DCP 1206: Probability
Lecture 26 — Parameter Estimation

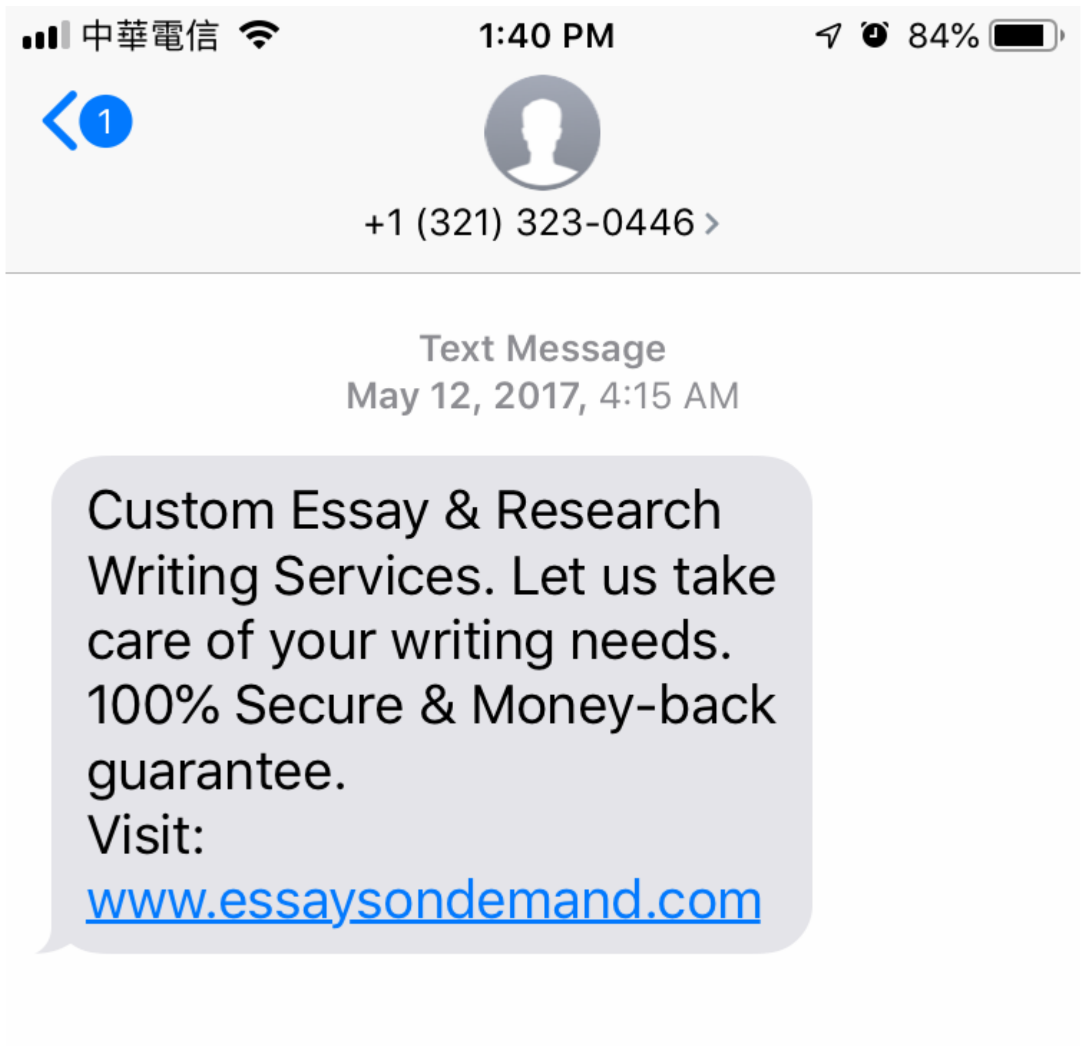
Ping-Chun Hsieh

December 20, 2019

Announcements

- HW6 is now on E3
  - Part I: Due on 12/27 (Friday)
  - Part II: Due on 1/3 (Friday)

Classic Example of MAP: Spam Filter



- Goal: Determine if a message is spam given the text
- Technique: MAP or also called "naive Bayes classification"

This Lecture

1. Maximum Likelihood Estimation (MLE)

2. Maximum A Posteriori Estimation (MAP)

- Reading material: N/A

Review: Find a Good Estimator Under Small n?

- 2 possible outcomes: Yes / No-Laughing
- $p = P(\text{outcome is “Yes”})$ is unknown
- Each toss is independent from other tosses

- Question: Suppose we observe “Yes, No-L, Yes, Yes, No-L”
  - How to estimate p in a principled way?
  - Is “sample mean” a good estimator?

Maximum Likelihood Estimation

Review: Maximum Likelihood for Bernoulli

- Example: Let $X_1, \cdots, X_n$ be a sequence of i.i.d. Bernoulli random variables with unknown mean
  - Suppose we observe $X_1 = x_1, \cdots, X_n = x_n$ $(x_i \in \{0,1\})$
  - Question: Under a guess of mean $= \theta$, $P(X_1 = x_1, \cdots, X_n = x_n; \theta) =$?
  - Question: Under what θ is $P(X_1 = x_1, \cdots, X_n = x_n; \theta)$ maximized?

$$P(X_1=x_1, X_2=x_2, \cdots, X_n=x_n; \theta) \leftarrow \text{Likelihood}$$

$$\ell(\theta) = \theta^{(x_1+x_2+\cdots+x_n)} \cdot (1-\theta)^{n-(x_1+x_2+\cdots+x_n)}$$

$$\text{log-likelihood} = \underbrace{(x_1+x_2+\cdots+x_n)}_{S} \cdot \ln\theta + (n-\underbrace{(x_1+x_2+\cdots+x_n)}_{S}) \cdot \ln(1-\theta)$$

$$\frac{d\ell(\theta)}{d\theta} = S \cdot \frac{1}{\theta} + (S-n) \cdot \frac{1}{1-\theta} = \frac{S(1-\theta)+(S-n)\theta}{\theta(1-\theta)} = \frac{S-n\theta}{\theta(1-\theta)} = 0$$

$$\theta^* = \frac{S}{n}$$

Sample mean

1. Maximum Likelihood Estimation

Maximum Likelihood Estimation (Formally)

- **Maximum Likelihood Estimation (MLE)**: Given observed data D, choose θ that maximizes the probability of observed data:

$$\theta_{\text{MLE}} = \arg\max_{\theta \in \Theta} P(D;\theta) = \arg\max_{\theta \in \Theta} \log P(D;\theta)$$

- **Question**: What if $D = \{X_i\}_{i=1}^N$ and the data samples are independent?

$$\theta_{\text{MLE}} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{N} \underline{P(X_i;\theta)} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{N} \log P(X_i;\theta)$$

- **Question**: What if $D = \{X_i\}_{i=1}^N$ and the data samples are i.i.d. Bernoulli?

$\theta_{\text{MLE}} =$ Sample mean.

What about continuous random variables?

Use density for MLE!

Example: MLE for Normal RVs

$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- Example: Let $X_1, \cdots, X_n$ be a sequence of i.i.d. normal random variables with unknown mean and variance
  - Suppose we observe $X_1 = x_1, \cdots, X_n = x_n$ ($x_i \in \mathbb{R}$)
  - Question: Under a guess of mean = θ (μ), $p(X_1 = x_1, \cdots, X_n = x_n; \theta)$ (μ, σ^2) = ?

Variance $= \sigma^2$

$$p(X_1=x_1,\cdots,X_n=x_n;\mu,\sigma^2) = \prod_{i=1}^{N} p(X_i=x_i;\mu,\sigma^2)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\overbrace{\text{log-likelihood}}^{\ell(\mu,\sigma^2)} = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

$$\frac{\partial \ell(\mu,\sigma^2)}{\partial \mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{N} 2(x_i-\mu)\cdot(-1) = 0$$

$$\frac{\partial \ell(\mu,\sigma^2)}{\partial \sigma} = \underbrace{-\frac{N}{2}\cdot\frac{1}{2\pi\sigma^2}\cdot 4\pi\sigma}_{-\frac{N}{\sigma}} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2 = 0$$

$$\mu_{MLE} = \frac{1}{N}\sum x_i$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum (x_i - \mu_{MLE})^2$$

Example: MLE for Normal RVs (Cont.)

- Example: Let $X_1, \cdots, X_n$ be a sequence of i.i.d. normal random variables with unknown mean and variance
  - Suppose we observe $X_1 = x_1, \cdots, X_n = x_n$
  - Question: Under what θ is $p(X_1 = x_1, \cdots, X_n = x_n; \theta)$ maximized?

MLE for Normal Random Variables (Formally)

- **MLE for Normal** : Given observed data $D = \{X_i\}_{i=1}^n$ of i.i.d. normal random variables , the MLE estimators $\theta_{\text{MLE}} = (\mu_{\text{MLE}}, \sigma^2_{\text{MLE}})$ are:

$$\mu_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ = sample mean

$$\sigma^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{\text{MLE}})^2$$

Is MLE always the same as the sample mean?

Check HW6 Problem 5!

Why is MLE a Good Estimator?

- Question: What kind of property do we want for MLE?

- **Consistency of MLE**: Let $X_1, \cdots, X_n$ be a sequence of i.i.d. random variables with model parameters $\theta \in \Theta$ (where the model is identifiable and Θ is assumed to be finite). Then, MLE converges to the true θ in probability:

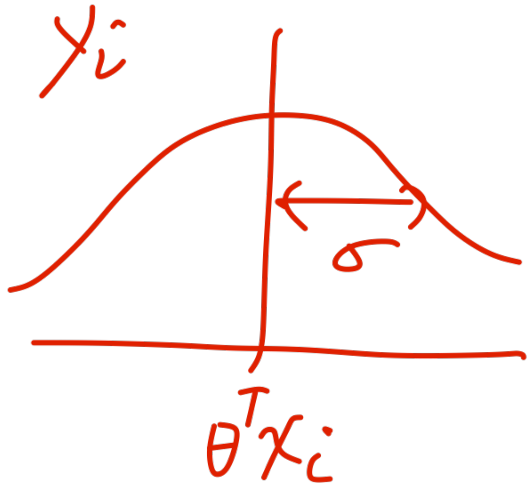
$$\theta_{\mathrm{MLE}} \xrightarrow{\mathrm{p}} \theta$$

- Remark: For the proof, please see the material on E3

Application: MLE for Linear Regression

- MLE is widely used in machine learning problems.
- Example: Linear regression
  - $\{(x_i, y_i)\}_{i=1}^n$ are i.i.d. samples and assume $y_i \sim \mathcal{N}(\theta^T x_i, \sigma^2)$

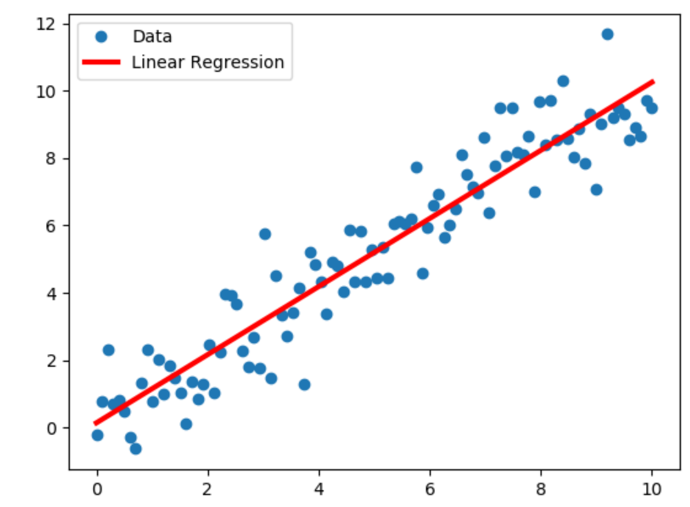
    unknown (θ) (known) (σ^2)
  - Question: Likelihood $p(\{y_1, \cdots, y_n\} \mid \{x_1, \cdots, x_n\}, \theta, \sigma) = ?$



Likelihood

$$= \prod_{i=1}^{n} p(y_i \mid x_i, \theta, \sigma)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}} \quad ✓$$



MLE for Linear Regression (Cont.)

- Example: Linear regression
  - $\{(x_i, y_i)\}_{i=1}^n$ are i.i.d. samples and assume $y_i \sim \mathcal{N}(\theta^T x_i, \sigma^2)$
  - Question: What is the MLE θ_{MLE}?

MLE for Linear Regression (Formally)

- **MLE for Linear Regression** : Given i.i.d. data samples $\{(x_i, y_i)\}_{i=1}^n$ with $y_i \sim \mathcal{N}(\theta^T x_i, \sigma^2)$, the MLE is

$$\theta_{\text{MLE}} = \underline{(X^T X)^{-1} X^T y}$$

where $y = [y_1, y_2, \cdots, y_n]^T$

$$X = [x_1, x_2, \cdots, x_n]^T$$

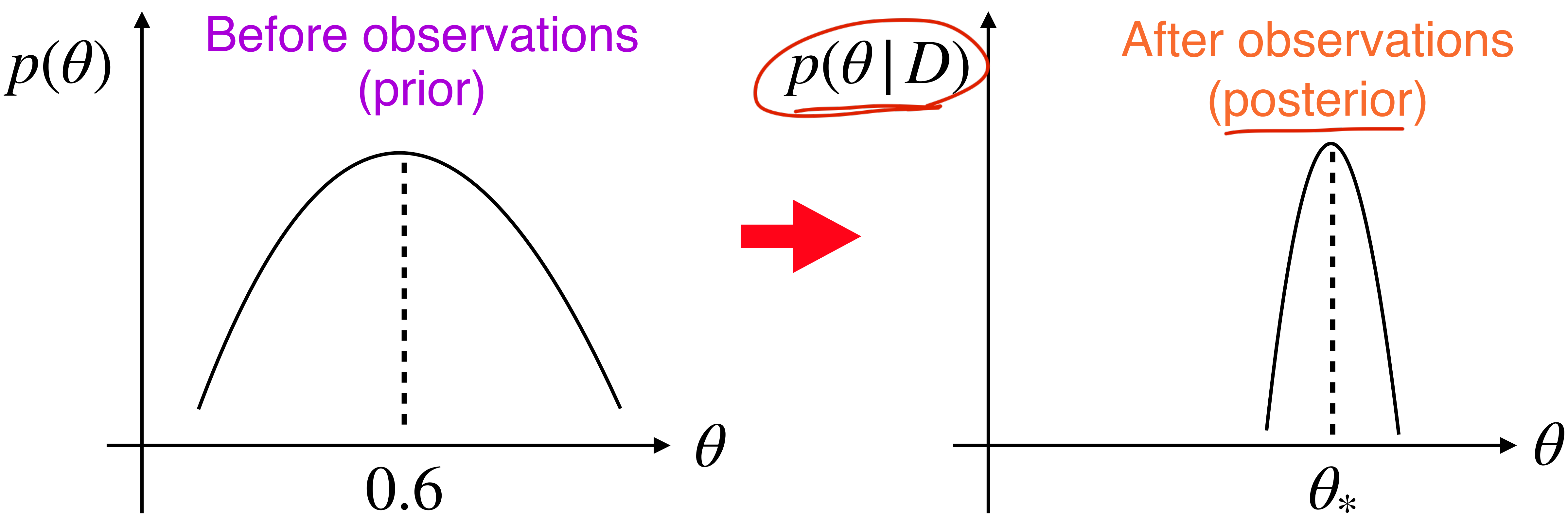
What if we have some prior knowledge about the possible value of θ?

Maximum a Posteriori Estimation

What if We Have Some Prior Knowledge?

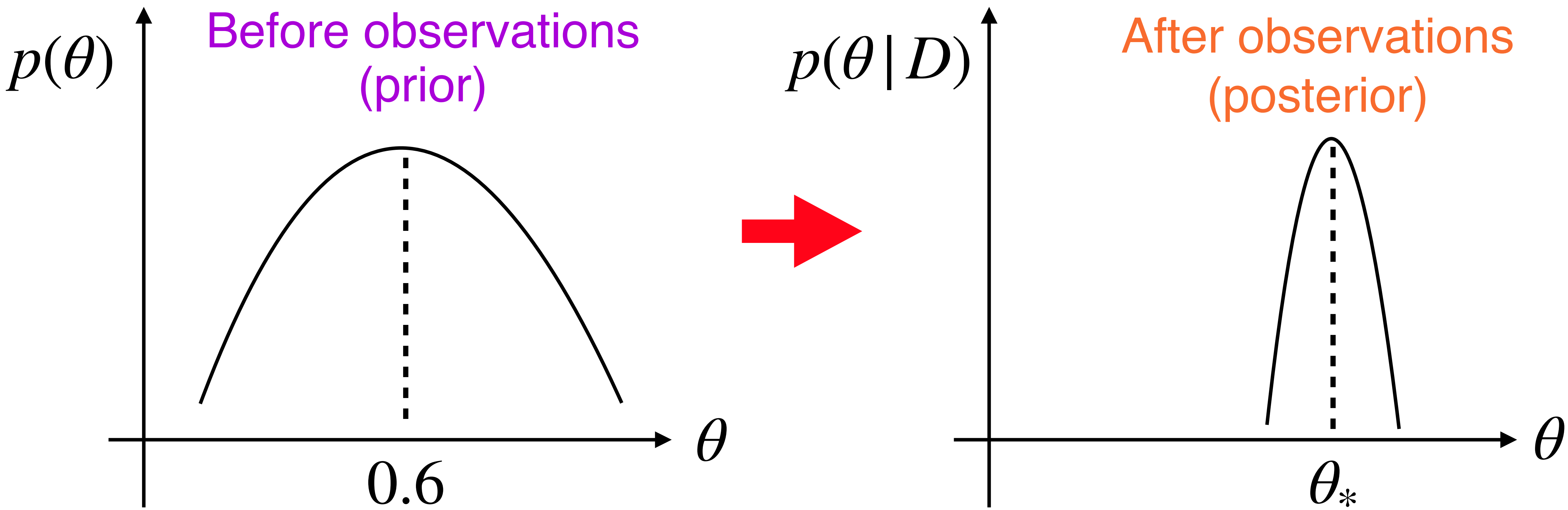
- 2 possible outcomes: Yes / No-Laughing
- $\theta_* = P(\text{outcome is "Yes"})$ is unknown
- Each toss is <u>independent</u> from other tosses

▸ Prior knowledge: Suppose we know θ_* is "close" to 0.6



▸ Idea: Instead of estimating a single θ, obtain a <u>distribution</u> over possible values of θ

How to Obtain Posterior Distribution?



Likelihood

prior

$$p(\theta\,|\,D) = \frac{p(D\,|\,\theta)p(\theta)}{p(D)} \propto \boxed{p(D\,|\,\theta)p(\theta)}$$

posterior

evidence

Recall: Maximum a Posteriori for Bernoulli

- Recall: HW1, Problem 6

Problem 6 (Inference via Bayes' Rule) (6+6+6=18 points)

Suppose we are given a coin with an unknown head probability $\theta \in \{0.3, 0.5, 0.7\}$. In order to infer the value θ, we experiment with the coin and consider Bayesian inference as follows: Define events $A_1 = \{\theta = 0.3\}$, $A_2 = \{\theta = 0.5\}$, $A_3 = \{\theta = 0.7\}$. Since initially we have no further information about θ, we simply consider the prior probability assignment to be $P(A_1) = P(A_2) = P(A_3) = 1/3$.

$P(A_1 \mid \text{observe HTT}) =$

$P(A_2 \mid \text{observe HTT}) =$

$P(A_3 \mid \text{observe HTT}) =$

MLE vs MAP (Formally)

- **Maximum Likelihood Estimation (MLE)**: Given observed data D, choose θ that maximizes the probability of observed data:

$$\theta_{\text{MLE}} = \arg\max_{\theta \in \Theta} P(D; \theta) = \arg\max_{\theta \in \Theta} \log P(D; \theta)$$

- **Maximum a Posteriori Estimation (MAP)**: Given observed data D and prior distribution $P(\theta)$, choose θ that maximizes the posterior probability:

$$\theta_{\text{MAP}} = \arg\max_{\theta \in \Theta} P(\theta \mid D) = \arg\max_{\theta \in \Theta} \log P(D \mid \theta) P(\theta)$$

- Question: When is MAP the same as MLE?

(Uniform prior)

$P(\theta) = \text{const}$

How to Choose a Prior Distribution?

- Question: What's the principle of choosing a prior?

1. Captures expert knowledge (if available)
2. Simple posterior update (most often)

- Example: Widely-used priors for simple posterior updates

non-informative / Jeffrey prior

1. Uniform prior ($p(\theta)$ is constant)
2. Conjugate prior (prior and posterior have the same form)

Review: Beta Distribution

• $\alpha = 1, \beta = 1$: uniform prior

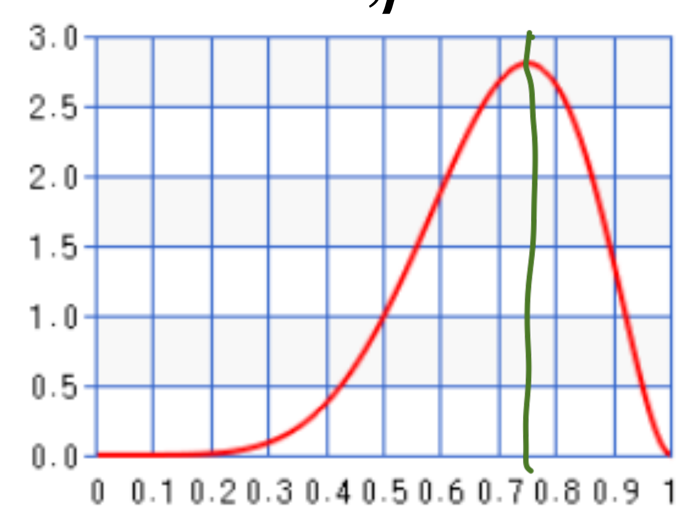
Beta Random Variables (Beta(α, β)): A random variable X is Beta with parameters α, β ($\alpha > 0, \beta > 0$) if its PDF is

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{, if } 0 < x < 1 \\ 0 & \text{, otherwise} \end{cases}$$

where $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$

- Question: Beta distribution is the conjugate prior for Bernoulli experiments. Why?

$\alpha = 7, \beta = 3$



$x = \frac{(7-1)}{(7-1)+(3-1)}$

Example: Beta Prior for Bernoulli Experiments

- Example: After tossing a coin with unknown head probability for 3 times, we observe "HTT". Consider a Beta(α, β) prior. What is the posterior?

$$p(\theta \mid D) \propto p(D \mid \theta) p(\theta)$$

$$p(\theta \mid HTT) \propto \underbrace{\theta^{1} \cdot (1-\theta)^{2}}_{\text{Likelihood}} \underbrace{\frac{1}{B(\alpha,\beta)} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}}_{\text{Beta prior}}$$

$$= \underbrace{\frac{1}{B(\alpha,\beta)} \cdot \theta^{(\alpha-1)+1} \cdot (1-\theta)^{(\beta-1)+2}}_{} \leftarrow \text{Beta posterior}$$

List of Conjugate Priors

Likelihood	Model parameters	Conjugate prior distribution
Bernoulli	p (probability)	Beta
Binomial	p (probability)	Beta
Negative binomial with known failure number, r	p (probability)	Beta
Poisson	λ (rate)	Gamma
Categorical	$\mathbf{p}$ (probability vector), k (number of categories; i.e., size of $\mathbf{p}$)	Dirichlet
Multinomial	$\mathbf{p}$ (probability vector), k (number of categories; i.e., size of $\mathbf{p}$)	Dirichlet
Hypergeometric with known total population size, N	M (number of target members)	Beta-binomial[4]
Geometric	p_0 (probability)	Beta

Likelihood	Model parameters	Conjugate prior distribution
Normal with known variance σ^2	μ (mean)	Normal
Normal with known precision τ	μ (mean)	Normal
Normal with known mean μ	σ^2 (variance)	Inverse gamma
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared
Normal with known mean μ	τ (precision)	Gamma
Normal[note 6]	μ and σ^2 Assuming exchangeability	Normal-inverse gamma

https://en.wikipedia.org/wiki/Conjugate_prior

Next Lecture

- Markov Chain

1-Minute Summary

1. Maximum Likelihood Estimation (MLE)

- MLE: $\theta_{\text{MLE}} = \underset{\theta \in \Theta}{\arg\max}\, P(D; \theta) = \underset{\theta \in \Theta}{\arg\max} \log P(D; \theta)$
- Bernoulli / Normal / Linear Regression

2. Maximum A Posteriori Estimation (MAP)

- MAP: $\theta_{\text{MAP}} = \underset{\theta \in \Theta}{\arg\max}\, P(\theta \mid D) = \underset{\theta \in \Theta}{\arg\max}\, P(D \mid \theta) P(\theta)$
- Conjugate priors