Statistical Estimation Theory (Lecture 13: Maximum Likelihood Parameter Estimation)

Dr. Laxminarayana S Pillutla
DA-IICT, Gandhinagar, Gujarat, India.
Email: laxminarayana_pillutla@daiict.ac.in

April 08, 2010
The material in this lecture is based on the book “An Introduction to Signal Detection and Estimation” by H. V. Poor.

Other Relevant References:

- “Fundamentals of Statistical Signal Processing Theory volume II: Detection Theory” by Steven M Kay
Motivation for ML based Parameter Estimation

• In the Bayesian parameter estimation approach, we considered estimation of an unknown quantity by modelling it as random.
• However in certain cases, the unknown parameter is not modelled as random.
  ◦ In this case we simply have an observation $Y \in \Gamma$ using which we need to estimate the parameter $\theta \in \Lambda$.

• The minimum variance unbiased estimator (MVUE) is a natural choice for non-random parameter estimation (see Section IV.C of the H. V. Poor’s book: “An Introduction to Signal Detection and Estimation”).
• However MVUE determination based on complete sufficient statistics may not be applicable in many situations: (i) Due to lack of useful complete sufficient statistics. (ii) Due to intractability of the required analysis.
• The ML based parameter estimation is a good alternative to MVUE estimation.
Maximum Likelihood (ML) Estimation

• Recall the MAP estimator we discussed earlier:

\[ \hat{\theta}_{\text{MAP}}(y) = \arg \max_{\theta \in \Lambda} p_\theta(y)w(\theta). \] (1)

• Computation of the MAP estimator above requires specification of the prior density \( w(\theta) \) on \( \Lambda \).

• In the absence of any prior information about the parameter, a natural choice of prior distribution is Uniform on \( \Lambda \).
  
  ◦ i.e., \( w(\theta) \) is constant on \( \Lambda \).

• Since \( w(\theta) \) is constant on \( \Lambda \), the MAP estimator in (1) may as well be determined such that the “likelihood function \( p_\theta(y) \)” is maximized (since constant wouldn’t affect the optimizer!!). The value of \( \theta \) that maximizes \( p_\theta(y) \) is called the “ML estimate”:

\[ \hat{\theta}_{\text{ML}}(y) = \arg \max_{\theta \in \Lambda} p_\theta(y). \] (2)
Remarks on ML Estimation

- There are (at least) two problems with the above argument:
  1. Assuming a uniform prior for the parameter \( \theta \) is not the same as assuming that the prior is unknown (or that the parameter is not random).
  2. In certain cases the parameter set \( \Lambda \) may not be bounded, therefore it may not be possible to construct uniform prior distribution (for example if \( \Lambda = \mathbb{R} \)).

- But ML estimation can indeed be motivated in more direct ways (not just as a surrogate to MAP estimation).
  - Indeed determining \( \theta \) that makes the observation most likely is in itself a legitimate and meaningful criterion.
  - The ML estimate in many cases can be shown to be asymptotically (i) unbiased and (ii) Gaussian.
  - It is also asymptotically efficient in the sense that it achieves the so called Cramer-Rao Lower Bound (CRLB).

- To understand above we define them formally...
Unbiased Estimator and MVUE

- An estimator $\hat{\theta}(y)$ of the parameter $\theta$ is said to be unbiased if

$$E_\theta \left[ \hat{\theta}(y) \right] = \theta ,$$  \hspace{1cm} (3)

where $E_\theta$ denotes expectation w.r.t. the density $p_\theta(y)$.

- The estimator $\hat{\theta}(y)$ is said to be asymptotically unbiased if (3) holds for a large sample size $N$ (i.e., $N \to \infty$).

- An unbiased estimator that minimizes the mean squared error for each $\theta \in \Lambda$ is called a **Minimum-variance unbiased estimator (MVUE)**.
Information Inequality and CRLB

- The CRLB can be obtained as a special case of the so called “Information Inequality”.
- The “Information Inequality” provides a fundamental limit on any estimator’s performance.
- It can be used as a benchmark against which we can compare the performance of arbitrary ad hoc estimator.
- Formally, the information inequality is stated in the following slide...
Information Inequality

Suppose that \( \hat{\theta} \) is an estimate of the parameter \( \theta \) in a family \( \{ P_\theta : \theta \in \Lambda \} \) and that the following conditions hold:

(C1) \( \Lambda \) is an open interval.

(C2) The family \( \{ P_\theta : \theta \in \Lambda \} \) has a corresponding family of densities \( \{ p_\theta : \theta \in \Lambda \} \), all of the members of which have the same support: i.e., the set \( \{ y | p_\theta(y) > 0 \} \) is the same for all \( \theta \in \Lambda \).

(C3) \( \frac{\partial}{\partial \theta} p_\theta(y) \) exists and is finite for all \( \theta \in \Lambda \) and all \( y \) in the support of \( p_\theta \).

(C4) \( \frac{\partial}{\partial \theta} \left[ \int_{\Gamma} h(y) p_\theta(y) \mu(dy) \right] \) exists and equals \( \int_{\Gamma} h(y) \frac{\partial}{\partial \theta} [p_\theta(y)] \mu(dy) \) for all \( \theta \in \Lambda \), for \( h(y) = \hat{\theta}(y) \) and \( h(y) = 1 \).

Then

\[
\text{Var}_\theta \left( \hat{\theta}(Y) \right) \geq \left( \frac{\frac{\partial}{\partial \theta} \left[ E_\theta \left\{ \hat{\theta}(Y) \right\} \right]}{I_\theta} \right)^2,
\]

where \( I_\theta \) is known as Fisher’s information for estimating \( \theta \) from \( Y \) and is given by
Information Inequality (ctd...)

\[ I_\theta = E \left\{ \left( \frac{\partial}{\partial \theta} \left[ \log (p_\theta(Y)) \right] \right)^2 \right\} . \]  (5)

Furthermore if the following condition also holds:

(C5) \( \frac{\partial^2}{\partial \theta^2} [p_\theta(y)] \) exists for all \( \theta \in \Lambda \) and \( y \) in the support of \( p_\theta \), and

\[
\int \frac{\partial^2}{\partial \theta^2} [p_\theta(y)] \mu(dy) = \frac{\partial^2}{\partial \theta^2} \left[ \int p_\theta(y) \mu(dy) \right],
\]

then

\[ I_\theta = -E_\theta \left\{ \frac{\partial^2}{\partial \theta^2} [\log (p_\theta(Y))] \right\} . \]  (6)
Remarks on Information Inequality

• The (4) is known as “Information Inequality”.
• Higher the Fisher’s Information $I_\theta$ for a given model, better the lower bound on estimation accuracy.
• The information inequality doesn’t say whether there exists an estimator that achieves the lower bound given by (4).
• An estimator that achieves equality in the information inequality (4) may exist only under certain Conditions (See the Reference Lehmann (1983) in the H. V. Poor’s book.)
Proof of Information Inequality

• By definition

\[ E_\theta \left\{ \hat{\theta}(Y) \right\} = \int_\Gamma \hat{\theta}(y)p_\theta(y)\mu(dy). \] (7)

• Differentiating (7) and using the condition C4:

\[
\frac{\partial}{\partial \theta} \left[ E_\theta \left\{ \hat{\theta}(Y) \right\} \right] = \frac{\partial}{\partial \theta} \left[ \int_\Gamma \hat{\theta}(y)p_\theta(y)\mu(dy) \right]
= \int_\Gamma \hat{\theta}(y) \frac{\partial}{\partial \theta} [p_\theta(y)] \mu(dy), \] (8)

where (8) follows from Condition C4 with \( h(y) = \hat{\theta}(y) \).

• Also the condition C4 with \( h(y) = 1 \) gives:

\[
\int_\Gamma \frac{\partial}{\partial \theta} [p_\theta(y)] = \frac{\partial}{\partial \theta} \left[ \int_\Gamma p_\theta(y)\mu(dy) \right]
= \frac{\partial}{\partial \theta}[1] = 0. \] (9)
Proof of Information Inequality (ctd...)

• In light of (9) we can write (8) as follows:

\[
\frac{\partial}{\partial \theta} \left[ E_\theta \left\{ \hat{\theta}(Y) \right\} \right] = \int_\Gamma \left( \hat{\theta}(y) - E_\theta \left\{ \hat{\theta}(Y) \right\} \right) \frac{\partial}{\partial \theta} [p_\theta(y)] \mu(dy)
\]

(since \( \frac{\partial}{\partial \theta} \left[ \log(p_\theta(y)) \right] = \frac{1}{p_\theta(y)} \frac{\partial}{\partial \theta} [p_\theta(y)] \))

\[
\therefore \frac{\partial}{\partial \theta} \left[ E_\theta \left\{ \hat{\theta}(Y) \right\} \right] = \int_\Gamma \left( \hat{\theta}(y) - E_\theta \left\{ \hat{\theta}(Y) \right\} \right) \times \\
\left( \frac{\partial}{\partial \theta} \left[ \log(p_\theta(y)) \right] \right) p_\theta(y) \mu(dy)
\]

• The above equation can therefore be expressed as

\[
\frac{\partial}{\partial \theta} \left[ E_\theta \left\{ \hat{\theta}(Y) \right\} \right] = E_\theta \left\{ \left( \hat{\theta}(Y) - E_\theta \left\{ \hat{\theta}(Y) \right\} \right) \left( \frac{\partial}{\partial \theta} \left[ \log(p_\theta(y)) \right] \right) \right\} .
\]
Cauchy-Schwarz (CS) Inequality

- The CS inequality states that for any two functions $f$ and $g$

$$
\left( \int |f g| d\mu \right)^2 \leq \int f^2 d\mu \int g^2 d\mu ,
$$

with equality if and only if $f = cg$ for some constant $c \in \mathbb{R}$.
Proof of Information Inequality (ctd...)

- Now to obtain the information inequality all we require is a simple application of CS inequality to (10). Doing so we obtain

\[
\left( \frac{\partial}{\partial \theta} \left[ E_{\theta} \left\{ \hat{\theta}(Y) \right\} \right] \right)^2 \leq E_{\theta} \left\{ \left( \hat{\theta}(Y) - E_{\theta} \left\{ \hat{\theta}(Y) \right\} \right)^2 \right\} \times E_{\theta} \left\{ \left( \frac{\partial}{\partial \theta} \left[ \log(p_{\theta}(y)) \right] \right)^2 \right\} \quad (12)
\]

\[
= E_{\theta} \left\{ \left( \hat{\theta}(Y) - E_{\theta} \left\{ \hat{\theta}(Y) \right\} \right)^2 \right\} I_{\theta}, \quad (13)
\]

where (13) follows by using the definition of \( I_{\theta} \) in (5).

- But by definition

\[
\text{Var}_{\theta} \left[ \hat{\theta}(Y) \right] = E_{\theta} \left\{ \left( \hat{\theta}(Y) - E_{\theta} \left\{ \hat{\theta}(Y) \right\} \right)^2 \right\} \quad (14)
\]
Proof of Information Inequality (ctd...)

- Therefore, from (13) and (14) we have

\[
\left( \frac{\partial}{\partial \theta} \left[ E_\theta \{ \hat{\theta}(Y) \} \right] \right)^2 \leq \text{Var} \left[ \hat{\theta}(Y) \right] I_\theta
\]

\[
\therefore \text{Var}_\theta \left[ \hat{\theta}(Y) \right] \geq \frac{\left( \frac{\partial}{\partial \theta} \left[ E_\theta \{ \hat{\theta}(Y) \} \right] \right)^2}{I_\theta}
\]

as desired in (4), which completes the proof of information inequality.
Proof of Information Inequality (ctd...)

To prove (6):
Note that,

\[
\frac{\partial^2}{\partial \theta^2} \left[ \log(p_\theta(y)) \right] = \frac{\partial}{\partial \theta} \left[ \frac{\partial}{\partial \theta} \left[ \log(p_\theta(y)) \right] \right]
\]

\[
= \frac{\partial}{\partial \theta} \left[ \frac{1}{p_\theta(y)} \frac{\partial}{\partial \theta} [p_\theta(y)] \right]
\]

\[
= -\frac{1}{(p_\theta(y))^2} \left( \frac{\partial}{\partial \theta} [p_\theta(y)] \right)^2 + \frac{1}{p_\theta(y)} \frac{\partial^2}{\partial \theta^2} [p_\theta(y)]
\]

\[
= -\left( \frac{1}{p_\theta(y)} \frac{\partial}{\partial \theta} [p_\theta(y)] \right)^2 + \frac{\partial^2}{\partial \theta^2} [p_\theta(y)]
\]

Therefore

\[
\frac{\partial^2}{\partial \theta^2} \left[ \log(p_\theta(y)) \right] = -\left( \frac{\partial}{\partial \theta} \left[ \log(p_\theta(y)) \right] \right)^2 + \frac{\partial^2}{\partial \theta^2} [p_\theta(y)]
\]
Proof of Information Inequality (ctd...)

- Applying $E_{\theta}$ operator on both sides of (16) we obtain:

\[
E_{\theta} \left\{ \frac{\partial^2}{\partial \theta^2} \left[ \log (p_{\theta}(y)) \right] \right\} = -E_{\theta} \left\{ \left( \frac{\partial}{\partial \theta} \left[ \log (p_{\theta}(y)) \right] \right)^2 \right\} + E_{\theta} \left\{ \frac{\partial^2}{\partial \theta^2} \left[ p_{\theta}(y) \right] \right\} \\
= -I_{\theta} + E_{\theta} \left\{ \frac{\partial^2}{\partial \theta^2} \left[ p_{\theta}(y) \right] \right\} p_{\theta}(y) \text{ (using the definition of $I_{\theta}$ in (5))}
\]

\[
\therefore I_{\theta} = -E_{\theta} \left\{ \frac{\partial^2}{\partial \theta^2} \left[ \log (p_{\theta}(y)) \right] \right\} + E_{\theta} \left\{ \frac{\partial^2}{\partial \theta^2} \left[ p_{\theta}(y) \right] \right\} \tag{17}
\]
Proof of Information Inequality (ctd...)

• Now consider

\[ E_{\theta} \left\{ \frac{\partial^2}{\partial \theta^2} \left[ \frac{p_\theta(y)}{p_\theta(y)} \right] \right\} = \int_\Gamma \frac{\partial^2}{\partial \theta^2} \left[ p_\theta(y) \right] p_\theta(y) \mu(dy) \]

\[ = \int_\Gamma \frac{\partial^2}{\partial \theta^2} \left[ p_\theta(y) \right] \mu(dy) \]

\[ = \frac{\partial^2}{\partial \theta^2} \left[ \int_\Gamma p_\theta(y) \mu(dy) \right] \quad \text{(from condition C5)} \]

\[ = \frac{\partial^2}{\partial \theta^2} [1] = 0 \quad \text{(18)} \]

• Using (18) in (17) we have

\[ I_\theta = -E_{\theta} \left\{ \frac{\partial^2}{\partial \theta^2} \left[ \log (p_\theta(y)) \right] \right\} , \]

as desired in (6), which completes the proof.
Special Case: CRLB for Unbiased Estimators

- When the estimator $\hat{\theta}(y)$ is unbiased, we have

$$E_\theta \left\{ \hat{\theta}(Y) \right\} = \theta$$

$$\therefore \frac{\partial}{\partial \theta} \left[ E_\theta \left\{ \hat{\theta}(Y) \right\} \right] = \frac{\partial}{\partial \theta} [\theta]$$

$$= 1.$$  \hspace{1cm} (19)

- Using (19) in the information inequality given by (4), we get the so called CRLB for an unbiased estimator:

$$\text{Var}_\theta \left[ \hat{\theta}(Y) \right] \geq \frac{1}{I_\theta},$$  \hspace{1cm} (20)

where $I_\theta$ is given by (5).
Necessary Condition to Achieve the Information Inequality Lower Bound

- An obvious question is “when does an estimator (unbiased estimator) achieve the Information inequality lower bound (CRLB)?”
- Recall that to prove information inequality we used the Cauchy-Schwarz (CS) inequality. The equality in CS inequality happens if and only if \( f(y) = cg(y) \). Using this condition for the information inequality case with \( f(y) = \frac{\partial}{\partial \theta} \left[ \log (p_\theta(y)) \right] \) and \( g(y) = \left( \hat{\theta}(y) - E_\theta \left\{ \hat{\theta}(Y) \right\} \right) \) we obtain

\[
\frac{\partial}{\partial \theta} \left[ \log (p_\theta(y)) \right] = K(\theta) \left( \hat{\theta}(y) - E_\theta \left\{ \hat{\theta}(Y) \right\} \right),
\]

(21)

with probability 1 under \( P_\theta \) for some constant \( K(\theta) \) (to be determined).
- Next let \( \Lambda = (a, b) \) (an open interval) and \( f(\theta) = E_\theta \left\{ \hat{\theta}(y) \right\} \).
Necessary Condition to Achieve the Information Inequality Lower Bound (ctd...)

- Thus from (21), \( \hat{\theta}(y) \) achieves the information inequality lower bound if and only if:

\[
\int_{a}^{\theta} \frac{\partial}{\partial \theta} \left[ \log (p_{\sigma}(y)) \right] d\sigma = \int_{a}^{\theta} K(\sigma) \left( \hat{\theta}(y) - f(\sigma) \right) d\sigma \quad \text{for } y \in \Gamma
\]

\[
[\log (p_{\sigma}(y))]_{a}^{\theta} = \int_{a}^{\theta} K(\sigma) \left( \hat{\theta}(y) - f(\sigma) \right) d\sigma
\]

\[
\log \left( \frac{p_{\theta}(y)}{p_{a}(y)} \right) = \int_{a}^{\theta} K(\sigma) \left( \hat{\theta}(y) - f(\sigma) \right) d\sigma \quad (22)
\]

\[
\therefore p_{\theta}(y) = \left( e^{\int_{a}^{\theta} K(\sigma)f(\sigma)d\sigma} \right) \left( e^{\hat{\theta}(y) \int_{a}^{\theta} K(\sigma)d\sigma} \right) p_{a}(y) \quad (23)
\]

- Thus if a given estimator achieves the equality in information inequality, then \( p_{\theta}(y) \) should satisfy (23).
**Determination of** \( K(\theta) \)

- Recall from (21), the equality holds in the information inequality, if

\[
\frac{\partial}{\partial \theta} \left[ \log (p_{\theta}(y)) \right] = K(\theta) \left( \hat{\theta}(y) - E_{\theta} \left\{ \hat{\theta}(Y) \right\} \right).
\]

- Of course, if above is true, then

\[
\text{Var}_{\theta} \left[ \hat{\theta}(Y) \right] = \frac{\left( \frac{\partial}{\partial \theta} \left[ E_{\theta} \left\{ \hat{\theta}(Y) \right\} \right] \right)^2}{I_{\theta}} \quad (24)
\]

- But

\[
I_{\theta} = E \left\{ \left( \frac{\partial}{\partial \theta} \left[ \log (p_{\theta}(Y)) \right] \right)^2 \right\} \quad (25)
\]
Determination of $K(\theta)$

- Now substituting (21) into (25), we obtain

\[
I_\theta = E_\theta \left\{ K^2(\theta) \left( \hat{\theta}(Y) - E_\theta \left\{ \hat{\theta}(Y) \right\} \right)^2 \right\}
\]
\[
= K^2(\theta) E_\theta \left\{ \left( \hat{\theta}(Y) - E_\theta \left\{ \hat{\theta}(Y) \right\} \right)^2 \right\}
\]
\[
= K^2(\theta) \text{Var}_\theta \left[ \hat{\theta}(Y) \right]
\]
\[
= K^2(\theta) \frac{\left( \frac{\partial}{\partial \theta} \left[ E_\theta \left\{ \hat{\theta}(Y) \right\} \right] \right)^2}{I(\theta)} \quad \text{(from (24))}
\]

\[
\therefore K^2(\theta) = \frac{I^2_\theta}{\left( \frac{\partial}{\partial \theta} \left[ E_\theta \left\{ \hat{\theta}(Y) \right\} \right] \right)^2}
\]

\[
\implies K(\theta) = \frac{I_\theta}{\left( \frac{\partial}{\partial \theta} \left[ E_\theta \left\{ \hat{\theta}(Y) \right\} \right] \right)}.
\]

(26)
Necessary Condition to Achieve CRLB

- Recall that for the case when the estimator is unbiased, the information inequality reduces to CRLB inequality.
- Therefore for an unbiased estimator \( \hat{\theta}(y) \) to achieve CRLB, we can derive conditions similar to (22) and (23) respectively by setting \( f(\sigma) = \sigma \). Accordingly, we obtain

\[
\log \left( \frac{p_{\theta}(y)}{p_{a}(y)} \right) = \int_{a}^{\theta} K(\sigma) \left( \hat{\theta}(y) - \sigma \right) d\sigma, \tag{27}
\]

where \( K(\sigma) \) can be computed from (26) by noting that \( E_{\theta} \left\{ \hat{\theta}(Y) \right\} = \theta \). Thus

\[
K(\sigma) = I_{\sigma}. \tag{28}
\]

- Using (28) in (27), we obtain

\[
\log \left( \frac{p_{\theta}(y)}{p_{a}(y)} \right) = \int_{a}^{\theta} I_{\sigma} \left( \hat{\theta}(y) - \sigma \right) d\sigma. \tag{29}
\]
Note on Necessary Condition to Achieve equality in Information Inequality

- **Note:**
  
i. The condition in (22), to be satisfied by any estimator to achieve lower bound in information inequality (CRLB) is only a necessary condition. They are not sufficient (i.e., if some arbitrary estimator $\hat{\theta}(y)$ satisfies (22) then it may not achieve the lower bound in information inequality. Likewise, if some unbiased estimator satisfies (29), then it need not achieve CRLB).
Likelihood Equation

• We now come back to ML estimation...
• Recall that maximizing the likelihood equation is same as maximizing $\log (p_\theta(y))$, therefore we can find the ML estimate as the one that maximizes $\log (p_\theta(y))$.
• Thus a necessary condition for the ML estimate to satisfy is (assuming $p_\theta(y)$ is differentiable),

$$\frac{\partial}{\partial \theta} [\log (p_\theta(y))] \bigg|_{\theta=\hat{\theta}_{\text{ML}}(y)} = 0.$$ (30)

• The (30) above is known as the Likelihood Equation.
• Note that (30) is only a necessary condition for an ML estimate to satisfy. It is not sufficient i.e., if some arbitrary $\theta$ satisfies the likelihood equation, then it need not be an ML estimate.
• But solutions to the likelihood equation in (30) still have useful properties, even though they not are really maxima of $p_\theta(y)$.
Properties of Solutions to Likelihood Equation

- Recall from our earlier discussion that for an unbiased estimator $\hat{\theta}$ to achieve CRLB, then it should satisfy

$$\log (p_\theta(y)) = \int_\sigma^\theta I_\sigma \left( \hat{\theta}(y) - \sigma \right) d\sigma + h(y),$$  \hspace{1cm} (31)

where we define $h(y) = \log (p_a(y))$.

- Differentiating (31) w.r.t. $\theta$ and evaluating at $\theta = \hat{\theta}_{\text{ML}}(y)$ we obtain:

$$\frac{\partial}{\partial \theta} \left[ \log (p_\theta(y)) \right] |_{\theta=\hat{\theta}_{\text{ML}}(y)} = I_\theta \left( \hat{\theta}(y) - \theta \right) |_{\theta=\hat{\theta}_{\text{ML}}(y)} = 0$$

$$\Rightarrow \hat{\theta}_{\text{ML}}(y) = \hat{\theta}(y).$$  \hspace{1cm} (32)

- Thus, if $\hat{\theta}$ achieves the CRLB, then it is the solution to the likelihood equation.
Properties of Solutions to Likelihood Equation (ctd...)

- In other words being a solution to the likelihood equation is indeed a necessary condition to achieve the CRLB.
  - However, it is not sufficient.
  - i.e., it is not always true that solutions to the likelihood equation will achieve the CRLB.
  - In particular, a solution to the likelihood equation may not even be unbiased.

- However, if \( \log(p_\theta(y)) \) has the form of (31), then the solution to the likelihood equation will achieve CRLB.

- When the solution to the likelihood equation is unbiased but does not achieve CRLB, then there may be other unbiased estimators that have smaller variance than \( \hat{\theta}_{ML}(y) \).

- Thus the solution to the likelihood equation can only sometimes be an MVUE.
Example 1: ML estimation of the Parameter of an Exponential Distribution

- Suppose that $\Gamma = \mathbb{R}^n$ and $\Lambda = (0, \infty)$ and $Y_1, \ldots, Y_n$ are i.i.d. exponential random variables with parameter $\theta$, i.e., $p_{\theta}(y) = \prod_{k=1}^{n} f_{\theta}(y_k)$ with

$$f_{\theta}(y_k) = \theta e^{-\theta y_k}.$$  \hfill (33)

- Accordingly

$$p_{\theta}(y) = \theta^n e^{-\theta \sum_{k=1}^{n} y_k} = \theta^n e^{-n \theta \overline{y}},$$  \hfill (34)

where we define $\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k$. 

©Dr. Laxminarayana Pillutla
Example 1: ML estimation of the Parameter of an Exponential Distribution (ctd...)

- Using the likelihood equation we can compute $\hat{\theta}_{\text{ML}}(y)$ as follows:

\[
\frac{\partial}{\partial \theta} \left[ \log (p_{\theta}(y)) \right] |_{\theta = \hat{\theta}_{\text{ML}}(y)} = 0
\]

\[\implies \frac{\partial}{\partial \theta} \left[ n \log(\theta) - n\theta \bar{y} \right] |_{\theta = \hat{\theta}_{\text{ML}}(y)} = 0
\]

\[\implies n \frac{n}{\hat{\theta}_{\text{ML}}(y)} = 0
\]

\[\implies \hat{\theta}_{\text{ML}}(y) = \frac{1}{\bar{y}}. \quad (35)
\]

- Since $\frac{\partial^2}{\partial \theta^2} \left[ \log (p_{\theta}(y)) \right] = \frac{-n}{\theta^2} < 0$, therefore $\hat{\theta}_{\text{ML}}(y)$ in (35) is a unique maximum.
Example 1: ML estimation of the Parameter of an Exponential Distribution (ctd...)

• Note that

\[ E_{\theta} \{ y_k \} = \int_{0}^{\infty} y_k f_{\theta}(y_k) = \int_{0}^{\infty} y_k \theta e^{-\theta y_k} = \frac{1}{\theta}. \quad (36) \]

• Also

\[ E_{\theta} \{ \bar{Y} \} = E_{\theta} \left\{ \frac{1}{n} \sum_{k=1}^{n} y_k \right\} \\
= \frac{1}{n} \sum_{k=1}^{n} E_{\theta} \{ y_k \} \\
= \frac{1}{n} \sum_{k=1}^{n} \frac{1}{\theta} \quad \text{(from (36))} = \frac{1}{\theta}. \quad (37) \]
Example 1: ML estimation of the Parameter of an Exponential Distribution (ctd...)

- From the calculations in the previous slide, it makes sense to estimate $\theta$ as $\frac{1}{\bar{y}}$.
- Also notice that as $n \to \infty$, $\bar{Y} = \frac{1}{n} \sum_{k=1}^{n} Y_k \overset{\text{a.s.}}{\to} E_{\theta} \{ Y_k \} = \frac{1}{\theta}$, by virtue of strong law of large numbers (SLLN). This of course implies (weak) convergence in probability i.e.,

$$P_{\theta} \left\{ \left| \bar{Y} - \frac{1}{\theta} \right| > \varepsilon \right\} \to 0$$

$$\Rightarrow \bar{Y} \overset{P_{\theta}}{\to} \frac{1}{\theta}$$

or $$\frac{1}{\bar{Y}} \overset{P_{\theta}}{\to} \theta.$$ (38)

i.e., the ML estimator of $\theta$ in (35) converges in probability under $P_{\theta}$ to $\frac{1}{\bar{Y}}$.
- This property of ML estimator is known as consistency.
- Indeed the consistency property of ML estimator for i.i.d. observations can be proved in a very general context (see Proposition IV.D.1 in thePoor’s book).
Example 1: ML estimation of the Parameter of an Exponential Distribution (ctd...)

- Next we check if the ML estimate in (35) is unbiased (i.e., we verify if \( E_\theta \left\{ \hat{\theta}_{\text{ML}}(Y) \right\} \rightarrow \theta \)). To this end, we compute \( E_\theta \left\{ \hat{\theta}_{\text{ML}}(Y) \right\} \rightarrow \theta \) as follows:

\[
E_\theta \left\{ \hat{\theta}_{\text{ML}}(Y) \right\} = E_\theta \left\{ \frac{1}{Y} \right\}. \quad (39)
\]

- To compute (39) we need the pdf of \( \bar{Y} \) denoted as \( p_{\bar{Y}}(\bar{y}) \). We compute this by computing the characteristic function of \( \bar{Y} \) and then taking inverse Fourier transform of it.

  - Recall from the lectures on pdf computation of one function of two random variables: “If \( X \) and \( Y \) are two independent r.v.'s then the pdf of \( X + Y \) is given by the convolution of their respective pdfs”.
  - In the current case \( \bar{Y} \) is nothing but the scaled sum of independent r.v.'s \( Y_k (k = 1, \cdots, n) \), hence the motivation to adopt the procedure based on Characteristic Functions.
Example 1: ML estimation of the Parameter of an Exponential Distribution (ctd...)

- Accordingly, the characteristic function of \( \bar{Y} \) denoted as \( \Phi_{\bar{Y}}(\omega) \) is given by:

\[
\Phi_{\bar{Y}}(\omega) = E_\theta \left[ e^{j\omega \bar{Y}} \right] \\
= E_\theta \left[ e^{j\omega \frac{1}{n} \sum_{k=1}^{n} Y_k} \right] \\
= \prod_{k=1}^{n} E_\theta \left\{ e^{j\omega \frac{Y_k}{n}} \right\} \quad \text{(since } Y_k \text{'s are independent)} \\
= \left[ E_\theta \left\{ e^{j\omega \frac{Y_k}{n}} \right\} \right]^n \quad \text{(since } Y_k \text{'s have common exponential distribution)} .
\]

(40)
Example 1: ML estimation of the Parameter of an Exponential Distribution (ctd...)

- Therefore, to obtain $\Phi_Y(\omega)$ we need to compute $E_\theta \left\{ e^{\frac{j\omega Y_k}{n}} \right\}$:

$$E_\theta \left\{ e^{\frac{j\omega Y_k}{n}} \right\} = \int_0^\infty e^{\frac{j\omega Y_k}{n}} \theta e^{-\theta y_k} dy_k$$

$$= \frac{1}{\left(1 - \frac{j\omega}{n\theta}\right)^n} \quad (41)$$

- Using (41) we can compute $\Phi_Y(\omega)$ in (40) to be

$$\Phi_Y(\omega) = \frac{1}{\left(1 - \frac{j\omega}{n\theta}\right)^n} \quad (42)$$
Example 1: ML estimation of the Parameter of an Exponential Distribution (ctd...)

- The characteristic function in (42) correspond to that of an Erlang-$n$ r.v. whose pdf (see Table at the end of Chapter 5 of Papoulis book) is given by

\[
p_{\overline{Y}}(\overline{y}) = \begin{cases} 
\frac{(n\theta)^n}{(n-1)!}(\overline{y})^{n-1}e^{-n\theta \overline{y}} & \text{if } \overline{y} \geq 0 \\
0 & \text{otherwise}
\end{cases} \quad (43)
\]

- Now $E_\theta \left\{ \frac{1}{\overline{Y}} \right\}$ in (39) can be computed as follows:

\[
E_\theta \left\{ \frac{1}{\overline{Y}} \right\} = \int_{-\infty}^{+\infty} \frac{1}{\overline{y}} p_{\overline{Y}}(\overline{y}) d\overline{y} = \int_{0}^{+\infty} \frac{1}{\overline{y}} \frac{(n\theta)^n}{\overline{y} (n-1)!}(\overline{y})^{n-1}e^{-n\theta \overline{y}} d\overline{y} = \cdots = \frac{n\theta}{(n-1)} \quad (44)
\]
Example 1: ML estimation of the Parameter of an Exponential Distribution (ctd...)

• From (44), it can be seen that

\[
E_{\theta} \left\{ \hat{\theta}_{\text{ML}}(Y) \right\} = E_{\theta} \left\{ \frac{1}{\bar{Y}} \right\} = \frac{n\theta}{(n - 1)} \quad (45)
\]

\[
\neq \theta
\]

• Thus the ML estimate in (35) is biased for any finite \(n\).

  ○ The quantity \( E_{\theta} \left\{ \hat{\theta}(Y) \right\} - \theta \) is in general called as the bias of the estimate \( \hat{\theta}(Y) \).

• However, it can be seen that

\[
\lim_{n \to \infty} E_{\theta} \left\{ \hat{\theta}_{\text{ML}}(Y) \right\} = \lim_{n \to \infty} \frac{n\theta}{(n - 1)} = \theta, \quad (46)
\]

i.e., \( \hat{\theta}_{\text{ML}} \) is asymptotically unbiased.
Example 1: ML estimation of the Parameter of an Exponential Distribution (ctd...)

- The property of asymptotic unbiasedness observed in the previous slide is indeed true under very general conditions for i.i.d. observations.
- Also note from (35) that

\[
\frac{\partial}{\partial \theta} \left[ \log \left( p_\theta(y) \right) \right] = n \left( \frac{1}{\theta} - \frac{1}{\hat{\theta}_{ML}(y)} \right),
\]

which is not of the form \( K(\theta) \left( \hat{\theta}_{ML}(y) - f(\theta) \right) \).

- Thus the ML estimate of \( \theta \) in (35) will not achieve the lower bound in information inequality.
Asymptotic Efficiency of ML Estimate

- Here, we demonstrate that the ML estimate in (35) is asymptotically efficient i.e., we show that as \( n \to \infty \), the \( \text{Var}_\theta \left[ \hat{\theta}_{ML}(Y) \right] \) approaches CRLB.
- To this end we first compute Fisher information \( I_\theta \) using the formula:

\[
I_\theta = -E_\theta \left\{ \frac{\partial^2}{\partial \theta^2} \left[ \log (p_\theta(y)) \right] \right\} \\
= -E_\theta \left\{ -\frac{n}{\theta^2} \right\} = \frac{n}{\theta^2}.
\]

(48)

- Therefore, the CRLB for any unbiased estimator is equal to

\[
\text{CRLB} = \frac{1}{I_\theta} = \frac{\theta^2}{n}.
\]

(49)
Asymptotic Efficiency of ML Estimate (ctd...)

- The variance of the ML estimator of $\theta$ in (35) is given by

$$\text{Var}_\theta \left[ \hat{\theta}_{\text{ML}}(Y) \right] = E\theta \left\{ \left( \hat{\theta}_{\text{ML}}(Y) \right)^2 \right\} - \left( E\theta \left\{ \hat{\theta}_{\text{ML}}(Y) \right\} \right)^2 \quad (50)$$

- In (44) we computed

$$E\theta \left\{ \hat{\theta}_{\text{ML}}(Y) \right\} = \frac{n\theta}{n - 1} \quad (51)$$

- The quantity $E\theta \left\{ \left( \hat{\theta}_{\text{ML}}(Y) \right)^2 \right\}$ can be computed as follows:

$$E\theta \left\{ \left( \hat{\theta}_{\text{ML}}(Y) \right)^2 \right\} = E\theta \left\{ \frac{1}{Y^2} \right\} = \int_{-\infty}^{+\infty} \frac{1}{y^2} p_Y(y) dy = \int_{0}^{+\infty} \frac{1}{y^2} \frac{(n\theta)^n}{(n - 1)!} (\frac{y}{n\theta})^{n-1} e^{-\frac{n\theta}{y}}$$
Asymptotic Efficiency of ML Estimate (ctd...)

(continued from previous slide ...)

\[ E_\theta \left\{ \left( \hat{\theta}_{ML}(\overline{Y}) \right)^2 \right\} = \frac{n^2 \theta^2}{(n-1)(n-2)} \quad (52) \]

- Using (51) and (52) in (50) we obtain

\[ \text{Var}_\theta \left[ \hat{\theta}_{ML}(\overline{Y}) \right] = \frac{n^2 \theta^2}{(n-1)^2(n-2)} \quad (53) \]

- From (49) and (53) we note that

\[ \lim_{n \to \infty} \frac{\text{Var}_\theta \left[ \hat{\theta}_{ML}(\overline{Y}) \right]}{\text{CRLB}} = \lim_{n \to \infty} \text{Var}_\theta \left[ \hat{\theta}_{ML}(\overline{Y}) \right] I_\theta \]

\[ = \lim_{n \to \infty} \frac{n^2 \theta^2}{(n-1)^2(n-2)} \times \frac{n}{\theta^2} \]
Asymptotic Efficiency of ML Estimate (ctd...)

(continued from previous slide ...)

\[
\lim_{n \to \infty} \frac{\text{Var}_\theta \left[ \hat{\theta}_{\text{ML}}(Y) \right]}{\text{CRLB}} = \lim_{n \to \infty} \frac{n^3}{(n-1)^2(n-2)} = 1 \quad (54)
\]

- From (54) we can conclude that the variance of \( \hat{\theta}_{\text{ML}}(Y) \) given by (35) equals CRLB asymptotically.
  - This property is known as asymptotic efficiency.
- Again the asymptotic efficiency property of an ML estimator can be proved under general conditions for i.i.d. observations.