Lecture 6: Statistical Inference

Yi, Yung (이융)

EE210: Probability and Introductory Random Processes
KAIST EE

MONTH DAY, 2021

Roadmap

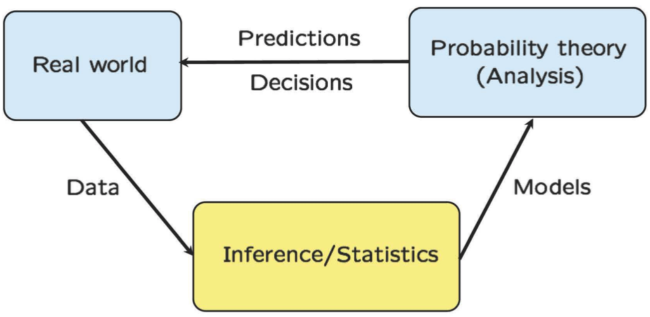
- Basics on Statistic Inference
- Framework of Bayesian Inference
- MAP (Maximum A Posteriori) Estimator
- LMS (Least Mean Squares) Estimator
- LLMS (Linear LMS) Estimator

- Framework of Classical Inference
- ML (Maximum Likelihood) Estimator

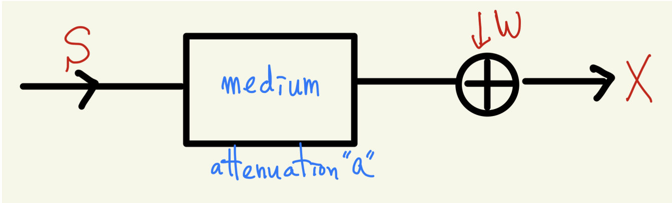
Roadmap

- Basics on Statistic Inference
- Framework of Bayesian Inference
- MAP (Maximum A Posteriori) Estimator
- LMS (Least Mean Squares) Estimator
- LLMS (Linear LMS) Estimator
- Framework of Classical Inference
- ML (Maximum Likelihood) Estimator

Inference: Big Picture



- Inference
  - Using data, probabilistic models or parameters for models are determined.
- Why building up models?
  - Analysis is possible, so that predictions and decisions are made.
- Recently, deep learning
  - Connecting big data and big model building



- $X = aS + W$
- Modeling building
  - know the original signal S, observe X
  - infer the model parameter a
- Variable estimation
  - know a, observe X
  - infer the original signal S
- Same mathematical structure, because the parameters in models are variables in many cases

- Hypothesis testing
  - Unknown: a few possible ones
  - Goal: small probability of incorrect decision
  - (Ex) Something detected on the radar. Is it a bird or an airplane?
- Estimation
  - Unknown: a value included in an infinite, typically continuous set
  - Goal: Finding the value close to the true value
  - (Ex) Biased coin with unknown probability of head $\theta \in [0, 1]$. Data of heads and tails. What is θ?
  - (Note) If you have the candidate values of $\theta = \{1/4, 1/2, 3/4\}$, then it's a hypothesis testing problem

Inference with Different Views: Bayesian vs. Classical (1)

- Biased coin with parameter θ (probability of head). Assume that $\theta \in \{1/4, 3/4\}$.

- Throw the coin 3 times and get (H, H, H). Goal: infer θ, 1/4 or 3/4?

- Distribution of θ (prior) e.g.,

$$\mathbb{P}(\theta = \frac{3}{4}) = 1/2, \quad \mathbb{P}(\theta = \frac{1}{4}) = 1/2$$

- Use Bayes' rule and find the posterior:

$$\mathbb{P}\Big[\theta = \frac{3}{4}\Big|(HHH)\Big] = \frac{27}{28}, \ \mathbb{P}\Big[\theta = \frac{1}{4}\Big|(HHH)\Big] = \frac{1}{28}$$

- Choose θ with larger posterior probability.
- *Bayesian approach* (Chapter 8)

- Find the probability of (H, H, H), if $\theta = \frac{1}{4}$ or $\frac{3}{4}$ (likelihood)

$$\mathbb{P}\Big[(HHH)|\theta = \frac{3}{4}\Big] = \Big(\frac{3}{4}\Big)^3$$

$$\mathbb{P}\Big[(HHH)|\theta = \frac{1}{4}\Big] = \Big(\frac{1}{4}\Big)^3$$

- Choose θ with a larger likelihood.
- *Classical approach* (Chapter 9)

Inference with Different Views: Bayesian vs. Classical (2)

Bayesian approach

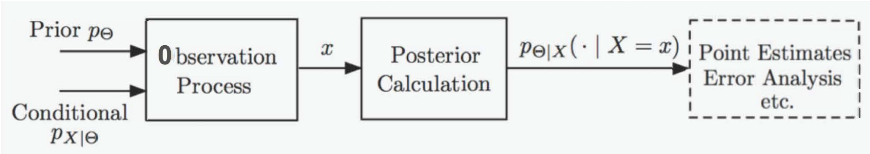
- Unknown: random variable with some distribution (prior)
- Unknown model as chosen randomly from a give model class
- Observed data x gives: posterior distribution $p_{\Theta|X}(\theta|x)$
- Choose θ with larger posterior probability (other methods exist)

Classical approach

- Unknown: deterministic value
- Unknown model as one of multiple probabilistic models
- Observed data x gives: likelihood $p(X; \theta)$
- Choose θ with larger likelihood (other methods exist)

- Who is the winner? A century-long debate (see p. 409 for discussion)

Framework of Bayesian Inference



- Unknown Θ
  - physical quantity or model parameter
  - random variable
  - prior distribution p_Θ and f_Θ
- Observations or measurements X
  - observation model $p_{X|\Theta}$ and $f_{X|\Theta}$
- That is, the joint distribution of X and Θ, $p_{X,\Theta}$ and $f_{X,\Theta}$, is given

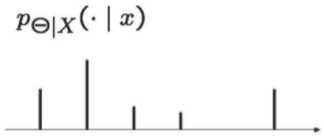
- Find the posterior distribution $p_{X|\Theta}$ and $f_{X|\Theta}$.
  - Use Bayes' rule
- Using the posterior distribution, apply one of the methods of choosing the final $\hat{\theta}$ for estimation and hypothesis testing.

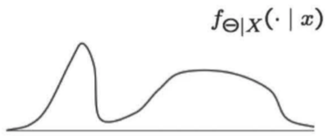
Roadmap

- Basics on Statistic Inference
- Framework of Bayesian Inference
- MAP (Maximum A Posteriori) Estimator
- LMS (Least Mean Squares) Estimator
- LLMS (Linear LMS) Estimator

- Framework of Classical Inference
- ML (Maximum Likelihood) Estimator

Point Estimation





- Given observation x, which θ are you going to choose?

M1. Choose the largest: Maximum a posteriori probability (MAP) rule

$$\boxed{\hat{\theta}_{\text{MAP}} = \arg\max_\theta p_{\Theta|X}(\theta|x), \quad \hat{\theta}_{\text{MAP}} = \arg\max_\theta f_{\Theta|X}(\theta|x)}$$

M2. Choose the mean: Conditional expectation, aka LMS (Least Mean Square)

$$\boxed{\hat{\theta}_{\text{LMS}} = \mathbb{E}[\Theta|X = x]}$$

- Why MAP and LMS are good? Not mathematically clear yet (later)

Estimator as a function

- Random observation: X
- Observation instance: x
- Estimate as a mapping from x to a number

$$\hat{\theta} = g(x), \quad \hat{\theta}_{\text{MAP}} = g_{\text{MAP}}(x), \quad \hat{\theta}_{\text{LMS}} = g_{\text{LMS}}(x)$$

- Estimator as a mapping from X to a random variable

$$\hat{\Theta} = g(X), \quad \hat{\Theta}_{\text{MAP}} = g_{\text{MAP}}(X), \quad \hat{\Theta}_{\text{LMS}} = g_{\text{LMS}}(X)$$

Example 1: Romeo and Juliet

- Romeo and Juliet start dating.
  - Romeo: late by $X \sim U[0, \theta]$.
- Unknown: θ modeled by a rv $\Theta \sim U[0, 1]$.

$$f_\Theta(\theta) = \begin{cases} 1, & 0 \le \theta \le 1 \\ 0, & \text{otherwise} \end{cases}$$

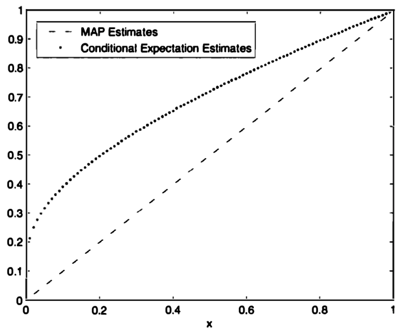
$$f_{X|\Theta}(x|\theta) = \begin{cases} \frac{1}{\theta}, & 0 \le x \le \theta \\ 0, & \text{otherwise} \end{cases}$$

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x|\theta)}{\int_0^1 f_\Theta(\theta') f_{X|\Theta}(x|\theta') d\theta'}$$
$$= \frac{1/\theta}{\int_x^1 \frac{1}{\theta'} d\theta'} = \frac{1}{\theta |\log x|}, \quad x \le \theta \le 1,$$

and $f_{\Theta|X}(\theta|x) = 0, \ \theta < x$ or $\theta > 1$.

- MAP rule
  - Given x, $f_{\Theta|X}(\theta|x)$ is decreasing in θ over $[x, 1]$.
  - $\hat{\theta}_{\text{MAP}} = x$.
- Conditional expectation estimator

$$\hat{\theta}_{\text{LMS}} = \mathbb{E}[\theta|X = x] = \int_x^1 \theta \frac{1}{\theta|\log x|} d\theta$$
$$= (1 - x)/|\log x|$$



Example 2: Biased Coin with Beta Prior (1)

- Biased coin with probability of head θ
- Unknown θ: modeled by Θ with some prior $f_\Theta(\theta)$
- Observation X: number of heads out of n tosses
- Posterior PDF

$$f_{\Theta|X}(\theta|k) = c f_\Theta(\theta) p_{X|\Theta}(k|\theta) = c \binom{n}{k} f_\Theta(\theta) \theta^k (1-\theta)^{n-k}, \ c \text{ the normalizing constant}$$

- If $\Theta \sim Beta(\alpha, \beta)$, what is $\hat{\theta}_{\text{MAP}}$?
- What is $Beta(\alpha, \beta)$?

Example 2: Biased Coin with Beta Prior (2)

Beta distribution

A continuous rv Θ follows a beta distribution with integer parameters $\alpha, \beta > 0$, if

$$f_\Theta(\theta) = \begin{cases} \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, & 0 < \theta < 1, \\ 0, & \text{otherwise}, \end{cases}$$

where $B(\alpha, \beta)$, called Beta function, is a normalizing constant, given by

$$B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$$

- A special case of $Beta(1,1)$ is $Uniform[0,1]$

Example 2: Biased Coin with Beta Prior (3)

- If $\Theta \sim Beta(\alpha,\beta)$, then $\Theta | \{X = k\} \sim Beta(k+\alpha, n-k+\beta)$
  - Very useful: Beta prior $\Longrightarrow$ Beta posterior
- Proof. For $Beta(\alpha,\beta)$ prior,

$$f_\Theta(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$f_{\Theta|X}(\theta|k) = c\binom{n}{k} f_\Theta(\theta)\theta^k(1-\theta)^{n-k} = \frac{d}{B(\alpha,\beta)} \cdot \theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}$$

where $d = c\binom{n}{k}$.

- Taking the logarithm,

$$\hat{\theta}_{\text{MAP}} = \arg\max_\theta \Big[(\alpha+k-1)\log\theta + (\beta+n-k+1)\log(1-\theta)\Big] = \frac{\alpha+k-1}{\alpha+\beta-2+n}$$

- When $\alpha = \beta = 1$ (i.e., $U[0,1]$ prior), $\hat{\theta}_{\text{MAP}} = \frac{k}{n}$

Example 3: Spam Filtering

- E-mail: spam (1) or legitimate (2), $\Theta \in \{1,2\}$, with prior $p_\Theta(1)$ and $p_\Theta(2)$.
- $\{w_1, w_2, \dots, w_n\}$: a collection of words which suggest "spam".
- For each i, a Bernoulli $X_i = 1$ if w_i appears and 0 otherwise.
- Observation model $p_{X_i|\Theta(x_i|1)}$ and $p_{X_i|\Theta(x_i|2)}$ are known. Conditioned on Θ, X_i are independent.
- Posterior PMF

$$\mathbb{P}\Big(\Theta = m|(x_1,\dots,x_n)\Big) = \frac{p_\Theta(m)\prod_{i=1}^n p_{X_i|\Theta}(x_i|m)}{\sum_{j=1,2} p_\Theta(j)\prod_{i=1}^n p_{X_i|\Theta}(x_i|j), \quad m = 1,2}$$

- MAP rule for this hypothesis testing problem. Decided that the message is spam if

$$p_\Theta(1)\prod_{i=1}^n p_{X_i|\Theta}(x_i|1) > p_\Theta(2)\prod_{i=1}^n p_{X_i|\Theta}(x_i|2)$$

MAP's Performance Guarantee

- MAP estimate is intuitive, but we need more mathematical support.
- Claim 1. For a given x, the MAP rule minimizes the probability of an incorrect decision.
- Claim 2. The MAP rule minimizes the overall probability of an incorrect decision, averaged over x.
- Proof. Let I and I_{map} be the indicator rv, representing the correct decision by any general estimator and the MAP, respectively.

$$\mathbb{E}[I|X=x] = \mathbb{P}\Big[g(X) = \Theta|X=x\Big] \leq \mathbb{P}\Big[g_{map}(X) = \Theta|X=x\Big] = \mathbb{E}[I_{map}|X=x]$$

Thus, Claim 1 holds. We now take the expectation of the above equations, the law of iterated expectations leads to Claim 2.

Roadmap

- Basics on Statistic Inference
- Framework of Bayesian Inference
- MAP (Maximum A Posteriori) Estimator
- LMS (Least Mean Squares) Estimator
- LLMS (Linear LMS) Estimator
- Framework of Classical Inference
- ML (Maximum Likelihood) Estimator

Least Mean Squares Estimator (1)

- Unknown: θ modeled by Θ with prior $f_\Theta(\cdot)$. Assume $\Theta \sim \mathit{Uniform}[4, 10]$.
- No observations available
- MAP estimate
  - Any value $\hat{\theta}_{map} \in [4, 10]$ (why? posterior = prior), not very useful
- What is your other choice?
  - Expectation: $\hat{\theta} = \mathbb{E}[\Theta] = 7$
  - looks reasonable, but why?
- Because it minimizes mean squared error (MSE)

$$\min_{\hat{\theta}} \mathbb{E}\left[(\Theta - \hat{\theta})^2\right] = \min_{\hat{\theta}} \left(\mathrm{var}(\Theta - \hat{\theta}) + \left(\mathbb{E}[\Theta - \hat{\theta}]\right)^2\right) = \min_{\hat{\theta}} \left(\mathrm{var}(\Theta) + \left(\mathbb{E}[\Theta - \hat{\theta}]\right)^2\right)$$

- minimized when $\hat{\theta} = \mathbb{E}[\Theta]$.

Least Mean Squares Estimator (2)

- Unknown: θ modeled by Θ with prior $f_\Theta(\cdot)$.
- Observation $X = x$ with model $f_{X|\Theta}(x|\theta)$
- Minimizing conditional mean squared error

$$\min_{\hat{\theta}} \mathbb{E}\Big[(\Theta - \hat{\theta})^2 | X = x\Big]$$

  - minimized when $\hat{\theta} = \mathbb{E}[\Theta | X = x]$.
  - LMS estimator $\hat{\Theta} = \mathbb{E}[\Theta | X]$
- Performance (MSE: Mean Squared Error)
  - When $X = x$, $\mathbb{E}\Big[(\Theta - \mathbb{E}[\Theta|X=x])^2 | X = x\Big] = \text{var}\Big(\Theta | X = x\Big)$
  - Averaged over X: $\mathbb{E}\Big[(\Theta - \mathbb{E}[\Theta|X])^2\Big] = \mathbb{E}\Big[\text{var}(\Theta|X=x)\Big]$

Example: Romeo and Juliet

- Romeo and Juliet start dating.
  - Romeo: late by $X \sim U[0, \theta]$.
- Unknown: θ modeled by a rv $\Theta \sim U[0, 1]$.

$$f_\Theta(\theta) = \begin{cases} 1, & 0 \le \theta \le 1 \\ 0, & \text{otherwise} \end{cases}$$

$$f_{X|\Theta}(x|\theta) = \begin{cases} \frac{1}{\theta}, & 0 \le x \le \theta \\ 0, & \text{otherwise} \end{cases}$$

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x|\theta)}{\int_0^1 f_\Theta(\theta') f_{X|\Theta}(x|\theta') d\theta'}$$

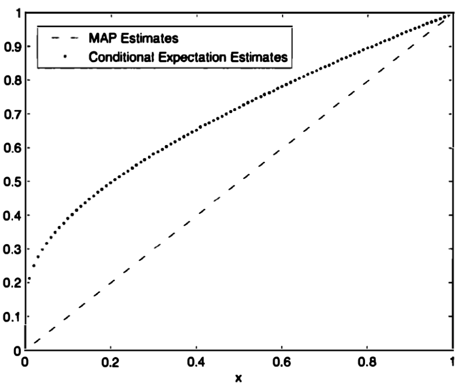
$$= \frac{1/\theta}{\int_x^1 \frac{1}{\theta'} d\theta'} = \frac{1}{\theta |\log x|}, \quad x \le \theta \le 1,$$

and $f_{\Theta|X}(\theta|x) = 0, \ \theta < x$ or $\theta > 1$.

- MAP rule
  - $\hat{\theta}_{\text{MAP}} = x$.
- LMS estimator

$$\hat{\theta}_{\text{LMS}} = \mathbb{E}[\theta | X = x] = \int_x^1 \theta \frac{1}{\theta |\log x|} d\theta$$

$$= (1 - x)/|\log x|$$



Example: Biased Coin with Beta Prior

- Remind. If $\Theta \sim Beta(\alpha, \beta)$, then $\Theta|\{X = k\} \sim Beta(k + \alpha, n - k + \beta)$
- Fact. If $\Theta \sim Beta(\alpha, \beta)$,

$$\mathbb{E}[\Theta] = \frac{1}{B(\alpha, \beta)} \int_0^1 \theta\theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = \frac{B(\alpha+1, \beta)}{B(\alpha, \beta)} = \frac{\alpha}{\alpha+\beta}$$

- Using the above fact,

$$\mathbb{E}[\Theta|X = k] = \frac{k+\alpha}{k+\alpha+n-k+\beta} = \frac{k+\alpha}{\alpha+\beta+n}$$

- For $\alpha = \beta = 1$ ($\Theta = Uniform[0, 1]$),

$$\mathbb{E}[\Theta|X = k] = \frac{k+1}{n+2}$$

Example: Signal Recovery from Noisy Measurement (1)

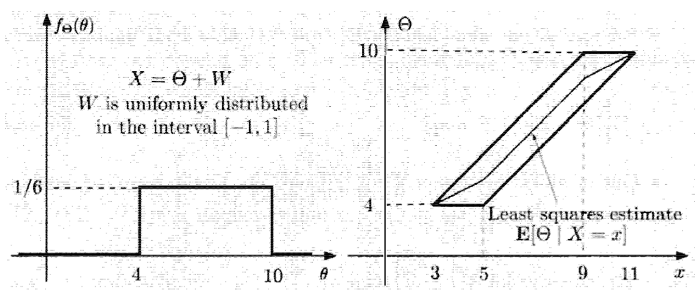
- Unknown: $\Theta \sim Uniform[4, 10]$
- Observe Θ with random error W as X. $W \sim Uniform[-1, 1]$

$$X = \Theta + W$$

- Given $\Theta = \theta$, $X = \theta + W \sim Uniform[\theta - 1, \theta + 1]$.

$$f_{\Theta,X}(\theta, x) = f_\Theta(\theta) f_{X|\Theta}(x|\theta) = \begin{cases} \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}, & \text{if } 4 \le \theta \le 10,\ \theta - 1 \le x \le \theta + 1, \\ 0, & \text{otherwise} \end{cases}$$

- $\hat{\theta}_{\mathrm{LMS}} = \mathbb{E}[\Theta|X = x]$ = midpoint of the corresponding vertical section



Example: Signal Recovery from Noisy Measurement (2)

- Unknown: $\Theta \sim \mathit{Uniform}[4, 10]$
- Observe Θ with random error W as X. $W \sim \mathit{Uniform}[-1, 1]$

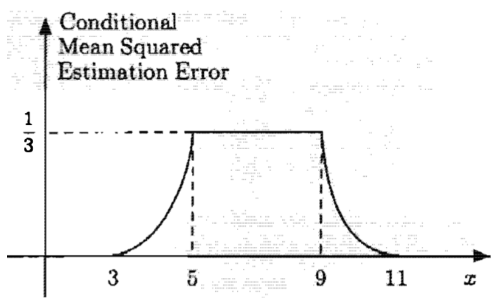
$$X = \Theta + W$$

- Given $\Theta = \theta$, $X = \theta + W \sim \mathit{Uniform}[\theta - 1, \theta + 1]$.

$$f_{\Theta,X}(\theta, x) = f_\Theta(\theta) f_{X|\Theta}(x|\theta) = \begin{cases} \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}, & \text{if } 4 \le \theta \le 10,\ \theta - 1 \le x \le \theta + 1, \\ 0, & \text{otherwise} \end{cases}$$

- Conditional MSE

$$\mathbb{E}\Big[(\Theta - \mathbb{E}[\Theta|X = x])^2 | X = x\Big]$$



Hardness of LMS Estimation

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x|\theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta') f_{X|\Theta}(x|\theta') d\theta'$$

- Observation model $f_{X|\Theta}(x|\theta)$ may not be always available
- Finding the posterior distribution is hard for multi-dimensional Θ
- Θ is very often high-dimensional, especially in the era of big data and deep learning
  - AlexNet in image recognition: 61M parameters (though not a Bayesian inference)
- Any alternative to LMS estimator?

Roadmap

- Basics on Statistic Inference
- Framework of Bayesian Inference
- MAP (Maximum A Posteriori) Estimator
- LMS (Least Mean Squares) Estimator
- LLMS (Linear LMS) Estimator

- Framework of Classical Inference
- ML (Maximum Likelihood) Estimator

Linear LMS (LLMS) Estimator: Approach

- Give up optimality, but choose a simple, but good one.
- General estimators $\hat{\Theta} = g(X)$, LMS estimator $\hat{\Theta}_{LMS} = \mathbb{E}[\Theta|X]$
- We consider a restricted class of $g(X)$: $\hat{\Theta} = \boxed{aX + b}$.
- Our goal is:

$$\min_{a,b} \mathbb{E}\left[(\Theta - aX - b)^2\right]$$

- Linear models are always the first choice for a simple design in engineering.

Linear LMS (LLMS) Estimator: Solution First

KAIST EE

LLMS

$$\hat{\Theta}_L = \mathbb{E}(\Theta) + \frac{\text{cov}(\Theta, X)}{\text{var}(X)}\Big(X - \mathbb{E}(X)\Big) = \mathbb{E}(\Theta) + \rho\frac{\sigma_\Theta}{\sigma_X}\Big(X - \mathbb{E}(X)\Big)$$

- No distributions on Θ and X: only means, variances, and covariances
- MSE $\mathbb{E}[(\hat{\Theta}_L - \Theta)^2]$? Assume $\mathbb{E}[\Theta] = \mathbb{E}[X] = 0$. $\mathbb{E}\Big[(\Theta - \rho\frac{\sigma_\Theta}{\sigma_X}X)^2\Big] = (1-\rho^2)\text{var}[\Theta]$
  - Uncertainty about Θ decreases by the factor of $1-\rho^2$
  - What happens if $|\rho| = 1$ or $\rho = 0$?

- If $\rho > 0$:
  - Baseline ($\mathbb{E}[\Theta]$) + correction term
  - If $X > \mathbb{E}[X] \implies \hat{\Theta}_L > \mathbb{E}[\Theta]$
  - If $X < \mathbb{E}[X] \implies \hat{\Theta}_L < \mathbb{E}[\Theta]$

- If $\rho = 0$ (uncorrelated):
  - Just baseline ($\mathbb{E}[\Theta]$)
  - $\hat{\Theta}_L = \mathbb{E}[\Theta]$
  - No use of data X

Linear LMS (LLMS) Estimator: Proof

KAIST EE

$$\hat{\Theta}_L = \mathbb{E}(\Theta) + \frac{\text{cov}(\Theta, X)}{\text{var}(X)}\Big(X - \mathbb{E}(X)\Big) \tag{1}$$

$$= \mathbb{E}(\Theta) + \rho\frac{\sigma_\Theta}{\sigma_X}\Big(X - \mathbb{E}(X)\Big) \tag{2}$$

$$\min_{a,b} \text{ERR}(a,b) = \min_{a,b} \mathbb{E}\Big[(\Theta - aX - b)^2\Big]$$

- Assume a was found.

$$\mathbb{E}\Big[(Y-b)^2\Big], \quad Y = \Theta - aX$$

- Minimized when $b = \mathbb{E}(Y) = \mathbb{E}(\Theta) - a\mathbb{E}(X)$.

$$\text{ERR}(a,b) = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \text{var}(Y)$$

$$= \text{var}[\Theta] + a^2\text{var}[X] - 2a\text{cov}(\Theta, X) \tag{3}$$

- (3) is minimized when $a = \frac{\text{cov}(\Theta,X)}{\text{var}[X]}$. Then,

$$\hat{\Theta}_L = aX + b = aX + \mathbb{E}(\Theta) - a\mathbb{E}(X)$$
$$= (1)$$

- Using $\rho = \frac{\text{cov}(\Theta,X)}{\sigma_\Theta\sigma_X}$, we get:

$$a = \frac{\rho\sigma_\Theta\sigma_X}{\sigma_X^2} = \frac{\rho\sigma_\Theta}{\sigma_X}$$

- Then, we have (2).

Example: Romeo and Juliet

- Romeo and Juliet start dating. Romeo: late by $X \sim U[0, \theta]$.
- Unknown: θ modeled by a rv $\Theta \sim U[0, 1]$.
- $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\Theta]] = \mathbb{E}[\Theta/2] = 1/4$
- Using $\mathbb{E}[\Theta] = 1/2$ and $\mathbb{E}[\Theta^2] = 1/3$,

$$\begin{aligned}\text{var}[X] &= \mathbb{E}[\text{var}[X|\Theta]] + \text{var}[\mathbb{E}[X|\Theta]] \\ &= \frac{1}{12}\mathbb{E}[\Theta^2] + \frac{1}{4}\text{var}[\Theta] = \frac{7}{144}\end{aligned}$$

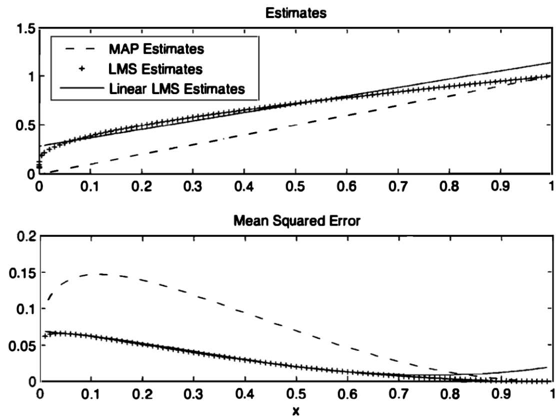
- $\text{cov}(\Theta, X) = \mathbb{E}[\Theta X] - \mathbb{E}[\Theta]\mathbb{E}[X]$

$$\begin{aligned}\mathbb{E}[\Theta X] &= \mathbb{E}[\mathbb{E}[\Theta X|\Theta]] = \mathbb{E}[\Theta \mathbb{E}[X|\Theta]] \\ &= \mathbb{E}[\Theta^2/2] = 1/6\end{aligned}$$

$$\text{cov}(\Theta, X) = 1/6 - 1/2 \cdot 1/4 = 1/24$$

- LLMS estimator is:

$$\begin{aligned}\hat{\Theta}_L &= \mathbb{E}(\Theta) + \frac{\text{cov}(\Theta, X)}{\text{var}(X)}\Big(X - \mathbb{E}(X)\Big) \\ &= \frac{1}{2} + \frac{1/24}{7/144}(X - \frac{1}{4}) = \frac{6}{7}X + \frac{2}{7}\end{aligned}$$



Example: Biased Coin with Uniform Prior

- Biased coin with probability of head θ
- Unknown $\Theta \sim uniform[0, 1]$,
  - $\mathbb{E}[\Theta] = 1/2$, $\text{var}[X] = 1/12$
- n tosses, X: number of heads.
- $p_{X|\Theta}(k|\theta)$: $Binomial(n, \theta)$
- $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\Theta]] = \mathbb{E}[n\Theta] = n/2$

$$\begin{aligned}\text{var}(X) &= \mathbb{E}[\text{var}(X|\Theta)] + \text{var}(\mathbb{E}[X|\Theta]) \\ &= \mathbb{E}[n\Theta(1-\Theta)] + \text{var}[n\Theta] \\ &= \frac{n}{2} - \frac{n}{3} + \frac{n^2}{12} = \frac{n(n+2)}{12}\end{aligned}$$

$$\text{cov}(\Theta, X) = \mathbb{E}[\Theta X] - \mathbb{E}[\Theta]\mathbb{E}[X] = \mathbb{E}[\Theta X] - n/4$$

$$\begin{aligned}\mathbb{E}[\Theta X] &= \mathbb{E}[\mathbb{E}[\Theta X|\Theta]] = \mathbb{E}[\Theta \mathbb{E}[X|\Theta]] \\ &= \mathbb{E}[n\Theta^2] = n/3\end{aligned}$$

$$\text{cov}(\Theta, X) = \tfrac{n}{3} - \tfrac{n}{4} = \tfrac{12}{n}$$

$$\hat{\Theta}_L = \frac{1}{2} + \frac{n/12}{n(n+2)/12}(X - \frac{n}{2}) = \frac{X+1}{n+2}$$

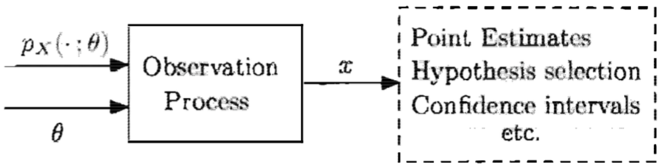
- What was the LMS estimator? $\frac{X+1}{n+2}$
- Same! Intuitive?

Yes, because the LMS esitmator was linear.

Roadmap

- Basics on Statistic Inference
- Framework of Bayesian Inference
- MAP (Maximum A Posteriori) Estimator
- LMS (Least Mean Squares) Estimator
- LLMS (Linear LMS) Estimator
- Framework of Classical Inference
- ML (Maximum Likelihood) Estimator

Framework of Classical Inference (1)



- Unknown θ
  - deterministic (not random) quantity (thus, no prior distribution)
  - No prior, No posterior probabilities
- Observations or measurements X
  - Random observation X's distribution just depends on θ
  - Notation: $p_X(x;\theta)$ and $f_X(x;\theta)$, θ-parameterized distribution of observations
- Choosing one among multiple probabilistic models
  - Each θ corresponds to a probabilistic model

Framework of Classical Inference (2)

- Problem types
  - Estimation
  - Hypothesis testing
  - Significance testing
- Key inference methods
  - ML (Maximum Likelihood) estimation
  - Linear regression
  - Likelihood ratio test
  - Significant testing
- Just a taste in this course due to time constraint.

Maximum Likelihood Estimation

- Random observation $x = (x_1, x_2, \ldots, x_n)$ of $X = (X_1, X_2, \ldots, X_n)$
  - Assume a scalar θ and a vector of observation in this lecture.
- $\boxed{\text{Likelihood } p_X(x_1, x_2, \ldots, x_n; \theta)}$
  - $p_X(x_1, x_2, \ldots, x_n; \theta)$
    - NOT the probability that the unknown parameter is equal to θ.
    - but, the probability that the observed value x arises when the parameter is θ.
  - ML (Maximum Likelihood) estimation

$$\hat{\theta}_{ml} = \arg\max_{\theta} p_X(x_1, x_2, \ldots, x_n; \theta)$$

- Very often, X_i are independent. Then, ML equals to maximizing the log-likelihood:

$$\log p_X(x_1, x_2, \ldots, x_n; \theta) = \log \prod_{i=1}^{n} p_{X_i}(x_i; \theta) = \sum_{i=1}^{n} \log p_{X_i}(x_i; \theta)$$

ML vs. MAP

- ML and MAP: How are they related?
- MAP in the Bayesian inference

$$\hat{\theta}_{map} = \arg\max_{\theta} p_{\Theta|X}(\theta|x) = \arg\max_{\theta} \frac{p_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{p_X(x)} = \frac{1}{p_X(x)} \arg\max_{\theta} p_{X|\Theta}(x|\theta)p_{\Theta}(\theta)$$

- ML in the classical inference

$$\hat{\theta}_{ml} = \arg\max_{\theta} p_X(x;\theta)$$

- $p_{X|\Theta}(x|\theta)$ in the Bayesian setting corresponds to $p_X(x;\theta)$ in the classical setting.
- When Θ is uniform (complete ignorance of Θ), MAP == ML

Example: Romeo and Juliet

- Romeo and Juliet start dating. Romeo: late by $X \sim U[0,\theta]$.
- Unknown: θ modeled by a rv $\Theta \sim U[0,1]$.
- MAP: $\hat{\theta}_{\text{MAP}} = x$
- LMS: $\hat{\theta}_{\text{LMS}} = (1-x)/|\log x|$
- LLMS: $\hat{\theta}_{\text{L}} = \frac{6}{7}x + \frac{2}{7}$
- ML: $\hat{\theta}_{\text{ML}} = \hat{\theta}_{\text{MAP}} = x$

Example: Estimation of Parameter of Exponential rv

- n identical, independent exponential rvs, $X_1, X_2, \ldots, X_n$ with parameter θ.
- Observation $x_1, x_2, \ldots, x_n$
- What is the ML estimate of θ?
- Reminder. $X \sim \exp(\lambda)$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad \mathbb{E}[X] = 1/\lambda$$

- Any guess? $\hat{\theta}_{\text{ML}} = \frac{n}{x_1 + x_2 \ldots x_n}$

$$\arg\max_{\theta} f_X(x;\theta) = \arg\max_{\theta} \prod_{i=1}^{n} \theta e^{-\theta x_i} = \arg\max_{\theta} \left(n \log\theta - \theta \sum_{i=1}^{n} x_i \right)$$

Questions?

1) What is statistical inference?
2) Draw the building blocks of Bayesian inference and explain how it works.
3) What are MAP and LMS estimators and their underlying philosophies?
4) What is LLMS estimator and why is it useful?
5) Compare the classical and Bayesian inference.
6) What is the ML estimator and how is it related to the MAP estimator?