12. Principles of Parameter Estimation

• consider the problem of estimating an unknown parameter of interest from a few of its noisy observations.

• Observations (measurement) are made on data that contain the desired parameter and undesired noise. Thus, for example, observations can be represented as

\[ X_i = \theta + n_i, \quad i = 1,2,\ldots,n. \]

Here \( \theta \) represents the unknown desired parameter, and \( n_i, \quad i = 1,2,\ldots,n \) represent random variables that may be dependent or independent from observation to observation.
Given $n$ observations $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, the estimation problem is to obtain the “best” estimator for the unknown parameter $\theta$ in terms of these observations.

Let us denote by $\hat{\theta}(X)$ the estimator for $\theta$. Obviously $\hat{\theta}(X)$ is a function of only the observations. “Best estimator” in what sense? Various optimization strategies can be used to define the term “best”.

Ideal solution would be when the estimate $\hat{\theta}(X)$ coincides with the unknown $\theta$. This of course may not be possible, and almost always any estimate will result in an error given by

$$e = \hat{\theta}(X) - \theta.$$  \hspace{1cm} (12-3)
Nonlinear estimators

- **Minimization of the mean square error (MMSE)**
  - Suppose $X$ and $\theta$ are two RVs, but we could only observe $X$. Based on $X$, we use a function $g(X)$ to obtain $\hat{\theta}$ as the estimation of $\theta$. As a measure, we try to minimize $E[(\theta - g(X))^2]$. The function is obtained as

$$\hat{\theta}_{\text{MMSE}}(X) = E[\theta \mid X] = \int \theta f(\theta \mid X)d\theta$$

- **Maximum likelihood (ML)**
  - Based on some observations of a random variable $X$, we estimate an unknown parameter $\theta$ by:

$$\hat{\theta}_{\text{ML}}(X) = \arg \max_\theta f(X \mid \theta)$$
Nonlinear estimators

- Maximum a posteriori probability (MAP)
  - Based on some observations of random variable $X$, we renew the pdf $f(\theta|X)$ and consider its peak value as the estimated value of $\theta$.

$$\hat{\theta}_{MAP}(X) = \arg\max_{\theta} f(\theta|X)$$
Some definitions

• **Unbiased estimator**
  – An estimator is unbiased if \( E[\hat{\theta}] = \theta \)
  Otherwise the estimator is biased estimator.

• **Consistent estimator**
  – An estimator is consistent if \( \sigma^2_{\hat{\theta}} \rightarrow 0 \) when \( n \rightarrow \infty \).
  – \( n \): the number of observations

• **Best estimator**
  – An estimator is called the best estimator if it minimizes the error
    \[
e = E[(\hat{\theta} - \theta)]
\]
Homework

The random variables

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \bar{v} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

are by definition the sample mean and the sample variance, respectively, of \( x_i \).

- If random variables \( X_i \) are independent with common mean \( \mu \) and variance \( \sigma^2 \), show that sample mean is an unbiased and consistent estimator of mean. Moreover, \( \bar{v} \) is an unbiased estimator of variance.
MMSE: an application

If a RV $y$ is to be estimated by a constant $c$ based on MSE principle we have the following:

$$e = E\{(y - c)^2\} = \int_{-\infty}^{\infty} (y - c)^2 f(y) dy \Rightarrow$$

We then minimize $e$ with respect to the unknown $c$.

$$\frac{\partial e}{\partial c} = -2 \int_{-\infty}^{\infty} (y-c) f(y) dy = 0 \Rightarrow c = \int_{-\infty}^{\infty} y f(y) dy, \text{ (Mean value)}$$
Maximum Likelihood (ML)

• ML assumes $X_1, X_2, \ldots, X_n$ has something to do with the unknown parameter $\theta$. More precisely, the joint p.d.f $f_X(x_1, x_2, \ldots, x_n; \theta)$ depends on $\theta$.

• ML assumes that the given sample data set is representative of the population $f_X(x_1, x_2, \ldots, x_n; \theta)$, and chooses that value for $\theta$ that most likely caused the observed data to occur.

• If $x_1, x_2, \ldots, x_n$ are given, $f_X(x_1, x_2, \ldots, x_n; \theta)$ is a function of $\theta$ alone, and the value of $\theta$ that maximizes the above p.d.f is the most likely value for $\theta$, and it is chosen as the ML estimate $\hat{\theta}_{ML}(X)$ for $\theta$. 

![Graph showing $f_X(x_1, x_2, \ldots, x_n; \theta)$ and $\hat{\theta}_{ML}(X)$]
Given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, the joint p.d.f $f_X(x_1, x_2, \ldots, x_n; \theta)$ represents the likelihood function, and the ML estimate can be determined either from the likelihood equation

$$\sup_{\hat{\theta}_{ML}} f_X(x_1, x_2, \ldots, x_n; \theta)$$

or using the log-likelihood function (sup in (12-4) represents the supremum operation)

$$L(x_1, x_2, \ldots, x_n; \theta) \triangleq \log f_X(x_1, x_2, \ldots, x_n; \theta). \quad (12-5)$$

If $L(x_1, x_2, \ldots, x_n; \theta)$ is differentiable and a supremum $\hat{\theta}_{ML}$ exists in (12-5), then that must satisfy the equation

$$\left. \frac{\partial \log f_X(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} \right|_{\theta = \hat{\theta}_{ML}} = 0. \quad (12-6)$$
Example 12.1: Let $X_i = \theta + w_i, \; i = 1 \rightarrow n,$ represent $n$ observations where $\theta$ is the unknown parameter of interest, and $w_i, \; i = 1 \rightarrow n,$ are zero mean independent normal r.vs with common variance $\sigma^2.$ Determine the ML estimate for $\theta$. Solution: Since $w_i$ are independent r.vs and $\theta$ is an unknown constant, we have $X_i$ s are independent normal random variables. Thus the likelihood function takes the form

$$f_X(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta).$$

(12-7)

Moreover, each $X_i$ is Gaussian with mean $\theta$ and variance $\sigma^2$ (Why?). Thus

$$f_{X_i}(x_i; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - \theta)^2 / 2\sigma^2}.$$ 

(12-8)

Substituting (12-8) into (12-7) we get the likelihood function to be...
\[
f_X(x_1, x_2, \ldots, x_n; \theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n} (x_i - \theta)^2 / 2\sigma^2}.
\]

(12-9)

It is easier to work with the log-likelihood function \( L(X; \theta) \) in this case. From (12-9)

\[
L(X; \theta) = \ln f_X(x_1, x_2, \ldots, x_n; \theta) = \frac{n}{2} \ln(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2},
\]

(12-10)

and taking derivative with respect to \( \theta \) as in (12-6), we get

\[
\left. \frac{\partial \ln f_X(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} \right|_{\theta = \hat{\theta}_{ML}} = 2 \sum_{i=1}^{n} \frac{(x_i - \theta)}{2\sigma^2} \right|_{\theta = \hat{\theta}_{ML}} = 0, \quad (12-11)
\]

or

\[
\hat{\theta}_{ML}(X) = \frac{1}{n} \sum_{i=1}^{n} X_i. \quad (12-12)
\]

Thus (12-12) represents the ML estimate for \( \theta \), which happens to be a linear estimator (linear function of the data) in this case.
Notice that the estimator is a r.v. Taking its expected value, we get

\[ E[\hat{\theta}_{ML}(x)] = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \theta, \]  

(12-13)
i.e., the expected value of the estimator does not differ from the desired parameter, and hence there is no bias between the two. Such estimators are known as **unbiased** estimators. Thus (12-12) represents an unbiased estimator for $\theta$.

Moreover the variance of the estimator is given by

\[
\text{Var} \left( \hat{\theta}_{ML} \right) = E[(\hat{\theta}_{ML} - \theta)^2] = \frac{1}{n^2} E \left\{ \left( \sum_{i=1}^{n} X_i - \theta \right)^2 \right\}
\]

\[
= \frac{1}{n^2} \left\{ \sum_{i=1}^{n} E(X_i - \theta)^2 + \sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n} E(X_i - \theta)(X_j - \theta) \right\}.
\]

The later terms are zeros since $X_i$ and $X_j$ are independent r.vs.
Then
\[
\text{Var} (\hat{\theta}_{ML}) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var} (X_i) = \frac{n \sigma^2}{n^2} = \frac{\sigma^2}{n}. 
\] (12-14)

Thus
\[
\text{Var}(\hat{\theta}_{ML}) \to 0 \quad \text{as} \quad n \to \infty, 
\] (12-15)

another desired property. We say such estimators (that satisfy (12-15)) are \textbf{consistent} estimators.

Next two examples show that ML estimator can be highly nonlinear.

**Example 12.2:** Let \( X_1, X_2, \ldots, X_n \) be independent, identically distributed uniform random variables in the interval \( (0, \theta) \) with common p.d.f
\[
f_{X_i}(x_i; \theta) = \frac{1}{\theta}, \quad 0 < x_i < \theta, \quad (12-16)
\]
where $\theta$ is an unknown parameter. Find the ML estimate for $\theta$.

Solution: The likelihood function in this case is given by

$$f_X(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta) = \frac{1}{\theta^n}, \quad 0 < x_i \leq \theta, \quad i = 1 \rightarrow n$$

$$= \frac{1}{\theta^n}, \quad 0 \leq \max(x_1, x_2, \ldots, x_n) \leq \theta. \quad (12-17)$$

From (12-17), the likelihood function in this case is maximized by the minimum value of $\theta$, and since $\theta \geq \max(X_1, X_2, \ldots, X_n)$, we get

$$\hat{\theta}_{ML}(X) = \max(X_1, X_2, \ldots, X_n) \quad (12-18)$$

to be the ML estimate for $\theta$. Notice that (18) represents a nonlinear function of the observations. To determine whether (12-18) represents an unbiased estimate for $\theta$, we need to evaluate its mean. To accomplish that in this case, it is easier to determine its p.d.f and proceed directly. Let $P_{\text{PILLAI}}^{14}$
\[ Z = \max( X_1, X_2, \cdots, X_n ) \]  
(12-19)

with \( X_i \) as in (12-16). Then

\[ F_z(z) = P[\max( X_1, X_2, \cdots, X_n ) \leq z] = P(X_1 \leq z, X_2 \leq z, \cdots, X_n \leq z) \]

\[ = \prod_{i=1}^{n} P(X_i \leq z) = \prod_{i=1}^{n} F_{X_i}(z) = \left( \frac{z}{\theta} \right)^n, \quad 0 < z < \theta, \]  
(12-20)

so that

\[ f_z(z) = \begin{cases} \frac{nz^{n-1}}{\theta^n}, & 0 < z < \theta, \\ 0, & \text{otherwise} \end{cases} \]  
(12-21)

Using (12-21), we get

\[ E[\hat{\theta}_{ML}(X)] = E(Z) = \int_0^\theta z f_z(z)dz = \frac{n}{\theta^n} \int_0^\theta z^n dz = \frac{n}{n+1} \frac{\theta^{n+1}}{\theta^n} = \frac{\theta}{(1+1/n)} . \]  
(12-22)

In this case \( E[\hat{\theta}_{ML}(X)] \neq \theta \), and hence the ML estimator is not an unbiased estimator for \( \theta \). However, from (12-22) as \( n \to \infty \)
\[
\lim_{n \to \infty} E[\hat{\theta}_{ML}(X)] = \lim_{n \to \infty} \frac{\theta}{(1 + 1/n)} = \theta,
\] (12-23)
i.e., the ML estimator is an asymptotically unbiased estimator. From (12-21), we also get
\[
E(Z^2) = \int_0^\theta z^2 f_z(z)dz = \frac{n}{\theta^n} \int_0^\theta z^{n+1}dz = \frac{n\theta^2}{n+2}
\] (12-24)
so that
\[
\text{Var}[\hat{\theta}_{ML}(X)] = E(Z^2) - [E(Z)]^2 = \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} = \frac{n\theta^2}{(n+1)(n+2)}.
\] (12-25)

Once again \( \text{Var}[\hat{\theta}_{ML}(X)] \to 0 \) as \( n \to \infty \), implying that the estimator in (12-18) is a consistent estimator.

**Example 12.3:** Let \( X_1, X_2, \ldots, X_n \) be i.i.d Gamma random variables with unknown parameters \( \alpha \) and \( \beta \). Determine the ML estimator for \( \alpha \) and \( \beta \).
Solution: Here \( x_i \geq 0 \), and
\[
f_X (x_1, x_2, \ldots, x_n; \alpha, \beta) = \frac{\beta^{n\alpha}}{\Gamma(\alpha)} \prod_{i=1}^{n} x_i^{\alpha-1} e^{-\beta \sum_{i=1}^{n} x_i}.
\] (12-26)

This gives the log-likelihood function to be
\[
L(x_1, x_2, \ldots, x_n; \alpha, \beta) = \log f_X (x_1, x_2, \ldots, x_n; \alpha, \beta)
= n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \left( \sum_{i=1}^{n} \log x_i \right) - \beta \sum_{i=1}^{n} x_i.
\] (12-27)

Differentiating \( L \) with respect to \( \alpha \) and \( \beta \) we get
\[
\frac{\partial L}{\partial \alpha} = n \log \beta - \frac{n}{\Gamma(\alpha)} \Gamma'(\alpha) + \sum_{i=1}^{n} \log x_i \bigg|_{\alpha, \beta = \hat{\alpha}, \hat{\beta}} = 0,
\] (12-28)
\[
\frac{\partial L}{\partial \beta} = \frac{n \alpha}{\beta} - \sum_{i=1}^{n} x_i \bigg|_{\alpha, \beta = \hat{\alpha}, \hat{\beta}} = 0.
\] (12-29)

Thus from (12-29)
\[
\hat{\beta}_{\text{ML}} (X) = \frac{\hat{\alpha}_{\text{ML}}}{\frac{1}{n} \sum_{i=1}^{n} x_i},
\] (12-30)
and substituting (12-30) into (12-28), it gives

$$\log \hat{\alpha}_{ML} - \frac{\Gamma'(\hat{\alpha}_{ML})}{\Gamma(\hat{\alpha}_{ML})} = \log \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right) - \frac{1}{n} \sum_{i=1}^{n} x_i. \quad (12-31)$$

Notice that (12-31) is highly nonlinear in $\hat{\alpha}_{ML}$.

In general the (log)-likelihood function can have more than one solution, or no solutions at all. Further, the (log)-likelihood function may not be even differentiable, or it can be extremely complicated to solve explicitly (see example 12.3, equation (12-31)).
Cramer - Rao Bound: Variance of any unbiased estimator $\hat{\theta}$ based on observations $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ for $\theta$ is lower bounded by

$$Var(\hat{\theta}) \geq \frac{1}{E\left(\frac{\partial \ln f_X (x_1, x_2, \ldots, x_n; \theta)}{\partial \theta}\right)^2} = \frac{-1}{E\left(\frac{\partial^2 \ln f_X (x_1, x_2, \ldots, x_n; \theta)}{\partial \theta^2}\right)}.$$ (12-32)

This important result states that the right side of (12-32) acts as a lower bound on the variance of all unbiased estimator for $\theta$. 
Naturally any unbiased estimator whose variance coincides with that in (12-32), must be the best. There are no better solutions! Such estimates are known as **efficient estimators**. Let us examine whether (12-12) represents an efficient estimator. Towards this using (12-11)

\[
\left( \frac{\partial \ln f_X(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} \right)^2 = \frac{1}{\sigma^4} \left( \sum_{i=1}^{n} (X_i - \theta) \right)^2; \quad (12-33)
\]

and

\[
E\left( \frac{\partial \ln f_X(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} \right)^2 = \frac{1}{\sigma^4} \left\{ \sum_{i=1}^{n} E[(X_i - \theta)^2] + \sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n} E[(X_i - \theta)(X_j - \theta)] \right\} = \frac{1}{\sigma^4} \sum_{i=1}^{n} \sigma^2 = \frac{n}{\sigma^2}, \quad (12-34)
\]

and substituting this into the first form on the right side of (12-32), we obtain the Cramer - Rao lower bound for this problem to be
But from (12-14) the variance of the ML estimator in (12-12) is the same as (12-35), implying that (12-12) indeed represents an efficient estimator in this case, the best of all possibilities!

It is possible that in certain cases there are no unbiased estimators that are efficient. In that case, the best estimator will be an unbiased estimator with the lowest possible Variance.
MAP estimator

What if the parameter of interest is a r.v with a-priori p.d.f \( f_\theta(\theta) \)? How does one obtain a good estimate for \( \theta \) based on the observations \( X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \)?

One technique is to use the observations to compute its a-posteriori probability density function \( f_{\theta|x}(\theta \mid x_1, x_2, \ldots, x_n) \). Of course, we can use the Bayes’ theorem in (11.22) to obtain this a-posteriori p.d.f. This gives

\[
f_{\theta|x}(\theta \mid x_1, x_2, \ldots, x_n) = \frac{f_{x|\theta}(x_1, x_2, \ldots, x_n \mid \theta) f_\theta(\theta)}{f_x(x_1, x_2, \ldots, x_n)}. \tag{12-36}
\]

Notice that (12-36) is only a function of \( \theta \), since \( x_1, x_2, \ldots, x_n \) represent given observations. Once again, we can look for
the most probable value of $\theta$ suggested by the above a-posteriori p.d.f. Naturally, the most likely value for $\theta$ is that corresponding to the maximum of the a-posteriori p.d.f (see Fig. 12.2). This estimator - maximum of the a-posteriori p.d.f is known as the MAP estimator for $\theta$.