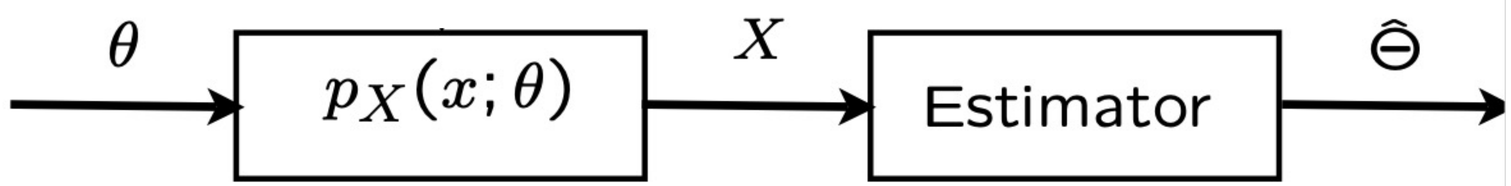
Bayesian Data Analysis

Prof. Pradeep Ravikumar
pradeepr@cs.cmu.edu

Recall: Frequentist Data Analysis

- Frequentist/Classical Data Analysis



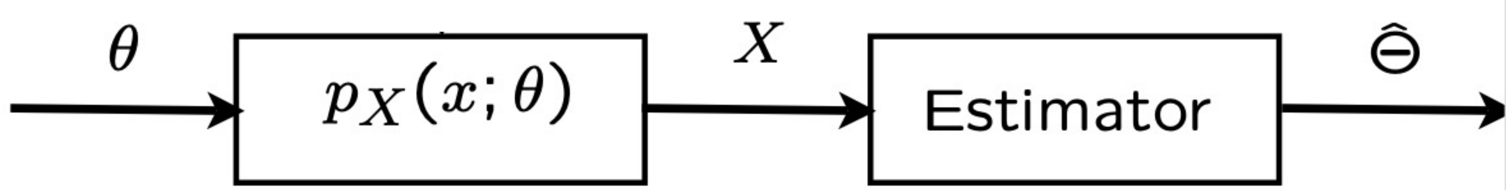
θ: unknown parameter (not a r.v.)

- E.g., $\theta =$ mass of electron

Example: we observe 10 coin flips from a biased coin, what is the bias of the coin θ?

Types of Data Analysis

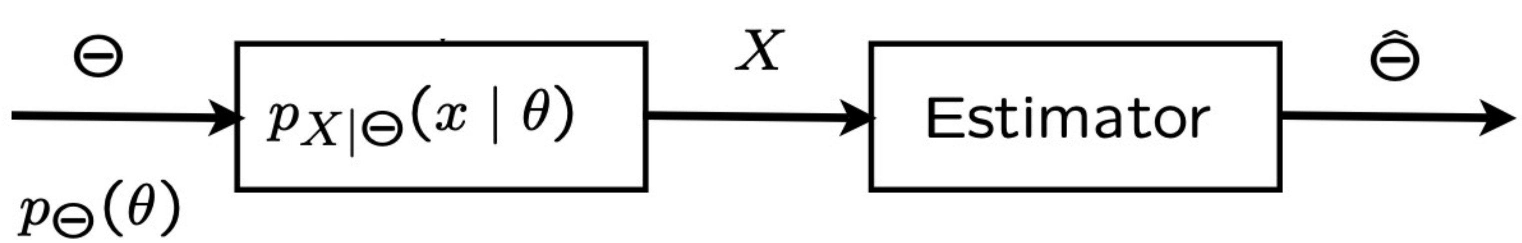
- Frequentist/Classical Data Analysis



θ: unknown parameter (not a r.v.)

  - E.g., $\theta =$ mass of electron

- Bayesian Data Analysis



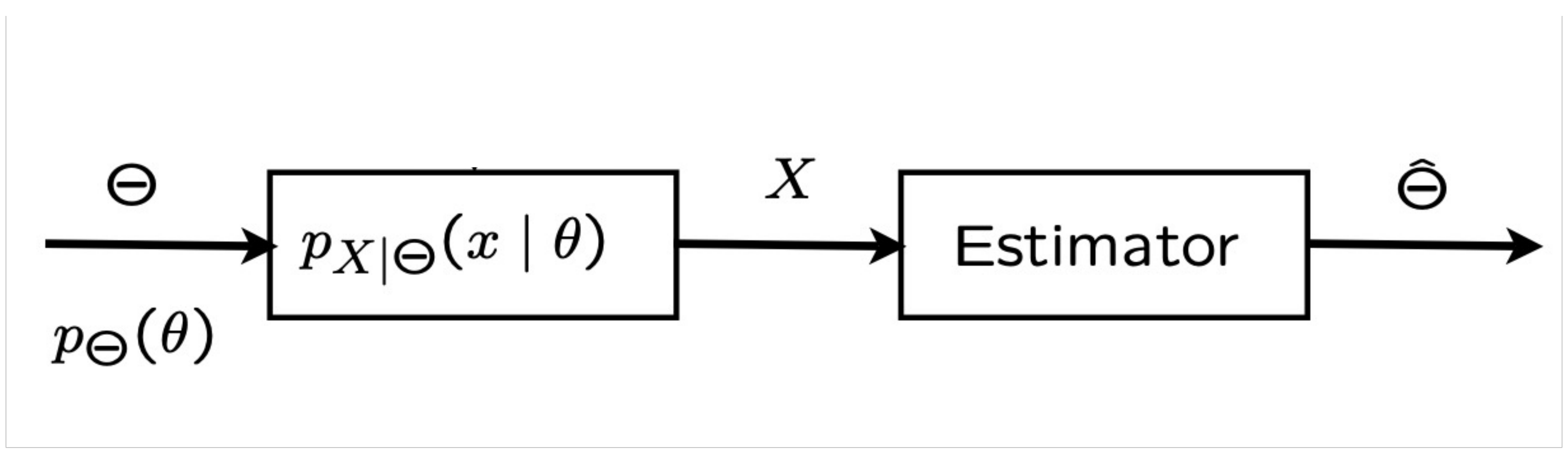
Parameter itself is random!

Bayesian vs Classical

- Classical: Say we have a model with mass of electron as parameter. We might not know the value, but it is nonetheless a constant.
- Bayesian: If we do not know its value completely, use a prior distribution reflecting what we do know.
- Classical: Prior Distribution seems arbitrary.
- Bayesian: Every statistical method makes some choice* ; might as well use a prior to codify these choices.
- Classical: Bayesian methods too difficult to compute (practical considerations)

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.



Recall: Your first consulting job

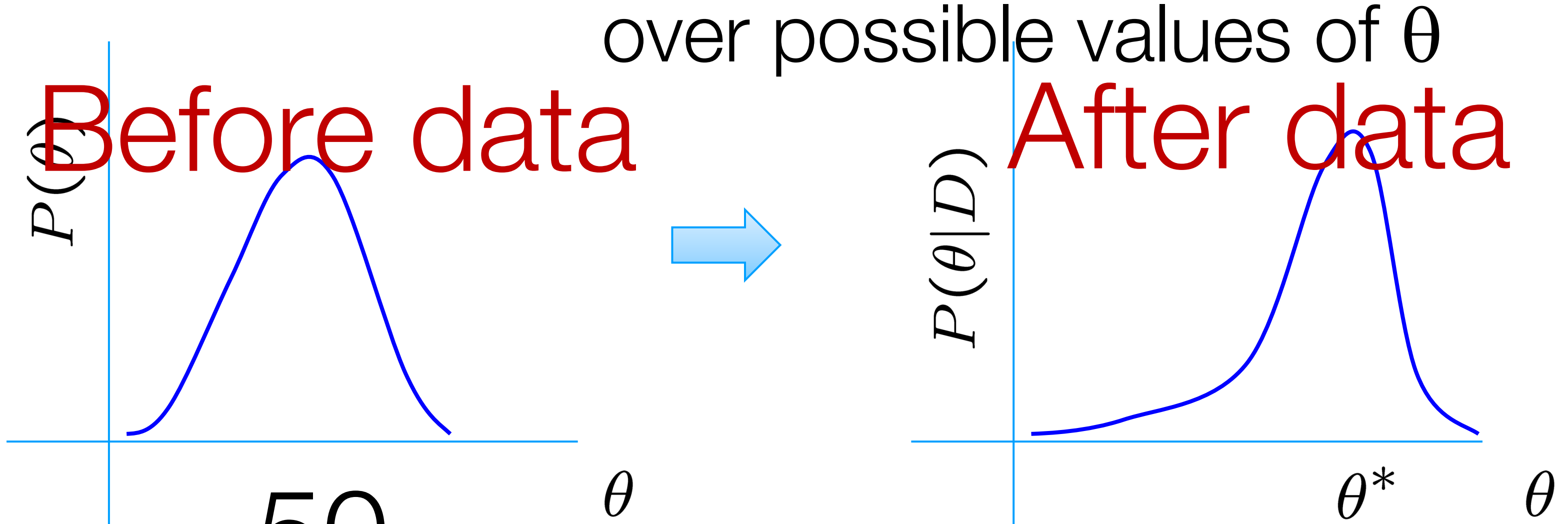
- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have a coin, if I flip it, what's the probability it will fall with the head up?
  - You say: Please flip it a few times:

  - You say: The probability is: **3/5** because… frequency of heads in all flips
  - **He says: But can I put money on this estimate?**
  - You say: ummm…. Maybe not.
    - Not enough flips (less than sample complexity)

What about prior knowledge?

- Billionaire says: Wait, I know that the coin is “close” to 50-50. What can you do for me now?
  - You say: I can learn it the Bayesian way…

- Rather than estimating a single θ, we obtain a distribution over possible values of θ



50

-

50

Bayesian Learning

- Use Bayes rule:

likelihood prior

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

Para
met
ers

Data

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$$

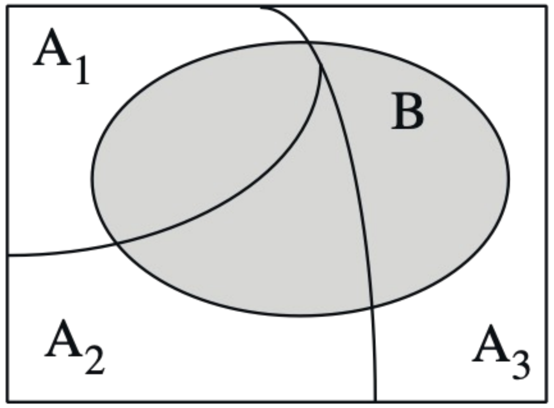
posterior likelihood prior

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

Bayes Rule

- Partition of sample space into A_1, A_2, A_3 :: "Prior Beliefs" $P(A_1)$, $P(A_2)$, $P(A_3)$

- We know $\mathbf{P}(B \mid A_i)$ for each i (for some event B)
- Wish to compute $\mathbf{P}(A_i \mid B)$
  - revise "beliefs", given that B occurred

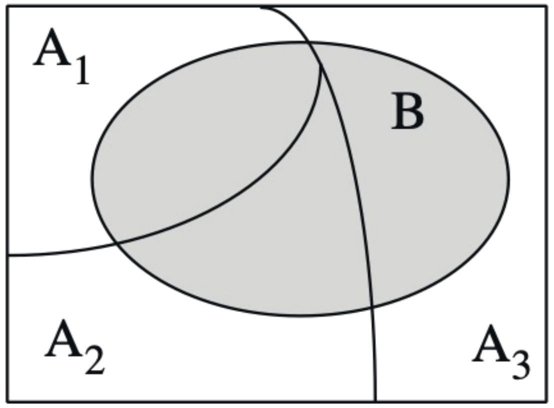


$$\mathbf{P}(A_i \mid B) = \frac{\mathbf{P}(A_i \cap B)}{\mathbf{P}(B)}$$

Bayes Rule

- Partition of sample space into A_1, A_2, A_3 :: "Prior Beliefs" $P(A_1)$, $P(A_2)$, $P(A_3)$

- We know $\mathbf{P}(B \mid A_i)$ for each i (for some event B)
- Wish to compute $\mathbf{P}(A_i \mid B)$
  - revise "beliefs", given that B occurred

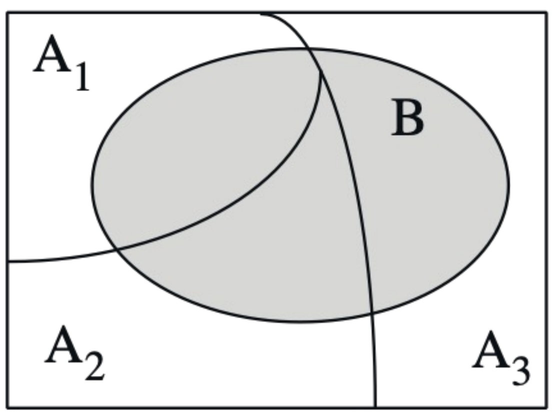


$$\mathbf{P}(A_i \mid B) = \frac{\mathbf{P}(A_i \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\mathbf{P}(B)}$$

Bayes Rule

- Partition of sample space into A_1, A_2, A_3 :: "Prior Beliefs" $P(A_1)$, $P(A_2)$, $P(A_3)$

- We know $\mathbf{P}(B \mid A_i)$ for each i (for some event B)
- Wish to compute $\mathbf{P}(A_i \mid B)$
  - revise "beliefs", given that B occurred

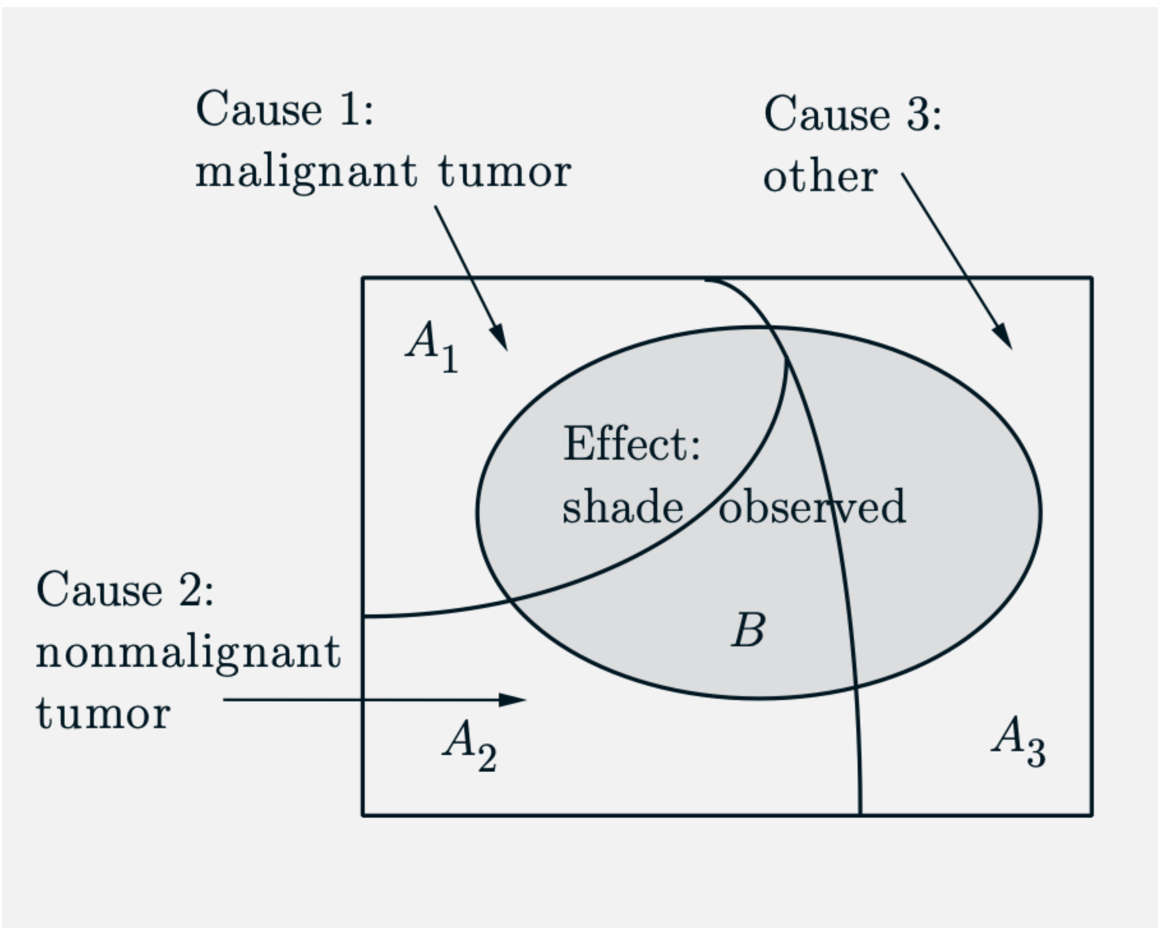


$$\mathbf{P}(A_i \mid B) = \frac{\mathbf{P}(A_i \cap B)}{\mathbf{P}(B)}$$

$$= \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\mathbf{P}(B)}$$

$$= \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\sum_j \mathbf{P}(A_j)\mathbf{P}(B \mid A_j)}$$

Bayes Rule



$$\begin{aligned}
\mathbf{P}(A_i \mid B) &= \frac{\mathbf{P}(A_i \cap B)}{\mathbf{P}(B)} \\
&= \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\mathbf{P}(B)} \\
&= \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\sum_j \mathbf{P}(A_j)\mathbf{P}(B \mid A_j)}
\end{aligned}$$

Given: (a) prior probabilities of causes,
(b) probability of effect given any of the causes

Obtain: probability of any of the causes given effect.

Example I

- A test for a certain rare disease is assumed to be correct 95% of the time: if a person has the disease, the test results are positive with probability 0.95, and if the person does not have the disease, the test results are negative with probability 0.95.

  A random person drawn from a certain population has probability 0.001 of having the disease. Given that the person just tested positive, what is the probability of having the disease?

  - A is the "cause" event that the person has the disease,
    B is the "effect" event that the test results are positive
  - Given :: $P(A)$, $P(A^c)$, and $P(B|A)$, $P(B|A^c)$
    Desired probability :: $P(A \mid B)$, "probability of cause given effect"

Example I

- A test for a certain rare disease is assumed to be correct 95% of the time: if a person has the disease, the test results are positive with probability 0.95, and if the person does not have the disease, the test results are negative with probability 0.95.

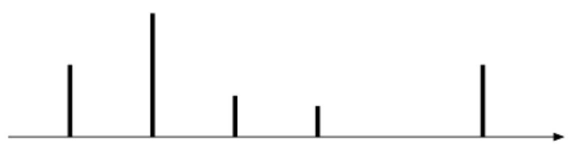
  A random person drawn from a certain population has probability 0.001 of having the disease. Given that the person just tested positive, what is the probability of having the disease?

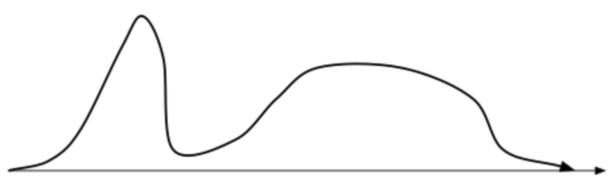
  - A is the "cause" event that the person has the disease,
    B is the "effect" event that the test results are positive

$$
\begin{aligned}
\mathbf{P}(A \mid B) &= \frac{\mathbf{P}(A)\mathbf{P}(B \mid A)}{\mathbf{P}(A)\mathbf{P}(B \mid A) + \mathbf{P}(A^c)\mathbf{P}(B \mid A^c)} \\
&= \frac{0.001 \cdot 0.95}{0.001 \cdot 0.95 + 0.999 \cdot 0.05} \\
&= 0.0187.
\end{aligned}
$$

Output of Bayesian Inference

- Posterior distribution:
  - pmf $p_{\Theta|X}(\cdot \mid x)$ or pdf $f_{\Theta|X}(\cdot \mid x)$





Example Ia

Romeo and Juliet start dating, but Juliet will be late on any date by a random amount X, uniformly distributed over the interval $[0, \theta]$. The parameter θ is unknown and is modeled as the value of a random variable Θ, uniformly distributed between zero and one hour. Assuming that Juliet was late by an amount x on their first date, how should Romeo use this information to update the distribution of Θ?

Example Ia

Romeo and Juliet start dating, but Juliet will be late on any date by a random amount X, uniformly distributed over the interval $[0, \theta]$. The parameter θ is unknown and is modeled as the value of a random variable Θ, uniformly distributed between zero and one hour. Assuming that Juliet was late by an amount x on their first date, 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

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

Example Ia

Romeo and Juliet start dating, but Juliet will be late on any date by a random amount X, uniformly distributed over the interval $[0, \theta]$. The parameter θ is unknown and is modeled as the value of a random variable Θ, uniformly distributed between zero and one hour. Assuming that Juliet was late by an amount x on their first date, 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

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

Bayes' Rule

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

Example Ib

Consider now a variation involving the first n dates. Assume that Juliet is late by random amounts $X_1, \dots, X_n$, which given $\Theta = \theta$, are uniformly distributed in the interval $[0, \theta]$, and conditionally independent.

Example Ib

Consider now a variation involving the first n dates. Assume that Juliet is late by random amounts $X_1, \ldots, X_n$, which given $\Theta = \theta$, are uniformly distributed in the interval $[0, \theta]$, and conditionally independent.

Let $X = (X_1, \ldots, X_n)$ and $x = (x_1, \ldots, x_n)$.

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

Conditional PDF

$$\overline{x} = \max\{x_1, \ldots, x_n\}.$$

Bayes' Rule

$$f_{\Theta|X}(\theta \,|\, x) = \begin{cases} \dfrac{c(\overline{x})}{\theta^n}, & \text{if } \overline{x} \le \theta \le 1, \\ 0, & \text{otherwise,} \end{cases} \qquad c(\overline{x}) = \frac{1}{\displaystyle\int_{\overline{x}}^{1} \frac{1}{(\theta')^n}\, d\theta'}.$$

Example II

Example : Spam Filtering. An email message may be "spam" or "legitimate." We introduce a parameter Θ, taking values 1 and 2, corresponding to spam and legitimate, respectively, with given probabilities $p_\Theta(1)$ and $p_\Theta(2)$. Let $\{w_1, \ldots, w_n\}$ be a collection of special words (or combinations of words) whose appearance suggests a spam message. For each i, let X_i be the Bernoulli random variable that models the appearance of w_i in the message ($X_i = 1$ if w_i appears and $X_i = 0$ if it does not). We assume that the conditional probabilities $p_{X_i|\Theta}(x_i \,|\, 1)$ and $p_{X_i|\Theta}(x_i \,|\, 2)$, $x_i = 0, 1$, are known. For simplicity we also assume that conditioned on Θ, the random variables $X_1, \ldots, X_n$ are independent.

Example II

Example : Spam Filtering. An email message may be "spam" or "legitimate." We introduce a parameter Θ, taking values 1 and 2, corresponding to spam and legitimate, respectively, with given probabilities $p_\Theta(1)$ and $p_\Theta(2)$. Let $\{w_1, \ldots, w_n\}$ be a collection of special words (or combinations of words) whose appearance suggests a spam message. For each i, let X_i be the Bernoulli random variable that models the appearance of w_i in the message ($X_i = 1$ if w_i appears and $X_i = 0$ if it does not). We assume that the conditional probabilities $p_{X_i|\Theta}(x_i \mid 1)$ and $p_{X_i|\Theta}(x_i \mid 2)$, $x_i = 0, 1$, are known. For simplicity we also assume that conditioned on Θ, the random variables $X_1, \ldots, X_n$ are independent.

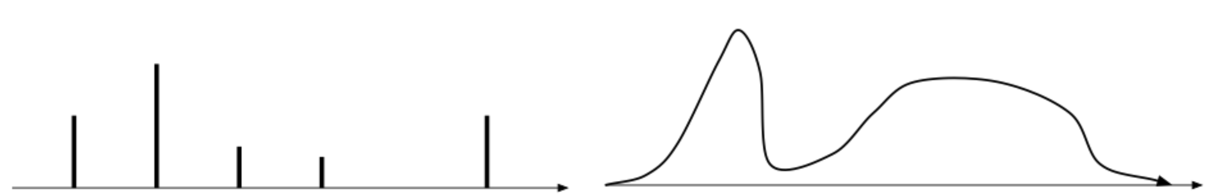
Bayes' Rule

$$\mathbf{P}(\Theta = m \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{p_\Theta(m) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i \mid m)}{\sum_{j=1}^{2} p_\Theta(j) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i \mid j)}, \qquad m = 1, 2.$$

Recall: Bayesian Inference

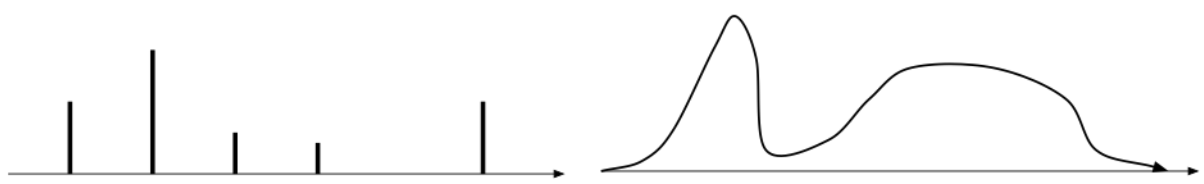
- We start with a prior distribution p_Θ or f_Θ for the unknown random variable Θ.
- We have a model $p_{X|\Theta}$ or $f_{X|\Theta}$ of the observation vector X.
- After observing the value x of X, we form the posterior distribution of Θ, using the appropriate version of Bayes' rule.

- Posterior distribution:
  - pmf $p_{\Theta|X}(\cdot \mid x)$ or pdf $f_{\Theta|X}(\cdot \mid x)$



Output of Bayesian Inference

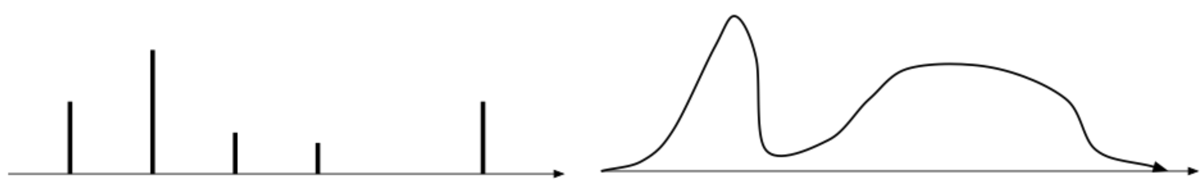
- Posterior distribution:
  - pmf $p_{\Theta|X}(\cdot \mid x)$ or pdf $f_{\Theta|X}(\cdot \mid x)$



- If interested in a single answer?

Output of Bayesian Inference

- Posterior distribution:
  - pmf $p_{\Theta|X}(\cdot \mid x)$ or pdf $f_{\Theta|X}(\cdot \mid x)$

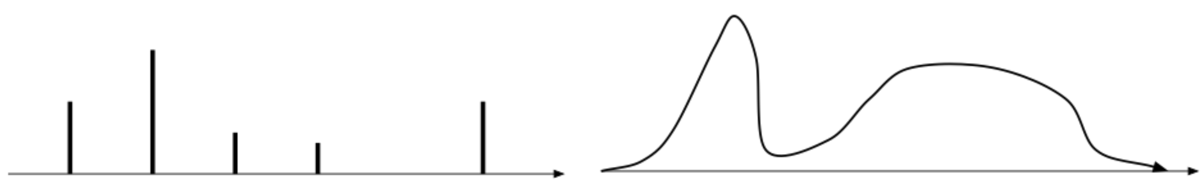


- If interested in a single answer:
  - Maximum a posteriori probability (MAP):
    - $p_{\Theta|X}(\theta^* \mid x) = \max_\theta p_{\Theta|X}(\theta \mid x)$
      minimizes probability of error;
      often used in hypothesis testing
    - $f_{\Theta|X}(\theta^* \mid x) = \max_\theta f_{\Theta|X}(\theta \mid x)$
  - Conditional expectation:

$$\mathbf{E}[\Theta \mid X = y] = \int \theta f_{\Theta|X}(\theta \mid x)\, d\theta$$

Output of Bayesian Inference

- Posterior distribution:
  - pmf $p_{\Theta|X}(\cdot \mid x)$ or pdf $f_{\Theta|X}(\cdot \mid x)$



- If interested in a single answer:
  - Maximum a posteriori probability (MAP):
    - $p_{\Theta|X}(\theta^* \mid x) = \max_\theta p_{\Theta|X}(\theta \mid x)$
      minimizes probability of error;
      often used in hypothesis testing
    - $f_{\Theta|X}(\theta^* \mid x) = \max_\theta f_{\Theta|X}(\theta \mid x)$
  - Conditional expectation:

$$\mathbf{E}[\Theta \mid X = y] = \int \theta f_{\Theta|X}(\theta \mid x)\, d\theta$$

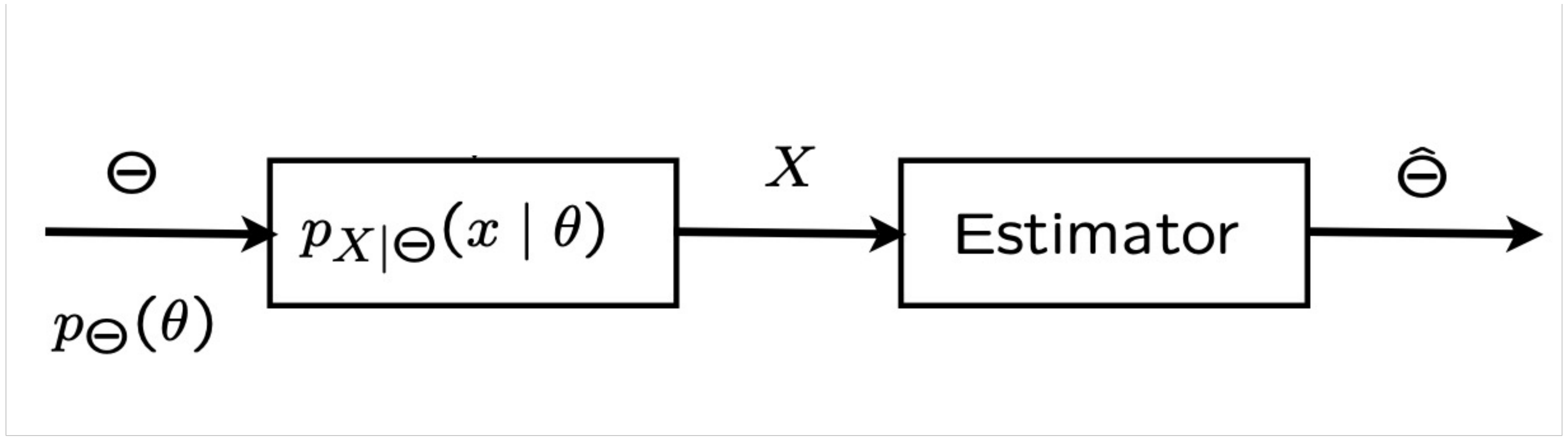
- Single answers can be misleading!

Point Estimation

$$\text{Parameter}: \ \Theta$$
$$\text{Observation}: \ X$$

Given observed value x of X, we want a single numerical value that represents our best guess of Θ

This single numerical value $\widehat{\Theta}$ is called a **point estimate.**



Point Estimation

$$\text{Parameter}: \ \Theta$$
$$\text{Observation}: \ X$$

Given observed value x of X, we want a single numerical value that represents our best guess of Θ

This single numerical value $\widehat{\Theta}$ is called a **point estimate.**

We get it by applying some function g to the observation x, $\widehat{\Theta} = g(x)$

Point Estimation

$$\text{Parameter}: \ \Theta$$

$$\text{Observation}: \ X$$

Given observed value x of X, we want a single numerical value that represents our best guess of Θ

This single numerical value $\widehat{\Theta}$ is called a **point estimate.**

We get it by applying some function g to the observation x, $\widehat{\Theta} = g(x)$

Point Estimator: $\widehat{\Theta}(X) = g(X)$

Note: a random variable

Point Estimators

- Note that $\widehat{\Theta}(X) = 1$ is an estimator as well!
  - Though not a particularly good one.
- In today's class:

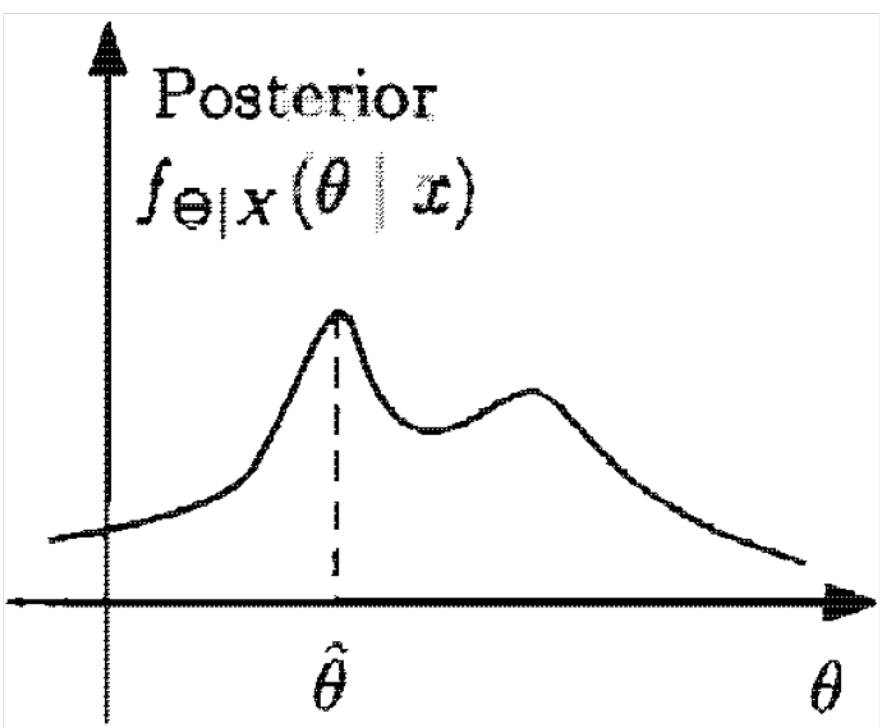
$$\text{MAP Estimator}: \ \widehat{\Theta}(X) = \arg\max_{\Theta} p_{\Theta}(\Theta|X)$$

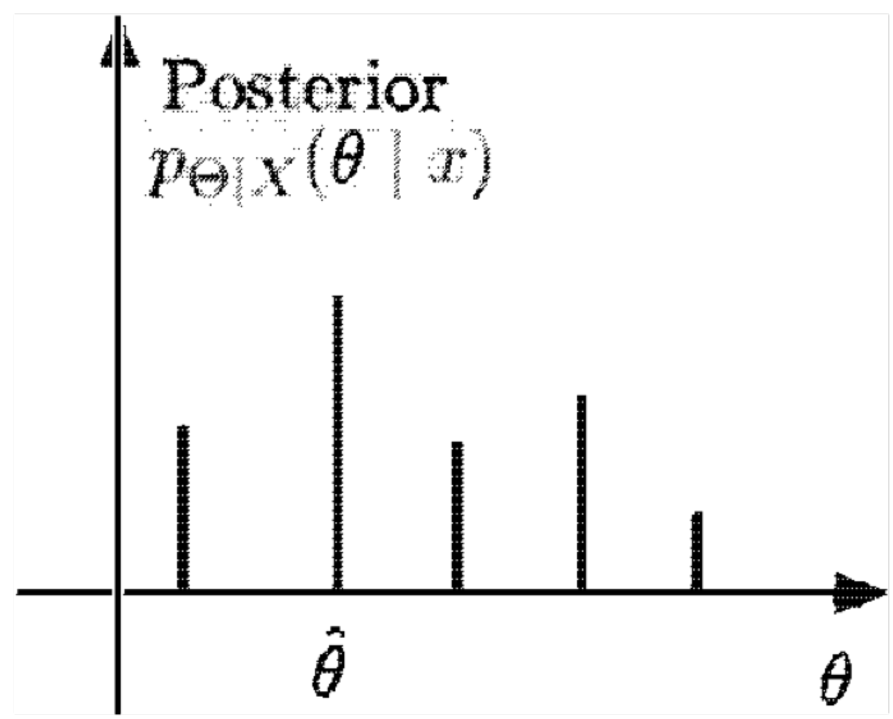
$$\text{Conditional Expectation}: \ \widehat{\Theta}(X) = \mathbb{E}(\Theta|X)$$

MAP Estimator

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

$$\hat{\theta} = \arg\max_{\theta} f_{\Theta|X}(\theta \mid x), \qquad (\Theta \text{ continuous}).$$





Optimality Properties

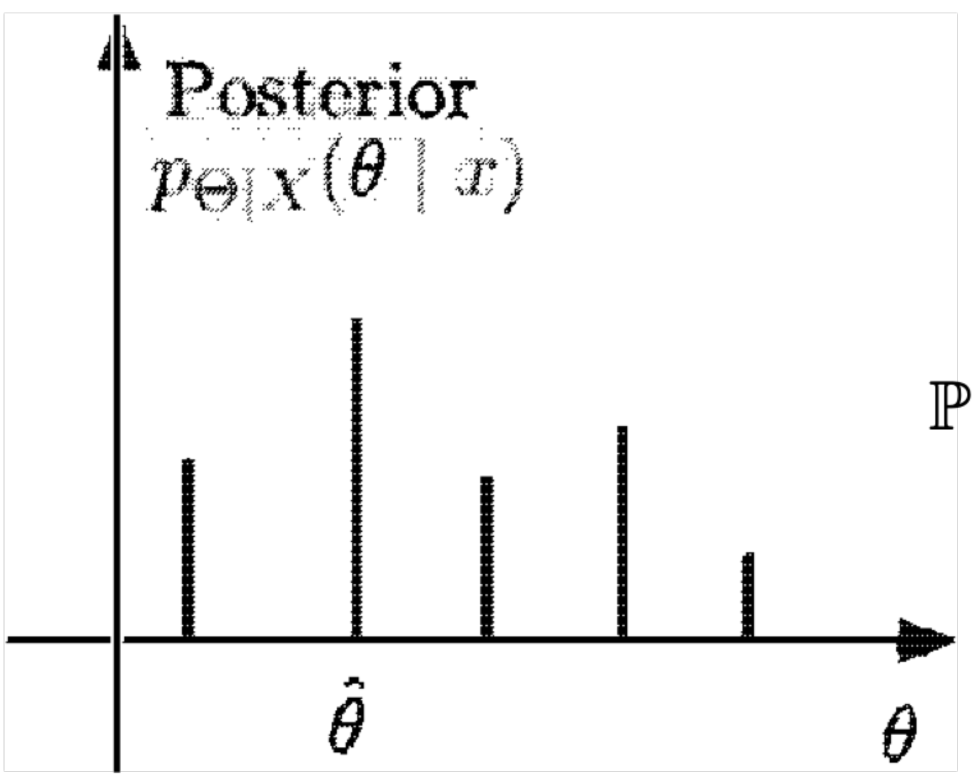
- Theorem: When Θ is discrete, the MAP estimator maximizes the probability of correct decision, given the observation x.

$$\mathbb{P}(\Theta = g(X)|X) \leq \mathbb{P}(\Theta = g_{\mathrm{map}}(X)|X)$$

Optimality Properties

- Theorem: When Θ is discrete, the MAP estimator maximizes the probability of correct decision, given the observation x.

$$\mathbb{P}(\Theta = g(X)|X) \leq \mathbb{P}(\Theta = g_{\text{map}}(X)|X)$$



$$\mathbb{P}(\Theta = g(X)|X = x) \leq \mathbb{P}(\Theta = g_{\text{map}}(X)|X = x)$$

MAP Rule

- Given the observation value x, the MAP rule selects a value $\hat{\theta}$ that maximizes over θ the posterior distribution $p_{\Theta|X}(\theta \,|\, x)$ (if Θ is discrete) or $f_{\Theta|X}(\theta \,|\, x)$ (if Θ is continuous).
- Equivalently, it selects $\hat{\theta}$ that maximizes over θ:

$$
\begin{array}{ll}
p_\Theta(\theta) p_{X|\Theta}(x \,|\, \theta) & \text{(if } \Theta \text{ and } X \text{ are discrete)}, \\
p_\Theta(\theta) f_{X|\Theta}(x \,|\, \theta) & \text{(if } \Theta \text{ is discrete and } X \text{ is continuous)}, \\
f_\Theta(\theta) p_{X|\Theta}(x \,|\, \theta) & \text{(if } \Theta \text{ is continuous and } X \text{ is discrete)}, \\
f_\Theta(\theta) f_{X|\Theta}(x \,|\, \theta) & \text{(if } \Theta \text{ and } X \text{ are continuous)}.
\end{array}
$$

... since the denominator depends only on x, and is the same for all \theta

Example I

Example : Consider example prev. where Juliet is late on the first date by a random amount X. The distribution of X is uniform over the interval $[0, \Theta]$, and Θ is an unknown random variable with a uniform prior PDF f_Θ over the interval $[0, 1]$. In that example, we saw that for $x \in [0, 1]$, the posterior PDF is

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

Example I

Example : Consider example prev. where Juliet is late on the first date by a random amount X. The distribution of X is uniform over the interval $[0, \Theta]$, and Θ is an unknown random variable with a uniform prior PDF f_Θ over the interval $[0, 1]$. In that example, we saw that for $x \in [0, 1]$, the posterior PDF is

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

For a given x, $f_{\Theta|X}(\theta \,|\, x)$ is decreasing in θ, over the range $[x, 1]$ of possible values of Θ. Thus, the MAP estimate is equal to 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.

Example II

Example : Here the parameter Θ takes values 1 and 2, corresponding to spam and legitimate messages, respectively, with given probabilities $p_\Theta(1)$ and $p_\Theta(2)$, and X_i is the Bernoulli random variable that models the appearance of w_i in the message ($X_i = 1$ if w_i appears and $X_i = 0$ if it does not). We have calculated the posterior probabilities of spam and legitimate messages as

$$\mathbf{P}(\Theta = m \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{p_\Theta(m) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i \mid m)}{\sum_{j=1}^{2} p_\Theta(j) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i \mid j)}, \qquad m = 1, 2.$$

Example II

Example : Here the parameter Θ takes values 1 and 2, corresponding to spam and legitimate messages, respectively, with given probabilities $p_\Theta(1)$ and $p_\Theta(2)$, and X_i is the Bernoulli random variable that models the appearance of w_i in the message ($X_i = 1$ if w_i appears and $X_i = 0$ if it does not). We have calculated the posterior probabilities of spam and legitimate messages as

$$\mathbf{P}(\Theta = m \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{p_\Theta(m)\prod_{i=1}^{n} p_{X_i|\Theta}(x_i \mid m)}{\sum_{j=1}^{2} p_\Theta(j)\prod_{i=1}^{n} p_{X_i|\Theta}(x_i \mid j)}, \qquad m = 1, 2.$$

Suppose we want to classify a message as spam or legitimate based on the corresponding vector $(x_1, \ldots, x_n)$. Then, the MAP rule decides that the message is spam if

Example II

Example : Here the parameter Θ takes values 1 and 2, corresponding to spam and legitimate messages, respectively, with given probabilities $p_\Theta(1)$ and $p_\Theta(2)$, and X_i is the Bernoulli random variable that models the appearance of w_i in the message ($X_i = 1$ if w_i appears and $X_i = 0$ if it does not). We have calculated the posterior probabilities of spam and legitimate messages as

$$\mathbf{P}(\Theta = m \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{p_\Theta(m) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i \mid m)}{\sum_{j=1}^{2} p_\Theta(j) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i \mid j)}, \qquad m = 1, 2.$$

Suppose we want to classify a message as spam or legitimate based on the corresponding vector $(x_1, \ldots, x_n)$. Then, the MAP rule decides that the message is spam if

$$\mathbf{P}(\Theta = 1 \mid X_1 = x_1, \ldots, X_n = x_n) > \mathbf{P}(\Theta = 2 \mid X_1 = x_1, \ldots, X_n = x_n),$$

or equivalently, if

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

MAP Estimate with Uniform priors

- Suppose we are doing Bayesian inference, and want to compute the MAP estimate.
- Suppose the prior is $p_\Theta(\theta)$, and the conditional PMF of data, $p_{X|\Theta}(x|\theta)$.
- Then the MAP Estimate:

$$\hat{\theta}_{\text{map}} = \arg\max_{\theta} \left\{ p_\Theta(\theta)\, p_{X|\Theta}(x|\theta) \right\}$$

MAP Estimate with Uniform priors

- Suppose we are doing Bayesian inference, and want to compute the MAP estimate.
- Suppose the prior is $p_\Theta(\theta)$, and the conditional PMF of data, $p_{X|\Theta}(x|\theta)$.
- Then the MAP Estimate:

$$\hat{\theta}_{\text{map}} = \arg\max_{\theta} \left\{ p_\Theta(\theta)\, p_{X|\Theta}(x|\theta) \right\}$$

- Suppose the prior $p_\Theta(\theta)$ is uniform. Then the MAP Estimate:

$$\hat{\theta}_{\text{map}} = \arg\max_{\theta} \left\{ p_{X|\Theta}(x|\theta) \right\}$$

MAP Estimate with Uniform priors

- Suppose we are doing Bayesian inference, and want to compute the MAP estimate.
- Suppose the prior is $p_\Theta(\theta)$, and the conditional PMF of data, $p_{X|\Theta}(x|\theta)$.
- Then the MAP Estimate:

$$\hat{\theta}_{\text{map}} = \arg\max_\theta \left\{ p_\Theta(\theta)\, p_{X|\Theta}(x|\theta) \right\}$$

- Suppose the prior $p_\Theta(\theta)$ is uniform. Then the MAP Estimate:

$$\hat{\theta}_{\text{map}} = \arg\max_\theta \left\{ p_{X|\Theta}(x|\theta) \right\}$$

- This is exactly the ML estimate, if we were doing classical statistics,

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

Compare: MLE and MAP

- Model, with unknown parameter(s): $X \sim p_X(x;\theta)$
- Pick θ that "makes data most likely"

$$\widehat{\theta}_{\mathsf{ML}} = \arg\max_{\theta} p_X(x;\theta)$$

- Compare to Bayesian MAP estimation:

$$\widehat{\theta}_{\mathsf{MAP}} = \max_{\theta} \frac{p_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{p_X(x)}$$

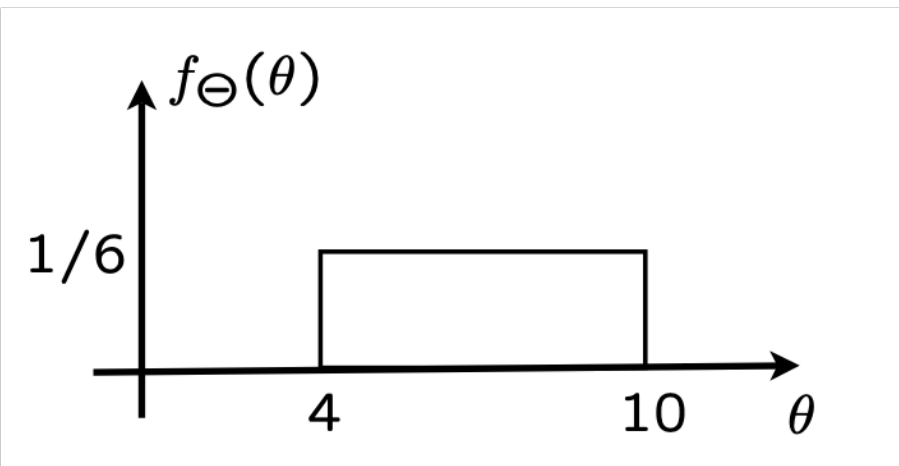
MLE vs MAP

You are no good when sample is small

You give a different answer for different priors

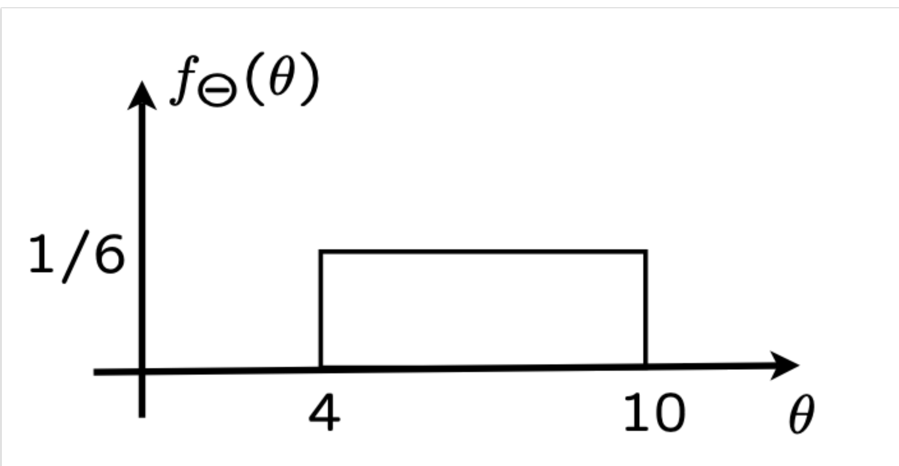
Least Mean Squares Estimation

- Estimation in the absence of information



Least Mean Squares Estimation

- Estimation in the absence of information

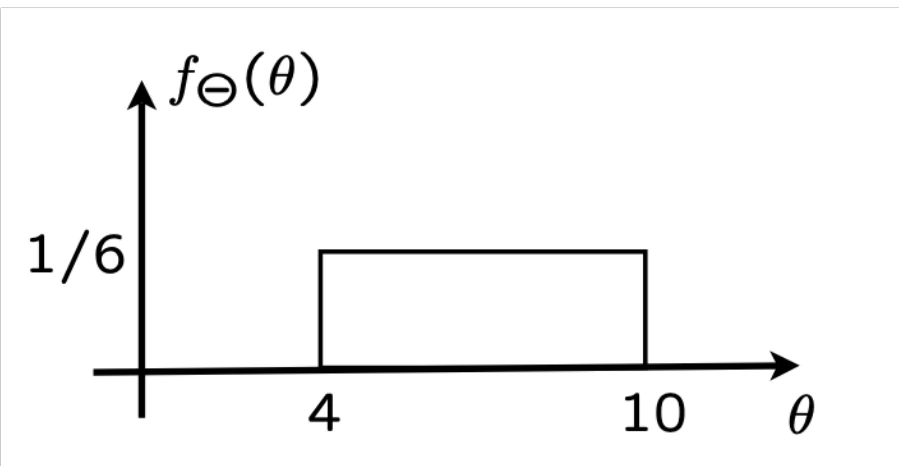


- find estimate c, to:

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

Least Mean Squares Estimation

- Estimation in the absence of information



- find estimate c, to:

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

- Optimal estimate: $c = \mathbf{E}[\Theta]$

- Optimal mean squared error:

$$\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta])^2\right] = \mathrm{Var}(\Theta)$$

LMS Estimation

- Two r.v.'s Θ, X
- we observe that $X = x$
  - new universe: condition on $X = x$
- $\mathbf{E}\left[(\Theta - c)^2 \mid X = x\right]$ is minimized by

  $c =$

LMS Estimation

- Two r.v.'s Θ, X
- we observe that $X = x$
  - new universe: condition on $X = x$
- $\mathbf{E}\left[(\Theta - c)^2 \mid X = x\right]$ is minimized by

  $c =$
- $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta \mid X = x])^2 \mid X = x\right]$

  $$\leq \mathbf{E}[(\Theta - g(x))^2 \mid X = x]$$

LMS Estimation

- Two r.v.'s Θ, X
- we observe that $X = x$
  - new universe: condition on $X = x$
- $\mathbf{E}\left[(\Theta - c)^2 \mid X = x\right]$ is minimized by

  $c =$
- $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta \mid X = x])^2 \mid X = x\right]$

$$\leq \mathbf{E}[(\Theta - g(x))^2 \mid X = x]$$

∘ $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta \mid X])^2 \mid X\right] \leq \mathbf{E}\left[(\Theta - g(X))^2 \mid X\right]$

LMS Estimation

- Two r.v.'s Θ, X
- we observe that $X = x$
  - new universe: condition on $X = x$
- $\mathbf{E}\left[(\Theta - c)^2 \mid X = x\right]$ is minimized by

  $c =$

- $$\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta \mid X = x])^2 \mid X = x\right] \le \mathbf{E}[(\Theta - g(x))^2 \mid X = x]$$

∘ $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta \mid X])^2 \mid X\right] \le \mathbf{E}\left[(\Theta - g(X))^2 \mid X\right]$

∘ $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta \mid X])^2\right] \le \mathbf{E}\left[(\Theta - g(X))^2\right]$ ... Law of iterated expectations

$\boxed{\mathbf{E}[\Theta \mid X] \text{ minimizes } \mathbf{E}\left[(\Theta - g(X))^2\right] \text{ over all estimators } g(\cdot)}$

Example I

Example : Consider example prev. where Juliet is late on the first date by a random amount X. The distribution of X is uniform over the interval $[0, \Theta]$, and Θ is an unknown random variable with a uniform prior PDF f_Θ over the interval $[0, 1]$. In that example, we saw that for $x \in [0, 1]$, the posterior PDF is

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

Example I

Example : Consider example prev. where Juliet is late on the first date by a random amount X. The distribution of X is uniform over the interval $[0, \Theta]$, and Θ is an unknown random variable with a uniform prior PDF f_Θ over the interval $[0, 1]$. In that example, we saw that for $x \in [0, 1]$, the posterior PDF is

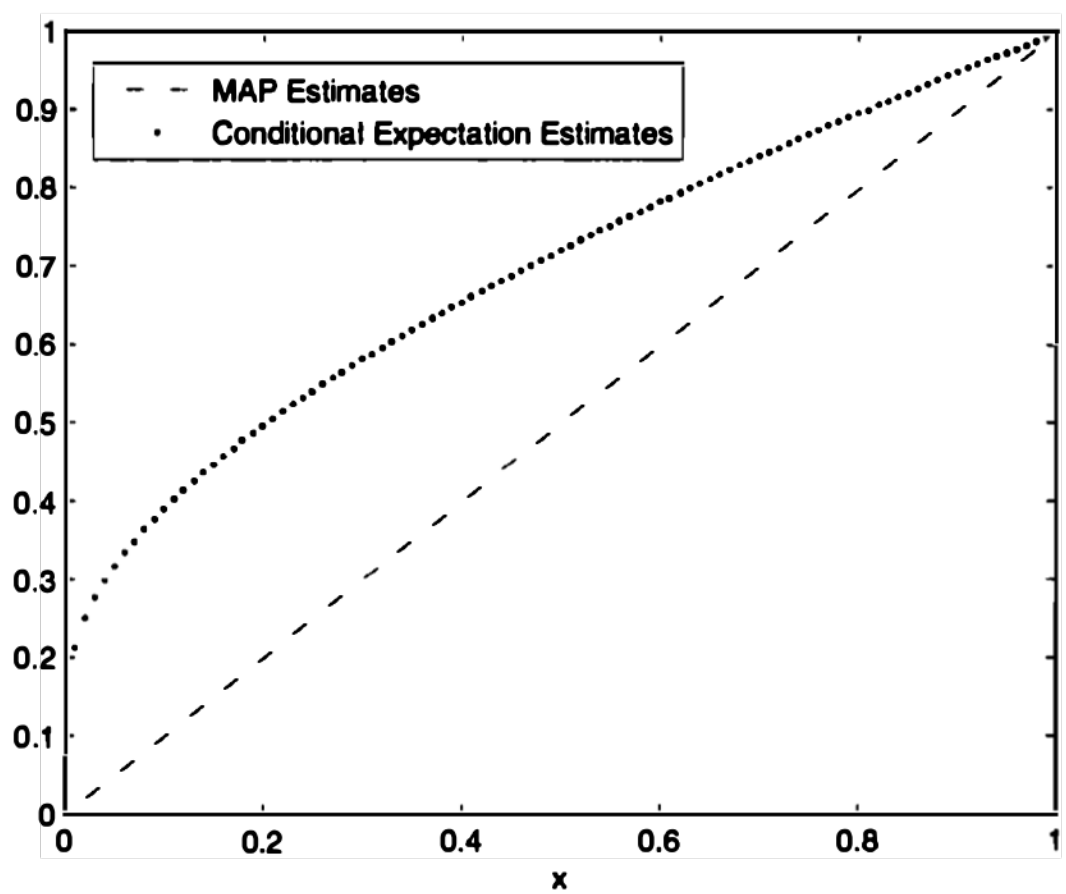
$$f_{\Theta|X}(\theta \,|\, x) = \begin{cases} \dfrac{1}{\theta \cdot |\log x|}, & \text{if } x \le \theta \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

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

Example I

Example : Consider example prev. where Juliet is late on the first date by a random amount X. The distribution of X is uniform over the interval $[0, \Theta]$, and Θ is an unknown random variable with a uniform prior PDF f_Θ over the interval $[0, 1]$. In that example, we saw that for $x \in [0, 1]$, the posterior PDF is

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



$$\widehat{\Theta}_{\text{MAP}}(x) = x$$

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