Parameter Estimation

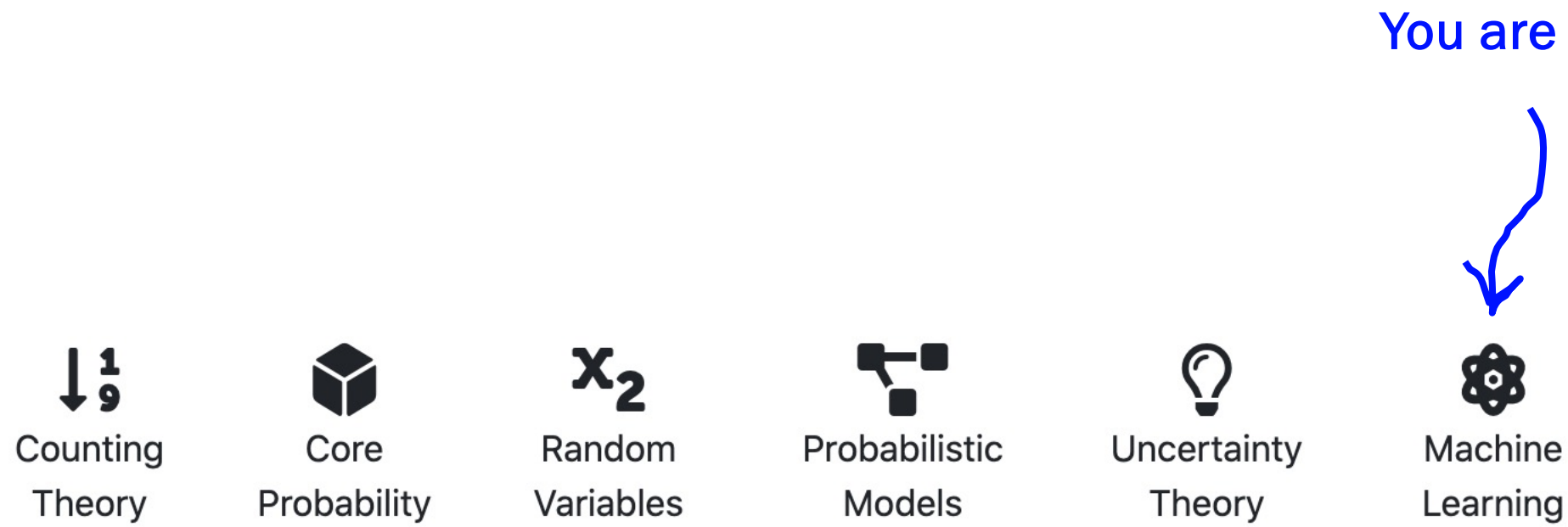
CS109, Stanford University

Announcements

We’re in the final stretch!

4 more lectures (including this one!

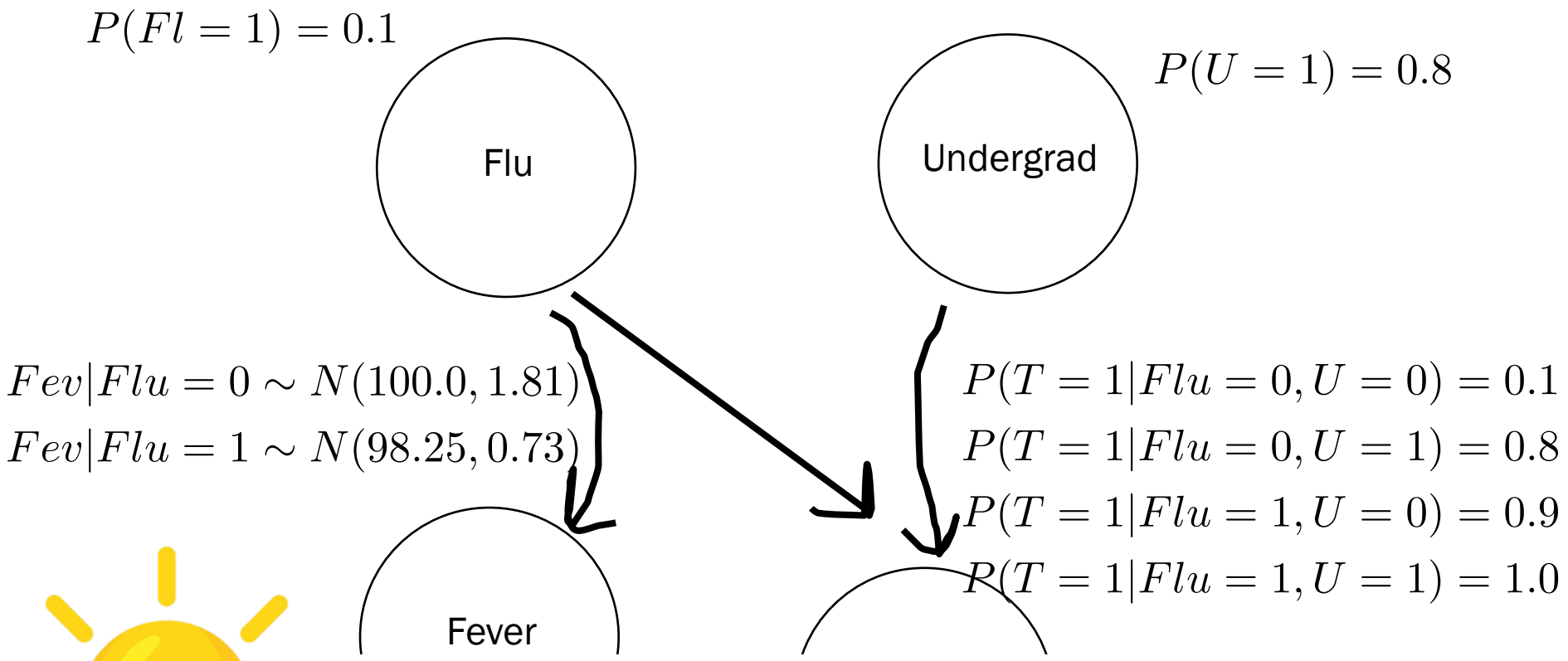
Where are we in CS109?



MACHINE LEARNING

Let’s remember our wonderful WebMD example!

Probabilistic Model



If you know the probability of each random variables given the ones that directly cause it, you can joint sample!

But where do those numbers come from?

Suspense

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?

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

Oh can we also learn the *structure* of the model too?

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

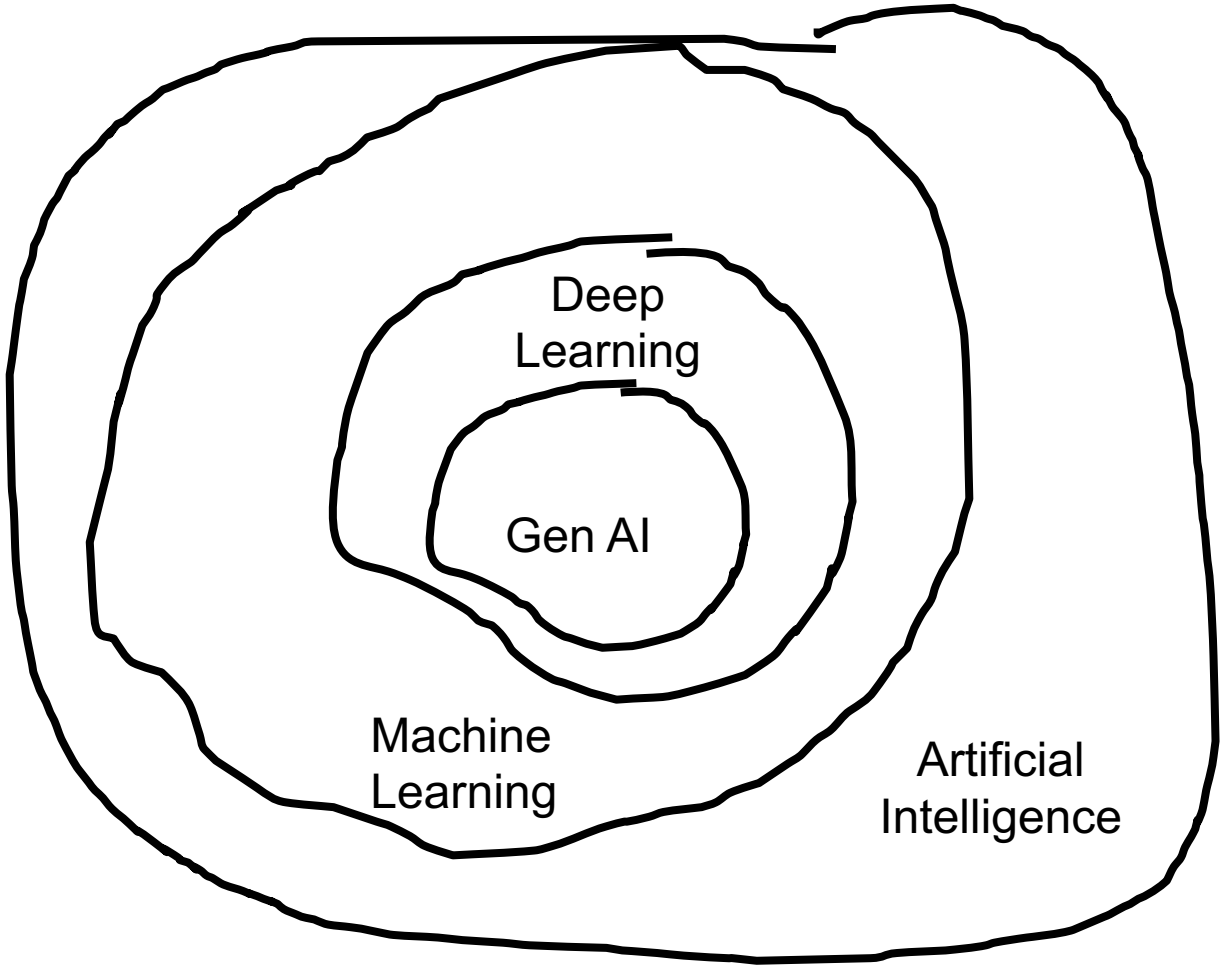
~~Oh can we also learn the *structure* of the model too?~~

I wish. Another day ☺

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

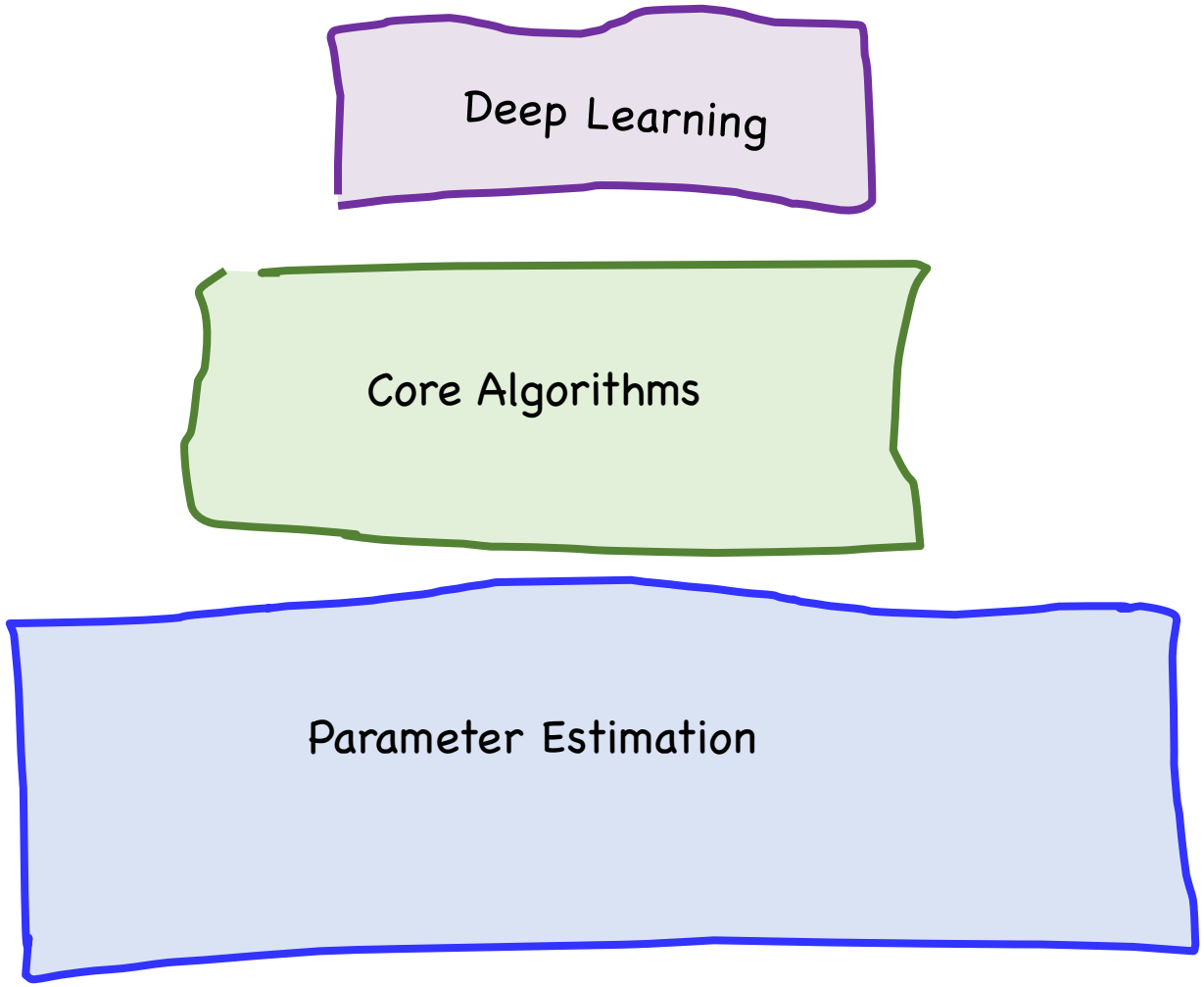
Machine Learning

AI and Machine Learning

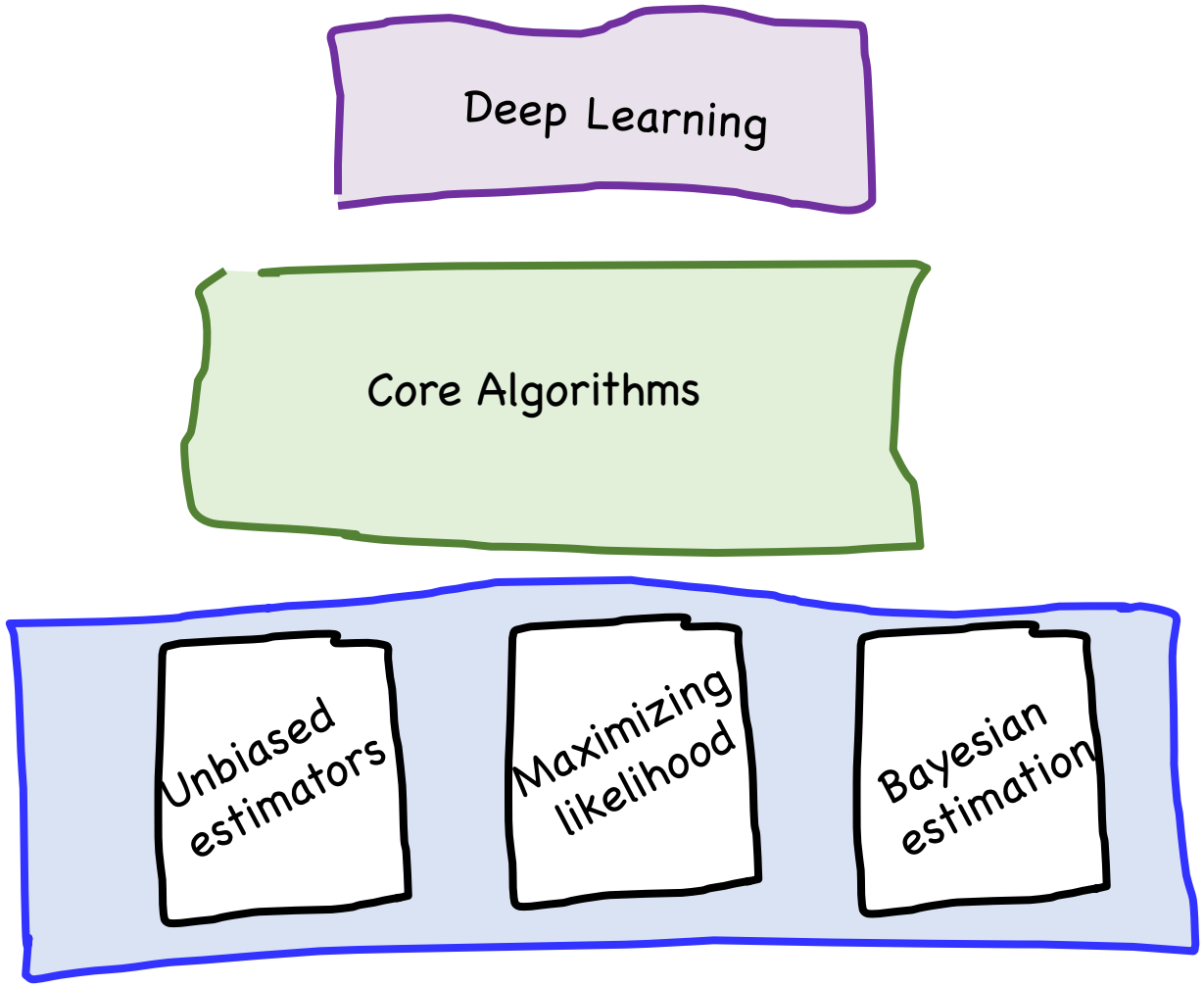


ML: Rooted in probability theory

Our Path



Our Path


Deep Learning
Core Algorithms
Unbiased estimators
Maximizing likelihood
Bayesian estimation

Jump Straight to Deep Learning?

Tensor Flow

Jump Straight to Deep Learning?

Understand the theory to help you debug

But another reason…

Computers struggle...

… especially for **human** problems.

Understand the theory
to push for **better systems**

Once upon a time…

…there was parameter estimation

What are Parameters?

Consider some probability distributions:

- Ber(p)
- Poi(λ)
- Uni(α, β)
- Normal(μ, σ^2)
- $Y = mX + b$
- etc…

$\theta = p$

$\theta = \lambda$

$\theta = (\alpha, \beta)$

$\theta = (\mu, \sigma^2)$

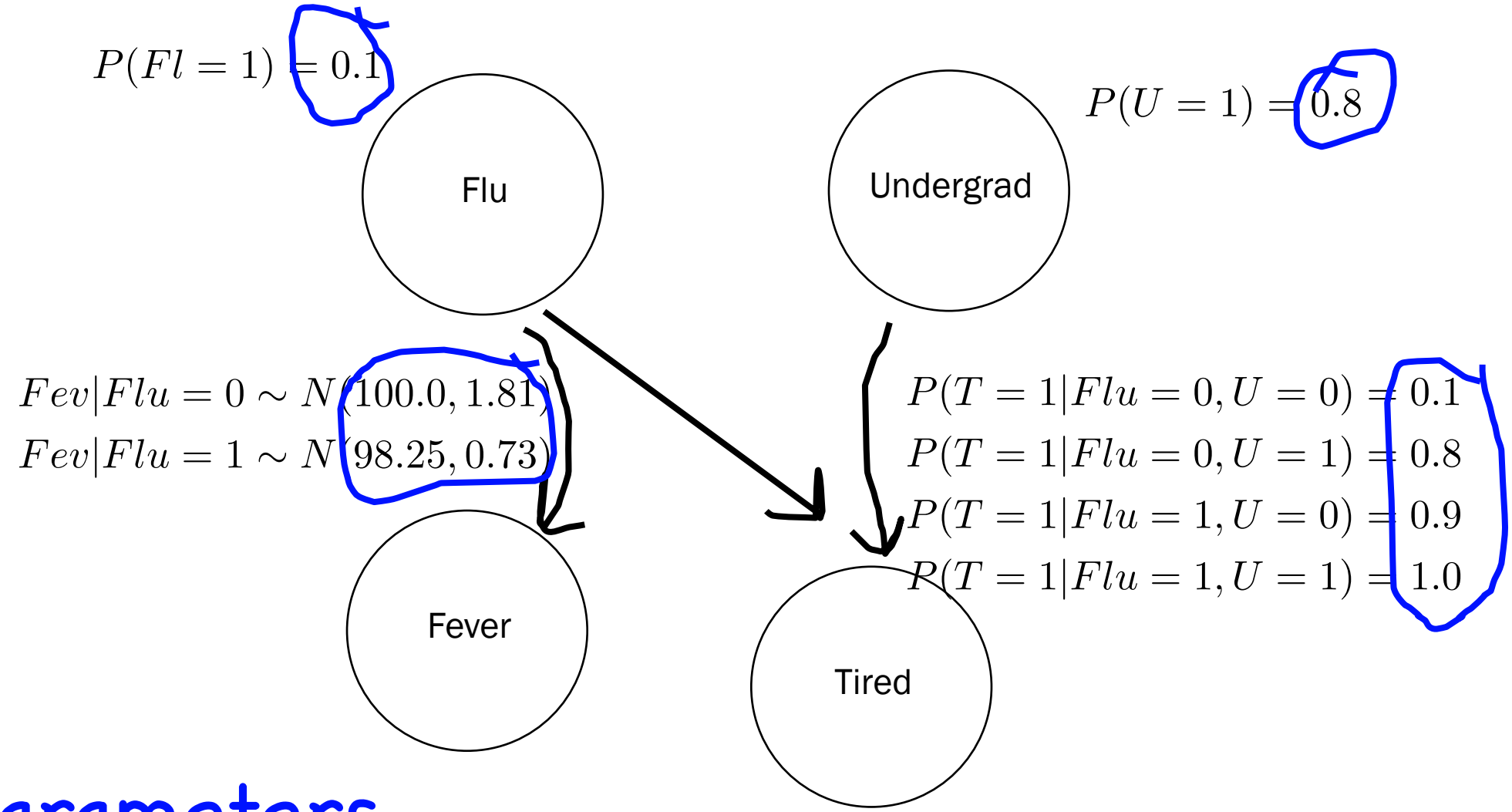
$\theta = (m, b)$

Call these “parametric models”

Given model, **parameters** yield actual distribution

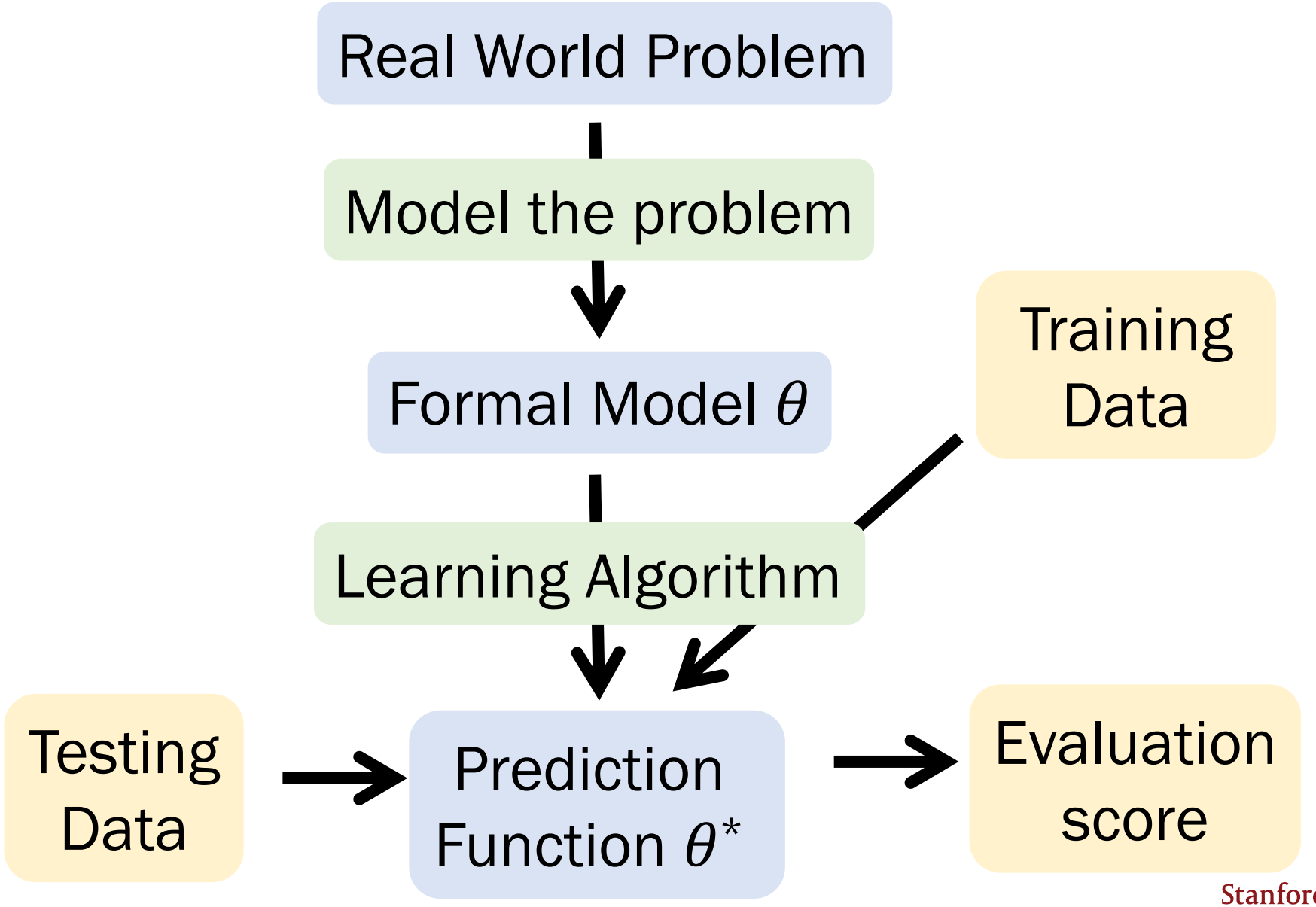
- Usually refer to parameters of distribution as θ
- Note that θ that can be a vector of parameters

What are Parameters?

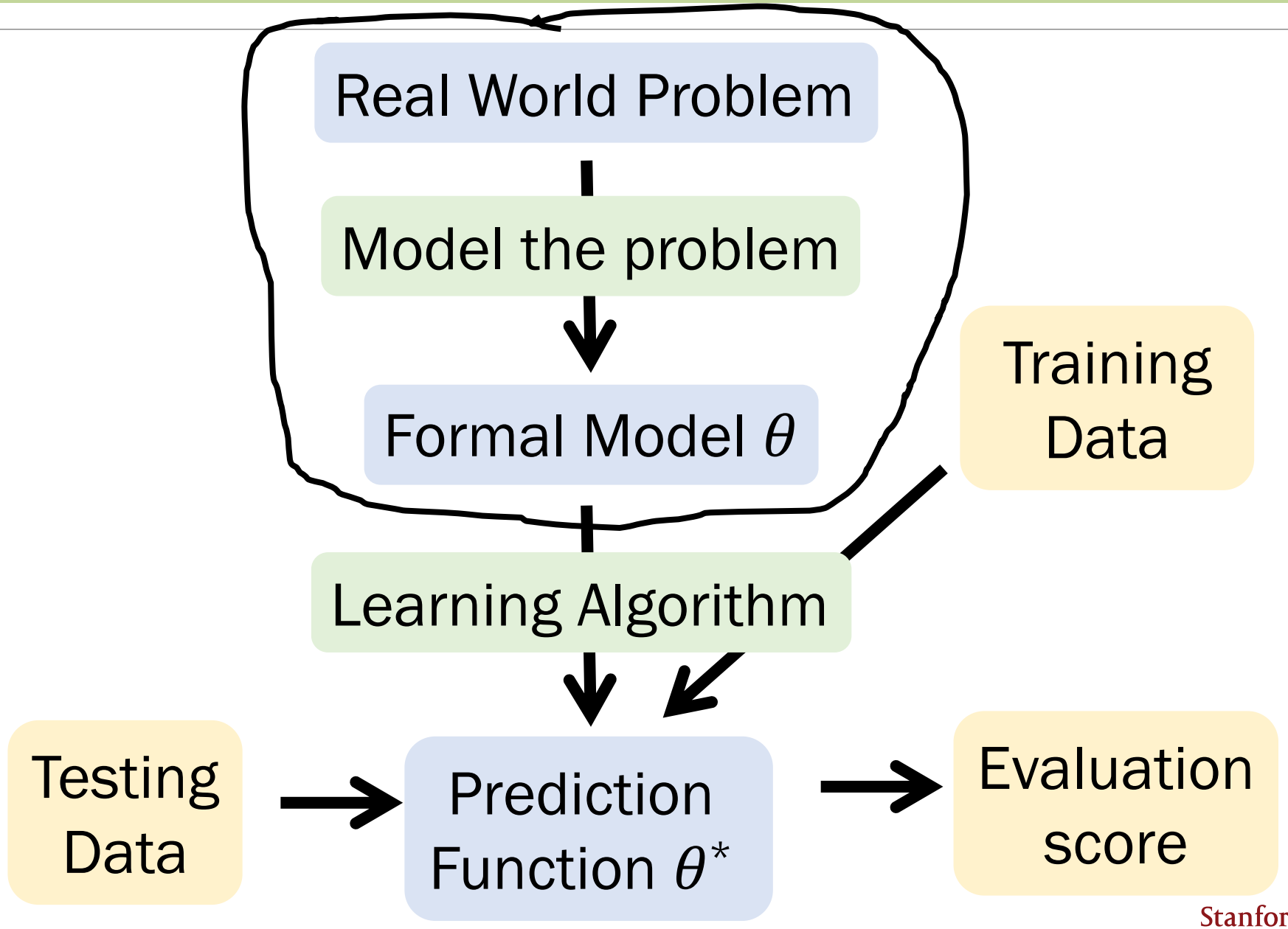


Parameters

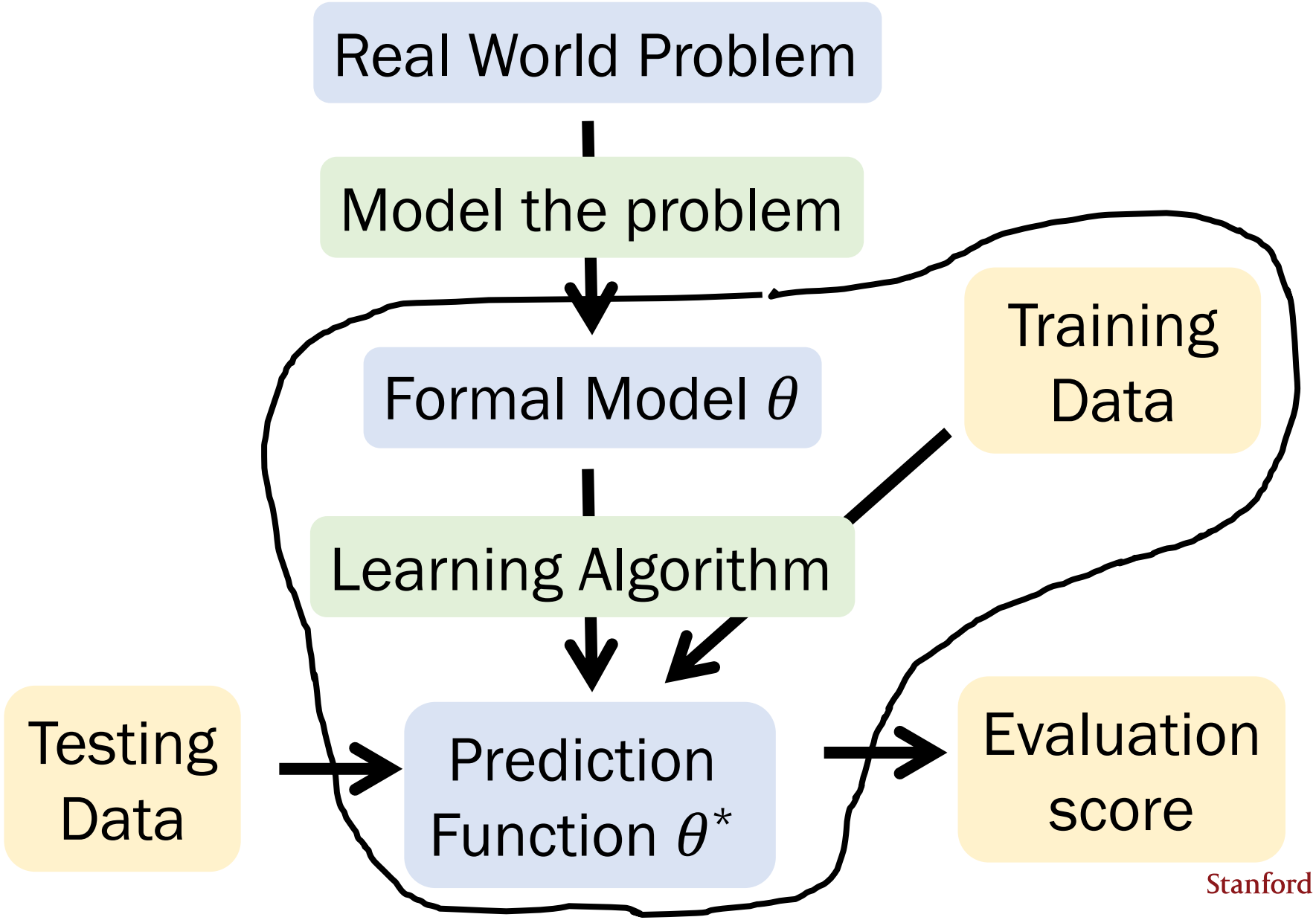
Why Do We Care?



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)



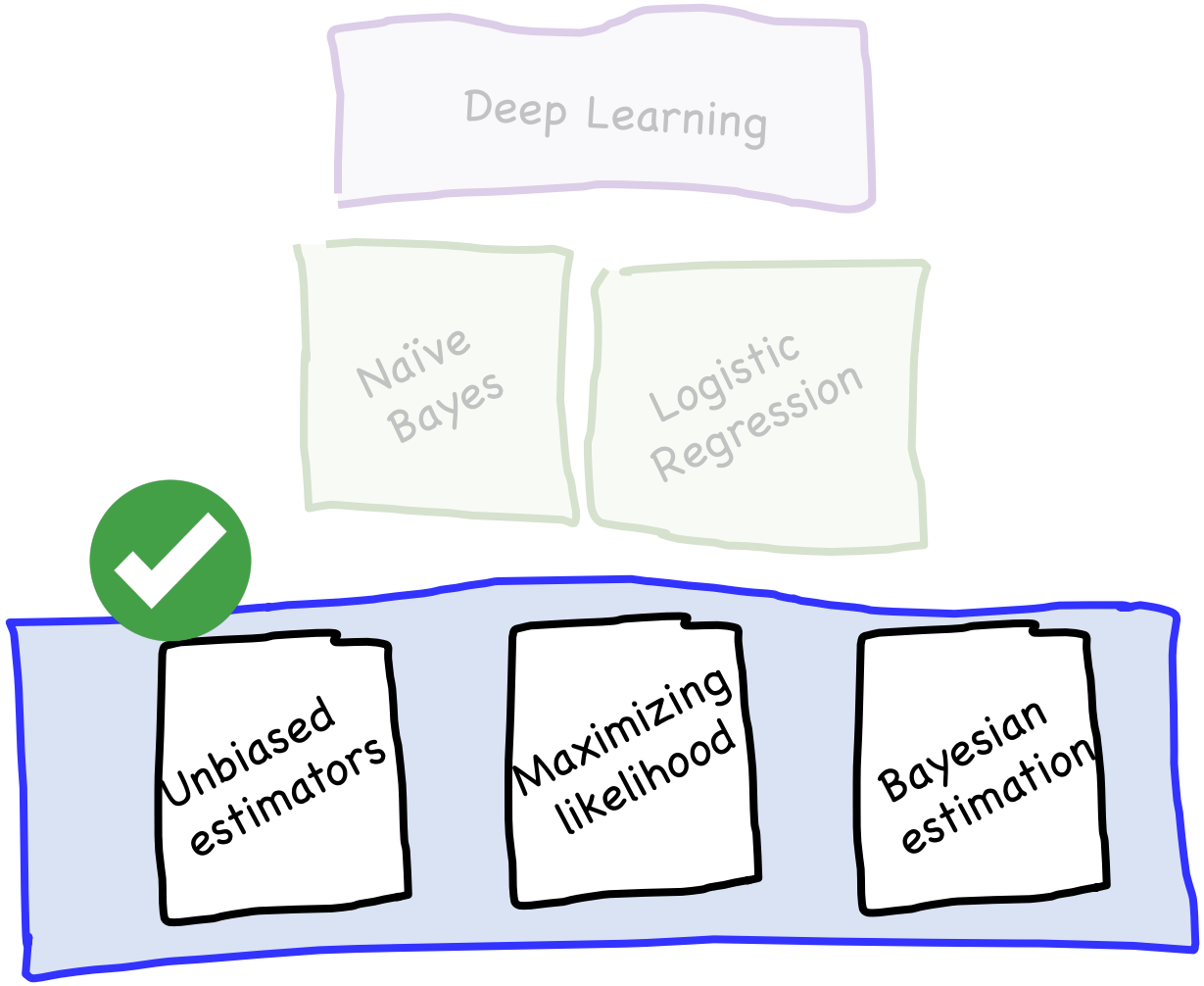
We've already seen some estimations

$X_1, X_2, \ldots, X_n$ are n i.i.d. random variables,
where X_i drawn from distribution F with $E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$.

Sample mean: $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ unbiased **estimate** of μ

Sample variance: $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$ unbiased **estimate** of σ^2

Parameter Estimation



Limited tool: how could we use that for fitting WebMD?

Great idea in Machine Learning


L-R
30
40
50
60

“I feel seen”

Insight: find the arguments that maximize measure of likelihood

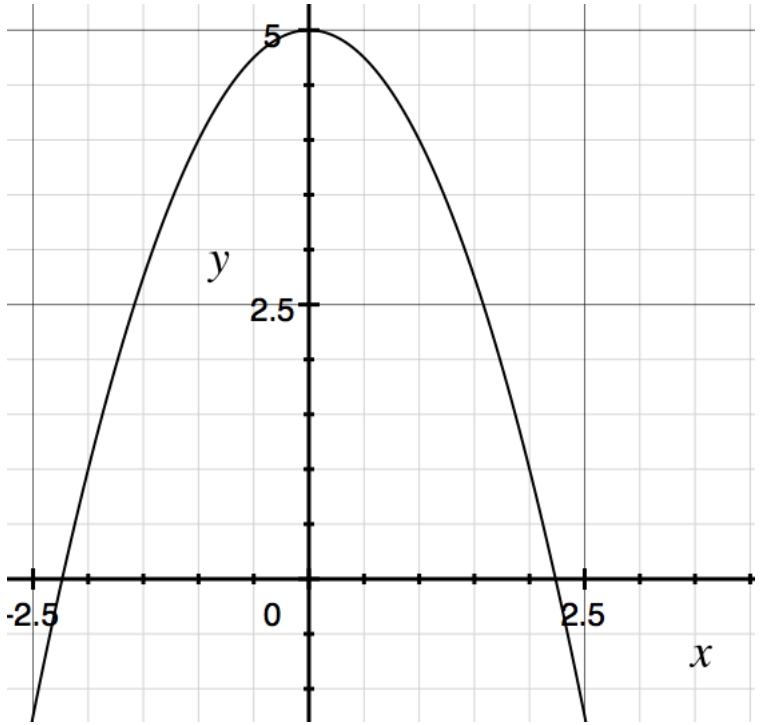
`argmax`

Argmax

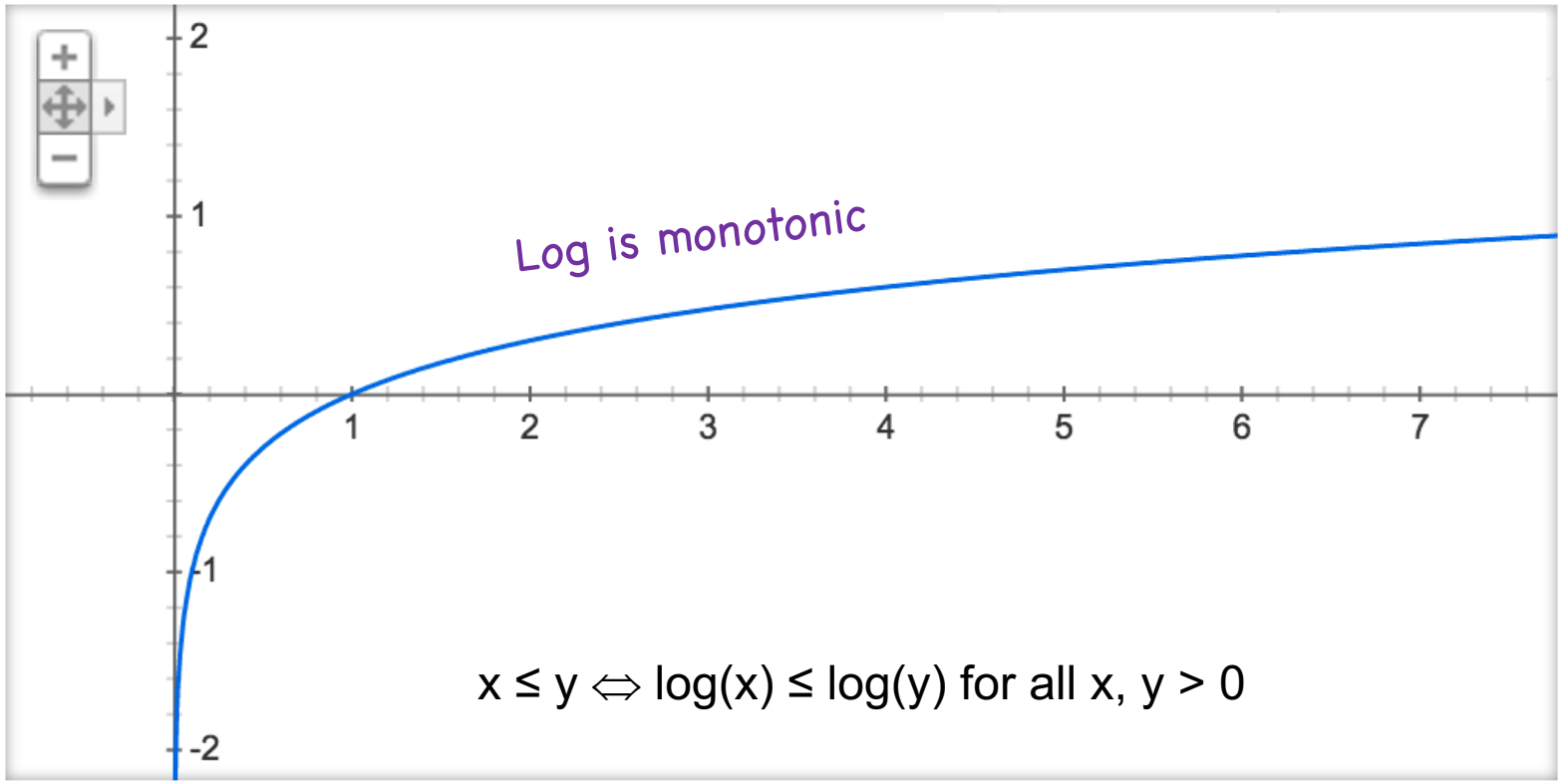
$$f(x) = -x^2 + 5$$

$$\max_x \; -x^2 + 5 = 5$$

$$\operatorname*{argmax}_x \; -x^2 + 5 = 0$$



Argmax of Log



Claim: $\underset{x}{\operatorname{argmax}}\ f(x) = \underset{x}{\operatorname{argmax}}\ \log f(x)$

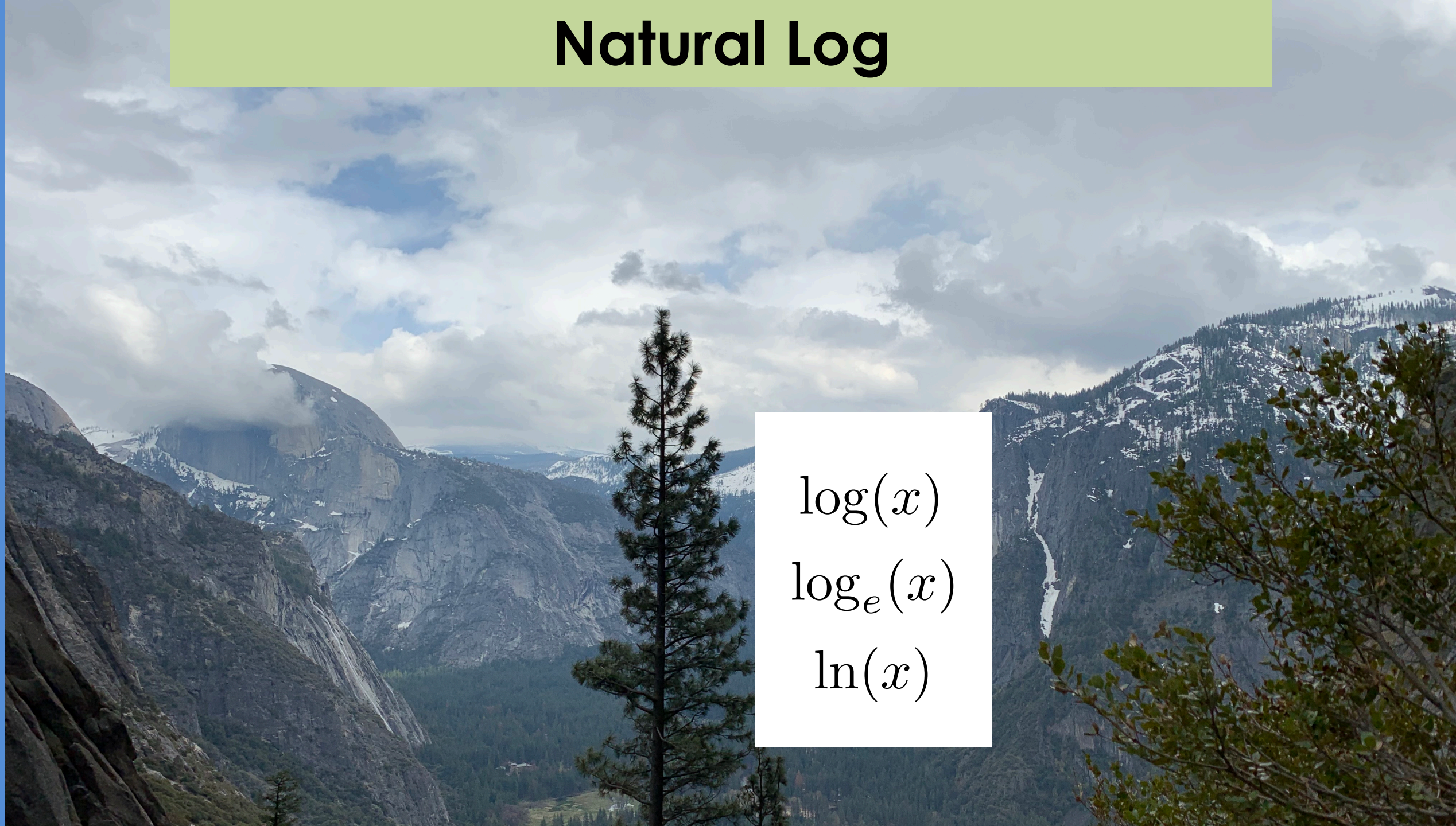
Argmax of Log

$$\underset{x}{\operatorname{argmax}}\ f(x) = \underset{x}{\operatorname{argmax}}\ \log f(x)$$

Log I Love You

$$\log(ab) = \log(a) + \log(b)$$

Natural Log



$$\log(x)$$

$$\log_e(x)$$

$$\ln(x)$$

Maximum Likelihood Algorithm

1. Decide on a model for the distribution of your samples. Define the PMF / PDF for your sample.

2. Write out the log likelihood function.

3. State that the optimal parameters are the argmax of the log likelihood function.

4. Use an optimization algorithm to calculate argmax

The Likelihood Function

n I.I.D. data points $x_1, x_2, \ldots, x_n$

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

This is just a product since X_i are I.I.D.

We explicitly specify parameter θ of distribution

Likelihood (of data given parameters):

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

Either the
PDF (continuous) or
PMF (discrete), or
joint if multiple variables per datapoint

Maximum Likelihood Algorithm

1. Decide on a model for the distribution of your samples. Define the PMF / PDF for your sample.

2. Write out the log likelihood function.

3. State that the optimal parameters are the argmax of the log likelihood function.

4. Use an optimization algorithm to calculate argmax

Story so far: We can chose parameters by finding the argmax of the log likelihood of our data

Maximum Likelihood

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

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

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ LL(\theta)$$

But how do we compute argmax?

Option #1: Straight optimization

Finding the argmax with calculus

$$\hat{x} = \arg\max_x f(x)$$

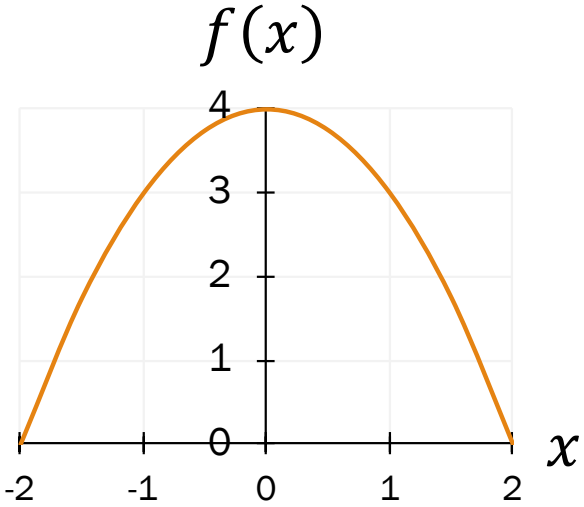
Let $f(x) = -x^2 + 4$, where $-2 < x < 2$.

Differentiate w.r.t. argmax's argument

$$\frac{d}{dx} f(x) = \frac{d}{dx}(-x^2 + 4) = -2x$$

Set to 0 and solve

$$-2x = 0 \quad \Rightarrow \quad \hat{x} = 0$$



Make sure $\hat{x}$ is a maximum

- Check $f(\hat{x} \pm \epsilon) < f(\hat{x})$
- Generally ignored in expository derivations
- We'll ignore it here too (and won't require it in class)
- arg min is defined similarly, relevant for gradient descent

General MLE Formula

Consider I.I.D. data: $X_1, X_2, ..., X_n$. Assume a model.

Use Maximum Likelihood to estimate parameters

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

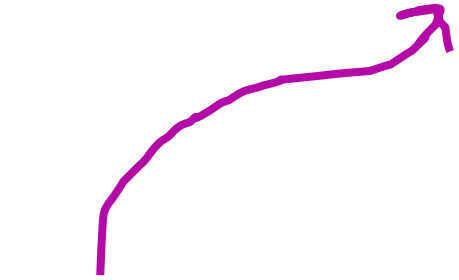
MLE for Poisson

$$X \sim \text{Poi}(\lambda)$$

MLE for Poisson

$$X \sim \text{Poi}(\lambda)$$

We observed the following samples:
[6, 1, 2, 1, 2, 3, 3, 2, 1, 3, 1, 3]



x_i

What is lambda?

Maximum Likelihood with Poisson

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

- $X_i \sim \text{Poi}(\lambda)$ **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

Maximum Likelihood with Poisson

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

- $X_i \sim \text{Poi}(\lambda)$ **Use Maximum Likelihood to estimate λ**
- Probability mass function can be written as: $f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

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

Maximum Likelihood with Poisson

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

- $X_i \sim \text{Poi}(\lambda)$ **Use Maximum Likelihood to estimate λ**
- Probability mass function can be written as: $f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Likelihood: $L(\lambda) = f(x_1 \dots x_n|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

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

4. Find the value of **λ** which maximizes log likelihood

Maximum Likelihood with Poisson

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

- $X_i \sim \text{Poi}(\lambda)$ **Use Maximum Likelihood to estimate λ**
- Probability mass function can be written as: $f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Likelihood: $L(\lambda) = f(x_1 \dots x_n|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Log-likelihood:

$$LL(\lambda) = \log \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \sum_{i=1}^{n} \log \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \sum_{i=1}^{n} -\lambda + x_i \log \lambda - \log x_i!$$

4. Find the value of λ which maximizes log likelihood

Maximum Likelihood with Poisson

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

- $X_i \sim \text{Poi}(\lambda)$ **Use Maximum Likelihood to estimate λ**
- Probability mass function can be written as: $f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Likelihood: $L(\lambda) = f(x_1 \dots x_n|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Log-likelihood:

$$LL(\lambda) = \log \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \sum_{i=1}^{n} \log \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \sum_{i=1}^{n} -\lambda + x_i \log \lambda - \log x_i!$$

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

$$\frac{\partial LL(\lambda)}{\partial \lambda} = \sum_{i=1}^{n} -1 + \frac{x_i}{\lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i \quad 0 = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i \quad \lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Isn’t that the same as
the sample mean?

Yes. For Poisson.

MLE of Poisson is the sample mean

MLE for Gaussian

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

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]
```

What are the parameters?

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

$$f(X_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}$$

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

$$LL(\theta) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}\right) = \sum_{i=1}^{n} \left[-\log(\sqrt{2\pi}\sigma) - (X_i-\mu)^2/(2\sigma^2)\right]$$

(using natural log)

$$= -\sum_{i=1}^{n} \log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n} [(X_i-\mu)^2/(2\sigma^2)]$$

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

$$f(X_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}$$

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

$$LL(\theta) = -\sum_{i=1}^{n} \log\left(\sqrt{2\pi}\sigma\right) - \sum_{i=1}^{n} [(X_i-\mu)^2/(2\sigma^2)]$$

with respect to μ

$$\frac{\partial LL(\theta)}{\partial \mu} = \sum_{i=1}^{n} [2(X_i-\mu)/(2\sigma^2)]$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0$$

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

$$f(X_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}$$

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

$$LL(\theta) = -\sum_{i=1}^{n} \log\left(\sqrt{2\pi}\sigma\right) - \sum_{i=1}^{n} [(X_i-\mu)^2/(2\sigma^2)]$$

with respect to μ

$$\frac{\partial LL(\theta)}{\partial \mu} = \sum_{i=1}^{n} [2(X_i-\mu)/(2\sigma^2)]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0$$

with respect to σ

$$\frac{\partial LL(\theta)}{\partial \sigma} = -\sum_{i=1}^{n}\frac{1}{\sigma} + \sum_{i=1}^{n} 2(X_i-\mu)^2/(2\sigma^3)$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2 = 0$$

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

$$f(X_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}$$

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

3. Solve resulting equations

Two equations, two unknowns:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0 \qquad -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i - \mu)^2 = 0$$

First, solve for μ_{MLE}:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i - \frac{1}{\sigma^2}\sum_{i=1}^{n}\mu = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} X_i = n\mu \quad \Rightarrow \quad \mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

unbiased

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

$$f(X_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}$$

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

3. Solve resulting equations

Two equations, two unknowns:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0 \qquad -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2 = 0$$

First, solve for μ_{MLE}:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i - \frac{1}{\sigma^2}\sum_{i=1}^{n}\mu = 0 \quad \Rightarrow \sum_{i=1}^{n} X_i = n\mu \quad \Rightarrow \mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

unbiased

Next, solve for σ_{MLE}:

$$\frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{n}{\sigma} \quad \Rightarrow \sum_{i=1}^{n}(X_i-\mu)^2 = \sigma^2 n \quad \Rightarrow \sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_{MLE})^2$$

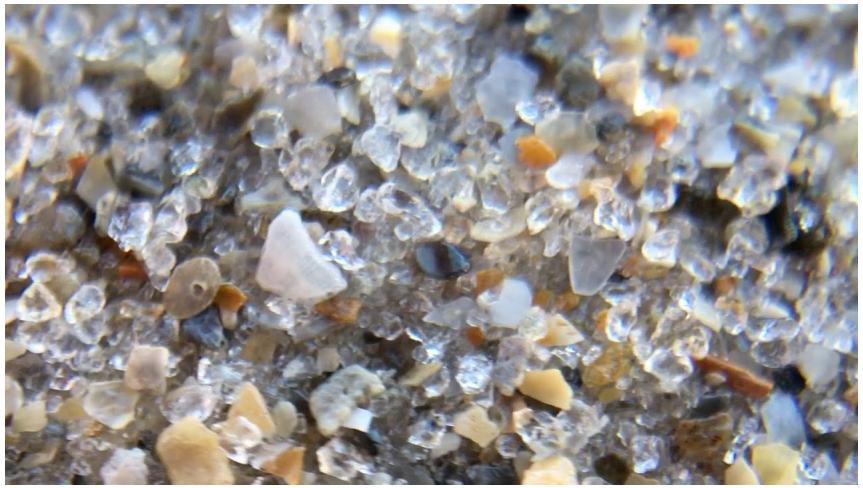
biased

Recreated the sample mean and (biased) version of variance

Can Apply it Even to a Novel Distribution!

```
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

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

https://chrispiech.github.io/probabilityForComputerScientists/en/examples/mle_pareto/

MLE for Bernoulli

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

Don’t we already have the Beta?

Yes! But this example is critical for developing towards deep learning.

Maximum Likelihood with Bernoulli

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = p_{MLE}$?

- Let $X_i \sim \text{Ber}(p)$.

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

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

?

$$f(X_i|p) = \begin{cases} p & \text{if } X_i = 1 \\ 1-p & \text{if } X_i = 0 \end{cases}$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

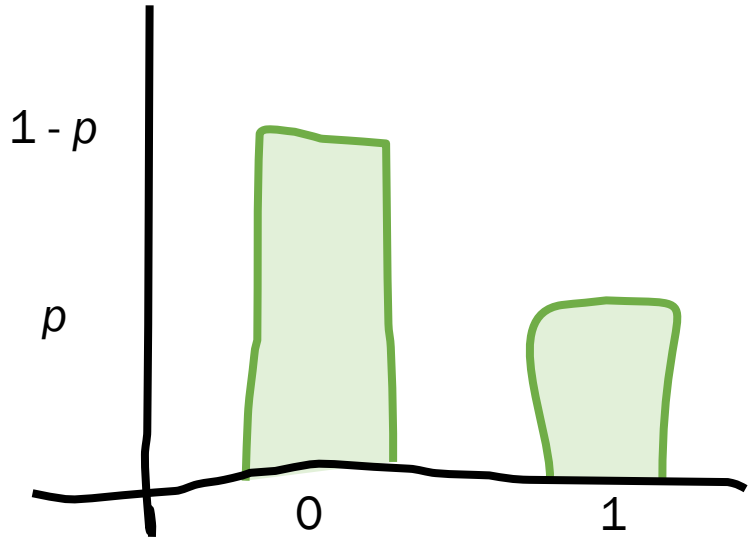
3. Solve resulting equations

Differentiable PMF for Bernoulli

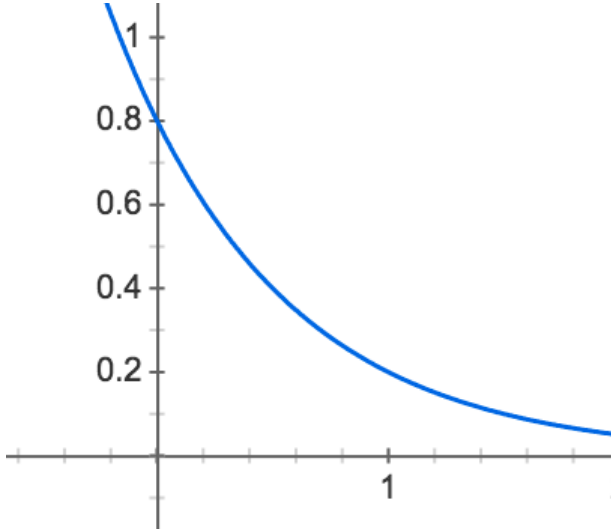
Consider I.I.D. random variables $X_1, X_2, ..., X_n$

- $X_i \sim \text{Ber}(p)$
- Probability mass function, $f(X_i = x_i | P = p)$

PMF of Bernoulli



PMF of Bernoulli (p = 0.2)



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

$$f(x_i|p = 0.2) = 0.2^{x_i}(1-0.2)^{1-x_i}$$

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

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

$$\prod_{i=1}^{n}$$

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

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

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

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

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

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

4. Find the value of p which maximizes log likelihood

Maximizing Likelihood with Bernoulli

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

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

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

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

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

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

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

Isn’t that the same as
unbiased estimator?

Yes. For Bernoulli.

Maximum Likelihood Algorithm

1. Decide on a model for the distribution of your samples. Define the PMF / PDF for your sample.

2. Write out the log likelihood function.

3. State that the optimal parameters are the argmax of the log likelihood function.

4. Use an optimization algorithm to calculate argmax

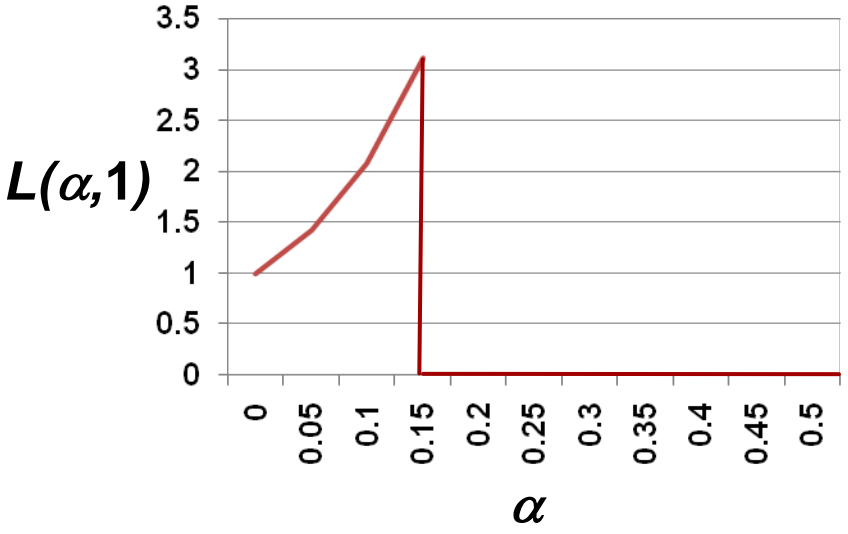
Its so general!

Understanding MLE with Uniform

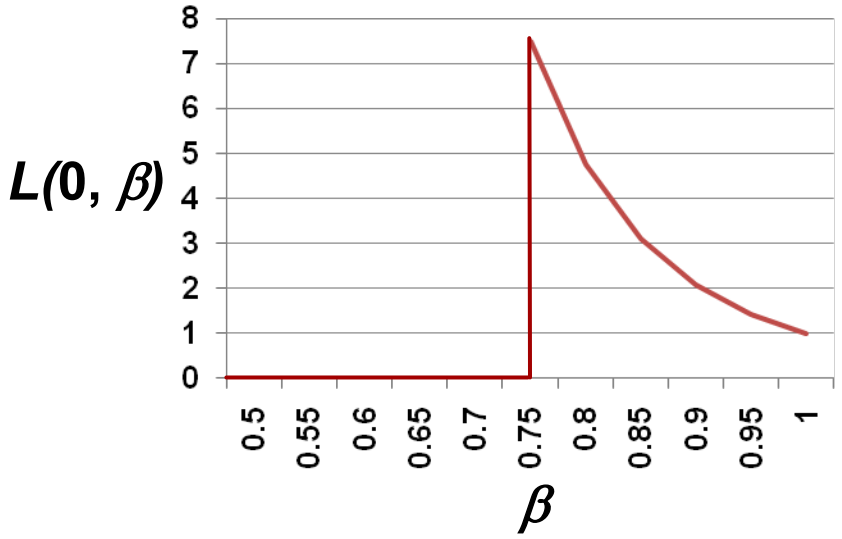
Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$

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

Likelihood: L(α,1)



Likelihood: L(0, β)



Small Samples = Problems

How do small samples affect MLE?

- In many cases, $\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$ = sample mean
  - Unbiased. Not too shabby…
- As seen with Normal, $\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{MLE})^2$
  - Biased. Underestimates for small *n* (e.g., 0 for n = 1)
- As seen with Uniform, $\alpha_{MLE} \geq \alpha$ and $\beta_{MLE} \leq \beta$
  - Biased. Problematic for small *n* (e.g., $\alpha = \beta$ when n = 1)
- Small sample phenomena intuitively make sense:
  - Maximum likelihood $\Rightarrow$ best explain data we've seen
  - Does not attempt to generalize to unseen data

Properties of MLE

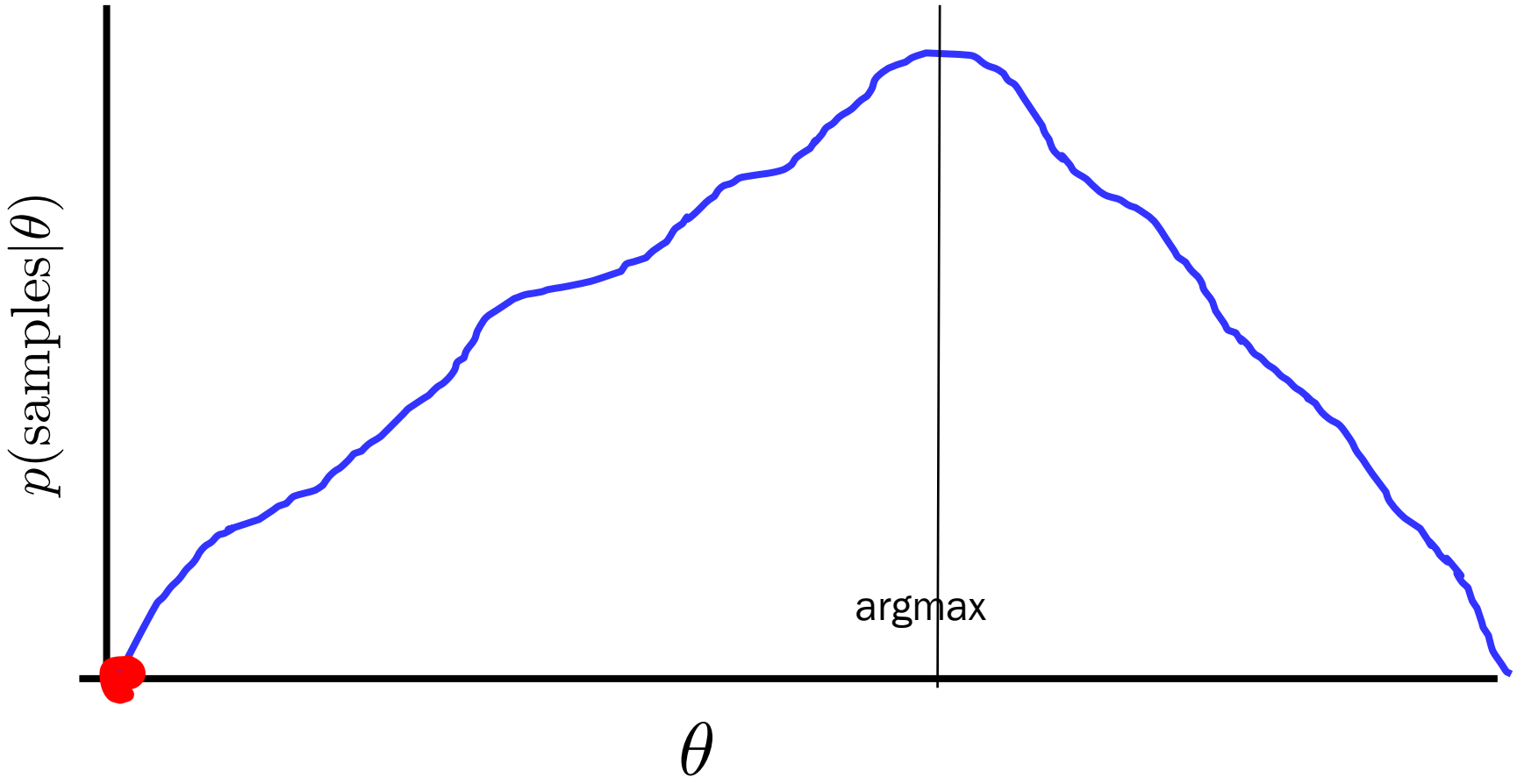
Maximum Likelihood Estimators are generally:

- **Asymptotically optimal** $\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$ for $\varepsilon > 0$
- **Potentially biased** (though asymptotically less so)
- **Often used in practice**

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

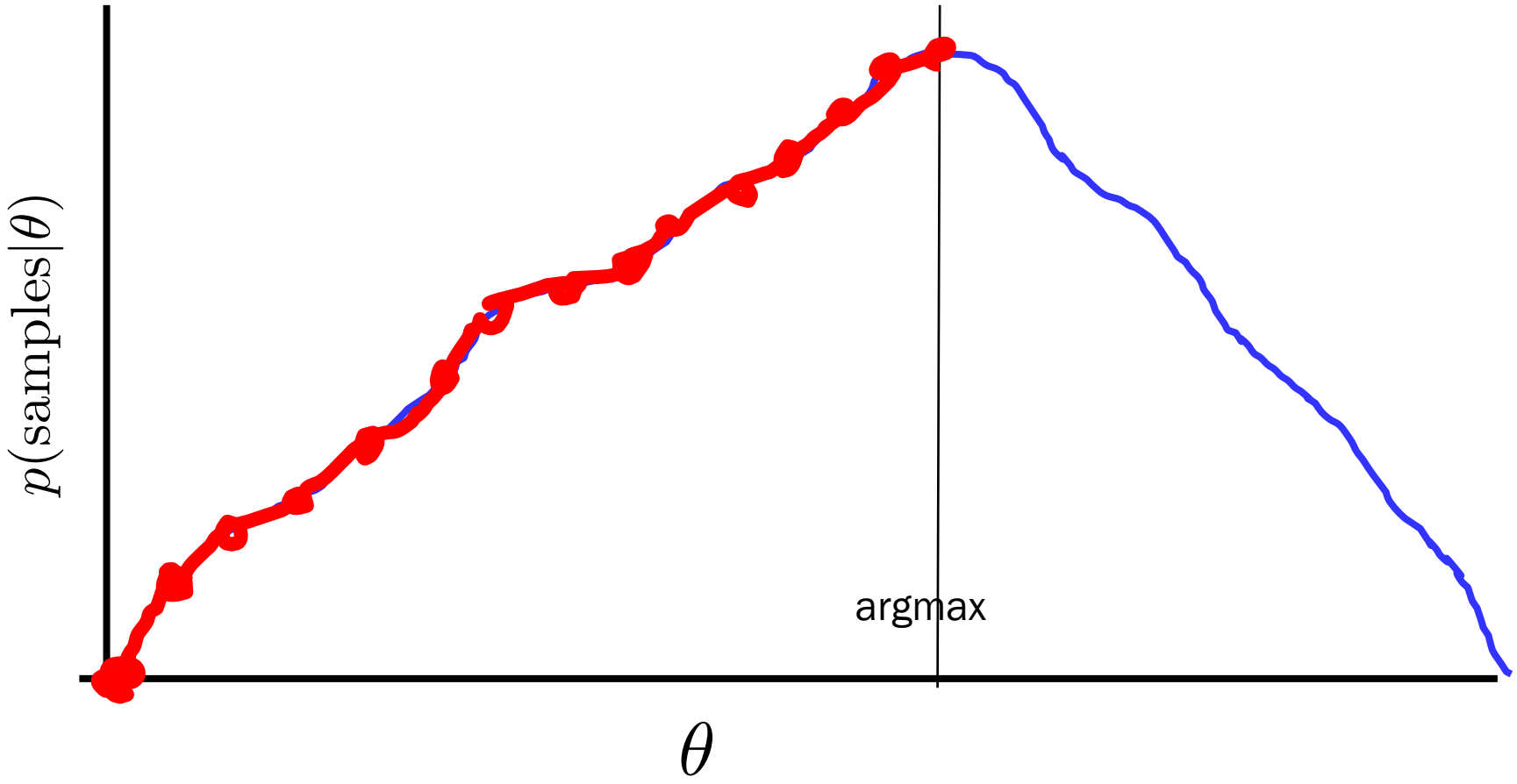
Optimization (argmax)
Option #2: Gradient Descent

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

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 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:

`gradient[j] = 0 for all` $0 \leq j \leq m$

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

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

Gradient Ascent

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

Repeat many times:

`gradient[j] = 0 for all` $0 \leq j \leq m$

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

$$\frac{dLL(\vec{\theta})}{d\mu_a} = \sum_i^n \frac{d}{d\mu_a}\left[-\frac{1}{2}\left(\frac{x_i - \mu_a}{\sigma_a}\right)^2\right]$$

$$= \sum_i^n 2\left(\frac{x_i - \mu_a}{\sigma_a}\right)\frac{1}{\sigma_a}$$

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

Gradient Ascent

Repeat many times

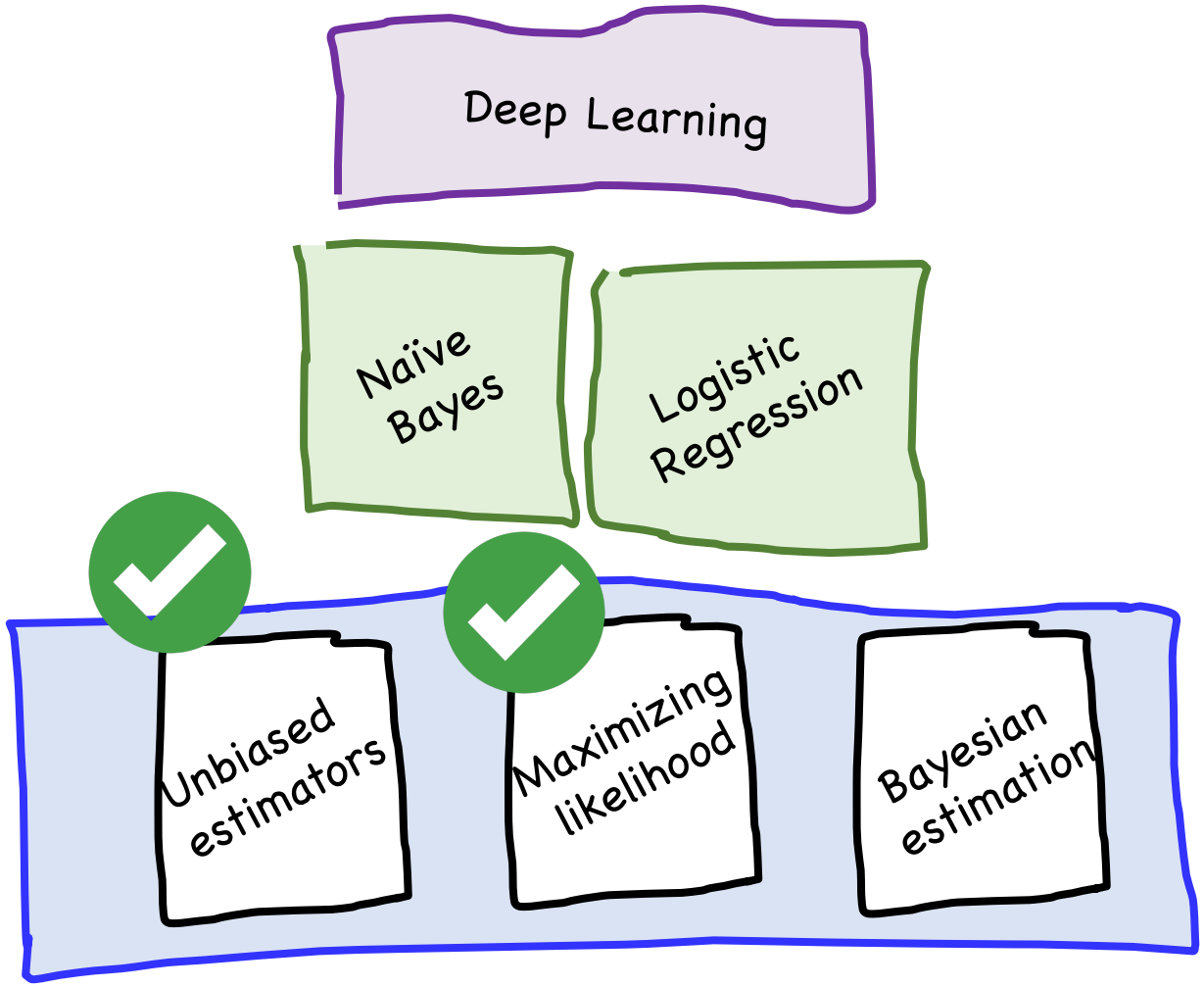
$$\theta_j^{\text{new}} = \theta_j^{\text{old}} + \eta \cdot \frac{\partial LL(\theta^{\text{old}})}{\partial \theta_j^{\text{old}}}$$

This is some **profound** life philosophy

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

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

Our Path


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

Likelihood Definition

Wikipedia:

Likelihood function

Article Talk

From Wikipedia, the free encyclopedia

The **likelihood function** (often simply called the **likelihood**) is the joint probability (or probability density) of observed data viewed as a function of the parameters of a statistical model.[1] [2] [3]

A generalized term for "PDF / PMF / Joint" of data as a function of parameters

MLE in a nutshell

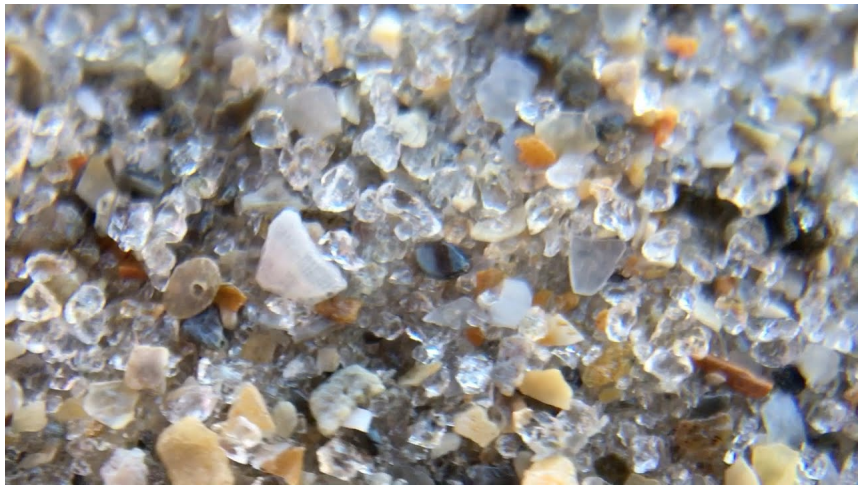
$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}}\ f(x^{(1)}, \ldots, x^{(n)}|\theta)$$

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

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 $X_1, X_2, \ldots, X_n$

- $X_i \sim$ 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 $X_1, X_2, \ldots, X_n$

- $X_i \sim$ 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:

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

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

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

Thinking Deeper About Two Envelopes

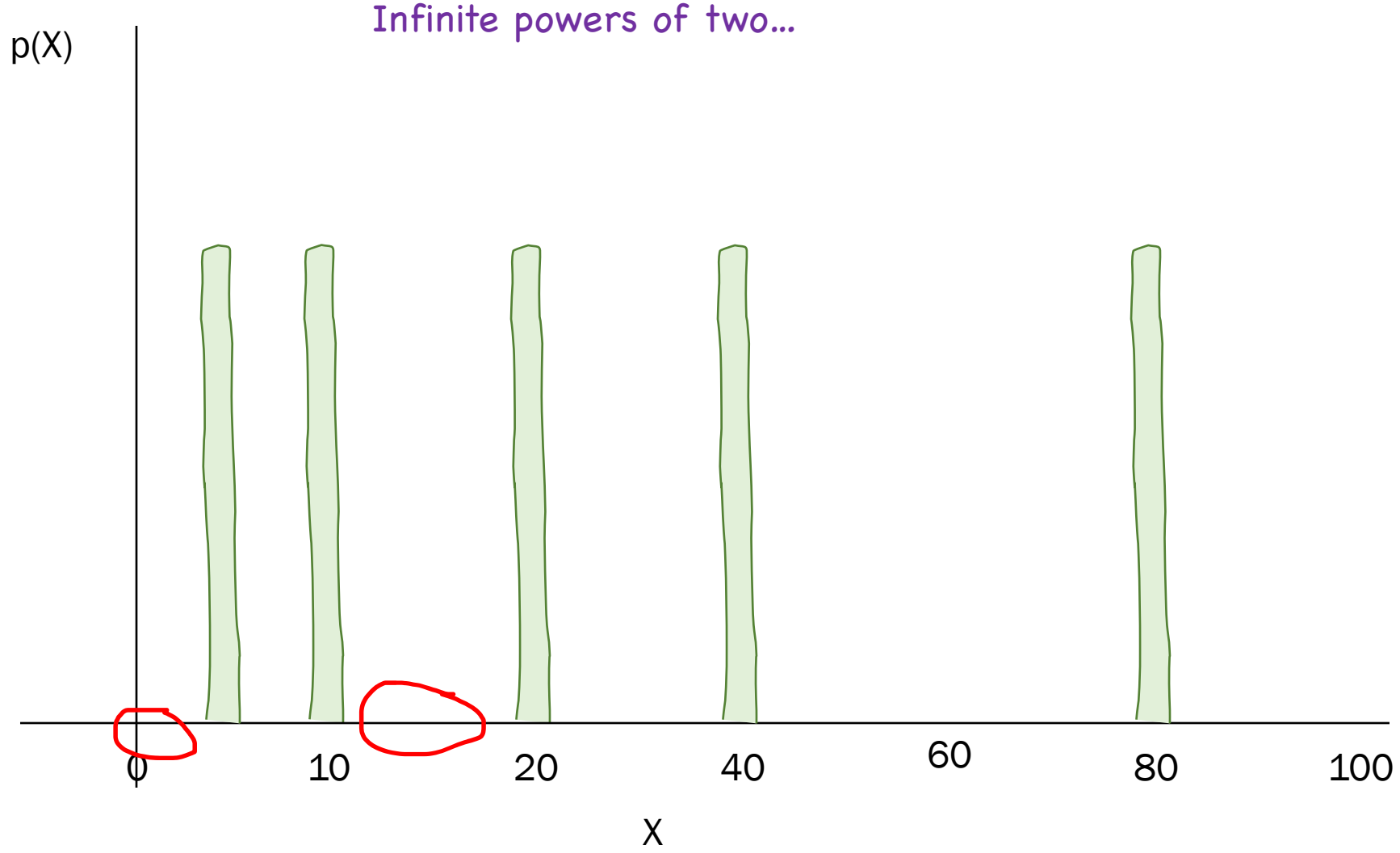
The "two envelopes" problem set-up

- 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?

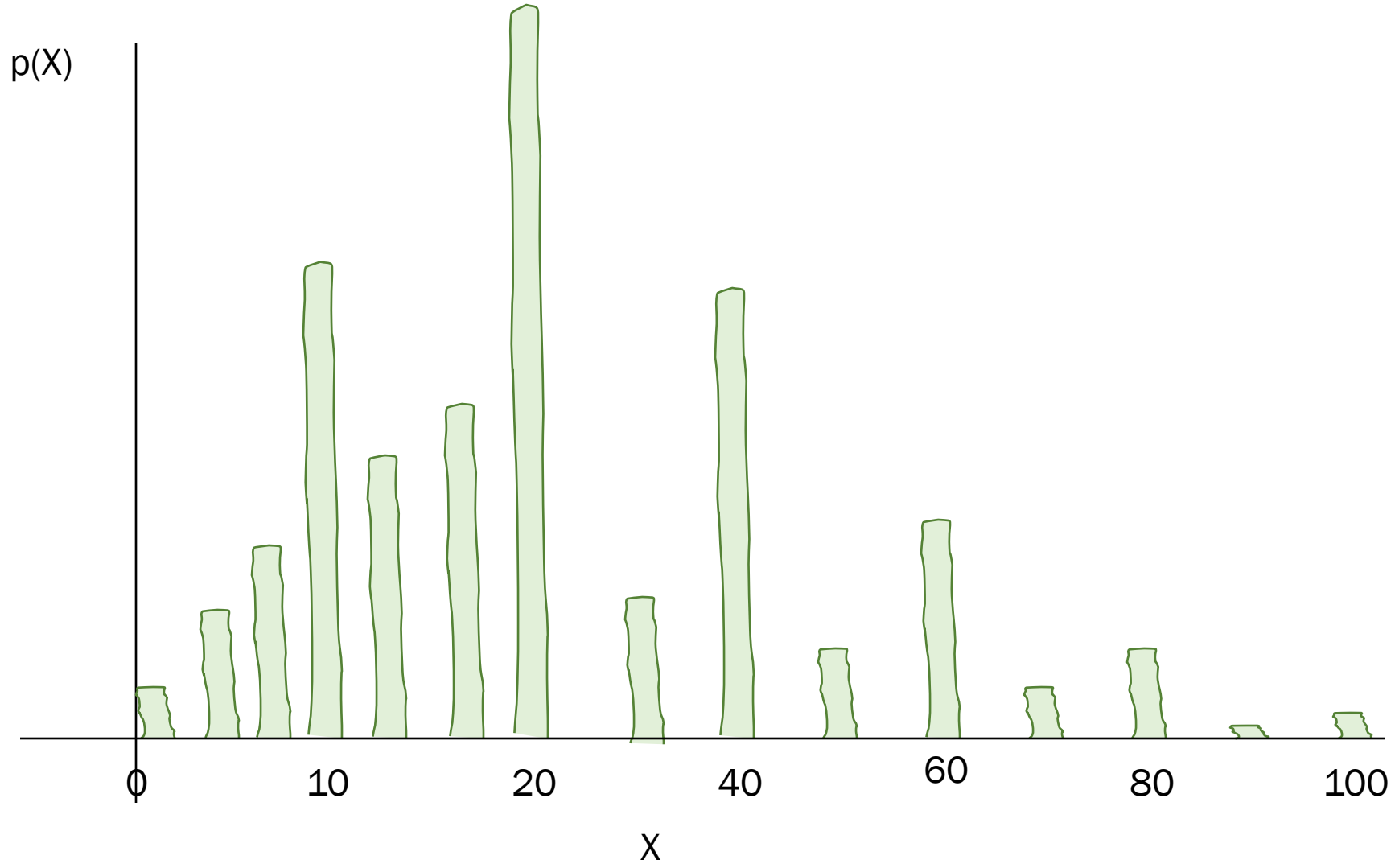


Subjectivity of Probability

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…

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

We have seen this play out before…

MLE vs Beta

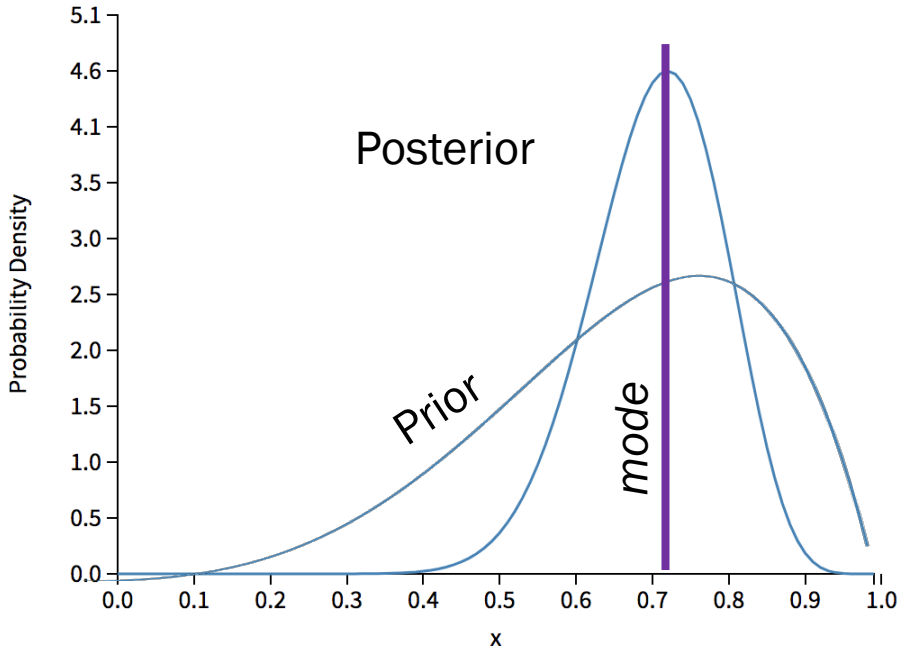
The medicine is tried on 20 patients. It “works” for 14 and “doesn’t work” for 6. What is your new belief that the drug works?

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:



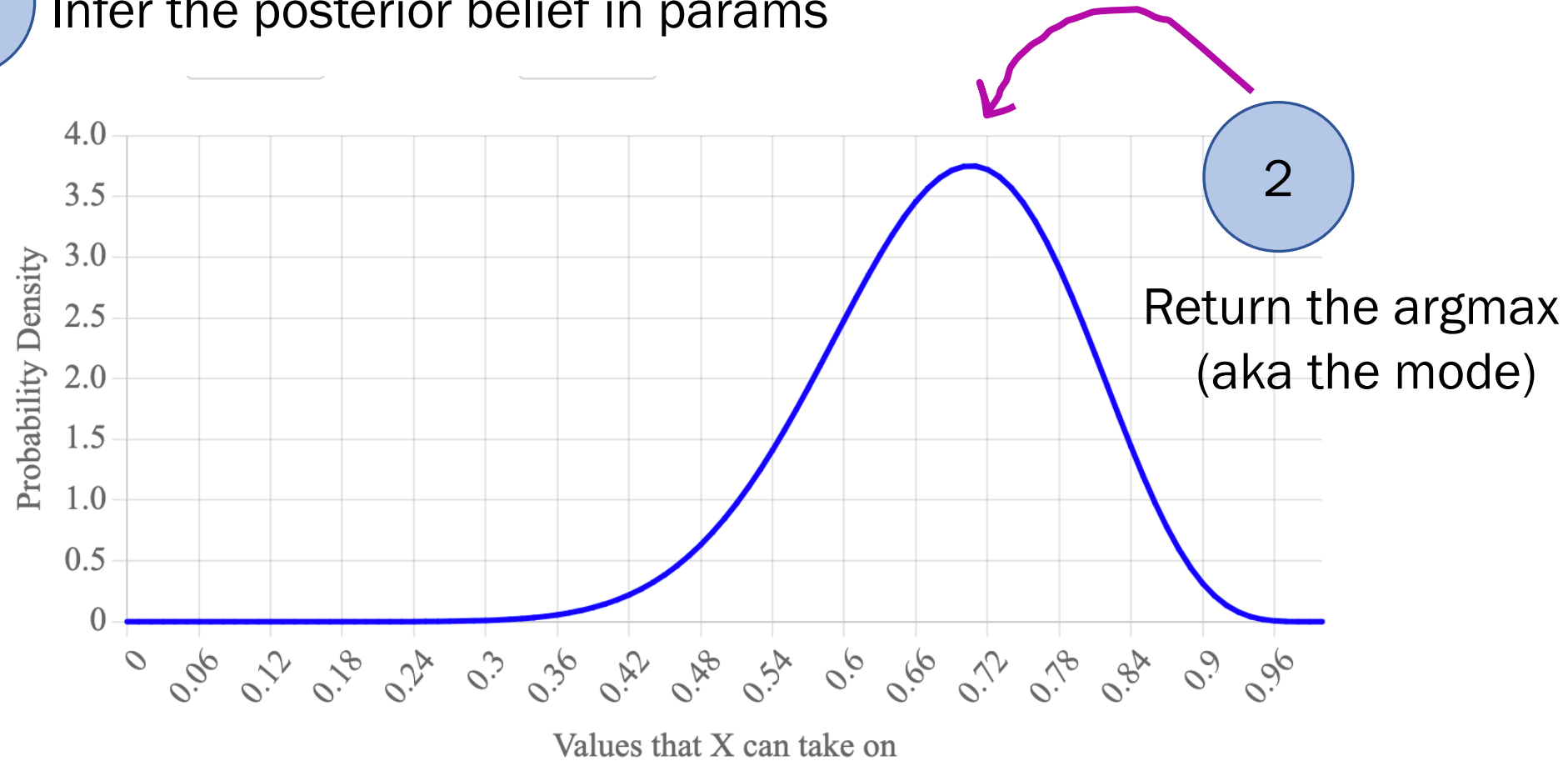
Could we use “Bayesian Inference” to estimate parameters?

Could we use “Bayesian Inference” to estimate parameters?

Yes! MAP!

Maximum A Posteriori (MAP)

1 Infer the posterior belief in params



2 Return the argmax (aka the mode)

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

The Laplace Prior!

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$(a = 2, b = 2)$
Saw 2 imaginary trials: 1 successes, 1 failures

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

Posterior Beta$(2 + n, 2 + m)$

MAP: $$p = \frac{n+1}{n+m+2}$$

MAP with a Laplace Prior: Why its justified to estimate p like this:

$$p_{\text{MAP}} = \frac{n+1}{n+m+2}$$

n = # observed successes

m = # observed fails

But MLE works for more than just estimating p

But now you need a prior belief for all params…

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	Prior 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 wont cover the Dirichlet or Inverse Gamma in lecture!

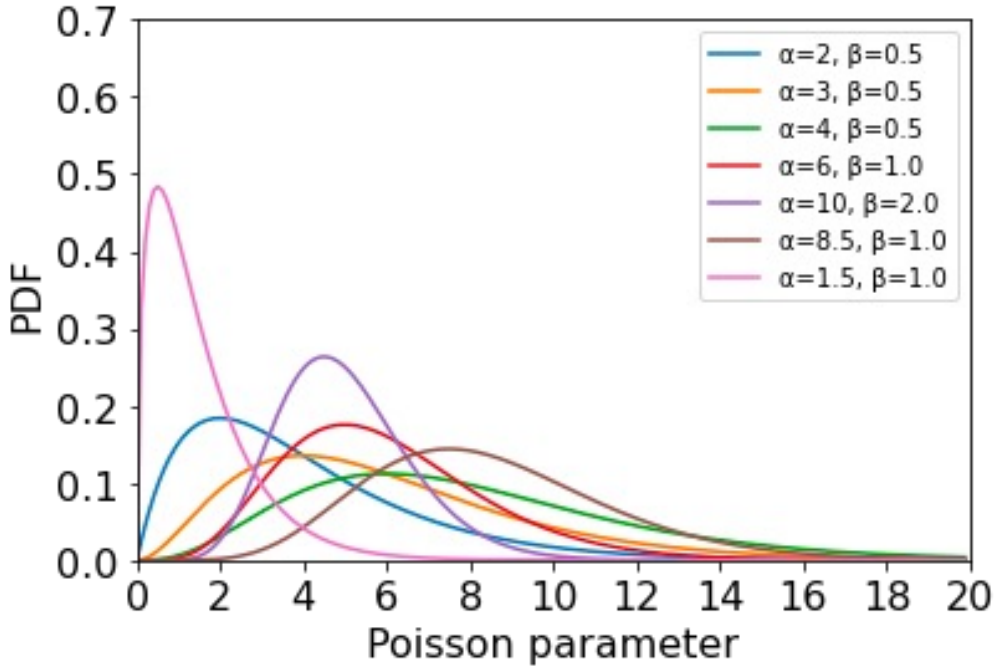
The Gamma Distribution

Random variable for Poisson param

$$\Lambda \sim \text{Gamma}(\alpha, \beta)$$

Saw $\alpha - 1$ total imaginary events

During β imaginary periods



PDF

$$f(\Lambda = \lambda) = K \cdot \lambda^{\alpha-1} e^{-\beta\lambda}$$

Mode

$$\text{mode}(\Lambda) = \frac{\alpha - 1}{\beta}$$

Good times with Gamma

$$f(\Lambda = \lambda) = K \cdot \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\text{mode}(\Lambda) = \frac{\alpha - 1}{\beta}$$

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 $\Lambda \sim \text{Gamma}(\alpha, \beta)$

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

Experiment Observe n events during next k time periods

Posterior On the board!

Good times with Gamma

$$f(\Lambda = \lambda) = K \cdot \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\text{mode}(\Lambda) = \frac{\alpha - 1}{\beta}$$

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 $\Lambda \sim \text{Gamma}(\alpha, \beta)$

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

Experiment Observe n events during next k time periods

Posterior

$$\lambda_{\text{MAP}} = \frac{\alpha + n - 1}{\beta + k}$$

MAP for Poisson

$$\lambda_{\text{MAP}} = \frac{\alpha + n - 1}{\beta + k}$$

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

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

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

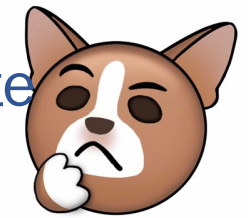
Now perform the experiment and see 11 events in next 2 time periods.

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

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

3. What is λ_{MAP}?

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



That was easy because Gamma is a conjugate…

Can we generalize?

Most important derivation 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

TODO: drop the g notation???

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

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

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:

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

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

MAP for Pareto

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

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

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

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

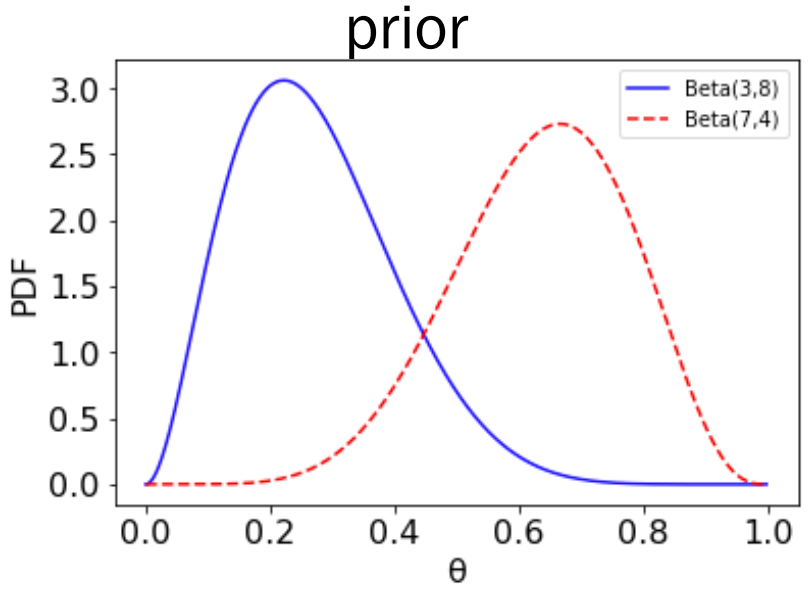
$$= K + \frac{-(\alpha - 2)^2}{6} + n \log \alpha - (\alpha + 1) \sum_{i=1}^{n} \log x_i$$

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

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

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: $\frac{6}{9}$



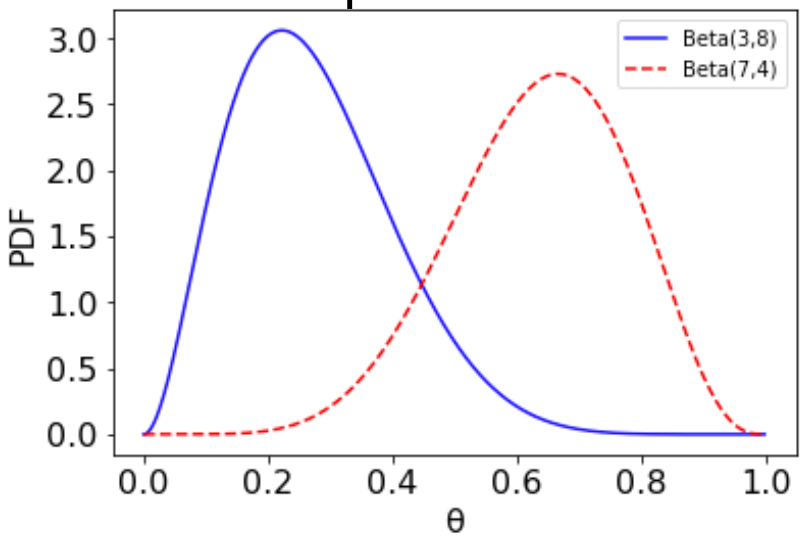
Now flip 100 coins and get 58 heads and 42 tails.

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

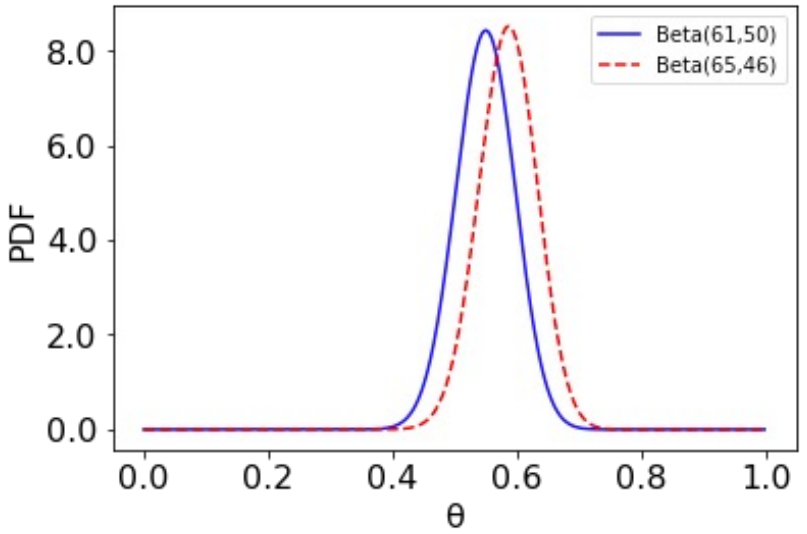


Now flip 100 coins and get 58 heads and 42 tails.

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

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

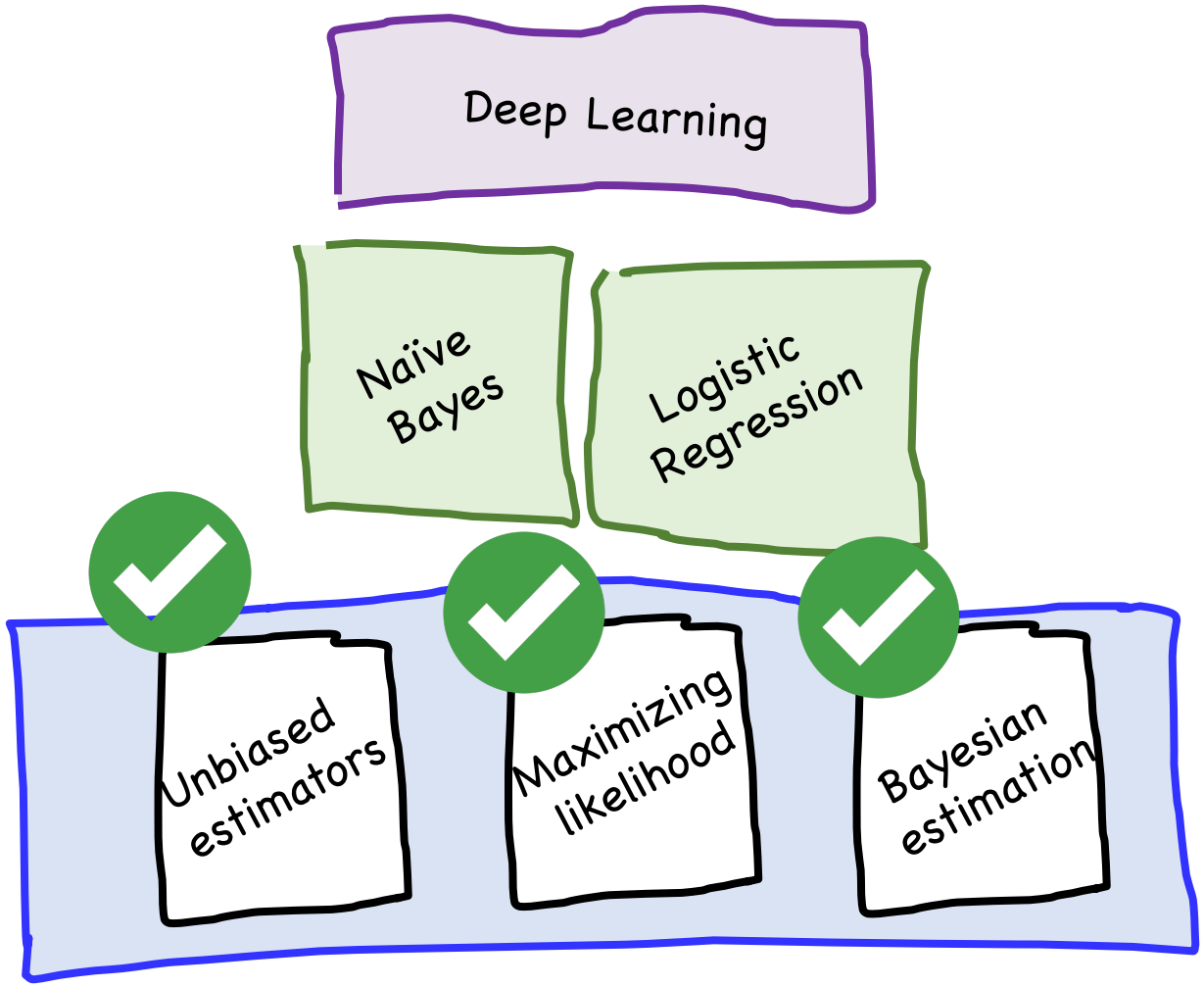
posterior



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

The last estimator has risen…

Our Path


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