Lecture 2 | Parameter Estimation

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

Three Steps:

1) Choose a parametric model (e.g. Gaussian)
   call the parameters θ.
2) collect a set of observations (samples) from X.

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

we assume x_i's are independently & identically distributed (iid)

3) Maximum Likelihood principle

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

ML estimate (MLE)

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

likelihood of the data D w.r.t. parameter θ. likelihood function

$\ast$ D is known, so $p(D|\theta)$ is a function of θ.
It is not a probability distribution. It doesn't have the same shape as the pdf.

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

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

log = natural logarithm (ln), log base e.

Data log likelihood

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

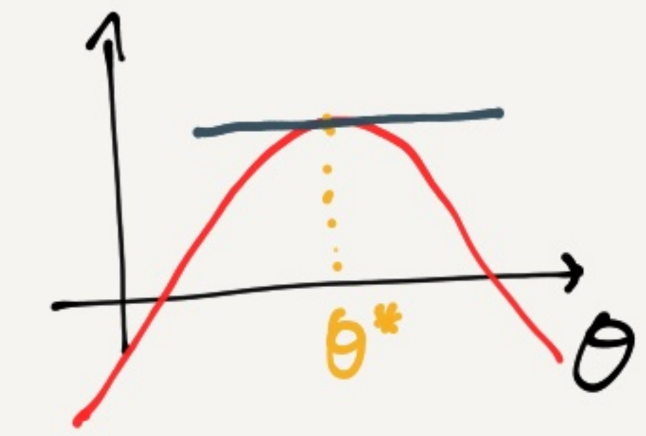
independence assumption

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

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

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

To get the ML solution



if θ is a scalar, at local maximum:

1) $\frac{\partial}{\partial \theta} \log p(D|\theta) = 0$ at θ^*
2) $\frac{\partial^2}{\partial \theta^2} \log p(D|\theta) < 0$ (at the max, it's concave)
3) check the boundary conditions of θ (i.e. it's a valid parameter)

if θ is a vector...

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

gradient

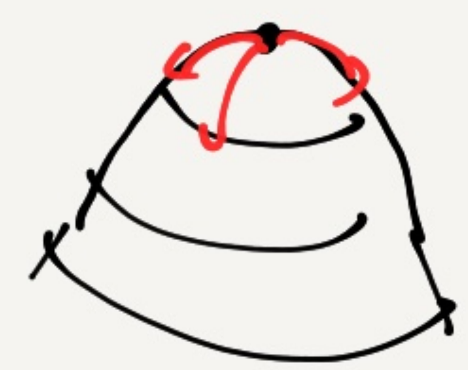
2) $\nabla^2_\theta \ell(\theta) \prec 0$ (negative definite)

Hessian

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

CS5487 Lecture Notes (2022A)
Prof. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

$H \prec 0$: H is negative definite: $\theta^T H \theta < 0, \ \forall \theta$. "mountain"



$H \succ 0$: H is positive definite:

$\theta^T H \theta > 0 \quad \forall \theta$. "bowl"

Example: Bernoulli

$\theta = \pi$, $0 \le \pi \le 1$, $\mathcal{X} = \{0, 1\}$

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

log-likelihood

$$\ell(\theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta) = \sum_{i=1}^{N} \log\left(\pi^{x_i} (1-\pi)^{1-x_i}\right)$$

$$= \sum_{i=1}^{N} \left[x_i \log \pi + (1-x_i) \log (1-\pi)\right]$$

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

"sufficient statistic" – $\ell(\theta)$ only depends on the data through this suff. statistic.

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

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

Solve for θ^* : compute the derivative & set to 0.

1) $\frac{\partial}{\partial \pi} \ell(\theta) = \frac{m}{\pi} + (N-m) \frac{1}{1-\pi} (-1) = 0$ $\quad \downarrow \pi(1-\pi)$

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

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

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

"fraction of 1s observed" (sample mean)

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

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

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

3) boundary condition:

$0 \le m \le N$

$\Rightarrow 0 \le \pi \le 1$ ✓

Example: Gaussian

$\theta = \mu$ (σ^2 known)

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

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

$$= \sum_i -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(x_i-\mu)^2$$

$$\boxed{\ell(\theta) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2}$$

what are the suff. stats? $\left\{\sum_{i=1}^{N} x_i, \sum_{i=1}^{N} x_i^2\right\}$

Solve for μ

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

$$\sum_i x_i - \sum_i \mu = 0$$

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

"sample mean"

$\theta = \sigma^2$ (μ known)

$$\frac{\partial}{\partial\sigma^2}\ell(\theta) = -\frac{N}{2}\frac{1}{\sigma^2} - \frac{1}{2}\left[\sum_i(x_i-\mu)^2\right]\frac{(-1)}{\sigma^4} = 0$$

$\sigma^4 \cdot 2$

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

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

"sample variance"

Multivariate Gaussian

$$\boxed{\begin{aligned}\hat{\mu} &= \frac{1}{N}\sum_{i=1}^{N} x_i \\ \hat{\Sigma} &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T\end{aligned}}$$

see the tutorial next week.

Estimators

The estimate is a number: $\mu_{ML} = \frac{1}{N}\sum_i x_i$

μ_{ML} ← value; x_i ← values

The estimator is a r.v. over many possible datasets.

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

$\hat{\mu}$ ← estimator also a r.v.; X_i ← r.v. for each sample

$X_i \sim p(x_i|\theta)$ (the true distribution)

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

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

1) Will it converge to the true value of Θ?

$$\text{Bias}(\hat{\Theta}) = E_{X_1,\dots,X_N}[\underset{\uparrow\ \text{estimator}}{\hat{\Theta}} - \underset{\uparrow\ \text{true value}}{\Theta}] = E(\hat{\Theta}) - \Theta$$

Measures expressiveness: if the bias is non-zero, we can never get the true parameter (even ∞ samples)

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

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

measuring the uncertainty/variability.

Example: Gaussian

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

Mean $E_{X_1\dots X_N}[\hat{\mu}] = E_{X_1\dots X_N}[\frac{1}{N}\sum_{i=1}^{N} x_i] = \frac{1}{N}\sum_{i=1}^{N} E_{X_i}[X_i] = \frac{N\mu}{N} = \mu$

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

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

$\frac{1}{N^2}$

$$= \frac{1}{N^2} E[(\sum_{i=1}^{N}(x_i - \mu))^2]$$

$$= \frac{1}{N^2} E[\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - \mu)(x_j - \mu)]$$

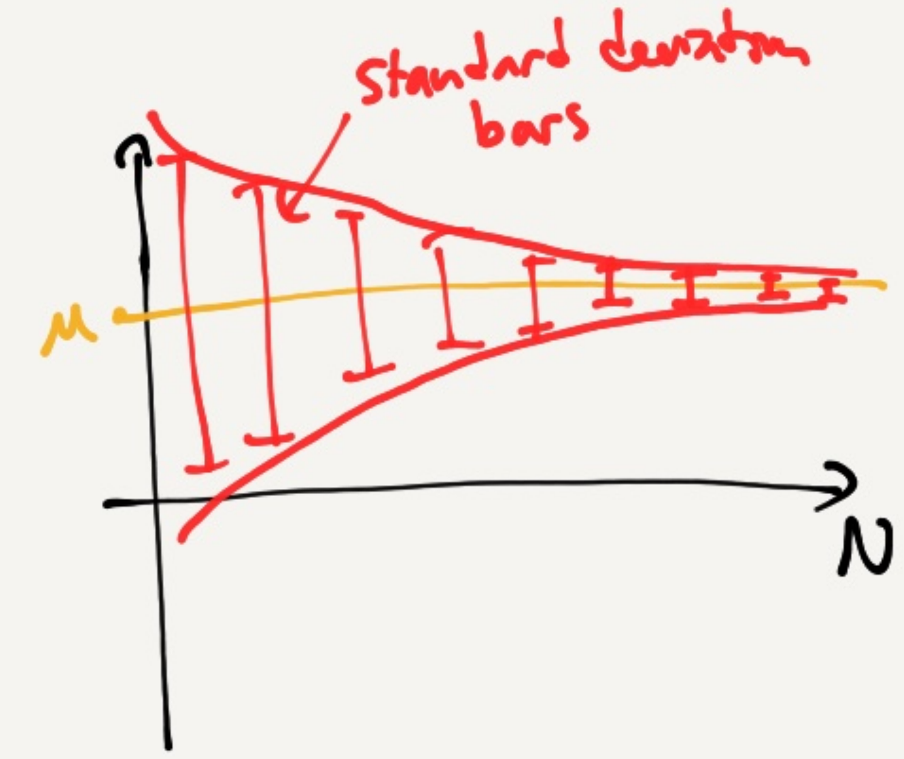
$i = j \Rightarrow \sigma^2$

$i \neq j \Rightarrow 0$

$$= \frac{1}{N^2}[\underset{\uparrow N}{\sum_{i=j}\sigma^2} + \underset{\uparrow 0}{\sum_{i\neq j} 0}] =$$

$$\boxed{\text{var}(\hat{\mu}) = \frac{1}{N}\sigma^2}$$

variance converges to zero as $N \to \infty$



for variance (PS 2-12)

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

To make it unbiased: $\tilde{\sigma}^2 = \frac{N}{N-1}\hat{\sigma}^2 = \frac{N}{N-1}\frac{1}{N}\sum_i (x_i - \mu)^2 = \frac{1}{N-1}\sum_i (x_i - \mu)^2$

Important Asymptotic Properties of MLE

1) **consistent** - asymptotically unbiased: as $N \to \infty$, the estimated value converges (in probability) to the true value.

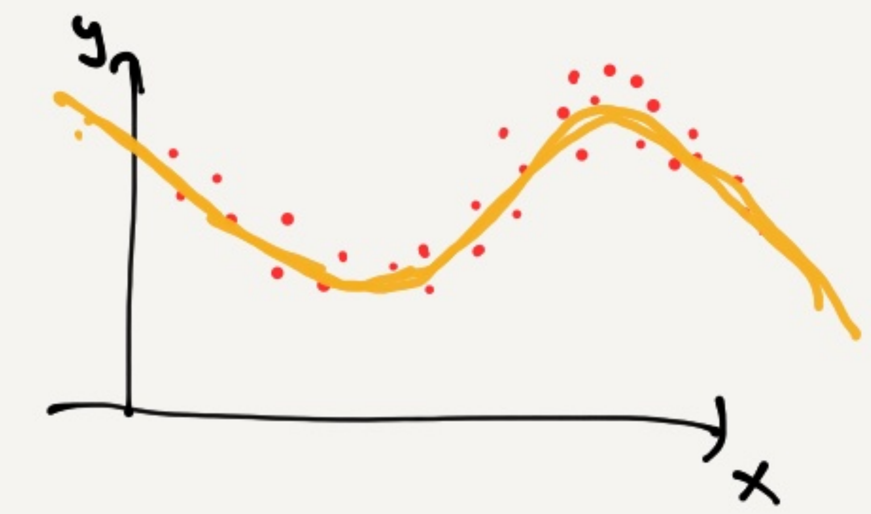
2) **efficient** - achieves the Cramér-Rao Lower Bound (CRLB) as $N \to \infty$.

CRLB is a theoretical bound on the variance of any unbiased estimator for a particular $p(x|\theta)$.

(no unbiased estimator can get lower variance).

MLE for regression (supervised learning)

$D = \{(x_i, y_i)\}$



$x \in \mathbb{R}$ input

polynomial function: $f(x,\theta) = \sum_{d=0}^{K} \theta_d x^d = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots \theta_K x^K$

$$= \underbrace{\begin{bmatrix} \theta_0 \\ \vdots \\ \theta_K \end{bmatrix}}_{\theta}^T \underbrace{\begin{bmatrix} 1 \\ x \\ \vdots \\ x^K \end{bmatrix}}_{\phi(x)} = \underbrace{\phi(x)^T \theta}_{\text{linear function of the parameters } \theta}$$

Observe a noisy output y_i for a given x_i:

$$y_i = f(x_i, \theta) + \epsilon_i$$

$\epsilon_i \sim N(0, \sigma^2)$ iid Gaussian noise.

pdf of y_i

$$p(y_i | x_i, \theta) = N(y_i | f(x_i, \theta), \sigma^2)$$

Estimate θ using MLE

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \log p(y_i | x_i, \theta)$$

$$\vdots$$

$$= \arg\min_{\theta} \sum_i (y_i - f(x_i, \theta))^2$$

Least-squares formulation

$$\vdots$$

$$= \arg\min_{\theta} \| y - \Phi^T \theta \|^2, \quad \Phi = \begin{bmatrix} \phi(x_1) & \dots & \phi(x_N) \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$\vdots$$

$$\boxed{\hat{\theta} = (\Phi \Phi^T)^{-1} \Phi y}$$

Notes:

1) ML is more general than L.S.

2) assumptions are explicit

   i) Gaussian additive noise

   ii) iid samples (iid noise)

   iii) $\mu = 0$, σ^2 variance

3) ML can describe other LS formulations

   i) weighted LS (PS 2-8)

   ii) regularized LS (Lecture 3)

   iii) L_p Norm (PS 2-9)