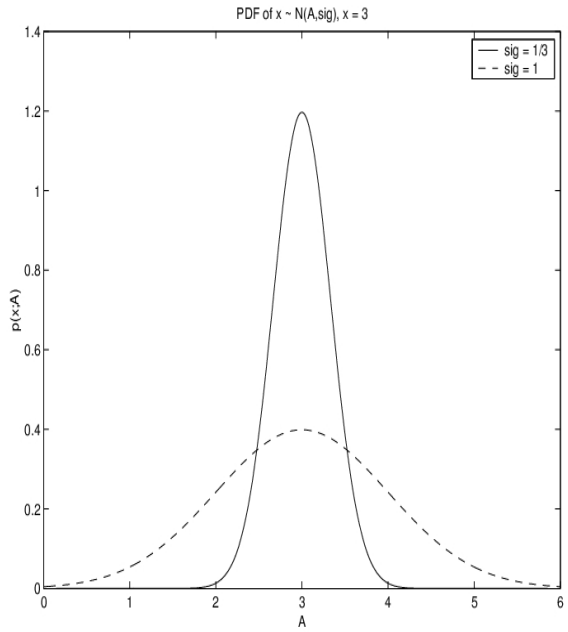
Cramér-Rao Bound (CRB) and Minimum Variance Unbiased (MVU) Estimation

Reading

- Kay-I, Ch. 3.

How accurately we can estimate a parameter θ depends on the pdf or pmf of the observation(s) x (i.e. on the likelihood function).

Example (Kay-I, Chapter 3): $x[0] = A + w[0]$, A unknown, $w[0] \in \mathcal{N}(0, \sigma^2)$.



Intuitively, sharpness of the pdf/pmf determines how accurately we can estimate A.

Cramér-Rao Bound - Regularity Assumptions

We make two regularity assumptions on $p(x;\theta)$:

(i) The set $A = \{x \mid p(x;\theta) > 0\}$ does not depend on θ. For all $x \in A$, $\theta \in \Theta$, $\partial/\partial\theta \log p(x;\theta)$ exists and is finite.
(Here, $\Theta \equiv$ the parameter space.)

(ii) If T is any statistic such that $\mathrm{E}_X(|T|) < \infty$ for all $\theta \in \Theta$, then integration and differentiation by θ can be interchanged in $\int T(x)\, p(x;\theta)\, dx$, i.e.

$$\frac{\partial}{\partial\theta}\left[\int T(x)\, p(x;\theta) dx\right] = \int T(x)\, \frac{\partial}{\partial\theta} p(x;\theta)\, dx \qquad (1)$$

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

Note: Checking assumption (ii) is not very practical. We need simple sufficient conditions on $p(x;\theta)$ 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 θ.

Notation: In this section, we adopt the (common) notation

$$\mathrm{E}_X[T(X)] = \int T(x)\; p(x;\theta)\, dx.$$

Generally, in these notes we will use either this notation or

$$\begin{aligned} \mathrm{E}_p[T(X)] &= \int T(x)\ p(x;\theta)\,dx \\ \mathrm{E}_{p(x;\theta)}[T(X)] &= \int T(x)\ p(x;\theta)\,dx. \end{aligned}$$

Observe that B & D adopt the (also common, but in statistics) notation $\mathrm{E}_\theta[T]$ for the above expectation, with goal to emphasize the dependence on the parameter θ — do not get confused by this difference. In the "classical" discussion here, only X is a random quantity, so we could as well drop the subscript and use $\mathrm{E}\,[T(X)]$.

Proposition. *If*

$$p(x;\theta) = h(x)\exp\{\eta(\theta)T(x) - A(\theta)\} \qquad \textit{(exponential family)}$$

and $\eta(\theta)$ has a nonvanishing continuous derivative on Θ, then (i) *and* (ii) *hold.*

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

$$\begin{aligned} \mathcal{I}(\theta) &= \mathrm{E}_X\left\{\left(\frac{\partial}{\partial\theta}\log p(X;\theta)\right)^2\right\} \\ &= \int\left(\frac{\partial}{\partial\theta}\log p(x;\theta)\right)^2 p(x;\theta)\,dx. \end{aligned}$$

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

Terminology:

$$\frac{\partial}{\partial \theta} \log p(x;\theta)$$

is known as the *score function* for θ.

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

$$\mathrm{E}_X \left| \frac{\partial}{\partial \theta} \log p(X;\theta) \right| < \infty.$$

Then

$$\mathrm{E}_X \Big(\frac{\partial}{\partial \theta} \log p(X;\theta) \Big) = 0 \text{ and, thus, } \mathcal{I}(\theta) = \mathrm{var}_X \left(\frac{\partial}{\partial \theta} \log p(X;\theta) \right).$$

Proof.

$$\mathrm{E}_X \left(\frac{\partial}{\partial \theta} \log p(X;\theta) \right) = \int \left\{ \left[\frac{\partial}{\partial \theta} p(x;\theta) \right] \Big/ p(x;\theta) \right\} p(x;\theta)\, dx$$

$$= \int \frac{\partial}{\partial \theta} p(x;\theta)\, dx = \frac{\partial}{\partial \theta} \int p(x;\theta)\, dx = 0.$$

Here, we have utilized the chain rule of differentiation:

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

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

Comments:

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

- The score function is equal to zero at the ML estimator of θ.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. measurements from a $\text{Poisson}(\lambda)$ distribution [$p(x_i; \lambda) = \lambda^{x_i}/(x_i!) \cdot \exp(-\lambda)$, see your distribution table]. Then

$$\begin{aligned} \frac{\partial}{\partial \lambda} \log p(\boldsymbol{x}; \lambda) &= \frac{\sum_{i=1}^n x_i}{\lambda} - n \\ \mathcal{I}(\lambda) &= \text{var}\left(\frac{\sum_{i=1}^n X_i}{\lambda} \right) = \frac{1}{\lambda^2} \cdot n\lambda = \frac{n}{\lambda}. \end{aligned}$$

Theorem 1. **(Information Inequality)** *Let $T(X)$ be any statistic such that $\text{var}_{p(x;\theta)}[T(X)] < \infty$ for all θ. Denote $\text{E}_{p(x;\theta)}[T(X)]$ by $\psi(\theta)$. Suppose that* (i) *and* (ii) *hold and $0 < \mathcal{I}(\theta) < \infty$. Then, for all θ*

$$\text{var}_{p(x;\theta)}[T(X)] \geq \frac{|\psi'(\theta)|^2}{\mathcal{I}(\theta)}. \tag{2}$$

where

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

Proof. Using (i) and (ii), we obtain

$$\psi'(\theta) = \int T(x) \frac{\partial p(x;\theta)}{\partial \theta} dx = \int T(x) \frac{\partial \log p(x;\theta)}{\partial \theta} p(x;\theta)\, dx.$$

Now

$$\psi'(\theta) = \operatorname{cov}\left[\frac{\partial \log p(X;\theta)}{\partial \theta}, T(X)\right].$$

[Recall: $\operatorname{cov}(P,Q) \stackrel{\triangle}{=} \mathrm{E}\left[(P - \mathrm{E}[P])(Q - \mathrm{E}[Q])\right]$.] Apply the Cauchy-Schwartz inequality

$$[\operatorname{cov}(P,Q)]^2 \leq \operatorname{var}(P) \cdot \operatorname{var}(Q)$$

to the random variables $\overbrace{\partial \log p(X;\theta)/\partial\theta}^{P}$ and $\overbrace{T(X)}^{Q}$:

$$|\psi'(\theta)|^2 \leq \operatorname{var}[T(X)] \cdot \operatorname{var}\left[\frac{\partial \log p(X;\theta)}{\partial \theta}\right].$$

The theorem follows because, by Lemma 1, $\operatorname{var}\left(\frac{\partial}{\partial\theta} \log p(X;\theta)\right) = \mathcal{I}(\theta)$. $\square$

Digression

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

$$\underbrace{\Big|}_{\text{determinant}} \quad \operatorname{cov}\Big(\quad \underbrace{\begin{bmatrix} P \\ Q \end{bmatrix}}_{\text{covariance matrix}} \quad \Big) \quad \underbrace{\Big|}_{\text{determinant}}$$

$$= \begin{vmatrix} \operatorname{var}(P) & \operatorname{cov}(P,Q) \\ \operatorname{cov}(P,Q) & \operatorname{var}(Q) \end{vmatrix} \geq 0$$

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, for arbitrary a and b, the following holds:

$$\operatorname{var}[a\,P + b\,Q] \geq 0$$

which can be rewritten as

$$[a, b]\, \operatorname{cov}\Big(\begin{bmatrix} P \\ Q \end{bmatrix} \Big) \begin{bmatrix} a \\ b \end{bmatrix} \geq 0, \quad \forall a, b$$

which, by definition of positive (semi)definiteness, implies that $\operatorname{cov}\Big(\begin{bmatrix} P \\ Q \end{bmatrix} \Big)$ is a positive semidefinite matrix.

(Back to the Main Track) Comments:

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

$$\mathrm{E}_{p(x;\theta)}[T(X)] = \psi(\theta) = \theta + \underbrace{b(\theta)}_{\text{bias}}$$

  and (2) can be viewed as a lower bound on the variance of $T(X)$:

$$\mathrm{var}_{p(x;\theta)}[T(X)] \geq \frac{|1 + b'(\theta)|^2}{\mathcal{I}(\theta)} \tag{3}$$

  (This result may not be very useful since it is hard to analytically compute bias in practice.) In this case, we can bound the MSE of $T(X)$ as follows:

$$\mathrm{MSE}[T(X)] = \mathrm{var}_{p(x;\theta)}[T(X)] + [b(\theta)]^2 \geq \frac{|1 + b'(\theta)|^2}{\mathcal{I}(\theta)} + [b(\theta)]^2. \tag{4}$$

- Since $\mathrm{E}_{p(x;\theta)}[T(X)] = \psi(\theta)$, we can view $T(X)$ as an unbiased estimator of $\psi = \psi(\theta)$; then (2) gives a lower bound on variance of $T(X)$, expressed in terms of the Fisher information $\mathcal{I}(\theta)$ for θ.

The lower bound depends on $T(X)$ through $\psi(\theta)$. If we consider *all* unbiased estimators $T(X)$ of $\psi(\theta) = \theta$, we obtain a *universal lower bound* given by the following.

Corollary 1. Suppose the conditions of the above theorem hold and $T(X)$ is an unbiased estimator of $\psi(\theta) = \theta$. Then

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

The function $1/\mathcal{I}(\theta)$ is often referred to as the *Cramér-Rao bound (CRB)* on the variance of an unbiased estimator of θ.

Proposition. *Suppose* $p(x;\theta)$ *satisfies, in addition to* (i) *and* (ii)*, the following condition:*
$p(x;\theta)$ *is twice differentiable and interchange between integration and differentiation is permitted.*
Then

$$\mathcal{I}(\theta) = -\mathrm{E}_{\,p(x;\theta)} \left\{ \frac{\partial^2}{\partial\theta^2} \log p(X;\theta) \right\}.$$

Proof.

$$\frac{\partial^2}{\partial\theta^2} \log p(x;\theta) \quad = \quad \frac{1}{p(x;\theta)} \cdot \frac{\partial^2}{\partial\theta^2} p(x;\theta) - \left(\frac{\partial}{\partial\theta} \log p(x;\theta) \right)^2$$

and apply expectation with respect to X to both sides [i.e. multiply by $p(x;\theta)$ and integrate]. □

The above results provides another way of computing the Fisher information number which may be more convenient that taking the expectation of the score squared.

Example. Back to the Poisson example:

$$\mathcal{I}(\lambda) = \mathrm{E}_X\left\{-\frac{\partial^2}{\partial\lambda^2}\log p(\boldsymbol{X};\lambda)\right\} = \frac{1}{\lambda^2}\mathrm{E}_X\Big(\sum_{i=1}^{n} X_i\Big) = \frac{n}{\lambda}$$

which is easier than the derivation on p. 5. In this case, $\overline{X} = (1/n)\cdot\sum_{i=1}^{n} X_i$ is the ML estimator of λ and it is unbiased. Since, in the Poisson case, $\mathrm{var}(X_i) = \lambda$, we have:

$$\mathrm{var}(\overline{X}) = \frac{\lambda}{n} = \mathrm{CRB}(\lambda) = \frac{1}{\mathcal{I}(\lambda)}$$

and, by Corollary 1, $\overline{X}$ is a minimum variance unbiased (MVU) estimator of λ.

Example: Let us continue with the same Poisson example but consider unbiased estimators $T(\boldsymbol{X})$ of $\psi = \lambda^2$. Here, $\psi(\lambda) = \lambda^2 \Longrightarrow \psi'(\lambda) = 2\lambda$ and

$$\mathrm{var}_{p(x;\lambda)}[T(\boldsymbol{X})] \geq \frac{|\psi'(\lambda)|^2}{\mathcal{I}(\lambda)} = \frac{4\,\lambda^2}{n/\lambda} = \frac{4\,\lambda^3}{n}.$$

Corollary 2. Suppose that the elements of $\boldsymbol{X} = [X_1, \ldots, X_n]^T$ are i.i.d. with density $p(x;\theta)$ and that the conditions (i) and (ii) hold. Define the "contribution" of a single measurement to the Fisher information:

$$\mathcal{I}_1(\theta) = \mathrm{E}\left\{\Big[\frac{\partial}{\partial\theta}\log p(X_1;\theta)\Big]^2\right\}.$$

(Here, we arbitrarily pick X_1 as our single measurement, its contribution to the Fisher information is equal to that of X_2 etc.) Then

$$\mathcal{I}(\theta) = n\,\mathcal{I}_1(\theta) \quad \text{and} \quad \mathrm{var}_{p(x;\theta)}[T(\boldsymbol{X})] \geq \frac{|\psi'(\theta)|^2}{n\,\mathcal{I}_1(\theta)}.$$

Proof.

$$\mathcal{I}(\theta) = \mathrm{var}\left(\frac{\partial}{\partial\theta}\log p(\boldsymbol{X};\theta)\right) = \mathrm{var}\left(\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log p(X_i;\theta)\right)$$

$$= \sum_{i=1}^{n}\mathrm{var}\left(\frac{\partial}{\partial\theta}\log p(X_i;\theta)\right) = n\,\mathcal{I}_1(\theta).$$

□

Example: Suppose that $X_1, \ldots, X_n$ are i.i.d. observations from a normal distribution with unknown mean θ and known variance σ^2. Note that the conditions (i) and (ii) hold. Then

$$\mathcal{I}_1(\theta) = \mathrm{E}\left\{\left(\frac{\partial}{\partial\theta}\log\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\cdot\exp\left[-\frac{(X_1-\theta)^2}{2\sigma^2}\right]\right\}\right)^2\right\}$$

$$= \mathrm{E}\left[\left(\frac{X_1-\theta}{\sigma^2}\right)^2\right] = \frac{1}{\sigma^2}$$

and, by Corollary 2,

$$\mathcal{I}(\theta) = n\,\mathcal{I}_1(\theta) = \frac{n}{\sigma^2}.$$

Observe that

$$\mathrm{var}\{\overline{X}\} = \frac{\sigma^2}{n} = \mathrm{CRB}(\theta) = \frac{1}{\mathcal{I}(\theta)}$$

and, therefore, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an MVU estimator of θ. Since $\overline{X}$ does not depend on σ^2, it is MVU for any σ^2. Hence, $\overline{X}$ is an MVU estimator of θ even if σ^2 is unknown.

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

Note: Efficient $\Rightarrow$ MVU. However, MVU $\not\Rightarrow$ efficient, because CRB is not always attainable by MVU estimators (at least not for finite samples, i.e. finite n).

Under certain regularity conditions, ML estimators attain the CRB asymptotically (i.e. for large n); hence they are *asymptotically efficient*, which is one of the main reasons for their popularity.

Proof of "Efficiency $\Rightarrow$ 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 equals the CRB for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, then it must be MVU.

In the following theorem, we give necessary and sufficient conditions for the CRB to be attainable. Previous examples in which MVU estimator attains the CRB were situations where $\boldsymbol{X}$ follows a one-parameter exponential family. This is not an accident.

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* θ*. Then* $p(x;\theta)$ *is a one-parameter exponential family with pdf/pmf of the form*

$$p(\boldsymbol{x};\theta) = h(\boldsymbol{x}) \exp[\eta(\theta)\, T(\boldsymbol{x}) - B(\theta)]. \tag{5}$$

Conversely, if $p(x;\theta)$ *is a one-parameter exponential family of the above form and* $\eta(\theta)$ *has a continuous nonvanishing derivative on* Θ*, then* $T(X)$ *achieves the CRB and is the MVU estimator of* $\mathrm{E}_{\boldsymbol{X}}[T(\boldsymbol{X})]$*. Hence,* $T(\boldsymbol{X})$ *is an efficient estimator of* $\mathrm{E}_{\boldsymbol{X}}[T(\boldsymbol{X})] = \psi(\theta)$*.*

Proof. See B & D, Theorem 3.4.2. □

Note that the above theorem gives both necessary and sufficient conditions for an efficient estimator.

One-Parameter Canonical Exponential Family. In handout # 1, we introduced the one-parameter canonical exponential family:

$$p(\boldsymbol{x};\eta) = h(\boldsymbol{x})\, \exp\left[\eta\, T(\boldsymbol{x}) - A(\eta)\right]$$

for which we know that

$$\mathrm{E}_{p(x;\eta)}[T(\boldsymbol{X})] = \frac{dA(\eta)}{d\eta}, \quad \mathrm{var}_{p(x;\eta)}[T(\boldsymbol{X})] = \frac{d^2A(\eta)}{d\eta^2}.$$

Therefore, in this case, Theorem 2 states that $T(\boldsymbol{X})$ is an efficient estimator of $\mathrm{E}_X[T(\boldsymbol{X})] = \frac{dA(\eta)}{d\eta}$; we can easily compute the variance of this estimator as well:

$$\mathcal{I}(\eta) = \mathrm{var}_{p(x;\eta)}\Big(\underbrace{T(\boldsymbol{X}) - \frac{dA(\eta)}{d\eta}}_{\text{score function}}\Big) = \mathrm{var}_{p(x;\eta)}\big(T(\boldsymbol{X})\big) = \frac{d^2A(\eta)}{d\eta^2}$$

and this variance, according to Theorem 2, is equal to the CRB of $\frac{dA(\eta)}{d\eta}$:

$$\mathrm{CRB}\Big(\frac{dA(\eta)}{d\eta}\Big) = \Big[\mathcal{I}\Big(\frac{dA(\eta)}{d\eta}\Big)\Big]^{-1} = \mathcal{I}(\eta).$$

$\mathrm{E}_{\boldsymbol{X}}[\boldsymbol{T(X)}] = \boldsymbol{\theta}$**.** Suppose now that we pick $\theta = \theta(\eta) = \frac{dA(\eta)}{d\eta}$; then, clearly, $T(\boldsymbol{X})$ is an efficient estimator of $\mathrm{E}_X[T(\boldsymbol{X})] = \frac{dA(\eta)}{d\eta} = \theta$ and

$$\underbrace{\frac{dA(\eta)}{d\eta}}_{\theta} = \underbrace{\frac{dB(\theta)}{d\theta}\Big|_{\theta=\frac{dA(\eta)}{d\eta}} \cdot \underbrace{\frac{d\theta(\eta)}{d\eta}}_{\mathrm{CRB}(\theta)=[\mathcal{I}(\theta)]^{-1}}}_{\text{chain rule}} \quad \Longrightarrow \quad \frac{dB(\theta)}{d\theta} = \theta \cdot \mathcal{I}(\theta).$$

This case is considered in Kay-I — differentiating the logarithm of (5) with respect to θ applying the above identity and using the fact that

$$\frac{d\eta}{d\theta} = \frac{1}{d\theta/d\eta} = \mathcal{I}(\theta)$$

yields the condition provided by Kay:

$$\frac{\partial \log p(\boldsymbol{x};\theta)}{\partial \theta} = \underbrace{\mathcal{I}(\theta)}_{\substack{\text{Fisher information} \\ \text{for } \theta}} [T(\boldsymbol{x}) - \theta] \qquad (6)$$

see Kay-I, App. 3A for Kay's proof. **To summarize: $T(x)$ is an efficient estimator of θ if and only if the score function corresponding to the underlying probabilistic model can be written in the form (6).**

Recall: We used the Cauchy-Schwartz inequality to derive the information inequality theorem. Proving the above result reduces to considering the case where equality holds in the Cauchy-Schwartz inequality — the score function needs to be an affine function of $T(\boldsymbol{x})$.

Comments:

- Theorem 2 tells us that $T(\boldsymbol{x})$ in (5) is an efficient estimator of its expectation. We could use either (5) or (6) (from Kay-I) to verify if an estimator is efficient.

- If we wish to answer the question if an efficient estimator of a particular parameter θ exists, then we should check (6). Sometimes (5) can be used even if $\mathrm{E}\left[T(\boldsymbol{x})\right] \neq$ *the parameter of interest*; in particular, this is the case when there is an affine relationship between $T(\boldsymbol{x})$ and the parameter of interest.
  Note: if $T(\boldsymbol{x})$ is an efficient estimator of its expectation $\mathrm{E}_{p(x;\theta)}[T(\boldsymbol{X})]$, this *does not* imply that a non-affine function of $T(\boldsymbol{x})$ is an efficient estimator of its expectation. For example, suppose that $T(\boldsymbol{x})$ is an efficient estimator of its expectation; then $1/T(\boldsymbol{x})$, say, will generally not be an efficient estimator of $\mathrm{E}_{p(x;\theta)}[1/T(\boldsymbol{X})]$.

Example: If $X_1, X_2, \ldots, X_n$ are i.i.d. $\mathrm{Poisson}(\lambda)$, then

$$\begin{aligned} p(x_1, \ldots, x_n; \lambda) &= \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \exp(-n\lambda) \\ &= \frac{1}{\prod_{i=1}^n x_i!} \exp\Bigg(\underbrace{\sum_{i=1}^n x_i}_{T(x)} \underbrace{\log \lambda}_{\eta} - \underbrace{n\lambda}_{A(\eta)} \Bigg). \end{aligned}$$

By utilizing the fact that the above pmf belongs to the one-parameter exponential family, we can find the Fisher information number for $\eta = \log \lambda$:

$$A(\eta) = n \, \exp(\eta) \quad \Longrightarrow \quad \mathcal{I}(\eta) = \frac{d^2 A(\eta)}{d\eta^2} = n \, \exp \eta = n \, \lambda$$

which is also equal to the CRB for $\frac{dA(\eta)}{d\eta} = n\exp(\eta) = n\lambda$. Finally, we also know that $T(x) = \sum_{i=1}^{n} x_i$ is an efficient estimator of $\frac{dA(\eta)}{d\eta} = n\lambda$, or

$$\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$

is an efficient estimator of λ and the CRB for λ is λ/n, which is consistent with the CRB result for Poisson mean parameter that we obtained before.

To show efficiency of $\overline{X}$ using (6), we first summarize the results that we have obtained earlier:

$$\mathrm{var}(\overline{X}) = \frac{\lambda}{n}, \quad \mathcal{I}(\lambda) = \frac{n}{\lambda}$$

and

$$\frac{\partial}{\partial\lambda} \log p(\boldsymbol{x};\lambda) = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n.$$

Indeed,

$$\frac{\partial \log p(\boldsymbol{x};\lambda)}{\partial\lambda} = \underbrace{\frac{n}{\lambda}}_{\mathcal{I}(\lambda)} \Big(\underbrace{\overline{x}}_{T(\boldsymbol{x})} - \lambda\Big).$$

Cramér-Rao Bound – Example

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

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

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

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

Since the observations are independent, we get

$$p(\boldsymbol{x};f) = \prod_{n=0}^{N-1} p(x[n];f).$$

Taking the log yields

$$\begin{aligned} \log p(\boldsymbol{x};f) &= \sum_{n=0}^{N-1} \log p(x[n];f) \\ &= -\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - s[n;f])^2 + \underbrace{\text{const}}_{\text{indep. of } f} \end{aligned}$$

Differentiate:

$$\frac{\partial \log p(\boldsymbol{x}; f)}{\partial f} = \frac{1}{\sigma^2} \cdot \sum_{n=0}^{N-1} \frac{\partial s[n; f]}{\partial f} \cdot (x[n] - s[n; f])$$

and once more:

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

The negative expected value of this expression is the Fisher information number

$$\begin{aligned} \mathcal{I}(f) &= -\mathrm{E}_{\,p(x;f)} \Big[\frac{\partial \log p(\boldsymbol{X}; f)}{\partial f} \Big] = \frac{1}{\sigma^2} \cdot \sum_{n=0}^{N-1} \left(\frac{\partial s[n; f]}{\partial f} \right)^2 \\ &= \mathrm{SNR} \cdot \sum_{n=0}^{N-1} \left[2\pi n \cdot \sin(2\pi f n + \phi) \right]^2 \end{aligned}$$

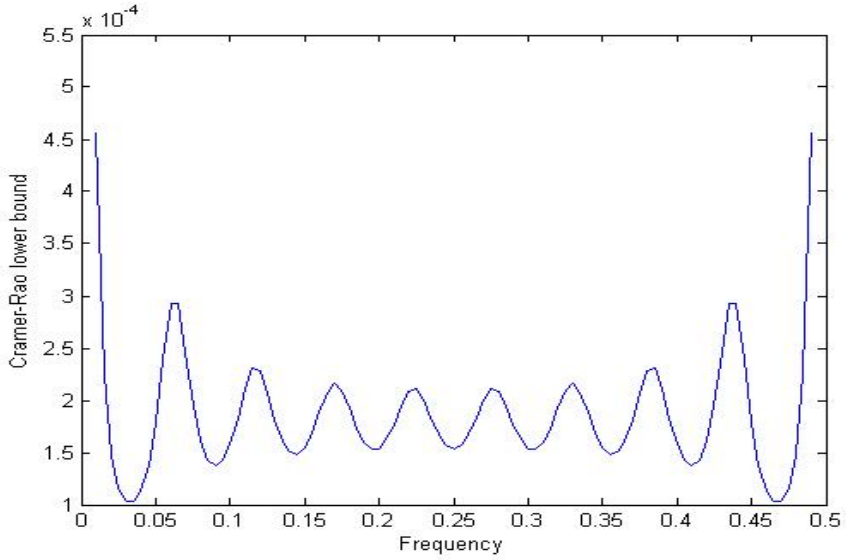
where $\mathrm{SNR} = A^2/\sigma^2$ is the signal-to-noise ratio. The CRB is $1/\mathcal{I}(f) \leq \mathrm{var}(f)$ for $\widehat{f}$ unbiased.

Cramér-Rao bound – Example (cont.)

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

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

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



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

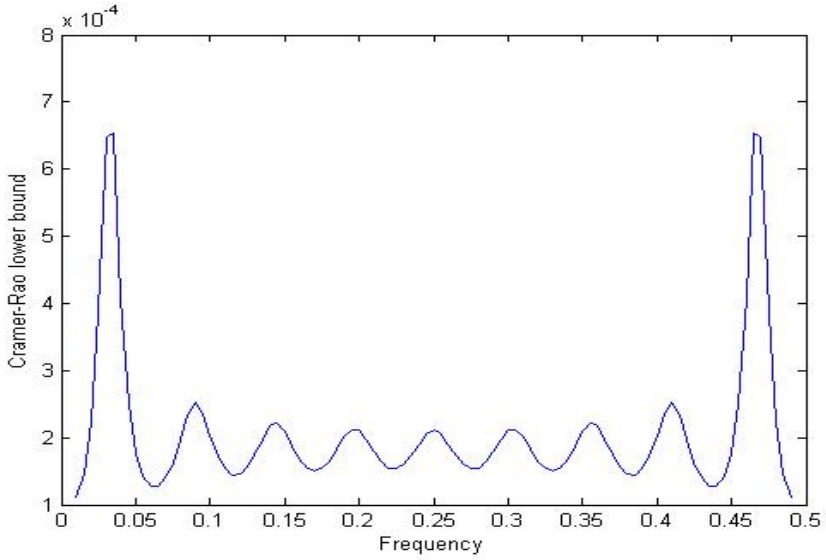
There are preferred frequencies!

As $f \searrow 0$, the CRB goes to infinity because, in this case, a slight change in frequency will not alter the signal significantly because $A\cos(2\pi f n)$ is a flat function of f in the neighborhood of $f = 0$. The CRB goes to infinity as $f \nearrow \frac{1}{2}$ as well.

Consider now the case where $\mathrm{SNR} = 1, N = 10,$ and $\phi =$

$-\pi/2$. Then

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



CRB for f as a function of f, for $\mathrm{SNR} = 1$, $N = 10$, 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, $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_d]^T$. We assume that the parameter space $\boldsymbol{\Theta}$ is an open subset of $\mathbb{R}^d$ and that $p(x; \boldsymbol{\theta})$ satisfies conditions (i) and (ii) when differentiation is with respect to $\theta_i,\ i = 1, 2, \ldots, d$.

The *Fisher information matrix (FIM)* is defined as

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

where $\mathcal{I}_{i,k}(\boldsymbol{\theta}) = \mathrm{E}_{\,p(x;\theta)} \left\{ \frac{\partial}{\partial \theta_i} \log p(X; \boldsymbol{\theta}) \, \frac{\partial}{\partial \theta_k} \log p(X; \boldsymbol{\theta}) \right\}$.

Proposition. *Under the above conditions*

(a)

$$\mathrm{E}_{\,p(\boldsymbol{x};\boldsymbol{\theta})} \left[\frac{\partial}{\partial \theta_i} \log p(X; \boldsymbol{\theta}) \right] = 0, \quad i = 1, 2, \ldots, d$$

$$\mathcal{I}_{i,k} = \mathrm{cov}_{p(\boldsymbol{x};\boldsymbol{\theta})} \left[\frac{\partial}{\partial \theta_i} \log p(X; \boldsymbol{\theta}), \frac{\partial}{\partial \theta_k} \log p(X; \boldsymbol{\theta}) \right]$$

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

$$\mathrm{E}_{\,p(\boldsymbol{x};\boldsymbol{\theta})} \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(X; \boldsymbol{\theta}) \right] = \underbrace{\mathbf{0}}_{d \times 1 \text{ vector of zeros}}$$

and

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{cov}_{p(\boldsymbol{x};\boldsymbol{\theta})} \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{X};\boldsymbol{\theta}) \right].$$

(b) *If $X_1, \ldots, X_n$ are i.i.d., then $\boldsymbol{X} = [X_1, \ldots, X_n]^T$ has Fisher information $n\mathcal{I}_1(\boldsymbol{\theta})$ where $\mathcal{I}_1(\boldsymbol{\theta})$ is the Fisher information due to a single observation X_1, say (or X_2 or . . .).*

(c) *If, in addition, $p(x;\boldsymbol{\theta})$ is differentiable and double integration and differentiation under the integral sign can be interchanged,*

$$[\mathcal{I}(\boldsymbol{\theta})]_{i,k} = -\mathrm{E}_{p(\boldsymbol{x};\boldsymbol{\theta})} \left[\frac{\partial^2}{\partial\theta_i \partial\theta_k} \log p(X;\boldsymbol{\theta}) \right], \quad i,k \in \{1,2,\ldots,d\}.$$

Example. Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ and $\boldsymbol{\theta} = [\mu, \sigma^2]^T$. Then

$$\begin{aligned}
\log p(x;\boldsymbol{\theta}) &= -\tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2 \\
\mathcal{I}_{11}(\boldsymbol{\theta}) &= -\mathrm{E}\left[\frac{\partial^2}{\partial\mu^2} \log[p(X;\boldsymbol{\theta})] \right] = \mathrm{E}\left[\sigma^{-2}\right] = \sigma^{-2} \\
\mathcal{I}_{12}(\boldsymbol{\theta}) &= -\mathrm{E}\left[\frac{\partial}{\partial\sigma^2}\frac{\partial}{\partial\mu} \log[p(X;\boldsymbol{\theta})] \right] \\
&= -\sigma^{-4}\mathrm{E}\left[X-\mu\right] = 0 = \mathcal{I}_{21}(\boldsymbol{\theta}) \\
\mathcal{I}_{22}(\boldsymbol{\theta}) &= -\mathrm{E}\left[\frac{\partial^2}{\partial(\sigma^2)^2} \log[p(X;\boldsymbol{\theta})] \right] = \sigma^{-4}/2.
\end{aligned}$$

Therefore

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & \sigma^{-4}/2 \end{bmatrix}. \tag{7}$$

Multiple I.I.D. Observations: How about n i.i.d. observations $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$ with $\boldsymbol{\theta} = [\mu, \sigma^2]^T$? Then, (7) implies that

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

and, consequently,

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

Decoupling: Note that the FIM in this example is diagonal. Therefore, CRB for μ remains the same whether or not σ^2 is known. Similarly, CRB for σ^2 is the same regardless of whether or not μ is known.

In general, the more parameters[1], 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. 2 hold and suppose that the Fisher information matrix*

[1] We have to compare nested models. Otherwise, we would be comparing apples and oranges.

$\mathcal{I}(\boldsymbol{\theta})$ is positive definite (hence nonsingular). Then, for $\mathrm{E}_{p(x;\theta)}[T(X)] = \psi(\boldsymbol{\theta})$,

$$\mathrm{var}_{p(x;\boldsymbol{\theta})}[T(X)] \geq \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \mathcal{I}(\boldsymbol{\theta})^{-1} \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

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

$$\boldsymbol{\psi}(\boldsymbol{\theta}) = \mathrm{E}_{p(\boldsymbol{x};\boldsymbol{\theta})}[\boldsymbol{T}(X)] = [\psi_1(\boldsymbol{\theta}), \ldots, \psi_d(\boldsymbol{\theta})]^T.$$

Then

$$\mathrm{cov}_{p(x;\boldsymbol{\theta})}[\boldsymbol{T}(X)] \geq \frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \mathcal{I}(\boldsymbol{\theta})^{-1} \frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}}.$$

Comments:

- $\boldsymbol{T}(X)$ in Theorem 3 will typically be an estimator [of $\boldsymbol{\psi}(\boldsymbol{\theta})$, say].

- $$\mathrm{cov}_{p(\boldsymbol{x};\boldsymbol{\theta})}[\boldsymbol{T}(X)] \geq \frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \mathcal{I}(\boldsymbol{\theta})^{-1} \frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}}.$$

  means that

$$\mathrm{cov}_{p(\boldsymbol{x};\boldsymbol{\theta})}[\boldsymbol{T}(X)] - \frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \mathcal{I}(\boldsymbol{\theta})^{-1} \frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} \geq 0$$

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

$$\boldsymbol{q}^T \boldsymbol{A} \boldsymbol{q} \geq 0 \quad \forall \boldsymbol{q}. \tag{8}$$

- If $\boldsymbol{T}(X)$ is an unbiased estimator of $\boldsymbol{\theta}$ [i.e. $\boldsymbol{\psi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$], then

$$\mathrm{cov}_{p(\boldsymbol{x};\boldsymbol{\theta})}[\boldsymbol{T}(X)] \geq \mathcal{I}(\boldsymbol{\theta})^{-1}.$$

- Suppose now that $\psi(\boldsymbol{\theta}) = \theta_i$ corresponding to $T_i(X)$, where $T_i(X)$ is the ith element of $\boldsymbol{T}(X)$ (and $\boldsymbol{T}(X)$ is an unbiased estimator of $\boldsymbol{\theta}$). Now, $\frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = [0, 0, \ldots, 0, \underbrace{1}_{i\text{th place}}, 0, \ldots, 0]^T$ and, consequently,

$$\mathrm{var}_{p(x;\boldsymbol{\theta})}[T_i(X)] \geq \underbrace{[\mathcal{I}(\boldsymbol{\theta})^{-1}]_{i,i}}_{(i,i)\text{th element of the CRB for } \boldsymbol{\theta}}.$$

- **Notation:** If

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

then

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

and

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

Multiparameter Exponential Family and Efficiency

Consider the canonical k-parameter exponential family:

$$p(x;\boldsymbol{\eta}) = \exp\Big[\underbrace{\sum_{i=1}^{d} T_i(x)\,\eta_i}_{\boldsymbol{T}(\boldsymbol{x})^T\boldsymbol{\eta}} - A(\boldsymbol{\eta})\Big] h(x)$$

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

$$\frac{\partial \log p(x;\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \boldsymbol{T}(x) - \frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}. \tag{9}$$

Hence, the Fisher information matrix is

$$\mathcal{I}(\boldsymbol{\eta}) = \mathrm{cov}_{p(x;\boldsymbol{\eta})}[\boldsymbol{T}(X)] = \underbrace{\frac{\partial^2 A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\,\partial \boldsymbol{\eta}^T}}_{d \times d \text{ matrix}}. \tag{10}$$

Theorem 4. *Each $T_i(X)$ is an MVU estimator of $\mathrm{E}_{p(x;\boldsymbol{\theta})}[T_i(X)]$.*

Comments:

- The claim in Theorem 4 is a different from stating that $T_i(X)$ is MVU for $\mathrm{E}_{p(\boldsymbol{x};\boldsymbol{\eta})}[T_i(X)] = \frac{\partial A(\boldsymbol{\eta})}{\partial \eta_i}$ if $\eta_k, k \neq i$ are known (which has already been stated in Theorem 2). How do we show this new claim?

  Proof. (of Theorem 4). Without loss of generality, let us focus on $i = 1$. Note that (10) and (9) imply:

$$\begin{aligned} \mathrm{var}_{p(x;\boldsymbol{\theta})}[T_1(X)] &= \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \eta_1^2} \\ \mathrm{E}_{p(\boldsymbol{x};\boldsymbol{\eta})}[T_1(X)] &= \psi(\boldsymbol{\eta}) = \frac{\partial A(\boldsymbol{\eta})}{\partial \eta_1}. \end{aligned}$$

  Therefore

$$\frac{\partial \psi(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T} = \frac{\partial A(\boldsymbol{\eta})}{\partial \eta_i \, \partial \boldsymbol{\eta}^T}$$

  is the first row of $\mathcal{I}(\boldsymbol{\eta}) = \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \, \partial \boldsymbol{\eta}^T}$, implying that

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

  and, finally,

$$\frac{\partial \psi(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T} \mathcal{I}(\boldsymbol{\eta})^{-1} \frac{\partial \psi(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \frac{\partial \psi(\boldsymbol{\eta})}{\partial \eta_1} = \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \eta_1^2} = \mathrm{var}_{p(x;\boldsymbol{\theta})}[T_1(X)]$$

i.e. we have achieved equality in Theorem 3 $\Longrightarrow$ $T_1(X)$ is MVU for $\mathrm{E}_{p(x;\boldsymbol{\theta})}[T_1(X)]$. $\square$

- Similar to the scalar case, the parametrization $\boldsymbol{\theta} = \frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}$ is considered in Kay-I, where it is shown that an unbiased estimator of $\boldsymbol{\theta}$ attains the CRB if and only if

$$\frac{\partial \log p(x;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \underbrace{\mathcal{I}(\boldsymbol{\theta})}_{\text{FIM for } \boldsymbol{\theta}} [\boldsymbol{T}(x) - \boldsymbol{\theta}] \qquad (11)$$

which generalizes (6). **To summarize: $\boldsymbol{T}(x)$ is an efficient estimator of $\boldsymbol{\theta}$ if and only if the score function corresponding to the underlying probabilistic model can be written in the form (11).**

Example: If $X_1, X_2, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, then $\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$ is the MVU estimator of μ and $\frac{1}{n} \cdot \sum_{i=1}^{n} X_i^2$ is the MVU estimator of $\mu^2 + \sigma^2$, which follows from

$$p(x_1, \ldots, x_n; \boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{n\mu^2}{2\sigma^2} \right)$$

$$\cdot \exp\left[-\frac{1}{2\sigma^2} \Big(\underbrace{n \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2}_{T_2(\boldsymbol{x})} - 2\mu n \cdot \underbrace{\overline{x}}_{T_1(\boldsymbol{x})} \Big) \right].$$

But, it *does not* follow that $\frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2$ is the MVU estimator of σ^2.

It seems that it would be quite difficult to use (11) to show the efficiency of

$$\boldsymbol{T}(\boldsymbol{X}) = \begin{bmatrix} \overline{X} \\ \frac{1}{n} \cdot \sum_{i=1}^{n} X_i^2 \end{bmatrix}$$

for estimating

$$\mathrm{E}_{p(\boldsymbol{x};\boldsymbol{\theta})}[\boldsymbol{T}(\boldsymbol{X})] = \boldsymbol{\theta} = \begin{bmatrix} \mu \\ \mu^2 + \sigma^2 \end{bmatrix}.$$

Therefore, keep Theorem 4 in mind, in addition to (11).

Gaussian CRB

Theorem 5. *Suppose that* $\boldsymbol{X}$ *has a* n*-variate Gaussian distribution,*

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

that is

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{C}|}} \exp\left[-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right].$$

Then, the (i,k)*th element of the FIM is given by*

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

This is a convenient general formula for analysis.

Proof. See Kay-I, App. 3C. □

Example: If $x[n] = s[n;\theta] + w[n]$ and $w[n]$ is AWGN with known variance σ^2 and $n = 1, 2, \ldots, N$. Then, we can write this model specification in a vector form:

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

$\boldsymbol{C}$ does not depend on θ (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: $x[n] = w[n]$, where $w[n]$, $n = 1, 2, \ldots, N$ is AWGN with variance σ^2. If $\theta = \sigma^2$, then

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

$\boldsymbol{C}(\theta) = \theta \, \boldsymbol{I} = \sigma^2 \, \boldsymbol{I}$.

$$\mathcal{I}(\sigma^2) = \frac{1}{2} \operatorname{tr} \left(\boldsymbol{C}^{-1} \frac{\partial \boldsymbol{C}}{\partial \sigma^2} \boldsymbol{C}^{-1} \frac{\partial \boldsymbol{C}}{\partial \sigma^2} \right) = \frac{1}{2 \, \sigma^4} \operatorname{tr}(\boldsymbol{I}) = \frac{N}{2 \, \sigma^4} \quad (12)$$

and, therefore,

$$\mathrm{CRB}(\sigma^2) = [\mathcal{I}(\sigma^2)]^{-1} = \frac{2 \, \sigma^4}{N}. \quad (13)$$

Say we wish to compute the CRB for σ:

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

which can also be computed using (2): $\psi(\sigma^2) = (\sigma^2)^{1/2}$, $\psi'(\sigma^2) = 1/2 \cdot (\sigma^2)^{-1/2}$, 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 a family of (generally biased) estimators of σ^2:

$$\widehat{\sigma}^2 = a \cdot \frac{1}{N}\sum_{n=0}^{N-1} x^2[n]$$

and found that $a_{\mathrm{OPT}} = \frac{N}{N+2}$ yields an estimator

$$\widehat{\sigma}^2_{\star} = a_{\mathrm{OPT}} \cdot \frac{1}{N}\sum_{n=0}^{N-1} x^2[n] = \frac{1}{N+2}\sum_{n=0}^{N-1} x^2[n]$$

whose MSE is the smallest within the family:

$$\mathrm{MSE}_{\mathrm{MIN}} = \frac{2\,\sigma^4}{N+2} < \frac{2\,\sigma^4}{N} = \mathrm{CRB}(\sigma^2)$$

see (13). (Note that $\widehat{\sigma}^2_\star$ is a *biased* estimator of σ^2 whereas the CRB is the lower bound on variance of *unbiased* estimators only.) Let us now apply the information inequality in (2) to the estimators in this family:

$$\mathrm{var}(\widehat{\sigma}^2) \geq \frac{|\psi'(\sigma^2)|^2}{\mathcal{I}(\sigma^2)} = \frac{2a^2\,\sigma^4}{N}$$

since $\psi(\sigma^2) = \mathrm{E}\,[\widehat{\sigma}^2] = a\sigma^2$ and $\psi'(\sigma^2) = a$. Now, (4) implies that

$$\mathrm{MSE}[\widehat{\sigma}^2] = \mathrm{var}[\widehat{\sigma}^2] + [b(\sigma^2)]^2 \geq \frac{2a^2\,\sigma^4}{N} + (a-1)^2\,\sigma^4$$

since $b(\sigma^2) = \psi(\sigma^2) - \sigma^2 = (a-1)\,\sigma^2$. Interestingly, the above inequality becomes equality for optimal $a = a_{\mathrm{OPT}} = \frac{N}{N+2}$. This MSE bound is not always attainable — we just happen to be lucky in this case.

Asymptotic CRB for WSS Processes

For the definition and properties of wide-sense stationary (WSS) signals, see handout # 10 for EE 420x (and pay attention to that handout's exposition of discrete-time processes, which we need here.)

The results presented here are based on the Whittle approximation, see e.g.

P. Whittle, "The analysis of multiple stationary time series," *J. R. Stat. Soc., Ser. B* vol. 15, pp. 125–139, 1953.

Theorem 6. *Assume wide-sense stationary Gaussian $x[n]$ is observed. The power spectral density (PSD) $P_{xx}(f;\theta)$ depends on unknown parameter vector $\boldsymbol{\theta}$. Then, for large N, the FIM is given by*

$$[\mathcal{I}]_{i,k} = \frac{N}{2} \int_{-1/2}^{1/2} \frac{\partial \log P_{xx}(f;\boldsymbol{\theta})}{\partial \theta_i} \cdot \frac{\partial \log P_{xx}(f;\boldsymbol{\theta})}{\partial \theta_k} \, df. \quad (14)$$

In practice, the above expression is valid if $N \gg$ the correlation time. (Recall: the correlation time of a random process is the time lag after which we can consider correlation between measurements negligible.) For stationary processes, it is often convenient to parametrize the power spectral density. Computing the exact CRB would require computing (and

differentiating) the autocorrelation matrix $\boldsymbol{C}(\boldsymbol{\theta})$ which may not yield simple closed-form solutions.

The discrete-frequency version of the above expression is also useful:

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{k=0}^{N-1} \frac{\partial \log[P_{xx}(f_k;\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \frac{\partial \log[P_{xx}(f_k;\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}^T} \qquad (15)$$

where

$$f_k = k/N, \quad k = 0, 1, \ldots, N-1.$$

For example, (15) may exist even when (14) does not. An example of such a case is the Doppler PSD (which goes to infinity, causing an integrability problem), see

A. Dogandžić and B. Zhang, "Estimating Jakes' Doppler power spectrum parameters using the Whittle approximation," *IEEE Trans. Signal Processing*, vol. 53, pp. 987–1005, Mar. 2005.

For scalar parameters

$$\mathcal{I}(\theta) = \frac{N}{2} \int_{-1/2}^{1/2} \left(\frac{\partial \log P_{xx}(f;\theta)}{\partial \theta} \right)^2 df.$$

Example 3.12 in Kay-I: Estimate the center frequency of a narrowband process:

$$P_{xx}(f; f_c) = Q(f - f_c) + Q(-f - f_c) + \sigma^2$$

where Q is a known function and σ^2 is AWGN variance. $f_{\rm c}$ takes values so that $Q(f - f_{\rm c})$ is always within $[0, 1/2]$. In our case, the asymptotic CRB expression simplifies to

$$\mathrm{CRB}(f_c) = \frac{1}{N \int_{-1/2}^{1/2} \left(\frac{\partial \log[Q(f)+\sigma^2]}{\partial f} \right)^2 df}.$$

For $Q(f) = \exp[-\frac{1}{2}(f/\sigma_f^2)]$, $\sigma_f \ll \frac{1}{2}$ (i.e. narrowband random process), and $Q(f) \gg \sigma^2$ (high SNR), we have

$$\mathrm{var}(\widehat{f_{\rm c}}) \geq \frac{12\sigma_f^4}{N}.$$

The narrower the PSDs (smaller σ_f^2), the better the accuracy.

Example: Range estimation, Kay-I, Example 3.13.

Digression: Complex Signals

Narrowband signal:

$$s(t) = A(t)\cos(\omega_0 t + \phi(t))$$

where $A(t)$ and $\phi(t)$ vary slowly compared to $\cos(\omega_0 t)$.

Complex representation (analytic signal)

$$\widetilde{s}(t) = A(t)e^{j\phi(t)}$$

can be generated using quadrature sampling.

In general: real band-pass signal can be represented as a *complex* low-pass signal (and sampled at a slower rate).

There is a need to process complex data, particularly in

- communications,
- radar/sonar,
- eddy-current nondestructive evaluation (NDE) etc.

A complex scalar is a 2-vector! We could work with real signals of twice the dimension, but it is customary to use the complex representation.

Complex Gaussian Distribution

Consider joint pdf of real and imaginary part of a complex vector $\boldsymbol{x}$:

$$\boldsymbol{x} = \boldsymbol{u} + j\boldsymbol{v}.$$

Assume $\boldsymbol{z} = [\boldsymbol{u}^T, \boldsymbol{v}^T]^T$. The $2n$-variate Gaussian pdf of the (real!) vector $\boldsymbol{z}$ is

$$p_{\boldsymbol{z}}(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^{2n}|\boldsymbol{C}|}} \exp\left[-\tfrac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_z)^T \boldsymbol{C}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_z)\right]$$

where

$$\boldsymbol{\mu}_z = \begin{bmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{uu} & \boldsymbol{C}_{uv} \\ \boldsymbol{C}_{vu} & \boldsymbol{C}_{vv} \end{bmatrix}.$$

That is

$$P\{\boldsymbol{z} \in A\} = \int_{\boldsymbol{z}\in A} p_z(\boldsymbol{z})\, d\boldsymbol{z}.$$

Complex Gaussian Distribution (cont.)

Suppose that $\boldsymbol{C}$ has a special structure:

$$\boldsymbol{C}_{uu} = \boldsymbol{C}_{vv} \quad \text{and} \quad \boldsymbol{C}_{uv} = -\boldsymbol{C}_{vu}.$$

(Note that $\boldsymbol{C}_{uv} = \boldsymbol{C}_{vu}^T$ by construction.) Then, we can define a complex Gaussian pdf:

$$p_X(\boldsymbol{x}) = \frac{1}{\pi^n |\boldsymbol{C}_x|} \exp\left[-(\boldsymbol{x}-\boldsymbol{\mu}_x)^H \boldsymbol{C}_x^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_x)\right]$$

where "H" denotes a complex conjugate transpose and

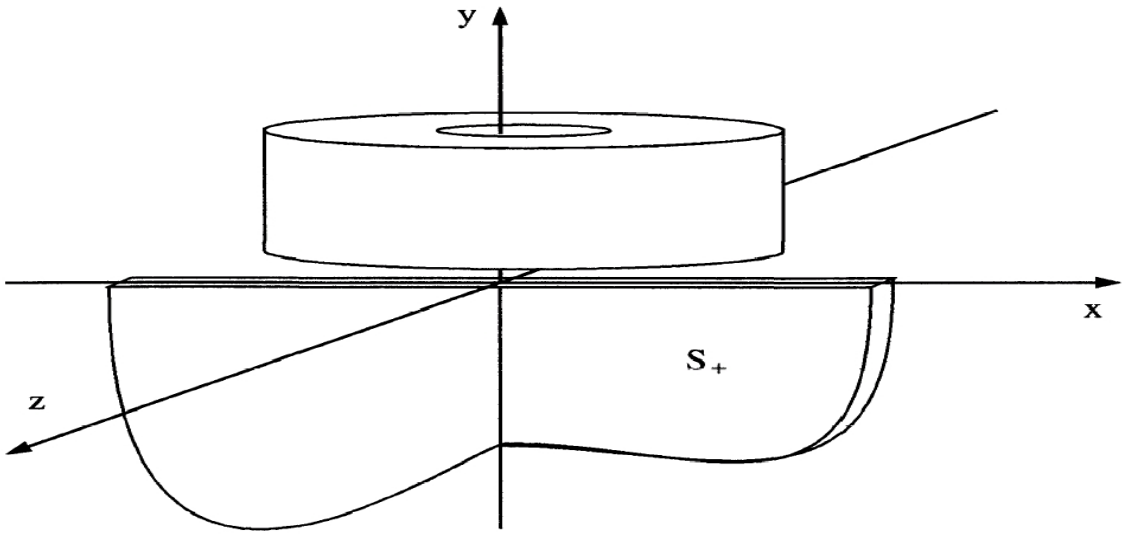
$$\begin{aligned}
\boldsymbol{\mu}_x &= \boldsymbol{\mu}_u + j\boldsymbol{\mu}_v \\
\boldsymbol{C}_x &= \mathrm{E}\left\{(\boldsymbol{x}-\boldsymbol{\mu}_x)(\boldsymbol{x}-\boldsymbol{\mu}_x)^H\right\} = 2\left(\boldsymbol{C}_{uu} + j\boldsymbol{C}_{vu}\right) \\
\boldsymbol{0} &= \mathrm{E}\left\{(\boldsymbol{x}-\boldsymbol{\mu}_x)(\boldsymbol{x}-\boldsymbol{\mu}_x)^T\right\}.
\end{aligned}$$

The (i,j)th element of the FIM for $\boldsymbol{x} \sim \mathcal{N}_{\mathrm{c}}(\boldsymbol{\mu}(\boldsymbol{\theta}), C(\boldsymbol{\theta}))$ is given by

$$\mathcal{I}_{i,j} = 2\mathrm{Re}\left\{\frac{\partial \boldsymbol{\mu}^H}{\partial \theta_i}\boldsymbol{C}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \theta_j}\right\} + \mathrm{tr}\left(\boldsymbol{C}^{-1}\frac{\partial \boldsymbol{C}}{\partial \theta_i}\boldsymbol{C}^{-1}\frac{\partial \boldsymbol{C}}{\partial \theta_j}\right)$$

see Kay-I, p. 525 — the proof is given in Appendix 15C of Kay-I.

Eddy-Current NDE Example (Crack Profile Inversion)



Coil above a conductor containing an open slot.

We adopt a complex deterministic signal-in-AWGN model:

$$x[n] = s[n, \boldsymbol{\theta}] + w[n], \quad n = 1, 2, \ldots, N$$

where

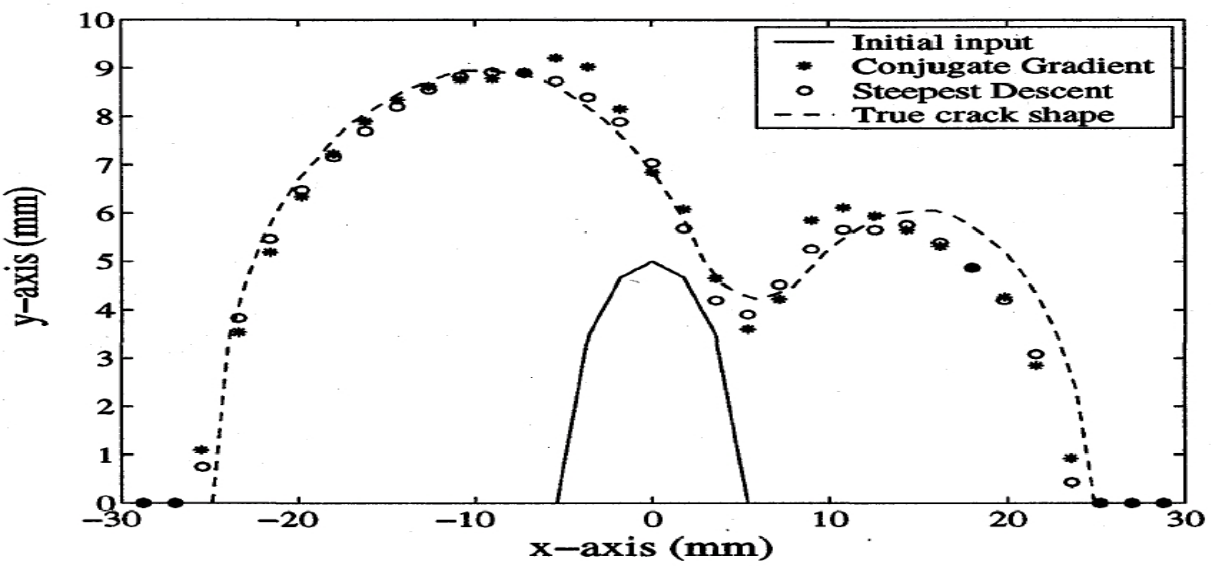
- $x[n] \equiv$ complex eddy-current impedance measurements collected at N locations (i.e. index n corresponds to a measurement location),
- $\boldsymbol{\theta} \equiv$ parameter vector describing the crack profile,
- $s[n, \boldsymbol{\theta}] \equiv$ model predictions for the crack profile described by $\boldsymbol{\theta}$. and

- additive complex white Gaussian noise (CWGN) with variance σ^2.

In

J.R. Bowler, W. Zhang, and A. Dogandžić, "Application of optimization methods to crack profile inversion using eddy current data," in *Rev. Progress Quantitative Nondestructive Evaluation,* D.O. Thompson and D.E. Chimenti (Eds.), Melville NY: Amer. Inst. Phys., vol. 22, 2003, pp. 742–749,

We utilize ML method to estimate $\boldsymbol{\theta}$. (In this case, ML estimation $\Longleftrightarrow$ nonlinear least-squares estimation).

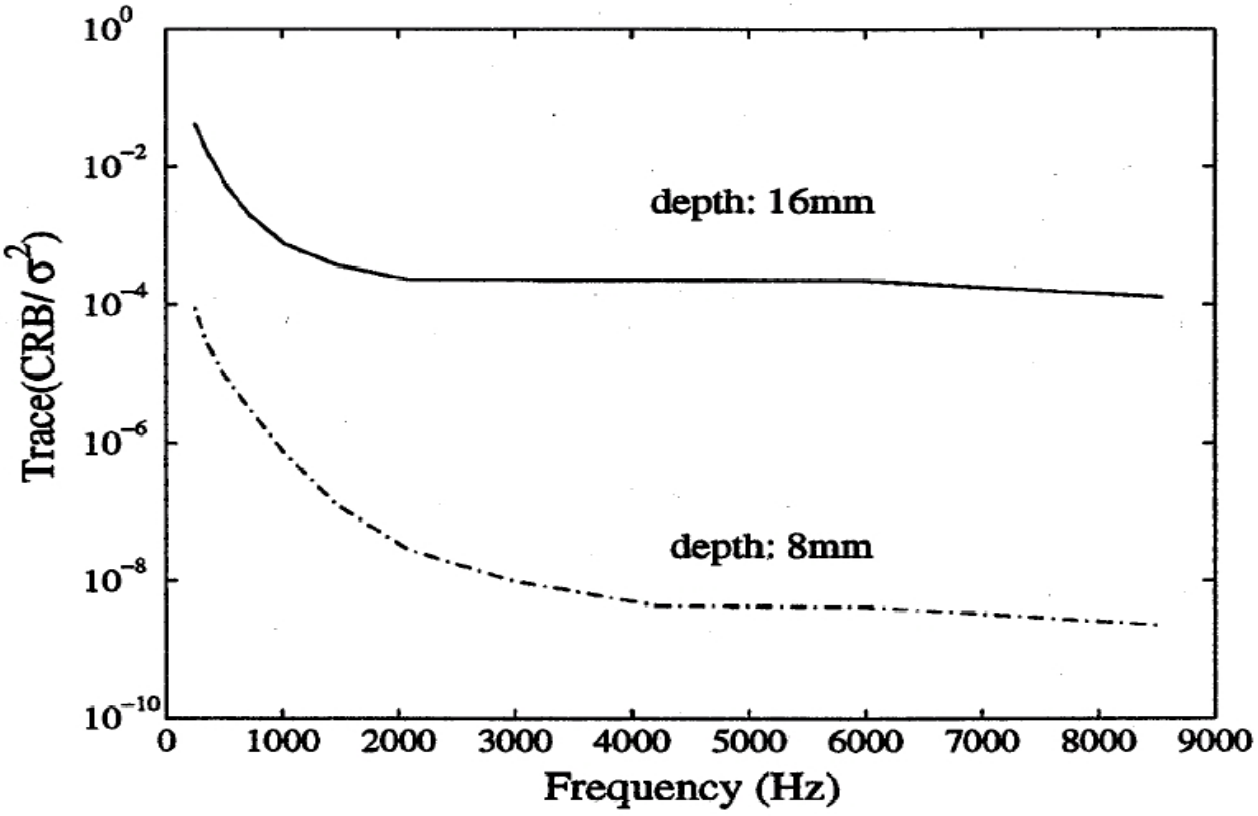


Nonlinear least-squares inversion of experimental impedance data to determine the crack shape using two optimization methods. We will discuss nonlinear least squares later in class. Let us focus here on the CRB.

CRB as a Tool for System Design

CRB vs. Frequency

Assume the noise variance σ^2 is constant across all the frequencies from 250 Hz to 8 kHz.

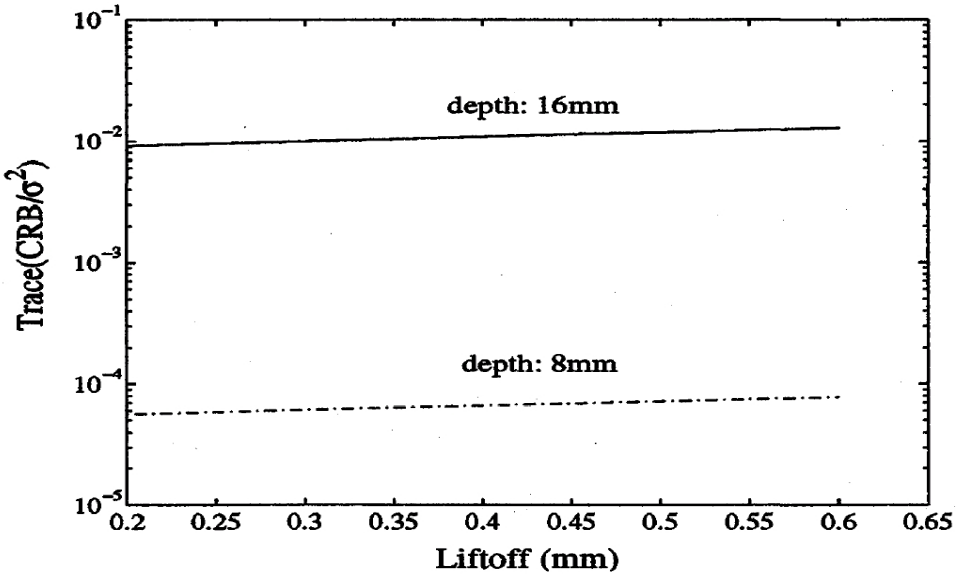


$\mathrm{trCRB}(\boldsymbol{\theta})/\sigma^2$ vs. frequency for crack depths 15 and 8 mm.

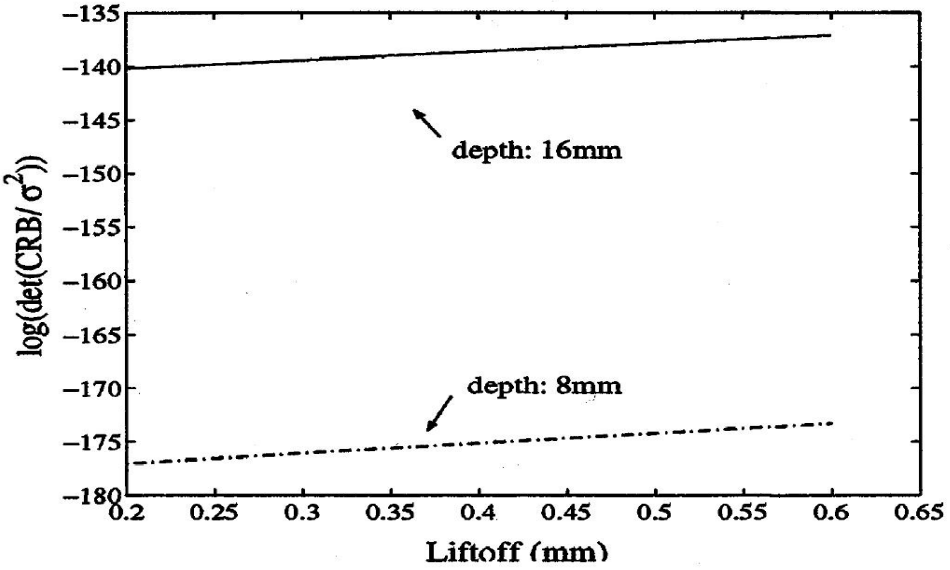
In the frequency domain in which our theory model is valid, the higher frequency the inversion is made at, the more accurate the inversion can be. As for different crack depths, the deeper the crack depth is, the larger the estimation error tends to be, which can be easily explained by the skin depth effect. Hence, to get better result, perform the inversion at as high frequency as possible.

CRB vs. Liftoff

Here, we assume that the noise variance σ^2 is constant across all liftoffs.



$\mathrm{trCRB}(\boldsymbol{\theta})/\sigma^2$ vs. liftoff for crack depths 15 and 8 mm.



$\det[\mathrm{CRB}(\boldsymbol{\theta})/\sigma^2]$ vs. liftoff for crack depths 15 and 8 mm.

(Approximately) exponential behavior of the CRB with liftoff.