CS 412

APRIL 19TH – MAXIMUM A POSTERIORI (BAYESIAN) ESTIMATION

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 (MAP)
- Classical/frequentist (MLE)

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.

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 $X = (X_1, \dots, X_n)$ 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.

- How would we know this? Our dataset!

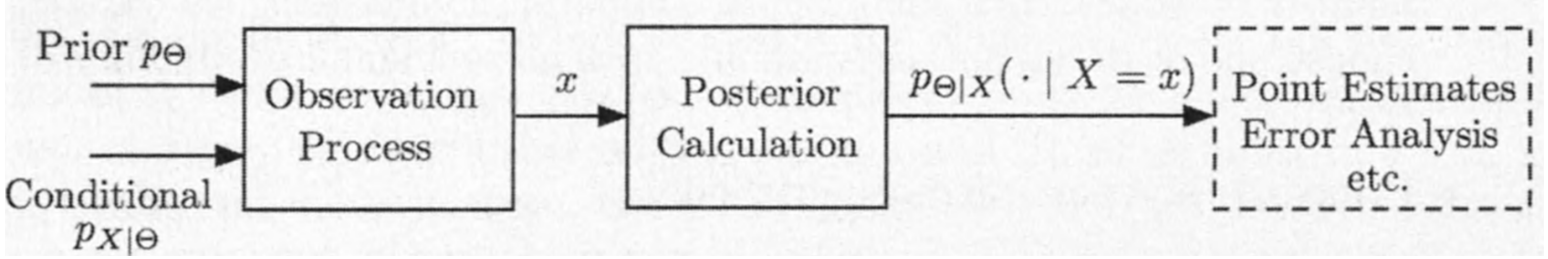
Equivalently, we assume that we know

- 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

We start with a prior distribution p_Θ or f_Θ for the unknown random variable Θ.

We have a model $p_{X|\Theta}$ or $f_{X|\Theta}$ of the observation vector X.

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:

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

- Θ discrete, X continuous:

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

- Θ continuous, X discrete:

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

- Θ continuous, X continuous:

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

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.

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

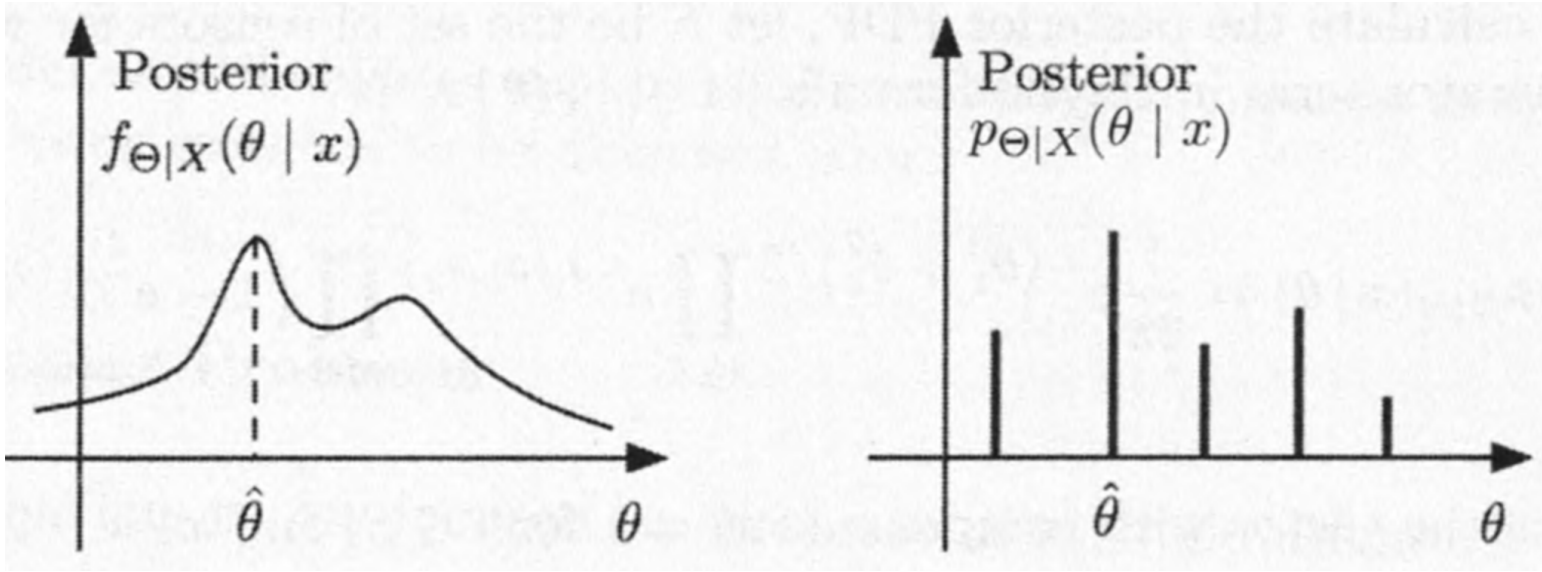
That is,

$\hat{\theta} = \underset{\theta}{\text{argmax}}\, p_{\Theta|X}(\theta|x)$, if Θ is discrete,

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

MAP

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



MAP

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 is more in line with our machine learning (PAC) outcomes

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 $\widehat{\Theta} = g(X)$

- its realized value equals $g(x)$ when $X = x$.

Two popular estimators

Two popular estimators:

- MAP: $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p_{\Theta|X}(\theta|x)$
- 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.

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 $[x, 1]$.

Meeting

Last slide: MAP gives $\hat{\theta} = x$.

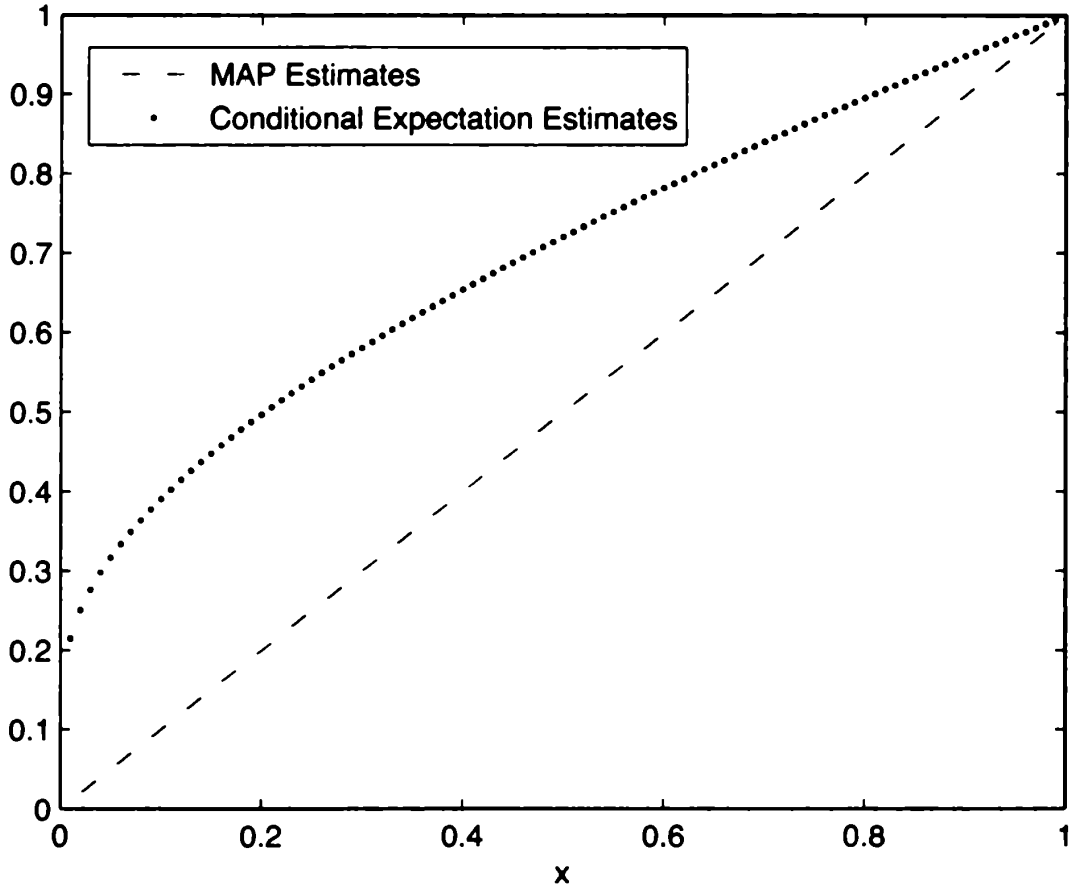
Note that this is an "optimistic" estimate.

- 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.

Hypothesis testing

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

$g\text{MAP}(x)$: the hypothesis selected by the MAP rule when $X = x$,

The probability of correct decision is

$$P(\Theta = g\text{MAP}(x) | X = x).$$

If $S_i = \{x: g\text{MAP}(x) = H_i\}$, then the overall probability of correct decision is

$$P\big(\Theta = g\text{MAP}(X)\big) = \sum_i P(\Theta = \theta_i, X \in S_i)$$

And the corresponding probability of error is

$$\sum_i P(\Theta \neq \theta_i, X \in S_i)$$

Example: binary classification

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.

Binary classification

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 $p_{X|\Theta}(x|2)$.

Binary classification

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.

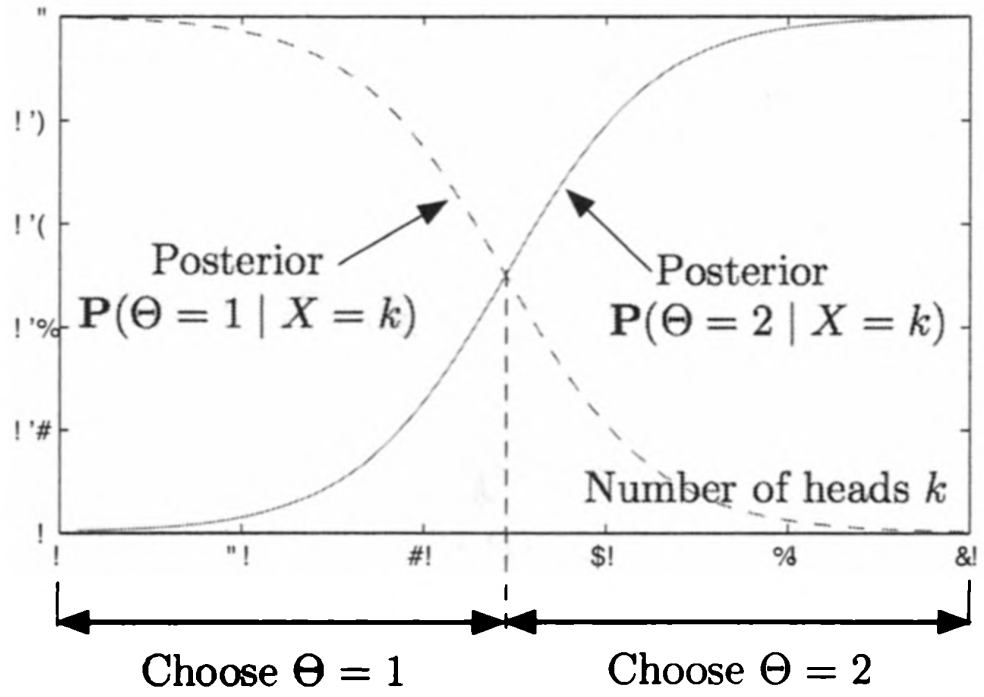
$X =$ the number of heads obtained.

MAP rule selects the hypothesis under which the observed outcome is most likely.

Binary classification

If $X = k$, we should decide $\Theta = 1$ if

$$p_1^k(1-p_1)^{n-k} > p_2^k(1-p_2)^{n-k}.$$



Binary classification

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.

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

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

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

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

// shifting doesn't change variance

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

// linearity of expectation

$$\geq var(\Theta)$$

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

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 $X = x$.

Estimation with observation

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 $\hat{\theta}$.

Estimation with observation

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 $4 \leq \theta \leq 10$ (and 0 otherwise).

$X|\Theta = \theta$ is uniform over $[\theta - 1, \theta + 1]$.

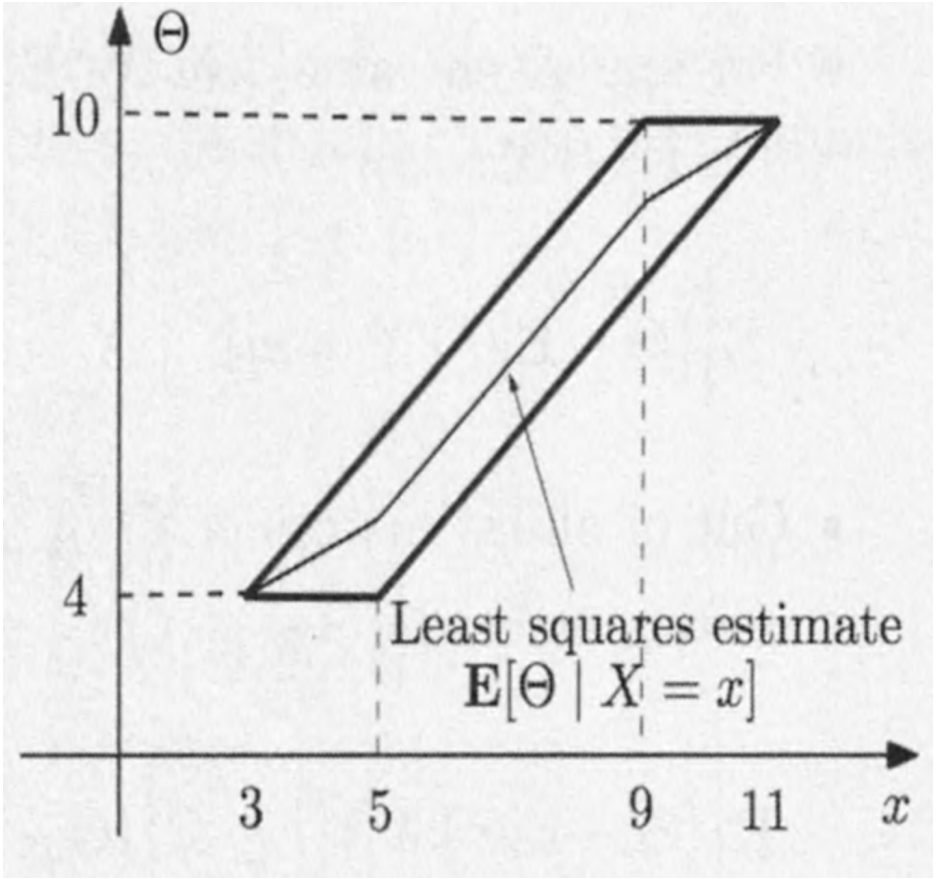
Joint PDF: $f_{\Theta,X}(\theta, x) = f_\Theta(\theta) f_{X|\Theta}(x|\theta) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$

- when $\theta \in [4,10]$ and $x \in [\theta - 1, \theta + 1]$.

Example

The joint PDF of Θ and X is uniform over the parallelogram.

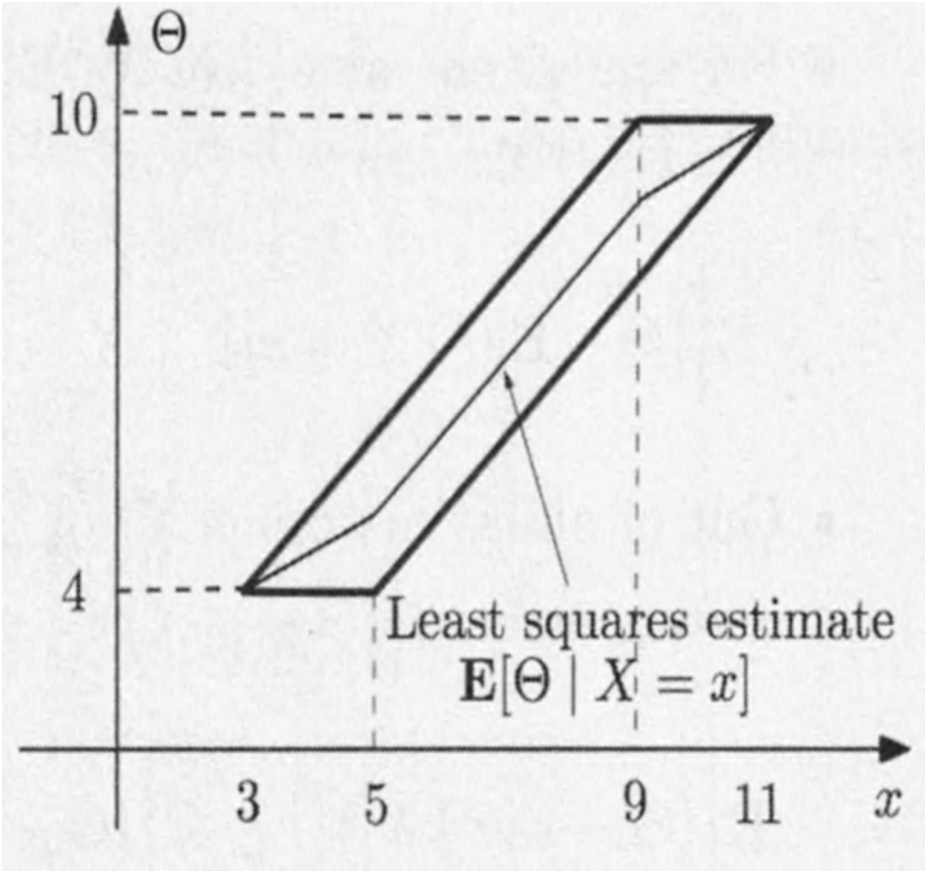
Given that $X = x$, the posterior PDF $f_{\Theta|X}$ is uniform on the corresponding vertical section of the parallelogram.



Example

Thus $E[G|X = x]$ is the midpoint of that section, which is a piecewise linear function of x.

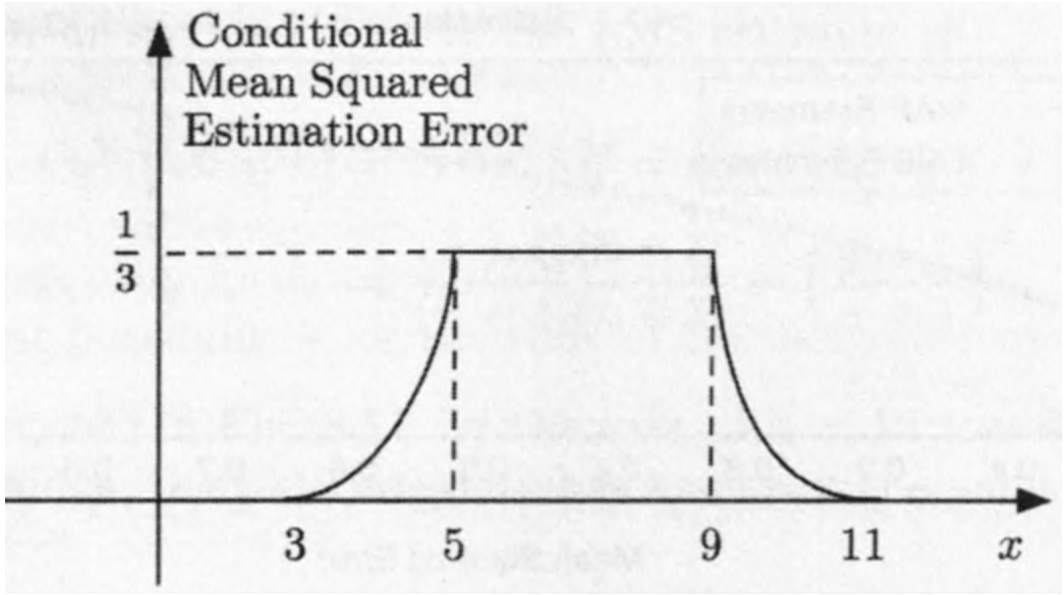
Conditioned on a particular value x of X, define the mean squared error as $E\left[(\Theta - E[\Theta|X])^2|X = x\right]$,



Example

The mean squared error
$E\left[(\Theta - E[\Theta|X])^2|X = x\right]$, 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: $\hat{\theta} = x$.

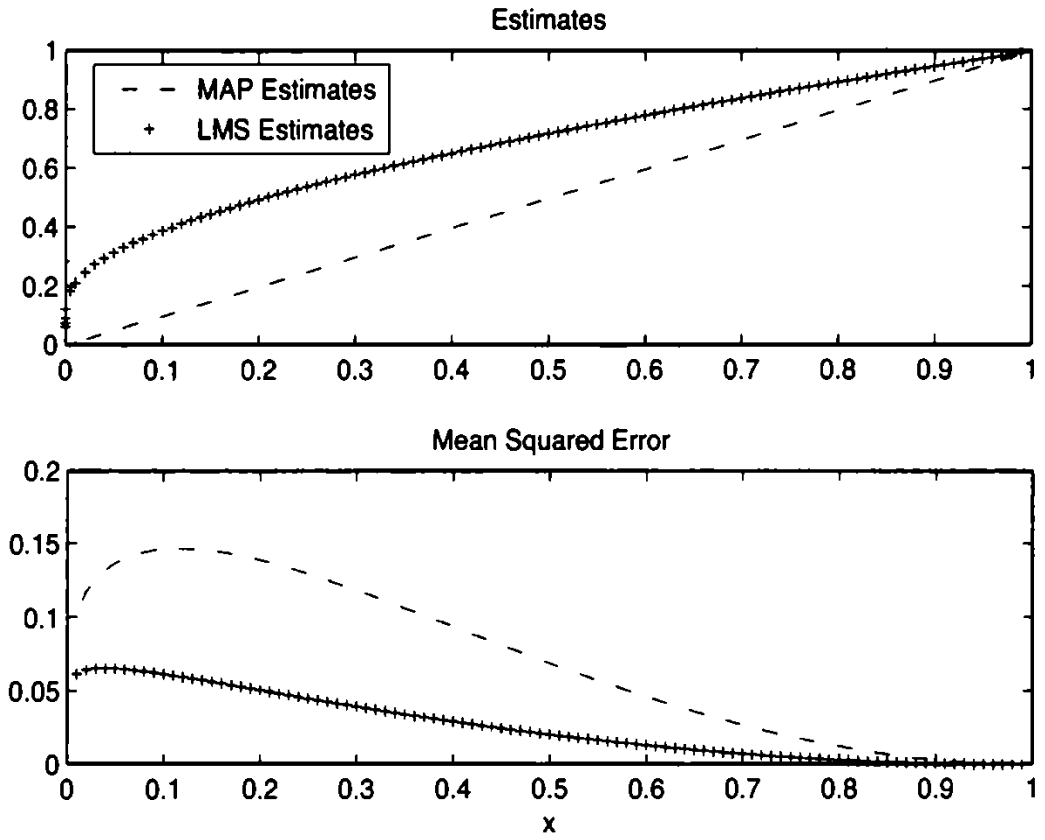
LMS: $\hat{\theta} = E[\Theta | X = x] = \int_x^1 \theta \frac{1}{\theta |\log x|} d\theta = \frac{1-x}{|\log x|}$

Let's calculate the conditional mean squared error for the MAP and the LMS estimates.

Bayesian Least Mean Square

MAP has smaller estimator.

LMS estimator has uniformly smaller mean squared error.



Bayesian Least Mean Square

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.

Next week

Next week

- Deep learning
- Potential strike starting on Tuesday
  - News from the university

Check your HW3 code

- Most of you are still failing because of function header mismatch

Make sure you know when your project meeting is

- Make sure you've done *some* work before you come to meet