MLE part I:

Example: Biased coin.

Assume we have a biased coin whose probability of head is unknown,

$$\begin{cases} P(X=H) = \theta \\ P(X=T) = 1-\theta. \end{cases}$$

we observe the following (iid): HHHHT.

Problem: estimate θ.

Intuitively, $\hat{\theta} = 4/5$.

supposed $\theta = 0.9$ vs $\theta = 0.1$:

If $\theta = 0.9$, $P(HHHHT \mid \theta = 0.9) = \theta^4(1-\theta)^1$

$= (0.9)^4(0.1)^1$

If $\theta = 0.1$, $P(HHHHT \mid \theta = 0.1) = (0.1)^4(0.9)^1$

Likelihood function (L):

$$L(\theta) = P(HHHHT \mid \theta) = \theta^4(1-\theta)^1$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\left(L(\theta)\right).$$

$$= \underset{\theta}{\operatorname{argmax}}\left(\theta^4(1-\theta)^1\right).$$

Maximizing likelihood function $\propto$ maximizing the log-likelihood function.

$$\ell(\theta) = \log L(\theta) = \log\left(\theta^4(1-\theta)^1\right)$$
$$= 4\log\theta + 1\log(1-\theta)$$

$$\hat{\theta} = \arg\max_{\theta}\underbrace{\left(4\log\theta + \log(1-\theta)\right)}_{\ell(\theta)}.$$

To find the θ that maximizes ℓ, need to set $\nabla\ell = 0$:

$$\nabla\ell(\theta) = \frac{4}{\theta} + \frac{-1}{1-\theta} = 0.$$

$$\Rightarrow \frac{4}{\theta} = \frac{1}{1-\theta} \Rightarrow \frac{1-\theta}{\theta} = \frac{1}{4} \Rightarrow \frac{1}{\theta} - 1 = \frac{1}{4}$$

$$\Rightarrow \frac{1}{\theta} = \frac{5}{4} \Rightarrow \boxed{\hat{\theta} = \frac{4}{5}}$$ as expected intuitively.

Parameter Estimation by MLE: Algorithm

Problem: Given set of observations $\{x_i\}_{i=1}^n$, drawn iid from unknown distribution P^*, from a parametric family of distributions: $\{p(\cdot \mid \theta) : \theta \in \Theta\}$. Estimate θ.

Algorithm:

1) Form the Likelihood function L:

$$L(\theta) = P(\{x_i\}_{i=1}^n \mid \theta) = \prod_{i=1}^n P(x_i \mid \theta)$$

2) Calculate $\ell(\theta) = \log L(\theta)$:

$$\ell(\theta) = \log(L(\theta)) = \log\left(\prod_{i=1}^n P(x_i \mid \theta)\right)$$
$$= \sum_{i=1}^n \log P(x_i \mid \theta)$$

3) Calculate the maximum likelihood Estimation:

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \ell(\theta) = \operatorname*{argmax}_{\theta} \left\{ \sum_{i=1}^n \log P(x_i \mid \theta) \right\}$$

4) Solve the optimization problem:

1) closed form, by setting $\nabla \ell(\theta) = 0$
2) Numerical method.

Example: Biased Coin, Revisited:

- Given $\{x_i\}_{i=1}^n$, where $x_i \in \{0, 1\}$ iid drawn from Bernoulli distribution:

$$\begin{cases} P(X=1) = \dfrac{\exp(w)}{1+\exp(w)} = \theta \\ P(X=0) = \dfrac{1}{1+\exp(w)} = 1-\theta \end{cases}$$

- Problem: Estimate w.

Use $P(X) = \dfrac{\exp(Xw)}{1+\exp(w)}$ → If $X=1 \Rightarrow \dfrac{\exp(w)}{1+\exp(w)}$

→ If $X=0 \Rightarrow \dfrac{1}{1+\exp(w)}$

$$\ell(w) = \log \overbrace{P(x_1, \dots, x_n \mid w)}^{L(\theta)} = \log \prod_{i=1}^n P(x_i \mid w)$$

$$= \sum_{i=1}^n \log P(x_i \mid w) = \sum_{i=1}^n \log\left(\frac{\exp(x_i w)}{1+\exp(w)}\right)$$

$$= \sum_{i=1}^n \left[x_i w - \log\left(1+\exp(w)\right)\right]$$

$$= \left(\sum_{i=1}^n x_i\right) w - n \log\left(1+\exp(w)\right)$$

$$= n\left(\bar{x} w - \log\left(1+\exp(w)\right)\right), \text{ where } \bar{x} = \sum_{i=1}^n \frac{x_i}{n}$$

$$\nabla \ell(w) = n\left(\bar{x} - \frac{\exp(w)}{1+\exp(w)}\right) = 0.$$

Recall: $\theta = \frac{\exp(w)}{1+\exp(w)}$

therefore $\hat{\theta} = \frac{\exp(\hat{w})}{1+\exp(\hat{w})} = \bar{x} = \frac{\#(1)}{\#(1)+\#(0)}$

$\Rightarrow \bar{x} = \frac{\exp(\hat{w})}{1+\exp(\hat{w})} \rightarrow \bar{x}\left(1+\exp(\hat{w})\right) = \exp(\hat{w})$

$\rightarrow \bar{x} + \bar{x}\exp(\hat{w}) = \exp(\hat{w})$

$\rightarrow \bar{x} = \exp(\hat{w}) - x\exp(\hat{w}) = \exp(\hat{w})(1-\bar{x})$

$\rightarrow \frac{\bar{x}}{1-\bar{x}} = \exp(\hat{w}) \rightarrow \boxed{\hat{w} = \log\left(\frac{\bar{x}}{1-\bar{x}}\right)}$

Example: Gaussian Distribution:

- Given $\{x_i\}_{i=1}^n$ iid drawn from Gaussian distribution $N(\mu, \sigma^2)$:

$$p(x \mid \underbrace{\theta}_{\mu, \sigma^2}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(x-\mu)^2\right), \quad \theta = \{\mu, \sigma\}.$$

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

$$= \sum_{i=1}^n \log\left[\left(\frac{1}{(2\pi)^{1/2}\sigma}\right) - \frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

$$= \sum_{i=1}^n \left[\log\left(\frac{1}{(2\pi)^{1/2}}\right) - \log(\sigma) - \frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

$$= \underbrace{-\frac{n}{2}\log(2\pi)}_{\text{constant}} - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2$$

(σ terms; μ term)

Finding $\hat{\mu}$

min instead of max bc of negative sign.

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}}\left[\sum_{i=1}^n (x_i-\mu)^2\right]. \quad \text{Find the derivative wrt } \mu.$$

$$\nabla_\mu\left(\sum_{i=1}^n (x_i-\mu)^2\right) = \sum_{i=1}^n 2(x_i-\mu)(-1) = \sum_{i=1}^n 2(\mu - x_i) \overset{\text{set}}{=} 0 = \sum_{i=1}^n (\mu - x_i)$$

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

= sample / empiricial mean is the MLE of the mean.

Finding $\hat{\sigma}$:

$$\hat{\sigma} = \underset{\sigma}{\operatorname{argmax}} \underbrace{\left[-n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \right]}_{\text{call this } f(\sigma).}$$

$$\nabla_\sigma f(\sigma) = \frac{-n}{\sigma} - \frac{(-2)}{2\sigma^3} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \overset{\text{set}}{=} 0.$$

$$\Rightarrow \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = \frac{n}{\sigma}$$

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

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

empirical/sample variance is the MLE of the variance!

Example: Uniform distribution.

- Given $\{x_i\}_{i=1}^n$ iid drawn from (continuous) uniform distribution $\text{Uniform}([0,\theta])$:

$$p(x|\theta) = \begin{cases} 1/\theta & \text{if } x \in [0,\theta] \\ 0, & \text{otherwise.} \end{cases}, \quad \theta > 0.$$

- What is the optimal θ?

$$\rightarrow p(x|\theta) = \frac{1}{\theta} \mathbb{I}(x \in [0,\theta])$$

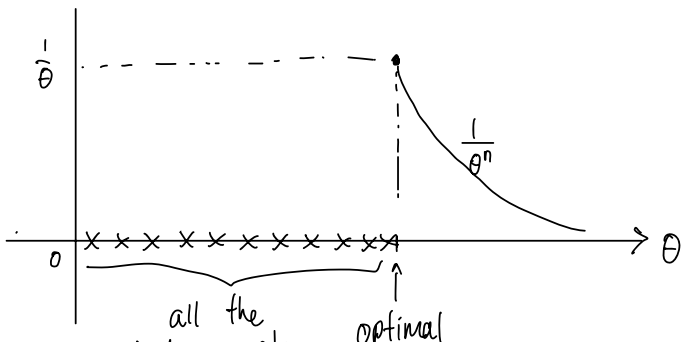
↑ indicator function: 1 if $[0,\theta]$, 0 otherwise.

$$L(\theta) = \prod_{i=1}^n p(x_i|\theta) = \prod_{i=1}^n \left(\frac{1}{\theta}\right) \mathbb{I}(x_i \in [0,\theta])$$

$$= \frac{1}{\theta^n} \prod_{i=1}^n \mathbb{I}(x_i \in [0,\theta])$$

$$= \begin{cases} \frac{1}{\theta^n} & \text{if } x_i \in [0,\theta] \ \forall\, i \\ 0, & \text{otherwise (i.e. any } x_i \text{ is not } \in [0,\theta]). \end{cases}$$

$$\Rightarrow \boxed{\hat{\theta} = \max\{x_1, \cdots, x_n\}}.$$



all the data points.

optimal point for θ, which covers the data points as tight as possible!

MLE for Regression:

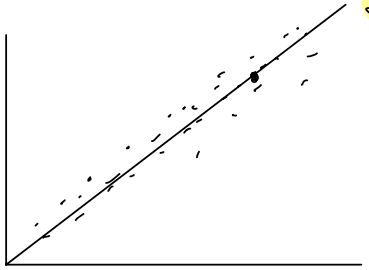
- Given $\{\vec{x}_i, y_i\}_{i=1}^n$, where $\vec{x}_i$, such that

$$P(y \mid \vec{x}_i, \vec{\theta}) = N\left(y \mid f(\vec{x}_i, \vec{\theta}), \sigma^2\right)$$ // Normal distribution.

↑ assume y is following the normal distribution.

$$\Rightarrow P(y \mid \vec{x}_i, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - f(\vec{x}_i, \vec{\theta}))^2}{2\sigma^2}\right)$$ Here x and θ are vectors.

- Estimate μ and σ^2.



$f(\vec{x}, \vec{\theta}) = \vec{\theta}\vec{x} + b$

mean_i is on the regression line,

y_i may not be on the line.

There is some noise (σ^2).

$$l(\theta, \sigma) = \sum_{i=1}^{n} \log P(y_i \mid \vec{x}_i, \vec{\theta})$$

$$= \sum_{i=1}^{n} \log \underbrace{\left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - f(x_i, \theta))^2}{2\sigma^2}\right)\right]}_{P(y_i \mid \vec{x}_i, \vec{\theta})}$$

$$= \left[\underbrace{-\frac{n}{2}\log(2\pi)}_{\text{constant}} - n \log \sigma - \sum_{i=1}^{n} \frac{1}{2\sigma^2}\left(y_i - \underbrace{f(\vec{x}_i, \theta)}_{\mu}\right)^2\right]$$

(σ: $n \log \sigma$ and $\frac{1}{2\sigma^2}$)

$$\hat{\theta} = \max_{\theta} \ell(\theta, \sigma) \Rightarrow \underbrace{\operatorname{argmin}_{\theta} \sum_{i=1}^{n} \left(y_i - f(\vec{x}_i, \hat{\theta}) \right)^2}_{\text{least square estimator.}}$$

$\therefore$ This is equivalent to minimizing the MSE.

$$\hat{\sigma}^2 = \max_{\sigma} \ell(\theta, \sigma)$$

$$= \max_{\sigma} \left(-n \log \sigma - \sum_{i=1}^{n} \frac{1}{2\sigma^2} \left(y_i - f(\vec{x}_i, \hat{\theta}) \right)^2 \right)$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(\vec{x}_i, \hat{\theta}) \right)^2$$

MLE for logistic regression

binary classification

- Given $\{\vec{x}_i, y_i\}_{i=1}^n$, where $y_i \in \{0, 1\}$, s.t.

$$p(y \mid x, \theta) = \frac{\exp(y \cdot f(x, \theta))}{1 + \exp(y \cdot f(x, \theta))}$$

- Estimate θ.

$$\left\{ \begin{array}{l} p(y = 1 \mid x, \theta) = \dfrac{\exp(1 \cdot f(x, \theta))}{1 + \exp(y \cdot f(x, \theta))} \\ p(y = 0 \mid x, \theta) = \dfrac{1}{1 + \exp(y \cdot f(x, \theta))} \end{array} \right\} \text{probability mass functions}$$

$$l(\theta) = \sum_{i=1}^{n} \log p(y_i \mid x_i, \theta).$$

$$= \sum_{i=1}^{n} \log \left[\frac{\exp(y_i \cdot f(x_i, \theta))}{1 + \exp(y_i \cdot f(x_i, \theta))} \right]$$

$$= \sum_{i=1}^{n} \left[y_i f(x_i, \theta) - \log\left(1 + \exp\left(f(x_i, \theta)\right)\right) \right]$$

$$\hat{\theta} = \arg\max_{\theta} \left\{ \sum_{i=1}^{n} \left[y_i f(x_i, \theta) - \log\left(1 + \exp\left(f(x_i, \theta)\right)\right) \right] \right\}$$

use numeric methods (such as gradient descent).

MLE part II: Theoretical properties.

- MLE estimator is an random variable

$$\hat{\theta} = \hat{\theta}(X_1 \cdots X_n) \qquad \{x_i\} \overset{iid}{\sim} p(\cdot \mid \theta_*)$$

- Evaluation metrics: Bias, variance, mean square error.
- unbiased estimators vs. Consistent estimators.

$$\text{Bias}(\hat{\theta}) = \left(\text{Expected value of } \hat{\theta}\right) - \left(\text{Actual value of } \theta\right)$$

$$= E\left[\hat{\theta}(X_1 \cdots X_n)\right] - \theta_* \quad (\theta_*: \text{true parameter})$$

$$= \int \hat{\theta}(X_1, \cdots, X_n) \underbrace{\prod_{i=1}^{n} p(x_i \mid \theta_*)}_{\text{pdf of } \hat{\theta}} dx - \theta_*$$

$$\text{Var}(\hat{\theta}) = E\left\{\left[\underbrace{\hat{\theta}(X_1, \cdots, X_n)}_{\text{estimator } \hat{\theta}} - \underbrace{E\left(\hat{\theta}(X_1, \cdots, X_n)\right)}_{\text{expectation of the estimator: } E(\hat{\theta})}\right]^2\right\}$$

Recall: $\text{Var}(X) \overset{\text{def}}{=} E\left\{(X - EX)^2\right\}$

If unbiased: $E(\hat{\theta}) = \theta_*$.

$$\text{MSE}(\hat{\theta}) = E\left\{\underbrace{\left(\hat{\theta}(X_1, \cdots, X_n) - \theta_*\right)^2}_{\text{squared error}}\right\}$$

(E: mean)

Bias-variance decomposition:

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

Proof:

mean squared error.

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta_*)^2\right] \quad \text{// definition of MSE}$$

$$= E\left\{\left[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta_*)\right]^2\right\}$$

subtract, add it back

variance, bias^2

$$= E\left[(\hat{\theta} - E(\hat{\theta}))^2 + (E(\hat{\theta}) - \theta_*)^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta_*)\right]$$

$= 0$.

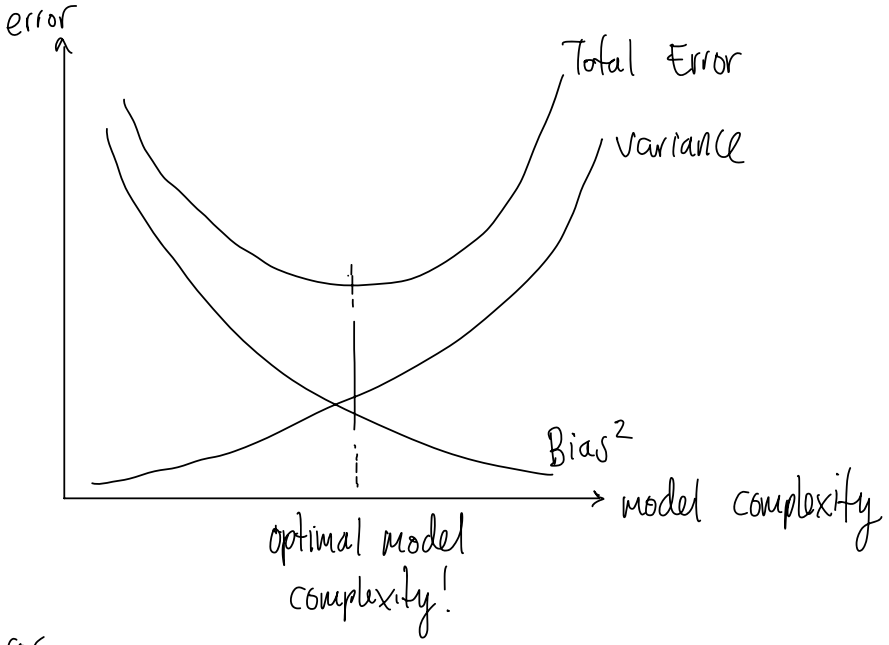
constant, move outside the expectation function.

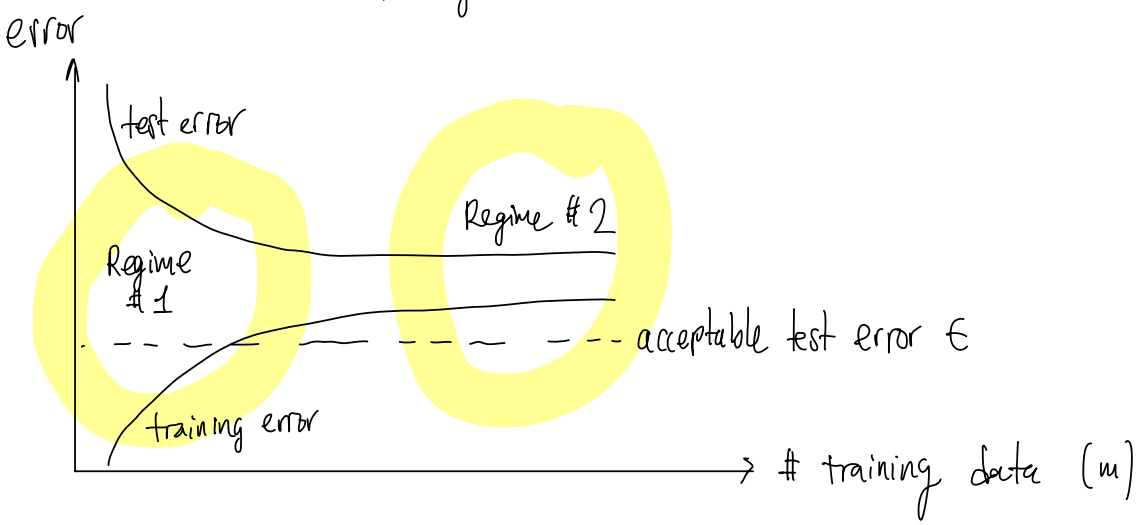
$$E\left[(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta_*)\right] = E\left[\hat{\theta} - E(\hat{\theta})\right](E(\hat{\theta}) - \theta_*)$$

bias

$$= \left[E(\hat{\theta}) - E(\hat{\theta})\right](E(\hat{\theta}) - \theta_*) \quad \text{// } E[E[X]] = E[X] \text{ (?)}$$

$$= 0.$$





Regime # 1 = High variance:

Symptoms: training error $\ll$ test error, training error $< \epsilon$, test error $> \epsilon$.

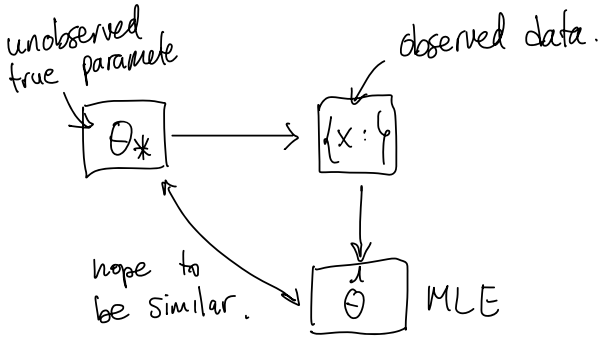
Remedy: increase m, reduce complexity of model, bagging.

Regime # 2 = High bias.

Symptoms: training error $> \epsilon$.

Remedy: increase model complexity, add features, boosting.

Framework:



If $\text{Bias}(\hat{\theta}) = 0$, $\hat{\theta}$ is "unbiased".

If $\text{MSE}(\hat{\theta}) \to 0$ as $n \to \infty$: $\hat{\theta}$ is "consistent". (implies)

If $\text{Bias} \to 0$ as $n \to \infty$: $\hat{\theta}$ is "asymptotically unbiased." (implies)

Example: Gaussian Distribution.

for $\{x_i\}_{i=1}^n \sim N(\mu, \sigma^2)$, MLE is:

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

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

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

"unbiased".

$$\hat{\sigma}^2 = \mathrm{Var}(\hat{\mu}) = E[(\hat{\mu} - \mu)^2] = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right)^2\right]$$

($\hat{\mu}$ calculated, $E(\hat{\mu}) = \mu$)

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

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

Assume independence:

$$\Rightarrow E[(x_i - \mu)]\,E[(x_j - \mu)] = 0$$

$$= \frac{n\sigma^2}{n^2} = \boxed{\frac{\sigma^2}{n}}$$

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

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

$E[\hat{\sigma}^2] \neq \sigma^2 \Rightarrow$ $\hat{\sigma}^2$ is a biased estimator for σ^2.

$$\hat{S}^2 = \underbrace{\frac{1}{n-1}}_{\substack{n-1 \text{ instead of } n.}} \sum_{i=1}^{n}(x_i - \hat{\mu})^2$$ is an unbiased estimator.

$\mathrm{bias}(\hat{\sigma}^2) \to 0$ as $n \to \infty \Rightarrow$ asymptotically unbiased.

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

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

Why MLE?

MLE is always consistent.

MLE is equivalent to minimizing Kullback-Leiber (KL) divergence (difference btwn two distributions, p and q):

$$KL(q \| p) \equiv E_q[\log q(x) - \log p(x)]$$

$$KL(p \| q) \neq KL(q \| p)$$

not symmetric!

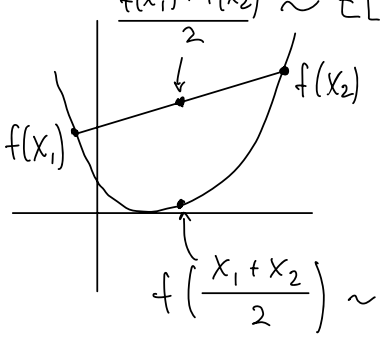
$$= \begin{cases} \sum_x q(x)\left(\log\left(\frac{q(x)}{p(x)}\right)\right) & \text{// discrete case.} \\ \int_x q(x)\left(\log\left(\frac{q(x)}{p(x)}\right)\right) dx & \text{// continous case.} \end{cases}$$

Claims

- $KL(q \| p) \geq 0$ for any q and p. (1)
- $KL(q \| p) = 0$ i.i.f. $q = p$. (2)

Proof of (1): use Jensen Inequality:

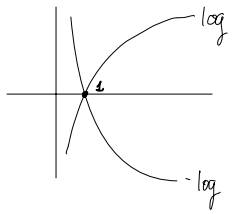
If $f(x)$ is convex, then $E_q[f(x)] \geq f(E_q[x])$



$$q(x) = \frac{dx_1 + dx_2}{2}$$

$$KL(q \| p) \equiv E_q\left[\log\left(\frac{q}{p}\right)\right]$$

↑ concave function



$$= E_q\left[\underbrace{-\log}_{-\log = \text{convex.}}\left(\frac{p(x)}{q(x)}\right)\right] \geq -\log\left(E_q\left[\frac{p(x)}{q(x)}\right]\right) \quad \text{// Jensen inequality}$$

$$= -\log\left(\sum_x q(x)\frac{p(x)}{q(x)}\right) \quad \text{// Definition of Expectation.}$$

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

This completes proof of (1). ∎

Proof of (2):

Suppose $KL(q \| p) = 0$. This means

$$0 = KL(q \| p) = -\log\left(E_q\left[\frac{p(x)}{q(x)}\right]\right) \Rightarrow \frac{p(x)}{q(x)} = \text{const.}$$

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

$$\Rightarrow \underbrace{1}_{\text{sum of probability mass} = 1} = \text{const} \cdot \underbrace{1}_{\text{sum of probability mass} = 1} \Rightarrow \text{const} = 1 \Rightarrow p = q.$$

How to connect MLE with KL Divergence?

$$KL(q \| p) = E_q\left[\log q(x) - \log p(x)\right]$$

Assume q is the data distribution, // Given

p is the "model". // want to optimize.

$$\min_\theta \left[KL(q \| p_\theta)\right] \Leftrightarrow \min_\theta E_q\left[\log q(x) - \log p(x)\right]$$

$$\Leftrightarrow \min_\theta -E_q\left[\log p_\theta(x)\right]$$

$$\Leftrightarrow \max_\theta E_q\left[\log p_\theta(x)\right]$$

$$= \max_\theta \left[\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i)\right] \quad \{x_i\} \sim q$$

average log-likelihood.

∴ Maximizing the log-likelihood is equivalent to minimizing the KL divergence btwn the data distribution q and the model p_θ.