1}$  with  $p < m$  such that  $\text{cov}(X', X')$  is positive definite.

The bottom line here is that we don't need to worry about the Fisher information matrix being non-invertible. If it is non-invertible, just excise one or more linearly dependent random variables to get a set of linearly independent random variables and  $I(\theta)$  will become invertible.

# Example: Fisher Information Matrix of Signal in AWGN

We assume the general system model

$$Y_k = s_k(\theta) + W_k \text{ for } k = 0, 1, \dots, n-1$$

where  $s_k(\theta) : \Lambda \mapsto \mathbb{R}$  is a deterministic signal with an unknown **vector** parameter  $\theta$  and where  $W_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ . We assume  $\sigma^2$  is a known parameter and that all of the regularity conditions are satisfied.

To compute the Fisher information matrix, we can write

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln p_Y(Y; \theta) = \frac{1}{\sigma^2} \sum_{k=0}^{n-1} \left\{ [Y_k - s_k(\theta)] \frac{\partial^2}{\partial \theta_i \partial \theta_j} s_k(\theta) - \left( \frac{\partial}{\partial \theta_i} s_k(\theta) \right) \left( \frac{\partial}{\partial \theta_j} s_k(\theta) \right) \right\}$$

Since  $E_{\theta}[Y_k] = s_k(\theta)$ , the  $ij$ th element of the FIM can be written as

$$I_{ij}(\theta) = -E_{\theta} \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln p_Y(Y; \theta) \right] = \frac{1}{\sigma^2} \sum_{k=0}^{n-1} \left( \frac{\partial}{\partial \theta_i} s_k(\theta) \right) \left( \frac{\partial}{\partial \theta_j} s_k(\theta) \right)$$

# Example: Fisher Information for Amplitude and Phase

Consider the case where

$$Y_k = a \cos(\omega k + \phi) + W_k \text{ for } k = 0, 1, \dots, n-1$$

where  $\omega$  is known,  $a > 0$  and  $\phi \in (-\pi, \pi)$  are unknown, and

$W_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$  with  $\sigma^2$  known. Let  $\theta = [a, \phi]^\top$  and compute the Fisher information matrix:

$$\begin{aligned} I_{00}(\theta) &= \frac{1}{\sigma^2} \sum_{k=0}^{n-1} \left( \frac{\partial}{\partial a} s_k(\theta) \right)^2 \\ &= \frac{1}{\sigma^2} \sum_{k=0}^{n-1} \cos^2(\omega k + \phi) = \frac{1}{\sigma^2} \sum_{k=0}^{n-1} \left( \frac{1}{2} + \cos(2(\omega k + \phi)) \right) \approx \frac{n}{2\sigma^2} \\ I_{11}(\theta) &= \frac{1}{\sigma^2} \sum_{k=0}^{n-1} \left( \frac{\partial}{\partial \phi} s_k(\theta) \right)^2 = \frac{1}{\sigma^2} \sum_{k=0}^{n-1} a^2 \sin^2(\omega k + \phi) \approx \frac{na^2}{2\sigma^2} \\ I_{01}(\theta) &= \frac{1}{\sigma^2} \sum_{k=0}^{n-1} \left( \frac{\partial}{\partial a} s_k(\theta) \right) \left( \frac{\partial}{\partial \phi} s_k(\theta) \right) = -\frac{1}{\sigma^2} \sum_{k=0}^{n-1} a \sin(\omega k + \phi) \cos(\omega k + \phi) \approx 0 \end{aligned}$$

# Example: Fisher Information for Amplitude and Phase

Remarks:

- ▶ The Fisher information matrix in this example is

$$I(\theta) = \frac{n}{2\sigma^2} \begin{bmatrix} 1 & 0 \\ 0 & a^2 \end{bmatrix}$$

- ▶ Clearly  $I(\theta)$  is positive definite when  $a > 0$ .
- ▶ Note that, since the observations are i.i.d.,  $I(\theta)$  satisfies the additive information property (as expected).
- ▶ We got lucky that the off-diagonal terms are (at least approximately) equal to zero here. The matrix inverse is easy to compute here. This will not be true in general.

# Information Inequality for Multiparameter Estimation

Under the multiparameter regularity conditions (see Lehmann TPE 1998 p. 127) and also assuming  $I(\theta)$  is positive definite, we can say that

$$\boxed{\text{cov}_\theta \left[ \hat{\theta}(Y) \right] \geq \beta^\top(\theta) I^{-1}(\theta) \beta(\theta)}$$

where the matrix inequality  $A \geq B$  means that  $A - B$  is positive semi-definite and

$$\beta^\top(\theta) := \left[ \frac{\partial}{\partial \theta_0} \mathbb{E}_\theta \left[ \hat{\theta}(Y) \right], \dots, \frac{\partial}{\partial \theta_{m-1}} \mathbb{E}_\theta \left[ \hat{\theta}(Y) \right] \right].$$

Note that  $\beta(\theta) \in \mathbb{R}^{m \times m}$  since  $\mathbb{E}_\theta \left[ \hat{\theta}(Y) \right] \in \mathbb{R}^m$ . What is  $\beta^\top(\theta)$  for an unbiased estimator?

It should be clear that this is equivalent to our scalar parameter result when we have only one parameter.

# Multiparameter Cramer-Rao Lower Bound

If we constrain our attention to unbiased estimators, the multiparameter Cramer-Rao lower bound (CRLB) can be simply expressed as

$$\boxed{\text{cov}_\theta \left[ \hat{\theta}(Y) \right] \geq I^{-1}(\theta)}$$

since

$$\beta^\top(\theta) := \left[ \frac{\partial}{\partial \theta_0} \text{E}_\theta \left[ \hat{\theta}(Y) \right], \dots, \frac{\partial}{\partial \theta_{m-1}} \text{E}_\theta \left[ \hat{\theta}(Y) \right] \right]$$

in the information inequality is just the  $m \times m$  identity matrix when the estimator is unbiased.



## Example: Amplitude and Phase Information Bound

We can compute the inverse of the Fisher information matrix easily:

$$I^{-1}(\theta) = \frac{2\sigma^2}{n} \begin{bmatrix} 1 & 0 \\ 0 & a^{-2} \end{bmatrix}$$

If we assume an unbiased estimator, then it is easy to show that

$$\beta^\top(\theta) = \left[ \frac{\partial}{\partial a} \mathbf{E}_\theta [\hat{\theta}(Y)], \frac{\partial}{\partial \phi} \mathbf{E}_\theta [\hat{\theta}(Y)] \right] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and the information inequality (the CRLB in this case) is simply

$$\text{cov}_\theta [\hat{\theta}(Y)] \geq \frac{2\sigma^2}{n} \begin{bmatrix} 1 & 0 \\ 0 & a^{-2} \end{bmatrix}$$

The diagonal elements of  $\text{cov}_\theta [\hat{\theta}(Y)]$  reveal the minimum variance for each parameter:  $\text{var}_a [\hat{a}(Y)] \geq \frac{2\sigma^2}{n}$  and  $\text{var}_\phi [\hat{\phi}(Y)] \geq \frac{2\sigma^2}{na^2}$ .

# Example: Unknown Amplitude, Phase, and Frequency

Let  $\theta = [a, \phi, \omega]^\top$ . We can compute

$$I_{02}(\theta) = \frac{1}{\sigma^2} \sum_{k=0}^{n-1} \left( \frac{\partial}{\partial a} s_k(\theta) \right) \left( \frac{\partial}{\partial \omega} s_k(\theta) \right) = -\frac{1}{\sigma^2} \sum_{k=0}^{n-1} ak \cos(\omega k + \phi) \sin(\omega k + \phi) \approx 0$$

$$I_{12}(\theta) = \frac{1}{\sigma^2} \sum_{k=0}^{n-1} \left( \frac{\partial}{\partial \phi} s_k(\theta) \right) \left( \frac{\partial}{\partial \omega} s_k(\theta) \right) = \frac{1}{\sigma^2} \sum_{k=0}^{n-1} a^2 k \sin^2(\omega k + \phi) \approx \frac{a^2}{2\sigma^2} \cdot \frac{(n-1)n}{2}$$

$$I_{22}(\theta) = \frac{1}{\sigma^2} \sum_{k=0}^{n-1} \left( \frac{\partial}{\partial \omega} s_k(\theta) \right)^2 = \frac{1}{\sigma^2} \sum_{k=0}^{n-1} a^2 k^2 \sin^2(\omega k + \phi) \approx \frac{a^2}{2\sigma^2} \cdot \frac{n(n-1)(2n-1)}{6}$$

where we have used the identities  $\sum_{k=0}^{n-1} k = \frac{n(n-1)}{2}$  and

$\sum_{k=0}^{n-1} k^2 = \frac{n(n-1)(2n-1)}{6}$ . The Fisher information matrix is then

$$I(\theta) = \frac{n}{2\sigma^2} \begin{bmatrix} 1 & 0 & 0 \\ 0 & a^2 & \frac{a^2(n-1)}{2} \\ 0 & \frac{a^2(n-1)}{2} & \frac{a^2(n-1)(2n-1)}{6} \end{bmatrix}$$

Note coupling between the unknown phase  $\phi$  and unknown frequency  $\omega$ .

## Example: Unknown Amplitude, Phase, and Frequency

If we consider unbiased estimators, then the information inequality (CRLB) will be

$$\text{cov}_\theta \left[ \hat{\theta}(Y) \right] \geq I^{-1}(\theta).$$

Although it is possible to symbolically invert  $I(\theta)$  in this case, let's look at a numerical example:  $n = 20$  and  $a^2 = \sigma^2 = 1$ . When we had only two unknown parameters  $\theta = [a, \phi]$ , the CRLB was

$$\text{cov}_\theta \left[ \hat{\theta}(Y) \right] \geq I^{-1}(\theta) = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$$

When we have three unknown parameters  $\theta = [a, \phi, \omega]$ , the CRLB can be computed as

$$\text{cov}_\theta \left[ \hat{\theta}(Y) \right] \geq I^{-1}(\theta) = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.371 & -0.029 \\ 0 & -0.029 & 0.003 \end{bmatrix}$$

Note the increase in  $\text{var}_\theta[\phi]$  as a consequence of the unknown frequency.

# Multiparam. Information Inequality: Nuisance Parameters

Lemma (Lehmann TPE 1998 pp. 127-128)

$$I_{ii}^{-1}(\theta) \leq [I^{-1}(\theta)]_{ii}$$

*with equality if and only if  $I_{ij}(\theta) = 0$  for all  $j \neq i$ .*

- ▶ As the example demonstrates and the Lemma confirms, the presence of additional unknown parameters never makes the problem of estimating a particular parameter easier. In most cases, additional unknown parameters make the estimation of a particular parameter more difficult.
- ▶ When these additional unknown parameters exist only to make the estimation of the desired parameters more difficult, they are called **nuisance parameters**.

# Conclusions

- ▶ **Information bound**: a very general lower bound on the variance of an estimator.
- ▶ The information bound applies to **biased or unbiased** estimators of a real-valued non-random scalar or vector parameter.
- ▶ Useful for finding MVU by “guessing and checking” as well as determining how well your estimator is working.
- ▶ The **Cramer-Rao lower bound** (CRLB) is a special case of the general bound and applies only to **unbiased** estimators.
- ▶ An unbiased estimator achieving the CRLB is efficient and MVU. The converse is not always true.
- ▶ Other bounds not covered here:
  - ▶ Chapman-Robbins inequality (finite differences)
  - ▶ Bhayyachayya inequality (higher order derivatives)
- ▶ Lots of extensions were not covered here. For example, the information inequality for functions of parameters, i.e.  $g(\theta)$ , or complex parameters/observations.