

10-301/601: Introduction to Machine Learning Lecture 25 – Learning Theory (Infinite Case)

Henry Chai

7/16/24

Front Matter

- Announcements
 - HW8 released 7/11, due 7/18 at 11:59 PM
 - **Recitation on 7/18 will be held online via Zoom**
 - See Piazza for the Zoom link
- Recommended Readings
 - Mitchell, Chapter 7.4

Recall: Statistical Learning Theory Corollary

- For a finite hypothesis set \mathcal{H} and arbitrary distribution p^* , given a training data set S s.t. $|S| = M$, 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 $|\mathcal{H}| = \infty$?

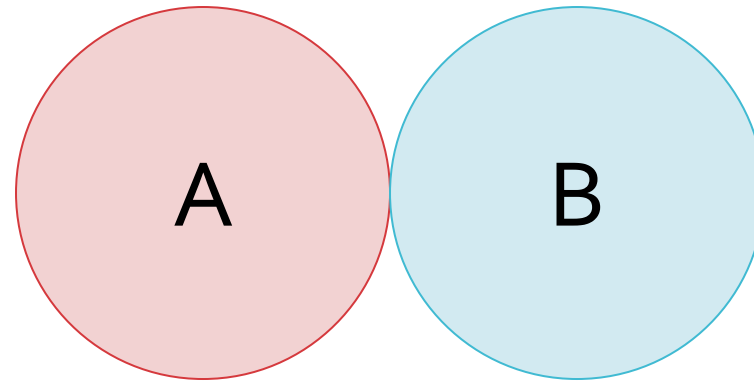
- For a finite hypothesis set \mathcal{H} and arbitrary distribution p^* , given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ have

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

with probability at least $1 - \delta$.

The Union Bound...

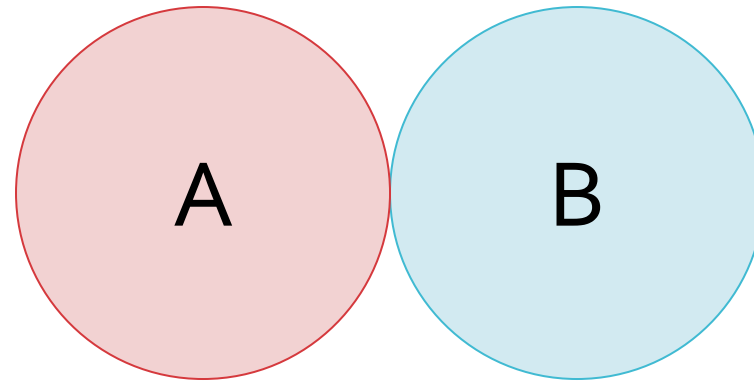
$$P\{A \cup B\} \leq P\{A\} + P\{B\}$$



The Union Bound is Bad!

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

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

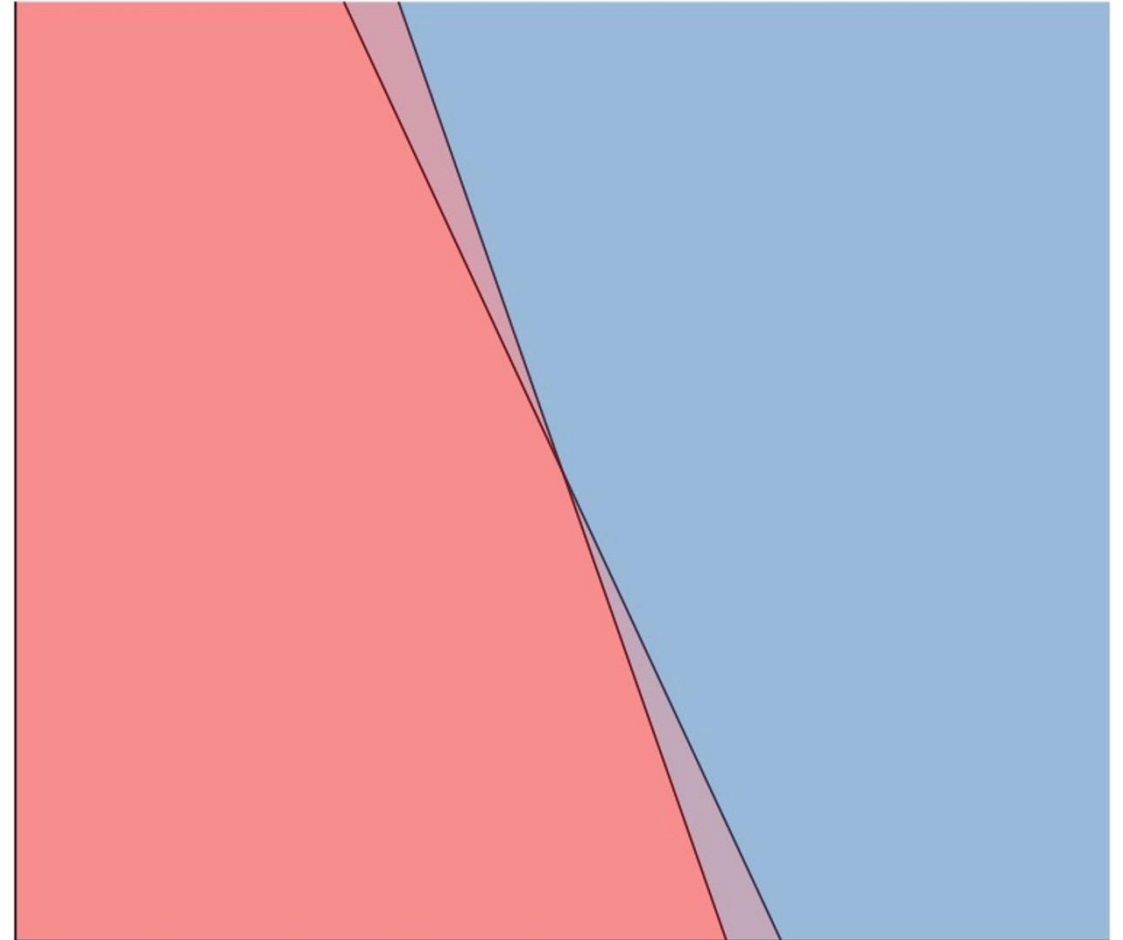


Intuition

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!

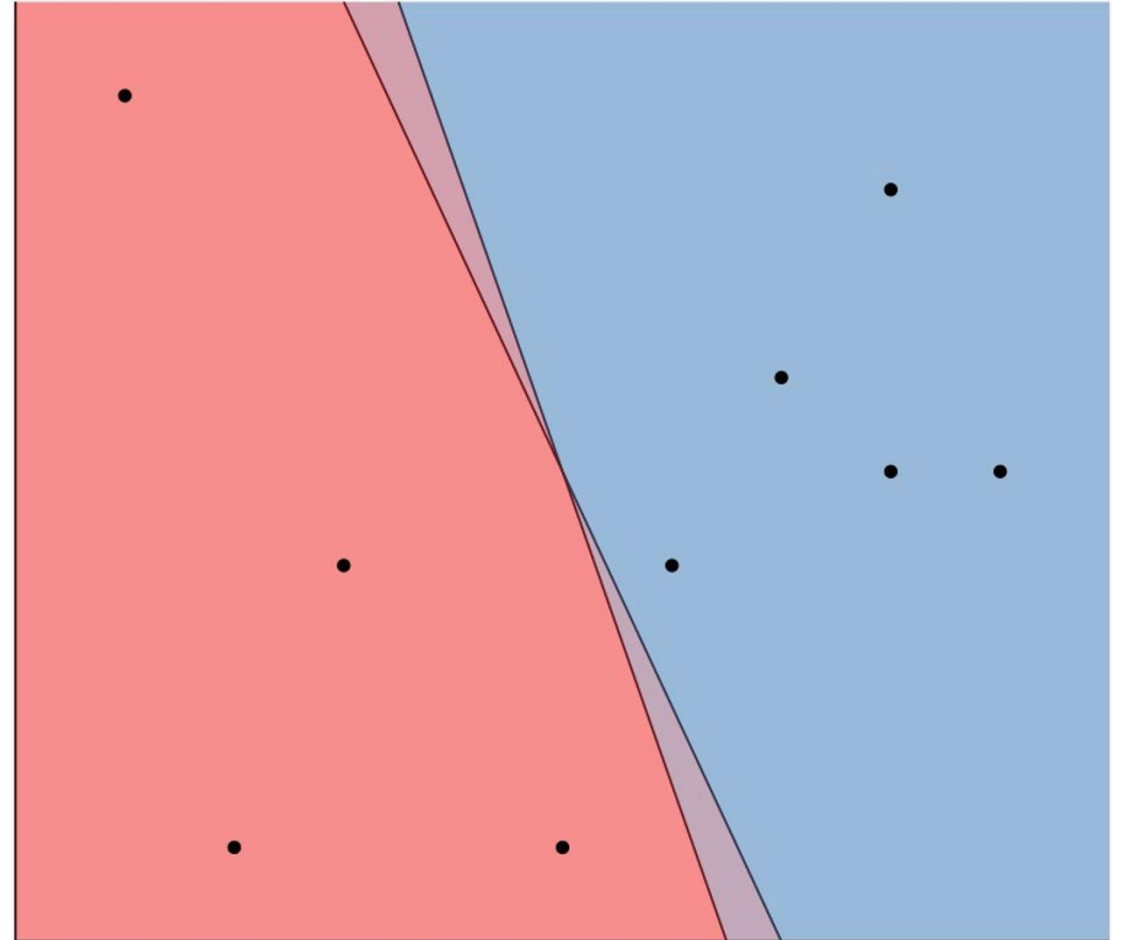


Intuition

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!

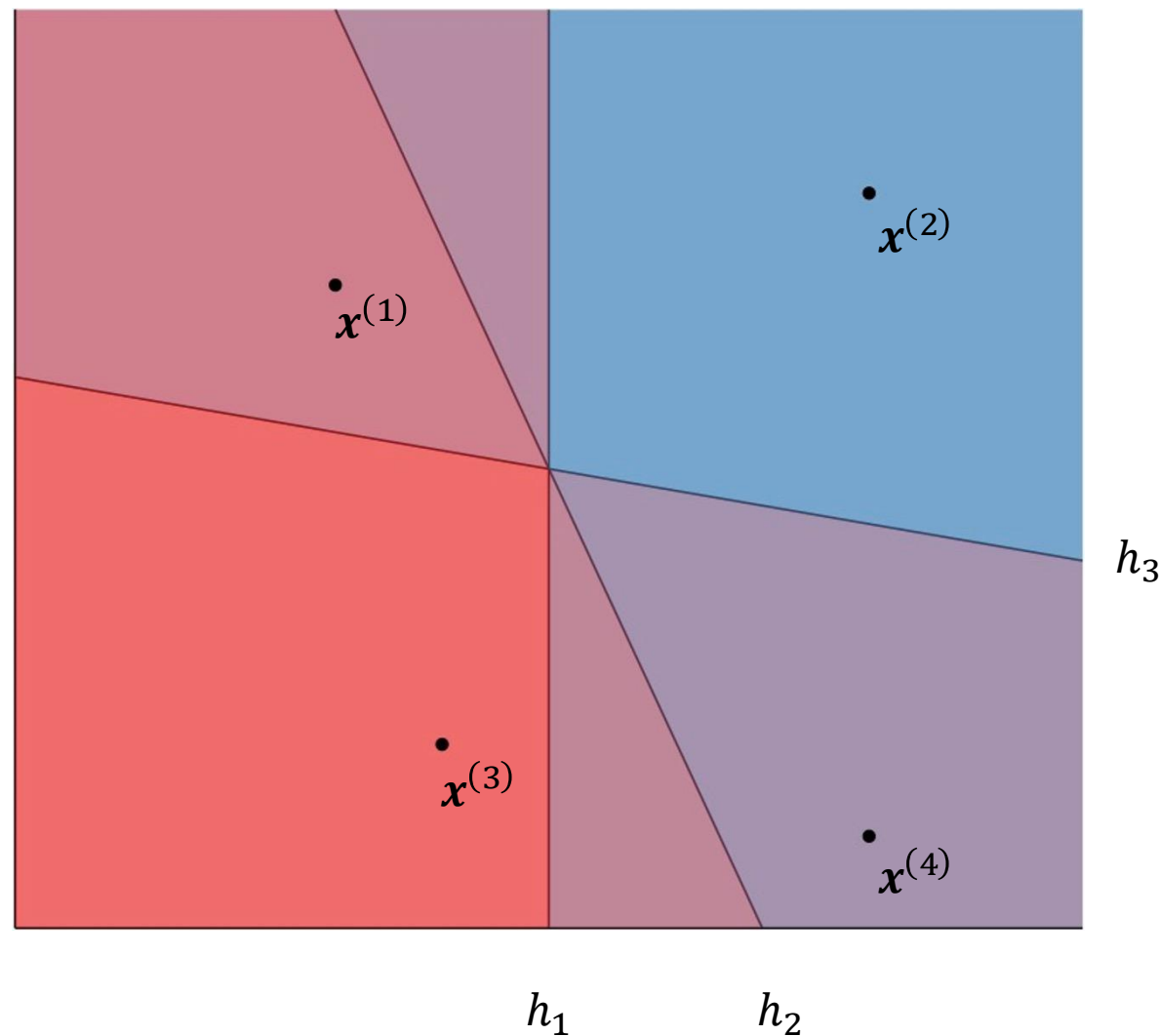


Labellings

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

Example: Labellings

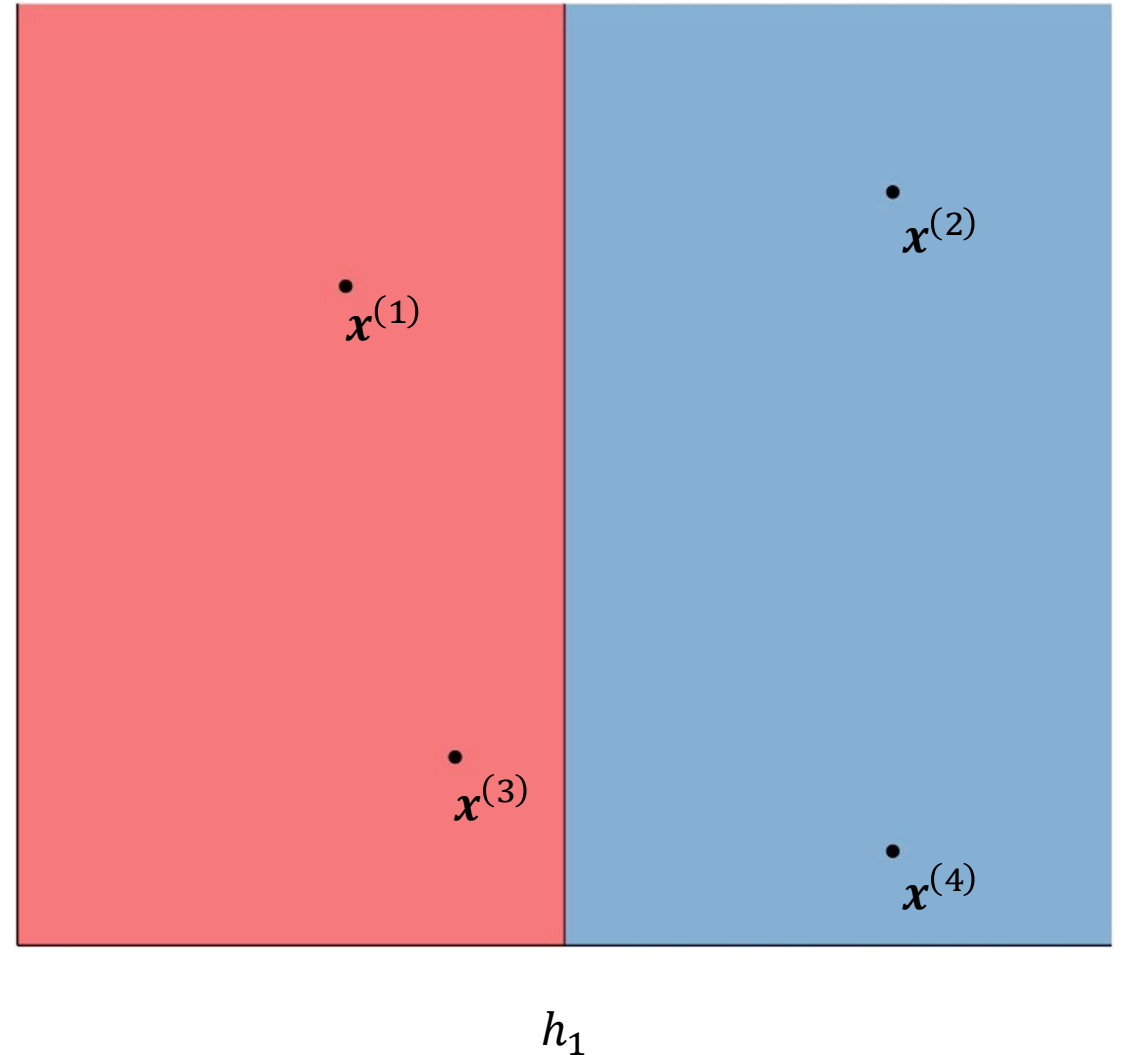
$$\mathcal{H} = \{h_1, h_2, h_3\}$$



Example: Labellings

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

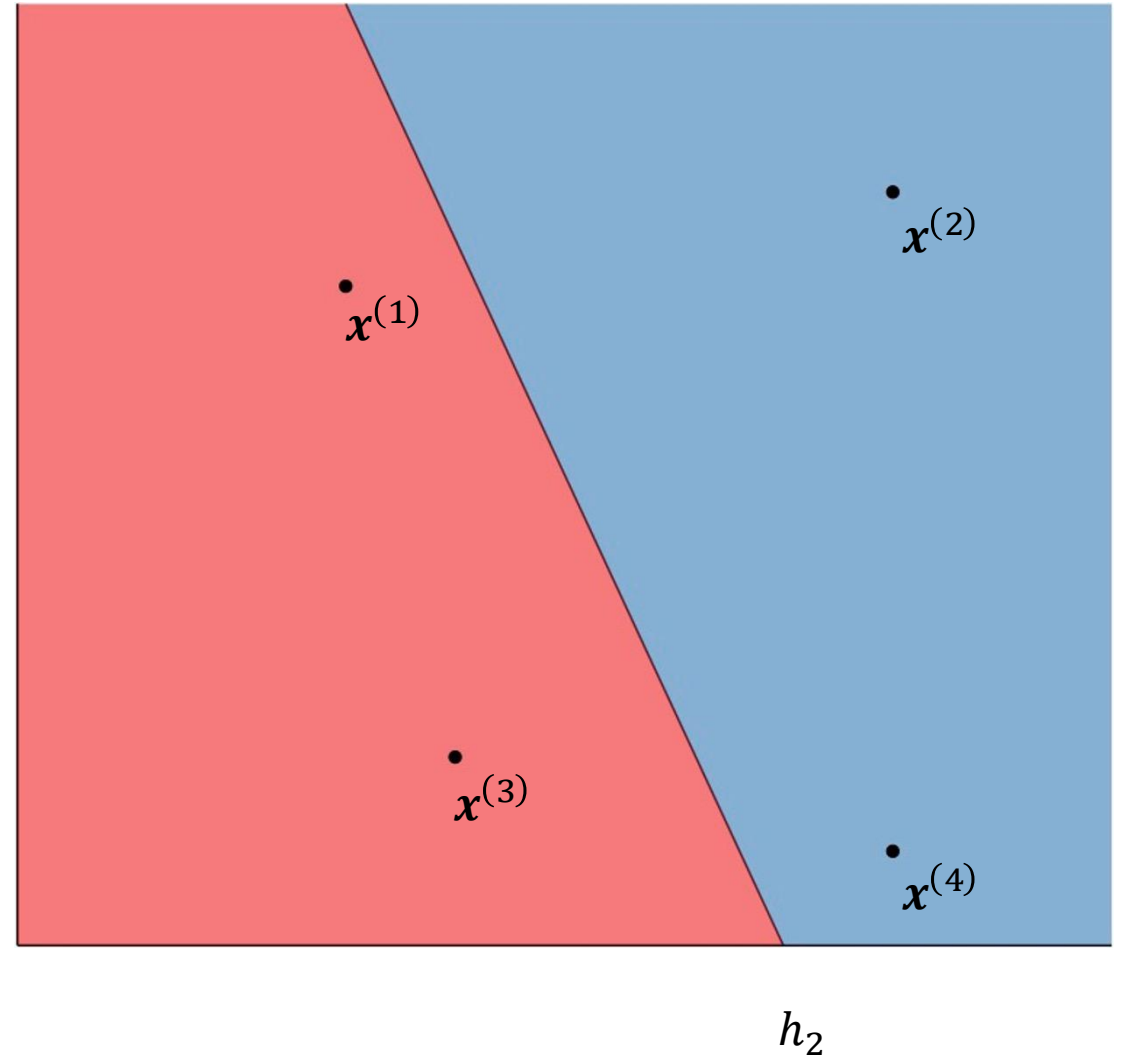
$$\begin{aligned} & \left(h_1(\mathbf{x}^{(1)}), h_1(\mathbf{x}^{(2)}), h_1(\mathbf{x}^{(3)}), h_1(\mathbf{x}^{(4)}) \right) \\ & = (-1, +1, -1, +1) \end{aligned}$$



Example: Labellings

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

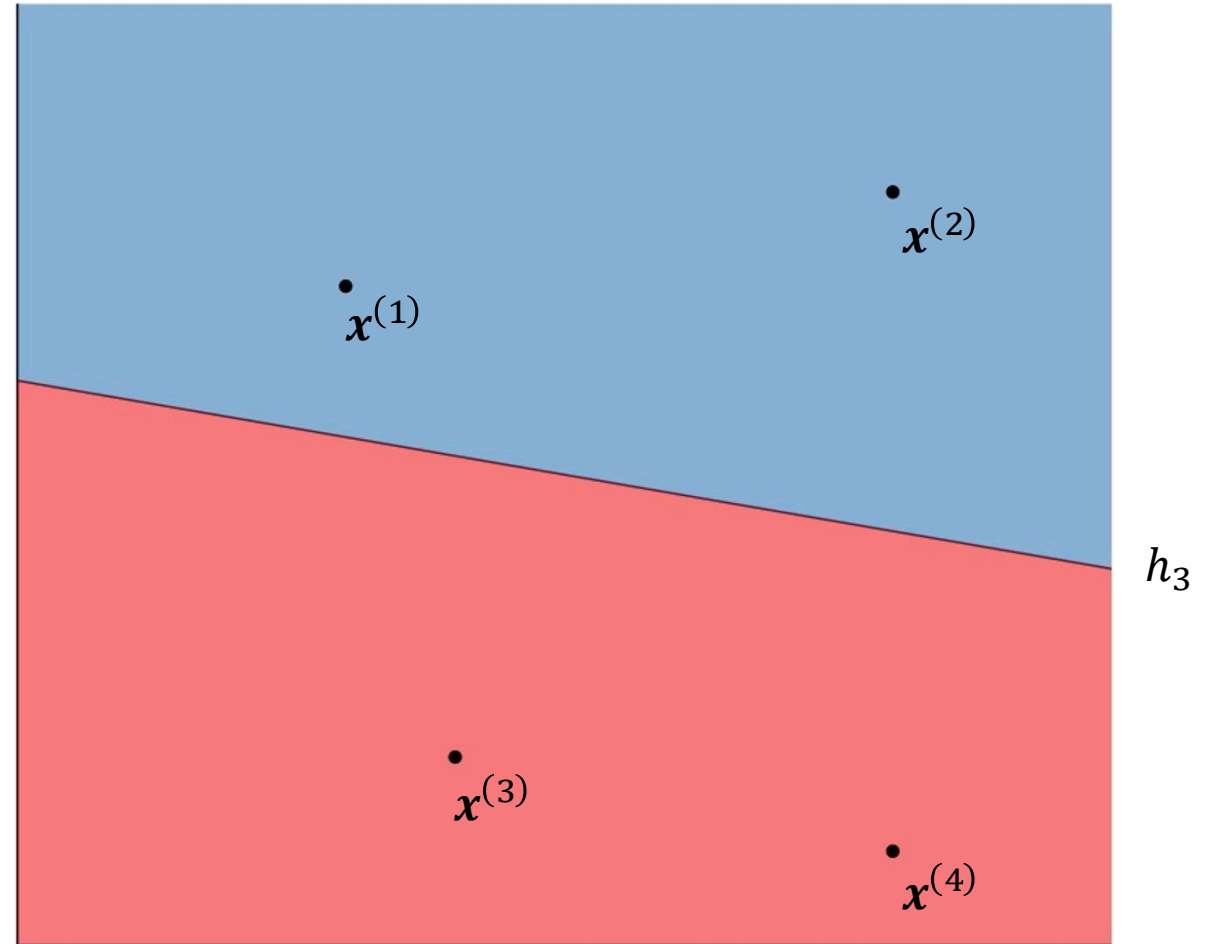
$$\begin{aligned} & \left(h_2(\mathbf{x}^{(1)}), h_2(\mathbf{x}^{(2)}), h_2(\mathbf{x}^{(3)}), h_2(\mathbf{x}^{(4)}) \right) \\ & = (-1, +1, -1, +1) \end{aligned}$$



Example: Labellings

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

$$\begin{aligned} & \left(h_3(\mathbf{x}^{(1)}), h_3(\mathbf{x}^{(2)}), h_3(\mathbf{x}^{(3)}), h_3(\mathbf{x}^{(4)}) \right) \\ & = (+1, +1, -1, -1) \end{aligned}$$

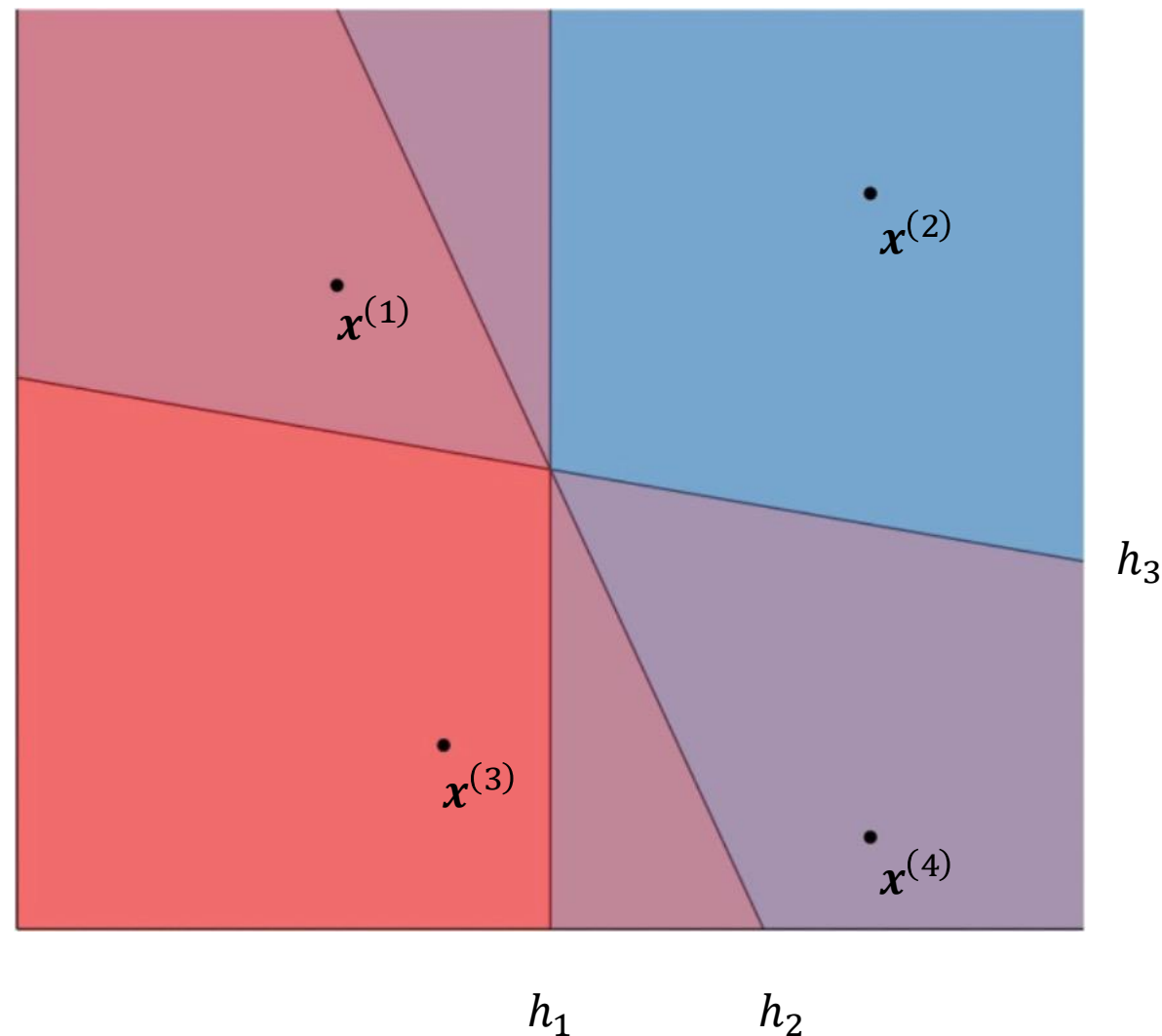


Example: Labellings

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

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

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

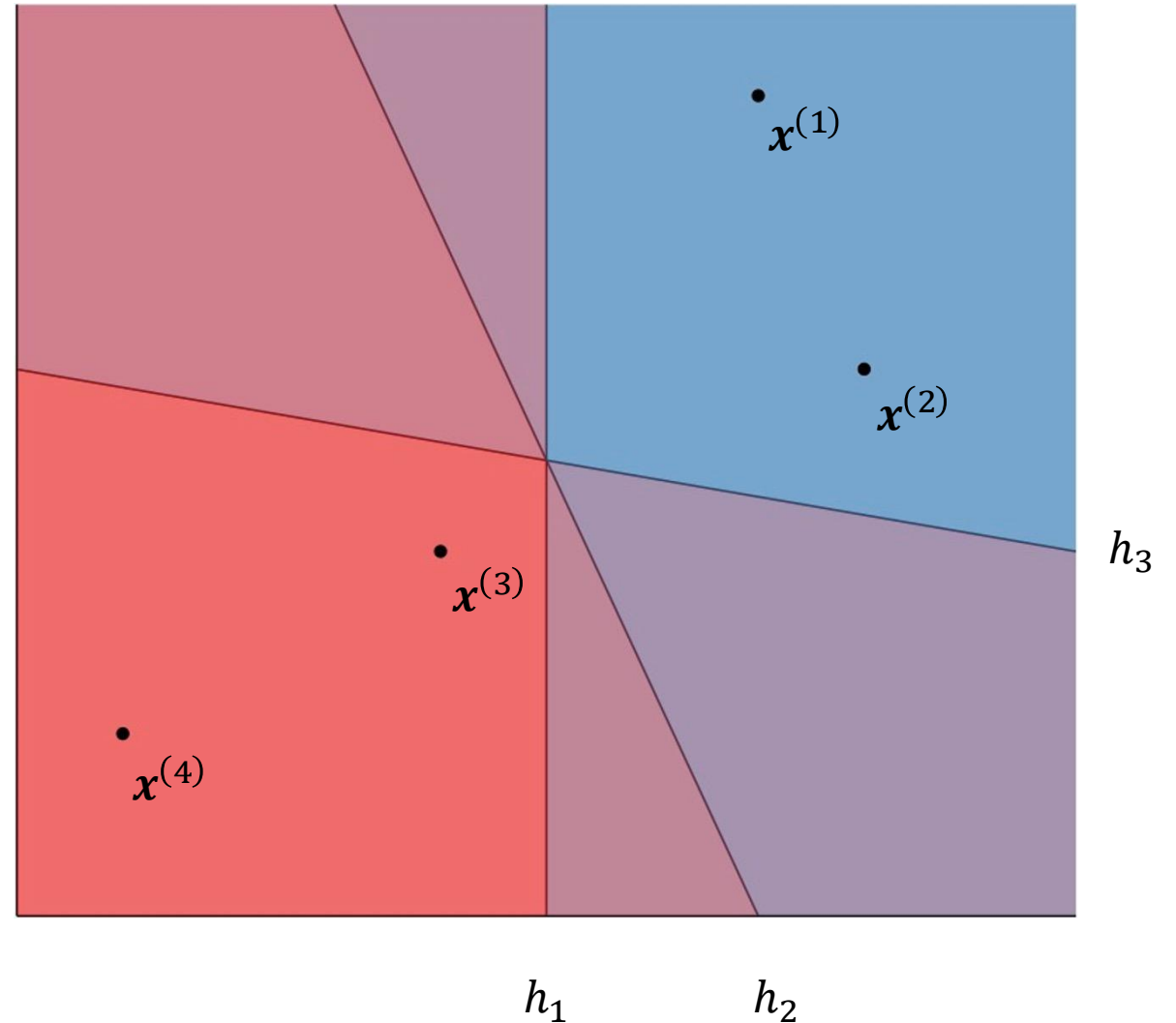


Example: Labellings

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

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

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



Growth Function

- The growth function of \mathcal{H} is the maximum number of distinct labellings \mathcal{H} can induce on *any* set of M data points:

$$g_{\mathcal{H}}(M) = \max_{S: |S|=M} |\mathcal{H}(S)|$$

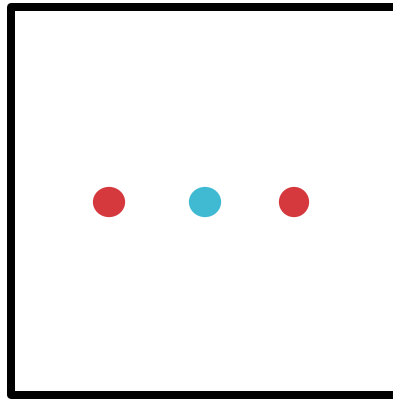
- $g_{\mathcal{H}}(M) \leq 2^M \forall \mathcal{H}$ and M
- \mathcal{H} shatters S if $|\mathcal{H}(S)| = 2^M$
- If $\exists S$ s.t. $|S| = M$ and \mathcal{H} shatters S , then $g_{\mathcal{H}}(M) = 2^M$

Growth Function: Example

- $\mathbf{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(3)$?

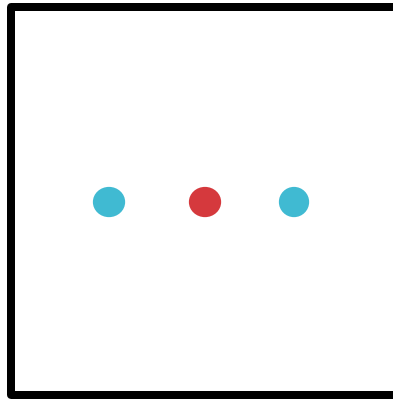
Growth Function: Example

- $\mathbf{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(3)$?



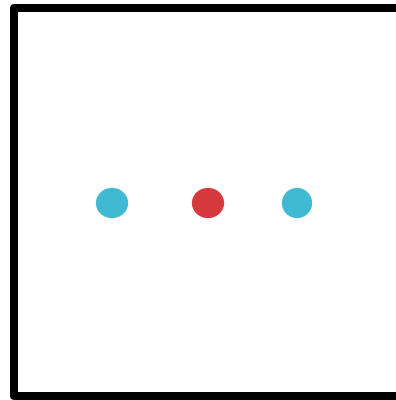
Growth Function: Example

- $\mathbf{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(3)$?

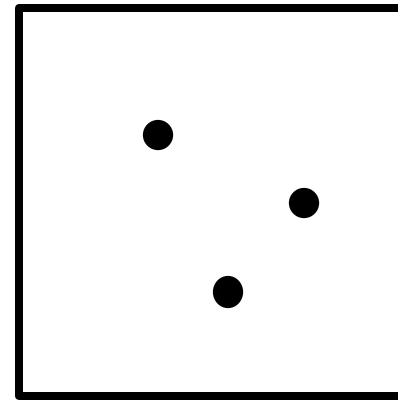


Growth Function: Example

- $\mathbf{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(3)$?



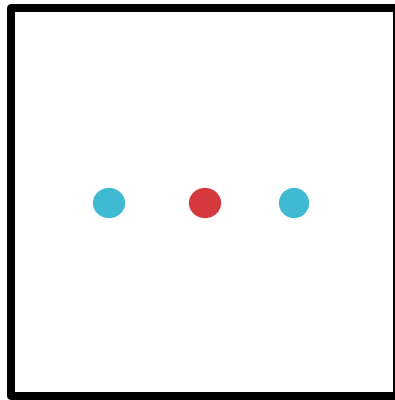
$$|\mathcal{H}(S_1)| = 6$$



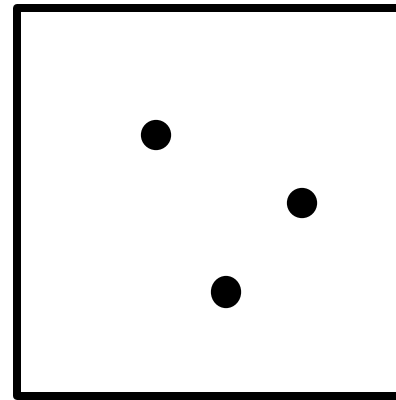
$$|\mathcal{H}(S_2)| = 8$$

Growth Function: Example

- $\mathbf{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- $g_{\mathcal{H}}(3) = 8 = 2^3$



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



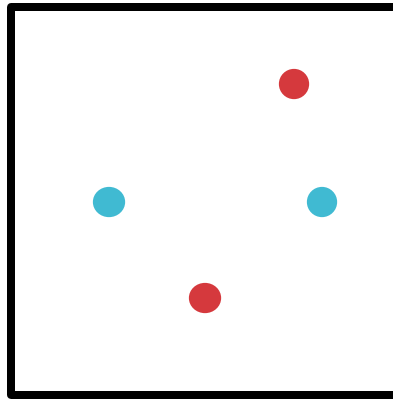
$$|\mathcal{H}(S_2)| = 8$$

Growth Function: Example

- $\mathbf{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(4)$?

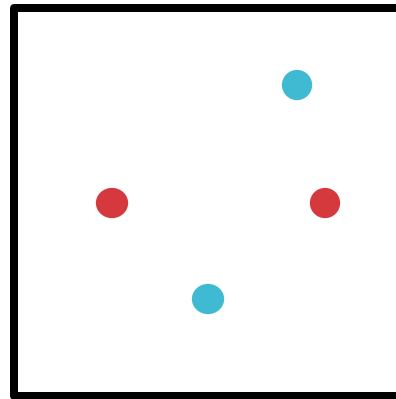
Growth Function: Example

- $\mathbf{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(4)$?



Growth Function: Example

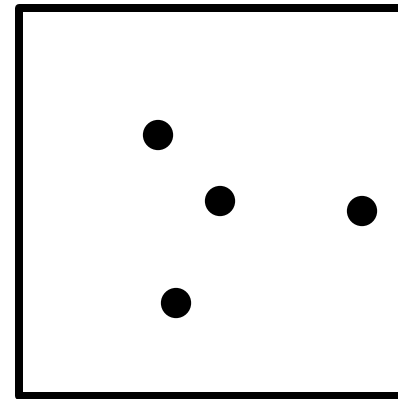
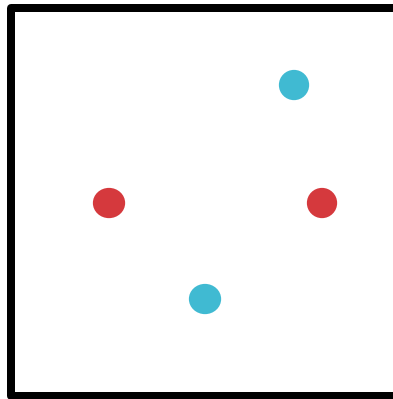
- $\mathbf{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(4)$?



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

Growth Function: Example

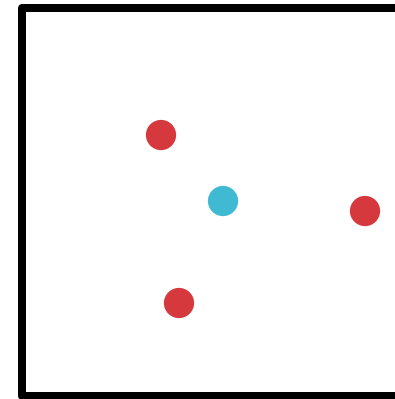
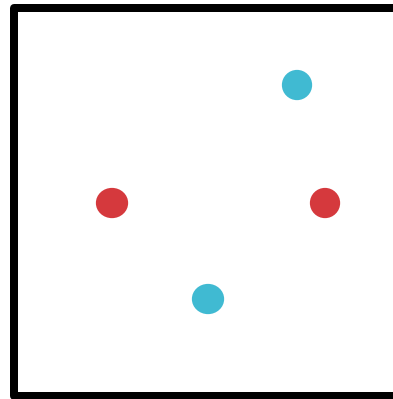
- $\mathbf{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(4)$?



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

Growth Function: Example

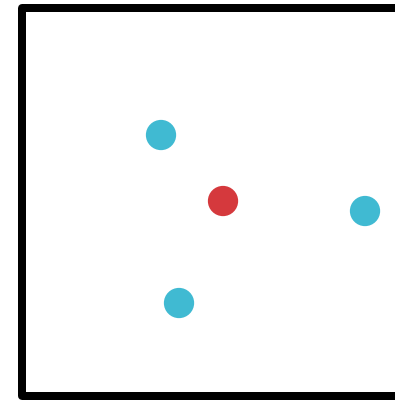
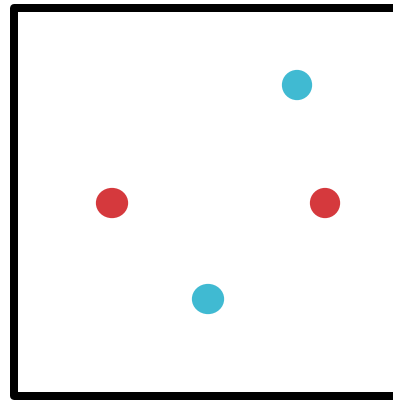
- $\mathbf{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(4)$?



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

Growth Function: Example

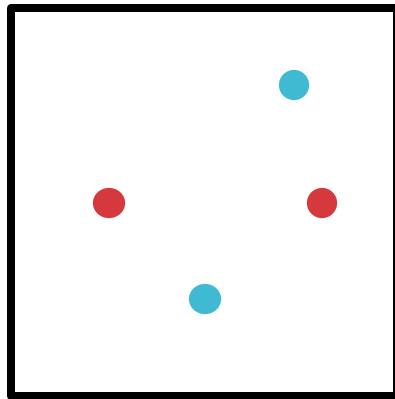
- $\mathbf{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(4)$?



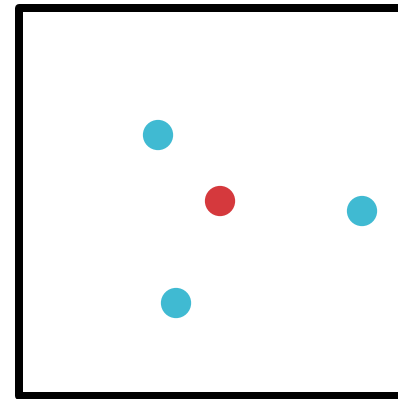
$$|\mathcal{H}(S_1)| = 14$$

Growth Function: Example

- $\mathbf{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- $g_{\mathcal{H}}(4) = 14 < 2^4$



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



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

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

$$M \geq \frac{2}{\epsilon} \left(\log_2(2g_{\mathcal{H}}(2M)) + \log_2\left(\frac{1}{\delta}\right) \right)$$

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

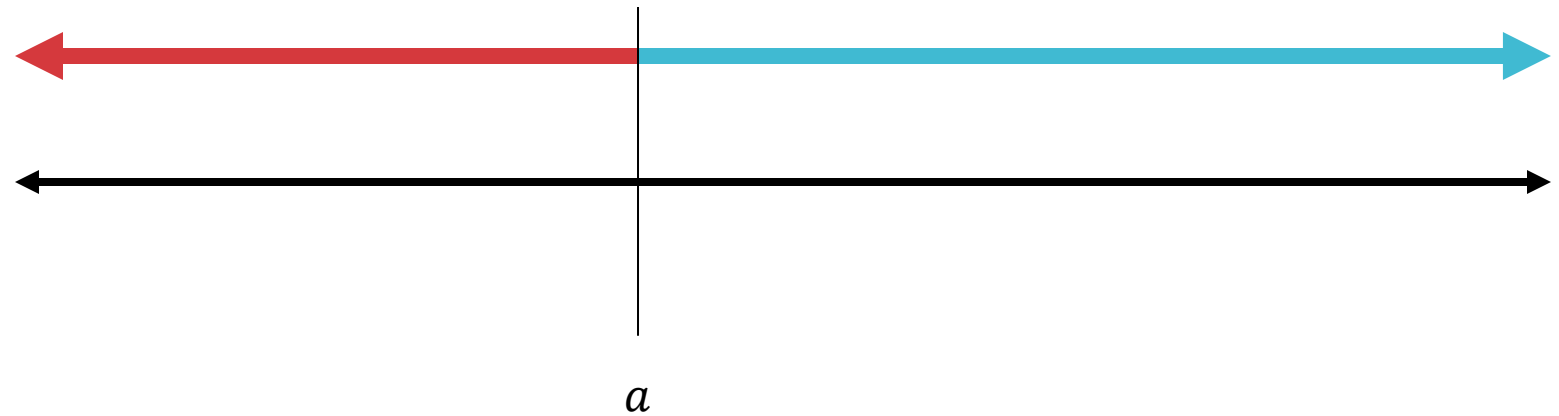
- M appears on both sides of the inequality...

Vapnik- Chervonenkis (VC)-Dimension

- $d_{VC}(\mathcal{H})$ = the largest value of M s.t. $g_{\mathcal{H}}(M) = 2^M$, i.e., the greatest number of data points that can be shattered by \mathcal{H}
 - If \mathcal{H} can shatter arbitrarily large finite sets, then $d_{VC}(\mathcal{H}) = \infty$
 - $g_{\mathcal{H}}(M) = O(M^{d_{VC}(\mathcal{H})})$ (Sauer-Shelah lemma)
- To prove that $d_{VC}(\mathcal{H}) = C$, you need to show
 1. \exists some set of C data points that \mathcal{H} can shatter and
 2. \nexists a set of $C + 1$ data points that \mathcal{H} can shatter

VC-Dimension: Example

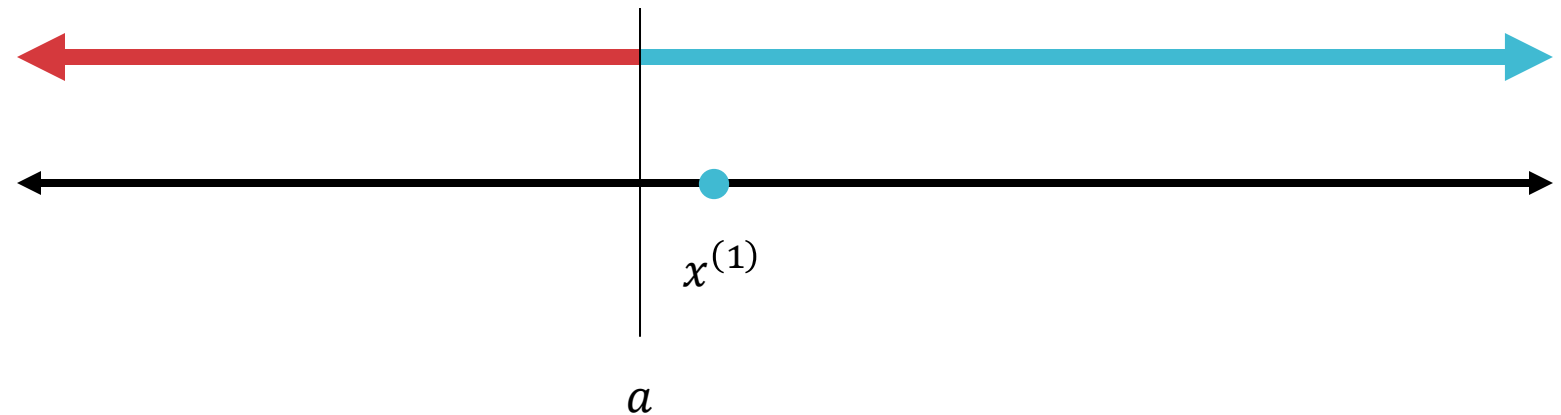
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

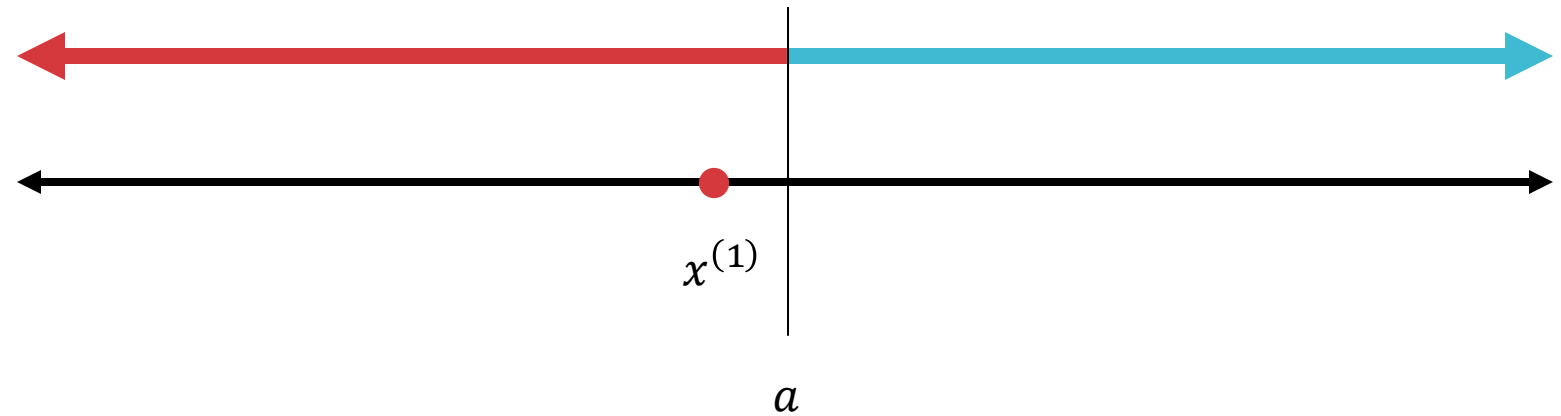
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

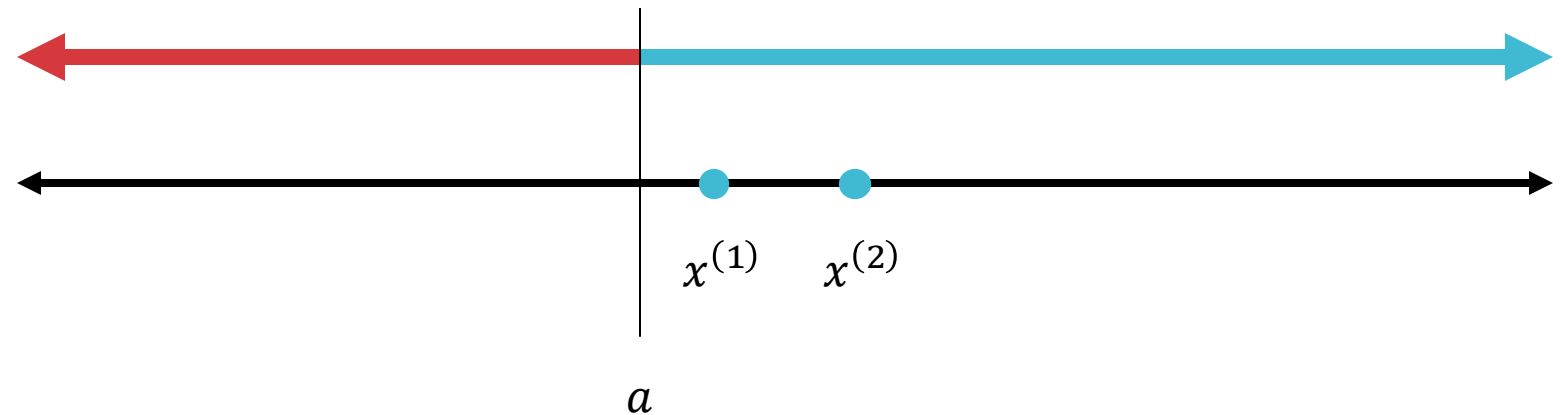
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

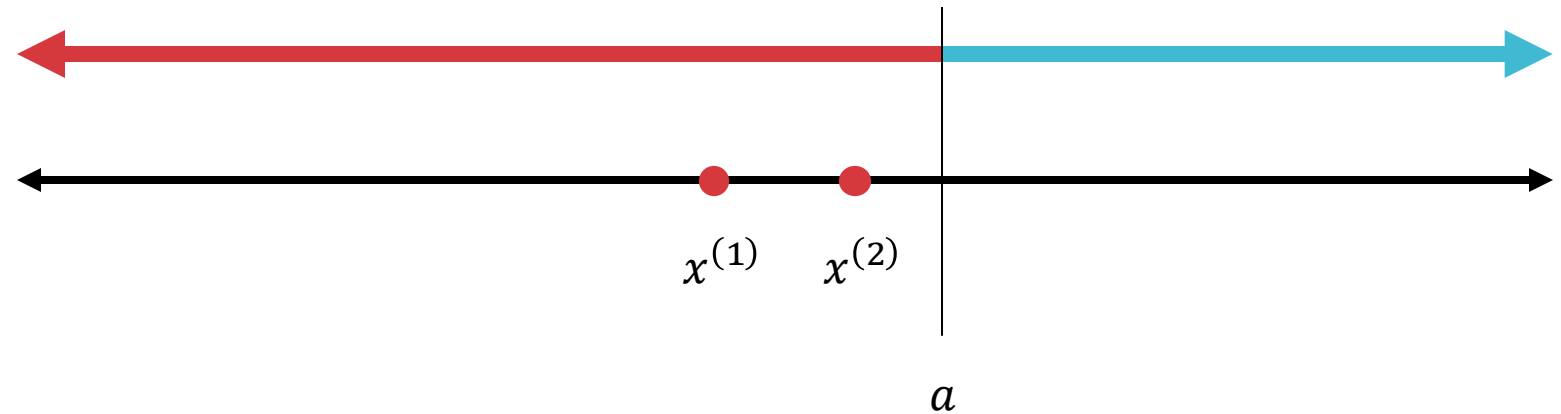
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

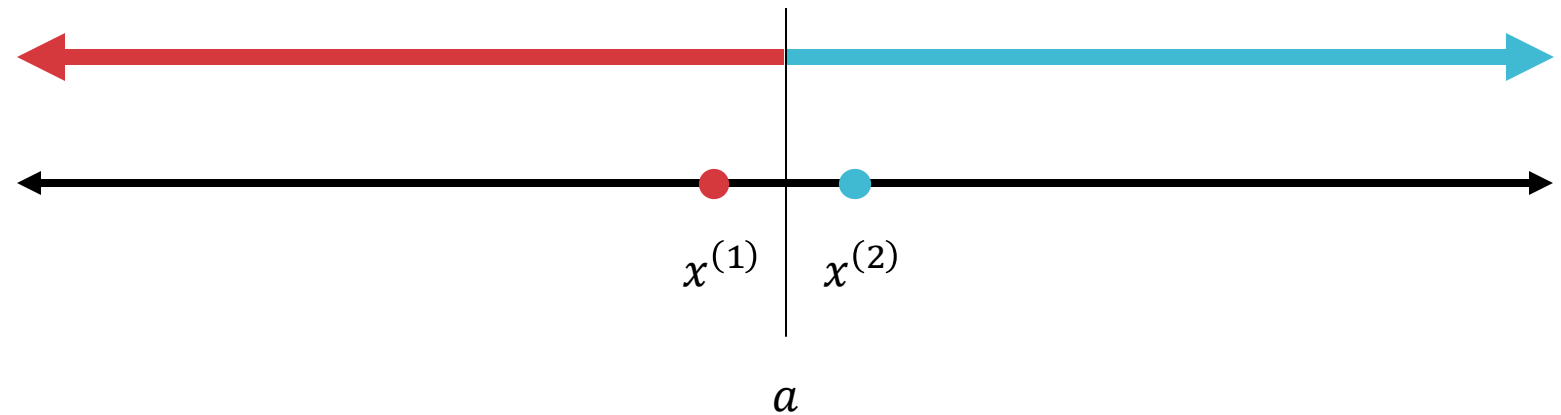
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

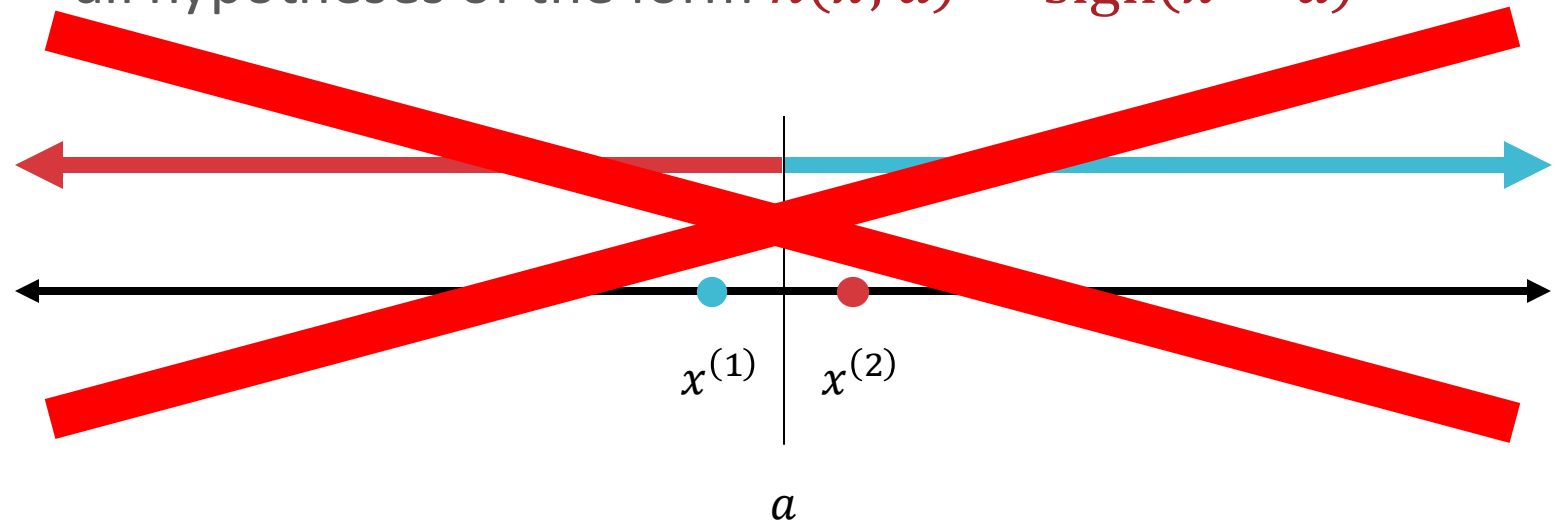
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

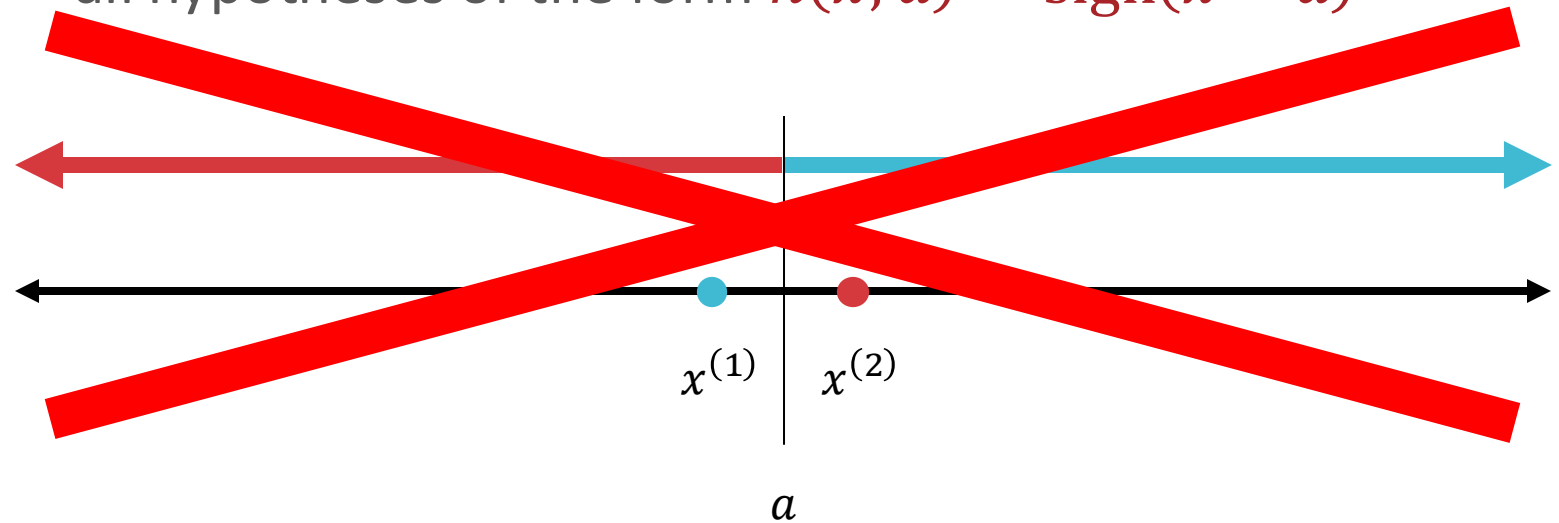
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

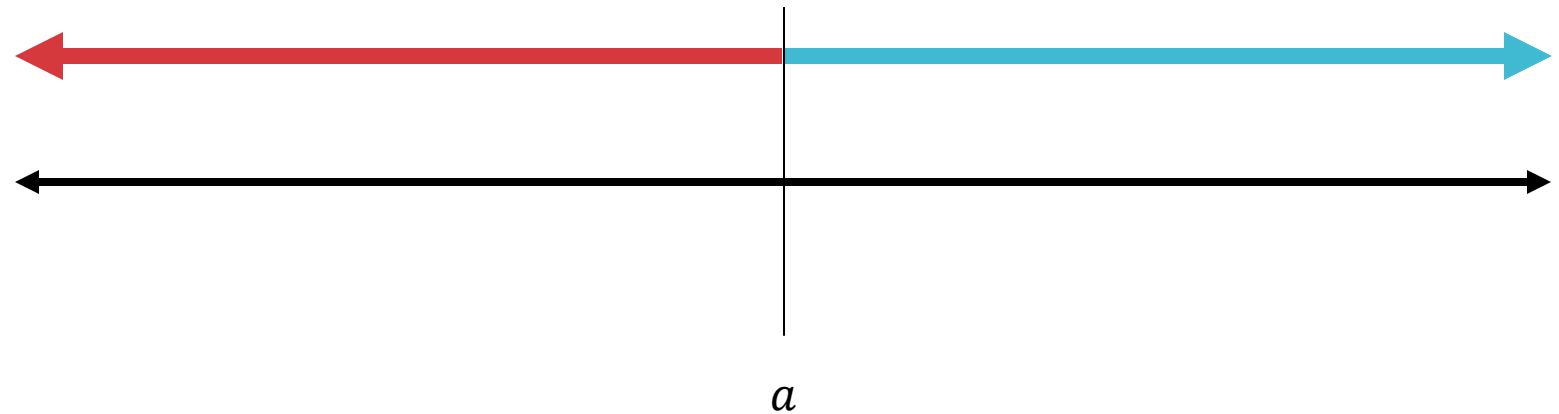
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

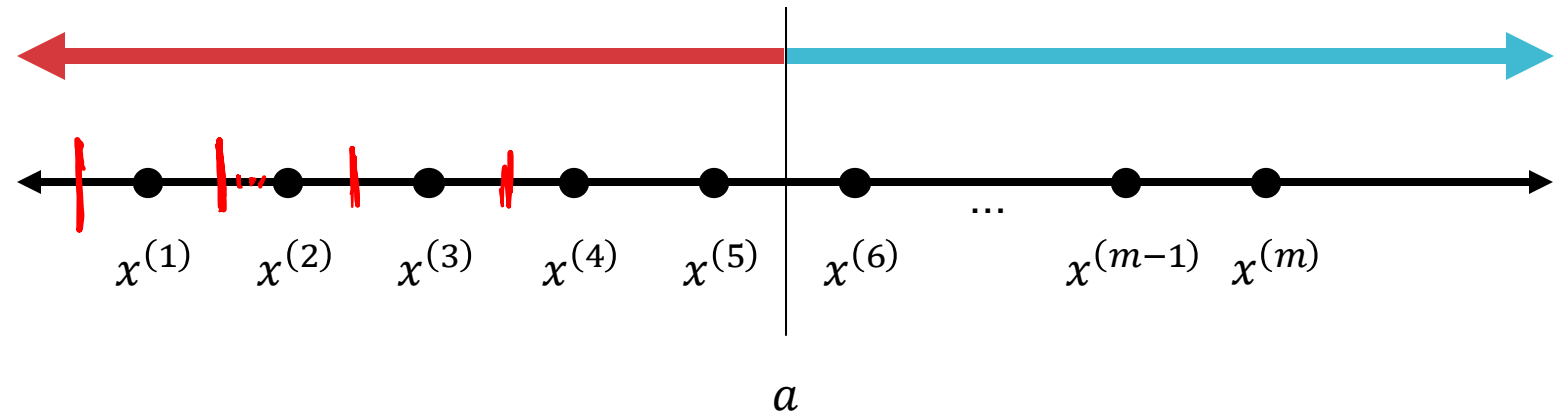
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- What is $g_{\mathcal{H}}(m)$?

VC-Dimension: Example

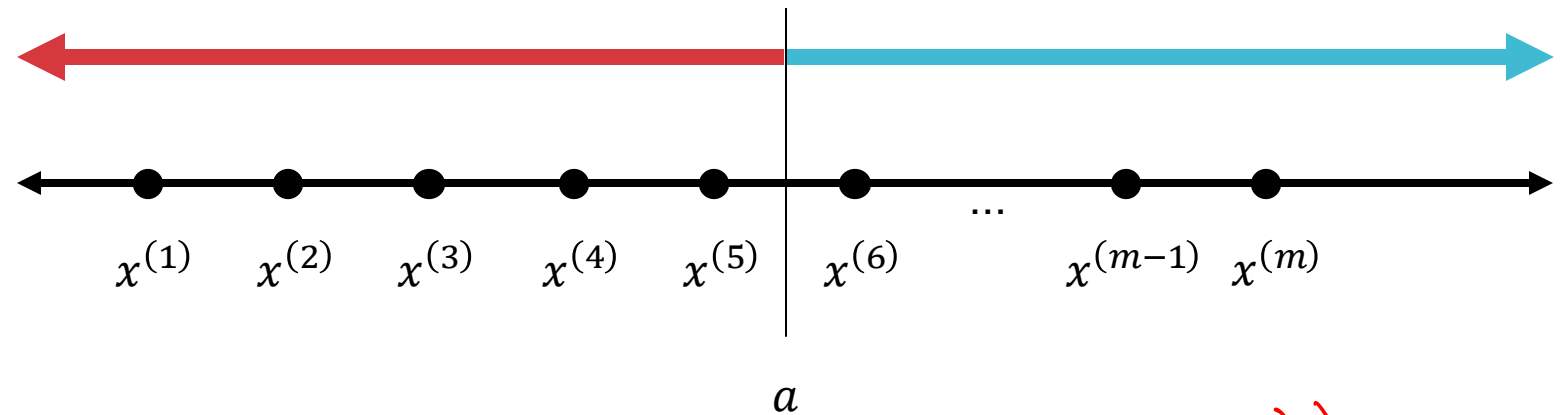
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- What is $g_{\mathcal{H}}(m)$?

VC-Dimension: Example

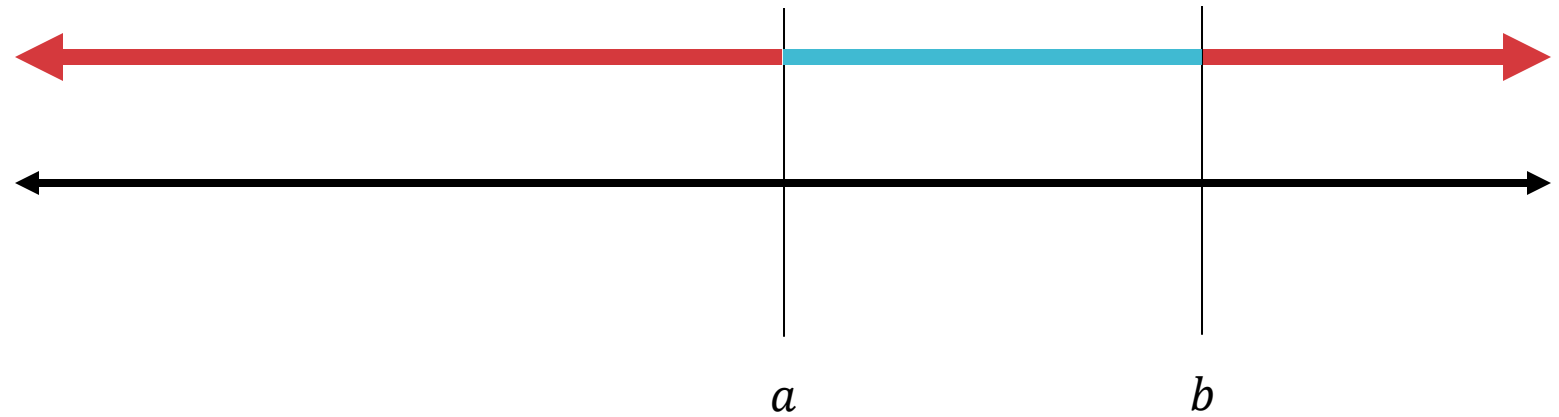
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- $g_{\mathcal{H}}(m) = m + 1 = O(m^1) \cong O(m^{\text{d}_{ve}(\mathcal{H})})$

VC-Dimension: Example

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



What are $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(m)$ for 1-dimensional positive *intervals*?

1 and $m + 1$

0%

2 and $m + 1$

0%

2 and $\frac{1}{2}(m^2 + m + 1)$

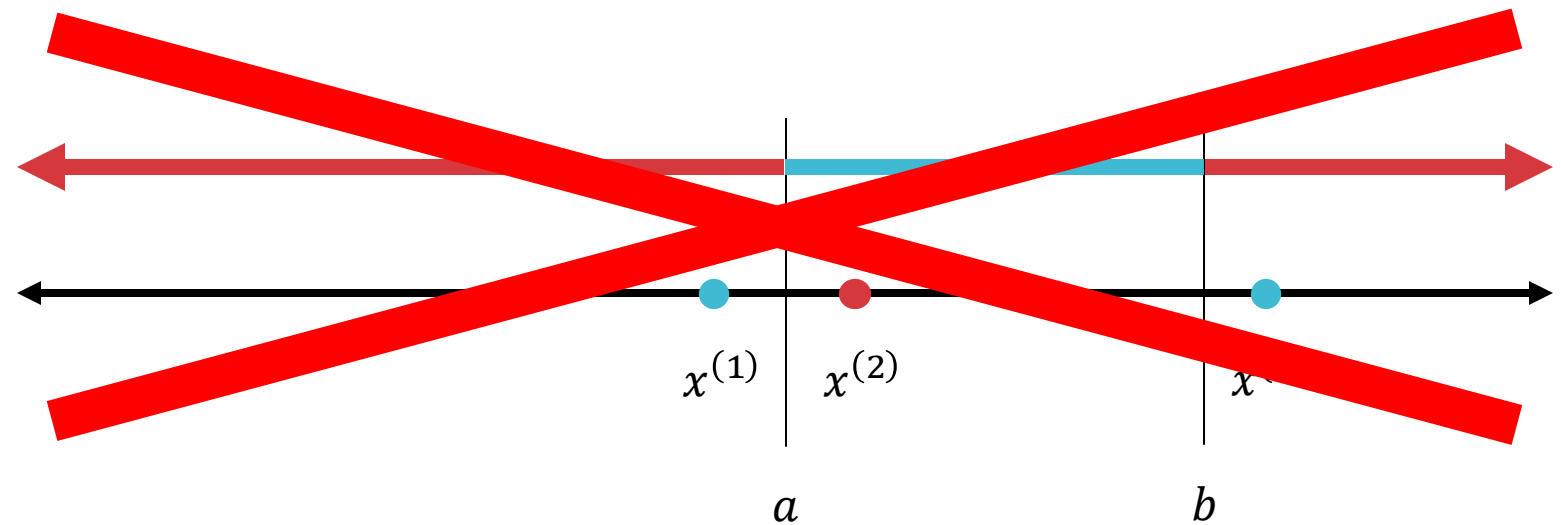
0%

3 and $\frac{1}{2}(m^2 + m + 1)$

0%

VC-Dimension: Example

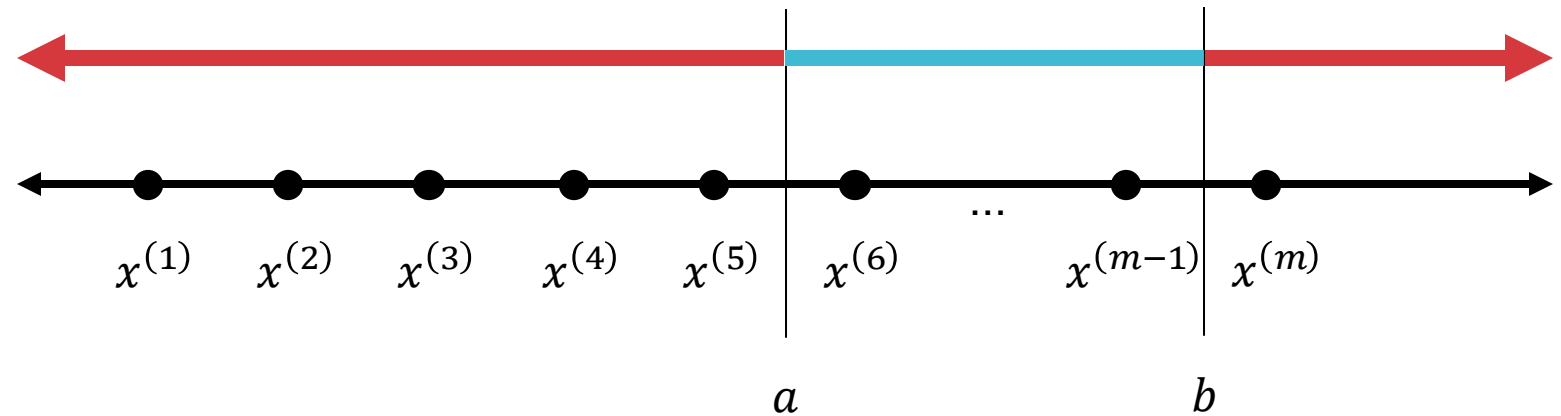
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive intervals



- What are $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(m)$?

VC-Dimension: Example

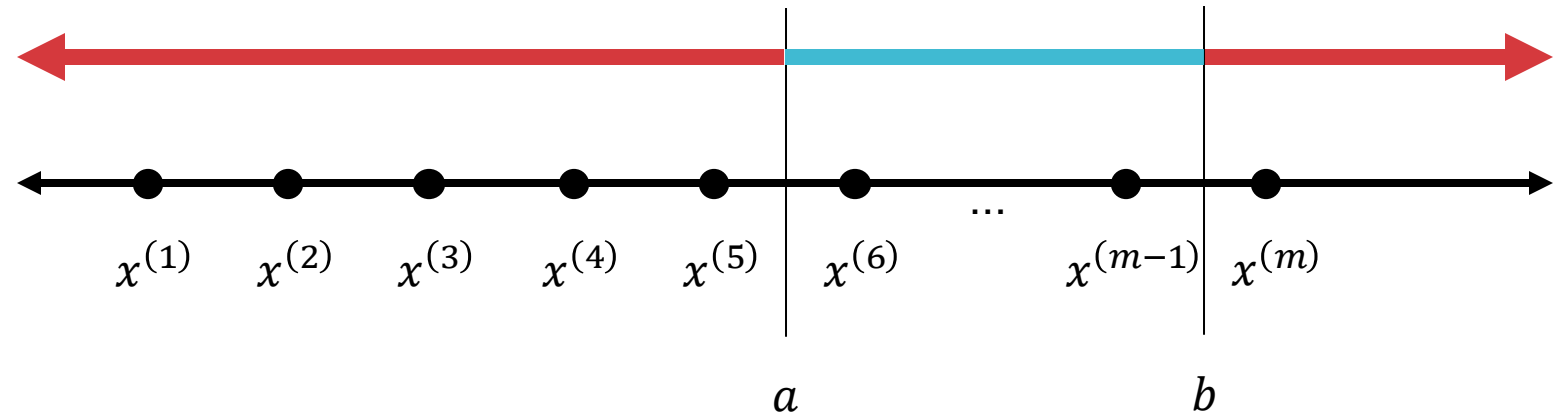
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive intervals



- What are $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(m)$?

VC-Dimension: Example

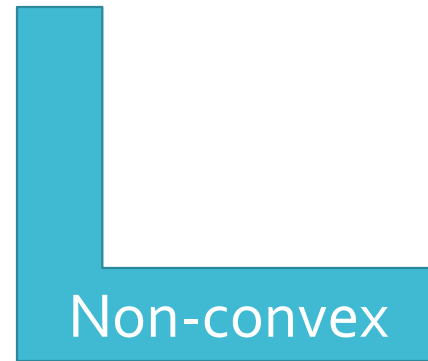
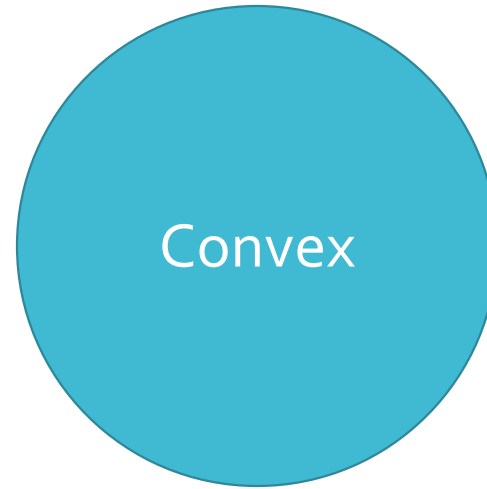
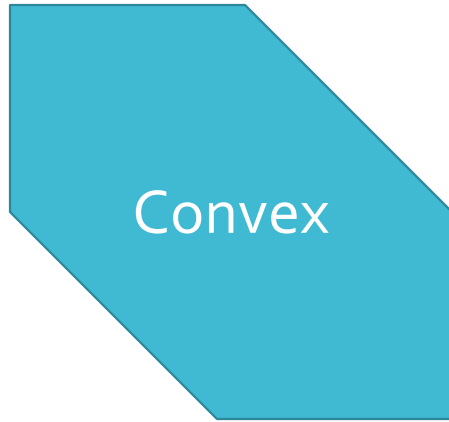
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive intervals



- $d_{VC}(\mathcal{H}) = 2$ and $g_{\mathcal{H}}(m) = \binom{m+1}{2} + 1 = O(m^2)$

Growth Function: Example

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



Growth Function: Example

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

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

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

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

- 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

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

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

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

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

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

Statistical Learning Theory Corollary

- 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

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

with probability at least $1 - \delta$.

Approximation Generalization Tradeoff

How well does
 h generalize?

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

How well does h
approximate c^* ?

Approximation Generalization Tradeoff

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

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

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

Key Takeaways

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