

## Lecture 9: Introduction to Statistical Inference

Yi, Yung (이웅)

EE210: Probability and Introductory Random Processes  
KAIST EE

August 31, 2021

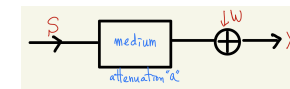
- (1) Overview on Statistical Inference
- (2) Bayesian Inference: Framework
- (3) Examples
- (4) MAP (Maximum A Posteriori) Estimator
- (5) LMS (Least Mean Squares) Estimator
- (6) LLMS (Linear LMS) Estimator
- (7) Classical Inference: ML Estimator

August 31, 2021 1 / 67

August 31, 2021 2 / 67

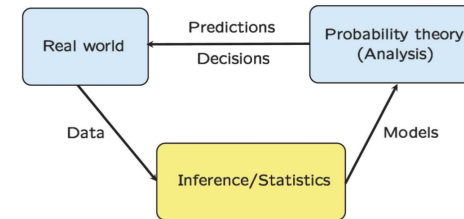
- (1) Overview on Statistical Inference
- (2) Bayesian Inference: Framework
- (3) Examples
- (4) MAP (Maximum A Posteriori) Estimator
- (5) LMS (Least Mean Squares) Estimator
- (6) LLMS (Linear LMS) Estimator
- (7) Classical Inference: ML Estimator

- Examples
  - Take 1000 voters uniformly at random, and count the popularity of each candidate to infer the true popularity.
  - COVID-19 has spread over a collection of people, and we collect a sample of COVID-19 infectees to infer the true source of infection.
  - When an original signal  $S$  is transmitted over the KAIST Wi-Fi connection, the received signal  $X$  becomes  $X = aS + W$ , where  $0 < a < 1$  and  $W \sim \mathcal{N}(0, 1)$ . If we have 10 samples of  $(S, X)$  values, what is the inferred value of  $a$ ?



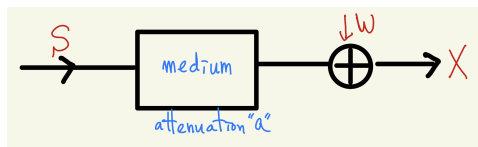
- Process of extracting information about an **unknown variable** or an **unknown model** from **noisy available data**

1. Samples are likely to be a good representation of the unknown
2. There exists uncertainty (i.e., noise) as to how well the sample represents the unknown
3. How to obtain samples has impact on inference (e.g., when we need to pay for online surveys)



Source: Introduction to Probability course by MIT

- 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



- $X = aS + W$
- Model 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  $\theta$ ?
  - (Note) If you have the candidate values of  $\theta = \{1/4, 1/2, 3/4\}$ , then it's a hypothesis testing problem

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

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

- Distribution of  $\theta$  (**prior**) e.g.,

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

- Use Bayes' rule and find the **posterior**:

$$\mathbb{P}\left[\theta = \frac{3}{4} \mid (HHH)\right] = \frac{27}{28}, \quad \mathbb{P}\left[\theta = \frac{1}{4} \mid (HHH)\right] = \frac{1}{28}$$

- Choose  $\theta$  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**)

$$\mathbb{P}\left[(HHH) \mid \theta = \frac{3}{4}\right] = \left(\frac{3}{4}\right)^3$$

$$\mathbb{P}\left[(HHH) \mid \theta = \frac{1}{4}\right] = \left(\frac{1}{4}\right)^3$$

- Choose  $\theta$  with a larger likelihood.
- **Classical approach** (Chapter 9)

(Note) There are other inference methods, and here we just show examples.

### Bayesian approach

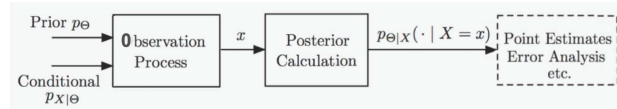
- **Unknown**: random variable with some distribution (prior)
- Unknown model as chosen randomly from a given model class
- Observed data  $x$  gives:
  - **posterior distribution**  $p_{\Theta|X}(\theta|x)$
- Choose  $\theta$  with larger posterior probability (**other methods exist**)

### Classical approach

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

- Fundamental difference about the nature of unknown models or variables
- Random variable or deterministic quantity
- Who is the winner? A century-long debate
- Example of debate: mass of the electron by noisy measurement
  - **Classical**. while unknown, it is a constant and there is no justification for modeling it as a random variable.
  - **Bayesian**. Prior distribution reflects our state of knowledge, e.g., some range of candidate values from our previous noisy measurements.
- Particular prior? too arbitrary vs. every statistical procedure's hidden choices
- Practical issues: Bayesian approach is often computationally intractable (multi-dimensional integrals)

- (1) Overview on Statistical Inference
- (2) **Bayesian Inference: Framework**
- (3) Examples
- (4) MAP (Maximum A Posteriori) Estimator
- (5) LMS (Least Mean Squares) Estimator
- (6) LLMS (Linear LMS) Estimator
- (7) Classical Inference: ML Estimator



- Unknown  $\Theta$ 
  - physical quantity or model parameter
  - random variable
  - **prior** distribution  $p_{\Theta}$  and  $f_{\Theta}$
- Observations or measurements  $X$ 
  - **observation model**  $p_{X|\Theta}$  and  $f_{X|\Theta}$
- That is, the **joint distribution** of  $X$  and  $\Theta$  ( $p_{X,\Theta}(x, \theta)$  and  $f_{X,\Theta}(x, \theta)$ ) is given
- Find the **posterior** distribution  $p_{\Theta|X}$  and  $f_{\Theta|X}$ , using Bayes' rule.
- The posterior distribution is the complete answer of the Bayesian inference.
- However, one may use it for further processing, depending on what he/she wants, e.g., point estimation.
- **Multiple** observations and **multiple** parameters are possible
  - $X = (X_1, \dots, X_n)$
  - $\Theta = (\Theta_1, \dots, \Theta_n)$

L9(2)

August 31, 2021 13 / 67

- $\Theta$ : discrete,  $X$ : discrete

$$p_{\Theta|X}(\theta|x) = \frac{p_{\Theta}(\theta)p_{X|\Theta}(x|\theta)}{p_X(x)}$$

$$p_X(x) = \sum_{\theta'} p_{\Theta}(\theta')p_{X|\Theta}(x|\theta')$$

- $\Theta$ : continuous,  $X$ : continuous

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

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

- $\Theta$ : discrete,  $X$ : continuous

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

$$f_X(x) = \sum_{\theta'} p_{\Theta}(\theta')f_{X|\Theta}(x|\theta')$$

- $\Theta$ : continuous,  $X$ : discrete

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

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

L9(2)

August 31, 2021 14 / 67

- (1) Overview on Statistical Inference
- (2) Bayesian Inference: Framework
- (3) **Examples**
- (4) MAP (Maximum A Posteriori) Estimator
- (5) LMS (Least Mean Squares) Estimator
- (6) LLMS (Linear LMS) Estimator
- (7) Classical Inference: ML Estimator

L9(3)

August 31, 2021 15 / 67

- Romeo and Juliet start dating, where Romeo is late by  $X \sim \mathcal{U}[0, \theta]$ .

- Unknown:  $\theta$  modeled by a rv  $\Theta \sim \mathcal{U}[0, 1]$ .

- Observation: Romeo was late by  $x$ .

- Prior and observation model (likelihood)

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

- Posterior

$$f_{\Theta|X}(\theta|x) = \frac{f_{\Theta}(\theta)f_{X|\Theta}(x|\theta)}{\int_0^1 f_{\Theta}(\theta')f_{X|\Theta}(x|\theta')d\theta'} = \begin{cases} \frac{1/\theta}{\int_x^1 \frac{1}{\theta'}d\theta'} = \frac{1}{\theta|\log x|}, & x \leq \theta \leq 1, \\ 0, & \theta < x \text{ or } \theta > 1 \end{cases}$$

L9(3)

August 31, 2021 16 / 67

- What happens if we have more observation samples?
  - Romeo was late  $n$  times by  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ ,  $X_i \sim \mathcal{U}[0, \theta]$ .
  - $X_1, \dots, X_n$  are conditionally independent, given  $\Theta = \theta$ .
  - Unknown:  $\theta$  modeled by a rv  $\Theta \sim \mathcal{U}[0, 1]$ .
  - Observation: Romeo was late  $n$  times by  $\mathbf{x} = (x_1, x_2, \dots, x_n)$
  - See Example 8.2 at pp. 414 for more detailed treatment.

- 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.
- Assumption: Conditioned on  $\Theta$ ,  $X_i$  are independent.
- Posterior PMF

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

- Biased coin with probability of head  $\theta$
- Unknown  $\theta$ : modeled by  $\Theta$  with some prior  $f_{\Theta}(\theta)$
- Observation  $X$ : number of heads out of  $n$  tosses
- **Question.** Suppose that you have freedom to choose the form of the prior distribution. What prior will you choose? Requirement of “good” priors?
- We will look at the prior whose distribution is something called the **Beta distribution**.

## Beta distribution

A continuous rv  $\Theta$  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

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

- See <https://youtu.be/8yaRt24qA1M> for the integration in the Beta function formula.
- A special case of Beta(1, 1) is  $\mathcal{U}[0, 1]$

- If  $\Theta \sim \text{Beta}(\alpha, \beta)$ , then  $\Theta | \{X = k\} \sim \text{Beta}(k + \alpha, n - k + \beta)$
- In other words, Beta prior  $\implies$  Beta posterior (why useful?)

**Proof.**

(a) First, the posterior pdf is given by:

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

(b) Next, for  $\text{Beta}(\alpha, \beta)$  prior,  $f_{\Theta}(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$ .

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

L9(3)

August 31, 2021 21 / 67

◦ Inference of a parameter  $\theta$

- **Single** observation
- $X$ : noisy observation of  $\theta$ , modeled as:  
 $X = \theta + W$ , where  $W \sim \mathcal{N}(0, \sigma^2)$
- Model  $\theta$  with a rv  $\Theta \sim \mathcal{N}(x_0, \sigma_0^2)$   
(normal prior)
- $\Theta$  and  $W$  are independent
- **Question.** Given an observation  $x$ , what is the posterior  $f_{\Theta|X}(\theta|x)$ ?

- **Multiple**  $n$  observations
- $n$  observations of  $\theta$ :  $W_i \sim \mathcal{N}(0, \sigma_i^2)$   
 $X_1 = \theta + W_1, \quad W_1 \sim \mathcal{N}(0, \sigma_1^2)$   
 $\vdots$   
 $X_n = \theta + W_n, \quad W_n \sim \mathcal{N}(0, \sigma_n^2)$
- Model  $\theta$  with  $\Theta \sim \mathcal{N}(x_0, \sigma_0^2)$
- $\Theta, W_1, \dots, W_n$  are independent
- **Question.** Given an observation  $x$ , what is the posterior  $f_{\Theta|X}(\theta|x)$ ?
  - $X = (X_1, \dots, X_n)$  and  $x = (x_1, \dots, x_n)$ ,

L9(3)

August 31, 2021 22 / 67

**Lemma.** Up to recaling, the pdf of the form  $e^{-\frac{1}{2}(ax^2 - 2bx + c)}$  is  $\mathcal{N}(\frac{b}{a}, \frac{1}{a})$ .

- **(Rough) Proof.** Note that the pdf of  $\mathcal{N}(\mu, \sigma^2)$ :  $f_X(x) = e^{-(x-\mu)^2/2\sigma^2}$  up to rescaling. Then,
  - $-\frac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2) = -\frac{1}{2}(ax^2 - 2bx + c)$
  - Thus,  $\sigma^2 = \frac{1}{a}$  and  $\frac{\mu}{\sigma^2} = b \implies \mu = b\sigma^2 = \frac{b}{a}$

L9(3)

August 31, 2021 23 / 67

**Theorem.** The product of two Gaussian pdfs  $\mathcal{N}(\mu_0, \nu_0)$  and  $\mathcal{N}(\mu_1, \nu_1)$  is  $\mathcal{N}\left(\frac{\nu_1\mu_0 + \nu_0\mu_1}{\nu_0 + \nu_1}, \frac{\nu_0\nu_1}{\nu_0 + \nu_1}\right)$ .

**Proof.** Using the Lemma in the previous slide, i.e., up to recaling, the pdf of the form  $e^{-\frac{1}{2}(ax^2 - 2bx + c)}$  is  $\mathcal{N}(\frac{b}{a}, \frac{1}{a})$ ,

$$\begin{aligned} & \exp(-(x - \mu_0)^2/2\nu_0) \times \exp(-(x - \mu_1)^2/2\nu_1) \\ &= \exp\left[-\frac{1}{2}\left(\left(\frac{1}{\nu_0} + \frac{1}{\nu_1}\right)x^2 - 2\left(\frac{\mu_0}{\nu_0} + \frac{\mu_1}{\nu_1}\right)x + c\right)\right] \\ &\implies \mathcal{N}\left(\nu\left(\frac{\mu_0}{\nu_0} + \frac{\mu_1}{\nu_1}\right), \overbrace{\frac{1}{\nu_0^{-1} + \nu_1^{-1}}}^{\nu}\right) = \mathcal{N}\left(\frac{\nu_1\mu_0 + \nu_0\mu_1}{\nu_0 + \nu_1}, \frac{\nu_0\nu_1}{\nu_0 + \nu_1}\right) \end{aligned}$$

L9(3)

August 31, 2021 24 / 67

**Theorem.** The product of  $n + 1$  Gaussian pdfs  $\mathcal{N}(\mu_0, \nu_0)$ ,  $\mathcal{N}(\mu_1, \nu_1), \dots, \mathcal{N}(\mu_n, \nu_n)$ , is  $\mathcal{N}(\mu, \nu)$ , where

$$\mu = \frac{\sum_{i=0}^n \frac{\mu_i}{\nu_i}}{\sum_{i=0}^n \frac{1}{\nu_i}}, \quad \nu = \frac{1}{\sum_{i=0}^n \frac{1}{\nu_i}}$$

L9(3)

August 31, 2021 25 / 67

- $n$  observations of  $\theta$ :  $W_i \sim \mathcal{N}(0, \sigma_i^2)$ , and  $\theta$  with the normal prior  $\Theta \sim \mathcal{N}(x_0, \sigma_0^2)$

$$X_i = \theta + W_i, \quad W_i \sim \mathcal{N}(0, \sigma_i^2), \quad i = 1, \dots, n$$

- $\Theta, W_1, \dots, W_n$  are independent and let  $X = (X_1, \dots, X_n)$ ,  $x = (x_1, \dots, x_n)$ .
- Our interest. The **posterior pdf**  $f_{\Theta|X}(\theta|x)$ .

- Prior.**  $f_{\Theta}(\theta) = c_1 \cdot \exp \left\{ -\frac{(\theta - x_0)^2}{2\sigma_0^2} \right\}$

- Observation model.** Noting that  $X_1, X_2, \dots, X_n$  are independent,

$$f_{X|\Theta}(x|\theta) = c_2 \cdot \exp \left\{ -\frac{(\theta - x_1)^2}{2\sigma_1^2} \right\} \cdots \exp \left\{ -\frac{(\theta - x_n)^2}{2\sigma_n^2} \right\}$$

L9(3)

August 31, 2021 26 / 67

- Numerator:**  $f_{\Theta}(\theta)f_{X|\Theta}(x|\theta) = c_1 c_2 \cdot \exp \left\{ -\sum_{i=0}^n \frac{(x_i - \theta)^2}{2\sigma_i^2} \right\}$ , which can be reexpressed as the following, using the **product of  $n + 1$  Gaussians**:

$$c_1 c_2 \cdot \exp \left\{ -\sum_{i=0}^n \frac{(x_i - \theta)^2}{2\sigma_i^2} \right\} = d \cdot \exp \left\{ -\frac{(\theta - m)^2}{2\nu} \right\},$$

$$\text{where } m = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}, \quad \nu = \frac{1}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

- Denominator:** just a constant, not a function of  $\theta$

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

L9(3)

August 31, 2021 27 / 67

- Thus, the posterior pdf  $f_{\Theta|X}(\theta|x) = a \cdot \exp \left\{ -\frac{(\theta - m)^2}{2\nu} \right\}$ , where

$$m = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}, \quad \nu = \frac{1}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

- Prior: Normal, Posterior: Normal**
- Special case when  $\sigma^2 = \sigma_0^2 = \sigma_1^2 = \dots = \sigma_n^2$ . Then,

$$m = \frac{x_0 + x_1 + \dots + x_n}{n + 1}, \quad \nu = \frac{\sigma^2}{n + 1}$$

- the prior mean  $x_0$  acts just as **another observation**.
- the standard deviation of the posterior goes to 0, at the rough rate of  $1/\sqrt{n}$ .

L9(3)

August 31, 2021 28 / 67

- **Recursive inference** is possible.
- Suppose that after  $X_1, \dots, X_n$  are observed, an additional observation  $X_{n+1}$  is observed.
- Instead of solving the inference problem from scratch, we can **view  $f_{\Theta|X_1, \dots, X_n}$  as our prior**, use the new observation to obtain the new posterior  $f_{\Theta|X_1, \dots, X_n, X_{n+1}}$
- In the example of parameter inference with the Normal prior, with the new observation  $x_{n+1} \sim \mathcal{N}(x_{n+1}, \sigma_{n+1}^2)$ , the posterior pdf is nothing but the Normal pdf of:

$$\text{mean} = \frac{(m/v) + (x_{n+1}/\sigma_{n+1}^2)}{(1/v) + (1/\sigma_{n+1}^2)}, \quad \text{variance} = \frac{1}{(1/v) + (1/\sigma_{n+1}^2)}$$

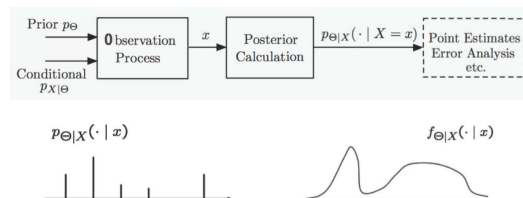
L9(3)

August 31, 2021 29 / 67

- (1) Overview on Statistical Inference
- (2) Bayesian Inference: Framework
- (3) Examples
- (4) **MAP (Maximum A Posteriori) Estimator**
- (5) LMS (Least Mean Squares) Estimator
- (6) LLMS (Linear LMS) Estimator
- (7) Classical Inference: ML Estimator

L9(4)

August 31, 2021 30 / 67



- **Point Estimate**

- Given observation  $x$ , which **single** value  $\theta$  are you going to choose as your inference result? People often want just the summary and a simple answer.
- Very often,  $\theta$ , our inference target, is by nature a single value, i.e., mass of the electron.

L9(4)

August 31, 2021 31 / 67



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

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

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

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

- Why MAP and LMS are good? Not mathematically clear yet (We will discuss later)
- Notation: The community uses  $\hat{\theta}$  to mean the estimated value, i.e., **hat** for estimated value.

L9(4)

August 31, 2021 32 / 67



- 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

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

From now on we focus on the MAP estimate, mainly based on the examples that we've discussed in the previous section.

Slide 16 for more details

- Romeo and Juliet start dating, where Romeo is late by  $X \sim \mathcal{U}[0, \theta]$ .
- Unknown:  $\theta$  modeled by a rv  $\Theta \sim \mathcal{U}[0, 1]$ .
- Observation: Romeo was late by  $x$ .
- **Question.** Given the observation sample  $x$ , what is  $\hat{\theta}_{\text{MAP}}$ ?
- **Intuition.** As  $x$  grows,  $\hat{\theta}_{\text{MAP}}$  decreases or increases? **Increases. Why?**
- Posterior:  $f_{\Theta|X}(\theta|x) = \begin{cases} \frac{1}{\theta|\log x|}, & x \leq \theta \leq 1, \\ 0, & \theta < x \text{ or } \theta > 1 \end{cases}$
- Given  $x$ ,  $f_{\Theta|X}(\theta|x)$  is decreasing in  $\theta$  over  $[x, 1]$ .  $\Rightarrow \hat{\theta}_{\text{MAP}} = x$ .

Slide 18 for more details

- 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.
- Assumption: Conditioned on  $\Theta$ ,  $X_i$  are independent.
- Posterior PMF

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

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

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

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

- If  $\Theta \sim \text{Beta}(\alpha, \beta)$ , then  $\Theta|X=k \sim \text{Beta}(k + \alpha, n - k + \beta)$
- $f_{\Theta|X}(\theta|k) \propto \theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}$

- MAP estimate: Taking the logarithm,

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

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

L9(4)

August 31, 2021 37 / 67

- The posterior pdf  $f_{\Theta|X}(\theta|X) = a \cdot \exp \left\{ -\frac{(\theta - m)^2}{2v} \right\}$ , where

$$m = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}, \quad v = \frac{1}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

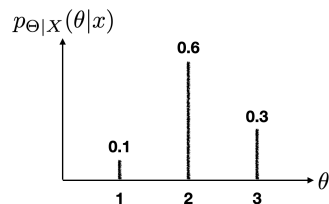
- The pdf is normal, so it is maximized when  $\theta = \text{mean}$ .
- Thus,  $\hat{\theta}_{\text{MAP}} = m$ .

L9(4)

August 31, 2021 38 / 67

## Why MAP Is Good? (1)

- MAP estimate is intuitive, but we need more mathematical evidence for its performance guarantee. We would trust its quality if it is **optimal in some sense**.



- MAP:  $\hat{\theta}_{\text{MAP}} = 2$

- Given  $X = x$ ,  $\theta$  that minimizes the probability of incorrect decision?

$$\hat{\theta}_{\text{MAP}} = \arg \min_{\hat{\theta}=1,2,3} \mathbb{P}(\hat{\theta} \neq \Theta|X=x)$$

- Average probability of incorrect decision

$$\begin{aligned} \mathbb{P}(\hat{\Theta} \neq \Theta) &= \sum_x \mathbb{P}(\hat{\Theta} \neq \Theta|X=x)p_X(x) \\ &= \sum_x \mathbb{P}(\hat{\theta} \neq \Theta|X=x)p_X(x) \\ &\geq \sum_x \mathbb{P}(\hat{\theta}_{\text{MAP}} \neq \Theta|X=x)p_X(x) \end{aligned}$$

L9(4)

August 31, 2021 39 / 67

## Why MAP Is Good? (2)

- **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  $I_{\text{MAP}}$  be the indicator rv, representing the correct decision by any general estimator and the MAP estimator, respectively.

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

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

L9(4)

August 31, 2021 40 / 67

- (1) Overview on Statistical Inference
- (2) Bayesian Inference: Framework
- (3) Examples
- (4) MAP (Maximum A Posteriori) Estimator
- (5) **LMS (Least Mean Squares) Estimator**
- (6) LLMS (Linear LMS) Estimator
- (7) Classical Inference: ML Estimator

L9(5)

August 31, 2021 41 / 67

- MAP: the estimate which maximizes the posterior pdf, which solves the following optimization problem (minimizing the prob. of incorrect decision):

$$\min_{\hat{\theta}} \mathbb{P}[\Theta \neq \hat{\theta} | X = x]$$

- What about applying other objective function? Like the following one (mean squared error)?

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

- Least Mean Square (LMS) Estimate

L9(5)

August 31, 2021 42 / 67

- Unknown:  $\theta$  modeled by  $\Theta$  with prior  $f_{\Theta}(\cdot)$ . Assume  $\Theta \sim \mathcal{U}[4, 10]$ .
- Assume that **no observations** available
- MAP estimate
  - Any value  $\hat{\theta}_{\text{MAP}} \in [4, 10]$  (why? posterior = prior), not very useful
- What is the other choice?
  - Expectation:  $\hat{\theta} = \mathbb{E}[\Theta] = 7$
  - looks reasonable, but why?
- First, it makes sense, but, second, it also minimizes the mean squared error (MSE)

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

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

L9(5)

August 31, 2021 43 / 67

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

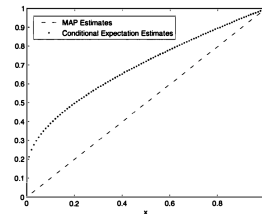
$$\min_{\hat{\theta}} \mathbb{E}[(\Theta - \hat{\theta})^2 | X = x]$$

- minimized when  $\hat{\theta} = \mathbb{E}[\Theta | X = x]$ .
- LMS estimator  $\hat{\Theta} = \mathbb{E}[\Theta | X]$
- What is the mean squared error of the LMS estimate?
  - When  $X = x$ ,  $\mathbb{E}[(\Theta - \mathbb{E}[\Theta | X = x])^2 | X = x] = \text{var}(\Theta | X = x)$
  - Averaged over  $X$ :  $\mathbb{E}[(\Theta - \mathbb{E}[\Theta | X])^2] = \mathbb{E}[\text{var}(\Theta | X)]$

L9(5)

August 31, 2021 44 / 67

- Romeo and Juliet start dating, where Romeo is late by  $X \sim \mathcal{U}[0, \theta]$ .
- Unknown:  $\theta$  modeled by a rv  $\Theta \sim \mathcal{U}[0, 1]$ .
- Observation: Romeo was late by  $x$ .
- Posterior:  $f_{\Theta|X}(\theta|x) = \begin{cases} \frac{1}{\theta|\log x|}, & x \leq \theta \leq 1, \\ 0, & \theta < x \text{ or } \theta > 1 \end{cases}$
- $\hat{\theta}_{\text{MAP}} = x$ .
- LMS estimator:  $\hat{\theta}_{\text{LMS}} = \mathbb{E}[\theta|X = x] = \int_x^1 \theta \frac{1}{\theta|\log x|} d\theta = (1-x)/|\log x|$



L9(5)

August 31, 2021 45 / 67

- Biased coin with prob. of head  $\theta$ . Unknown  $\theta$  modeled by  $\Theta$  with prior  $f_{\Theta}(\theta)$ .
- Observation  $X$ : number of heads out of  $n$  tosses
- If  $\Theta \sim \text{Beta}(\alpha, \beta)$ , then  $\Theta|X = k \sim \text{Beta}(k + \alpha, n - k + \beta)$

- MAP estimate

$$\hat{\theta}_{\text{MAP}} = \frac{\alpha + k - 1}{\alpha + \beta - 2 + n}$$

- For  $\alpha = \beta = 1$  ( $\mathcal{U}[0, 1]$  prior),

$$\hat{\theta}_{\text{MAP}} = \frac{k}{n}$$

- Fact. If  $\Theta \sim \text{Beta}(\alpha, \beta)$ ,

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

- LMS estimate:

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

- For  $\alpha = \beta = 1$  ( $\mathcal{U}[0, 1]$  prior):  $\mathbb{E}[\Theta|X = k] = \frac{k+1}{n+2}$

L9(5)

August 31, 2021 46 / 67

- The posterior pdf  $f_{\Theta|X}(\theta|x) = a \cdot \exp\left\{-\frac{(\theta - m)^2}{2v}\right\}$ , where

$$m = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}, \quad v = \frac{1}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

- The pdf is normal, so it is maximized when  $\theta = \text{mean}$ .
- Thus,  $\hat{\theta}_{\text{MAP}} = m$ .
- What is the LMS estimate?

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

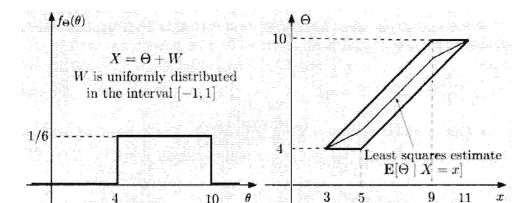
L9(5)

August 31, 2021 47 / 67

- Send signal  $\theta$  with the uniform noise  $W \sim \mathcal{U}[-1, 1]$ . Observe  $X$
- $X = \Theta + W$ , where model  $\theta$  with  $\Theta \sim \mathcal{U}[4, 10]$
- Given  $\Theta = \theta$ ,  $X = \theta + W \sim \mathcal{U}[\theta - 1, \theta + 1]$ .

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

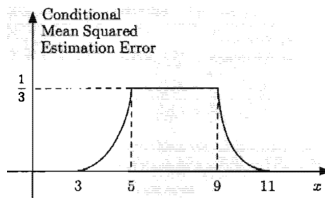
$\hat{\theta}_{\text{LMS}} = \mathbb{E}[\Theta|X = x]$ : midpoint of the corresponding vertical section



L9(5)

August 31, 2021 48 / 67

- What is conditional MSE?  $\mathbb{E}[(\Theta - \mathbb{E}[\Theta|X = x])^2|X = x]$
- Given  $X = 3$ , it's the variance of  $\mathcal{U}[4, 4] = 0$
- Given  $X = 5$ , it's the variance of  $\mathcal{U}[4, 6] = (6 - 4)^2/12 = 1/3$
- The rising pattern between  $X = 3$  and  $X = 5$  is quadratic. This is because the expectation increases linearly, where the variance increases in a quadratic manner.



L9(5)

August 31, 2021 49 / 67

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

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

- Observation model  $f_{X|\Theta}(x|\theta)$  may not be always available
- Finding the posterior distribution is hard for multi-dimensional  $\Theta$
- $\Theta$  is very often high-dimensional, especially in the era of big data and deep learning
  - AlexNet in image recognition: 61M parameters
  - GPT-3 in natural language processing: 175B parameters
- Any alternative to LMS estimator?

L9(5)

August 31, 2021 50 / 67

- (1) Overview on Statistical Inference
- (2) Bayesian Inference: Framework
- (3) Examples
- (4) MAP (Maximum A Posteriori) Estimator
- (5) LMS (Least Mean Squares) Estimator
- (6) LLMS (Linear LMS) Estimator
- (7) Classical Inference: ML Estimator

L9(6)

August 31, 2021 51 / 67

- 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  $g(X)$ 
  - Estimator:  $\hat{\Theta} = aX + b$ .
  - Estimate: Given  $X = x$ ,  $\hat{\theta} = ax + b$ .

- Our goal is to try our best within this restricted class:

$$\min_{a,b} \mathbb{E}[(\Theta - aX - b)^2|X = x], \quad \min_{a,b} \mathbb{E}[(\Theta - aX - b)^2]$$

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

L9(6)

August 31, 2021 52 / 67

## LLMS

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

where the correlation coefficient  $\rho = \frac{\text{cov}(\Theta, X)}{\sigma_\Theta \sigma_X}$ .

- **No need of distributions** on  $\Theta$  and  $X$ : only means, variances, and covariances

- |   |  |
|---|--|
| <ul style="list-style-type: none"> <li>◦ If <math>\rho &gt; 0</math> :           <ul style="list-style-type: none"> <li>- Baseline (<math>\mathbb{E}[\Theta]</math>) + correction term</li> <li>- If <math>X &gt; \mathbb{E}[X] \Rightarrow \hat{\Theta}_L &gt; \mathbb{E}[\Theta]</math></li> <li>- If <math>X &lt; \mathbb{E}[X] \Rightarrow \hat{\Theta}_L &lt; \mathbb{E}[\Theta]</math></li> </ul> </li> </ul> | <ul style="list-style-type: none"> <li>◦ If <math>\rho = 0</math> (uncorrelated):           <ul style="list-style-type: none"> <li>- Just baseline (<math>\mathbb{E}[\Theta]</math>)</li> <li>- <math>\hat{\Theta}_L = \mathbb{E}[\Theta]</math></li> <li>- No use of data <math>X</math></li> </ul> </li> </ul> |
|---|--|

L9(6)

August 31, 2021 53 / 67

- MSE  $\mathbb{E}[(\hat{\Theta}_L - \Theta)^2]$

- Assume  $\mathbb{E}[\Theta] = \mathbb{E}[X] = 0$  (for simplicity). Then,  $\text{MSE} = \mathbb{E}[(\Theta - \rho \frac{\sigma_\Theta}{\sigma_X} X)^2]$
- Note that  $\text{var}[\Theta] = \sigma_\Theta^2 = \mathbb{E}(\Theta^2)$  and  $\text{var}[X] = \sigma_X^2 = \mathbb{E}(X^2)$

$$\begin{aligned} \mathbb{E}[(\Theta - \rho \frac{\sigma_\Theta}{\sigma_X} X)^2] &= \text{var}(\Theta - \rho \frac{\sigma_\Theta}{\sigma_X} X) \\ &= \text{var}(\Theta) + \left(\rho \frac{\sigma_\Theta}{\sigma_X}\right)^2 \text{var}(X) - 2\left(\rho \frac{\sigma_\Theta}{\sigma_X}\right) \text{cov}(\Theta, X) = (1 - \rho^2) \text{var}[\Theta] \end{aligned}$$

- Uncertainty about  $\Theta$  after observation **decreases** by the factor of  $1 - \rho^2$
- What happens if  $|\rho| = 1$  or  $\rho = 0$ ?

$$\hat{\Theta}_L = \mathbb{E}(\Theta) + \rho \frac{\sigma_\Theta}{\sigma_X} (X - \mathbb{E}(X))$$

L9(6)

August 31, 2021 54 / 67

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

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

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

- Assume  $a$  was found.

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

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

**Slide pp. 43**

$$\begin{aligned} \text{ERR}(a, b) &= \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \text{var}(Y) \\ &= \text{var}[\Theta] + a^2 \text{var}[X] - 2a \text{cov}(\Theta, X) \end{aligned} \quad (3)$$

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

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

- 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).

**Slides 17, 35, and 45 for more details**

- Romeo and Juliet start dating, where Romeo is late by  $X \sim \mathcal{U}[0, \theta]$ .
- Unknown:  $\theta$  modeled by a rv  $\Theta \sim \mathcal{U}[0, 1]$ .
- Random observation:  $X$
- $\hat{\Theta}_{\text{MAP}} = X$ , and  $\hat{\Theta}_{\text{LMS}} = (1 - X)/\log X$ .
- **Question.** What is the LLMS estimator  $\hat{\Theta}_L$ ?

L9(6)

August 31, 2021 55 / 67

L9(6)

August 31, 2021 56 / 67

- $\hat{\Theta}_L = \mathbb{E}(\Theta) + \frac{\text{cov}(\Theta, X)}{\text{var}(X)} (X - \mathbb{E}(X))$
- $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\Theta]] = \mathbb{E}[\Theta/2] = 1/4$
- Using  $\mathbb{E}[\Theta] = 1/2$  and  $\mathbb{E}[\Theta^2] = 1/3$ ,  

$$\text{var}[X] = \mathbb{E}[\text{var}[X|\Theta]] + \text{var}[\mathbb{E}[X|\Theta]]$$

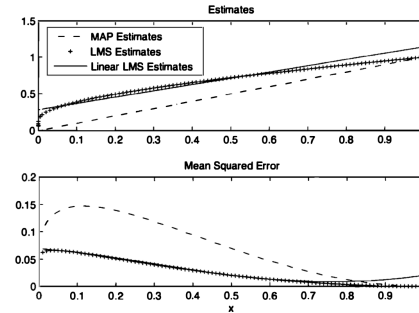
$$= \frac{1}{12}\mathbb{E}[\Theta^2] + \frac{1}{4}\text{var}[\Theta] = \frac{7}{144}$$
- $\text{cov}(\Theta, X) = \mathbb{E}[\Theta X] - \mathbb{E}[\Theta]\mathbb{E}[X]$   

$$\mathbb{E}[\Theta X] = \mathbb{E}[\mathbb{E}[\Theta X|\Theta]] = \mathbb{E}[\Theta \mathbb{E}[X|\Theta]]$$

$$= \mathbb{E}[\Theta^2/2] = 1/6$$

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

$$\hat{\Theta}_L = \frac{1}{2} + \frac{1/24}{7/144} (X - \frac{1}{4}) = \frac{6}{7}X + \frac{2}{7}$$



L9(6)

August 31, 2021 57 / 67

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

$$\text{var}(X) = \mathbb{E}[\text{var}(X|\Theta)] + \text{var}(\mathbb{E}[X|\Theta])$$

$$= \mathbb{E}[n\Theta(1 - \Theta)] + \text{var}[n\Theta]$$

$$= \frac{n}{2} - \frac{n}{3} + \frac{n^2}{12} = \frac{n(n+2)}{12}$$

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

$$\mathbb{E}[\Theta X] = \mathbb{E}[\mathbb{E}[\Theta X|\Theta]] = \mathbb{E}[\Theta \mathbb{E}[X|\Theta]]$$

$$= \mathbb{E}[n\Theta^2] = n/3$$

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

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

- $\hat{\Theta}_{\text{MAP}} = \frac{X}{n}$
- $\hat{\Theta}_{\text{LMS}} = \frac{X+1}{n+2}$
- $\hat{\Theta}_L = \hat{\Theta}_{\text{LMS}}$ ! Intuitive?
- Yes, because the LMS estimator was linear.

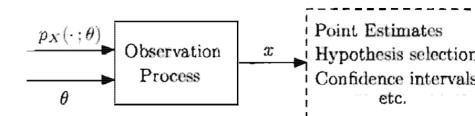
L9(6)

August 31, 2021 58 / 67

- (1) Overview on Statistical Inference
- (2) Bayesian Inference: Framework
- (3) Examples
- (4) MAP (Maximum A Posteriori) Estimator
- (5) LMS (Least Mean Squares) Estimator
- (6) LLMS (Linear LMS) Estimator
- (7) **Classical Inference: ML Estimator**

L9(7)

August 31, 2021 59 / 67



- Unknown  $\theta$ 
  - **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  $\theta$
  - Notation:  $p_X(x; \theta)$  and  $f_X(x; \theta)$ ,  $\theta$ -parameterized distribution of observations
- Choosing one among multiple probabilistic models
  - Each  $\theta$  corresponds to a probabilistic model

L9(7)

August 31, 2021 60 / 67

- Problem types
  - **Estimation**:  $\theta$ : prob. of head?
  - Hypothesis testing:  $\theta = 1/2$  or  $\theta = 1/4$ ?
  - Significance testing:  $\theta = 1/2$  or not?
- Key inference methods
  - **ML (Maximum Likelihood) estimation**
  - Linear regression
  - Likelihood ratio test
  - Significant testing
- Just a taste in this course.

L9(7)

August 31, 2021 61 / 67

- Random observation  $x = (x_1, x_2, \dots, x_n)$  of  $X = (X_1, X_2, \dots, X_n)$ 
  - Assume a scalar  $\theta$  and a vector of multiple observations in this lecture.

- **Likelihood**  $p_X(x_1, x_2, \dots, x_n; \theta)$

- $p_X(x_1, x_2, \dots, x_n; \theta)$ 
  - The probability that the observed value  $x$  arises when the parameter is  $\theta$ .
- **ML (Maximum Likelihood) estimation**

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} p_X(x_1, x_2, \dots, x_n; \theta)$$

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

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

L9(7)

August 31, 2021 62 / 67

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

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

- ML in the classical inference

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

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

Slides 17, 35, 45, and 56 for more details

- Romeo and Juliet start dating. Romeo: late by  $X \sim U[0, \theta]$ .
- Unknown:  $\theta$  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:  $\hat{\theta}_{\text{ML}} = \hat{\theta}_{\text{MAP}} = x$

L9(7)

August 31, 2021 63 / 67

L9(7)

August 31, 2021 64 / 67



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

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

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

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

Questions?

L9(7)

August 31, 2021 65 / 67

L9(7)

August 31, 2021 66 / 67

- 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?

L9(7)

August 31, 2021 67 / 67