CS5487 Lecture Notes (2024A)
Prof. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

Lecture 2: Parameter Estimation

How do we find a prob. dist. for a r.v. X?

Three step:

1) Choose a parametric model (e.g. Gaussian), call parameters θ. $p(x_i|\theta)$

2) Assemble a collection of samples (observations) from X:

$$D = \{x_1, \ldots, x_N\}$$

We assume x_i's are independent samples of X.
(i.i.d. = independent & identically distributed)

3) Maximum Likelihood principle:

"the optimal parameter θ^* is that which maximizes the likelihood (probability) of the training data D."

ML estimate:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\; p(D|\theta)$$

likelihood of data w.r.t. parameter θ
"likelihood function"

$$= \underset{\theta}{\operatorname{argmax}}\; \underbrace{\log p(D|\theta)}_{\ell(\theta)\,=\,\text{log-likelihood function (LL)}}$$

$$= \underset{\theta}{\operatorname{argmin}}\; \underbrace{-\log p(D|\theta)}_{\text{negative log-likelihood (NLL) loss function.}}$$

Note: log = natural log (ln) = log base e.

Note: D is known, so $p(D|\theta)$ is a function of θ.
This is not a pdf, and does not have the same shape as the $p(x)$.

Data LL

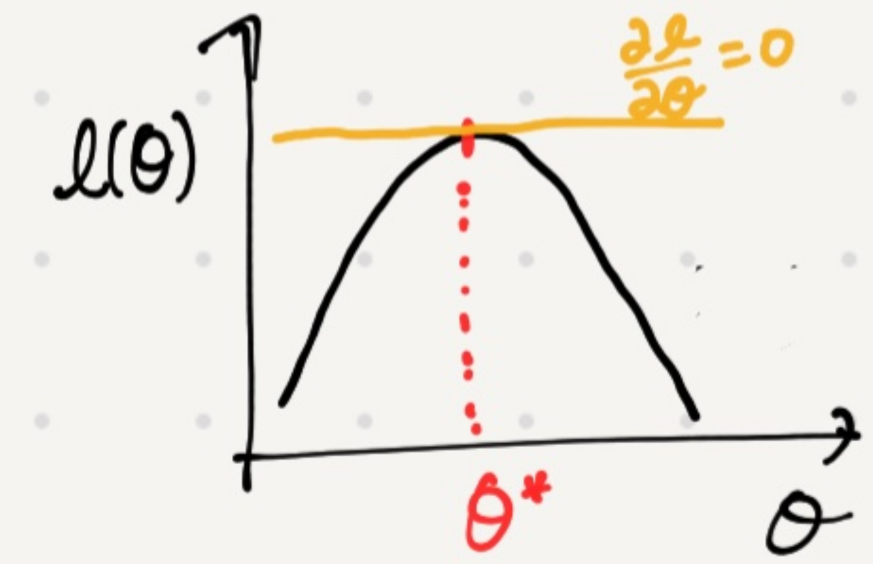
$$\ell(\theta) = \log p(D|\theta)$$

independence

$$= \log \underbrace{\prod_{i=1}^{N}}\, p(x_i|\theta)$$

$\log(ab) = \log a + \log b$

$$= \sum_{i=1}^{N} \log p(x_i|\theta)$$

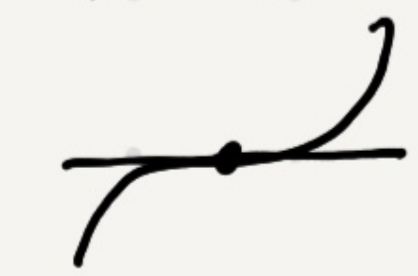


To get ML solution:

If θ is a scalar, at local optimum:

1) $\frac{\partial}{\partial \theta} \log p(D|\theta) = 0$ at θ^*

2) $\frac{\partial^2}{\partial \theta^2} \log p(D|\theta) < 0$ (concave at θ^*)



3) check boundary conditions on θ (if necessary)

If θ is a vector...

1) $\nabla_\theta \ell(\theta) = \begin{bmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \theta_p} \end{bmatrix} = 0$

gradient

2) $\nabla^2_\theta \ell(\theta) \prec 0$ (negative definite)

Hessian describes the curvature at θ.

$$\nabla^2_\theta \ell = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_p} & \cdots & \frac{\partial^2 \ell}{\partial \theta_p^2} \end{bmatrix}$$

Negative defn ($H \prec 0$):
$\theta^T H \theta < 0, \ \forall \theta$.
$\Rightarrow$ every direction will decrease the gradient
$\Rightarrow$ top of hill

Example: Bernoulli

$\theta = \pi \quad , \quad 0 \leq \pi \leq 1 \qquad p(x_i|\pi) = \pi^{x_i}(1-\pi)^{1-x_i}$

$D = \{x_1, \dots, x_N\}$

$$\ell(\theta) = \sum_{i=1}^{N} \log p(x_i|\theta)$$

$$= \sum_i \log \pi^{x_i}(1-\pi)^{1-x_i}$$

$\log(ab) = \log a + \log b$

$$= \sum_i \log \pi^{x_i} + \log(1-\pi)^{1-x_i}$$

$\log a^b = b \log a$

$$= \sum_i \underbrace{x_i} \log \pi + \underbrace{(1-x_i)} \log(1-\pi)$$

$$= \underbrace{\left(\sum_{i=1}^{N} x_i\right)}_{\#\text{ of 1s}} \log \pi + \underbrace{\left(\sum_{i=1}^{N}(1-x_i)\right)}_{\#\text{ of 0s}} \log(1-\pi)$$

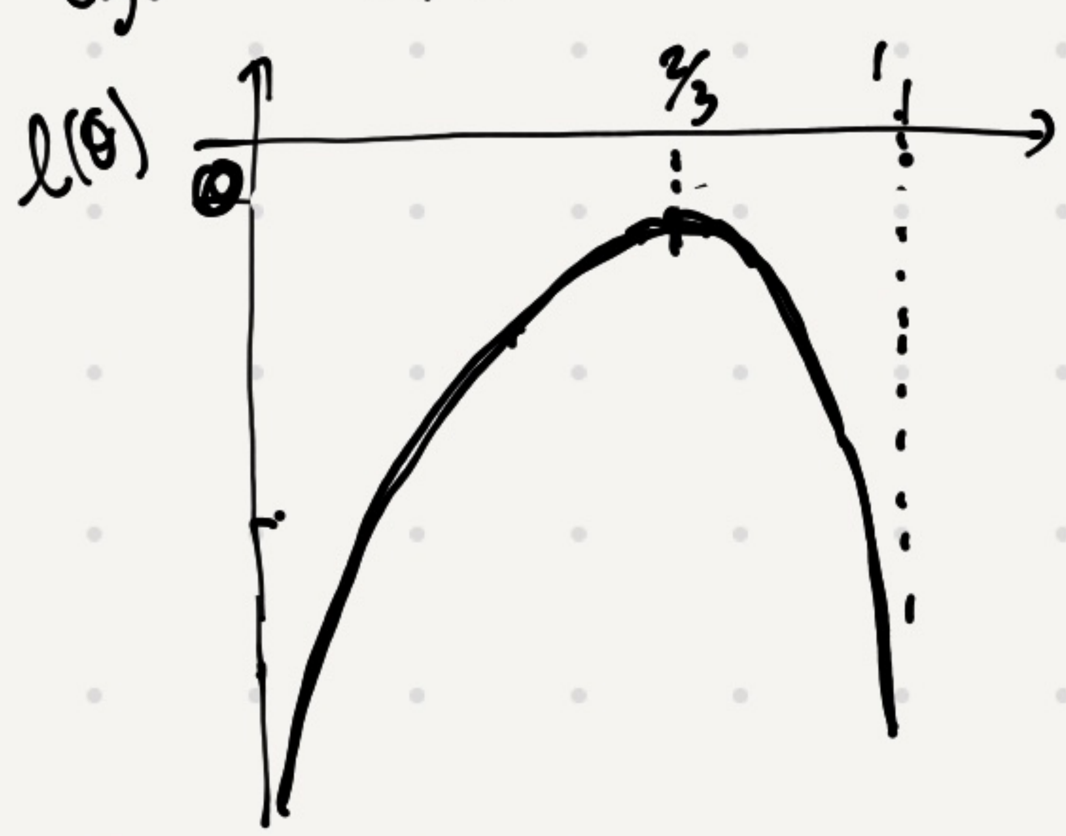
$m \qquad\qquad N-m$

↑ m "Sufficient statistic" - $\ell(\theta)$ only depends on the N observations through this term.

$\Rightarrow m = \sum_{i=1}^{N} x_i$

$$\ell(\theta) = m \log \pi + (N-m)\log(1-\pi)$$

e.g. $m=4$, $N=6$



1) $$\frac{\partial}{\partial \pi}\ell(\pi) = \frac{m}{\pi} + \frac{N-m}{1-\pi}\underbrace{\frac{\partial}{\partial \pi}(1-\pi)}_{-1} = 0$$

$\pi(1-\pi)$

$$\Rightarrow m(1-\pi) + (N-m)\pi(-1) = 0$$

$$m - \cancel{m\pi} - N\pi + \cancel{m\pi} = 0$$

$$m - N\pi = 0 \Rightarrow \boxed{\hat{\pi} = \frac{m}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

"Fraction of 1s we saw"

Sample mean

2) $$\frac{\partial^2}{\partial \pi^2}\ell(\theta) = \frac{\partial}{\partial \pi}\left(\underbrace{\frac{\partial}{\partial \pi}\ell(\theta)}\right) = \frac{\partial}{\partial \pi}\left(\underbrace{\frac{m}{\pi}}_{\pi^{-1}} - \underbrace{\frac{N-m}{1-\pi}}_{(1-\pi)^{-1}}\right)$$

$$= \frac{m}{\pi^2}(-1) - \frac{N-m}{(1-\pi)^2}(-1)\underbrace{\frac{\partial}{\partial \pi}(1-\pi)}_{(-1)}$$

$$= -\underbrace{\frac{m}{\pi^2}}_{>0} - \underbrace{\frac{N-m}{(1-\pi)^2}}_{>0} < 0 \quad ✓$$

3) Boundary condition:

$$0 \leq m \leq N \Rightarrow 0 \leq \pi \leq 1 \quad ✓$$

Example: Gaussian

1) $\theta = \mu$ (mean), σ^2 is known

$D = \{x_1, \dots, x_N\}$

$$\ell(\theta) = \sum_{i=1}^{N} \log p(x_i|\theta)$$

$$= \sum_i \left[\underbrace{-\frac{1}{2}\log 2\pi} \underbrace{-\frac{1}{2}\log \sigma^2} \underbrace{-\frac{1}{2\sigma^2}} (x_i-\mu)^2 \right]$$

$$= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\underbrace{\sum_{i=1}^{N}(x_i-\mu)^2}_{\sum_i (x_i^2 - 2\mu x_i + \mu^2)}$$

what are the suff. stats?

$\sum_i x_i^2$, $\sum_i x_i$

• $\frac{\partial \ell}{\partial \mu} = \frac{\partial \ell}{\partial \mu}\left(-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right) = \underbrace{-\frac{1}{\cancel{2}\sigma^2}}_{\times}\sum_i \cancel{2}(x_i-\mu)\cdot\underbrace{\frac{\partial}{\partial\mu}(x_i-\mu)}_{-1\ \times} = 0$

$$\Rightarrow \sum_{i=1}^{N}(x_i-\mu) = 0$$

$$\sum_i x_i - N\mu = 0 \Rightarrow \boxed{\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

"sample mean"

2) $\theta = \sigma^2$ (μ is known)

• $\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2}\frac{1}{\sigma^2} - \frac{1}{2}\frac{1}{\sigma^4}(-1)\sum_{i=1}^{N}(x_i-\mu)^2 = 0$ $\times\ \sigma^4$

$$-\frac{N}{\cancel{2}}\sigma^2 + \frac{1}{\cancel{2}}\sum_{i=1}^{N}(x_i-\mu)^2 = 0$$

$$\Rightarrow \boxed{\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

"sample variance"

M.v. Gaussian

solution on tutorial...

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$$

Estimators

The **estimate** ($\hat{\mu}$) is a number given a dataset D.

The dataset is random, then the **estimator** is a r.v.

Dataset r.v.: $X_1, \dots, X_N$, $X_i \sim p(x_i|\theta)$

Estimator: $f(X_1, \dots, X_N) = \frac{1}{N}\sum_{i=1}^{N} X_i$

r.v. for each sample according to the true distribution.

ML estimate: $\hat{\mu} = f(X_1,\dots,X_N)\Big|_{X_1=x_1,\dots,X_N=x_n}$

$$= \frac{1}{N}\sum_i x_i$$

- Since the estimator is a r.v., we can calculate its mean & variance → quantify how good the estimator is.

Bias & Variance: $\hat{\theta} = f(X_1,\dots,X_N)$

1) Will it converge to the true value θ?

measures expressiveness

$$\text{Bias}(\hat{\theta}) = E_{X_1\dots X_N}[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta$$

true value, mean of $\hat{\theta}$

If the bias is non-zero, we can never get the true value! (even with infinite # of samples).

2) How long will it take to converge? (How many samples?)

$$\text{Var}(\hat{\theta}) = E_{X_1\dots X_N}[(\hat{\theta} - E\hat{\theta})^2]$$

measures uncertainty / variability.

Example: Gaussian

Estimator: $\hat{\mu} = \frac{1}{N}\sum_i X_i$, $X_i \sim N(\mu, \sigma^2)$

1) **mean**: $E_{X_1\dots X_N}[\hat{\mu}] = E_{X_1\dots X_N}\left[\frac{1}{N}\sum_i X_i\right] = \frac{1}{N}\sum_i E_{X_i}[X_i] = \frac{1}{N}\sum_i \mu$

$$= \frac{1}{N} N\mu = \mu$$

$$\Rightarrow \boxed{\text{Bias}(\hat{\mu}) = \mu - \mu = 0} \quad ✓$$

2) $\text{var}(\hat{\mu}) = E_{X_1\dots X_N}[(\hat{\mu} - E\hat{\mu})^2] = E\left[\left(\frac{1}{N}\sum_i X_i - \mu\right)^2\right]$

$\frac{1}{N}N\mu$

$\frac{1}{N}\left(\sum_i (x_i - \mu)\right)$

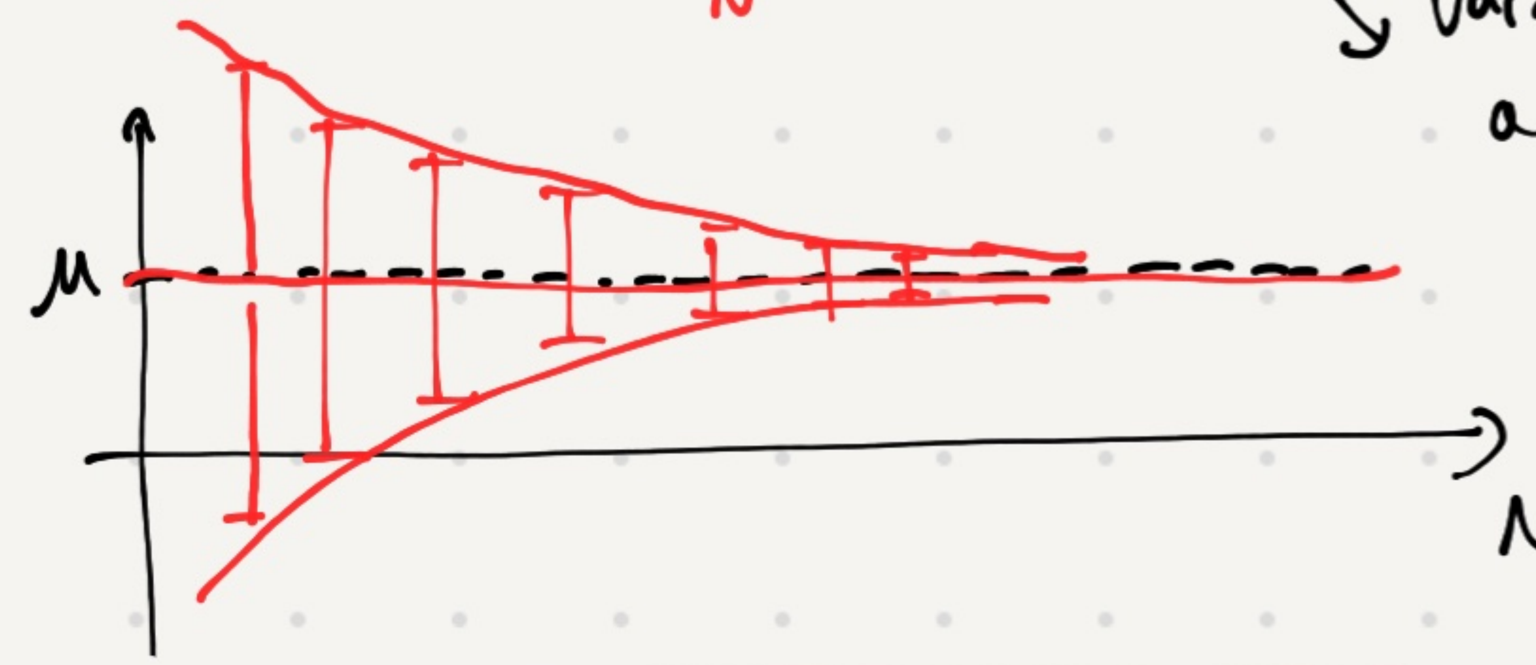
$$= \frac{1}{N^2} E\left[\left(\sum_i (x_i - \mu)\right)^2\right]$$

$$= \frac{1}{N^2} E\left[\sum_i\sum_j (x_i-\mu)(x_j-\mu)\right]$$

$$= \frac{1}{N^2}\sum_i\sum_j E[(x_i-\mu)(x_j-\mu)] \Rightarrow \begin{cases} i=j: E[(x_i-\mu)^2] = \sigma^2 \\ i\neq j: E[(x_i-\mu)(x_j-\mu)] = 0 \end{cases}$$

$$= \frac{1}{N^2}\left(\underbrace{\sum_{i=j}}_{N} \sigma^2\right) = \boxed{\frac{1}{N}\sigma^2}$$

↘ variance converges to 0 as $N \to \infty$.



Another example (PS 2-12)

$$E[\hat{\sigma}^2] = \underbrace{\frac{N-1}{N}}\sigma^2 \Rightarrow \text{Bias}(\hat{\sigma}^2) = -\frac{1}{N}\sigma^2$$

$$\hat{\hat{\sigma}}^2 = \frac{N}{N-1}\hat{\sigma}^2 = \frac{N}{N-1}\cdot\frac{1}{N}\sum_i (x_i-\mu)^2 = \boxed{\frac{1}{N-1}\sum_i (x_i-\mu)^2}$$

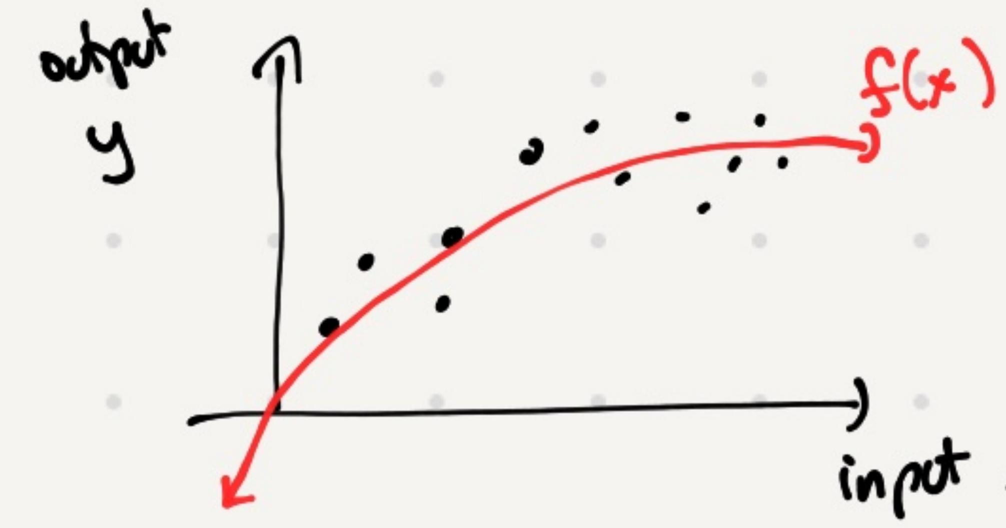
unbiased estimator of σ^2.

Important Asymptotic Properties of MLE

1) consistent - As $N \to \infty$, the estimated value converges to the true value. (asymptotically unbiased)

2) efficient - Achieves the Cramér-Rao Lower Bound (CRLB) as $N \to \infty$.
CRLB is a theoretical bound on the variance of any unbiased estimator for a given $p(x|\theta)$.
(i.e. no unbiased estimator achieves lower variance than MLE)

MLE for Regression (Supervised Learning)



Assume $f(x)$ is a polynomial.

$$f(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots \theta_d x^d$$

$$f(x,\theta) = \underbrace{\begin{bmatrix}\theta_0\\ \vdots \\ \theta_d\end{bmatrix}}_{\theta}^T \underbrace{\begin{bmatrix}1\\ x\\ x^2\\ \vdots \\ x^d\end{bmatrix}}_{\phi(x)} = \theta^T \phi(x)$$

Observe a noisy version y given an input x:

$$\underset{\uparrow \atop \text{r.v.}}{y} = \underbrace{f(x,\theta)}_{\text{value}} + \underset{\uparrow \atop \text{r.v.}}{\epsilon}, \quad \epsilon \sim N(0, \sigma^2) \text{ iid}$$

equivalently, $p(y|x,\theta) = N(y \mid f(x,\theta), \sigma^2)$

Given dataset $D = \{(x_i, y_i)\}_{i=1}^N$, estimate θ.

MLE:

$$\theta^* = \underset{\theta}{\arg\max} \sum_i \log p(y_i | x_i, \theta) \leftarrow \text{MLE}$$

tutorial $\vdots$

$$= \underset{\theta}{\arg\min} \sum_i (y_i - f(x_i,\theta))^2 \Leftarrow \text{least-squares formulation}$$

$\vdots$

$$\theta^* = (\Phi \Phi^T)^{-1} \Phi y, \quad \Phi = [\phi(x_1) \dots \phi(x_N)], \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$