I225E Statistical Signal Processing

9. Maximum Likelihood Estimation

MAWALIM and UNOKI

candylim@jaist.ac.jp and unoki@jaist.ac.jp

School of Information Science

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY,
HOKURIKU 1990

Maximum Likelihood Estimation

What if MVUE (minimum variance unbiased estimator) does not exist or unknown?

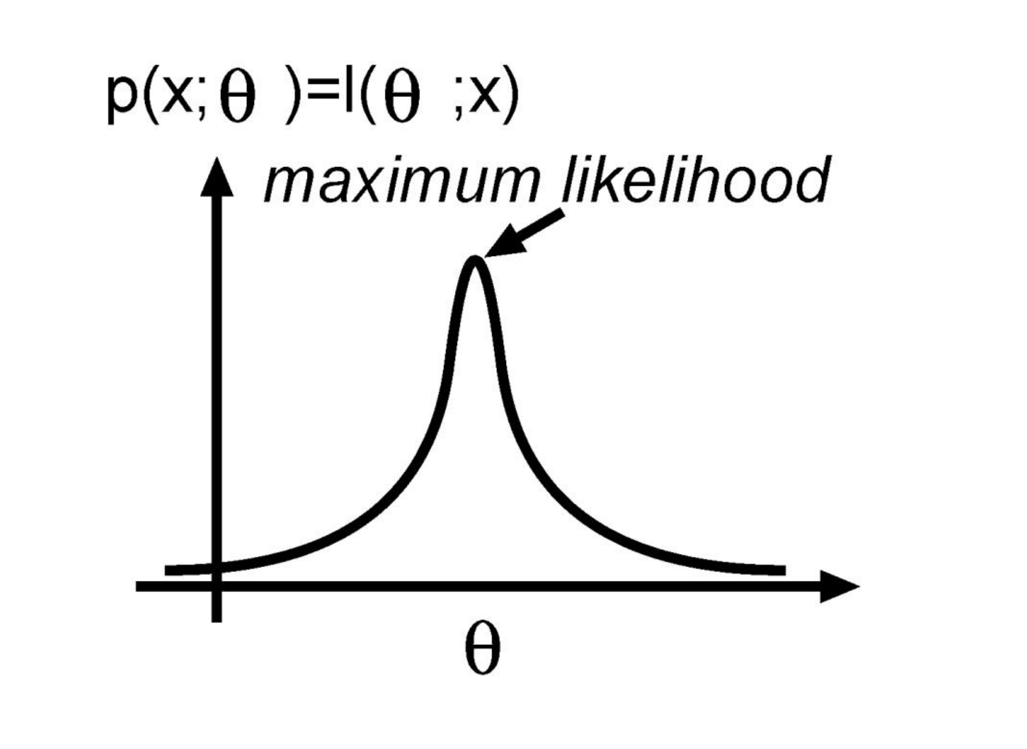
⟹ **Maximum Likelihood Estimation**

[Features]

1. Easy to implement
2. Optimal for large enough data records
3. Under certain conditions, asymptotically efficient
4. In other words, converges to MVUE

⟹ Applied to various practical problems.

Random variable $X \sim p(x;\theta)$ is observed. Viewing x as fixed and θ as variable, we call $l(\theta;x) = p(x;\theta)$ as the likelihood of θ (given x).



Definition

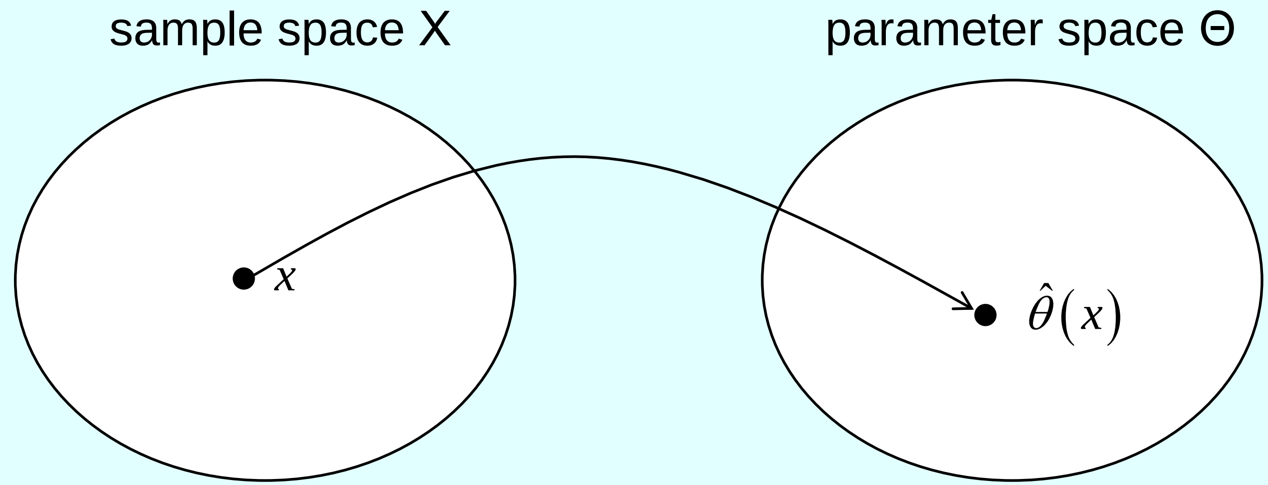
$\hat{\theta}$ is called ***maximum likelihood estimator*** if

$$\forall x, \quad l\left(\hat{\theta}; x\right) = \max_{\theta \in \Theta} l(\theta; x).$$

This is equivalent to $\hat{\theta}(x) = \arg\max_{\theta \in \Theta} l(\theta; x)$

Note:

MLE (maximum likelihood estimator) selects the value of θ such that the observed x corresponds to the most probable outcome. Likelihood can be viewed as a density function for θ conditioned on $X = x$. However, classical estimator views θ as nonrandom.



$$\hat{\theta}_{\mathrm{ML}}(x) = \arg\max_{\theta} P(x \mid \theta)$$

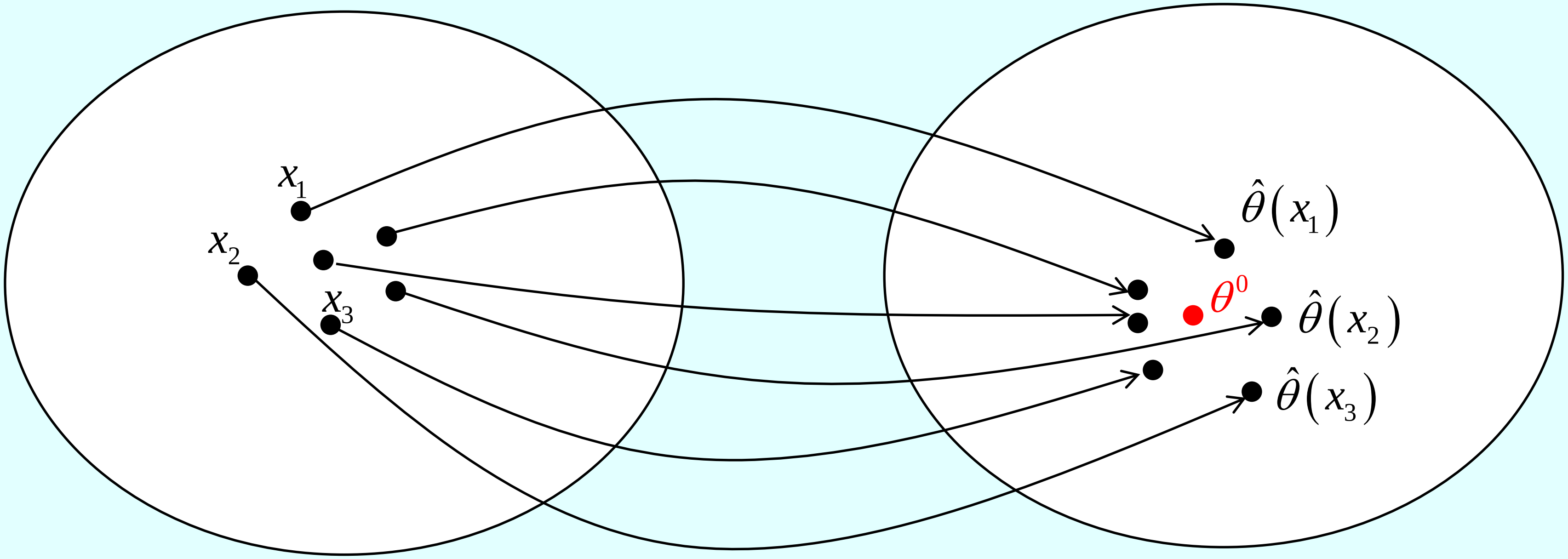
$$= \arg\max_{\theta} \log P(x \mid \theta)$$

ML is …

- Asymptotically unbiased (i.e., approaches to a true value).
- Asymptotically efficient (i.e., achieves minimum variance, CRLB).

sample space X

parameter space Θ



KL divergence between true density *p*(*x*) and parametrized density *q*(*x*|*θ*):

$$D_{\mathrm{KL}}\left[p(x);q(x|\theta)\right]=\int dx\,p(x)\log\frac{p(x)}{q(x|\theta)}$$

$$=\mathrm{E}\left[\log p(x)\right]-\mathrm{E}\left[\log q(x|\theta)\right]$$

Minimization of KL divergence $D_{\mathrm{KL}}\left[p(x);q(x|\theta)\right]$

⟷

Maximization of $\mathrm{E}\left[\log q(x|\theta)\right]$

Sampling approximation:

$$\mathrm{E}\left[\log q(x|\theta)\right]\approx\frac{1}{N}\sum_{i=1}^{N}\log q(x_i|\theta)$$

Sampling approximation:

$$\mathrm{E}\left[\log q\left(x \mid \hat{\theta}\right)\right]-\frac{1}{N}\sum_{i=1}^{N}\log q\left(x_i \mid \hat{\theta}\right)\approx-\left(\hat{\theta}-\theta^0\right)^{\mathrm{T}} E\left[\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\log q\left(x \mid \hat{\theta}\right)\right]\left(\hat{\theta}-\theta^0\right)$$

$$=\left(\hat{\theta}-\theta^0\right)^{\mathrm{T}} I\left(\hat{\theta}\right)\left(\hat{\theta}-\theta^0\right)$$

Fisher information:

$$I\left(\hat{\theta}\right)\equiv\mathrm{E}\left[\frac{\partial\log q\left(x \mid \hat{\theta}\right)}{\partial\theta}\frac{\partial\log q\left(x \mid \hat{\theta}\right)}{\partial\theta^{\mathrm{T}}}\right]=\mathrm{E}\left[-\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\log q\left(x \mid \hat{\theta}\right)\right]$$

In the limit of large samples (infinite *N*), the ML estimator is unbiased and efficient.

$$\hat{\theta} \sim \mathrm{N}\left(\theta^0,\frac{1}{N}I^{-1}\left(\hat{\theta}\right)\right)$$

ML is …

- Asymptotically unbiased (i.e., approaches to a true value).
- Asymptotically efficient (i.e., achieves minimum variance, CRLB).

Suppose a random variable $X \sim p(x;\theta)$, where θ is fixed but unknown. Assume that $p(x;\theta)$ satisfies the “regularity” condition:

$$\mathrm{E}\left[\frac{\partial}{\partial\theta}\log p(x|\theta)\right]=0,$$

where the expectation is with respect to $p(x;\theta)$. Then the variance of any unbiased estimator $\hat{\theta}$ satisfies

$$\mathrm{Var}\left[\hat{\theta}\right]\geq\frac{1}{I(\theta)}$$

Fisher information:

$$I(\theta)\equiv\mathrm{E}\left[\left(\frac{\partial\log p(x|\theta)}{\partial\theta}\right)^2\right]=\mathrm{E}\left[-\frac{\partial^2\log p(x|\theta)}{\partial\theta^2}\right]$$

Suppose a random variable $X \sim p(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is fixed but unknown. Assume that $p(x|\boldsymbol{\theta})$ satisfies the “regularity” condition:

$$\mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log p\left(x \mid \boldsymbol{\theta}\right)\right] = 0,$$

where the expectation is with respect to $p(x;\theta)$. Then the variance of any unbiased estimator $\widehat{\boldsymbol{\theta}}$ satisfies

$$\mathrm{Cov}\left[\hat{\boldsymbol{\theta}}\right] \geq \mathbf{I}^{-1}\left(\boldsymbol{\theta}\right)$$

Fisher information matrix:

$$\left\{\mathbf{I}\left(\boldsymbol{\theta}\right)\right\}_{ij} \equiv \mathrm{E}\left[\frac{\partial \log p\left(x \mid \boldsymbol{\theta}\right)}{\partial \theta_i} \frac{\partial \log p\left(x \mid \boldsymbol{\theta}\right)}{\partial \theta_j}\right] = \mathrm{E}\left[-\frac{\partial^2 \log p\left(x \mid \theta\right)}{\partial \theta_i \partial \theta_j}\right]$$

Computing the MLE

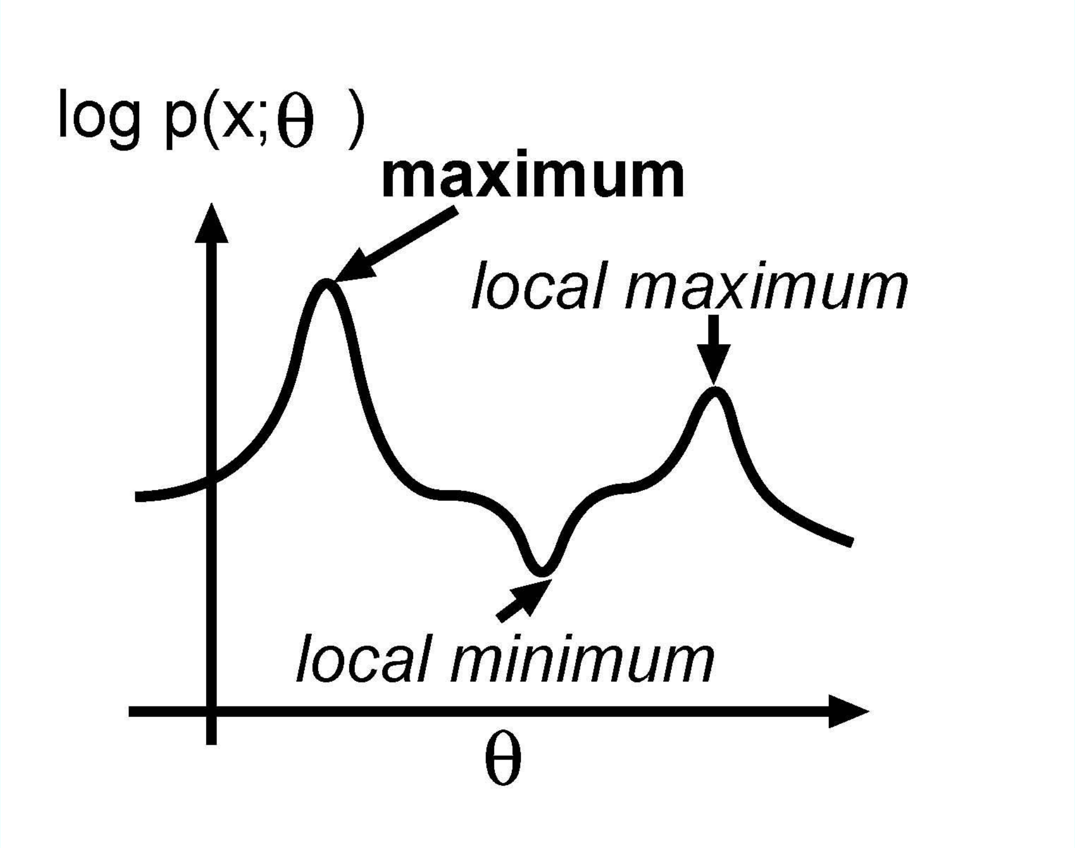
1. Since many models we work with have an exponential form, it is often convenient to maximize the log-likelihood $\ln l(\theta; x)$.

2. If the likelihood function is differentiable, $\hat{\theta}(x)$ is a solution of $\frac{\partial}{\partial\theta} \ln l(\theta; x) = 0$. We need to verify that such a solution is in fact a local max and not a local min or a saddle point.

$\Longrightarrow$ This can be checked whether the Hessian $\frac{\partial^2}{\partial\theta\partial\theta^T} \ln l(\theta; x)$ is negative semidefinite at $\hat{\theta}(x)$.

Computing the MLE

3. If several local maxima exist, MLE is the one with largest likelihood.



Example 1

Suppose $\boldsymbol{X} = [X[0], X[1], \cdots, X[N-1]]^T$, where $X[n] \sim N(\mu, \sigma^2)$, $n = 0, \cdots, N-1$. Find the MLE $\hat{\mu}$ for μ.

$$p(\boldsymbol{x};\mu) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[n]-\mu)^2\right]$$

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

$$\ln p(\boldsymbol{x};\mu) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-\mu)^2$$

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

$$\rightarrow \sum_{n=0}^{N-1}(x[n]-\mu) = 0$$

Hence, MLE is $\hat{\mu} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$

Example 2

Suppose $\boldsymbol{X} = [X[0], X[1], \cdots, X[N-1]]^T$, where $X[n] \sim N(\mu, \sigma^2)$, $n = 0, \cdots, N-1$. Find the MLE $\hat{\theta}$ for $\theta = [\mu, \sigma^2]$.

$$\ln p(\boldsymbol{x};\theta) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-\mu)^2$$

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

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

Since $\hat{\theta} = [\hat{\mu}, \hat{\sigma}^2]$ should satisfy local maximal condition,

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

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

Therefore,

$$\hat{\mu} = \frac{1}{N}\sum_{n=0}^{N-1} X[n]$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=0}^{N-1}(X[n] - \hat{\mu})^2$$

Asymptotic Property

Suppose $X \sim p(x;\theta)$. Let $\hat{\theta}$ be the MLE of θ based on n i.i.d. (independent and identically distributed) realization $X[0], X[1], \cdots, X[N-1]$ of X. Under certain regularity conditions, distribution of $\hat{\theta}$ asymptotically converges as

$$\hat{\theta} \sim N(\theta, \boldsymbol{I}^{-1}(\theta)) \text{ as } N \rightarrow \infty.$$

Here, $\boldsymbol{I}(\theta)$ is the Fisher information matrix evaluated at the true θ.

Hence,

- $E\{\hat{\theta}\} \rightarrow \theta \Longrightarrow$ MLE is asymptotically unbiased.
- $Cov\left(\hat{\theta}\right) \rightarrow I^{-1}(\theta) \Longrightarrow$ MLE is asymptotically efficient.

Note: Regularity conditions are:

- Existence of first and second derivatives of log-likelihood function $\ln l(\theta; x)$.
- $E\left\{\frac{\partial \ln p(x;\theta)}{\partial \theta}\right\} = 0.$

Confirmation using Example 2

Suppose $\boldsymbol{X} = [X[0], X[1], \cdots, X[N-1]]^{\,T}$, where $X[n] \sim N(\mu, \sigma^2)$, $n = 0, \cdots, N-1$. Maximum likelihood estimator $\hat{\theta} = [\hat{\mu}, \hat{\sigma}^2]$ are given by

$$\hat{\mu} = \frac{1}{N}\sum_{n=0}^{N-1} X[n]$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=0}^{N-1}(X[n] - \hat{\mu})^2$$

Since random variable $\sum_{n=0}^{N-1}\left(\frac{X[n]-\bar{X}}{\sigma}\right)^2$

(where $\bar{X} = \frac{1}{N}\sum_{n=0}^{N-1} X[n]$) has chi-square distribution with $N-1$ degrees of freedom (χ^2_{N-1}-distribution), its mean and variance are given by $N-1$ and $2(N-1)$. Because of $\frac{N}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{N-1}$,

$$E[\hat{\sigma}^2] = \frac{N-1}{N}\sigma^2,$$

$$Var(\hat{\sigma}^2) = \left(\frac{\sigma^2}{N^2}\right)^2 \{2(N-1)\}$$

Hence,

$$E[\hat{\theta}] = \begin{bmatrix} \mu \\ \frac{N-1}{N}\sigma^2 \end{bmatrix} \xrightarrow{(N\to\infty)} \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \theta$$

$$Cov(\hat{\theta}) = \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2(N-1)}{N^2}\sigma^4 \end{bmatrix} \xrightarrow{(N\to\infty)} \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix} = I^{-1}(\theta)$$

This shows that $\hat{\theta} = [\hat{\mu}, \hat{\sigma}^2]$ converges asymptotically to an efficient estimator.

Practical Techniques

In practical situations, maximum likelihood estimator cannot be always obtained in explicit form. The likelihood function needs to be maximized via iterative procedure.

- Newton-Raphson method
- EM (Expectation-Maximization) algorithm