Application: Tail Bounds

CSE 312 Winter 26
Lecture 21

Markov's Inequality

Let X be a random variable supported (only) on non-negative numbers. For any $t > 0$

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

Chebyshev's Inequality

Let X be a random variable. For any $t > 0$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbf{Var}(X)}{t^2}$$

(Multiplicative) Chernoff Bound

Let $X_1, X_2, \ldots, X_n$ be *independent* Bernoulli random variables. *Let* $X = \sum X_i$, and $\mu = \mathbb{E}[X]$. For any $0 \leq \delta \leq 1$

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{3}\right) \text{ and } \mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{2}\right)$$

But Wait! There's More

For this class, please limit yourself to:
Markov, Chebyshev, and Chernoff, as stated in these slides…

But for your information. There's more.

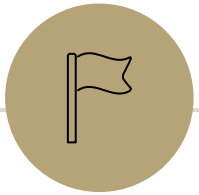
Trying to apply Chebyshev, but only want a "one-sided" bound (and tired of losing that almost-factor-of-two)Try Cantelli's Inequality

In a position to use Chernoff, but want additive distance to the mean instead of multiplicative? They got one of those.

Have a sum of independent random variables that aren't indicators, but are bounded, you better believe Wikipedia's got one

Have a sum of random **matrices** instead of a sum of random numbers. Not only is that a thing you can do, but the eigenvalue of the matrix concentrates

There's a whole book of these!



One More Bound

The Union Bound

The Union bound

Union Bound

For any events E, F

$$\mathbb{P}(\boldsymbol{E} \cup \boldsymbol{F}) \leq \mathbb{P}(\boldsymbol{E}) + \mathbb{P}(\boldsymbol{F})$$

Proof? $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$

And $\mathbb{P}(E \cap F) \geq 0$.

Concentration Applications

A common pattern:

Figure out “what could possibly go wrong” – often these are dependent.

Use a concentration inequality for each of the things that could go wrong.

Union bound over everything that could go wrong.

Frogs

There are 20 frogs on each location in a 5x5 grid. Each frog will independently jump to the left, right, up, down, or stay where it is with equal probability. A frog at an edge of the grid magically warps to the corresponding edge (pac-man-style).

Bound the probability that at least one square ends up with at least 36 frogs.

These events are dependent – adjacent squares affect each other!

Frogs (answer)

For an arbitrary location:

There are 100 frogs who could end up there (those above, below, left, right, and at that location). Each with probability $.2$. Let X be the number that land at the location we're interested in.

$$\mathbb{P}(X \geq 36) = \mathbb{P}(X \geq (1+\delta)20) \leq \exp\left(-\frac{\left(\frac{4}{5}\right)^2 \cdot 20}{3}\right) \leq 0.015$$

There are 25 locations. Since all locations are symmetric, by the union bound the probability of at least one location having 36 or more frogs is at most $25 \cdot 0.015 \leq 0.375$.

Takeaways

When an operation is expensive, you can often design an experiment so that $\mathbb{E}[X]$ is the value you want to learn.

Designing a poll

An ML example: stochastic gradient descent

The "gradient" is the direction you should update parameters to decrease error.

Computing a gradient is time-consuming (look at every datapoint). Looking at a random subset of data points is faster, and the expected value of your estimate is the target gradient. But can you guarantee you're close…

Concentration inequalities let you bound "how close" you'll be for big n.

Concentration lets you bound probabiltities for general n

[random-pivot] Quicksort runs in time O(n log n) with probability 1-1/n

The proof involves counting the number of "good" pivots, which is a binomial random variable…dependent on $\log n$, where n is the size of your array.

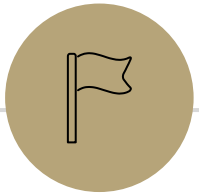
Tail Bounds – Takeaways

Useful when an experiment is complicated and you just need the probability to be small (you don't need the exact value).

Choosing a minimum n for a poll – don't need exact probability of failure, just to make sure it's small.

Designing probabilistic algorithms – just need a guarantee that they'll be extremely accurate

Learning more about the situation (e.g. learning variance instead of just mean, knowing bounds on the support of the starting variables) usually lets you get more accurate bounds.



Applications

Privacy Preservation

A real-world example (adapted from *The Ethical Algorithm* by Kearns and Roth; based on protocol by Warner [1965]).

And gives a sense of how randomness is actually used to protect privacy.

Privacy Preservation with Randomness

You're working with a social scientist. They want to get accurate data on the rate at which people cheat on their romantic partners.

We know about polling accuracy!

Do a poll, call up a random sample of adults and ask them "have you ever cheated on your romantic partner?"

Use a tail-bound to estimate the needed number n get a guaranteed good estimate, right?

You do that, and somehow, no one says they cheated.

What’s the problem?

People lie.

Or they might be concerned about you keeping this data.

Databases can be leaked (or infiltrated. Or subpoenaed).

You don’t want to hold this data, and the people you’re calling don’t want you to hold this data.

Doing Better With Randomness

You don’t really need to know **who** was cheating. Just how many people were.

Here’s a protocol:

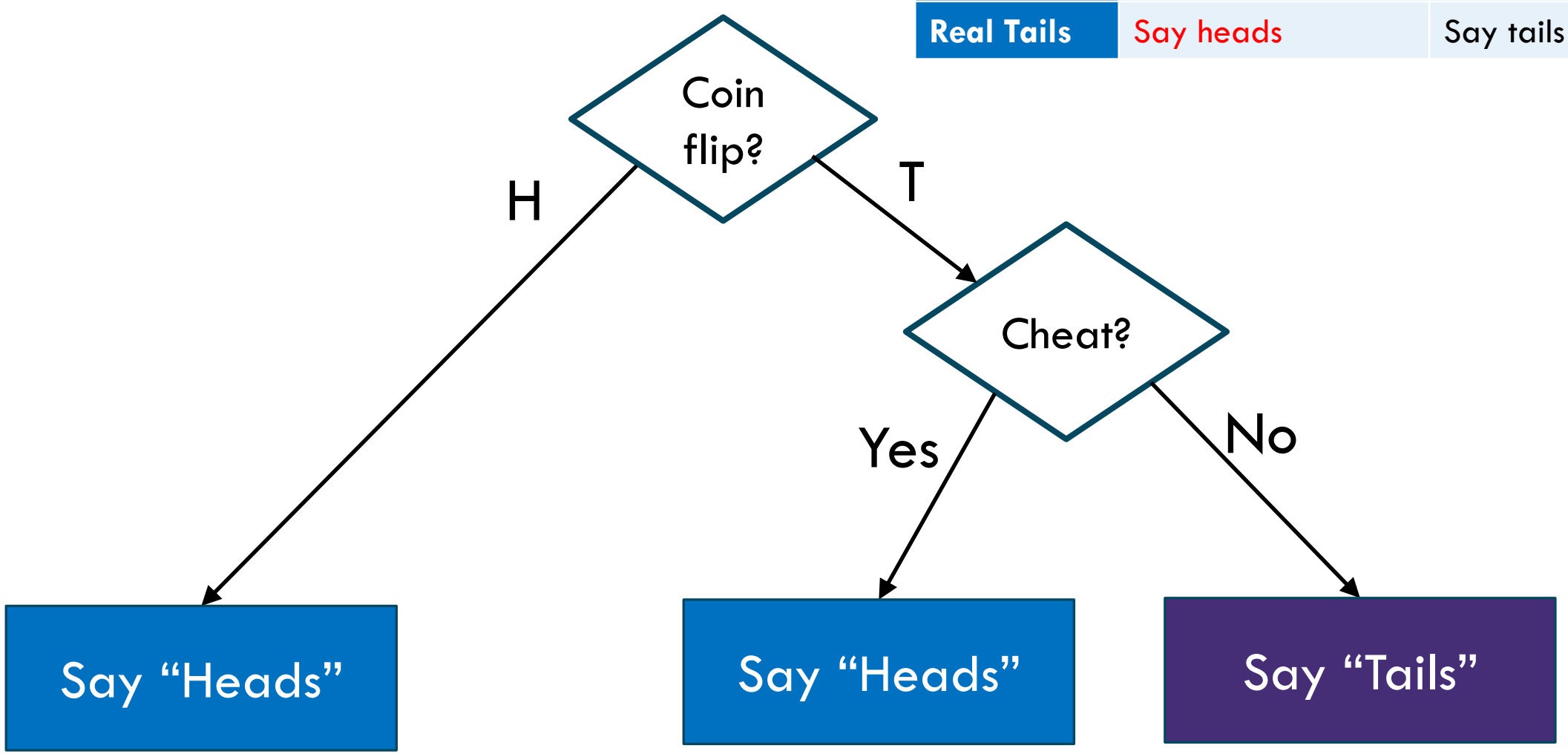
Please flip a coin.

If the coin is heads, or you have ever cheated, please tell me “heads”

If the coin is tails and you have not ever cheated, please tell me “tails”

Warner's Protocol

Result	Yes (cheated)	No (never cheated)
Real Heads	Say heads	Say heads
Real Tails	Say heads	Say tails



Will it be private?

If you are someone who has cheated, and you report heads can that be used against you? Not substantially – just say “no the coin came up heads!”

You discover your partner said heads, what’s the probability that they cheated?

Will it be private? (computation)

If you are someone who has cheated on your spouse, and you report heads can that be used against you? Not substantially – just say "no the coin came up heads!"

$$\mathbb{P}(C|H) = \frac{\mathbb{P}(H|C)\cdot\mathbb{P}(C)}{\mathbb{P}(H)} = \frac{1\cdot\mathbb{P}(C)}{\frac{1}{2}\mathbb{P}(\overline{C}) + 1\cdot\mathbb{P}(C)}$$

Is this a substantial change?

No. For real world values (~15%) of $\mathbb{P}(C)$, the probability estimate would increase (to ~26%). But that isn't too damaging.

But will it be accurate?

But we've lost our data haven't we? People answered a different question. Can we still estimate how many people cheated?

Suppose you asked 100 people the "heads/tails" question, and 60 people said "heads." What do you predict would be the number of people who cheated on a partner?

Can you generalize your idea for n people polled, and X the number of people that said "heads"?

But will it be accurate? (definition)

But we've lost our data haven't we? People answered a different question. Can we still estimate how many people cheated?

Suppose you poll n people, and let X be the number of people who said "heads" We'll find an estimate Y of the number of people who cheated in the sample, and let p be the true probability of cheating in the population. What should Y be? Can we draw a margin of error around Y?

$$\mathbb{P}(X_i = 1) = \frac{1}{2} + \frac{1}{2} \cdot p$$

$$\mathbb{E}[X] = \frac{n}{2} + \frac{1}{2}\mathbb{E}[Y]$$

We'll define Y to be: $Y = 2\left(X - \frac{n}{2}\right)$.

This is a **definition**, based on how the $\mathbb{E}[Y]$ should relate to the $\mathbb{E}[X]$.

But will it be accurate? (Variance)

$Y = 2\left(X - \frac{n}{2}\right)$

$\mathrm{Var}(X) = \mathrm{Var}(\sum X_i) = \sum \mathrm{Var}(X_i)$

$\mathrm{Var}(X_i)$? It's an indicator with parameter $p + (1-p) \cdot \frac{1}{2} = \frac{1}{2} + \frac{p}{2}$

So $\mathrm{Var}(X_i) = \left(\frac{1}{2} + \frac{p}{2}\right)\left(\frac{1}{2} - \frac{p}{2}\right)$

$\mathrm{Var}(Y) = 4\mathrm{Var}(X) = 4n\mathrm{Var}(X_i) = 4n\left(\frac{1}{2} + \frac{p}{2}\right)\left(\frac{1}{2} - \frac{p}{2}\right) \leq \frac{4n}{4} = n$

The variance is 4 times as much as it would have been for a non-anonymous poll.

Can we use Chernoff?

(Multiplicative) Chernoff Bound

Let $X_1, X_2, \ldots, X_n$ be *independent* Bernoulli random variables.
Let $X = \sum X_i$, and $\mu = \mathbb{E}[X]$. For any $0 \leq \delta \leq 1$

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{3}\right) \text{ and } \mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{2}\right)$$

What happens with $\mathbf{n = 1000}$ people?

What range will we be within at least 95% of the time?

A different inequality

If we try to use Chernoff, we'll hit a frustrating block.

Since μ depends on p, p appears in the formula for δ. And we wouldn't get an absolute guarantee unless we could plug in a p.

And it'll turn out that as $p \to 0$ that $\delta \to \infty$ so we don't say anything then.

Luckily, there's always another bound…

☹ Can't bound δ without bounding p

The right tail is the looser bound, so ensuring the right tail is less than 2.5% gives us the needed guarantee.

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{3}\right) = \exp\left(-\frac{\delta^2 1000p}{3}\right) \leq .025$$

$$-\frac{\delta^2 1000p}{3} \leq \ln(.025)$$

$$-\delta^2 \leq \frac{3\cdot\ln(.025)}{1000p}$$

$$\delta \geq \sqrt{\frac{-3\ln(.025)}{1000p}}$$

As $p \to 0$, $\delta \to \infty$ – we're not actually making a claim anymore.

Hoeffding's Inequality

Hoeffding's Inequality

Let $X_1, X_2, \ldots, X_n$ be *independent* RVs, each with range $[0,1]$.
Let **$\bar{X} = \sum X_i/n$, and $\mu = \mathbb{E}[\bar{X}]$. For any $t \geq 0$**

$$\mathbb{P}(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2\exp(-2nt^2)$$

$|X - \mathbb{E}[X]| \geq t$ if and only if $|Y - \mathbb{E}[Y]| \geq 2t$. Why?

$$Y = 2\left(X - \frac{n}{2}\right) \text{ or } X = \frac{Y+n}{2}$$

$$|X - \mathbb{E}[X]|$$
$$= \left|\frac{Y+n}{2} - \mathbb{E}\left[\frac{Y+n}{2}\right]\right|$$
$$= \left|\frac{Y+n}{2} - \mathbb{E}\left[\frac{Y}{2}\right] - \frac{n}{2}\right|$$
$$= \left|\frac{Y}{2} - \mathbb{E}\left[\frac{Y}{2}\right]\right|$$
$$= \frac{1}{2}|Y - \mathbb{E}[Y]|$$

So $|X - \mathbb{E}[X]| \geq t$ if and only if $\frac{1}{2}|Y - \mathbb{E}[Y]| \geq t$ iff $|Y - \mathbb{E}[Y]| \geq 2t$.

Use Hoeffding's Inequality

Hoeffding's Inequality

Let $X_1, X_2, \ldots, X_n$ be *independent* RVs, each with range $[0,1]$.
Let $\bar{X} = \sum X_i/n$, and $\mu = \mathbb{E}[\bar{X}]$. For any $t \geq 0$

$$\mathbb{P}(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2\exp(-2nt^2)$$

How close will we be with *n=1000 with probability at least* .95*?*

$|X - \mathbb{E}[X]| \geq t$ if and only if $|Y - \mathbb{E}[Y]| \geq 2t$.

Margin of Error

$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq t) = \mathbb{P}(|X - \mathbb{E}[X]| \geq t/2) \leq 2\exp(-2nt^2) \leq .05$

For $n = 1000$, we get:

$2\exp\left(-2n\left(\frac{t}{2}\right)^2\right) \leq .05 \Rightarrow -\frac{2000t^2}{4} \leq \ln(.025) \Rightarrow t \leq .086.$

$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq .086) \leq .05$

So our margin of error is about 8.6%.

To get a margin-of-error of 5% need $2\exp\left(-2n\left(\frac{.05}{2}\right)^2\right) \leq .05$

$n \geq 2952$

How much do we lose?

We lose a factor of two in the length of the margin (equivalently, we'd need to talk to 4 times as many people to have the same confidence.

You can also control this tradeoff.

Want more accuracy? Make it roll a die: report 1 if cheated (truth o/w)

Want more security? Make it Bernoulli with probability $p \gg \frac{1}{2}$ or cheated have the same report (e.g. report "die roll 1 [and didn't cheat]" or "die roll 2-6 [or did cheat]"

Will people actually admit to anything?

Yes, they actually will.

Ask Me Anything: Assessing Academic Dishonesty

By Nathan Brunelle and John R. Hott

When Nathan was at University of Virginia, he and a colleague ran this protocol, asking students whether they had cheated in their course.

They estimated about 40% of students cheated in one of their courses.

In The Real World

Injecting randomness to preserve privacy is a real thing.

Instead of having everyone flip a coin, “random noise” can be inserted after all the data has been collected.

Differential privacy is being used to protect the 2020 Census data.

The overall count of people in each state is exact (well, exactly the data they collected). But the data per block or per city has been randomized to protect against revealing who lives where.

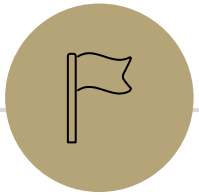
This video nicely explains what’s involved. Notice that the accuracy guarantees come in the same “inside-margin-of-error-with-probability” guarantees we’ve been giving for our randomness (just much stronger).

More applications?

Concentration inequalities can help you tell the accuracy of predictions you make (or that you learn from data).

Next, we'll see "maximum likelihood estimation", a method of making predictions from data.

Later, we'll hint at some other ML-related uses of running an experiment, designed such that $\mathbb{E}[X]$ is some unknown-to-you value that's too expensive to compute.



Maximum Likelihood Estimation

Asking The Opposite Question

So far:

Given rules for an experiment.

Given the event/outcome we're interested in.

You calculate/estimate/bound what the probability is.

Estimation Problems:

Given **some** of the rules of the experiment.

Given what happened.

You estimate what the rest of the rules of the experiment were.

Example

Suppose you flip a coin independently 10 times, and you see

HTTTHHTHHH

What is your estimate of the probability the coin comes up heads?

A. Something less than 0.5
B. 0.5
C. Something between 0.5 and 0.6
D. 0.6
E. Something more than 0.6

Maximum Likelihood Estimation (idea)

Idea: we got the results we got.

High probability events happen more often than low probability events.

So, guess the rules that maximize the probability of the events we saw (relative to other choices of the rules).

Since that event happened, might as well guess the set of rules for which that event was a ‘high probability event’.

Maximum Likelihood Estimation (definition)

Formally, we are trying to estimate a parameter of the experiment (here: the probability of a coin flip being heads).

The likelihood of an event E under a parameter θ is

$\mathcal{L}(E;\theta)$ is $\mathbb{P}(E)$ when the experiment is run with θ

We'll use the notation $\mathbb{P}(E;\theta)$ for probability when run with parameter θ where the semicolon means "extra rules" rather than conditioning

We will choose $\hat{\theta} = \mathbf{argmax}_{\theta}\ \mathcal{L}(E;\theta)$

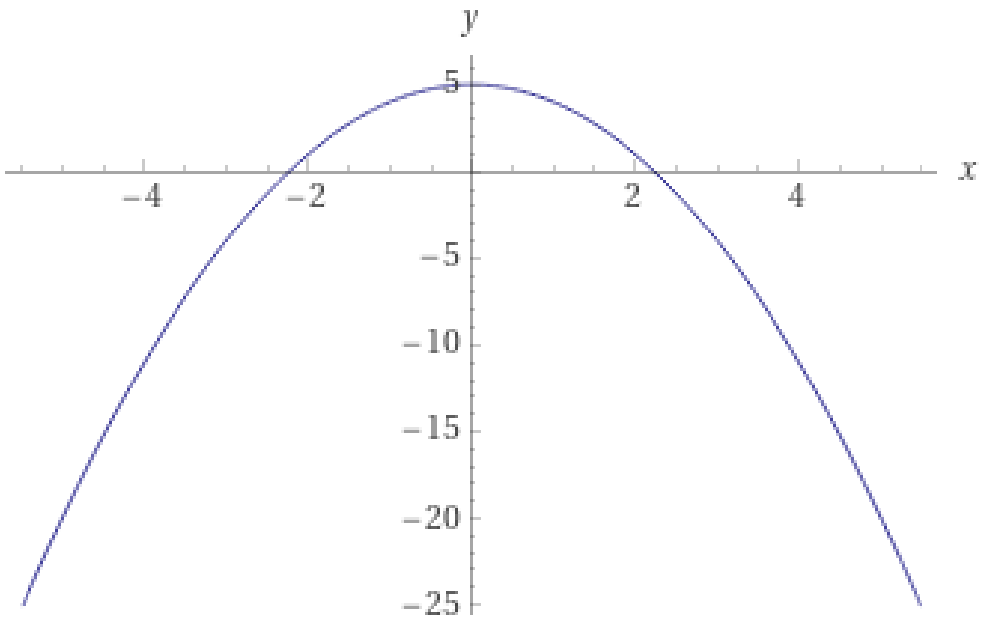
$\mathbf{argmax}$ is the argument that produces the maximum so the θ that causes $\mathcal{L}(E;\theta)$ to be maximized.

Argmax?

For example

$\mathbf{argmax}_x\ (5 - \mathrm{x}^2) = 0$

$\max_x(5 - x^2) = 5$, the input (argument) which produces 5 is 0, so the argmax is 0



Notation comparison

$\mathbb{P}(X|Y)$ probability of **event** X, conditioned on the **event** Y having happened (Y is a subset of the sample space).

$\mathbb{P}(X;\theta)$ probability of X, where to properly define our probability space we need to know the extra piece of information θ**.** Since θ isn't an event, this is not conditioning.

$\mathcal{L}(X;\theta)$ the likelihood of event X, given that an experiment was run with parameter θ. Likelihoods don't have all the properties we associate with probabilities (e.g. summing them up doesn't give 1) and this isn't conditioning on an event (θ is a parameter/rule of how the event could be generated).

MLE

Maximum Likelihood Estimator

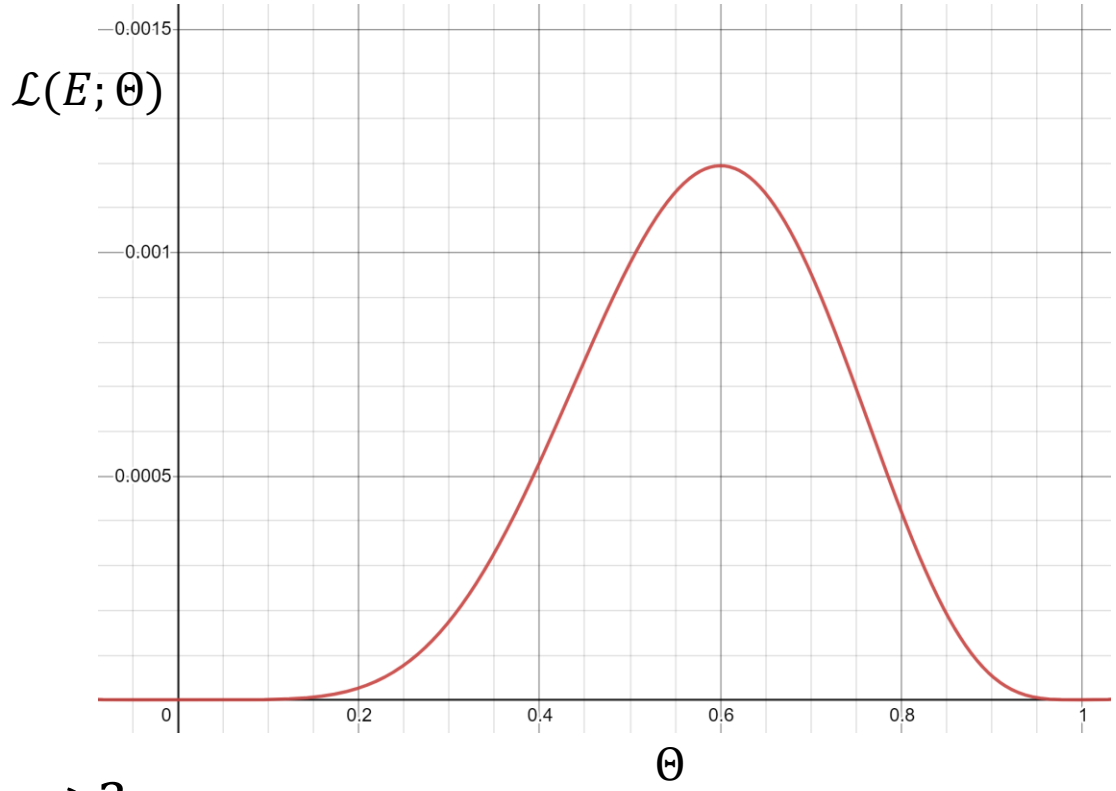
The maximum likelihood estimator of the parameter θ is:

$$\hat{\theta} = \text{argmax}_{\theta} \ \mathcal{L}(E; \theta)$$

θ is a variable, $\hat{\theta}$ is a number (or formula given the event).

We'll also use the notation $\hat{\theta}_{\text{MLE}}$ if we want to emphasize how we found this estimator.

The Coin Example



$\mathcal{L}(\text{HTTTHHHTHHH}\,;\theta) = \theta^6(1-\theta)^4$

Where is θ maximized?

How do we usually find a maximum?

Calculus!!

$$\frac{d}{d\theta}\theta^6(1-\theta)^4 = 6\theta^5(1-\theta)^4 - 4\theta^6(1-\theta)^3$$

Set equal to 0 and solve

$$6\theta^5(1-\theta)^4 - 4\theta^6(1-\theta)^3 = 0 \Rightarrow 6(1-\theta) - 4\theta = 0 \Rightarrow -10\theta = -6 \Rightarrow \theta = \frac{3}{5}$$

The Coin Example (maximizer)

For this problem, θ must be in the closed interval $[0,1]$. Since $\mathcal{L}()$ is a continuous function, the maximum must occur at and endpoint or where the derivative is 0.

Evaluate $\mathcal{L}(\cdot;0) = 0, \mathcal{L}(\cdot;1) = 0$

at $\theta = 0.6$ we get a positive value,

so $\theta = 0.6$ is the maximizer on the interval $[0,1]$.

Maximizing a Function

CLOSED INTERVALS

Set derivative equal to 0 and solve.

Evaluate likelihood at endpoints and any critical points.

Maximum value must be maximum on that interval.

SECOND DERIVATIVE TEST

Set derivative equal to 0 and solve.

Take the second derivative. If negative everywhere, then the critical point is the maximizer.