ENGG2430A Probability and Statistics for Engineers

Chapter 8: Bayesian Statistical Inference

Instructor: Shengyu Zhang

Statistical inference

- Statistical inference is the process of extracting information about an unknown variable or an unknown model from available data.

- Two main approaches
  - Bayesian statistical inference
  - Classical statistical inference

Statistical inference

- Main categories of inference problems
  - parameter estimation
  - hypothesis testing
  - significance testing

Statistical inference

- Most important methodologies
  - maximum a posteriori (MAP)
  - probability rule,
  - least mean squares estimation,
  - maximum likelihood,
  - regression,
  - likelihood ratio tests

Bayesian versus Classical Statistics

- Two prominent schools of thought
  - Bayesian
  - Classical/frequentist.
- Difference: *What's the nature of the unknown models or variables?*
- Bayesian: they are treated as random variables with known distributions.
- Classical/frequentist: they are treated as deterministic but unknown quantities.

Bayesian

- When trying to infer the nature of an unknown model, it views the model as chosen randomly from a given model class.
- *Introduce a random variable* Θ that characterizes the model,
- Postulate a *prior distribution* $p_\Theta(\theta)$.
- Given observed data x, one can use Bayes' rule to derive a *posterior distribution* $p_{\Theta|X}(\theta|x)$.
  - This captures all information that x can provide about θ.

Classical/frequentist

- View the unknown quantity θ as an unknown constant.
- Strives to develop an estimate of θ.
- We are dealing with multiple candidate probabilistic models, one for each possible value of θ.

Model versus Variable Inference

- Model inference: the object of study is a real phenomenon or process,…
- …for which we wish to construct or validate a model on the basis of available data
  - e.g., do planets follow elliptical trajectories?
- Such a model can then be used to make predictions about the future, or to infer some hidden underlying causes.

Model versus Variable Inference

- Variable inference: we wish to estimate the value of one or more unknown variables by using some related, possibly noisy information
  - e.g., what is my current position, given a few GPS readings?

Statistical Inference Problems

- Estimation: a model is fully specified, except for an unknown, possibly multidimensional, parameter θ, which we wish to estimate.
- This parameter can be viewed as either a random variable …
  - Bayesian approach
- …or as an unknown constant
  - classical approach.
- Objective: to estimate θ.

Statistical Inference Problems

- Binary hypothesis testing:
  - start with two hypotheses
  - use the available data to decide which of the two is true.
- m-ary hypothesis testing: there is a finite number m of competing hypotheses.
  - Evaluation: typically by error probability.
- Both Bayesian and classical approaches are possible.

Content

- Bayesian inference, the posterior distribution
- Point estimation, hypothesis testing, MAP
- Bayesian least mean squares estimation
- Bayesian linear least mean squares estimation

Bayesian inference

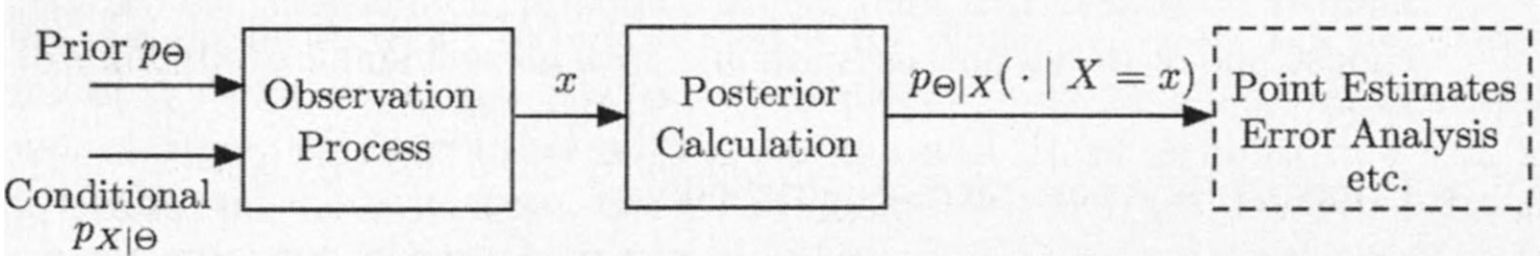
- In Bayesian inference, the unknown quantity of interest is modeled as a random variable or as a finite collection of random variables.
  - We usually denote it by Θ.
- We aim to extract information about Θ, based on observing a collection $X = (X_1, \ldots, X_n)$ of related random variables.
  - called observations, measurements, or an observation vector.

Bayesian inference

- We assume that we know the joint distribution of Θ and X.
- Equivalently, we assume that we know
  - A prior distribution p_Θ or f_Θ, depending on whether Θ is discrete or continuous.
  - A conditional distribution $p_{X|\Theta}$ or $f_{X|\Theta}$, depending on whether X is discrete or continuous.

Bayesian inference

- After a particular value x of X has been observed, a complete answer to the Bayesian inference problem is provided by the posterior distribution $p_{\Theta|X}$ or $f_{\Theta|X}$.
  - It encapsulates everything there is to know about Θ, given the available information.



Summary of Bayesian Inference

1. We start with a prior distribution p_Θ or f_Θ for the unknown random variable Θ.
2. We have a model $p_{X|\Theta}$ or $f_{X|\Theta}$ of the observation vector X.
3. After observing the value x of X, we form the posterior distribution of Θ, using the appropriate version of Bayes' rule.

Bayes' rule: summary

- Depending on discrete or continuous Θ and X, there are four versions of Bayes' rule.
- They are syntactically all similar.

The Four Versions of Bayes' Rule

- Θ discrete, X discrete:

$$p_{\Theta|X}(\theta \,|\, x) = \frac{p_\Theta(\theta) p_{X|\Theta}(x \,|\, \theta)}{\sum_{\theta'} p_\Theta(\theta') p_{X|\Theta}(x \,|\, \theta')}.$$

- Θ discrete, X continuous:

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

- Θ continuous, X discrete:

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

- Θ continuous, X continuous:

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

Example: meeting

- Romeo and Juliet meeting: Juliet will be late on any date by a random amount X, uniformly distributed over the interval $[0, \theta]$.
- θ is unknown and is modeled as the value of a random variable uniformly distributed in $[0,1]$.
- Assume that Juliet was late by an amount x on their first date.
- *Question: How should Romeo use this information to update the distribution of* θ*?*

- Prior PDF: $f_\Theta(\theta) = \begin{cases} 1 & \text{if } 0 \le \theta \le 1, \\ 0 & \text{otherwise.} \end{cases}$

- Conditional PDF of the observation:

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

- $f_\Theta(\theta) = 1$ if $0 \le \theta \le 1$
- $f_{X|\Theta}(x|\theta) = 1/\theta$ if $0 \le x \le \theta$
- Use Bayes' rule: the posterior PDF is

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

$$= \frac{1/\theta}{\int_x^1 \frac{1}{\theta'} d\theta'} = \frac{1}{\theta \cdot |\log x|}, \quad \text{if } 0 \le x \le \theta \le 1$$

  - and $f_{\Theta|X}(\theta|x) = 0$ otherwise.

Example: Inference of common mean of normal

- Suppose that $X_1, \ldots, X_n$ are independent normal r.v. with
  - an unknown common mean,
  - and known variances $\sigma_1^2, \ldots, \sigma_n^2$.
- Suppose that the common mean follows the a normal prior $N(x_0, \sigma_0^2)$.
- Then $X_i = \Theta + W_i$, where
  - Θ, W_i are a independent normal r.v.
  - Θ follows $N(x_0, \sigma_0^2)$, W_i follows $N(0, \sigma_i^2)$.

- Last slide: $X_i = \Theta + W_i$.
  - Θ follows $N(x_0, \sigma_0^2)$, W_i follows $N(0, \sigma_i^2)$.
- Prior PDF: $f_\Theta(\theta) = c_1 \exp\left\{-\frac{(\theta - x_0)^2}{2\sigma_0^2}\right\}$,
- Model: $f_{X|\Theta}(x|\theta) = c_2 \exp\left\{-\frac{(x_1-\theta)^2}{2\sigma_1^2}\right\} \ldots \exp\left\{-\frac{(x_n-\theta)^2}{2\sigma_n^2}\right\}$
  - c_1 and c_2 are constants.
- By Bayes' rule: $f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x|\theta)}{\int_0^1 f_\Theta(\theta') f_{X|\Theta}(x|\theta') d\theta'}$
  - Note: The denominator doesn't depend on θ.

Numerator

- $f_\Theta(\theta) f_{X|\Theta}(x|\theta) = c_1 c_2 \exp\left\{-\sum_{i=0}^{n} \frac{(x_i-\theta)^2}{2\sigma_i^2}\right\}$.
- The exponent is a quadratic form, thus can be written as

$$d \cdot \exp\left\{-\frac{(\theta - m)^2}{2v}\right\}$$

for some constant d, where

$$m = \left(\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}\right) / \left(\sum_{i=0}^{n} \frac{1}{\sigma_i^2}\right),\ v = 1 / \left(\sum_{i=0}^{n} \frac{1}{\sigma_i^2}\right)$$

.

- Thus $f_{\Theta|X}(\theta|x) \propto \exp\left\{-\frac{(\theta-m)^2}{2v}\right\}$
- So the posterior PDF $f_{\Theta|X}(\theta|x)$ is normal with mean m and variance v.
- Recall prior: $\Theta \sim N\left(x_0, \sigma_0^2\right)$.
- A remarkable property: *the posterior distribution of Θ is in the same family as the prior distribution*,
  - the family of normal distributions.

- This property opens up the possibility of efficient recursive inference.
- Suppose that after $X_1, \ldots, X_n$ are observed, an additional observation X_{n+1} is obtained.
- Instead of solving the inference problem from scratch, we can view $f_{\Theta|X_1,\ldots,X_n}$ as our prior, and use the new observation to obtain the new posterior $f_{\Theta|X_1,\ldots,X_n,X_{n+1}}$.

- Thus the new posterior is normal distribution with mean

$$\left(\frac{m}{v}+\frac{x_{n+1}}{\sigma_{n+1}^2}\right)/\left(\frac{1}{v}+\frac{1}{\sigma_{n+1}^2}\right)$$

and variance

$$1/\left(\frac{1}{v}+\frac{1}{\sigma_{n+1}^2}\right).$$

Content

- Bayesian inference, the posterior distribution
- Point estimation, hypothesis testing, MAP
- Bayesian least mean squares estimation
- Bayesian linear least mean squares estimation

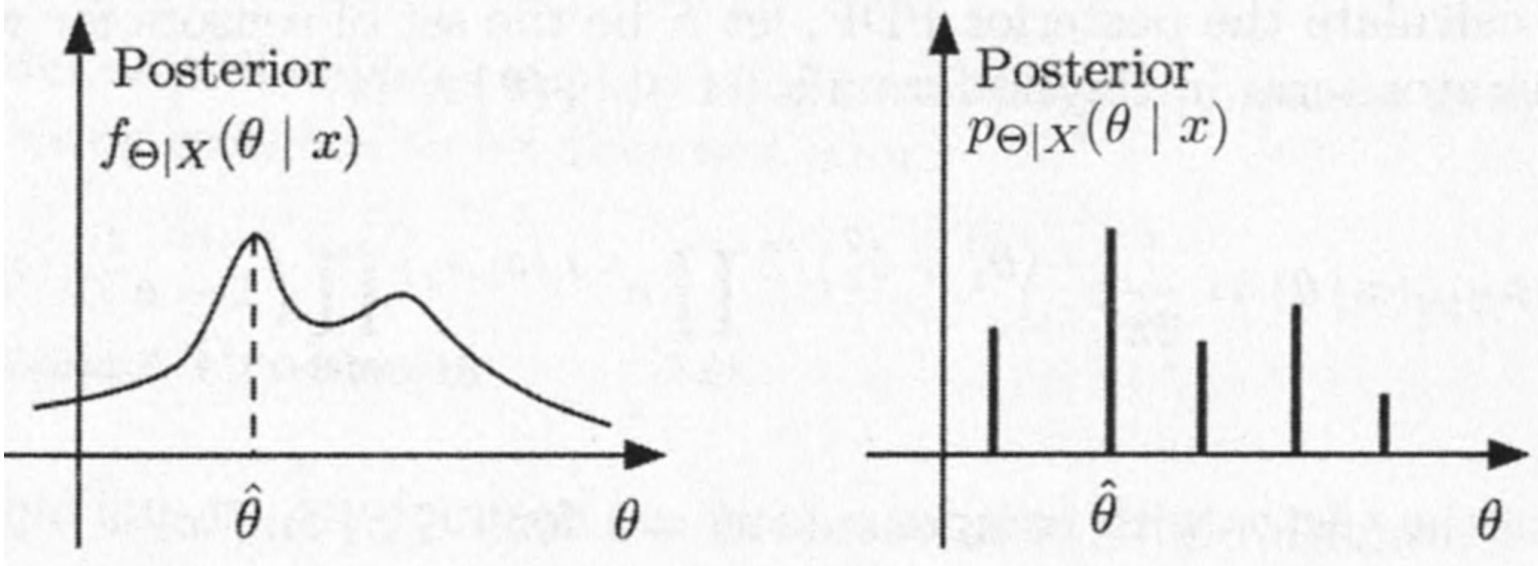
MAP

- Given the value x of the observation, we select a value of θ, denoted $\hat{\theta}$, that maximizes the posterior distribution
  - $p_{\Theta|X}(\theta|x)$ if Θ is discrete
  - $p_{\Theta|X}(\theta|x)$ if Θ is continuous
- That is,

  $\hat{\theta} = \arg\max_{\theta} p_{\Theta|X}(\theta|x)$, if Θ is discrete,

  $\hat{\theta} = \arg\max_{\theta} f_{\Theta|X}(\theta|x)$, if Θ is continuous.

- This is called the Maximum a Posteriori probability (MAP) rule.



- When Θ is discrete, the MAP rule has an important optimality property.
- Since it chooses θ to be the most likely value of Θ, it *maximizes the probability of correct decision for any given value* x.
- This implies that it also maximizes (over all decision rules) the overall (averaged over all possible values x) probability of correct decision.

Computational shortcut

- Recall posterior: $p_{\Theta|X}(\theta|x) = \frac{p_\Theta(\theta)p_{X|\Theta}(x|\theta)}{\sum_{\theta'} p_\Theta(\theta')p_{X|\Theta}(x|\theta')}$
- An important computational shortcut.
- The denominator is independent of θ.
- Thus, to maximize the posterior, we only need to maximize the numerator $p_\Theta(\theta)p_{X|\Theta}(x|\theta)$
  - or similar expressions if Θ and/or X are continuous.
- Calculation of the denominator is unnecessary.

Example

- $X_1, \ldots, X_n$ are independent normal r.v. with
  - an unknown common mean $\Theta \sim N(x_0, \sigma_0^2)$,
  - and known variances $\sigma_1^2, \ldots, \sigma_n^2$.
- Posterior: $f_{\Theta|X}(\theta|x) \propto \exp\left\{-\frac{(\theta-m)^2}{2v}\right\}$ with

  $m = \left(\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}\right) / \left(\sum_{i=0}^{n} \frac{1}{\sigma_i^2}\right)$, $v = 1/\left(\sum_{i=0}^{n} \frac{1}{\sigma_i^2}\right)$
- The MAP estimate: $\hat{\theta} = m$.
  - because the normal PDF is maximized at its mean

Point Estimation

- Point estimate: a value that represents our best guess of the value of Θ.
- Estimate: the numerical value $\hat{\theta}$ that we choose on observation x.
- The value of $\hat{\theta}$ is to be determined by applying some function g to the observation x, resulting in $\hat{\theta} = g(x)$.
- Estimator: the random variable $\hat{\Theta} = g(X)$
  - its realized value equals $g(x)$ when $X = x$.

Two popular estimators

- Two popular estimators:
  - MAP: $\hat{\theta} = \arg\max_{\theta} p_{\Theta|X}(\theta|x)$
  - Conditional Expectation: $\hat{\theta} = \mathbf{E}[\Theta|X = x]$.
- Conditional expectation estimator is also called least mean squares (LMS) estimator.
  - It minimizes the mean squared error over all estimators.
  - To be elaborated later.

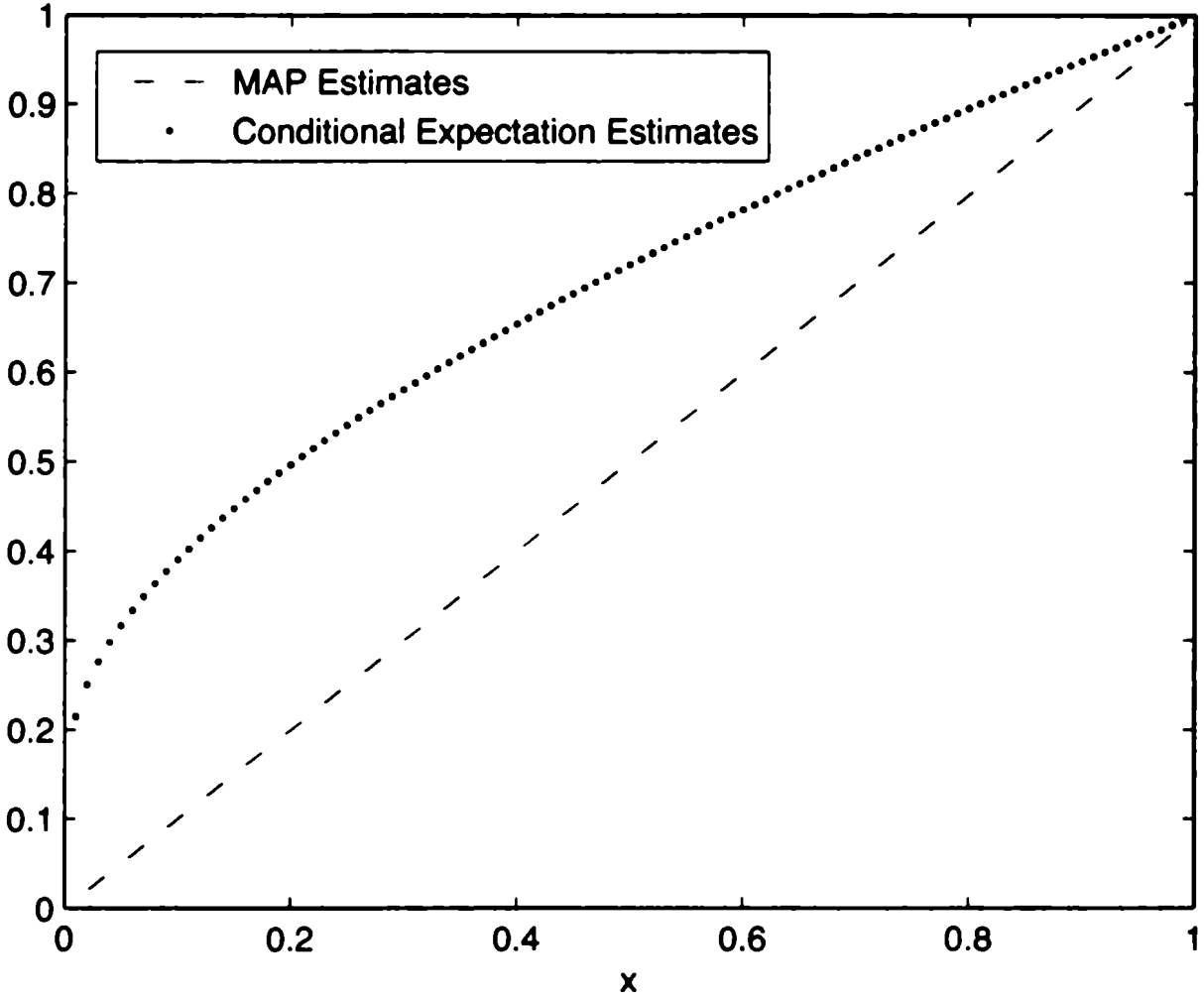
Example: Romeo and Juliet meeting

- Juliet is late on the first date by a random amount X.
- The distribution of X is uniform over $[0, \Theta]$.
- Θ is an unknown random variable with a uniform prior PDF f_Θ over the interval $[0,1]$.
- Recall: $f_{\Theta|X}(\theta|x) = \frac{1}{\theta \cdot |\log x|}$, if $0 \le x \le \theta \le 1$
- MAP: $\hat{\theta} = x$, because $f_{\Theta|X}(\theta|x)$ is decreasing in θ over the range $[x, 1]$.

- Last slide: MAP gives $\hat{\theta} = x$.
- Note that this is an "optimistic" estimate.
  - If Juliet is late by a small amount on the first date $(x \approx 0)$, the estimate of future lateness is also small.
- Conditional expectation: less optimistic.

$$\mathbf{E}[\Theta | X = x] = \int_x^1 \theta \frac{1}{\theta |\log x|} d\theta = \frac{1-x}{|\log x|}.$$

MAP vs. conditional expectation



Hypothesis testing

- Θ takes one of m values, $\theta_1, \ldots, \theta_m$.
  - m is usually a small integer; often $m = 2$.
- The ith hypothesis: the event $H_i \stackrel{\text{def}}{=} \{\Theta = \theta_i\}$.
- Once the value x of X is observed, we may use Bayes' rule to calculate the posterior probabilities

$$P(\Theta = \theta_i | X = x) = P_{\Theta|X}(\theta_i|x),$$

for each i.

- MAP: select the hypothesis H_i with the *largest posterior probability* $P(\Theta = \theta_i | X = x)$.

- Equivalently, it selects a hypothesis H_i with the largest $P_\Theta(\theta_i) P_{X|\Theta}(x|\theta_i)$ (if X is discrete) or $P_\Theta(\theta_i) f_{X|\Theta}(x|\theta_i)$ (if X is continuous).
  - Computational shortcut

Correct probability

- $g_{\mathrm{MAP}}(x)$: the hypothesis selected by the MAP rule when $X = x$,
- The probability of correct decision is
$$P(\Theta = g_{\mathrm{MAP}}(x) | X = x).$$
- If $S_i = \{x: g_{\mathrm{MAP}}(x) = H_i\}$, then the overall probability of correct decision is
$$P\left(\Theta = g_{\mathrm{MAP}}(X)\right) = \sum_i P(\Theta = \theta_i, X \in S_i)$$
- And the corresponding probability of error is
$$\sum_i P(\Theta \neq \theta_i, X \in S_i)$$

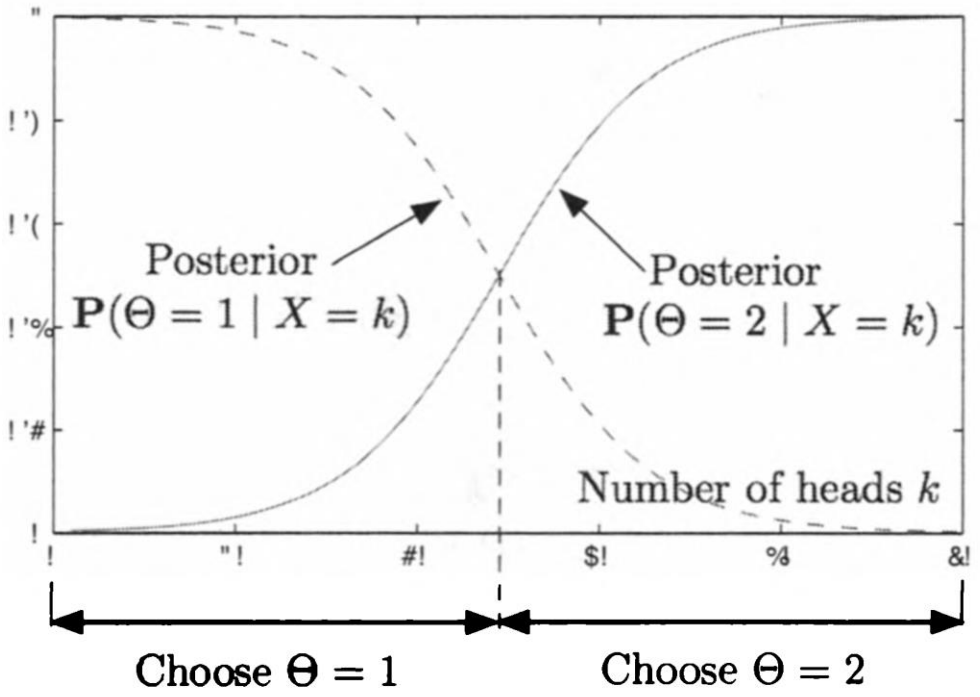
Example: binary hypothesis testing

- Two biased coins, with probabilities of heads equal to p_1 and p_2, respectively.
- We choose a coin at random: either coin is equally likely to be chosen.
  - This gives the prior
- We want to infer its identity, based on the outcome of a single toss.

- Let $\Theta = 1$ and $\Theta = 2$ be the hypotheses that coin 1 or 2, respectively, was chosen.
- $X = \begin{cases} 1 & \text{if head,} \\ 0 & \text{if tail.} \end{cases}$
- MAP: compare $p_\Theta(1)p_{X|\Theta}(x|1)$ and $p_\Theta(2)p_{X|\Theta}(x|2)$, and take the larger one.
- Since $p_\Theta(1) = p_\Theta(2) = 1/2$, we just need to compare $p_{X|\Theta}(x|1)$ and $p_{X|\Theta}(x|2)$.

- For instance, the outcome is tail.
- $P(tail|\Theta = 1) = 1 - p_1$,
  $P(tail|\Theta = 2) = 1 - p_2$.
- So MAP rule selects the H_i with smaller p_i.
- We can also toss the selected coin n times.
- X = the number of heads obtained.
- MAP rule selects the hypothesis under which the observed outcome is most likely.

- If $X = k$, we should decide $\Theta = 1$ if
$$p_1^k(1-p_1)^{n-k} > p_2^k(1-p_2)^{n-k}.$$



- The character of the MAP rule, as illustrated in the above figure, is typical of decision rules in binary hypothesis testing problems.
- It is specified by a *partition of the observation space* into the two disjoint sets in which each of the two hypotheses is chosen.
- In this example, the MAP rule is specified by a single threshold k^*:
- Accept $\Theta = 1$ if $k \leq k^*$, and accept $\Theta = 2$ otherwise.

Content

- Bayesian inference, the posterior distribution
- Point estimation, hypothesis testing, MAP
- Bayesian least mean squares estimation
- Bayesian linear least mean squares estimation

Estimation without observation

- Considering the simpler problem of estimating Θ with a constant $\hat{\theta}$, in the absence of an observation X.
- The estimation error: $\hat{\theta} - \Theta$
- The mean squared error: $E\left[\left(\hat{\theta} - \Theta\right)^2\right]$
- *Question: What's the minimum* $E\left[\left(\hat{\theta} - \Theta\right)^2\right]$ *(over choices of* $\hat{\theta}$*)?*
- *Answer:* $var[\Theta]$, *achieved when* $\hat{\theta} = E[\Theta]$.

proof

- $E\left[\left(\hat{\theta} - \Theta\right)^2\right]$

$= var\left(\Theta - \hat{\theta}\right) + \left(E\left[\Theta - \hat{\theta}\right]\right)^2$ // def of var()

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

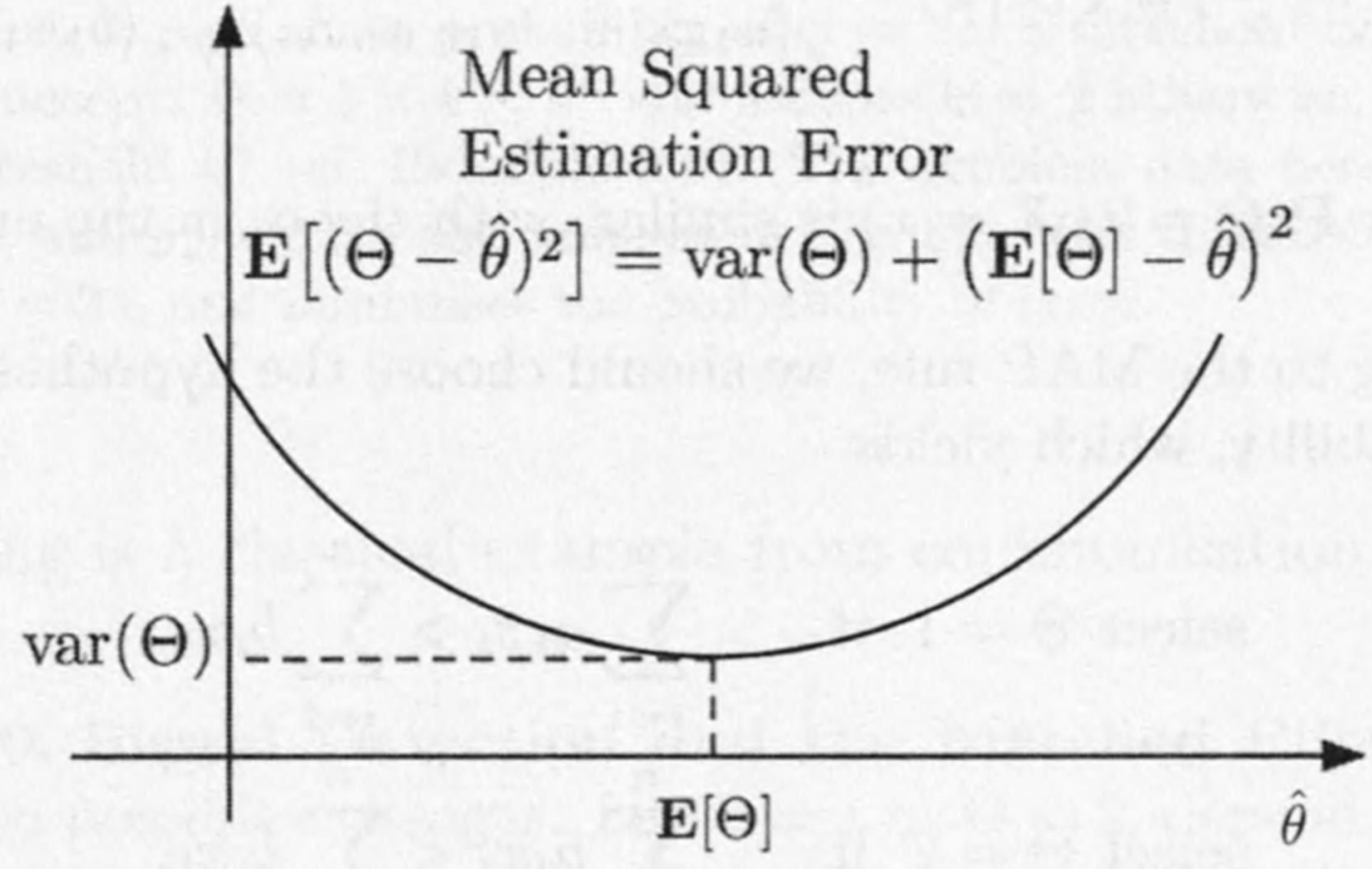
// shifting doesn't change variance

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

// linearity of expectation

$\geq var(\Theta)$

// "=" achieved when $\hat{\theta} = E[\Theta]$.


Mean Squared
Estimation Error
$\mathbf{E}\left[(\Theta - \hat{\theta})^2\right] = \mathrm{var}(\Theta) + \left(\mathbf{E}[\Theta] - \hat{\theta}\right)^2$
$\mathrm{var}(\Theta)$
$\mathbf{E}[\Theta]$
$\hat{\theta}$

Estimation with observation

- Now suppose that we have observation X.
- We still like to estimate Θ to minimize the mean squared error.
- Note that once we know the value x of X, the situation is identical to the one considered earlier, …
- …except that we are now in a new universe: everything is conditioned on $X = x$.

- We can therefore adapt our earlier conclusion.

- And assert that the conditional expectation $E[\Theta|X = x]$ minimizes the conditional mean squared error $E\left[\left(\Theta - \hat{\theta}\right)^2|X = x\right]$ over all constants $\hat{\theta}$.

- Generally, the (unconditional) mean squared estimation error associated with an estimator $g(X)$ is defined as

$$E\left[\left(\Theta - g(X)\right)^2\right].$$

- View $E[\Theta|X]$ as an estimator/function of X, the preceding analysis shows that out of all possible estimators.
- The mean squared estimation error is minimized when

$$g(X) = E[\Theta|X].$$

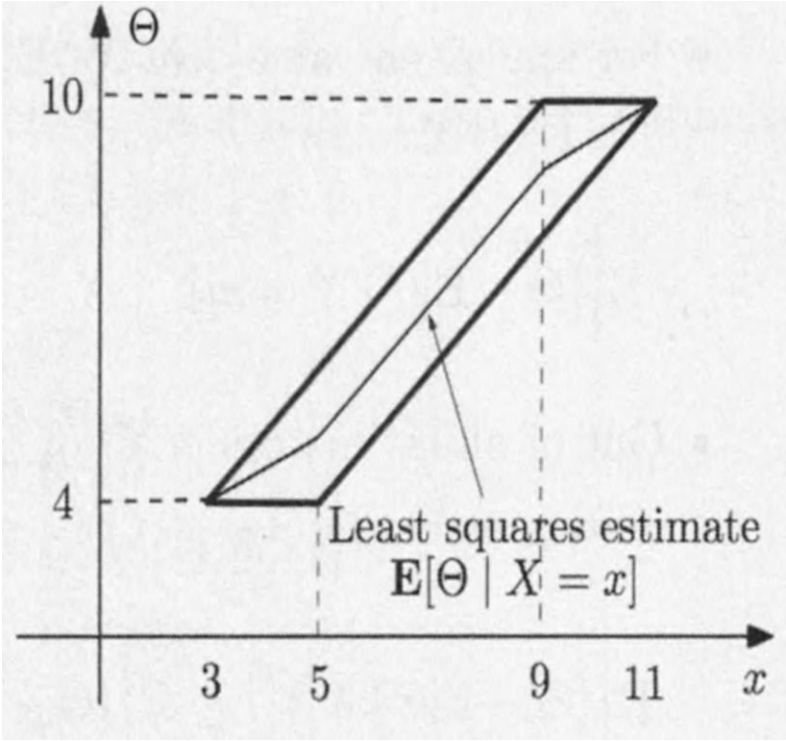
Example

- Θ: uniform over $[4,10]$
- Independent noise W: uniform over $[-1,1]$
- We observe Θ with error W:

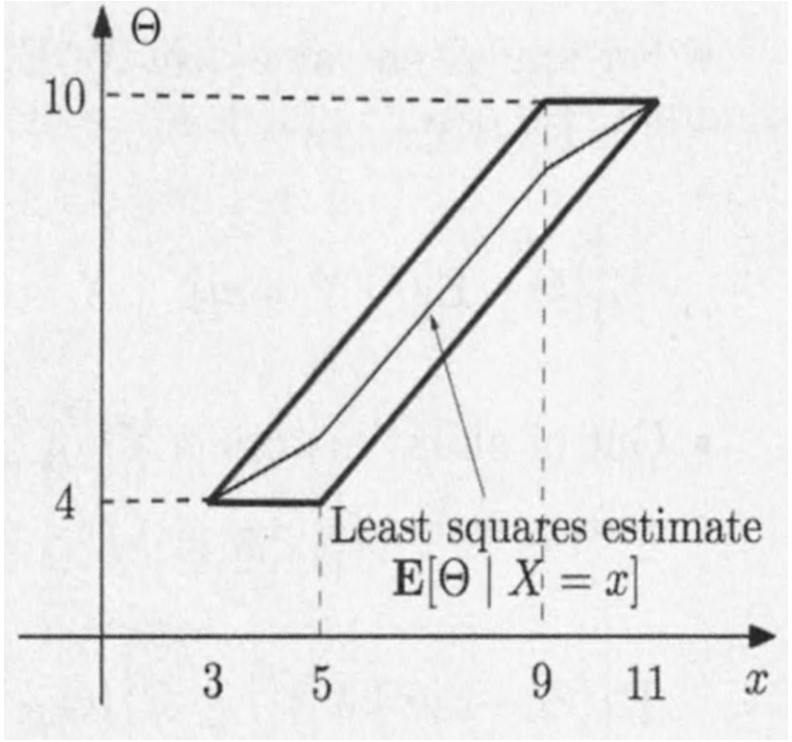
$$X = \Theta + W$$

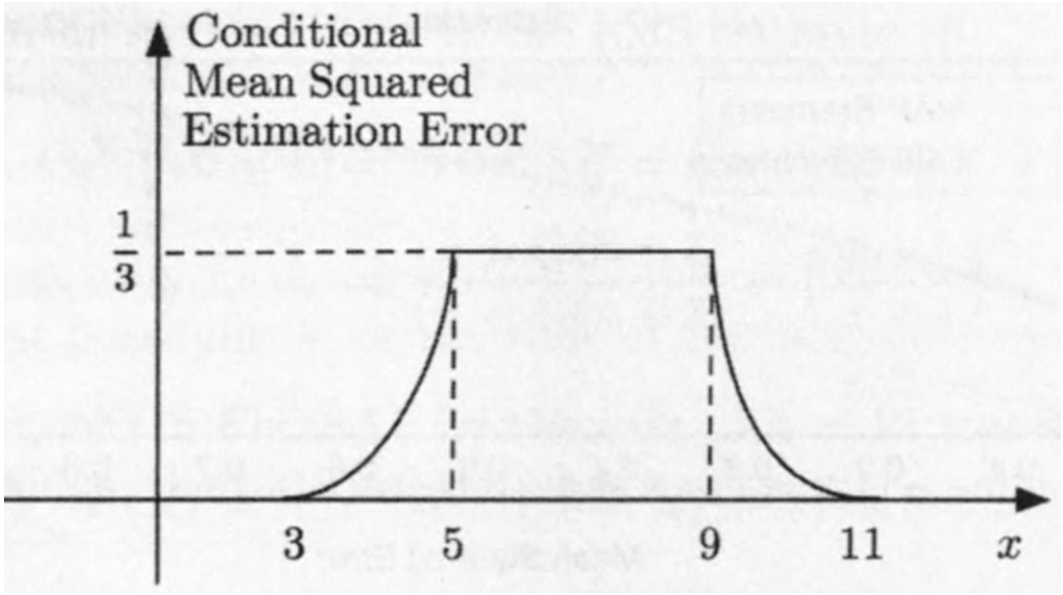
- $f_\Theta(\theta) = 1/6$ if $4 \le \theta \le 10$ (and 0 otherwise).
- $X|\Theta = \theta$ is uniform over $[\theta - 1, \theta + 1]$.
- Joint PDF: $f_{\Theta,X}(\theta, x) = f_\Theta(\theta) f_{X|\Theta}(x|\theta) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$
  - when $\theta \in [4,10]$ and $x \in [\theta - 1, \theta + 1]$.

- The joint PDF of Θ and X is uniform over the parallelogram.
- Given that $X = x$, the posterior PDF $f_{\Theta|X}$ is uniform on the corresponding vertical section of the parallelogram.



- Thus $E[G|X = x]$ is the midpoint of that section, which is a piecewise linear function of x.
- Conditioned on a particular value x of X, define the mean squared error as $E[(\Theta - E[\Theta|X])^2|X = x]$,





- The mean squared error $E[(\Theta - E[\Theta|X])^2|X = x]$, equals the conditional variance of Θ.
- It is a function of x, illustrated in the above figure.

Example: meeting

- Juliet is late on the first date by a random amount X that is uniformly distributed over $[0, \Theta]$.
- Θ: uniform prior over the interval $[0,1]$.
- MAP: $\hat{\theta} = x$.
- LMS: $\hat{\theta} = E[\Theta|X = x] = \int_x^1 \theta \frac{1}{\theta|\log x|} d\theta = \frac{1-x}{|\log x|}$
- Let's calculate the conditional mean squared error for the MAP and the LMS estimates.

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

$$=\int_x^1\left(\hat{\theta}-\theta\right)^2\frac{1}{\theta|\log x|}d\theta$$

$$=\int_x^1\left(\hat{\theta}^2-2\hat{\theta}\theta+\theta^2\right)\frac{1}{\theta|\log x|}d\theta$$

$$=\hat{\theta}^2-\hat{\theta}\,\frac{2(1-x)}{\theta|\log x|}+\frac{1-x^2}{2|\log x|}.$$

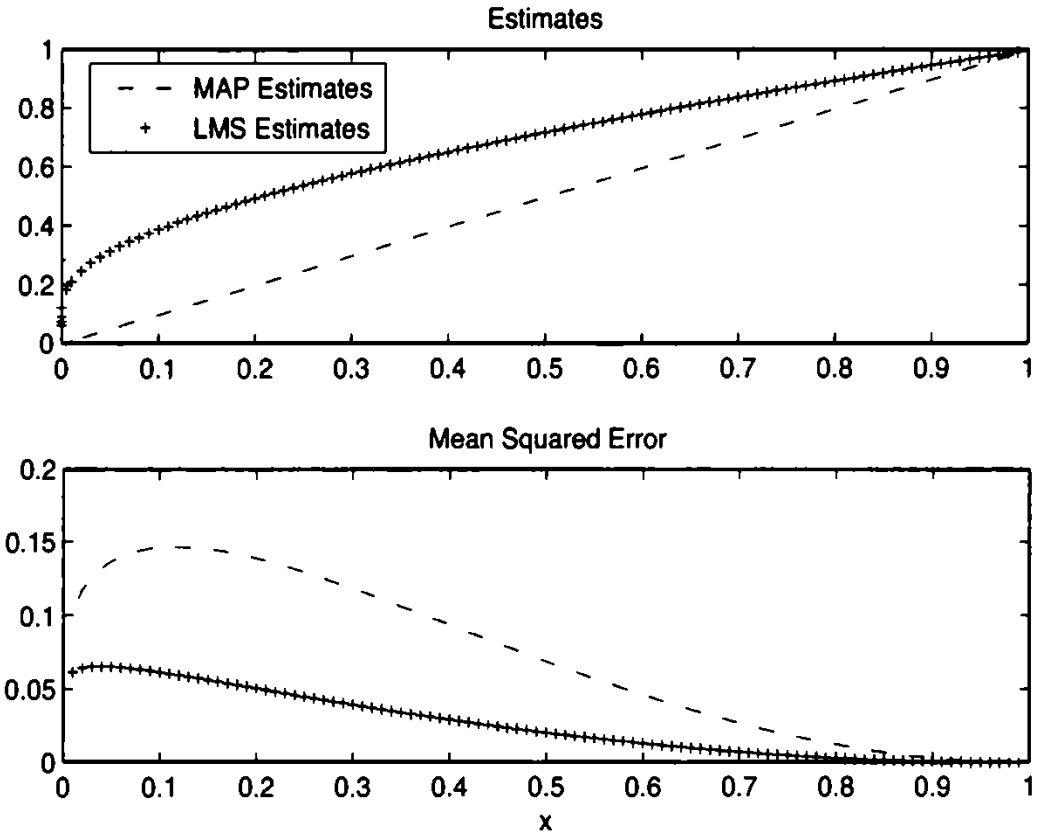
- $E\left[\left(\hat{\theta}-\Theta\right)^2|X=x\right]=\hat{\theta}^2-\hat{\theta}\,\frac{2(1-x)}{\theta|\log x|}+\frac{1-x^2}{2|\log x|}.$
- MAP: $\hat{\theta}=x.$

$$E\left[\left(\hat{\theta}-\Theta\right)^2|X=x\right]=x^2+\frac{3x^2-4x+1}{2|\log x|}$$

- LMS: $\hat{\theta}=\frac{1-x}{|\log x|}.$

$$E\left[\left(\hat{\theta}-\Theta\right)^2|X=x\right]=\frac{1-x^2}{2|\log x|}-\left(\frac{1-x}{\log x}\right)^2$$

- MAP has smaller estimator.
- LMS estimator has uniformly smaller mean squared error.



Properties of estimation error

- Denote $\hat{\Theta} = E[\Theta|X]$, $\tilde{\Theta} = \hat{\Theta} - \Theta$
  - The LMS estimator and the associated estimation error, respectively.
- $E[\tilde{\Theta}] = E[\hat{\Theta} - \Theta] = E[E[\Theta|X]] - E[\Theta] = 0$
- $E[\tilde{\Theta}|X = x] = E[\hat{\Theta} - \Theta|X = x] = E[\hat{\Theta}|X = x] - E[\Theta|X = x] = E[\Theta|X = x] - E[\Theta|X = x] = 0.$

- $\widehat{\Theta} = E[\Theta|X]$, $\widetilde{\Theta} = \widehat{\Theta} - \Theta$
- $E[\widehat{\Theta}\widetilde{\Theta}] = E\left[E[\widehat{\Theta}\widetilde{\Theta}|X]\right]$ // iterated expectation
  $= E\left[\widehat{\Theta}E[\widetilde{\Theta}|X]\right]$ // $\widehat{\Theta}$ depends only on X
  $= 0$ // $E[\widetilde{\Theta}|X = x] = 0, \forall x$. So $E[\widetilde{\Theta}|X] = 0$
- $Cov(\widehat{\Theta}\widetilde{\Theta}) = E[\widehat{\Theta}\widetilde{\Theta}] - E[\widehat{\Theta}]E[\widetilde{\Theta}] = 0 - 0 = 0$.
- Therefore, by considering the variance of both sides in $\Theta = \widetilde{\Theta} + \widehat{\Theta}$, we have

$$var(\Theta) = var(\widehat{\Theta}) + var(\widetilde{\Theta})$$

Content

- Bayesian inference, the posterior distribution
- Point estimation, hypothesis testing, MAP
- Bayesian least mean squares estimation
- Bayesian linear least mean squares estimation

- LMS estimator is sometimes hard to compute, and we need alternatives.
- We derive an estimator that minimizes the mean squared error within a restricted class of estimators: linear functions of the observations.
- This estimator may result in higher mean squared error.
- But it has a significant computational advantage.
  - It requires simple calculations, involving only means, variances, and covariances of the parameters and observations.

- A linear estimator of a random variable Θ, based on observations $X_1, \ldots, X_n$, has the form

$$\widehat{\Theta} = a_1 X_1 + \cdots + a_n X_n + b$$

- Given a particular choice of the scalars $a_1, \ldots, a_n, b$, the corresponding mean squared error is

$$\mathrm{E}\left[\left(\widehat{\Theta} - a_1 X_1 - \cdots - a_n X_n - b\right)^2\right]$$

- The linear LMS estimator chooses $a_1, \ldots, a_n, b$ to minimize the above expression.

- We first develop the solution for the case where $n = 1$, and then generalize.
- The estimator is $\widehat{\Theta} = aX + b$ and the mean squared error is $\mathrm{E}\left[\left(\widehat{\Theta} - aX - b\right)^2\right]$.
- We are interested in finding a and b that minimize this error.

- If a is chosen, then it's easy to find the optimal b:
- Choose a constant b to estimate the random variable $\Theta - aX$.
- By the discussion in previous section, the best choice is $b = E[\Theta - aX] = E[\Theta] - aE[X]$.
- Thus it remains to minimize

$$\mathrm{E}[(\Theta - aX - E[\Theta] + aE[X])^2]$$

which is $var(\Theta - aX)$.

- $$var(\Theta - aX)$$
$$= var(\Theta) + a^2 var(X) + 2 \cdot cov(\Theta, -aX)$$
$$= var(\Theta) + a^2 var(X) - 2a \cdot cov(\Theta, X)$$
- This is minimized when $a = \frac{cov(\Theta,X)}{var(X)} = \rho \frac{\sigma_\Theta}{\sigma_X}$
  - σ_Θ and σ_X: standard deviation of Θ and X, respectively.
  - $\rho = \frac{cov(\Theta,X)}{\sigma_\Theta \sigma_X}$: the correlation coefficient.

- With this choice of a, the estimator

$$\widehat{\Theta} = aX + b = aX + E[\Theta] - aE[X]$$
$$= a(X - E[X]) + E[\Theta]$$
$$= \rho \frac{\sigma_\Theta}{\sigma_X}(X - E[X]) + E[\Theta].$$

- And the mean squared estimation error is

$$var\left(\Theta - \widehat{\Theta}\right) = \left(1 - \rho^2\right) var(\Theta)$$

Example

- Juliet is late by an amount X uniformly distributed over $[0, \Theta]$, and Θ is a random variable with a uniform prior PDF $f_\Theta(\theta)$ over the interval $[0,1]$.
- Let us derive the linear LMS estimator of Θ based on X.
- By law of iterated expectation,

$$E[X] = E\big[E[X|\Theta]\big] = E[\Theta/2] = \frac{E[\Theta]}{2} = \frac{1}{4}$$

- By law of total variance,

$$var(X) = E[var(X|\Theta)] + var(E[X|\Theta])$$
$$= E\left[\frac{\Theta^2}{12}\right] + var\left(\frac{\Theta}{2}\right)$$
$$= \frac{1}{12}\int_0^1 \theta^2 d\theta + \frac{1}{4}\frac{(1-0)^2}{12} = \frac{7}{144}$$

- Now we compute $cov(\Theta, X)$.
- $E[\Theta X] = E\left[E[\Theta X|\Theta]\right] = E\left[\Theta E[X|\Theta]\right]$
$$= E[\Theta^2/2] = 1/6$$

- $cov(\Theta, X) = E[\Theta X] - E[\Theta]E[X] = \frac{1}{6} - \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{24}$
- The linear LMS estimator is
- $\widehat{\Theta} = E[\Theta] + \frac{cov(\Theta,X)}{var(X)}(X - E[X])$

$$= \frac{1}{2} + \frac{1/24}{7/144}\left(X - \frac{1}{4}\right) = \frac{6}{7}X + \frac{2}{7}$$