10-301/601: Introduction to Machine Learning Lecture 15 – Learning Theory (Infinite Case)

Henry Chai & Matt Gormley

10/23/23

Front Matter

- Announcements
  - HW5 released 10/9, due 10/27 (Friday) at 11:59 PM
  - Exam 3 scheduled
    - Tuesday, December 12th from 5:30 PM to 8:30 PM
  - Sign up for peer tutoring! See Piazza for more details

Recall - Theorem 1: Finite, Realizable Case

- For a *finite* hypothesis set $\mathcal{H}$ such that $c^* \in \mathcal{H}$ (*realizable*) 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 1: Finite, Realizable Case

- For a *finite* hypothesis set $\mathcal{H}$ such that $c^* \in \mathcal{H}$ (*realizable*) 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 $R(h) \leq \epsilon$

- Making the bound tight and solving for ϵ gives…

Statistical Learning Theory Corollary

- For a *finite* hypothesis set $\mathcal{H}$ such that $c^* \in \mathcal{H}$ (*realizable*) and arbitrary distribution p^*, given a training dataset S where $|S| = M$, 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 $1 - \delta$.

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

Statistical Learning Theory Corollary

- For a *finite* hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, given a training dataset S where $|S| = M$, all $h \in \mathcal{H}$ have

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

with probability at least $1 - \delta$.

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

- For a *finite* hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, given a training data set S where $|S| = M$, all $h \in \mathcal{H}$ have

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

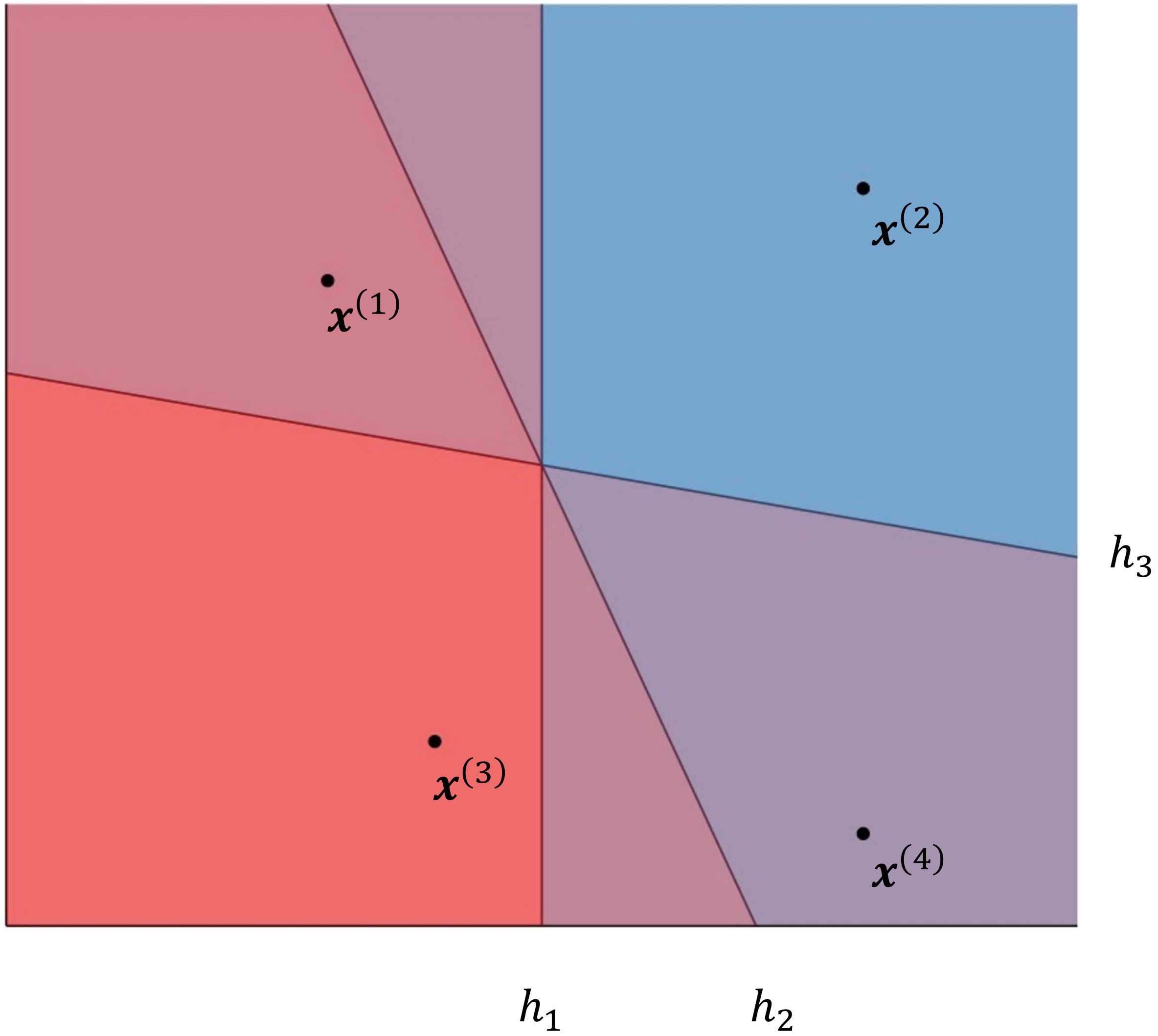
with probability at least $1 - \delta$.

Labellings

- Given some finite set of data points $S = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}\}$ and some hypothesis $h \in \mathcal{H}$, applying h to each point in S results in a **labelling**
  - $[h(\boldsymbol{x}^{(1)}), \dots, h(\boldsymbol{x}^{(M)})]$ is a vector of M +1's and -1's (recall: our discussion of PAC learning assumes binary classification)
- Given $S = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}\}$, 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 $\mathcal{H}(S) = \{[h(\boldsymbol{x}^{(1)}), \dots, h(\boldsymbol{x}^{(M)})] \mid h \in \mathcal{H}\}$

Example: Labellings

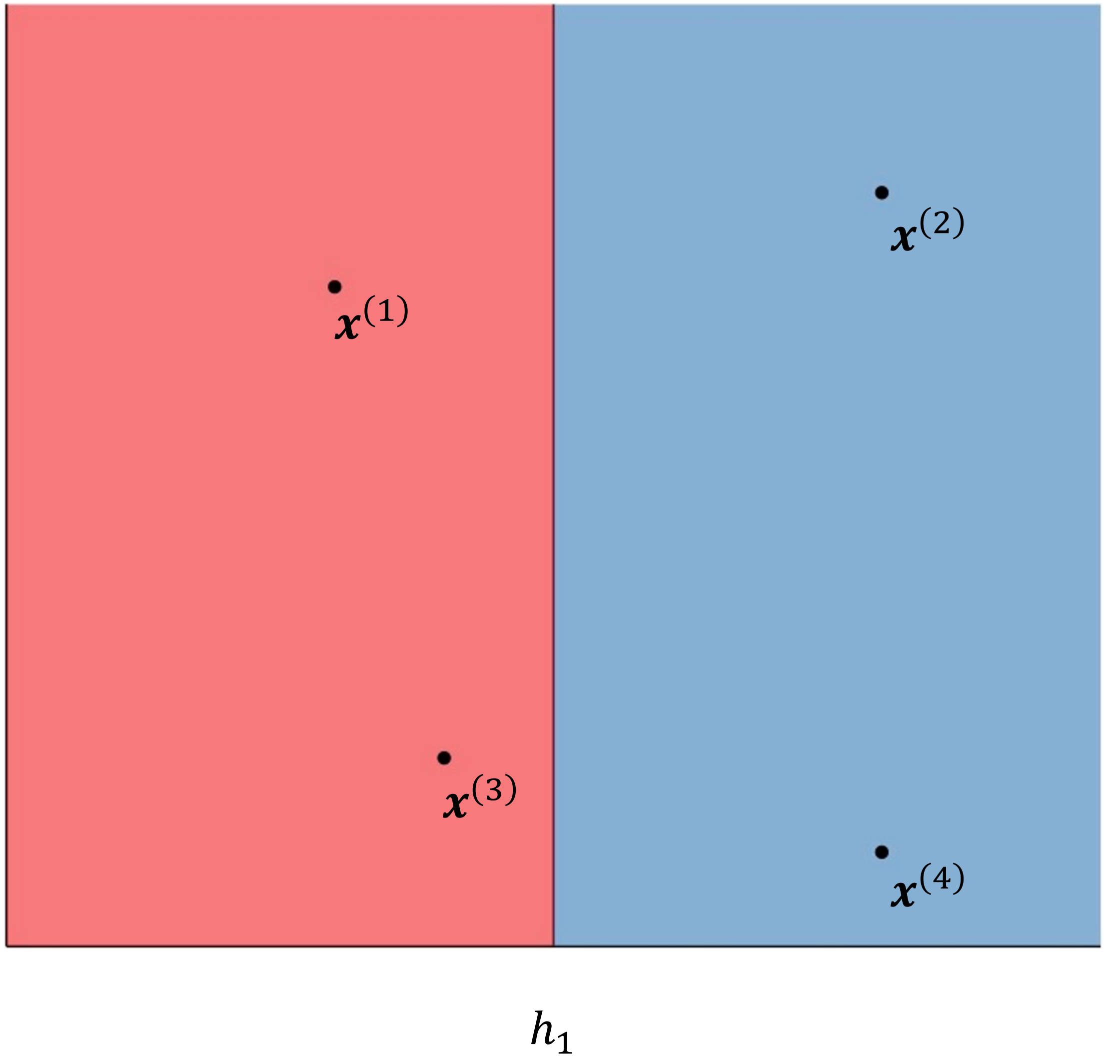
$\mathcal{H} = \{h_1, h_2, h_3\}$



Example: Labellings

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

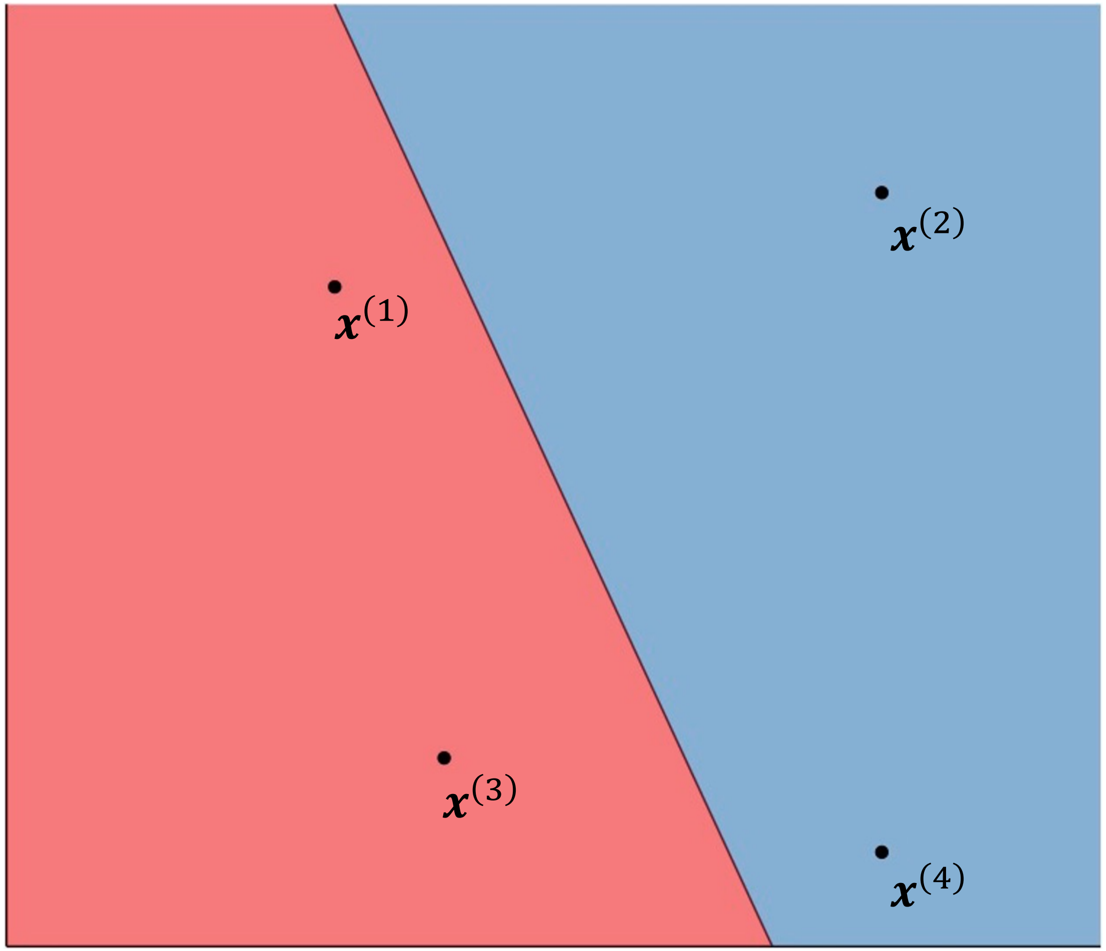
$[h_1(\boldsymbol{x}^{(1)}), h_1(\boldsymbol{x}^{(2)}), h_1(\boldsymbol{x}^{(3)}), h_1(\boldsymbol{x}^{(4)})]$
$= (-1, +1, -1, +1)$



Example: Labellings

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

$[h_1(\boldsymbol{x}^{(1)}), h_1(\boldsymbol{x}^{(2)}), h_1(\boldsymbol{x}^{(3)}), h_1(\boldsymbol{x}^{(4)})]$
$= (-1, +1, -1, +1)$

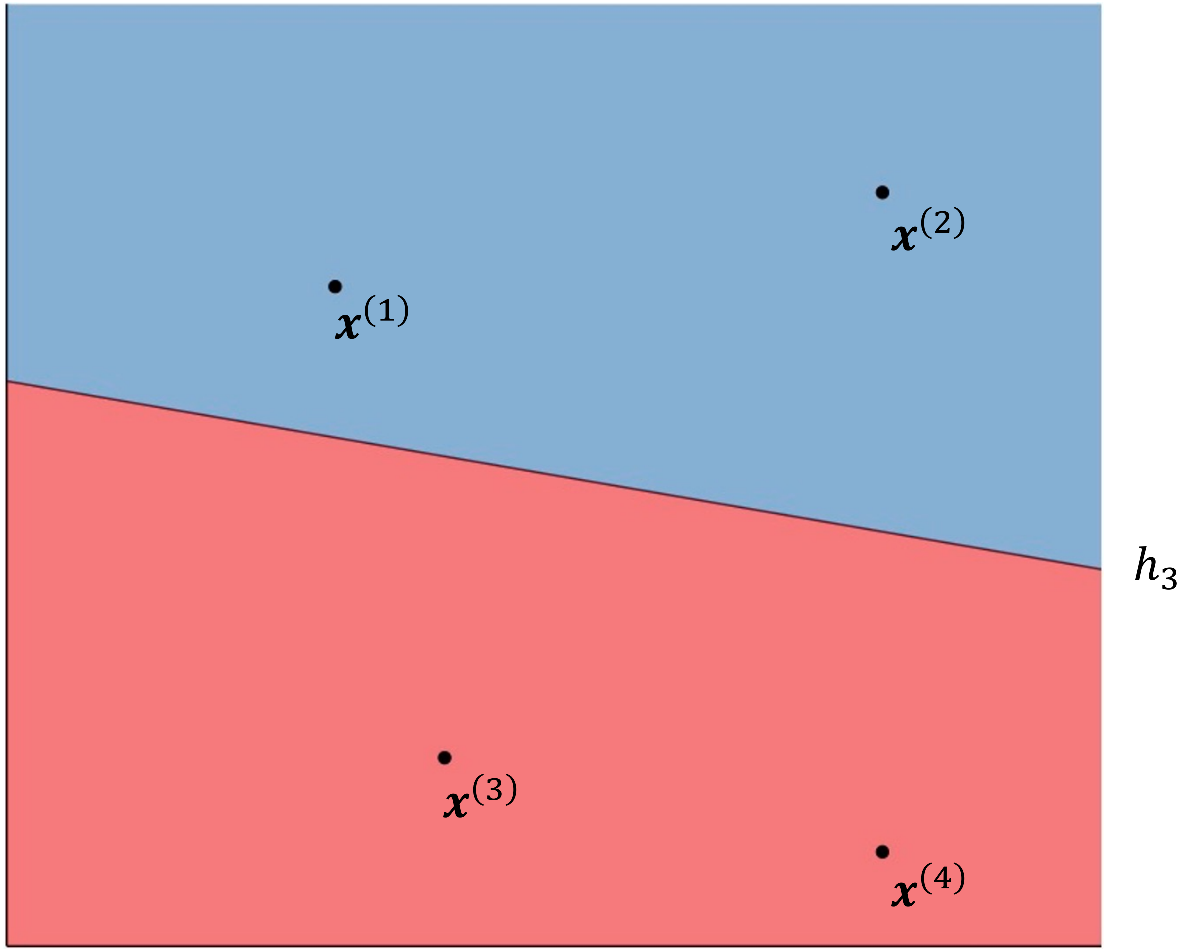


h_2

Example: Labellings

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

$$\left[h_1\left(\boldsymbol{x}^{(1)}\right), h_1\left(\boldsymbol{x}^{(2)}\right), h_1\left(\boldsymbol{x}^{(3)}\right), h_1\left(\boldsymbol{x}^{(4)}\right)\right]$$
$$= (+1, +1, -1, -1)$$



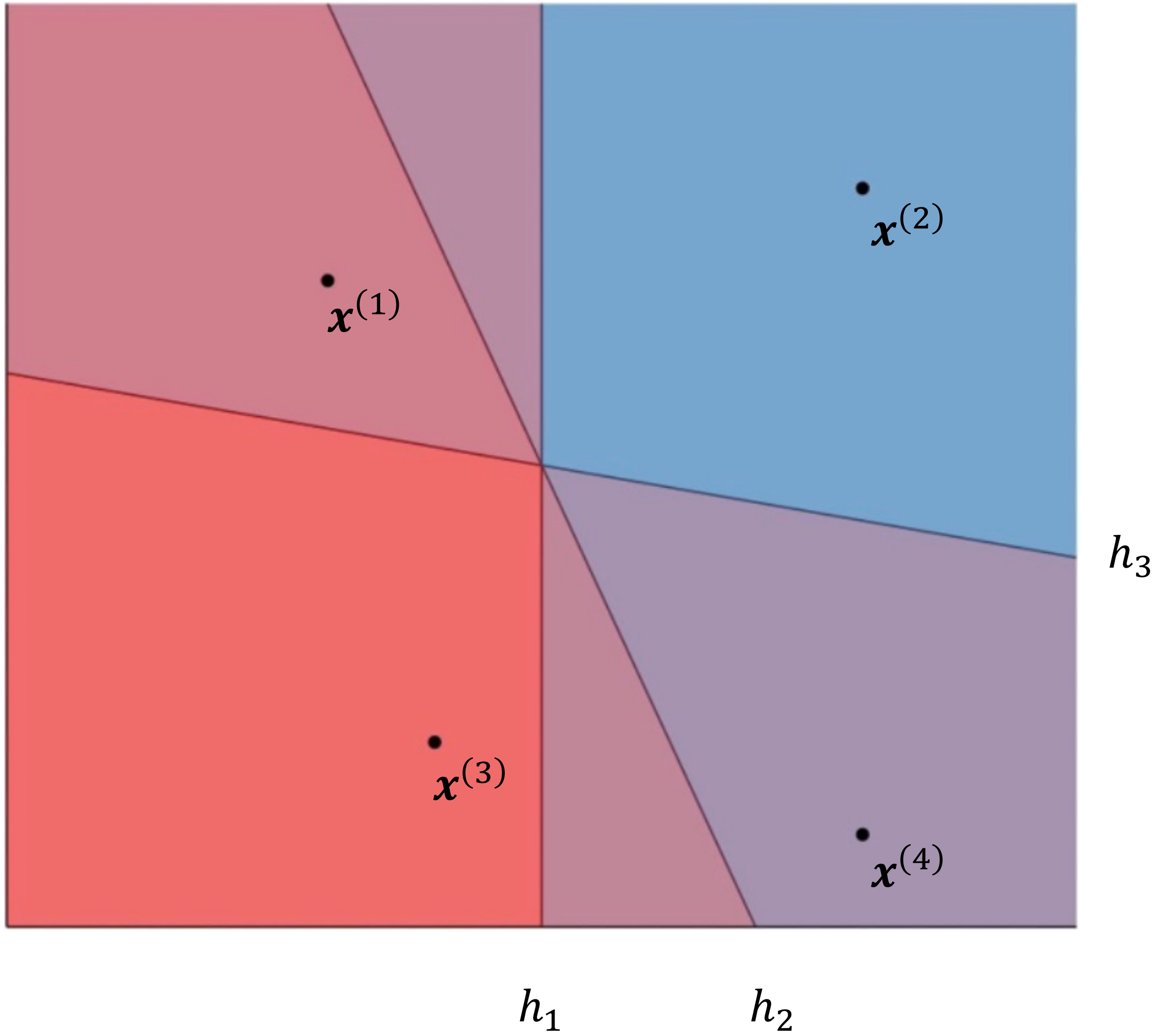
Example: Labellings

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

$\mathcal{H}(S)$

$= \{[+1, +1, -1, -1], [-1, +1, -1, +1]\}$

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



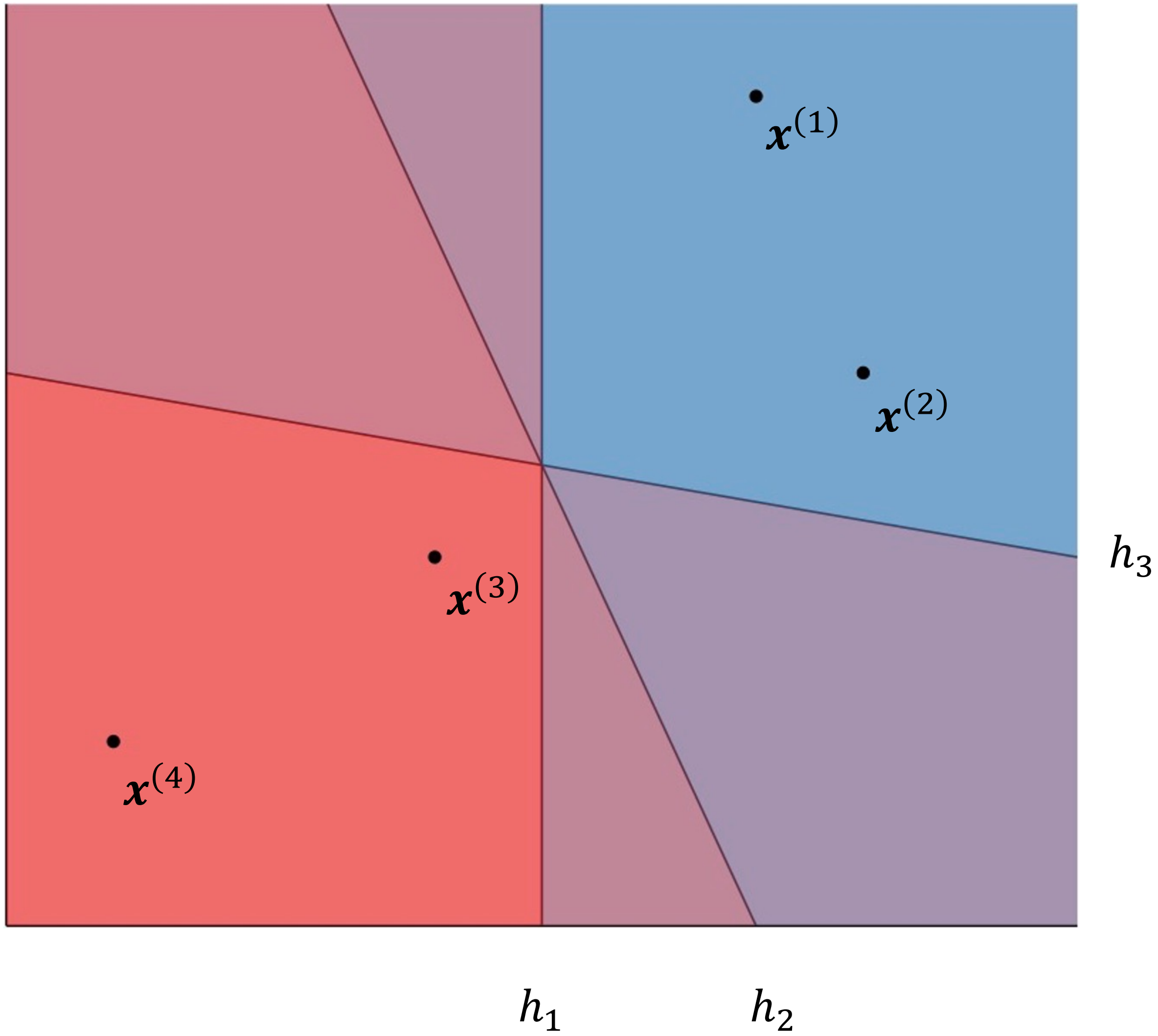
Example: Labellings

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

$\mathcal{H}(S)$

$= \{[+1, +1, -1, -1]\}$

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

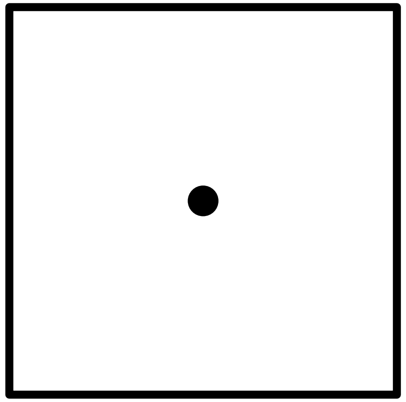


VC-Dimension

- $\mathcal{H}(S)$ 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

VC-Dimension: Example

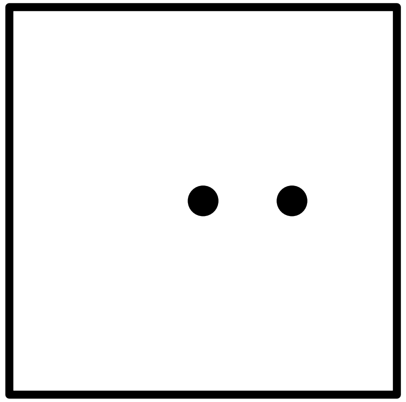
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?



S

VC-Dimension: Example

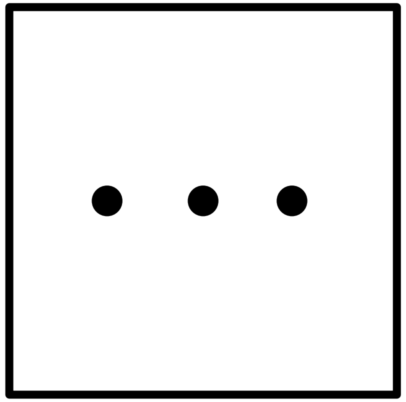
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?



S

VC-Dimension: Example

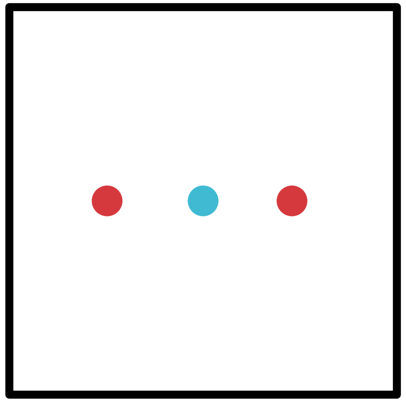
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



S

VC-Dimension: Example

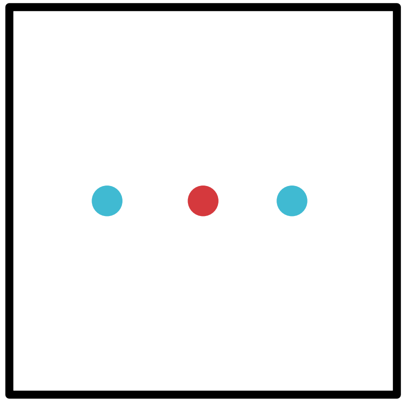
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



S

VC-Dimension: Example

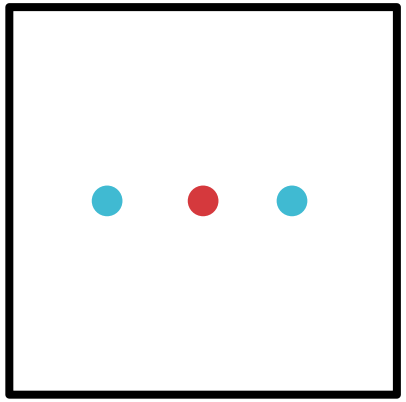
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



S

VC-Dimension: Example

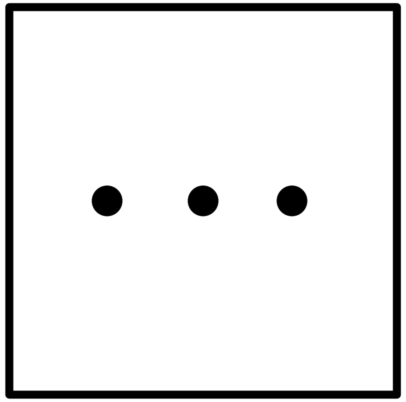
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter ***some*** set of 3 points?



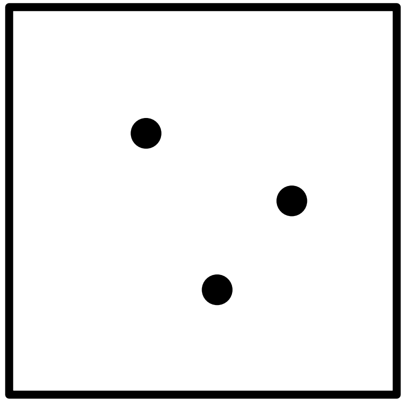
S

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?



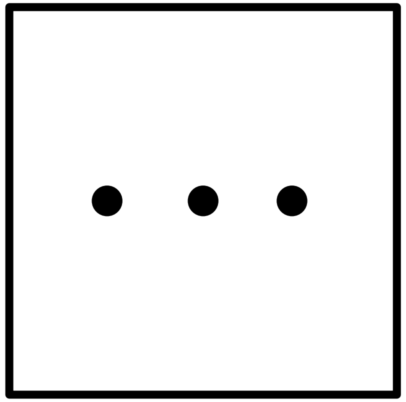
S_1



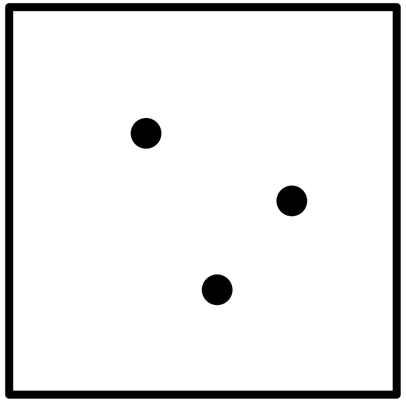
S_2

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?



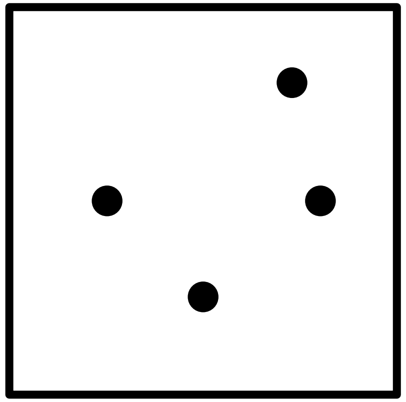
$|\mathcal{H}(S_1)| = 6$



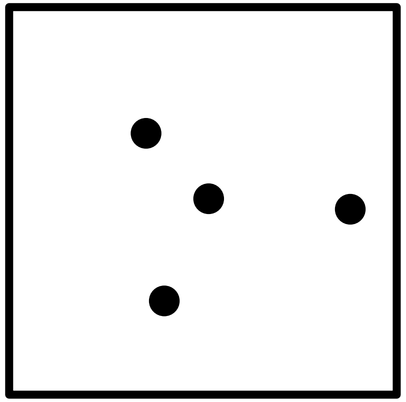
$|\mathcal{H}(S_2)| = 8$

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?



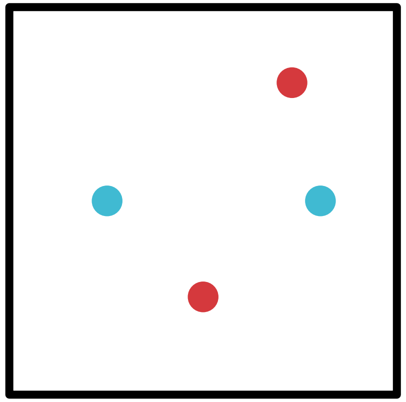
S_1
All points on the convex hull



S_2
At least one point inside the convex hull

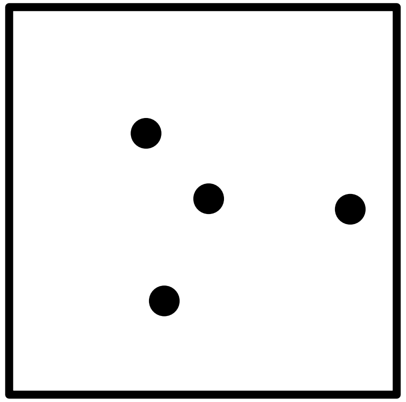
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ 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

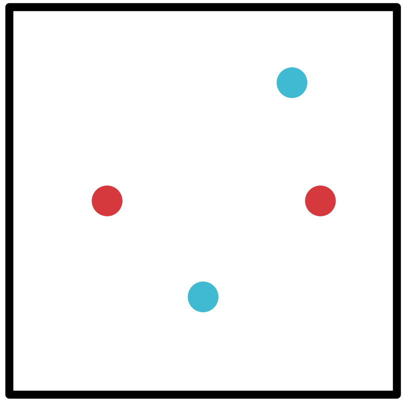


S_2

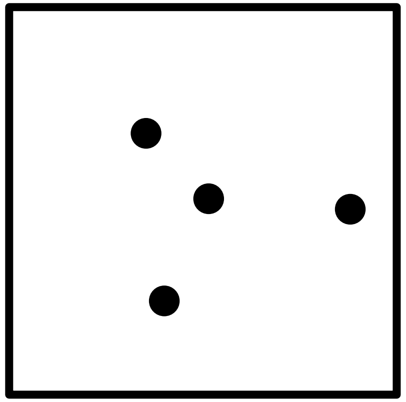
At least one point inside the convex hull

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?



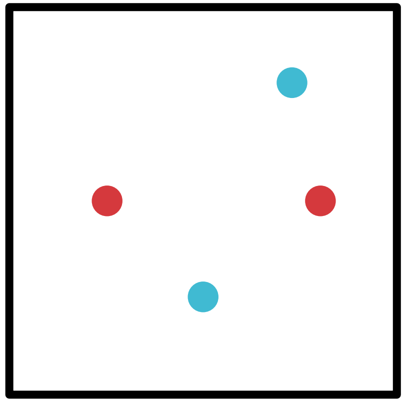
S_1
All points on the convex hull



S_2
At least one point inside the convex hull

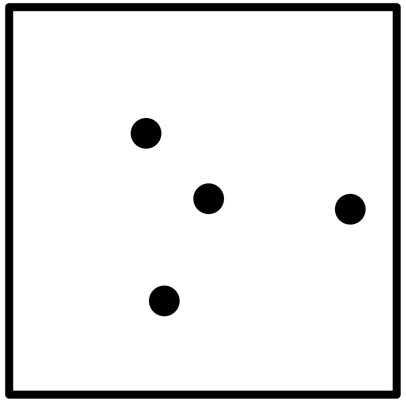
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ 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

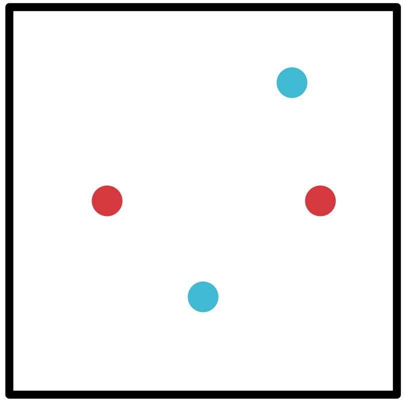


S_2

At least one point inside the convex hull

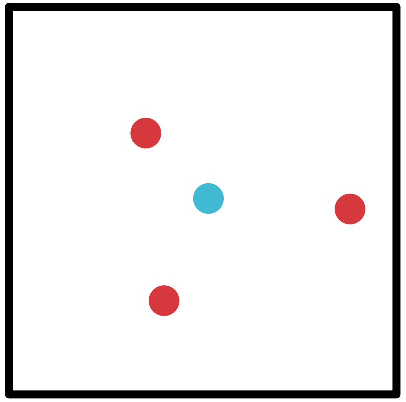
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ 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

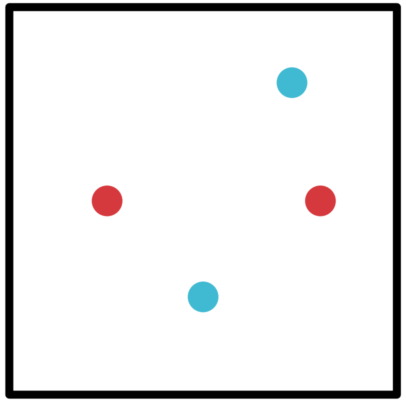


S_2

At least one point inside the convex hull

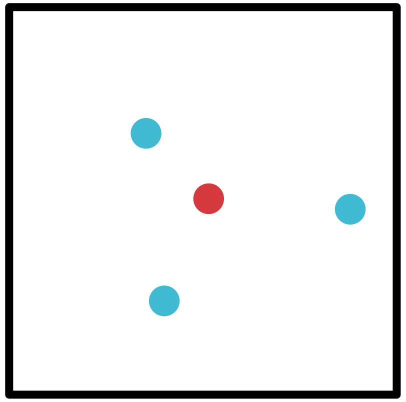
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ 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

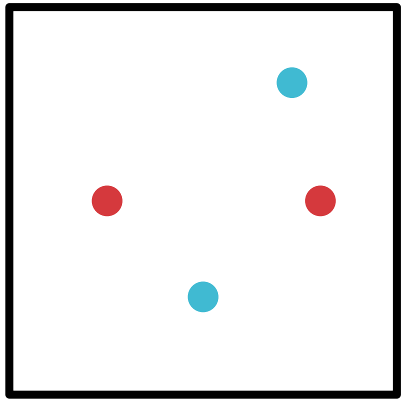


S_2

At least one point inside the convex hull

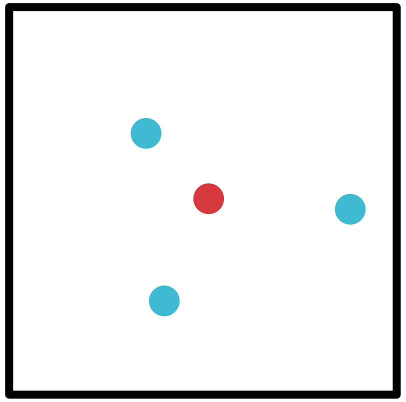
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ 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?



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

All points on the convex hull

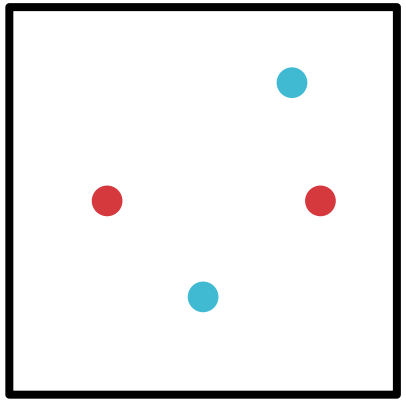


$|\mathcal{H}(S_2)| = 14$

At least one point inside the convex hull

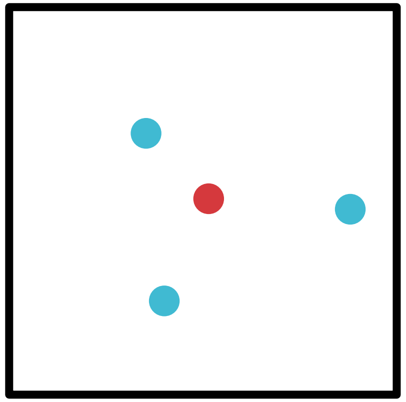
VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^2$ 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



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

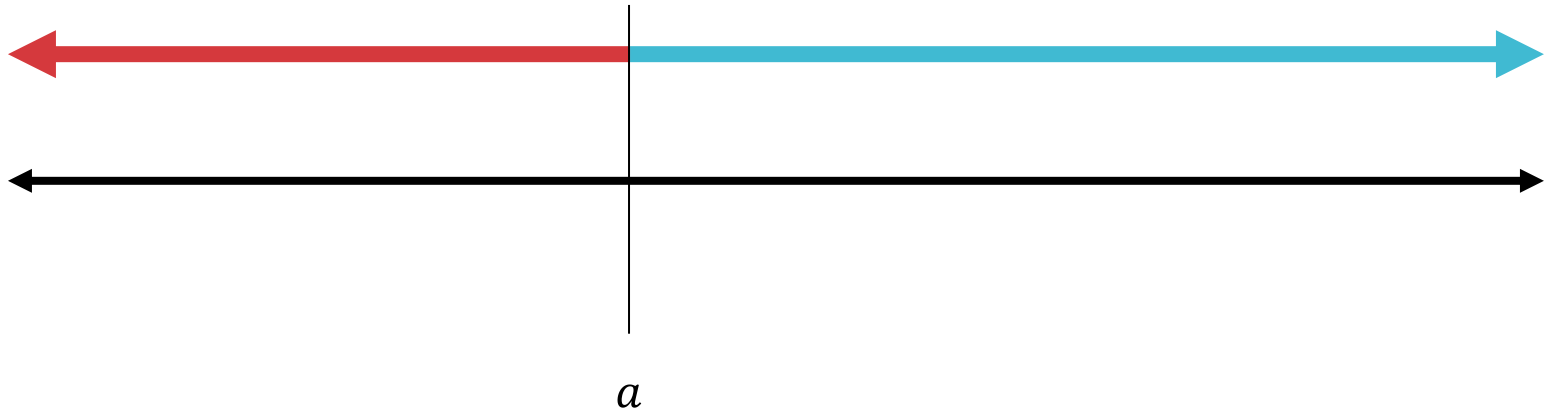
At least one point inside the convex hull

VC-Dimension: Example

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

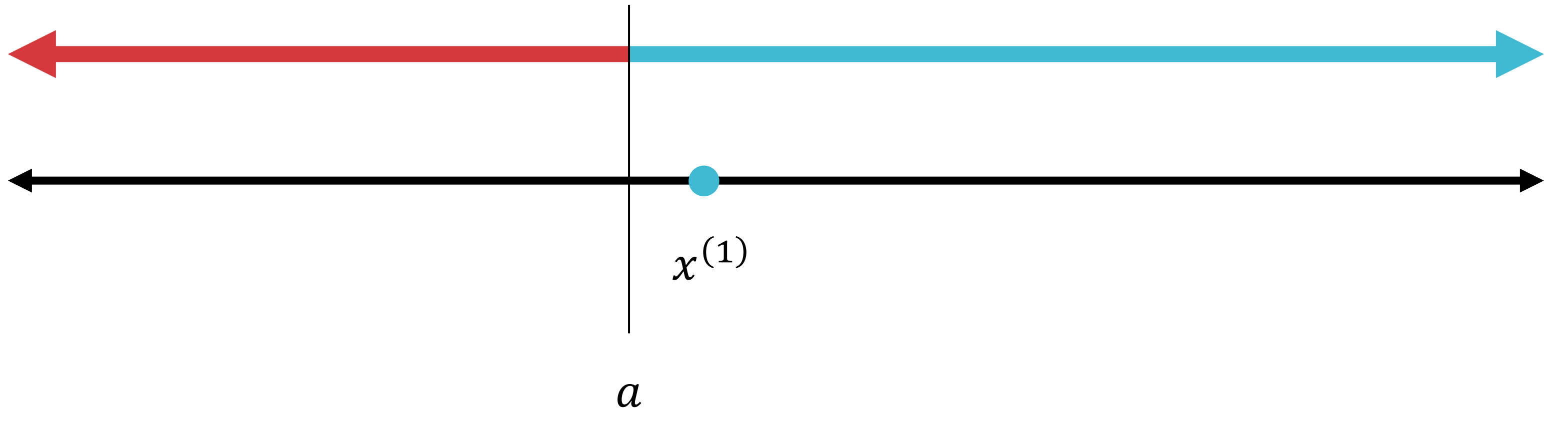
VC-Dimension: Example

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



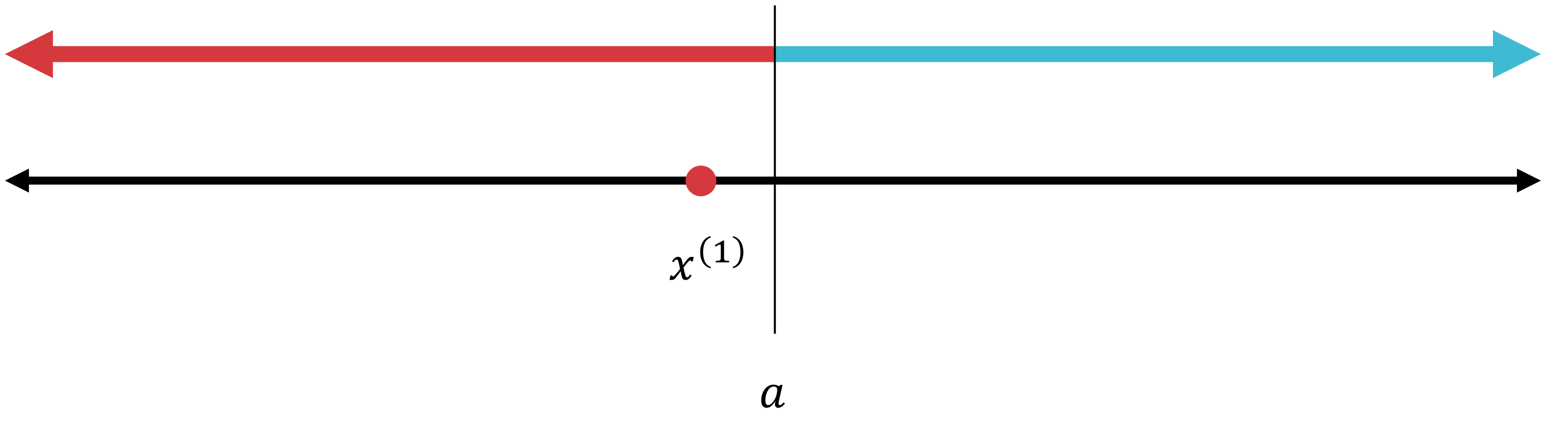
VC-Dimension: Example

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



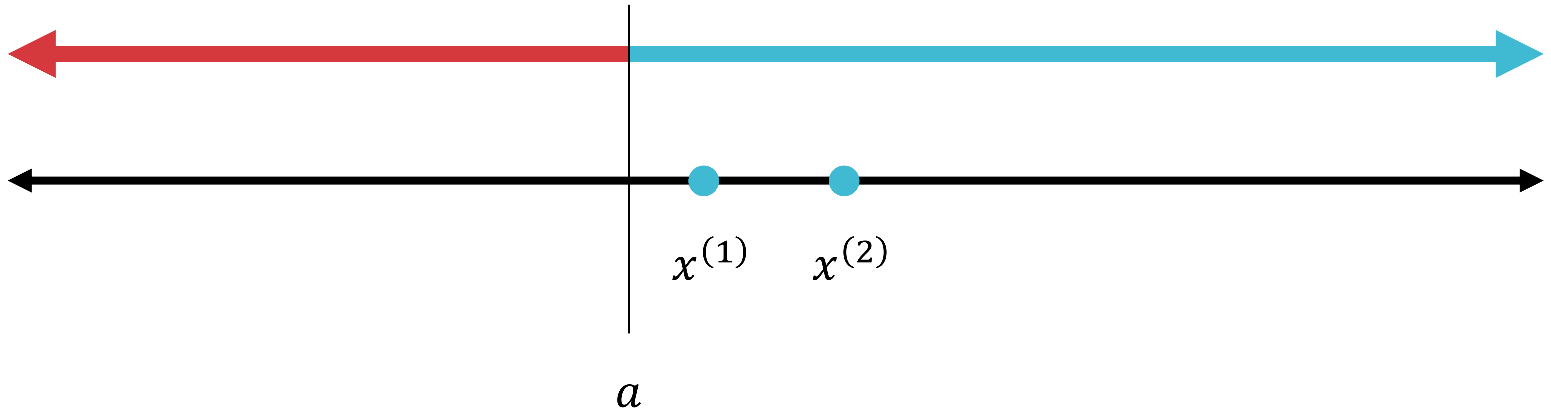
VC-Dimension: Example

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



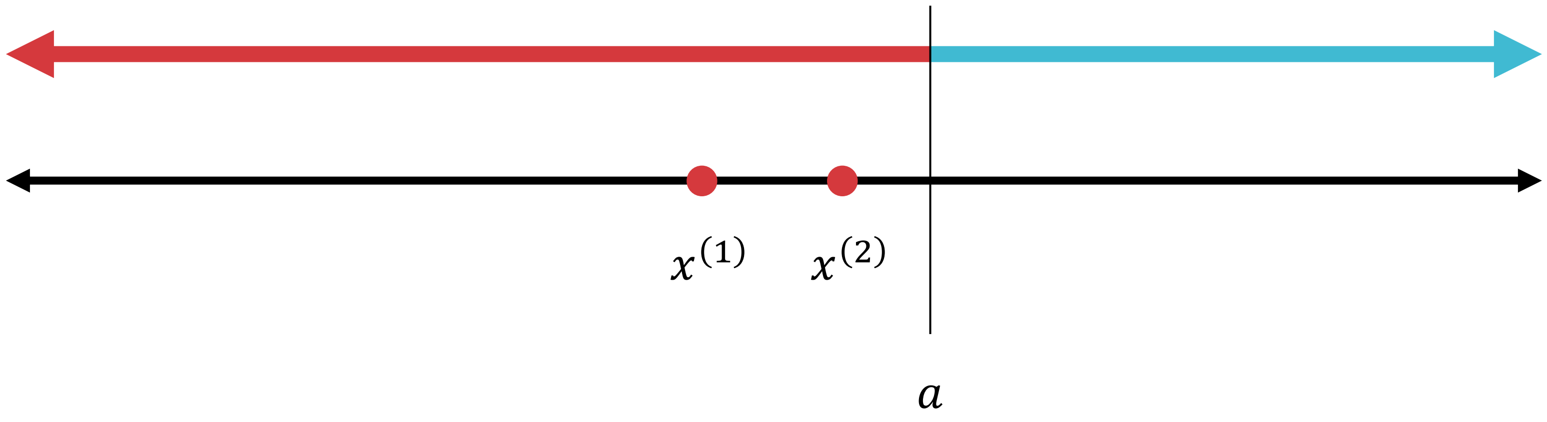
VC-Dimension: Example

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



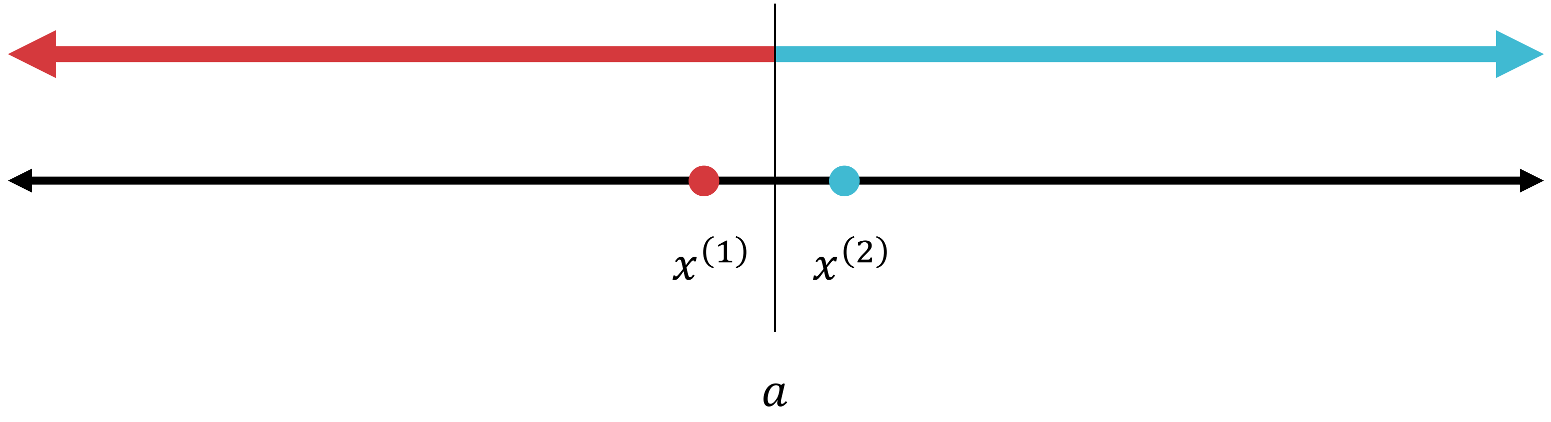
VC-Dimension: Example

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



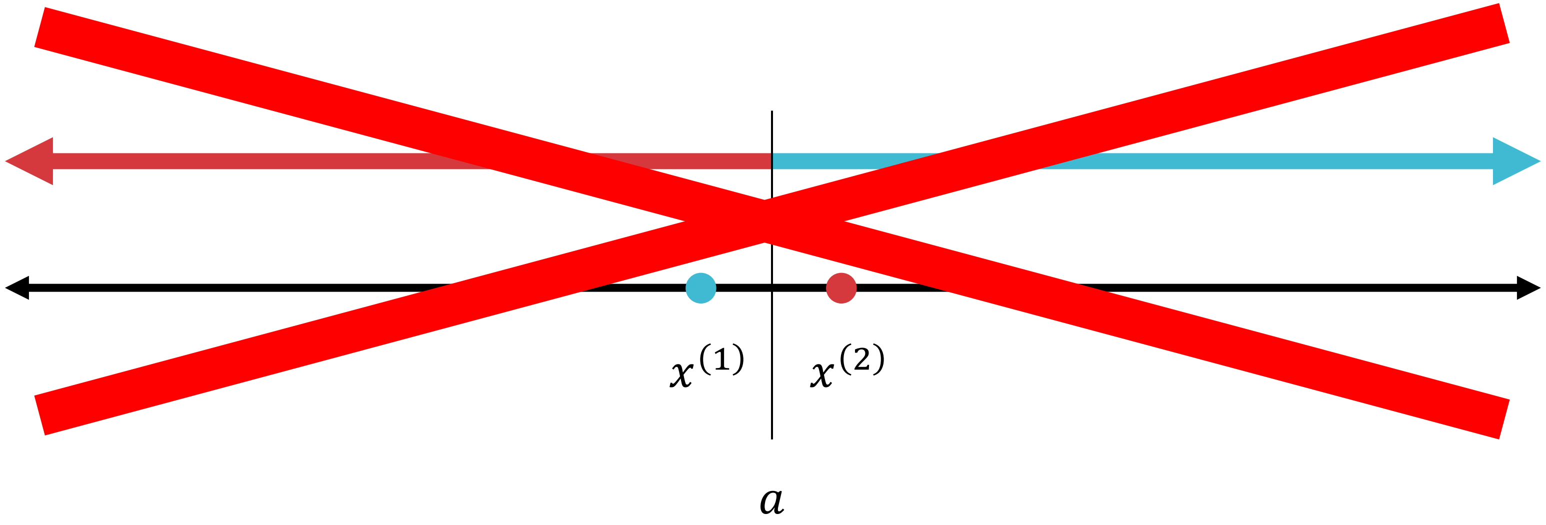
VC-Dimension: Example

- $x \in \mathbb{R}$ 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

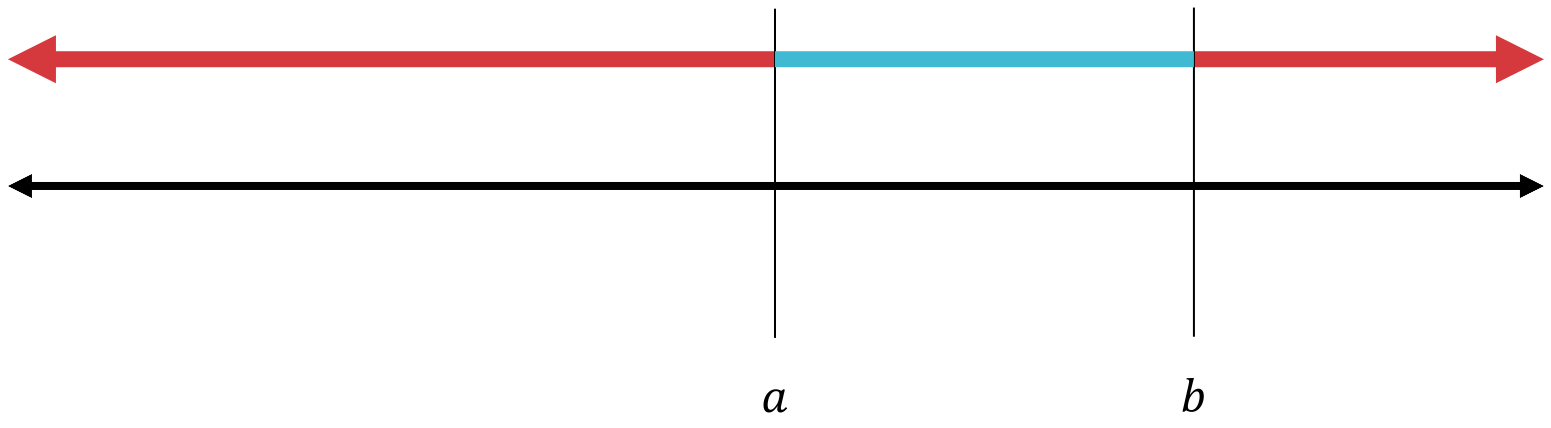
- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

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



Theorem 3: Vapnik-Chervonenkis (VC)-Bound

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

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

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

Statistical Learning Theory Corollary 3

- Infinite, realizable case: for any hypothesis set $\mathcal{H}$ such that $c^* \in \mathcal{H}$ and arbitrary distribution p^*, given a training dataset S where $|S| = M$, 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 arbitrary 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 arbitrary distribution p^*, given a training dataset S where $|S| = M$, 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?

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

How well does h approximate c^*?

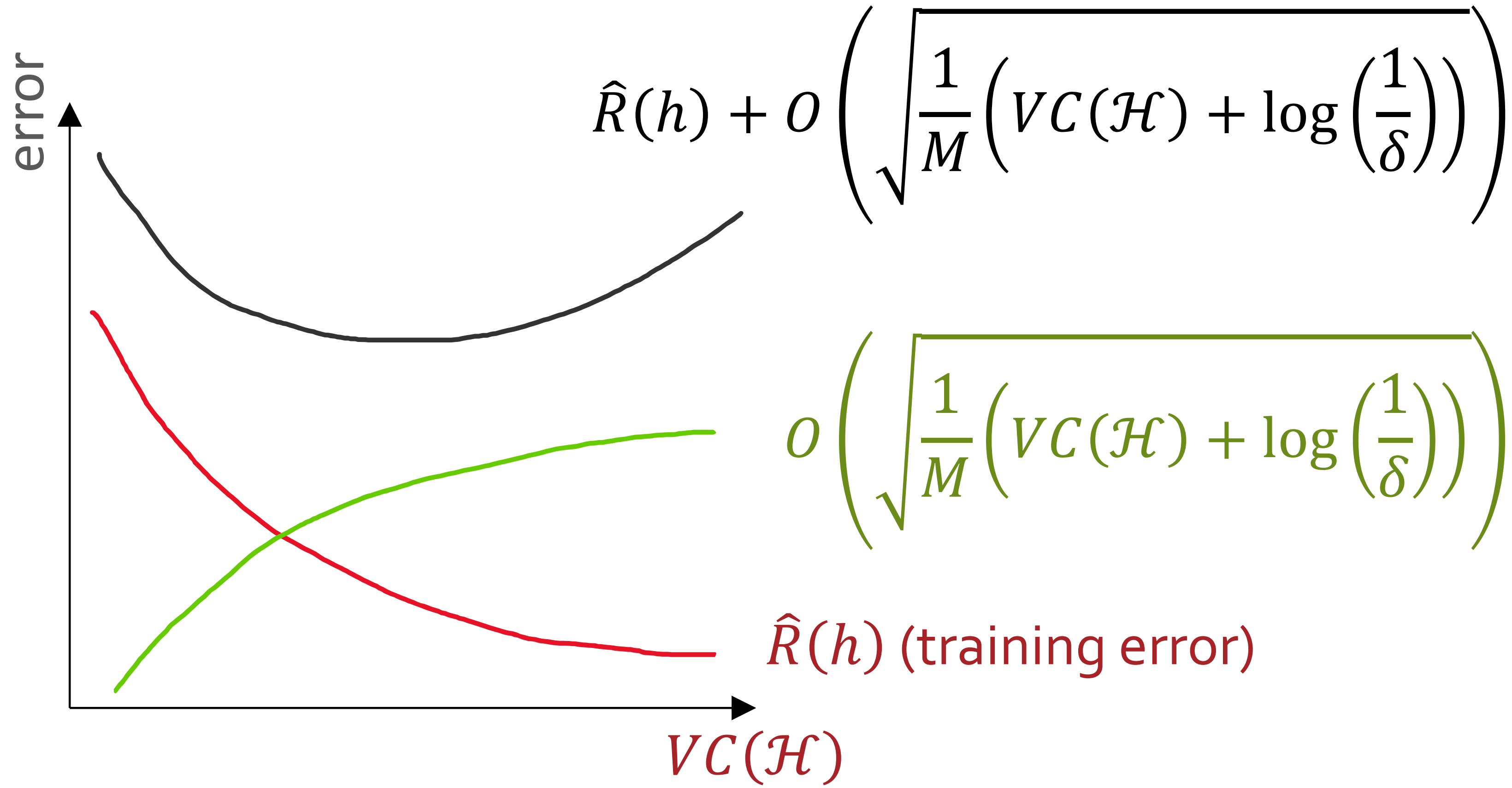
Approximation Generalization Tradeoff

Increases as
$VC(\mathcal{H})$ increases

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

Decreases as
$VC(\mathcal{H})$ increases

Learning Theory and Model Selection



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

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)

- Given independent, identically distributed observations (iid) $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$ from a parametrized probability distribution, MLE sets the parameters by maximizing the likelihood of the data:

$$\theta^{MLE} = \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D} \mid \theta) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p\left(x^{(i)} \mid \theta\right)$$

- Intuition: assign as much of the (finite) probability mass to the observed data *at the expense of unobserved data*

Recall: Maximum Likelihood Estimation (MLE)

- Given independent, identically distributed observations (iid) $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$ from a parametrized probability distribution, MLE sets the parameters by maximizing the *log*-likelihood of the data:

$$\theta^{MLE} = \underset{\theta}{\operatorname{argmax}} \log p(\mathcal{D} \mid \theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log p\left(x^{(i)} \mid \theta\right)$$

- Intuition: assign as much of the (finite) probability mass to the observed data *at the expense of unobserved data*

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

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

Maximum a Posteriori (MAP) Estimation

- Insight: sometimes we have *prior* information we want to incorporate into parameter estimation
- Idea: use Bayes rule to reason about the *posterior* distribution over the parameters

Maximum a Posteriori (MAP) Estimation

1. Specify the *generative story*, i.e., the data generating distribution, including a *prior distribution*
2. Maximize the log-posterior of $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$

$$\ell_{MAP}(\theta) = \log p(\theta) + \sum_{i=1}^{N} \log p\left(x^{(i)} | \theta\right)$$

3. Solve in *closed form*: take partial derivatives, set to 0 and solve