

PAC helps us express bounds on overfitting

- PAC = Probably Approximately Correct criterion

$$P \left(\left| R(h) - \hat{R}(h) \right| \leq \epsilon, \forall h \in \mathcal{H} \right) \geq 1 - \delta$$

for some ϵ (difference between true and empirical risk)
and δ (probability of “failure”)

why the name?

Bounds (finite cases)

- Finite, realizable case:

$$R(h) \leq \frac{1}{M} \left(\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right) \right)$$

- Finite, agnostic case:

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2M} \left(\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right) \right)}$$

- for all $h \in \mathcal{H}$, with M examples, failure probability δ

For infinite hypothesis classes, growth function

- Def'n: the **growth function** $S_{\mathcal{H}}(M)$ is the maximum number of **distinct** $h \in \mathcal{H}$ for a dataset of size M
- for halfspaces in 2D,

| | | | | | | |
|----------------------|---|---|---|----|-----------|-----|
| M | 1 | 2 | 3 | 4 | 5 | ... |
| $S_{\mathcal{H}}(M)$ | 2 | 4 | 8 | 14 | ≤ 26 | ... |

- for larger M , it turns out $S_{\mathcal{H}}(M) = O(M^3)$

not obvious!

Two kinds of behavior

PAC learnable (can't memorize more than d points)

- For many hypothesis classes \mathcal{H} , similar behavior:

$$S_{\mathcal{H}}(M) = \begin{cases} 2^M & M \leq d & \text{shattered} \\ \ll 2^M & M > d & \text{not shattered} \end{cases}$$

- e.g., intervals (or rectangles or hyperrectangles)
- e.g., bounded-depth decision trees
- e.g., fixed-architecture neural networks

- For many other classes \mathcal{H} , instead $S_{\mathcal{H}}(M) = 2^M$ for all M

- e.g., unbounded-depth decision trees
- e.g., unbounded-size neural networks

can shatter a set of each size

Not PAC learnable (can memorize at any $|\mathcal{D}|$)

Sauer's lemma

d is called the **VC-dimension** of \mathcal{H}

- Suppose $S_{\mathcal{H}}(M) = 2^M$ for $M \leq d$, but $S_{\mathcal{H}}(d+1) < 2^{d+1}$
 - ➔ Then $S_{\mathcal{H}}(M) = O(M^d)$

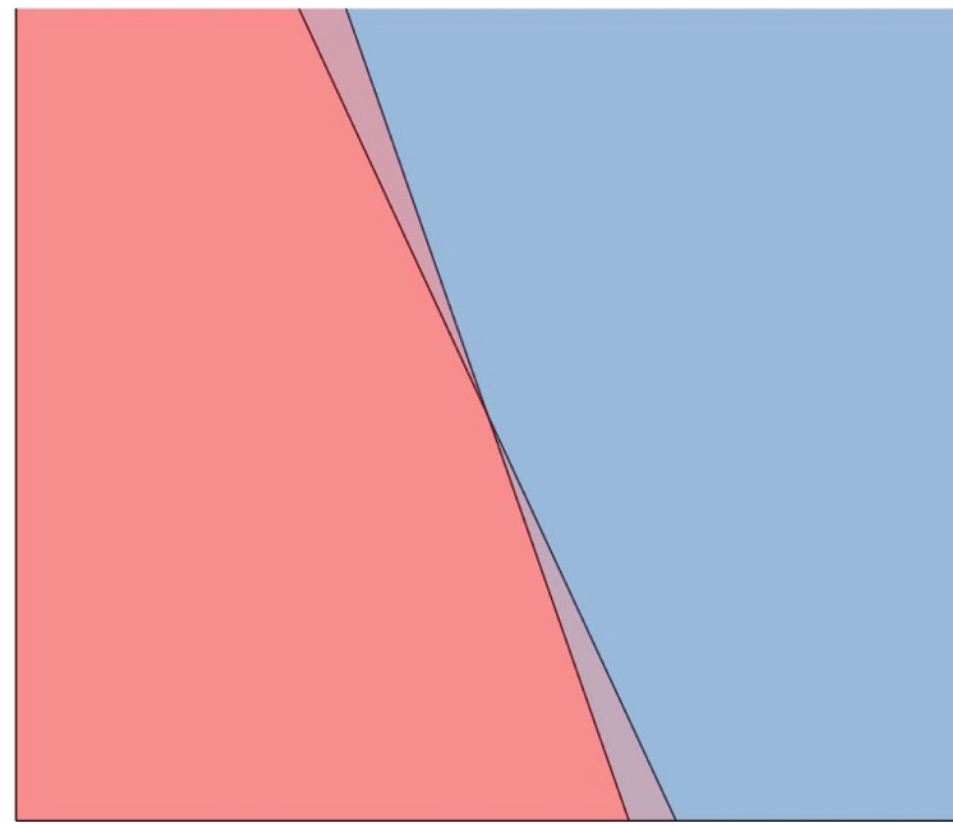
“Suppose we grow exponentially (i.e., shatter) only up to $M = d$. Then for $M > d$ we grow polynomially, with degree d .”

related results derived multiple times: Sauer, Shelah, Perles, Vapnik/Chervonenkis

Finding the VC-dimension

- We defined VC-dimension of \mathcal{H} , $VC(\mathcal{H})$, as the size of the largest set S that \mathcal{H} can shatter
 - If \mathcal{H} can shatter arbitrarily large sets, $VC(\mathcal{H}) = \infty$
- To prove that $VC(\mathcal{H}) = d$, need to show
 1. \exists some set of d data points that \mathcal{H} can shatter and
 2. \nexists a set of $d + 1$ data points that \mathcal{H} can shatter

VC-dimension example

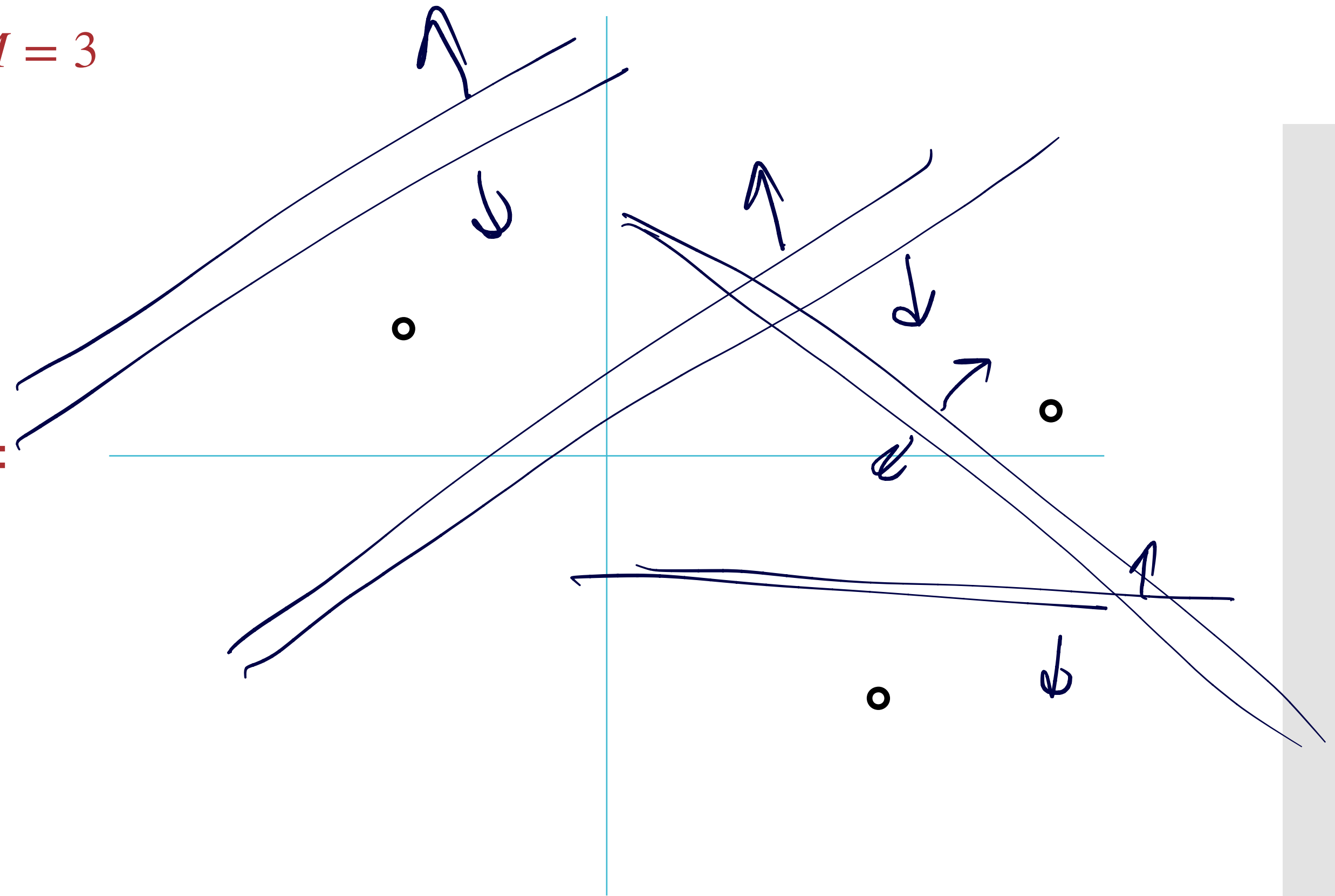


\mathcal{H} = halfspaces

i.e., linear separators

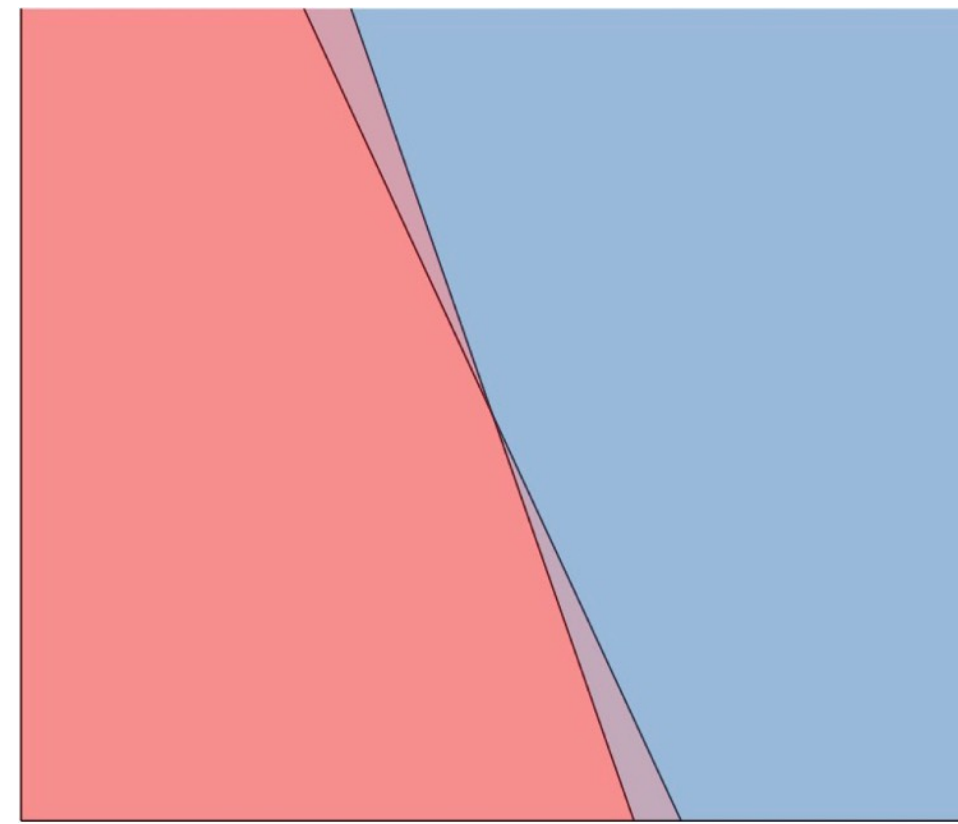
$M = 3$

$\mathcal{D} =$



- Can shatter a dataset of size 3

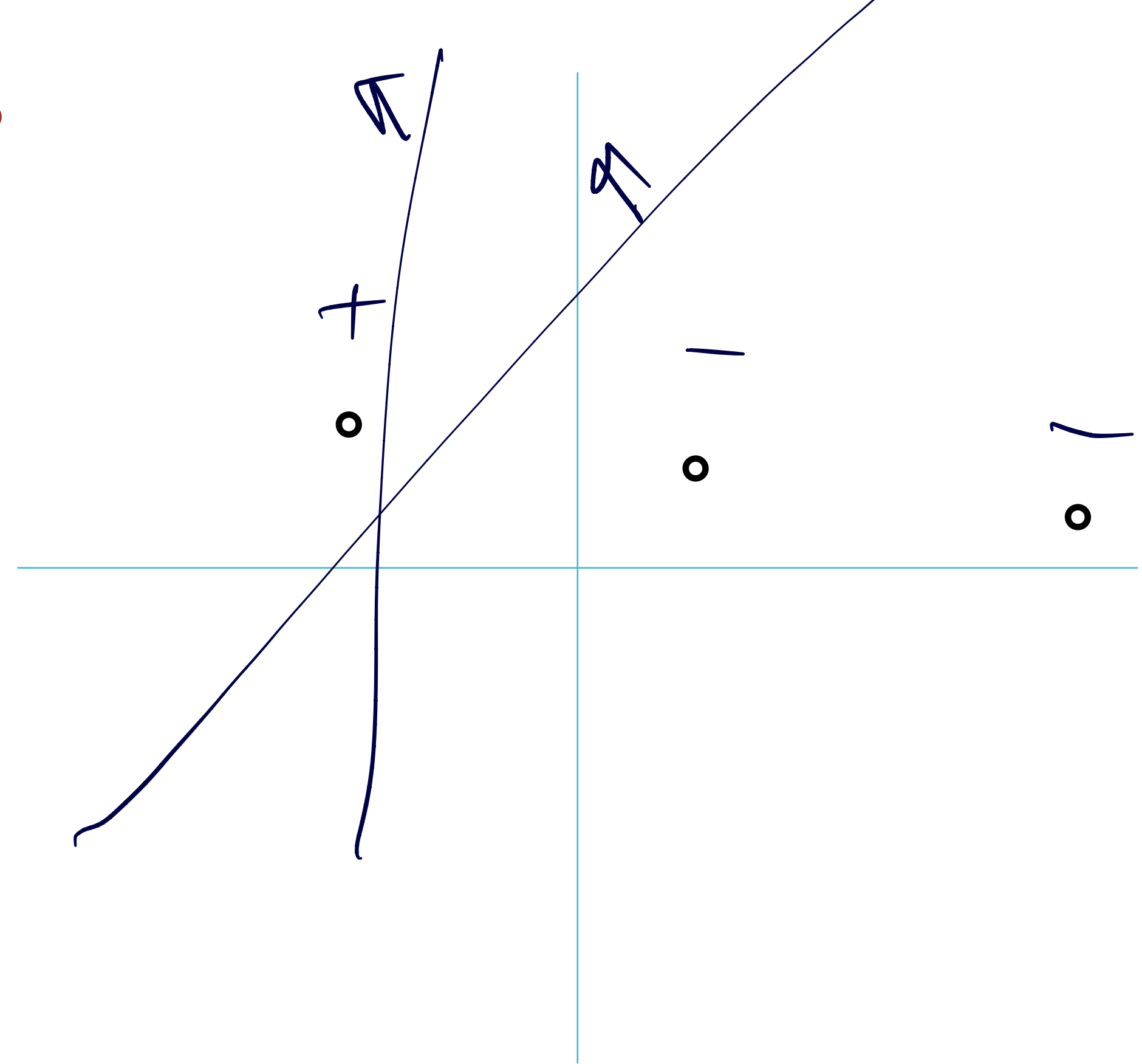
VC-dimension example



\mathcal{H} = halfspaces

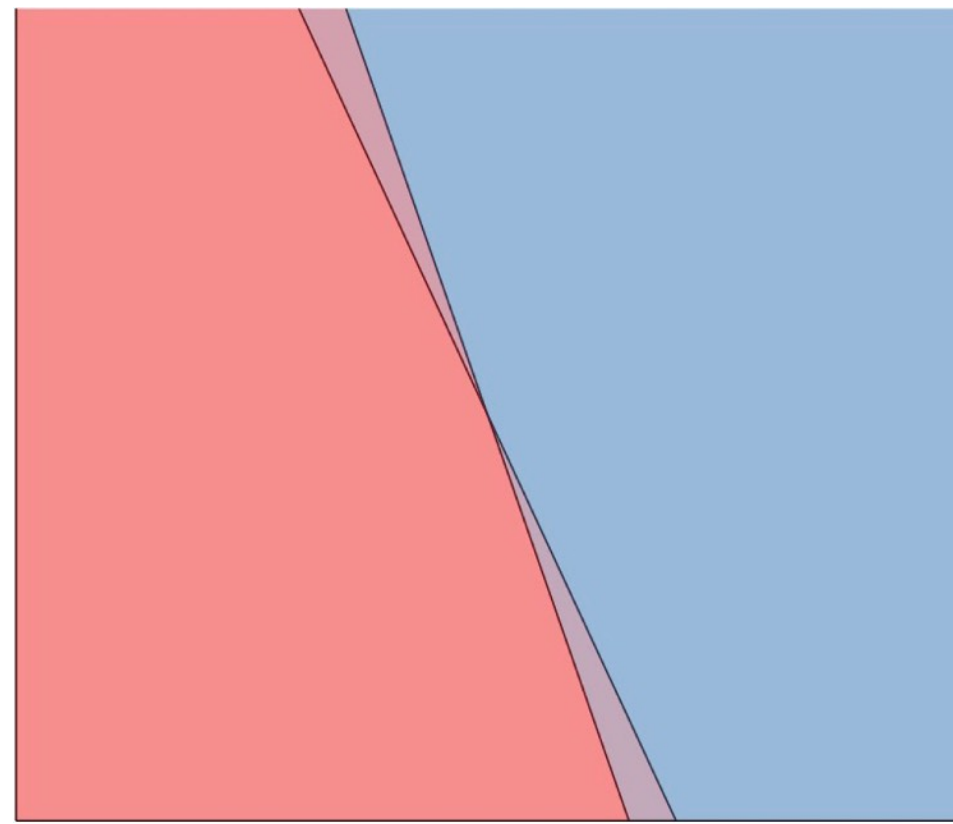
$$M = 3$$

$$\mathcal{D} =$$



- But not this one — only 6 distinct hypotheses $< 2^3$
- Tempting to say $VC(\mathcal{H}) < 3$ — but would be *wrong*

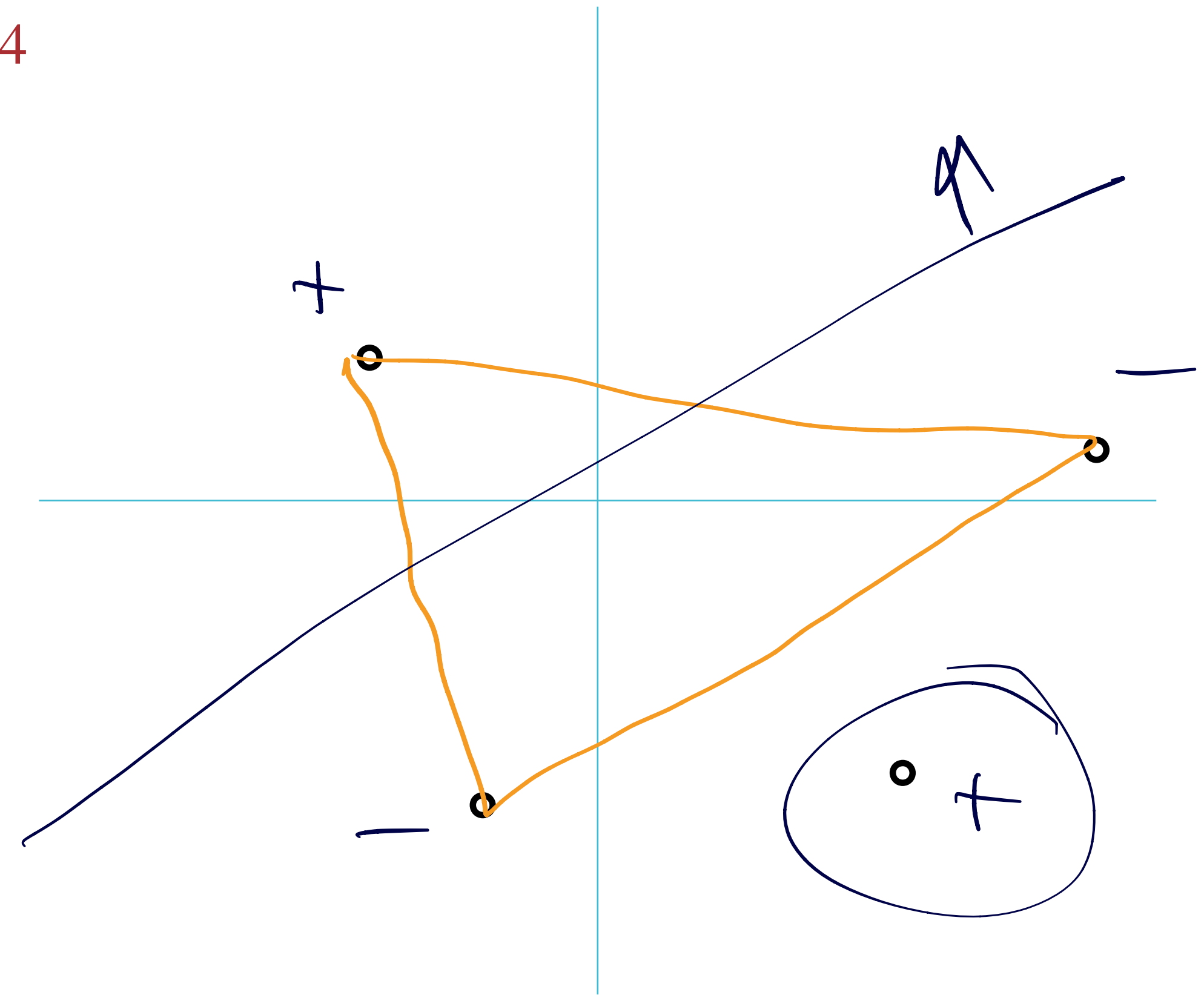
VC-dimension example



\mathcal{H} = halfspaces

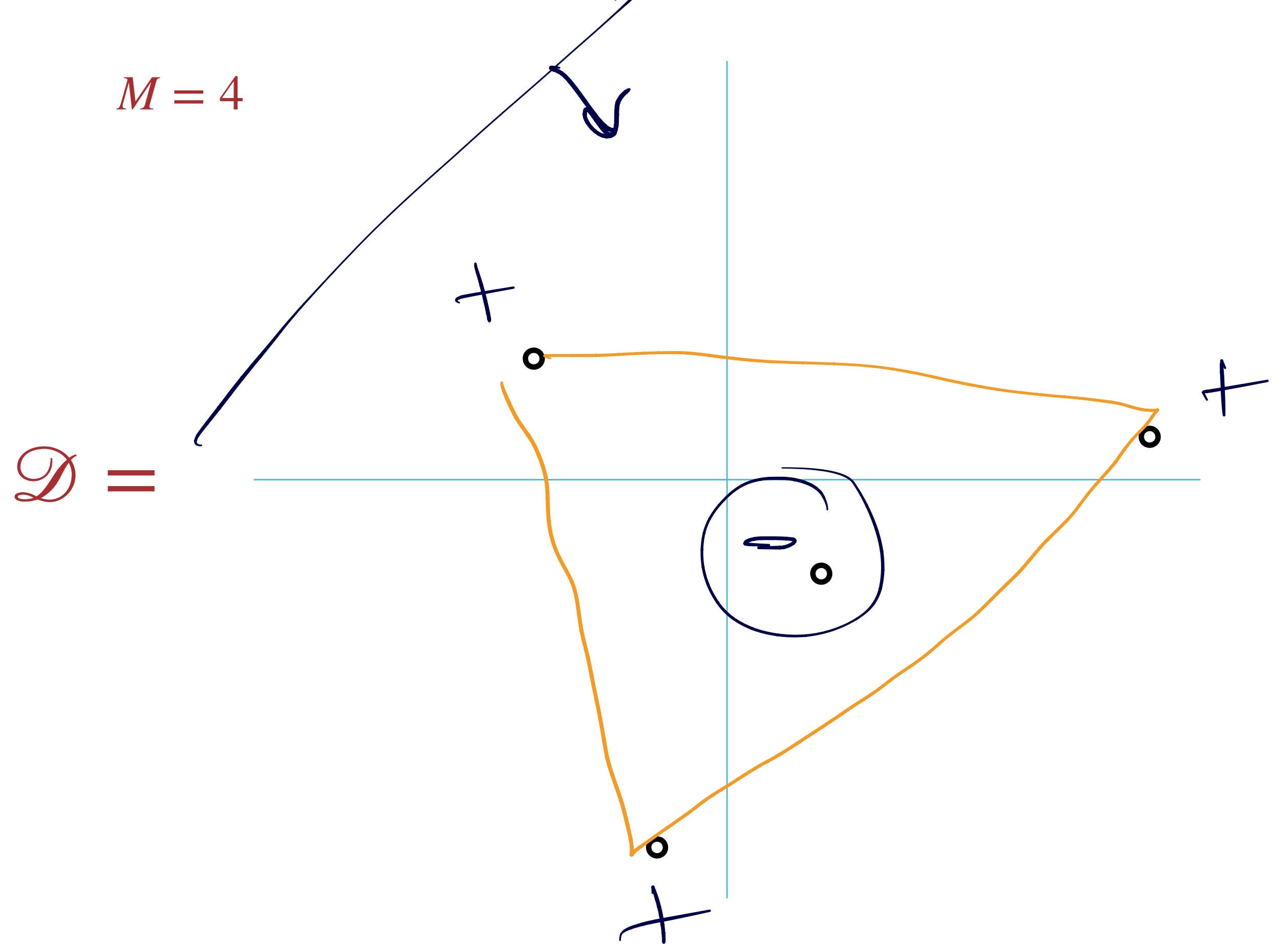
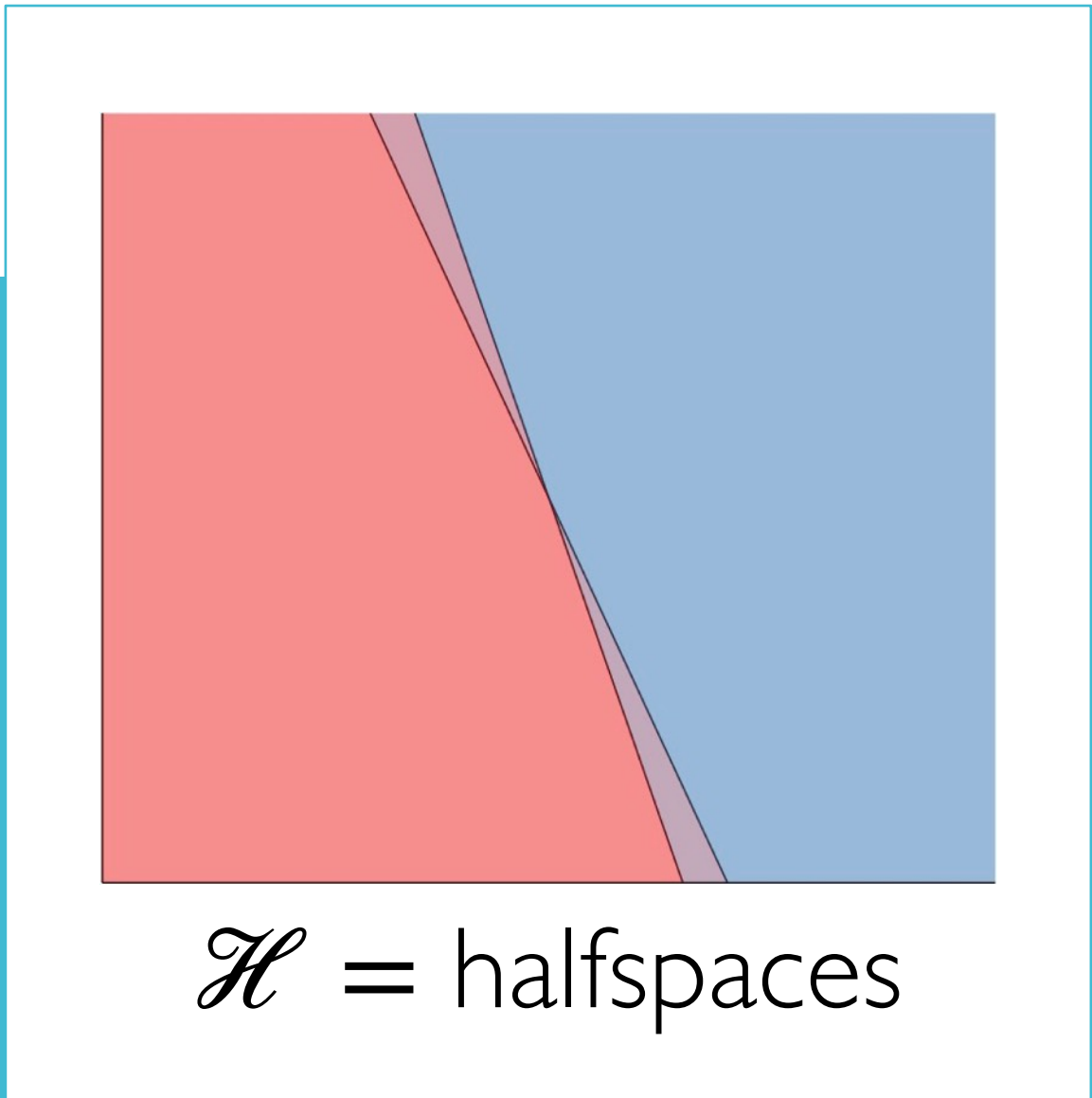
$M = 4$

$\mathcal{D} =$

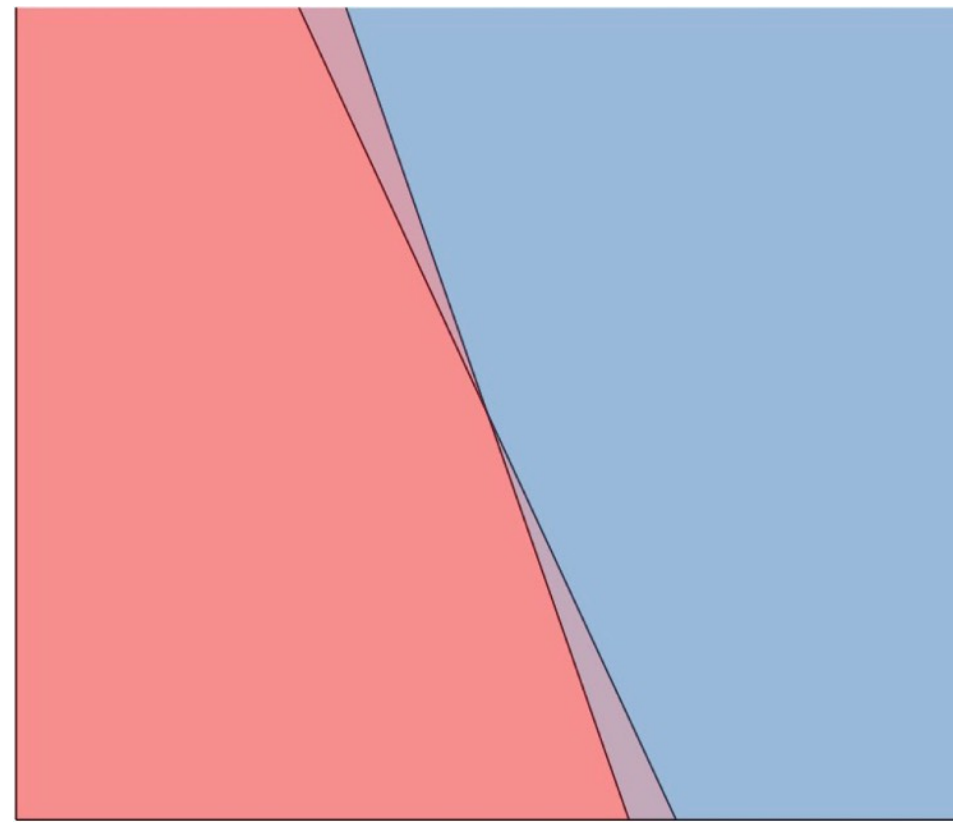


- Can't shatter this dataset of size 4

VC-dimension example



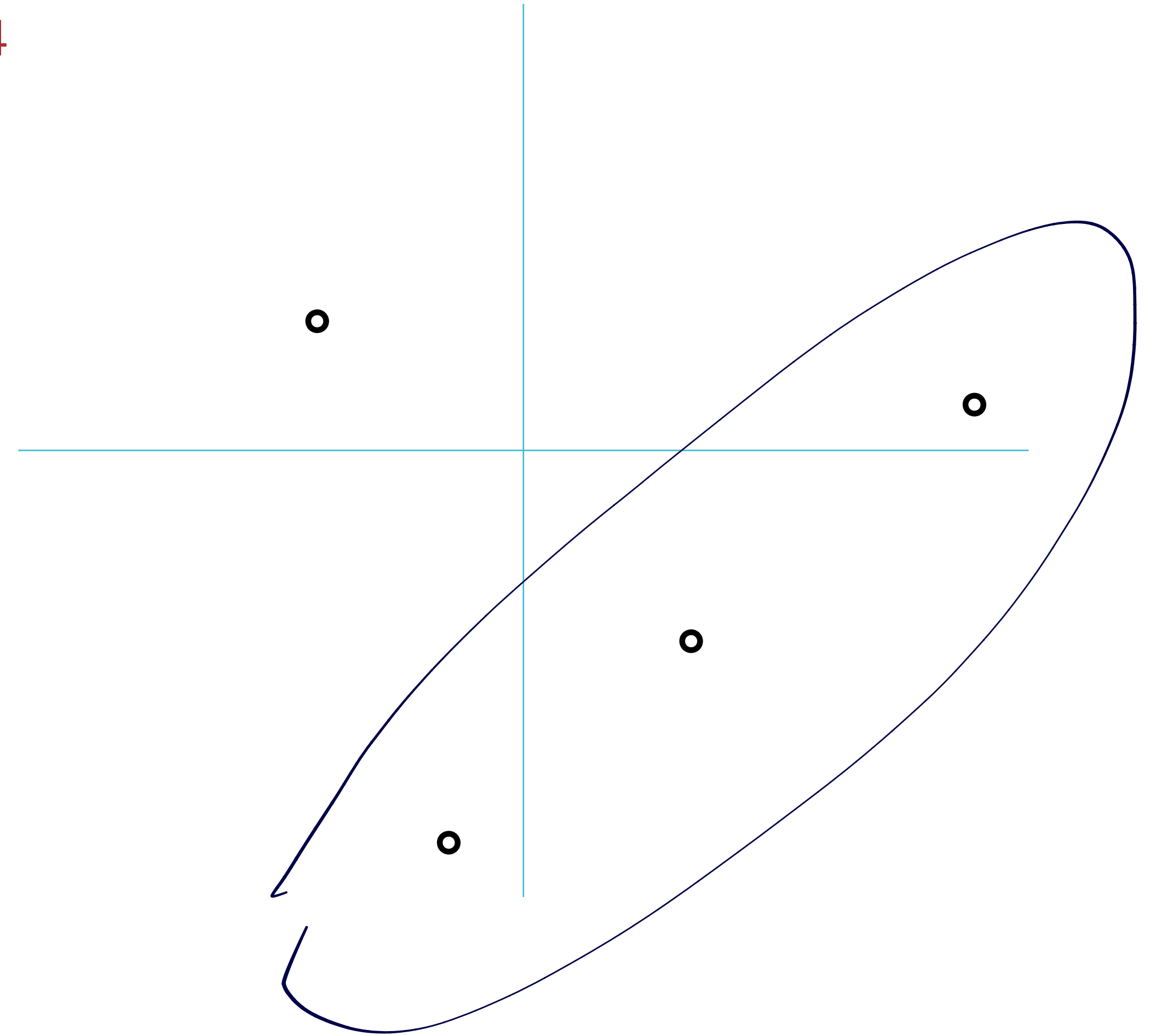
- Nor this one



\mathcal{H} = halfspaces

$$M = 4$$

$$\mathcal{D} =$$



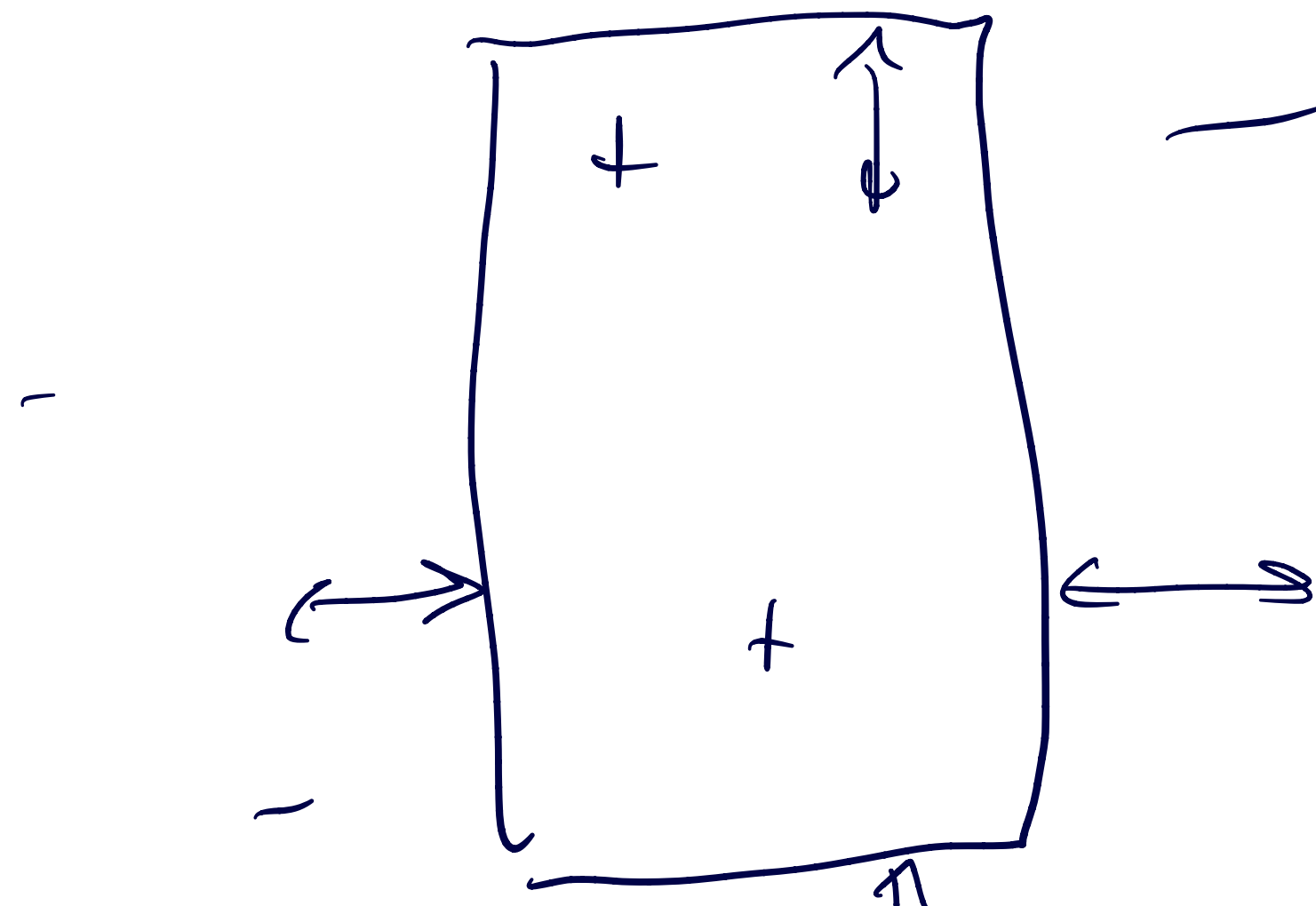
VC-dimension example

- Nor this one — and this covers all the cases

Halfspaces (linear separators)

- Just argued that halfspaces in 2D have $VC = 3$
- In general, halfspaces in d dimensions: $VC = d + 1$

More VC- dimension examples



- Try this at home: what is $VC(\mathcal{H})$ for
 - \mathcal{H} = half-lines where positive class is on right
 - \mathcal{H} = real intervals, positive when $x \in (a, b)$
 - \mathcal{H} = axis-parallel rectangles in 2D (+ on interior)

Learning objectives

- You should be able to...
 - Identify properties of a learning setting, assumptions needed to ensure low generalization error
 - Distinguish true error, train error (and test, validation errors)
 - Define PAC: what is approximately correct and what occurs with high probability
 - Apply sample complexity bounds to real-world learning examples
 - Theoretically motivate regularization

Probabilistic machine learning

- Start from *generative story* of our data
 - ▶ e.g., first step $x^{(i)} \sim N(\mu, I)$ where $\mu \in \mathbb{R}^d$
 - ▶ ... then second step $y^{(i)} \sim N(w^\top x^{(i)} + b, 0.1^2)$
 - ▶ or maybe instead $x_j^{(i)} \sim \text{Bernoulli}(\pi_j)$ so $x^{(i)} \in \{0,1\}^d$
 - ▶ or maybe instead $x^{(i)} \sim$ an arbitrary unknown dist'n
 - ▶ or maybe instead $y^{(i)} \sim \text{Bernoulli}(\sigma(w^\top x^{(i)} + b))$
- Each step in the story has *parameters* (e.g., μ, π_j, w, b , the arbitrary distribution of $x^{(i)}$)
 - ▶ our goal is to learn (some of) the parameters

Likelihood

- Collect a training dataset, and use generative story to write out *likelihood*: $P(\text{data} \mid \text{parameters})$, considered as a function of **parameters** (hold data fixed)
 - ▶ or maybe *conditional likelihood*:
 $P(\text{some data} \mid \text{rest of data, parameters})$
 - ▶ useful if it lets us skip learning some of the parameters
- E.g., if $x^{(i)} \sim N(\mu, I)$ (with no $y^{(i)}$),
likelihood = $\prod_{i=1}^M \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\|x-\mu\|^2}$
- E.g., if $y^{(i)} \sim \text{Bernoulli}(\sigma(w^\top x^{(i)} + b))$,
conditional likelihood = $\prod_{i=1}^M p_i^{y^{(i)}} (1 - p_i)^{(1-y^{(i)})}$ where
 $p_i = \sigma(w^\top x^{(i)} + b)$

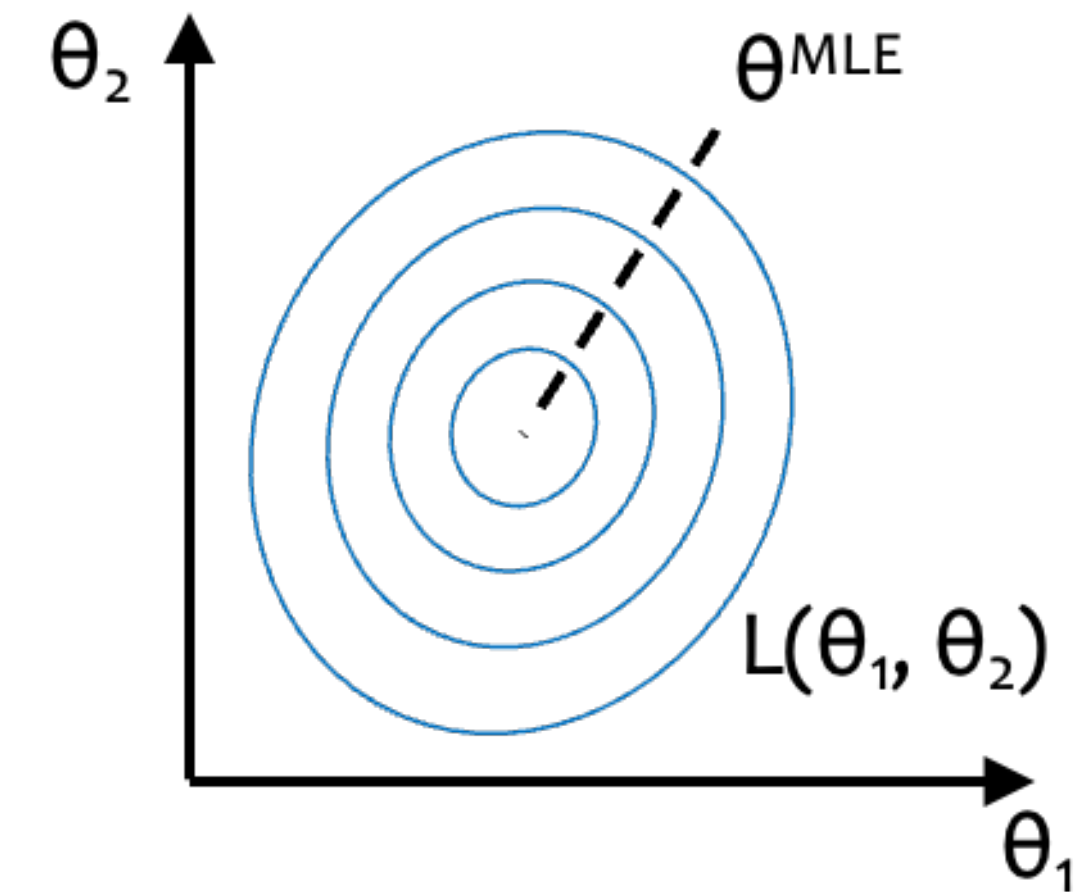
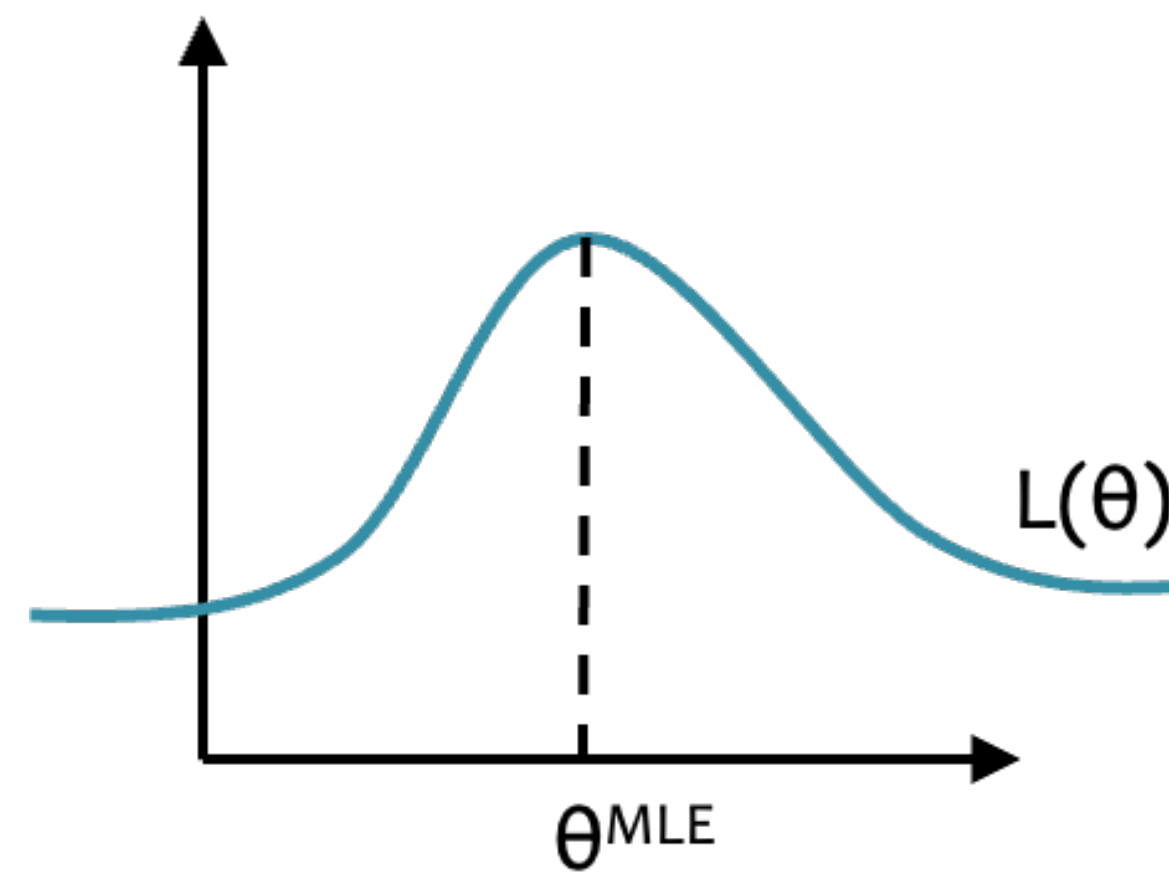
Estimation

- Finally, use the likelihood function to find the best possible parameters
 - ▶ so far, maximum likelihood estimation

MLE

Principle of *Maximum Likelihood Estimation*:
Choose the parameters that maximize the
likelihood of the data

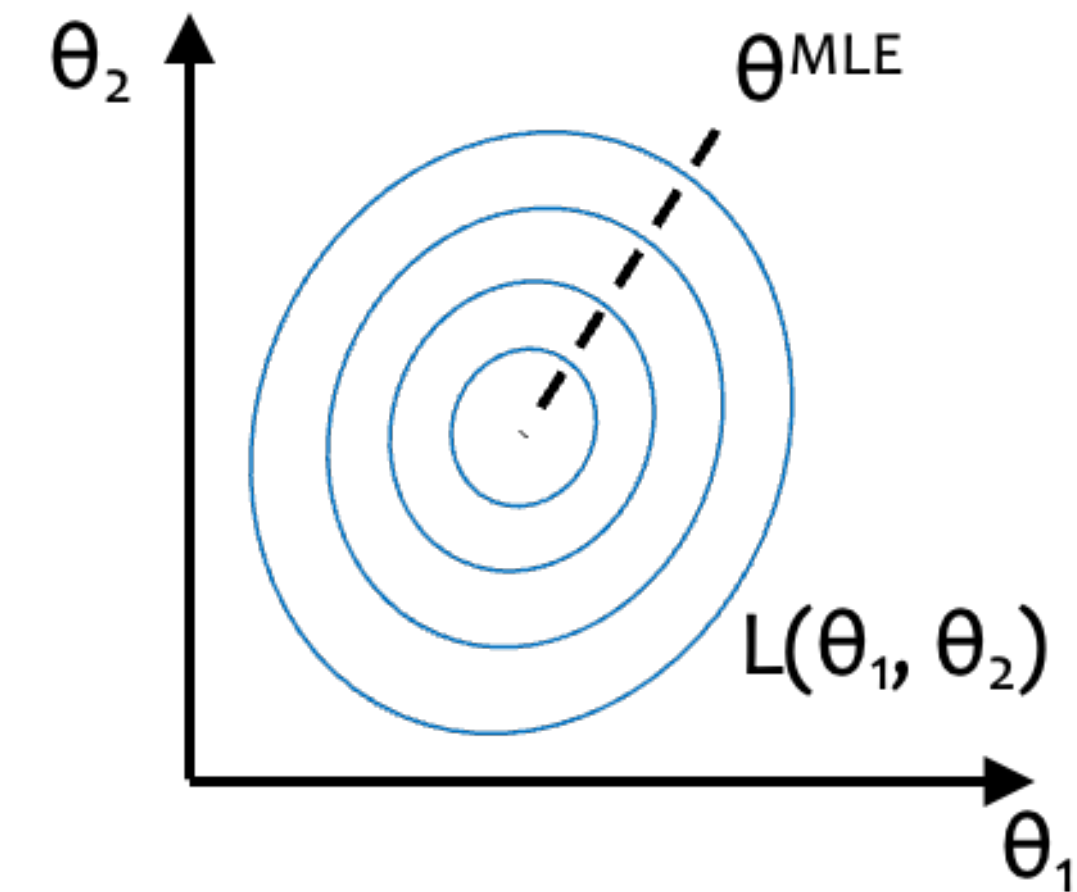
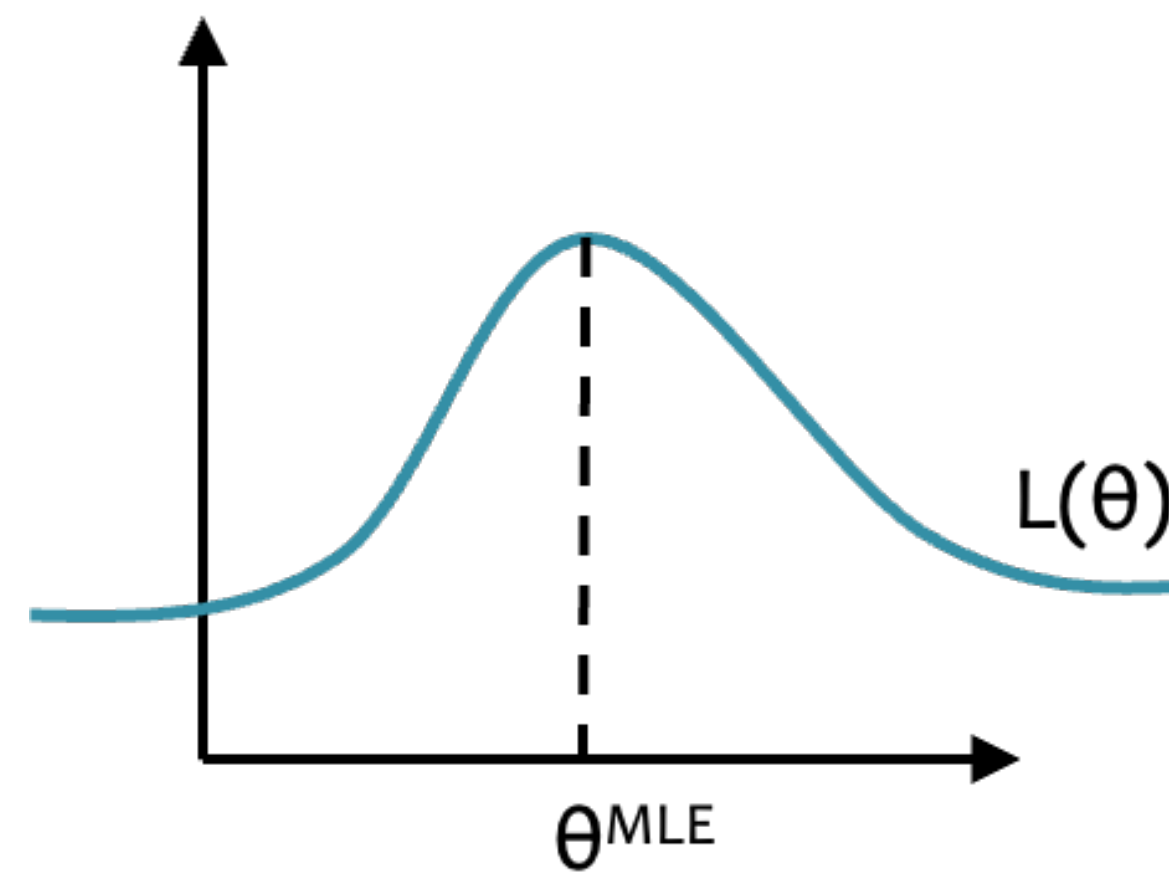
$$\theta^{\text{MLE}} = \arg \max_{\theta} \prod_{i=1}^M p(\mathbf{x}^{(i)} | \theta)$$



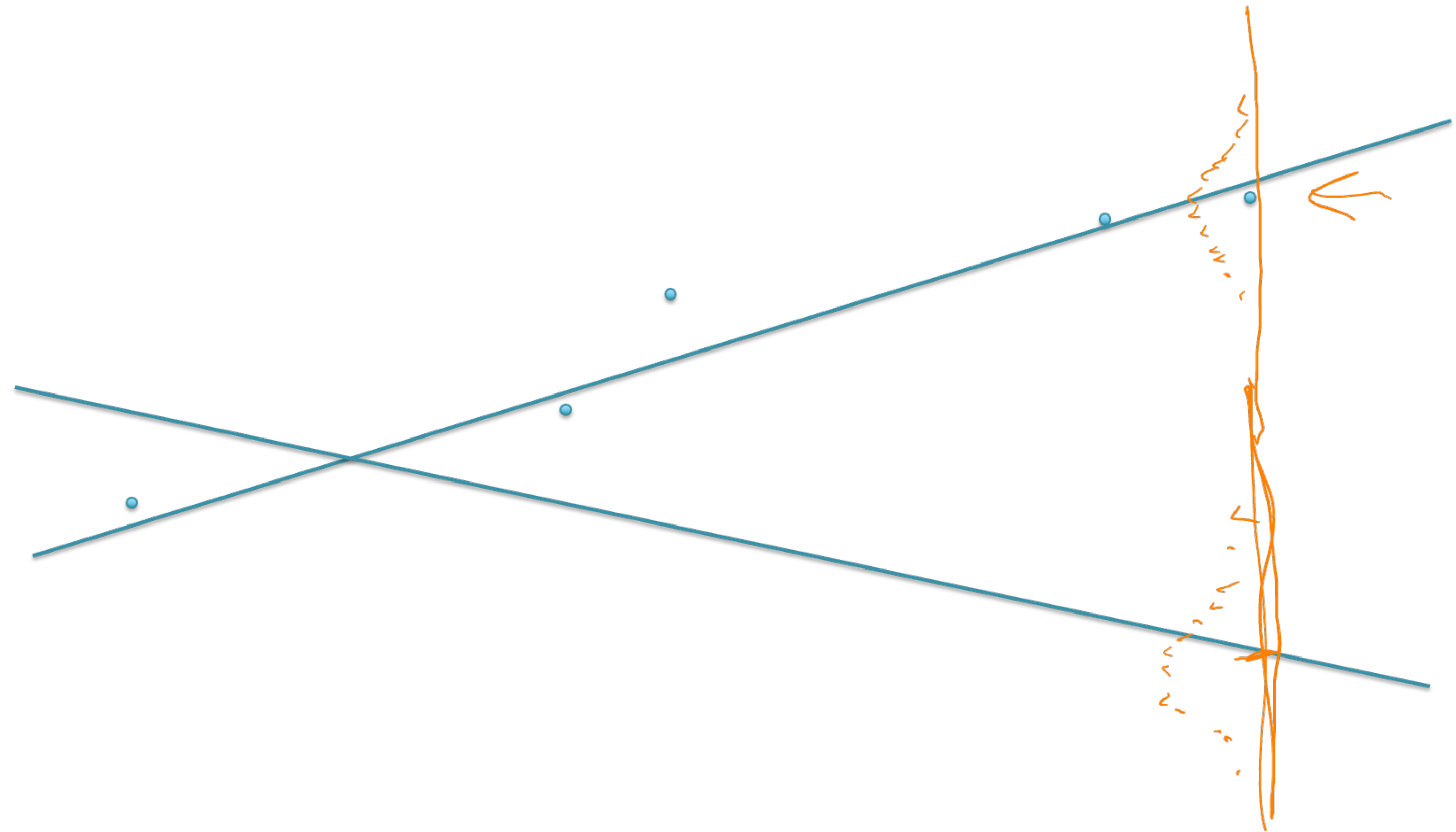
MLE

Principle of *Maximum Likelihood Estimation*:
Choose the parameters that maximize the
conditional likelihood of the data

$$\theta^{\text{MLE}} = \arg \max_{\theta} \prod_{i=1}^M p(\mathbf{x}^{(i)} | y^{(i)}, \theta)$$



Why MLE?



- There's only a fixed amount of total probability (sum-to-one constraint)
- MLE pushes as much probability as possible to things we've observed
 - ▶ ...and steals it from things we didn't observe

Another possibility

- MLE is great, but we'll look at another strategy today: maximum a posteriori estimation

Probability reminders



- Three common ways to describe probability:
 - ▶ probability distribution
 - ▶ probability density
 - ▶ measure


Distribution

- *Idea: assign numbers to atomic events so that, of times we assign (say) 0.23, the event happens 23% of them*
- A probability distribution function p maps discrete atomic events to real numbers, with
 - ▶ $p(a) \geq 0$ for all atomic events a
 - ▶ $\sum_{a \in U} p(a) = 1$ summed over universe U (entire table)
- E.g., uniform on n atoms: $p(a) = \frac{1}{n}$ for each
- Notation: $p_X, p_Y, p_{X,Z}$ — but often skip subscripts

| $Z \rightarrow$ | 0 | 1 |
|-------------------|-----|-----|
| $X, Y \downarrow$ | | |
| 00 | 0.1 | 0.2 |
| 01 | 0.2 | 0.0 |
| 10 | 0.1 | 0.3 |
| 11 | 0.1 | 0.0 |

also called pmf

atomic event = smallest
division of outcomes

e.g.,  = 2, not

 \in {even #s}

*Idea: assign numbers to
atomic events so that, of times
we assign (say) 0.23, the
event happens 23% of them*

Distribution

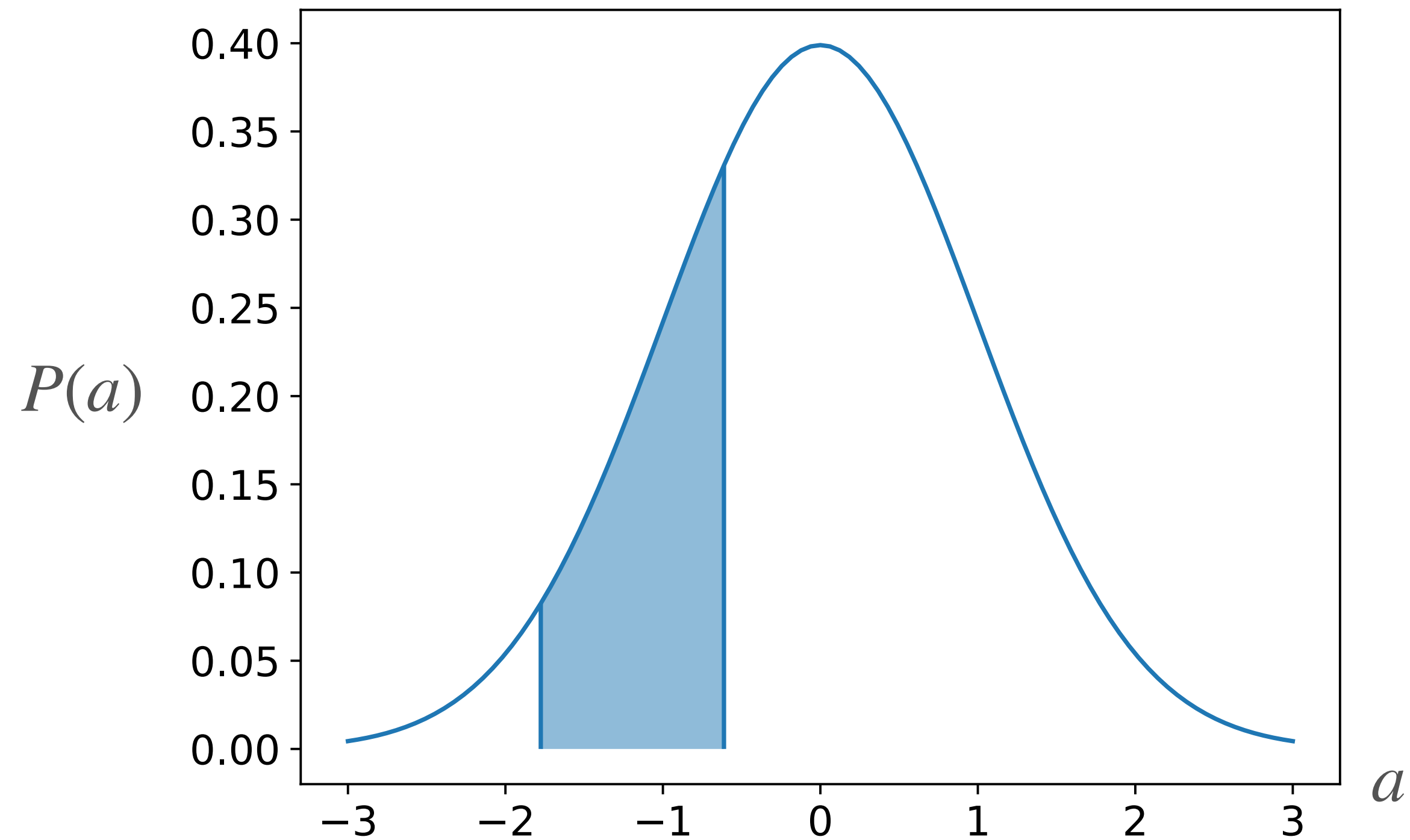
- A probability distribution function p maps discrete atomic events to real numbers, with
 - ▶ $p(a) \geq 0$ for all atomic events a
 - ▶ $\sum_{a \in U} p(a) = 1$ summed over universe U (entire table)
- E.g., uniform on n atoms: $p(a) = \frac{1}{n}$ for each
- Notation: $p_X, p_Y, p_{X,Z}$ — but often skip subscripts

$P(X, Y, Z)$

| $Z \rightarrow$ | 0 | 1 |
|-------------------|-----|-----|
| $X, Y \downarrow$ | | |
| 00 | 0.1 | 0.2 |
| 01 | 0.2 | 0.0 |
| 10 | 0.1 | 0.3 |
| 11 | 0.1 | 0.0 |

also called pmf

Density

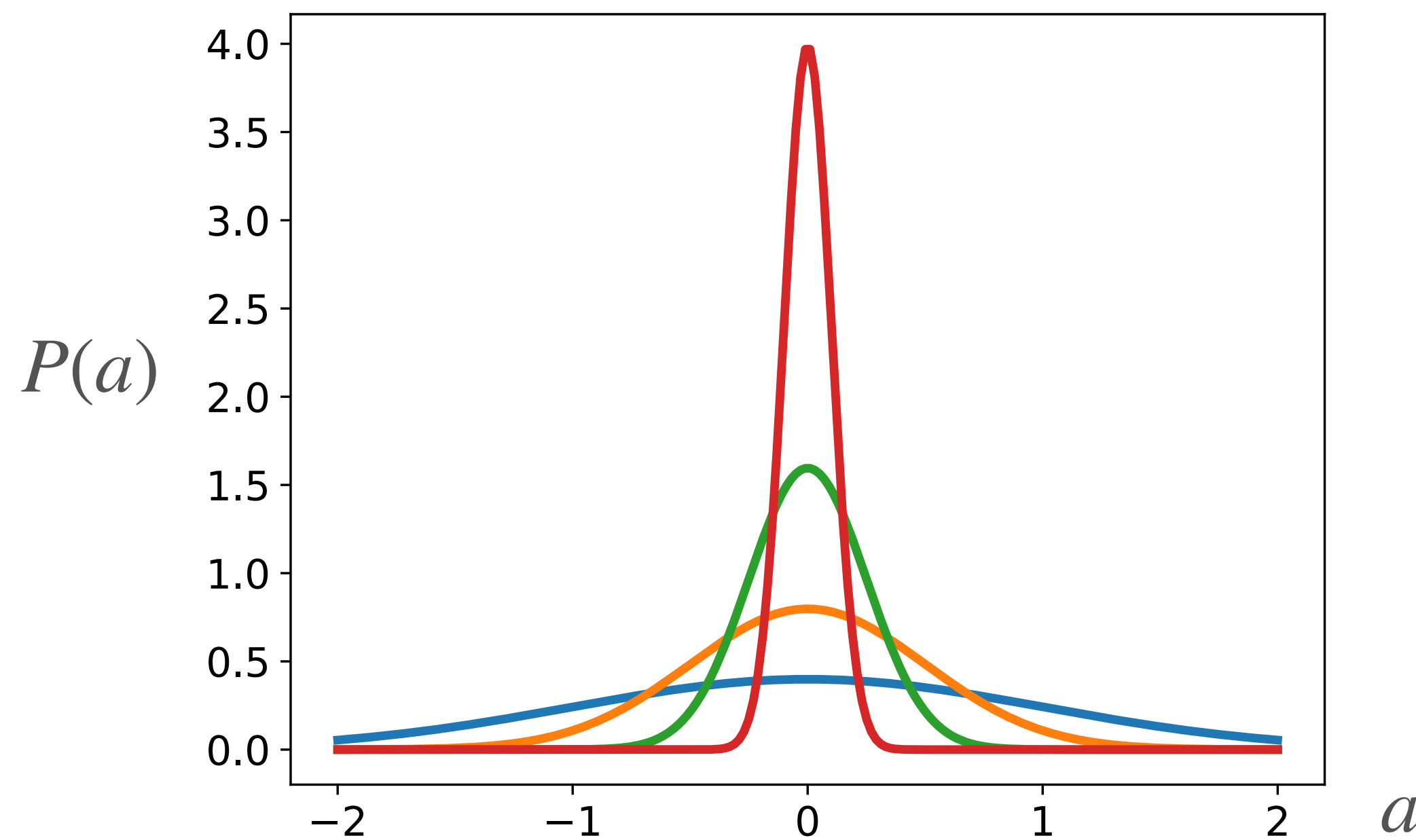


- *Idea: assign numbers to atomic events, so that integrating over atomic events yields a probability*
- A probability density function p maps continuous atomic events to real numbers, with
 - ▶ $p(a) \geq 0$ for all atomic events a
 - ▶ $\int_{a \in U} p(a) da = 1$ over universe U

PDF

- Some statisticians thought they were clever making PDF be the acronym for both
 - ▶ and we use $p(\cdot)$ or $P(\cdot)$ for either
 - ▶ and say distribution even if we mean density (or measure)
 - ▶ worse, can see different uses in the same equation (!)
- Makes some sense since a lot of the same rules happen to apply to both representations
- But they are *not the same thing!*

***Not the
same thing***



- E.g., for a density, $p(0.37) \neq$ probability of seeing 0.37
 - ▶ in fact, probability of any atomic event = 0!
- For a discrete distribution, $p(a) \leq 1$ for each outcome (atomic event) a , but for a continuous density, $p(a)$ can be arbitrarily large
- Density acts like the *derivative* of probability

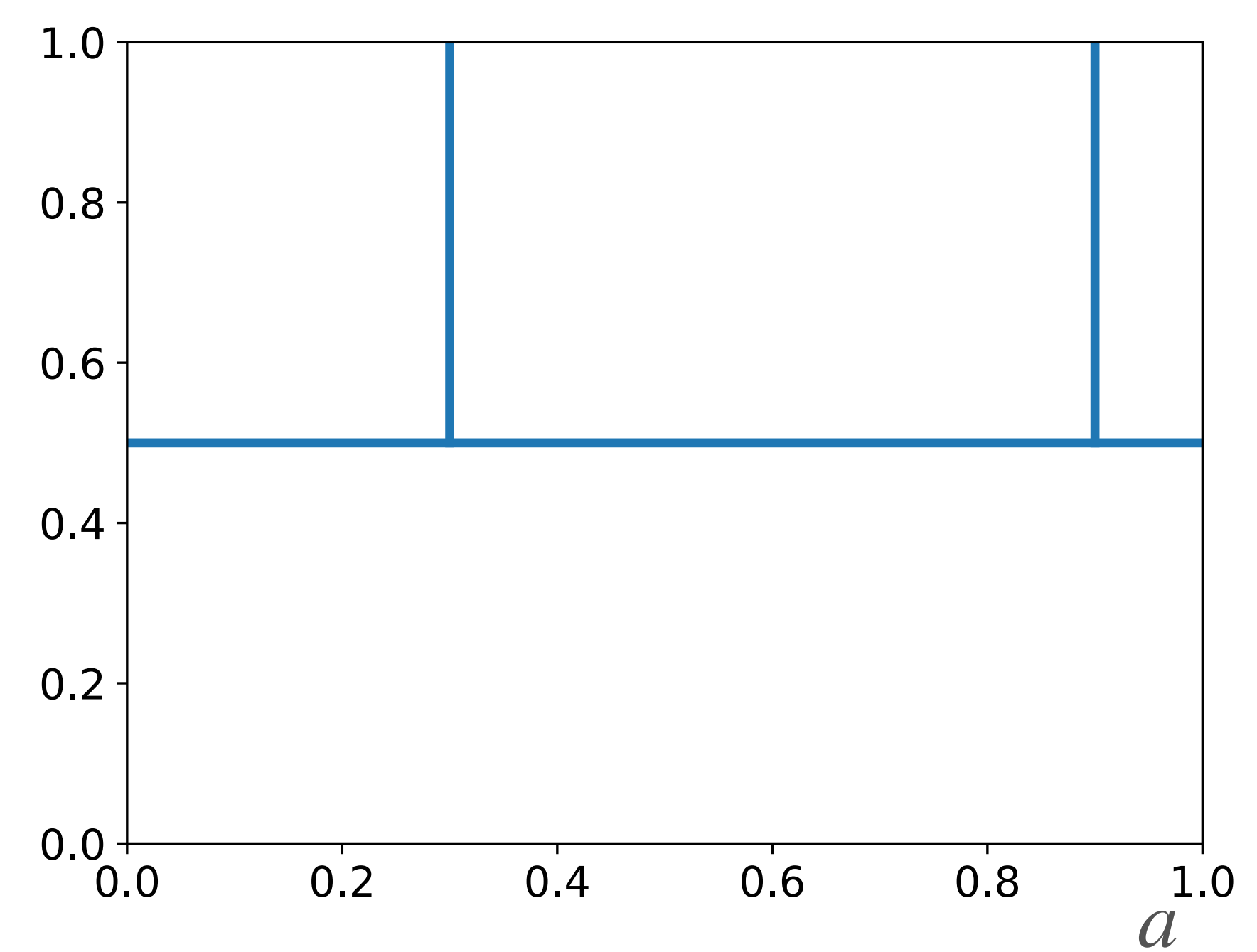
$$p(E) = \int_{a \in E} p(a) da \text{ for an event } E$$



Measure

won't typically use this
in 10-301/601

$P(a)$



- *Idea: directly define the result of integration, enforcing only rules of nonnegativity, subadditivity, and $P(U) = 1$*
- Generalizes both discrete and continuous PDFs
- Notation is like continuous: $p(E) = \int_{a \in E} \mu(a) da$
- But OK to have atomic events with $p(a) > 0$
 - ▶ imagine a Gaussian bump that's infinitely narrow centered at a (for physicists among us: a Dirac delta)
 - ▶ e.g., $p(x) = 0.5 + 0.25 \delta(x - 0.3) + 0.25 \delta(x - 0.9)$
 $x \in [0, 1]$

Conditional probability

$$p(z|x,y)$$

~~$$p(x,y|z)$$~~

| $Z \rightarrow$ | 0 | 1 |
|-------------------|-----|-----|
| $X, Y \downarrow$ | | |
| 00 | 0.8 | 0.2 |
| 01 | 0.2 | 0.8 |
| 10 | 0.7 | 0.3 |
| 11 | 1.0 | 0.0 |

- Describe w/ *conditional* distribution, density, or measure
 - ▶ same as above, except $p(\cdot)$ takes conditioning event as another input, $p(X, Y | Z = z)$
 - ▶ sums or integrates to 1 over X, Y for each atomic event z (but not other way)
- Discrete and continuous versions, just like nonconditional (joint or marginal) probabilities
- Can think of $p(X, Y | Z = z)$ as a function of just x, y (with z fixed and compiled in), or as a function of both x, y and z

- ▶ to emphasize latter, write $p(X, Y | Z)$

$$p(z|x,y)$$

Bayes rule

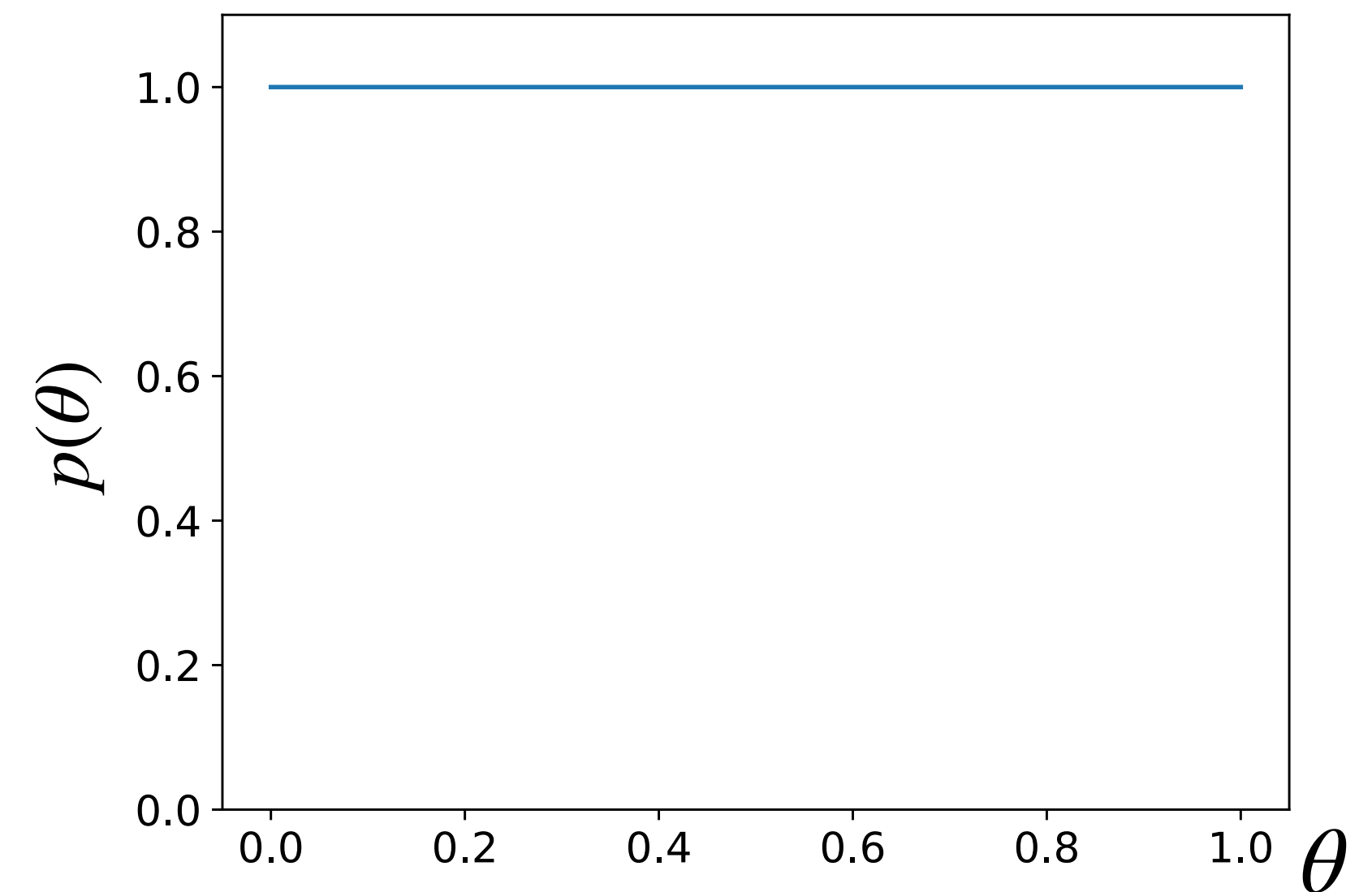
- By factorization:
 - ▶ $p(X) p(Y | X) = p(X, Y) = p(Y) p(X | Y)$
- Divide through by $p(X)$:
 - ▶ $p(Y | X) = p(X | Y) p(Y) / p(X)$ ← *Bayes rule*

Working with probabilities

- You should know how to work with probabilities — readings / OH / Piazza can help if you're rusty
- Discrete and continuous random variables, events; marginal, joint, and conditional probabilities; expectation and conditional expectation; mean and variance; independence and conditional independence
- Familiarity with common distributions: e.g., Gaussian, Bernoulli, multinomial
- Manipulating probability tables or continuous densities
 - ▶ including conditioning and conditional tables/densities
- Most important tools: factorization; Bayes rule; linearity of expectations
 - ▶ also broadcasting; change of variables (Jacobian rule)

Example

- We have a biased coin: $p(\text{heads})$ is θ
- Parameter $\theta \in [0,1]$ selected uniformly
 - ▶ the above is what we know before seeing any data (*prior*)
- We now see 5 flips: X_1, X_2, \dots, X_5 are HHTHT
- None of the flips depend on any of the others (they are conditionally independent given θ)
- We want the distribution of θ after seeing $X_{1:5}$ (*posterior*)
 - ▶ terminology: prior and posterior are always relative to a given inference — often the posterior from one inference becomes the prior for the next



Bayes rule

prior, likelihood, fit

$$P(\text{HHTHT} | \theta) = \theta \cdot \theta \cdot (1 - \theta) \cdot \theta \cdot (1 - \theta)$$

$$P(\theta | \text{HHTHT}) = \frac{P(\text{HHTHT} | \theta) P(\theta)}{P(\text{HHTHT})}$$

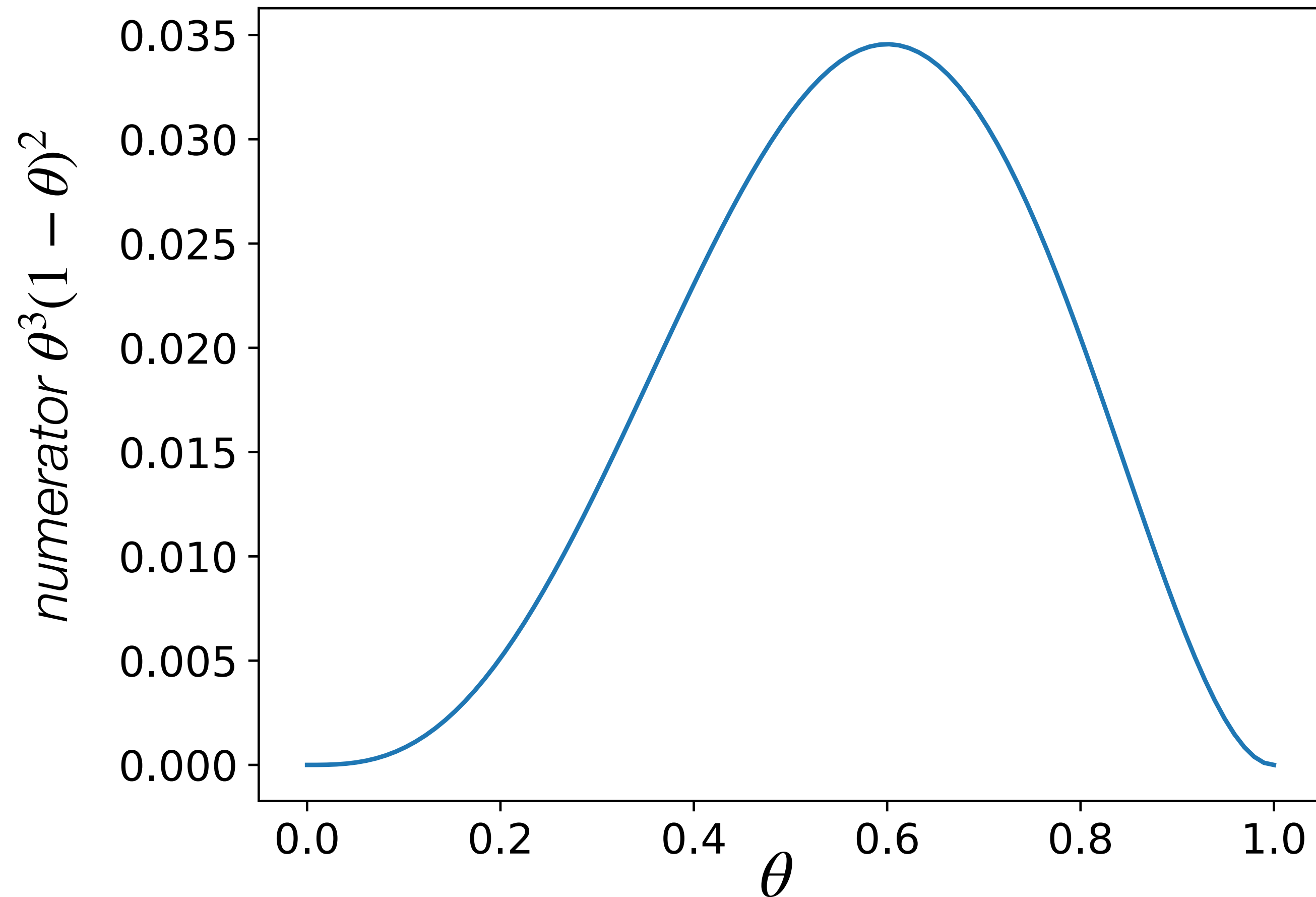
likelihood

fit

prior
uniform
= 1

$$\frac{\theta^3 (1 - \theta)^2}{Z}$$

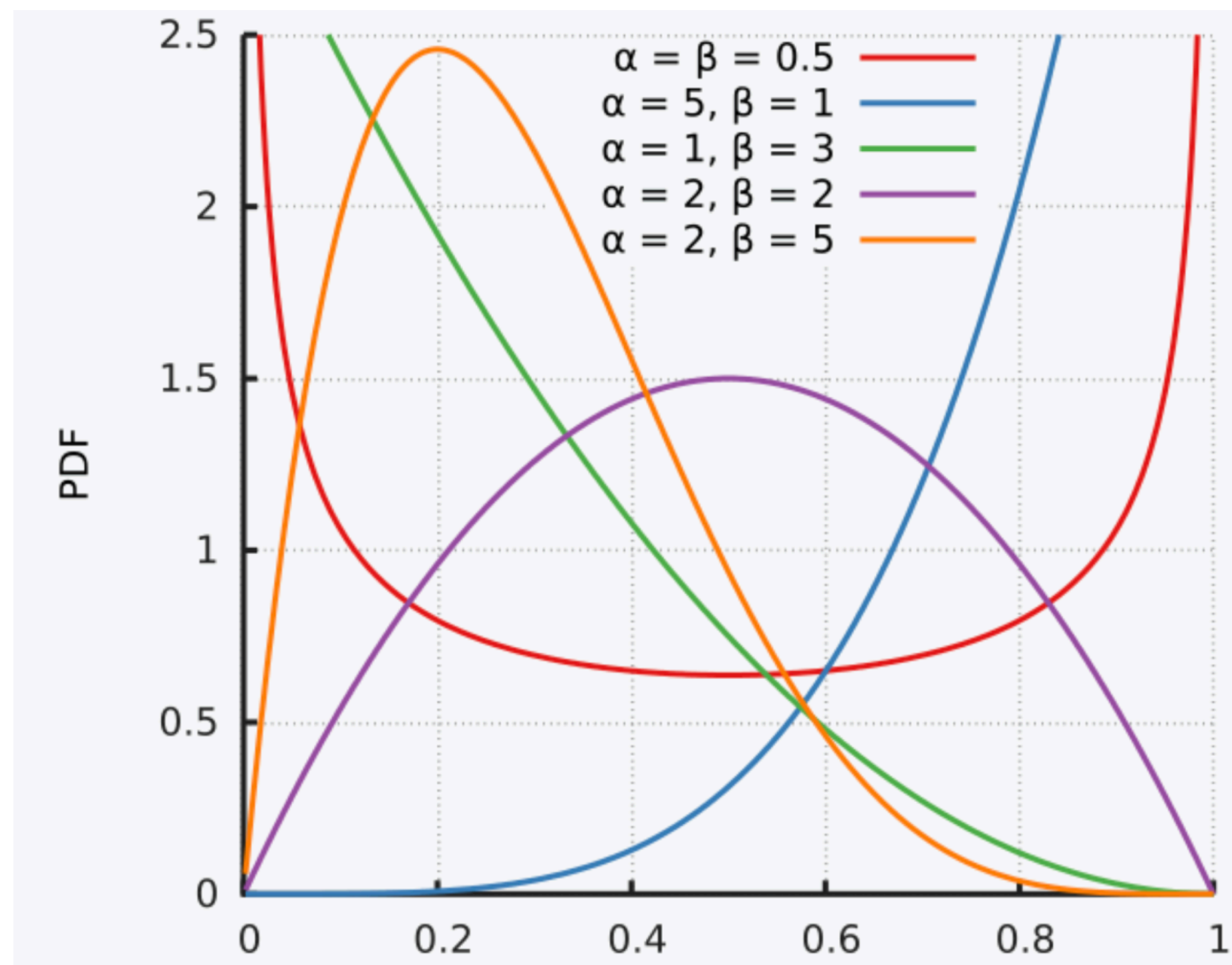
Finding the posterior



Posterior =
 $60 \theta^3 (1-\theta)^2$

- Numerator depends on θ but denominator does not
 - ▶ so, we can find denominator using constraint that the density has to integrate to 1
 - ▶ this is a common strategy: often a pain to find directly

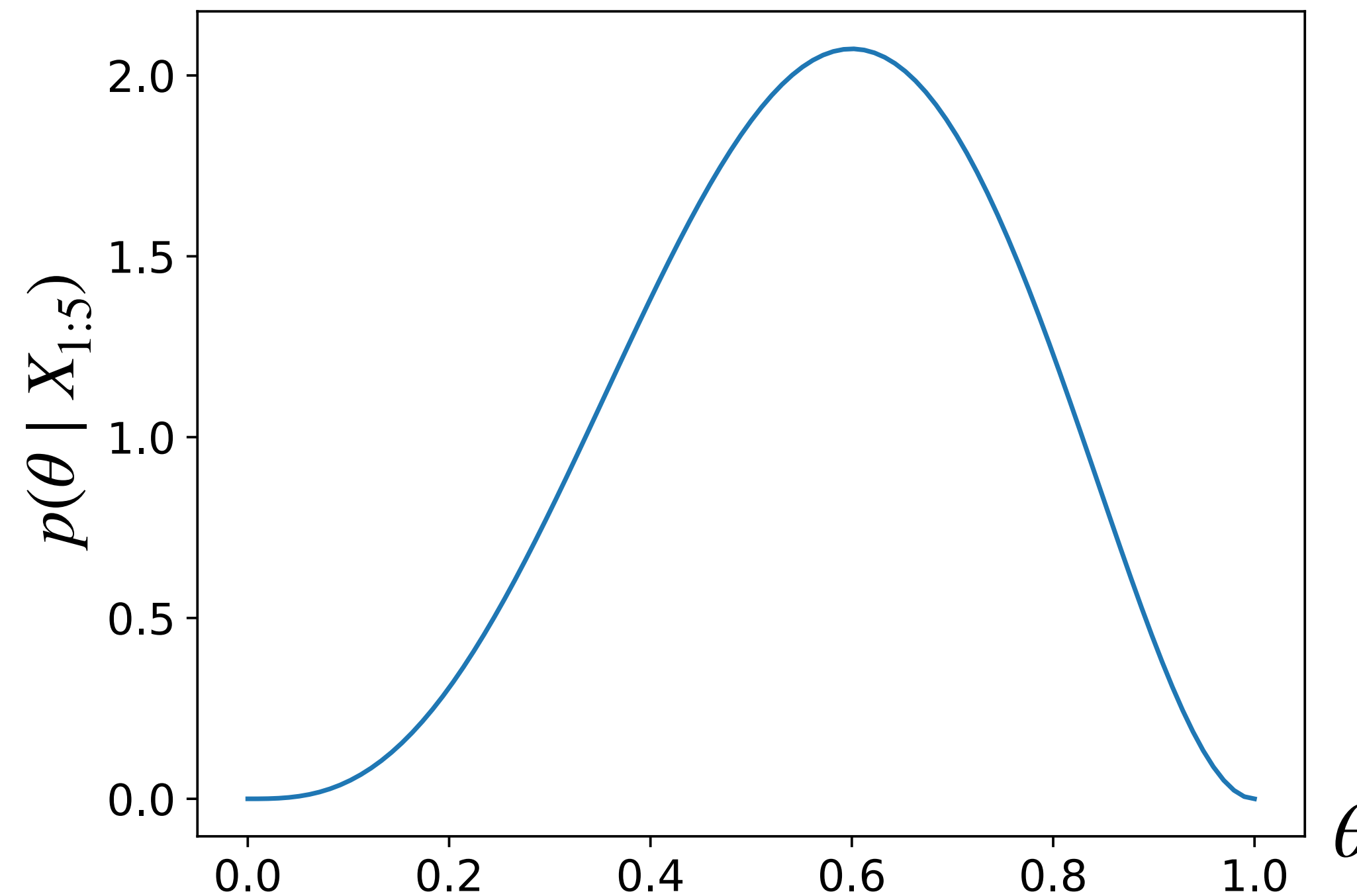
Beta



plot credit: Wikipedia

- Posterior is $P(\theta | X_{1:5}) = 60 \theta^3 (1 - \theta)^2$
- This is a **Beta** distribution
 - ▶ in general, $\propto \theta^{\alpha-1} (1 - \theta)^{\beta-1}$, parameters $\alpha, \beta > 0$:
 - ▶ α moves mass toward $\theta = 1$, β moves it toward $\theta = 0$
 - ▶ larger α, β make a sharper distribution

MAP



- Highest point of posterior (the *mode*) is $\theta = \frac{3}{5}$
 - ▶ if we have to report a single θ , this is a good one
 - ▶ called *maximum a posteriori* or *MAP* estimate
- Can also acknowledge uncertainty: e.g., report the whole distribution, or list of 100 samples from the distribution, or mean and standard deviation, or an interval that contains 95% of its probability

Finding the MAP

$$\frac{d}{d\theta} (\ln 60 + 3 \ln \theta + 2 \ln (1 - \theta))$$

$$= \frac{3}{\theta} + \frac{2}{1-\theta} (-1)$$

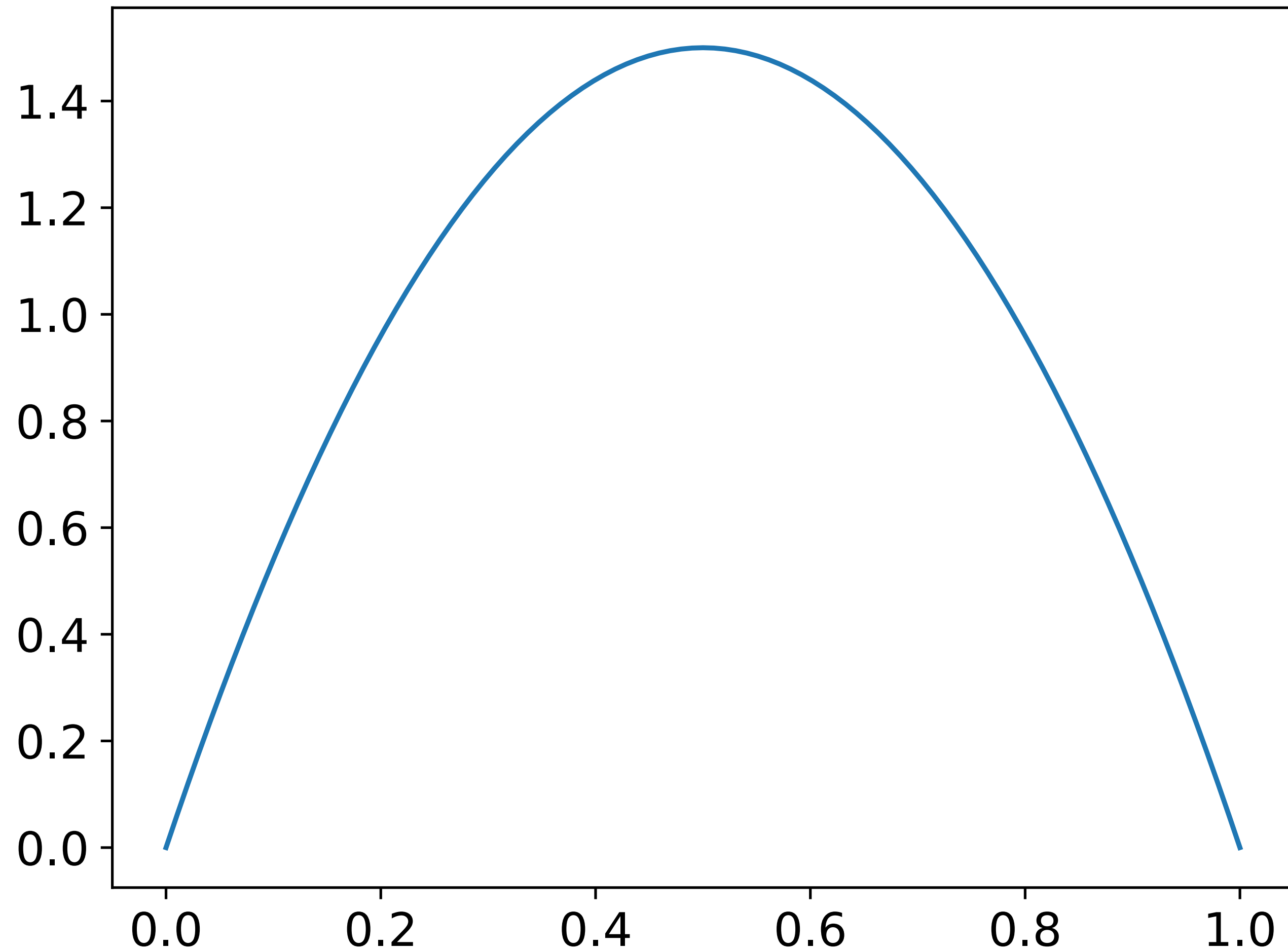
$$3(1-\theta) = 2\theta \quad 3 = 5\theta \quad \theta = \frac{3}{5}$$

- Posterior = $60 \theta^3 (1 - \theta)^2$
- To find mode, differentiate and set to 0
 - ▶ becomes easier if we take log before derivative
 - ▶ log turns *product* over examples into *sum* over examples, sum rule is simpler than product rule

Statistics and estimators

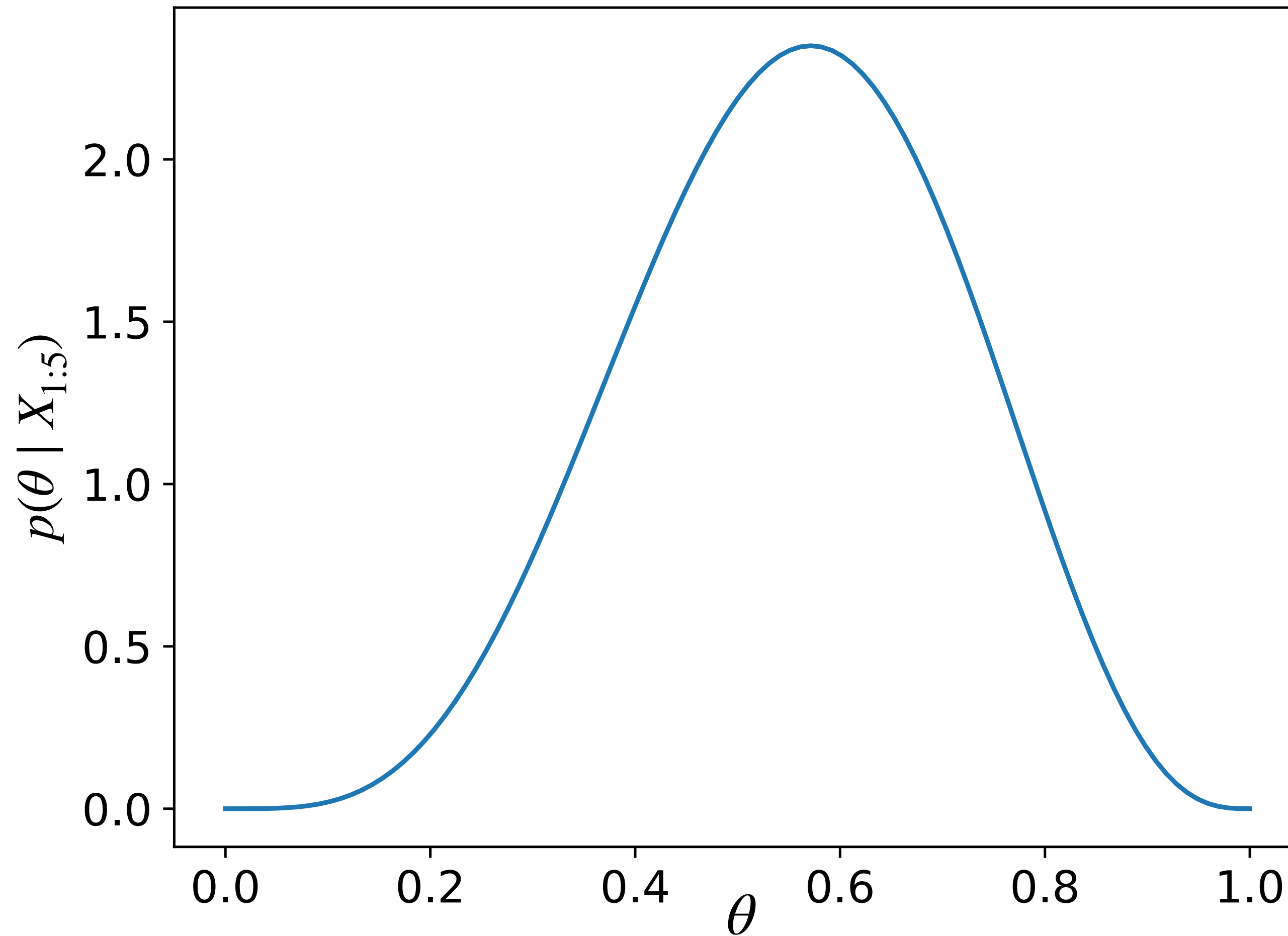
- Very reasonable answer: $\frac{3}{5}$ of our data were heads, so we say $p(H) = \frac{3}{5}$
- Observed fraction of heads is called a *statistic*: a function of our data
 - ▶ other statistics: fraction of tails; $\ln \frac{1 + \#heads}{1 + \#tails}$
- A statistic that we use to find θ is called an *estimator*
 - ▶ in this case, $\mathbb{E}(\text{observed fraction}) = \text{true } \theta$
 - ▶ so our estimator is *unbiased*
- Value of estimator is *estimate* — here $\frac{3}{5}$

***Changing
the prior
changes the
estimate***



- Prior is $p(\theta) = 6\theta(1 - \theta)$ (a Beta distribution)
- Same data: X_1, X_2, \dots, X_5 are HHTHT
- Find posterior and MAP for θ after seeing $X_{1:5}$

New MAP



→ $P(\theta | X_{1:5}) = 280 \theta^4 (1 - \theta)^3$ mode is $\theta = \frac{4}{7}$

Influence of prior

- Same data, two priors → two different estimates of θ
 - ▶ 2nd prior was narrower
 - ▶ pulls estimate toward its center at $\frac{1}{2}$
 - ▶ uniform prior in a sense pays more attention to the data (likelihood term), doesn't favor one θ over another

Conjugate priors

- Something interesting happened
 - ▶ prior was uniform = Beta(1, 1), posterior was Beta(3, 4)^{4, 3}
 - ▶ prior was Beta(2, 2), posterior was Beta(4, 5)^{5, 4}
- Not an accident: Beta is the **conjugate prior** for Bernoulli
 - ▶ i.e., when parameter θ has a Beta prior, and when data is Bernoulli(θ), then posterior for θ will also be Beta
- Happens somewhat often:
 - ▶ $x \sim N(\mu, \sigma^2)$: conjugate prior for μ is Normal
 - ▶ $x \sim \text{Exponential}(\lambda)$: conjugate prior for λ is Gamma
- Useful since it simplifies inference:
 - ▶ e.g., observe h Heads and t Tails: add h to α and t to β

Maximum likelihood

- C.f. maximum likelihood estimation (MLE):
 - ▶ in posterior $p(\theta) p(X_{1:5} | \theta) / p(X_{1:5})$, MAP ignored denominator (can recover by normalizing)
 - ▶ MLE also ignores prior $p(\theta)$, leaving only likelihood
$$\arg \max_{\theta} p(X_{1:5} | \theta)$$
$$= \arg \max_{\theta} p(X_1 | \theta) p(X_2 | \theta) \dots p(X_5 | \theta)$$
$$= \arg \max_{\theta} [\ln p(X_1 | \theta) + \dots + \ln p(X_5 | \theta)]$$
- Like setting prior to uniform, but works even when uniform prior doesn't make sense

MLE vs. MAP

- Both are *consistent*:
 - ▶ under weak assumptions, parameter \rightarrow true value as $|\text{data}| \rightarrow \infty$
- Different advantages:
 - ▶ MLE: no need to think about prior
 - ▶ MAP: a reasonable prior can often make estimates more reliably accurate by reducing variance
 - ▶ helps w/ low data
 - ▶ which can be a problem even in huge datasets (!)

Another MAP example

- New distribution: Exponential(λ)
 - ▶ PDF is $\lambda e^{-\lambda x}$ for $x \geq 0$
 - ▶ parameter $\lambda > 0$
- Data: $x^{(i)} \sim \text{Exponential}(\lambda)$ for $i = 1, \dots, M$
- Find MLE: and MAP
 - ▶ write down log-likelihood of $\{x^{(i)}\}$ (\neq prior)
 - ▶ compute derivative wrt parameter λ
 - ▶ set to 0, solve for λ
 - ▶ (ideally, check that we found global maximum)

$$\lambda \sim e^{-\lambda} \quad \lambda > 0$$

$$\lambda e^{-\lambda x} \text{ for } x \geq 0$$

- ▶ write down log-likelihood of $\{x^{(i)}\}$
- ▶ compute derivative wrt parameter λ
- ▶ set to 0, solve for λ

$$\begin{aligned}
 \ln L &= \sum_{i=1}^M \ln \left[\lambda e^{-\lambda x^{(i)}} \right] \\
 &= \sum_{i=1}^M \left(\ln \lambda - \lambda x^{(i)} \right) \\
 \frac{d}{d\lambda} \ln L &= \sum_{i=1}^M \left(\frac{1}{\lambda} - x^{(i)} \right) \\
 &= \frac{M}{\lambda} - \sum_{i=1}^M x^{(i)} \\
 &= 0
 \end{aligned}$$

MLE for Linear Regression

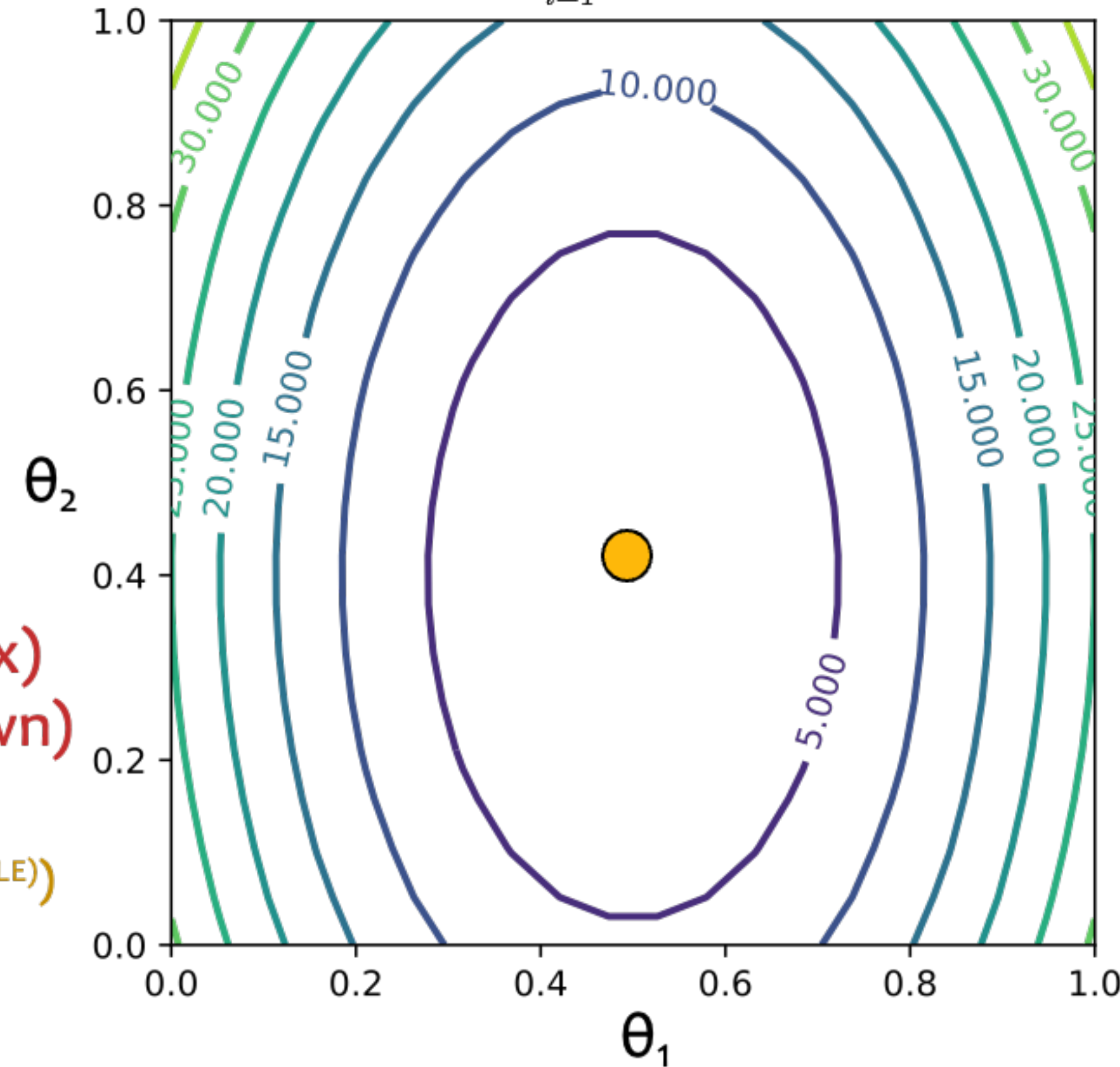
Optimization Method #2: Closed Form

1. Evaluate

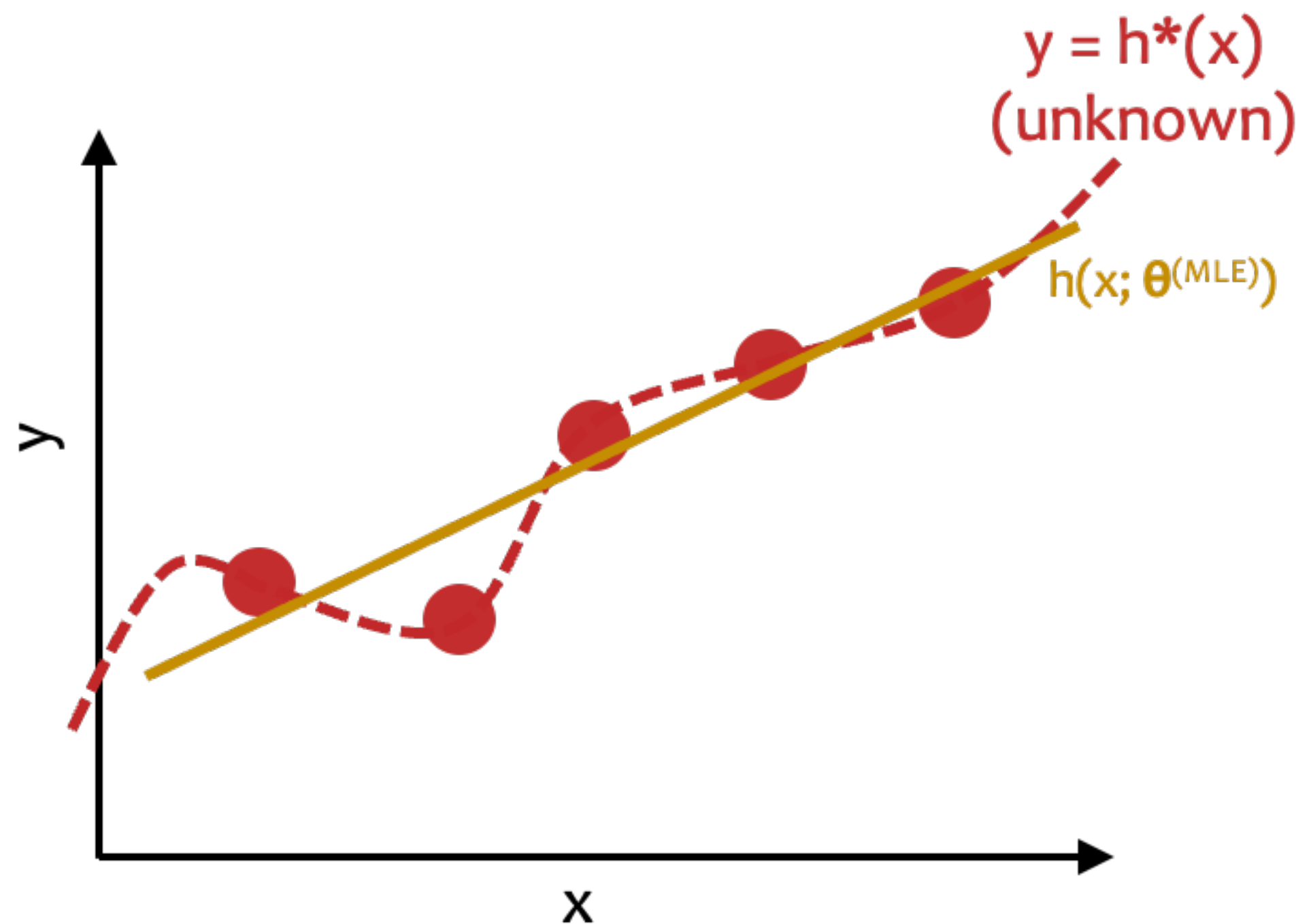
$$\theta^{\text{MLE}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

2. Return θ^{MLE}

$$J(\theta) = J(\theta_1, \theta_2) = \frac{1}{N} \sum_{i=1}^N (y^{(i)} - \theta^T \mathbf{x}^{(i)})^2 + 6(\theta_1 - 0.4)^2$$



| t | θ_1 | θ_2 | $J(\theta_1, \theta_2)$ |
|-----|------------|------------|-------------------------|
| MLE | 0.59 | 0.43 | 0.2 |



1. You'll work through the view of linear regression as a probabilistic model in the homework!
2. You'll also see how L1 and L2 regularization is equivalent to MAP estimation

Philosophy of probability

- Bayesian view (explanations so far):
 - ▶ parameters like θ are random variables, just like data
 - ▶ prior and posterior describe our state of knowledge about θ
- Frequentist view (also very reasonable):
 - ▶ parameters like θ are constants (properties of the world) not random variables
 - ▶ doesn't make sense to ask for a prior or posterior distribution over θ since it's not random
- \exists other philosophies of probability besides frequentist and Bayesian, but these are the two best known

To MAP or not to MAP?

- MAP doesn't make sense for frequentists! (no prior or posterior)
 - ▶ MLE does, and so does regularized MLE
 - ▶ regularized MLE is mathematically equivalent to MAP if regularizer = $\log(\text{prior})$, but justification is different
 - ▶ and MLE can use regularizers that aren't $\log(\text{prior})$
- Frequentist's justifications for either:
 - ▶ consistency, convergence rates, robustness
 - ▶ e.g., find estimator that $\rightarrow \theta$ quickly, even for worst case true distribution of data (minimax convergence rate)

Difference of opinion

- Bayesian critique of frequentist:
 - ▶ you are ignoring useful information (your incoming prior knowledge), so you are leaving performance on the table
- Frequentist critique of Bayesian:
 - ▶ you don't get the best possible minimax convergence rates, so you are leaving performance on the table
- Both are right
 - ▶ optimize different objectives → disagree on best answer
- ML'ers are typically very pragmatic: use Bayesian or frequentist reasoning depending on which leads to a practical computational solution