

# Maximum A Posteriori (MAP)

CS 109  
Lecture 22  
May 16th, 2016

Previously in CS109...

# Game of Estimators



Non spoiler: this didn't happen

# Side Plot



argmax

argmax of log

Mother of  
optimizations?

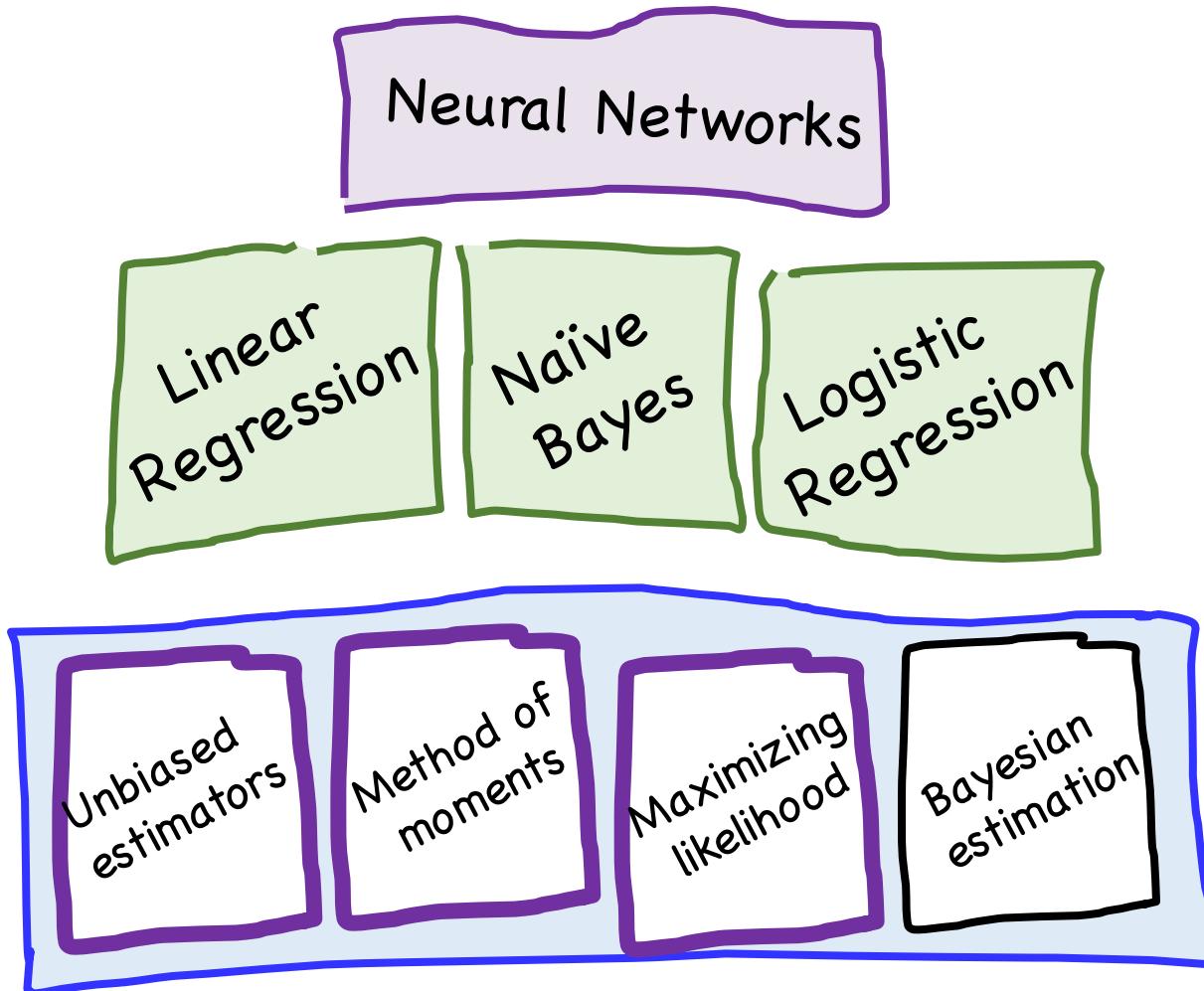
# Reviving an Old Story Line



The Multinomial Distribution  $\text{Mult}(p_1, \dots, p_k)$

$$p(x_1, \dots, x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

# Machine Learning So Far



# Maximum Likelihood of Data

- Consider  $n$  I.I.D. random variables  $X_1, X_2, \dots, X_n$ 
  - $X_i$  is a sample from density function  $f(X_i | \theta)$

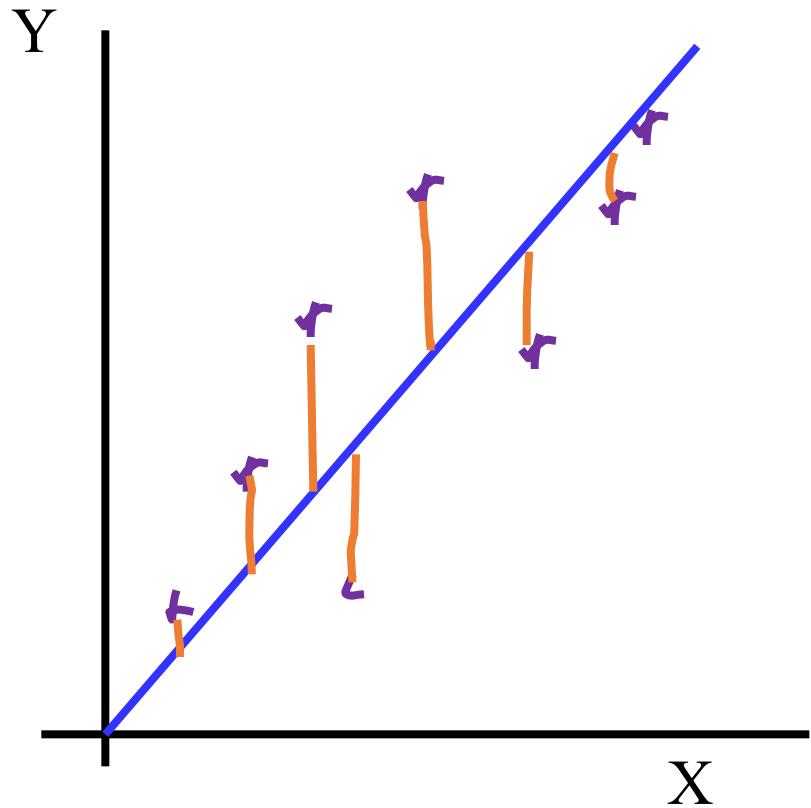
$$L(\theta) = \prod_{i=1}^n f(X_i | \theta)$$

$$LL(\theta) = \log L(\theta) = \log \prod_{i=1}^n f(X_i | \theta) = \sum_{i=1}^n \log f(X_i | \theta)$$

$$\theta_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} LL(\theta)$$

# MLE to Linear Regression

How do you fit this line?



Assume:

$$Y = \theta X + Z$$

$$Z \sim N(0, \sigma^2)$$

Calculate MLE of  $\theta$

$$\hat{\theta} = \operatorname{argmin}_{\theta} \sum_{i=1}^m (Y_i - \theta X_i)^2$$

This is an algorithm called linear regression. Learn more about it later...

# Watch it Online

MLE to Linear Regression

MLE  
linear transform

$$Y = \theta X + Z \quad Z \sim N(0, \sigma^2)$$
$$\text{LL}(\theta) = \log \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(Y_i - \theta X_i)^2}{2\sigma^2}} f(x_i)$$
$$= \sum_{i=1}^n \log \frac{1}{\sigma \sqrt{2\pi}} + \log \left( e^{-\frac{(Y_i - \theta X_i)^2}{2\sigma^2}} \right) + \log f(x_i)$$
$$= \sum_{i=1}^n \log \frac{1}{\sigma} - \frac{(Y_i - \theta X_i)^2}{2\sigma^2} + \log f(x_i)$$

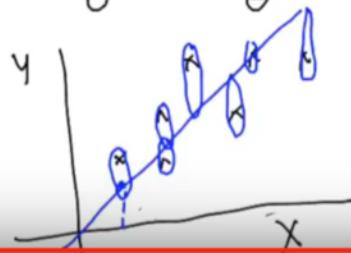
( $x_1, y_1$ )  
( $x_2, y_2$ )  
...  
( $x_n, y_n$ )

Data

$Y | X \theta \sim N(\theta X, \sigma^2)$

$$\hat{\theta} = \operatorname{argmax}_{\theta} \text{LL}(\theta) = \operatorname{argmax}_{\theta} - \sum_{i=1}^n \frac{(Y_i - \theta X_i)^2}{2\sigma^2}$$
$$= \operatorname{argmin}_{\theta} \sum_{i=1}^n (Y_i - \theta X_i)^2$$

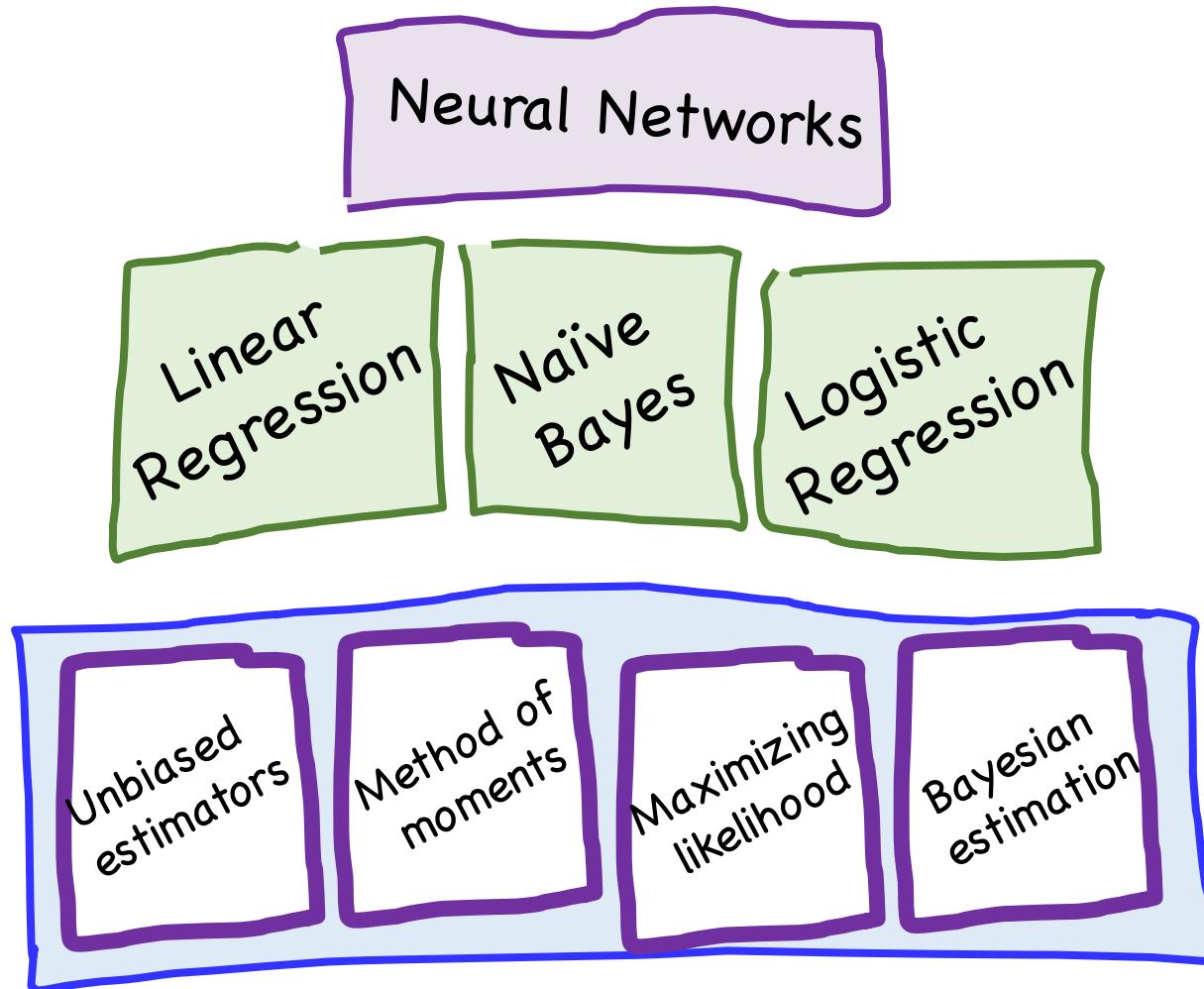
sum of  
squared  
errors



# Episode 22

# The Song of The Last Estimator

# The Song of the Last Estimator



Something rotten  
in the world of MLE

Foreshadowing..

# Need a Volunteer

So good to see  
you again!



# Two Envelopes

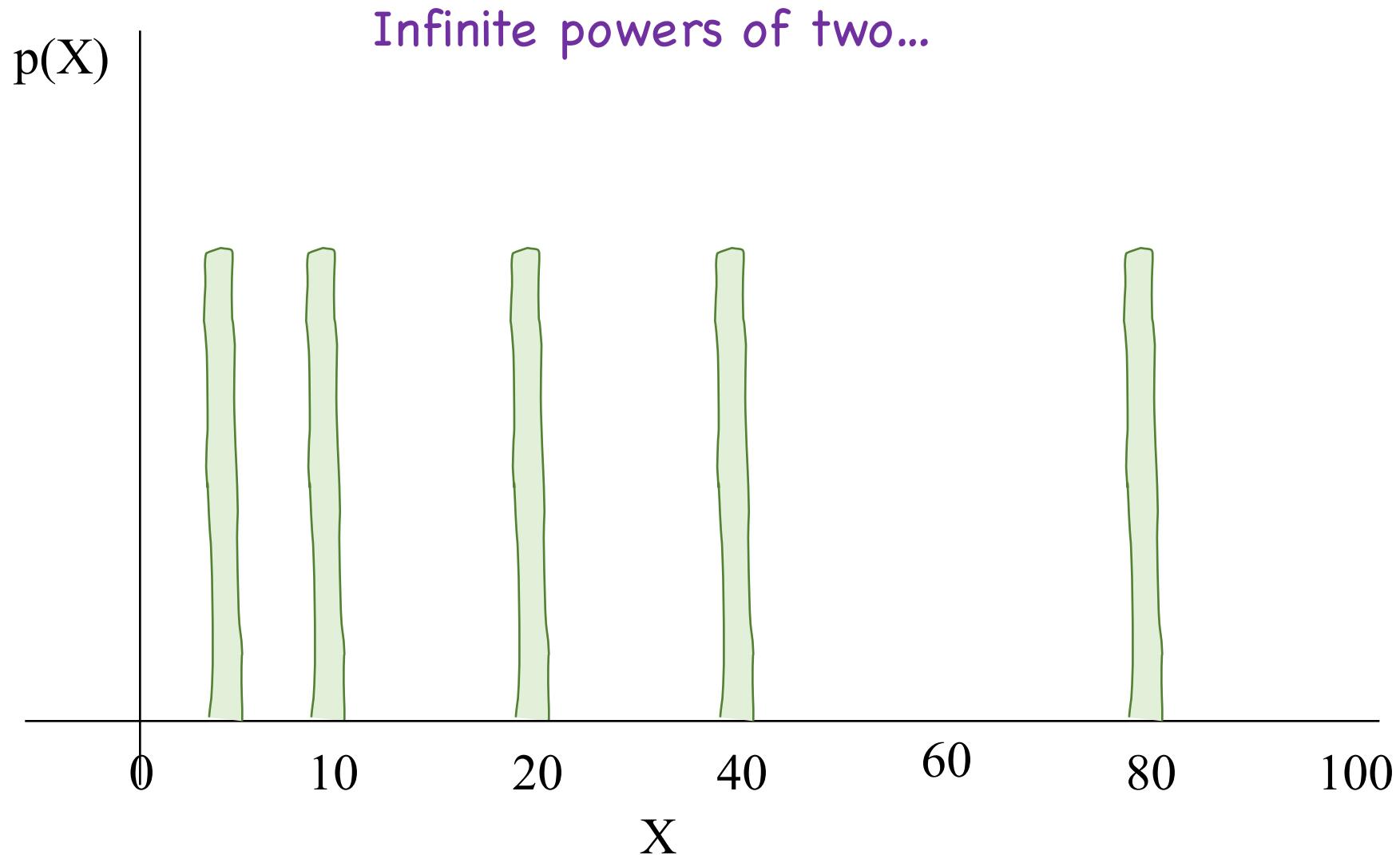
- I have two envelopes, will allow you to have one
  - One contains  $\$X$ , the other contains  $\$2X$
  - Select an envelope
    - Open it!
  - Now, would you like to switch for other envelope?
  - To help you decide, compute  $E[\$ \text{ in other envelope}]$ 
    - Let  $Y = \$ \text{ in envelope you selected}$ 
$$E[\$ \text{ in other envelope}] = \frac{1}{2} \cdot \frac{Y}{2} + \frac{1}{2} \cdot 2Y = \frac{5}{4}Y$$
  - Before opening envelope, think either equally good
  - So, what happened by opening envelope?
    - And does it really make sense to switch?

# Thinking Deeper About Two Envelopes

- The “two envelopes” problem set-up
  - Two envelopes: one contains  $\$X$ , other contains  $\$2X$
  - You select an envelope and open it
    - Let  $Y = \$$  in envelope you selected
    - Let  $Z = \$$  in other envelope
- $E[Z | Y] = \frac{1}{2} \cdot \frac{Y}{2} + \frac{1}{2} \cdot 2Y = \frac{5}{4}Y$ 

---
- $E[Z | Y]$  above assumes all values  $X$  (where  $0 < X < \infty$ ) are equally likely
  - Note: there are infinitely many values of  $X$
  - So, not true probability distribution over  $X$  (doesn’t integrate to 1)

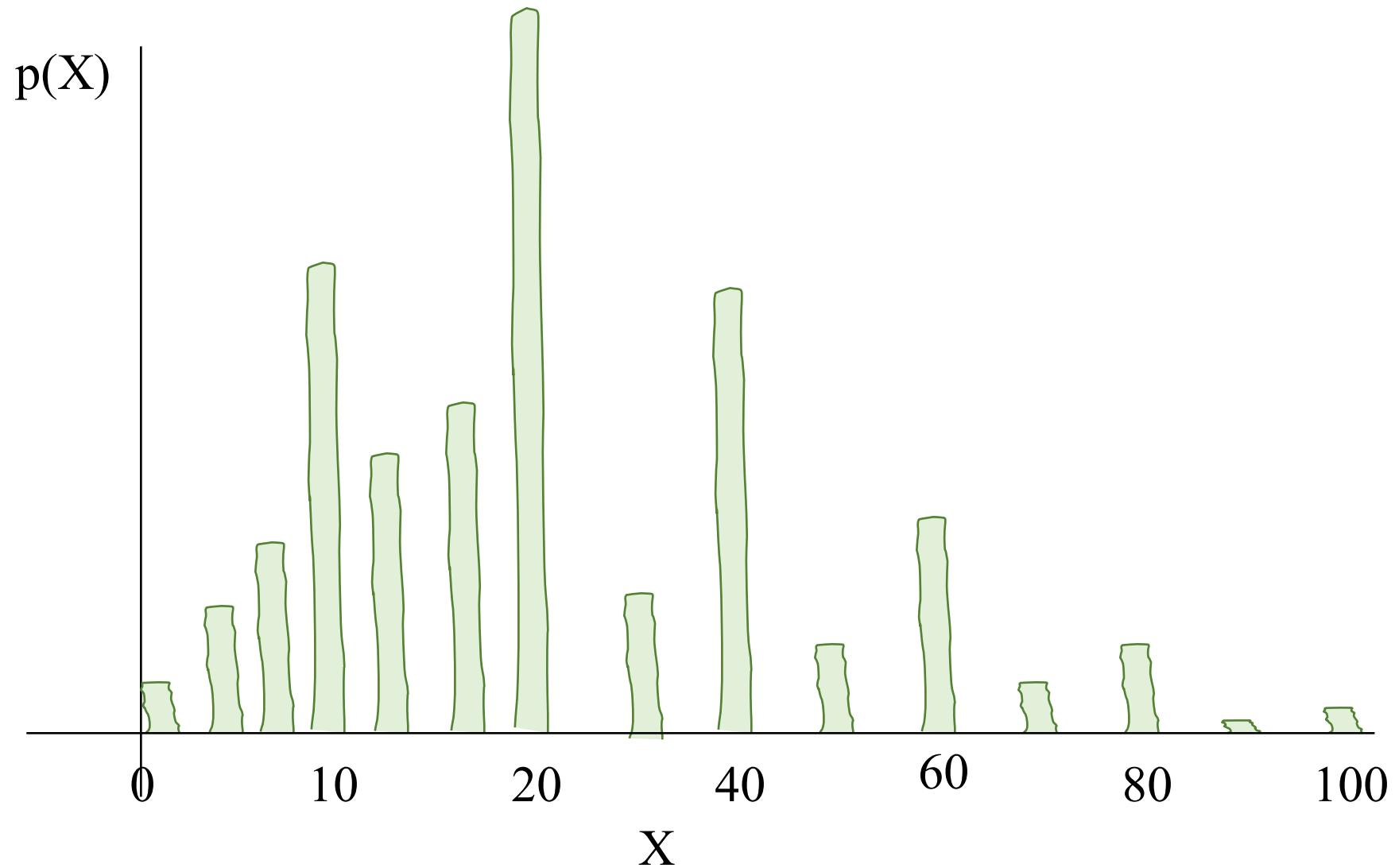
# All Values are Equally Likely?



# Subjectivity of Probability

- Belief about contents of envelopes
  - Since implied distribution over  $X$  is not a true probability distribution, what is our distribution over  $X$ ?
    - *Frequentist*: play game infinitely many times and see how often different values come up.
    - Problem: I only allow you to play the game *once*
  - Bayesian probability
    - Have prior belief of distribution for  $X$  (or anything for that matter)
    - Prior belief is a *subjective* probability
      - By extension, all probabilities are subjective
    - Allows us to answer question when we have no/limited data
      - E.g., probability a coin you've never flipped lands on heads
    - As we get more data, prior belief is “swamped” by data

# Subjectivity of Probability



# The Envelope, Please

- *Bayesian*: have prior distribution over  $X$ ,  $P(X)$ 
  - Let  $Y = \$$  in envelope you selected
  - Let  $Z = \$$  in other envelope
  - Open your envelope to determine  $Y$
  - If  $Y > E[Z | Y]$ , keep your envelope, otherwise switch
    - No inconsistency!
  - Opening envelope provides data to compute  $P(X | Y)$  and thereby compute  $E[Z | Y]$
  - Of course, there's the issue of how you determined your prior distribution over  $X$ ...
    - Bayesian: Doesn't matter how you determined prior, but you *must* have one (whatever it is)
    - Imagine if envelope you opened contained \$20.01

# The Dreaded Half Cent



Envelope Summary:  
Probabilities are beliefs  
Incorporating prior beliefs is useful

Especially for one shot learning...

# One Shot Learning

Single training example:

ବୁ

Test set:

a	ଶ	ଅ	ଶ
କୁ	ଅ	ପ୍ଲ	କୁପ୍ଲ
ମ	କୁ	ଇ	ମ
ମ	ଅ	କୁ	ମୁ

# Priors for Parameter Estimation?

# Flash Back: Bayes Theorem

- Bayes' Theorem ( $\theta$  = model parameters,  $D$  = data):

“Posterior”      “Likelihood”      “Prior”

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

- Likelihood: you've seen this before (in context of MLE)
  - Probability of data given probability model (parameter  $\theta$ )
- Prior: before seeing any data, what is belief about model
  - I.e., what is *distribution* over parameters  $\theta$
- Posterior: after seeing data, what is belief about model
  - After data  $D$  observed, have posterior distribution  $p(\theta | D)$  over parameters  $\theta$  conditioned on data. Use this to predict new data.

# Computing $P(\theta | D)$

- Bayes' Theorem ( $\theta$  = model parameters,  $D$  = data):

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

- We have prior  $P(\theta)$  and can compute  $P(D | \theta)$
- But how do we calculate  $P(D)$ ?
  - Complicated answer:  $P(D) = \int P(D | \theta) P(\theta) d\theta$
  - Easy answer: It does not depend on  $\theta$ , so ignore it
    - Just a constant that forces  $P(\theta | D)$  to integrate to 1

Most important slide of today

# Maximum A Posteriori

- Recall Maximum Likelihood Estimator (MLE) of  $\theta$

$$\theta_{MLE} = \arg \max_{\theta} \prod_{i=1}^n f(X_i | \theta)$$

- Maximum A Posteriori (MAP) estimator of  $\theta$ :

$$\begin{aligned}\theta_{MAP} &= \arg \max_{\theta} f(\theta | X_1, X_2, \dots, X_n) = \arg \max_{\theta} \frac{f(X_1, X_2, \dots, X_n | \theta) g(\theta)}{h(X_1, X_2, \dots, X_n)} \\ &= \arg \max_{\theta} \frac{\left( \prod_{i=1}^n f(X_i | \theta) \right) g(\theta)}{h(X_1, X_2, \dots, X_n)} = \arg \max_{\theta} g(\theta) \prod_{i=1}^n f(X_i | \theta)\end{aligned}$$

where  $g(\theta)$  is prior distribution of  $\theta$ .

- As before, can often be more convenient to use log:

$$\theta_{MAP} = \arg \max_{\theta} \left( \log(g(\theta)) + \sum_{i=1}^n \log(f(X_i | \theta)) \right)$$

- MAP estimate is the mode of the posterior distribution

# Maximum A Posteriori

Estimated  
parameter

$$\theta_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} \left( \log(g(\theta)) + \sum_{i=1}^n \log(f(X_i|\theta)) \right)$$

Chose the value of theta  
that maximizes:

Log prior

Sum of  
log likelihood

Gotta get that intuition

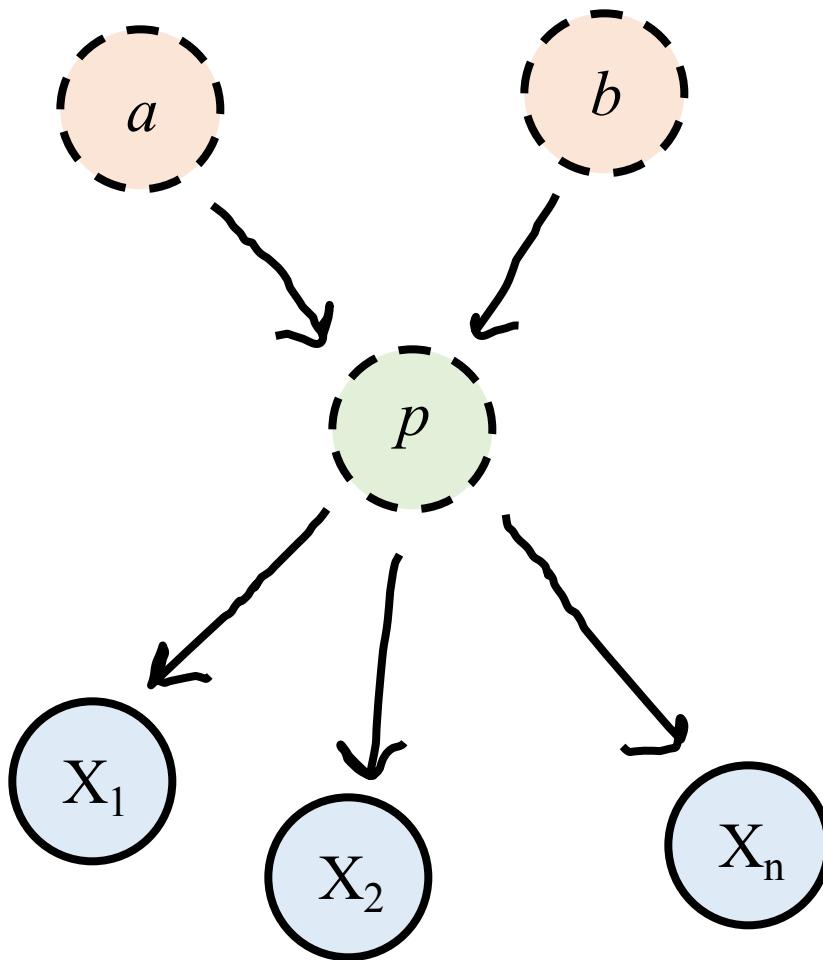
# $P(\theta | D)$ For Beta and Bernoulli

- Prior:  $\theta \sim \text{Beta}(a, b)$ ;  $D = \{n \text{ heads}, m \text{ tails}\}$

$$\begin{aligned} f_{\theta|D}(\theta = p | D) &= \frac{f_{D|\theta}(D | \theta = p) f_{\theta}(\theta = p)}{f_D(D)} \\ &= \frac{\binom{n+m}{n} p^n (1-p)^m \cdot \frac{p^{a-1} (1-p)^{b-1}}{C_1}}{C_2} = \frac{\binom{n+m}{n}}{C_1 C_2} p^n (1-p)^m \cdot p^{a-1} (1-p)^{b-1} \\ &= C_3 p^{n+a-1} (1-p)^{m+b-1} \end{aligned}$$

- Estimate  $p$ 
  - $\theta_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} f(\theta | D)$   
 $= \underset{\theta}{\operatorname{argmax}} (n + a - 1) \log \theta + (m + b - 1) \log(1 - \theta)$
- By definition,  $f(\theta | D)$  is  $\text{Beta}(a + n, b + m)$

# Hyper Parameters



Hyperparameter  
 $a, b$  are fixed

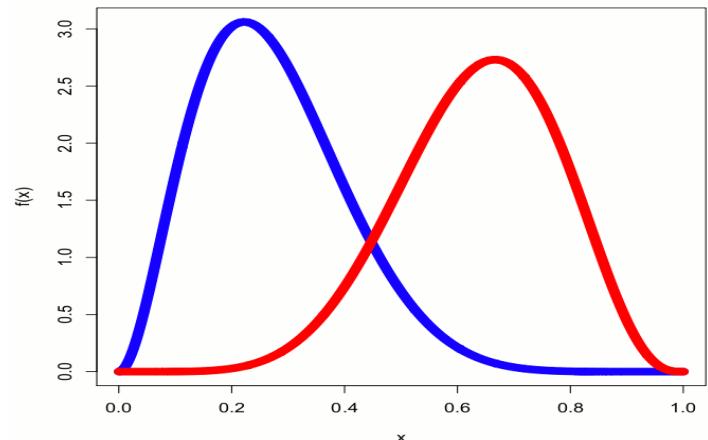
Prior  
 $p \sim \text{Beta}(a, b)$

Data distribution  
 $X_i \sim \text{Bern}(p)$

MAP will estimate the most likely value of  $p$  for this model

# Where'd Ya Get Them $P(\theta)$ ?

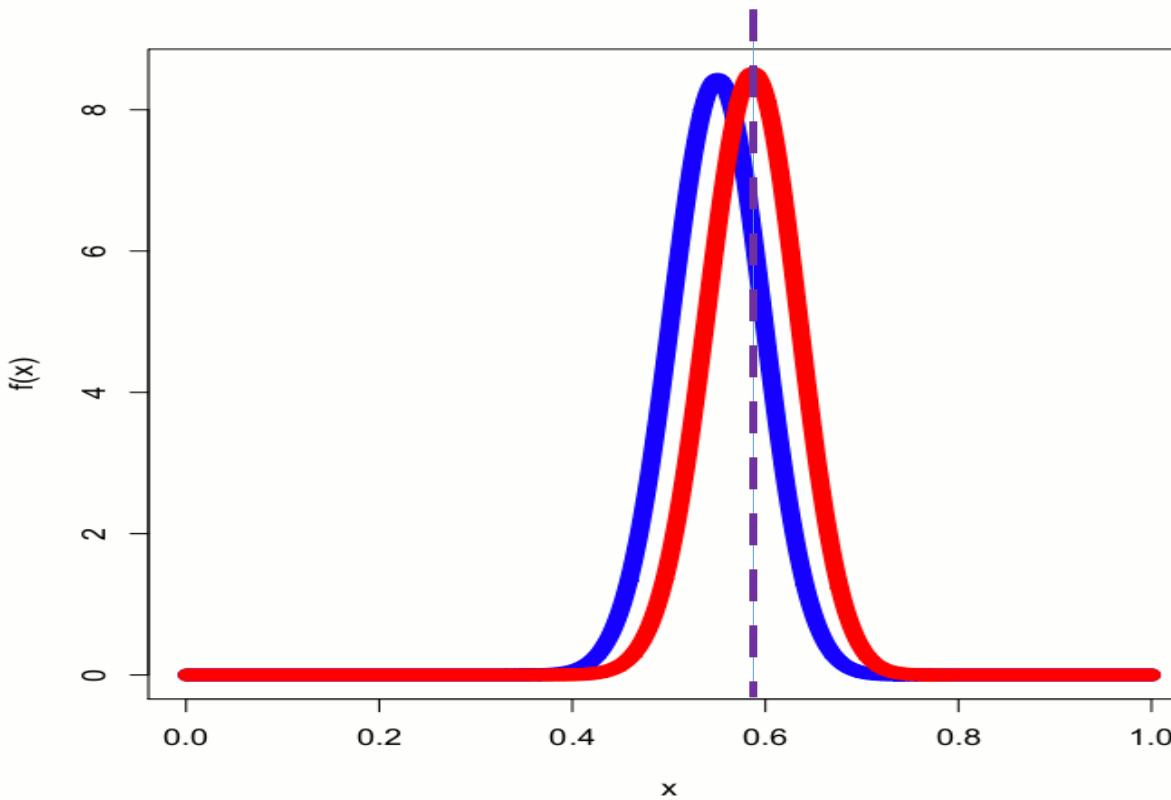
- $\theta$  is the probability a coin turns up heads
- Model  $\theta$  with 2 different priors:
  - $P_1(\theta)$  is Beta(3,8) (blue)
  - $P_2(\theta)$  is Beta(7,4) (red)
- They look pretty different!



- Now flip 100 coins; get 58 heads and 42 tails
  - What do posteriors look like?

# It's Like Having Twins

argmax returns the mode

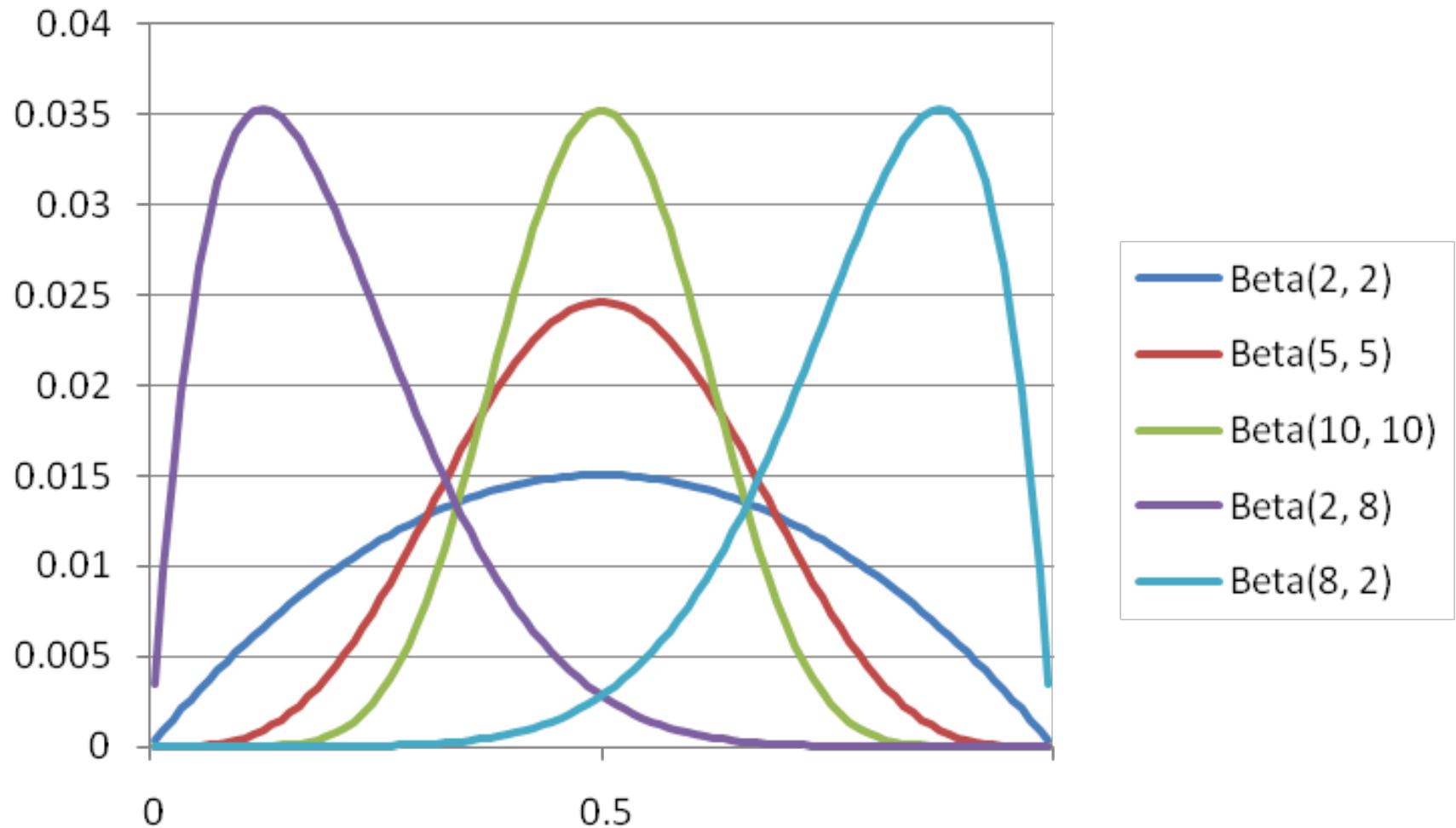


- As long as we collect enough data, posteriors will converge to the true value!

# Conjugate Distributions Without Tears

- Just for review...
- Have coin with unknown probability  $\theta$  of heads
  - Our prior (subjective) belief is that  $\theta \sim \text{Beta}(a, b)$
  - Now flip coin  $k = n + m$  times, getting  $n$  heads,  $m$  tails
  - Posterior density:  $(\theta | n \text{ heads}, m \text{ tails}) \sim \text{Beta}(a+n, b+m)$ 
    - Beta is conjugate for Bernoulli, Binomial, Geometric, and Negative Binomial
  - $a$  and  $b$  are called “hyperparameters”
    - Saw  $(a + b - 2)$  imaginary trials, of those  $(a - 1)$  are “successes”
  - For a coin you never flipped before, use  $\text{Beta}(x, x)$  to denote you think coin likely to be fair
    - How strongly you feel coin is fair is a function of  $x$

# Mo' Beta



# Gonna Need Priors

Parameter	Distribution for Parameter
Bernoulli $p$	Beta
Binomial $p$	Beta
Poisson $\lambda$	Gamma
Exponential $\lambda$	Gamma
Multinomial $p_i$	Dirichlet
Normal $\mu$	Normal
Normal $\sigma^2$	Inverse Gamma

Don't need to know Inverse Gamma. But it will know you...

# Multinomial is Multiple Times the Fun

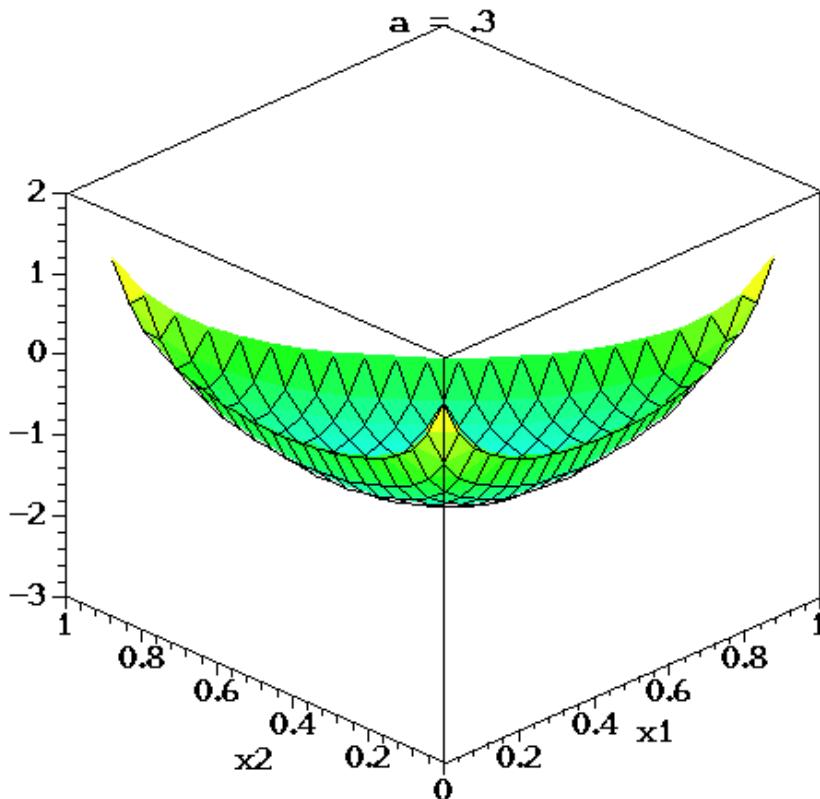
- Dirichlet( $a_1, a_2, \dots, a_m$ ) distribution
  - Conjugate for Multinomial
    - Dirichlet generalizes Beta in same way Multinomial generalizes Bernoulli

$$f(X_1 = x_1, X_2 = x_2, \dots, X_m = x_m) = K \prod_{i=1}^m x_i^{a_i - 1}$$

- Intuitive understanding of hyperparameters:
  - Saw  $\sum_{i=1}^m a_i - m$  imaginary trials, with  $(a_i - 1)$  of outcome  $i$
- Updating to get the posterior distribution
  - After observing  $n_1 + n_2 + \dots + n_m$ , new trials with  $n_i$  of outcome  $i \dots$
  - ... posterior distribution is Dirichlet( $a_1 + n_1, a_2 + n_2, \dots, a_m + n_m$ )

# Best Short Film in the Dirichlet Category

- And now a cool animation of  $\text{Dirichlet}(a, a, a)$ 
  - This is actually *loa* density (but you get the idea...)



Thanks  
Wikipedia!

# Example: Estimating Die Parameters



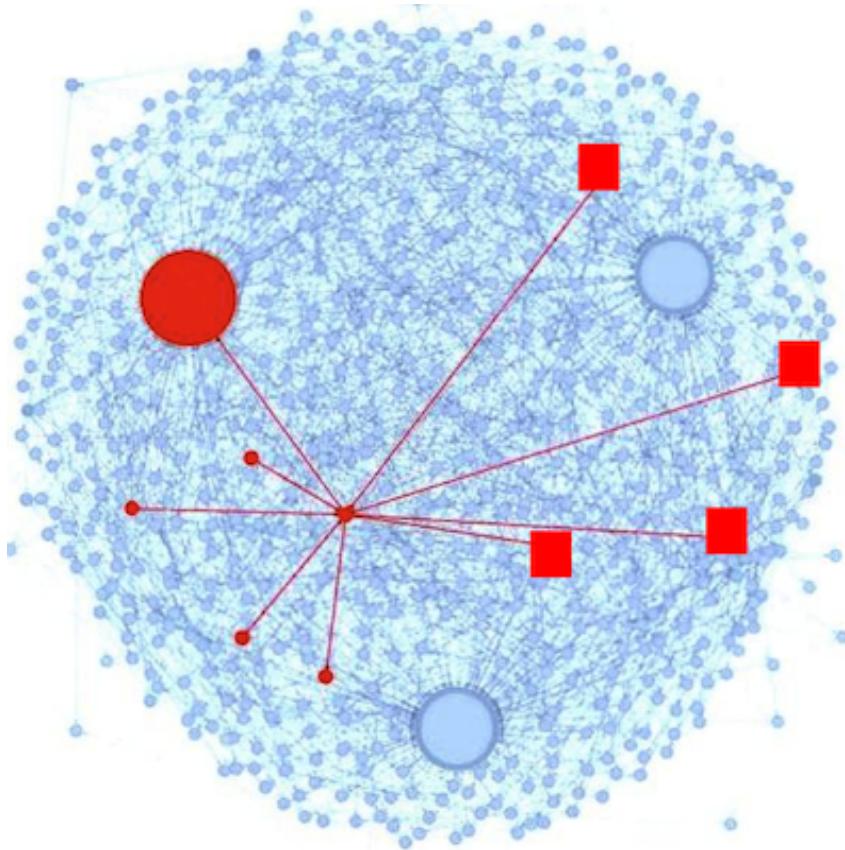
# Your Happy Laplace

- Recall example of 6-sides die rolls:
  - $X \sim \text{Multinomial}(p_1, p_2, p_3, p_4, p_5, p_6)$
  - Roll  $n = 12$  times
  - Result: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes
    - MLE:  $p_1=3/12$ ,  $p_2=2/12$ ,  $p_3=0/12$ ,  $p_4=3/12$ ,  $p_5=1/12$ ,  $p_6=3/12$
  - Dirichlet prior allows us to pretend we saw each outcome  $k$  times before. MAP estimate:  $p_i = \frac{X_i + k}{n + mk}$ 
    - Laplace’s “law of succession”: idea above with  $k = 1$
    - Laplace estimate:  $p_i = \frac{X_i + 1}{n + m}$
    - Laplace:  $p_1=4/18$ ,  $p_2=3/18$ ,  $p_3=1/18$ ,  $p_4=4/18$ ,  $p_5=2/18$ ,  $p_6=4/18$
    - No longer have 0 probability of rolling a three!

# Good Times with Gamma

- $\text{Gamma}(k, \theta)$  distribution
  - Conjugate for Poisson
    - Also conjugate for Exponential, but we won't delve into that
  - Intuitive understanding of hyperparameters:
    - Saw  $k$  total imaginary events during  $\theta$  prior time periods
  - Updating to get the posterior distribution
    - After observing  $n$  events during next  $t$  time periods...
    - ... posterior distribution is  $\text{Gamma}(k + n, \theta + t)$
    - Example:  $\text{Gamma}(10, 5)$
    - Saw 10 events in 5 time periods. Like observing at rate = 2
    - Now see 11 events in next 2 time periods →  $\text{Gamma}(21, 7)$
    - Equivalent to updated rate = 3

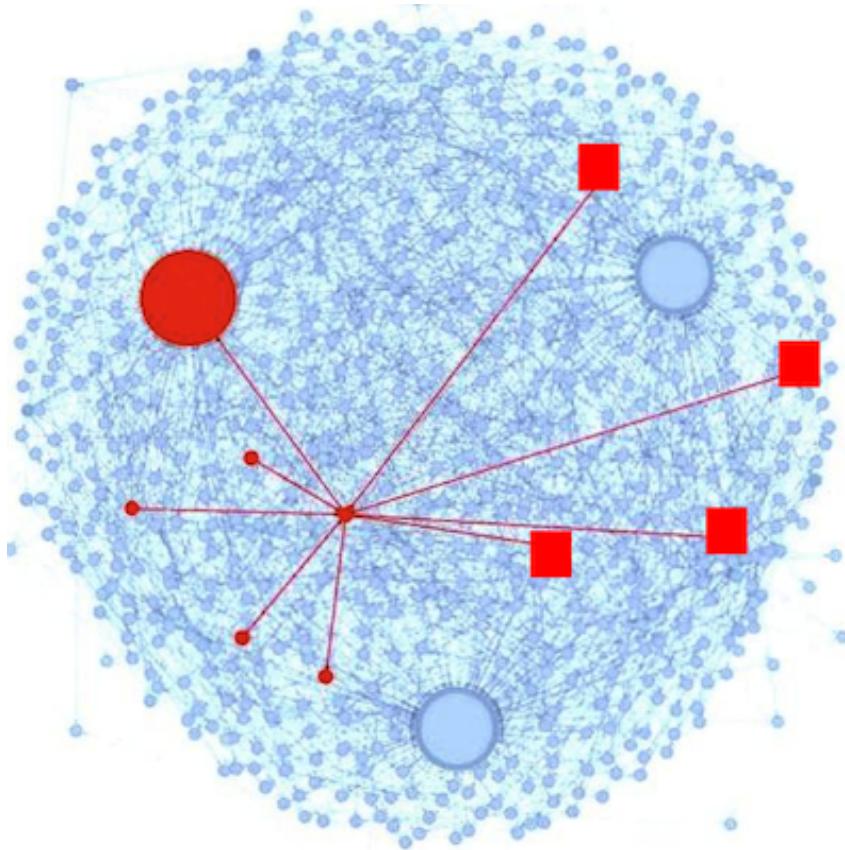
# Is Peer Grading Accurate Enough?



Peer Grading on Coursera  
HCI.

31,067 peer grades for  
3,607 students.

# Is Peer Grading Accurate Enough?



= hyperparameter

1. Defined random variables for:
  - True grade ( $s_i$ ) for assignment  $i$
  - Observed ( $z_i$ ) score for assign  $i$
  - Bias ( $b_j$ ) for each grader  $j$
  - Variance ( $r_j$ ) for each grader  $j$
2. Designed a probabilistic model that defined the distributions for all random variables

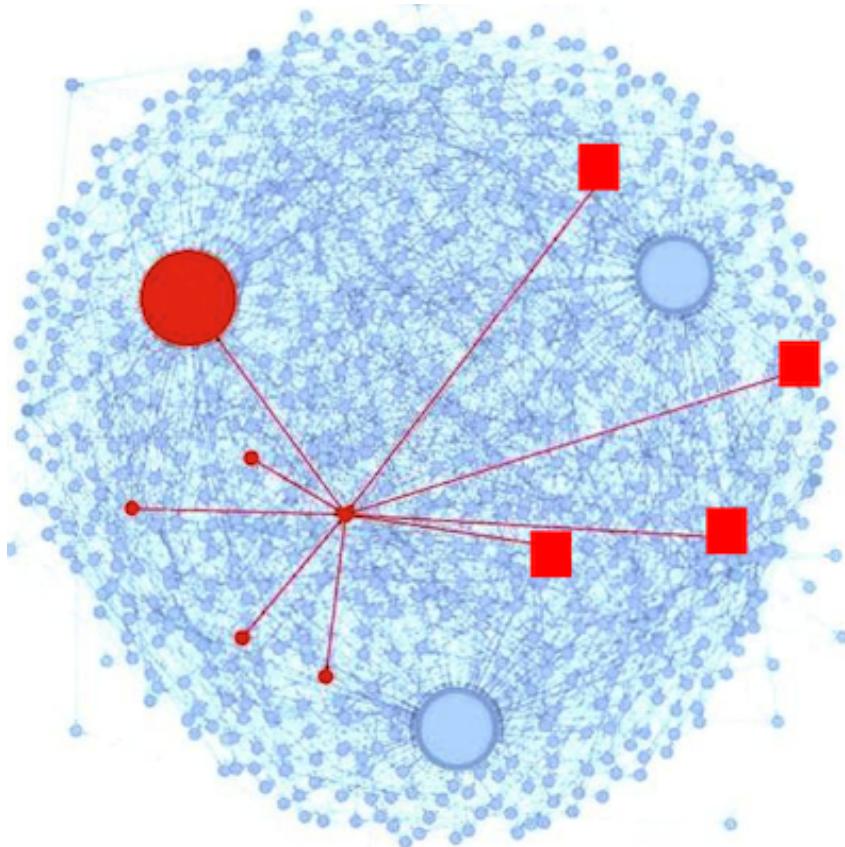
$$z_i^j \sim \mathcal{N}(\mu = s_i + b_j, \sigma = \sqrt{r_j})$$

$$s_i \sim N(\mu_0, \sigma_0)$$

$$b_i \sim N(0, \eta_0)$$

$$r_i \sim \text{InvGamma}(\alpha_0, \theta_0)$$

# Is Peer Grading Accurate Enough?



1. Defined random variables for:
  - True grade ( $s_i$ ) for assignment  $i$
  - Observed ( $z_i$ ) score for assign  $i$
  - Bias ( $b_j$ ) for each grader  $j$
  - Variance ( $r_j$ ) for each grader  $j$
2. Designed a probabilistic model that defined the distributions for all random variables
3. Found variable assignments using MAP estimation given the observed data

Inference or Machine Learning

The last estimator has risen...

Next time: Machine Learning algorithms





# It's Normal to Be Normal

- $\text{Normal}(\mu_0, \sigma_0^2)$  distribution
  - Conjugate for Normal (with unknown  $\mu$ , known  $\sigma^2$ )
  - Intuitive understanding of hyperparameters:
    - A priori, believe true  $\mu$  distributed  $\sim N(\mu_0, \sigma_0^2)$
  - Updating to get the posterior distribution
    - After observing  $n$  data points...
    - ... posterior distribution for  $\mu$  is:

$$N\left( \left( \frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n x_i}{\sigma^2} \right) \Bigg/ \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right), \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} \right)$$