Machine Learning and Data Mining

Lecture 1.2: Statistical Learning – MLE / MAP

CS 434

HW1 Out!

HW1 is out now. Due in 2 weeks. Has two parts:

- Implement k Nearest Neighbors and K-Fold Cross Validation and reason about our dataset/preprocessing (80% of the points)

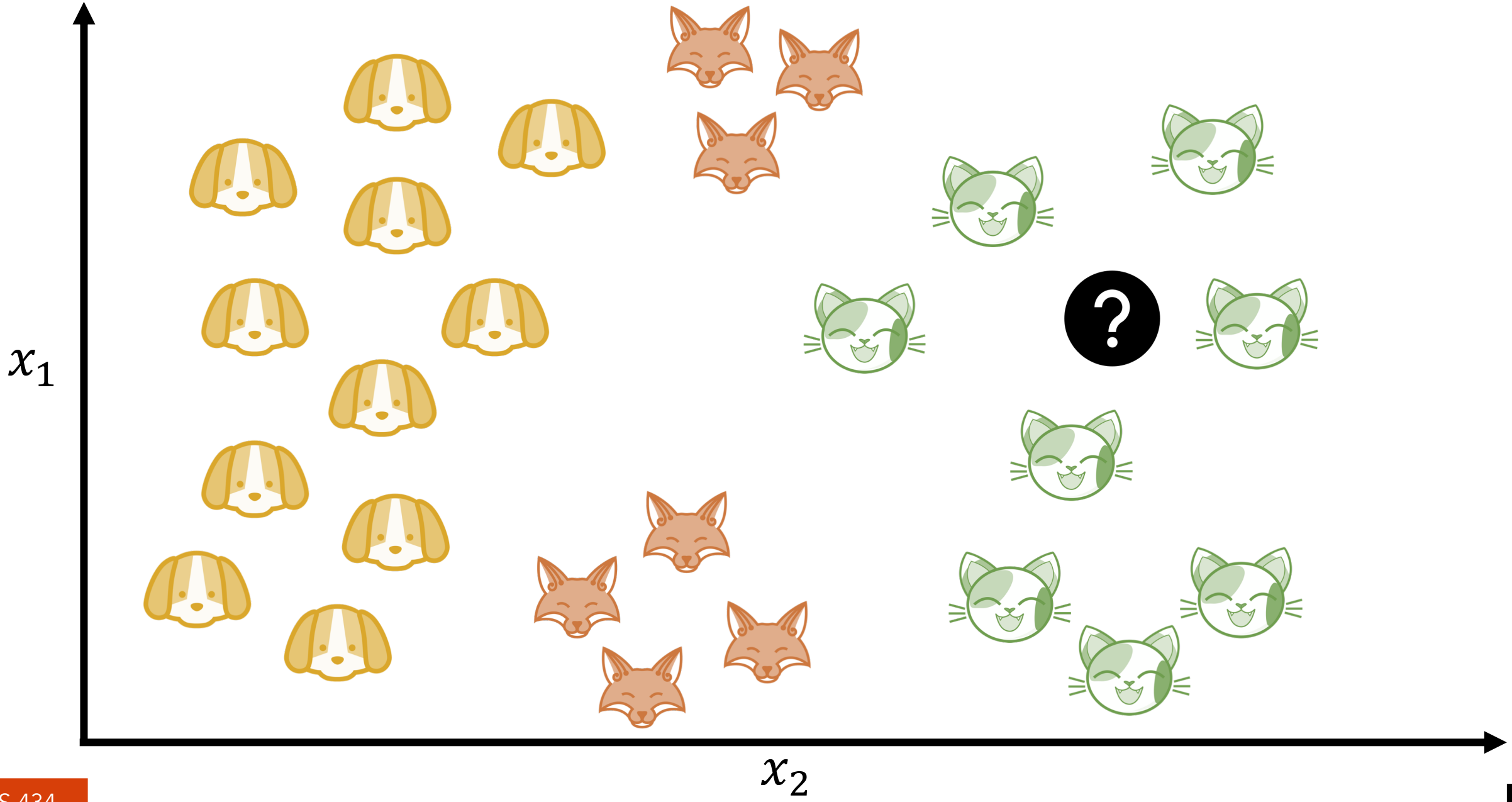
- Some math to fit parameters to a Poisson-Gamma model. We will cover the principles involved today with different distributions. (20% of the points)

Involves submission to a class Kaggle competition for the test set. Bonus points for beating the TA entry. Bragging rights for being in the top 5.

RECAP

From Last Lecture

Intuition for your First Machine Learning Algorithm

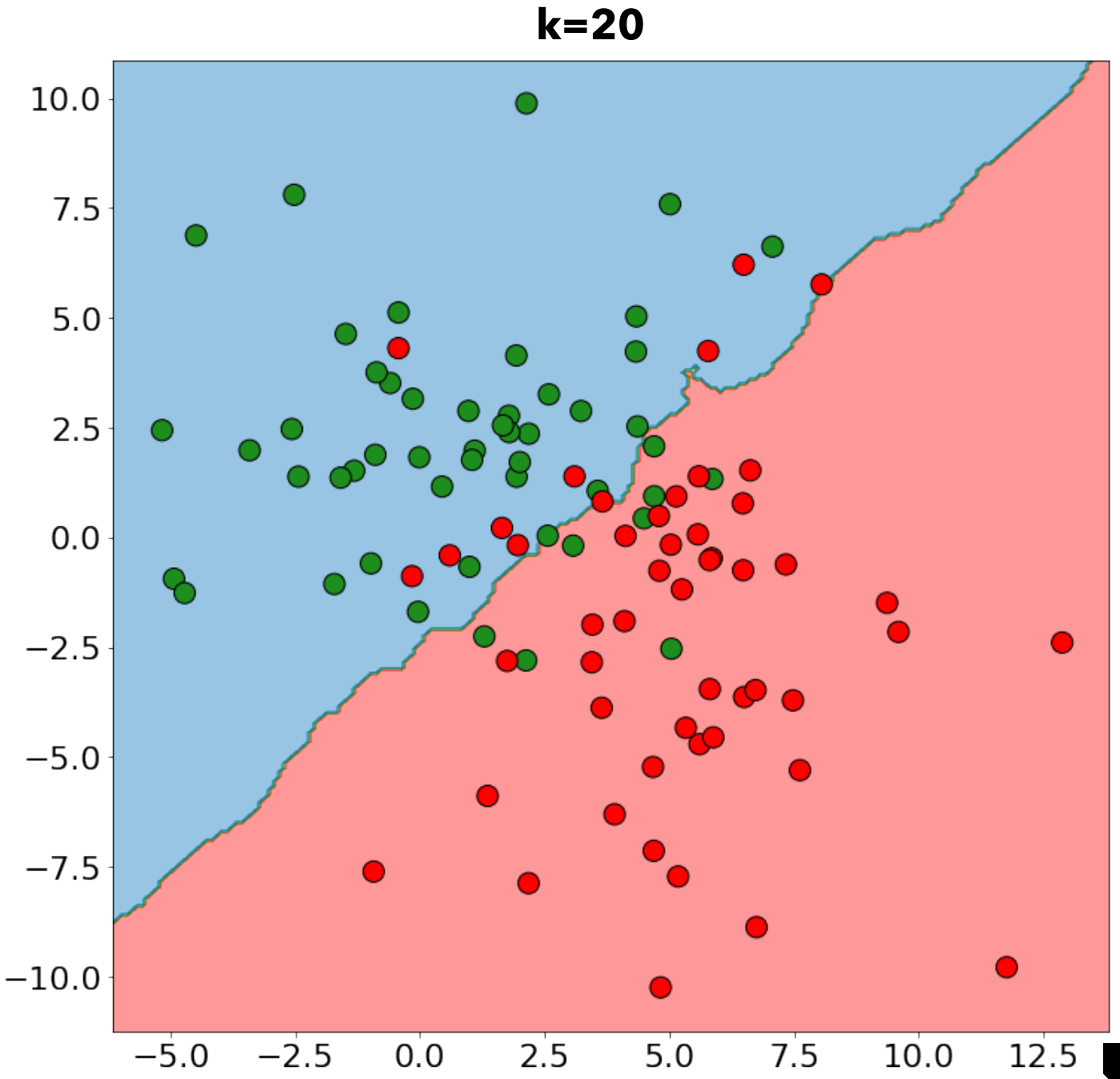


More on the Effect of k on Decision Boundaries

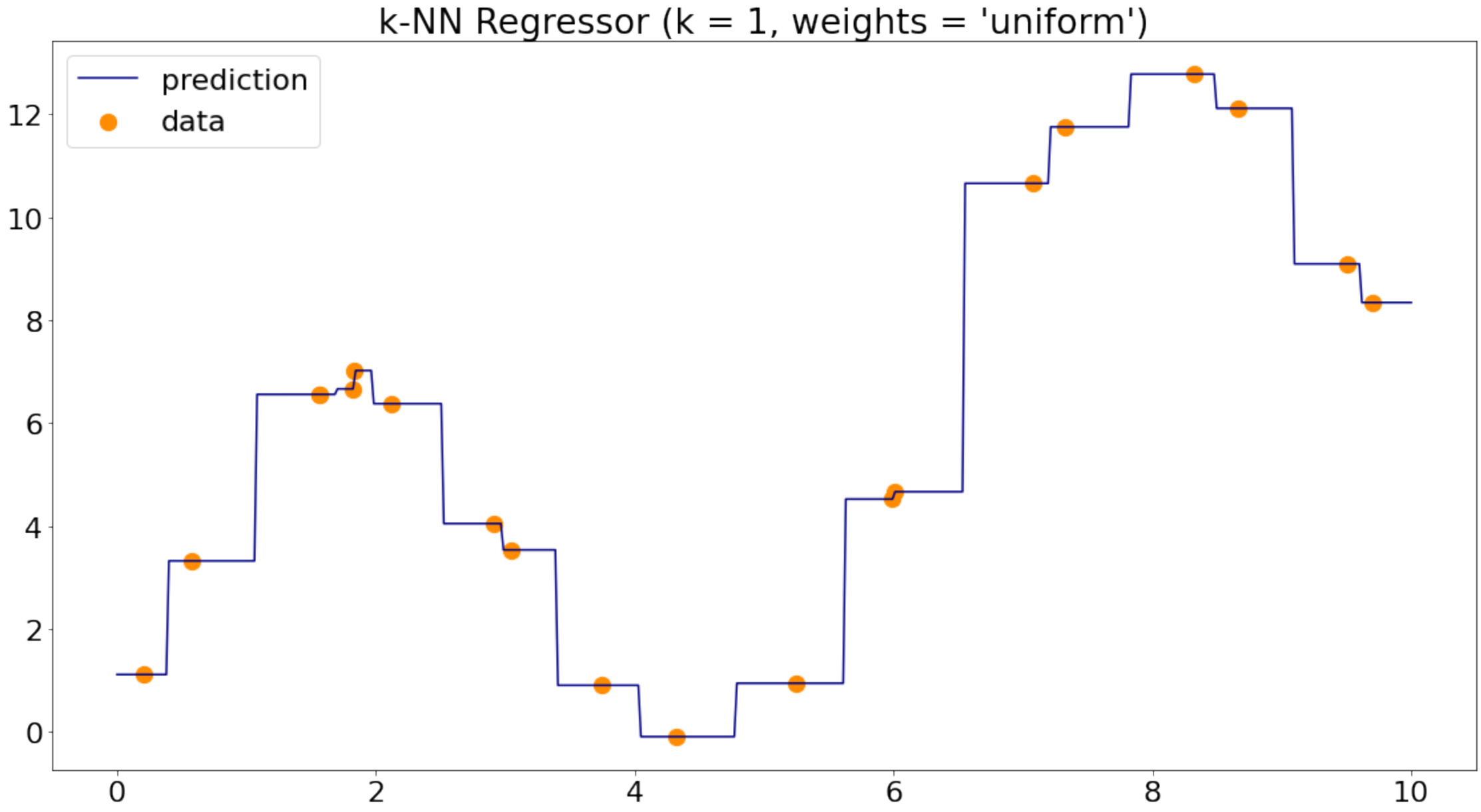
If your dataset has noise or many examples, these decision boundaries can become nasty!

Noisy training examples might lead to poor generalization to new data.

We can try to combat this by increasing the number of neighbors we examine.



Formalizing k-Nearest Neighbors for Regression



Weighted k-Nearest Neighbor

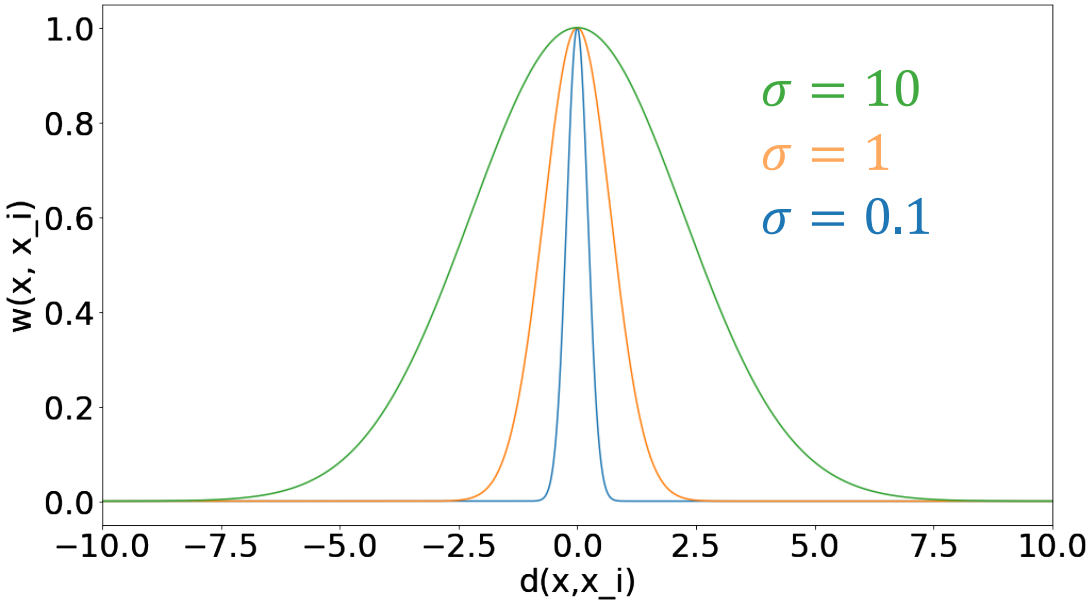
What makes a good weight function?

- Gets small as distance from point gets large

One good candidate:

$$w(x, x_i) = e^{-\frac{d(x,x_i)^2}{\sigma}}$$

σ is sometimes referred to as "bandwidth"



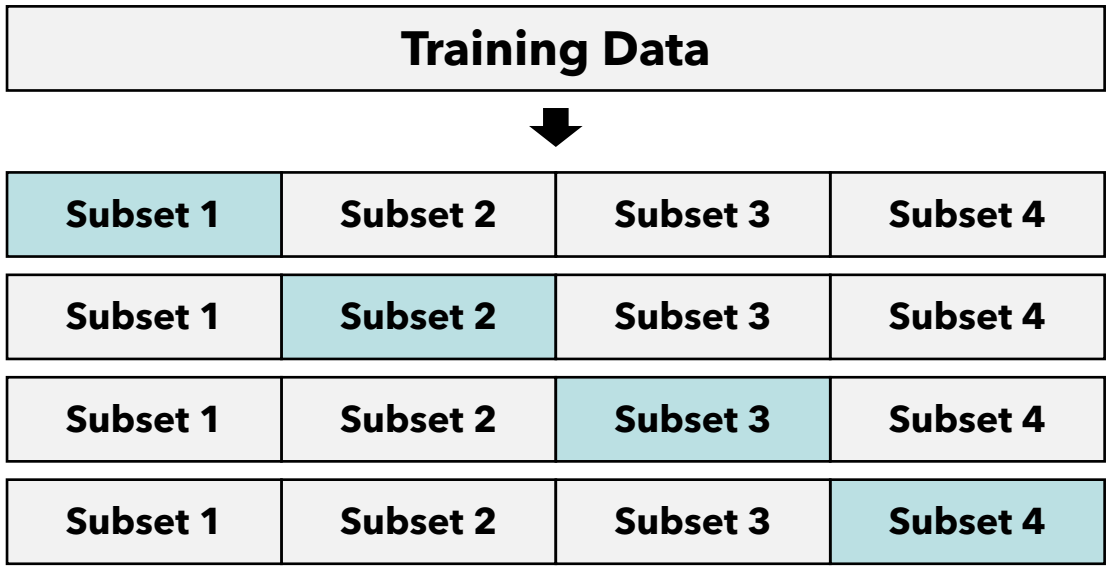
Does these curves look familiar?

Gaussians my dear Watson. Gaussians!

$$\mathcal{N}(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma}}$$

Model Selection – K-fold Cross Validation

Idea: Make many validation sets and evaluate model choices by average performance over them.



Example: 4-fold cross validation

Train on subsets 2,3,4. Test on subset 1 → $perf_1$

Train on subsets 1,3,4. Test on subset 2 → $perf_2$

Train on subsets 1,2,4. Test on subset 3 → $perf_3$

Train on subsets 1,2,3. Test on subset 4 → $perf_4$

Estimated validation performance is $\frac{1}{4}(\sum_i perf_i)$

Today's Learning Objectives

Be able to answer:

- What are basic concepts and definitions in probability theory?
- What is maximum likelihood estimation (MLE)?
  - What is a likelihood function?
  - What is the MLE estimate for a Bernoulli variable?
  - What is bias and variance of an estimator?
  - What is the MLE estimate for a Normal variable?
- What is a maximum a posteriori estimate (MAP)?
  - What is a prior and posterior? What is a

Why do we care about probability?

The world is a ***very*** uncertain place

- 30 years of Artificial Intelligence research danced around this fact

- Then some AI researchers decided to use some ideas from the eighteenth century to model uncertainty…

  - **For us:** Really useful view of machine learning (both in practice for algorithm design and for deriving theoretic results about learning)

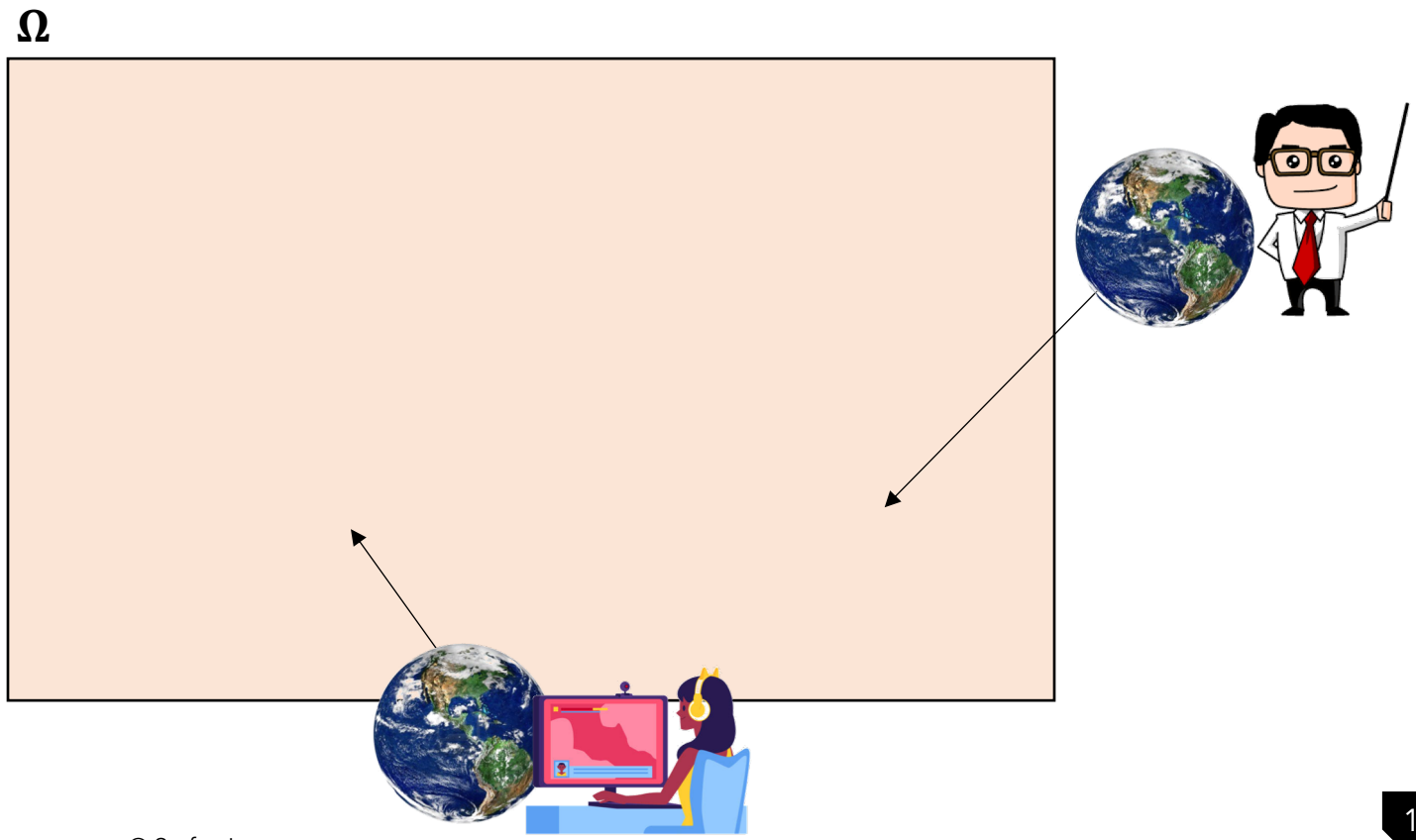
Slide adapted from Dhruv Batra. Comic a modification from 2016 webcomic Shen Comics

Brief Probability Review

Humans are sort of trash at probability. Doubly so once equations get involved.

Easiest to imagine a set of worlds where every possible outcome occurs.
Example: Consider the probability of me having various careers.

Sample Space Ω is the set of all possible outcomes.



Sample Space Ω is the set of all possible outcomes.

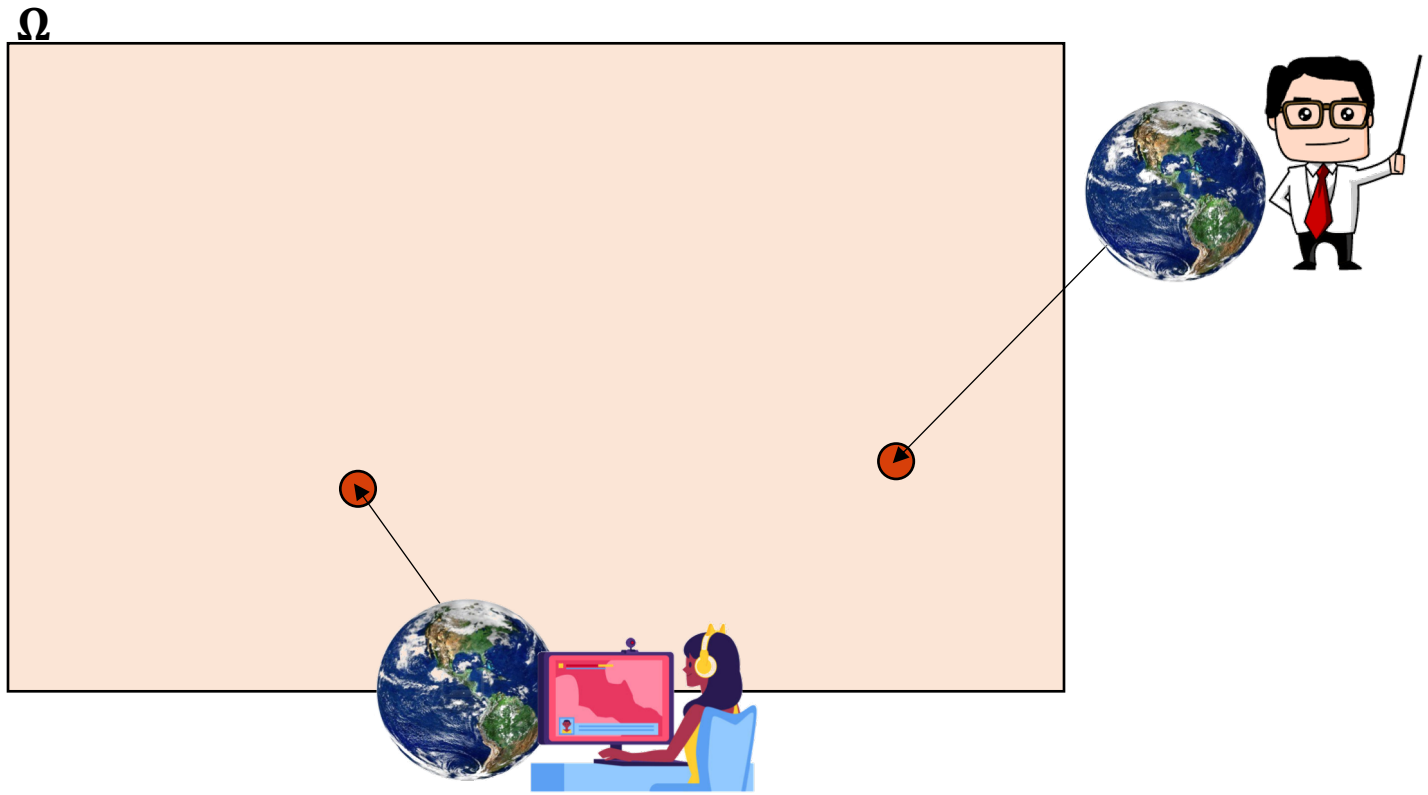
An **Event A** is a subset of Ω

$$A = \{streamer\}$$

Or

$$A = \{streamer, farmer\}$$

Think of this as an or



Brief Probability Review

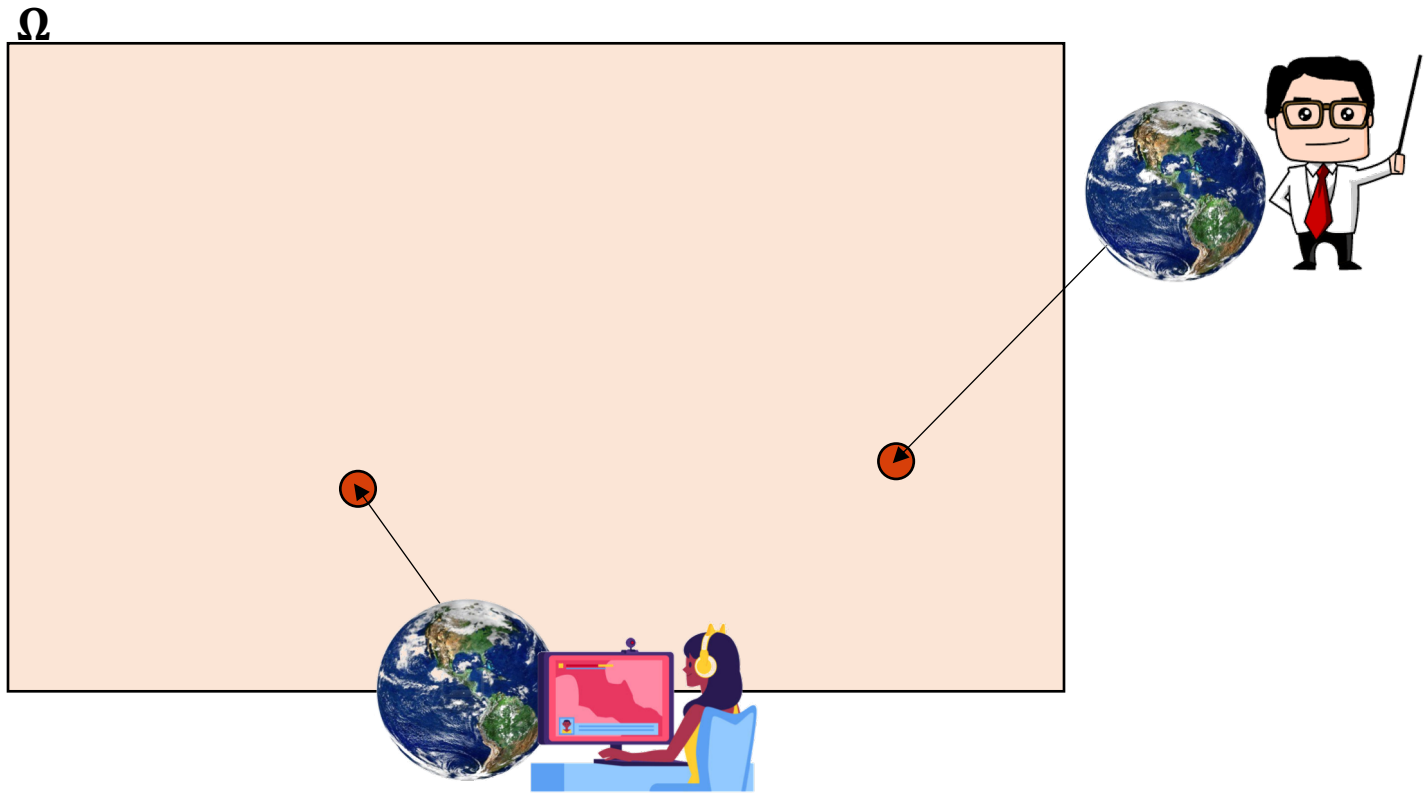
Sample Space Ω is the set of all possible outcomes.
An **Event A** is a subset of **Ω,** e.g. A={streamer} or {streamer, farmer}

Assign a real number
to each event P(A).
Call it a probability if:

1) $\boldsymbol{P(E) \geq 0 \; \forall E}$

2) $\boldsymbol{P(\Omega) = 1, \qquad P(\{\}) = 0}$

3*) $\boldsymbol{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$



*This is not the σ-addivity axiom typically cited as the 3rd axiom but is roughly equivalent. Good enough for us.

Brief Probability Review

Sample Space Ω is the set of all possible outcomes.
An **Event A** is a subset of **Ω,** e.g. A={streamer} or {streamer, farmer}

Assign a real number
to each event P(A).
Call it a probability if:

1) $P(E) \geq 0 \; \forall E$

2) $P(\Omega) = 1, \qquad P(\{\}) = 0$

3*) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Brief Probability Review

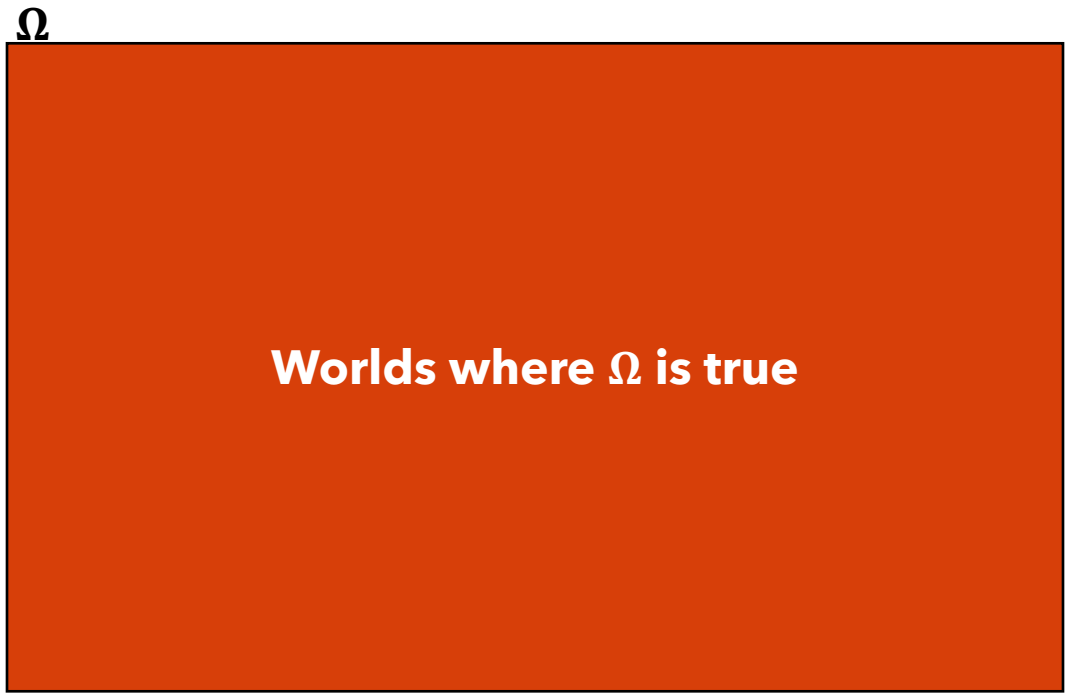
Sample Space Ω is the set of all possible outcomes.
An **Event A** is a subset of **Ω,** e.g. A={streamer} or {streamer, farmer}

Assign a real number
to each event P(A).
Call it a probability if:

1) $\boldsymbol{P(E) \geq 0 \; \forall E}$

2) $\boldsymbol{P(\Omega) = 1,}$ $\qquad$ $\boldsymbol{\mathbf{P}(\{\}) = 0}$

3*) $\boldsymbol{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$



Worlds where any event occurs

Brief Probability Review

Sample Space Ω is the set of all possible outcomes.
An **Event A** is a subset of **Ω,** e.g. A={streamer} or {streamer, farmer}

Assign a real number
to each event P(A).
Call it a probability if:

1) $\boldsymbol{P(E) \geq 0 \; \forall E}$

2) $\boldsymbol{P(\Omega) = 1}$, $\mathbf{P(\{\}) = 0}$

3*) $\boldsymbol{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$

Worlds where none of the
possible events occur?

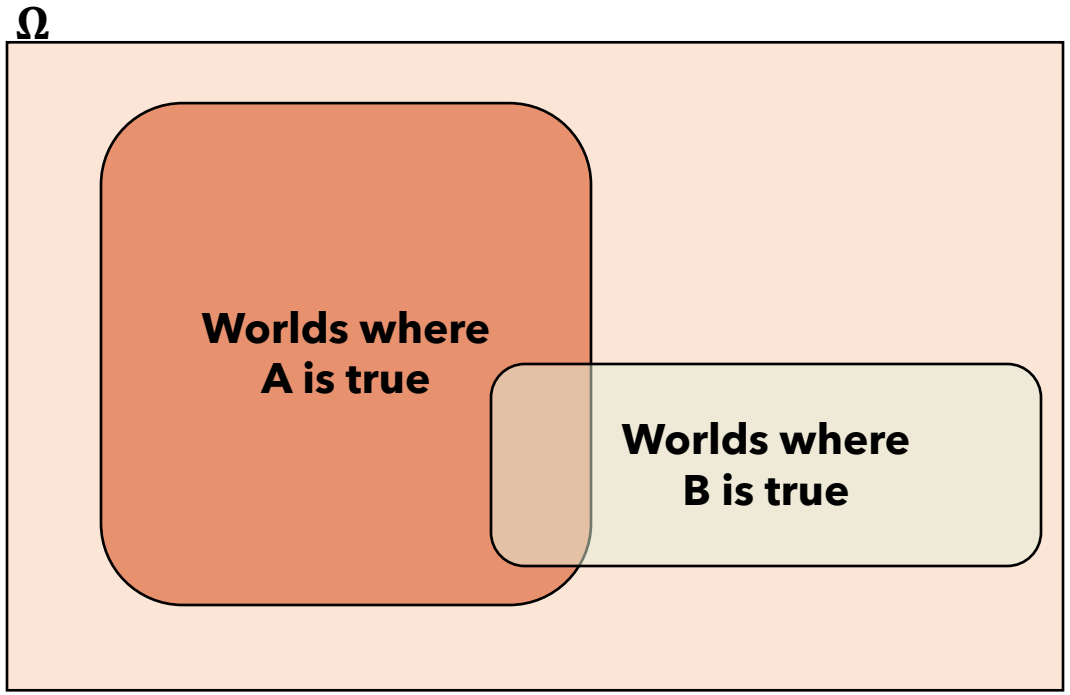
Sample Space Ω is the set of all possible outcomes.
An **Event A** is a subset of **Ω,** e.g. A={streamer} or {streamer, farmer}

Assign a real number to each event P(A).
Call it a probability if:

1) $P(E) \geq 0 \; \forall E$

2) $P(\Omega) = 1, \qquad P(\{\}) = 0$

3*) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Don't want to double count the area of intersection!

Brief Probability Review

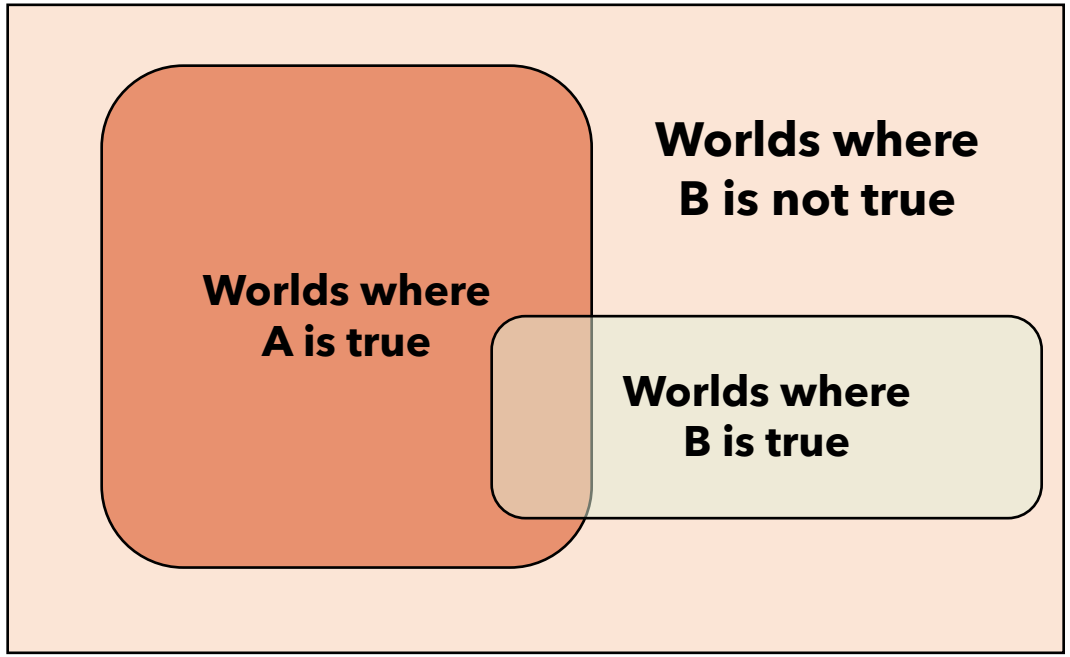
Marginalization: Can "sum out" events from joint distributions

Discrete:

$$P(A) = \sum_{b \in Val(B)} P(A, B = b)$$

Continuous:

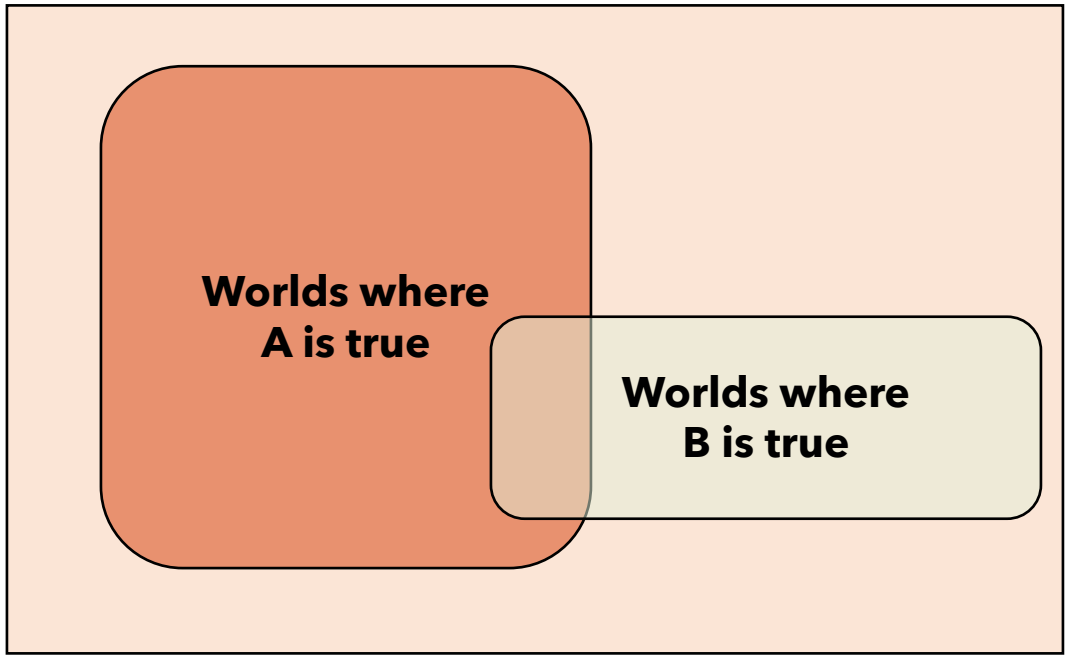
$$P(A) = \int_{b \in Val(B)} P(A, B = b)\, db$$



$$P(A) = P(A, B = False) + P(A, B = True)$$

Conditional Distributions: P(A|B) – Of worlds where B occurs, what fraction also have A?

$$P(A|B) = \frac{P(A,B)}{P(B)}$$



Chain Rule: $$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$

Bayes Rule: A useful way to reverse direction of conditional distributions. Combines chain rule and definition of conditional distribution. Simple yet fundamental.

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

A **Random Variable** X is a mapping between events in **Ω** to numbers.

X(A)	0	1	2	3	4
Event A	Streamer	Professor	Telemarketer	Programmer	…

Coin flip example. What does the random variable Y represent?

Y(B)	2	1	2	0	…
Event B	H,H,T	T,T,H	H,T,H	T,T,T	…

A **Random Variable** X is a mapping between events in Ω to numbers.

Discrete random variables take on one of a finite set of values

- Can reason about the probability of specific values

Continuous random variables take on one of an infinite set of values (e.g. height)

- Probability of any specific value is pretty much zero
- Easier to talk about probability of being in a range

Probability densities describe a mapping from values of X to probabilities

Discrete random variable **X** has
probability mass function (pmf) p_X

$$p_X(x) = P(\{w \mid X(w) = x\})$$

$$0 \leq p_X(x) \leq 1 \; \forall x \in X$$

$$\sum_{x \in Val(X)} p_X(x) = 1$$

Brief Probability Review

Consider drawing a card from a standard 52 card deck:

$$\text{Sample Space } \boldsymbol{\Omega} = \{\boldsymbol{A\spadesuit}, \boldsymbol{A\clubsuit}, \boldsymbol{A\heartsuit}, \boldsymbol{A\diamondsuit}, \boldsymbol{2\spadesuit}, \boldsymbol{2\clubsuit}, \boldsymbol{2\heartsuit}, \boldsymbol{2\diamondsuit}, \dots, \boldsymbol{K\spadesuit}, \boldsymbol{K\clubsuit}, \boldsymbol{K\heartsuit}, \boldsymbol{K\diamondsuit}\}$$

Let's define a discrete random variable **X** such that:

$$X(\boldsymbol{A\spadesuit}) = \boldsymbol{X}(\boldsymbol{A\clubsuit}) = \boldsymbol{X}(\boldsymbol{A\heartsuit}) = \boldsymbol{X}(\boldsymbol{A\diamondsuit}) = \mathbf{1}$$
$$X(\boldsymbol{2\spadesuit}) = \boldsymbol{X}(\boldsymbol{2\clubsuit}) = \boldsymbol{X}(\boldsymbol{2\heartsuit}) = \boldsymbol{X}(\boldsymbol{2\diamondsuit}) = \mathbf{2}$$
$$\dots$$
$$X(\boldsymbol{K\spadesuit}) = \boldsymbol{X}(\boldsymbol{K\clubsuit}) = \boldsymbol{X}(\boldsymbol{K\heartsuit}) = \boldsymbol{X}(\boldsymbol{K\diamondsuit}) = \mathbf{13}$$

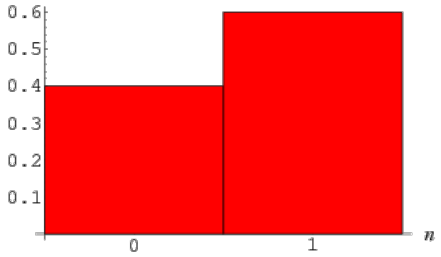
Check the probability of a specific outcome:

$$P_X(x = 13) = P(\{\, w \mid X(w) = 13 \,\}) = \frac{4}{52}$$

Common discrete distributions:

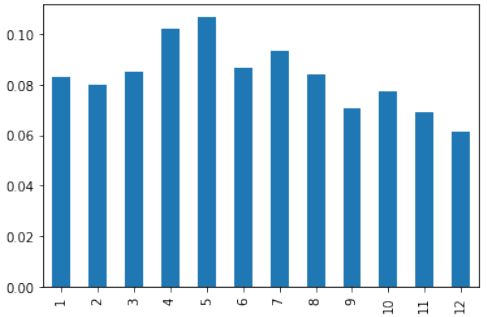
Binary random variable $X \in \{0,1\}$ is distributed according to a $\boldsymbol{Bernoulli}(\theta)$ if:

$$p_X(x) = \theta^x(1-\theta)^{(1-x)}$$



Discrete random variable $X \in \{0,1,\ldots,c\}$ is distributed according to a $\boldsymbol{Categorical}(\theta_1,\ldots,\theta_c)$ if:

$$p_X(x) = \theta_x$$



Probability densities describe a mapping from values of X to probabilities

Discrete random variable **X** has **probability mass function (pmf)** p_X

$$p_X(x) = P(\{w \mid X(w) = x\})$$

$$0 \leq p_X(x) \leq 1 \; \forall x \in X$$

$$\sum_{x \in Val(X)} p_X(x) = 1$$

Continuous random variable **X** has **probability density function (pdf)** $\mathbf{f_X}$

$$\int_{x \in A} f_x(x)\, dx = P(\{\, w \mid X(w) \in A \,\})$$

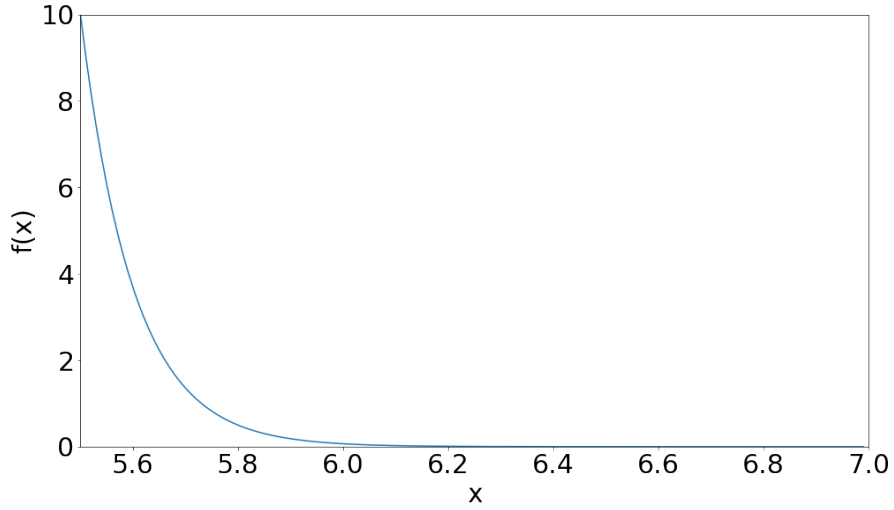
Note that f can be greater than 1

$$f_X(x) \geq 0 \; \forall x \in X$$

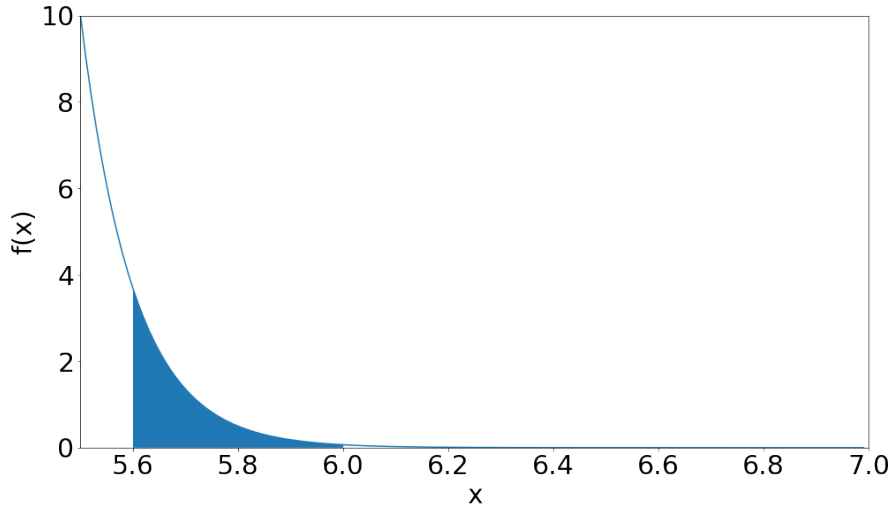
$$\int_{x \in Val(X)} f_x(x)\, dx = 1$$

Sample Space $\boldsymbol{\Omega} = (0, \infty)$ be the width of pipes from a factory in cm

$$f_X(x) = \begin{cases} 0 & x < 5.5 \\ 10e^{-10(x-5.5)} & x \geq 5.5 \end{cases}$$



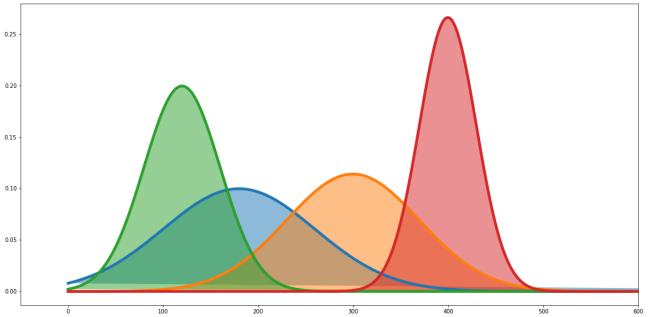
$$P(\{w|5.6 < X(w) \leq 6)$$
$$= \int_{5.6}^{6} f_X(x)dx = 0.361$$



Common continuous distributions:

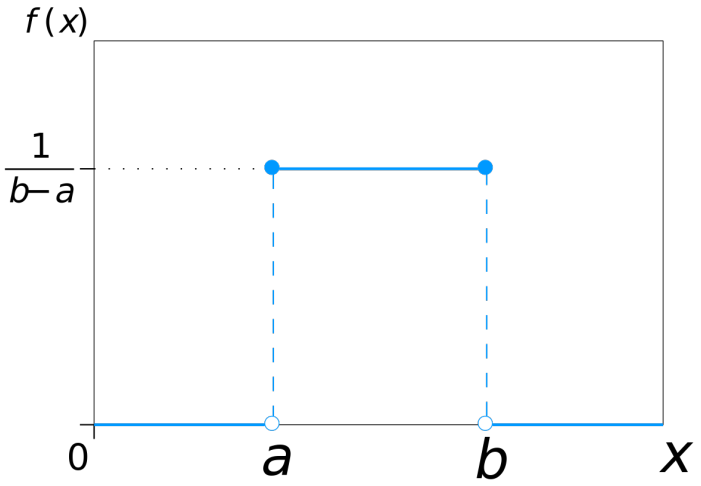
Continuous random variable $X \in (-\infty, \infty)$ is distributed according to a $\boldsymbol{Gausian}(\mu, \sigma)$ if:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Discrete random variable $X \in [a, b]$ is distributed according to a $\boldsymbol{Uniform}(a, b)$, b>a if:

$$p_X(x) = \frac{1}{b-a}$$



Expectation (Discrete Random Variable X):

Value of g(·) at x

$$E_X[g(x)] = \sum_{x \in Val(X)} p_X(x) g(x)$$

Probability of X=x

Can think of it as an average of $g(x)$ when $x's$ are sample from p_X

or equivalently

as a weighted average of $g(x)$ with weight $p_X(x)$

Expectation (Continuous Random Variable X):

Value of $g(\cdot)$ at x

$$E_X[g(x)] = \int_{x \in Val(X)} f_x(x) g(x)\, dx$$

Probability density of X=x

Can think of it as an average of $g(x)$ when $x's$ are sample from f_X

or equivalently

as a weighted average of $g(x)$ with weight $f_X(x)$

Brief Probability Review

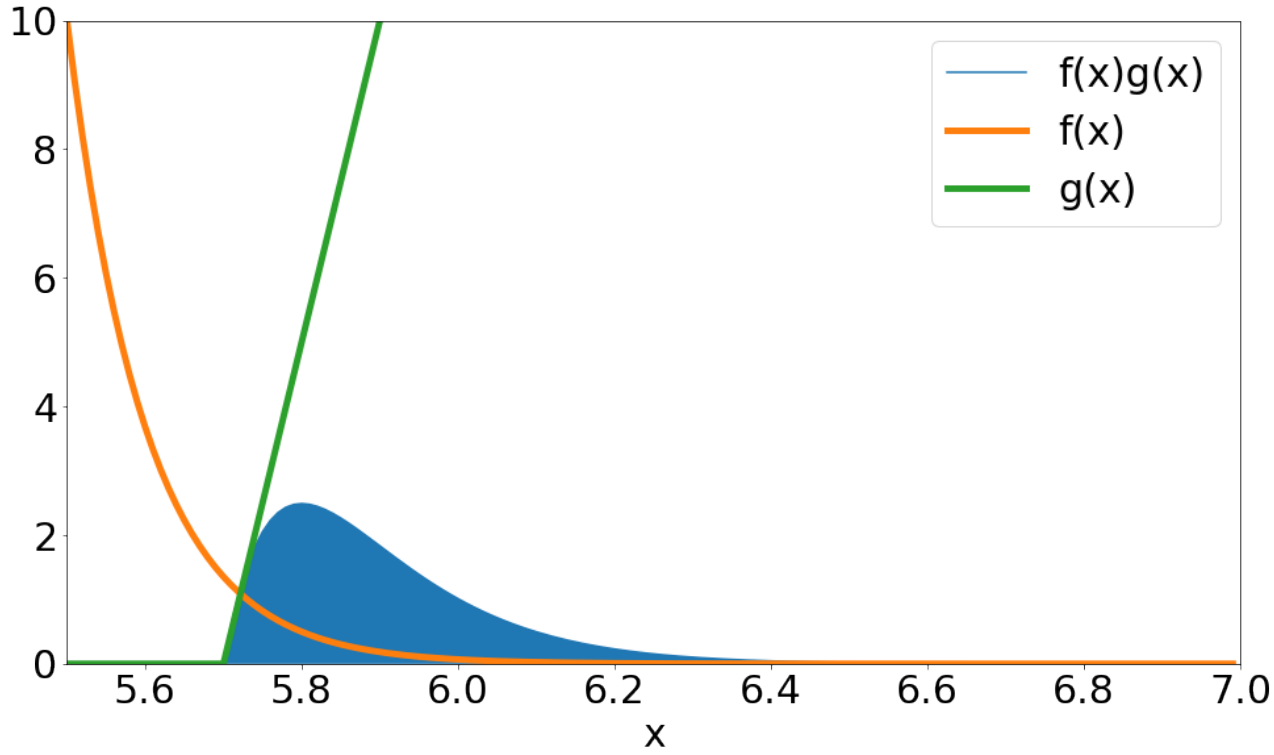
Sample Space $\boldsymbol{\Omega} = (0, \infty)$ be the width of pipes from a factory in cm.
Suppose you have to pay a penalty of \$50 for every mm over 5.7.

What is the average penalty for a pipe produced from your factory?

$$f_X(x) = \begin{cases} 0 & x < 5.5 \\ 10e^{-10(x-5.5)} & x \geq 5.5 \end{cases}$$

$$g(x) = \begin{cases} 0 & x \leq 5.7 \\ 50(x - 5.7) & x > 5.7 \end{cases}$$

$$E_X[g(x)] = \int_{-\infty}^{\infty} f_X(x)g(x)dx \approx \$0.67$$



Properties of Expectation

$$E_X[aX + bY] = aE_x[X] + bE_X[Y]$$

$$E_{X,Y}[XY] = E_x[X]E_Y[Y] \quad if \; X \perp\!\!\!\perp Y$$

X and Y are independent

A couple of special expectations:

$$E_X[x] = mean\ of\ X$$

$$\begin{aligned} Var[x] &= E_X[(x - E_X[x])^2] \\ &= E_X[x^2 - 2XE_X[x] + E_X[x]^2] \\ &= E_X[x^2] - 2E_X[x]E_X[X] + E_X[x]^2 \\ &= E_X[x^2] - E_X[x]^2 \end{aligned}$$

Let X be a Gaussian random variable with mean 0 and variance 1:

$Var[x] = E[x^2] - E[x]^2$ but mean is zero so $Var[x] = E[x^2]$ and the variance is given as 1

$$E_X[ax^2 + bx + c] = aE_X[x^2] + bE_X[x] + c$$

$E_X[x]$ is the mean = 0

A Philosophic Debate: Interpreting Probabilities

What does it mean when we say P(A)?

Frequentist View:

- Normalized limiting frequency of a repeating, non-deterministic event
- If I keep sampling worlds, then P(A) is $\lim_{N \to \infty} \frac{\#(\boldsymbol{A \text{ is } true})}{N}$

Makes sense for things I can repeat, like coin flips. But what about things like P("earth explodes on January 1, 2000") that only happen once?

Bayesian View:

- P(A) is your "belief" about A

Slide adapted from Dhruv Batra.

View our data as random variables:

- X (input) and Y (output) are random variables and the dataset is drawn from their joint distribution: $x_i, y_i \sim P(X, Y)$

- Datasets are formed by repeatedly sampling from $P(X, Y)$

IID: Independently and Identically Distributed

- Assume training and testing are drawn IID from $P(X, Y)$
  - Implies *some* hope of performing well on test

  - Imagine if that weren't the case? The test set would be systematically different than the training set… bad news.

Today's Learning Objectives

Be able to answer:

- ~~What are basic concepts and definitions in probability theory?~~

- What is maximum likelihood estimation (MLE)?
  - What is a likelihood function?
  - What is the MLE estimate for a Bernoulli variable?
  - What is bias and variance of an estimator?
  - What is the MLE estimate for a Normal variable?

- What is a maximum a posteriori estimate (MAP)?
  - What is a prior and posterior? What is a

Your First Probabilistic Learning Algorithm: MLE

We will often want to "learn" parameters for our models that make our models "agree" with some observed data as much as possible.

Probability gives us a nice tool for reasoning about this – namely, that we would want to find parameters such that the observed data has high probability under our models.

We will start with very simple models today but will be reusing these techniques a few times throughout the course.

Your First Probabilistic Learning Algorithm: MLE

After taking this course, you drop out of OSU and join an illegal betting company focused on **underground mascot fights** because you want to start making some **real money with ML**.

Your new boss asks you:

- If Benny Beaver and the Oregon Duck face off, what is the probability Benny wins?

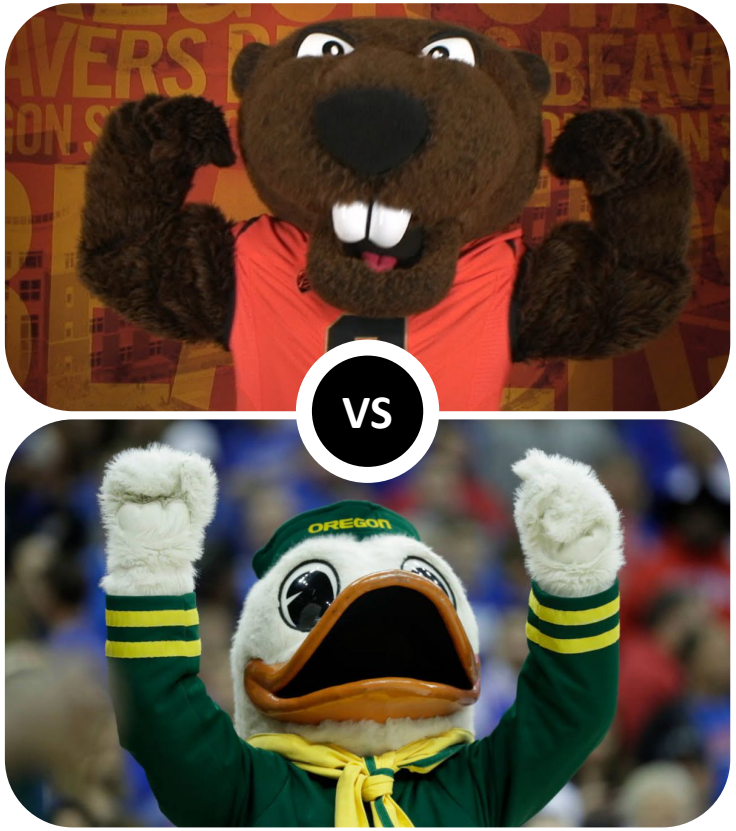
Knowing that data is everything, you ask:

- What happened in the past?

W, W, L, L, W

You say P(Benny Wins) = ?

Why?



This slide in no way condones or encourages underground mascot fights or illegal gambling. Even if we would win

Your First Probabilistic Learning Algorithm: MLE

The intuitive answer is that 3/5 "fits the data" the best. We'll formalize that notion this lecture.

Maximum Likelihood Estimation – Find parameters that make the observed data most likely.

1. Assume a probabilistic model of how the data was generated $x \sim P(x; \theta)$ parameterized by some set of parameters θ

2. Find $\hat{\theta}_{MLE}$ that maximizes the probability (or likelihood) of generating the training data under the probabilistic model.

Why MLE?

- Often leads to "natural" or intuitive parameter estimates
- MLE is optimal if model class is correct (e.g. Normal model for normally distributed data)

Our "generative story" for mascot fights

Assume: Benny has some constant probability of winning any given fight – call this theta.

$P(win) = \theta$. Let's say $\theta = 0.7$. How does this number relate to the outcomes we observe?

$rand(0,1) \leq 0.7$	$rand(0,1) \leq 0.7$	$rand(0,1) \leq 0.7$	$rand(0,1) \leq 0.7$
$0.8 \leq 0.7 \Rightarrow False$	$0.4 \leq 0.7 \Rightarrow True$	$0.9 \leq 0.7 \Rightarrow False$	$0.29 \leq 0.7 \Rightarrow True$
			
$x_1 = 0$	$x_2 = 1$	$x_3 = 0$	$x_4 = 1$

This slide in no way condones or encourages underground mascot fights or illegal gambling. Even if we would win

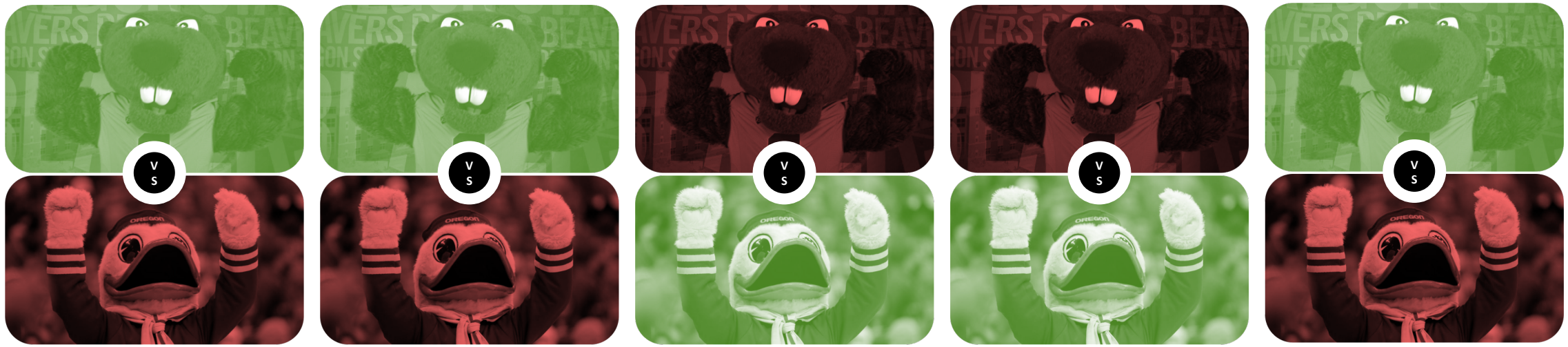
Maximum Likelihood Estimation (MLE)

MLE answers the question: What value of theta makes the data we observed MOST likely?

$$P(win) = \theta$$

$P(W, W, L, L, W \mid \theta = 0.1)$ would be small

$P(W, W, L, L, W \mid \theta = 0.99)$ would be very small



This slide in no way condones or encourages underground mascot fights or illegal gambling. Even if we would win.

Baby Steps: Estimating the Weight of a Coin with MLE

Dataset: Suppose you observe a coin flipped **6** times such that $\boldsymbol{D} = \{\boldsymbol{H}, \boldsymbol{T}, \boldsymbol{H}, \boldsymbol{H}, \boldsymbol{H}, \boldsymbol{T}\}$

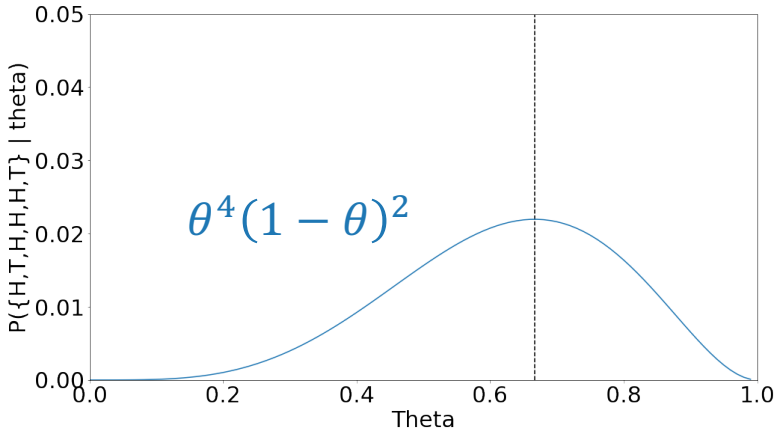
Model Assumption: Assume the probability of getting heads on the coin is some unknown θ.

Write out likelihood of the training data as a function of parameters θ:

$$\mathcal{L}(\theta) = P(D \mid \theta) = P(\{H,T,H,H,H,T\}|\theta) = P(H|\theta)P(T|\theta)P(H|\theta)P(H|\theta)P(H|\theta)P(T|\theta)$$

$$= \theta(1-\theta)\theta\theta\theta(1-\theta) \quad = \theta^4(1-\theta)^2$$

Find $\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}}\ \mathcal{L}(\theta)$**:**



Any guesses what the peak of the graph is?

$$\frac{4}{6} = \frac{\#Heads}{\#Heads + \#Tails}$$

Is this generally true?

Question Break!

What assumption did we make when we set $P(\{H,T,H,H,H,T\}|\theta) = P(H|\theta)P(T|\theta)P(H|\theta)P(H|\theta)P(H|\theta)P(T|\theta)$?

A θ was greater than 0.

B θ was greater than 0.5

C Each coin flip was independent

D Each coin flip was dependent

General Case: Estimating the Weight of a Coin with MLE

Dataset: Let X_i be a binary random variable (H→1, T→0). Given a dataset $D = \{x_i\}_{i=1}^{N}$ representing a coin being flipped N times. (e.g. {H,T,H,H,H,T} → {1,0,1,1,1,0})

Model Assumption: Assume that X is distributed according to a Bernoulli distribution with parameter θ such that: $P(x) = \theta^x(1-\theta)^{1-x}$

General Case: Estimating the Weight of a Coin with MLE

Dataset: Let X_i be a binary random variable (H→1, T→0). Given a dataset $D = \{x_i\}_{i=1}^{N}$ representing a coin being flipped N times. (e.g. {H,T,H,H,H,T} → {1,0,1,1,1,0})

Model Assumption: Assume that X is distributed according to a Bernoulli distribution with parameter θ such that: $P(x) = \theta^x(1-\theta)^{1-x}$

Write out likelihood of the training data as a function of parameters θ:

$$\mathcal{L}(\theta) = P(D \mid \theta) = \prod_{i=1}^{N} P(x_i \mid \theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

Collect all the like terms and sum their exponents

$$= \theta^{\Sigma x_i}(1-\theta)^{\Sigma(1-x_i)}$$

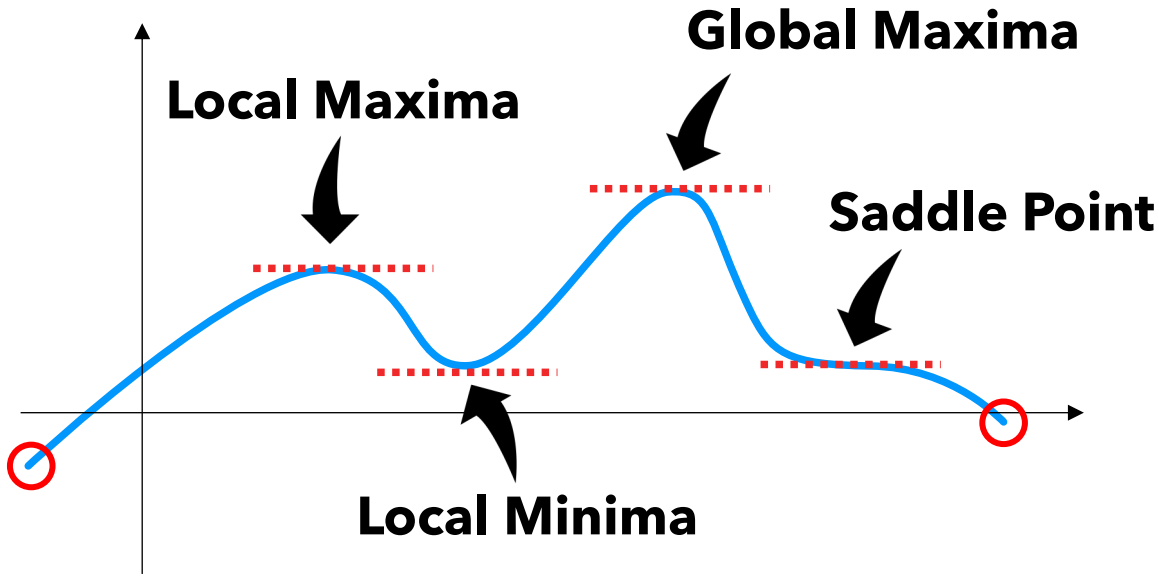
Now our goal is to find θ that maximizes this

$$= \theta^{\#Heads}(1-\theta)^{\#Tails}$$

Find $\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}}\ \mathcal{L}(\theta)$**:** How can we do this analytically?

Finding Maxima of Functions

Recall from Calculus: Maxima, minima, and saddle-points in a function are "critical points" where the derivative is zero. (Also have to consider end points if asked to find a maxima within a finite range)



Can tell minima/maxima/saddles apart with 2nd derivative test to check for local concavity.

General Case: Estimating the Weight of a Coin with MLE

If $\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}} \quad \mathcal{L}(\theta) = \theta^{\#H}(1-\theta)^{\#T}$, then $\mathcal{L}'(\hat{\theta}_{MLE}) = 0$

Taking $\frac{d}{d\theta}\ \theta^{\#H}(1-\theta)^{\#T}$ gets messy. Lots of polynomial terms due to the exponents.

Log-likelihood is often used instead – pesky exponents get turned into scalar multipliers and products turn to sums. **Recall that** $\log(x^a) = a\log(x)$, $\log(ab) = \log(a) + \log(b)$, **and** $\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$

$$\mathcal{LL}(\theta) = \#H * log\theta + \#T * \log(1-\theta)$$

Does $\underset{\theta}{argmax}\ \mathcal{LL}(\theta) = \underset{\theta}{argmax}\ \mathcal{L}(\theta)$**?** Yes! Log is monotonically increasing so argmax are preserved.

General Case: Estimating the Weight of a Coin with MLE

If $\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}}\ \mathcal{LL}(\boldsymbol{\theta}) = \#H * log\theta + \#T * \log(1-\theta)$, then $\mathcal{LL}'(\hat{\theta}_{MLE}) = 0$

$$\frac{d}{d\theta} \#H * log\theta + \#T * \log(1-\theta) = \frac{\#H}{\theta} - \frac{\#T}{1-\theta}$$

$$\frac{\#H}{\theta} - \frac{\#T}{1-\theta} = 0$$

$$(1-\theta)\#H - \theta\#T = 0$$

$$\#H - \theta(\#H + \#T) = 0$$

$$\hat{\theta}_{MLE} = \frac{\#H}{\#H + \#T}$$

Question Break!

Checking in. How'd we do?

A I understood what just happened

B I think I understood what just happened

C I might not understand what just happened

D I do not understand what just happened

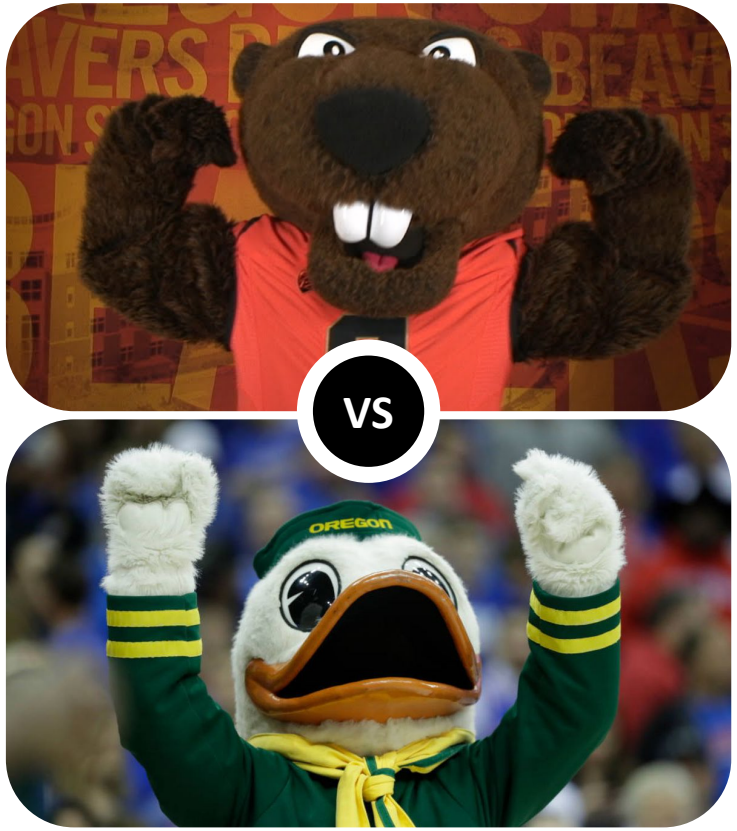
Your First Probabilistic Learning Algorithm

What happened in the past? **W, W, L, L, W**

You say "P(Benny Wins) = 3/5 and I can prove it!

Your boss says: What if I told you they won 300 out of 500?

You say "Still 3/5 and I can prove it!"



Your boss says: Which is better? 3/5 or 300/500?

You say: Uh…. Same estimate? More the merrier?

Your boss says: Why am I paying you so much?

Two concepts: **Bias** of your estimate – whether your estimate is right on average.
Variance of your estimate– how much variability your estimator will have.

This slide in no way condones or encourages underground mascot fights or illegal gambling. Even if we would win.

Bias and Variance of our MLE Estimator for Bernoulli Parameter

Dataset: Let X_i be a binary random variable (H→1, T→0).
Given a dataset $D = \{x_i\}_{i=1}^{N}$

$$\hat{\theta}_{MLE} = \frac{\#H}{\#H + \#T} = \frac{1}{n}\sum_i x_i$$

Bias: Over all possible datasets, what would I expect my error in estimating the true parameters to be on average.

$$E_D\left[\hat{\theta}_{MLE} - \theta\right] = E_D\left[\hat{\theta}_{MLE}\right] - E_D[\theta]$$

$$= E_D\left[\frac{1}{n}\sum_i x_i\right] - \theta$$

Linearity of expectation →

$$= \frac{1}{n}\left[\sum_i E_{X_i}[x_i]\right] - \theta$$

Call this an unbiased estimator as bias is 0 → no systematic error →

$$= \frac{1}{n}\left[\sum_i \theta\right] - \theta = 0$$

Bias and Variance of our MLE Estimator for Bernoulli Parameter

Dataset: Let X_i be a binary random variable (H→1, T→0).
Given a dataset $D = \{x_i\}_{i=1}^{N}$

$$\hat{\theta}_{MLE} = \frac{\#H}{\#H + \#T} = \frac{1}{n}\sum_i x_i$$

Variance: Over all possible datasets, how do I expect my estimate to vary from its mean

$$Var_D\left[\hat{\theta}_{MLE}\right] = Var_D\left[\frac{1}{n}\sum_i x_i\right]$$

$Var[aX + bY] = a^2Var[X] + b^2Var[Y] + 2abCov[X, Y]$
Covariance is zero here as each x_i is independent.

$$= \frac{1}{n^2}Var_D\left[\sum_i x_i\right] = \frac{1}{n^2}\sum_i Var_{X_i}[x_i]$$

$Var[x] = \theta(1-\theta)$ for Bernoulli X
(Not shown here, you can do it)

$$= \frac{1}{n^2}\sum_i \theta(1-\theta)$$

Our estimator has lower variance as n gets larger!

$$= \frac{\theta(1-\theta)}{n}$$

Question Break!

For a fixed value of n, the variance still changes as a function of the theta. When does the numerator get large?

Be prepared to explain why your answer makes intuitive sense in terms of observing coin flips. E.g., why would there be higher or lower variance given the same number of observations.

$$Var[\hat{\theta}_{MLE}] = \frac{\theta(1-\theta)}{n}$$

A When θ is close to 1

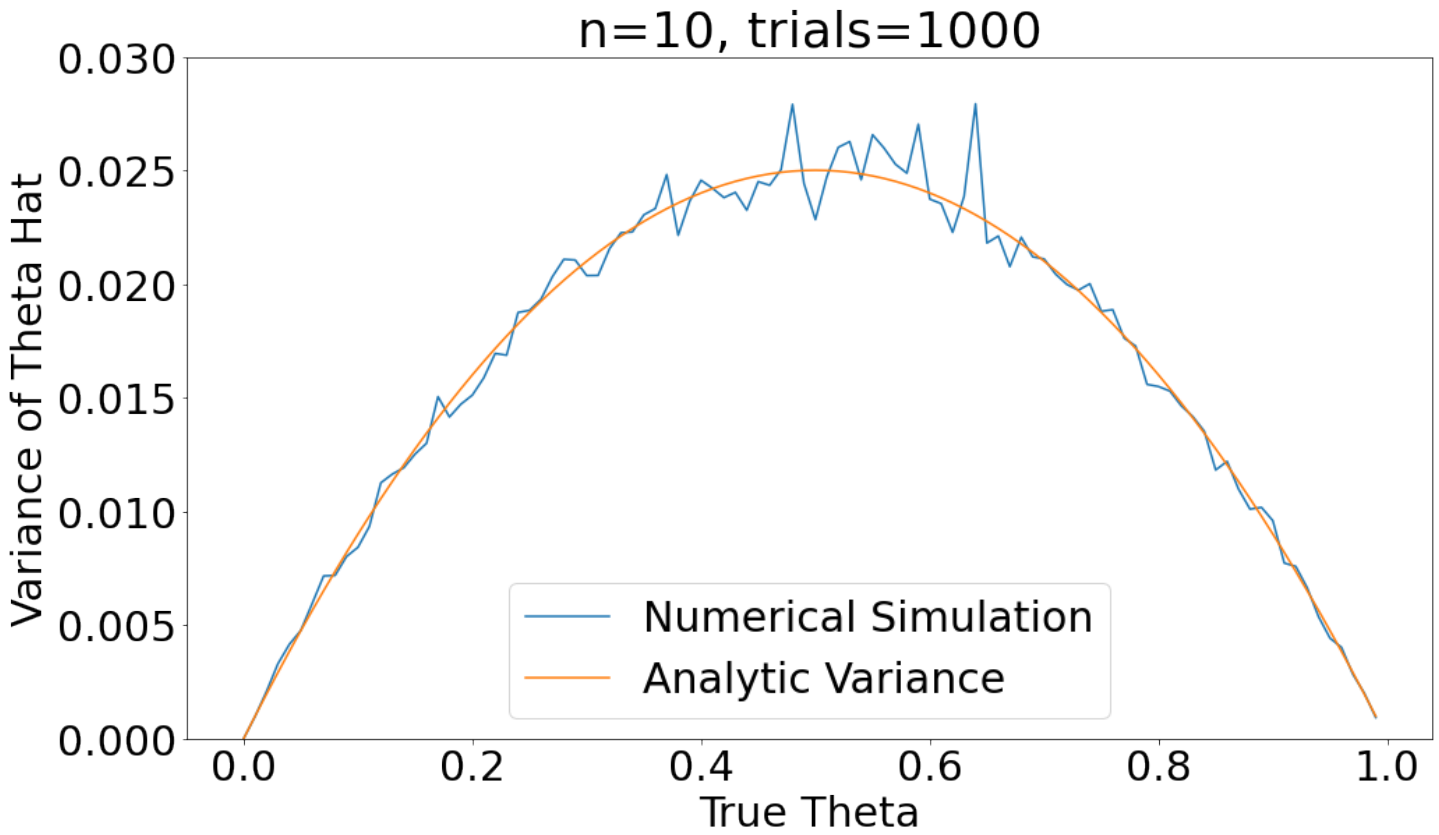
B When θ is close to 0

C When θ is close to 0.5

D When n gets small

Numerical Simulation vs Analytic Variance

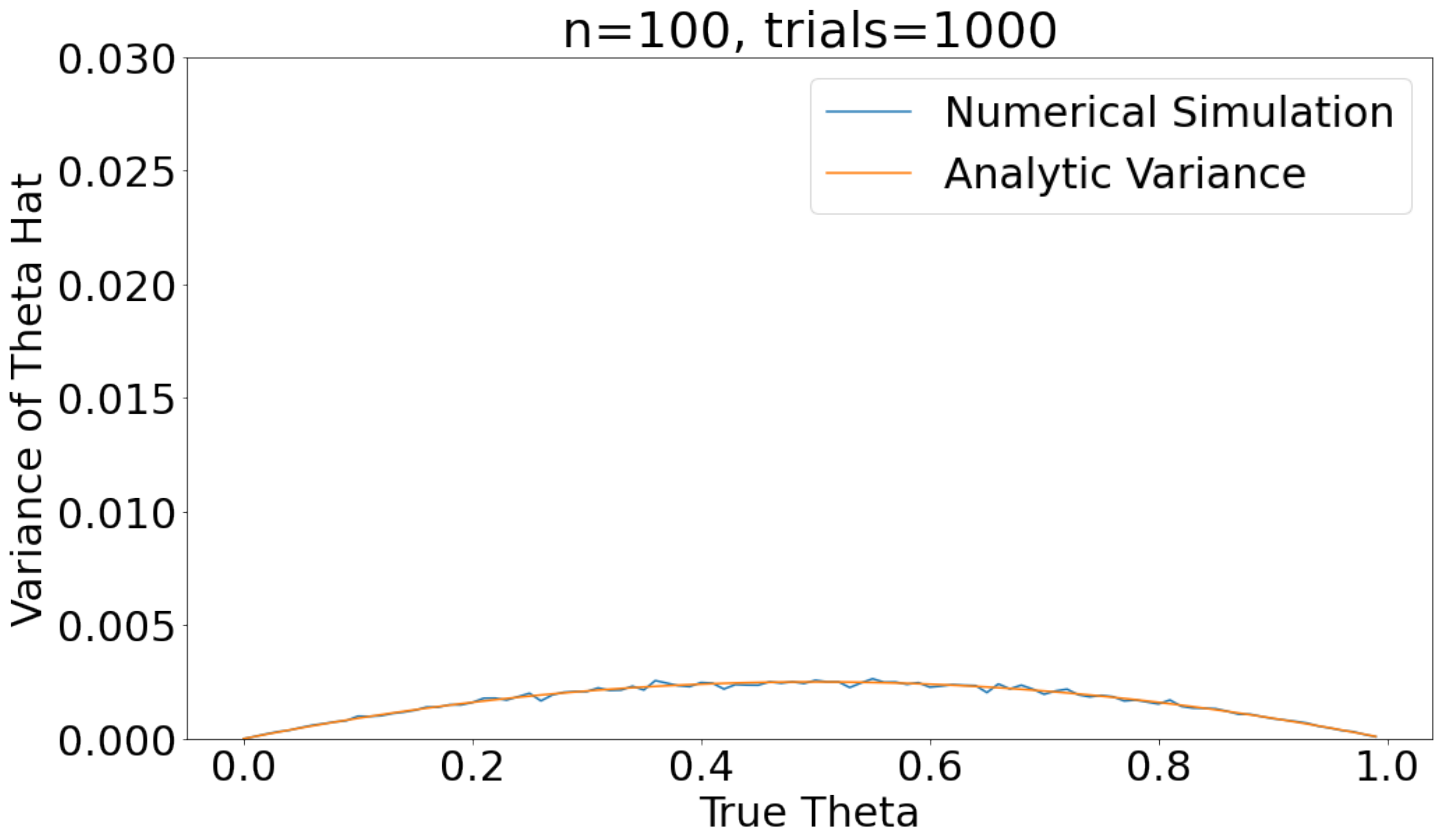
```python
import numpy as np
import matplotlib
font = {'family' : 'normal',
        'weight' : 'normal',
        'size'   : 30}

matplotlib.rc('font', **font)

import matplotlib.pyplot as plt


for n in [10, 100, 1000]:

  theta = np.arange(0,1,0.01)

  trials = 1000

  error = []
  tht = []
  for t in theta:
    sumer = 0
    for _ in range(trials):
      est = np.mean(np.random.rand(n)<=t)
      sumer += (t-est)**2
    error.append(sumer/trials)
    tht.append(t)

  plt.figure(figsize=(16,9))
  plt.plot(tht,error, label="Numerical Simulation")
  plt.plot(theta, theta*(1-theta)/n, label="Analytic Variance")
  plt.xlabel("True Theta")
  plt.ylabel("Variance of Theta Hat")
  plt.title("n={}, trials={}".format(n, trials))
  plt.ylim(0, 0.030)
  plt.legend()
```



$$Var_D[\hat{\theta}_{MLE}] = \frac{\theta(1-\theta)}{n}$$

Numerical Simulation vs Analytic Variance

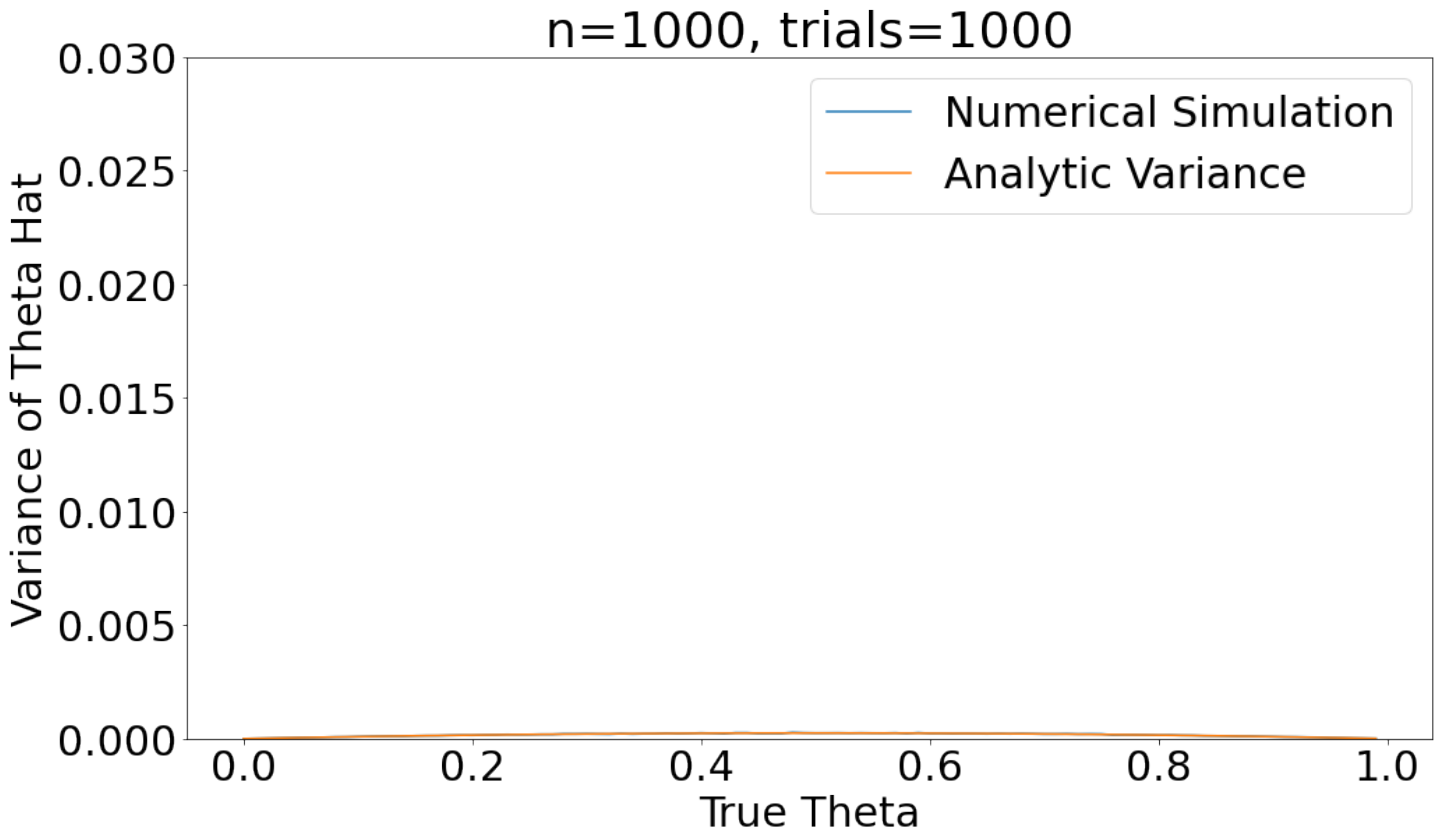
```python
import numpy as np
import matplotlib
font = {'family' : 'normal',
        'weight' : 'normal',
        'size'   : 30}

matplotlib.rc('font', **font)

import matplotlib.pyplot as plt


for n in [10, 100, 1000]:

  theta = np.arange(0,1,0.01)

  trials = 1000

  error = []
  tht = []
  for t in theta:
    sumer = 0
    for _ in range(trials):
      est = np.mean(np.random.rand(n)<=t)
      sumer += (t-est)**2
    error.append(sumer/trials)
    tht.append(t)

  plt.figure(figsize=(16,9))
  plt.plot(tht,error, label="Numerical Simulation")
  plt.plot(theta, theta*(1-theta)/n, label="Analytic Variance")
  plt.xlabel("True Theta")
  plt.ylabel("Variance of Theta Hat")
  plt.title("n={}, trials={}".format(n, trials))
  plt.ylim(0, 0.030)
  plt.legend()
```



$$Var_D\left[\hat{\theta}_{MLE}\right] = \frac{\theta(1-\theta)}{n}$$

Numerical Simulation vs Analytic Variance

```python
import numpy as np
import matplotlib
font = {'family' : 'normal',
        'weight' : 'normal',
        'size'   : 30}

matplotlib.rc('font', **font)

import matplotlib.pyplot as plt


for n in [10, 100, 1000]:

  theta = np.arange(0,1,0.01)

  trials = 1000

  error = []
  tht = []
  for t in theta:
    sumer = 0
    for _ in range(trials):
      est = np.mean(np.random.rand(n)<=t)
      sumer += (t-est)**2
    error.append(sumer/trials)
    tht.append(t)

  plt.figure(figsize=(16,9))
  plt.plot(tht,error, label="Numerical Simulation")
  plt.plot(theta, theta*(1-theta)/n, label="Analytic Variance")
  plt.xlabel("True Theta")
  plt.ylabel("Variance of Theta Hat")
  plt.title("n={}, trials={}".format(n, trials))
  plt.ylim(0, 0.030)
  plt.legend()
```



$$Var_D\left[\hat{\theta}_{MLE}\right] = \frac{\theta(1-\theta)}{n}$$

Your First Probabilistic Learning Algorithm

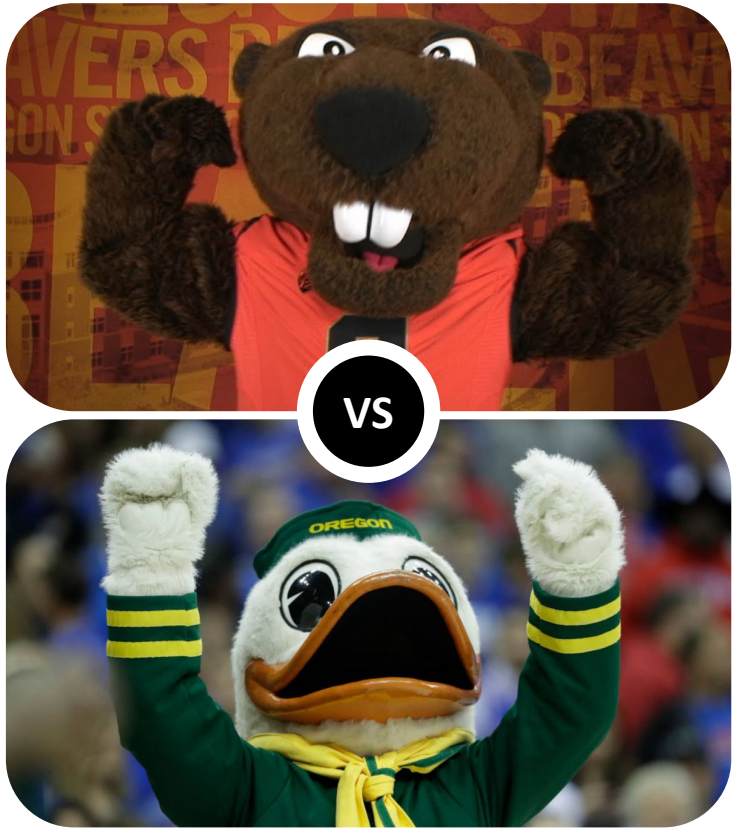
What happened in the past? **W, W, L, L, W**

Your boss says: Which is better? 3/5 or 300/500?

~~**You say:** Uh…. Same estimate? More the merrier?~~

You say: Same estimate but we can be much more sure it isn't just random variation. And of course I can prove it!

Your boss says: This is why I pay you the big bucks.



This slide in no way condones or encourages underground mascot fights or illegal gambling. Even if we would win.

Your First Probabilistic Learning Algorithm: MLE

What is MLE for?

- Fitting the parameters of probabilistic models. You assume some "generative story" of how your data was produced and then find parameters that maximize the likelihood of observed data.

Maximum Likelihood Estimation – Find parameters that make the observed data most likely.

1. Assume a probabilistic model of how the data was generated $x \sim P(x; \theta)$ parameterized by some set of parameters θ

2. Find $\hat{\theta}_{MLE}$ that maximizes the probability (or likelihood) of generating the training data under the probabilistic model.

So far, we've seen MLE for Bernoulli random variables… let's try Normal ones

General Case: Estimating the Mean and Variance of Normal Distribution

Dataset: Let X_i be a continuous random variable. Given a dataset $D = \{x_i\}_{i=1}^{N}$ representing some repeated independent measurement.

Model Assumption: Assume that X is distributed according to a Normal distribution with parameter $\theta = (\mu, \sigma)$ distributed according to a $\mathcal{N}(\mu, \sigma)$.

$$\underset{Normal(\mu,\sigma)}{f_X(x)} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Write out log-likelihood of parameters θ and then find its maximum:

$$P(D \mid \theta) = \prod_{i=1}^{N} P(x_i \mid \theta) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$logP(D \mid \theta) = Nlog(1) - \mathrm{N}\log\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2\sigma^2}\sum_{i}(x_i - \mu)^2$$

General Case: Estimating the Mean and Variance of Normal Distribution

Write out log-likelihood of parameters θ and then find its maximum:

$$logP(D \mid \theta) = Nlog(1) - \mathrm{N}\log\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2\sigma^2}\sum_{i}^{N}(x_i - \mu)^2$$

Take derivative of parameters and find where they equal 0:

$$\frac{d\ logP(D \mid \theta)}{d\mu} = \frac{1}{\sigma^2}\sum_{i}^{N}(x_i - \mu) = \frac{1}{\sigma^2}\left(N\mu - \sum_{i=1}^{N} x_i\right) \rightarrow \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Just the sample mean

$$\frac{d\ logP(D \mid \theta)}{d\sigma} = \frac{N}{\sigma} - \frac{1}{\sigma^3}\sum_{i}^{N}(x_i - \mu)^2 = 0$$

$$= N\sigma^2 - \sum_{i}^{N}(x_i - \mu)^2 \rightarrow \sigma = \sqrt{\frac{1}{N}\sum_{i}^{N}(x_i - \mu)^2}$$

Just the sample standard deviation

Today's Learning Objectives

Be able to answer:

- ~~What are basic concepts and definitions in probability theory?~~
- ~~What is maximum likelihood estimation (MLE)?~~
  - ~~What is a likelihood function?~~
  - ~~What is the MLE estimate for a Bernoulli variable?~~
  - ~~What is bias and variance of an estimator?~~
  - ~~What is the MLE estimate for a Normal variable?~~
- What is a maximum a posteriori estimate (MAP)?
  - What is a prior and posterior? What is a

Your First Probabilistic Learning Algorithm

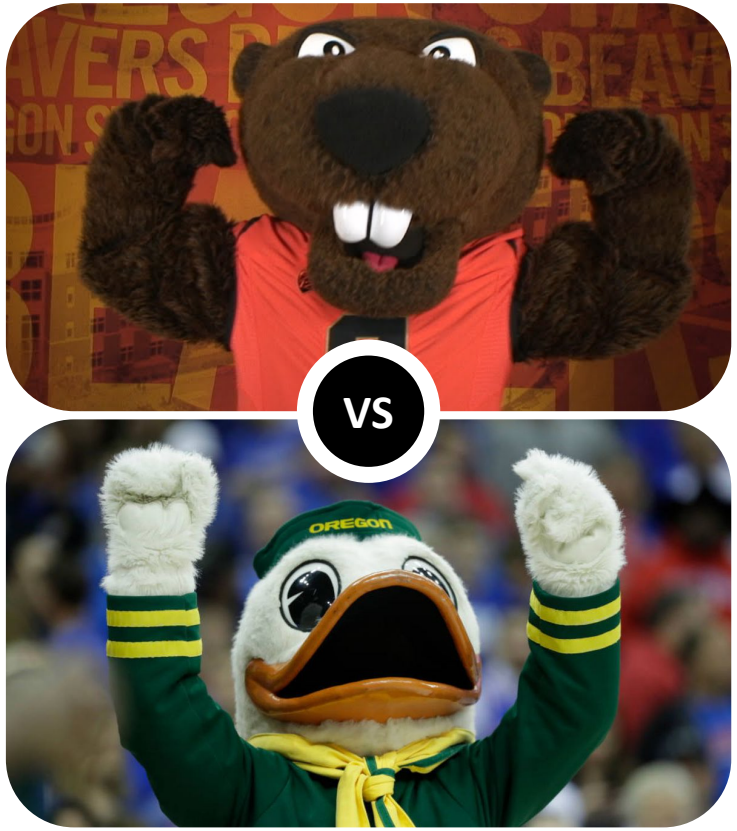
What happened in the past? **W, W, L, L, W**

Your boss says: Which is better? 3/5 or 300/500?

~~**You say:** Uh…. Same estimate? More the merrier?~~

You say: Same estimate but we can be much more sure it isn't just random variation. And of course I can prove it!

Your boss says: This is why I pay you the big bucks.



Your boss says: What if I'm pretty sure the duck is using performance enhancing drugs this time?

You say: Ah… Bayesian it is then…

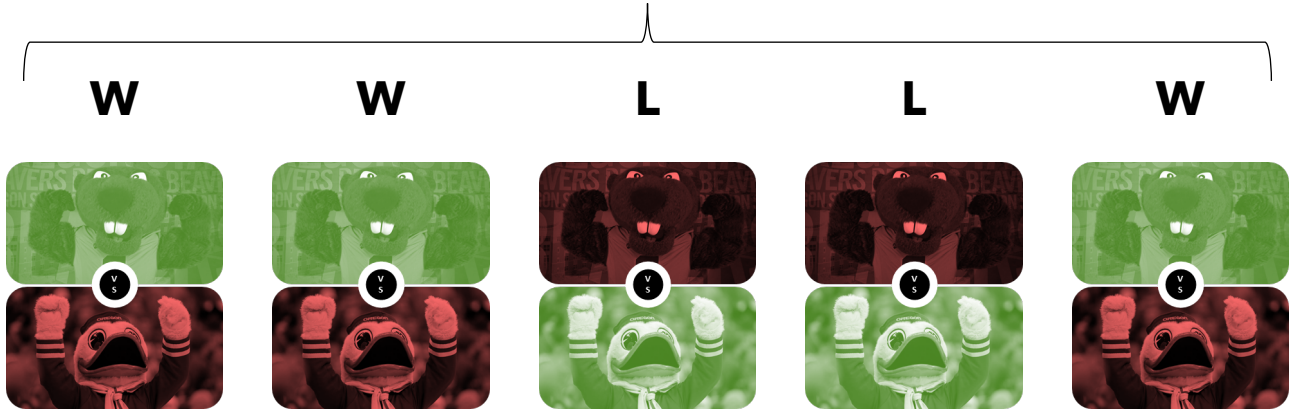
This slide in no way condones or encourages underground mascot fights or illegal gambling. Even if we would win.

An Informal Way to Add Prior Belief

Assume: We believe the duck may be using performance enhancing drugs for next fight. How can we reason about this belief alongside our observations of history?

Real data we've observed yields

$$\hat{\theta}_{MLE} = \frac{\#wins}{total} = \frac{3}{5}$$



This slide in no way condones or encourages underground mascot fights or illegal gambling. Even if we would win

An Informal Way to Add Prior Belief

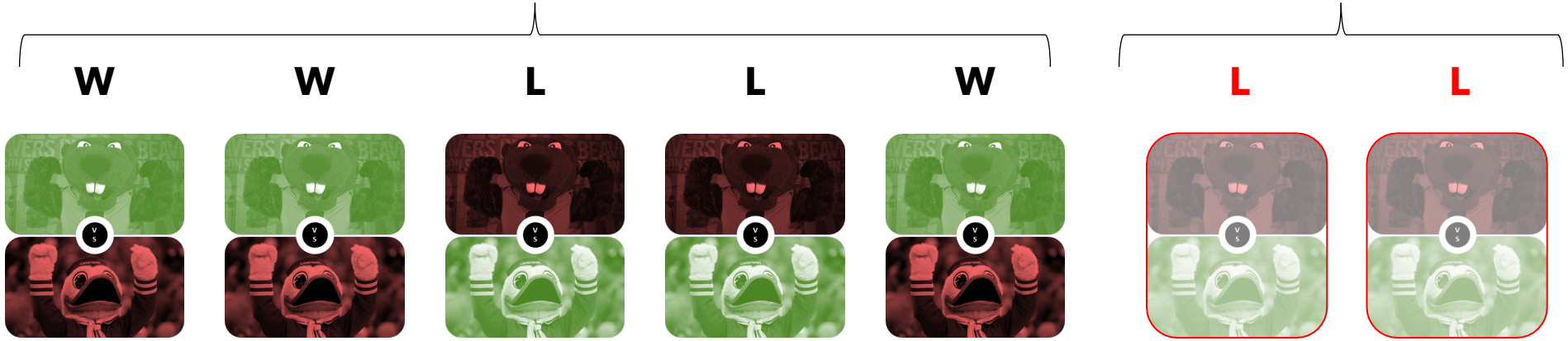
Assume: We believe the duck may be using performance enhancing drugs for next fight. How can we reason about this belief alongside our observations of history?

Real data we've observed yields

$$\hat{\theta}_{MLE} = \frac{\#wins}{total} = \frac{3}{5}$$

Add fake observation to express our prior belief and find MAP:

$$\hat{\theta}_{MAP} = \frac{\#wins + \#fakewins}{total} = \frac{3}{7}$$



This slide in no way condones or encourages underground mascot fights or illegal gambling. Even if we would win

An Informal Way to Add Prior Belief

I can express different priors and different confidences in those priors by changing the outcome/number of fake observations

Real data we've observed yields

$$\hat{\theta}_{MLE} = \frac{\#wins}{total} = \frac{3}{5}$$

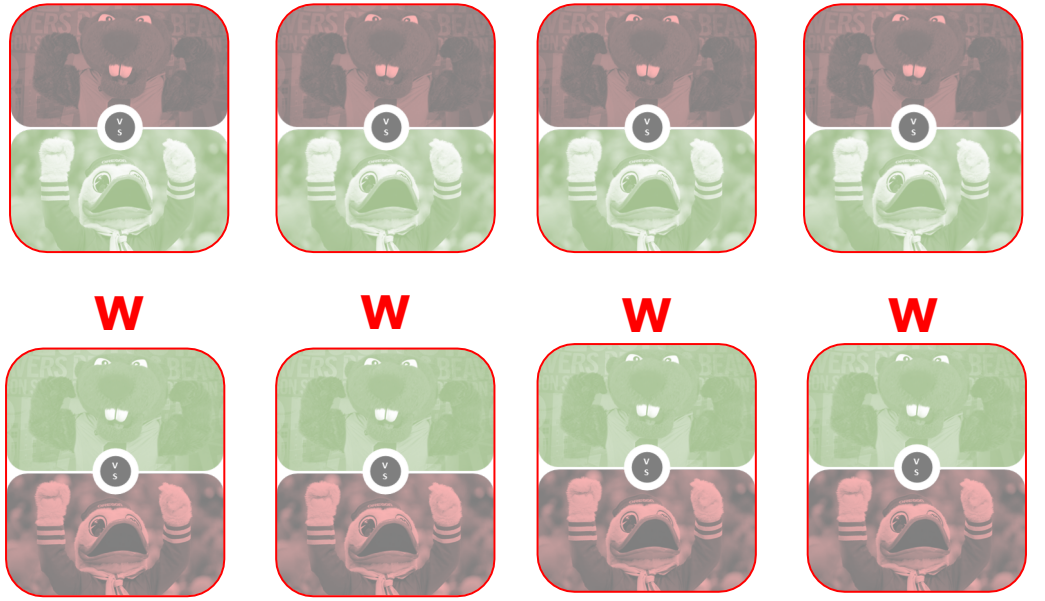
W W L L W

Add fake observation to express our prior belief and find MAP:

$$\hat{\theta}_{MAP} = \frac{\#wins + \#fakewins}{total} = \frac{7}{13}$$

L L L L

W W W W



This slide in no way condones or encourages underground mascot fights or illegal gambling. Even if we would win.

An Informal Way to Add Prior Belief

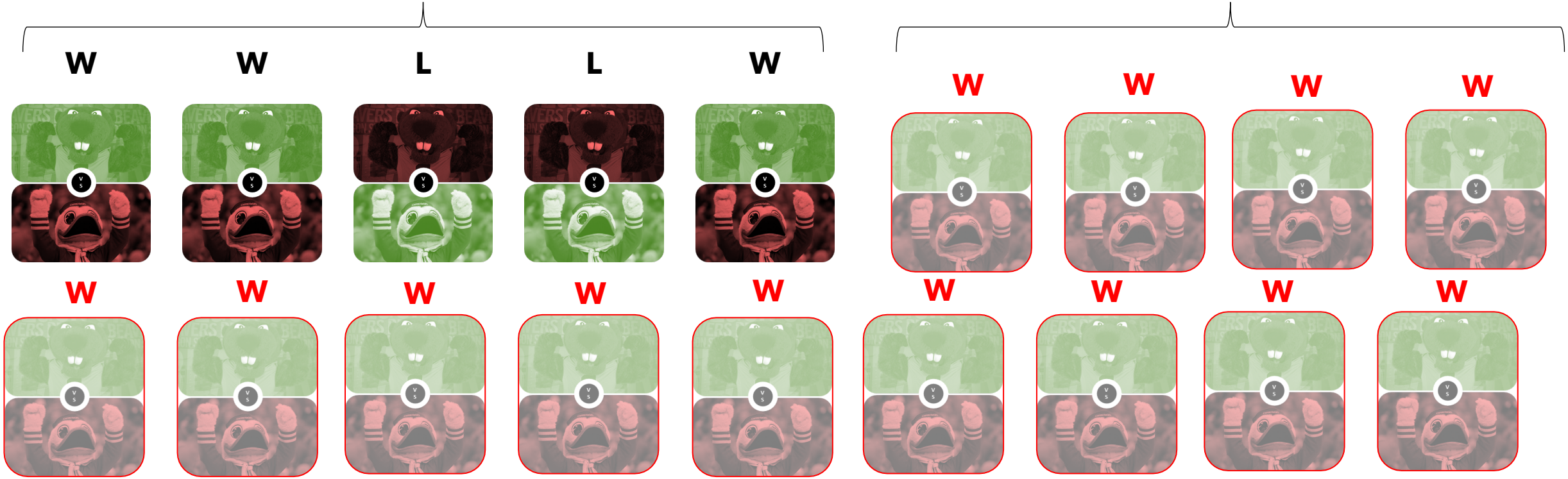
I can express different priors and different confidences in those priors by changing the outcome/number of fake observations

Real data we've observed yields

$$\hat{\theta}_{MLE} = \frac{\#wins}{total} = \frac{3}{5}$$

Add fake observation to express our prior belief and find MAP:

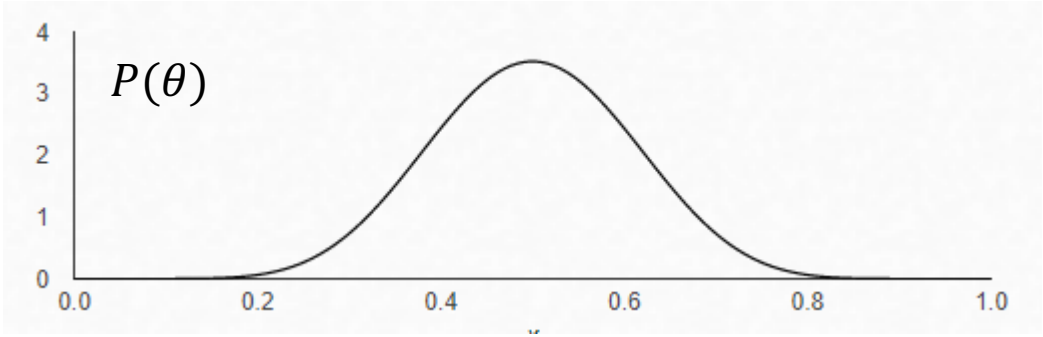
$$\hat{\theta}_{MAP} = \frac{\#wins + \#fakewins}{total} = \frac{16}{18}$$



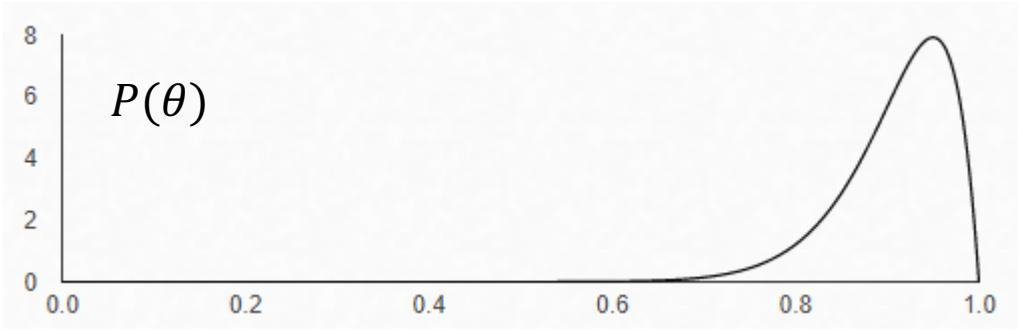
This slide in no way condones or encourages underground mascot fights or illegal gambling. Even if we would win

A Formal Way to Add Prior Beliefs

An alternative way to reason about what we were doing is to assume some prior distribution for the value of our parameter θ itself:



Expresses belief that theta is around 0.5, but not too certain.



Expresses belief that theta is near 0.9 and is fairly certain.

https://homepage.divms.uiowa.edu/~mbognar/applets/beta.html

Your First Probabilistic Learning Algorithm

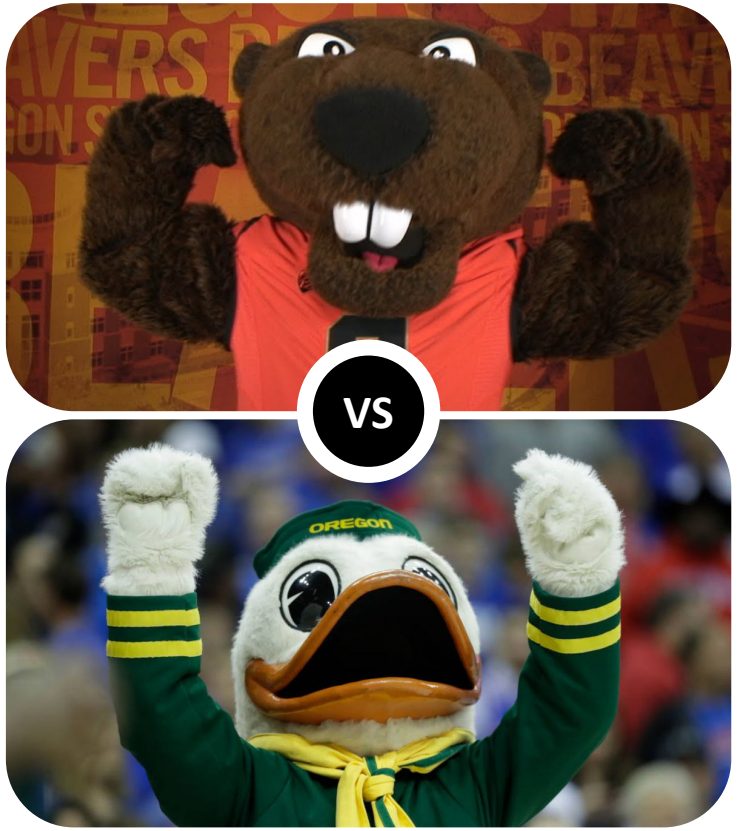
What happened in the past? **W, W, L, L, W**

Your boss says: Which is better? 3/5 or 300/500?

~~**You say:** Uh…. Same estimate? More the merrier?~~

You say: Same estimate but we can be much more sure it isn't just random variation. And of course I can prove it!

Your boss says: This is why I pay you the big bucks.



Your boss says: What if I'm pretty sure the duck is using performance enhancing drugs this time?

You say: Ah… Bayesian it is then…

This slide in no way condones or encourages underground mascot fights or illegal gambling. Even if we would win.

Likelihood $P(D|\theta)$:

- We've already seen this in MLE. The likelihood of the data given parameters θ

Priors $P(\theta)$:

- Express beliefs about parameters before any observations are included
- Can act as regularizers (more about this next lecture)

Posterior $P(\theta|D) \propto P(D|\theta)P(\theta)$:

- Distribution over θ after considering both the dataset and the prior

Conjugate Priors:

- Prior is conjugate to likelihood if it leads to the same distribution as posterior
- Gives closed form representation of posterior!
  - Using non-conjugates can get computationally messy.

We will very often be most interested in the posterior.

Posterior
Belief about theta after evidence and prior belief combined

Likelihood
How likely theta makes the dataset

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

Prior
Belief about theta provided by human

Constant with respect to theta so typically will not address

Important to note – the posterior is a distribution over theta!

Your Second Statistical Learning Algorithm: MAP

Maximum A Posteriori – Find parameters that make the observed data most likely but consider a prior over the parameters.

1. Assume a prior distribution over θ, $P(\theta)$

2. Assume a probabilistic model of how the data is generated:
   -- parameter $\theta \sim P(\theta)$ and then data $x \sim P(x|\theta)$

3. Find $\hat{\theta}_{MAP}$ that maximize the posterior $P(\theta|D) \propto P(D|\theta)P(\theta)$

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

Why MAP?

- Rigorous framework to combine observations (likelihood) with beliefs (prior)

Question Break!

Assuming a dataset D and a parameter being estimated θ, label these distributions:

$$P(\theta), \qquad P(D|\theta), \qquad P(\theta|D)$$

A Posterior, Prior, Likelihood

B Prior, Posterior, Likelihood

C Likelihood, Posterior, Prior

D Prior, Likelihood, Posterior

A Note on Beta Distributions

Introducing the **Beta Distribution**
(continuous from 0 to 1)

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} = Beta(\beta_H, \beta_T)$$

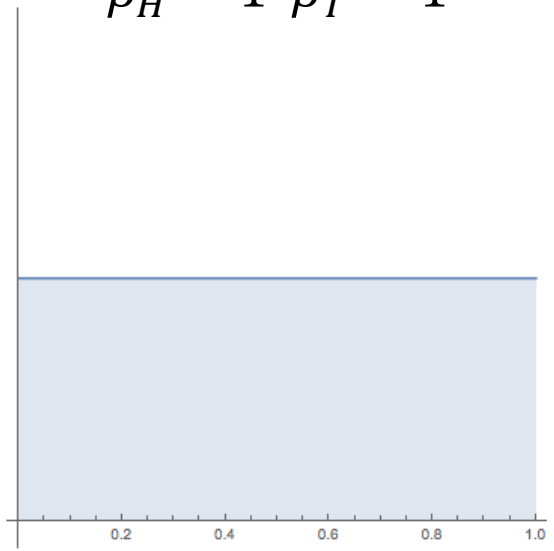
$$\beta_H > 0 \;\; \beta_T > 0$$

A normalizer so it integrates to 1.

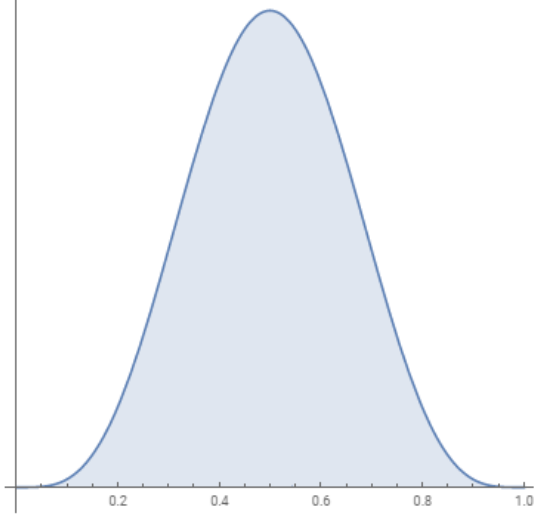
$$\text{Mode of Beta} = \frac{\beta_H - 1}{\beta_H + \beta_T - 2}$$

$$if \;\; \beta_H > 1 \;\; \beta_T > 1$$

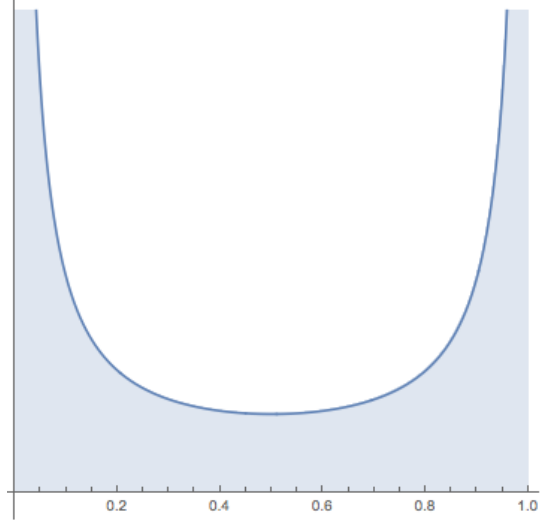
$\beta_H = 1 \;\; \beta_T = 1$



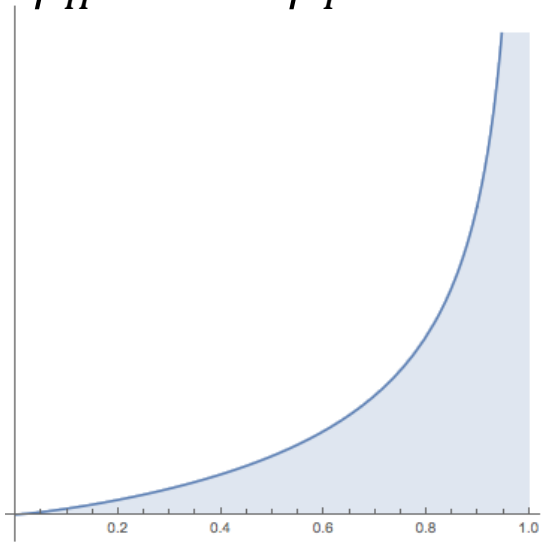
$\beta_H = 5 \;\; \beta_T = 5$



$\beta_H = 0.01 \;\; \beta_T = 0.01$



$\beta_H = 2.20 \;\; \beta_T = 0.40$



General Case: Estimating the Bias of a Coin with MAP and a Beta Prior

Dataset: Let X_i be a binary random variable (H→1, T→0). Given a dataset $D = \{x_i\}_{i=1}^{N}$ representing a coin being flipped N times. (e.g. {H,T,H,H,H,T} → {1,0,1,1,1,0})

Model Assumption: Assume that X is distributed according to a Bernoulli distribution with parameter θ distributed according to a $Beta(\beta_H, \beta_T)$.

Write out posterior of parameters θ and then find its maximum:

$$\boldsymbol{P(\theta|D)} \propto P(D \mid \theta)P(\theta) = P(\theta)\prod_{i=1}^{N} P(x_i \mid \theta)$$

We know this term. Same as back when we did MLE.

$$= P(\theta)\ \theta^{\#H}(1-\theta)^{\#T}$$

$$\underset{Beta(\beta_H,\beta_T)}{P(\theta)} = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H,\beta_T)}$$

$$\propto \theta^{\#H+\beta_H-1}(1-\theta)^{\#T+\beta_T-1}$$

$$\propto Beta(\#H+\beta_H, \#T+\beta_T)$$

General Case: Estimating the Bias of a Coin with MAP and a Beta Prior

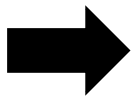
Dataset: Let X_i be a binary random variable (H→1, T→0). Given a dataset $D = \{x_i\}_{i=1}^{N}$ representing a coin being flipped N times. (e.g. {H,T,H,H,H,T} → {1,0,1,1,1,0})

Model Assumption: Assume that X is distributed according to a Bernoulli distribution with parameter θ distributed according to a $Beta(\beta_H, \beta_T)$.

Write out posterior of parameters θ and then find its maximum:

$$\boldsymbol{P(\theta|D)} = Beta(\#H + \beta_H, \#T + \beta_T)$$

$$\text{Mode of Beta} = \frac{\beta_H - 1}{\beta_H + \beta_T - 2} \quad if\ \beta_H > 1\ \ \beta_T > 1$$



$$\boldsymbol{\hat{\theta}_{MAP}} = \frac{\#H + \beta_H - 1}{\#H + \beta_H + \#T + \beta_T - 2}$$

Works the Long Way Too. No magic.

Write out posterior of parameters θ:

$$\boldsymbol{logP(\theta|D)} \propto \log P(\theta) + \sum_{i=1}^{N} \log P(x_i|\theta)$$

$$= \log P(\theta) + \#H * log\,\theta + \#T log(1-\theta)$$

$$= (\#H + \beta_H - 1) * log\,\theta + (\#T + \beta_T - 1) log(1-\theta) - \log B(\beta_H, \beta_T)$$

Find maximum by setting derivative to 0:

$$\frac{d\ logP(D|\theta)P(\theta)}{d\theta} = \frac{\#H + \beta_H - 1}{\theta} + \frac{\#T + \beta_T - 1}{1-\theta}$$

$$\boldsymbol{\hat{\theta}_{MAP}} = \frac{\#H + \beta_H - 1}{\#H + \beta_H + \#T + \beta_T - 2}$$

General Case: Estimating the Bias of a Coin with MAP and a Beta Prior

β_H and β_T seem to behave like "fake" heads/tails here!

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{MAP}} = \frac{\#H + \beta_H - 1}{\#H + \beta_H + \#T + \beta_T - 2}$$

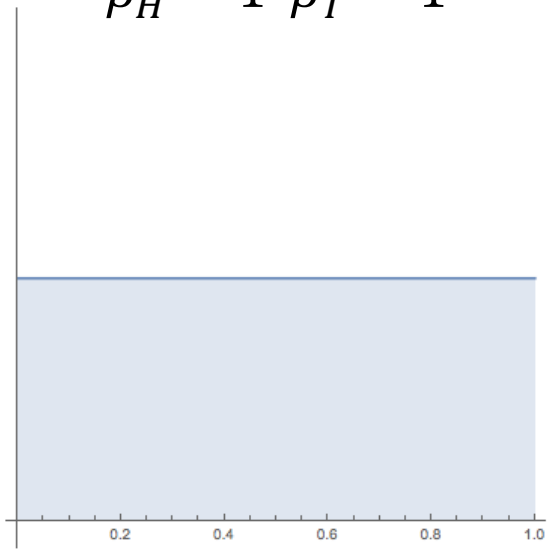
If $\beta_H = \beta_T = 1$ → returns to the MLE estimate:

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{MLE}} = \frac{\#H}{\#H + \#T}$$

This setting is a case of a "uniform prior" such that:

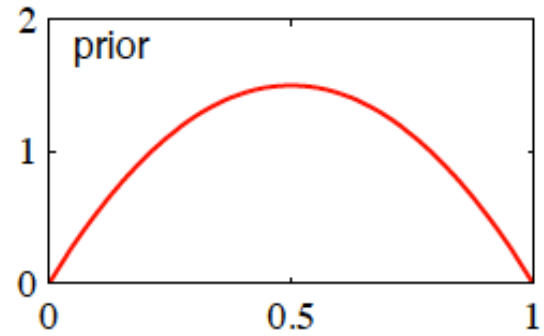
$$P(\theta|D) \propto P(D|\theta)P(\theta) = P(D|\theta)$$

$\beta_H = 1 \;\; \beta_T = 1$

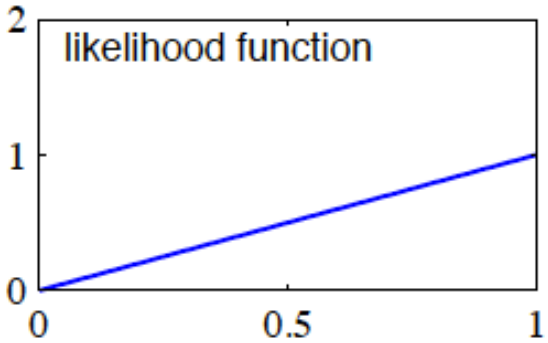


Effect of Prior on Coin Flip

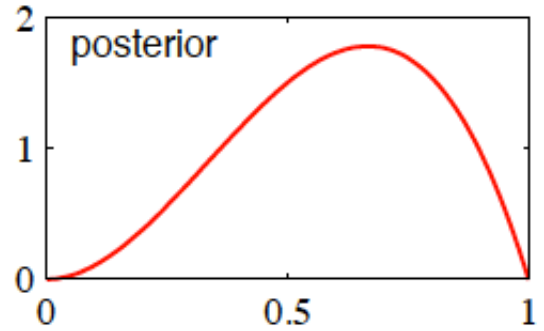
- **Prior = Beta(2,2)**
  - $\theta_{prior} = 0.5$



- **Dataset = {H}**
  - $L(\theta) = \theta$
  - $\theta_{MLE} = 1$



- **Posterior = Beta(3,2)**
  - $\theta_{MAP} = (3-1)/(3+2-2) = 2/3$



Question Break!

What happens as the number of observed data points gets larger?

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{MAP}} = \frac{\#H + \beta_H - 1}{\#H + \beta_H + \#T + \beta_T - 2}$$

A The fake heads/tails from the prior matter more

B The fake heads/tails from the prior matter less

C You lost me with this Bayesian stuff

D

Conjugate Priors

Likelihood/prior distribution pairs that when multiplied result in a new version of the prior distribution.

Some well-known and useful ones:

likelihood $p(x \mid \theta)$	conjugate prior $p_0(\theta)$	posterior $p(\theta \mid x)$
$\text{Normal}(\theta, \sigma)$	$\text{Normal}(\mu_0, \sigma_0)$	$\text{Normal}(\mu_1, \sigma_1)$
$\text{Binomial}(N, \theta)$	$\text{Beta}(r, s)$	$\text{Beta}(r + n, s + N - n)$
$\text{Poisson}(\theta)$	$\text{Gamma}(r, s)$	$\text{Gamma}($ [redacted] $)$
$\text{Multinomial}(\theta_1, \ldots, \theta_k)$	$\text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$	$\text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_k + n_k)$

Question Break!

What is the difference between a MLE and MAP estimate?

A MLE makes assumptions about the parameters before observing data

B One has more vowels

C There are no differences

D MAP maximizes the posterior (likelihood * prior) whereas MLE maximizes the likelihood

Today's Learning Objectives

Be able to answer:

- ~~What are basic concepts and definitions in probability theory?~~
- ~~What is maximum likelihood estimation (MLE)?~~
  - ~~What is a likelihood function?~~
  - ~~What is the MLE estimate for a Bernoulli variable?~~
  - ~~What is bias and variance of an estimator?~~
  - What is the MLE estimate for a Normal variable?
- ~~What is a maximum a posteriori estimate (MAP)?~~
  - ~~What is a prior and posterior? What is a~~

Next time

Next Time: We will consider conditional models and head into least-squares linear regression.