



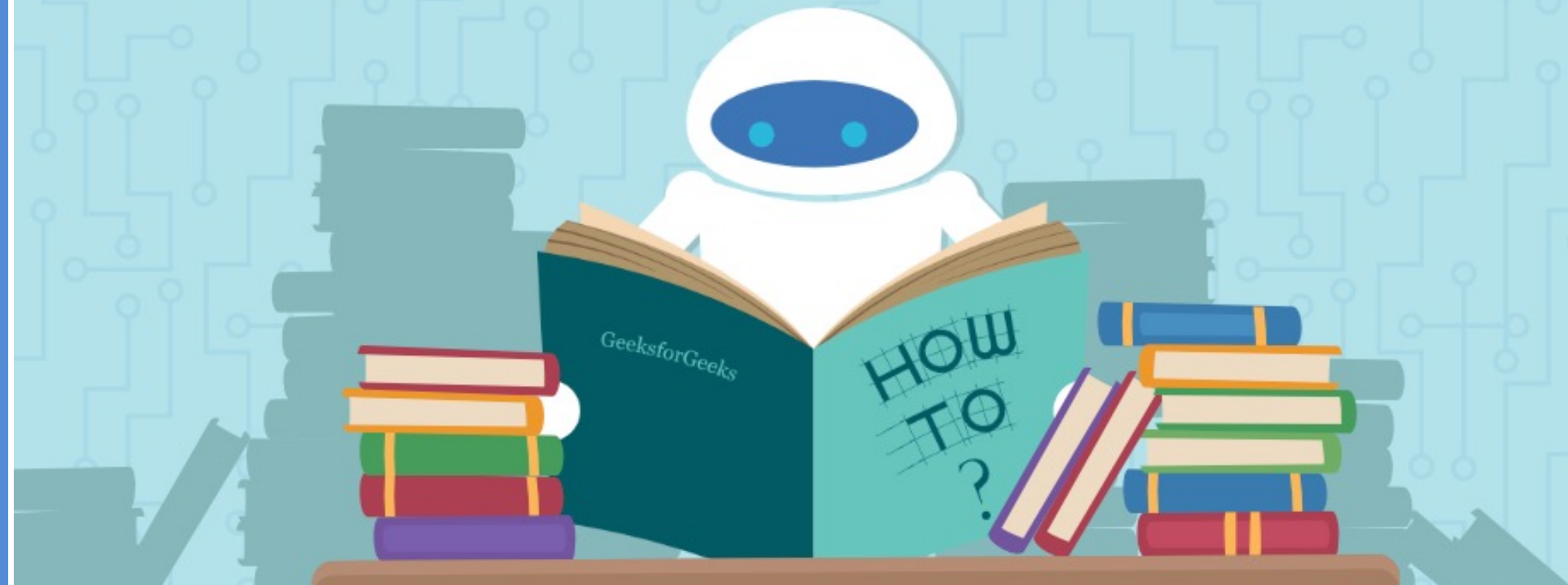
Maximum A Posteriori

Chris Piech

CS109, Stanford University

<Review>

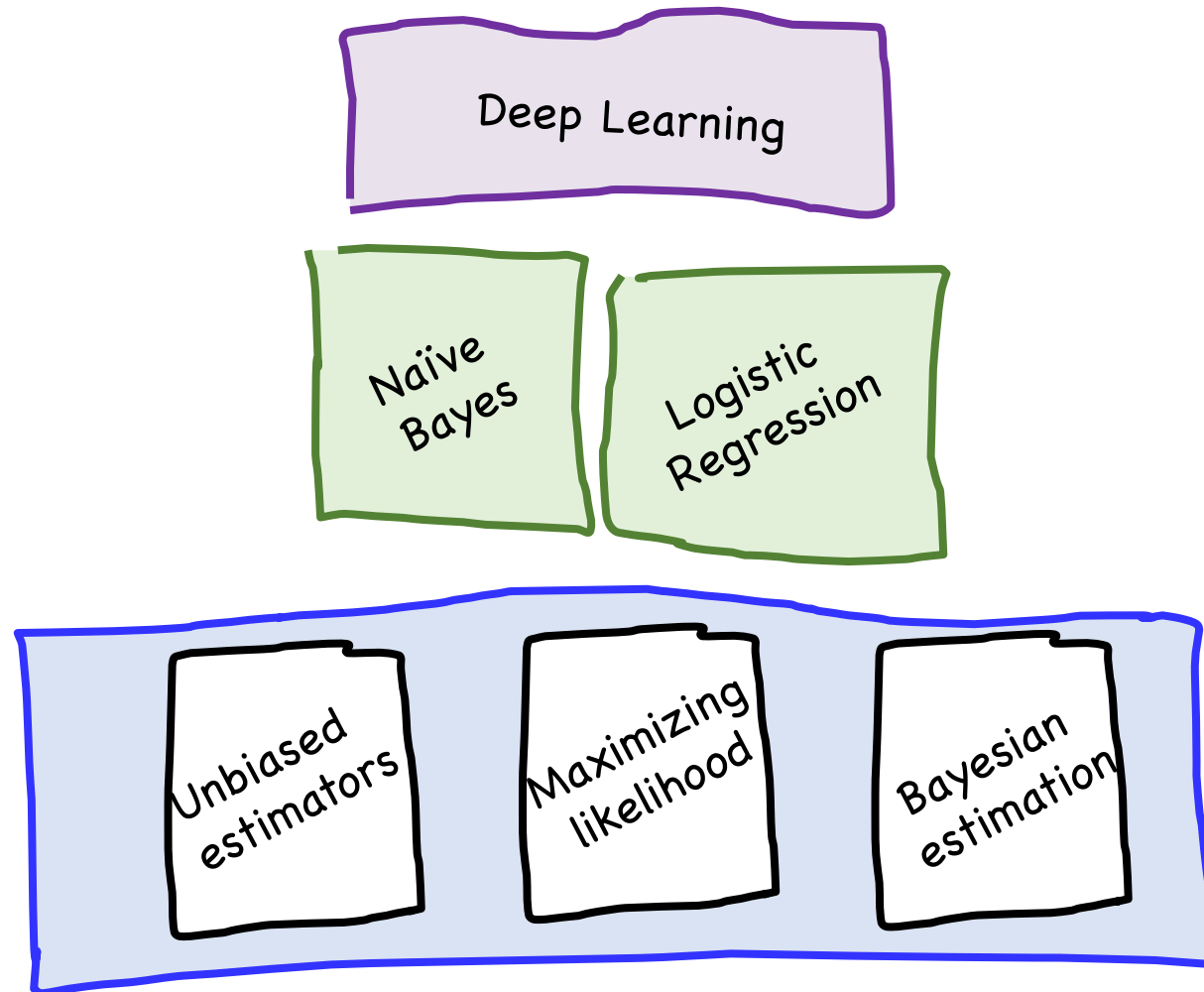
MACHINE LEARNING



At this point, if you are given a *model*,
with all the involved probabilities, you
can make predictions

But what if you want to *learn* the probabilities in the model?

Our Path



What are Parameters?

Consider some probability distributions:

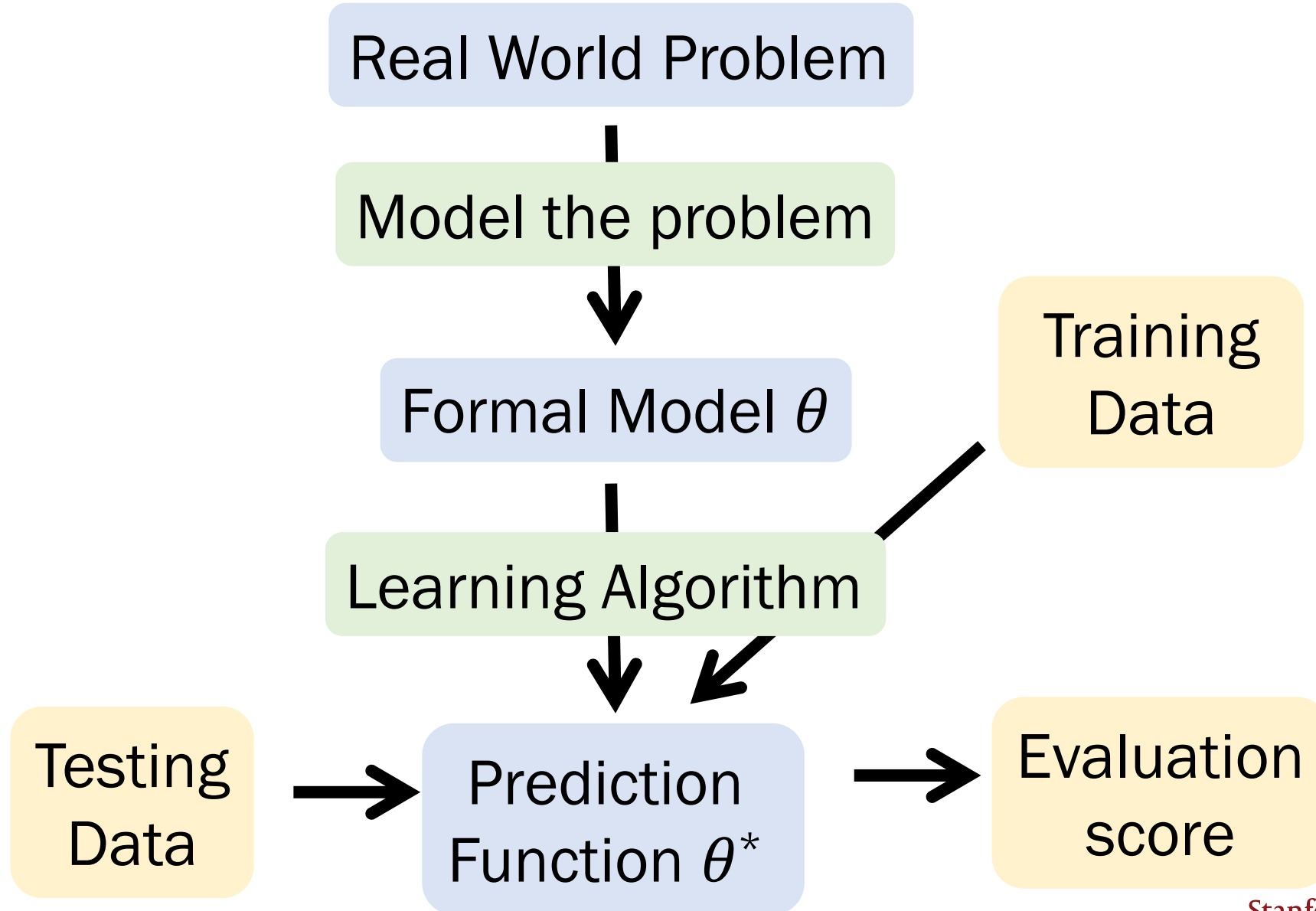
- $\text{Ber}(p)$ $\theta = p$
- $\text{Poi}(\lambda)$ $\theta = \lambda$
- $\text{Uni}(\alpha, \beta)$ $\theta = (\alpha, \beta)$
- $\text{Normal}(\mu, \sigma^2)$ $\theta = (\mu, \sigma^2)$
- $Y = mX + b$ $\theta = (m, b)$
- etc...

Call these “parametric models”

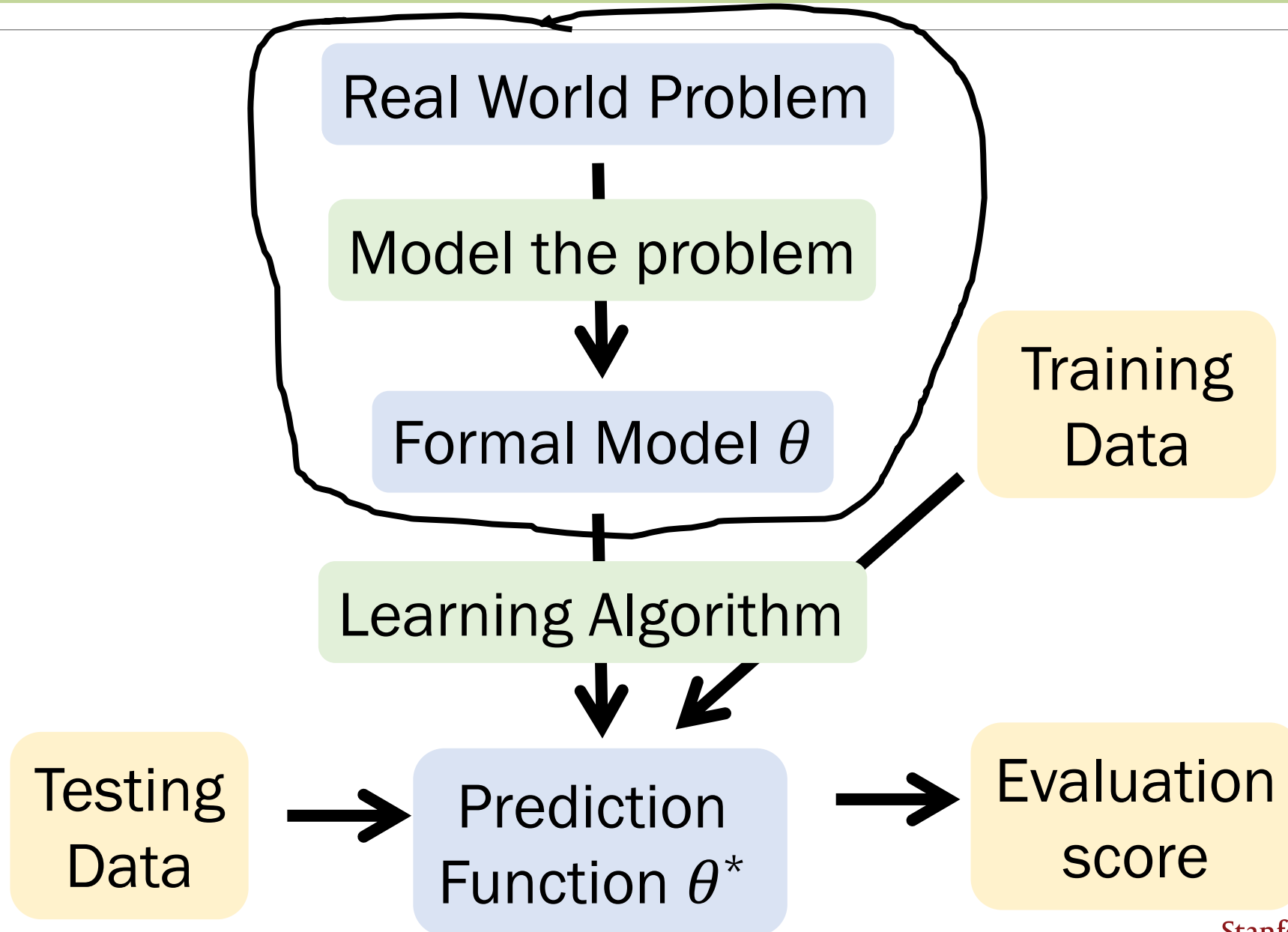
Given model, **parameters** yield actual distribution

- Usually refer to parameters of distribution as θ
- Note that θ that can be a vector of parameters

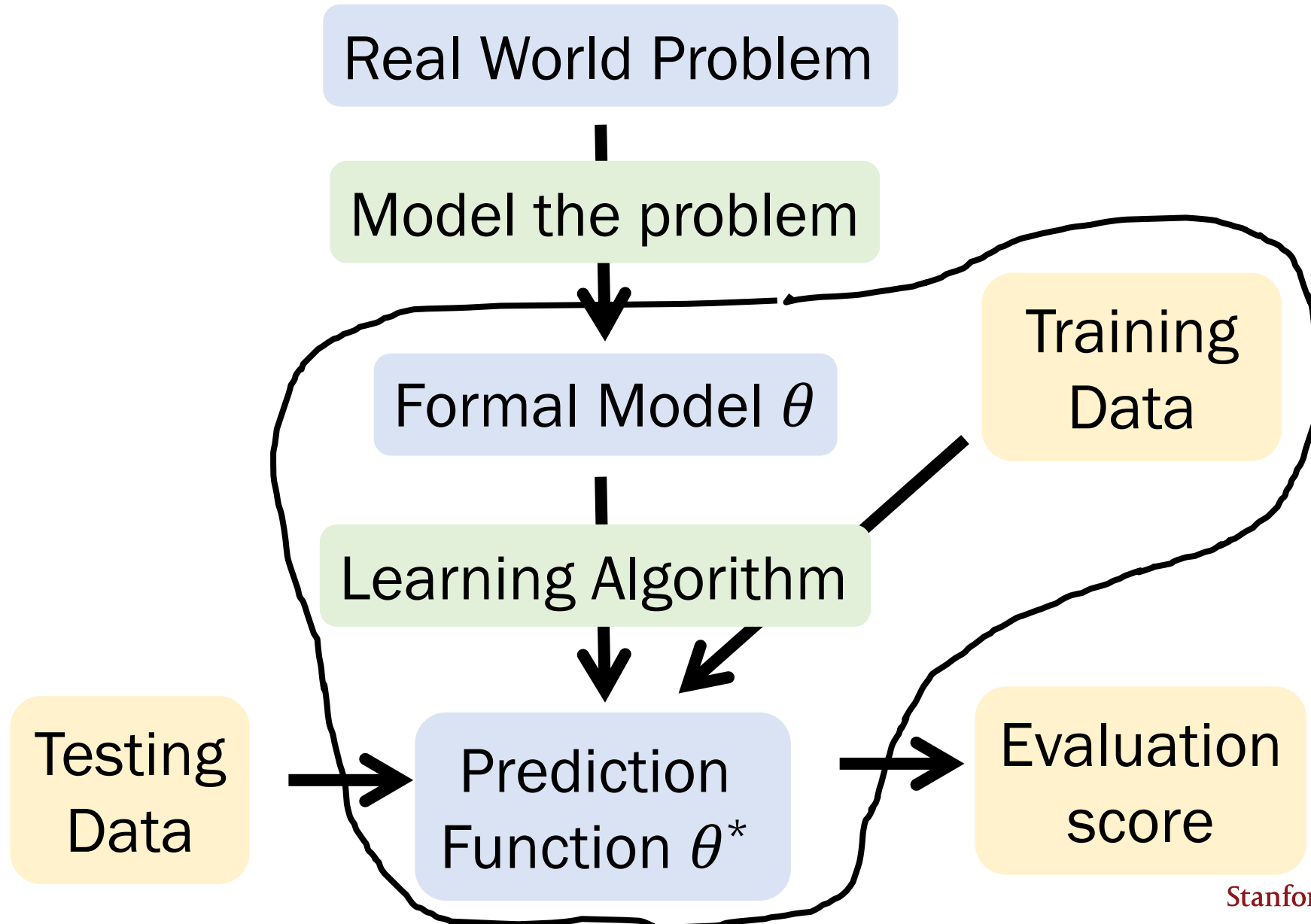
Why Do We Care?



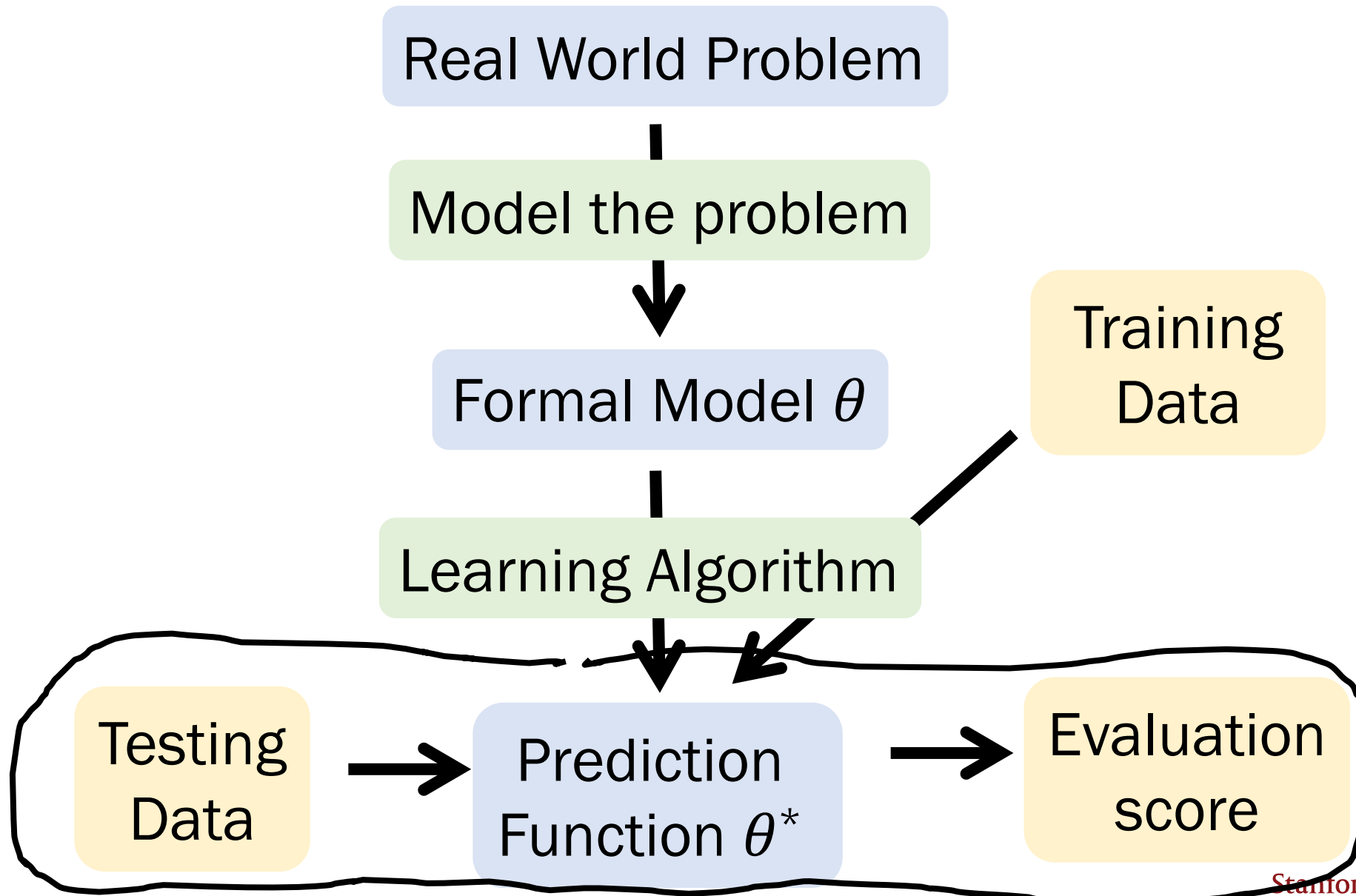
Modelling



Parameter Estimation (aka Training)



Testing



MLE Idea: Chose params that make the data look likely

Data = [6.3 , 5.5 , 5.4, 7.1, 4.6, 6.7, 5.3 , 4.8, 5.6, 3.4, 5.4, 3.4, 4.8, 7.9, 4.6, 7.0, 2.9, 6.4, 6.0 , 4.3]

Estimate the Parameters

Parameter μ :

Parameter σ :

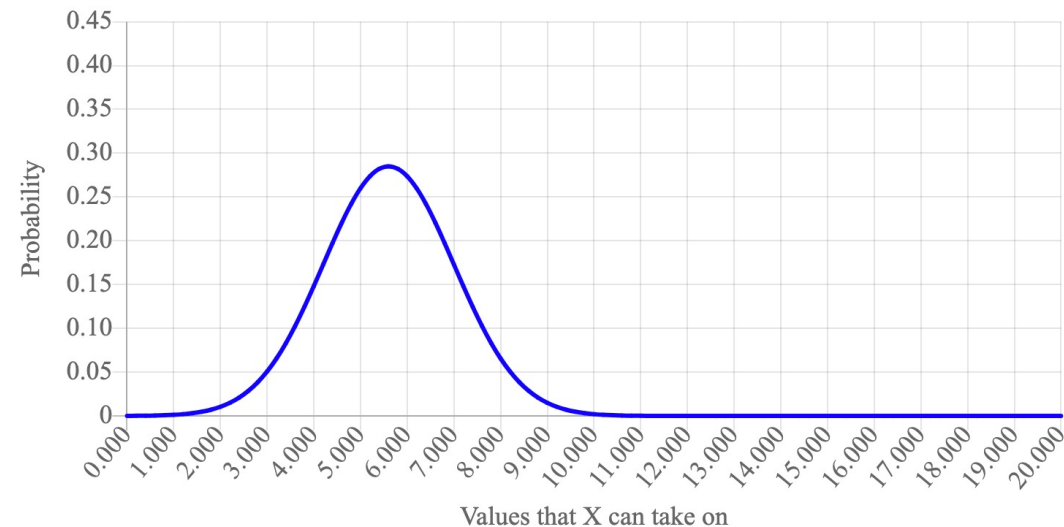
Likelihood

Likelihood: 1.9542923784106326e-15

Log Likelihood: -301.9

Best Seen: -301.9

PDF Graph





The Likelihood Function



n I.I.D. data points X_1, X_2, \dots, X_n

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

This is just a product since X_i are independent

We explicitly specify parameter θ of distribution

The Likelihood Function (full)



n I.I.D. data points X_1, X_2, \dots, X_n

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

This is just a product since X_i are independent

We explicitly specify parameter θ of distribution

The Likelihood Function (vectors)



n I.I.D. data points X_1, X_2, \dots, X_n

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

This is just a product since X_i are independent

We explicitly specify parameter θ of distribution

The Likelihood Function (vectors)



n I.I.D. data points $\vec{X}_1, \vec{X}_2, \dots, \vec{X}_n$

$$L(\vec{\theta}) = \prod_{i=1}^n f(\vec{X}_i = \vec{x}_i | \vec{\Theta} = \vec{\theta})$$

This is just a product since X_i are independent

We explicitly specify parameter $\vec{\theta}$ of distribution

Log Likelihood

$$LL(\theta) = \log L(\theta)$$

Computing the MLE

$$\theta_{MLE} = \arg \max_{\theta} LL(\theta)$$

General approach for finding θ_{MLE} , the MLE of θ :

1. Determine formula for $LL(\theta)$

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

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

$$\frac{\partial LL(\theta)}{\partial \theta}$$

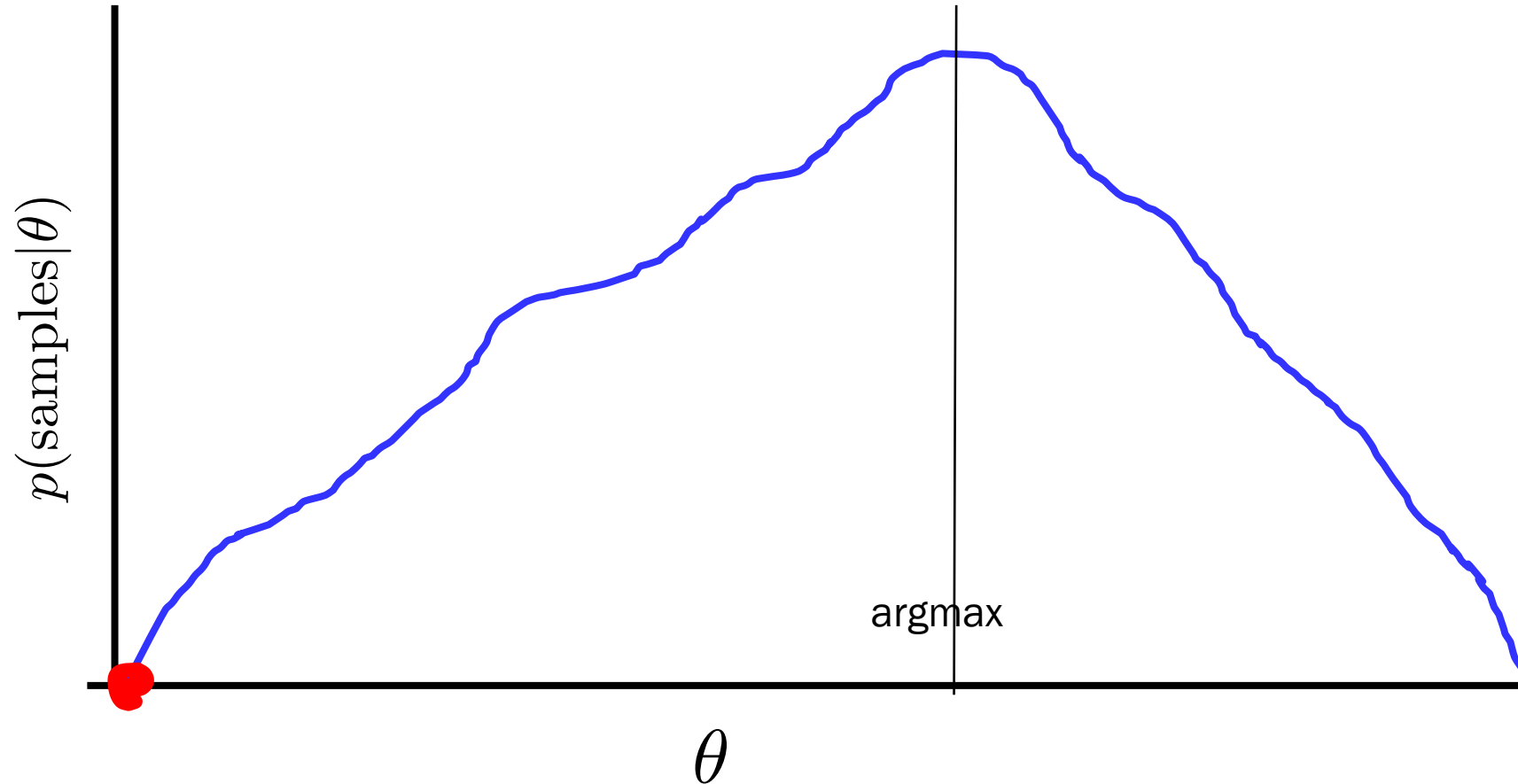
3. Solve

argmax

Estimate of theta

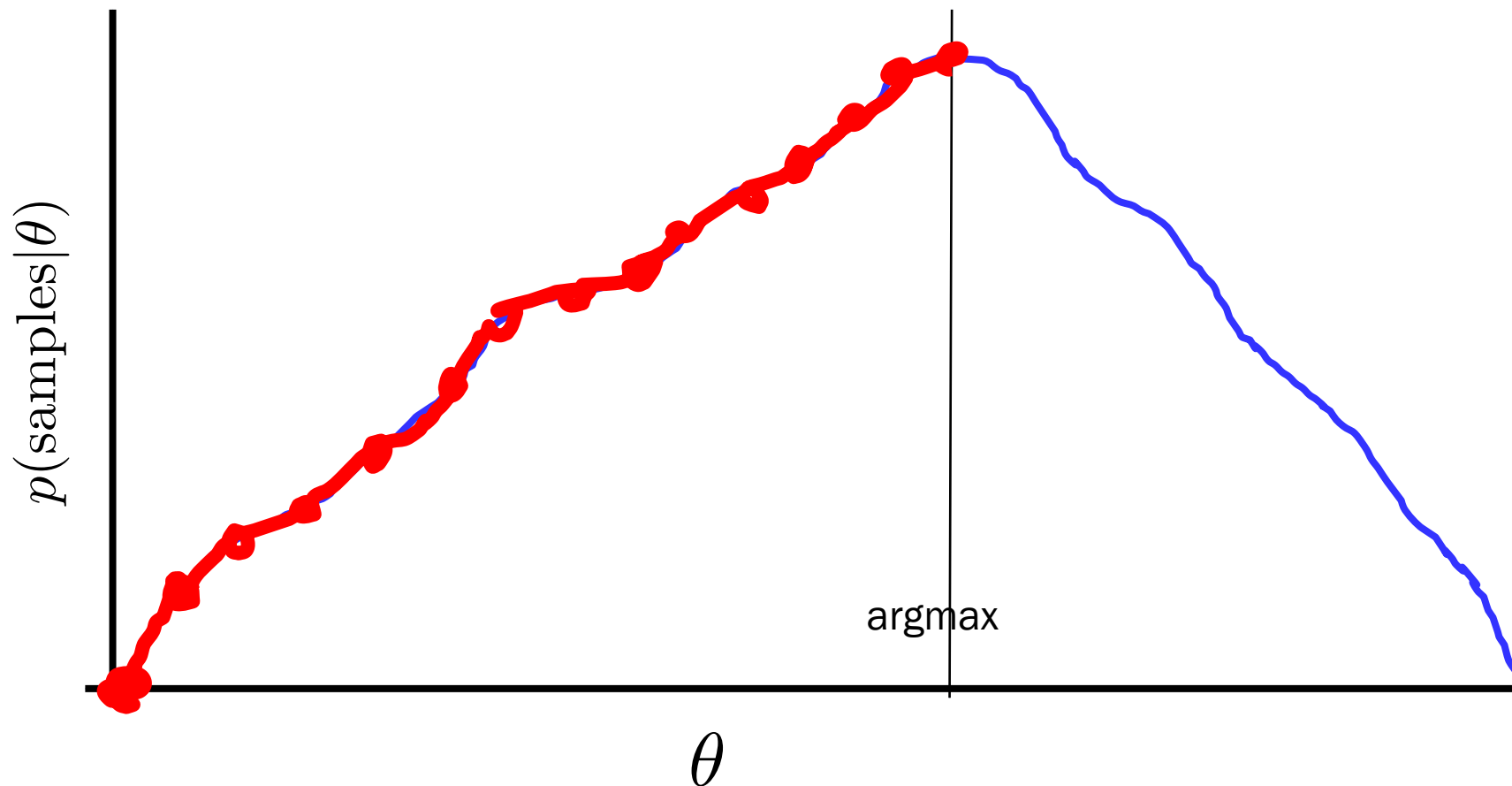
$LL(\theta)$ is often easier to differentiate than $L(\theta)$.

Gradient Ascent



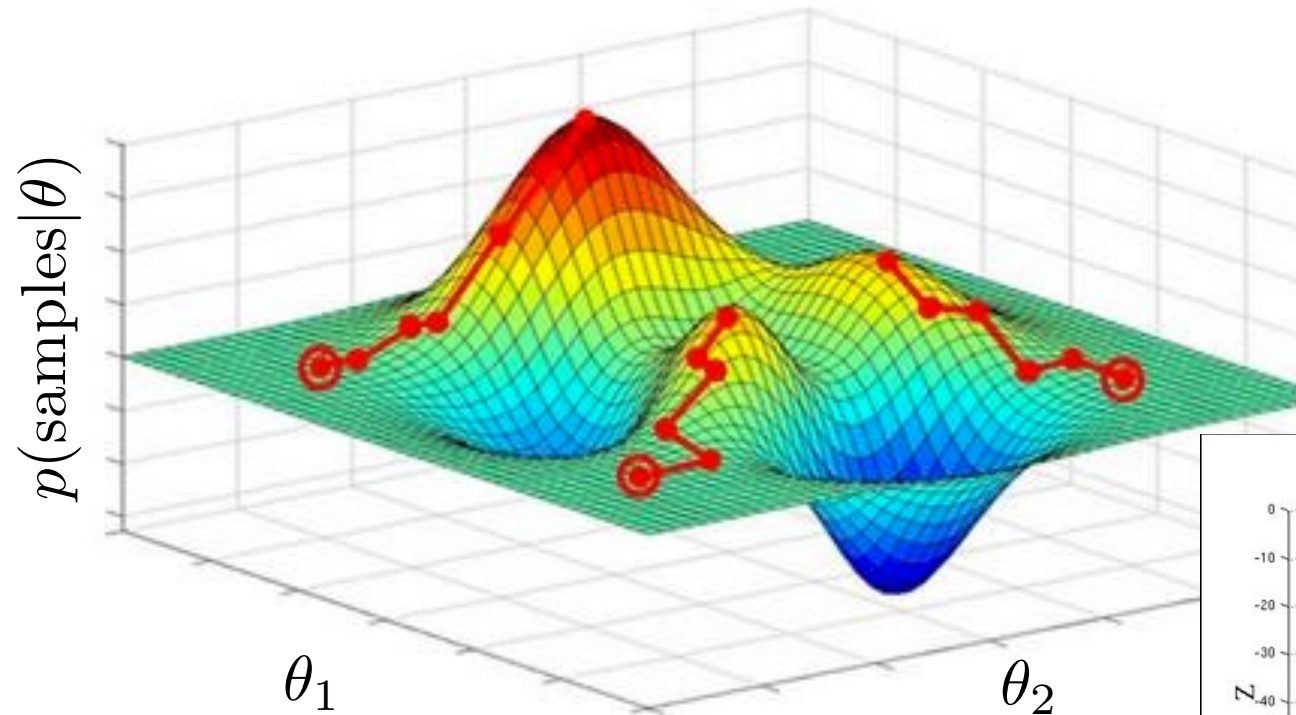
Walk uphill and you will find a local maxima
(if your step size is small enough)

Gradient Ascent

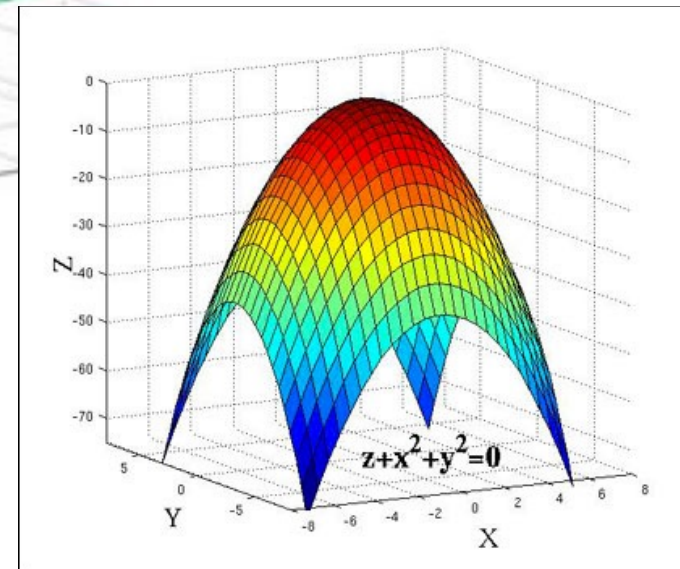


Walk uphill and you will find a local maxima
(if your step size is small enough)

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

Repeat many times

$$\theta_j^{\text{new}} = \theta_j^{\text{old}} + \eta \cdot \frac{\partial LL(\theta^{\text{old}})}{\partial \theta_j^{\text{old}}}$$

This is some **profound** life philosophy

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



Gradient **descent** is your bread and butter algorithm for optimization
(use argmax of neg LL)

<End Review>

Imagine you need to estimate λ for a Poisson

Data = [12, 10, 11, 10, ... , 12]

Assume I.I.D.

Assume Poisson with the same λ

What is your estimate for λ ? Use MLE!

Maximizing Likelihood with Poisson

- Consider I.I.D. dataset X_1, X_2, \dots, X_n
 - $X_i \sim \text{Poi}(\lambda)$. **Use Maximum Likelihood to estimate λ .**

1. What is the likelihood of one X_i

2. What is the likelihood of all the *data*

3. What is the log-likelihood all the *data*

4. Find the value of λ which maximizes log likelihood

Maximizing Likelihood with Poisson

- Consider I.I.D. dataset X_1, X_2, \dots, X_n
 - $X_i \sim \text{Poi}(\lambda)$. **Use Maximum Likelihood to estimate λ .**
 - Probability mass function: $f(X_i|\lambda) = \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}$

2. What is the likelihood of all the *data*

3. What is the log-likelihood all the *data*

4. Find the value of λ which maximizes log likelihood

Maximizing Likelihood with Poisson

- Consider I.I.D. dataset X_1, X_2, \dots, X_n
 - $X_i \sim \text{Poi}(\lambda)$. **Use Maximum Likelihood to estimate λ .**
 - Probability mass function: $f(X_i|\lambda) = \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}$
 - Likelihood:

$$L(\theta) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}$$

3. What is the log-likelihood all the *data*

4. Find the value of λ which maximizes log likelihood

Maximizing Likelihood with Poisson

- Consider I.I.D. dataset X_1, X_2, \dots, X_n
 - $X_i \sim \text{Poi}(\lambda)$. **Use Maximum Likelihood to estimate λ .**

- Probability mass function: $f(X_i|\lambda) = \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}$

- Likelihood:

$$L(\theta) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}$$

- Log-likelihood:

$$LL(\theta) = \sum_{i=1}^n \log\left(\frac{e^{-\lambda} \lambda^{X_i}}{X_i!}\right) = \sum_{i=1}^n (-\lambda \log e + X_i \log \lambda - \log X_i!) = -n\lambda + \log(\lambda) \sum_{i=1}^n X_i - \sum_{i=1}^n \log(X_i!)$$

4. Find the value of λ which maximizes log likelihood

Maximizing Likelihood with Poisson

- Consider I.I.D. dataset X_1, X_2, \dots, X_n
 - $X_i \sim \text{Poi}(\lambda)$. **Use Maximum Likelihood to estimate λ .**

- Probability mass function: $f(X_i|\lambda) = \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}$

- Likelihood:

$$L(\theta) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}$$

- Log-likelihood:

$$LL(\theta) = \sum_{i=1}^n \log\left(\frac{e^{-\lambda} \lambda^{X_i}}{X_i!}\right) = \sum_{i=1}^n (-\lambda \log e + X_i \log \lambda - \log X_i!) = -n\lambda + \log(\lambda) \sum_{i=1}^n X_i - \sum_{i=1}^n \log(X_i!)$$

- Differentiate λ , and argmax:

$$\frac{\partial LL(\theta)}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^n X_i \quad \rightarrow \quad \lambda_{MLE} = \frac{1}{n} \sum_{i=1}^n X_i$$

Imagine you need to estimate λ for a Poisson

- Data = [12, 10, 11, 10, ... , 12]
- Assume I.I.D.
- Assume Poisson with the same λ
- What is your estimate for λ ? Use MLE!

$$\lambda_{MLE} = \frac{1}{n} \sum_{i=1}^n X_i$$

Maximizing Likelihood with Bernoulli

- Consider I.I.D. random variables X_1, X_2, \dots, X_n
 - $X_i \sim \text{Ber}(p)$. **Use Maximum Likelihood to estimate p .**

1. What is the likelihood of one X_i

2. What is the likelihood of all the *data*

3. What is the log-likelihood all the *data*

4. Find the value of p which maximizes log likelihood

Maximizing Likelihood with Bernoulli

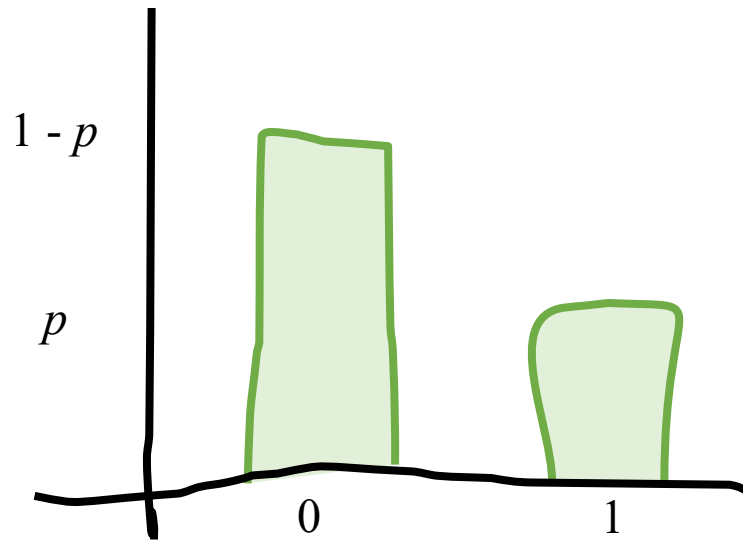
- Consider I.I.D. random variables X_1, X_2, \dots, X_n
 - $X_i \sim \text{Ber}(p)$
 - Probability mass function, $f(X_i | p)$:



Maximizing Likelihood with Bernoulli

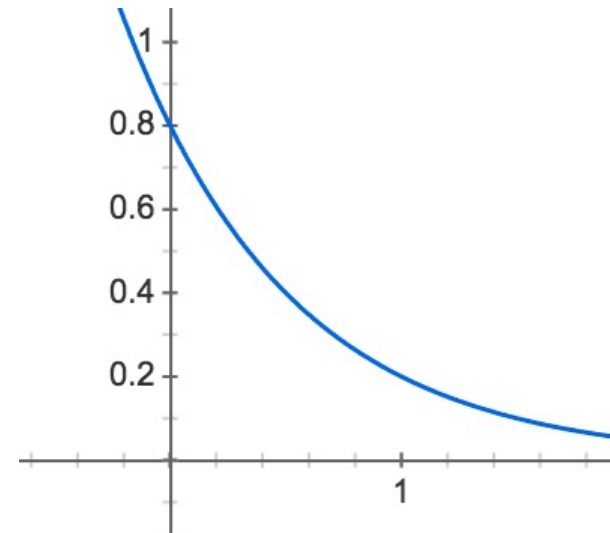
- Consider I.I.D. random variables X_1, X_2, \dots, X_n
 - $X_i \sim \text{Ber}(p)$
 - Probability mass function, $f(X_i | p)$:

PMF of Bernoulli



$$f(X_i | p) = p^{x_i} (1-p)^{1-x_i}$$

PMF of Bernoulli ($p = 0.2$)



$$f(x) = 0.2^x (1 - 0.2)^{1-x}$$

Bernoulli PMF

$$X \sim \text{Ber}(p)$$



$$f(X = x|p) = p^x (1 - p)^{1-x}$$

Maximizing Likelihood with Bernoulli

- Consider I.I.D. random variables X_1, X_2, \dots, X_n
 - $X_i \sim \text{Ber}(p)$. **Use Maximum Likelihood to estimate p .**
 - Probability mass function, $f(X_i | p)$, can be written as:

$$f(X_i | p) = p^{x_i} (1-p)^{1-x_i} \quad \text{where } x_i = 0 \text{ or } 1$$

2. What is the likelihood of all the *data*

3. What is the log-likelihood all the *data*

4. Find the value of p which maximizes log likelihood

Maximizing Likelihood with Bernoulli

- Consider I.I.D. random variables X_1, X_2, \dots, X_n
 - $X_i \sim \text{Ber}(p)$. **Use Maximum Likelihood to estimate p .**
 - Probability mass function, $f(X_i | p)$, can be written as:
$$f(X_i | p) = p^{x_i} (1-p)^{1-x_i} \quad \text{where } x_i = 0 \text{ or } 1$$
 - Likelihood: $L(\theta) = \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i}$

3. What is the log-likelihood all the *data*

4. Find the value of p which maximizes log likelihood

Maximizing Likelihood with Bernoulli

- Consider I.I.D. random variables X_1, X_2, \dots, X_n
 - $X_i \sim \text{Ber}(p)$. **Use Maximum Likelihood to estimate p .**
 - Probability mass function, $f(X_i | p)$, can be written as:

$$f(X_i | p) = p^{x_i} (1-p)^{1-x_i} \quad \text{where } x_i = 0 \text{ or } 1$$

- Likelihood: $L(\theta) = \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i}$
- Log-likelihood:

$$LL(\theta) = \sum_{i=1}^n \log(p^{X_i} (1-p)^{1-X_i}) = \sum_{i=1}^n [X_i (\log p) + (1-X_i) \log(1-p)]$$

4. Find the value of p which maximizes log likelihood

Maximizing Likelihood with Bernoulli

- Consider I.I.D. random variables X_1, X_2, \dots, X_n
 - $X_i \sim \text{Ber}(p)$. **Use Maximum Likelihood to estimate p .**

- Probability mass function, $f(X_i | p)$, can be written as:

$$f(X_i | p) = p^{x_i} (1-p)^{1-x_i} \quad \text{where } x_i = 0 \text{ or } 1$$

- Likelihood: $L(\theta) = \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i}$

- Log-likelihood:

$$LL(\theta) = \sum_{i=1}^n \log(p^{X_i} (1-p)^{1-X_i}) = \sum_{i=1}^n [X_i (\log p) + (1-X_i) \log(1-p)]$$

- Differentiate w.r.t. p , and set to 0:

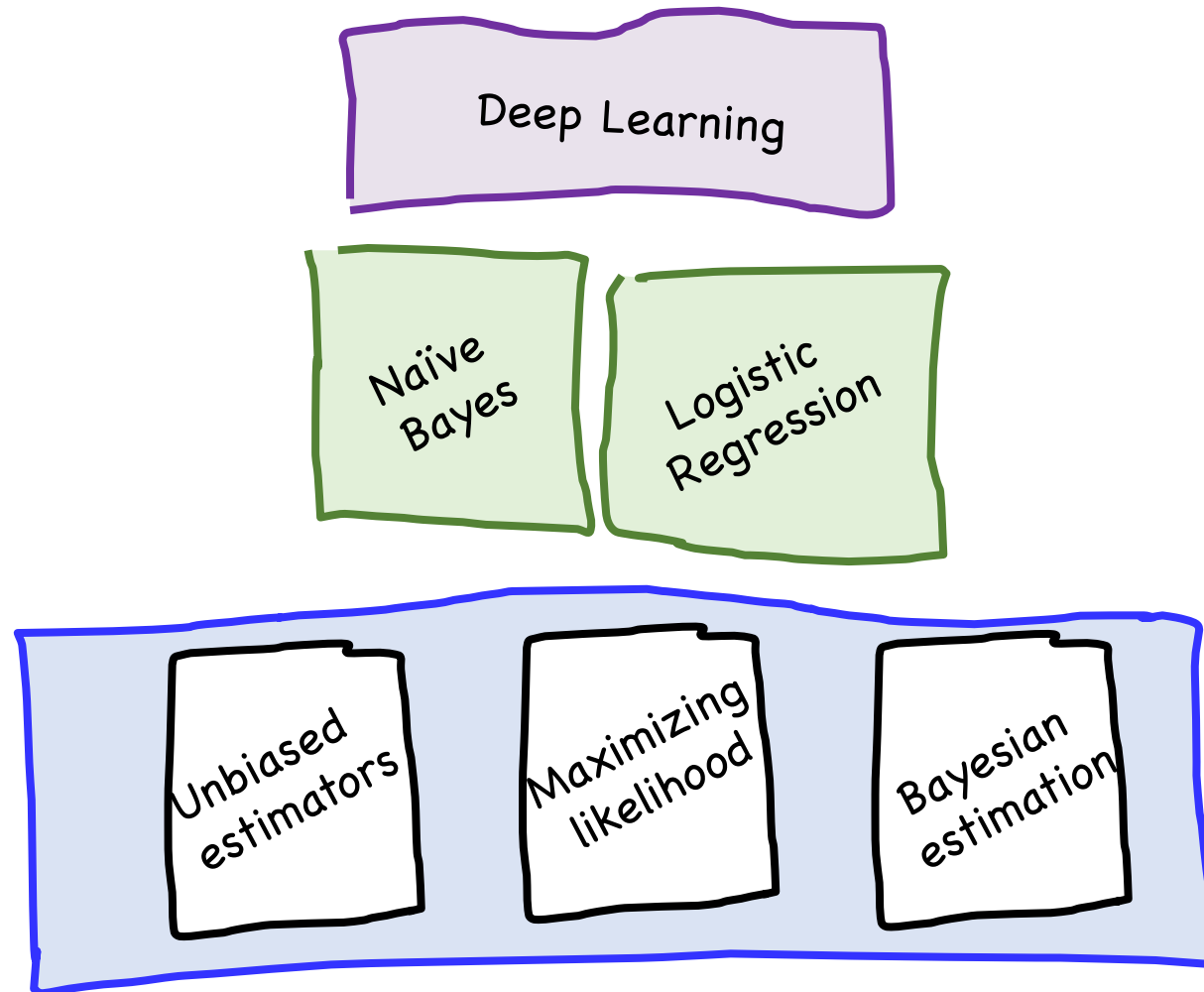
$$= Y(\log p) + (n-Y) \log(1-p) \quad \text{where } Y = \sum_{i=1}^n X_i$$

$$\frac{\partial LL(p)}{\partial p} = Y \frac{1}{p} + (n-Y) \frac{-1}{1-p} = 0 \quad \Rightarrow \quad p_{MLE} = \frac{Y}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

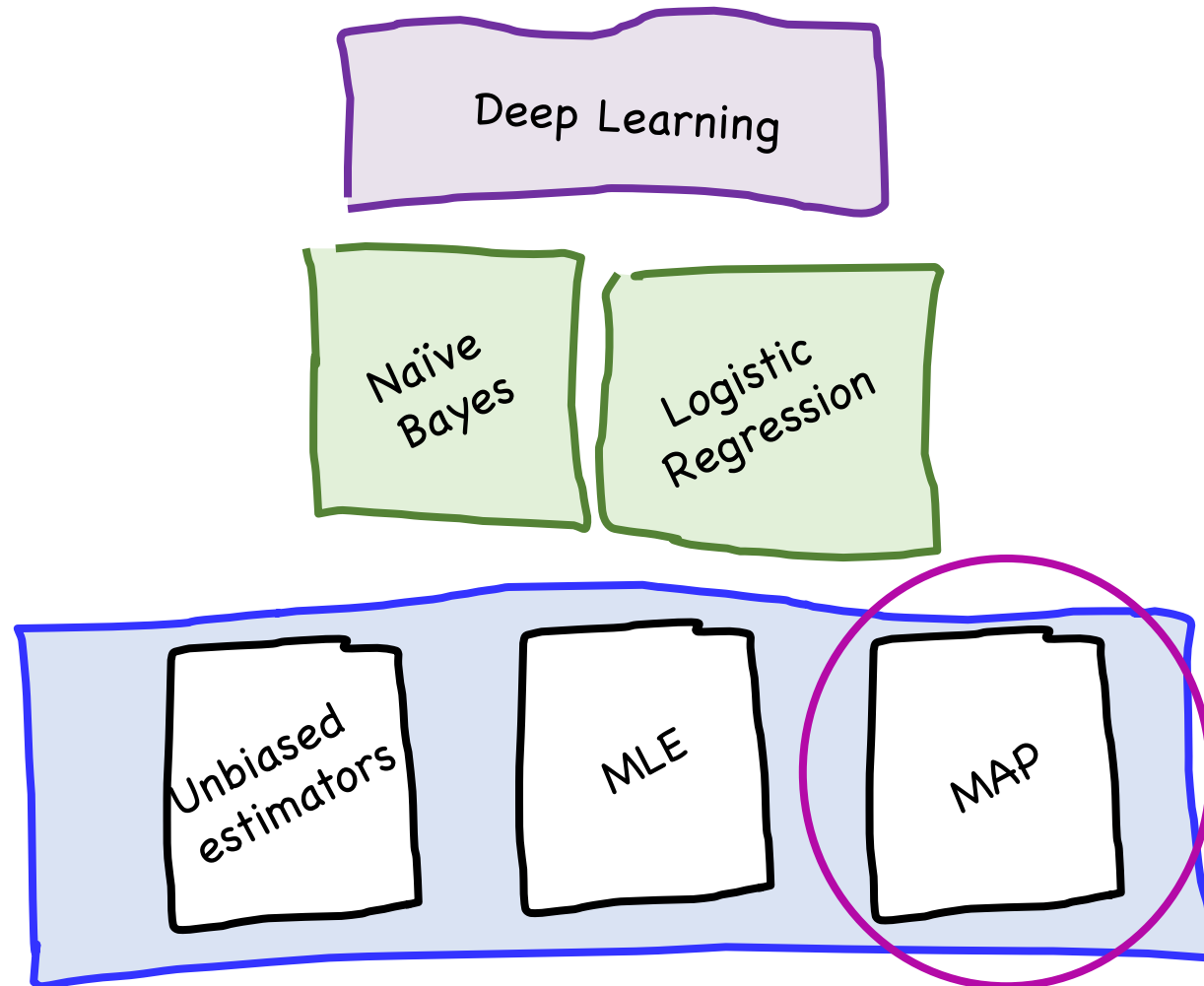
Machine Learning:
Learn parameters (mostly with MLE) for
probabilistic models.

Optimization (argmax)
Option #2: Gradient Descent

Our Path



Our Path



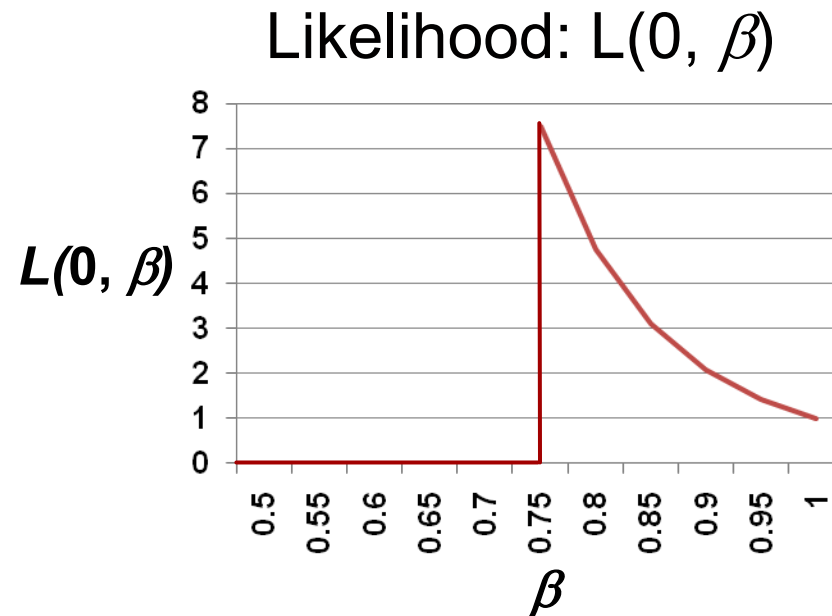
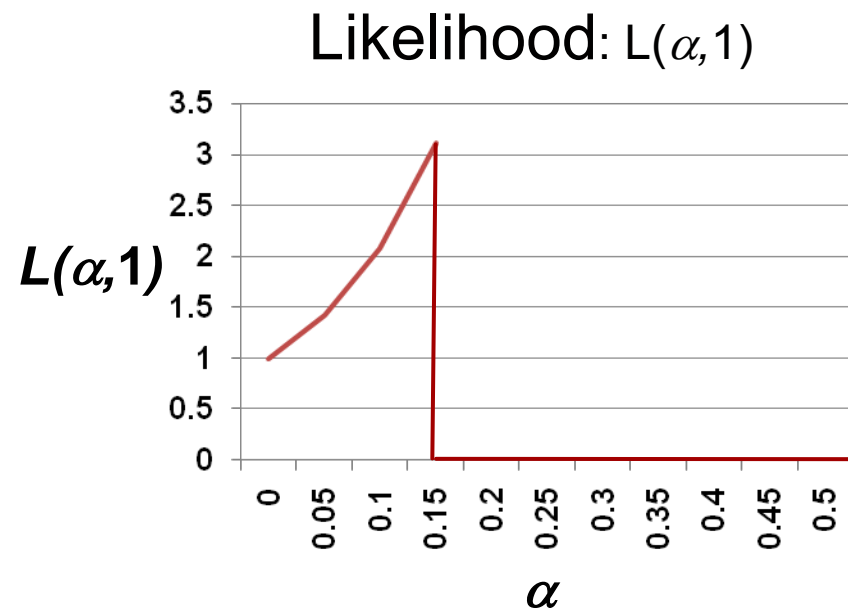
Pedagogic Pause

Something rotten
in the world of MLE

MLE Could Benefit from Priors

Consider I.I.D. random variables X_1, X_2, \dots, X_n

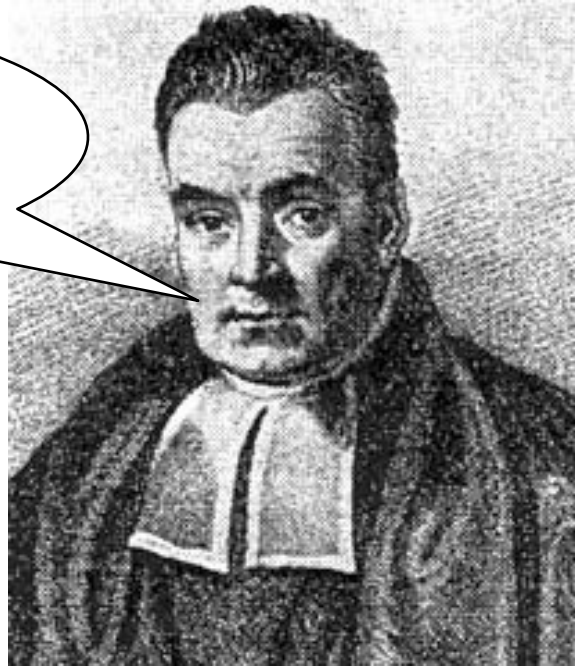
- $X_i \sim \text{Uni}(0, 1)$
- Observe data:
 - 0.15, 0.20, 0.30, 0.40, 0.65, 0.70, 0.75



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 $Y = \$$ 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

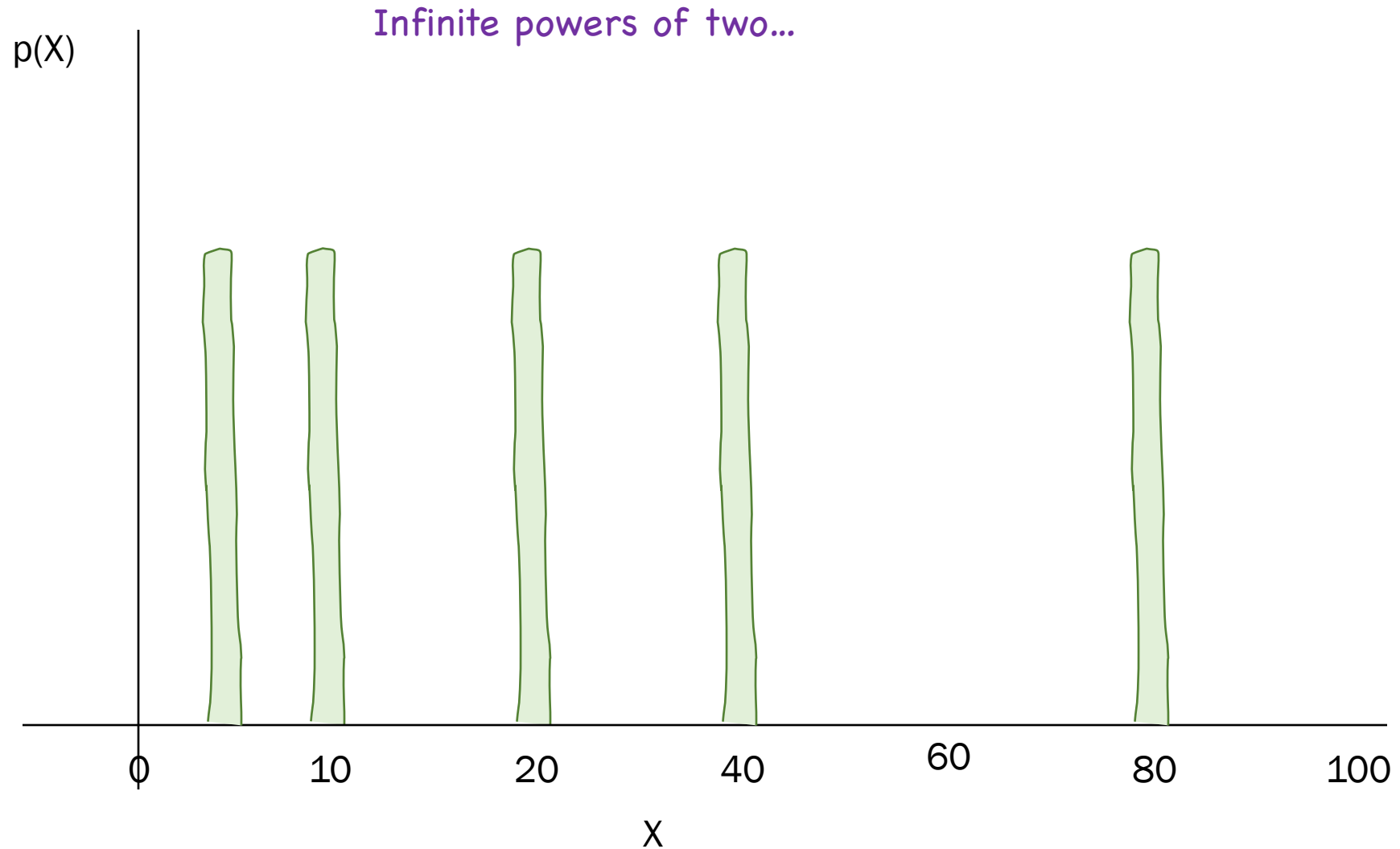
The “two envelopes” problem set-up

- 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?

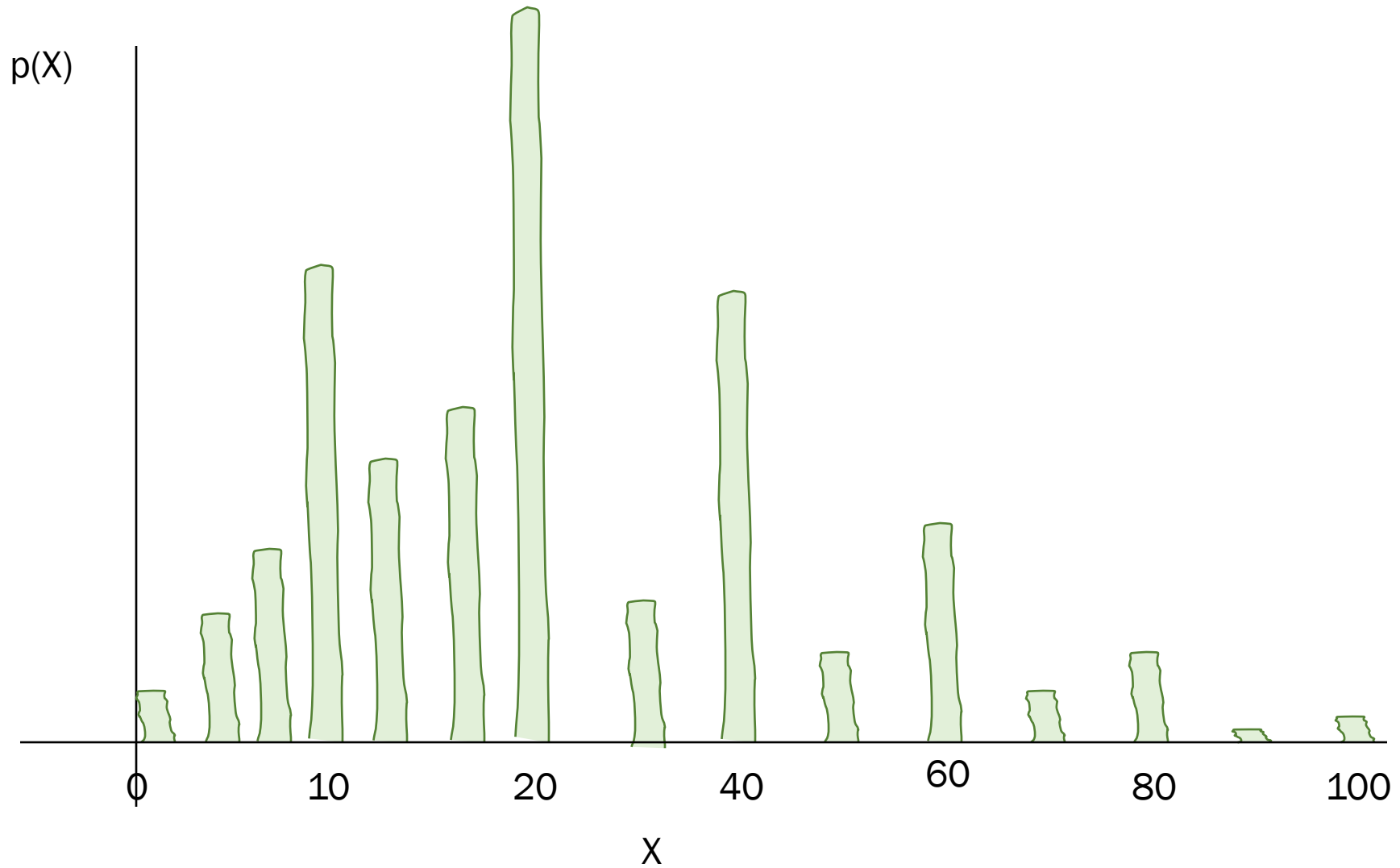


Subjectivity of Probability

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

Envelope Summary:
Probabilities are beliefs.
Incorporating prior beliefs is useful

Priors for Parameter Estimation?

New Philosophy: Chose the most likely parameters given the data:

MLE vs MAP

Data: $x^{(1)}, \dots, x^{(n)}$

Maximum Likelihood Estimation

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta} f(X^{(1)} = x^{(1)}, \dots, X^{(n)} = x^{(n)} | \theta)$$

Maximum A Posteriori

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} f(\Theta = \theta | X^{(1)} = x^{(1)}, \dots, X^{(n)} = x^{(n)})$$

MLE vs MAP

Data: $x^{(1)}, \dots, x^{(n)}$

Maximum Likelihood Estimation

$$\begin{aligned}\hat{\theta}_{MLE} &= \operatorname{argmax}_{\theta} f(X^{(1)} = x^{(1)}, \dots, X^{(n)} = x^{(n)} | \theta) \\ &= \operatorname{argmax}_{\theta} \left(\sum_i \log f(X^{(i)} = x^{(i)} | \theta) \right)\end{aligned}$$

Maximum A Posteriori

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} f(\Theta = \theta | X^{(1)} = x^{(1)}, \dots, X^{(n)} = x^{(n)})$$

Notation Shorthand

MAP, without shorthand

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} f(\Theta = \theta | X^{(1)} = x^{(1)}, \dots, X^{(n)} = x^{(n)})$$

Our shorthand notation

θ is shorthand for the event: $\Theta = \theta$

$x^{(i)}$ is shorthand for the event: $X^{(i)} = x^{(i)}$

MAP, now with shorthand

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} f(\theta | x^{(1)}, \dots, x^{(n)})$$

MLE vs MAP

Data: $x^{(1)}, \dots, x^{(n)}$

Maximum Likelihood Estimation

$$\begin{aligned}\hat{\theta}_{MLE} &= \operatorname{argmax}_{\theta} f(x^{(1)}, \dots, x^{(n)} | \theta) \\ &= \operatorname{argmax}_{\theta} \left(\sum_i \log f(x^{(i)} | \theta) \right)\end{aligned}$$

Maximum A Posteriori

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} f(\theta | x^{(1)}, \dots, x^{(n)})$$

Most important slide of today

Maximum A Posteriori


data: $x^{(1)}, \dots, x^{(n)}$ $\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} f(\theta|x^{(1)}, \dots, x^{(n)})$

likelihood

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} \frac{f(x^{(1)}, x^{(2)}, \dots, x^{(n)}|\theta)g(\theta)}{h(x^{(1)}, x^{(2)}, \dots, x^{(n)})}$$

posterior

prior



Maximum A Posteriori

data: $x^{(1)}, \dots, x^{(n)}$ $\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} f(\theta | x^{(1)}, \dots, x^{(n)})$

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} \frac{g(\theta) f(x^{(1)}, x^{(2)}, \dots, x^{(n)} | \theta)}{h(x^{(1)}, x^{(2)}, \dots, x^{(n)})}$$

$$= \operatorname{argmax}_{\theta} \frac{g(\theta) \prod_{i=1}^n f(x^{(i)} | \theta)}{h(x^{(1)}, x^{(2)}, \dots, x^{(n)})}$$

$$= \operatorname{argmax}_{\theta} g(\theta) \prod_{i=1}^n f(x^{(i)} | \theta)$$

$$= \operatorname{argmax}_{\theta} \left(\log(g(\theta)) + \sum_{i=1}^n \log(f(x^{(i)} | \theta)) \right)$$



Maximum A Posteriori



Estimated
parameter

Log prior

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

Chose the value of theta
that maximizes:

Sum of
log likelihood

MLE vs MAP

Data: $x^{(1)}, \dots, x^{(n)}$

Maximum Likelihood Estimation

$$\begin{aligned}\hat{\theta}_{MLE} &= \operatorname{argmax}_{\theta} f(x^{(1)}, \dots, x^{(n)} | \theta) \\ &= \operatorname{argmax}_{\theta} \left(\sum_i \log f(x^{(i)} | \theta) \right)\end{aligned}$$

Maximum A Posteriori

$$\begin{aligned}\hat{\theta}_{MAP} &= \operatorname{argmax}_{\theta} f(\theta | x^{(1)}, \dots, x^{(n)}) \\ &= \operatorname{argmax}_{\theta} \left(\log(g(\theta)) + \sum_{i=1}^n \log(f(x^{(i)} | \theta)) \right)\end{aligned}$$

Gotta get that intuition

$P(\theta | D)$ For Bernoulli

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

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} f(\theta|\text{data}) = \underset{\theta}{\operatorname{argmax}} f(\text{data}|\theta)g(\theta)$$

This is the beta PDF

$$= \underset{\theta}{\operatorname{argmax}} \log g(\theta) + \log f(\text{data}|\theta)$$

This is ???

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

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

$$\begin{aligned}\hat{\theta}_{MAP} &= \underset{\theta}{\operatorname{argmax}} f(\theta \mid \text{data}) &&= \underset{\theta}{\operatorname{argmax}} f(\text{data} \mid \theta)g(\theta) \\ &\text{This is the beta PDF} \swarrow && \\ &= \underset{\theta}{\operatorname{argmax}} \log g(\theta) + \log f(\text{data} \mid \theta) && \nwarrow \text{Product of thetas and (1-theta)s} \\ &= \underset{\theta}{\operatorname{argmax}} \log \left[\frac{1}{\beta} \theta^{a-1} (1-\theta)^{b-1} \right] \\ &\quad + n \log f(\text{heads} \mid \theta) \\ &\quad + m \log f(\text{tails} \mid \theta) \\ &= \underset{\theta}{\operatorname{argmax}} \log \frac{1}{\beta} + (a-1) \log \theta + (b-1) \log(1-\theta) + n \log \theta + m \log(1-\theta) \\ &= \underset{\theta}{\operatorname{argmax}} (a-1+n) \log \theta + (b-1+m) \log(1-\theta)\end{aligned}$$

$P(\theta | D)$ For Bernoulli

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

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} f(\theta | \text{data})$$

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

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

That's the mode of the updated beta

Beta(a, b) is a conjugate prior for the probability of success in Bernoulli and Binomial distributions.

$$f(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}$$

Prior

Beta(a, b)

Saw $a + b - 2$ imaginary trials: $a - 1$ successes, $b - 1$ failures

Experiment

Observe $n + m$ new trials: n successes, m failures

Posterior

Beta($a + n, b + m$)

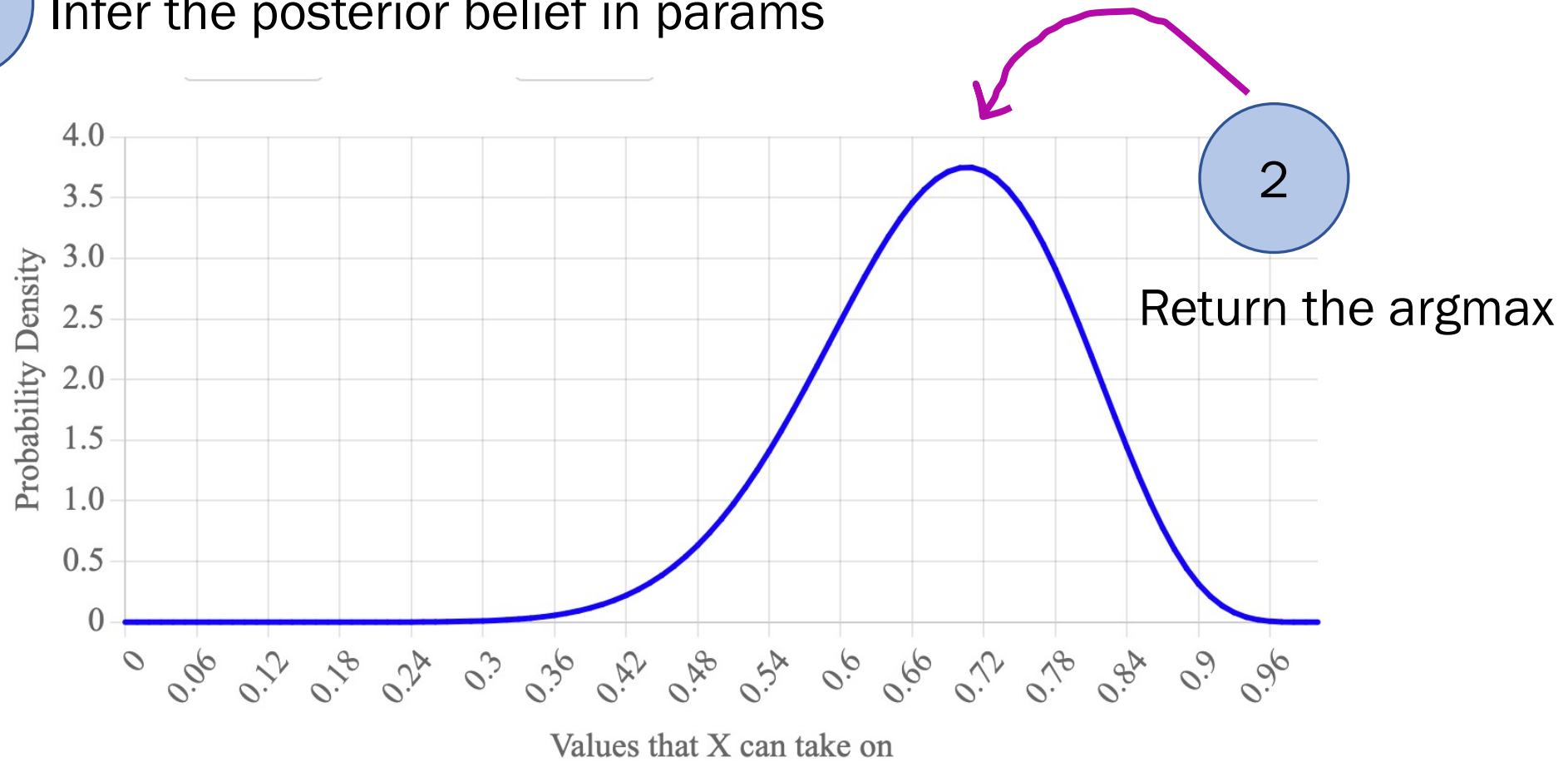
MAP:

$$p = \frac{a + n - 1}{a + b + n + m - 2}$$

$P(\theta | D)$ For Bernoulli

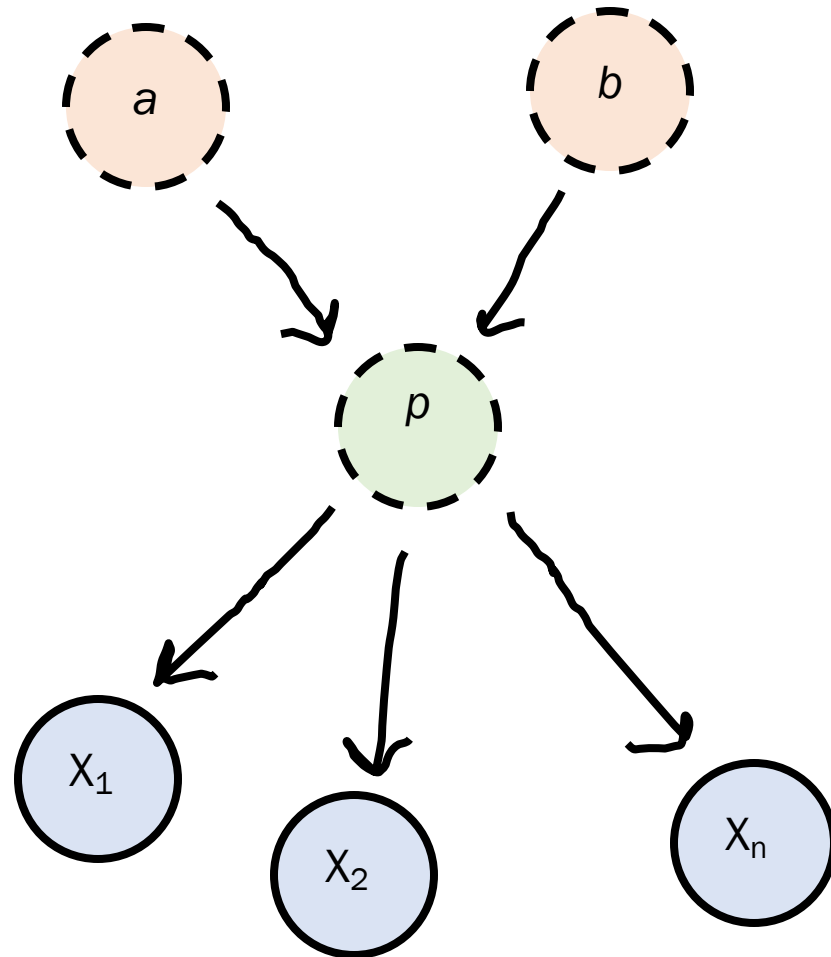
1

Infer the posterior belief in params



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

Conjugate distributions

MAP
estimator:

$$\theta_{MAP} = \arg \max_{\theta} f(\theta | X_1, X_2, \dots, X_n)$$

The **mode** of the
posterior distribution of θ

Distribution parameter	Conjugate distribution
Bernoulli p	Beta
Binomial p	Beta
Multinomial p_i	Dirichlet
Poisson λ	Gamma
Exponential λ	Gamma
Normal μ	Normal
Normal σ^2	Inverse Gamma

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

CS109: We'll only focus on MAP for
Bernoulli/Binomial p , Multinomial p_i , and Poisson λ .

Multinomial is Multiple times the fun

Dirichlet(a_1, a_2, \dots, a_m) is a conjugate for Multinomial.

- Generalizes Beta in the same way Multinomial generalizes Bernoulli/Binomial:

$$f(x_1, x_2, \dots, x_m) = \frac{1}{B(a_1, a_2, \dots, a_m)} \prod_{i=1}^m x_i^{a_i-1}$$

Prior

Dirichlet(a_1, a_2, \dots, a_m)

Saw $(\sum_{i=1}^m a_i) - m$ imaginary trials, with $a_i - 1$ of outcome i

Experiment

Observe $n_1 + n_2 + \dots + n_m$ new trials, with n_i of outcome i

Posterior

Dirichlet($a_1 + n_1, a_2 + n_2, \dots, a_m + n_m$)

MAP:

$$p_i = \frac{a_i + n_i - 1}{(\sum_{i=1}^m a_i) + (\sum_{i=1}^m n_i) - m}$$

Your Happy Laplace

Laplace gives a classic Dirichlet

Prior Dirichlet($a_1 = 2, a_2 = 2, \dots, a_m = 2$)
Saw m imaginary trials, with 1 of outcome i

Experiment Observe $n_1 + n_2 + \dots + n_m$ new trials, with n_i of outcome i

Posterior Dirichlet($2 + n_1, 2 + n_2, \dots, 2 + n_m$)

MAP:

$$p_i = \frac{n_i + 1}{\left(\sum_{i=1}^m n_i\right) + m}$$

Back to our happy Laplace

Consider our previous 6-sided die.

- Roll the dice $n = 12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

Recall θ_{MLE} : $p_1 = 3/12, p_2 = 2/12, p_3 = 0/12, \triangle!$
 $p_4 = 3/12, p_5 = 1/12, p_6 = 3/12$

What are your Laplace estimates for each roll outcome?



Back to our happy Laplace

Consider our previous 6-sided die.

- Roll the dice $n = 12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

Recall θ_{MLE} : $p_1 = 3/12, p_2 = 2/12, p_3 = 0/12, \triangle!$
 $p_4 = 3/12, p_5 = 1/12, p_6 = 3/12$

What are your Laplace estimates for each roll outcome?

$$p_i = \frac{n_i + 1}{(\sum_{i=1}^m n_i) + m}$$

$p_1 = 4/18, p_2 = 3/18, p_3 = 1/18, \checkmark$
 $p_4 = 4/18, p_5 = 2/18, p_6 = 4/18$

Laplace smoothing:

- Easy to implement/remember
- Avoids estimating a parameter of 0

Good times with Gamma

Gamma(α, β) is a conjugate for Poisson.

- Also conjugate for Exponential, but we won't delve into that
- Mode of gamma: $(\alpha - 1)/\beta$

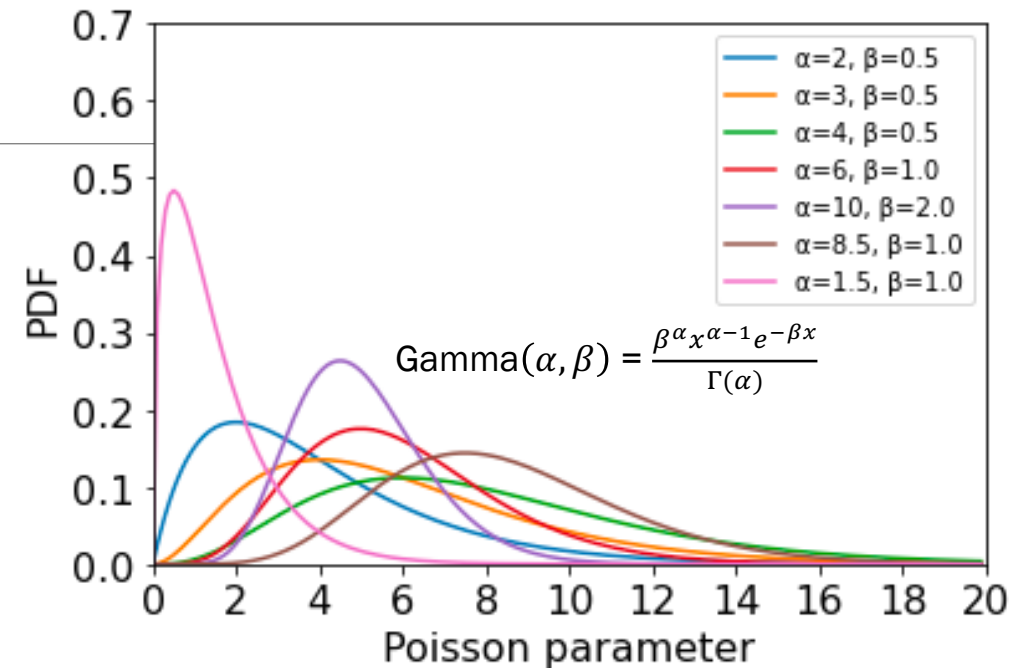
Prior $\theta \sim \text{Gamma}(\alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$

Saw $\alpha - 1$ total imaginary events during β prior time periods

Experiment Observe n events during next k time periods

Posterior $(\theta | n \text{ events in } k \text{ periods}) \sim \text{Gamma}(\alpha + n, \beta + k)$

MAP: $\theta_{MAP} = \frac{\alpha + n - 1}{\beta + k}$



MAP for Poisson

Gamma(α, β)
is conjugate for Poisson Mode: $\frac{\alpha-1}{\beta}$

Let λ be the average # of successes in a time period.

1. What does it mean to have a prior of $\theta \sim \text{Gamma}(11, 5)$?

Observe 10 imaginary events in 5 time periods, i.e., observe at Poisson rate = 2

Now perform the experiment and see 11 events in next 2 time periods.

2. Given your prior, what is the posterior distribution?
3. What is θ_{MAP} ?



MAP for Poisson

Gamma(α, β)
is conjugate for Poisson Mode: $\frac{\alpha-1}{\beta}$

Let λ be the average # of successes in a time period.

1. What does it mean to have a prior of $\theta \sim \text{Gamma}(11, 5)$?

Observe 10 imaginary events in 5 time periods, i.e., observe at Poisson rate = 2

Now perform the experiment and see 11 events in next 2 time periods.

2. Given your prior, what is the posterior distribution?

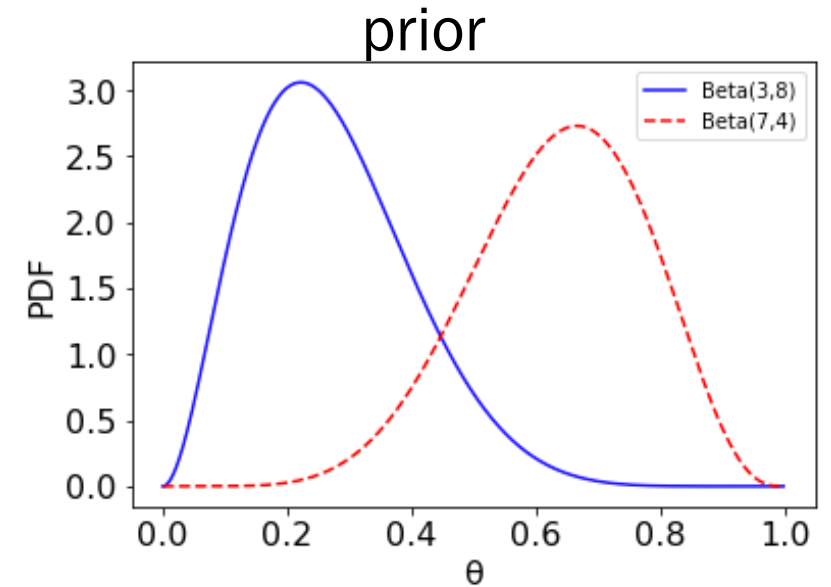
$(\theta | n \text{ events in } k \text{ periods}) \sim \text{Gamma}(22, 7)$

3. What is θ_{MAP} ?

$\theta_{MAP} = 3$, the updated Poisson rate

Where'd you get them priors?

- Let θ be the probability a coin turns up heads.
- Model θ with 2 different priors:
 - Prior 1: **Beta(3,8)**: 2 imaginary heads, 7 imaginary tails mode: $\frac{2}{9}$
 - Prior 2: **Beta(7,4)**: 6 imaginary heads, 3 imaginary tails mode: $\frac{6}{9}$



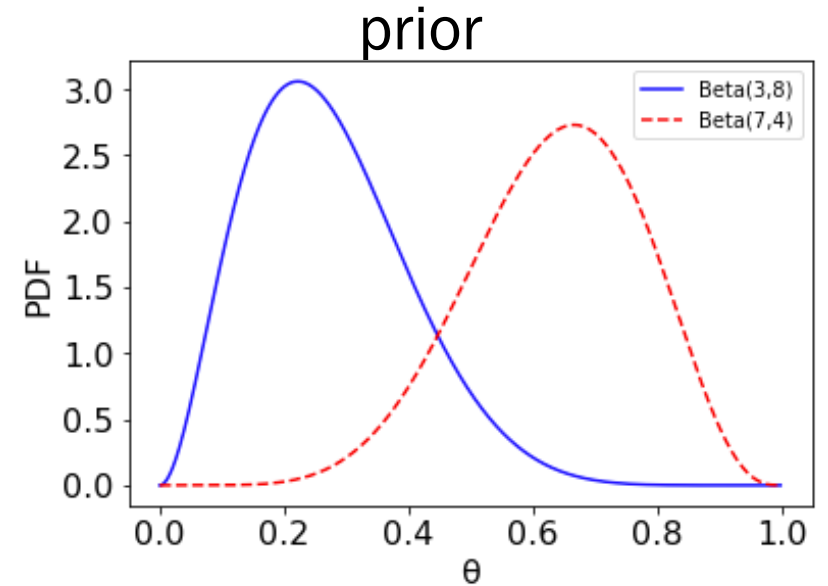
Now flip 100 coins and get 58 heads and 42 tails.

1. What are the two posterior distributions?
2. What are the modes of the two posterior distributions?



Where'd you get them priors?

- Let θ be the probability a coin turns up heads.
- Model θ with 2 different priors:
 - Prior 1: **Beta(3,8)**: 2 imaginary heads, 7 imaginary tails mode: $\frac{2}{9}$
 - Prior 2: **Beta(7,4)**: 6 imaginary heads, 3 imaginary tails mode: $\frac{6}{9}$

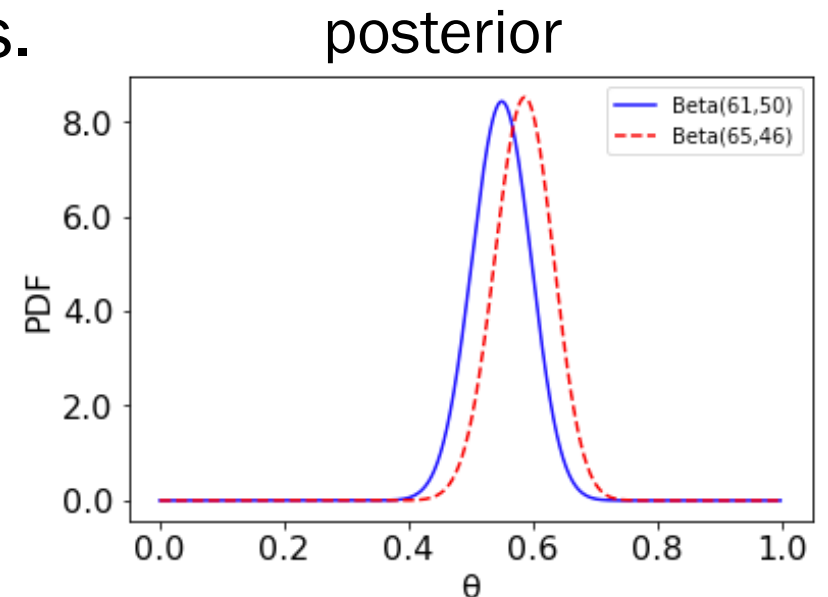


Now flip 100 coins and get 58 heads and 42 tails.

Posterior 1: **Beta(61,50)** mode: $\frac{60}{109}$

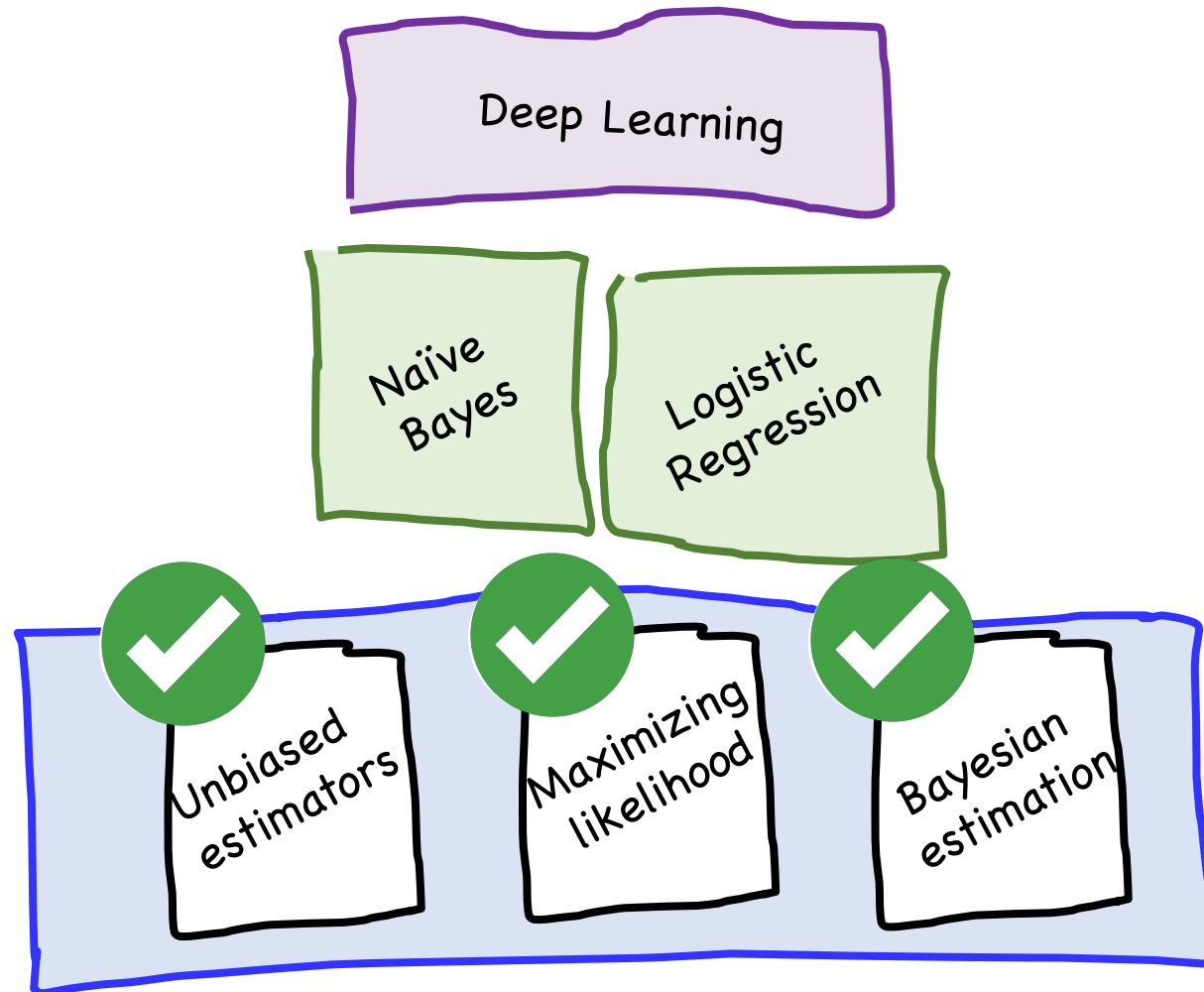
Posterior 2: **Beta(65,46)** mode: $\frac{64}{109}$

Provided we collect enough data,
posteriors will converge to the true value.



The last estimator has risen...

Our Path



Two philosophies for parameter estimation

MLE

Chose theta to make the data likely.

Don't need a prior on theta.



MAP

Chose theta that is most likely given the data

Need a prior on theta