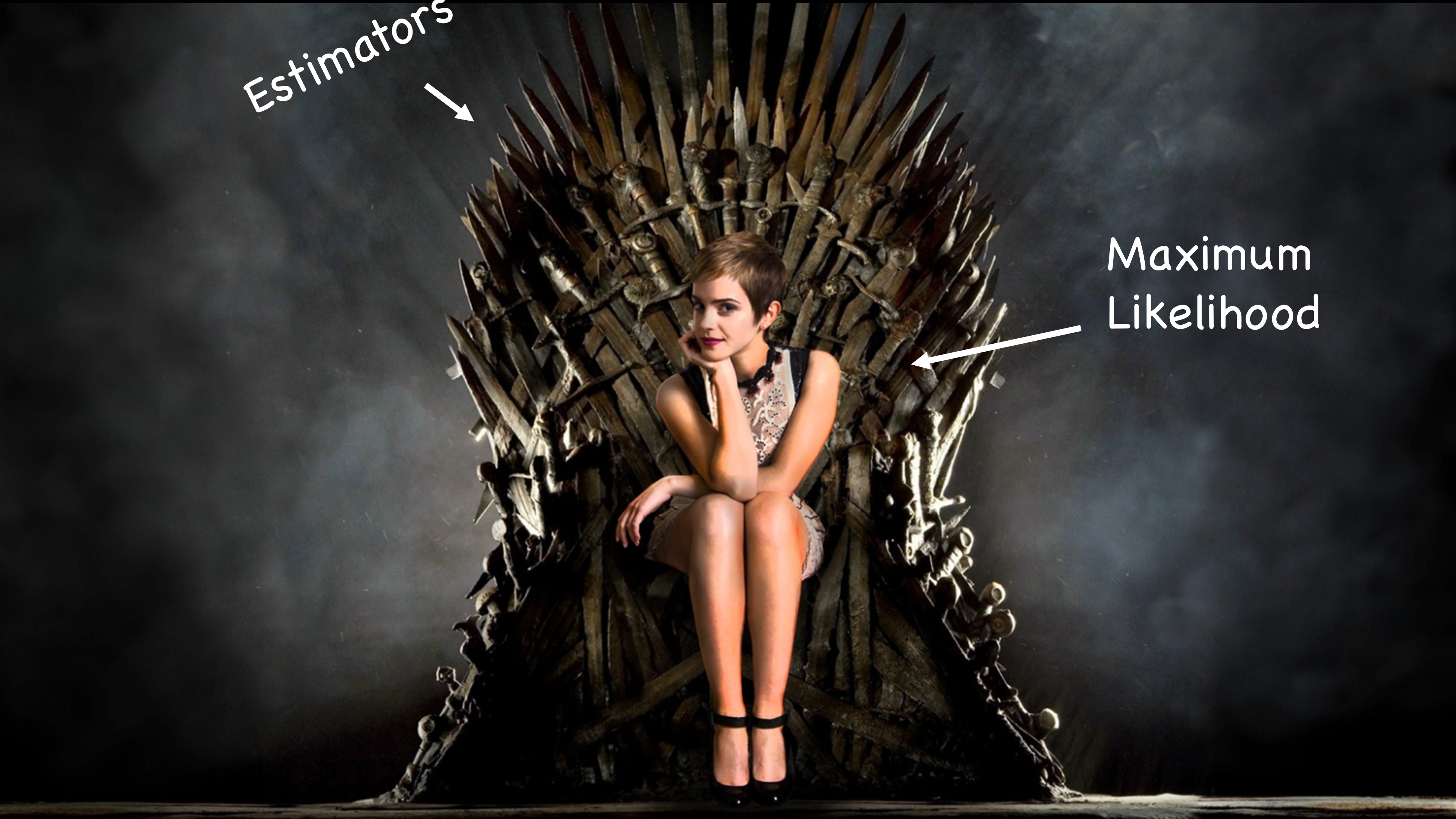
Maximum A Posteriori

Chris Piech
CS109, Stanford University

Previously in CS109…

Game of Estimators



Non spoiler: this didn't happen in game of thrones

Maximum Likelihood of Data

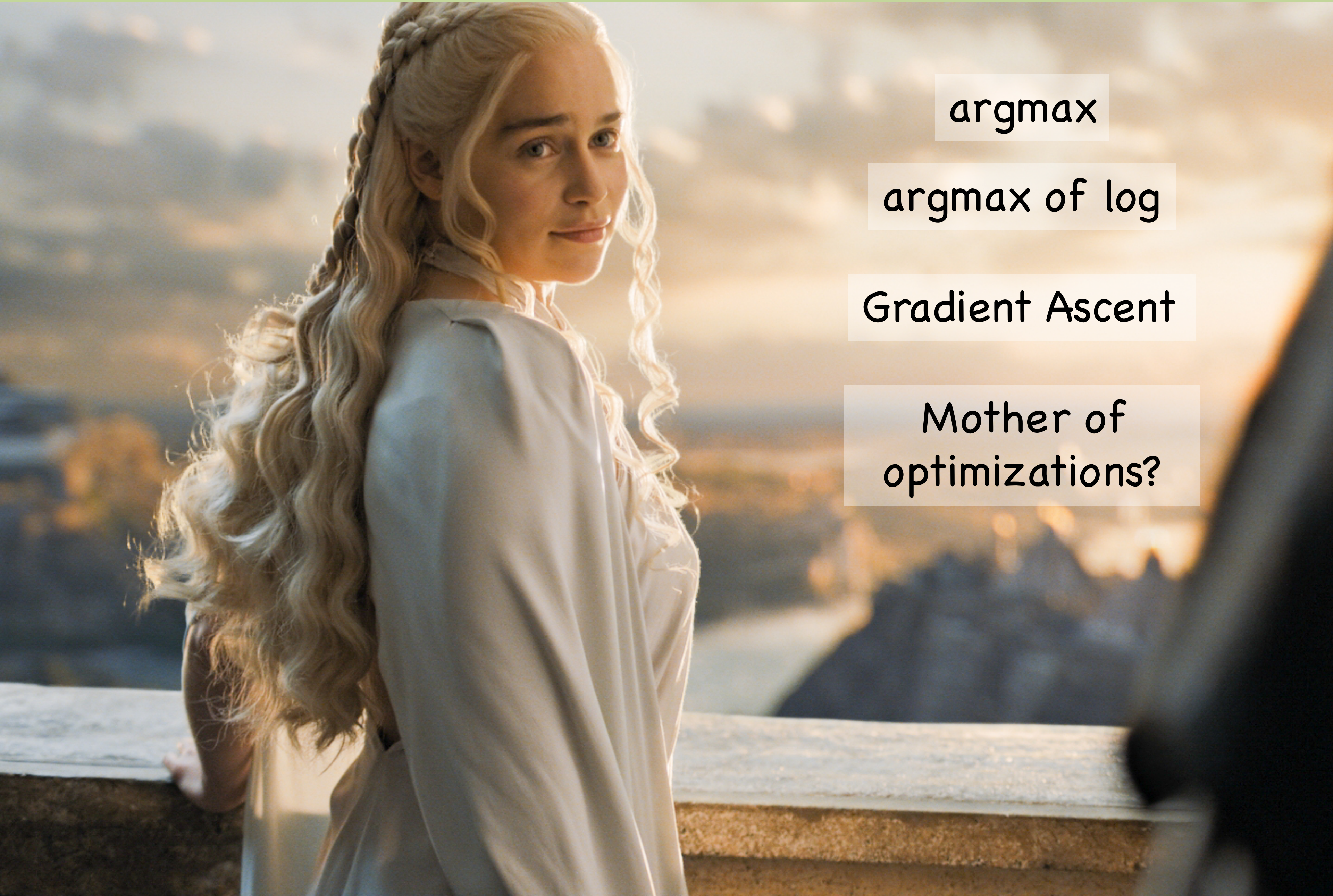
- Consider *n* I.I.D. random variables $X_1, X_2, \ldots, X_n$
  - X_i is a sample from density function $f(X_i \mid \theta)$

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

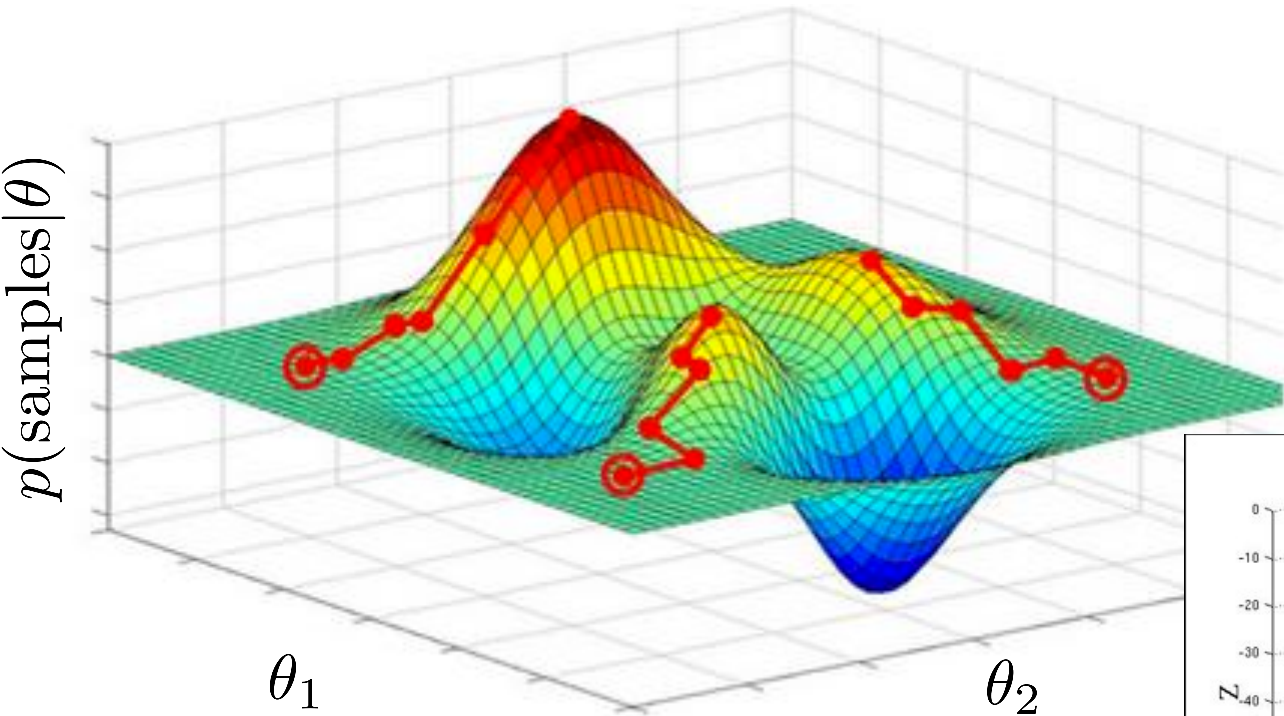
$$LL(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(X_i \mid \theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta)$$

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

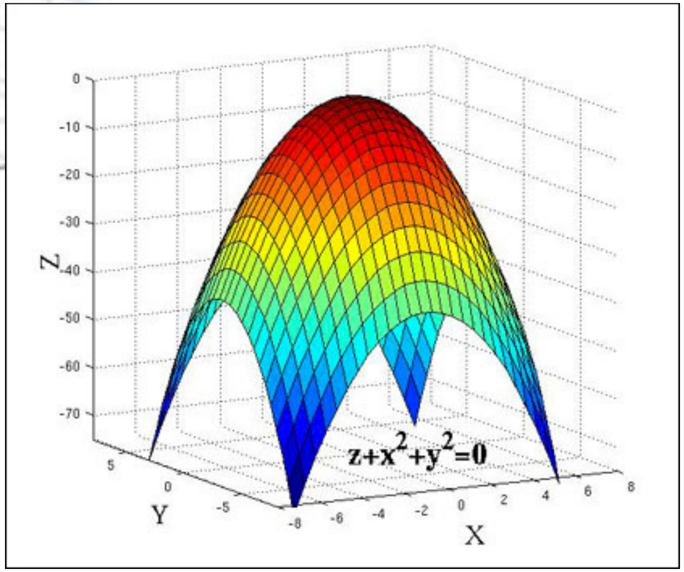
Side Plot


argmax
argmax of log
Gradient Ascent
Mother of optimizations?

Gradient Ascent



Especially good if function is convex



Walk uphill and you will find a local maxima
(if your step size is small enough)

Gradient Ascent

Initialize: $\theta_j = 0$ for all $0 \le j \le m$

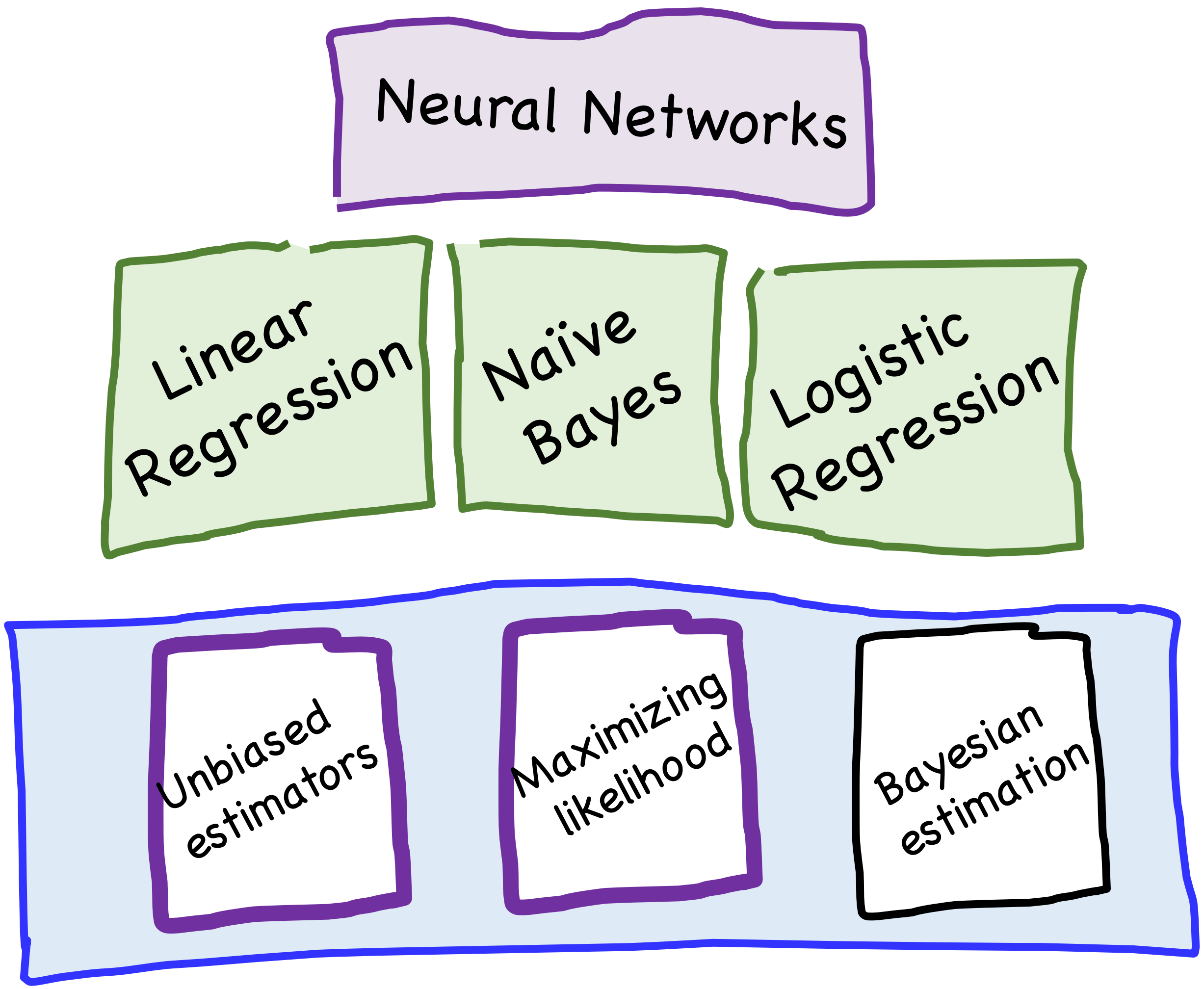
Repeat many times:

gradient[j] = 0 for all $0 \le j \le m$

Calculate all gradient[j]*'s based on data and current setting of theta*

θ_j += η * gradient[j] for all $0 \le j \le m$

Our Path

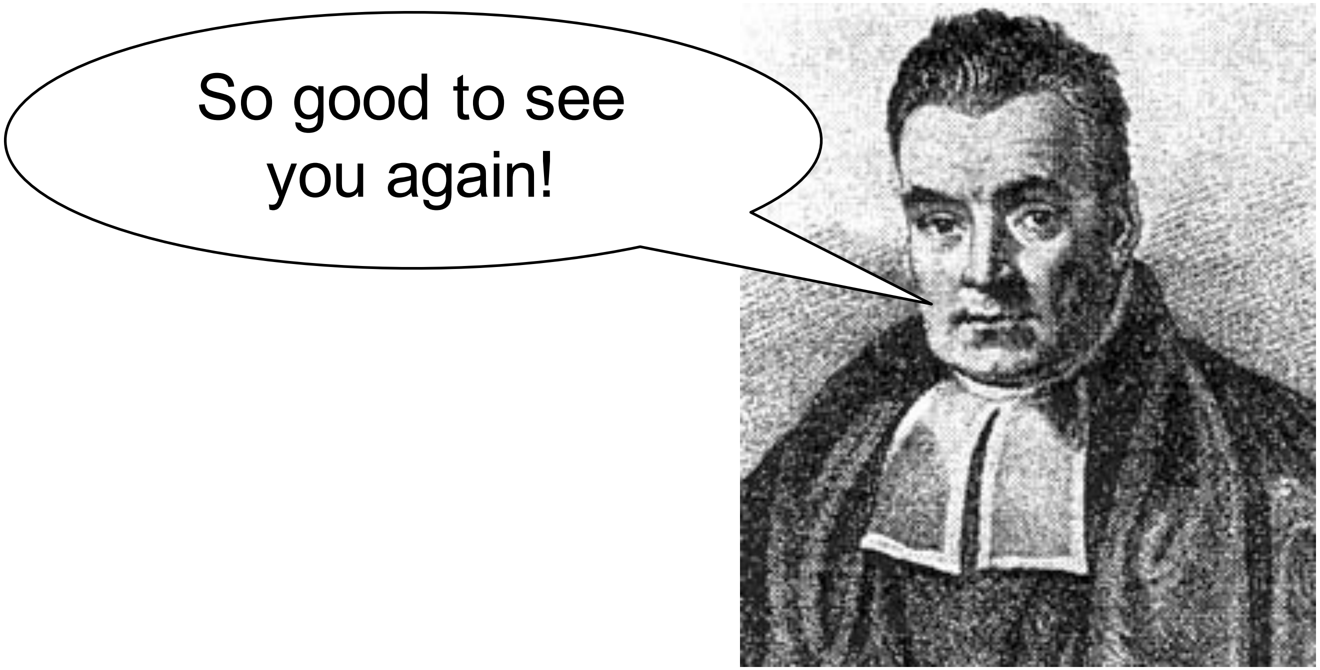

Neural Networks
Linear Regression
Naïve Bayes
Logistic Regression
Unbiased estimators
Maximizing likelihood
Bayesian estimation

Episode 2
The Song of The Last Estimator

Something rotten
in the world of MLE

Foreshadowing..

Need a Volunteer



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[\$ in other envelope]
    - Let Y = \$ 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 <u>equally</u> good
  - So, what happened by opening envelope?
    - And does it really make sense to switch?

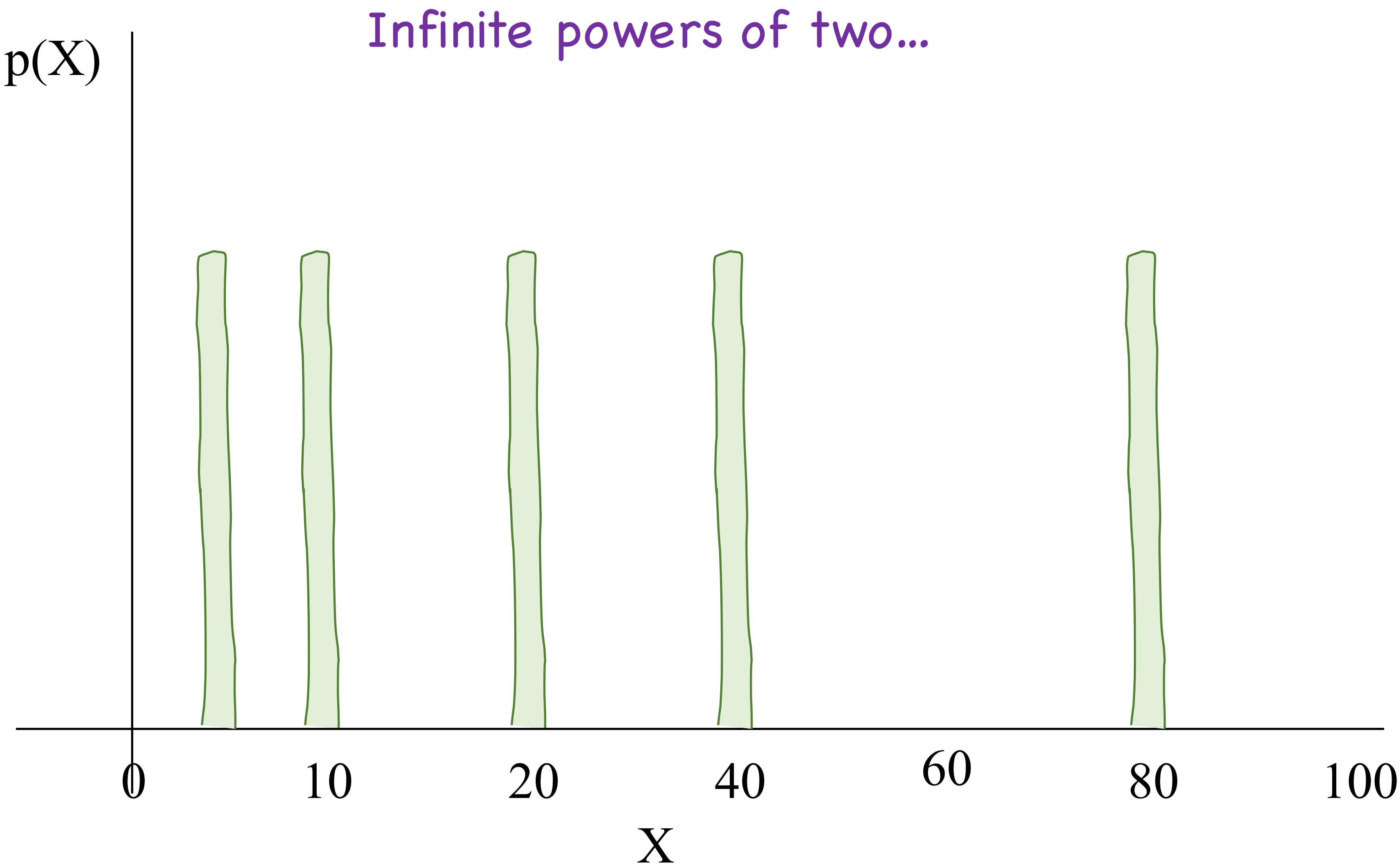
Thinking Deeper About Two Envelopes

- 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

$$E[Z \mid Y] = \frac{1}{2} \cdot \frac{Y}{2} + \frac{1}{2} \cdot 2Y = \frac{5}{4} Y$$

  - 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)

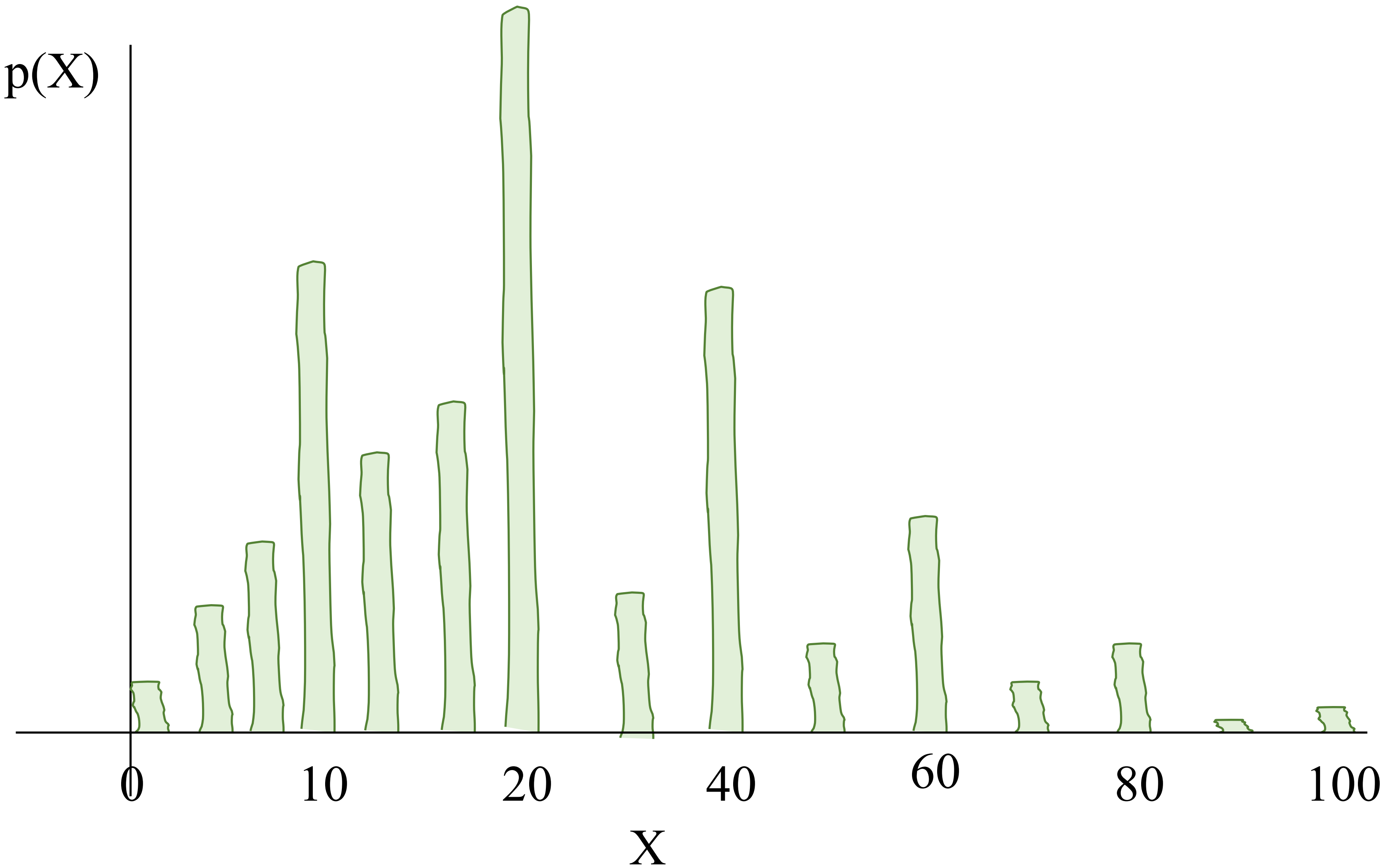
All Values are Equally Likely?



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

Envelope Summary:
Probabilities are beliefs
Incorporating prior beliefs is useful

Priors for Parameter Estimation?

Flash Back: Bayes Theorem

- Bayes' Theorem (θ = model parameters, D = data):

$$\underbrace{P(\theta \mid D)}_{\text{"Posterior"}} = \frac{\overbrace{P(D \mid \theta)}^{\text{"Likelihood"}} \, \overbrace{P(\theta)}^{\text{"Prior"}}}{P(D)}$$

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

Computing $P(\theta \mid D)$

- Bayes' Theorem (θ = model parameters, D = data):

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

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

Maximum A Posteriori

Maximum Likelihood Estimation

$$\theta_{\mathrm{MLE}} = \underset{\theta}{\operatorname{argmax}}\ f(X_1, X_2, \ldots, X_n|\theta)$$

$$= \underset{\theta}{\operatorname{argmax}}\ \sum_i \log f(X_i|\theta)$$

Maximum A Posteriori

$$\theta_{\mathrm{MAP}} = \underset{\theta}{\operatorname{argmax}}\ f(\theta|X_1, X_2, \ldots, X_n)$$

Most important slide of today

Maximum A Posteriori

- Recall Maximum Likelihood Estimator (MLE) of θ

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

- Maximum A Posteriori (MAP) estimator of θ:

$$\theta_{MAP} = \arg\max_{\theta} f(\theta \mid X_1, X_2, \ldots, X_n) = \arg\max_{\theta} \frac{f(X_1, X_2, \ldots, X_n \mid \theta)\, g(\theta)}{h(X_1, X_2, \ldots, X_n)}$$

$$= \arg\max_{\theta} \frac{\left(\prod_{i=1}^{n} f(X_i \mid \theta)\right) g(\theta)}{h(X_1, X_2, \ldots, X_n)} = \arg\max_{\theta} g(\theta) \prod_{i=1}^{n} f(X_i \mid \theta)$$

where g(θ) is prior distribution of θ.

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

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

  - MAP estimate is the mode of the posterior distribution

Maximum A Posteriori

Estimated parameter

Log prior

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

Sum of log likelihood

Chose the value of theta that maximizes:

MLE vs MAP

Maximum Likelihood Estimation

$$\theta_{\text{MLE}} = \underset{\theta}{\text{argmax}} \; f(X_1, X_2, \ldots, X_n|\theta)$$
$$= \underset{\theta}{\text{argmax}} \; \sum_i \log f(X_i|\theta)$$

Maximum A Posteriori

$$\theta_{\text{MAP}} = \underset{\theta}{\text{argmax}} \; f(\theta|X_1, X_2, \ldots, X_n)$$
$$= \underset{\theta}{\text{argmax}} \; \Big(\log g(\theta) + \sum_i \log f(X_i|\theta) \Big)$$

Gotta get that intuition

P(θ | D) For Bernoulli

- Prior: θ ~ Beta(*a*, *b*); D = {*n* heads, *m* tails}
- Estimate *p, aka* θ

$$\theta_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}}\ f(\theta|X_1, X_2, \ldots, X_n)$$

$$= \underset{\theta}{\operatorname{argmax}}\ \Big(\log g(\theta) + \sum_i \log f(X_i|\theta)\Big)$$

$$= \underset{\theta}{\operatorname{argmax}} \log\left[\frac{1}{\beta}\theta^{a-1}(1-\theta)^{b-1}\right]$$
$$+n \log f(\text{heads}|\theta)$$
$$+m \log f(\text{tails}|\theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \log\frac{1}{\beta} + (a-1)\log\theta + (b-1)\log(1-\theta) + n\log\theta + m\log(1-\theta)$$

$$= \underset{\theta}{\operatorname{argmax}}(a-1+n)\log\theta + (b-1+m)\log(1-\theta)$$

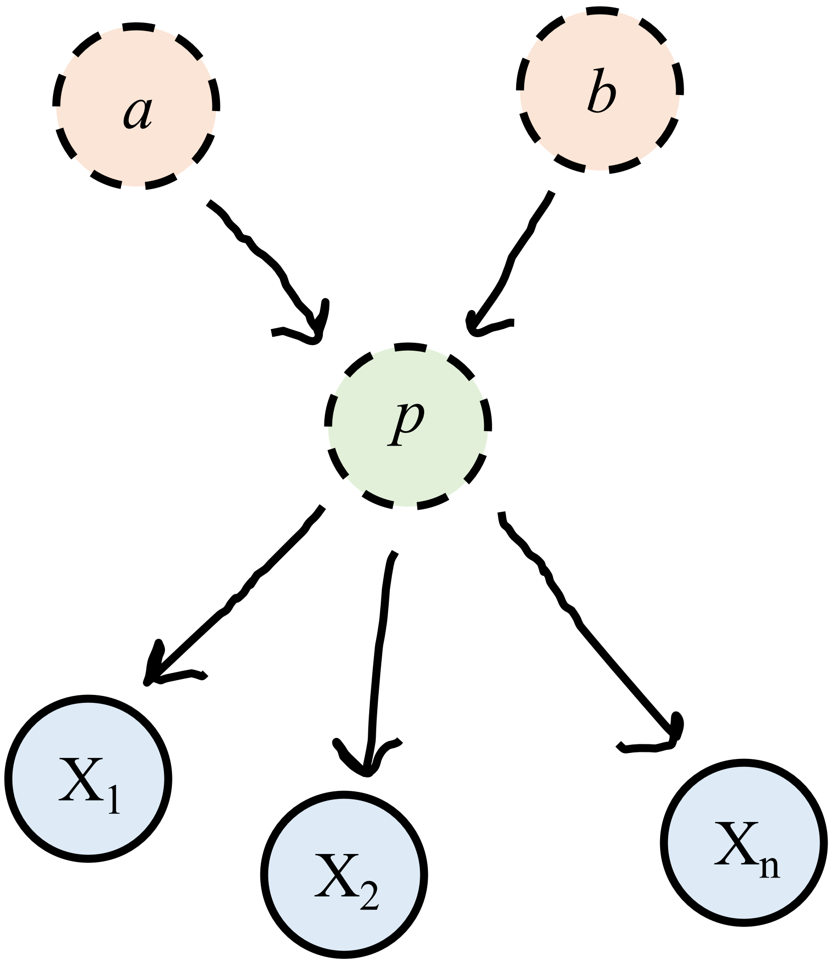
$P(\theta \mid D)$ For Bernoulli

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

$$\begin{aligned}\theta_{\text{MAP}} &= \underset{\theta}{\operatorname{argmax}}\ f(\theta|D) \\ &= \underset{\theta}{\operatorname{argmax}}\ (n+a-1)\log\theta + (m+b-1)\log(1-\theta)\end{aligned}$$

$$\theta_{\text{MAP}} = \frac{n+a-1}{n+m+a+b-2}$$

Hyper Parameters



Hyperparameter

a, b are fixed

Prior

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

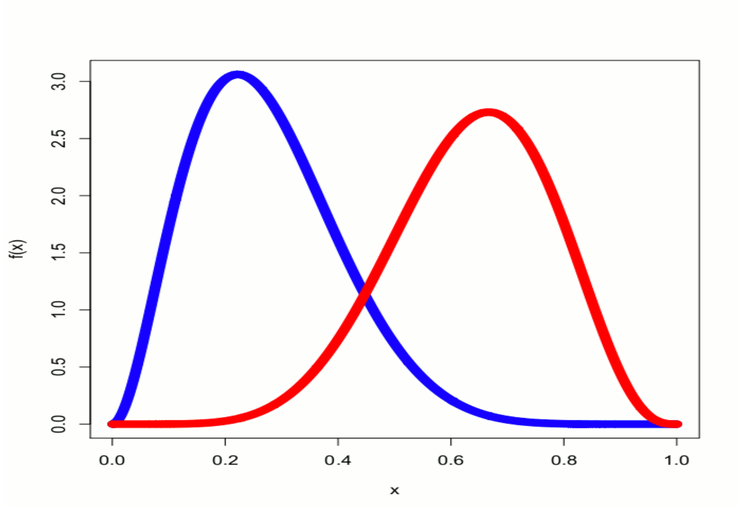
Data distribution

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

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

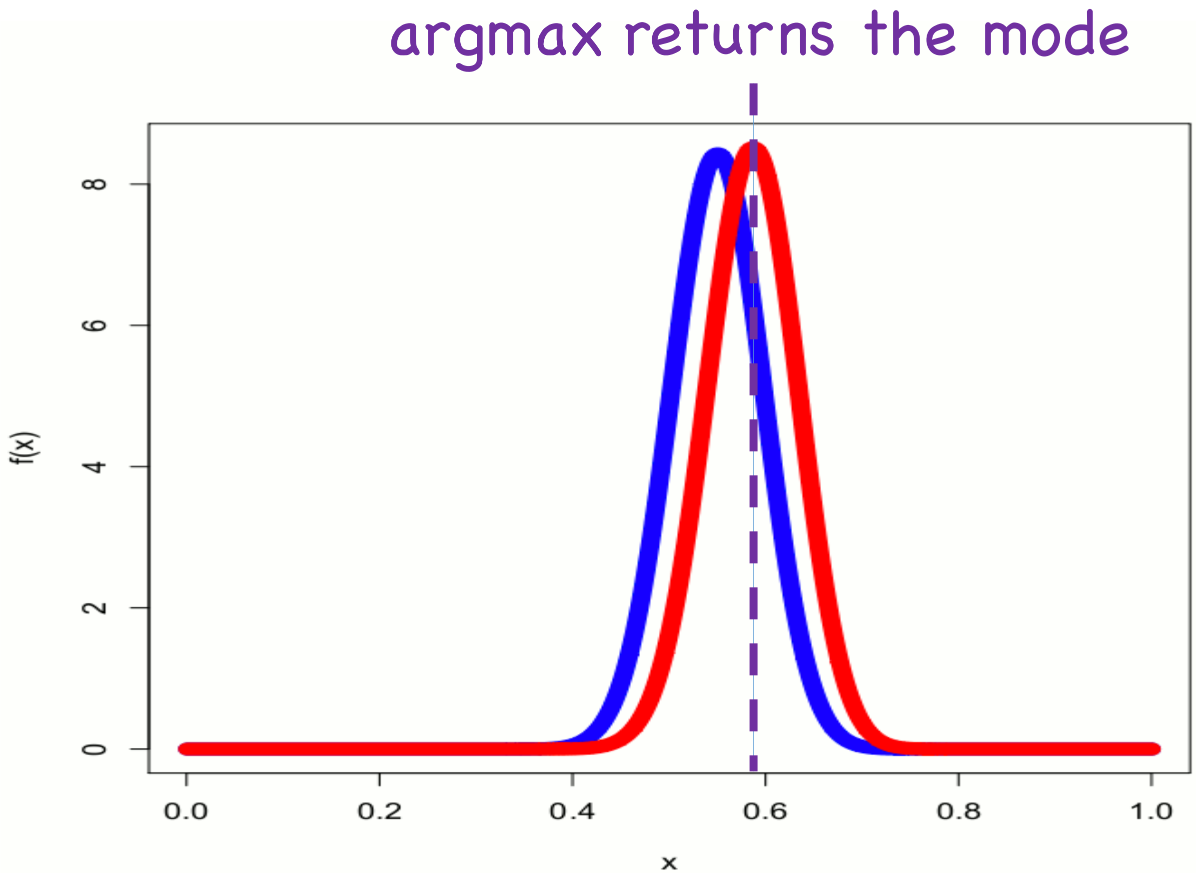
Where'd Ya Get Them P(θ)?

- θ is the probability a coin turns up heads
- Model θ 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

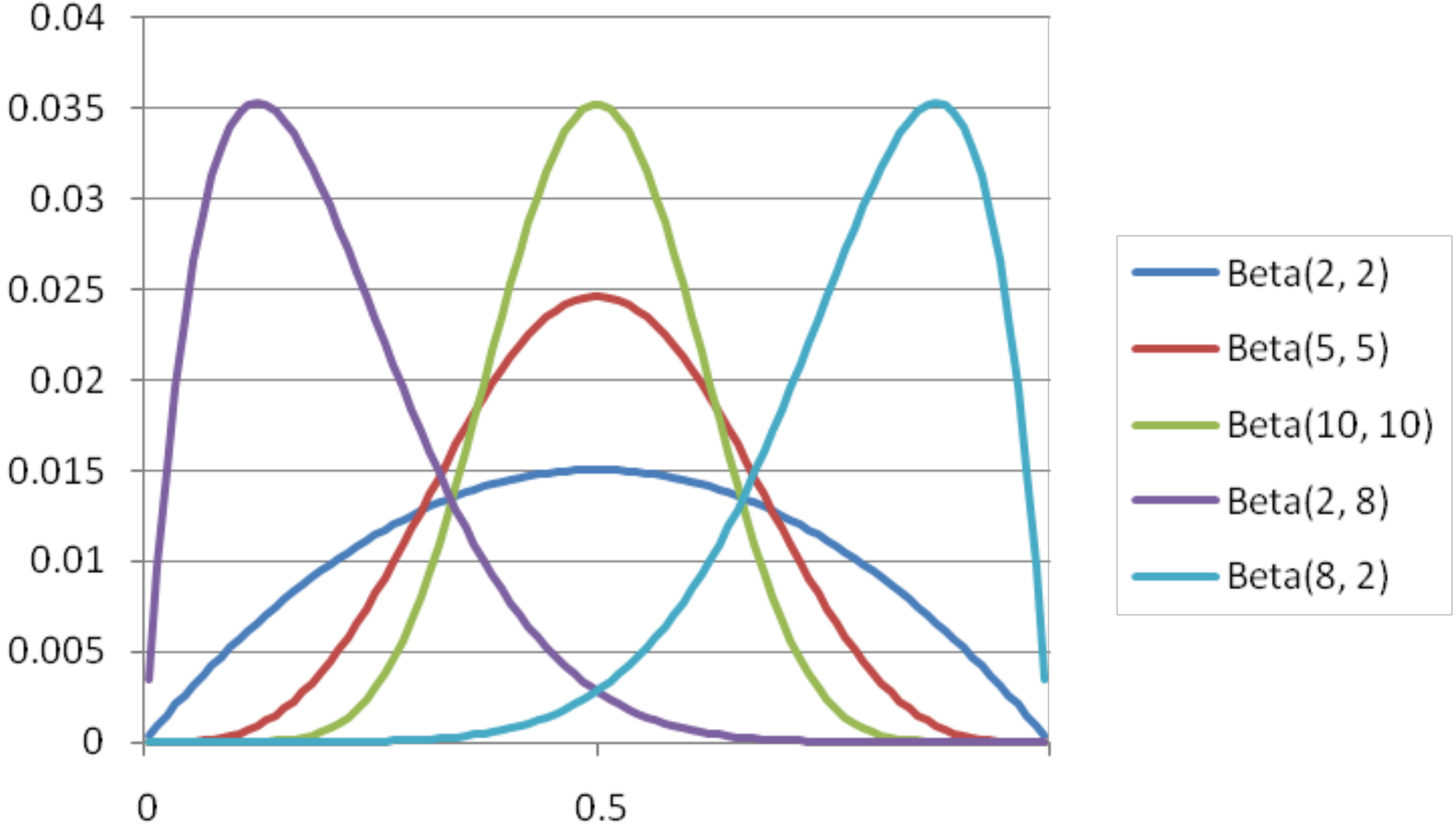


- As long as we collect enough data, posteriors will converge to the true value!

Conjugate Distributions Without Tears

- Just for review…
- Have coin with unknown probability θ of heads
  - Our <u>prior</u> (subjective) belief is that $\theta \sim \text{Beta}(a, b)$
  - Now flip coin $k = n + m$ times, getting n heads, m tails
  - Posterior density: $(\theta \mid n \text{ heads}, m \text{ tails}) \sim \text{Beta}(a+n, b+m)$
    - 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 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



0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
0.5
Beta(2, 2)
Beta(5, 5)
Beta(10, 10)
Beta(2, 8)
Beta(8, 2)

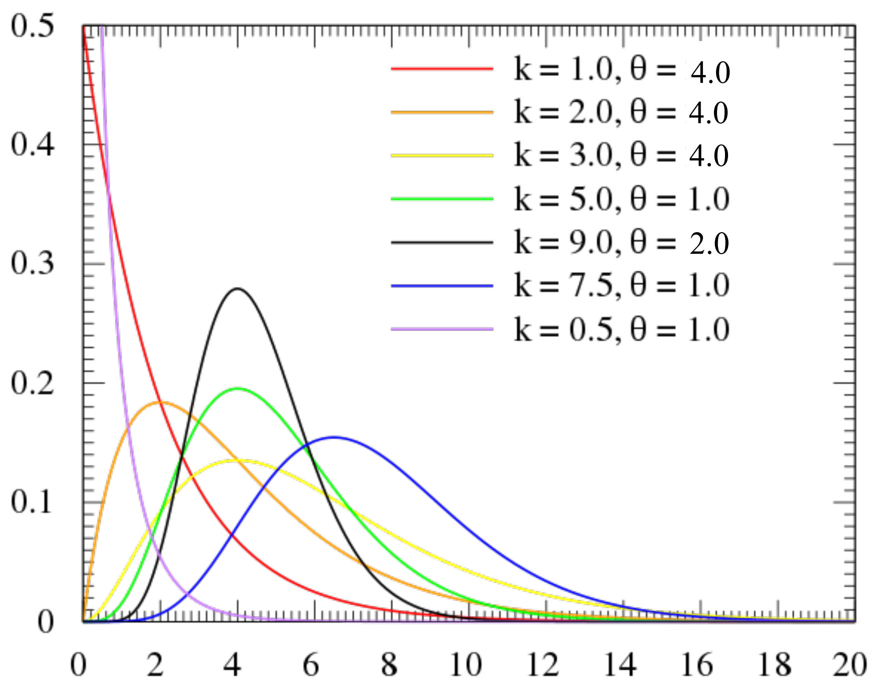
Gonna Need Priors

Parameter	Distribution for Parameter
Bernoulli p	Beta
Binomial p	Beta
Poisson λ	Gamma
Exponential λ	Gamma
Multinomial p_i	Dirichlet
Normal μ	Normal
Normal σ^2	Inverse Gamma

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

Good Times with Gamma

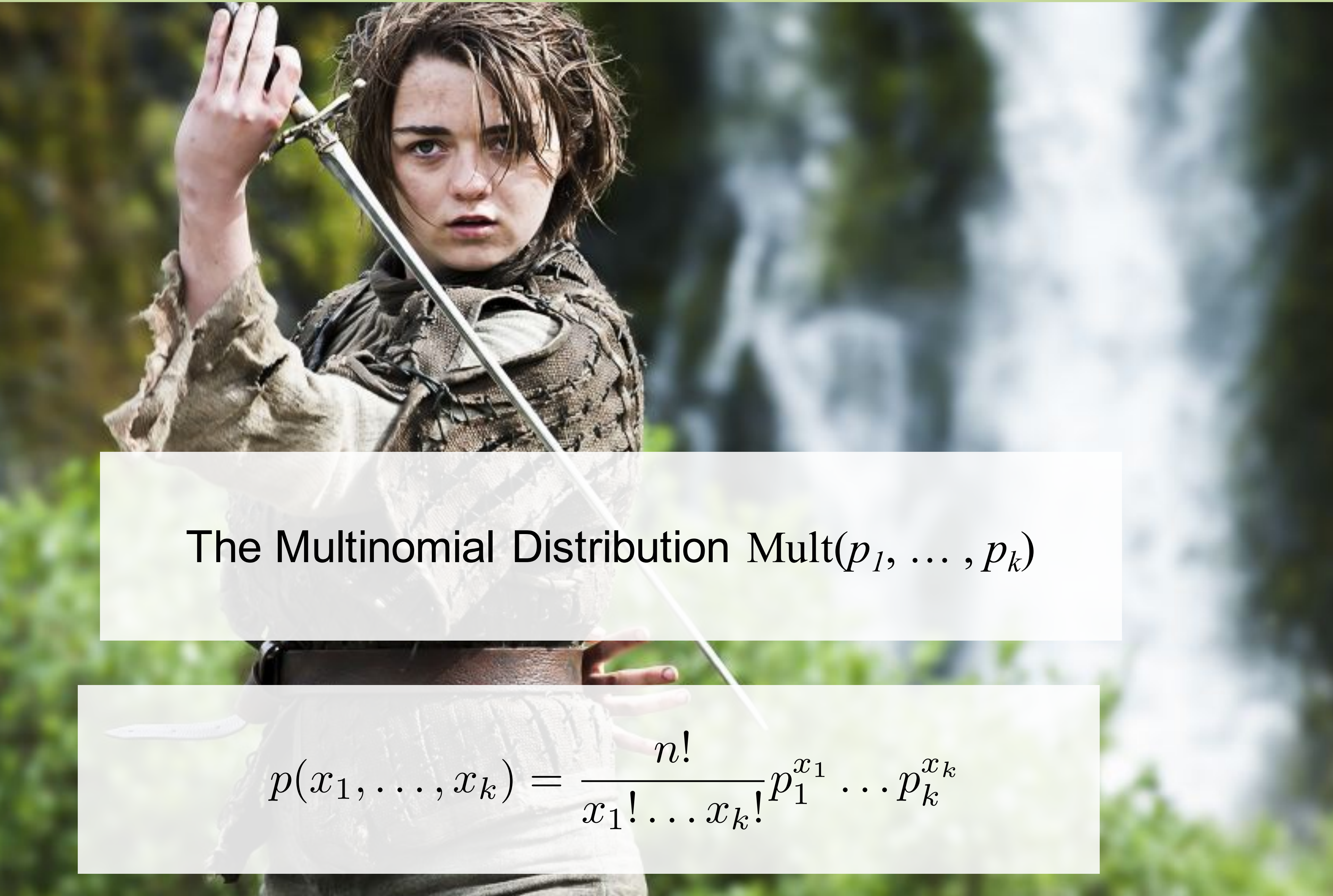
- Gamma(k, θ) distribution
  - Conjugate for Poisson Rate
    - Also conjugate for Exponential, but we won't delve into that
  - Intuitive understanding of hyperparameters:
    - Saw k total imaginary events during θ prior time periods



Good Times with Gamma

- Gamma(k, θ) distribution
  - Conjugate for Poisson Rate
    - Also conjugate for Exponential, but we won't delve into that
  - Intuitive understanding of hyperparameters:
    - Saw k total imaginary events during θ prior time periods
  - Updating with observations
    - After observing n events during next t time periods…
    - ... posterior distribution is Gamma(k + n, θ + t)
    - …MAP estimator for Poisson with Gamma prior is (k+n)/(θ + t)
    - Example: Prior for rate is Gamma(10, 5)
    - Saw 10 events in 5 time periods. Like observing at rate = 2
    - Now see 11 events in next 2 time periods → Gamma(21, 7)
    - MAP rate = 3

Reviving an Old Story Line



The Multinomial Distribution $\mathrm{Mult}(p_1, \ldots, p_k)$

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

Multinomial is Multiple Times the Fun

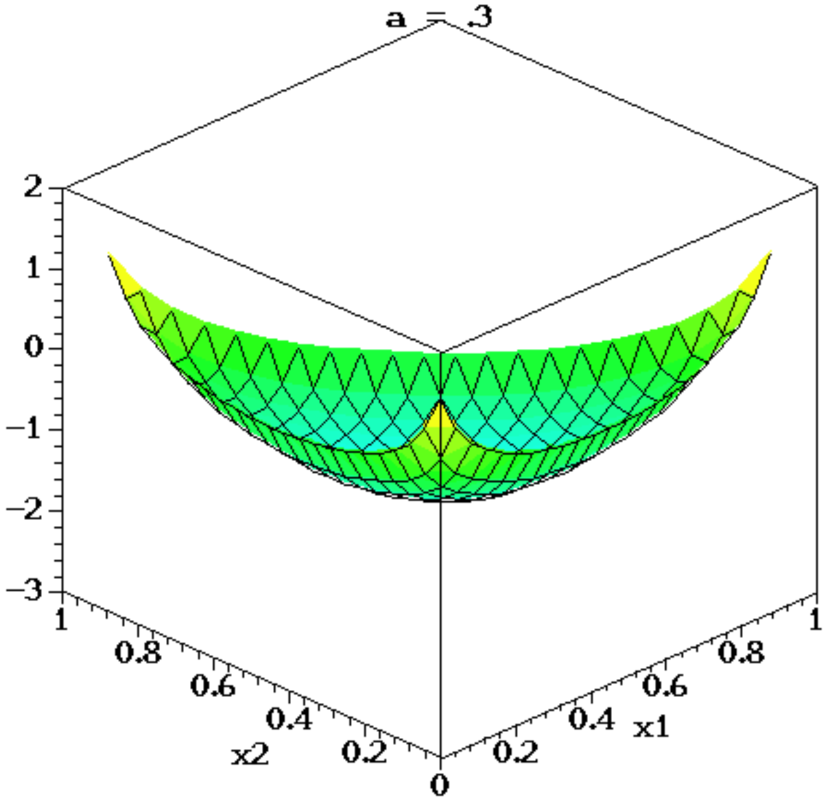
- Dirichlet(a_1, a_2, ..., a_m) distribution
  - Conjugate for Multinomial
    - Dirichlet generalizes Beta in same way Multinomial generalizes Bernoulli

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

  - 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 $n_1 + n_2 + \ldots + n_m$, new trials with n_i of outcome i...
    - ... posterior distribution is Dirichlet($a_1 + n_1$, $a_2 + n_2$, ..., $a_m + n_m$)

Best Short Film in the Dirichlet Category

- And now a cool animation of Dirichlet(*a*, *a*, *a*)
  - This is actually *log* density (but you get the idea…)



Thanks Wikipedia!

Example: Estimating Die Parameters

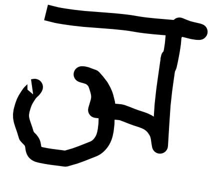
Your Happy Laplace

- Recall example of 6-sides die rolls:
  - X ~ 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!

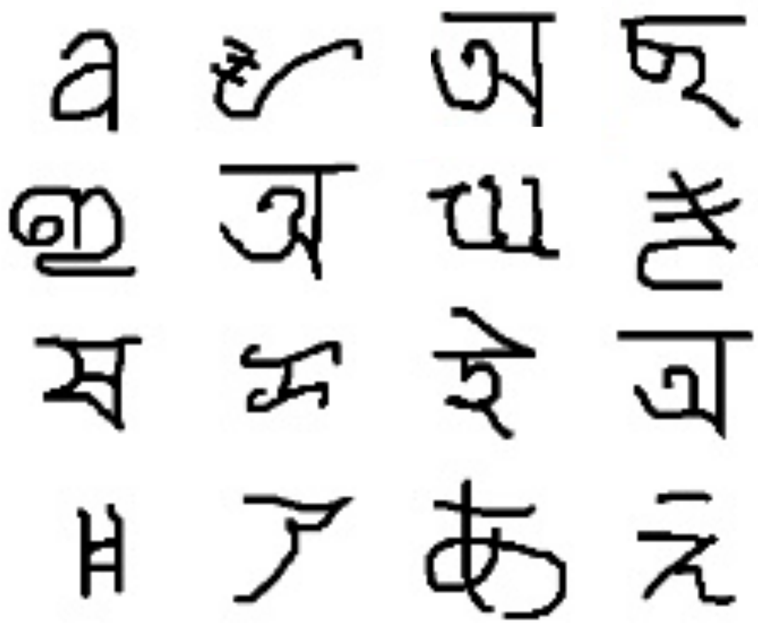
The last estimator has risen…

One Shot Learning

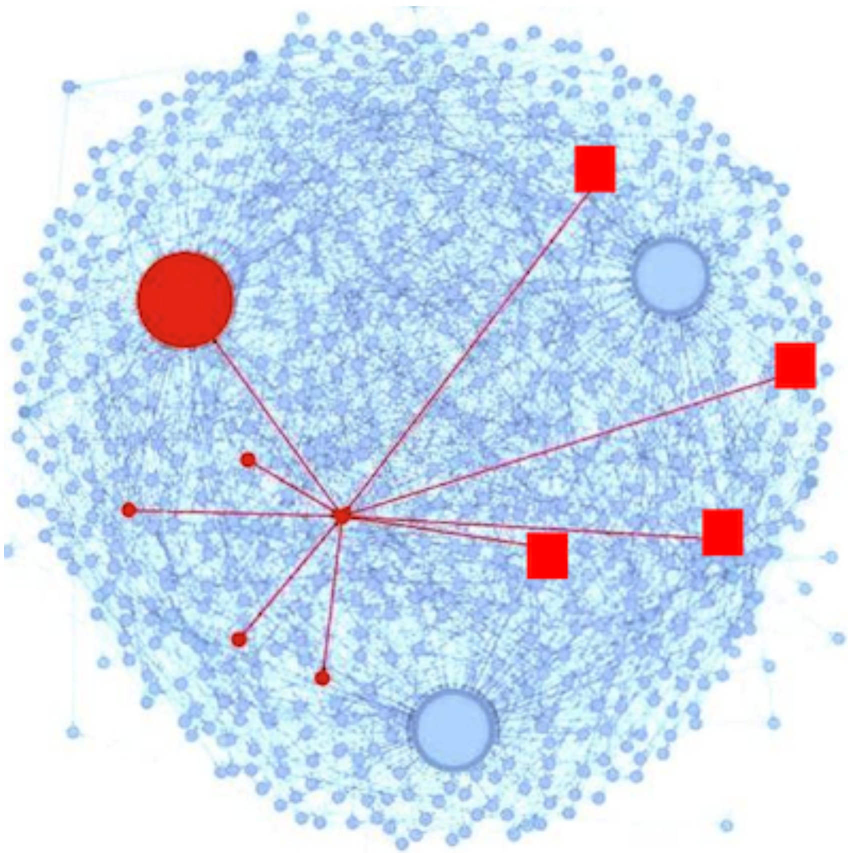
Single training example:



Test set:



Is Peer Grading Accurate Enough?

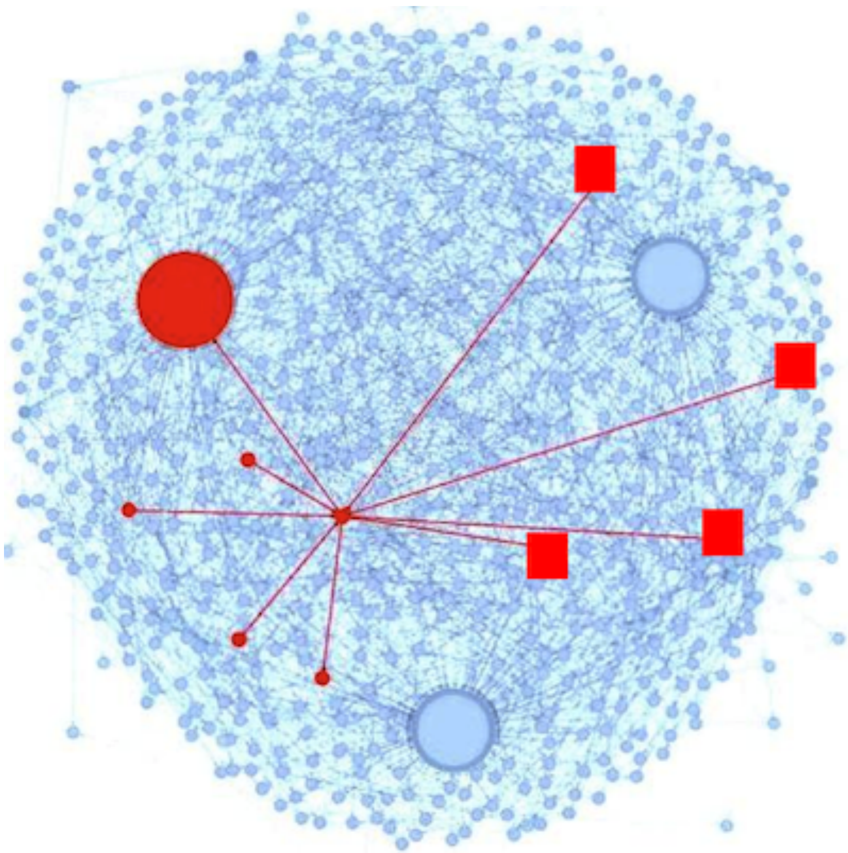


Peer Grading on Coursera HCI.

31,067 peer grades for 3,607 students.

Tuned Models of Peer Assessment. C Piech, J Huang, A Ng, D Koller

Is Peer Grading Accurate Enough?



1. Defined random variables for:
- True grade (s_i) for assignment i
- Observed (z_i^j) 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

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

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

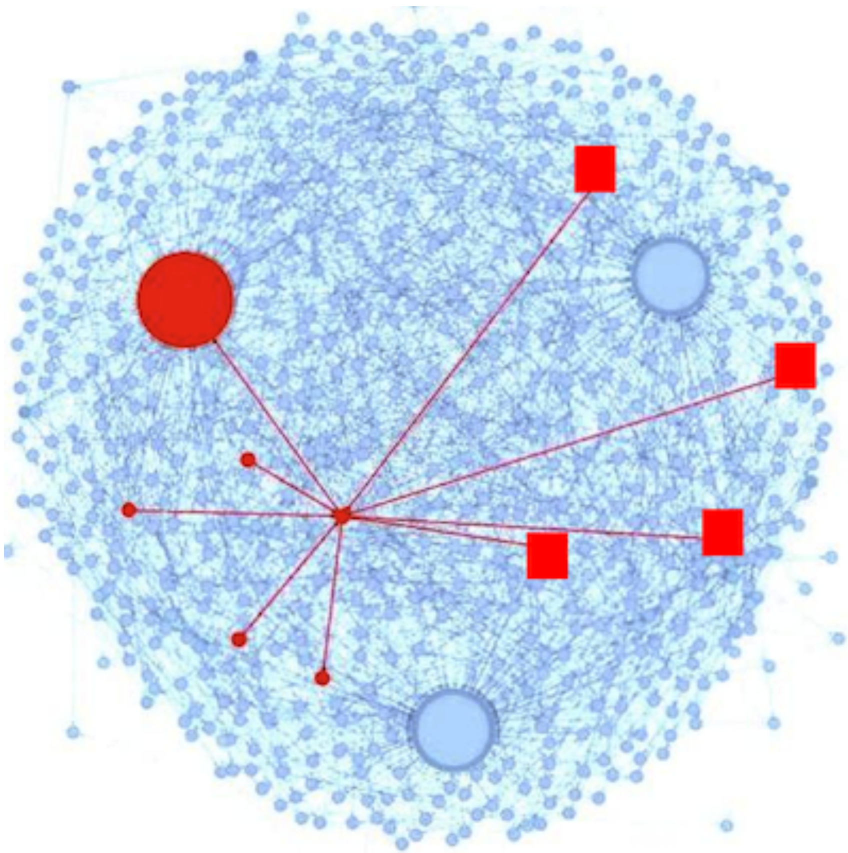
$$b_i \sim N(0, \eta_0)$$

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

= hyperparameter

Tuned Models of Peer Assessment. C Piech, J Huang, A Ng, D Koller

Is Peer Grading Accurate Enough?



1. Defined random variables for:
- True grade (s_i) for assignment i
- Observed (z_i^j) score for assign i
- Bias (b_j) for each grader j
- Variance (r_j) for each grader j

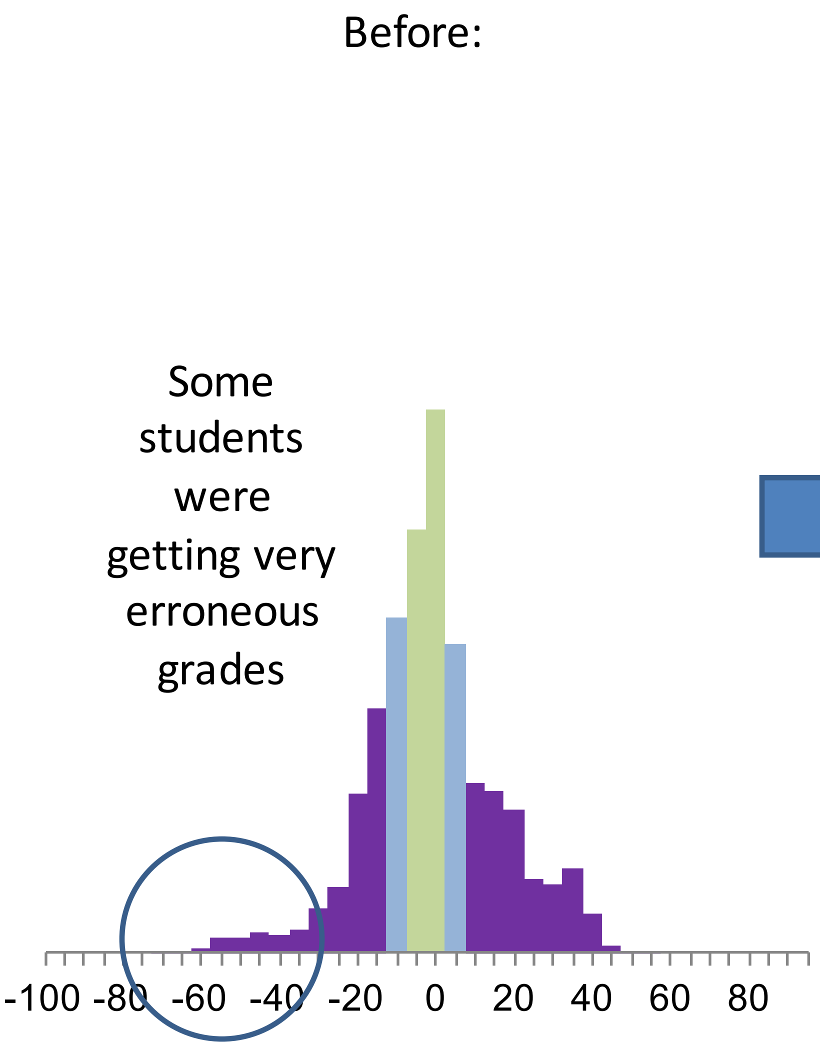
2. Designed a probabilistic model that defined the distributions for all random variables

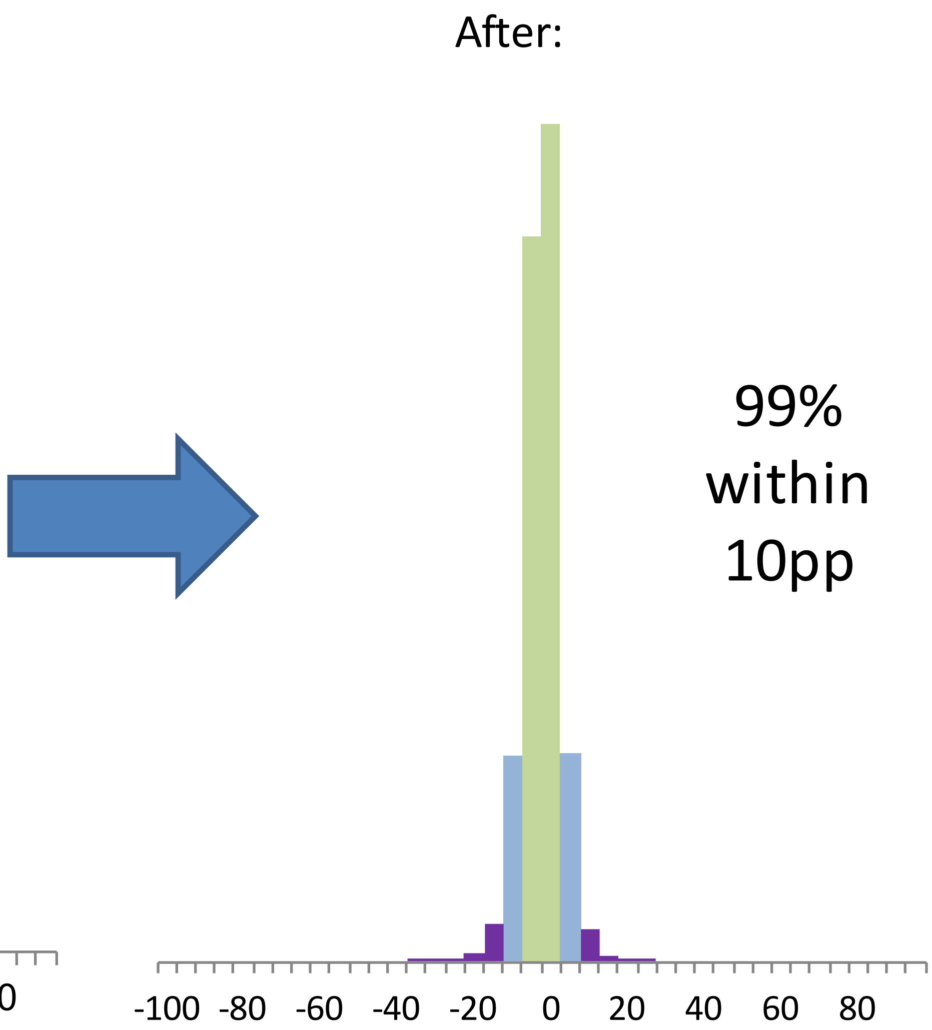
3. Found variable assignments using MAP estimation given the observed data

Inference or Machine Learning

Tuned Models of Peer Assessment. C Piech, J Huang, A Ng, D Koller

Improved Accuracy





Error is based on ground truth assignments. Results are across all assignments (~10,000 submissions)

Next time: Machine Learning algorithms

It's Normal to Be Normal

- Normal(μ_0, σ_0^2) distribution
  - Conjugate for Normal (with unknown μ, known σ^2)
  - Intuitive understanding of hyperparameters:
    - A priori, believe true μ distributed ~ N(μ_0, σ_0^2)
  - Updating to get the posterior distribution
    - After observing *n* data points…
    - … posterior distribution for μ is:

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