Maximum Likelihood Estimation (MLE)

1) MVUE might be difficult to synthesize
   → might not be efficient ($\neq$ CRLB) & hence CRLB equality condition cannot be used.
   → Synthesis using "sufficient statistics" we will cover later (has its own problems)

2) MLE works always → nice asymptotic performance guarantees (e.g. min variance, unbiasedness, consistency)

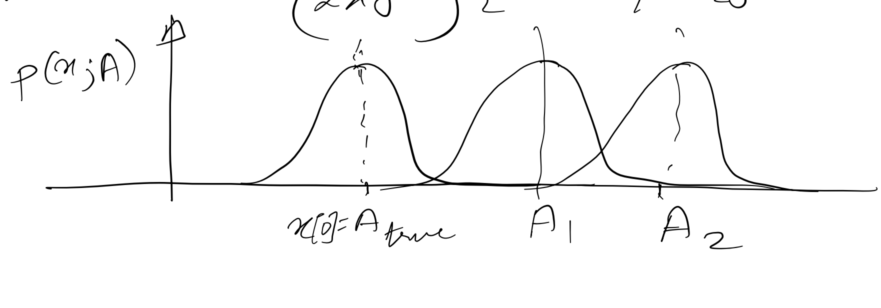
Let $p(x;\theta)$ $(x \in \mathbb{R}^N ; \theta \in S)$ be the joint pdf / pmf of $\{x[0], \cdots, x[N-1]\}$. Then the likelihood fn. is defined as $p(x = x_0 ; \theta) : S \to \mathbb{R}$.

Log-likelihood fn. : $\ln p(x = x_0 ; \theta)$

If in the Gaussian case; $x = x_0$ was observed:

$x[0] = A + w[0]$ $\qquad$ $w[0] \sim$ WGN

$$p(x;A) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(x - A)^2\right]$$



So MLE $\hat{\theta} = \arg\max_{\theta \in S} \ln p(x = x_0; \theta)$

For the example above: $\frac{\partial \ln p(x;A)}{\partial A} = \frac{1}{\sigma^2}(x[0] - A) = 0$

$\Rightarrow \hat{A}_{MLE} = x[0]$.

Similarly for $x[n] = A + w[n]$ $n = 0, 1, \cdots, N-1$

$\hookrightarrow$ WGN

$\hat{A}_{MLE} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$

In the above examples $\hat{A}_{MVUE} = \hat{A}_{MLE}$ can be verified - since $\hat{A}_{MVUE}$ could be calculated from CRLB.

However MLE is easy to find even if MVUE is hard.

Ex:

$x[n] = A + w[n]$ $n = 0, \cdots, N-1$ & $\underline{A > 0}$

& $w[n] \sim N(0, \underline{A})$

$$p(x;A) = \frac{1}{(2\pi A)^{N/2}} \exp\left[-\frac{1}{2A}\sum_{n=0}^{N-1}(x[n] - A)^2\right]$$

$$\frac{\partial \ln p(x;A)}{\partial A} = -\frac{N}{2A} + \frac{1}{A}\sum_{n=0}^{N-1}(x[n] - A) + \frac{1}{2A^2}\sum_{n=0}^{N-1}(x[n] - A)^2$$

Check "reg" condition holds so CRLB applicable but it is difficult to see whether

$$\frac{\partial \ln p(x, A)}{\partial A} = I^{-1}(A)(\hat{A} - A) \text{ ??}$$

In any case, CRLB can be computed

$$\mathrm{var}(\hat{A}) \geq \frac{A^2}{N(A+\frac{1}{2})}.$$

: But no MVU estimator

<u>Now consider MLE:</u>

$$\frac{\partial \ln p(x;A)}{\partial A} = 0 \Rightarrow \hat{A}^2 + \hat{A} - \frac{1}{N}\sum_{n=0}^{N-1} x^2[n] = 0$$

Solving & using $\hat{A} > 0 \rightarrow \hat{A}_{MLE} = -\frac{1}{2} + \underbrace{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}}$

Q. So how good is this MLE? Is it unbiased? MVUE? Is it efficient? We did not (could not) even compute $\mathrm{var}(\hat{A}_{MLE})$!!

<u>FACT</u>: If an efficient estimator exists, the MLE will produce it. (Exercise)

<u>Proof</u>: $\frac{\partial \ln p(x;\theta)}{\partial \theta} = 0$ gives MLE.

But by CRLB an efficient estimator will satisfy $\frac{\partial \ln p(x;\theta)}{\partial \theta} = I^{-1}(\theta)\,[\hat{\theta} - \theta]$ — (1)

clearly (1) $= 0$ will produce $\hat{\theta} = \theta$.

\# <u>Asymptotic guarantees later.</u>

MLE works easily for vector parameters:

$$\theta = [\theta_1 \cdots \theta_p]; \quad \left[\frac{\partial \ln p(x;\theta)}{\partial \theta_1} \cdots\cdots \frac{\partial \ln p(x;\theta)}{\partial \theta_p}\right] = 0_{1\times p}$$

Ex: $x[n] = A + w[n] \qquad n = 0, \dots, N-1$

$\hookrightarrow N(0, \sigma^2)$ & σ^2 unknown.

$\theta = [A \ \sigma^2]^T$.

1. $$\frac{\partial \ln p(x;\theta)}{\partial A} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] + A)$$

2. $$\frac{\partial \ln p(x;\theta)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2$$

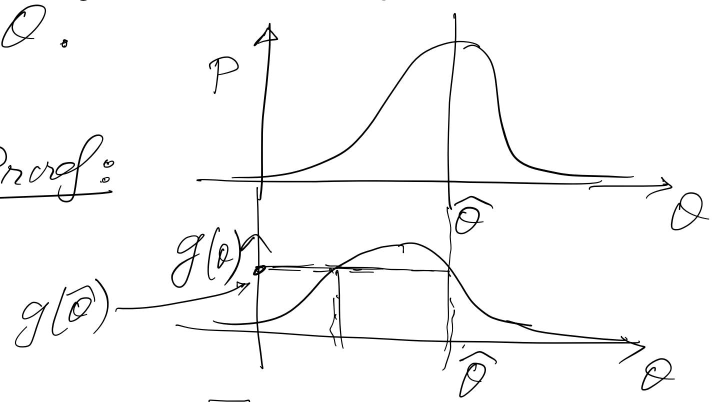
For (1) $= 0$, $\hat{A}_{MLE} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$

For (2) $= 0$, $\hat{\sigma}^2 = \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \hat{A}_{MLE})^2$

Invariance Property of MLE

FACT: The MLE of the parameter $\alpha = g(\theta)$ where pdf is $p(x;\theta)$ (parameterized by θ) is given by $\hat{\alpha} = g(\hat{\theta})$ where $\hat{\theta}$ is MLE of θ.

Sketch of Proof:



$$\left.\begin{array}{l} \max_{\alpha} \ p(x;\theta) \\ \text{where } \alpha = g(\theta) \end{array}\right] \equiv \min_{\alpha,\theta} \overbrace{\{p(x;\theta) + \lambda[\alpha - g(\theta)]\}}^{H(\alpha,\theta)}$$

$$\frac{\partial H}{\partial \alpha} = 0; \quad \frac{\partial H}{\partial \theta} = 0$$

yields $\lambda = 0$, $\frac{\partial p}{\partial \theta} = 0 \rightarrow \hat{\theta}$, $g(\hat{\theta})$ are the solutions.

Ex: $x[n] \sim N(0, \sigma^2)$ $\rightarrow \sigma^2$ unknown $\rightarrow$ Power in dB to be estimated.

$$P = 10 \log_{10} \sigma^2$$

$$p(x, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n]\right]$$

$$\frac{\partial \ln p(x; \sigma^2)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=0}^{N-1} x^2[n] = 0$$

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

. From invariance property:

$$\hat{P} = 10 \log_{10} \hat{\sigma}^2 = 10 \log_{10} \frac{1}{N} \sum_{n=0}^{N-1} x^2[n]$$

Recall:

Defn: A seq of r.v.s $X_1, X_2, \ldots$ converges in prob. to a r.v. X, if for every $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|X_n - X| \geq \varepsilon) = 0$$

or

$$\lim_{n \to \infty} P(|X_n - X| < \varepsilon) = 1$$

They converge almost surely to a r.v. X if for every $\varepsilon > 0$,

$$P\left(\lim_{n \to \infty} |X_n - X| < \varepsilon\right) = 1$$

Then converge in distribution to a r.v. X if $\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$ at all points where $F_X(x)$ is continuous.

FACT: A.S.C $\Rightarrow$ Conv. in prob $\Rightarrow$ Conv. in distribution

FACT 1: Weak Law of Large Nos:
Let $X_1, X_2 \dots$ be iid r.v. with $EX_i = \mu$ and $\mathrm{Var}\, X_i = \sigma^2 < \infty$. Define $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then for every $\varepsilon > 0$

$$\lim_{n \to \infty} P(|\overline{X}_n - \mu| < \varepsilon) = 1$$

i.e. $\overline{X}_n \xrightarrow{P} \mu$.

FACT 2: Strong law of large nos:
Let $X_1, X_2 \dots$ be iid r.v.s with $EX_i = \mu$ and $\mathrm{Var}\, X_i = \sigma^2 < \infty$, define $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$
Then for every $\varepsilon > 0$,

$$P\left(\lim_{n \to \infty} |\overline{X}_n - \mu| < \varepsilon\right) = 1$$

i.e. $\overline{X}_n \xrightarrow{a.s} \mu$.

FACT 2 $\Rightarrow$ FACT 1
(FACT 1 is still stated separately mainly for historical reasons & easy proof)

Central Limit Theorem: Let $X_1, X_2, \cdots$ be a seq. of iid r.v.s with $E X_i = \mu < \infty$ and $0 < Var X_i = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Let $F_n(x)$ denote the cdf of $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$. Then for any $x \in (-\infty, \infty)$,

$$\lim_{n \to \infty} F_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

i.e. $F_n(x) \xrightarrow{d} N(0, 1)$

Defn: A sequence of estimators $\hat{\theta}_N = \theta_N(x[0], \cdots, x[N-1])$ is a consistent sequence of estimators of θ if for every $\varepsilon > 0$, & $\forall \theta$,

$$\lim_{N \to \infty} P_\theta(|\hat{\theta}_N(x) - \theta| < \varepsilon) = 1$$

$$\equiv \lim_{N \to \infty} P_\theta(|\hat{\theta}_N(x) - \theta| \geq \varepsilon) = 0$$

$\hat{\theta}_N(x)$ converges in prob. to θ (true) for each θ.

Ex: $x[n] = A + w[n]$ & $w[n] \sim N(0, 1)$

Consider the seq. of sample means

$\bar{x}_N = \frac{1}{N} \sum x[i]$ $\qquad \bar{x}_N \sim N\left(A, \frac{1}{N}\right)$

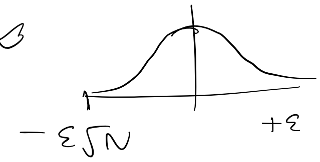
Hence $\quad P_A\left(|\bar{x}_N - A| < \varepsilon\right)$

$$= \int_{A-\varepsilon}^{A+\varepsilon} \left(\frac{N}{2\pi}\right)^{1/2} e^{-(N/2)(\bar{x}_N - A)^2} \, d\bar{x}_N = \int_{-\varepsilon}^{+\varepsilon} \left(\frac{N}{2\pi}\right)^{1/2} e^{-\left(\frac{N}{2}\right)(y^2)} \, dy$$

$(y = \bar{x}_N - \theta)$

$$= \int_{-\varepsilon\sqrt{N}}^{+\varepsilon\sqrt{N}} \left(\frac{1}{2\pi}\right)^{1/2} e^{-\left(\frac{1}{2}\right)t^2} \, dt = P\left(-\varepsilon\sqrt{N} < Z < \varepsilon\sqrt{N}\right) \quad Z \sim N(0,$$

$(t = y\sqrt{n})$ $\qquad \to 1$ as $N \to \infty$



Hence $\bar{x}_N$ is consistent estimator for A.

Usually a sufficient condition is used to check:

Recall, from Chebychev's Inequality,

$$P_\theta\left(|\hat{\theta}_N - \theta| \geq \varepsilon\right) \leq \frac{E_\theta\left[(\hat{\theta}_N - \theta)^2\right]}{\varepsilon^2}$$

But $\quad E_\theta\left[(\hat{\theta}_N - \theta)^2\right] = \mathrm{Var}_\theta\, \hat{\theta}_N + \left[\mathrm{Bias}_\theta\, \hat{\theta}_N\right]^2$

Hence if $\quad \mathrm{Var} \to 0$ & $\mathrm{Bias} \to 0$ as $N \to \infty$

then $\quad \hat{\theta}_N \xrightarrow{P} \theta$

Thm: If $\hat{\theta}_n$ is a sequence of estimators of a parameter θ satisfying

1> $\lim\limits_{N \to \infty} Var_\theta(\hat{\theta}_N) = 0$

2> $\lim\limits_{N \to \infty} Bias_\theta(\hat{\theta}_N) = 0$

$\Bigg\}\ \forall\theta$ (1)

then $\hat{\theta}_N$ is a consistent sequence of estimators of θ.

Ex: Easy to verify for the example above

$E_\theta \bar{x}_N = \theta \qquad Var_\theta(\bar{x}_N) = \frac{1}{N}$

Consistency of MLEs:

Thm: Let $x[1] \ldots x[n]$ be iid $p(x[i];\theta)$ and let $L(\theta; x) = \prod_{i=1}^{n} p(x[i];\theta)$ be the likelihood fn. Let $\hat{\theta}_n$ denote the MLE of θ. Under "some regularity conditions" on $p(x[i];\theta)$, for every $\varepsilon > 0$ & $\forall\theta$,

$$\lim_{n \to \infty} P_\theta(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0$$

i.e. $\hat{\theta}_n$ is a consistent estimator of θ

Outline of proof:

Let θ_0 be the true parameter value.

Let $L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln p(x[i]; \theta)$ ← depends on sample

& $L(\theta) = \int_X [\ln p(x;\theta)]\, p(x;\theta_0)\, dx$

$= E_{\theta_0}[\ln p(x;\theta)]$

Then by L.L.N $\forall \theta$,

$$L_n(\theta) \longrightarrow E_{\theta_0}[\ln p(x;\theta)] = L(\theta)$$

does not depend on sample

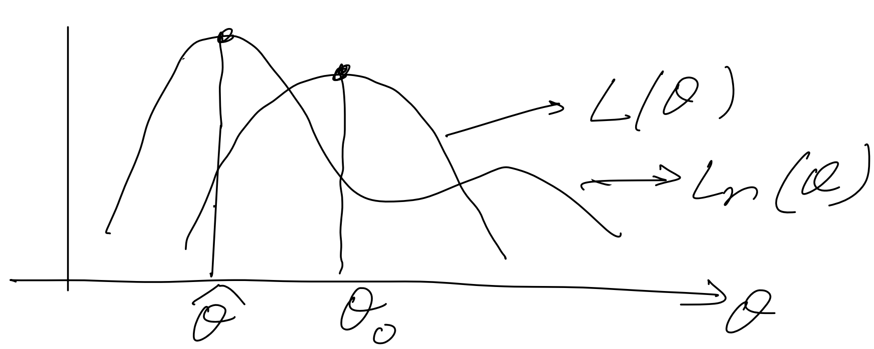
Claim: $L(\theta) \leq L(\theta_0)$ $\forall \theta$

[Infact $P(L(\theta) < L(\theta_0)) = 1$]

Proof: Exercise

Then: proof follow from:

1. $\hat{\theta}_n$ is the maximizer of $L_n(\theta)$ (MLE)
2. θ_0 is the maximizer of $L(\theta)$ (by claim
3. $\forall \theta$, $L_n(\theta) \to L(\theta)$ by LLN.



Q. What about asymptotic variance of the estimator?

Def$^{\underline{n}}$: For an estimator $\hat{\theta}_n$ of θ, if $k_n(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2)$, where $\{k_n\}$ is a sequence of constants, then σ^2 is called the asymptotic variance of $\hat{\theta}_n$.

Example: For n iid $N(\mu, \sigma^2)$ observations $X_1, \dots, X_n$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Note: We need this $\sqrt{n}$, otherwise variance goes to 0.

Def$^{\underline{n}}$: A sequence of estimators $\hat{\theta}_n$ is asymptotically efficient for θ if $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$ where $I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2} \ln p(x;\theta)\right]$

(i.e. asymptotically $\hat{\theta}_n$ is unbiased
& " $n \operatorname{Var}(\hat{\theta}_n)$ achieves CRLB)

$\hookrightarrow$ Why? Recall $\operatorname{Var}(\hat{\theta}_n) \to 0$.

Thm: Let $x[1], \cdots, x[n]$ be iid $p(x[i]; \theta_0)$, $\hat{\theta}$ be MLE for θ_0. Under "reg. conditions" on $p(x; \theta_0)$

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N(0, I^{-1}(\theta_0)) \quad \text{where}$$

$I^{-1}(\theta_0)$ is the CRLB. $\Leftrightarrow$ $\hat{\theta}$ is consistent & asymptotically efficient estimator of θ_0.

Proof: Let $L(\theta) = \sum_{i=1}^{n} \ln p(x_i; \theta)$ be the likelihood function & let θ_0 be the true value of θ.

Taylor-series of $\frac{\partial L}{\partial \theta}$ around θ_0:

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial \theta}\Big|_{\theta_0} + (\theta - \theta_0) \frac{\partial^2 L}{\partial \theta^2}\Big|_{\theta_0}$$

For finding the MLE: $\hat{\theta}$, we put

$$\frac{\partial L}{\partial \theta} = 0 \Rightarrow (\hat{\theta} - \theta_0) = \frac{-\frac{\partial L}{\partial \theta}\Big|_{\theta_0}}{\frac{\partial^2 L}{\partial \theta^2}\Big|_{\theta_0}}$$

or

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{-\frac{1}{\sqrt{n}} \frac{\partial L}{\partial \theta}\Big|_{\theta_0}}{\frac{1}{n} \frac{\partial^2 L}{\partial \theta^2}\Big|_{\theta_0}}$$

Now:

$$\left.\begin{aligned} -\frac{1}{\sqrt{n}} \frac{\partial L}{\partial \theta}\Big|_{\theta_0} &\xrightarrow{d} N(0, I(\theta_0)) \\ \frac{1}{n} \frac{\partial^2 L}{\partial \theta^2}\Big|_{\theta_0} &\xrightarrow{P} I(\theta_0) \end{aligned}\right\} \text{Proof skipped}$$

Hence $\sqrt{n}\,(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, \dfrac{1}{I(\theta_0)}\right)$