10-301/601: Introduction to Machine Learning Lecture 15 – Learning Theory

Geoff Gordon

with thanks to Matt Gormley & Henry Chai

Statistical learning theory setup

1. Data points are generated i.i.d. from some *unknown* distribution
   $$\boldsymbol{x}^{(n)} \sim p^*(\boldsymbol{x})$$
2. Labels are generated from some *unknown* function
   $$y^{(n)} = c^*\left(\boldsymbol{x}^{(n)}\right) \in \{-1, +1\}$$ note: ***binary*** classification
3. The learning algorithm chooses the hypothesis (classifier) with lowest *training* error rate from a specified hypothesis set, $\mathcal{H}$
4. Goal: return a hypothesis (or classifier) with low *true* error rate (measure on *test* set)

Types of Error

- True error rate
  - Actual quantity of interest for learning
  - How well your hypothesis will perform on new samples
- Training error rate
  - Used to choose $h \in \mathscr{H}$ (e.g., fit model parameters)
  - May be a very optimistic estimate of true error

Types of Error

- True error rate
  - Actual quantity of interest for learning
  - How well your hypothesis will perform on new samples
- Training error rate
  - Used to choose $h \in \mathscr{H}$ (e.g., fit model parameters)
  - May be a very optimistic estimate of true error
- Test error rate
  - Used to evaluate hypothesis performance
  - Good estimate of true error (w/ enough test data)
- Validation error rate
  - Used to help choose $\mathscr{H}$ (e.g., set hyperparameters)
  - Somewhat optimistic estimate of true error

Error rate is also called risk

- True error rate = (true) ***risk*** — unknown

$$R(h) = \mathbb{E}$$

- Training error rate = ***empirical risk*** — we can measure this

$$\hat{R}(h) = \mathbb{E}$$

Three classifiers

1. The *true classifier*, c^*: best answer but may be unachievable

2. The (*true*) *risk minimizer* (best achievable answer):

$$h^* = \operatorname*{argmin}_{h \in \mathscr{H}} R(h)$$

3. The *empirical risk minimizer* (the only one of the three that we can actually know)

$$\hat{h} = \operatorname*{argmin}_{h \in \mathscr{H}} \hat{R}(h)$$

Three classifiers

1. The *true classifier*, c^*: best answer but may be unachievable

2. The (*true) risk minimizer* (best achievable answer):

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} R(h)$$

3. The *empirical risk minimizer* (the only one of the three that we can actually know)

$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \hat{R}(h)$$

ERM = the learning method that picks a hypothesis by minimizing empirical risk

Overfitting

- Recall: ***overfitting*** = difference between true error rate and training error rate =
- Goal for today: ***predict*** and ***control*** overfitting for ERM
  - by finding (and proving) conditions that keep $\hat{R}(h)$ close to $R(h)$

Bound on overfitting

- PAC = **P**robably **A**pproximately **C**orrect criterion

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

for some ϵ (difference between true and empirical risk)

and δ (probability of "failure")

why the name?

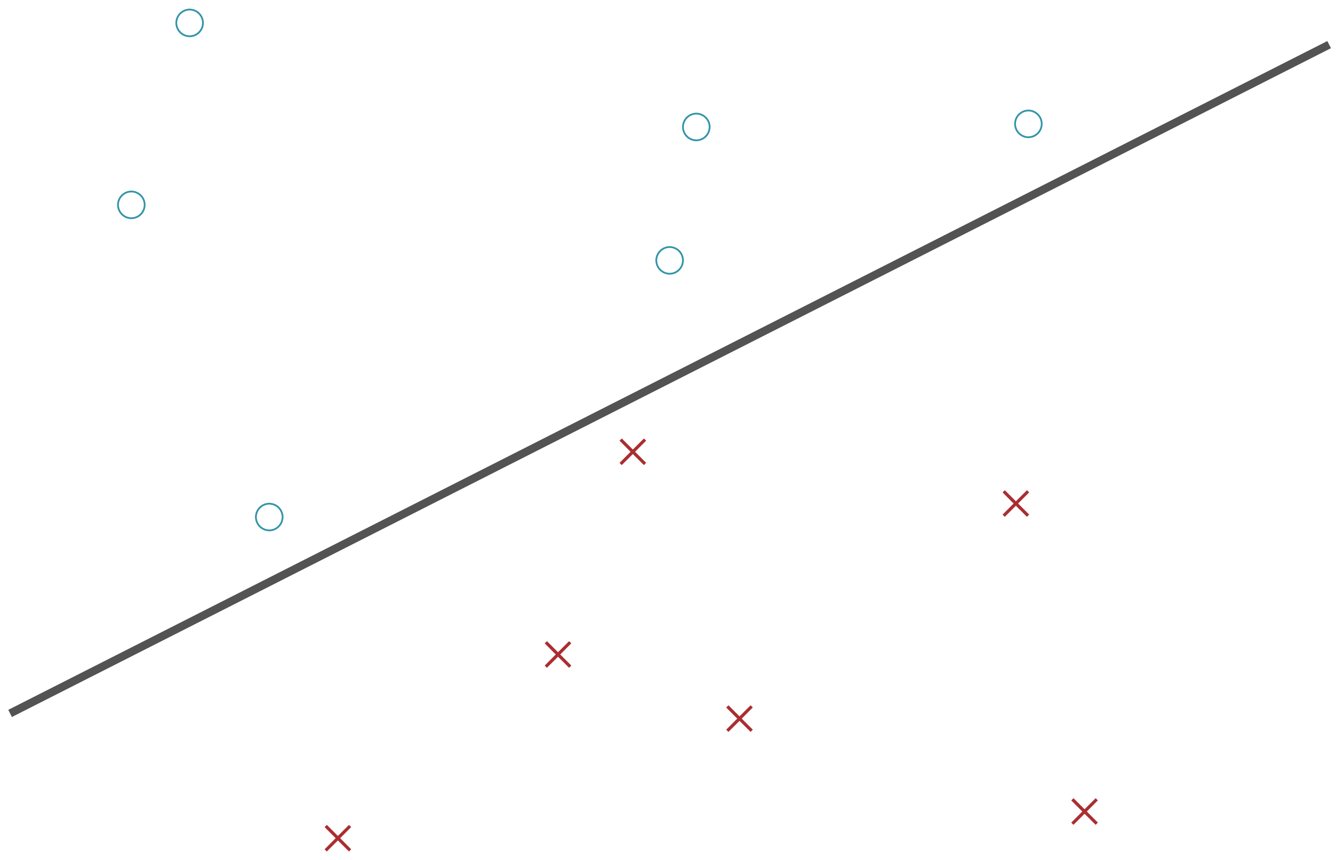
Sample Complexity

- We will do ERM on some $\mathscr{H}$ with M training examples
- We want to satisfy the PAC criterion with ***small*** ϵ and δ
- Chief levers: $\mathscr{H}$ and M
- Sample complexity (of ERM on $\mathscr{H}$) = the M we need in order to satisfy the PAC criterion for a given ϵ and δ

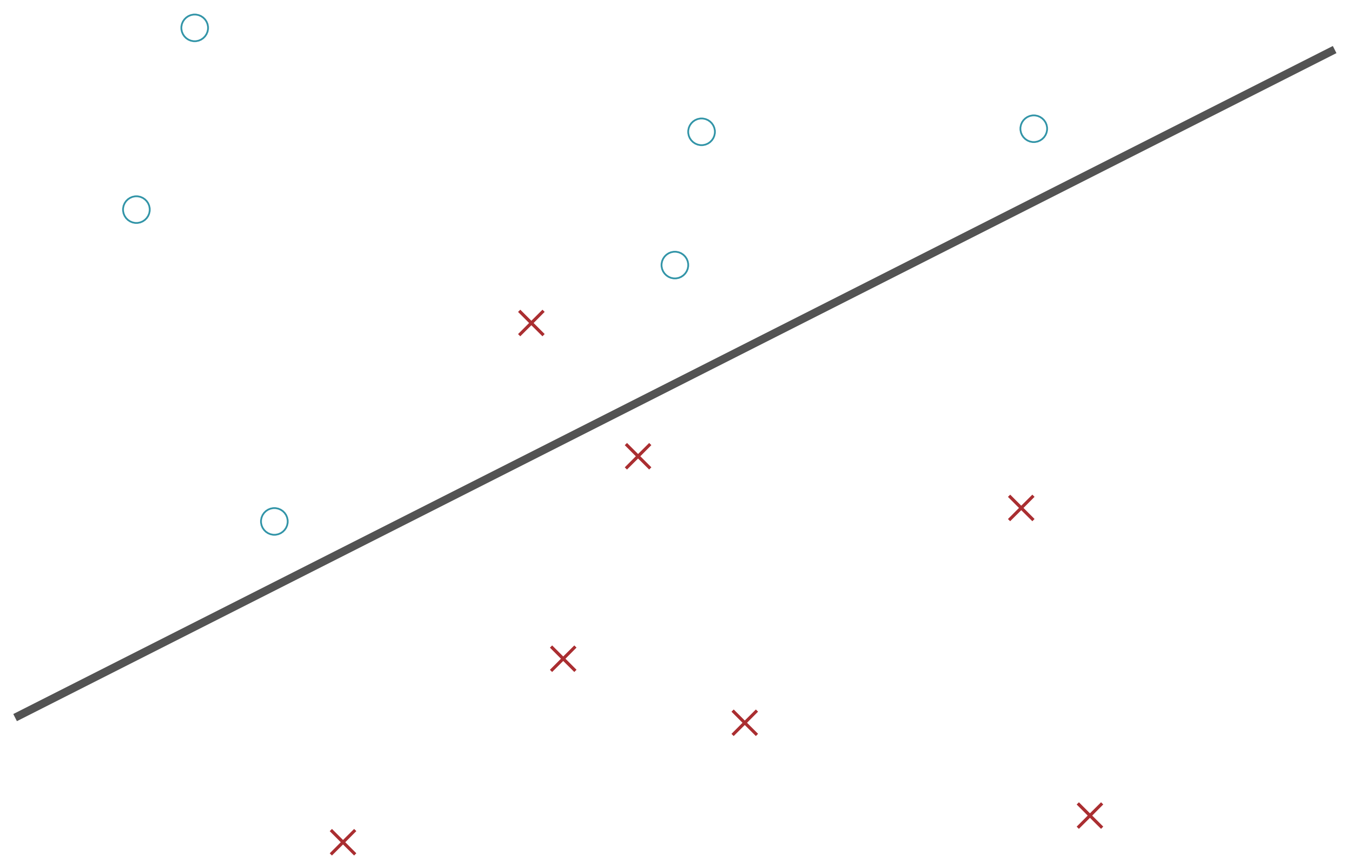
Four cases

- Realizable vs. Agnostic
  - Realizable $\rightarrow c^* \in \mathscr{H}$
  - Agnostic $\rightarrow c^*$ might or might not be in $\mathscr{H}$
- Finite vs. Infinite
  - Finite $\rightarrow \left|\mathscr{H}\right| < \infty$
  - Infinite $\rightarrow \left|\mathscr{H}\right| = \infty$

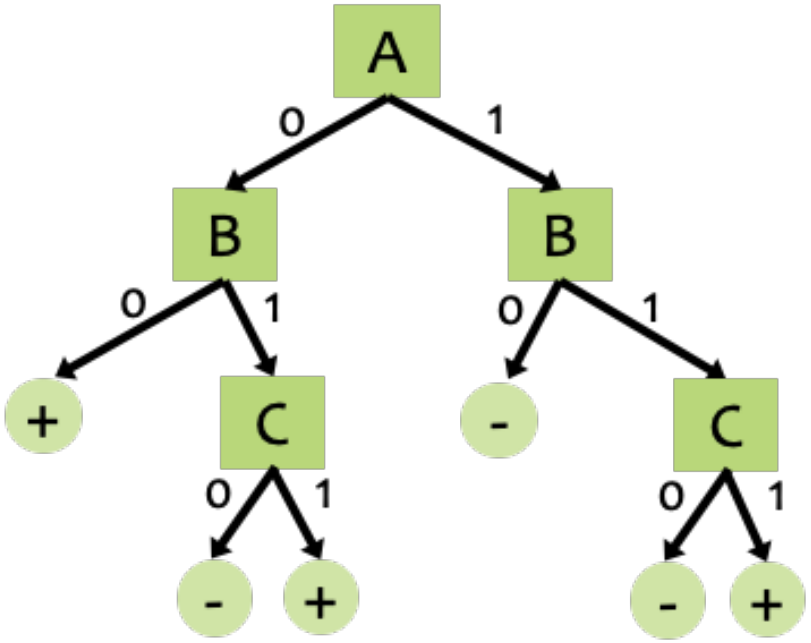
Realizable vs. agnostic

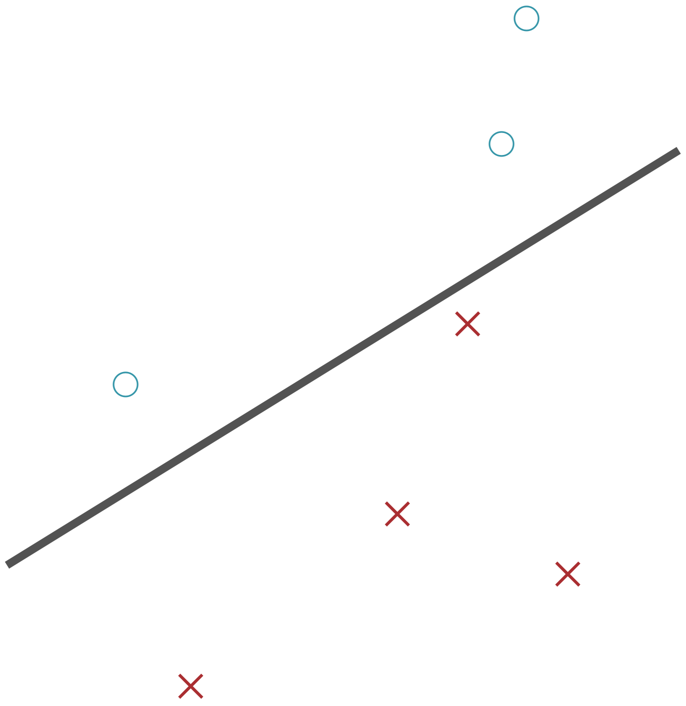


Realizable vs. agnostic



Finite vs. infinite $|\mathcal{H}|$





- Decision trees of bounded depth on discrete attributes vs. linear separators in 2D

Theorem 1: Finite, Realizable Case

For a finite hypothesis set $\mathcal{H}$ s.t. $c^* \in \mathcal{H}$ and arbitrary distribution p^*, if the number of labelled training data points satisfies

$$M \geq \frac{1}{\epsilon}\left(\ln\left(|\mathcal{H}|\right) + \ln\left(\frac{1}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

We will prove this over the next few slides

Theorem 1: Finite, Realizable Case

Bound is linear in $\frac{1}{\epsilon}$

thesis set $\mathscr{H}$ s.t. $c^* \in \mathscr{H}$ and arbitrary

f the number of labelled training data

points sfies

$$M \geq \frac{1}{\epsilon}\left(\ln\left(|\mathscr{H}|\right) + \ln\left(\frac{1}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathscr{H}$ with

$\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

We will prove this over the next few slides

Theorem 1: Finite, Realizable Case

Bound is linear in $\frac{1}{\epsilon}$

but only logarithmic in $|\mathcal{H}|$ and $\frac{1}{\delta}$

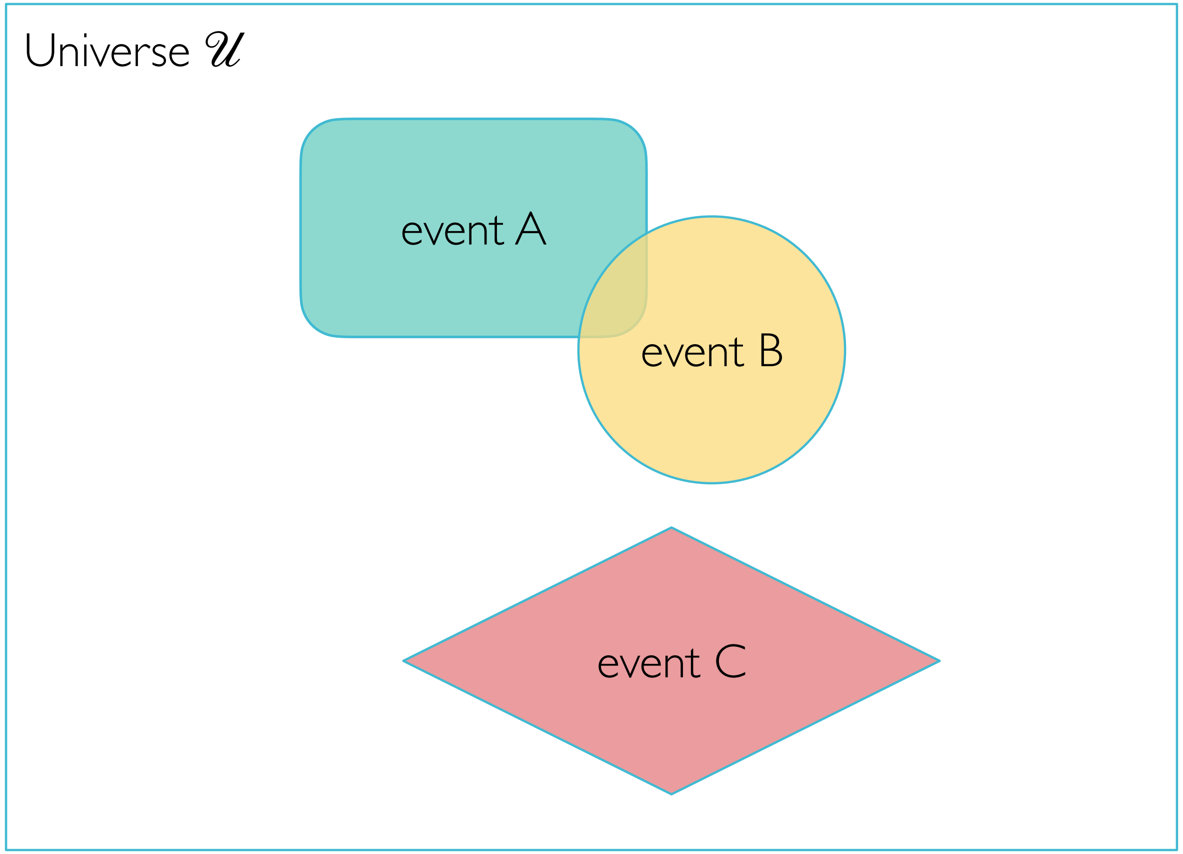
... thes... nd arbitrary ...f the... raining data points ... sfies

$$M \geq \frac{1}{\epsilon}\left(\ln\left(|\mathcal{H}|\right) + \ln\left(\frac{1}{\delta}\right)\right)$$

then with probability at least $1-\delta$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

We will prove this over the next few slides

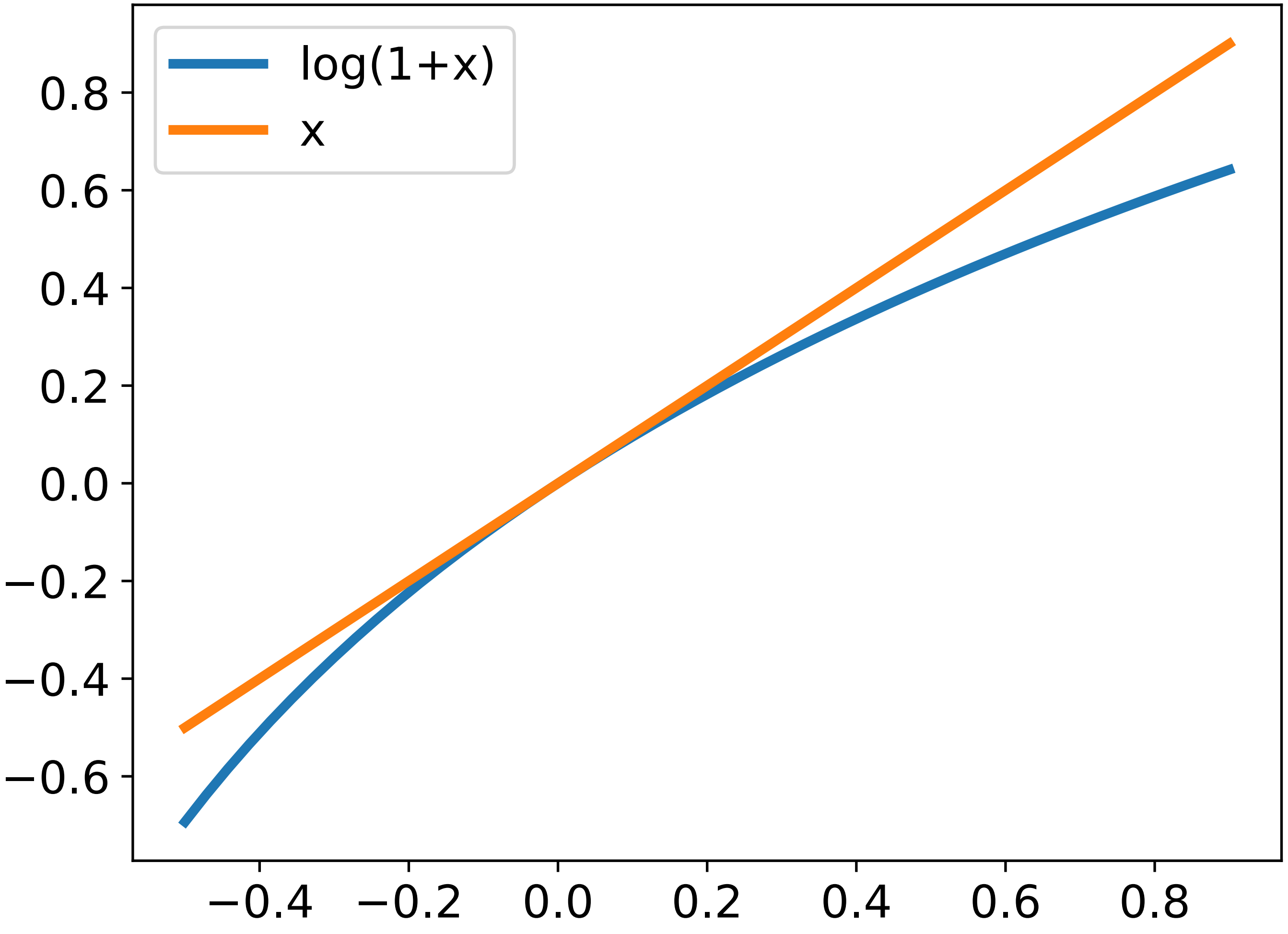
Union bound



$$P(A \text{ or } B \text{ or } C) \le P(A) + P(B) + P(C)$$

$P(\text{some event happens}) \le \text{sum of probabilities}$

Bound on log



$$\ln(1+x) \leq x$$

Notation for conclusion

- Theorem said: if

$$M \geq \frac{1}{\epsilon}\left(\ln\left(|\mathscr{H}|\right) + \ln\left(\frac{1}{\delta}\right)\right)$$

then with probability at least $1 - \delta$, all $h \in \mathscr{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

- Write E for event: $\exists h \in \mathscr{H}$ with $\hat{R}(h) = 0$, $R(h) > \epsilon$
- Theorem's conclusion is $P(E) < \delta$

Bound for one hypothesis

- Consider some h with $R(h) > \epsilon$. What's $P(\hat{R}(h) = 0)$?

Bound for k hypotheses

- Suppose there are k hypotheses with $R(h) > \epsilon$. What's $P(E)$, i.e., $P(\hat{R}(h_1) = 0$ or $\hat{R}(h_2) = 0$ or $\ldots)$?

Theorem will be true if $P(E) \leq \delta$

Solve for M

we have

$$P(E) < |\mathscr{H}|(1 - \epsilon)^M$$

Theorem 1: corollary

- For a finite hypothesis set $\mathscr{H}$ s.t. $c^* \in \mathscr{H}$ and arbitrary distribution p^*, given a training data set S s.t. $|S| = M$, all $h \in \mathscr{H}$ with $\hat{R}(h) = 0$ have

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

with probability at least $1 - \delta$.

Recall Theorem 1 said $M \geq \frac{1}{\epsilon}\left(\ln\left(|\mathscr{H}|\right) + \ln\left(\frac{1}{\delta}\right)\right)$

Theorem 2: finite, agnostic case

- For a finite hypothesis set $\mathscr{H}$ and arbitrary distribution p^*, if the number of labelled training data points satisfies

$$M \geq \frac{1}{2\epsilon^2}\left(\ln\left(\left|\mathscr{H}\right|\right) + \ln\left(\frac{2}{\delta}\right)\right)$$

  then with probability at least $1 - \delta$, all $h \in \mathscr{H}$ satisfy

$$\left|R(h) - \hat{R}(h)\right| \leq \epsilon$$

- Bound is inversely ***quadratic*** in ϵ, e.g., halving ϵ means we need four times as many labelled training data points

Theorem 2: finite, agnostic case

- For a finite hypothesis set $\mathscr{H}$ and arbitrary distribution p^*, if the number of labelled training data points satisfies

$$M \geq \frac{1}{2\epsilon^2}\left(\ln\left(\left|\mathscr{H}\right|\right) + \ln\left(\frac{2}{\delta}\right)\right)$$

then with probability at least $1-\delta$, all $h \in \mathscr{H}$ satisfy

$$\left|R(h) - \hat{R}(h)\right| \leq \epsilon$$

- Bound is inversely ***quadratic*** in ϵ, e.g., halving ϵ means we need four times as many labelled training data points
- Again, making the bound tight and solving for ϵ gives…

Theorem 2: corollary

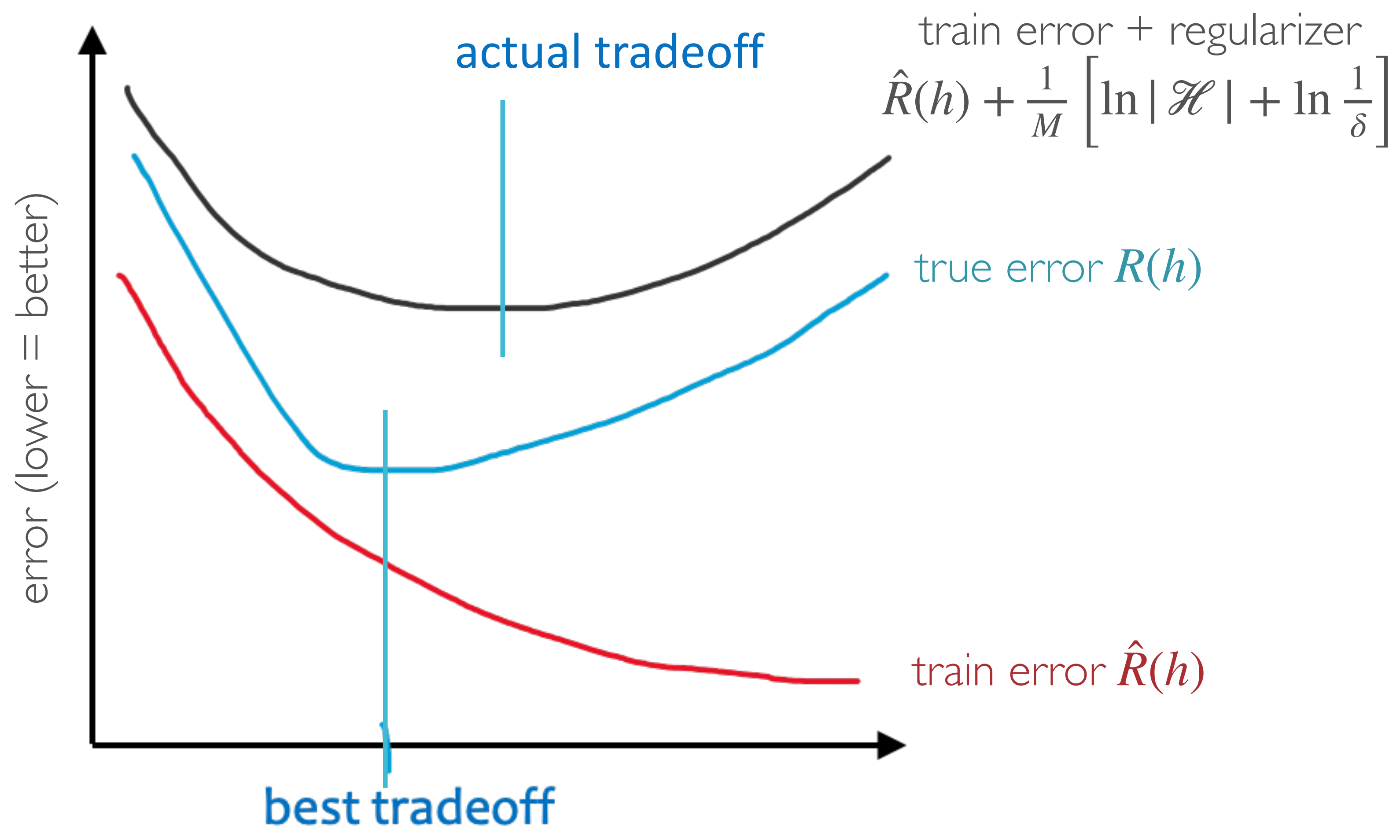
- For a finite hypothesis set $\mathscr{H}$ and arbitrary distribution p^*, given a training data set S s.t. $\left|S\right| = M$, all $h \in \mathscr{H}$ have

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

with probability at least $1 - \delta$.

Learning theory & model selection

Key point: we want to trade off low training error vs. keeping $\mathscr{H}$ simple



Ex: $\mathscr{H}$ = conjunctions on d binary attributes: $\ln|\mathscr{H}| = d\ln 3$
Expert sorts attributes, most likely to be relevant first
We allow conjunctions on first d of them:

- training error $\downarrow$ as d increases
- regularizer $\uparrow$ as d increases
- stop when PAC bound is smallest (best tradeoff)

What happens when $|\mathscr{H}| = \infty$?

- Bounds:

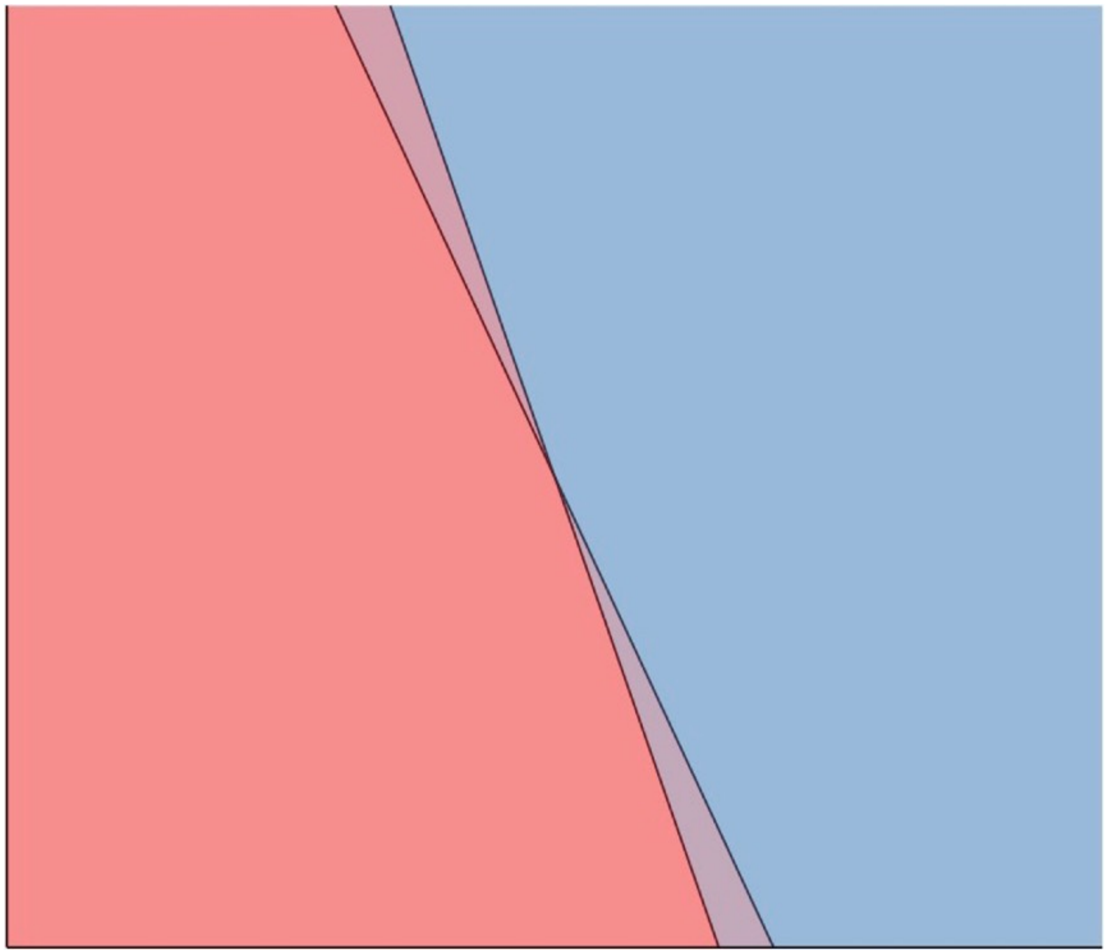
$$R(h) \leq \frac{1}{M}\left(\ln\left(|\mathscr{H}|\right) + \ln\left(\frac{1}{\delta}\right)\right)$$

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

with probability at least $1 - \delta$.

Intuition

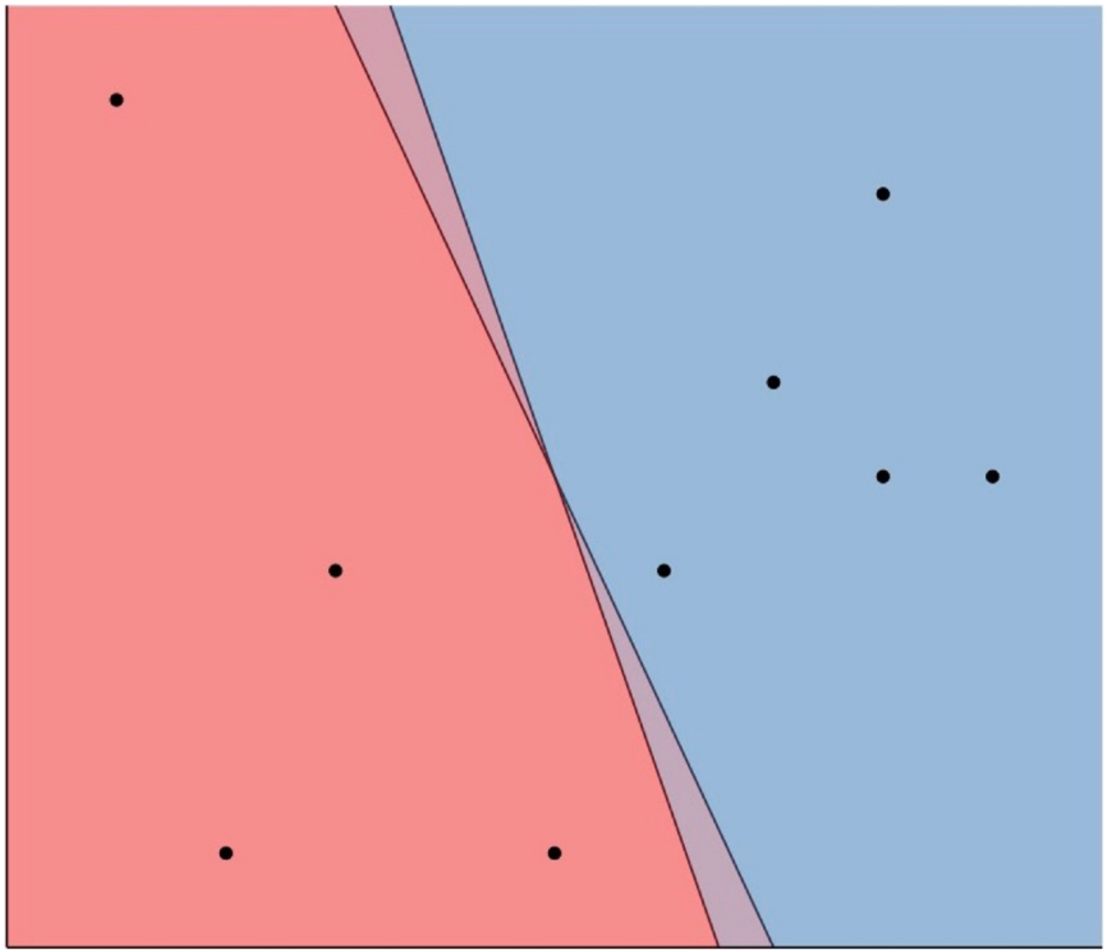
For "nice" infinite hypothesis sets $\mathcal{H}$, many hypotheses in $\mathcal{H}$ will behave similarly



Intuition

For "nice" infinite hypothesis sets $\mathscr{H}$, many hypotheses in $\mathscr{H}$ will behave similarly

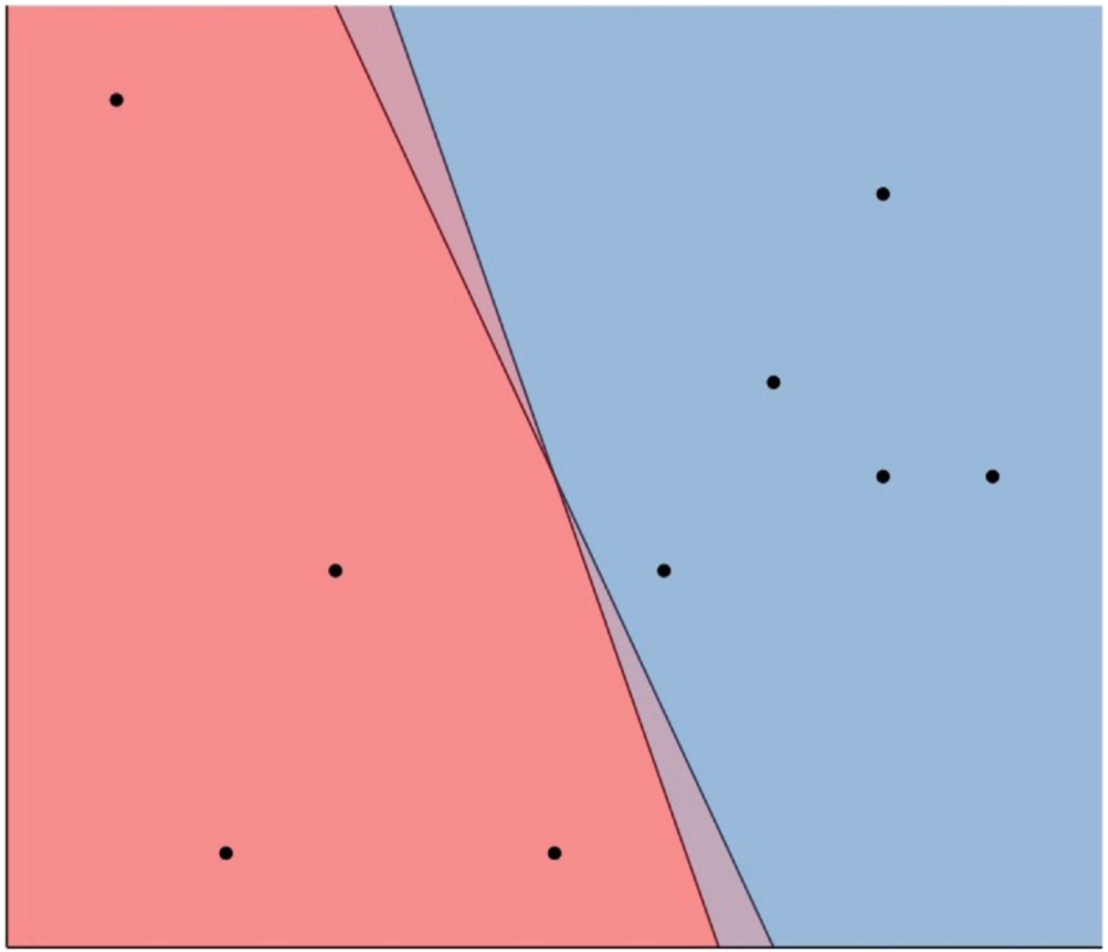
Relative to this dataset, these two hypotheses are *identical*!



Intuition

For "nice" infinite hypothesis sets $\mathscr{H}$, many hypotheses in $\mathscr{H}$ will behave similarly

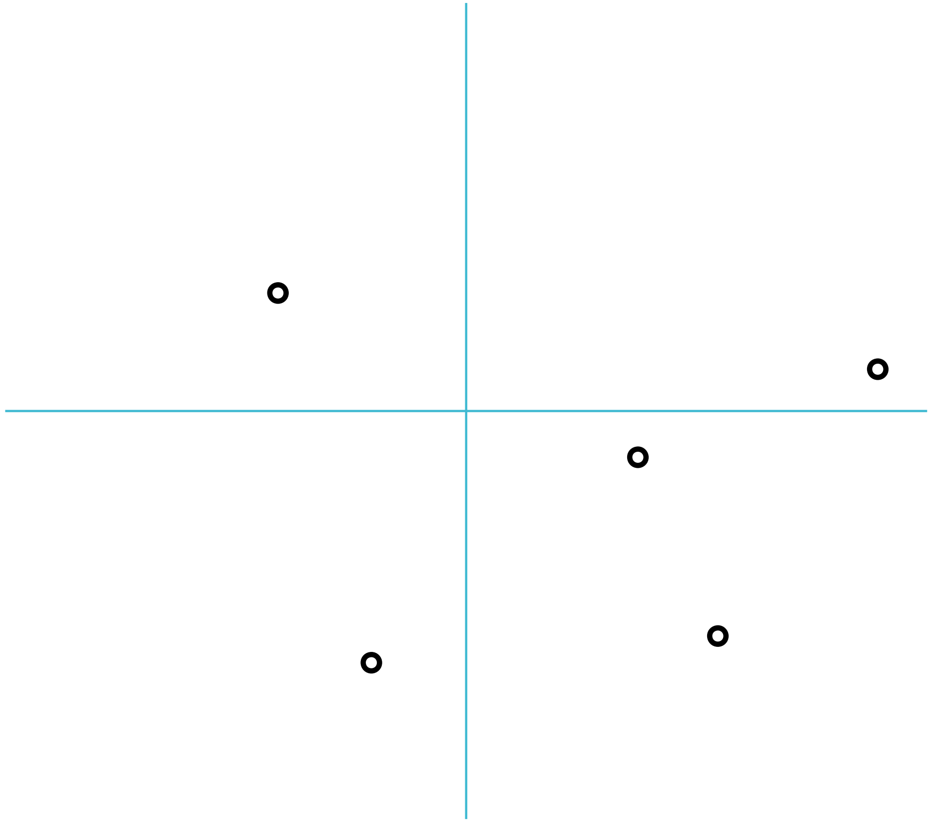
Relative to this dataset, these two hypotheses are *identical*!



Idea: instead of using full size of $\mathscr{H}$, count how many ***actually distinct*** hypotheses there are

How many distinct $h \in \mathscr{H}$ can there be?

$M = 5$



- What's the largest possible number of distinct hypotheses on M points?

What does our bound tell us if $|\mathscr{H}| = 2^m$?

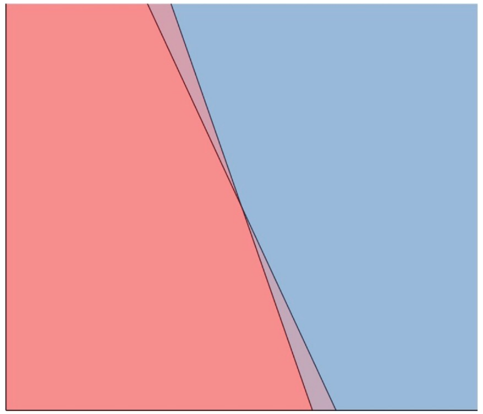
$$R(h) \leq \frac{1}{M}\left(\ln\left(|\mathscr{H}|\right) + \ln\left(\frac{1}{\delta}\right)\right)$$

What does our bound tell us if $|\mathscr{H}| = 2^m$?

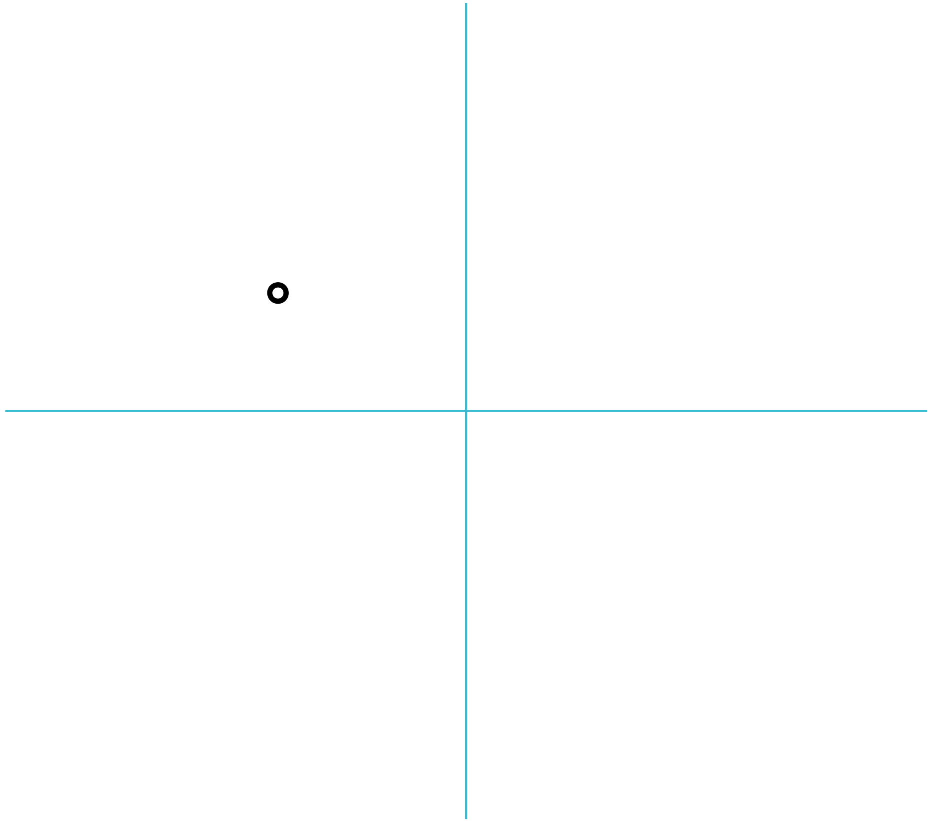
$$R(h) \le \frac{1}{M}\left(\ln\left(|\mathscr{H}|\right) + \ln\left(\frac{1}{\delta}\right)\right)$$

Vacuous! Need a tighter count of $|\mathscr{H}|$

Not surprising since $|\mathscr{H}| = 2^m$ is a kind of memorization learner

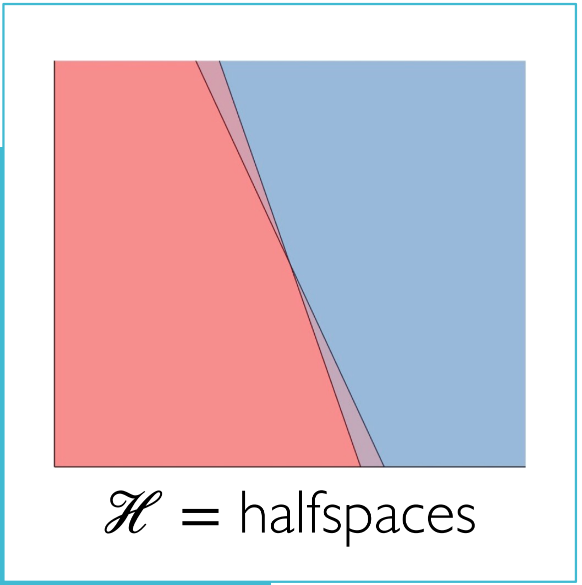


$M = 1$

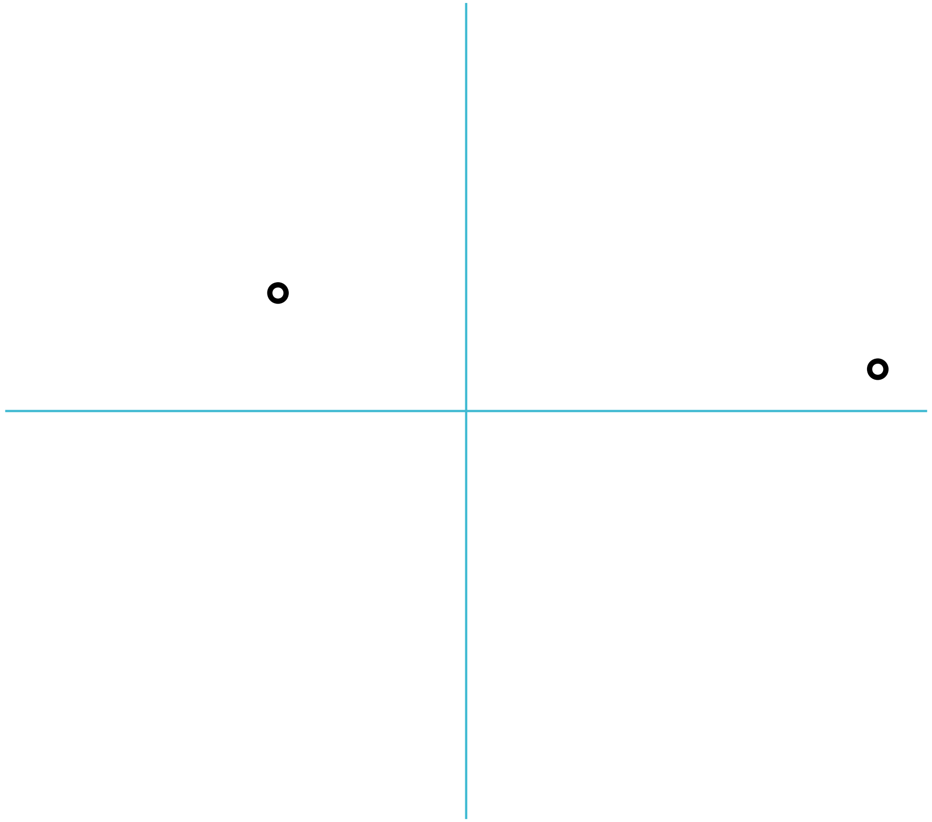
$\mathscr{D} =$ 

Counting h

- Fix $\mathscr{H}$
- Consider datasets $\mathscr{D}$ of size $M = 1, 2, \ \dots$
- For each dataset, count how many ***actually distinct*** hypotheses $h \in \mathscr{H}$ there are: $|\mathscr{H}(\mathscr{D})|$

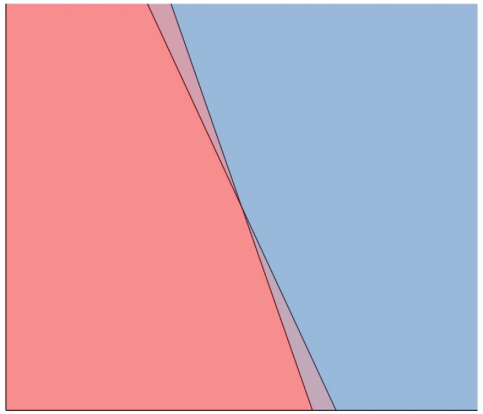


$M = 2$

$\mathscr{D} =$ 

Counting h

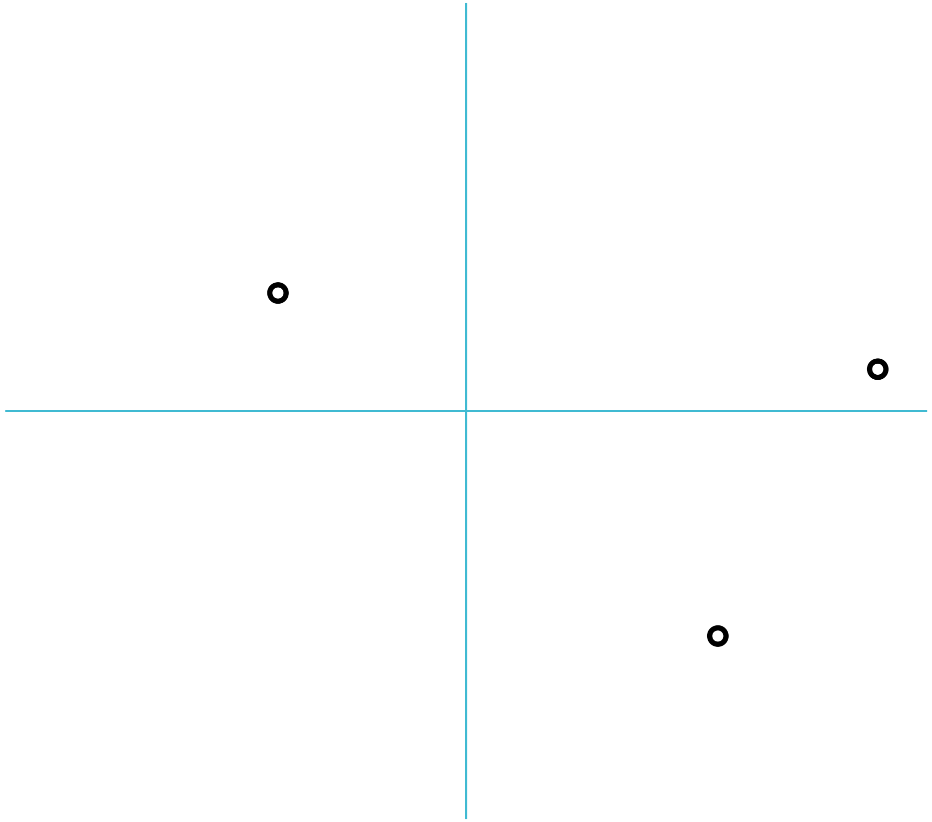
- Fix $\mathscr{H}$
- Consider datasets $\mathscr{D}$ of size $M = 1, 2, \ldots$
- For each dataset, count how many ***actually distinct*** hypotheses $h \in \mathscr{H}$ there are: $|\mathscr{H}(\mathscr{D})|$



$\mathscr{H}$ = halfspaces

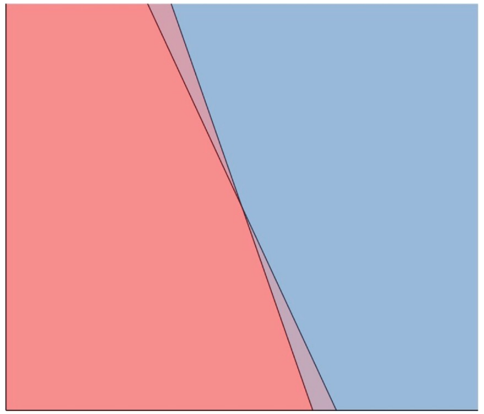
$M = 3$

$\mathscr{D} =$



Counting h

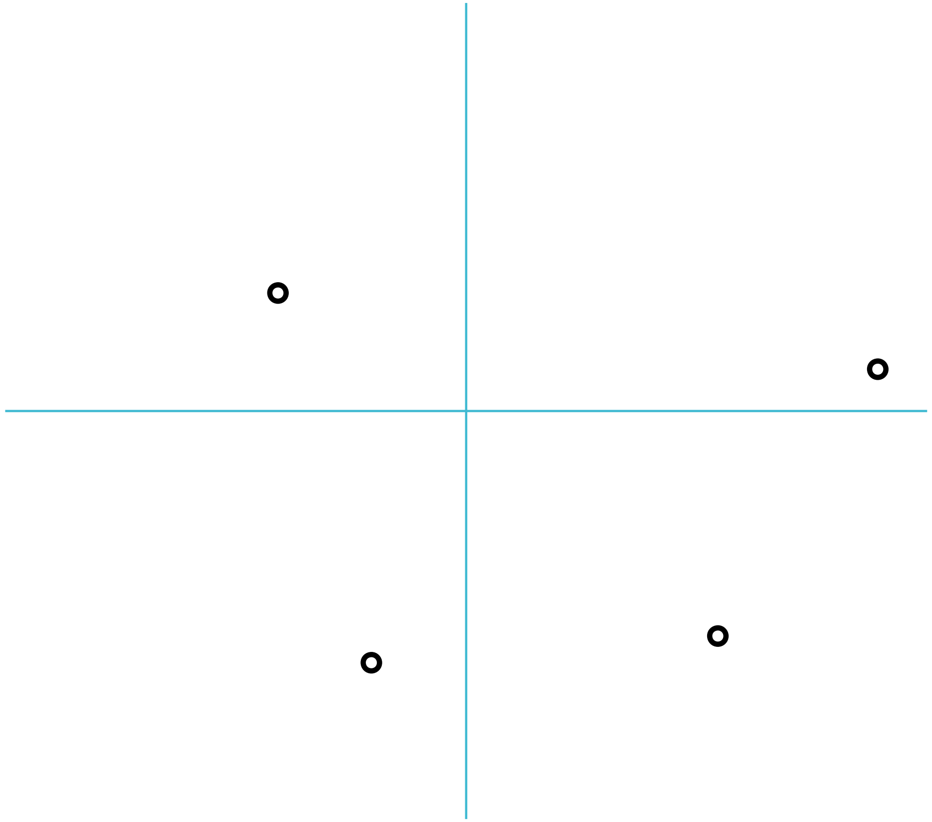
- Fix $\mathscr{H}$
- Consider datasets $\mathscr{D}$ of size $M = 1, 2, \ldots$
- For each dataset, count how many ***actually distinct*** hypotheses $h \in \mathscr{H}$ there are: $|\mathscr{H}(\mathscr{D})|$



$\mathscr{H}$ = halfspaces

$M = 4$

$\mathscr{D} =$



Counting h

- Fix $\mathscr{H}$
- Consider datasets $\mathscr{D}$ of size $M = 1, 2, \ \ldots$
- For each dataset, count how many ***actually distinct*** hypotheses $h \in \mathscr{H}$ there are: $|\mathscr{H}(\mathscr{D})|$

Growth function

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

M	1	2	3	4	5	…
$S_{\mathscr{H}}(M)$	2	4	8	14	≤ 26	…

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

not obvious!

Growth function

Def'n: ***shattering***

M	1	2	3	4	5	…
$S_{\mathscr{H}}(M)$	2	4	8	14	≤ 26	…

← shattered   not shattered →

$\mathscr{H}$ ***shatters*** a set of points if it can classify them all possible ways

Two kinds of behavior

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

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

  - e.g., intervals (or rectangles or hyperrectangles)
  - e.g., bounded-depth decision trees
  - e.g., fixed-architecture neural networks
- For many other classes $\mathscr{H}$, instead $S_{\mathscr{H}}(M) = 2^M$ for all M
  - e.g., unbounded-depth decision trees
  - e.g., unbounded-size neural networks

can shatter a set of each size

Two kinds of behavior

Learnable (can't memorize more than d points)

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

$$S_{\mathcal{H}}(M) = \begin{cases} 2^M & M \leq d \\ \ll 2^M & M > d \end{cases}$$

shattered

not shattered

  - 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 learnable (can memorize at any $|\mathcal{D}|$)

Sauer's lemma

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

"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

Sauer's lemma

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

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

  ➡ Then $S_{\mathscr{H}}(M) = O(M^d)$

"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

What do our bounds tell us w/ Sauer's lemma?

- Finite realizable case:

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

- Infinite realizable case:

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

What do our bounds tell us w/ Sauer's lemma?

- Finite agnostic case:

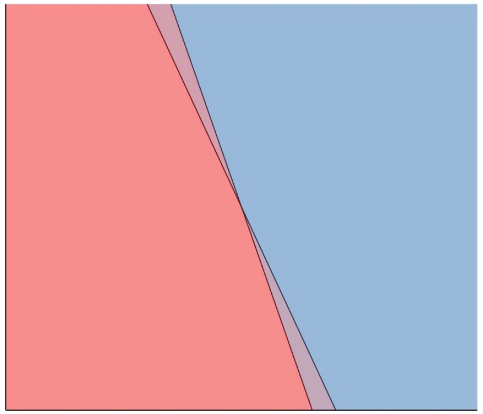
$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2M}\left(\ln\left(|\mathscr{H}|\right) + \ln\left(\frac{2}{\delta}\right)\right)}$$

- Infinite agnostic case:

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

Finding the VC-dimension

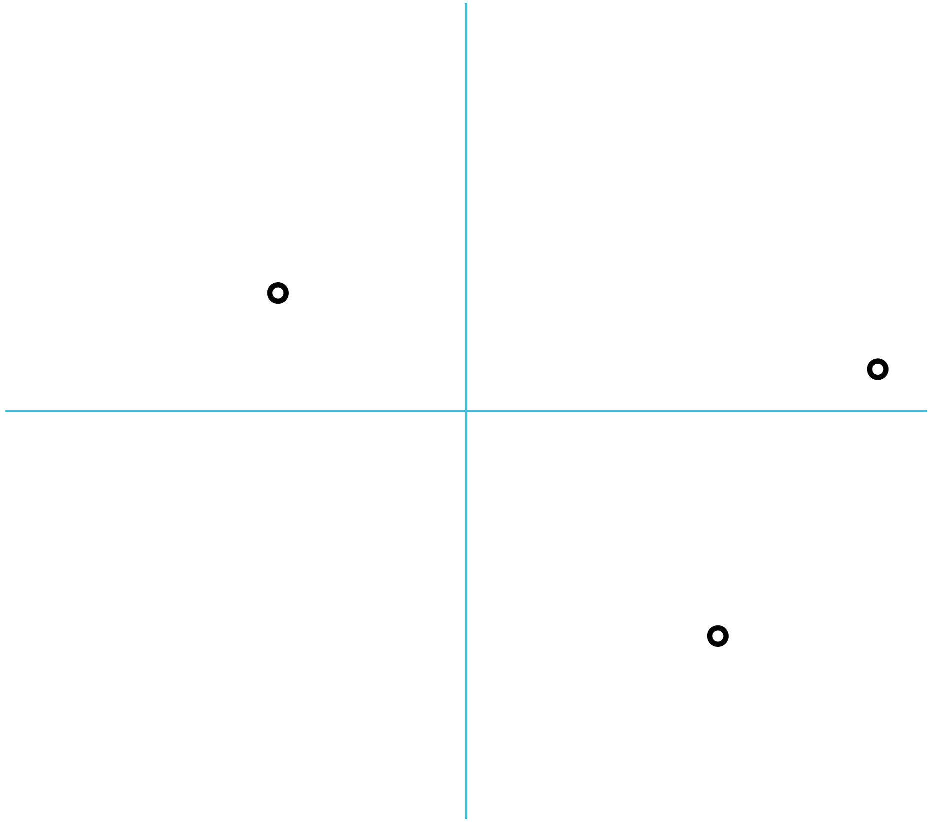
- We defined **VC-dimension** of $\mathscr{H}$, $VC\left(\mathscr{H}\right)$, as the size of the largest set S that $\mathscr{H}$ can shatter
  - If $\mathscr{H}$ can shatter arbitrarily large sets, $VC(\mathscr{H}) = \infty$
- To prove that $VC\left(\mathscr{H}\right) = d$, need to show
  1. $\exists$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



$\mathcal{H}$ = halfspaces

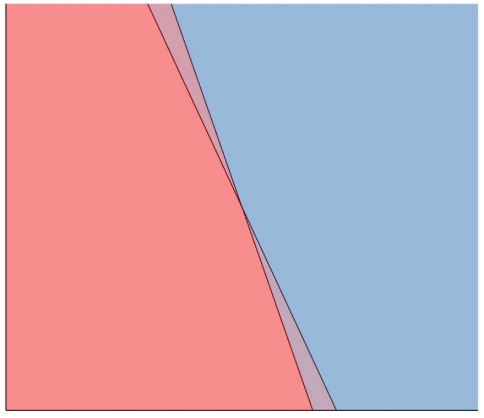
$M = 3$

$\mathcal{D} =$



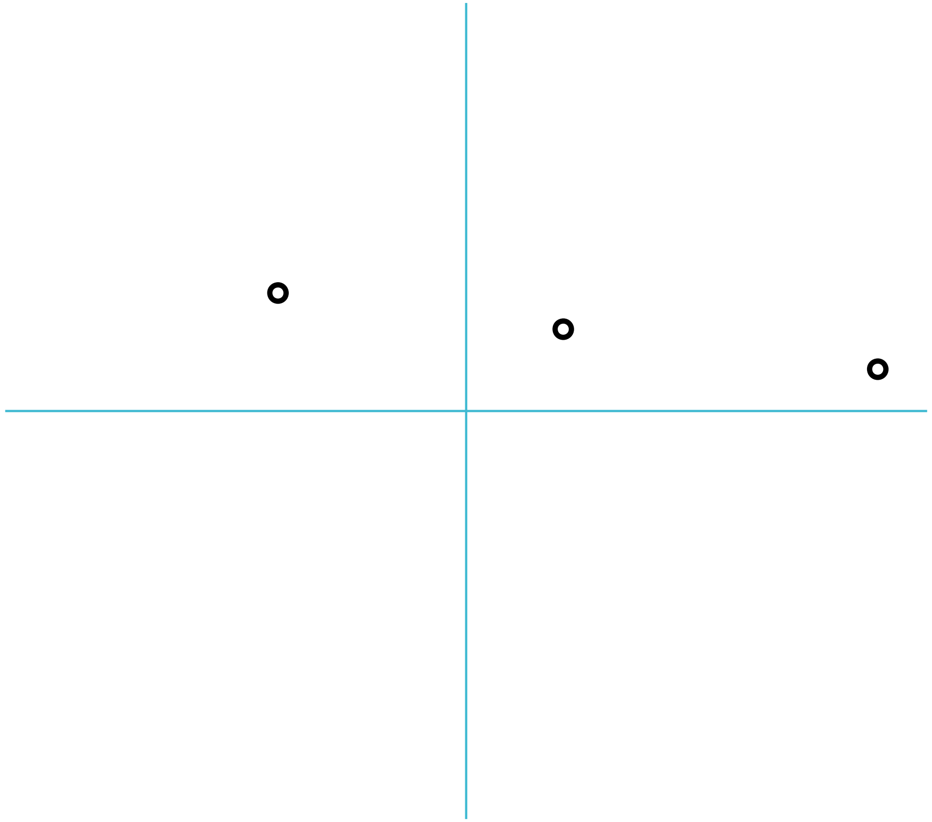
VC-dimension example

- Before, we looked at this dataset of size 3



$M = 3$

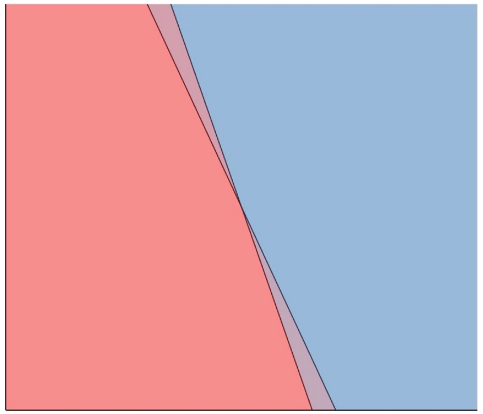
$\mathcal{D} =$



VC-dimension example

- But what if we had looked at this one?

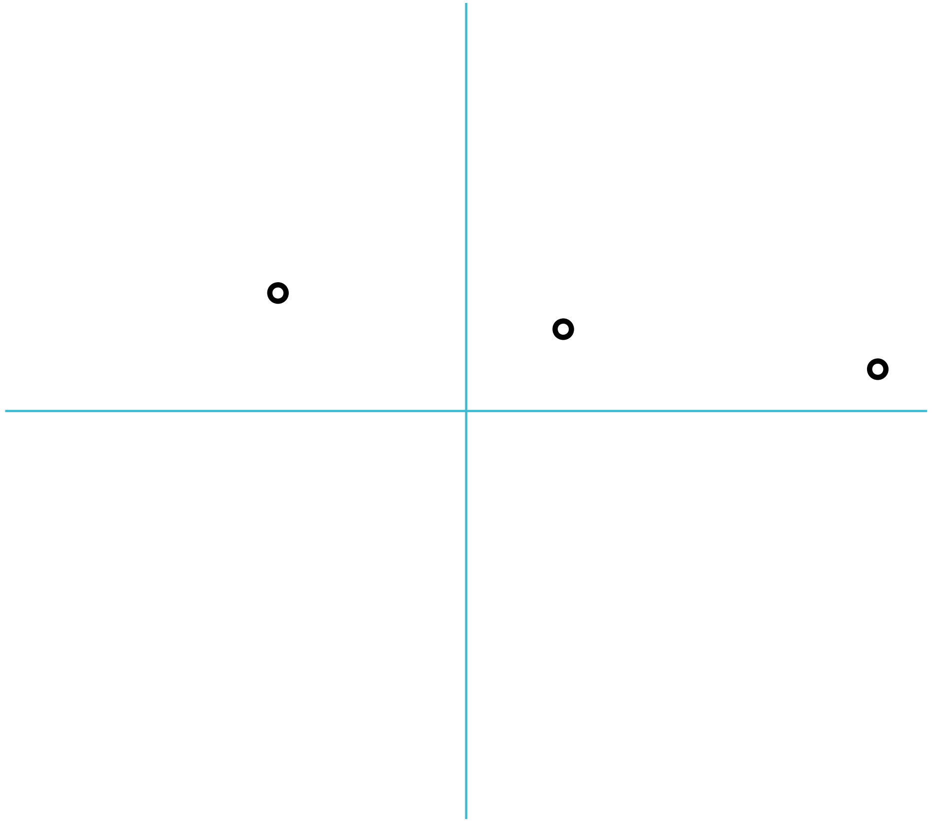
VC-dimension example



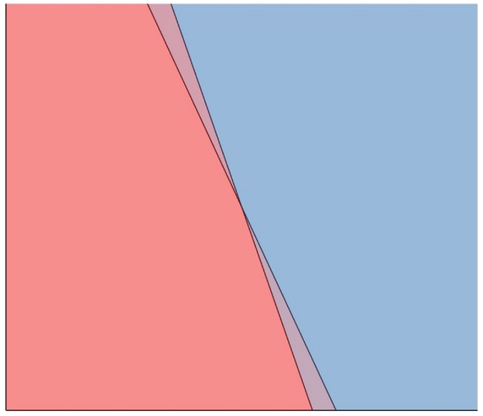
$\mathscr{H}$ = halfspaces

$M = 3$

$\mathscr{D} =$



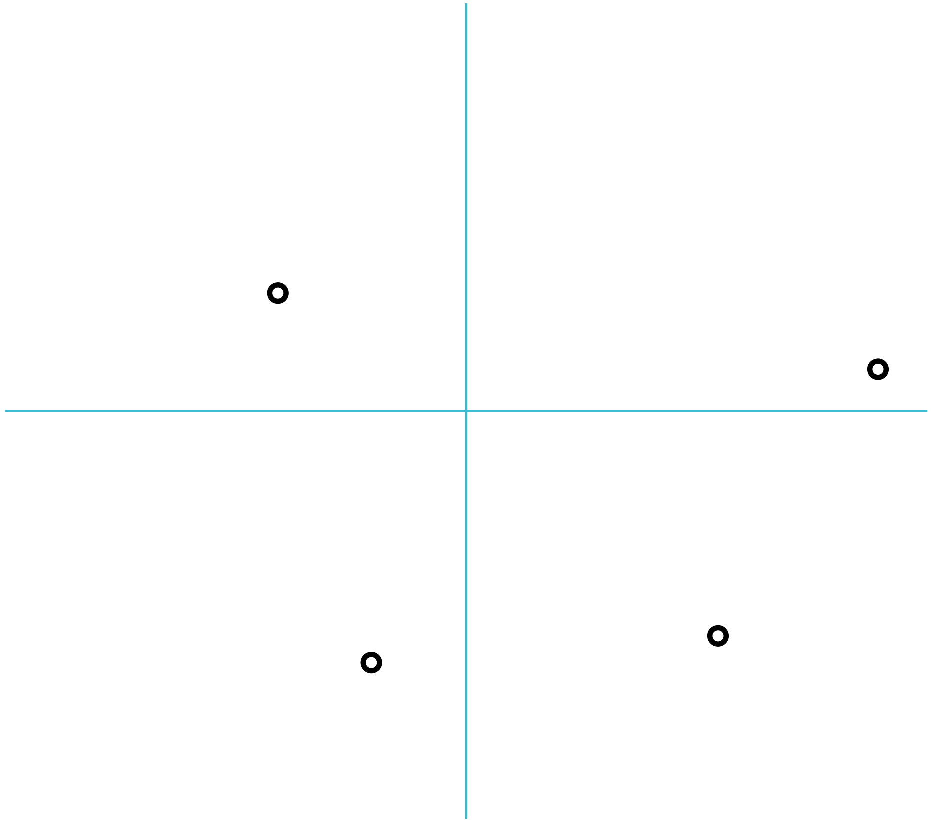
- Only 6 distinct hypotheses $< 2^3$
- Tempting to say $\text{VC}(\mathscr{H}) < 3$ — but would be ***wrong***



VC-dimension example

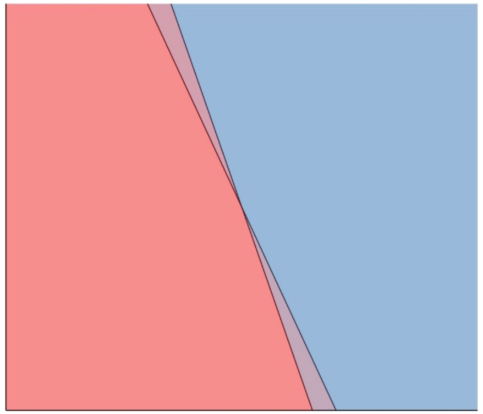
$M = 4$

$\mathcal{D} =$



- Similarly, looked at this dataset of size 4

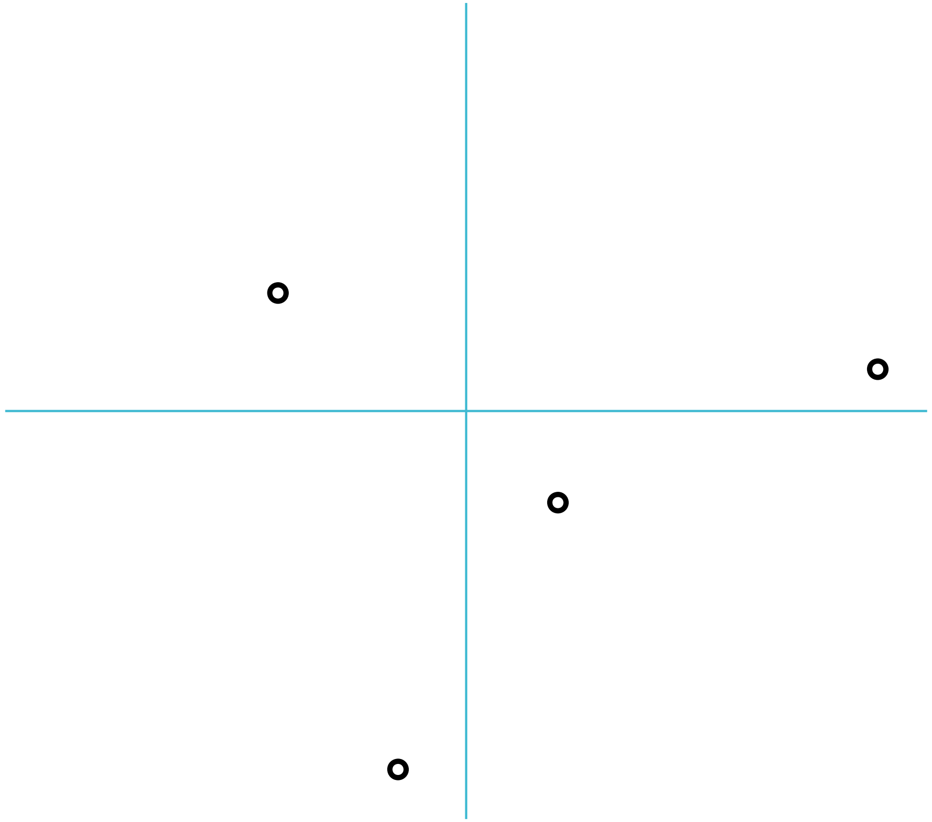
VC-dimension example



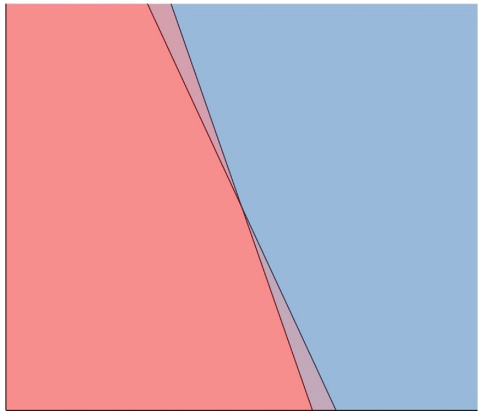
$\mathscr{H}$ = halfspaces

$M = 4$

$\mathscr{D} =$



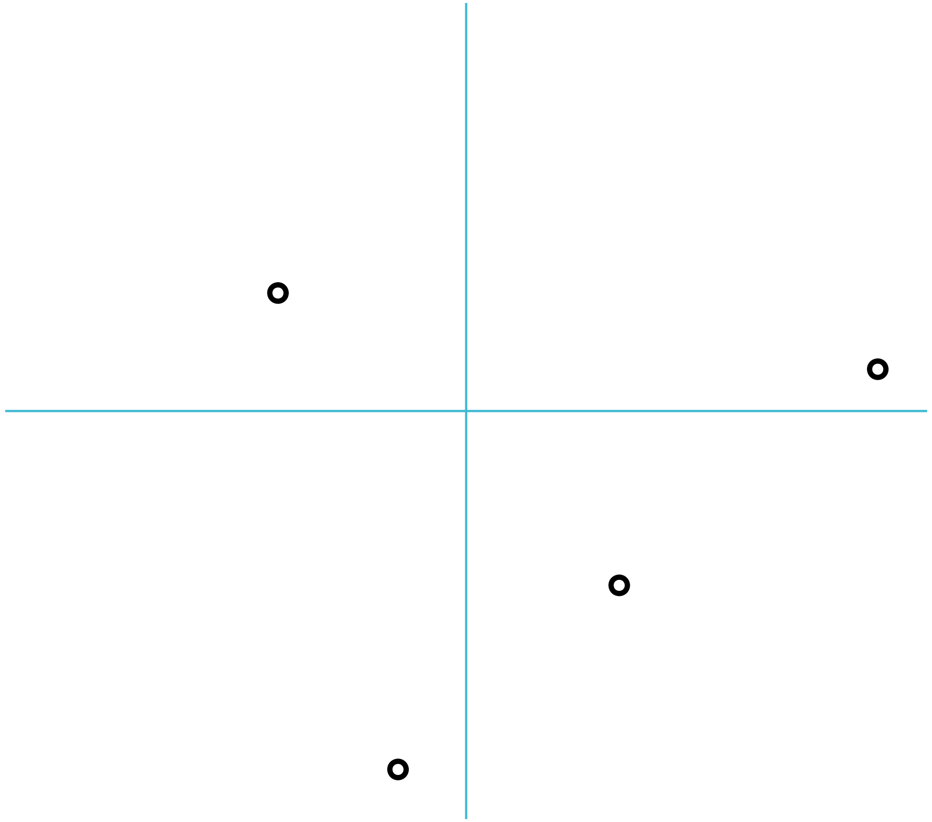
- But really should have checked this one as well



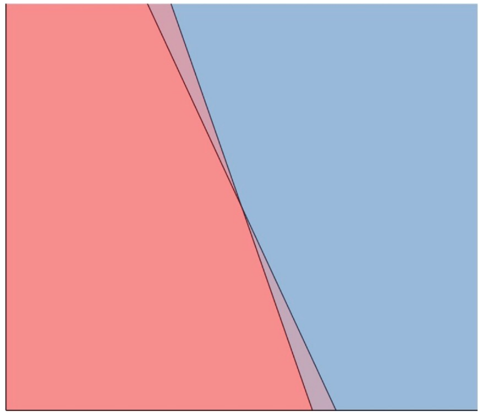
VC-dimension example

$M = 4$

$\mathscr{D} =$

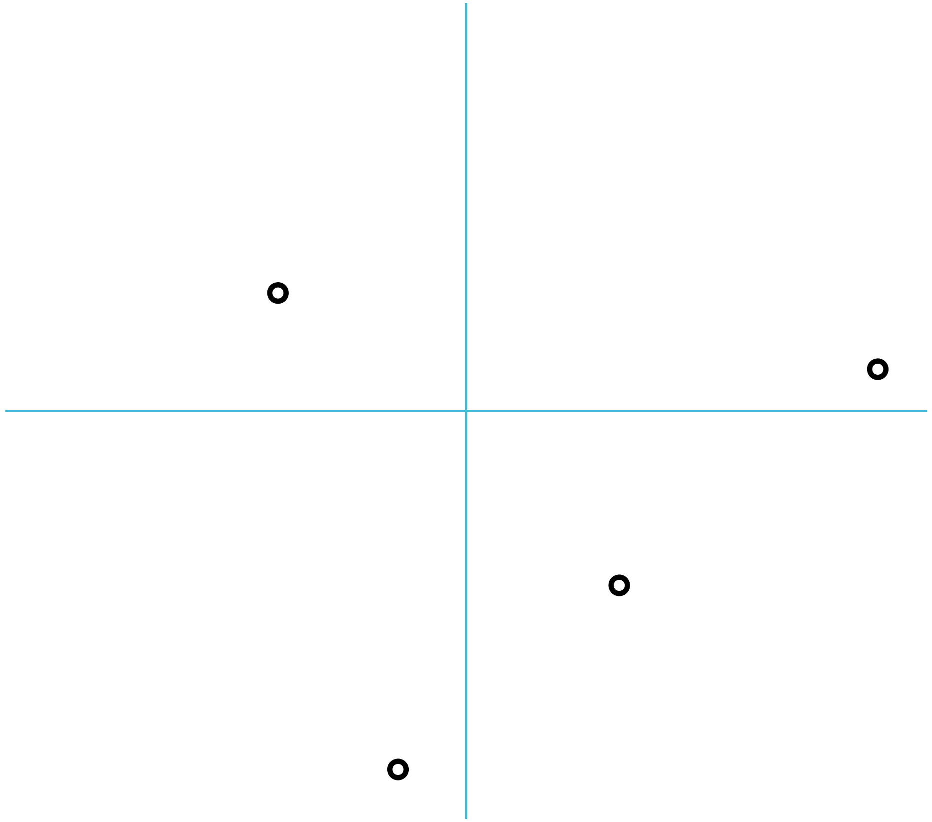


- And this one



$M = 4$

$\mathscr{D} =$



VC-dimension example

- And this one

fortunately it turns out we're OK: no matter how we set up the points, can't shatter 4 of them

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 $VC(\mathscr{H})$ for
  - $\mathscr{H}$ = half-lines *where positive class is on right*
  - $\mathscr{H}$ = real intervals, positive when $x \in (a, b)$
  - $\mathscr{H}$ = 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