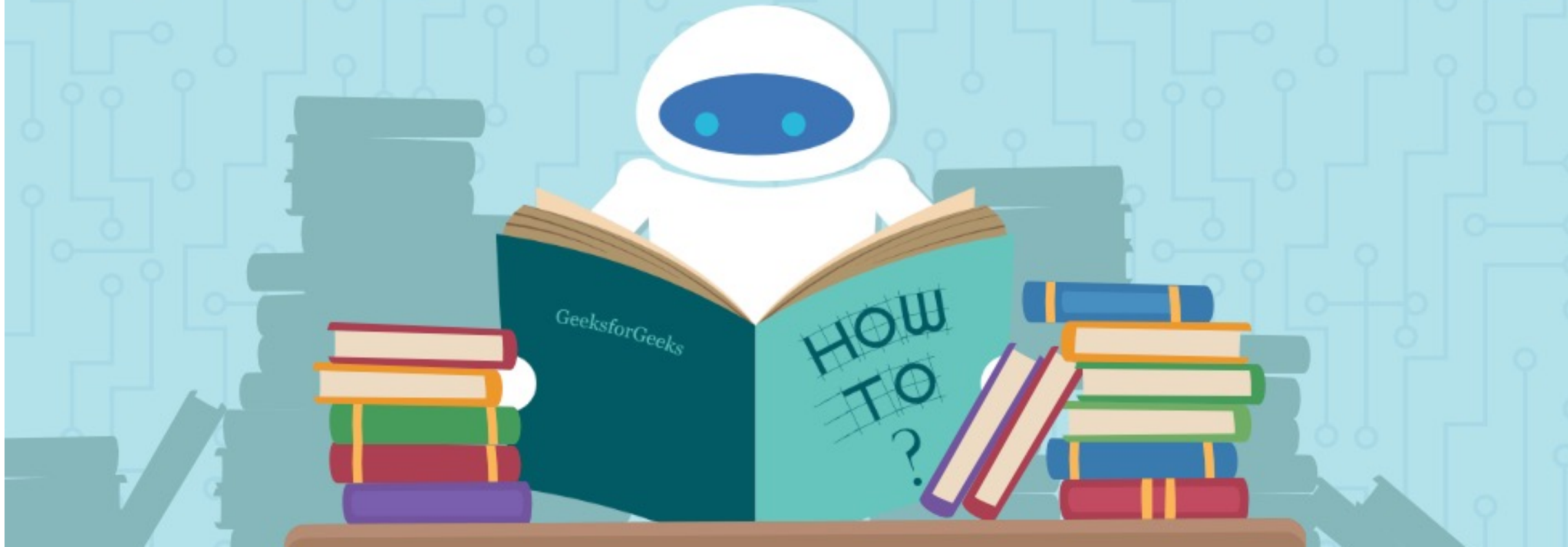
Maximum A Posteriori

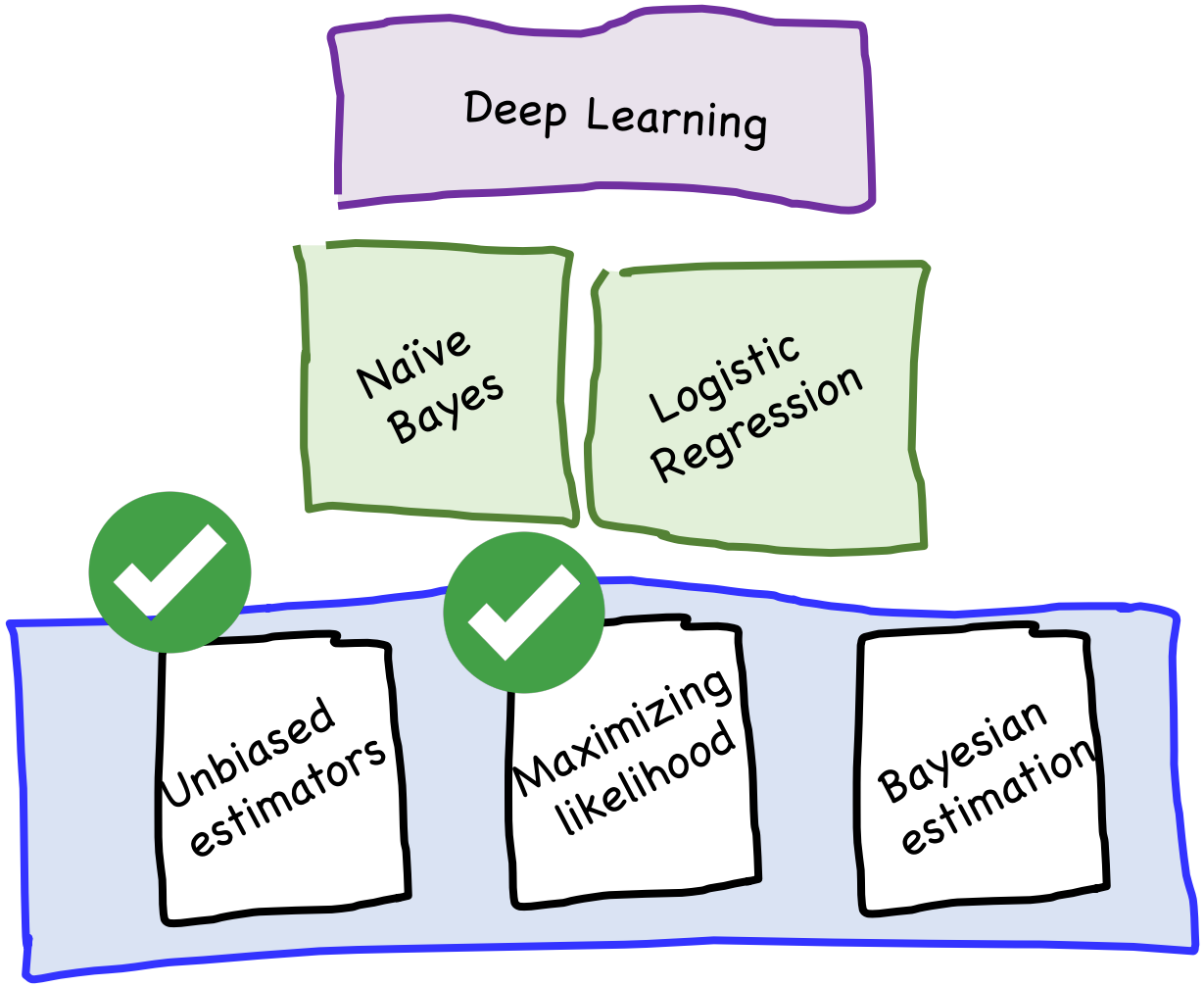
Chris Piech
CS109, Stanford University

<Review>

MACHINE LEARNING



Our Path

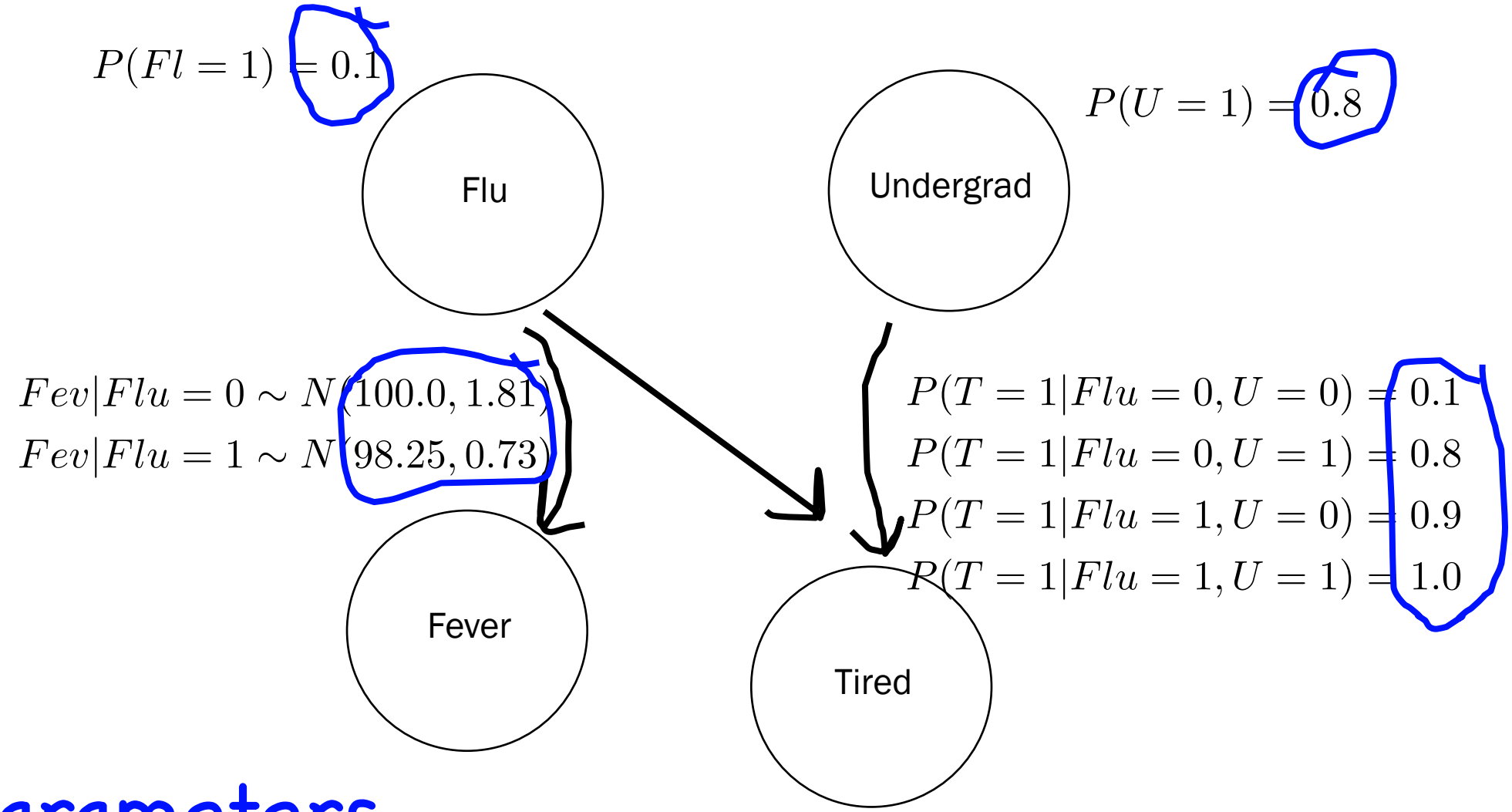

Deep Learning
Naïve Bayes
Logistic Regression
Unbiased estimators
Maximizing likelihood
Bayesian estimation

What are Parameters?

Consider some probability distributions:

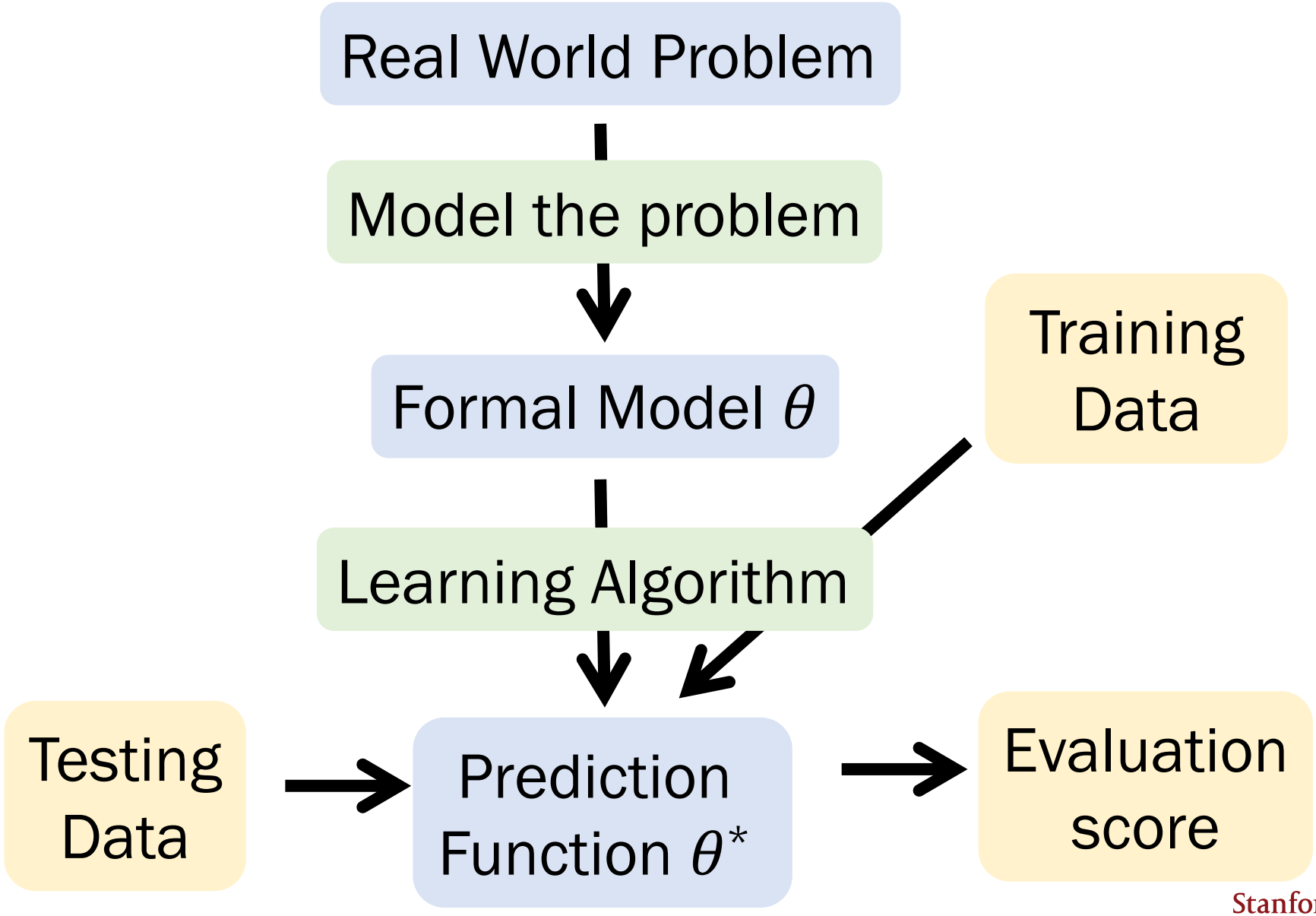
- Ber(p) $\quad \theta = p$
- Poi(λ) $\quad \theta = \lambda$
- Uni(α, β) $\quad \theta = (\alpha, \beta)$
- Normal(μ, σ^2) $\quad \theta = (\mu, \sigma^2)$
- $Y = mX + b$ $\quad \theta = (m, b)$

What are Parameters?

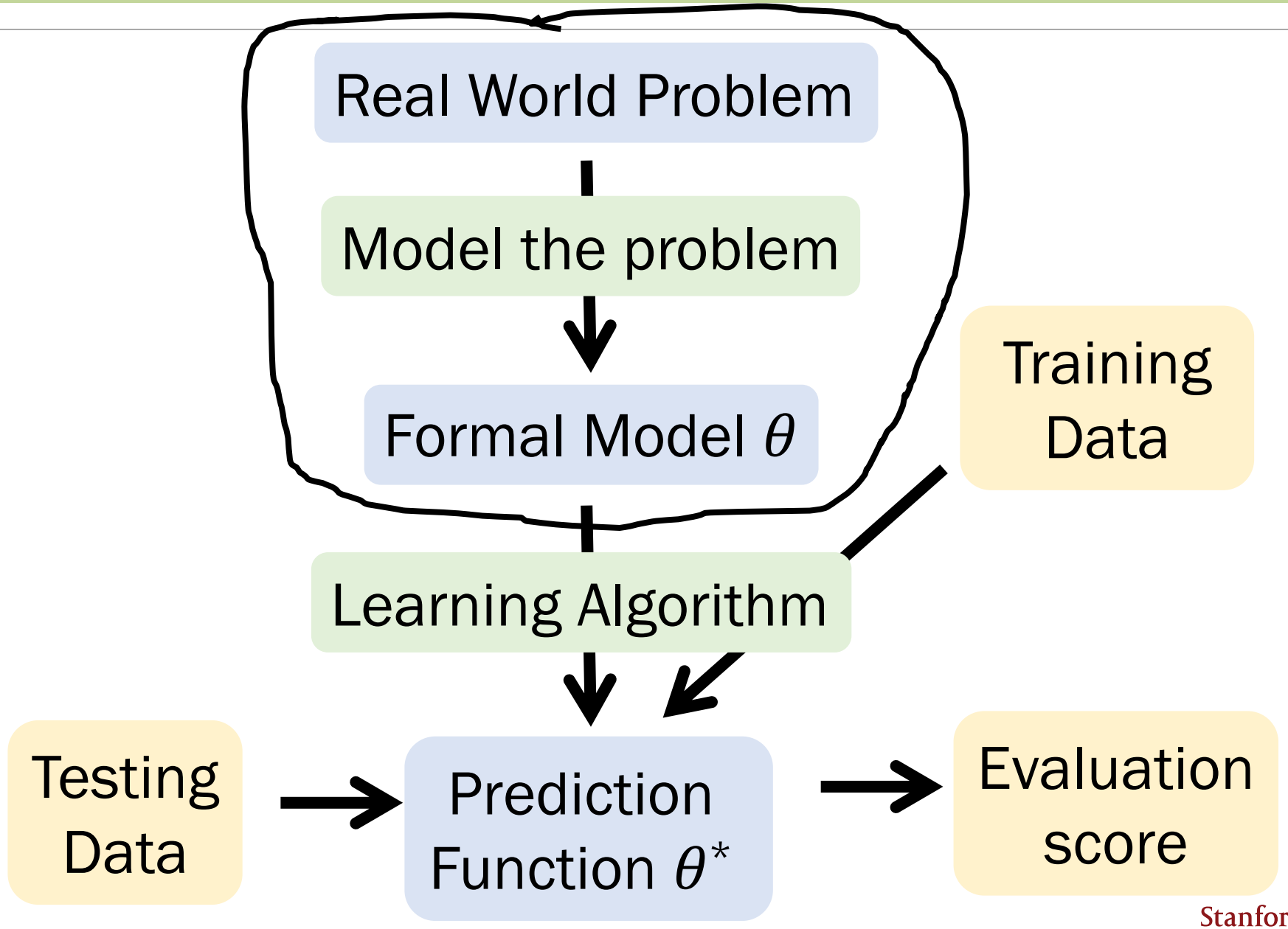


Parameters

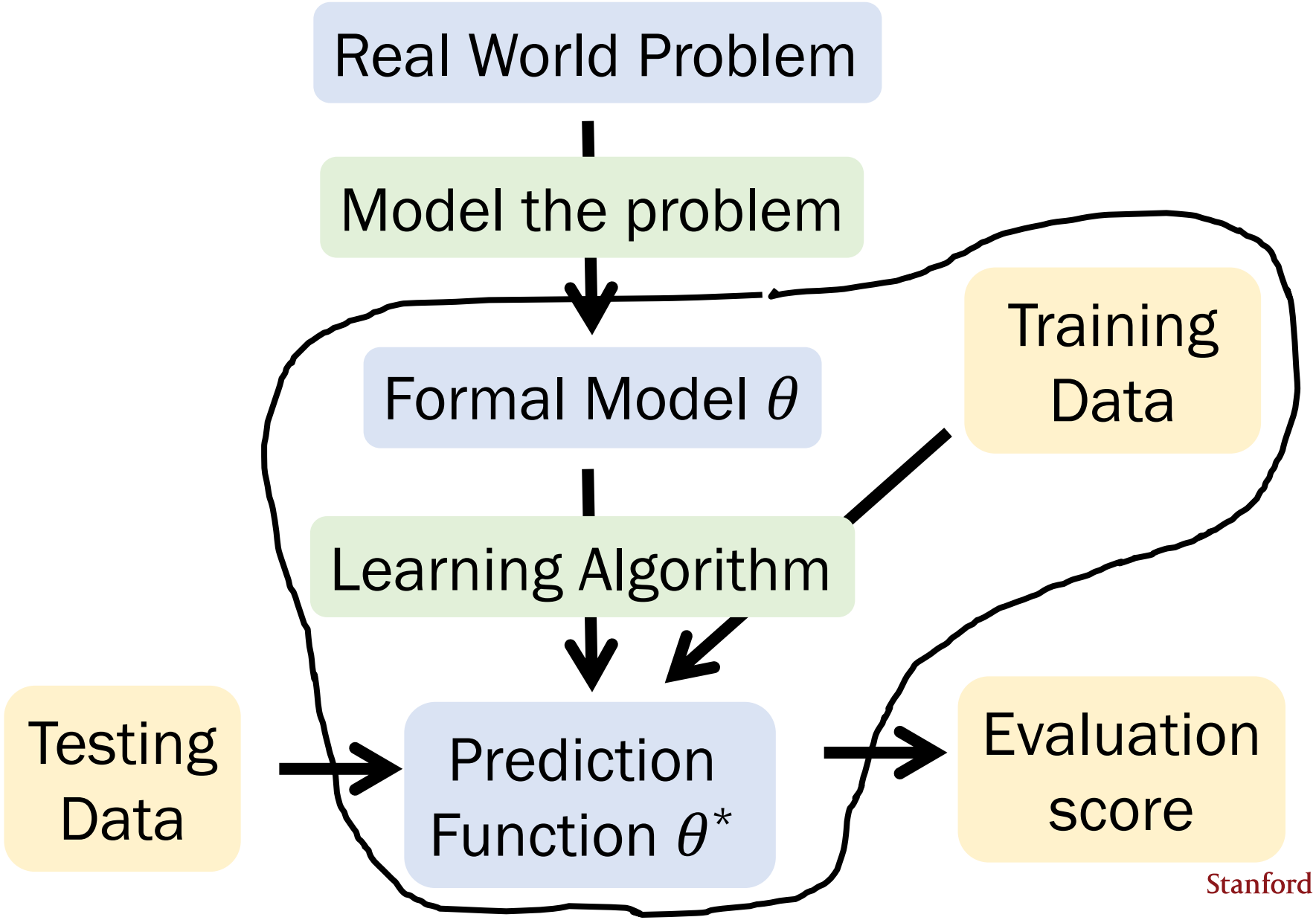
Why Do We Care?



Modelling


Real World Problem
Model the problem
Formal Model θ
Training Data
Learning Algorithm
Testing Data
Prediction Function θ^*
Evaluation score

Parameter Estimation (aka Training)



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

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: 5.6 Parameter σ: 1.4

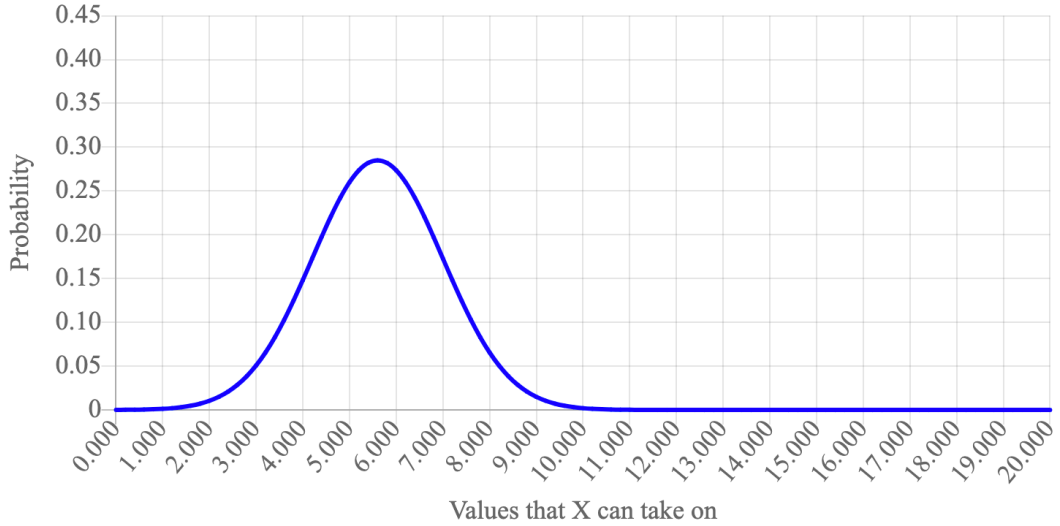
Likelihood

Likelihood: 1.9542923784106326e-15

Log Likelihood: -301.9

Best Seen: -301.9

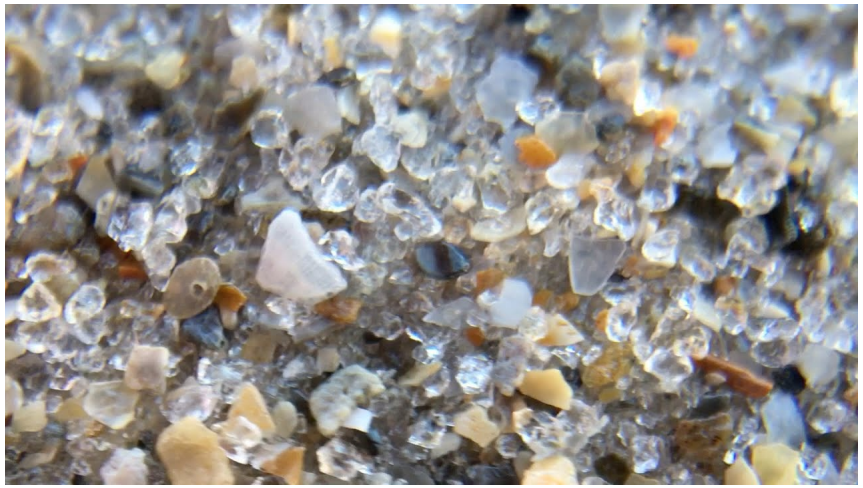
PDF Graph



MLE for a Pareto

```
observations = [1.677, 3.812, 1.463, 2.641, 1.256, 1.678, 1.157,
1.146, 1.323, 1.029, 1.238, 1.018, 1.171, 1.123, 1.074, 1.652,
1.873, 1.314, 1.309, 3.325, 1.045, 2.271, 1.305, 1.277, 1.114,
1.391, 3.728, 1.405, 1.054, 2.789, 1.019, 1.218, 1.033, 1.362,
1.058, 2.037, 1.171, 1.457, 1.518, 1.117, 1.153, 2.257, 1.022,
1.839, 1.706, 1.139, 1.501, 1.238, 2.53 , 1.414, 1.064, 1.097,
1.261, 1.784, 1.196, 1.169, 2.101, 1.132, 1.193, 1.239, 1.518,
2.764, 1.053, 1.267, 1.015, 1.789, 1.099, 1.25 , 1.253, 1.418,
1.494, 1.015, 1.459, 2.175, 2.044, 1.551, 4.095, 1.396, 1.262,
1.351, 1.121, 1.196, 1.391, 1.305, 1.141, 1.157, 1.155, 1.103,
1.048, 1.918, 1.889, 1.068, 1.811, 1.198, 1.361, 1.261, 4.093,
2.925, 1.133, 1.573]
```

```
def estimate_alpha(observations):
    print('your code here')
```



We know sand is distributed as a pareto with PDF

$$f(x) = \frac{\alpha}{x^{\alpha+1}}$$

$$\alpha_{\text{mle}} = \frac{n}{\sum_i \log x_i}$$

MLE for a Pareto

- Consider I.I.D. random variables $X_1, X_2, ..., X_n$
  - $X_i \sim$ Pareto(α). **Use Maximum Likelihood to estimate α.**

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

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

3. Find the value of α which maximizes log likelihood

MLE for a Pareto

- Consider I.I.D. random variables $X_1, X_2, ..., X_n$
  - $X_i \sim$ Pareto(α). **Use Maximum Likelihood to estimate α.**
  - Likelihood:

$$L(\alpha) = \prod_{i=1}^{n} \frac{\alpha}{x_i^{\alpha+1}}$$

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

3. Find the value of α which maximizes log likelihood

MLE for a Pareto

- Consider I.I.D. random variables $X_1, X_2, ..., X_n$
  - $X_i \sim$ Pareto(α). **Use Maximum Likelihood to estimate α.**
  - Likelihood:

$$L(\alpha) = \prod_{i=1}^{n} \frac{\alpha}{x_i^{\alpha+1}}$$

  - Log-likelihood:

$$LL(\alpha) = \sum_{i=1}^{n} \log \alpha - (\alpha + 1) \log x_i = n \log \alpha - (\alpha + 1) \sum_{i=1}^{n} \log x_i$$

3. Find the value of α which maximizes log likelihood

MLE for a Pareto

- Consider I.I.D. random variables $X_1, X_2, ..., X_n$
  - $X_i \sim$ Pareto(α). **Use Maximum Likelihood to estimate α.**
  - Likelihood:

$$L(\alpha) = \prod_{i=1}^{n} \frac{\alpha}{x_i^{\alpha+1}}$$

  - Log-likelihood:

$$LL(\alpha) = \sum_{i=1}^{n} \log \alpha - (\alpha + 1) \log x_i = n \log \alpha - (\alpha + 1) \sum_{i=1}^{n} \log x_i$$

  - Chose **α** to be the argmax of LL:

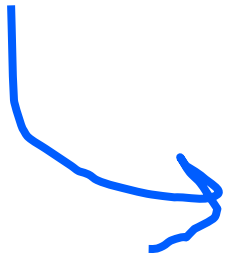
$$\frac{\partial LL(\alpha)}{\partial \alpha} = \frac{n}{\alpha} - \sum_{i=1}^{n} \log x_i$$


arg max

argmax of log liklihood

Recall for the pareto

$$LL(\alpha) = n \log \alpha - (\alpha + 1) \sum_{i=1}^{n} \log x_i$$



All argmax methods require the derivative

$$\frac{\partial LL(\alpha)}{\partial \alpha} = \frac{n}{\alpha} - \sum_{i=1}^{n} \log x_i$$

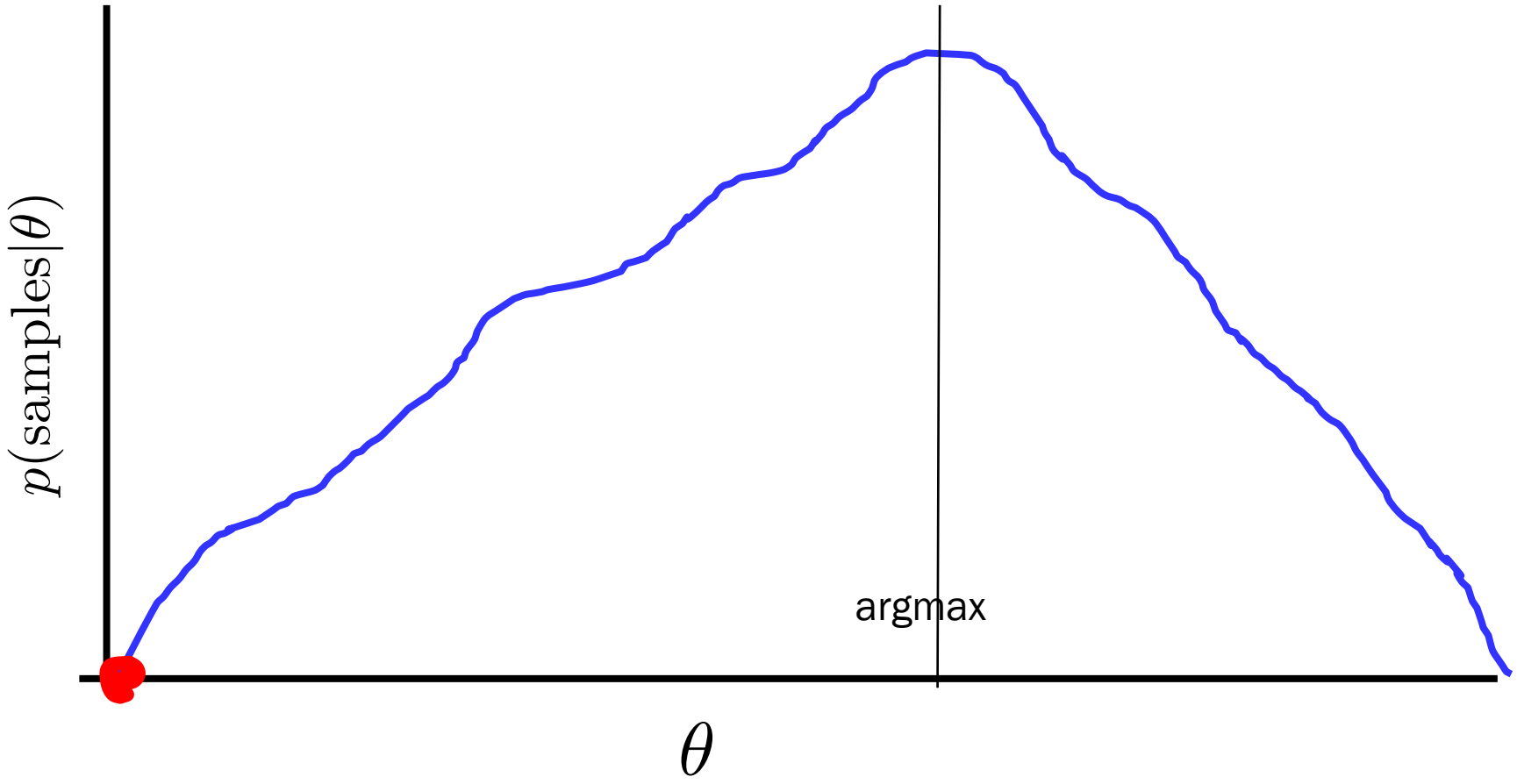
Argmax Option #1: set the derivative to 0, and solve for alpha

End Review


Disney
THE
LION KING II
SIMBA'S · PRIDE

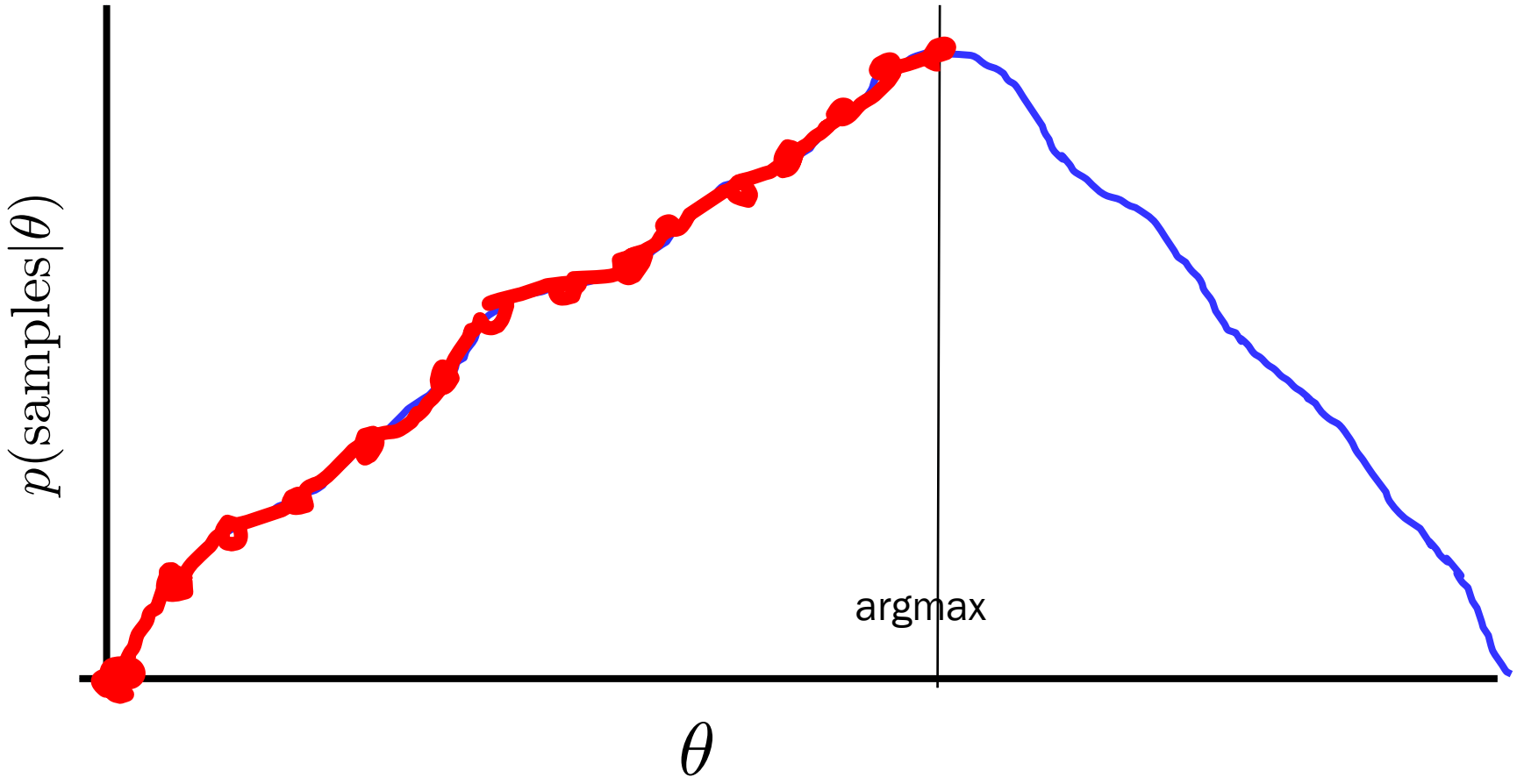
Optimization (argmax)
Option #2: Gradient Descent

Gradient Ascent



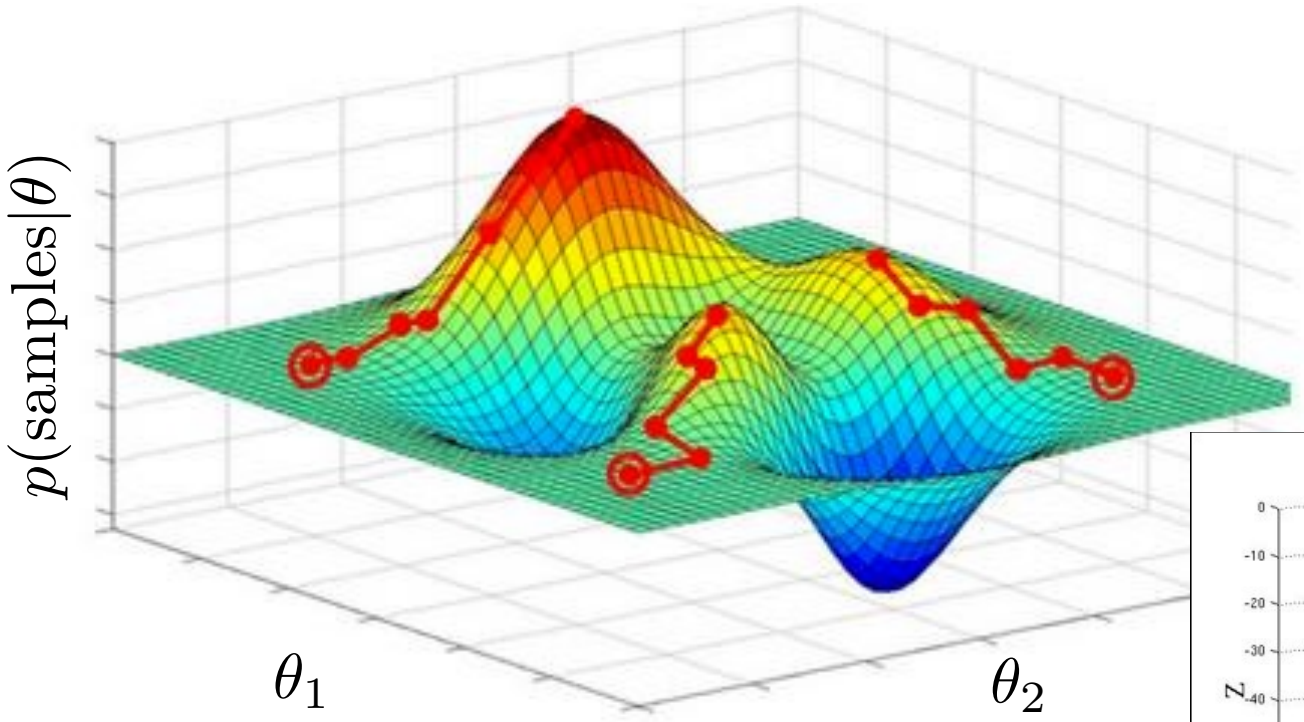
Walk uphill and you will find a local maxima
(if your step size is small enough)

Gradient Ascent

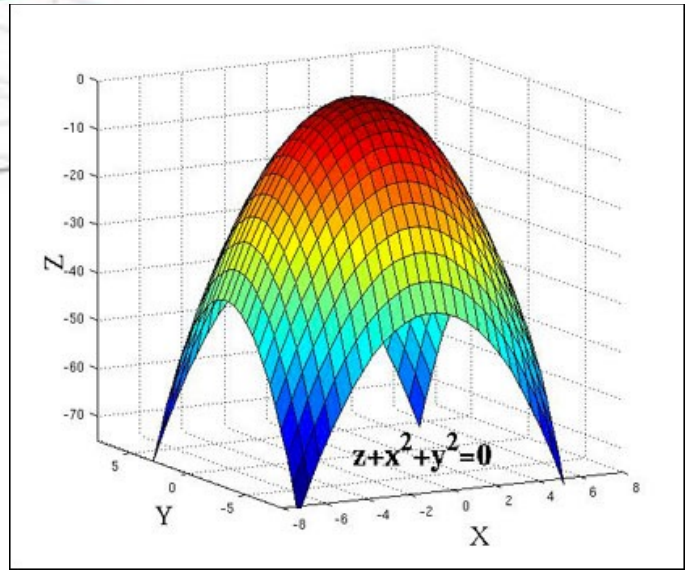


Walk uphill and you will find a local maxima
(if your step size is small enough)

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

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

Step size constant

This is some **profound** life philosophy

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

Gradient Ascent

Initialize: θ_j **= random for all** $0 \leq j \leq m$

Calculate all θ_j

Gradient Ascent

Initialize: θ_j **= random for all** $0 \leq j \leq m$

Repeat many times:

Calculate all `gradient[j]`*'s based on data*

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

Gradient Ascent for Pareto

```
Initialize: alpha = some random start
```

```
Repeat many times:
```

```
# Calculate gradient_alpha based on data
```

$$\frac{\partial LL(\alpha)}{\partial \alpha} = \frac{n}{\alpha} - \sum_{i=1}^{n} \log x_i$$

```
alpha += η * gradient_alpha
```

Gradient Ascent for Pareto

```
Initialize: alpha = some random start
```

```
Repeat many times:

    # Calculate gradient_alpha based on data
    gradient_alpha = n / alpha
    for x_i in data:
        gradient_alpha -= math.log(x_i)

    alpha += η * gradient_alpha
```

Gradient **descent** is the bread and butter algorithm for optimization
(use neg LL to switch ascent to descent)

Pedagogic Pause

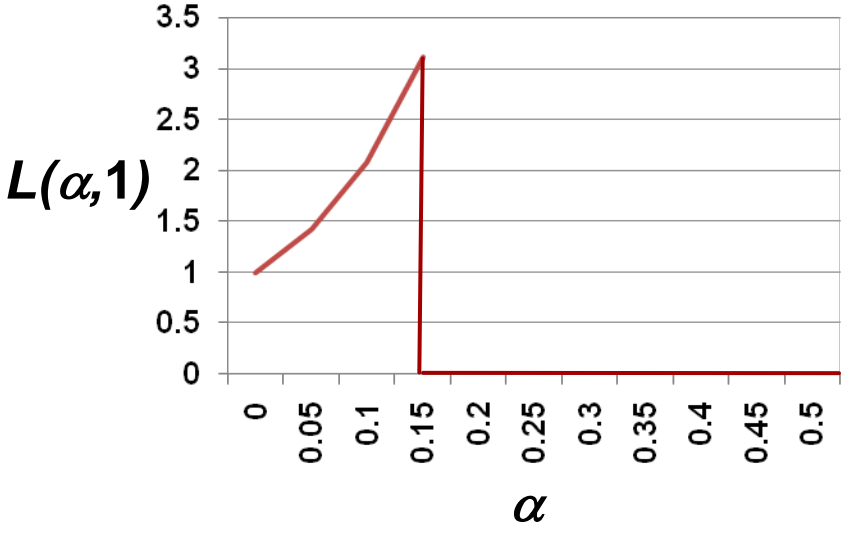
Something rotten
in the world of MLE

MLE Could Benefit from Priors

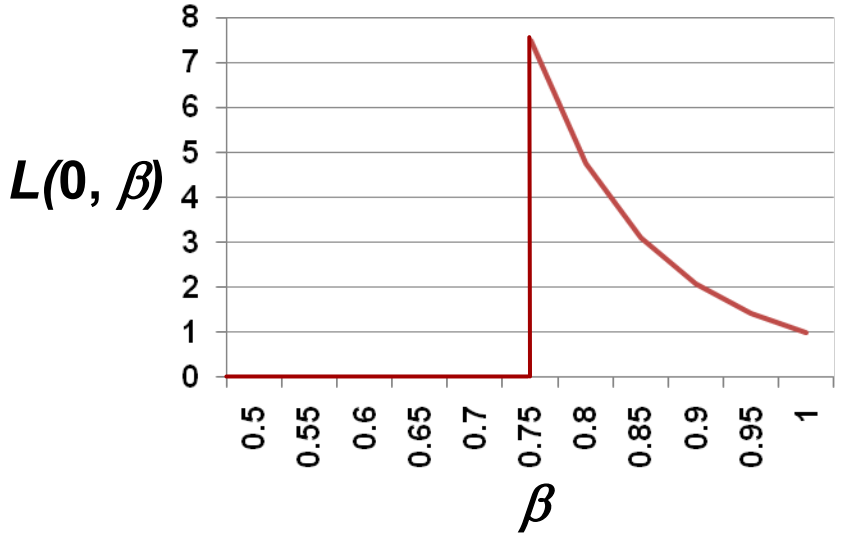
Consider I.I.D. random variables $X_1, X_2, ..., X_n$

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

Likelihood: $L(\alpha, 1)$



Likelihood: $L(0, \beta)$



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 equally good
- So, what happened by opening envelope?
  - And does it really make sense to switch?

Thinking Deeper About Two Envelopes

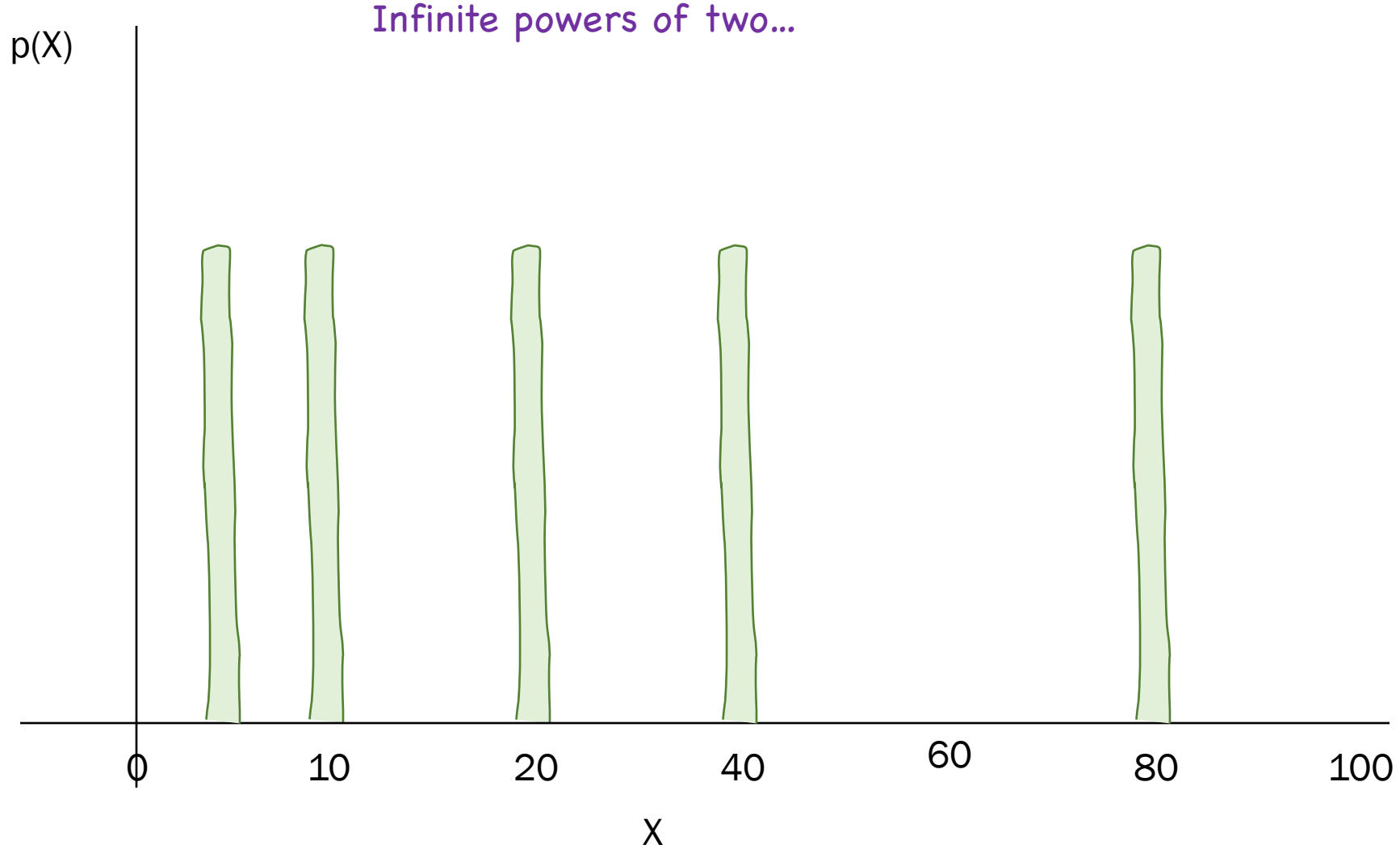
The "two envelopes" problem set-up

- 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

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

- 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?

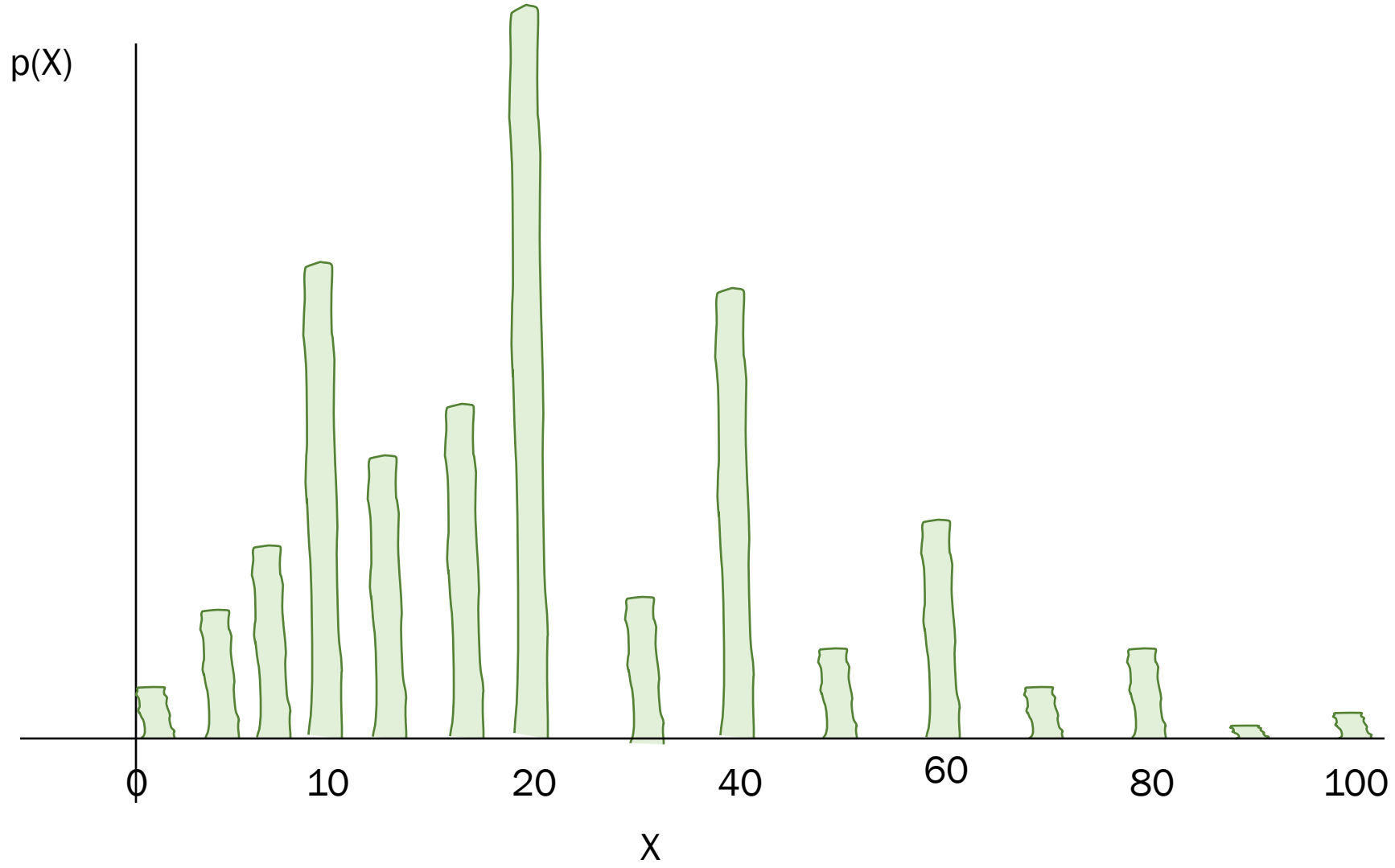


Subjectivity of Probability

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

We have seen this play out before…

MLE vs Beta

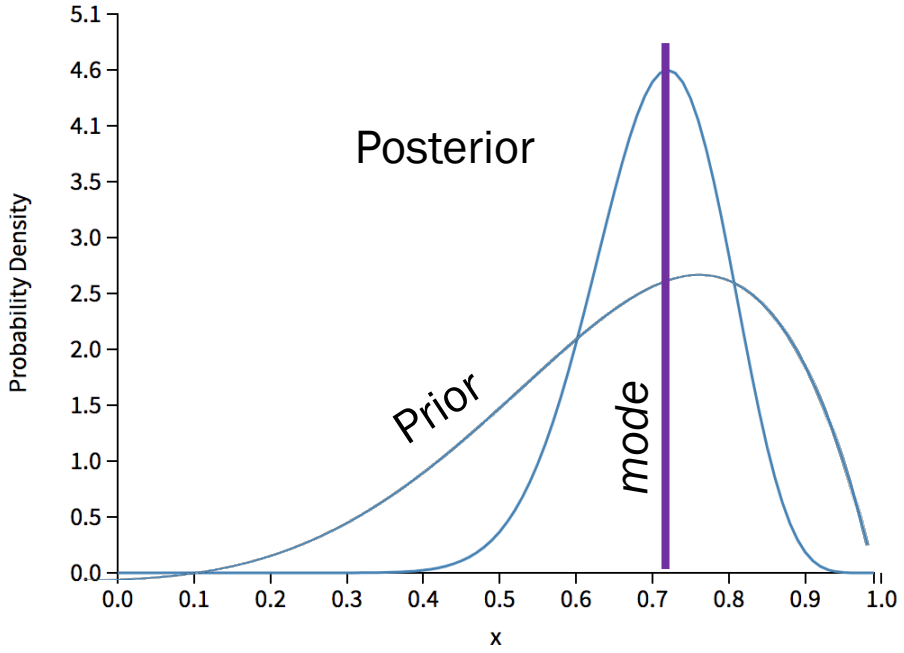
The medicine is tried on 20 patients. It "works" for 14 and "doesn't work" for 6. What is your new belief that the drug works?

In other words I have 20 IID samples from a Bernoulli. Estimate p. The data is [1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0]

MLE estimate:

$$p \approx \frac{14}{20} = 0.7$$

Beta estimate:



Priors for Parameter Estimation?

MLE vs MAP

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

Maximum Likelihood Estimation

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

$$= \underset{\theta}{\operatorname{argmax}} \left(\sum_i \log f(X^{(i)} = x^{(i)} | \theta) \right)$$

Maximum A Posteriori

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

Notation Shorthand

MAP, without shorthand

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

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

$$\hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}}\ f(\theta | x^{(1)}, \ldots, x^{(n)})$$

MLE vs MAP

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

Maximum Likelihood Estimation

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}}\ f(x^{(1)}, \ldots, x^{(n)}|\theta)$$

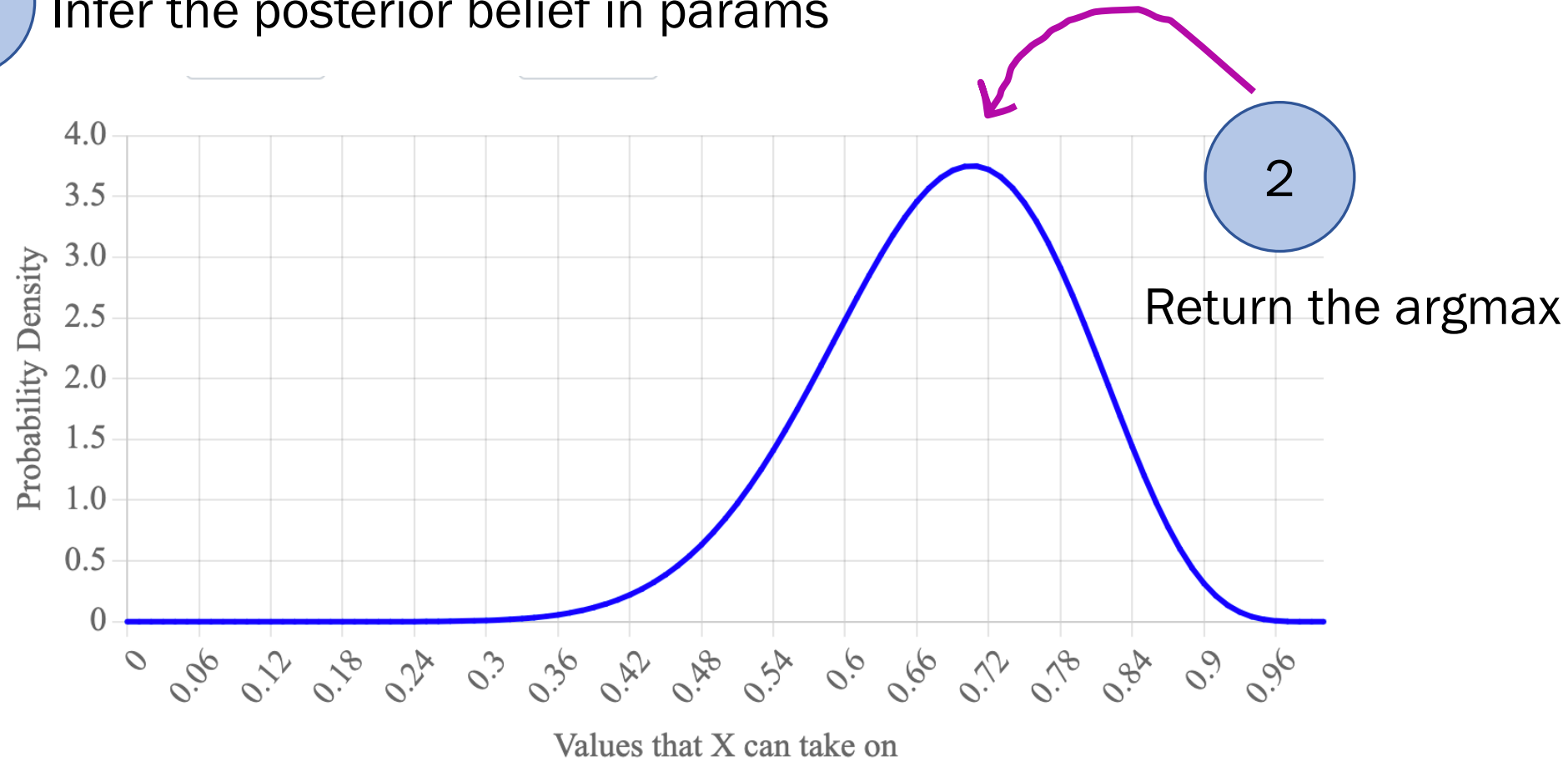
$$= \underset{\theta}{\text{argmax}} \left(\sum_i \log f(x^{(i)}|\theta) \right)$$

Maximum A Posteriori

$$\hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}}\ f(\theta|x^{(1)}, \ldots, x^{(n)})$$

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

1 Infer the posterior belief in params



Quick MAP for Bernoulli and Binomial

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

But MLE works for more than just estimating p

Conjugate distributions

MAP estimator:

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

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

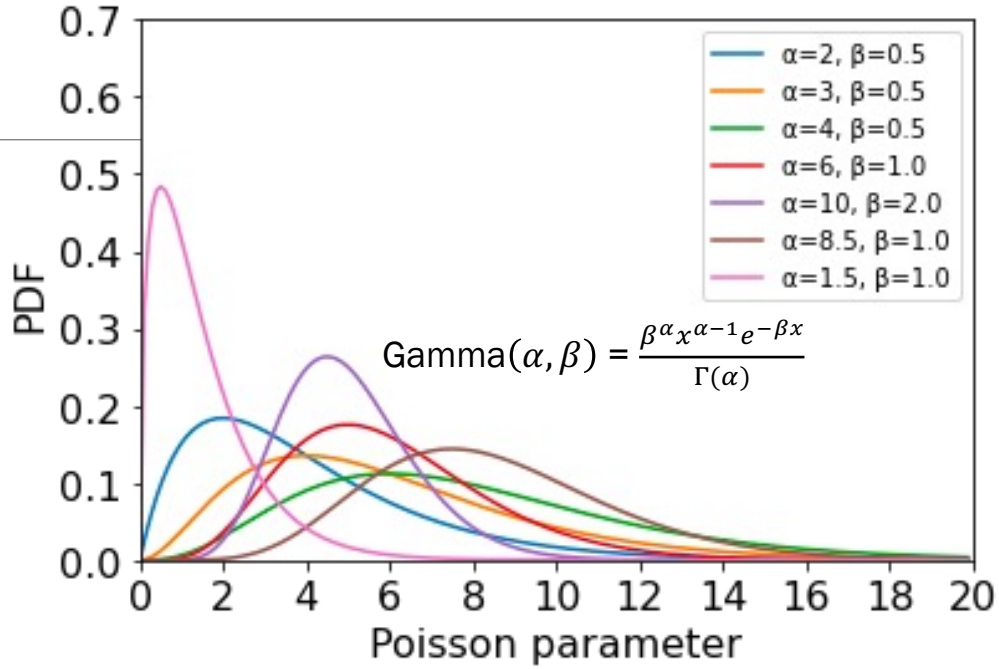
Don't need to know Inverse Gamma… but it will know you ☺

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

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

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

Multinomial is Multiple times the fun

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

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

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

Prior Dirichlet$(a_1, a_2, \ldots, a_m)$
Saw $(\sum_{i=1}^{m} a_i) - m$ imaginary trials, with $a_i - 1$ of outcome i

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

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

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

Your Happy Laplace

Laplace gives a classic Direchlet

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

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

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

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}: $p_1 = 3/12, p_2 = 2/12, p_3 = 0/12,$ ⚠
$p_4 = 3/12, p_5 = 1/12, p_6 = 3/12$

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}: $p_1 = 3/12, p_2 = 2/12, p_3 = 0/12,$ ⚠️
$p_4 = 3/12, p_5 = 1/12, p_6 = 3/12$

What are your Laplace estimates for each roll outcome?

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

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

Laplace smoothing:
- Easy to implement/remember
- **Avoids estimating a parameter of 0**

Can we generalize?

Most important slide of today

Maximum A Posteriori

data: $x^{(1)}, \ldots, x^{(n)}$

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

---

likelihood

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

posterior

prior

Maximum A Posteriori

data: $x^{(1)}, \ldots, x^{(n)}$ $\quad \hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}}\ f(\theta | x^{(1)}, \ldots, x^{(n)})$

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

$$= \underset{\theta}{\text{argmax}}\ \frac{g(\theta) \prod_{i=1}^{n} f(x^{(i)} | \theta)}{h(x^{(1)}, x^{(2)}, \ldots, x^{(n)})}$$

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

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

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

Maximum Likelihood Estimation

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}}\ f(x^{(1)}, \ldots, x^{(n)}|\theta)$$

$$= \underset{\theta}{\text{argmax}} \left(\sum_i \log f(x^{(i)}|\theta) \right)$$

Maximum A Posteriori

$$\hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}}\ f(\theta|x^{(1)}, \ldots, x^{(n)})$$

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

Gotta get that intuition

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

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

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

This is the beta PDF

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

This is ???

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

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

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

This is the beta PDF

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

Product of thetas and (1-theta)s

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

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

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

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

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

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

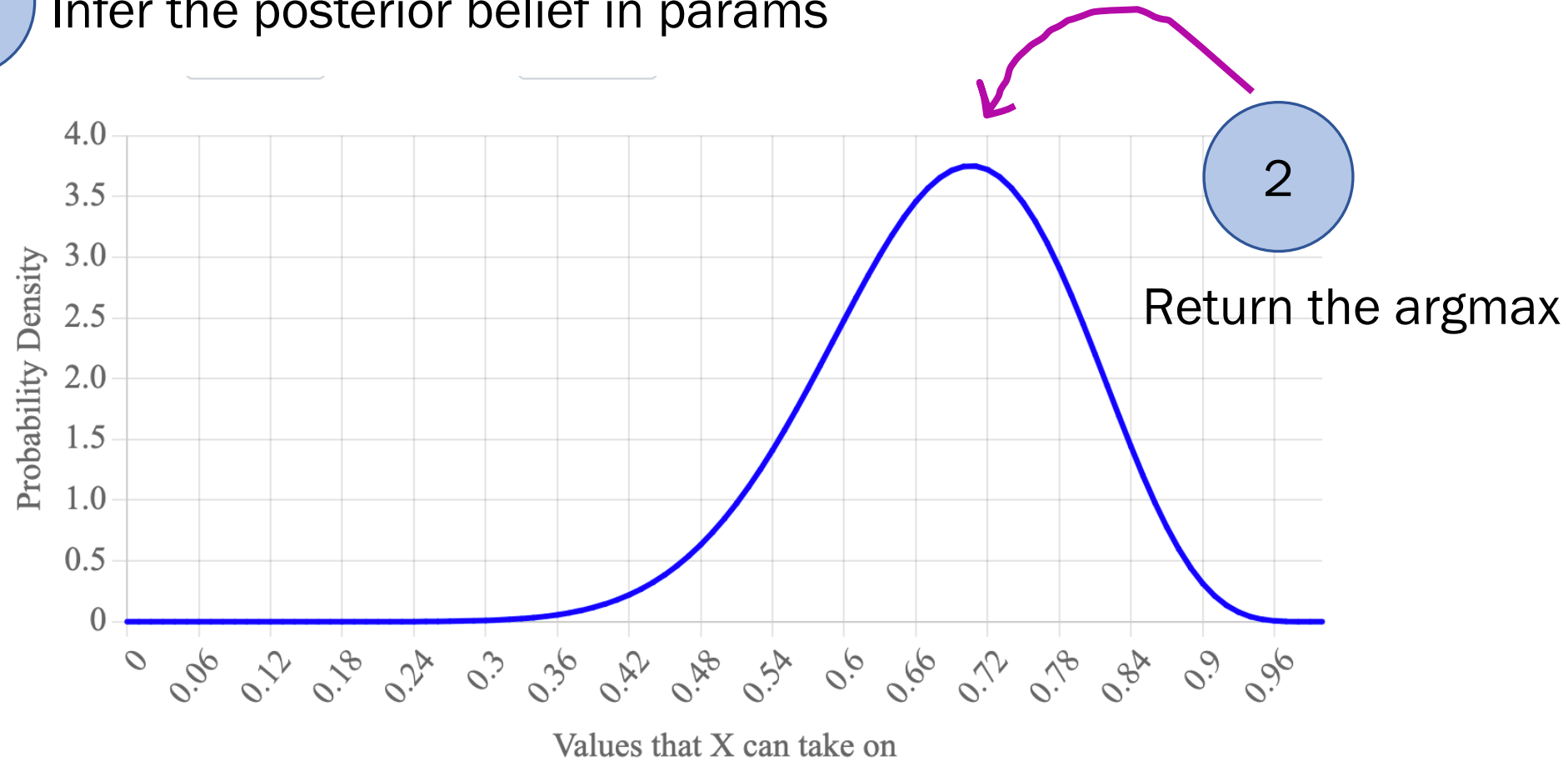
$$= \underset{\theta}{\operatorname{argmax}} (a-1+n)\log\theta + (b-1+m)\log(1-\theta)$$

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

That's the mode of the updated beta

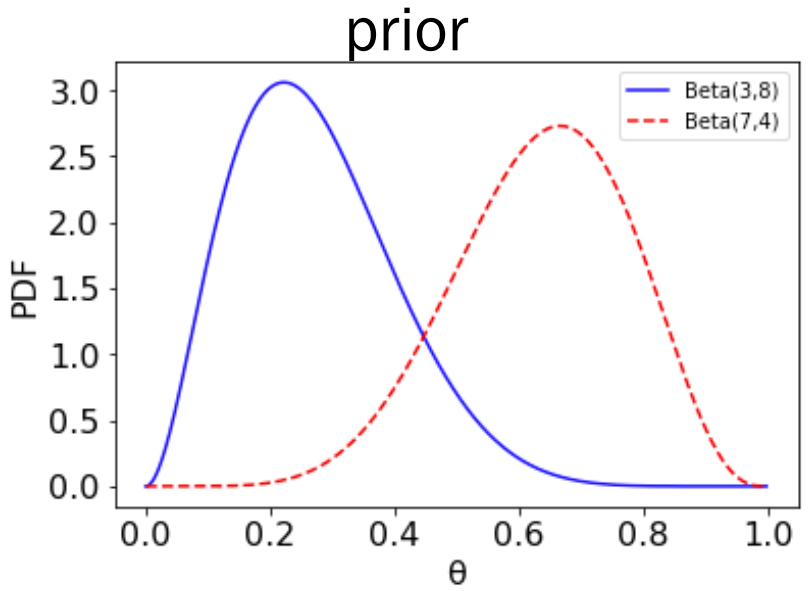
$P(\theta \mid D)$ For Bernoulli

1 Infer the posterior belief in params



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: $\frac{6}{9}$



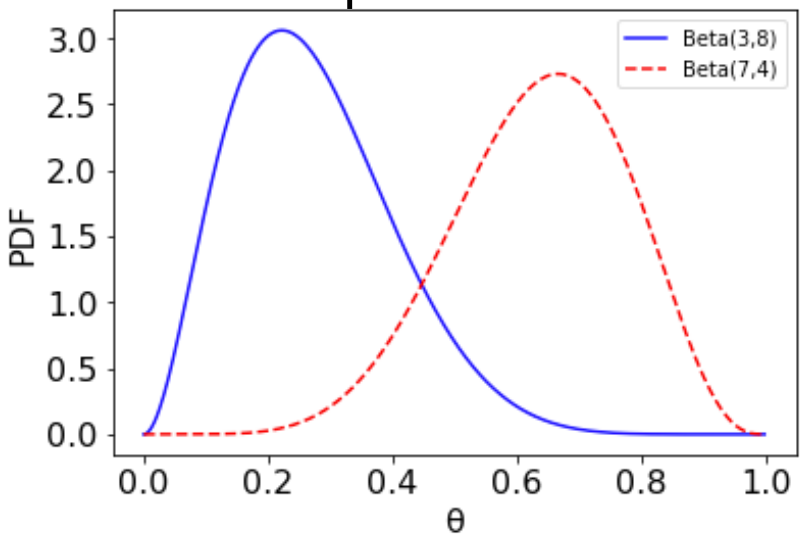
Now flip 100 coins and get 58 heads and 42 tails.

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: $\frac{6}{9}$

prior

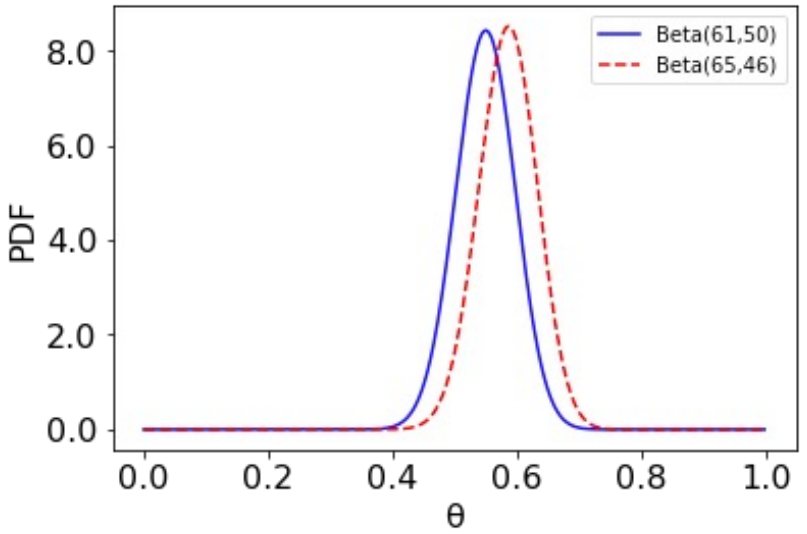


Now flip 100 coins and get 58 heads and 42 tails.

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

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

posterior

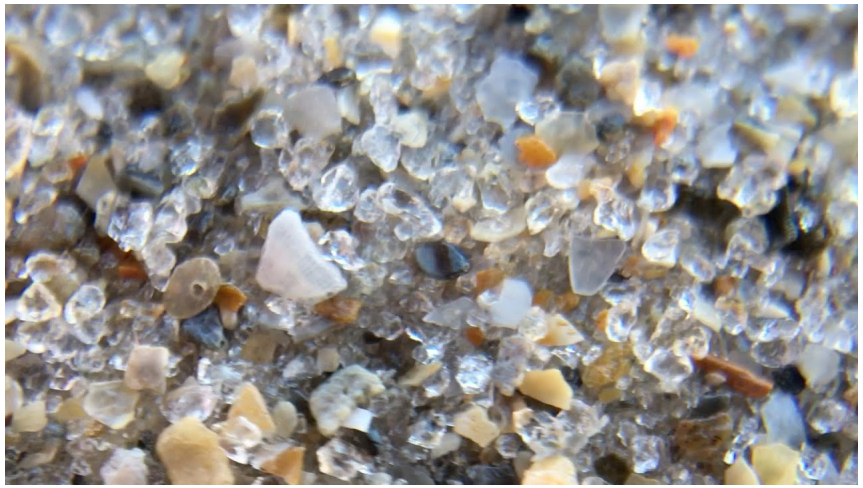


Provided we collect enough data, posteriors will converge to the true value.

MLE for a Pareto

```
observations = [1.677, 3.812, 1.463, 2.641, 1.256, 1.678, 1.157,
1.146, 1.323, 1.029, 1.238, 1.018, 1.171, 1.123, 1.074, 1.652,
1.873, 1.314, 1.309, 3.325, 1.045, 2.271, 1.305, 1.277, 1.114,
1.391, 3.728, 1.405, 1.054, 2.789, 1.019, 1.218, 1.033, 1.362,
1.058, 2.037, 1.171, 1.457, 1.518, 1.117, 1.153, 2.257, 1.022,
1.839, 1.706, 1.139, 1.501, 1.238, 2.53 , 1.414, 1.064, 1.097,
1.261, 1.784, 1.196, 1.169, 2.101, 1.132, 1.193, 1.239, 1.518,
2.764, 1.053, 1.267, 1.015, 1.789, 1.099, 1.25 , 1.253, 1.418,
1.494, 1.015, 1.459, 2.175, 2.044, 1.551, 4.095, 1.396, 1.262,
1.351, 1.121, 1.196, 1.391, 1.305, 1.141, 1.157, 1.155, 1.103,
1.048, 1.918, 1.889, 1.068, 1.811, 1.198, 1.361, 1.261, 4.093,
2.925, 1.133, 1.573]
```

```
def estimate_alpha(observations):
    print('your code here')
```



We know sand is distributed as a pareto with PDF

$$f(x) = \frac{\alpha}{x^{\alpha+1}}$$

Prior: $\alpha \sim N(\mu = 2.5, \sigma^2 = 3)$

MLE for a Pareto

- $X_i \sim \text{Pareto}(\alpha)$. **Use MAP to estimate α.**
- MAP function:

$$= \log g(\alpha) + n \log \alpha - (\alpha + 1) \sum_{i=1}^{n} \log x_i$$

$$= \log \frac{1}{\sqrt{3}\sqrt{2\pi}} e^{\frac{-(\alpha-2)^2}{6}} + n \log \alpha - (\alpha + 1) \sum_{i=1}^{n} \log x_i$$

$$= K + \frac{-(\alpha-2)^2}{6} + n \log \alpha - (\alpha + 1) \sum_{i=1}^{n} \log x_i$$

- Choose **α** which is the argmax of this function

$$\frac{\partial \text{MAP}(\alpha)}{\partial \alpha} = -2\alpha + 4 + \frac{n}{\alpha} - \sum_{i=1}^{n} \log x_i$$

Gradient Descent for Pareto

```
Initialize: alpha = some random start
```

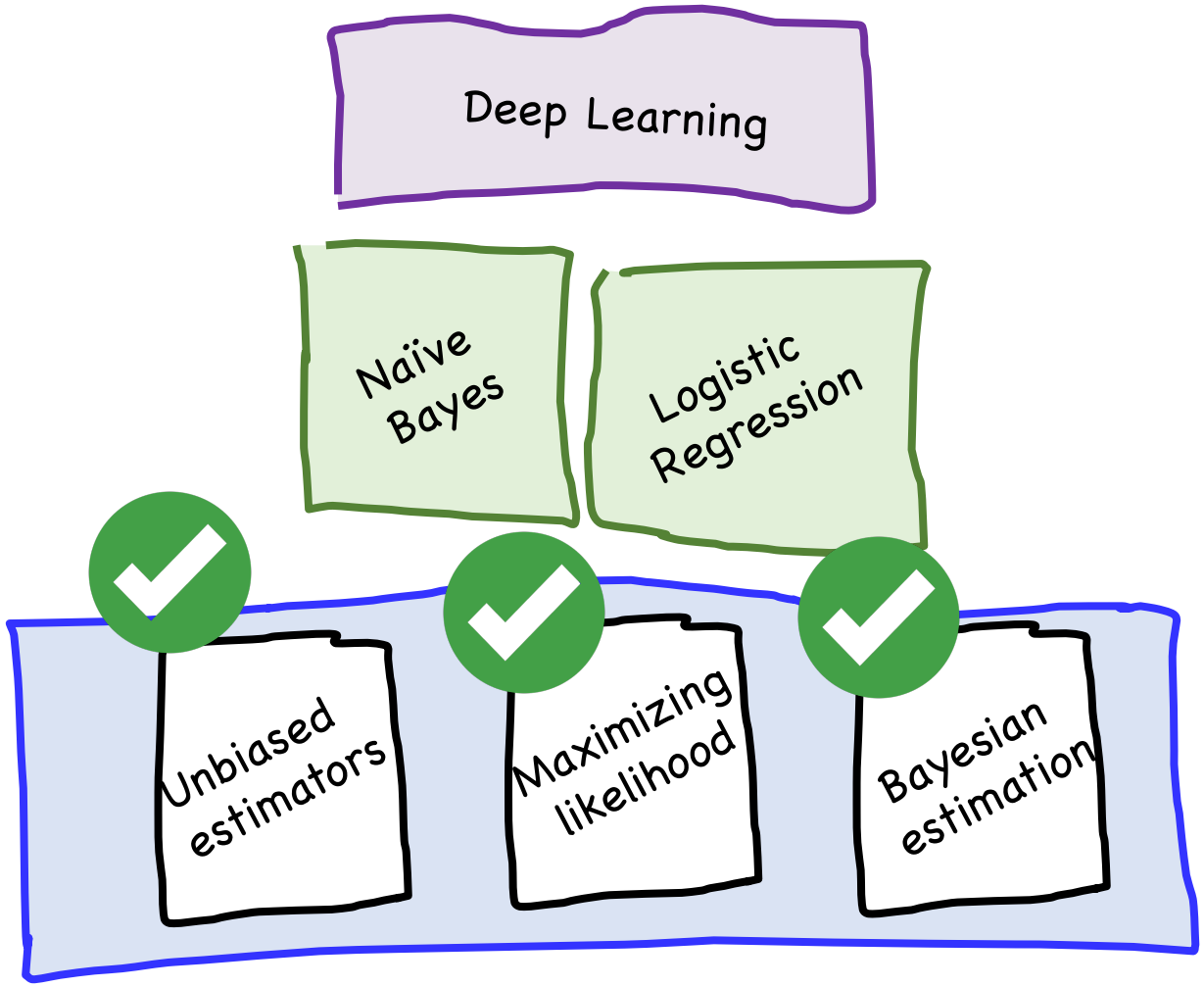
```
Repeat many times:

    # Calculate gradient_alpha based on data
    gradient_alpha = -2 * alpha + 4 + n / alpha
    for x_i in data:
        gradient_alpha -= math.log(x_i)

    alpha -= η * gradient_alpha
```

The last estimator has risen…

Our Path


Deep Learning
Naïve Bayes
Logistic Regression
Unbiased estimators
Maximizing likelihood
Bayesian estimation