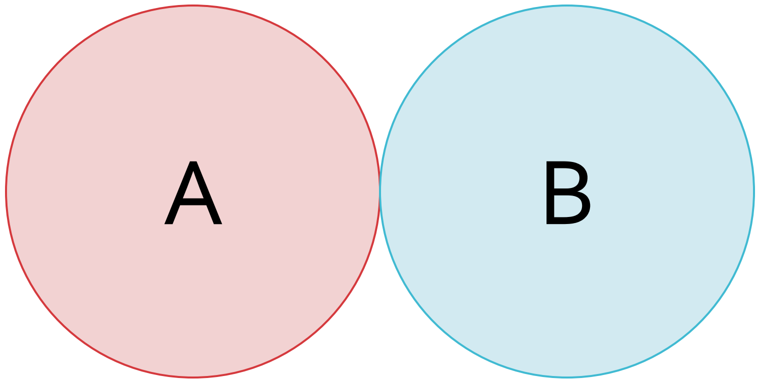
10-301/601: Introduction to Machine Learning Lecture 18 – Learning Theory (Infinite Case)

Henry Chai

5/28/25

The Union Bound…

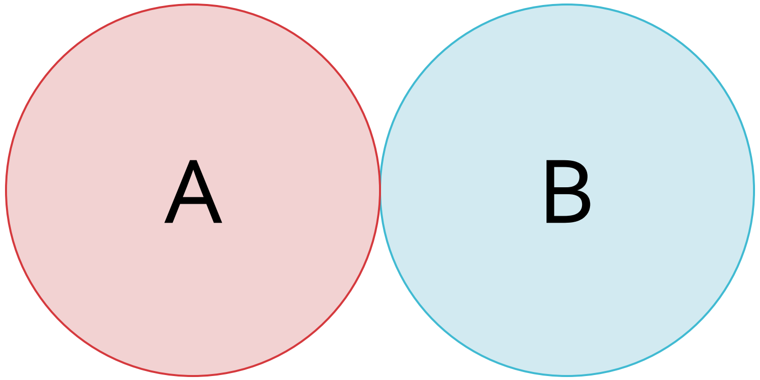
$$P\{A \cup B\} \leq P\{A\} + P\{B\}$$



The Union Bound is Bad!

$$P\{A \cup B\} \leq P\{A\} + P\{B\}$$

$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}$$

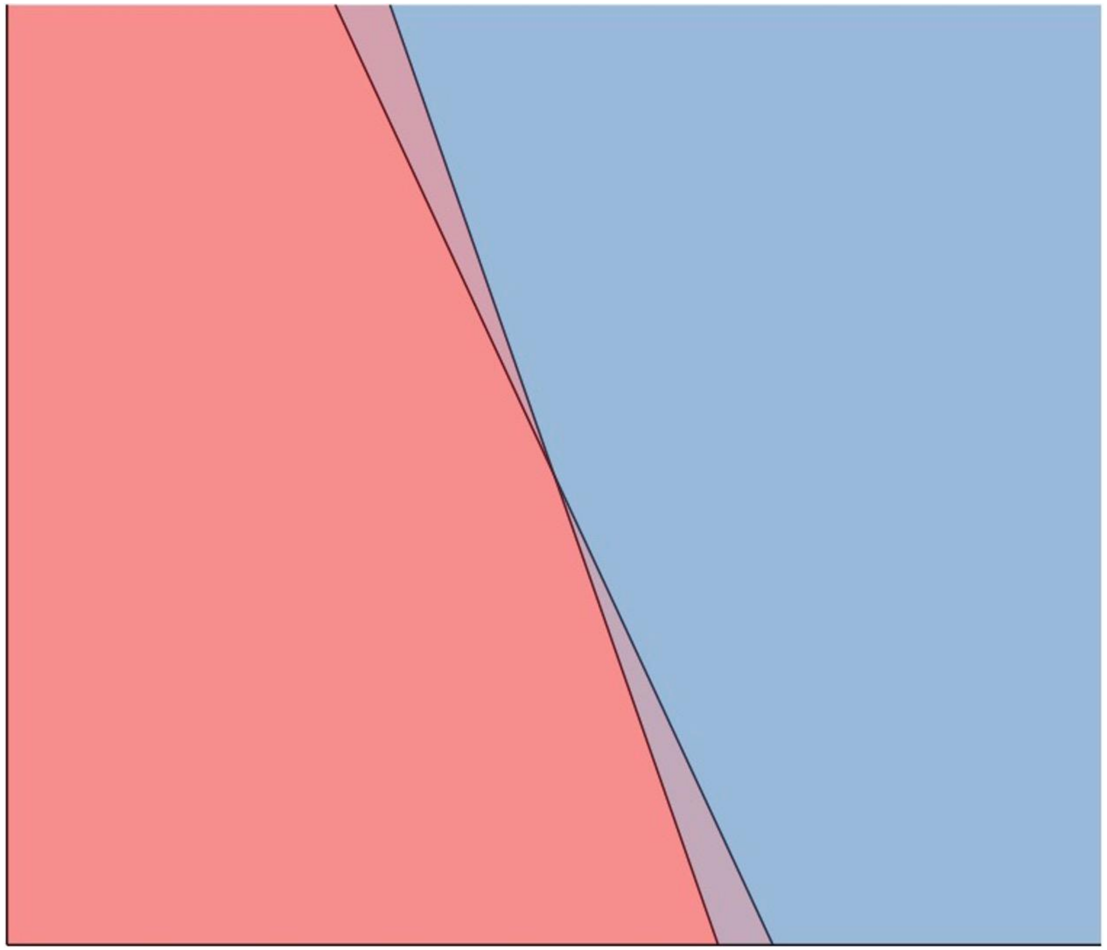


Intuition

If two hypotheses $h_1, h_2 \in \mathcal{H}$ are very similar, then the events

- "h_1 is consistent with the first m training data points"
- "h_2 is consistent with the first m training data points"

will overlap a lot!

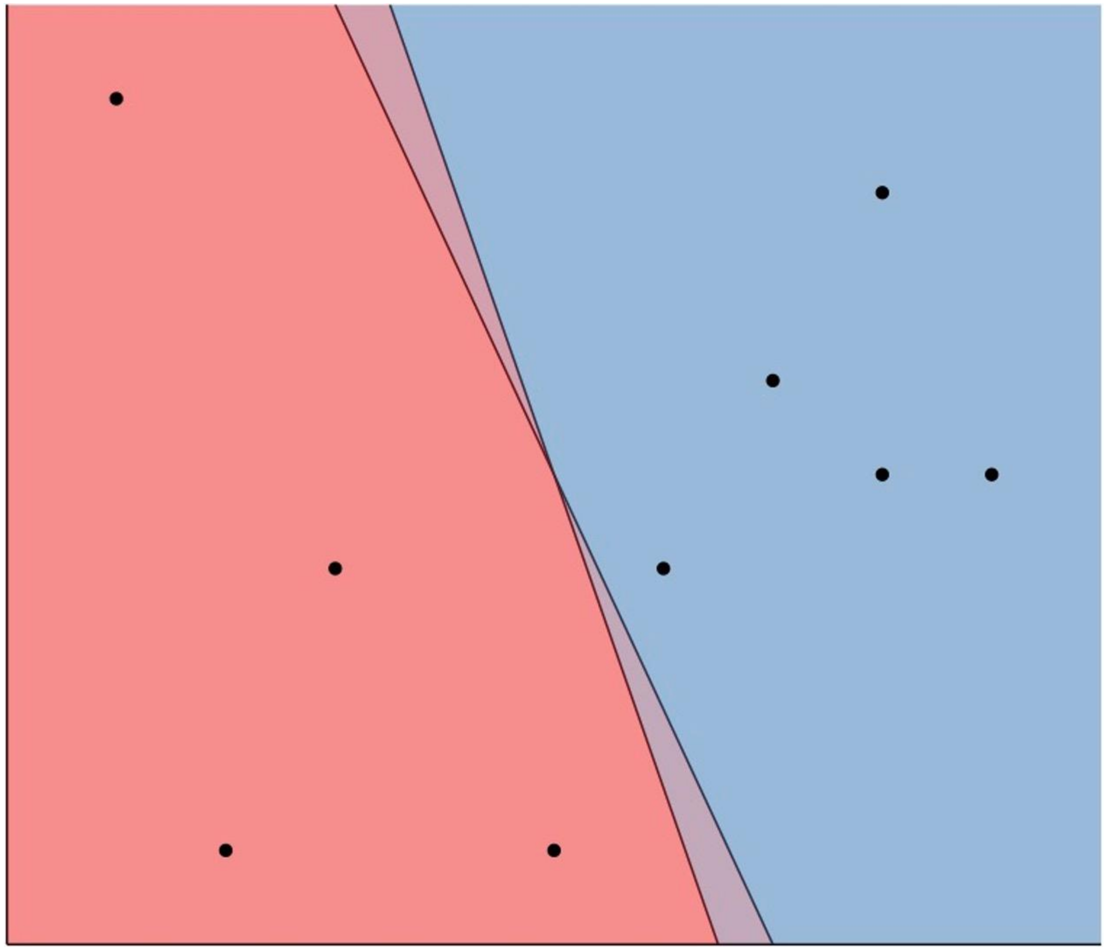


Intuition

If two hypotheses $h_1, h_2 \in \mathcal{H}$ are very similar, then the events

- "h_1 is consistent with the first m training data points"
- "h_2 is consistent with the first m training data points"

will overlap a lot!



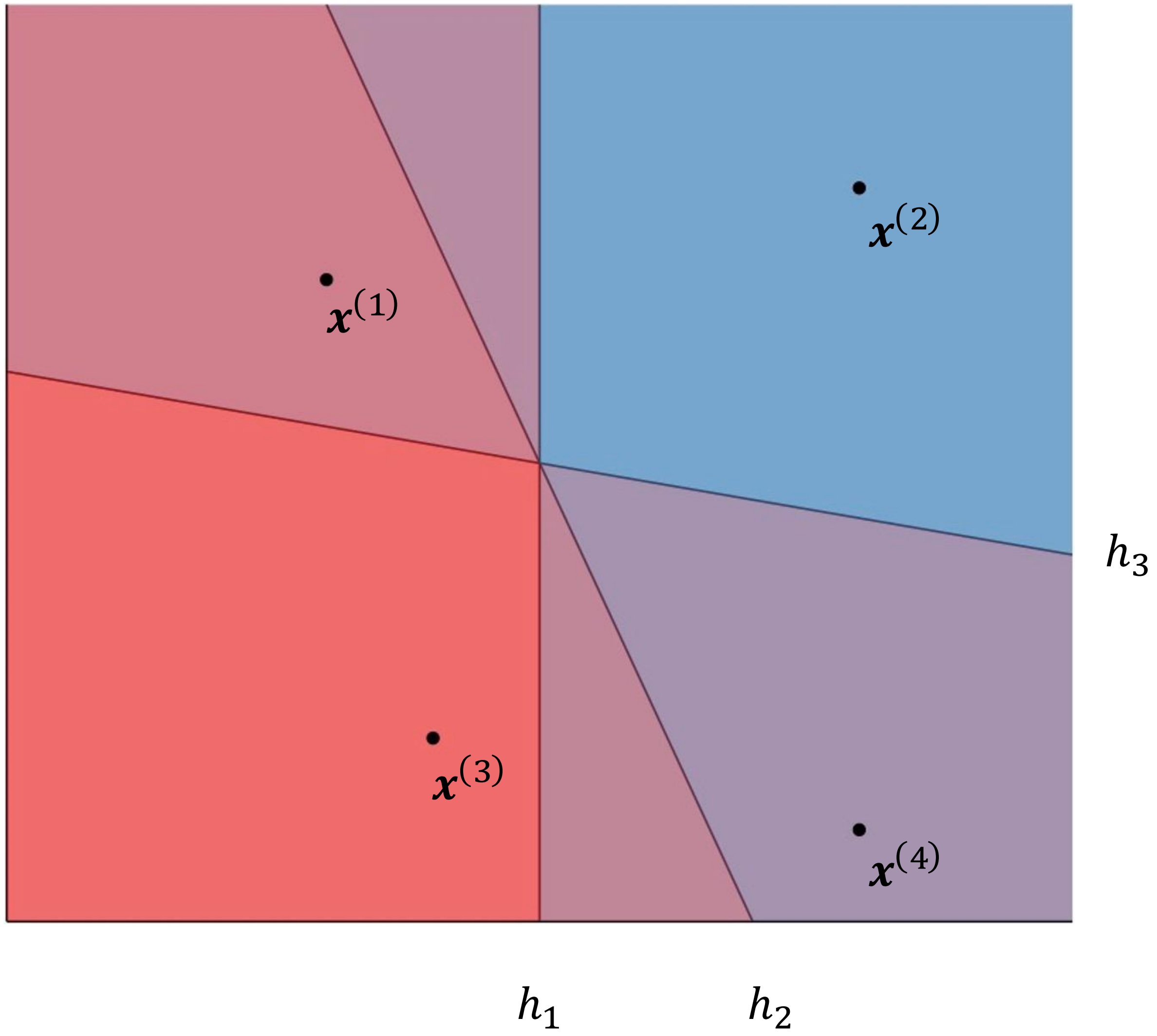
Labellings

- Given some finite set of data points $S = \left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}\right)$ and some hypothesis $h \in \mathcal{H}$, applying h to each point in S results in a **labelling**
  - $\left(h\left(\boldsymbol{x}^{(1)}\right), \dots, h\left(\boldsymbol{x}^{(M)}\right)\right)$ is a vector of M +1's and -1's
- Given $S = \left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}\right)$, 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

$$\mathcal{H}(S) = \left\{\left(h\left(\boldsymbol{x}^{(1)}\right), \dots, h\left(\boldsymbol{x}^{(M)}\right)\right) \middle| h \in \mathcal{H}\right\}$$

Example: Labellings

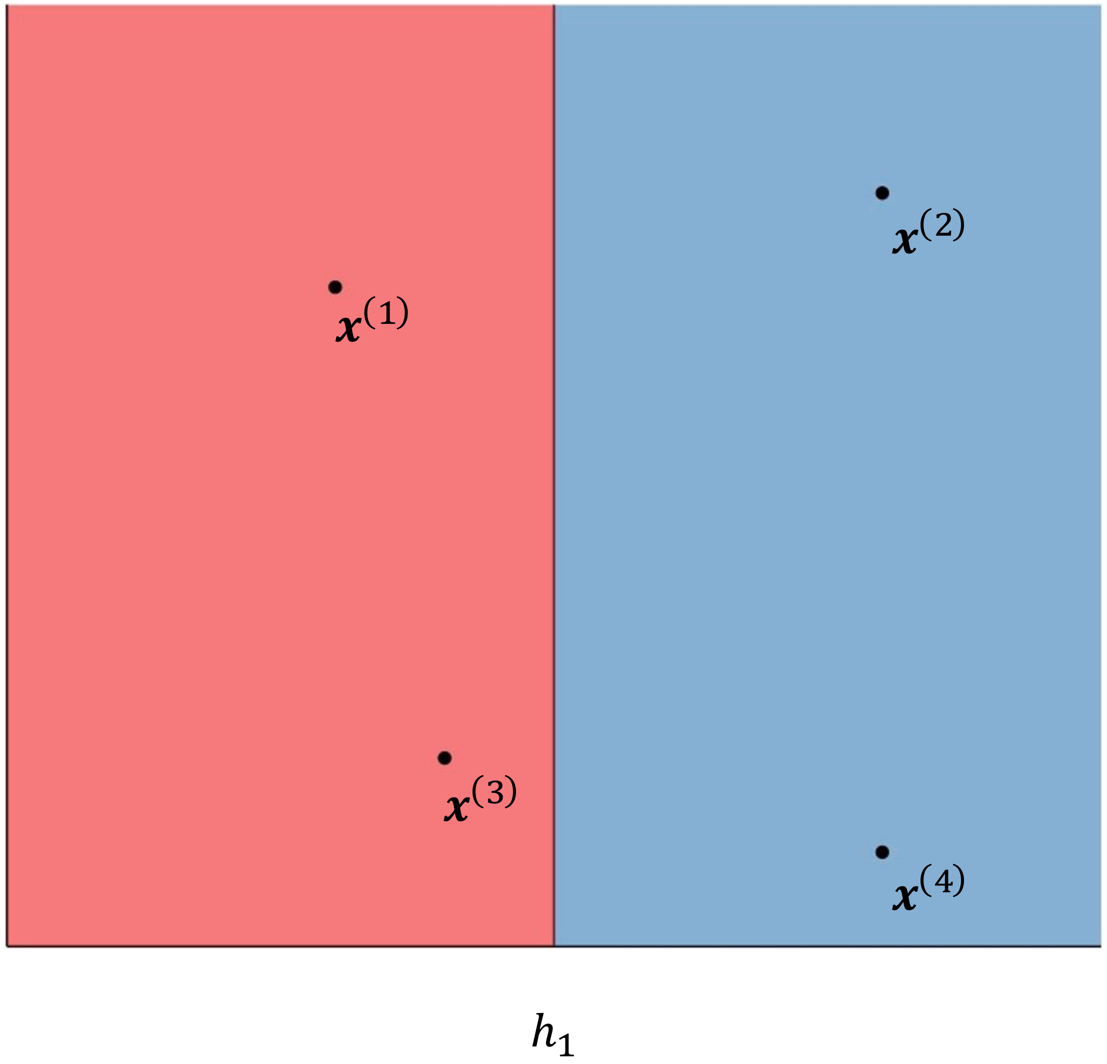
$\mathcal{H} = \{h_1, h_2, h_3\}$



Example: Labellings

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

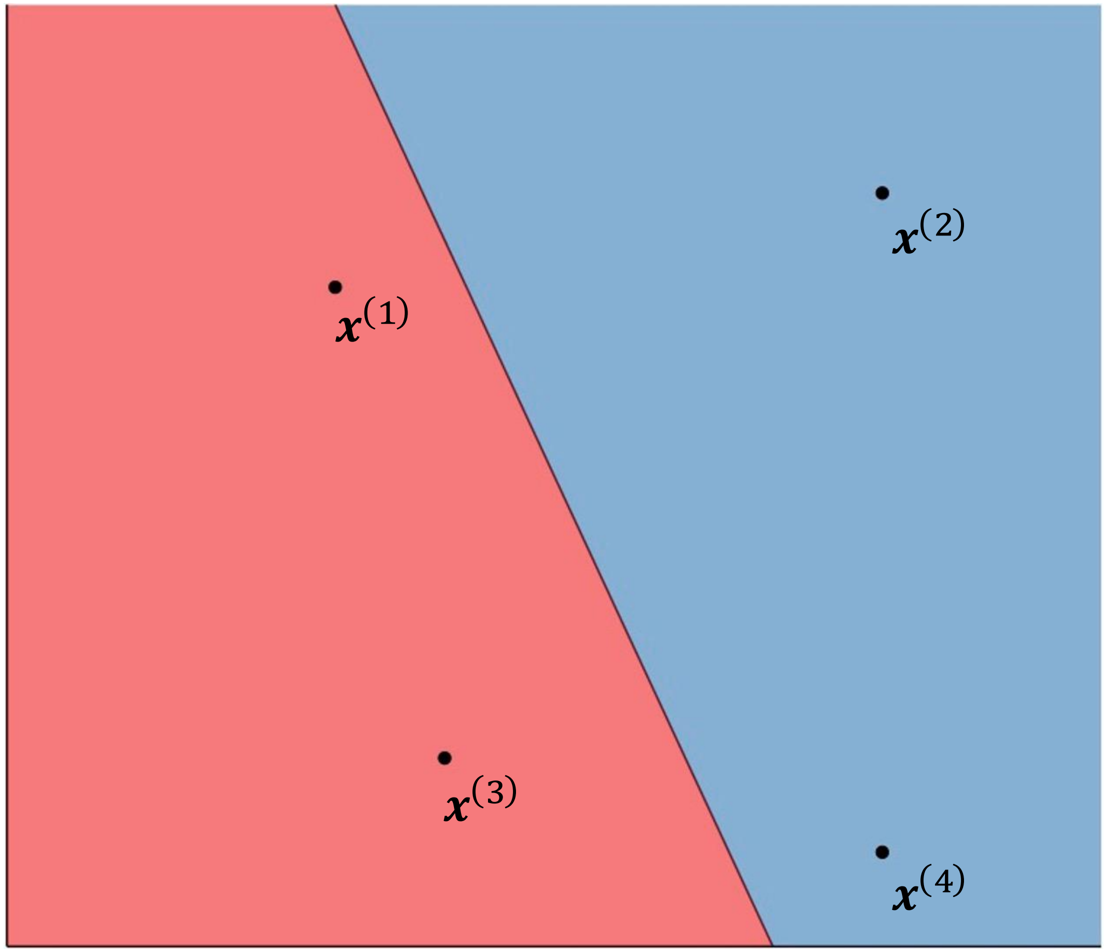
$$\left(h_1\left(\boldsymbol{x}^{(1)}\right), h_1\left(\boldsymbol{x}^{(2)}\right), h_1\left(\boldsymbol{x}^{(3)}\right), h_1\left(\boldsymbol{x}^{(4)}\right)\right)$$
$$= (-1, +1, -1, +1)$$



Example: Labellings

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

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

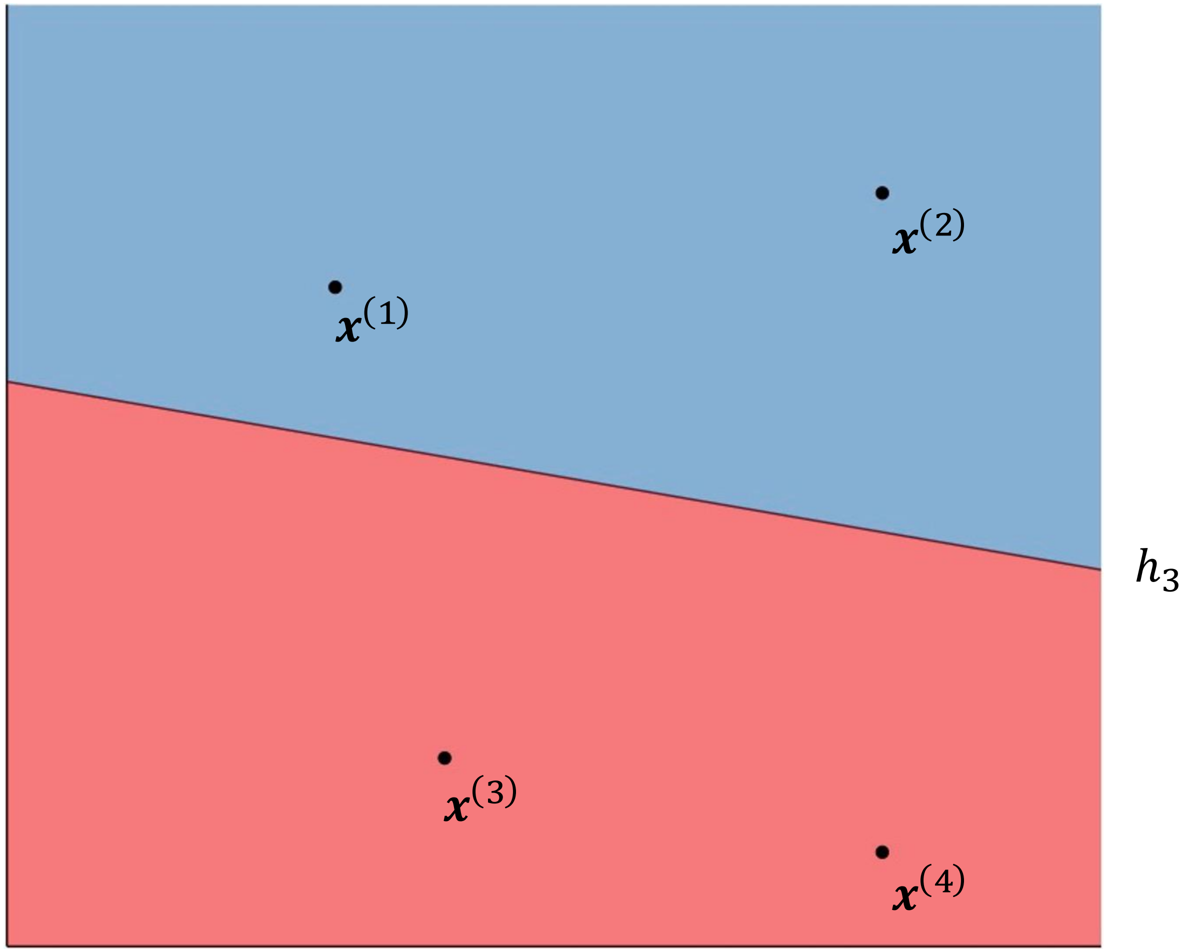


h_2

Example: Labellings

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

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

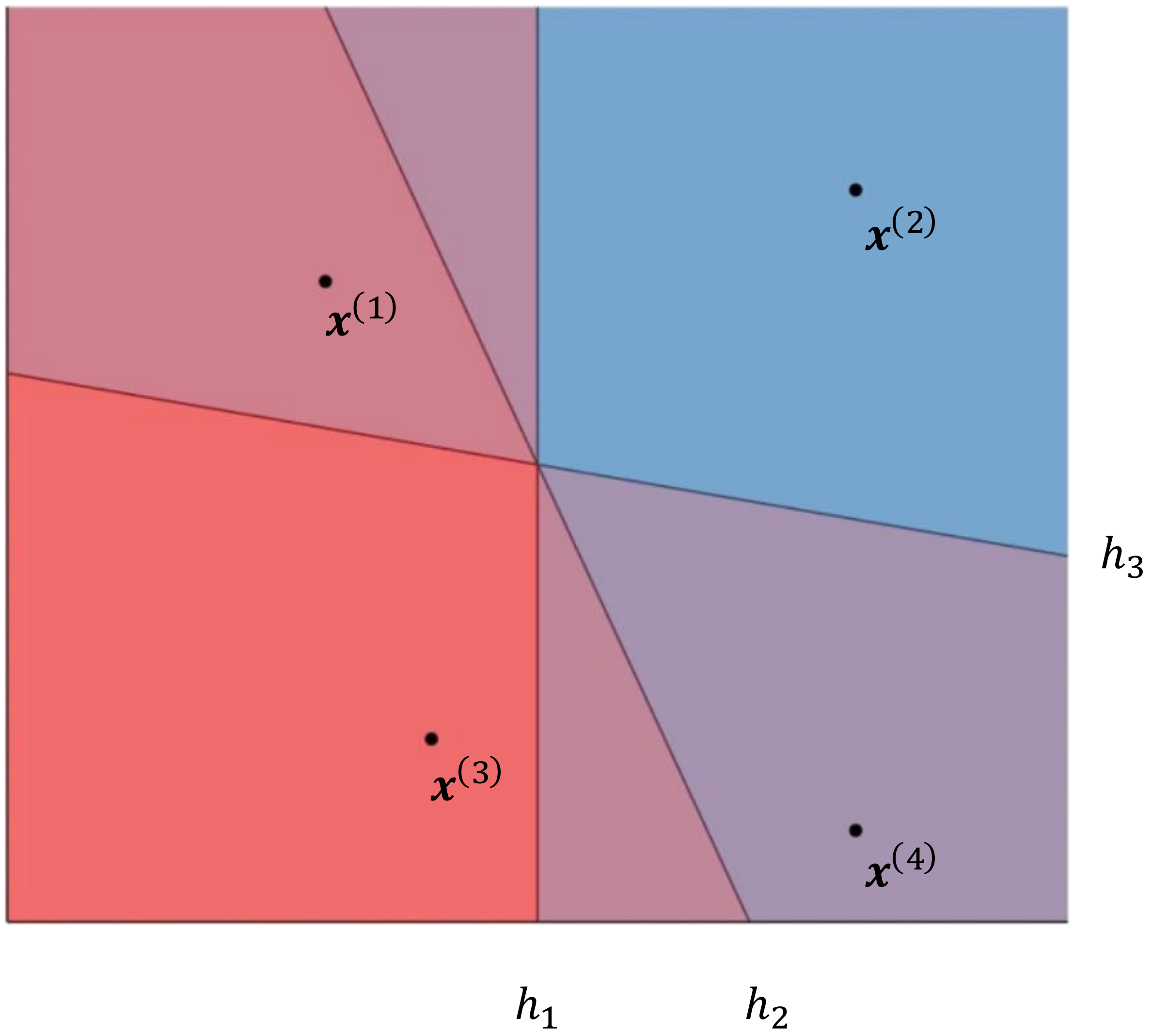


Example: Labellings

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

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

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

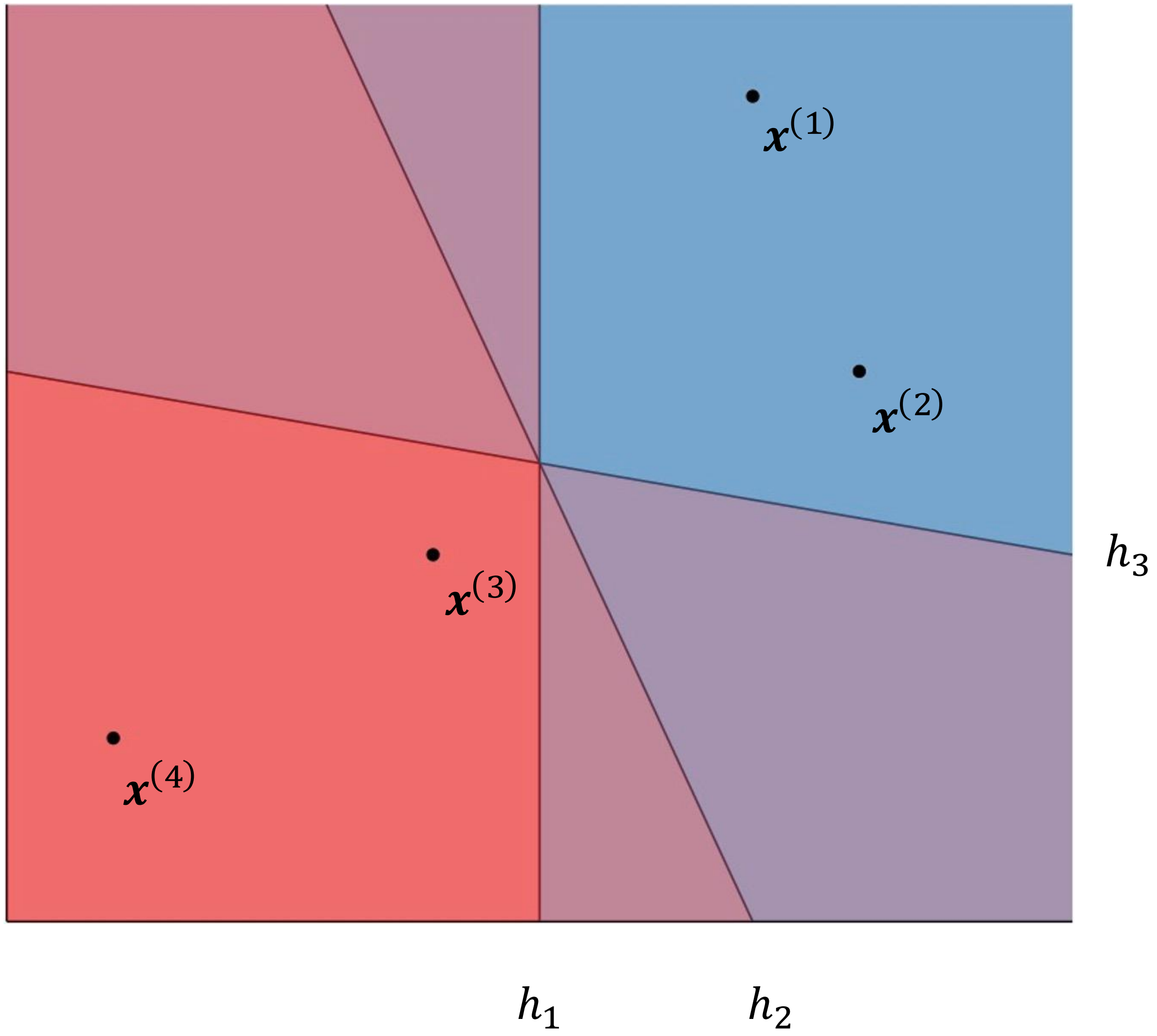


Example: Labellings

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

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

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

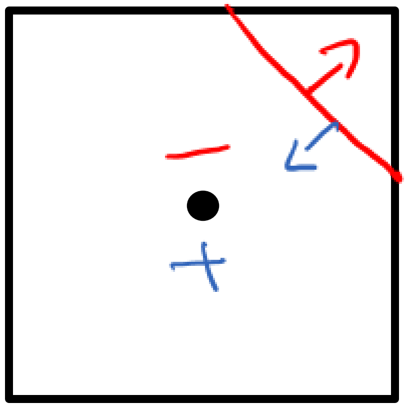


VC-Dimension

- $\mathcal{H}(S)$ 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 $d_{VC}(\mathcal{H}) = \infty$
- 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

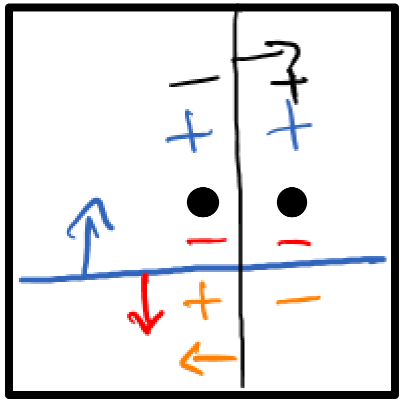
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?



S

VC-Dimension: Example

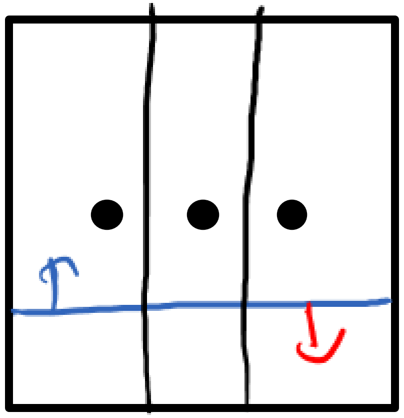
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?



S

VC-Dimension: Example

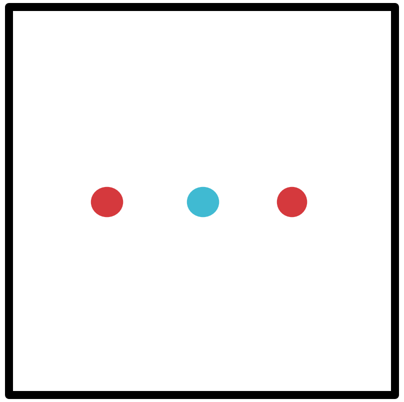
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



S

VC-Dimension: Example

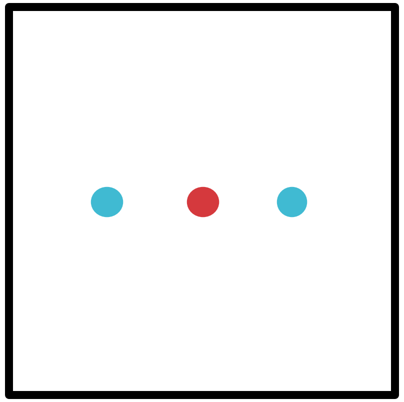
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



S

VC-Dimension: Example

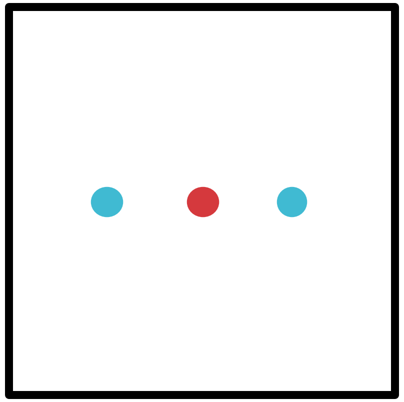
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



S

VC-Dimension: Example

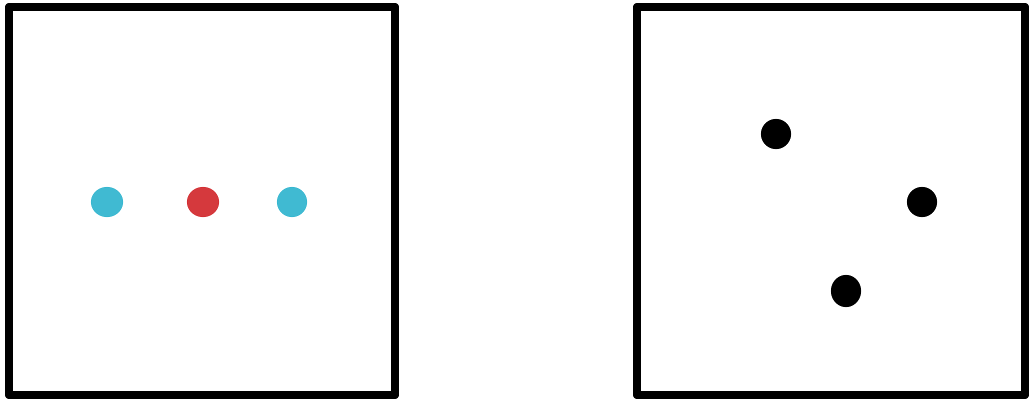
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter ***some*** set of 3 points?



S

VC-Dimension: Example

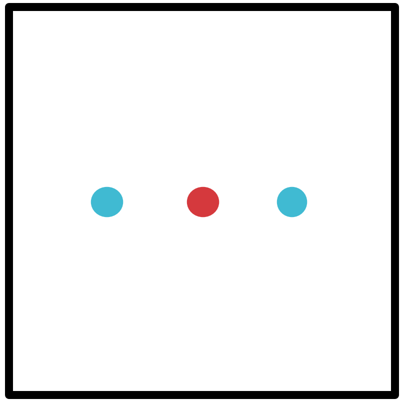
- $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $VC(\mathcal{H})$?
  - Can $\mathcal{H}$ shatter some set of 1 point?
  - Can $\mathcal{H}$ shatter some set of 2 points?
  - Can $\mathcal{H}$ shatter some set of 3 points?



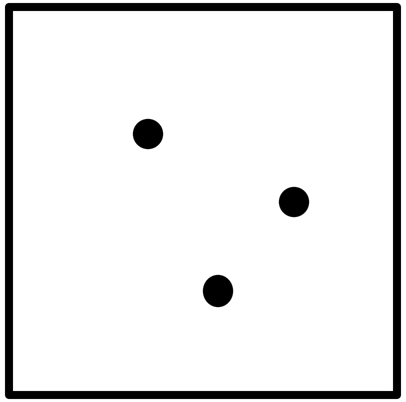
S_1 S_2

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?



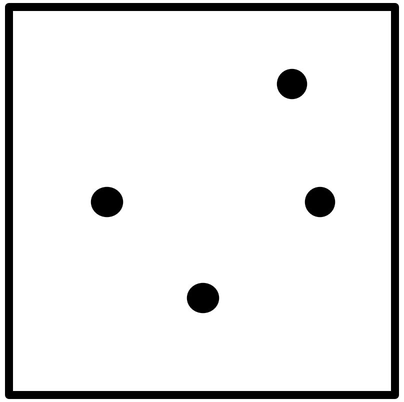
$|\mathcal{H}(S_1)| = 6$



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

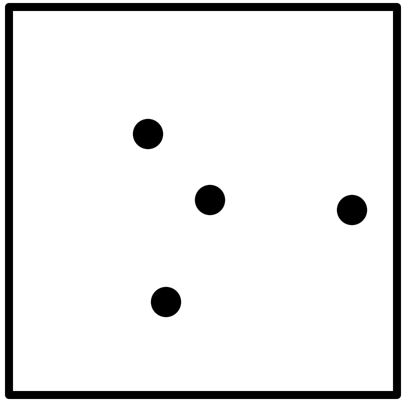
VC-Dimension: Example

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

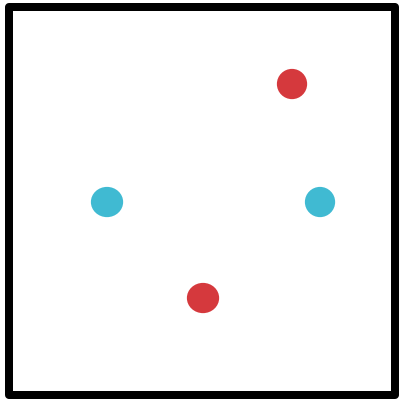


S_2

At least one point inside the convex hull

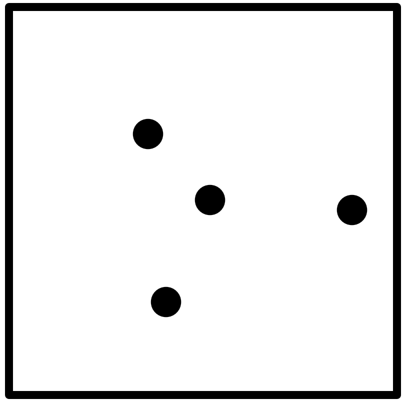
VC-Dimension: Example

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

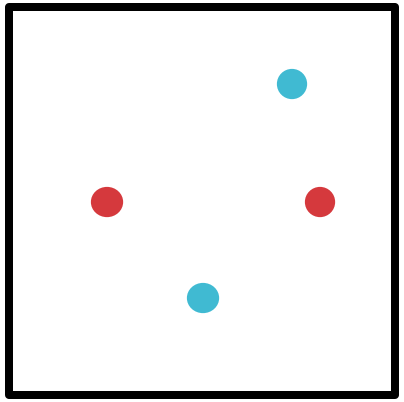


S_2

At least one point inside the convex hull

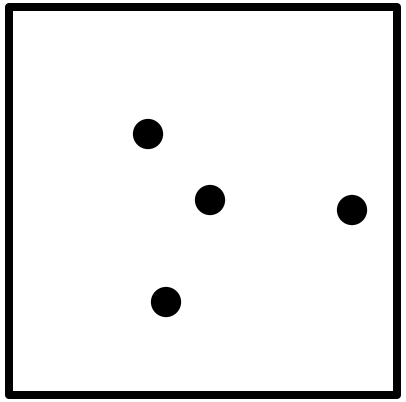
VC-Dimension: Example

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

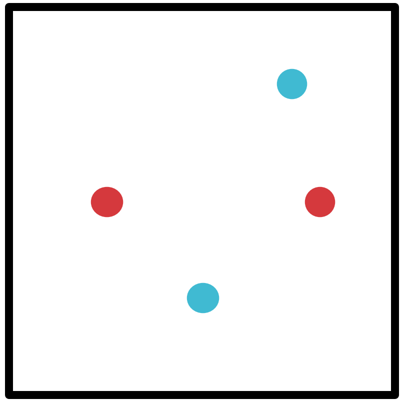


S_2

At least one point inside the convex hull

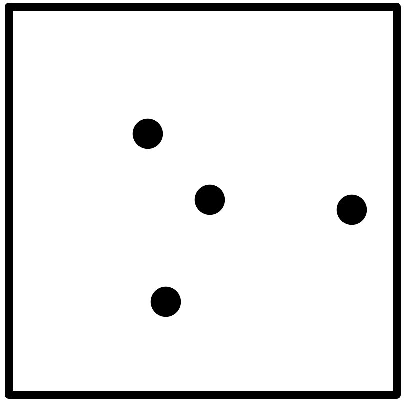
VC-Dimension: Example

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

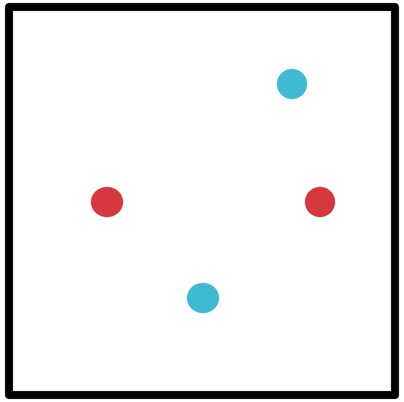


S_2

At least one point inside the convex hull

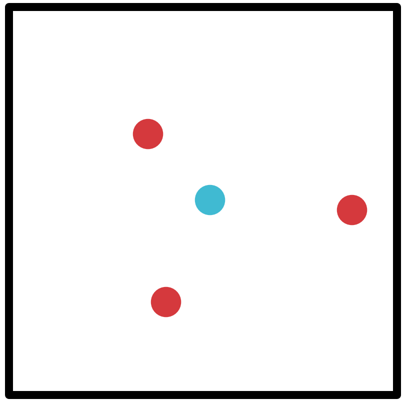
VC-Dimension: Example

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

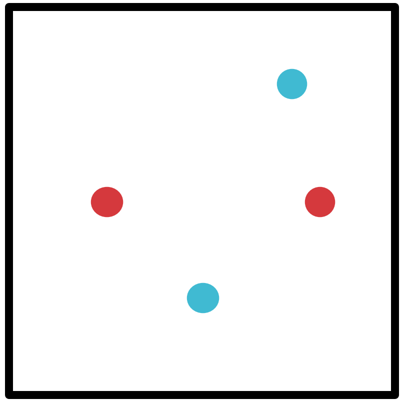


S_2

At least one point inside the convex hull

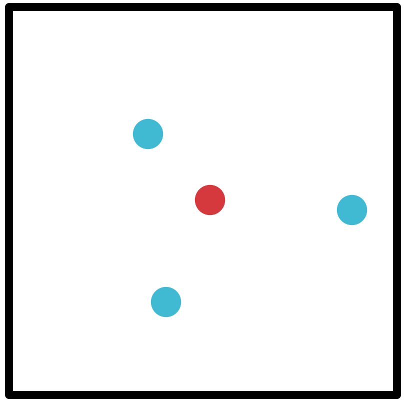
VC-Dimension: Example

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

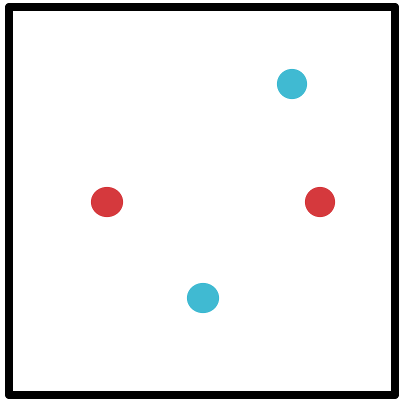


S_2

At least one point inside the convex hull

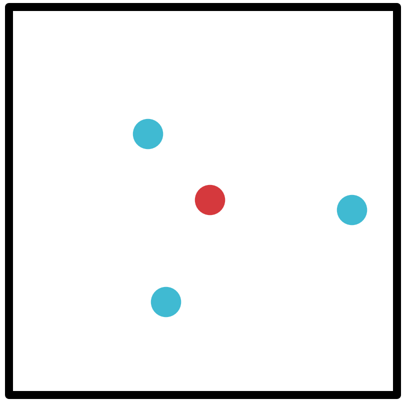
VC-Dimension: Example

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

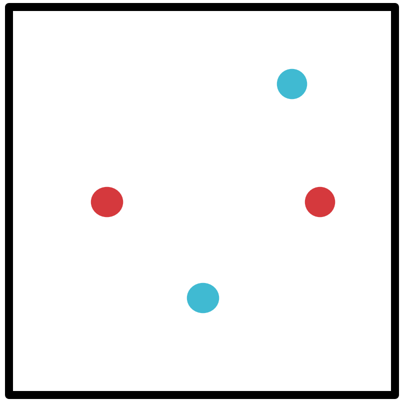


$|\mathcal{H}(S_2)| = 14$

At least one point inside the convex hull

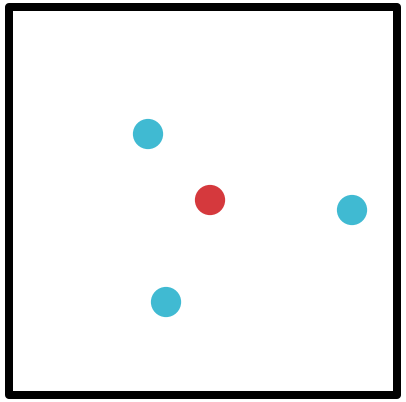
VC-Dimension: Example

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



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

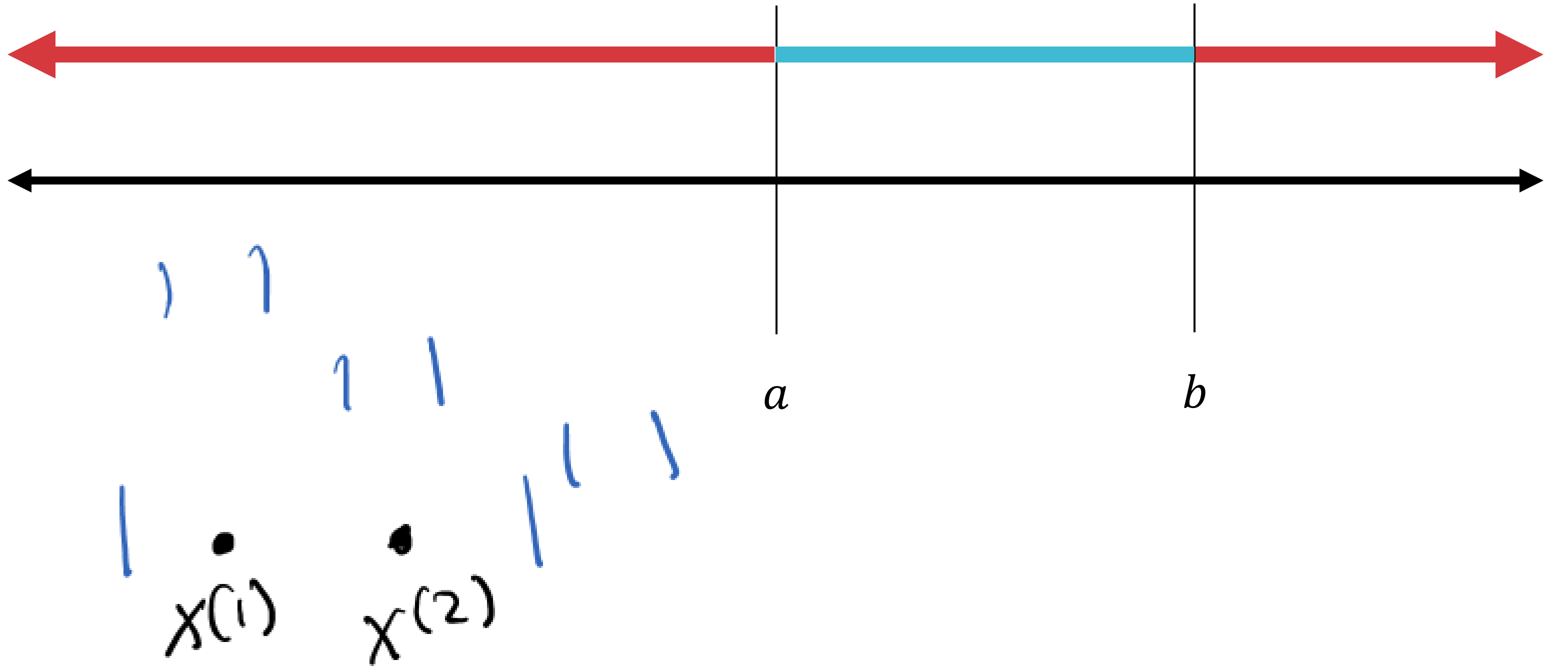
At least one point inside the convex hull

VC-Dimension: Example

- $\boldsymbol{x} \in \mathbb{R}^d$ 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 intervals



What is the VC-dimension of $\mathcal{H}$ = all 1-dimensional positive intervals?

0

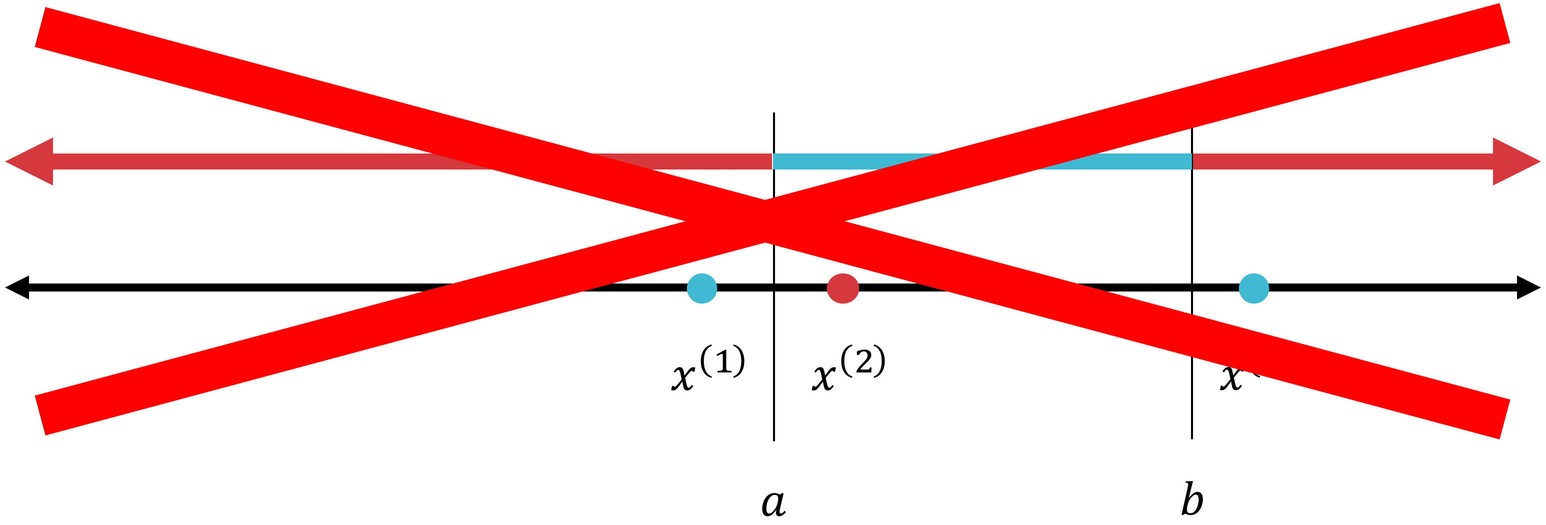
1

2

3

VC-Dimension: Example

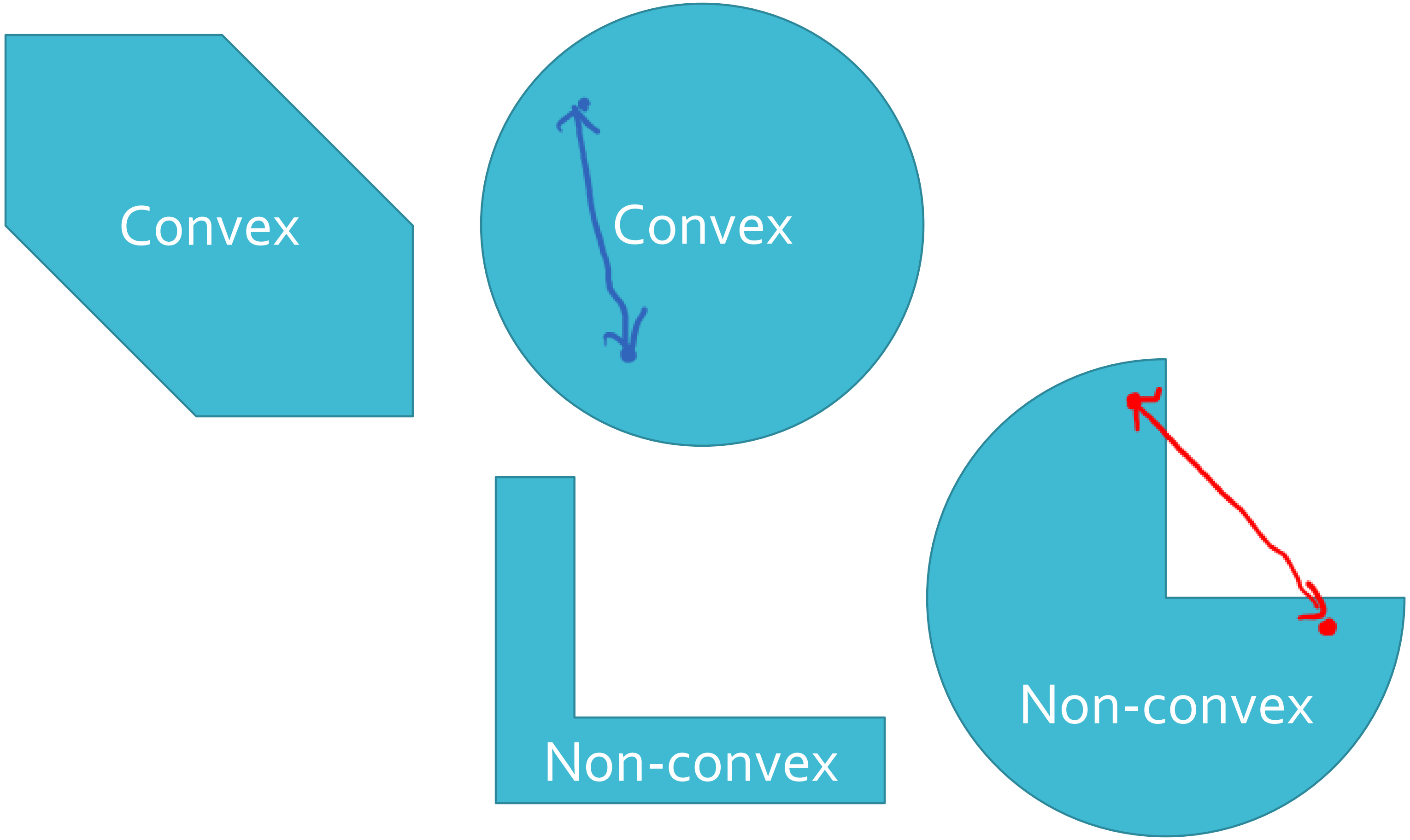
- $x \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive intervals



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

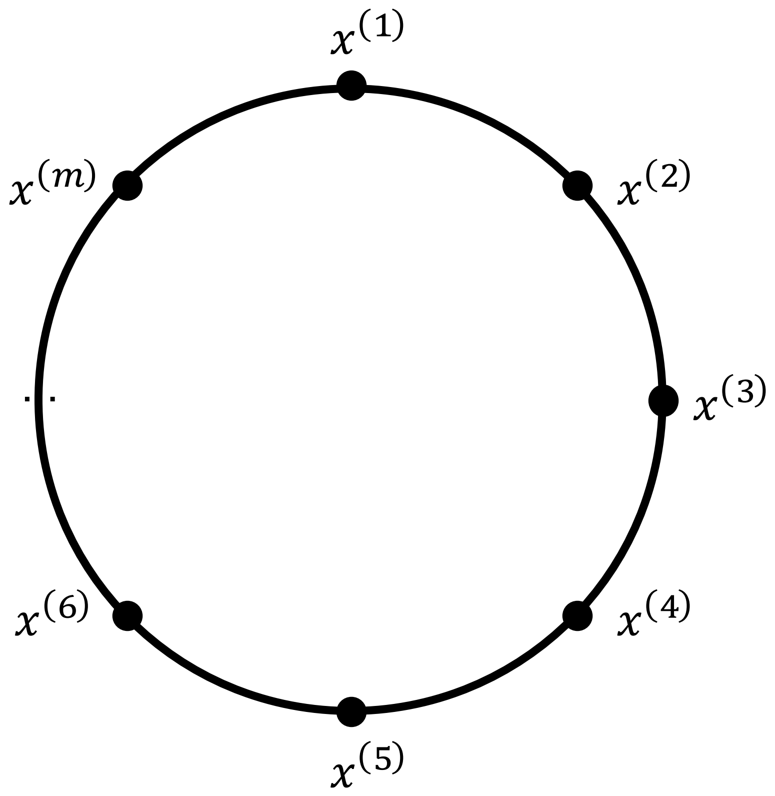
VC-Dimension: Example

- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional positive convex sets



VC-Dimension: Example

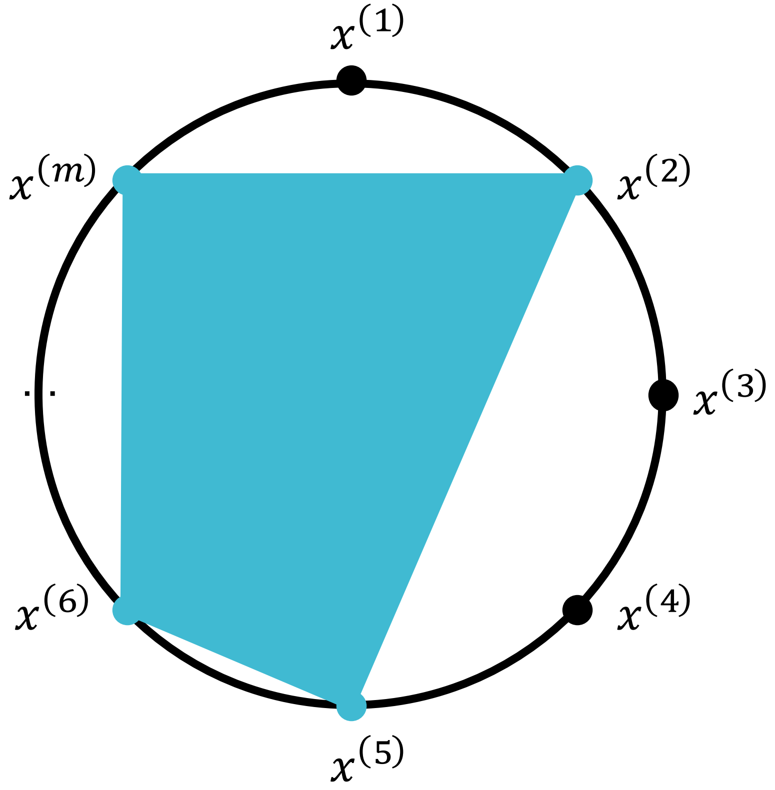
- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional positive convex sets
- What is $d_{VC}(\mathcal{H})$?



VC-Dimension: Example

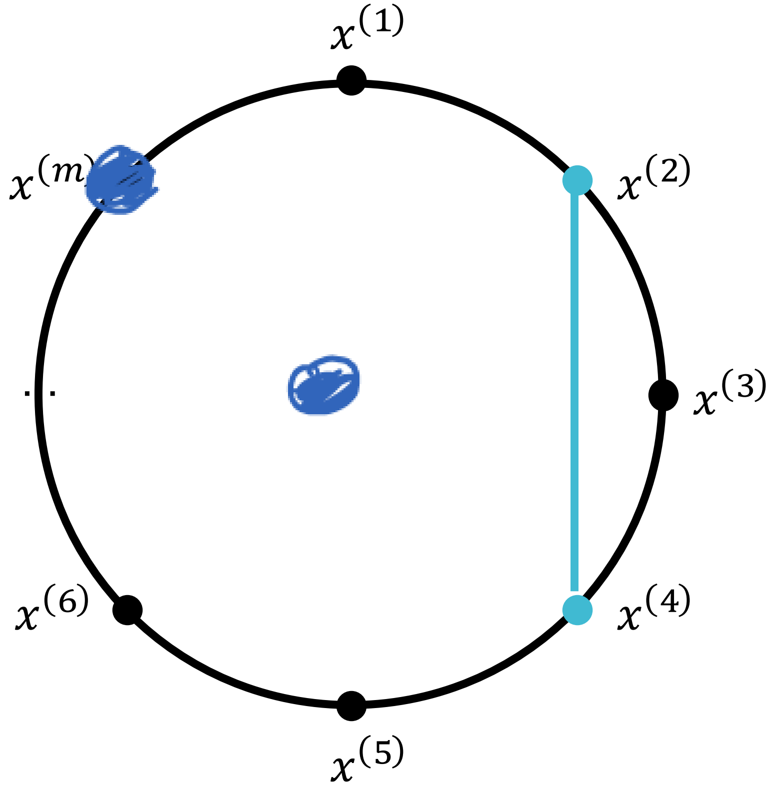
- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional positive convex sets

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



VC-Dimension: Example

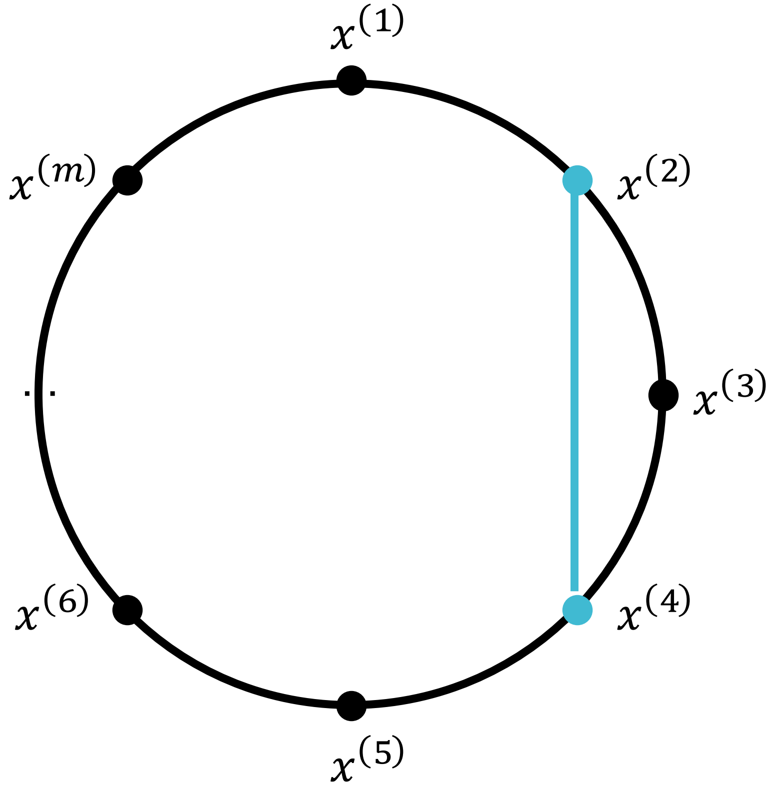
- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional positive convex sets
- What is $d_{VC}(\mathcal{H})$?



VC-Dimension: Example

- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional positive convex sets

- $d_{VC}(\mathcal{H}) = \infty$!



Theorem 3: Vapnik-Chervonenkis (VC)-Bound

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

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

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 $\mathcal{H}$ 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

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

$$|R(h) - \hat{R}(h)| \le O\left(\sqrt{\frac{1}{M}\left(VC(H) + \ln(1/\delta)\right)}\right)$$

$$-O\left(\sqrt{\frac{1}{M}\left(VC(H) + \ln(1/\delta)\right)}\right) \le R(h) - \hat{R}(h) \le O\left(\sqrt{\frac{1}{M}\left(VC(H) + \ln(1/\delta)\right)}\right)$$

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?

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

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

Key Takeaways

- For infinite hypothesis sets, use the VC-dimension (or the growth function) as a measure of complexity
  - Computing $d_{VC}(\mathcal{H})$
  - Sample complexity and statistical learning theory style bounds using $d_{VC}(\mathcal{H})$