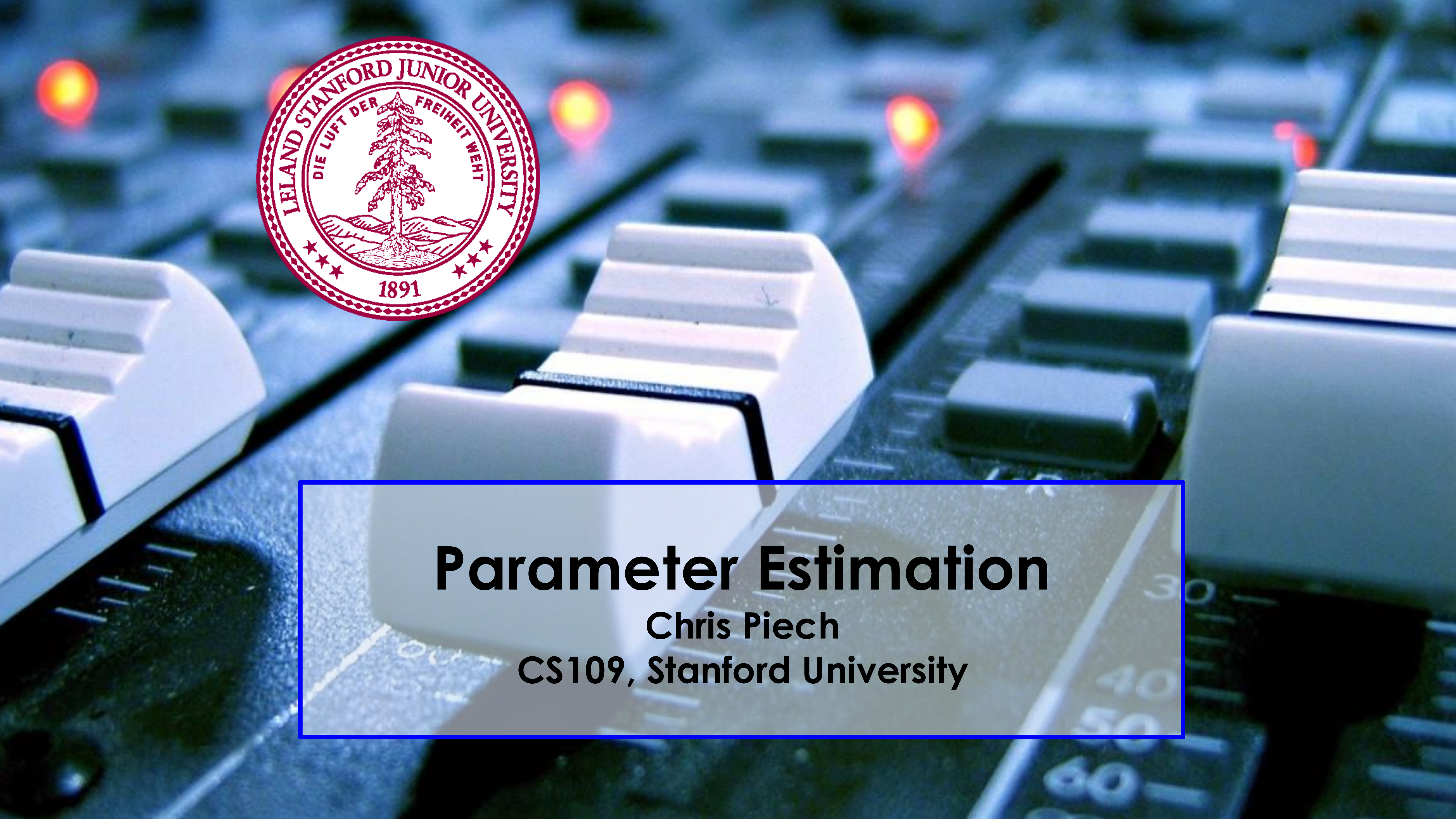


Parameter Estimation

Chris Piech

CS109, Stanford University

Where are we in CS109?

You are here

Counting Theory

Core Probability

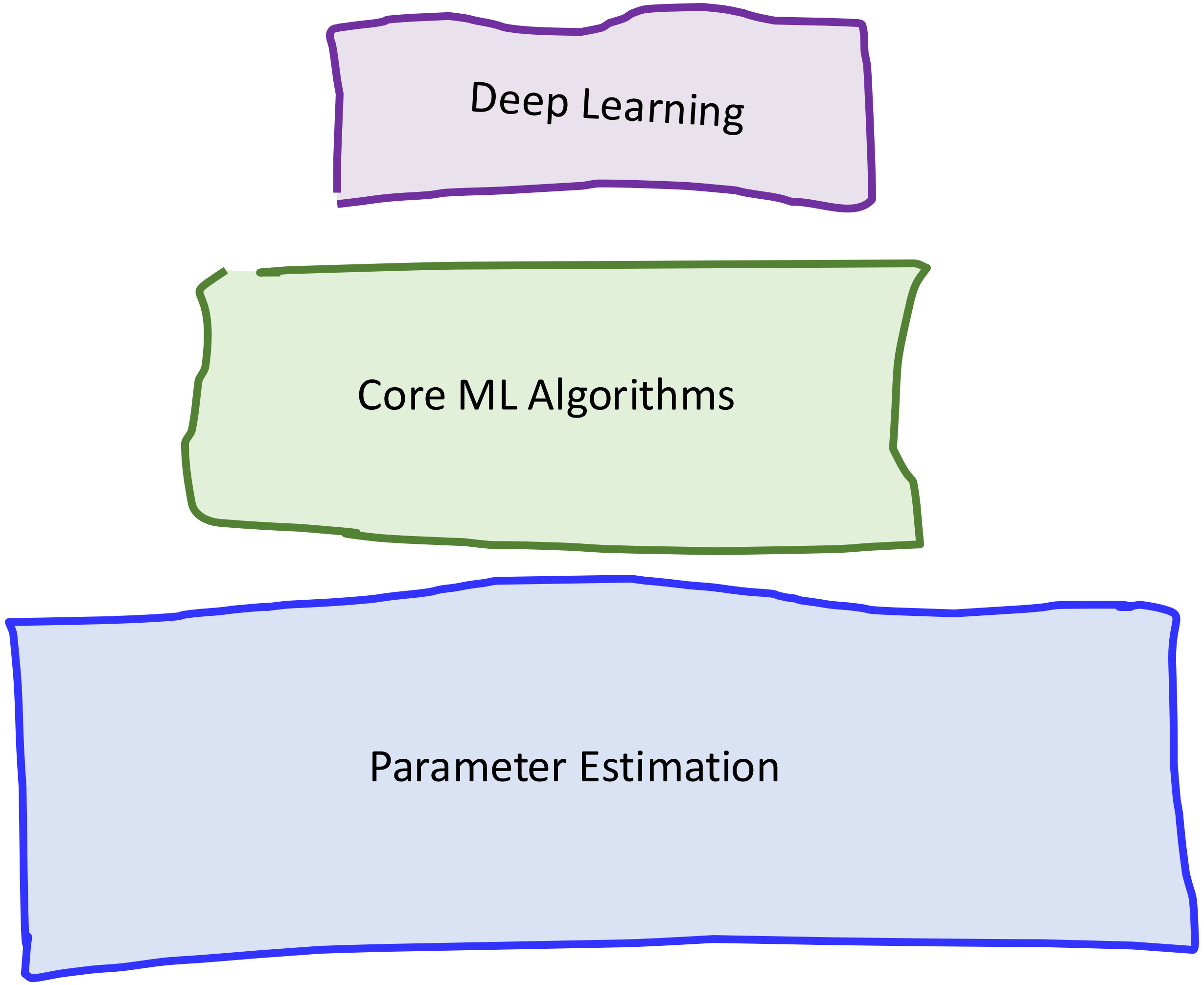
Random Variables

Probabilistic Models

Uncertainty Theory

Machine Learning

Our Path

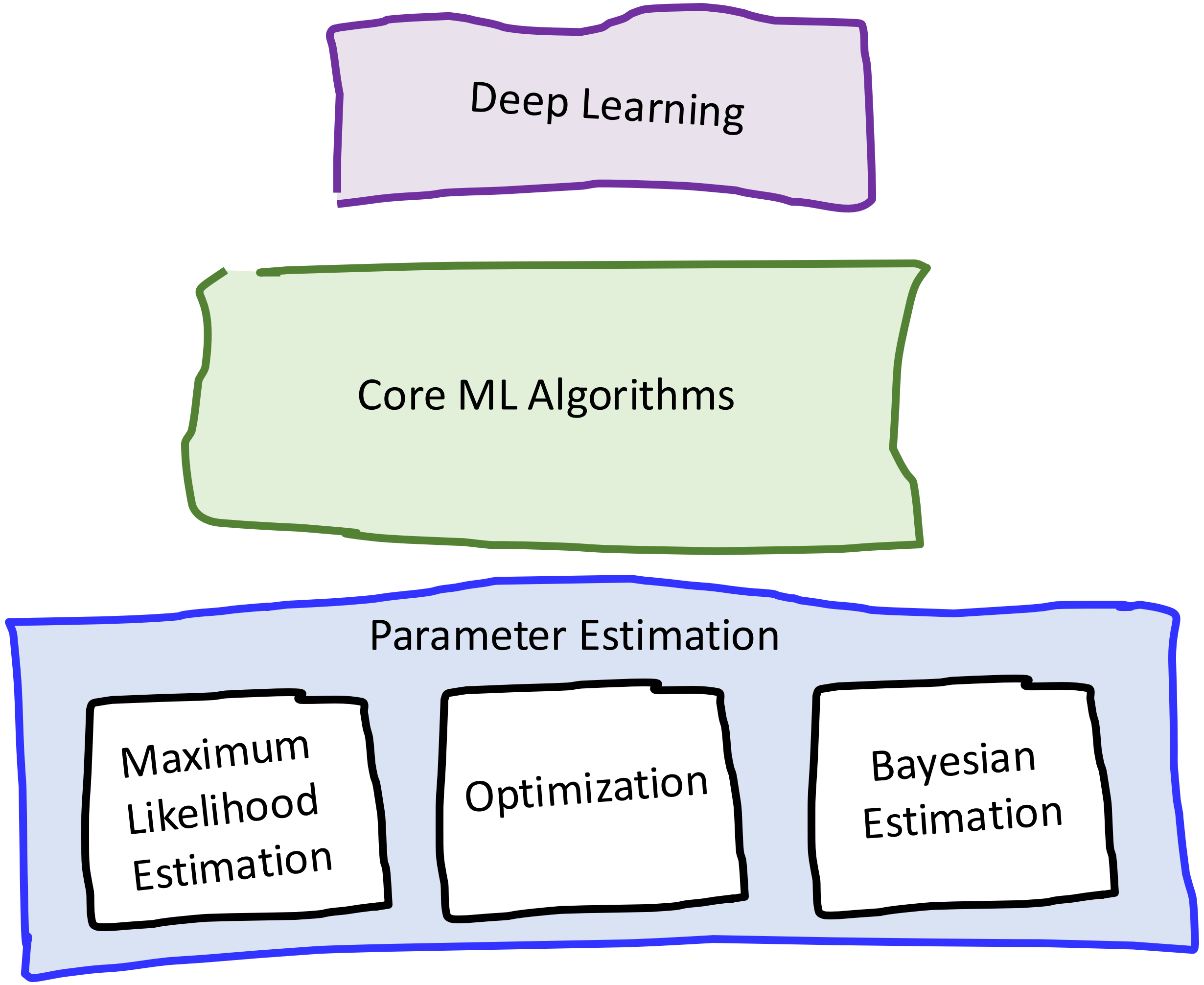


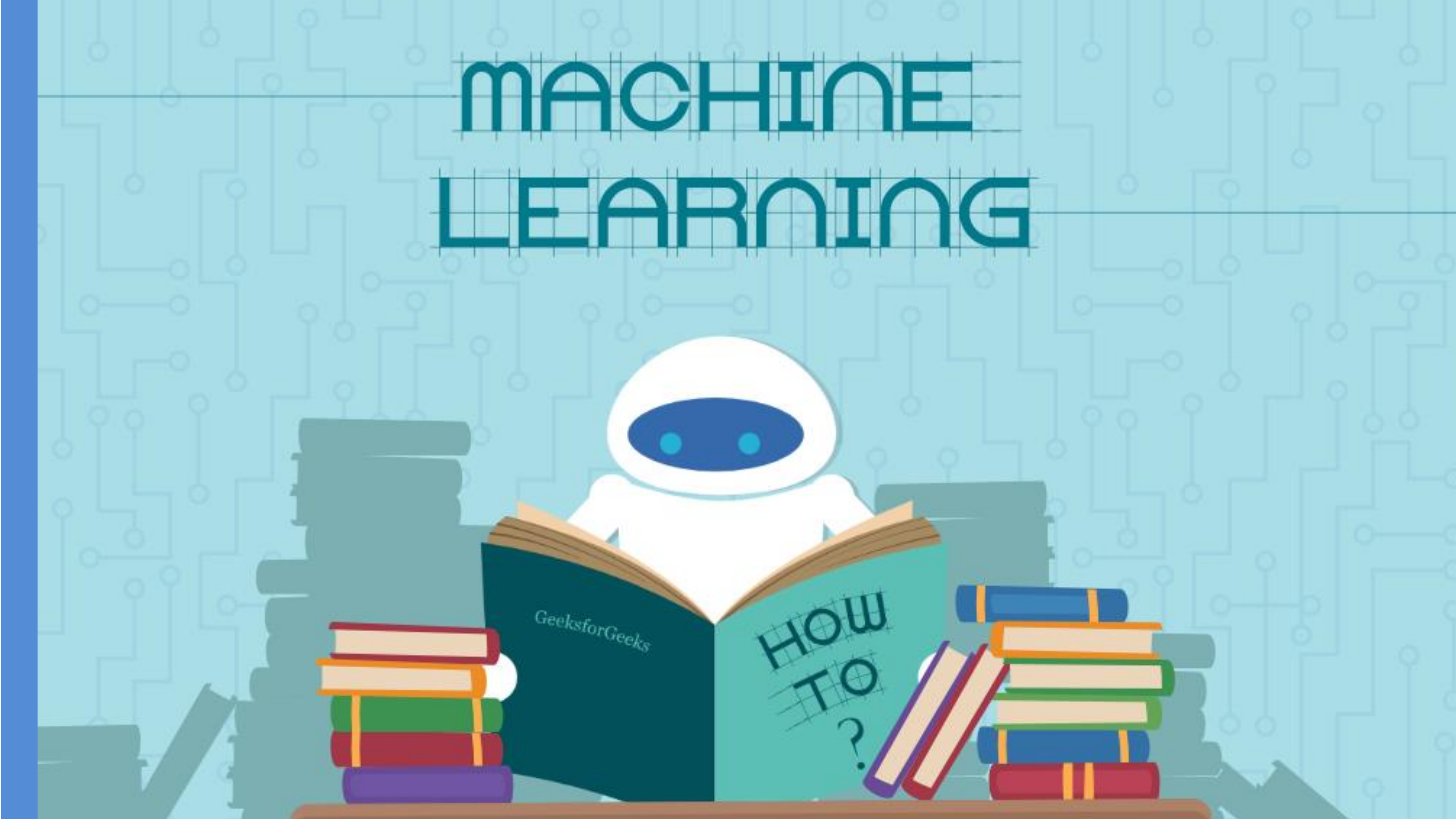
Deep Learning

Core ML Algorithms

Parameter Estimation

Our Path


Deep Learning
Core ML Algorithms
Parameter Estimation
Maximum Likelihood Estimation
Optimization
Bayesian Estimation


MACHINE
LEARNING
GeeksforGeeks
HOW
TO
?

Why Not Jump Straight to Deep Learning?

Review

Shorthand for Equality Events

Our shorthand notation

$$f(x|\theta)$$

$\boxed{x}$ Is shorthand for the event $\boxed{X = x}$

$\boxed{\theta}$ Is shorthand for the event $\boxed{\Theta = \theta}$

Full Notation

$$f(X = x|\Theta = \theta)$$

End Review

Remember This Problem?

A doctor on call can receive calls independently at all hours of the day.

The average rate of calls per day is 6.

What is the probability the doctor receives more than 8 calls today?

Remember This Problem?

A doctor on call can receive calls independently at all hours of the day.

The average rate of calls per day is 6.

What is the probability the doctor receives more than 8 calls today?

Solution:

Let X be the number of calls the doctor receives today. $X \sim \text{Poi}(\lambda = 6)$.

$$P(X > 8) = 1 - \sum_{k=0}^{8} P(X = k) = 1 - \sum_{k=0}^{8} \frac{\lambda^k e^{-\lambda}}{k!}$$

Remember This Problem?

A doctor on call can receive calls independently at all hours of the day.

The average rate of calls per day is 6.

What is the probability the doctor receives more than 8 calls today?

We also had to choose the right **parameter** (λ)

We chose the Poisson as our "**model**"

Solution:

Let X be the number of calls the doctor receives today. $X \sim \text{Poi}(\lambda = 6)$.

$$P(X > 8) = 1 - \sum_{k=0}^{8} P(X = k) = 1 - \sum_{k=0}^{8} \frac{\lambda^k e^{-\lambda}}{k!}$$

What If You Aren't Given A Parameter?

A doctor on call can receive calls independently at all hours of the day.

~~The average rate of calls per day is 6.~~

What is the probability the doctor receives more than 8 calls today?

Today, We Learn:

Today, We Learn:

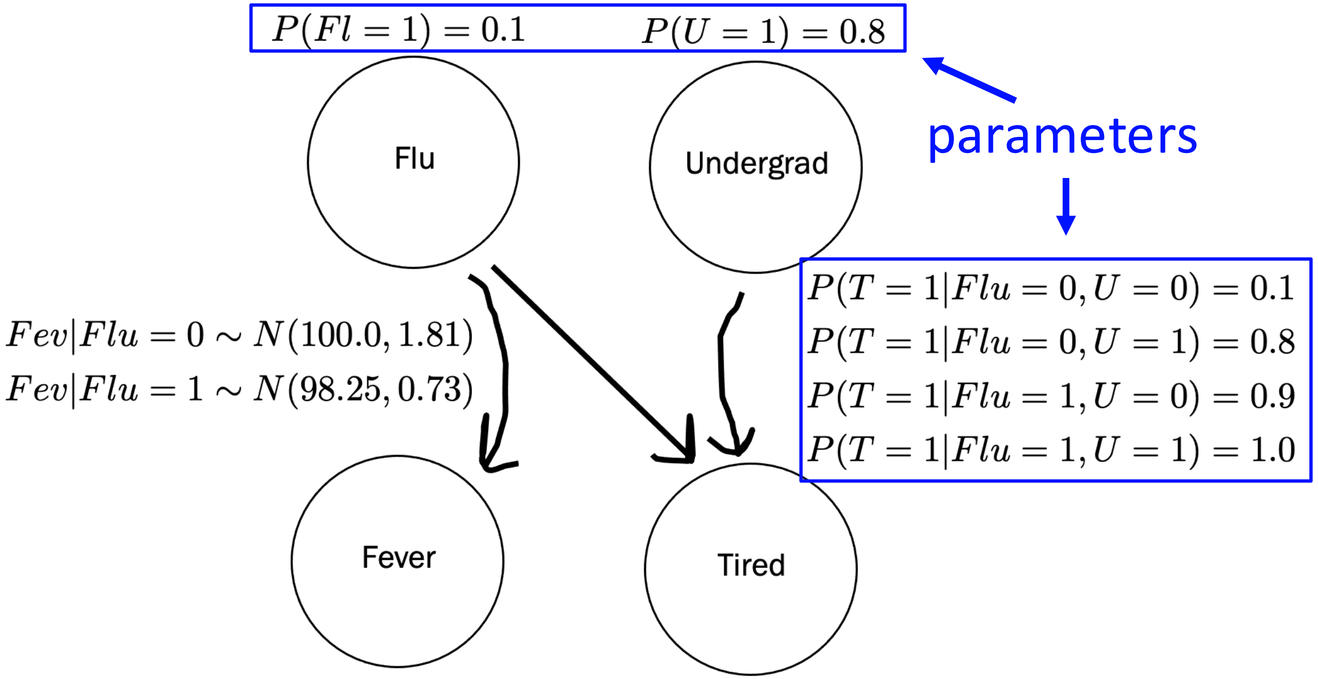
Where do parameters come from???

At this point: if you have a *model*
and all necessary *parameters*,
you can make predictions

But what if you need to *learn* the parameters for the model?

Parameter Estimation Is The Foundation

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



$P(Fl=1)=0.1$ $\quad P(U=1)=0.8$

$Fev|Flu=0 \sim N(100.0, 1.81)$

$Fev|Flu=1 \sim N(98.25, 0.73)$

$P(T=1|Flu=0,U=0)=0.1$

$P(T=1|Flu=0,U=1)=0.8$

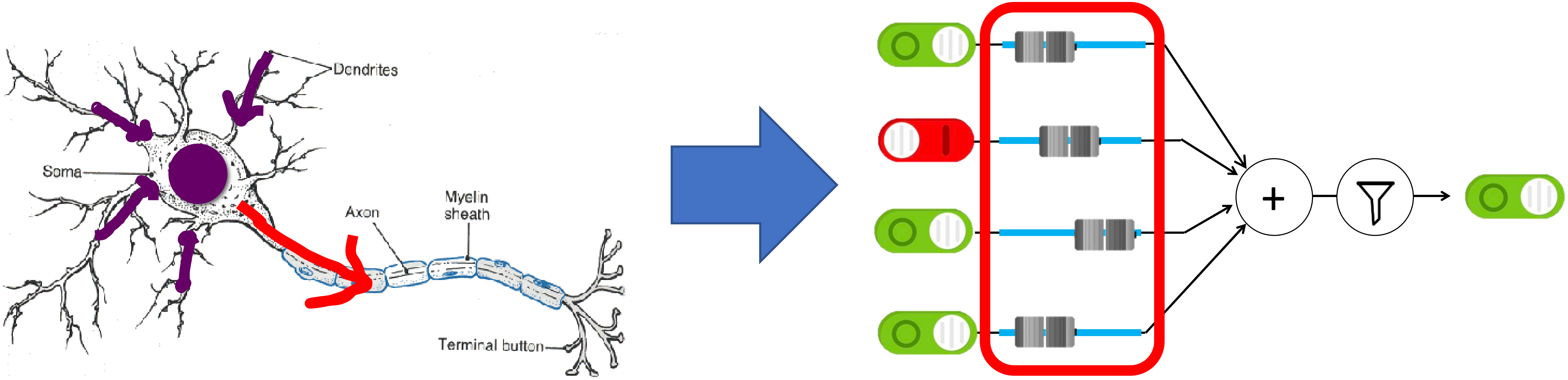
$P(T=1|Flu=1,U=0)=0.9$

$P(T=1|Flu=1,U=1)=1.0$

The strategy for solving this…

…also works for more complex models

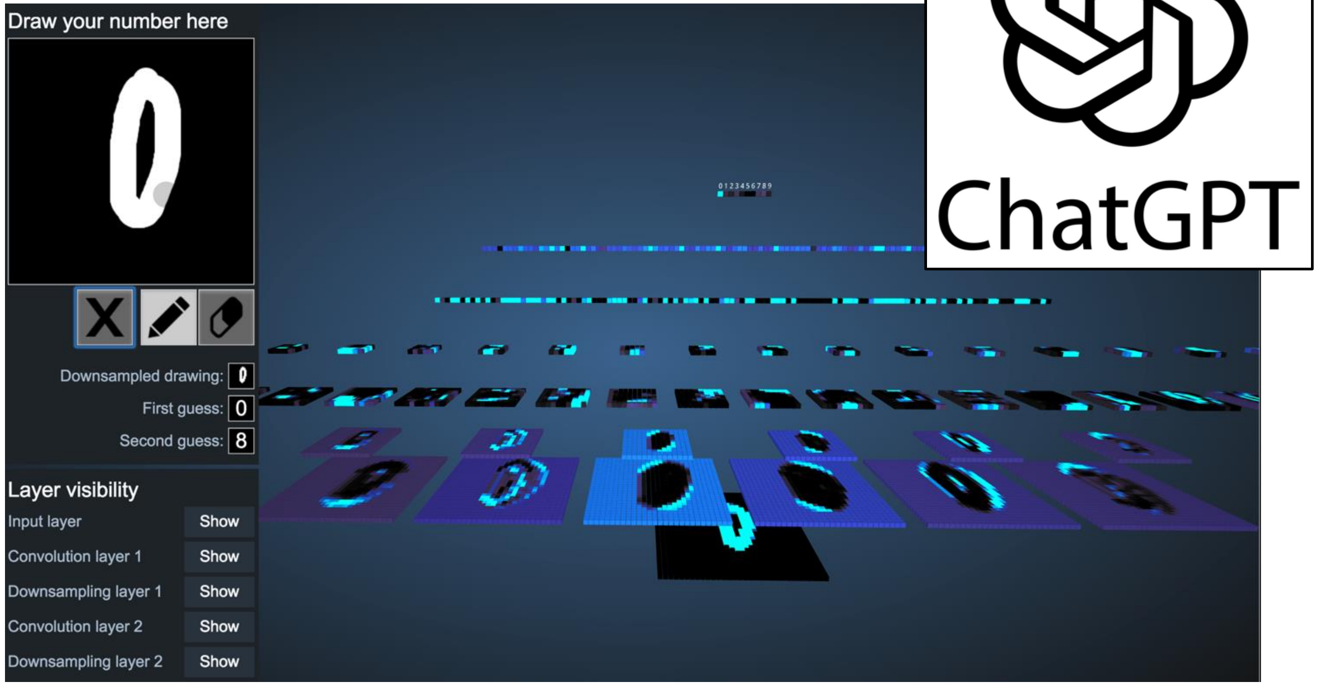
Deep Learning: Neurons With Parameters



Neural Networks get their intelligence from their parameters.

Parameter Estimation Is The Foundation

$$X \sim \mathrm{Poi}(\lambda = ???)$$



Draw your number here

Downsampled drawing: 0

First guess: 0

Second guess: 8

Layer visibility

Input layer	Show
Convolution layer 1	Show
Downsampling layer 1	Show
Convolution layer 2	Show
Downsampling layer 2	Show

ChatGPT

The strategy for solving this…

…is the same as “model training” for this

How Do We Estimate A Parameter From Data?

A doctor on call can receive calls independently at all hours of the day.

~~The average rate of calls per day is 6.~~

Each day last week, the doctor counted how many calls she received in a day.

Here is a list of the number of calls each day: [10, 4, 7, 6, 8, 4, 5].

What is the probability the doctor receives more than 8 calls today?

How Do We Estimate A Parameter From Data?

A doctor on call can receive calls independently at all hours of the day.

~~The average rate of calls per day is 6.~~

Each day last week, the doctor counted how many calls she received in a day.

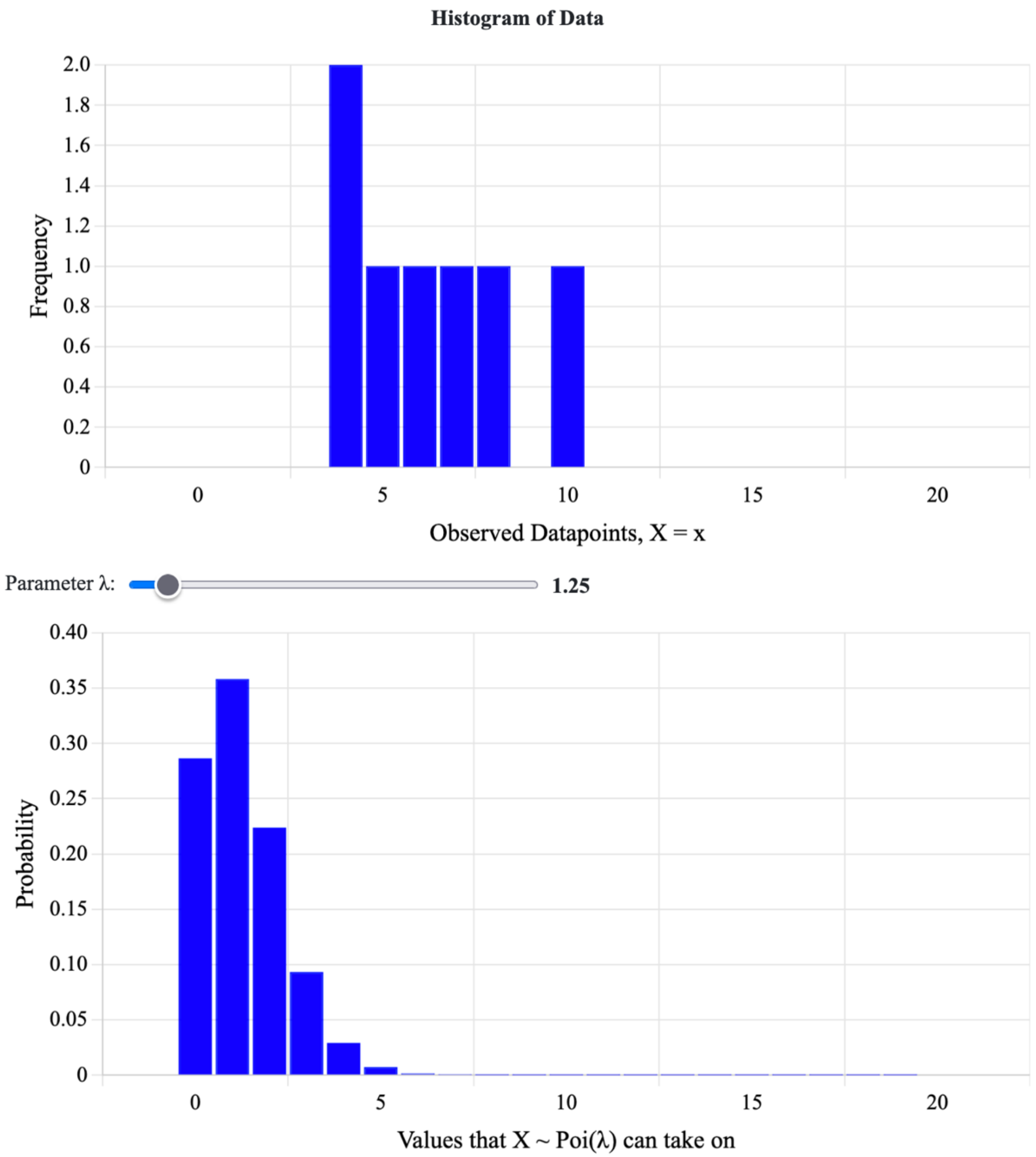
Here is a list of the number of calls each day: [10, 4, 7, 6, 8, 4, 5].

What is the probability the doctor receives more than 8 calls today?

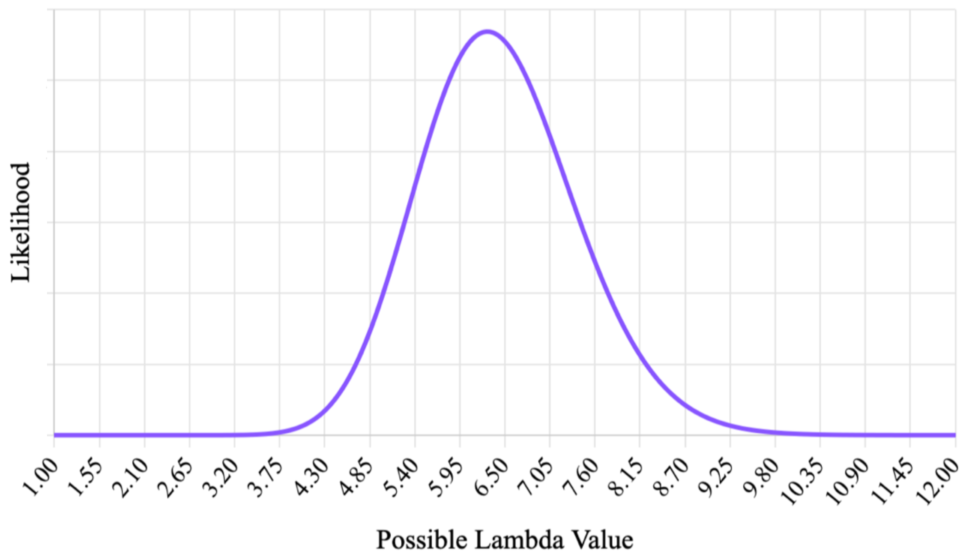
Let X be the number of calls the doctor receives today. $X \sim \text{Poi}(\lambda = \textbf{???})$.

(We'll assume the data comes from an actual Poisson process with some true λ)

To The Course Reader!



Parameter λ: 1.25



How To Choose The "Best" Parameters: MLE

We want to choose the parameter value that maximizes the probability of the data.

How To Choose The “Best” Parameters: MLE

We want to choose the parameter value that maximizes the probability of the data.

Maximum **Likelihood** **Estimation!**

How To Choose The "Best" Parameters: MLE

We want to choose the parameter value that maximizes the probability of the data.

Maximum **Likelihood** **Estimation!**

How do we quantify "probability of the data"?

The Likelihood Function

Definition: The probability of our observed data.

$$L(\theta) = \mathrm{P}(\mathrm{data})$$

θ is shorthand for parameter(s)
(here, $\theta = \lambda$)

The Likelihood Function

Definition: The probability of our observed data.

If we had a single observation, $X = x$:

$$\begin{aligned} L(\theta) &= \mathrm{P}(X = x) \\ &= \mathrm{P}(x) \end{aligned}$$

(in our example, this would just be the Poisson PMF)

The Likelihood Function

Definition: The ***joint*** probability of our observed data.

For a list of observations, $[x_1, x_2, \ldots, x_n]$:

$$L(\theta) = \mathrm{P}(x_1, x_2, \ldots, x_n)$$

The Likelihood Function

Definition: The ***joint*** probability of our observed data.

For a list of observations, $[x_1, x_2, \dots, x_n]$:

$$L(\theta) = \mathrm{P}(x_1, x_2, \dots, x_n)$$

If we assume independence
between datapoints,
the joint becomes a product

$$= \prod_{i=1}^{n} \underline{\mathrm{P}(x_i)}$$

(still the Poisson PMF)

🔑

The Likelihood Function

Definition: The ***joint*** probability of our observed data, *as a function of the model's parameters.*

For a list of observations, $[x_1, x_2, \ldots, x_n]$:

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

Generally, this could be a PMF, or PDF, or some joint distribution

How To Choose The “Best” Parameters: MLE

We want to choose the parameter value that maximizes the probability of the data:

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

How To Choose The “Best” Parameters: MLE

We want to choose the parameter value that maximizes the probability of the data:

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

To put words into math:

$$\underset{\substack{\uparrow \\ \text{Our best estimate}}}{\hat{\theta}} = \underset{\theta}{\operatorname{argmax}}\ L(\theta)$$

Quick math review!

Argmax

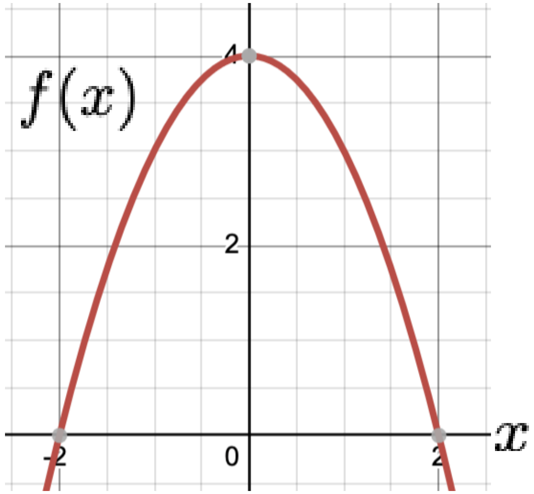
$\hat{x} = \arg\max_{x} f(x)$ = “the value of x that maximizes f(x)”

Argmax

$\hat{x} = \arg\max_{x} f(x)$ = "the value of x that maximizes f(x)"

$f(x) = -x^2 + 4$

$\max f(x) = \ ?$



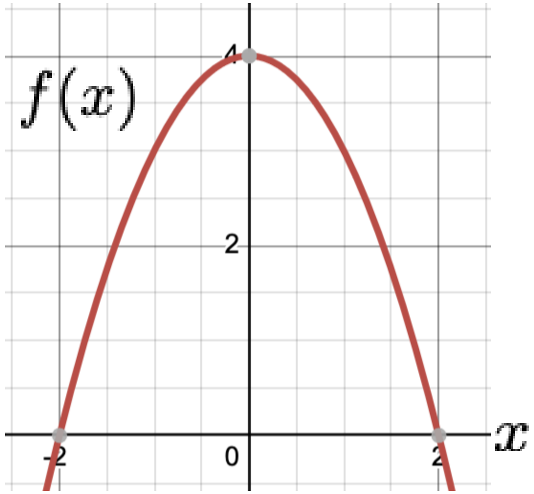
Argmax

$\hat{x} = \arg\max_{x} f(x)$ = "the value of x that maximizes f(x)"

$$f(x) = -x^2 + 4$$

$$\max f(x) = 4$$

$$\arg\max_{x} f(x) = ?$$



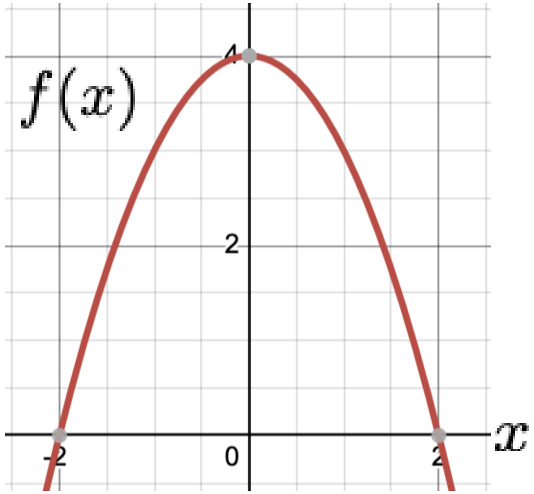
Argmax

$\hat{x} = \arg\max_{x} f(x)$ = “the value of x that maximizes f(x)”

$$f(x) = -x^2 + 4$$

$$\max f(x) = 4$$

$$\arg\max_{x} f(x) = 0$$



Because at x = 0, f(x) is maximized!

But how do we compute argmax?

Option #1: Take the derivative

(Option #2 is next Monday’s lecture)

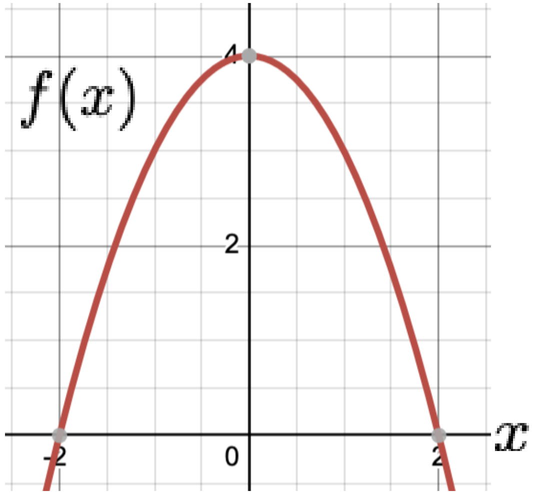
Argmax

$\hat{x} = \arg\max_x f(x)$ = "the value of x that maximizes f(x)"

$$f(x) = -x^2 + 4$$

$$\max f(x) = 4$$

$$\arg\max_x f(x) = 0$$



How to compute argmax? Take the derivative!

The derivative always equals 0 at the argmax.

https://www.mathsisfun.com/calculus/derivative-plotter.html

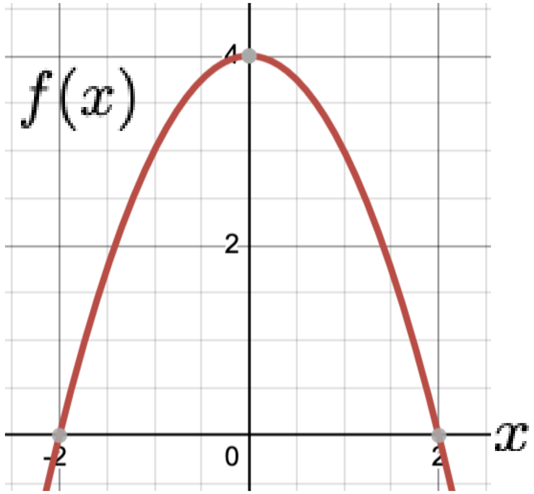
Argmax

$\hat{x} = \arg\max_x f(x)$ = "the value of x that maximizes f(x)"

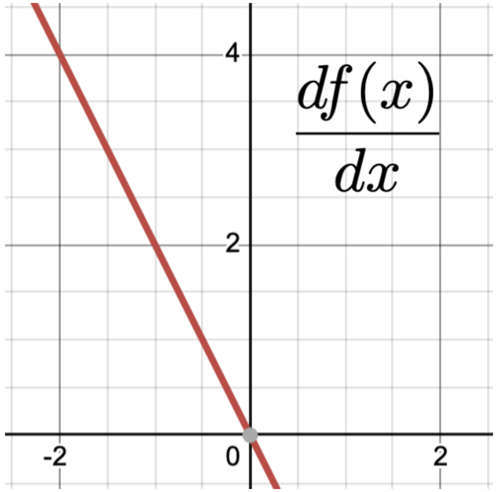
$$f(x) = -x^2 + 4$$

$$\max f(x) = 4$$

$$\arg\max_x f(x) = 0$$



$$\frac{df(x)}{dx} = -2x$$



How to compute argmax? Take the derivative!

The derivative always equals 0 at the argmax.

https://www.mathsisfun.com/calculus/derivative-plotter.html

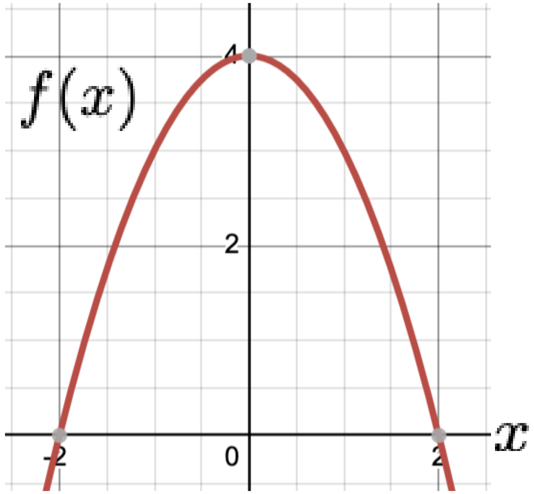
Argmax

$\hat{x} = \arg\max_{x} f(x)$ = "the value of x that maximizes f(x)"

$$f(x) = -x^2 + 4$$

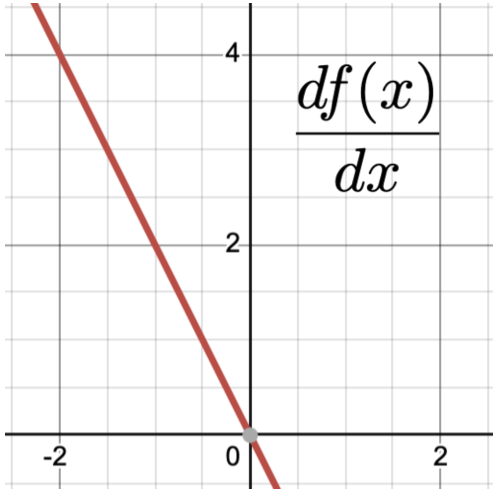
$$\max f(x) = 4$$

$$\arg\max_{x} f(x) = 0$$



$$\frac{df(x)}{dx} = -2x$$

-2x = 0 when x = 0



How to compute argmax? Take the derivative!

The derivative always equals 0 at the argmax.

https://www.mathsisfun.com/calculus/derivative-plotter.html

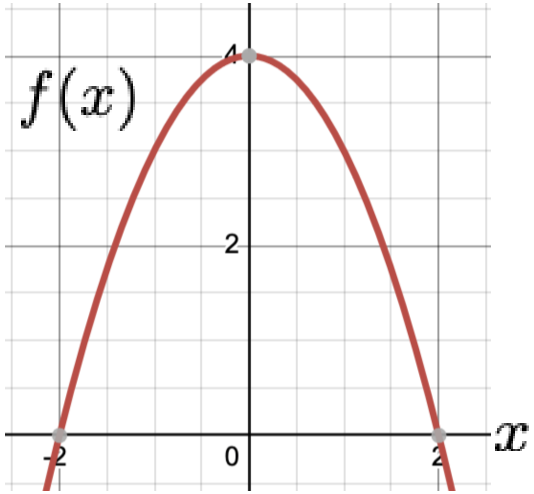
Argmax

$\hat{x} = \arg\max_{x} f(x)$ = "the value of x that maximizes f(x)"

$$f(x) = -x^2 + 4$$

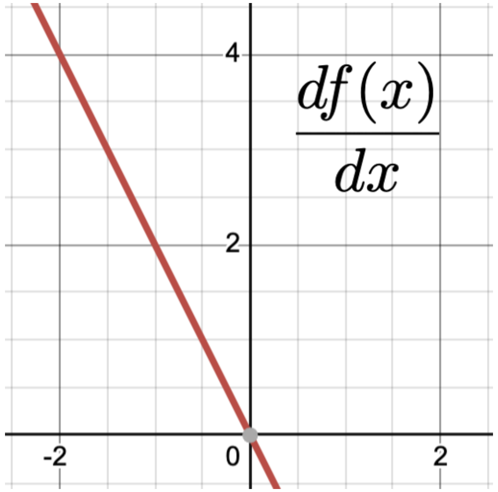
$$\max f(x) = 4$$

$$\arg\max_{x} f(x) = 0$$



$$\frac{df(x)}{dx} = -2x$$

-2x = 0 when x = 0



How to compute argmax? Take the derivative!

The derivative always equals 0 at the argmax.

(technically, we should check the 2nd derivative to make sure $\hat{x}$ is a maximum)

https://www.mathsisfun.com/calculus/derivative-plotter.html

Fun Facts About logs

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

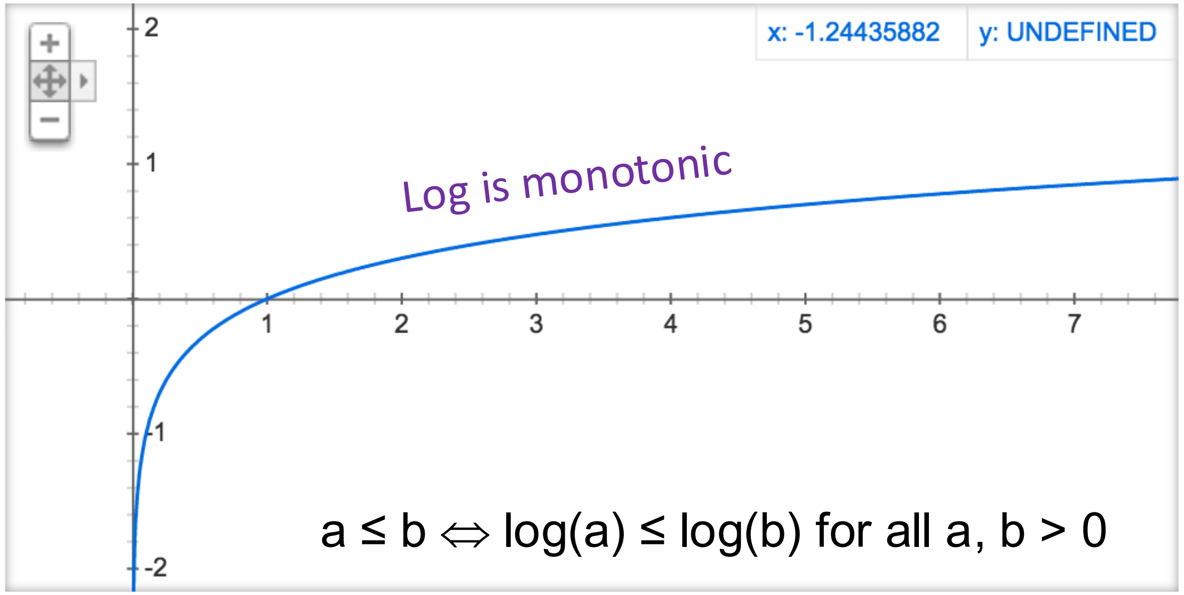
We can break down log of a product into the sum of logs.

Fun Facts About logs

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

We can break down log of a product into the sum of logs.

Graph for log(x)



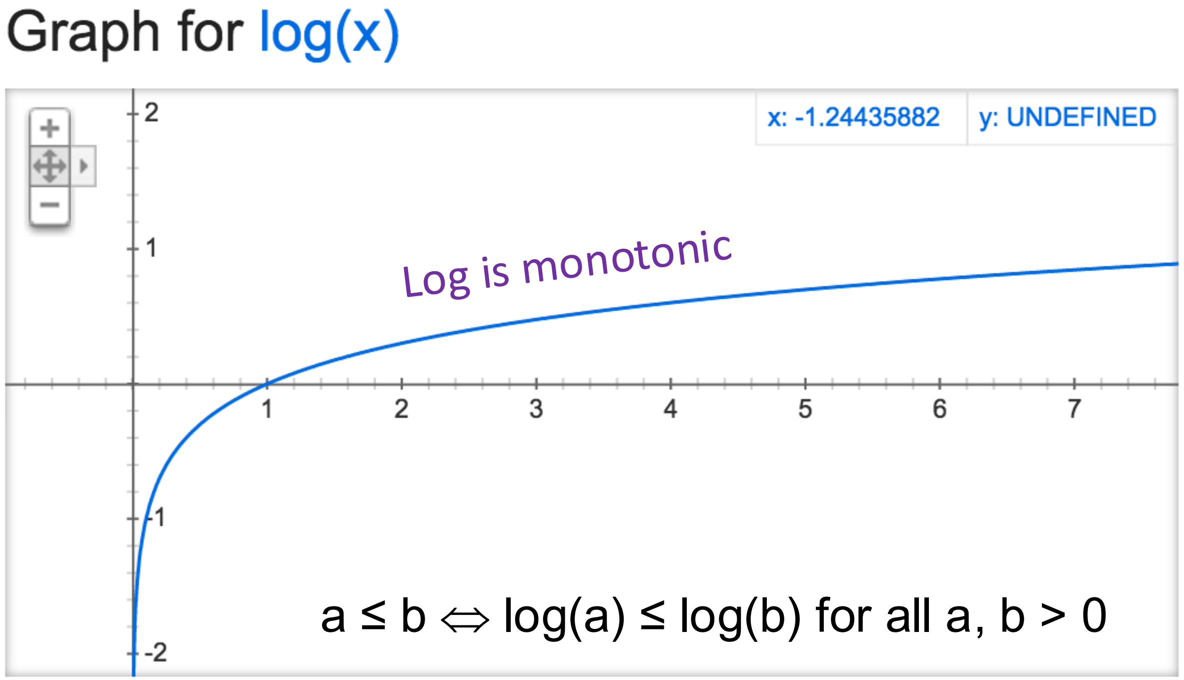
Log is monotonic

a ≤ b ⇔ log(a) ≤ log(b) for all a, b > 0

Fun Facts About logs

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

We can break down log of a product into the sum of logs.



Graph for log(x)

x: -1.24435882 y: UNDEFINED

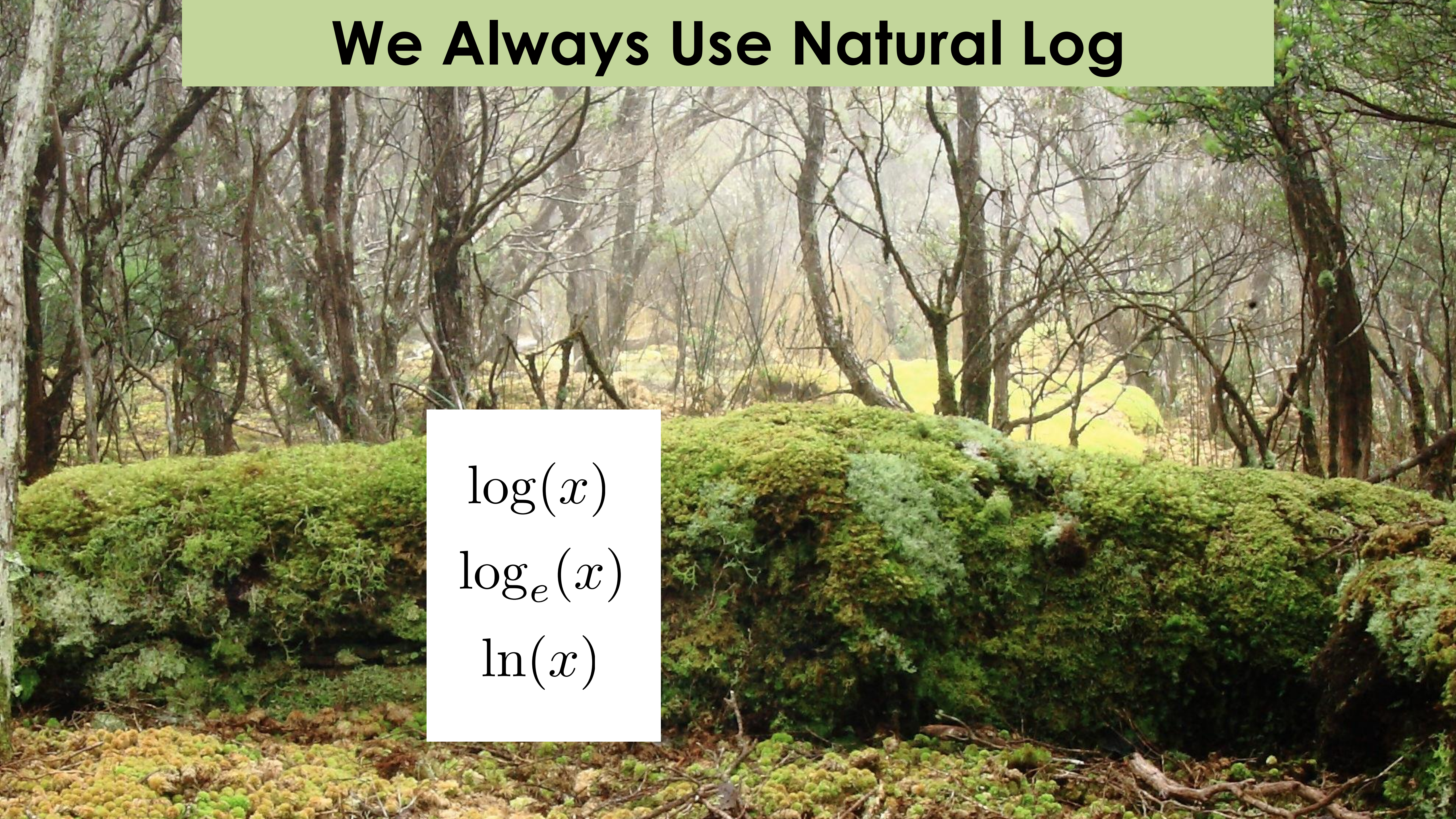
Log is monotonic

a ≤ b ⇔ log(a) ≤ log(b) for all a, b > 0

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

We can take the log of what we want the argmax of, and find the argmax of that instead

We Always Use Natural Log



$$\log(x)$$

$$\log_e(x)$$

$$\ln(x)$$

Maximum Likelihood with Poisson

Data: We have IID observations [$X_1 = x_1$, $X_2 = x_2$, ..., $X_n = x_n$]. Model: $X_i \sim \mathrm{Poi}(\lambda)$

Goal: Estimate λ by MLE

Step 0: Write out the likelihood of the data: $L(\lambda)$

Then, find the value of λ which maximizes $L(\lambda)$

Maximum Likelihood with Poisson

Data: We have IID observations [$X_1 = x_1$, $X_2 = x_2$, ..., $X_n = x_n$]. Model: $X_i \sim \mathrm{Poi}(\lambda)$

Goal: Estimate λ by MLE

PMF: $$f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \longrightarrow L(\lambda) = f(x_1 \ldots x_n|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

Then, find the value of λ which maximizes L(λ)

Maximum Likelihood with Poisson

Data: We have IID observations $[X_1 = x_1, X_2 = x_2, ..., X_n = x_n]$. Model: $X_i \sim \text{Poi}(\lambda)$

Goal: Estimate λ by MLE

PMF: $f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$ → $L(\lambda) = f(x_1 \dots x_n|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

To compute $\hat{\lambda} = \underset{\lambda}{\text{argmax}}\, L(\lambda)$

1. **Take log of likelihood.**
2. **Take derivative w.r.t. λ.**
3. **Set equal to 0 and solve for λ.**

Maximum Likelihood with Poisson

Data: We have IID observations [$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$]. Model: $X_i \sim \text{Poi}(\lambda)$

Goal: Estimate λ by MLE

PMF: $f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$ → $L(\lambda) = f(x_1 \ldots x_n|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

To compute $\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, L(\lambda)$

1. **Take log of likelihood.**
2. **Take derivative w.r.t. λ.**
3. **Set equal to 0 and solve for λ.**

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

$$= \sum_{i=1}^{n} -\lambda + x_i \log \lambda - \log x_i!$$

Maximum Likelihood with Poisson

Data: We have IID observations $[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n]$. Model: $X_i \sim \text{Poi}(\lambda)$

Goal: Estimate λ by MLE

PMF: $f(x_i|\lambda) = \dfrac{e^{-\lambda}\lambda^{x_i}}{x_i!} \longrightarrow L(\lambda) = f(x_1 \ldots x_n|\lambda) = \displaystyle\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

To compute $\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, L(\lambda)$

1. **Take log of likelihood.**
2. **Take derivative w.r.t. λ.**
3. **Set equal to 0 and solve for λ.**

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

$$= \sum_{i=1}^{n} -\lambda + x_i \log \lambda - \log x_i!$$

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

Maximum Likelihood with Poisson

Data: We have IID observations $[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n]$. Model: $X_i \sim \text{Poi}(\lambda)$

Goal: Estimate λ by MLE

PMF: $f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$ → $L(\lambda) = f(x_1 \ldots x_n|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$

To compute $\hat{\lambda} = \underset{\lambda}{\text{argmax}}\, L(\lambda)$

1. **Take log of likelihood.**
2. **Take derivative w.r.t. λ.**
3. **Set equal to 0 and solve for λ.**

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$$

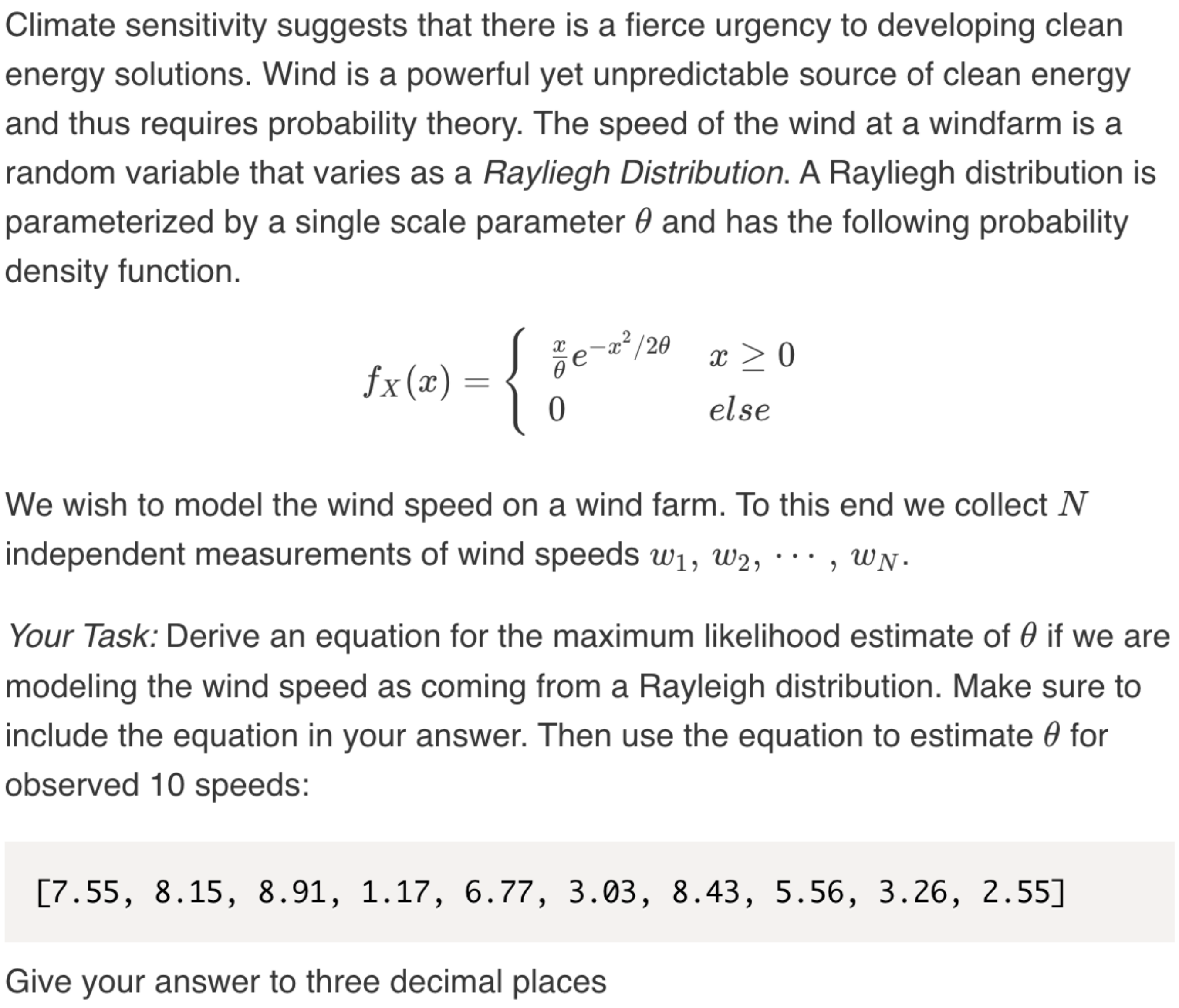
$$LL(\lambda) = \log \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \sum_{i=1}^{n} \log \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

$$= \sum_{i=1}^{n} -\lambda + x_i \log \lambda - \log x_i!$$

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

MLE problems all follow this same formula

MLE of the Wind



Climate sensitivity suggests that there is a fierce urgency to developing clean energy solutions. Wind is a powerful yet unpredictable source of clean energy and thus requires probability theory. The speed of the wind at a windfarm is a random variable that varies as a *Rayliegh Distribution*. A Rayliegh distribution is parameterized by a single scale parameter θ and has the following probability density function.

$$f_X(x) = \begin{cases} \frac{x}{\theta} e^{-x^2/2\theta} & x \geq 0 \\ 0 & else \end{cases}$$

We wish to model the wind speed on a wind farm. To this end we collect N independent measurements of wind speeds $w_1, w_2, \cdots, w_N$.

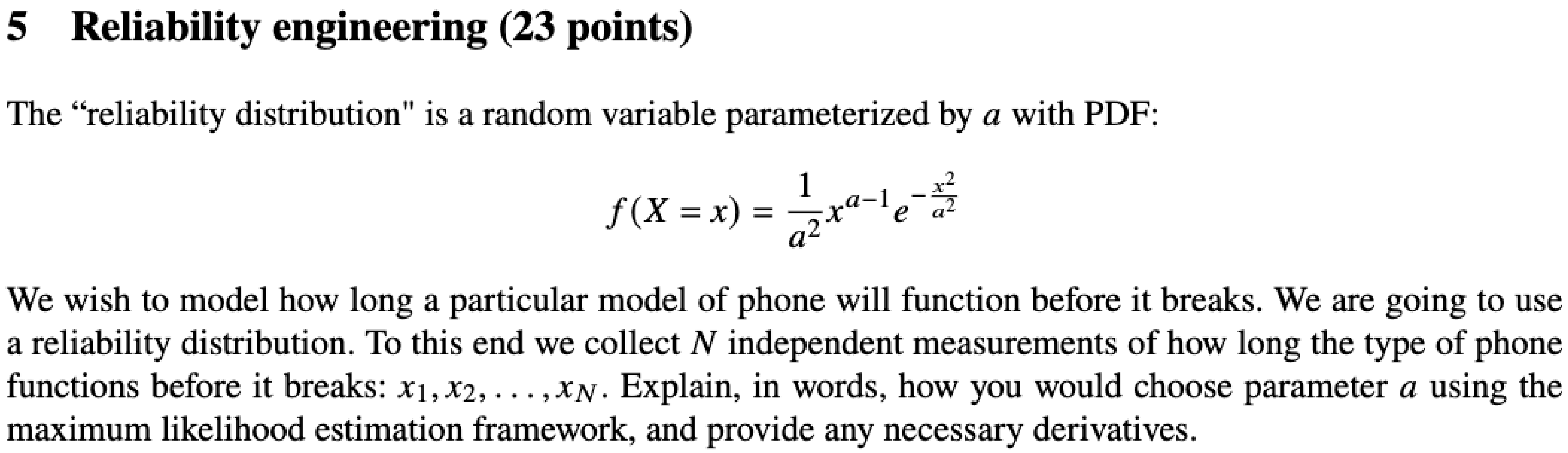
Your Task: Derive an equation for the maximum likelihood estimate of θ if we are modeling the wind speed as coming from a Rayleigh distribution. Make sure to include the equation in your answer. Then use the equation to estimate θ for observed 10 speeds:

```
[7.55, 8.15, 8.91, 1.17, 6.77, 3.03, 8.43, 5.56, 3.26, 2.55]
```

Give your answer to three decimal places

On pset 7!

MLE Is Guaranteed on the Final Exam



5 Reliability engineering (23 points)

The “reliability distribution" is a random variable parameterized by a with PDF:

$$f(X = x) = \frac{1}{a^2} x^{a-1} e^{-\frac{x^2}{a^2}}$$

We wish to model how long a particular model of phone will function before it breaks. We are going to use a reliability distribution. To this end we collect N independent measurements of how long the type of phone functions before it breaks: $x_1, x_2, \ldots, x_N$. Explain, in words, how you would choose parameter a using the maximum likelihood estimation framework, and provide any necessary derivatives.

You’ll get to practice in section too

MLE Plotline

1. Decide on a model for the distribution of your data. Identify the PMF / PDF.

2. Write out the likelihood function.

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

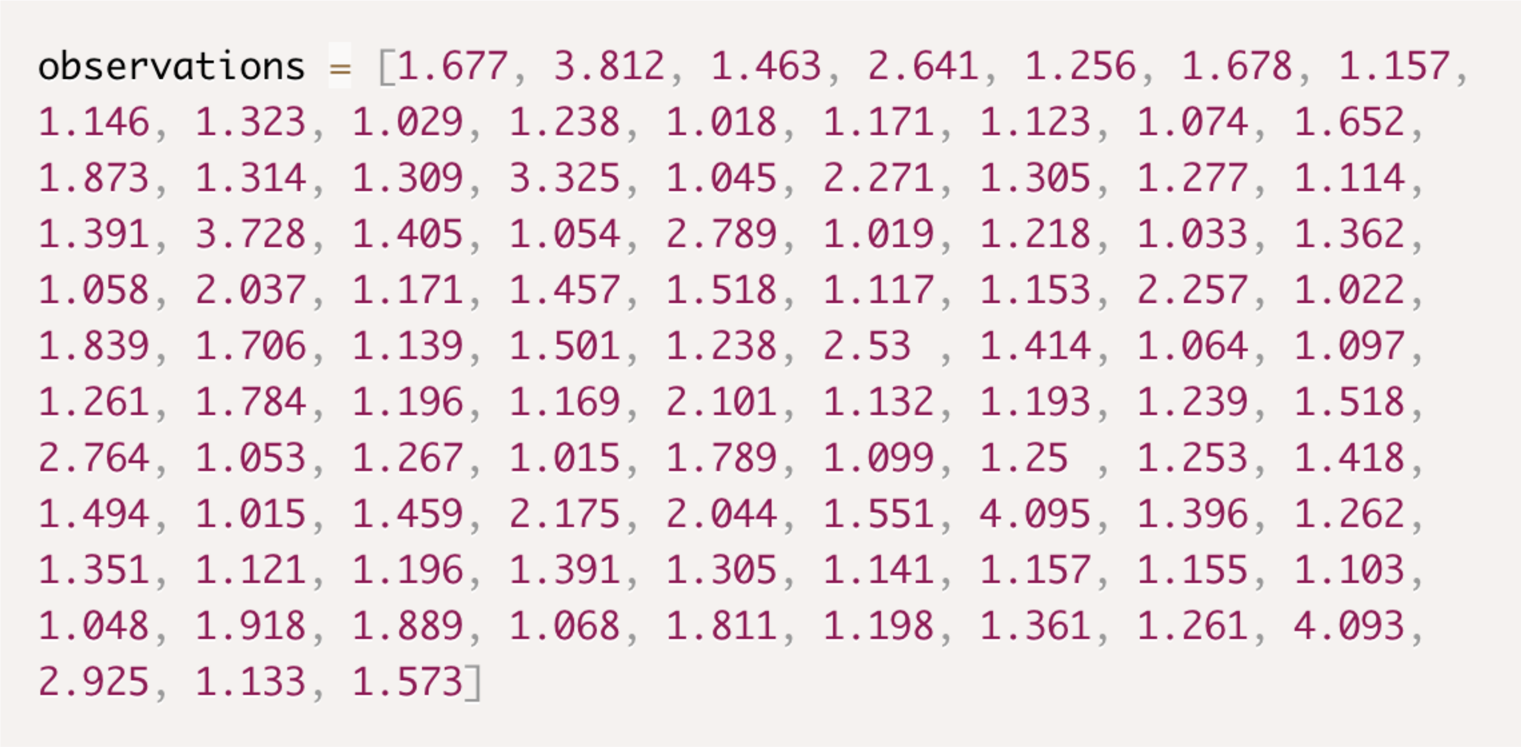
4. Find the argmax.

4a. Take log

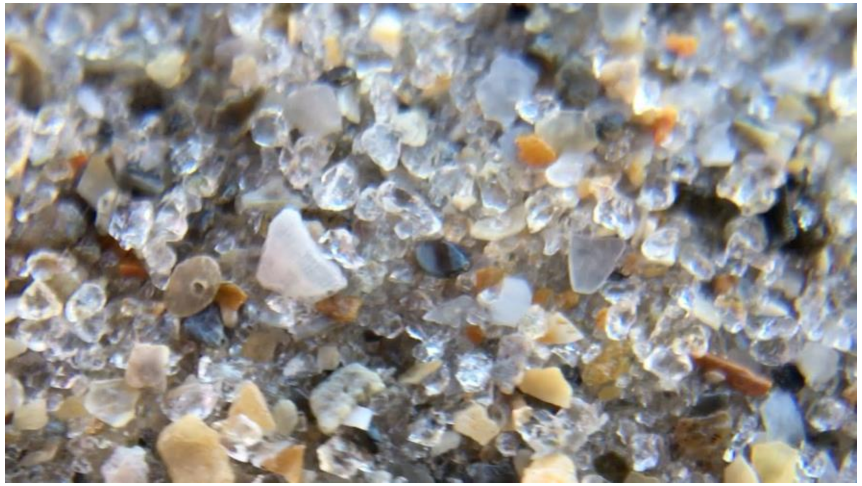
4b. Take derivative w.r.t. the parameter

4c. Set derivative equal to 0 and solve

Let’s Practice on the 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]
```



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

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

MLE for Bernoulli

A new drug is given to 20 patients. It "works" for 14 and "doesn't work" for 6.
What is your best estimate of the probability the drug will work for the next patient?

In other words:

We have 20 I.I.D. observations of $X_i \sim \text{Bern}(p)$. We want to estimate p.
The data: [1,1,0,1,1,1,0,0,1,0,1,1,1,0,1,0,1,1,1,1]

Goal: Estimate p by MLE

Step 0: Write out the likelihood of the data: $L(\lambda)$

MLE for Bernoulli

A new drug is given to 20 patients. It “works” for 14 and “doesn’t work” for 6.
What is your best estimate of the probability the drug will work for the next patient?

In other words:

We have 20 I.I.D. observations of $X_i \sim \text{Bern}(p)$. We want to estimate p.
The data: [1,1,0,1,1,1,0,0,1,0,1,1,1,0,1,0,1,1,1,1]

Goal: Estimate p by MLE

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

MLE for Bernoulli

A new drug is given to 20 patients. It “works” for 14 and “doesn’t work” for 6.
What is your best estimate of the probability the drug will work for the next patient?

In other words:

We have 20 I.I.D. observations of $X_i \sim \text{Bern}(p)$. We want to estimate p.
The data: [1,1,0,1,1,1,0,0,1,0,1,1,1,0,1,0,1,1,1,1]

Goal: Estimate p by MLE

PMF: $$f(x_i|p) = \begin{cases} p & \text{if } x_i = 1 \\ 1-p & \text{if } x_i = 0 \end{cases} \longrightarrow L(p) = f(x_1, \ldots, x_n|p) = \prod_{i=1}^{n} \text{???}$$

MLE for Bernoulli

A new drug is given to 20 patients. It “works” for 14 and “doesn’t work” for 6.
What is your best estimate of the probability the drug will work for the next patient?

In other words:

We have 20 I.I.D. observations of $X_i \sim \text{Bern}(p)$. We want to estimate p.
The data: [1,1,0,1,1,1,0,0,1,0,1,1,1,0,1,0,1,1,1,1]

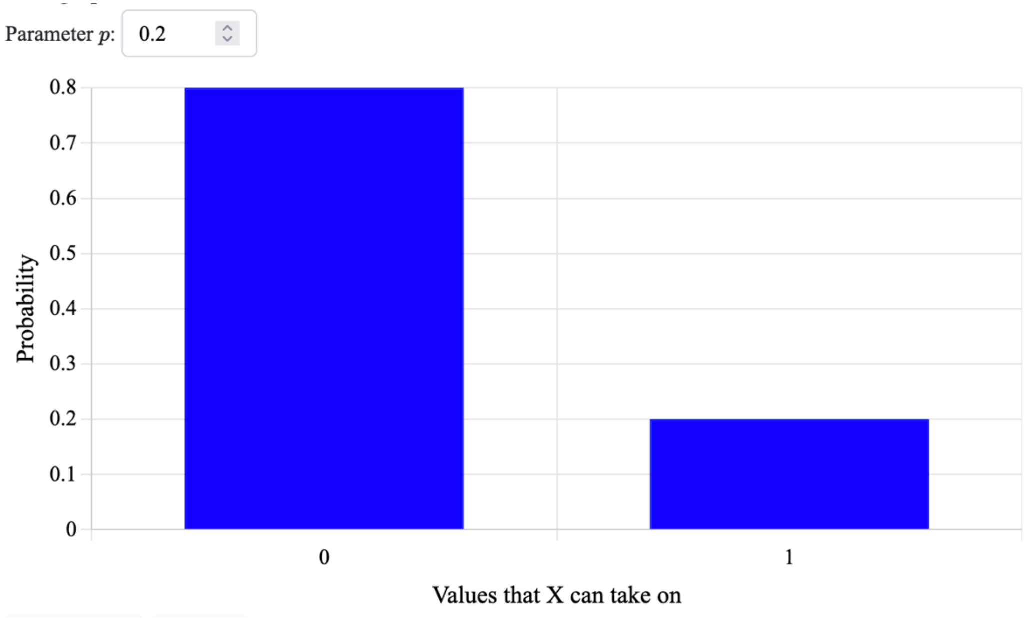
Goal: Estimate p by MLE

PMF: $$f(x_i|p) = \begin{cases} p & \text{if } x_i = 1 \\ 1-p & \text{if } x_i = 0 \end{cases} \longrightarrow L(p) = f(x_1, \ldots, x_n|p) = \prod_{i=1}^{n} \text{???}$$

We Need A New Bernoulli PMF...

PMF of Bernoulli

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

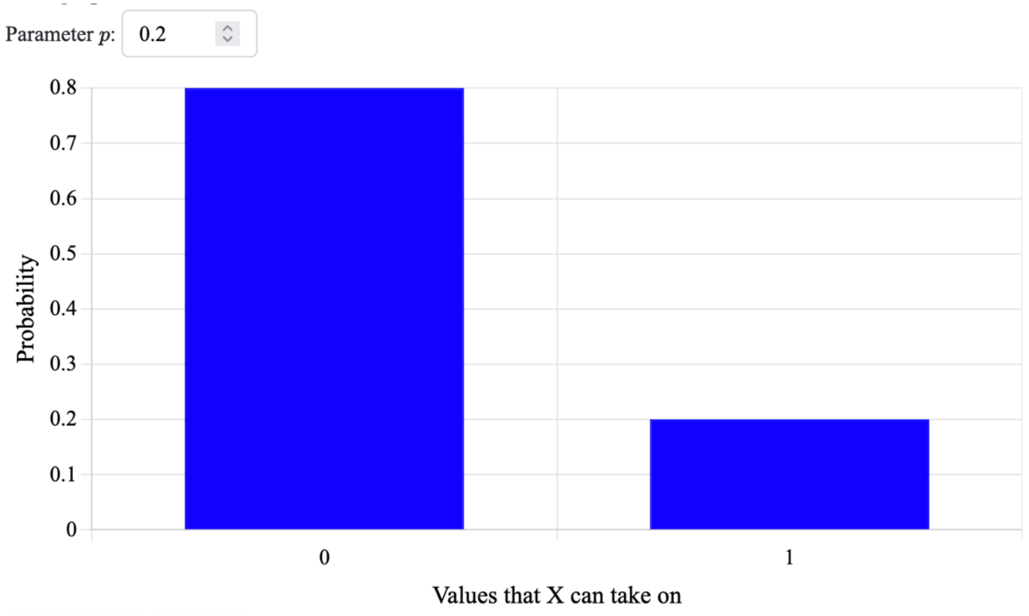


Parameter p: 0.2

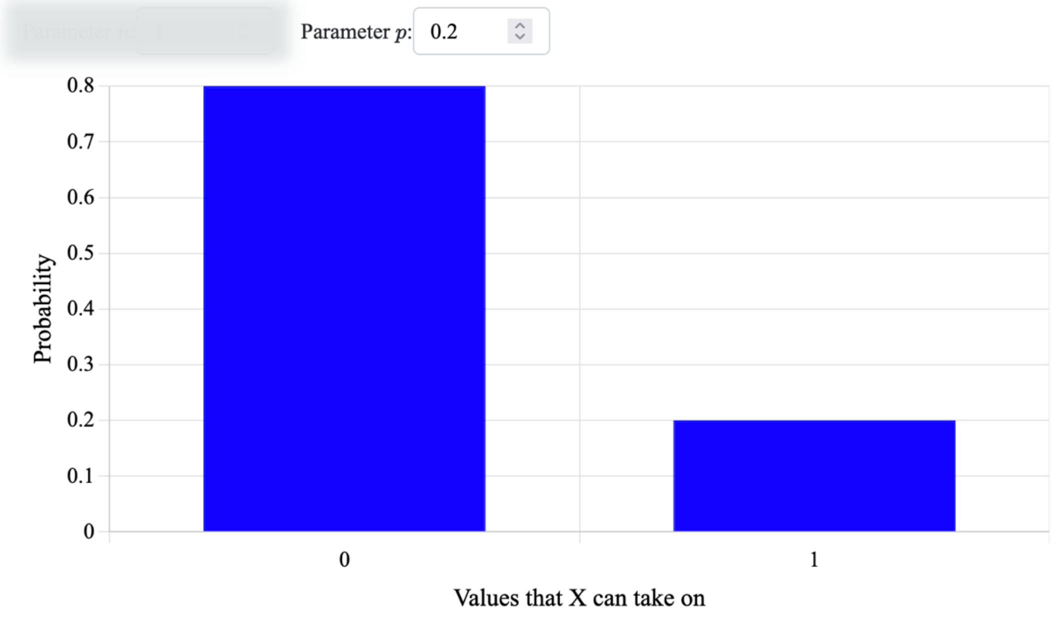
We Need A New Bernoulli PMF…

PMF of Bernoulli

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



Parameter p: 0.2

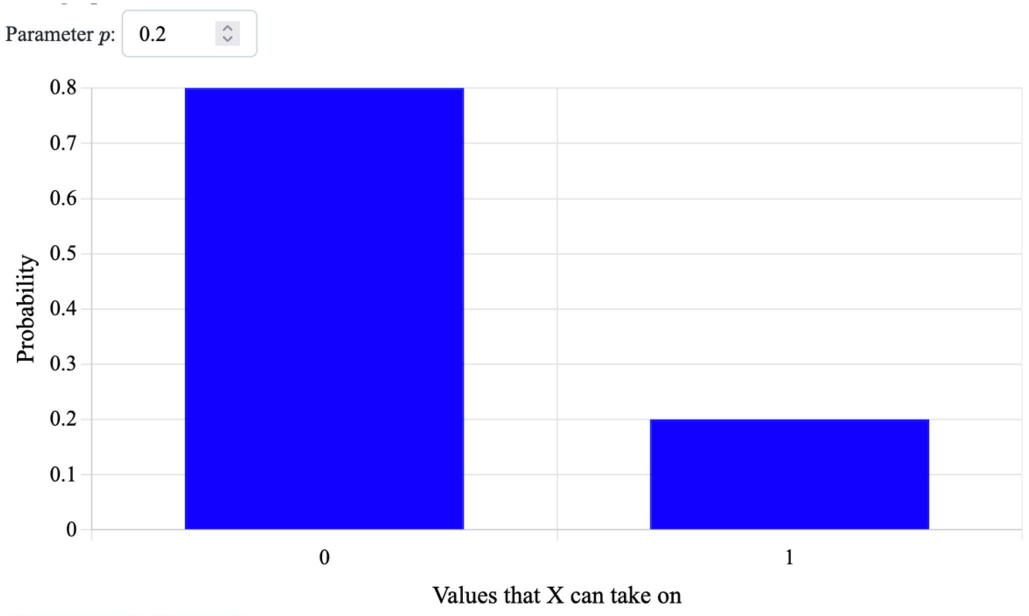


Parameter p: 0.2

We Need A New Bernoulli PMF…

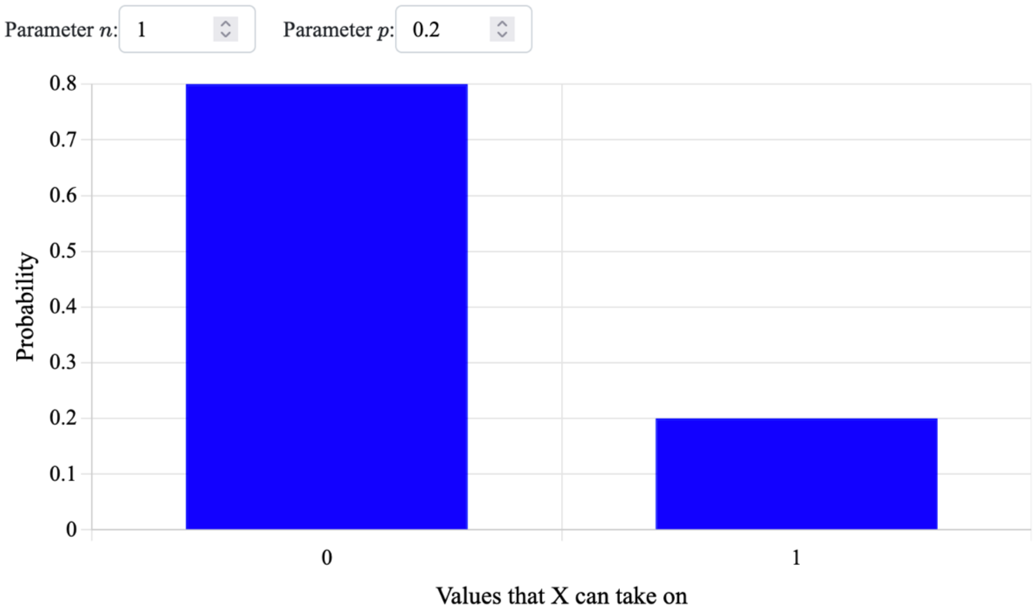
PMF of Bernoulli

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


Parameter p: 0.2

PMF of Binomial, *with n* = 1 *?!*

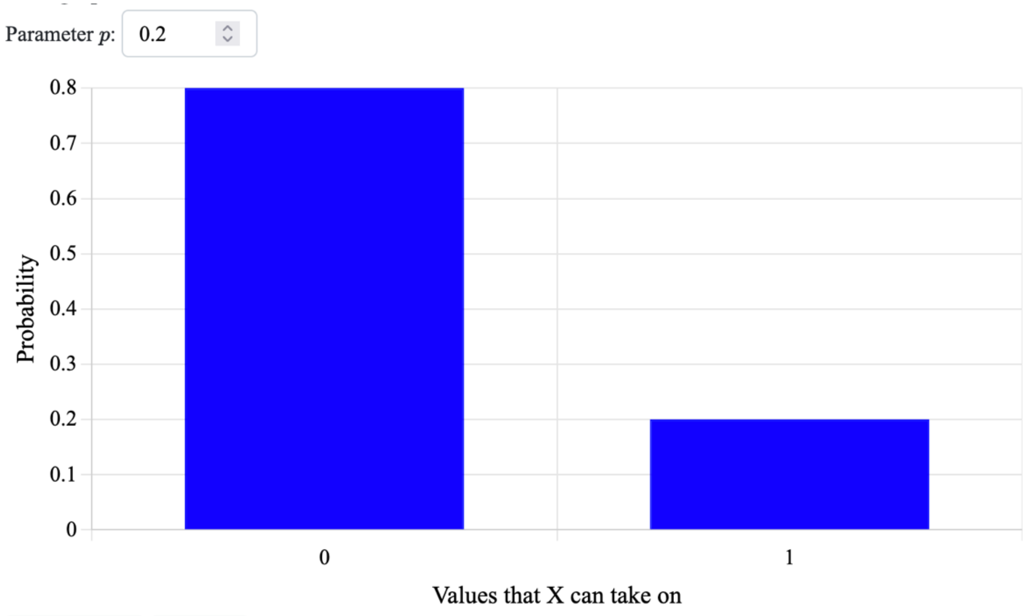
$$f(x_i|p) = p^{x_i}(1-p)^{1-x_i}$$


Parameter n: 1
Parameter p: 0.2

We Need A New Bernoulli PMF…

PMF of Bernoulli

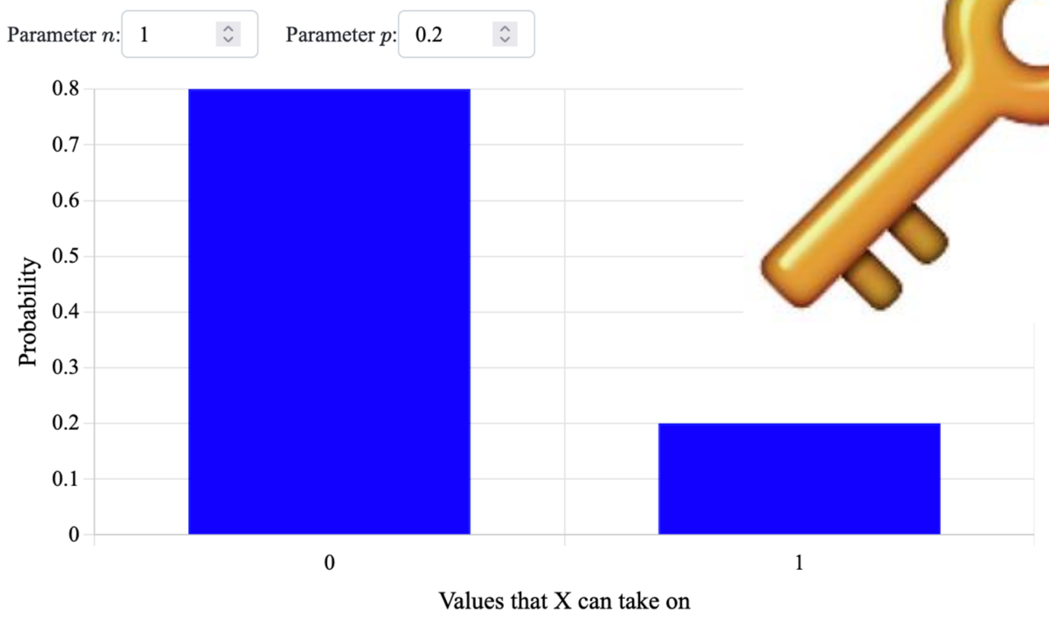
$$f(x_i|p) = \begin{cases} p & \text{if } x_i = 1 \\ 1-p & \text{if } x_i = 0 \end{cases}$$



Parameter p: 0.2

PMF of Binomial, *with n* = 1 *?!*

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



Parameter n: 1

Parameter p: 0.2

When we need to take the derivative, we use this “smooth” / “continuous” PMF for Bernoulli!

MLE for Bernoulli

We have 20 I.I.D. observations of $X_i \sim \text{Bern}(p)$. We want to estimate p.

Goal: Estimate *p* by MLE

PMF: $$\xcancel{f(x_i|p) = \begin{cases} p & \text{if } x_i = 1 \\ 1-p & \text{if } x_i = 0 \end{cases}} \longrightarrow L(p) = f(x_1, ..., x_n|p) = \prod_{i=1}^{n} \text{???}$$

MLE for Bernoulli

We have 20 I.I.D. observations of $X_i \sim \mathrm{Bern}(p)$. We want to estimate p.

Goal: Estimate p by MLE

PMF: $\boxed{f(x_i|p) = p^{x_i}(1-p)^{1-x_i}}$ $\longrightarrow$ $L(p) = f(x_1, ..., x_n|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$

"Smooth" Bernoulli PMF

MLE for Bernoulli

We have 20 I.I.D. observations of $X_i \sim \mathrm{Bern}(p)$. We want to estimate p.

Goal: Estimate p by MLE

PMF: $f(x_i|p) = p^{x_i}(1-p)^{1-x_i}$ → $L(p) = f(x_1, ..., x_n|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$

To compute $\hat{p} = \underset{p}{\operatorname{argmax}}\, L(p)$ **:**

1. **Take log of likelihood.**
2. **Take derivative w.r.t. p.**
3. **Set equal to 0 and solve for p.**

MLE for Bernoulli

We have 20 I.I.D. observations of $X_i \sim \mathrm{Bern}(p)$. We want to estimate p.

Goal: Estimate p by MLE

PMF: $f(x_i|p) = p^{x_i}(1-p)^{1-x_i} \longrightarrow L(p) = f(x_1, \ldots, x_n|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$

To compute $\hat{p} = \underset{p}{\operatorname{argmax}}\, L(p)$**:**

1. **Take log of likelihood.**
2. **Take derivative w.r.t. p.**
3. **Set equal to 0 and solve for p.**

$$
\begin{aligned}
LL(p) &= \sum_{i=1}^{n} \log\left[p^{x_i}(1-p)^{1-x_i}\right] \\
&= \sum_{i=1}^{n} x_i \log(p) + (1-x_i)\log(1-p)
\end{aligned}
$$

Take Derivative: $LL(p) = \sum_{i=1}^{n} x_i \log p + (1 - x_i) \log(1 - p)$

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

Take the derivative w.r.t. *p*

$$= \sum_{i=1}^{n} \boxed{\frac{\partial}{\partial p} x_i \log p} + \frac{\partial}{\partial p} (1 - x_i) \log(1 - p)$$

Derivative of a sum

$$= \sum_{i=1}^{n} \frac{x_i}{p} + \boxed{\frac{\partial}{\partial p} (1 - x_i) \log(1 - p)}$$

Derivative of log(*p*) = 1/*p*

$$= \sum_{i=1}^{n} \frac{x_i}{p} - \frac{1 - x_i}{1 - p}$$

Derivative of log(*p*) = -1/(1-*p*)

MLE for Bernoulli

We have 20 I.I.D. observations of $X_i \sim \mathrm{Bern}(p)$. We want to estimate p.

Goal: Estimate p by MLE

PMF: $f(x_i|p) = p^{x_i}(1-p)^{1-x_i}$ $\longrightarrow$ $L(p) = f(x_1, ..., x_n|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$

To compute $\hat{p} = \underset{p}{\operatorname{argmax}}\, L(p)$**:**

1. **Take log of likelihood.**
2. **Take derivative w.r.t. p.**
3. **Set equal to 0 and solve for p.**

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

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

MLE for Bernoulli

We have 20 I.I.D. observations of $X_i \sim \mathrm{Bern}(p)$. We want to estimate p.

Goal: Estimate p by MLE

PMF: $f(x_i|p) = p^{x_i}(1-p)^{1-x_i}$ $\longrightarrow$ $L(p) = f(x_1, \ldots, x_n|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$

To compute $\hat{p} = \underset{p}{\operatorname{argmax}}\, L(p)$

1. **Take log of likelihood.**
2. **Take derivative w.r.t. p.**
3. **Set equal to 0 and solve for p.**

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

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

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

Solving for *p*: Full Walkthrough

$$0 = \sum_{i=1}^{n} \frac{x_i}{\hat{p}} - \frac{1 - x_i}{1 - \hat{p}}$$

$$= \sum_{i=1}^{n} \frac{x_i}{\hat{p}} - \sum_{i=1}^{n} \frac{1 - x_i}{1 - \hat{p}}$$

$$= \frac{y}{\hat{p}} - \frac{n - y}{1 - \hat{p}}$$

$$\frac{n - y}{1 - \hat{p}} = \frac{y}{\hat{p}}$$

$$\hat{p}(n - y) = y(1 - \hat{p})$$

$$\hat{p}n - \hat{p}y = y - \hat{p}y$$

$$\hat{p}n = y$$

Let $\sum_{i=1}^{n} x_i = y$ To make life easier

And $\sum_{i=1}^{n} 1 - x_i = \sum_{i=1}^{n} 1 - \sum_{i=1}^{n} x_i = n - y$

$$\hat{p} = \frac{1}{n} y$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i$$

MLE for Bernoulli

We have 20 I.I.D. observations of $X_i \sim \mathrm{Bern}(p)$. We want to estimate p.

Goal: Estimate *p* by MLE

PMF: $f(x_i|p) = p^{x_i}(1-p)^{1-x_i}$ $\longrightarrow$ $L(p) = f(x_1, ..., x_n|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$

To compute $\hat{p} = \underset{p}{\operatorname{argmax}}\, L(p)$

1. **Take log of likelihood.**
2. **Take derivative w.r.t. *p*.**
3. **Set equal to 0 and solve for *p*.**

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

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

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

Don’t we already have the Beta?

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

MLE for Bernoulli vs. The Beta

A new drug is given to 20 patients. It “works” for 14 and “doesn’t work” for 6. What is your best estimate of the probability that the drug will work for the next patient?

In other words:

We have 20 I.I.D. observations of $X_i \sim \text{Bern}(p)$. We want to estimate p.
The data: [1,1,0,1,1,1,0,0,1,0,1,1,1,0,1,0,1,1,1,1]

MLE estimate: $p \approx \frac{14}{20} = 0.7$

MLE for Bernoulli vs. The Beta

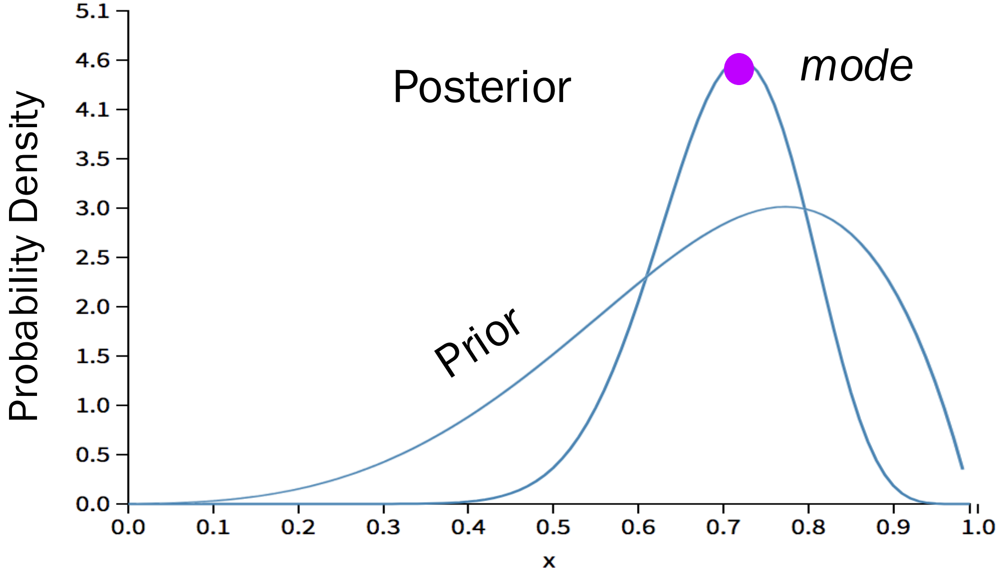
A new drug is given to 20 patients. It “works” for 14 and “doesn’t work” for 6. What is your best estimate of the probability that the drug will work for the next patient?

In other words:

We have 20 I.I.D. observations of $X_i \sim \text{Bern}(p)$. We want to estimate p.
The data: [1,1,0,1,1,1,0,0,1,0,1,1,1,0,1,0,1,1,1,1]

MLE estimate: $p \approx \frac{14}{20} = 0.7$

With a $\text{Beta}(a = 14 + 1, b = 6 + 1)$:



MLE for Bernoulli vs. The Beta

A new drug is given to 20 patients. It "works" for 14 and "doesn't work" for 6. What is your best estimate of the probability that the drug will work for the next patient?

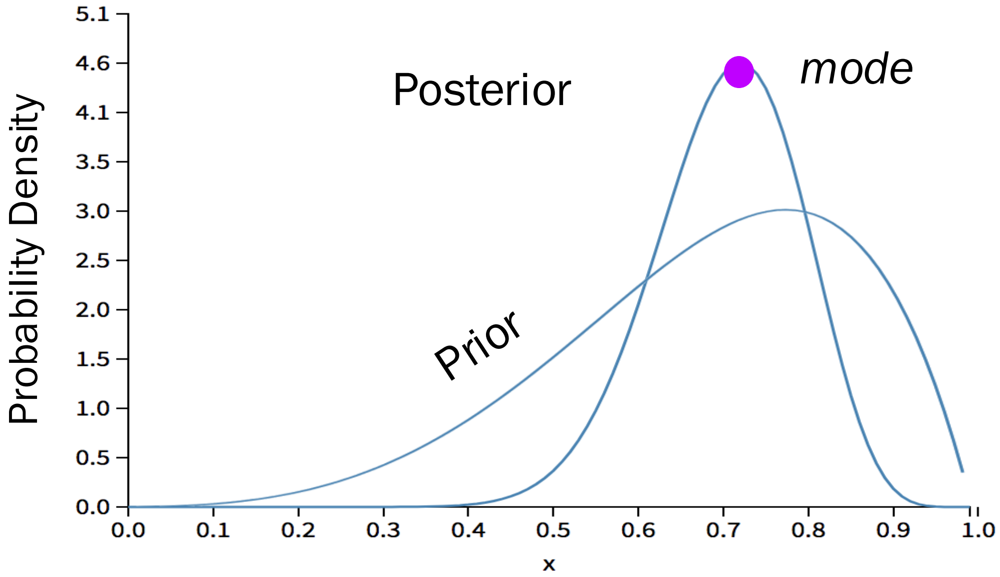
In other words:

We have 20 I.I.D. observations of $X_i \sim \text{Bern}(p)$. We want to estimate p.
The data: [1,1,0,1,1,1,0,0,1,0,1,1,1,0,1,0,1,1,1,1]

MLE estimate: $$p \approx \frac{14}{20} = 0.7$$

When do we want
a **point estimate**
vs. an entire **distribution**?

With a $\text{Beta}(a = 14 + 1, b = 6 + 1)$:



MLE for Gaussian

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

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 datset of n i.i.d. random variables $X_1, X_2, \ldots, X_n$. $X_i \sim N(\mu, \sigma^2)$. Find $\hat{\mu}$, $\hat{\sigma}^2$.

Step 0: Write out the likelihood: $L(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$

Then, find the values of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ which maximize $L(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$

Maximum Likelihood with Normal

Consider a datset of n i.i.d. random variables $X_1, X_2, \ldots, X_n$. $X_i \sim N(\mu, \sigma^2)$. Find $\hat{\mu}$, $\hat{\sigma}^2$.

$$f(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) \xrightarrow{\text{Write } L(\theta)} L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)$$

Then, find the values of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ which maximize $L(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$

Maximum Likelihood with Normal

Consider a datset of n i.i.d. random variables $X_1, X_2, \ldots, X_n$. $X_i \sim N(\mu, \sigma^2)$. Find $\hat{\mu}$, $\hat{\sigma}^2$.

$$f(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(\frac{-(x_i-\mu)^2}{2\sigma^2}\Big) \xrightarrow{\text{Write } L(\theta)} L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(\frac{-(x_i-\mu)^2}{2\sigma^2}\Big)$$

To compute $\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}}\, L(\mu,\sigma^2)$

1. **Take log of likelihood.**
2. **Take derivative w.r.t. each param.**
3. **Set equal to 0 and solve for each param.**

Maximum Likelihood with Normal

Consider a datset of n i.i.d. random variables $X_1, X_2, \dots, X_n$. $X_i \sim N(\mu, \sigma^2)$. Find $\hat{\mu}, \hat{\sigma}^2$.

$$f(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right) \xrightarrow{\text{Write } L(\theta)} L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right)$$

To compute $\hat{\mu}, \hat{\sigma}^2 = \underset{\mu, \sigma^2}{\operatorname{argmax}}\, L(\mu, \sigma^2)$

1. Take log of likelihood.
2. Take derivative w.r.t. each param.
3. Set equal to 0 and solve for each param.

$$LL(\theta) = -\sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

Maximum Likelihood with Normal

Consider a datset of n i.i.d. random variables $X_1, X_2, \dots, X_n$. $X_i \sim N(\mu, \sigma^2)$. Find $\hat{\mu}, \hat{\sigma}^2$.

$$f(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) \xrightarrow{\text{Write } L(\theta)} L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)$$

To compute $\hat{\mu}, \hat{\sigma}^2 = \operatorname*{argmax}_{\mu,\sigma^2} L(\mu, \sigma^2)$

1. **Take log of likelihood.**
2. **Take derivative w.r.t. each param.**
3. **Set equal to 0 and solve for each param.**

$$LL(\theta) = -\sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

First, estimate $\boldsymbol{\mu}$:

$$\frac{\partial}{\partial\mu} LL(\theta) = \sum_{i=1}^{n} \frac{2\cdot(x_i-\mu)}{2\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

Maximum Likelihood with Normal

Consider a datset of n i.i.d. random variables $X_1, X_2, \ldots, X_n$. $X_i \sim N(\mu, \sigma^2)$. Find $\hat{\mu}$, $\hat{\sigma}^2$.

$$f(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) \xrightarrow{\text{Write } L(\theta)} L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)$$

To compute $\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}}\, L(\mu, \sigma^2)$

1. **Take log of likelihood.**
2. **Take derivative w.r.t. each param.**
3. **Set equal to 0 and solve for each param.**

$$LL(\theta) = -\sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

First, estimate $\boldsymbol{\mu}$:

$$\frac{\partial}{\partial\mu}LL(\theta) = \sum_{i=1}^{n} \frac{2\cdot(x_i-\mu)}{2\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Maximum Likelihood with Normal

Consider a datset of n i.i.d. random variables $X_1, X_2, \ldots, X_n$. $X_i \sim N(\mu, \sigma^2)$. Find $\hat{\mu}$, $\hat{\sigma}^2$.

$$f(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) \xrightarrow{\text{Write } L(\theta)} L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)$$

To compute $\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\arg\max}\, L(\mu, \sigma^2)$

1. **Take log of likelihood.**
2. **Take derivative w.r.t. each param.**
3. **Set equal to 0 and solve for each param.**

$$LL(\theta) = -\sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Then, estimate $\boldsymbol{\sigma}^2$:

$$\begin{aligned}\frac{\partial}{\partial\sigma}LL(\theta) &= -\sum_{i=1}^{n}\frac{1}{\sigma} + \sum_{i=1}^{n}\frac{2(x_i-\mu)^2}{2\sigma^3} \\ &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2\end{aligned}$$

Maximum Likelihood with Normal

Consider a datset of n i.i.d. random variables $X_1, X_2, \ldots, X_n$. $X_i \sim N(\mu, \sigma^2)$. Find $\hat{\mu}$, $\hat{\sigma}^2$.

$$f(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) \xrightarrow{\text{Write } L(\theta)} L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)$$

To compute $\hat{\mu}, \hat{\sigma}^2 = \operatorname*{argmax}_{\mu,\sigma^2} L(\mu,\sigma^2)$

1. **Take log of likelihood.**
2. **Take derivative w.r.t. each param.**
3. **Set equal to 0 and solve for each param.**

$$LL(\theta) = -\sum_{i=1}^{n} \log\frac{1}{\sqrt{2\pi}\sigma} - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

Then, estimate $\boldsymbol{\sigma^2}$:

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

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

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

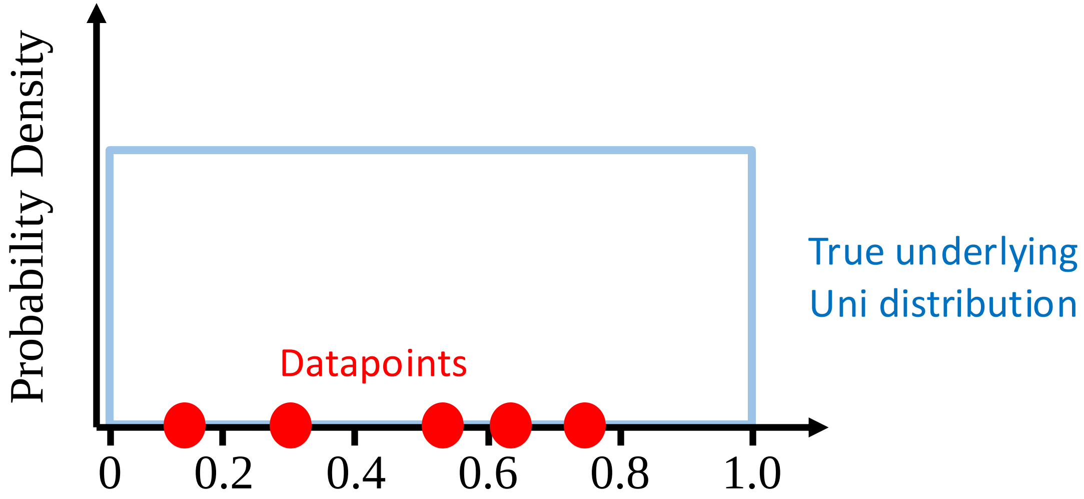
MLE gives a biased estimate for variance…

MLE For Uniform

Consider a n i.i.d. random variables $X_1, X_2, \dots, X_n$. $X_i \sim \text{Uni}(0, 1)$.

We observe $[0.15, 0.31, 0.54, 0.62, 0.75]$.

What parameter values for $\text{Uni}(\alpha, \beta)$ will MLE choose?

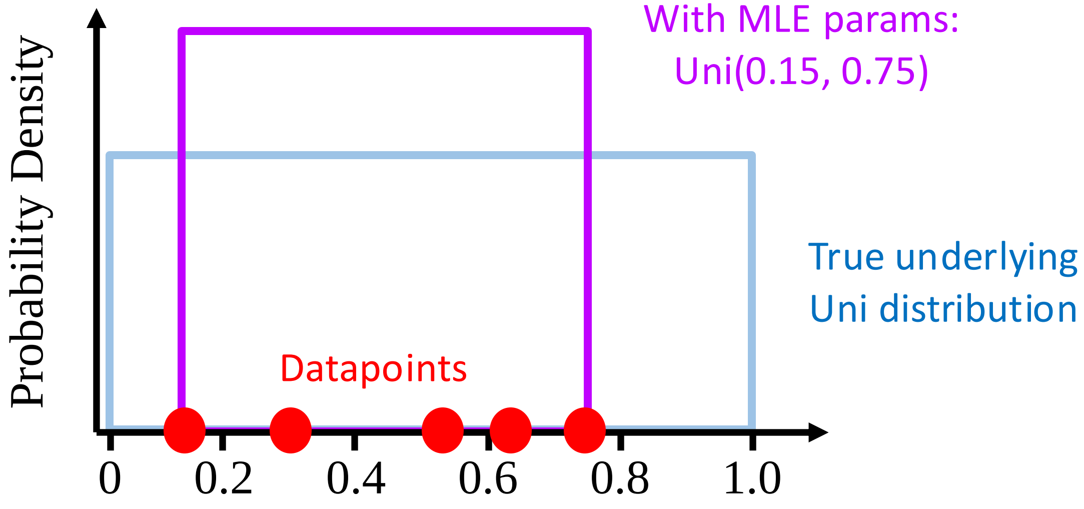


MLE For Uniform

Consider a n i.i.d. random variables $X_1, X_2, \ldots, X_n$. $X_i \sim \text{Uni}(0, 1)$.

We observe $[0.15, 0.31, 0.54, 0.62, 0.75]$.

What parameter values for $\text{Uni}(\alpha, \beta)$ will MLE choose?

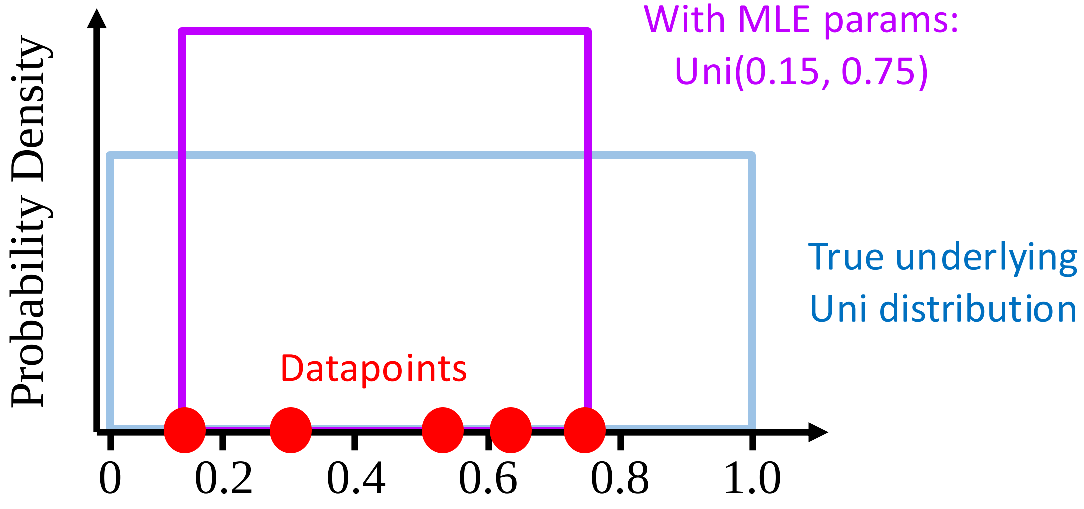


MLE For Uniform

Consider a n i.i.d. random variables $X_1, X_2, \ldots, X_n$. $X_i \sim \text{Uni}(0, 1)$.

We observe $[0.15, 0.31, 0.54, 0.62, 0.75]$.

What parameter values for $\text{Uni}(\alpha, \beta)$ will MLE choose?



MLE overfits to the data!

We also have no way to incorporate a prior belief…

Pros and Cons of MLE

Pros:

1. We can understand reasoning behind some parameter estimates (often the mean)

2. Once we know MLE estimate formula, we can apply it over and over again

3. *Asymptotically optimal:* in the limit (with infinite data), we converge to the true value

4. You don’t need to think about what a good prior belief is for each parameter

Pros and Cons of MLE

Pros:

1. We can understand reasoning behind some parameter estimates (often the mean)
2. Once we know MLE estimate formula, we can apply it over and over again
3. *Asymptotically optimal:* in the limit (with infinite data), we converge to the true value
4. You don't need to think about what a good prior belief is for each parameter

Cons:

1. *Biased estimates:* no guarantee that we don't tend to over- or under- estimate
2. *Overfitting*: doesn't attempt to generalize to unseen data
3. ^ This is problematic for small *n!*
4. No option for a prior belief for params

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

MLE Plotline

1. Decide on a model for the distribution of your data. Identify the PMF / PDF.

2. Write out the likelihood function.

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

4. Find the argmax.

4a. Take log

4b. Take derivative w.r.t. the parameter

4c. Set derivative equal to 0 and solve

MLE Plotline

1. Decide on a model for the distribution of your data. Identify the PMF / PDF.

2. Write out the likelihood function.

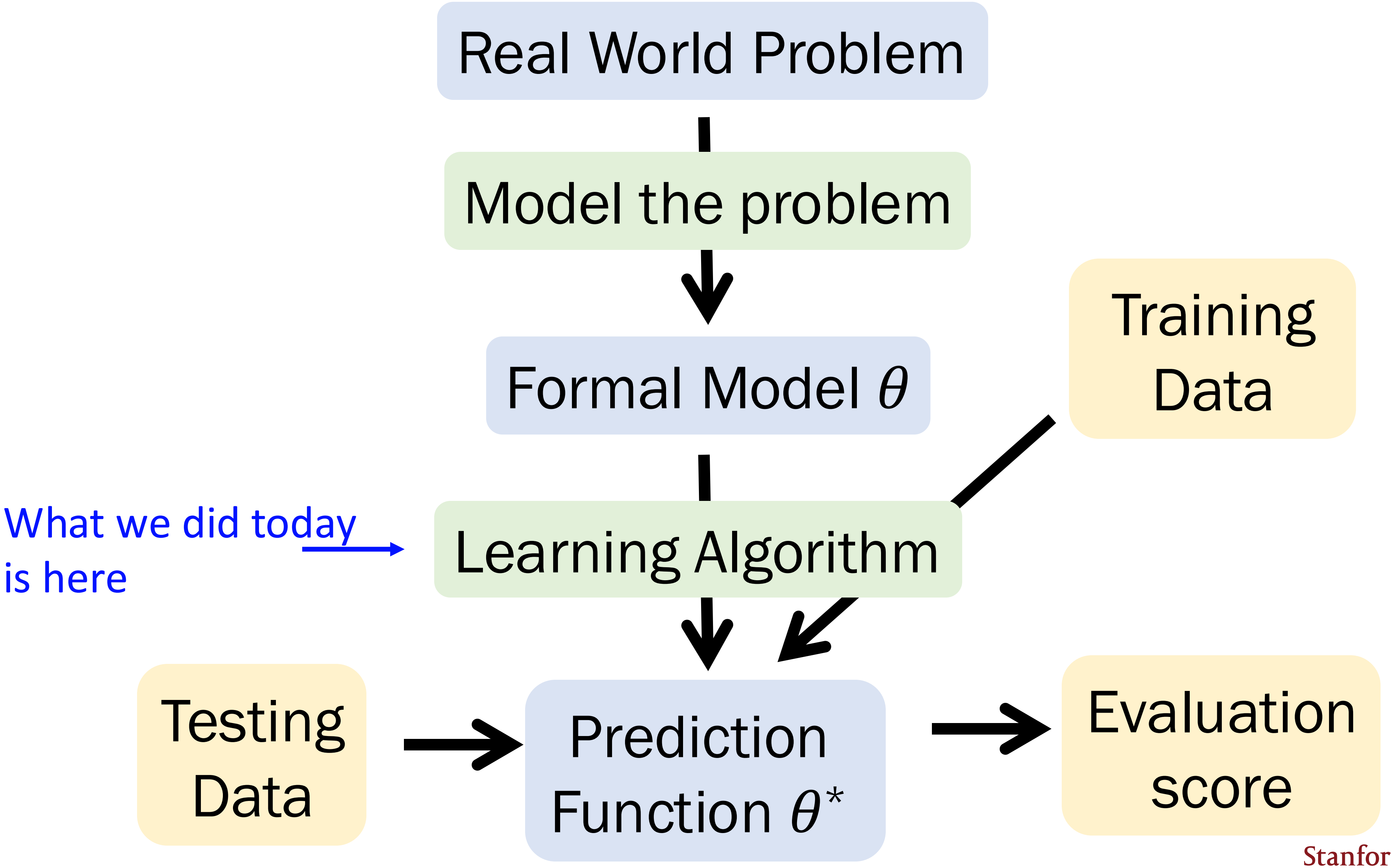
3. State that the optimal parameters are the argmax of the likelihood function.

4. Find the argmax.

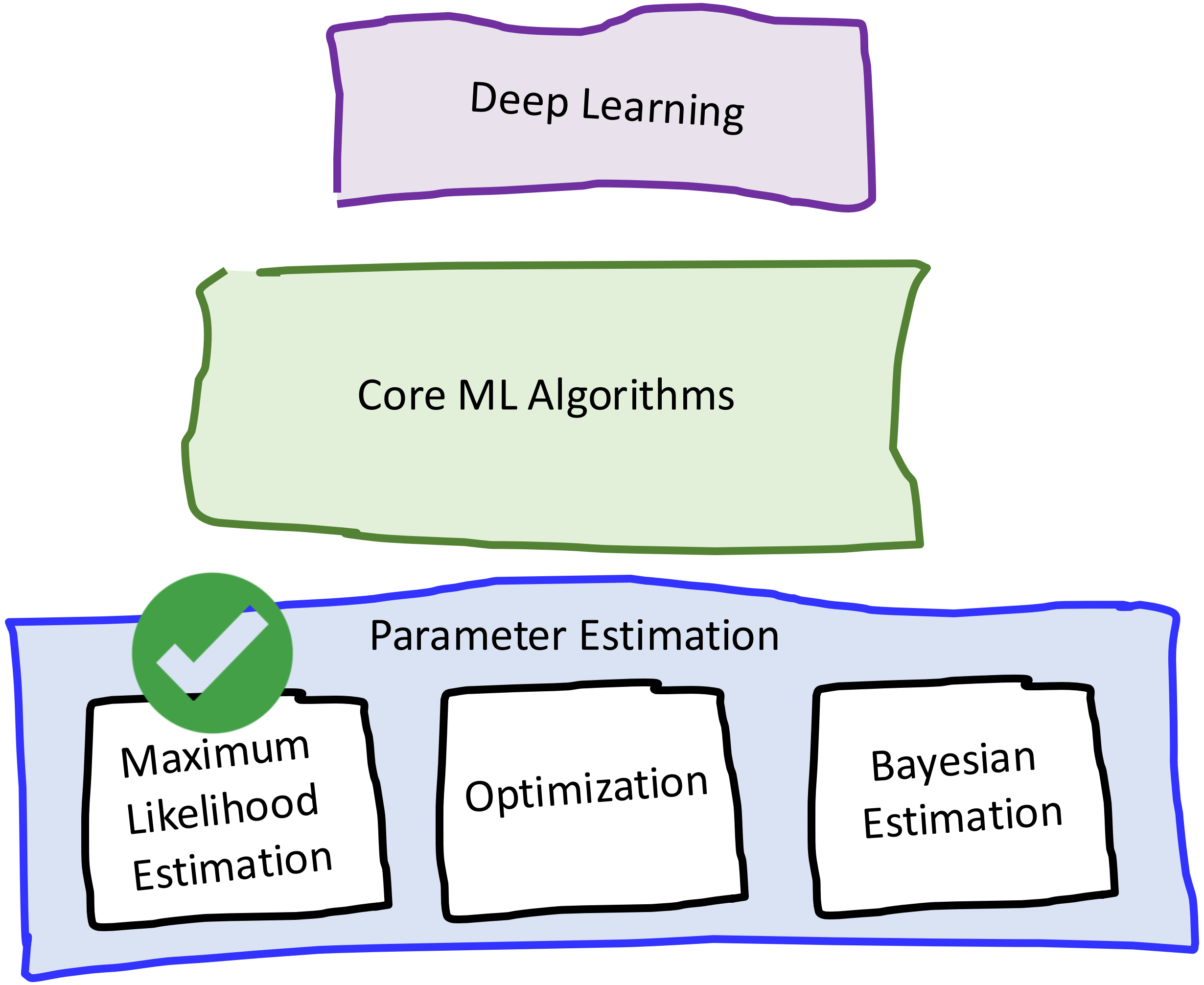
Next lecture:

can we make computers do this for us?

Next Week: Full Machine Learning Workflow



Our Path



Deep Learning

Core ML Algorithms

Parameter Estimation

✓ Maximum Likelihood Estimation

Optimization

Bayesian Estimation