12. Principles of Parameter Estimation

The purpose of this lecture is to illustrate the usefulness of the various concepts introduced and studied in earlier lectures to practical problems of interest. In this context, consider the problem of estimating an unknown parameter of interest from a few of its noisy observations. For example, determining the daily temperature in a city, or the depth of a river at a particular spot, are problems that fall into this category.

Observations (measurement) are made on data that contain the desired nonrandom parameter θ and undesired noise. Thus, for example,

$$\text{Observation} = \text{signal (desired part)} + \text{noise}, \qquad (12\text{-}1)$$

or, the i th observation can be represented as

$$X_i = \theta + n_i, \quad i = 1,2,\cdots,n. \qquad (12\text{-}2)$$

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

Let us denote by $\hat{\theta}(X)$ the estimator for θ. 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 θ. This of course may not be possible, and almost always any estimate will result in an error given by

$$e = \hat{\theta}(X) - \theta . \tag{12-3}$$

One strategy would be to select the estimator $\hat{\theta}(X)$ so as to minimize some function of this error - such as - minimization of the mean square error (MMSE), or minimization of the absolute value of the error etc.

A more fundamental approach is that of the **principle of Maximum Likelihood** (ML).

The underlying assumption in any estimation problem is

that the available data $X_1, X_2, \cdots, X_n$ has something to do with the unknown parameter θ. More precisely, we assume that the joint p.d.f of $X_1, X_2, \cdots, X_n$ given by $f_X(x_1, x_2, \cdots, x_n; \theta)$ depends on θ. The method of maximum likelihood assumes that the given sample data set is representative of the population $f_X(x_1, x_2, \cdots, x_n; \theta)$, and chooses that value for θ that most likely caused the observed data to occur, i.e., once observations $x_1, x_2, \cdots, x_n$ are given, $f_X(x_1, x_2, \cdots, x_n; \theta)$ is a function of θ alone, and the value of θ that maximizes the above p.d.f is the most likely value for θ, and it is chosen as the ML estimate $\hat{\theta}_{ML}(X)$ for θ (Fig. 12.1).

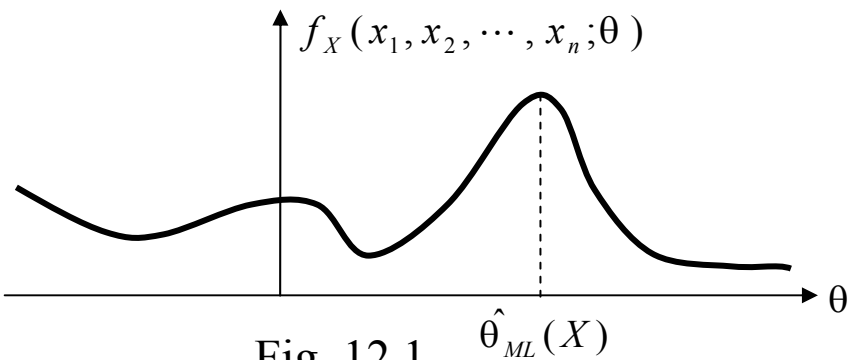

Fig. 12.1

Given $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n,$ the joint p.d.f $f_X(x_1, x_2, \cdots, x_n; \theta)$ represents the likelihood function, and the ML estimate can be determined either from the likelihood equation

$$\sup_{\hat{\theta}_{ML}} \; f_X(x_1, x_2, \cdots, x_n; \theta) \tag{12-4}$$

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

$$L(x_1, x_2, \cdots, x_n; \theta) \stackrel{\Delta}{=} \log f_X(x_1, x_2, \cdots, x_n; \theta). \tag{12-5}$$

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

$$\left. \frac{\partial \log f_X(x_1, x_2, \cdots, x_n; \theta)}{\partial \theta} \right|_{\theta = \hat{\theta}_{ML}} = 0. \tag{12-6}$$

We will illustrate the above procedure through several examples:

Example 12.1: Let $X_i = \theta + w_i, \quad i = 1 \to n,$ represent n observations where θ is the unknown parameter of interest, and $w_i, \quad i = 1 \to n,$ are zero mean independent normal r.vs with common variance σ^2. Determine the ML estimate for θ.

Solution: Since w_i are independent r.vs and θ is an unknown constant, we have X_i s are independent normal random variables. Thus the likelihood function takes the form

$$f_X(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta). \tag{12-7}$$

Moreover, each X_i is Gaussian with mean θ and variance σ^2 (Why?). Thus

$$f_{X_i}(x_i; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - \theta)^2 / 2\sigma^2}. \tag{12-8}$$

Substituting (12-8) into (12-7) we get the likelihood function to be

$$f_X(x_1, x_2, \cdots, x_n;\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(x_i-\theta)^2/2\sigma^2}. \qquad (12\text{-}9)$$

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

$$L(X;\theta) = \ln f_X(x_1, x_2, \cdots, x_n;\theta) = \frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2\sigma^2}, \qquad (12\text{-}10)$$

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

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

or

$$\hat{\theta}_{ML}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i. \qquad (12\text{-}12)$$

Thus (12-12) represents the ML estimate for θ, 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

$$E[\hat{\theta}_{ML}(x)] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \theta, \tag{12-13}$$

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 θ. Moreover the variance of the estimator is given by

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

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

The later terms are zeros since X_i and X_j are independent r.vs.

Then

$$Var(\hat{\theta}_{ML}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \qquad (12\text{-}14)$$

Thus

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

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

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

Example 12.2: Let $X_1, X_2, \cdots, X_n$ be independent, identically distributed uniform random variables in the interval $(0, \theta)$ with common p.d.f

$$f_{X_i}(x_i;\theta) = \frac{1}{\theta}, \quad 0 < x_i < \theta, \qquad (12\text{-}16)$$

where θ is an unknown parameter. Find the ML estimate for θ.

Solution: The likelihood function in this case is given by

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

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

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

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

to be the ML estimate for θ. Notice that (18) represents a nonlinear function of the observations. To determine whether (12-18) represents an unbiased estimate for θ, 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

$$Z = \max(\, X_1, X_2, \cdots, X_n) \tag{12-19}$$

with X_i as in (12-16). Then

$$F_Z(z) = P[\max(\, X_1, X_2, \cdots, X_n) \le z] = P(X_1 \le z, X_2 \le z, \cdots, X_n \le z)$$

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

so that

$$f_Z(z) = \begin{cases} \dfrac{n z^{\,n-1}}{\theta^{\,n}}, & 0 < z < \theta, \\ 0, & \text{otherwise} \quad . \end{cases} \tag{12-21}$$

Using (12-21), we get

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

In this case $E[\hat{\theta}_{ML}(X)] \neq \theta$, and hence the ML estimator is not an unbiased estimator for θ. However, from (12-22) as $n \to \infty$

$$\lim_{n\to\infty} E[\hat{\theta}_{ML}(X)] = \lim_{n\to\infty} \frac{\theta}{(1+1/n)} = \theta, \qquad (12\text{-}23)$$

i.e., the ML estimator is an asymptotically unbiased estimator. From (12-21), we also get

$$E(Z^2) = \int_0^\theta z^2 f_Z(z)dz = \frac{n}{\theta^n}\int_0^\theta z^{n+1}dz = \frac{n\theta^2}{n+2} \qquad (12\text{-}24)$$

so that

$$Var[\hat{\theta}_{ML}(X)] = E(Z^2) - [E(Z)]^2 = \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} = \frac{n\theta^2}{(n+1)^2(n+2)}. \qquad (12\text{-}25)$$

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

Example 12.3: Let $X_1, X_2, \cdots, X_n$ be i.i.d Gamma random variables with unknown parameters α and β. Determine the ML estimator for α and β.

Solution: Here $x_i \geq 0,$ and

$$f_X(x_1, x_2, \cdots, x_n; \alpha, \beta) = \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n} \prod_{i=1}^{n} x_i^{\alpha-1} \; e^{-\beta \sum_{i=1}^{n} x_i}. \qquad (12\text{-}26)$$

This gives the log-likelihood function to be

$$L(x_1, x_2, \cdots, x_n; \alpha, \beta) = \log f_X(x_1, x_2, \cdots, x_n; \alpha, \beta)$$

$$= n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1)\left(\sum_{i=1}^{n} \log x_i \right) - \beta \sum_{i=1}^{n} x_i. \qquad (12\text{-}27)$$

Differentiating L with respect to α and β we get

$$\frac{\partial L}{\partial \alpha} = n \log \beta - \frac{n}{\Gamma(\alpha)} \Gamma'(\alpha) + \sum_{i=1}^{n} \log x_i \Big|_{\alpha, \beta = \hat{\alpha}, \hat{\beta}} = 0, \qquad (12\text{-}28)$$

$$\frac{\partial L}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i \Big|_{\alpha, \beta = \hat{\alpha}, \hat{\beta}} = 0. \qquad (12\text{-}29)$$

Thus from (12-29)

$$\hat{\beta}_{ML}(X) = \frac{\hat{\alpha}_{ML}}{\frac{1}{n} \sum_{i=1}^{n} x_i}, \qquad (12\text{-}30)$$

and substituting (12-30) into (12-28), it gives

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

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)).

Best Unbiased Estimator:

Referring back to example 12.1, we have seen that (12-12) represents an unbiased estimator for θ with variance given by (12-14). It is possible that, for a given *n*, there may be other

unbiased estimators to this problem with even lower variances. If such is indeed the case, those estimators will be naturally preferable compared to (12-12). In a given scenario, is it possible to determine the lowest possible value for the variance of *any* unbiased estimator? Fortunately, a theorem by Cramer and Rao (Rao 1945; Cramer 1948) gives a complete answer to this problem.

Cramer - Rao Bound: Variance of any unbiased estimator $\hat{\theta}$ based on observations $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ for θ must satisfy the lower bound

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

This important result states that the right side of (12-32) acts as a lower bound on the variance of *all* unbiased estimator for θ, provided their joint p.d.f satisfies certain regularity restrictions. (see (8-79)-(8-81), Text).

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)

$$\left(\frac{\partial \ln f_X(x_1, x_2, \cdots, x_n;\theta)}{\partial\theta}\right)^2 = \frac{1}{\sigma^4}\left(\sum_{i=1}^{n}(X_i - \theta)\right)^2; \qquad (12\text{-}33)$$

and

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

$$= \frac{1}{\sigma^4}\sum_{i=1}^{n}\sigma^2 = \frac{n}{\sigma^2}, \qquad (12\text{-}34)$$

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

$$\frac{\sigma^2}{n}. \qquad (12\text{-}35)$$

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.

How does one find such an unbiased estimator?

Fortunately Rao-Blackwell theorem (page 335-337, Text) gives a complete answer to this problem.

Cramer-Rao bound can be extended to multiparameter case as well (see page 343-345,Text).

So far, we discussed nonrandom parameters that are unknown. 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 θ based on the observations $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$?

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

$$f_{\theta|X}(\theta \mid x_1, x_2, \cdots, x_n) = \frac{f_{X|\theta}(x_1, x_2, \cdots, x_n \mid \theta) f_\theta(\theta)}{f_X(x_1, x_2, \cdots, x_n)}. \qquad \text{(12-36)}$$

Notice that (12-36) is only a function of θ, since $x_1, x_2, \cdots, x_n$ represent given observations. Once again, we can look for

the most probable value of θ suggested by the above a-posteriori p.d.f. Naturally, the most likely value for θ 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 θ.

It is possible to use other optimality criteria as well. Of course, that should be the subject matter of another course!

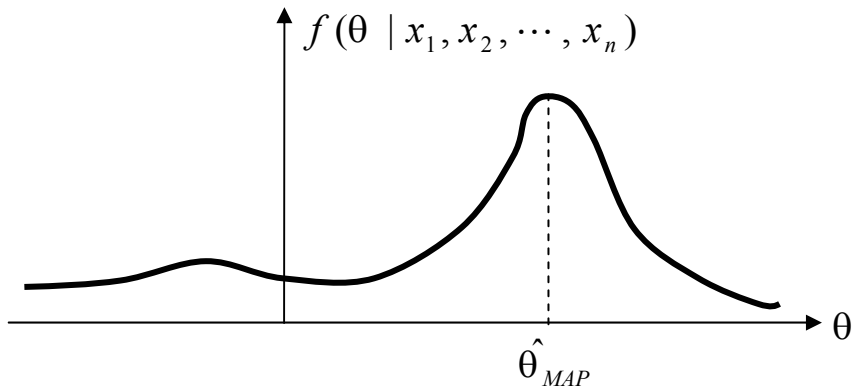

Fig. 12.2