Maximum Likelihood Estimation: Theoretical Properties

Qiang Liu
UT Austin

- MLE estimator is random variable (as a function of the random data):

$$\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n), \qquad \{x_i\} \overset{iid}{\sim} p(\cdot \mid \theta_*).$$

- Evaluation metrics: Bias, variance, mean square error (MSE).
- Unbiased estimators vs. consistent estimators.

$$\mathrm{Bias}(\hat{\theta}) = E_{\theta^*}\left[\hat{\theta}(x_1 \cdots x_n)\right] - \theta^*$$

If $\mathrm{Bias}(\hat{\theta}) = 0$, we call $\hat{\theta}$ "unbiased".

$\mathrm{MSE}(\hat{\theta}) \to 0$, as $n \to \infty$, $\hat{\theta}$ "consistent".

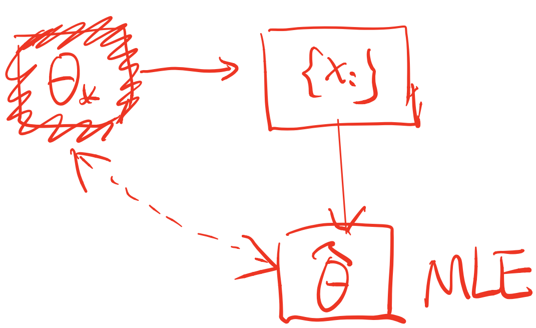
$\mathrm{Bias}(\hat{\theta}) \to 0$, as $n \to \infty$, "Asymptotic Unbiased"

2

$$\mathrm{MSE}(\hat{\theta}) = \left(\mathrm{Bias}(\hat{\theta})\right)^2 + \mathrm{Var}(\hat{\theta})$$

Proof:
$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= E\left[(\hat{\theta} - \theta_*)^2\right] \\
&= E\left[\left(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta_*\right)^2\right] \\
&= E\Big[(\hat{\theta} - E(\hat{\theta}))^2 + (E(\hat{\theta}) - \theta_*)^2 \\
&\qquad + \underbrace{2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta_*)}_{=0}\Big] \\
&= \mathrm{Var}(\hat{\theta}) + (\mathrm{bias}(\hat{\theta}))^2
\end{aligned}$$

$$\begin{aligned}
E\left[(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta_*)\right] &= E\left[\hat{\theta} - E(\hat{\theta})\right]\left(E(\hat{\theta}) - \theta^*\right) \\
&= \left(E(\hat{\theta}) - E(\hat{\theta})\right)\left(E(\hat{\theta}) - \theta^*\right) \\
&= 0
\end{aligned}$$



- **Example:** For $\{x_i\}_{i=1}^n \sim \mathcal{N}(\mu, \sigma^2)$, MLE is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu})^2.$$

"Unbiased"

$$\boxed{\mathrm{Bias}(\hat{\mu})} = E[\hat{\mu}] - \mu = E\left[\frac{1}{n}\sum_{i=1}^n x_i\right] - \mu = \frac{1}{n}\sum_{i=1}^n E[x_i] - \mu$$

$$= \frac{1}{n}\sum_{i=1}^n \mu - \mu = \boxed{0}$$

$$\boxed{\mathrm{var}(\hat{\mu})} = E[(\hat{\mu} - \mu)^2]$$

$$= E\left[\left(\frac{1}{n}\sum_{i=1}^n x_i - \mu\right)^2\right] = E\left[\frac{1}{n^2}\left(\sum_{i=1}^n (x_i - \mu)^2 + \sum_{i \neq j}(x_i - \mu)(x_j - \mu)\right)\right]$$

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

$$\boxed{= \frac{1}{n}\sigma^2} + \sum E[(x_i - \mu)] E[(x_j - \mu)]$$

$$= 0$$

$$\mathrm{MSE}(\hat{\mu}) = (\mathrm{Bias}(\hat{\mu}))^2 + \mathrm{var}(\hat{\mu}) = \frac{\sigma^2}{n}.$$

If $n \to \infty$, $\mathrm{MSE}(\hat{\mu}) \to 0 \Rightarrow$ Consistent.

$$E[\hat{\sigma}^2] \neq \sigma^2 \quad , \quad \hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \hat{\mu})^2$$

"Biased"

$$E[\hat{S}^2] = \sigma^2$$

$\mathrm{Bias}(\hat{\sigma}^2) \to 0$. $n \to \infty$ (Asymp. Unbiased)

$\mathrm{var}(\hat{\sigma}^2) \to 0$ $n \to \infty$

$\mathrm{MSE}(\hat{\sigma}^2) \to 0$ $n \to \infty$ (consistent)

Why MLE?

- MLE is equivalent to minimizing Kullback-Leibler (KL) Divergence.

$$\mathrm{KL}(q \,\|\, p) = \mathbb{E}_q[\log q(x) - \log p(x)]. = \begin{cases} \sum_x q(x)\left(\log\left(\frac{q(x)}{p(x)}\right)\right) \\ \int q(x) \log\left(\frac{q(x)}{p(x)}\right) dx \end{cases}$$

- $\mathrm{KL}(q \,\|\, p) \geq 0$ for any q and p.
- $\mathrm{KL}(q \,\|\, p) = 0$ if and only if $q = p$.

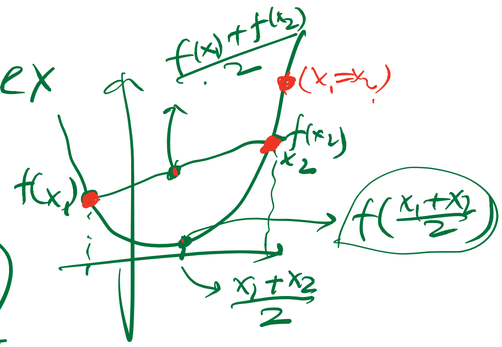
$$KL(q\,\|\,P) \neq KL(P\,\|\,q)$$

Jensen's Inequality:

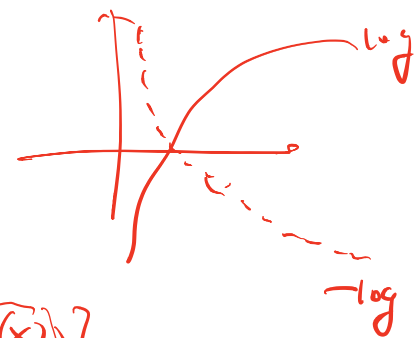
If $f(x)$ is convex

then

$$E_q[f(x)] \geq f(E_q(x))$$



$$q(x) = \frac{\delta_{x_1} + \delta_{x_2}}{2}$$



$$\underbrace{KL(q\,\|\,P)}_{\geq 0} = E_q\left[\log\left(\frac{q}{p}\right)\right] = E_q\left[-\log\left(\frac{P(x)}{q(x)}\right)\right]$$

$$\geq -\log\left(E_q\left[\frac{P(x)}{q(x)}\right]\right)$$

$$= -\log\left(\sum_x q(x) \frac{P(x)}{q(x)}\right)$$

$$= -\log\left(\sum_x P(x)\right) = -\log(1) = 0$$

$$\Rightarrow \frac{P(x)}{q(x)} = \text{const}$$

$$\Rightarrow P(x) = \text{const} \cdot q(x)$$

$$\Rightarrow \sum_x P(x) = \text{const} \cdot \sum_x q(x)$$

$$\Rightarrow 1 = \text{const} \cdot 1 \quad \Rightarrow \text{const} = 1$$

$$\Rightarrow p = q$$

$$MLE \Longleftrightarrow KL$$

$$KL(q\,\|\,P) = E_q[\log q(x) - \log P(x)]$$

Assume q is data distribution (Given)

P_θ is "model" $\theta \in \Theta$

$$\min_\theta(KL(q\,\|\,P_\theta)) \Leftrightarrow \min_\theta E_q[\log q(x) - \log P_\theta(x)]$$

$$\Leftrightarrow \min_{\theta} -E_q[\log P_\theta(x)]$$

$$\Leftrightarrow \boxed{\max_{\theta} E_q[\log P_\theta(x)]}$$

$$\approx \max_{\theta} \frac{1}{n}\sum_{i=1}^{n} \log P_\theta(x_i) \qquad \{x_i\} \sim q$$

Avg log-likelihood.