Topic 12: Bayesian Statistical Inference

Lecture Outline

- Introduction
  - *Probability* vs. *Statistics*
  - Bayesian vs. Classical Statistics
- Bayesian Statistical Inference
  - Bayesian estimation
  - Bayesian hypothesis testing (Bayesian detection)
- Maximum a Posteriori (MAP) Rule
  - MAP estimation and MAP detection
- Least Mean Squares (LMS) and *Linear* LMS Estimation

Reading : Textbook 8.1-8.4

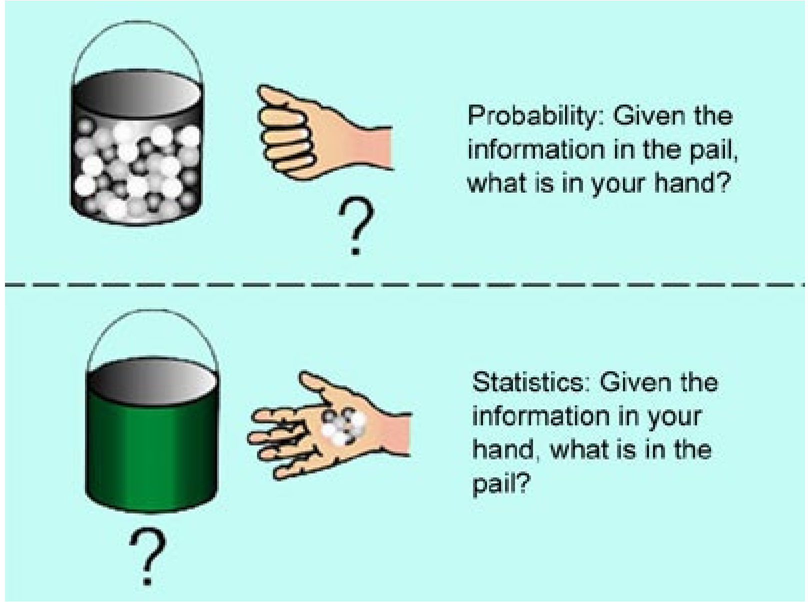
Probability vs. Statistics

- Probability – an axiomatic mathematical theory

- Statistical inference – estimate (or predict) something (*unknown* variables/model parameters/reality) based on the *observed data*
  - Statistical inference – many “methods” have been proposed, depending on multitude of factors such as on the performance, e.g. *minimum MSE* or *minimum error probability*, or more generally *minimum loss*, the designer would like to achieve

工程系統設計(如手機通訊演算法)，或是透過人工智慧處理的預估與分類問題幾乎都是在得知某些事件(如*已知量測訊號、蒐集到的資料)*的前提下作出**決策判斷**

Probability vs. Statistics



Probability: Given the information in the pail, what is in your hand?

Statistics: Given the information in your hand, what is in the pail?

Example:

- (Probability) 給定環境資訊，計算特定事件會發生的機率

  Ex1: Flipping a *fair coin* two times, the probability of two “heads” is 1/4

- (Statistics) 給定特定事件(觀察結果)，推論出環境資訊為何?

  Ex: 有一銅板但不知其出現正面機率, 要如何估計出此出現正面機率? 你的直覺做法為何呢?

Bayesian vs. Classical Statistics

統計學的兩大門派：**Bayesian vs. Classical**

- Bayesian vs. Classical Statistics
  - **Bayesian**: Unknown parameter (model) is treated as a random variable. In this case, we need to assume a proper distribution, i.e. the *prior distribution*, for the unknown parameter
  - **Classical**: Unknown parameter (model) is treated as a deterministic quantity

Both Bayesian and classical methods may give identical results, particularly when the *prior* does not provide useful information

Statistical Inference Problems

Problems of statistical interference can be divided into two types: *estimation problem* or detection (*hypothesis testing) problem*

- Estimation (or, regression in machine learning terminology)
  - Estimation problem involves with deciding ***continuous-valued*** parameter(s)

    若欲估計的參數是連續實數，此時被稱作是 estimation 或是 regression 的問題
  - Ex: We employ a polynomial model to predict tomorrow’s stock value. Then, we need to find the coefficients of the specified polynomial

- Detection (or hypothesis testing) (or classification, in machine learning terminology)
  - Dection problem involves with deciding finite ***discrete-valued*** parameter(s)

    若欲破解的參數為離散可數，此時被稱作是 detection 或是 regression 的問題
  - Ex: A smart phone decides whether “0” or “1” is transmitted in digital communications

Statistical Inference Problems

兩大門派都各自有對付 estimation 和 detection 的手段

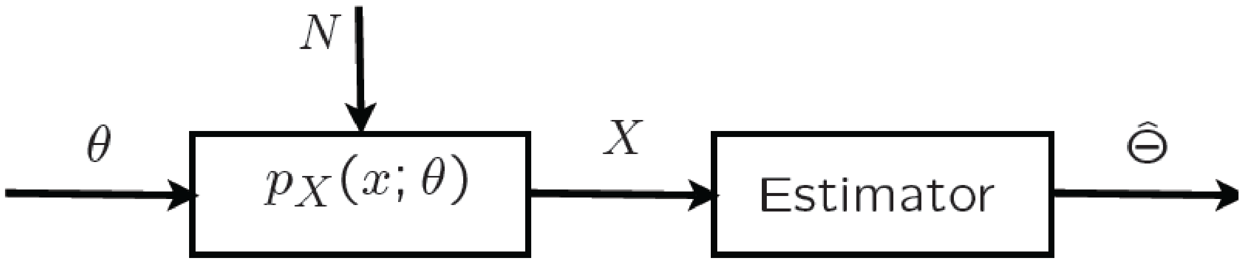
- Bayesian Statistical Inference → Chapter 8
  - Bayesian Estimation
    1) Maximum a Posteriori Estimation (MAP estimation) → Section 8.1, 8.2
    2) Least Mean-Square Error Estimation (LSE or MMSE) → Section 8.3, 8.4
  - Bayesian Detection
    1) Maximum a Posteriori detection (MAP detection) → Section 8.1, 8.2

上述這三種 Baysian Inference 都牽涉到計算 Posterior Probability/Density

- Classical Statistics → Chapter 9
  - Classical Estimation / Classical Hypothesis Testing

研究所課程【檢測與估計】(detection and estimation) 有更為深入的探討!

Classical Statistics



- Observed data "X" are noisy (corrupted by "*N*")
- θ: unknown deterministic (continuous) parameter

  $\hat{\Theta}$: an estimator of θ that depends on X
- Example: $X_i = \theta + N_i$

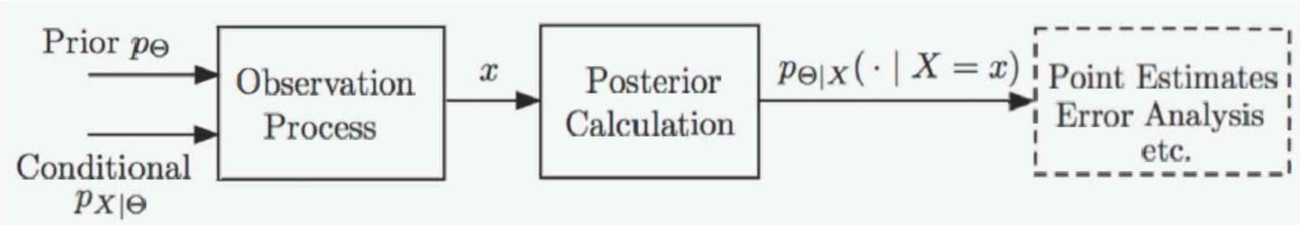
  N_i: i.i.d. with zero mean and variance σ^2

  Given observations: $X_1, X_2, \cdots X_n$

  An estimate of θ *:* $\widehat{\Theta} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$

  Performance?

Bayesian Statistics



- The desired but unknown variable Θ:
  - In the Bayesian framework, Θ is modeled as a RV (may be discrete or continuous)
  - We need to assume a **prior** distribution $p_\Theta(\theta)$
    (*a priori*，*事前機率 for discrete* Θ 、*事前機率密度 for continuous* Θ)
- *Goal*: Estimate Θ using the observed data X
  - Bayesian approach needs posterior distribution $p_{\Theta|X}(\theta|x)$ to update our understanding about Θ ($p_{\Theta|X}(\theta|x)$ can be *事後機率* for discrete Θ 、*事後機率密度* for continuous Θ)
  - Finding the posterior distribution $p_{\Theta|X}(\theta|x)$ relies on
    - ✓ Bayes' rule
    - ✓ A system model

    A system model describes how observation X is mathematically related to Θ, which specifically provides the likelihood function $L(\theta) \equiv p_{X|\Theta}(x|\theta)$, when given a fixed observed value at $X = x$

Example (8.2, p.414)

- **“A”** is late in a date. The late time is an RV X, uniformly distributed over the interval $[0, \theta]$.
- The parameter θ is unknown and is modeled as an RV Θ, which is uniformly distributed over [0,1].
- After one date, we observe an event of $X=x$, say, 0.32 hours. How do we use this information to update the distribution of Θ?

(Sol) We look for $f_{\Theta|X}(\theta|x)$.

Bayes' Rules (1)

Bayes rule is of critical importance in the MAP detection/estimation problem

- There are 4 versions of Bayes' rules (textbook p. 181 and p. 413), depending on whether
  - unknown variable Θ is discrete or continuous
  - observation (data) X is discrete or continuous

- Hypothesis Testing (unknown discrete parameter Θ)
  - *Discrete data X*

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

  - *****Continuous data X***

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

Bayes' Rules (2)

- Estimation (unknown continuous parameter Θ)
  - ➢ *****Discrete data X***

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

Ex: A coin with unknown bias (probability of head)

-- Observe X heads in *n* tosses

  - ➢ *Continuous data X*

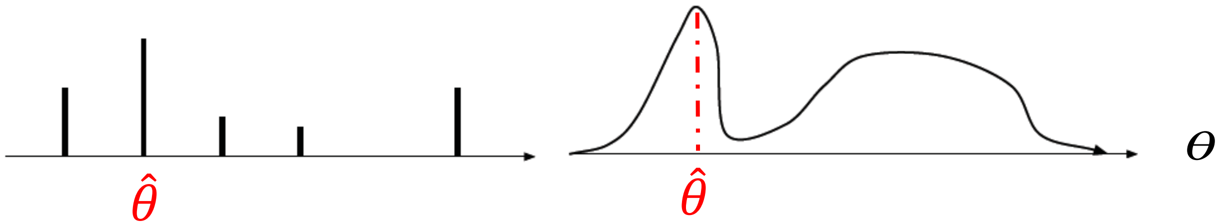
$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

Bayesian Inference Procedure

1) Start with a prior distribution, $p_\Theta(\theta)$, for the unknown random variable Θ
2) A model that describes the relation between the observation vector *X* and the unknown Θ, which allows for calculation of the likelihood function $p_{X|\Theta}(x|\theta)$
3) After observing the value x of X, evaluate the posterior distribution of $p_{\Theta|X}(\theta|x)$, using the appropriate version of Bayes' rule. (p.413)

Maximum a Posteriori Probability (MAP) Rule

- Having obtained the posterior distribution of $p_{\Theta|X}(\theta|x)$, we find the value of θ with the maximum posterior prob./pdf
  - pmf $p_{\Theta|X}(\cdot \mid x)$ or pdf $f_{\Theta|X}(\cdot \mid x)$



- *For discrete* Θ, this is called MAP detection (MAP hypothesis testing)

$$\hat{\theta} = \arg\max_{\theta} p_{\Theta|X}(\theta \mid x)$$

- *For continuous* Θ, this is called MAP estimation

$$\hat{\theta} = \arg\max_{\theta} f_{\Theta|X}(\theta \mid x)$$

MAP Rule vs. Conditional Expectation

- **Conditional Expectation**: $E(\Theta|X=x)$, the MMSE estimator (in page 16, Topic 10), i.e., the least mean squares estimator (Sec 8.3)
- *Ex*: (8.7, p.424) "**A**" is late in a date, …

(i) MAP:

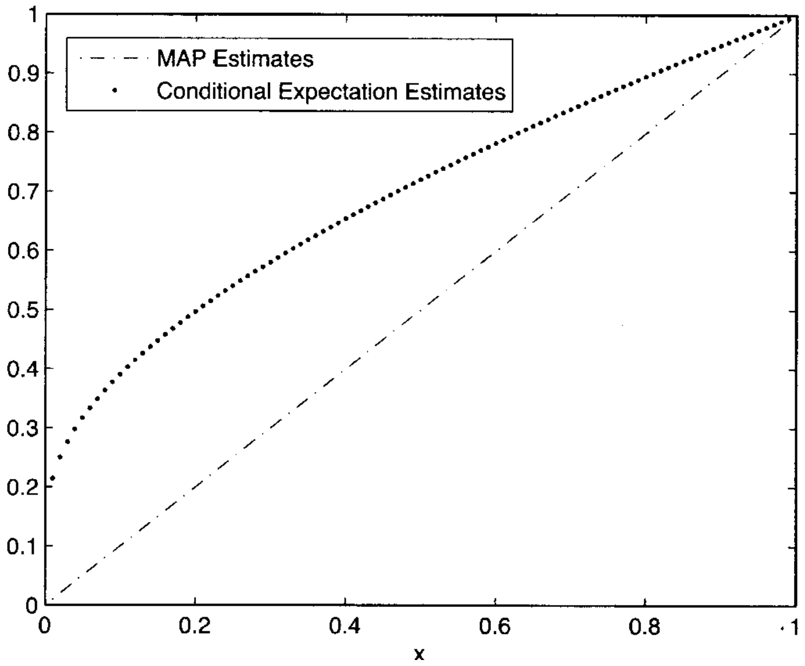
$$f_{\Theta|X}(\theta \mid x) = \frac{1}{\theta \cdot |\log x|}, \quad \text{if } x \le \theta \le 1$$

➔ $\hat{\theta} = x$.

(ii) Conditional Mean:

$$E(\Theta \mid X = x) = \int_x^1 \theta \frac{1}{\theta \cdot |\log x|} d\theta$$

$$= \frac{1-x}{|\log x|}$$



More on MAP Hypothesis Testing

Problem formulation:

RV Θ takes one of m values, θ_1, …, θ_m. Once the measurement RV X is observed (with value x), we'd like to decide which hypothesis (one out of θ_1, …, θ_m) is true.

- MAP detection rule: Pick up θ_i based on the maximum $p_{\Theta|X}(\theta_i\,|x)$.
- The case of m =2 is called the MAP binary hypothesis test (*null* and *alternate*)

$$\mathrm{H}_0\text{: } \Theta = \theta_1$$

$$\mathrm{H}_1\text{: } \Theta = \theta_2$$

- Tie: If there is a *tie*, either can be selected arbitrarily.
- *MAP detection is the decision rule that minimizes the probability of incorrect decision*

More on MAP Hypothesis Testing

- *MAP detection is the decision rule that minimizes the probability of error decision. (p. 420)*

Remarks:

- This theorem lays the foundations for *signal detection* in modern digital communication systems (4G, 5G, 6G and beyond) and for *objects classification* in machine learning/artificial intelligence (AI) applications!
- In most cases, we are interested not only in designing MAP criterion, but also in knowing the corresponding min. probability of error decision.
  See Example 3.8, Example 8.9, and Problem 4 of HW 5.

Example (8.9, p.426)

- **Two coins:**

  Coin 1 (Θ =1): p_1 (head) = 0.46

  Coin 2 (Θ =2): p_2 (head) = 0.52

Let $p_\Theta(\theta=1) = p_\Theta(\theta=2) = 0.5$. And X is the number of observed heads in n tosses. That is, the outcome of one toss, X=1 (head) or 0 (tail).

(a) Decide which coin is selected with one observation, say, "tail" (X=0).

(b) Decide which coin is selected with n tosses and k heads appear (X=k).

Sol) Calculate $p_{\Theta|X}(\theta|x) = p_\Theta(\theta)p_{X|\Theta}(x|\theta)$.

Now, because $p_\Theta(1) = p_\Theta(2)$ we only need to calculate and compare $p_{X|\Theta}(x|\theta)$ for θ =1 and θ =2.

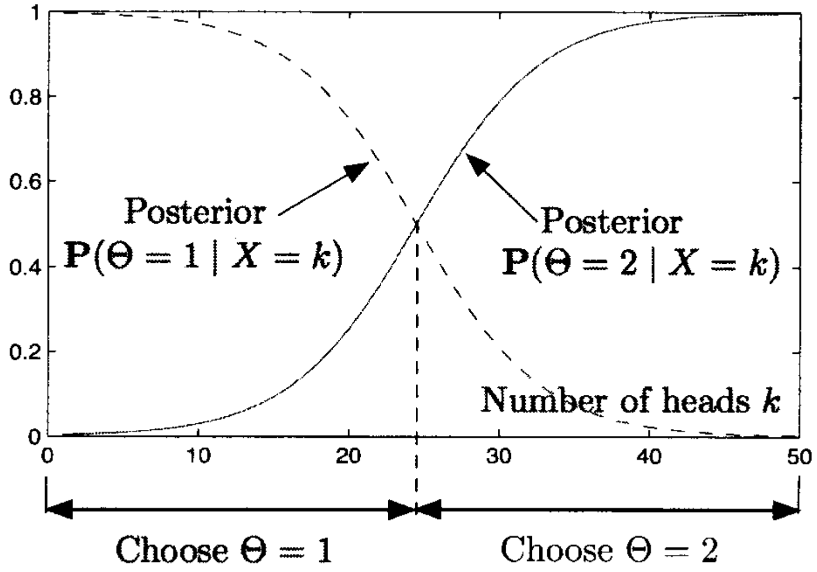
(a) $p_{X|\Theta}(x\text{=tail}|\theta\text{=}1) = 1\text{-}0.46 = 0.54$

$p_{X|\Theta}(x\text{=tail}|\theta\text{=}2) =1\text{-}0.52 = 0.48$

Sol) (b) *n* tosses and *k* heads, geometric distribution

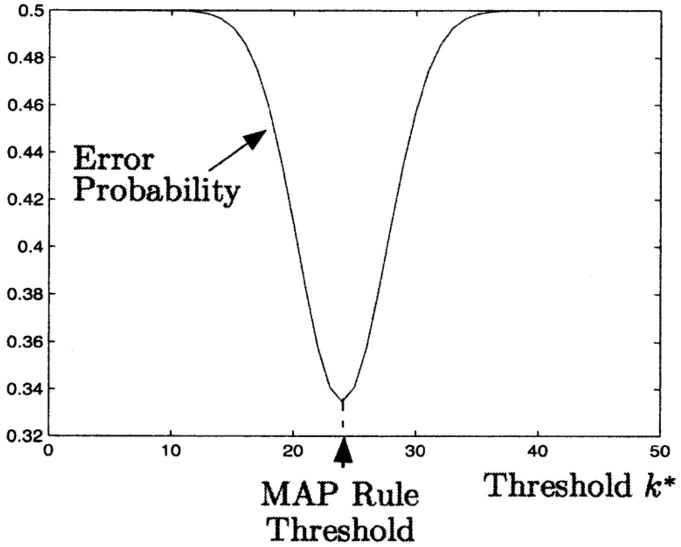
Prob: $p_{X|\Theta}(x=k|\theta=1) = p_1{}^k(1-p_1)^{n-k}$

$p_{X|\Theta}(x=k|\theta=2) = p_2{}^k(1-p_2)^{n-k}$



Error analysis: threshold k^*:

$$P(error) = P(\Theta = 1, X > k^*) + P(\Theta = 2, X \le k^*)$$



Least Mean Squares Estimation – No Observation

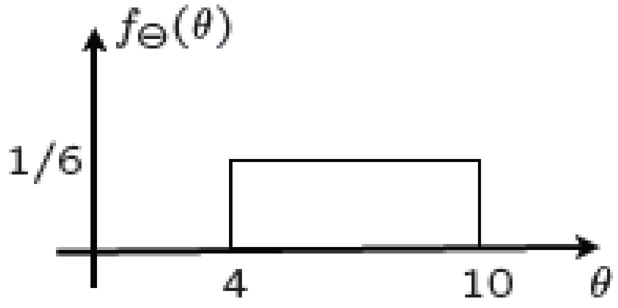
- Estimate a **random value** using **a constant**.

Goal: Find $c = g(X)$ that minimizes the **mean squared error**

Ex: Estimate R.V. Θ using c.

$$\text{minimize } E[(\Theta - c)^2]$$

➔ $\hat{c} = E[\Theta]$



(pf) $E[(\Theta - c)^2]$

$$= E[\Theta^2] - 2cE[\Theta] + c^2$$

minimize $-2cE[\Theta] + c^2$ → take derivative $-E[\Theta] + c = 0$

- Optimal MSE in this case:

$$E[(\Theta - E[\Theta])^2] = \mathrm{var}(\Theta)$$

Least Mean Squares Estimation – Based on X

- Two RV's Θ and X; estimate Θ based on an observation $X = x$.

Goal: Find $g(X)$ that minimizes the **mean squared error**

$$\text{minimize } E[(\Theta - g(X))^2 | X = x]$$

➔ We have proved in Topic 10 that $\hat{\theta}_{LMS} = E[\Theta | X = x]$

- This is true for any x value of X. Thus, the **least mean squares (LMS) estimator** of Θ is **conditional mean**: $E[\Theta|X]$
- That is, out of all estimators $g(X)$ of Θ based on X, $E[\Theta|X]$ gives the smallest mean squared error.

$$E[(\Theta - E[\Theta|X])^2] \leq E[(\Theta - g(X))^2]$$

- Finding $E[\Theta|X]$ requires?

Example (8.11, p.432) (1/2)

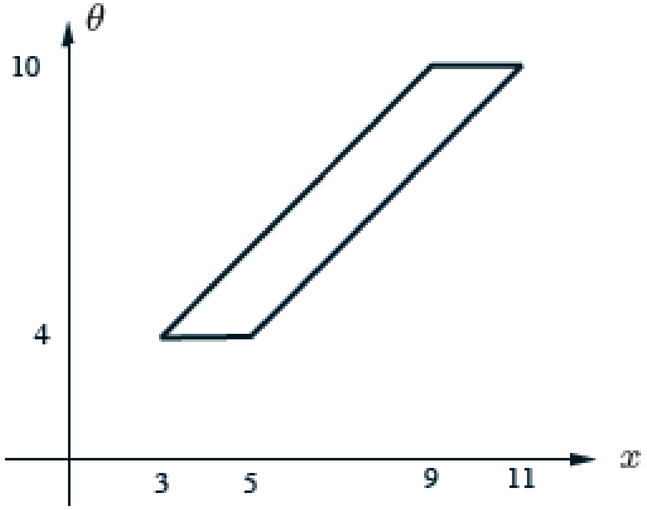
- **Two RV's:** Θ and $X = \Theta + W$

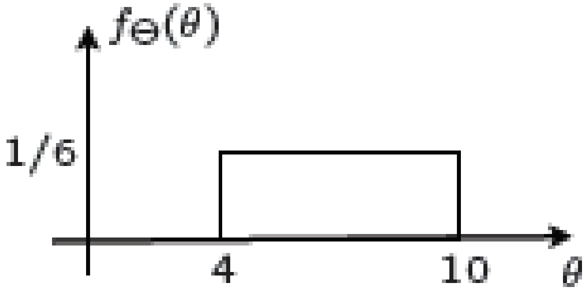
  Θ : uniform over [4,10]

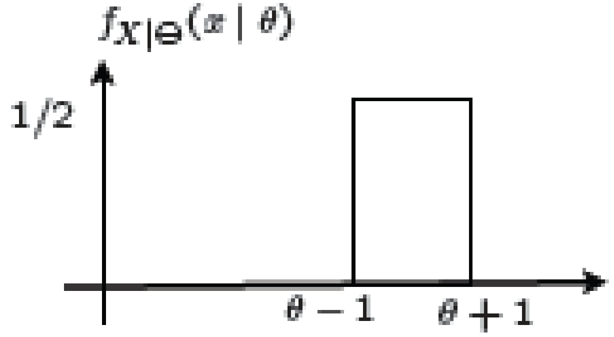
  W: uniform over [-1,1] indep of Θ

  What is $E[\Theta|X]$?

Sol)







Example (8.11, p.432) (2/2)

Sol) (a) $f_{\Theta,X}(\theta,x) = f_{\Theta}(\theta)f_{X|\Theta}(x|\theta) = \frac{1}{6}\frac{1}{2} = \frac{1}{12}$

$f_{\Theta|X}(\theta|x) = {}^{f_{\Theta,X}(\theta,x)}/_{f_X(x)}$

$= {}^{\frac{1}{12}}/_{f_X(x)} = \text{uiform}$

Thus, pick up an x, all nonzero-prob θ values form a vertical section $[x\text{-}1, x+1]$. (X-W)

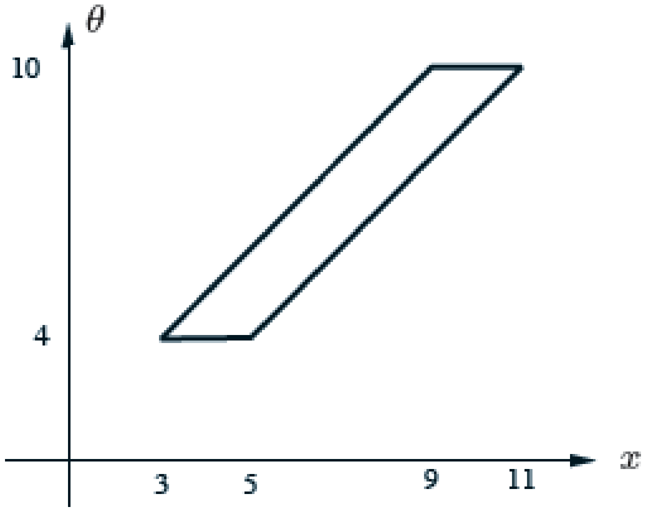
$E[\Theta|X = x]$ is the midpoint of that section.

(b) Mean squared error:

$E[(\Theta - E[\Theta|X = x])^2]$

$x \in [5,9],\ \ \text{MSE} = {}^{2^2}/_{12} = \frac{1}{3}$

$x \in [3,5],\ \ \text{MSE} = {}^{(x+1-4)^2}/_{12} = {}^{(x-3)^2}/_{12}$



Properties of the Estimation Error

Let

$$\hat{X} = E[X|Y], \quad \text{and} \quad \tilde{X} = X - \hat{X}$$

denote the least squares estimator and the associated estimation error $\tilde{X}$, respectively. Both of the above are random variables.

The following properties hold:

- MMSE estimator is unbiased: $E[\hat{X}] = E[X]$ (or, equiv., $E[\tilde{X}] = 0$)
  (unbiased 的定義：估計結果與原本所欲估計參數有相同期望值)
- Estimator is uncorrelated with error: $\text{cov}(\hat{X}, \tilde{X}) = 0$
- Power conservation: $\text{var}(X) = \text{var}(\hat{X}) + \text{var}(\tilde{X})$

Linear Least Mean Squares Estimation

- Two RV's Θ and X; estimate Θ based on an observation X.

  The function $g(X)$ that minimizes the **mean squared error**

  $$\text{minimize } E[(\Theta - g(X))^2]$$

  is given by $\hat{\theta}_{LMS} = E[\Theta|X]$

  - This conditional mean very often is nonlinear in X, or does not have closed-form expression

- It is desirable to find $g(X)$ that is constrained to be linear in X

  $$g(X) = aX + b$$

  We wish to find a and b such that $E[(\Theta - aX - b)^2]$ is minimized.

  This is called *linear* LMS estimation or *linear* MMSE (LMMSE).

Linear Least Mean Squares Estimation

- We wish to find a and b such that $E[(\Theta - aX - b)^2]$ is minimized.

  - Fixed a, we have the best b given by $b = E[\Theta] - aE[X]$
  - With this b, it remains to minimize

  $E[(\ \Theta - aX - (E[\Theta] - aE[X])\ )^2]$, which is exactly $\mathrm{var}(\Theta - aX)$

$$\mathrm{var}(\Theta - aX) = \sigma_\Theta^2 + a^2\sigma_X^2 - 2a \cdot \mathrm{cov}(\Theta, X)$$

  - The best a minimizing the above is

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

We therefore have

$$\widehat{\Theta}_{LLMS} = E[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$

Linear Least Mean Squares Estimation

- The corresponding MSE is

$$\mathrm{var}(\Theta - aX) = (1-\rho^2)\sigma_\Theta^2$$

- We can re-arrange the LMS estimator as

$$\frac{\widehat{\Theta}_{LLMS} - E[\Theta]}{\sigma_\Theta} = \rho \cdot \frac{X - E[X]}{\sigma_X}$$

This allows an interesting interpretation:

- The *normalized* X and $\hat{\theta}_{LLMS}$ is proportional to each other, subject to a scaling factor ρ
- This is reasonable as
  1) We are searching a linear relation
  2) The similarity between X and Θ is described by the correlation coefficient ρ

Example 8.15, p. 439

- **“A”** is late in a date. The late time is an RV X, uniformly distributed over the interval $[0, \theta]$.
- The parameter θ is unknown and is modeled as an RV Θ, which is uniformly distributed over [0,1].
- What is the linear LMS estimator of Θ based on X?

$$\mathrm{var}(X) = E[\mathrm{var}(X|\Theta)] + \mathrm{var}(E[X|\Theta]) = \frac{7}{144}$$

$$\mathrm{cov}(\Theta, X) = E[\Theta X] - E[\Theta]E[X] = \frac{1}{6} - \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{24}$$

Multiple Observations with Single Parameter

- In many applications, we have more than one observations $X_1, X_2, \dots, X_n$ in order to estimate one single parameter Θ

We still can to find g(X) that is constrained to be linear in $X_1, X_2, \dots, X_n$

$$g(X) = a_1X_1 + a_2X_2 + \cdots + a_nX_n + b$$

We wish to find $a_1, a_2, \dots, a_n$ and b such that

$E[(\Theta - (a_1X_1 + a_2X_2 + \cdots + a_nX_n + b\ ))^2]$ is minimized.

These coefficients $a_1, a_2, \dots, a_n$ and b can be determined by setting to zero its partial derivatives with respective to $a_1, a_2, \dots, a_n$ and b