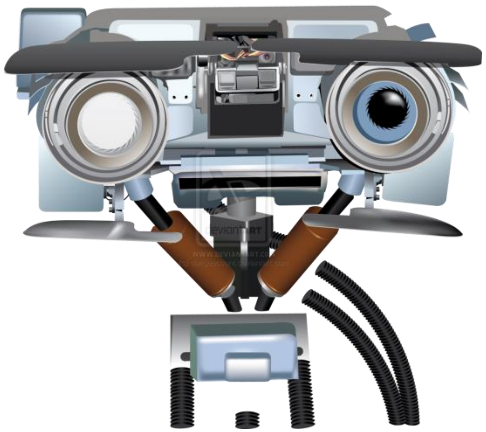
Using Zoom for Lectures

- **Sign in using:**
  – your name
- **Please mute both:**
  – your video cameras for the entire lecture
  – your audio/mics unless asking or answering a question
- **Asking/answering a question, option 1:**
  – click on Participants
  – use the hand icon to raise your hand
  – I will call on you and ask you to unmute yourself
- **Asking/answering a question, option 2:**
  – click on Chat
  – type your question, and I will answer it

Today: Outline

- **Frequentist *vs.* Bayesian**

- **Reminder:** PS4, due Mar 30 (no late submissions)
  PS4 self score, due Apr 3



Recap: Maximum Likelihood

for Linear Regression

So far, we have treated outputs as noiseless

- Defined cost function as “distance to true output”
- An alternate view:
  - data (x,y) are generated by unknown process
  - however, we only observe a noisy version
  - how can we model this uncertainty?

- Alternative cost function?

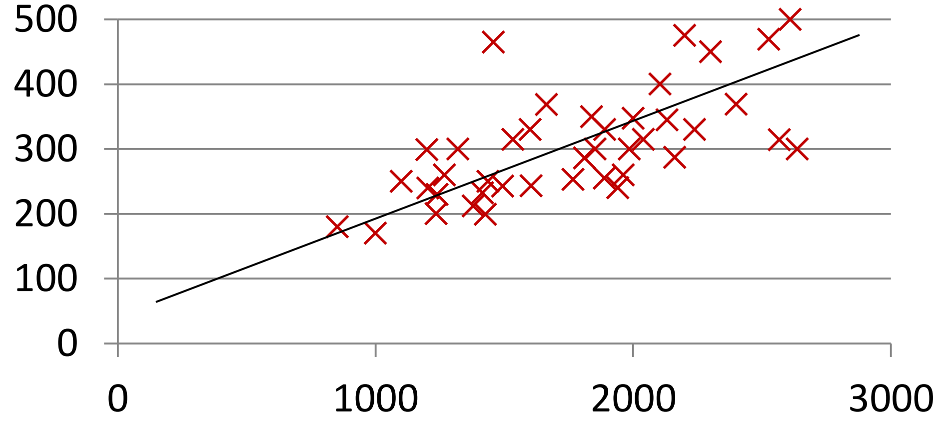
How to model uncertainty in data?

Hypothesis:

$$h_\theta(x) = \theta^T x$$

θ: parameters

$D = \left(x^{(i)}, y^{(i)}\right)$: data

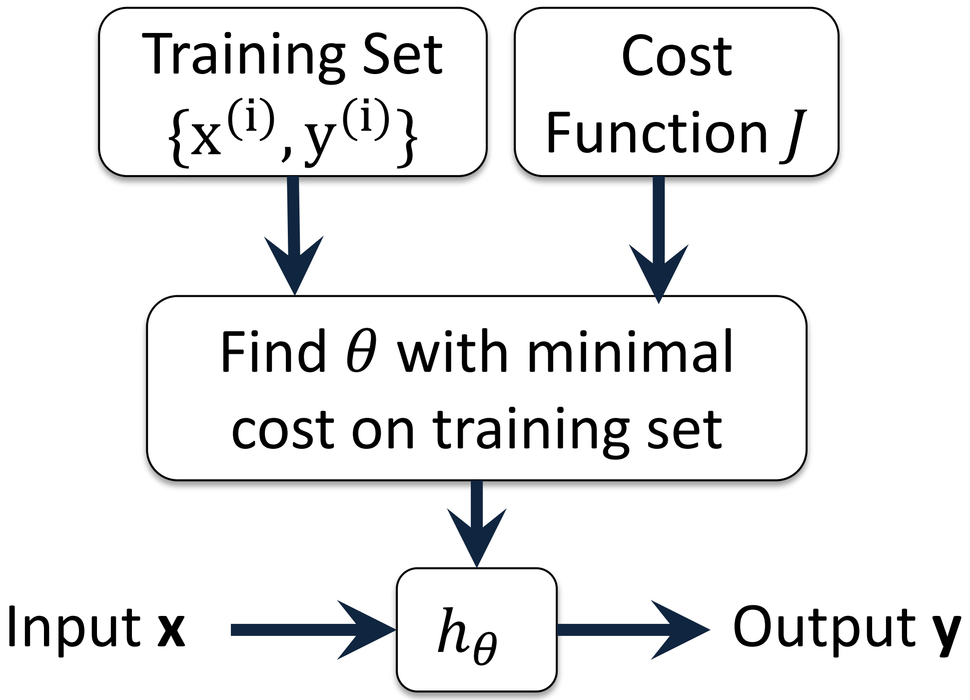


New cost function:

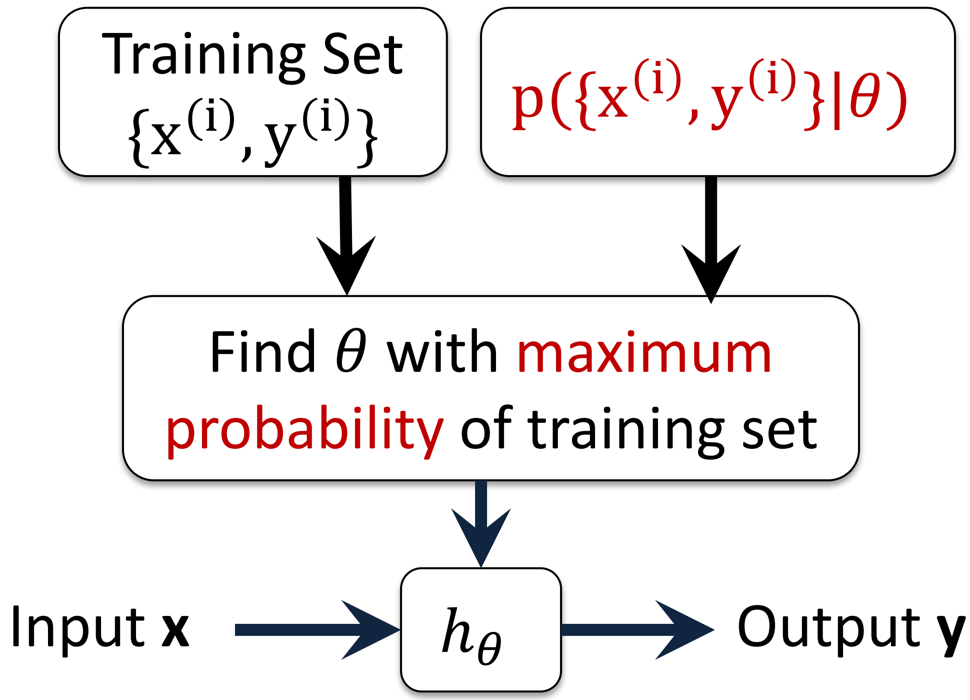
maximize probability of data given model:

$$\mathrm{p}((\mathrm{x}^{(\mathrm{i})}, \mathrm{y}^{(\mathrm{i})})|\theta)$$

Recall: Cost Function



Alternative View:
“Maximum Likelihood”



Maximum Likelihood: Example

- Intuitive example: Estimate a coin toss

  I have seen 3 flips of heads, 2 flips of tails, what is the chance of head (or tail) of my next flip?

- Model:

  Each flip is a Bernoulli random variable X

  X can take only two values: 1 (head), 0 (tail)

  $$p(X = 1) = \theta, \quad p(X = 0) = 1 - \theta$$

- θ is a parameter to be identified from data

Maximum Likelihood: Example

- 5 (independent) trials

$X_1 = 1$ $X_2 = 0$ $X_3 = 1$ $X_4 = 1$ $X_5 = 0$

- Likelihood of all 5 observations:

$$p(X_1, \dots, X_5 | \theta) = \theta^3 (1 - \theta)^2$$

- Intuition

ML chooses θ such that likelihood is maximized

Maximum Likelihood: Example

- 5 (independent) trials

$X_1 = 1$ $X_2 = 0$ $X_3 = 1$ $X_4 = 1$ $X_5 = 0$

- Likelihood of all 5 observations:

$$p(X_1, \ldots, X_5 | \theta) = \theta^3 (1 - \theta)^2$$

- Solution

$$\theta_{ML} = \frac{3}{(3 + 2)}$$

Frequentist Approach

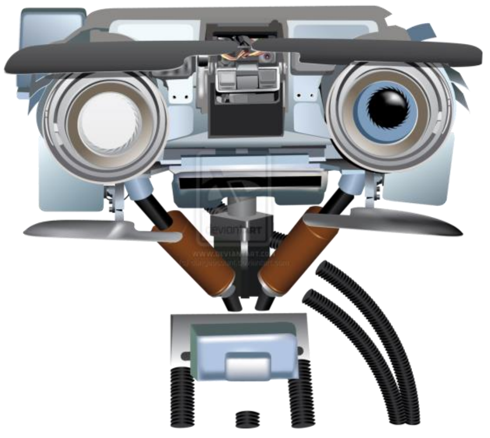
i.e. fraction of heads in total number of trials

Frequentist *vs.* Bayesian

- What is probability?
  - Related to the frequencies of related events
    Frequentists
  - Related to our own certainty/uncertainty of events
    Bayesians

Frequentist *vs.* Bayesian

- Thus we analyze:
  - Variation of data in terms of fixed model parameters
    Frequentists
  - Variation of beliefs about parameters in terms of fixed observed data
    Bayesians



Bayesian Methods

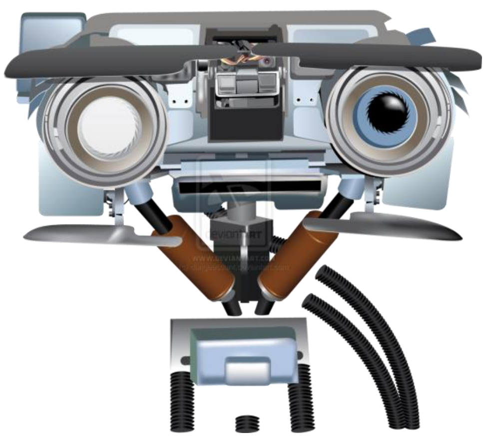
CS542 Machine Learning

Bayesian Methods

- Before, we derived cost functions from maximum likelihood, then added regularization terms to these cost functions

- Can we derive regularization directly from probabilistic principles?

- Yes! Use Bayesian methods

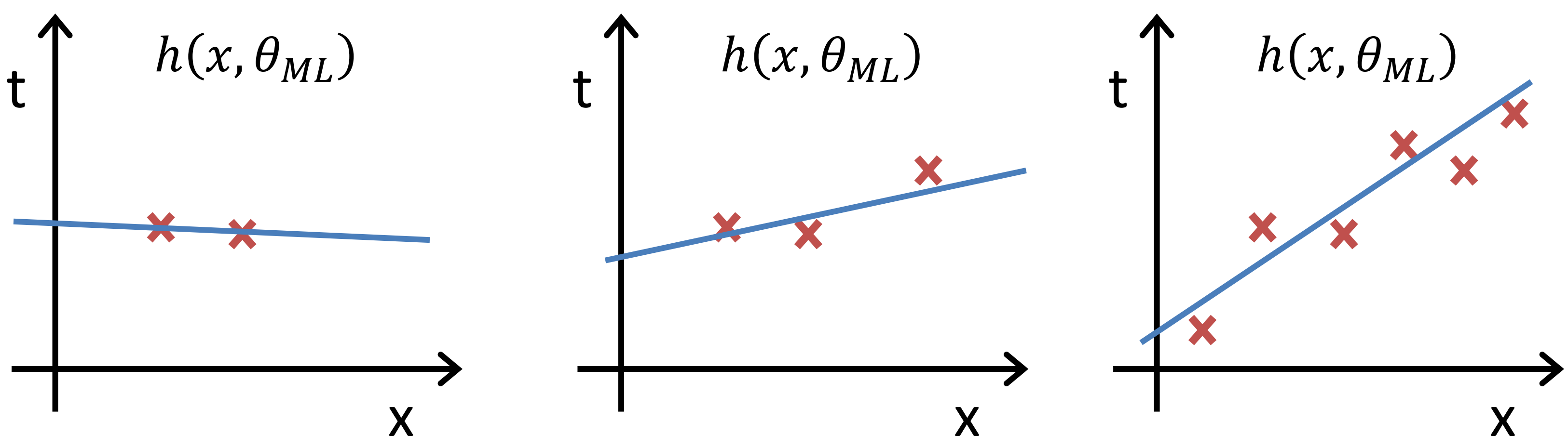


Bayesian Methods

Motivation

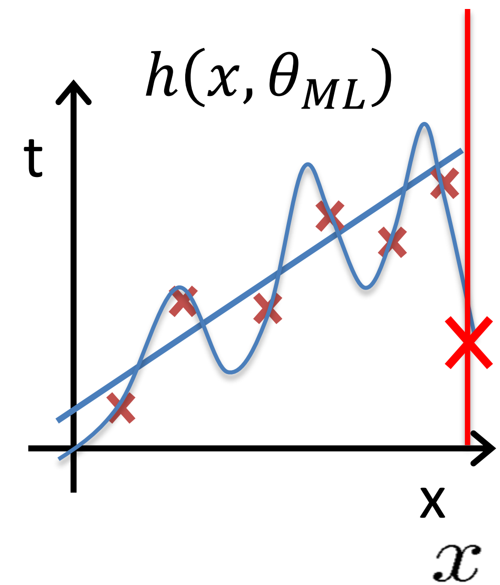
Problem with Maximum Likelihood: Bias

- ML estimates are biased
- Especially a problem for small number of samples, or high input dimensionality
- Suppose we sample 2,3,6 points from the same dataset, use ML to fit regression parameters



Problem with Maximum Likelihood: Overfitting

- ML estimates cannot be used to choose complexity of model
  - E.g. suppose we want to estimate the number of basis functions
  - Choose K=1?
  - Or K=15?



- ML will always choose K that best fits training data (in this case, K=15)

- Solution: use a Bayesian method--define a prior distribution over the parameters (results in regularization)

Bayesian *vs.* Frequentist

Frequentist: maximize data likelihood

$$p(D|model) = p(D|\theta)$$

Bayesian: treat θ as random variable, maximize posterior

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$ ***Baye's Rule***

$p(D|\theta)$ is the data likelihood, $p(\theta)$ is the prior over the model parameters

Bayesian Method

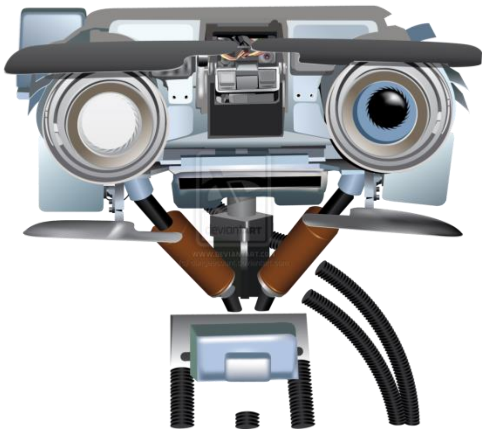
Treat θ as random variable, maximize posterior

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

Likelihood $p(D|\theta)$ is the same as before, as in Maximum Likelihood

Prior $p(\theta)$ is a new distribution we model; specifies which parameters are more likely *a priori*, before seeing any data

$p(D)$ does not depend on θ, constant when choosing θ with the highest posterior probability



Prior over Model Parameters

Intuition

Will he score?

Score! Score! Miss Score!

Your estimate of $\theta = p(score)$?

Will he score?

Score! Score! Miss Score!

- Prior information: player= LeBron James
- Your estimate of $\theta = p(score)$?
- Prior $p(\theta)$ reflects prior knowledge, *e.g.*, $\theta \approx 1$

Prior Distribution

Prior distributions $p(\theta)$ are probability distributions of model parameters based on some a priori knowledge about the parameters.

Prior distributions are independent of the observed data.

Coin Toss Example

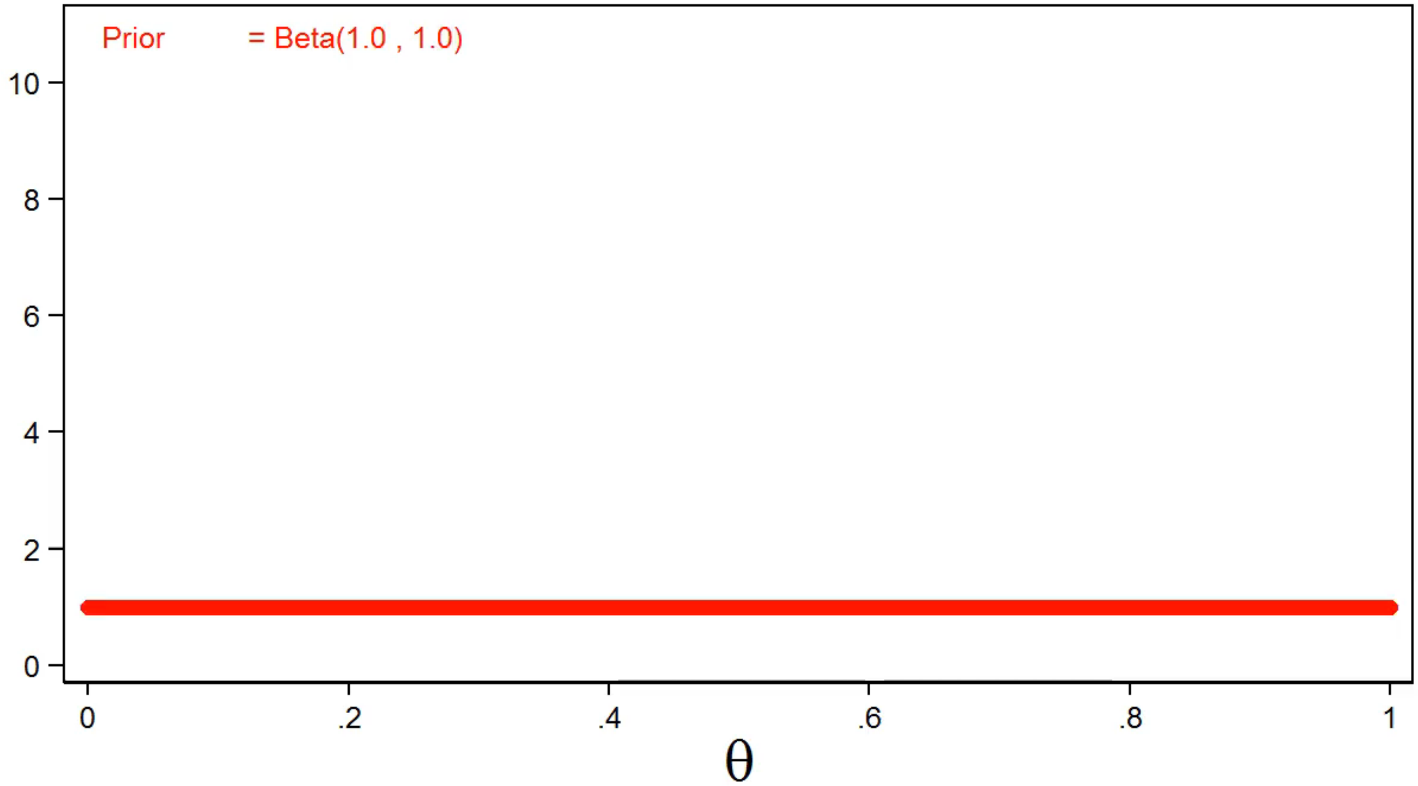
What is the probability of heads (θ)?

Beta Prior for θ

$$P(\theta) = Beta(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}$$

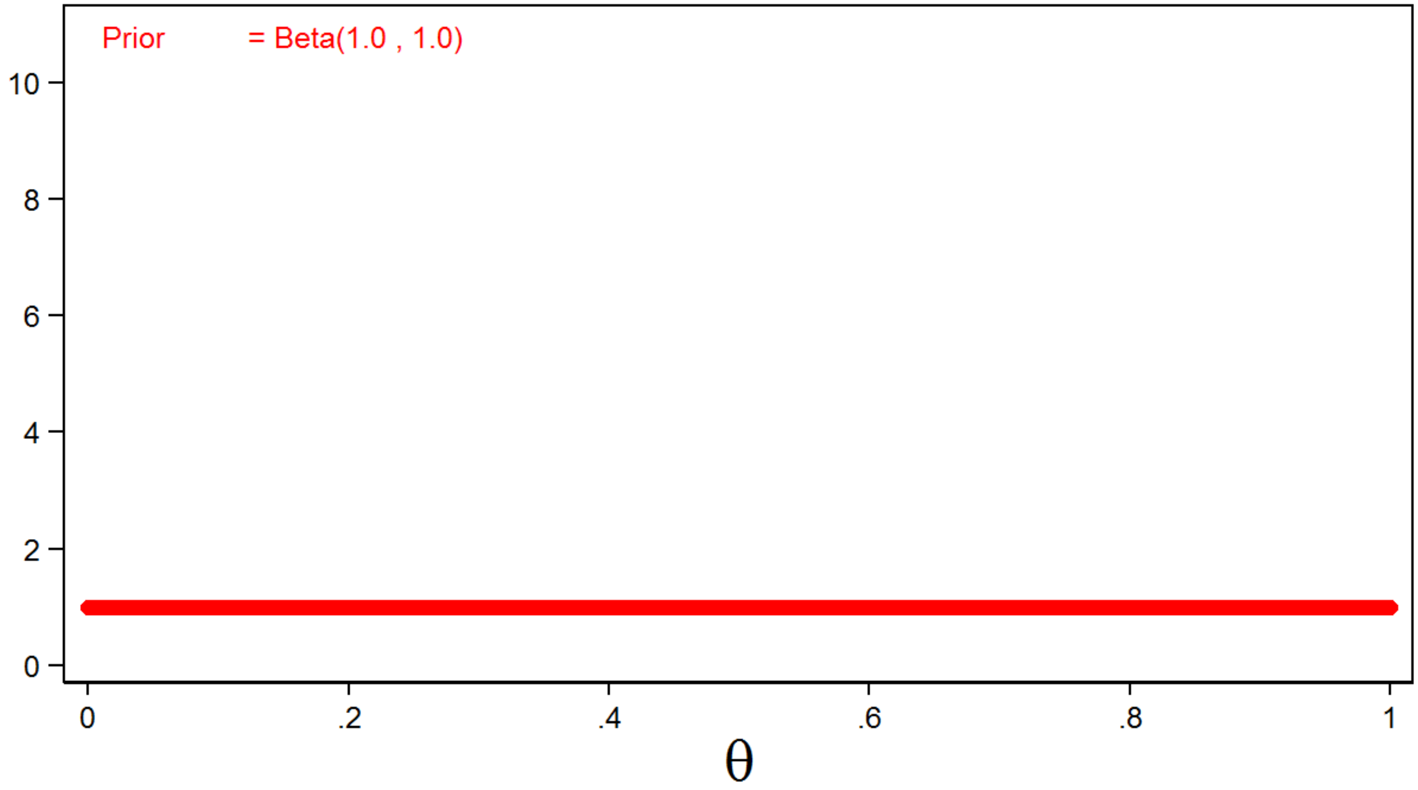
Beta Prior for θ

$$P(\theta) = Beta(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}$$



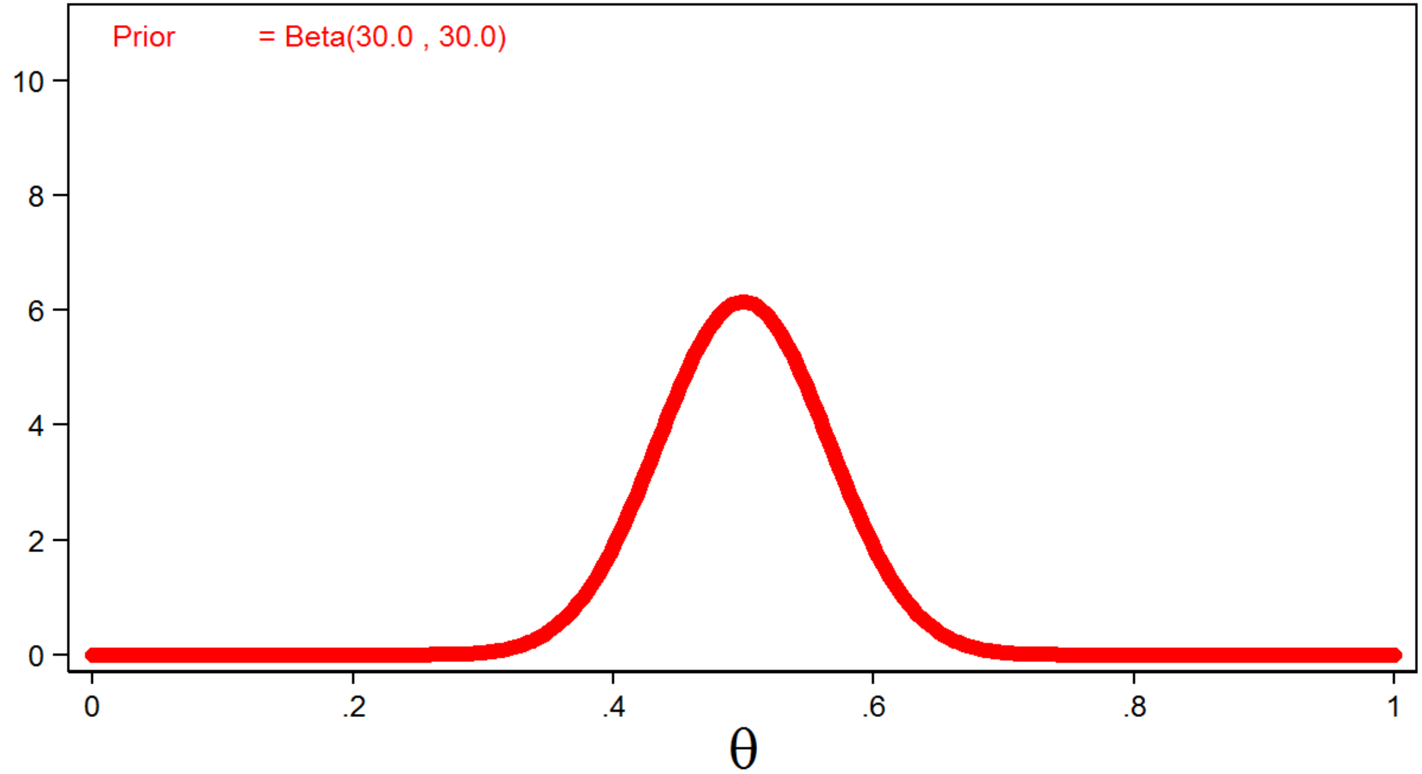
Uninformative Prior

$$P(\theta) = Beta(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}$$



Informative Prior

$$P(\theta) = Beta(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}$$



Coin Toss Experiment

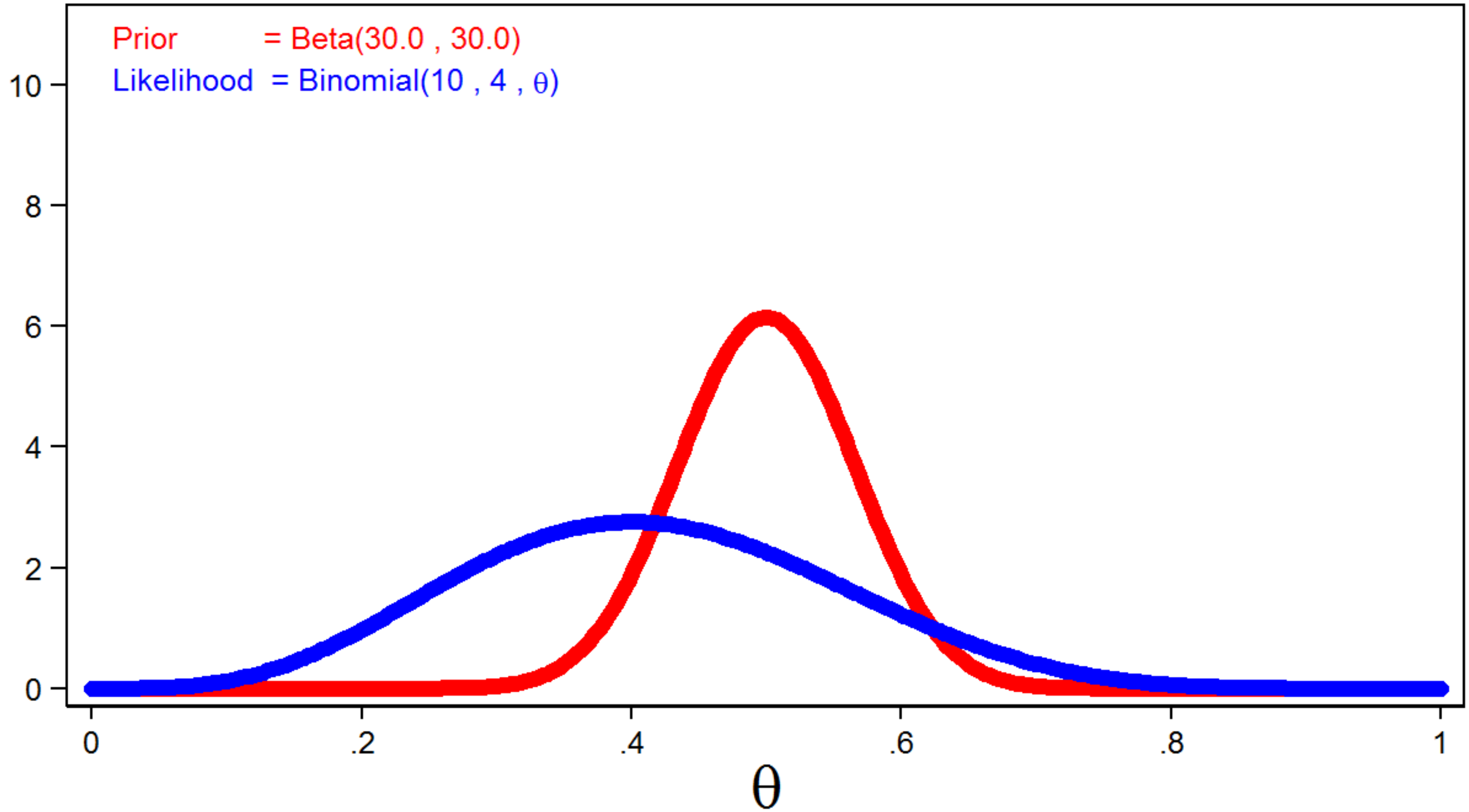
- $n = 10$ coin tosses
- $y = 4$ number of heads

Likelihood Function for the Data

$$P(y|\theta) = Binomial(n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{(n-y)}$$

Prior and Likelihood

$$P(y|\theta) = Binomial(n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{(n-y)}$$



Posterior Distribution

$$Posterior = Prior \times Likelihood$$

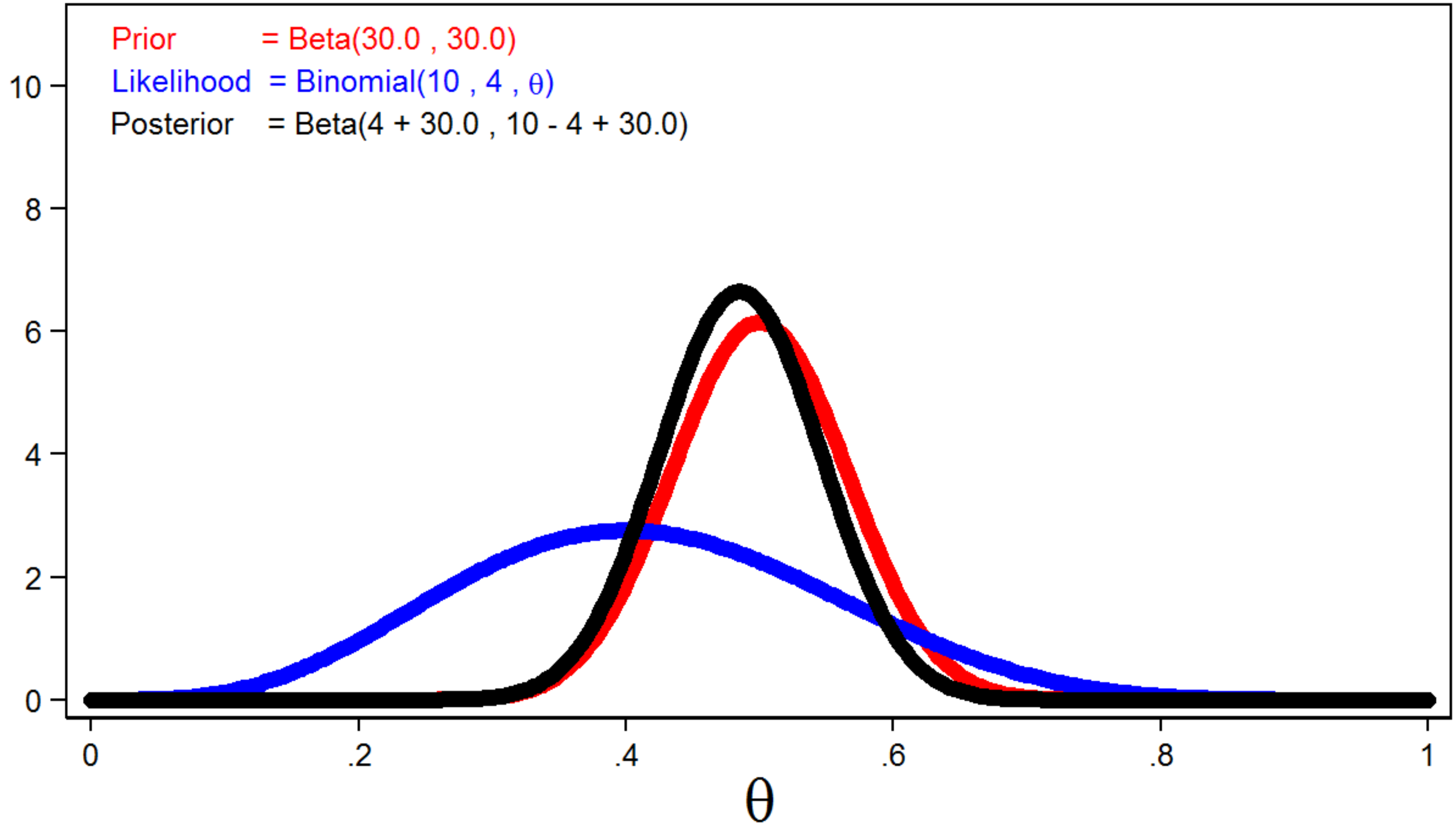
$$P(\theta|y) = P(\theta)P(y|\theta)$$

$$P(\theta|y) = Beta(\alpha, \beta) \times Binomial(n, \theta)$$

$$= Beta(y + \alpha, n - y + \beta)$$

This is why we chose the Beta distribution as our prior, posterior is also a Beta distribution: conjugate prior.

Posterior Distribution



Priors

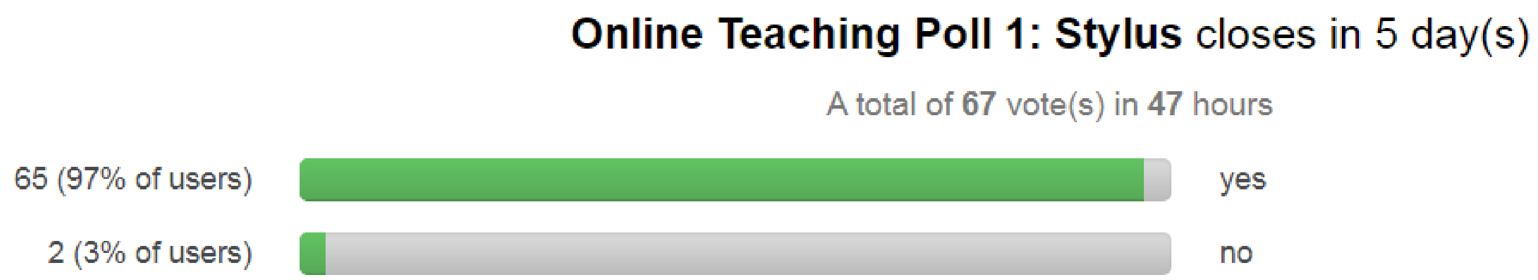
- Example: Brightness of a star

  - Informative Prior:
    Based on all other stars in the sky

  - Non-informative Prior
    Make all brightness values equally probable

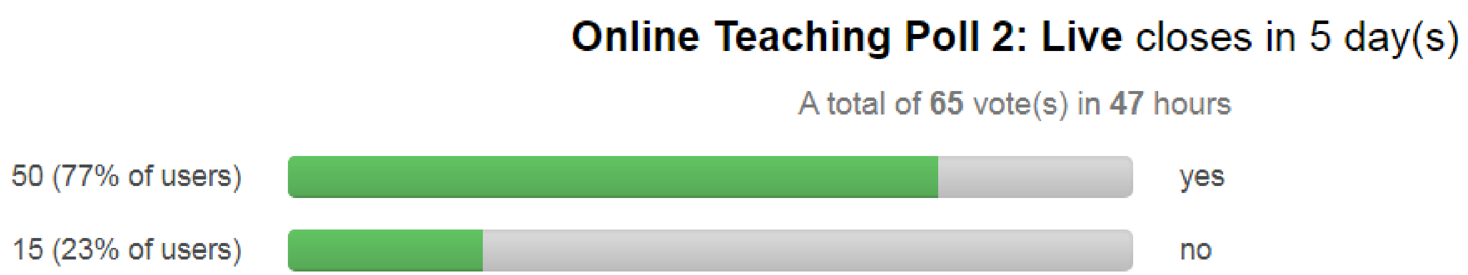
Poll 1: Stylus

- Will continue to use the stylus



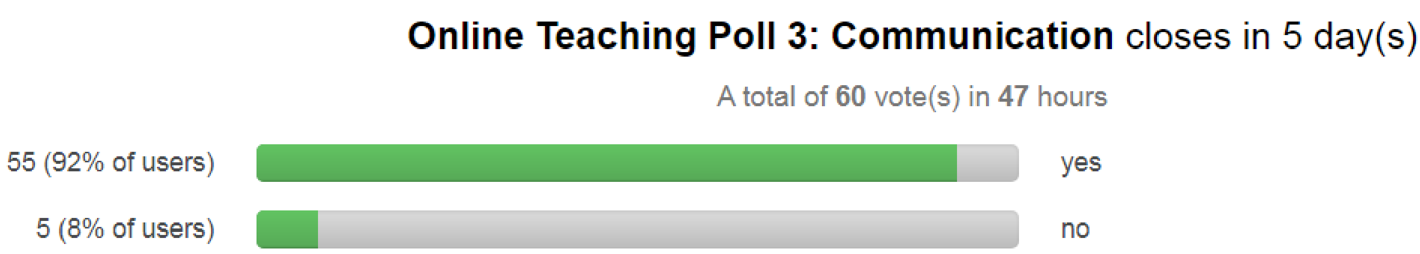
Poll 2: Live

- Will continue to give live lectures and post recorded videos



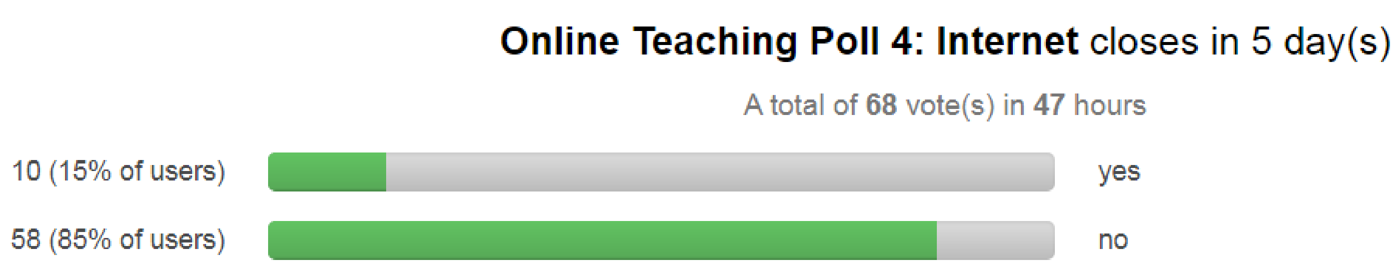
Poll 3: Communication

- We have increased our daily frequency of replied to Piazza posts
- Link for online zoom office hours: https://bostonu.zoom.us/j/741468463
- Same usual schedule
- Waiting room style



Poll 4: Internet

- Will not make the midterm have any multimedia content.
- If you voted that you have internet problems: Please email me immediately to discuss their nature and how BU can help.



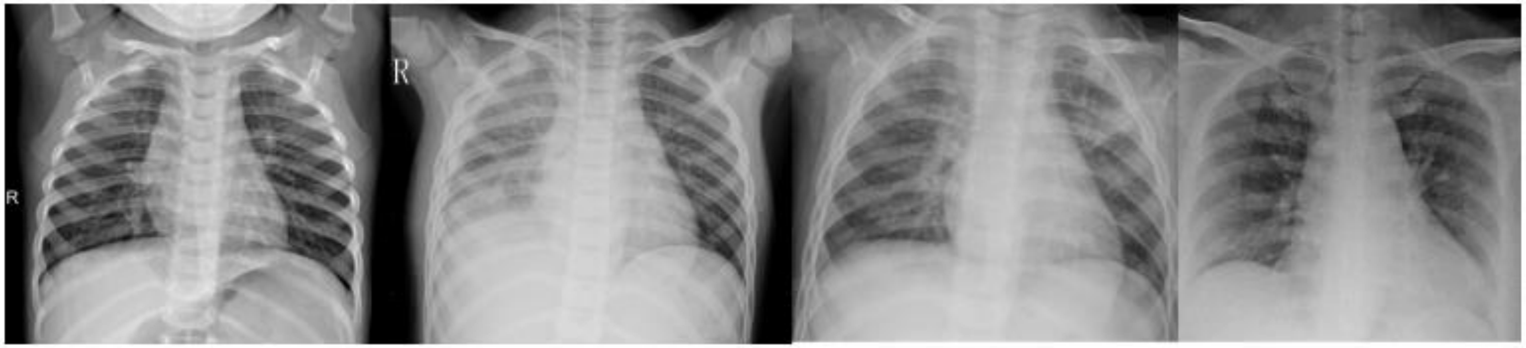
Please do respond to polls, it is instrumental for the course in its current online setting

Midterm Exam

- **Date: Wed Apr 8**
- **Administering the exam:**
  - During lecture time
  - Open: Video camera + Microphone + Share screen
  - Open exam pdf
  - Take photos of your solutions on paper
  - Submit a pdf of the photos on a google form, just like you submit assignments
  - Confirm we received your submission before you leave (through private chat)
- **What do you need?**
  - Internet + Pen/pencil + Empty sheets of paper (~10)
  - New question, new page
  - Cell phone to take photos of your solutions at the end for submission

Class Challenge

- Classification of X-rays for COVID-19



PA Chest X-rays of admitted patients. From Left to Right: Healthy, Bacterial, Viral, COVID19.