10-301/601: Introduction to Machine Learning Lecture 16 – Learning Theory (Infinite Case)

Henry Chai

10/26/22

Q & A:

Why is the answer C?

- Let $\mathcal{H}$ be the set of all conjunctions over M Boolean variables, $\boldsymbol{x} \in \{0,1\}^M$; examples of conjunctions are
  - $h(\boldsymbol{x}) = x_1(1 - x_2)x_4x_{10}$
  - $h(\boldsymbol{x}) = (1 - x_3)(1 - x_4)x_8$

$|H| = 3^{10}$

$\neq 10^3$

- Assuming $c^* \in \mathcal{H}$, if $M = 10$, $\epsilon = 0.1$, and $\delta = 0.01$, at least how many labelled examples do we need to satisfy the PAC criterion using Theorem 1?

A. 1 **(TOXIC)**

B. $10(2 \ln 10 + \ln 100) \approx 92$

C. $10(3 \ln 10 + \ln 100) \approx 116$

D. $10(10 \ln 2 + \ln 100) \approx 116$

E. $10(10 \ln 3 + \ln 100) \approx 156$

F. $100(2 \ln 10 + \ln 10) \approx 691$

G. $100(3 \ln 10 + \ln 10) \approx 922$

H. $100(10 \ln 2 + \ln 10) \approx 924$

I. $100(10 \ln 3 + \ln 10) \approx 1329$

Q & A:

How does the statistical learning theory corollary follow from this theorem?

- 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

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

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

Q & A:

How does the statistical learning theory corollary follow from this theorem?

- 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

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

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

- Solving for ϵ gives...

Q & A:

How does the statistical learning theory corollary follow from this theorem?

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

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

with probability at least $1 - \delta$.

Front Matter

- Announcements
  - HW5 released 10/13, due 10/27 (tomorrow) at 11:59 PM
  - HW6 released 10/27 (tomorrow), due 11/4 at 11:59 PM
    - Only two late days allowed on HW6
  - Exam 2 on 11/10, two weeks from tomorrow (more details to follow)
    - All topics between Lecture 8 and Lecture 17 (next Monday's lecture) are in-scope
    - Exam 1 content may be referenced but will not be the primary focus of any question
  - Exam 3 scheduled
    - Thursday, December 15^{th} from 9:30 AM to 11:30 AM
  - Sign up for peer tutoring! See Piazza for more details

Recall - 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

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

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

Recall - Theorem 2: Finite, Agnostic Case

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

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

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

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

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

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

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

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

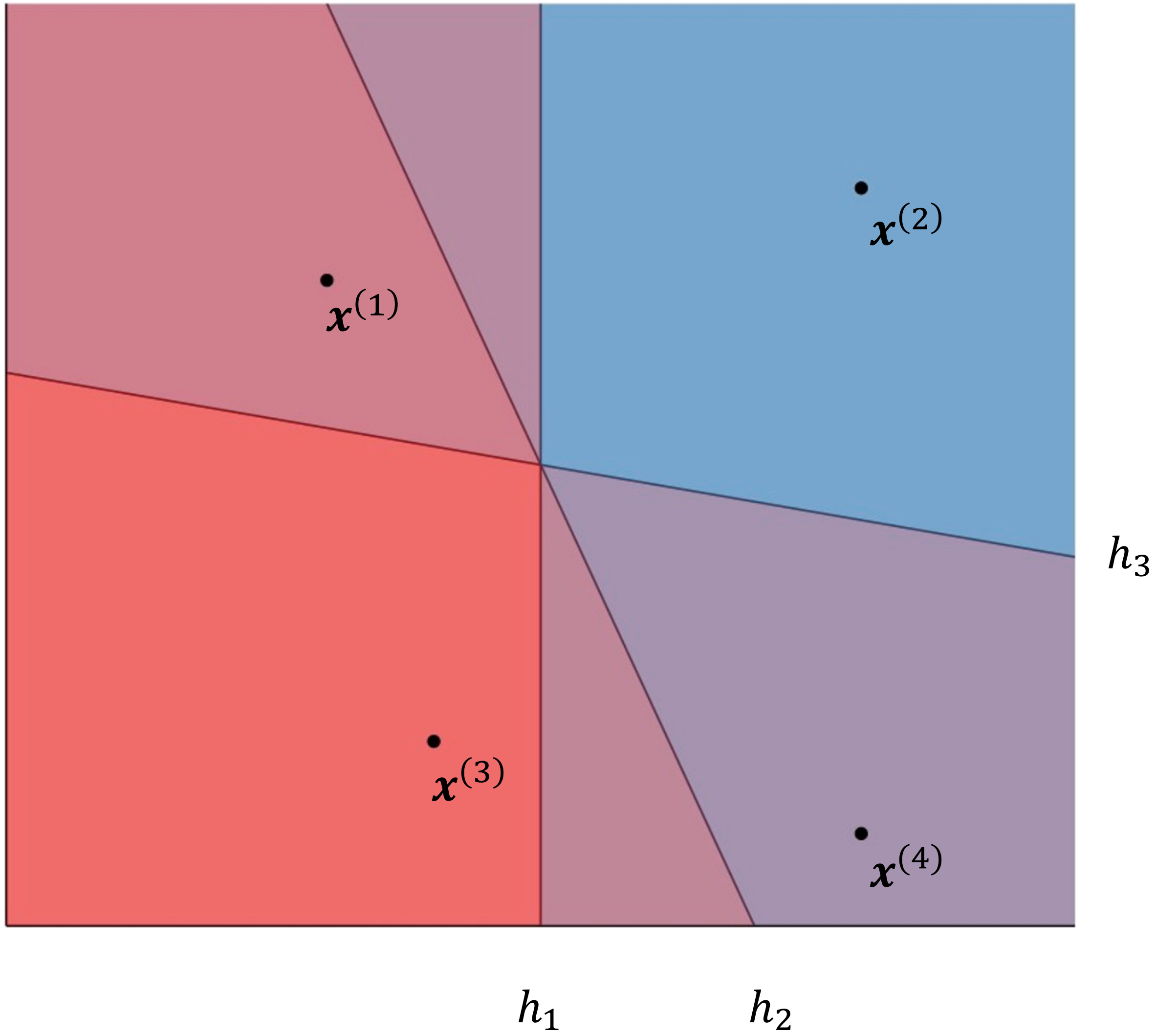
- Insight: $|\mathcal{H}|$ measures how complex our hypothesis set is
- Idea: define a different measure of hypothesis set complexity

Labellings

- Given some finite set of data points $S = \left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}\right)$ and some hypothesis $h \in \mathcal{H}$, applying h to each point in S results in a **labelling**
  - $\left(h\left(\boldsymbol{x}^{(1)}\right), \ldots, h\left(\boldsymbol{x}^{(M)}\right)\right)$ is a vector of M +1's and -1's
  - **Important note:** our discussion of PAC learning assumes binary classification
- Given $S = \left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}\right)$, each hypothesis in $\mathcal{H}$ induces a labelling but not necessarily a unique labelling
  - The set of labellings induced by $\mathcal{H}$ on S is
    $\mathcal{H}(S) = \left\{\left(h\left(\boldsymbol{x}^{(1)}\right), \ldots, h\left(\boldsymbol{x}^{(M)}\right)\right) \middle| h \in \mathcal{H}\right\}$

Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

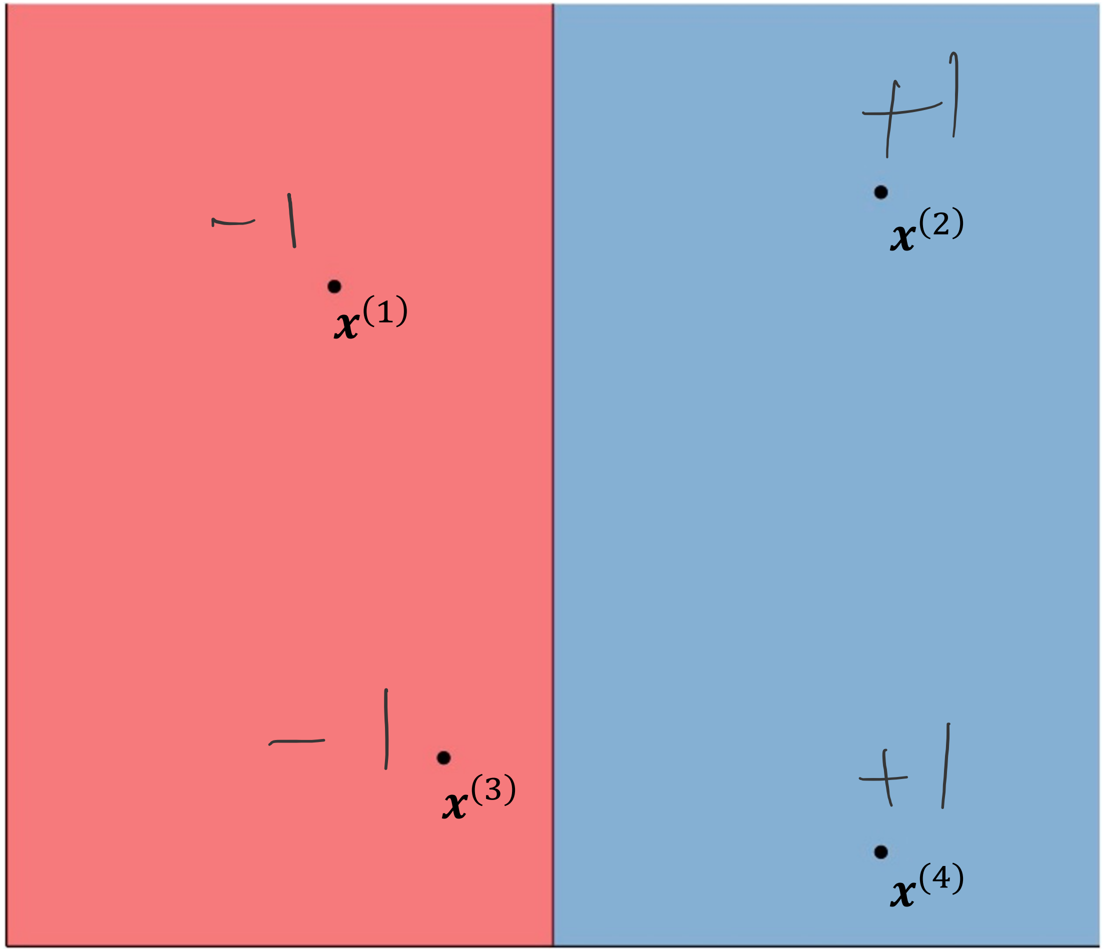


Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$$\left(h_1\left(\boldsymbol{x}^{(1)}\right), h_1\left(\boldsymbol{x}^{(2)}\right), h_1\left(\boldsymbol{x}^{(3)}\right), h_1\left(\boldsymbol{x}^{(4)}\right)\right)$$

$$= \underbrace{(-1, +1, -1, +1)}$$

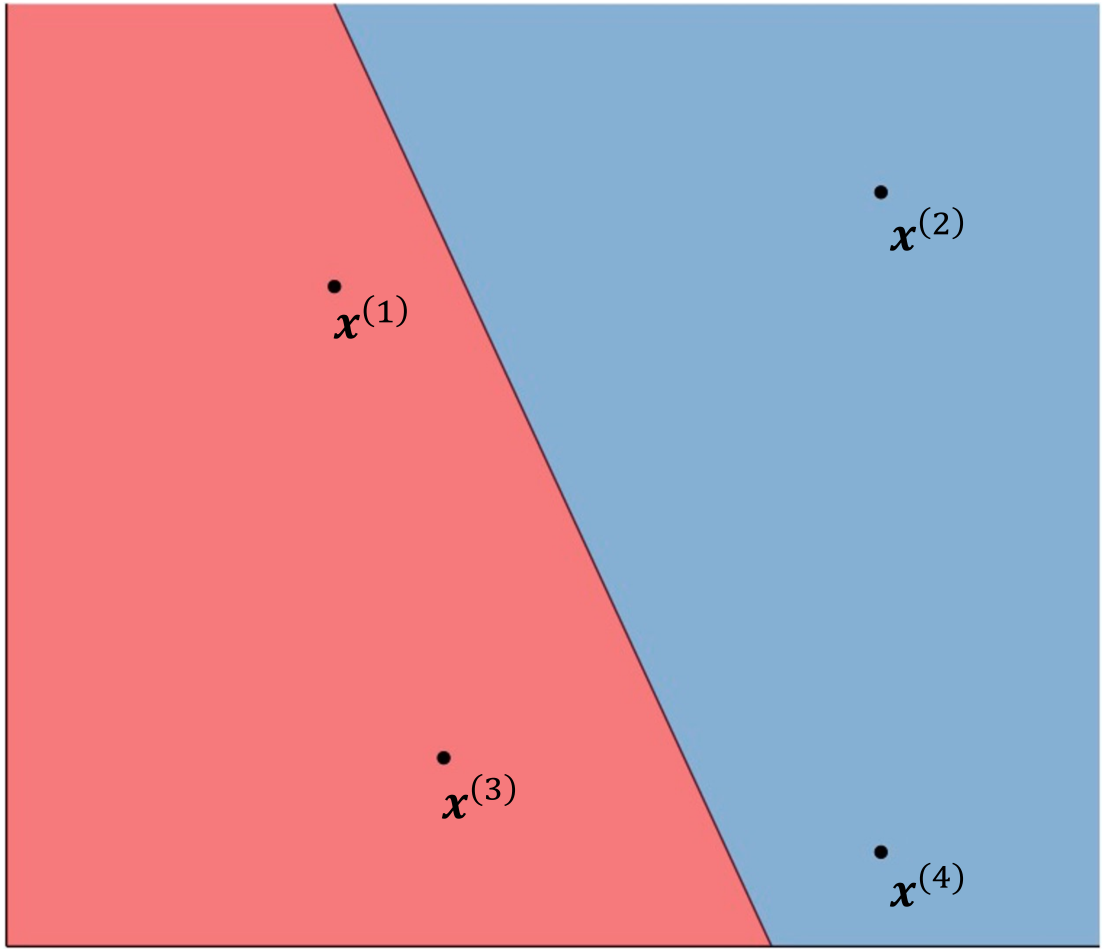


h_1

Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$$\left(h_2\left(\boldsymbol{x}^{(1)}\right), h_2\left(\boldsymbol{x}^{(2)}\right), h_2\left(\boldsymbol{x}^{(3)}\right), h_2\left(\boldsymbol{x}^{(4)}\right)\right)$$
$$= (-1, +1, -1, +1)$$

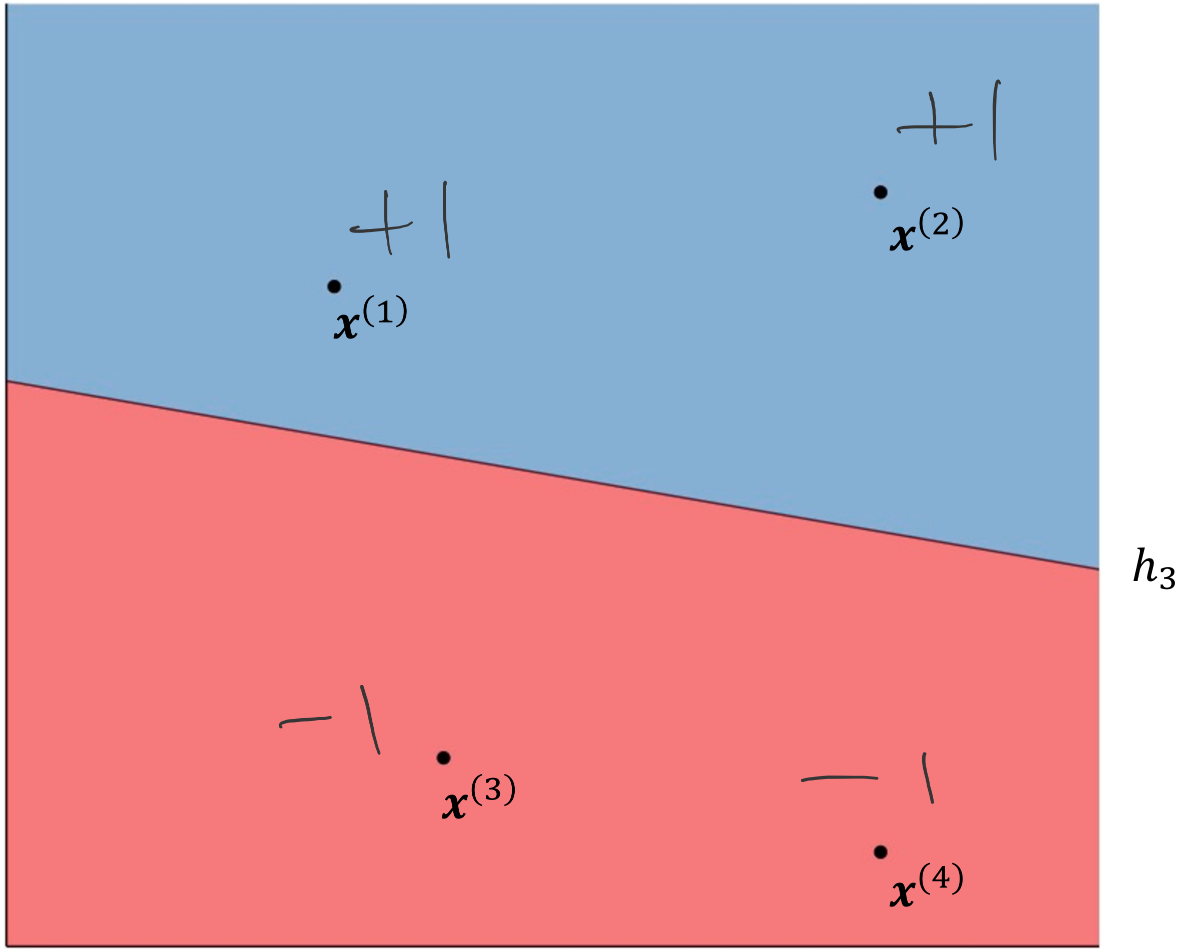


h_2

Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$$\left(h_3\left(\boldsymbol{x}^{(1)}\right), h_3\left(\boldsymbol{x}^{(2)}\right), h_3\left(\boldsymbol{x}^{(3)}\right), h_3\left(\boldsymbol{x}^{(4)}\right)\right)$$
$$= \underbrace{(+1, +1, -1, -1)}$$

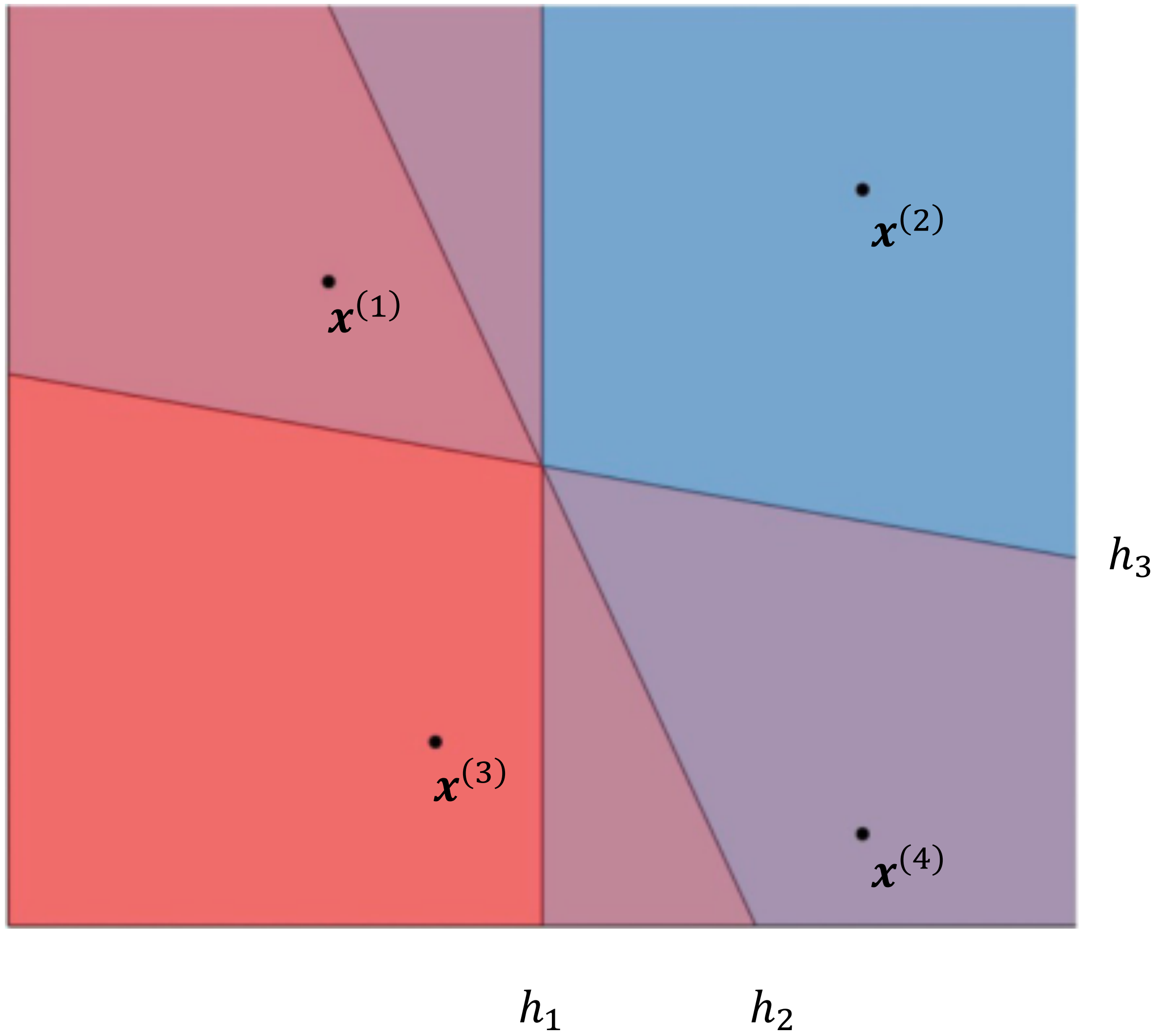


Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$\mathcal{H}(S)$
$= \{(+1, +1, -1, -1), (-1, +1, -1, +1)\}$

$|\mathcal{H}(S)| = 2$

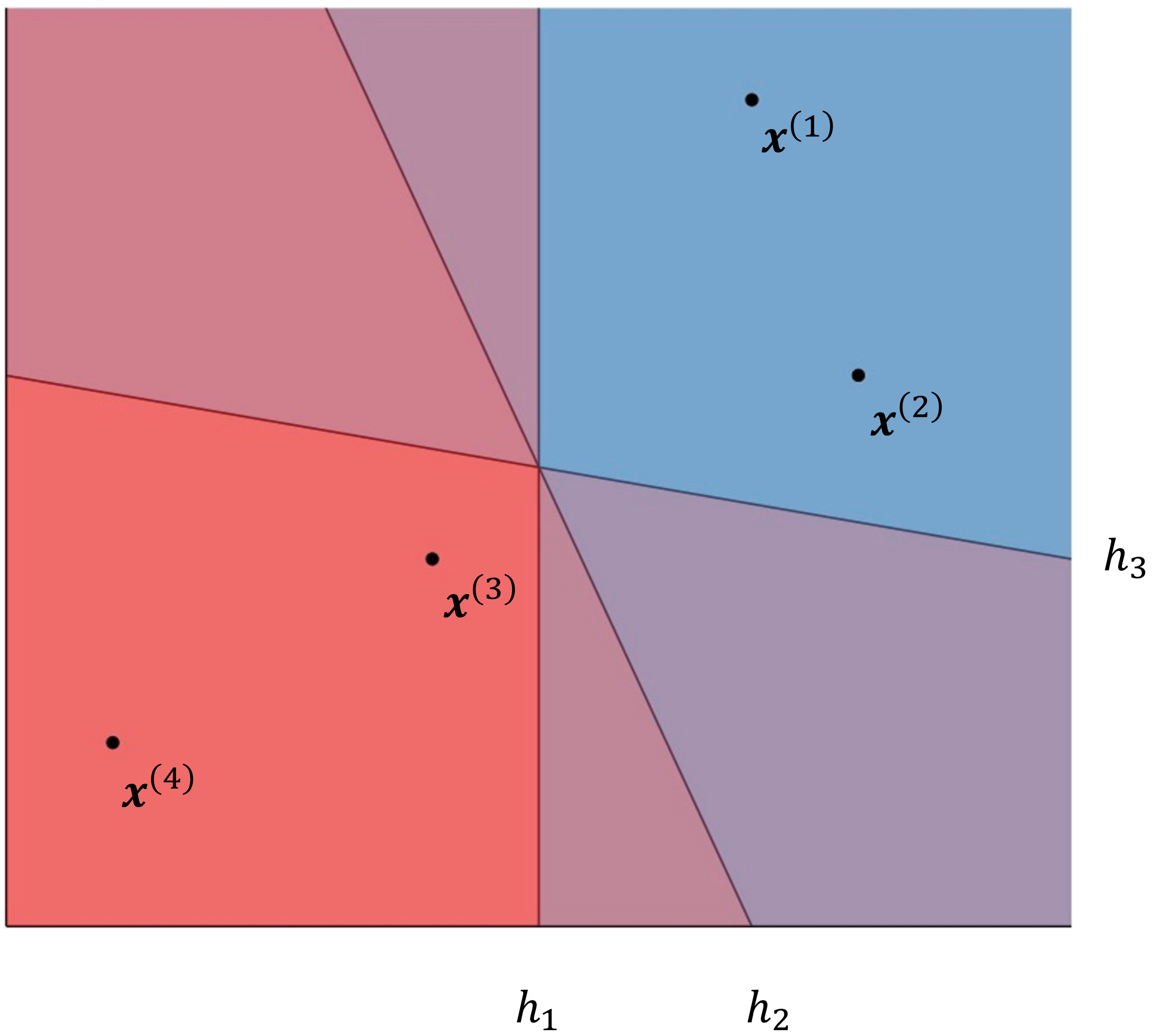


Example: Labellings

$\mathcal{H} = \{h_1, h_2, h_3\}$

$\mathcal{H}(S) = \{(+1, +1, -1, -1)\}$

$|\mathcal{H}(S)| = 1$



VC-Dimension

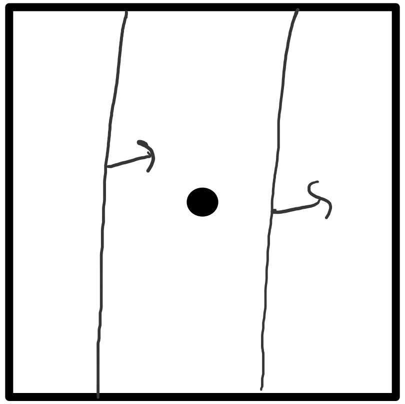
you get to pick S!

- $\mathcal{H}(S)$ is the set of all labellings induced by $\mathcal{H}$ on S
  - If $|S| = M$, then $|\mathcal{H}(S)| \leq 2^M$
  - $\mathcal{H}$ **shatters** S if $|\mathcal{H}(S)| = 2^M$
- The **VC-dimension** of $\mathcal{H}$, $VC(\mathcal{H})$, is the size of the largest set S that can be shattered by $\mathcal{H}$.
  - If $\mathcal{H}$ can shatter arbitrarily large finite sets, then $VC(\mathcal{H}) = \infty$
- To prove that $VC(\mathcal{H}) = d$, you 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

you don't get to pick !!

VC-Dimension: Example

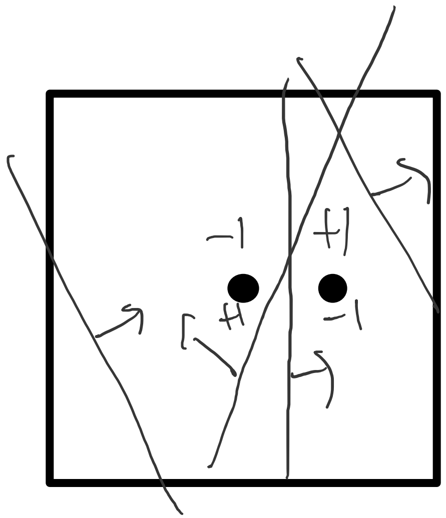
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?



S

VC-Dimension: Example

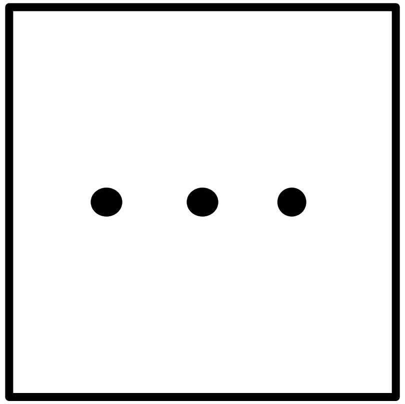
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
    - Can $\mathcal{H}$ shatter some set of 1 point? ✓
    - Can $\mathcal{H}$ shatter some set of 2 points?



S

VC-Dimension: Example

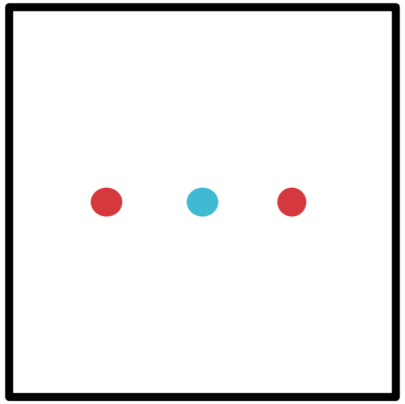
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point? ✓
  - Can $\mathcal{H}$ shatter some set of 2 points? ✓
  - Can $\mathcal{H}$ shatter some set of 3 points?



S

VC-Dimension: Example

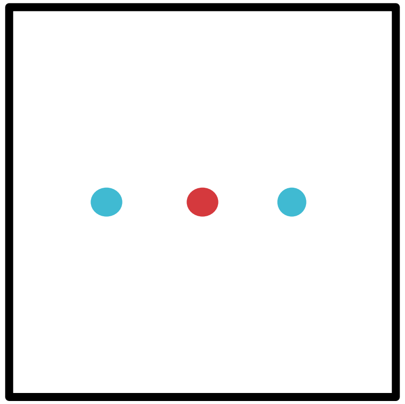
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



S

VC-Dimension: Example

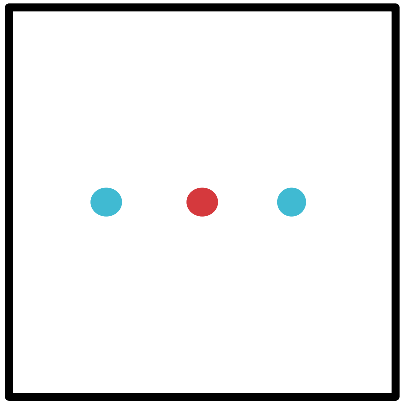
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



S

VC-Dimension: Example

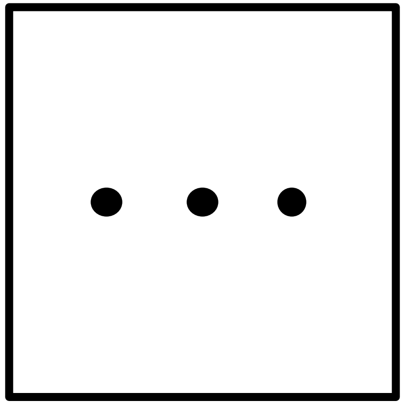
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter ***some*** set of 3 points?



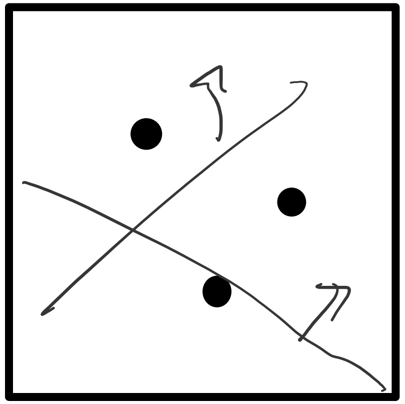
S

VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



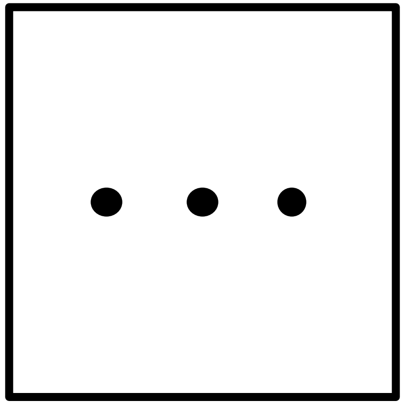
S_1



S_2

VC-Dimension: Example

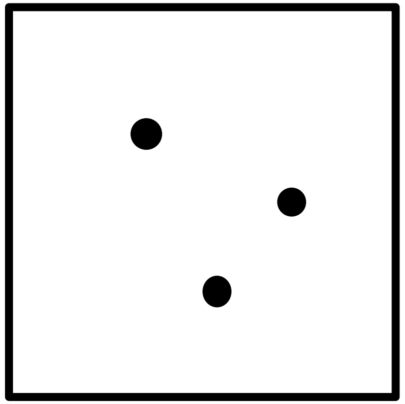
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



$|\mathcal{H}(S_1)| = 6$

$= 8 - 2$

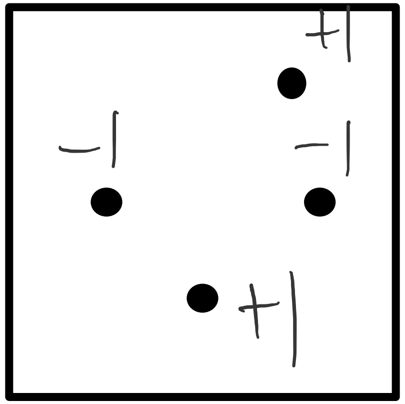
$< 8 = 2^3$



$|\mathcal{H}(S_2)| = 8$

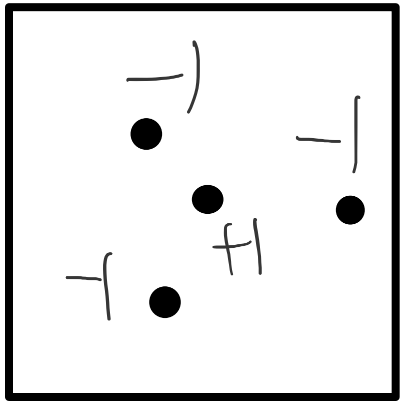
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
    - Can $\mathcal{H}$ shatter some set of 1 point? ✓
    - Can $\mathcal{H}$ shatter some set of 2 points? ✓
    - Can $\mathcal{H}$ shatter some set of 3 points? ✓
    - Can $\mathcal{H}$ shatter some set of 4 points?



S_1

All points on the convex hull

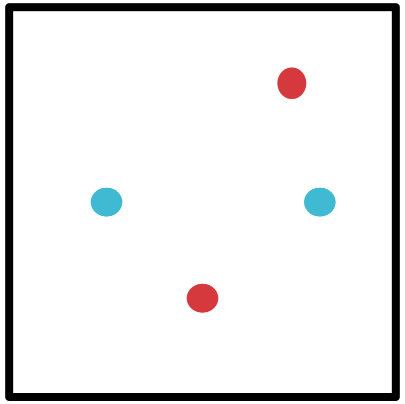


S_2

At least one point inside the convex hull

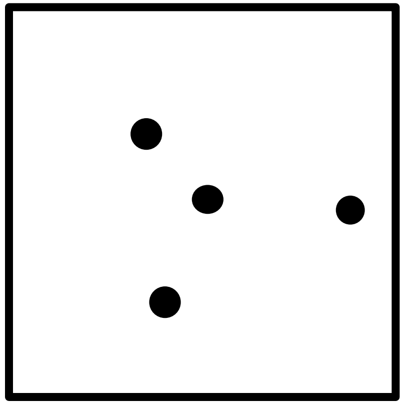
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



S_1

All points on the convex hull

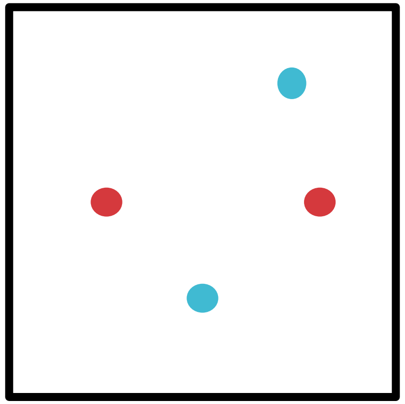


S_2

At least one point inside the convex hull

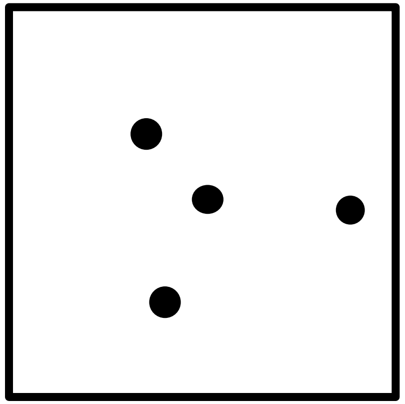
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



S_1

All points on the convex hull

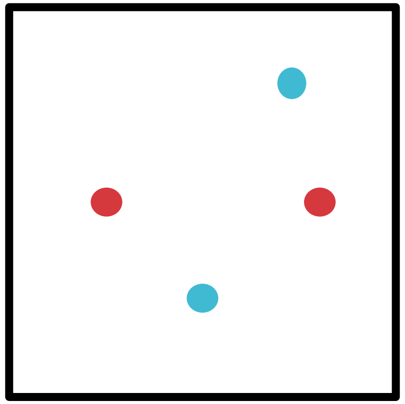


S_2

At least one point inside the convex hull

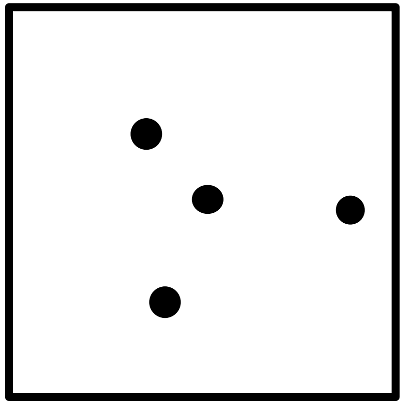
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



$|\mathcal{H}(S_1)| = 14$

All points on the convex hull

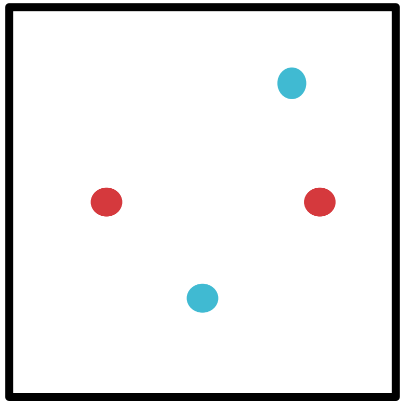


S_2

At least one point inside the convex hull

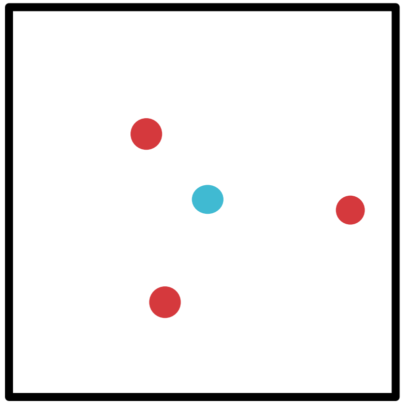
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ $=$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



$|\mathcal{H}(S_1)| = 14$

All points on the convex hull

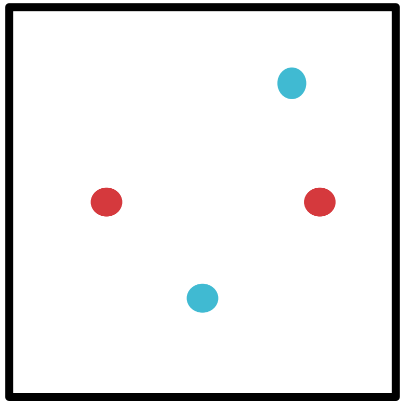


S_2

At least one point inside the convex hull

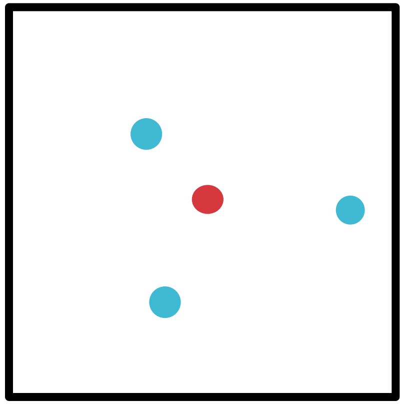
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
    - Can $\mathcal{H}$ shatter some set of 1 point?
    - Can $\mathcal{H}$ shatter some set of 2 points?
    - Can $\mathcal{H}$ shatter some set of 3 points?
    - Can $\mathcal{H}$ shatter some set of 4 points?



$|\mathcal{H}(S_1)| = 14$

All points on the convex hull

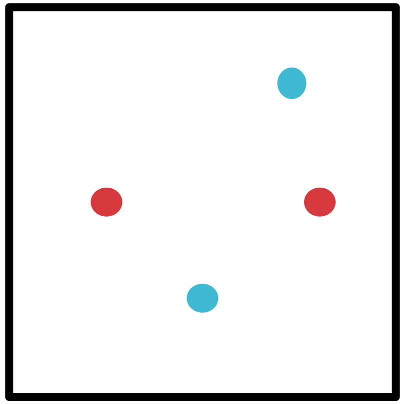


S_2

At least one point inside the convex hull

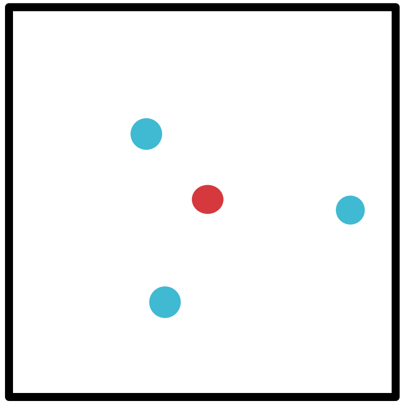
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



$|\mathcal{H}(S_1)| = 14 < 2^4$

All points on the convex hull

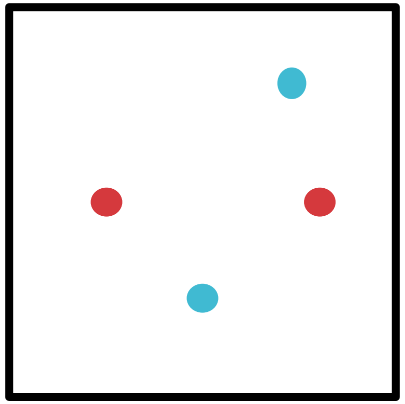


$|\mathcal{H}(S_2)| = 14$

At least one point inside the convex hull

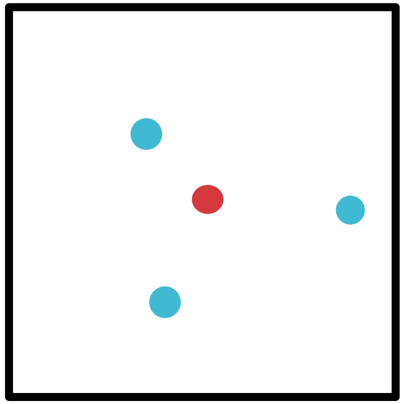
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- $VC(\mathcal{H}) = 3$
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?
  - Can $\mathcal{H}$ shatter some set of 4 points?



$|\mathcal{H}(S_1)| = 14$

All points on the convex hull



$|\mathcal{H}(S_2)| = 14$

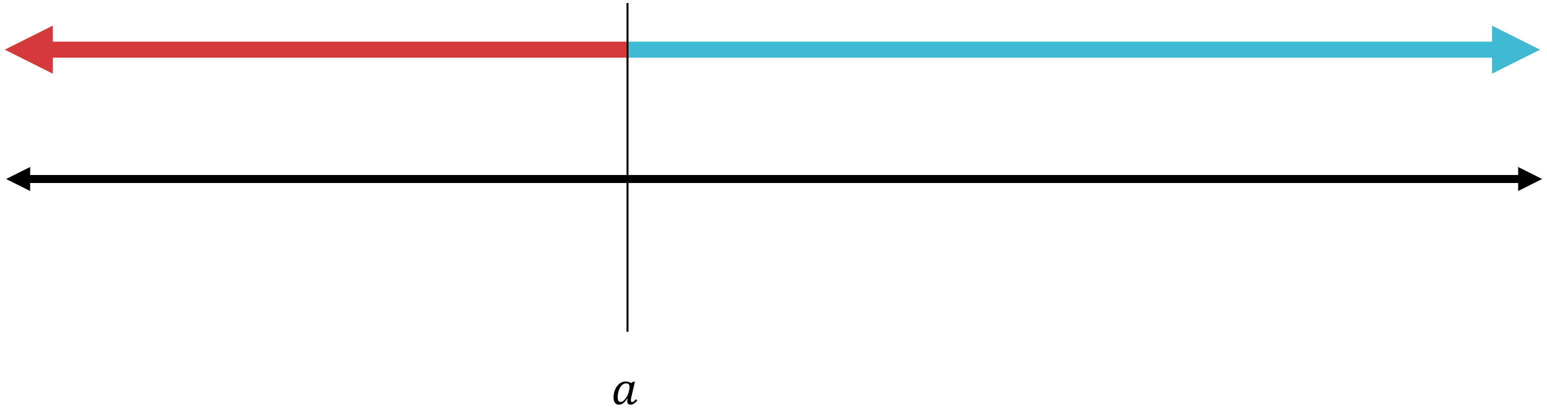
At least one point inside the convex hull

VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all d-dimensional linear separators
- $VC(\mathcal{H}) = d + 1$

VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$

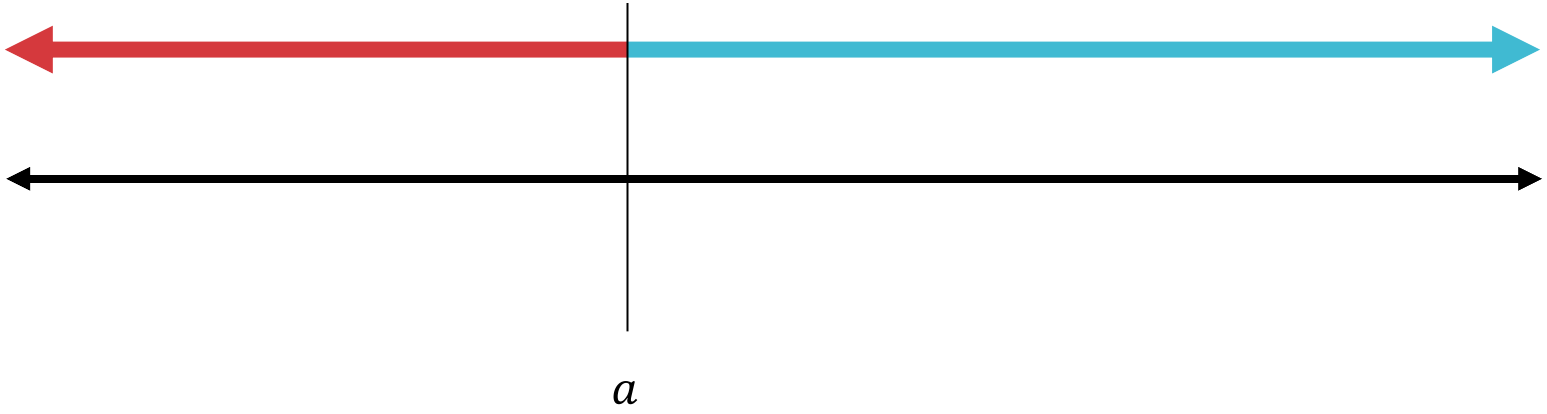


Poll Question 1:

What is $VC(\mathcal{H})$?

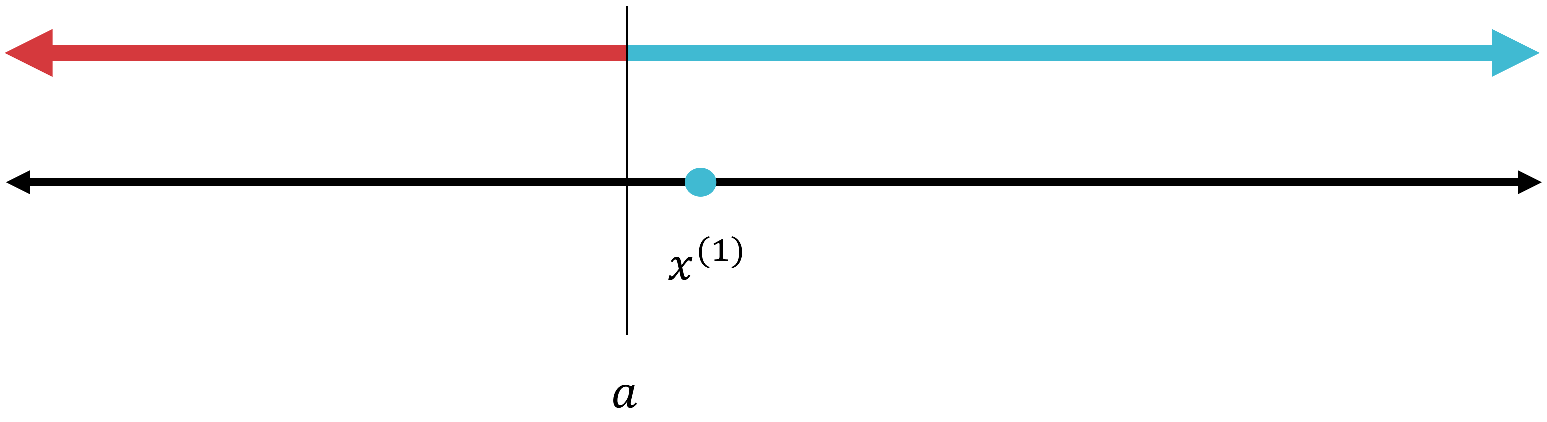
A. -1 **(TOXIC)**
B. 0
C. 1
D. 2
E. 3

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



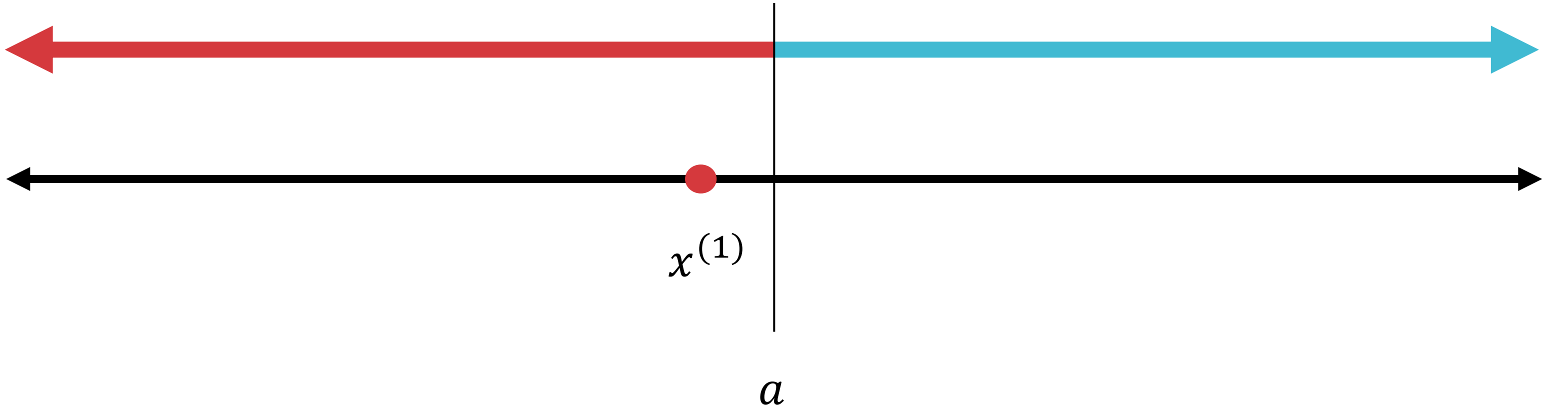
VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



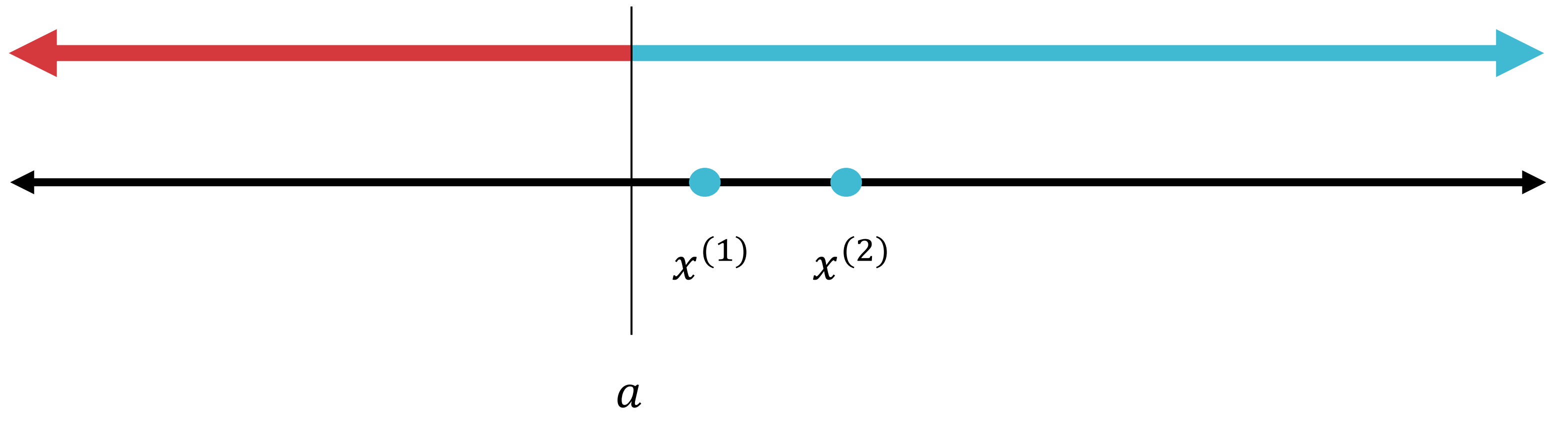
VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



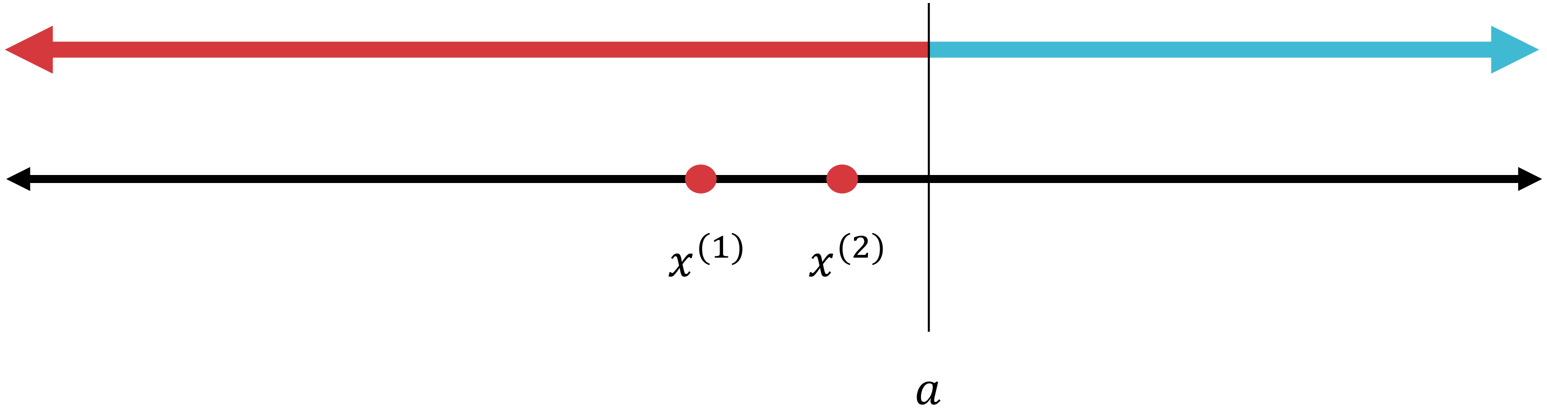
VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



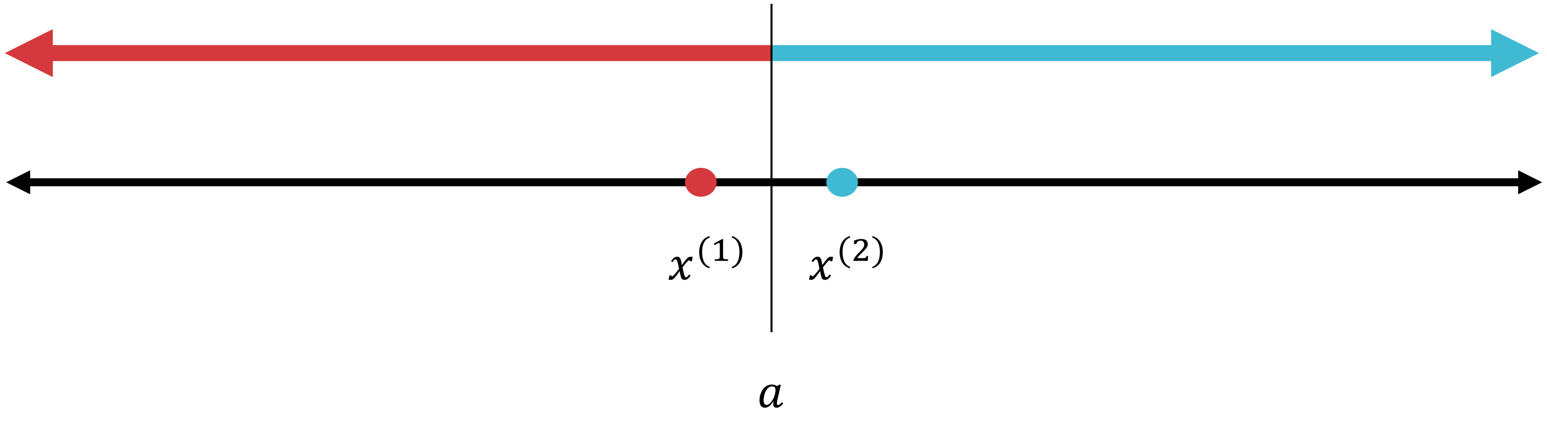
VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



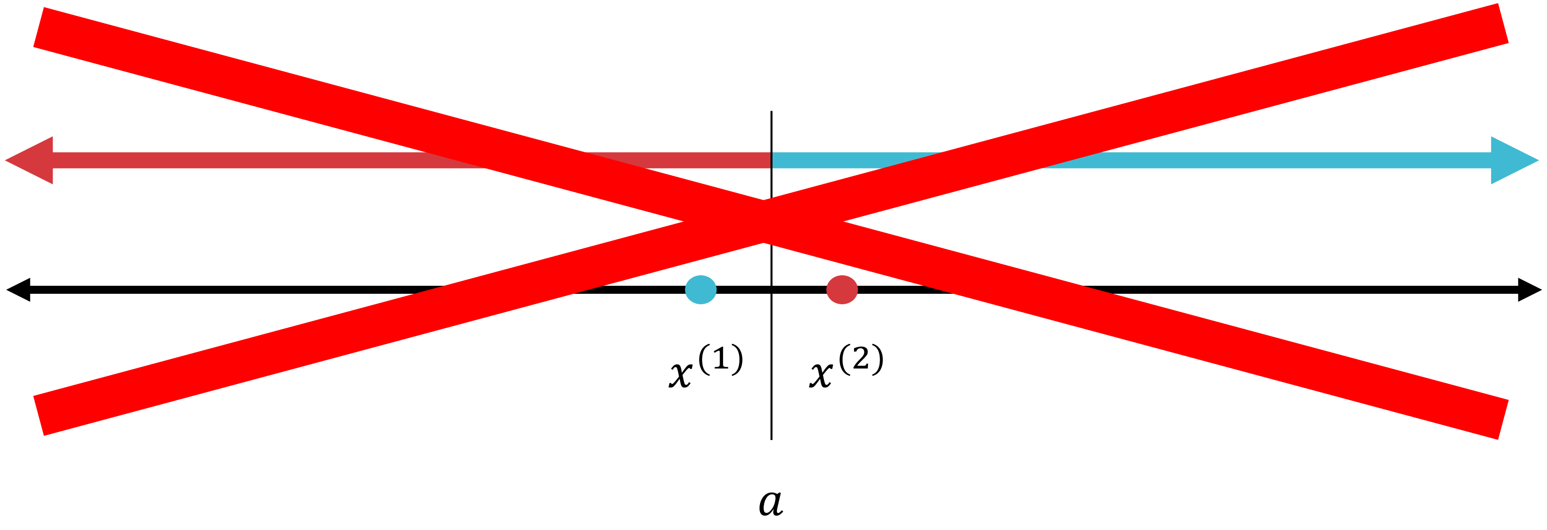
VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



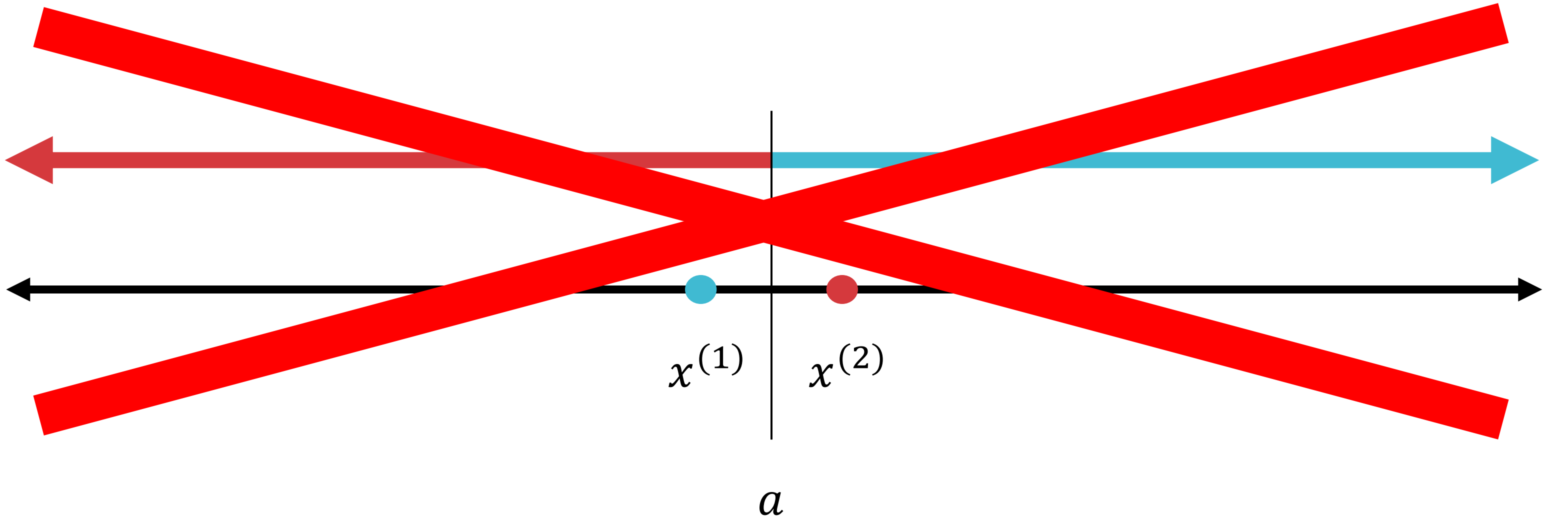
VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



VC-Dimension: Example

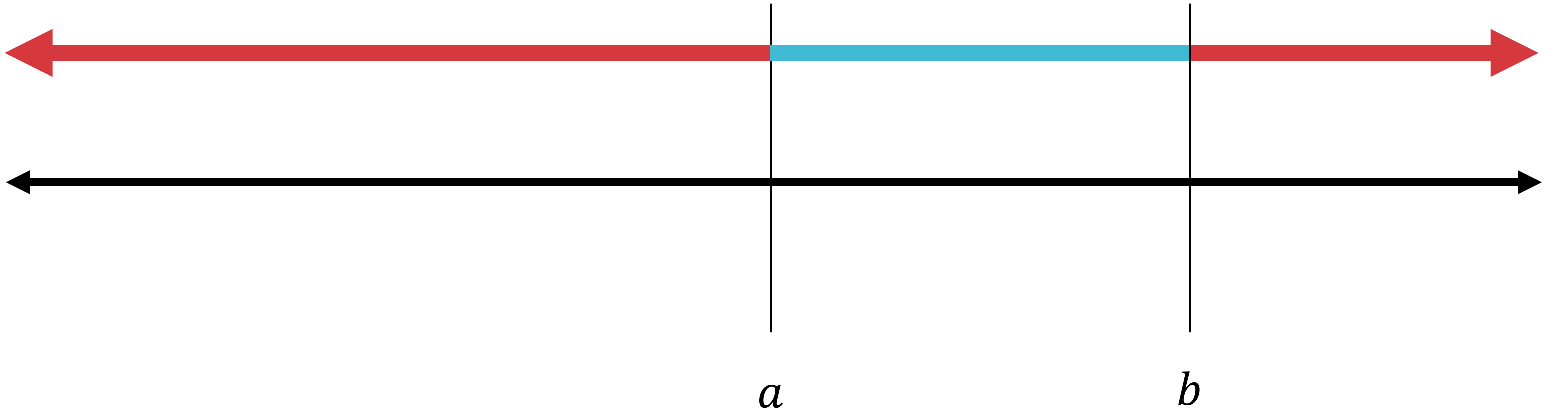
- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- $VC(\mathcal{H}) = 1$

VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive intervals

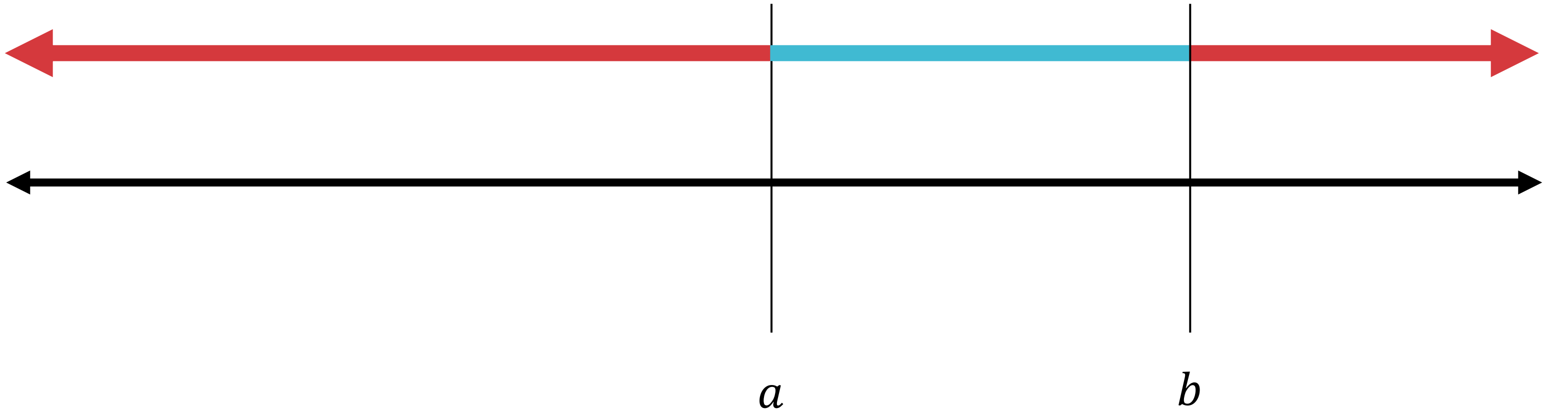


Poll Question 2:

What is $VC(\mathcal{H})$?

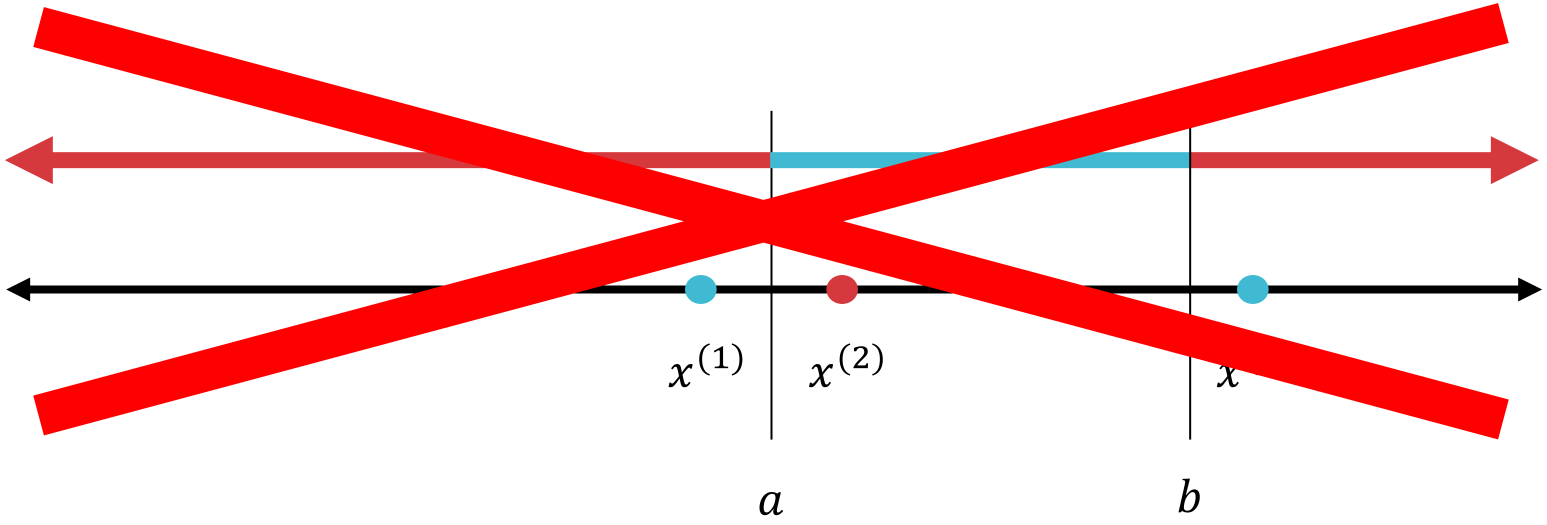
A. 0
B. 1
C. 1.5 **(TOXIC)**
D. 2
E. 3

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive intervals



VC-Dimension: Example

- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive intervals



- $VC(\mathcal{H}) = 2$

Theorem 3: Vapnik-Chervonenkis (VC)-Bound

where $c^* \in \mathcal{H}$

- Infinite, realizable case: for any hypothesis set $\mathcal{H}$ and distribution p^*, if the number of labelled training data points satisfies

instead of $|H|$

$$M = \underset{\sim}{O}\left(\frac{1}{\epsilon}\underbrace{\left(VC(\mathcal{H})\log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

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

Statistical Learning Theory Corollary 3

- Infinite, realizable case: for any hypothesis set $\mathcal{H}$ and distribution p^*, given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have

$$R(h) \leq O\left(\frac{1}{M}\left(VC(\mathcal{H}) \log\left(\frac{M}{VC(\mathcal{H})}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$$

with probability at least $1 - \delta$.

Theorem 4: Vapnik-Chervonenkis (VC)-Bound

- Infinite, agnostic case: for any hypothesis set $\mathcal{H}$ and distribution p^*, if the number of labelled training data points satisfies

$$M = O\left(\frac{1}{\epsilon^2}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)\right)$$

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

Statistical Learning Theory Corollary 4

- Infinite, agnostic case: for any hypothesis set $\mathcal{H}$ and distribution p^*, given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ have

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{M}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

with probability at least $1 - \delta$.

Approximation Generalization Tradeoff

How well does h generalize?

$$R(h) \leq \underbrace{\hat{R}(h)}_{\text{How well does } h \text{ approximate } c^*?} + O\left(\overbrace{\sqrt{\frac{1}{M}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}}^{\text{How well does } h \text{ generalize?}}\right)$$

How well does h approximate c^*?

Approximation Generalization Tradeoff

$$R(h) \leq \underbrace{\hat{R}(h)}_{\text{Decreases as } VC(\mathcal{H}) \text{ increases}} + O\left(\overbrace{\sqrt{\frac{1}{M}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}}^{\text{Increases as } VC(\mathcal{H}) \text{ increases}}\right)$$

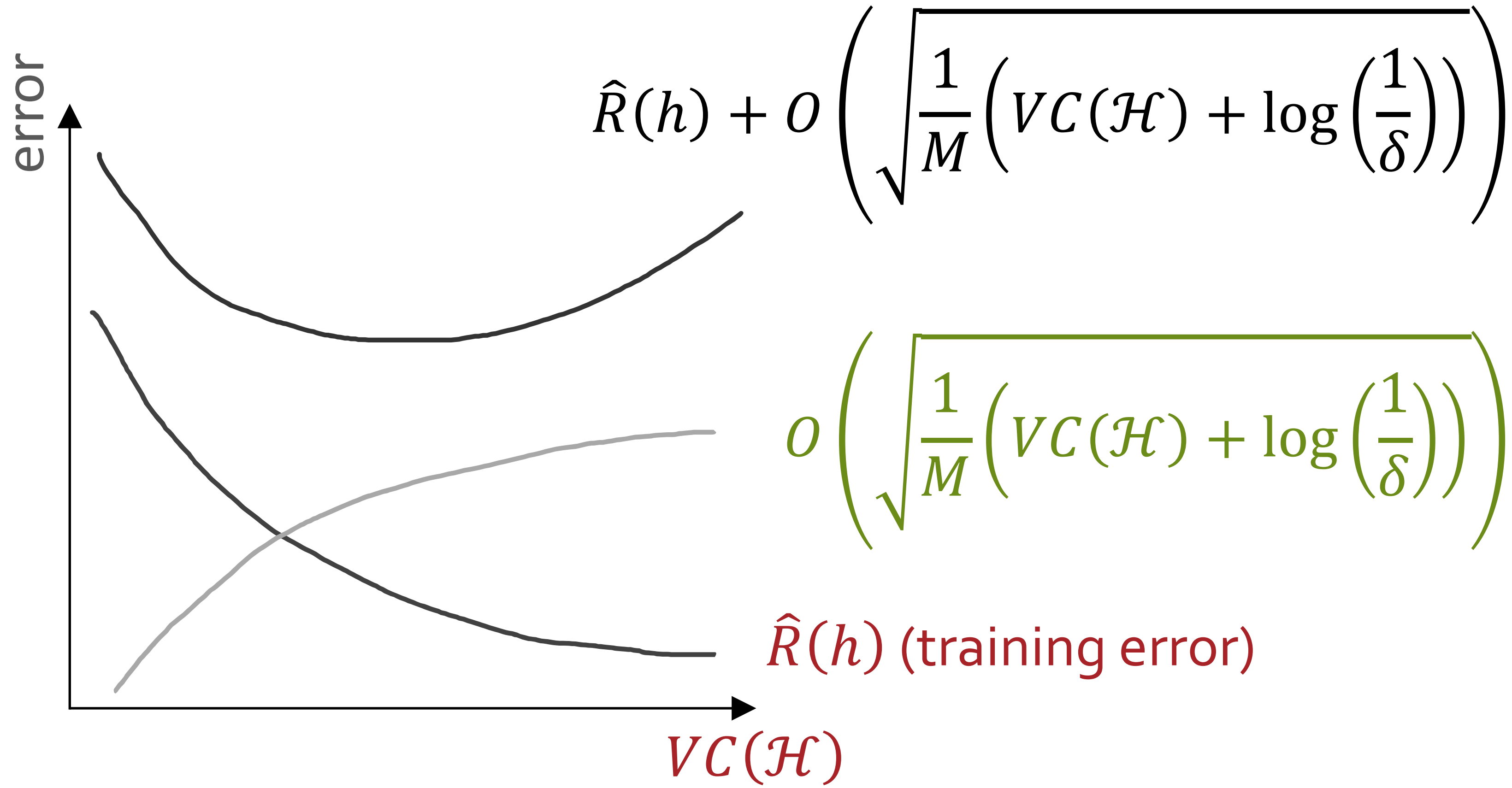
Can we use this corollary to guide model selection?

- Infinite, agnostic case: for any hypothesis set $\mathcal{H}$ and distribution p^*, given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ have

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{M}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

with probability at least $1 - \delta$.

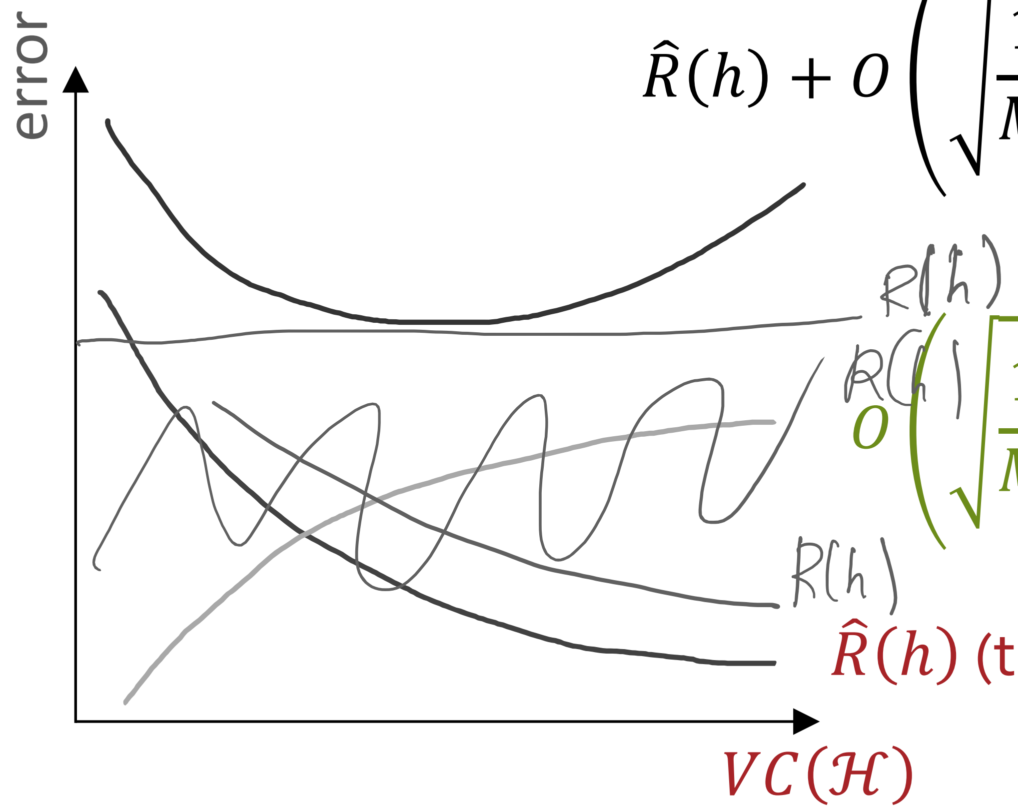
Learning Theory and Model Selection



- How can we find this "best tradeoff" for linear separators?
- Use a regularizer! By (effectively) reducing the number of features our model considers, we reduce its VC-dimension.

Learning Theory and Model Selection

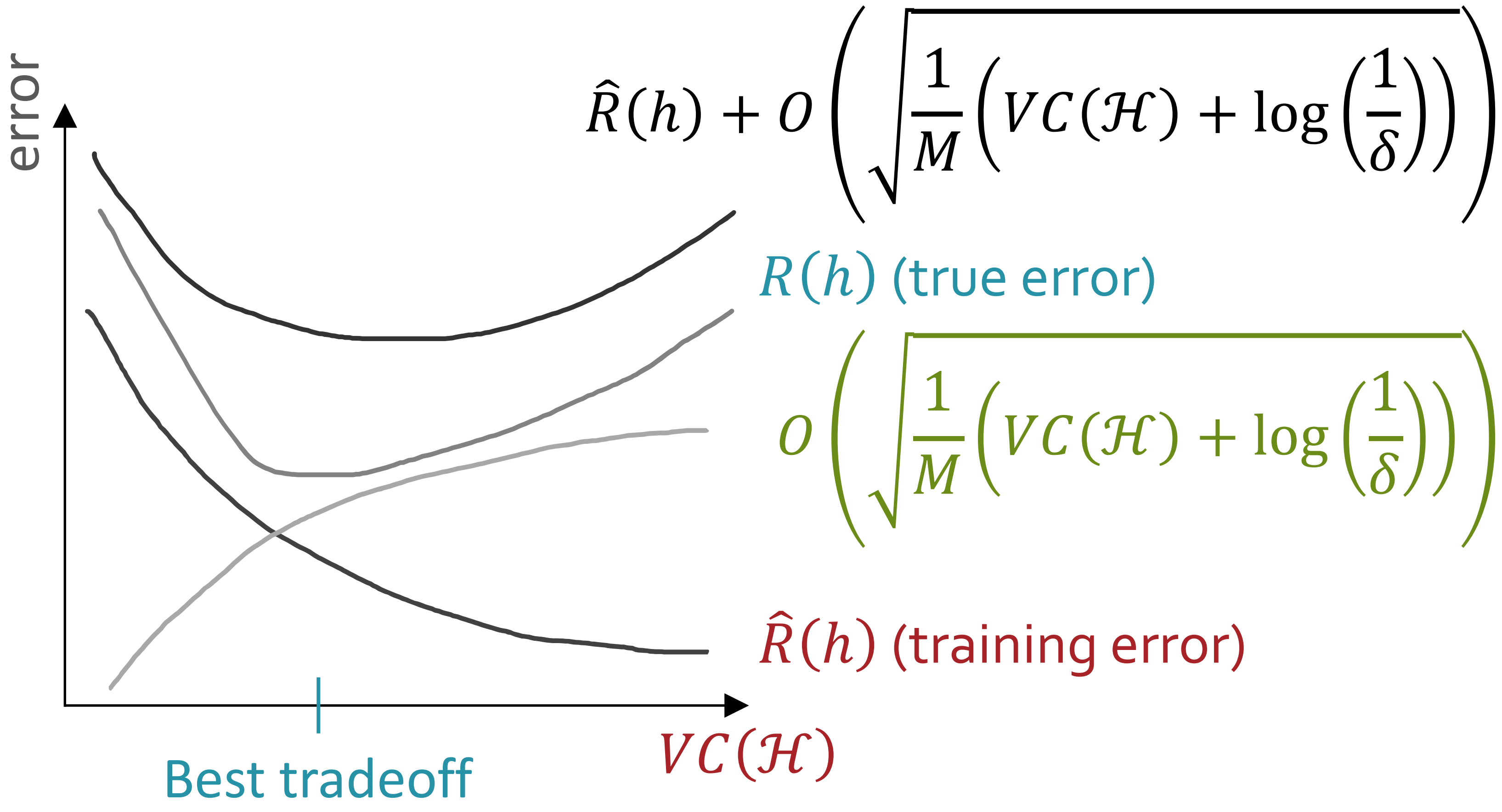
$$\hat{R}(h) + O\left(\sqrt{\frac{1}{M}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$



$$O\left(\sqrt{\frac{1}{M}\left(VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

$\hat{R}(h)$ (training error)

Learning Theory and Model Selection



- How can we find this "best tradeoff" for linear separators?
- Use a regularizer! By (effectively) reducing the number of features our model considers, we reduce its VC-dimension.

Learning Theory Learning Objectives

You should be able to…

- Identify the properties of a learning setting and assumptions required to ensure low generalization error
- Distinguish true error, train error, test error
- Define PAC and explain what it means to be approximately correct and what occurs with high probability
- Apply sample complexity bounds to real-world machine learning examples
- Theoretically motivate regularization

Poll Question 3:

What questions do you have?

You should be able to…

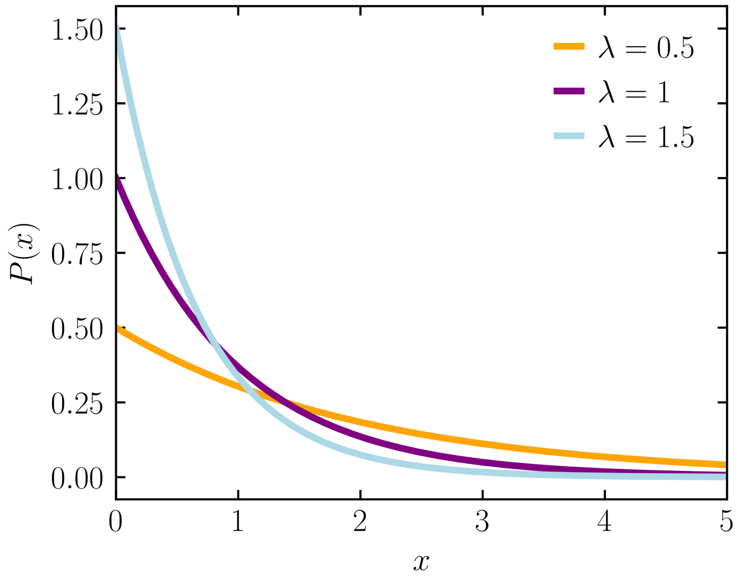
- Identify the properties of a learning setting and assumptions required to ensure low generalization error
- Distinguish true error, train error, test error
- Define PAC and explain what it means to be approximately correct and what occurs with high probability
- Apply sample complexity bounds to real-world machine learning examples
- Theoretically motivate regularization

Recall: Probabilistic Learning

- Previously:
  - (Unknown) Target function, $c^*: \mathcal{X} \rightarrow \mathcal{Y}$
  - Classifier, $h : \mathcal{X} \rightarrow \mathcal{Y}$
  - Goal: find a classifier, h, that best approximates c^*
- Now:
  - (Unknown) Target *distribution*, $y \sim p^*(Y|\boldsymbol{x})$
  - Distribution, $p(Y|\boldsymbol{x})$
  - Goal: find a distribution, p, that best approximates p^*

Recall: Maximum Likelihood Estimation (MLE)

- Insight: every valid probability distribution has a finite amount of probability mass as it must sum/integrate to 1
- Idea: set the parameter(s) so that the likelihood of the samples is maximized
- Intuition: assign as much of the (finite) probability mass to the observed data *at the expense of unobserved data*
- Example: the exponential distribution



Source: https://en.wikipedia.org/wiki/Exponential_distribution#/media/File:Exponential_probability_density.svg

Bernoulli Distribution MLE

- A Bernoulli random variable takes value 1 with probability ϕ and value 0 with probability $1-\phi$
- The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^x(1-\phi)^{1-x}$$

Coin Flipping MLE

- A Bernoulli random variable takes value 1 (or heads) with probability ϕ and value 0 (or tails) with probability $1-\phi$
- The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^x (1-\phi)^{1-x}$$

- Given N iid samples $\{x^{(1)}, \dots, x^{(N)}\}$, the log-likelihood is

$$\ell(\phi) = \log\left(\prod_{i=1}^{N} p(x^{(i)}; \phi)\right)$$

$$= \sum_{i=1}^{N} \log\left(p(x^{(i)}; \phi)\right)$$

$$= \sum_{i=1}^{N} \log\left(\phi^{x^{(i)}} (1-\phi)^{1-x^{(i)}}\right)$$

$$= \sum_{i=1}^{N} x^{(i)} \log(\phi) + (1-x^{(i)}) \log(1-\phi)$$

$$= N_1 \log(\phi) + N_0 \log(1-\phi)$$

where N_i = # of i's in my dataset

Coin Flipping MLE

- A Bernoulli random variable takes value 1 (or heads) with probability ϕ and value 0 (or tails) with probability $1-\phi$
- The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^x(1-\phi)^{1-x}$$

- The partial derivative of the log-likelihood is

$$\ell(\phi) = N_1 \log(\phi) + N_0 \log(1-\phi)$$

$$\frac{\partial \ell}{\partial \phi} = \frac{N_1}{\phi} + \frac{N_0}{1-\phi}(-1)$$

$$\Rightarrow \frac{N_1}{\hat{\phi}} - \frac{N_0}{1-\hat{\phi}} = 0 \Rightarrow \frac{N_1}{\hat{\phi}} = \frac{N_0}{1-\hat{\phi}}$$

$$\Rightarrow N_1(1-\hat{\phi}) = N_0\hat{\phi} \Rightarrow N_1 = (N_1+N_0)\hat{\phi}$$

$$\Rightarrow \hat{\phi} = N_1 / N_1 + N_0$$

Coin Flipping MLE

- A Bernoulli random variable takes value 1 (or heads) with probability ϕ and value 0 (or tails) with probability $1-\phi$
- The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^x (1-\phi)^{1-x}$$

- The partial derivative of the log-likelihood is

$$\frac{N_1}{\hat{\phi}} - \frac{N_0}{1-\hat{\phi}} = 0 \rightarrow \frac{N_1}{\hat{\phi}} = \frac{N_0}{1-\hat{\phi}}$$

$$\rightarrow N_1\left(1-\hat{\phi}\right) = N_0\hat{\phi} \rightarrow N_1 = \hat{\phi}(N_0 + N_1)$$

$$\rightarrow \hat{\phi} = \frac{N_1}{N_0 + N_1}$$

- where N_1 is the number of 1's in $\{x^{(1)}, \dots, x^{(N)}\}$ and N_0 is the number of 0's