10-701: Introduction to Machine Learning

Lecture 20 – Learning Theory (Infinite Case)

Hoda Heidari

* Slides adopted from F24 offering of 10701 by Henry Chai.

Statistical Learning Theory Model

1. Data points are generated i.i.d. from some *unknown* distribution

$$\boldsymbol{x}^{(n)} \sim p^*(\boldsymbol{x})$$

2. Labels are generated from some *unknown* function

$$y^{(n)} = c^*\left(\boldsymbol{x}^{(n)}\right)$$

3. The learning algorithm chooses the hypothesis (or classifier) with lowest *training* error rate from a specified hypothesis set, $\mathcal{H}$

4. Goal: return a hypothesis (or classifier) with low *true* error rate

Types of Risk (a.k.a. Error)

- Expected *risk* of a hypothesis h (a.k.a. true error)

true error $R(h) = P_{\boldsymbol{x} \sim p^*}\left(c^*(\boldsymbol{x}) \neq h(\boldsymbol{x})\right)$

- Empirical risk of a hypothesis h (a.k.a. training error)

empirical error / training error

$$\hat{R}(h) = P_{\boldsymbol{x} \sim \boxed{\mathcal{D}}}\left(c^*(\boldsymbol{x}) \neq h(\boldsymbol{x})\right)$$

$$= \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\left(c^*\left(\boldsymbol{x}^{(n)}\right) \neq h\left(\boldsymbol{x}^{(n)}\right)\right)$$

$$= \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\left(y^{(n)} \neq h\left(\boldsymbol{x}^{(n)}\right)\right)$$

where $\boxed{\mathcal{D}} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N}$ is the training data set with $\boldsymbol{x}^{i}$ denoting a point sampled uniformly at random from p^*

Three Hypotheses of Interest

1. The *true function*, c^*

2. The *expected risk minimizer*,

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

3. The *empirical risk minimizer*,

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

Key Question

- Given a hypothesis with zero/low training error, what can we say about its true error?

Sample Complexity & PAC Learnability

PAC

$1-\delta$ $\quad$ ε

$\mathcal{H}$

- A hypothesis class is PAC-learnable if for every $\epsilon, \delta \in (0, 1)$, there exists a sample size $\underbrace{m(\epsilon, \delta)}$ polynomial in $1/\epsilon$ and $1/\delta$, such that with m i.i.d. samples from ANY distribution p^* the algorithm outputs a hypothesis whose generalization error is at most ϵ with probability at least $1 - \delta$.

Generalization error $= R(h) - \hat{R}(h)$

PAC-learnability Results

- Four cases
  - Realizable vs. Agnostic
    - Realizable → $c^* \in \mathcal{H}$
    - Agnostic → c^* might or might not be in $\mathcal{H}$
  - Finite vs. Infinite
    - Finite → $|\mathcal{H}| < \infty$ ← last lecture
    - Infinite → $|\mathcal{H}| = \infty$ ← today's lecture

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

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

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

- Making the bound tight (setting the two sides equal to each other) and solving for ϵ gives...

Proof Steps

$\in \mathcal{H}$

1. Consider a hypothesis h with $R(h) > \epsilon$. Show that the probability of $\hat{R}(h) = 0$ is bounded.
2. Suppose there are k hypothesis $h \in \mathcal{H}$ with $R(h) > \epsilon$. Use $\boxed{\text{union bound}}$ to upper bound the likelihood of at least one of them having $\hat{R}(h) = 0$.

$$\left\{ \begin{array}{l} P(\hat{R}(h_1) = 0 \vee \hat{R}(h_2) = 0 \vee \cdots \vee \hat{R}(h_k) = 0) \leq \\ P(\hat{R}(h_1) = 0) + P(\hat{R}(h_2) = 0) + \cdots + P(\hat{R}(h_k) = 0) \quad k(1-\epsilon)^M \end{array} \right.$$

3. Upper bound k with $|\mathcal{H}|$.
4. Set the above upper bound to be less than or equal δ to obtain the statement of the theorem.

Statistical Learning Theory Corollary: Finite, Realizable Case

- 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

$\hat{R}(h) = 0$

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

with probability at least $1 - \delta$.

Statistical Learning Theory Corollary: Finite, Agnostic Case

- 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

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

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

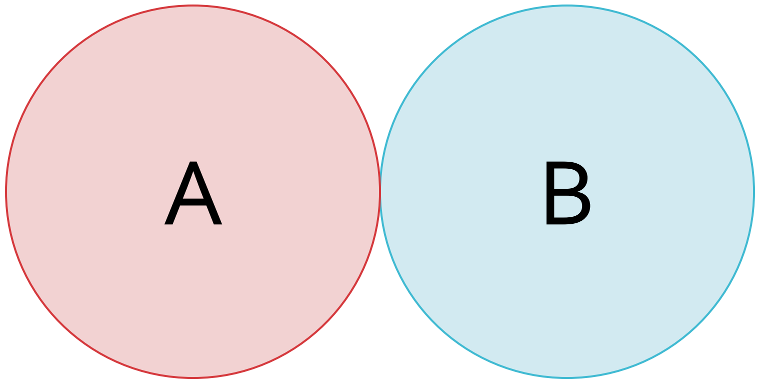
$$R(h) \leq \hat{R}(h) + \underbrace{\sqrt{\frac{1}{2M}\left(\ln(|\mathcal{H}|) + \ln\left(\frac{2}{\delta}\right)\right)}}_{\infty}$$

with probability at least $1 - \delta$.

The Union Bound is not tight!

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

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



Events of interest in our proof

- $\hat{R}(h_1) = 0$
- $\hat{R}(h_2) = 0$

$\vdots$

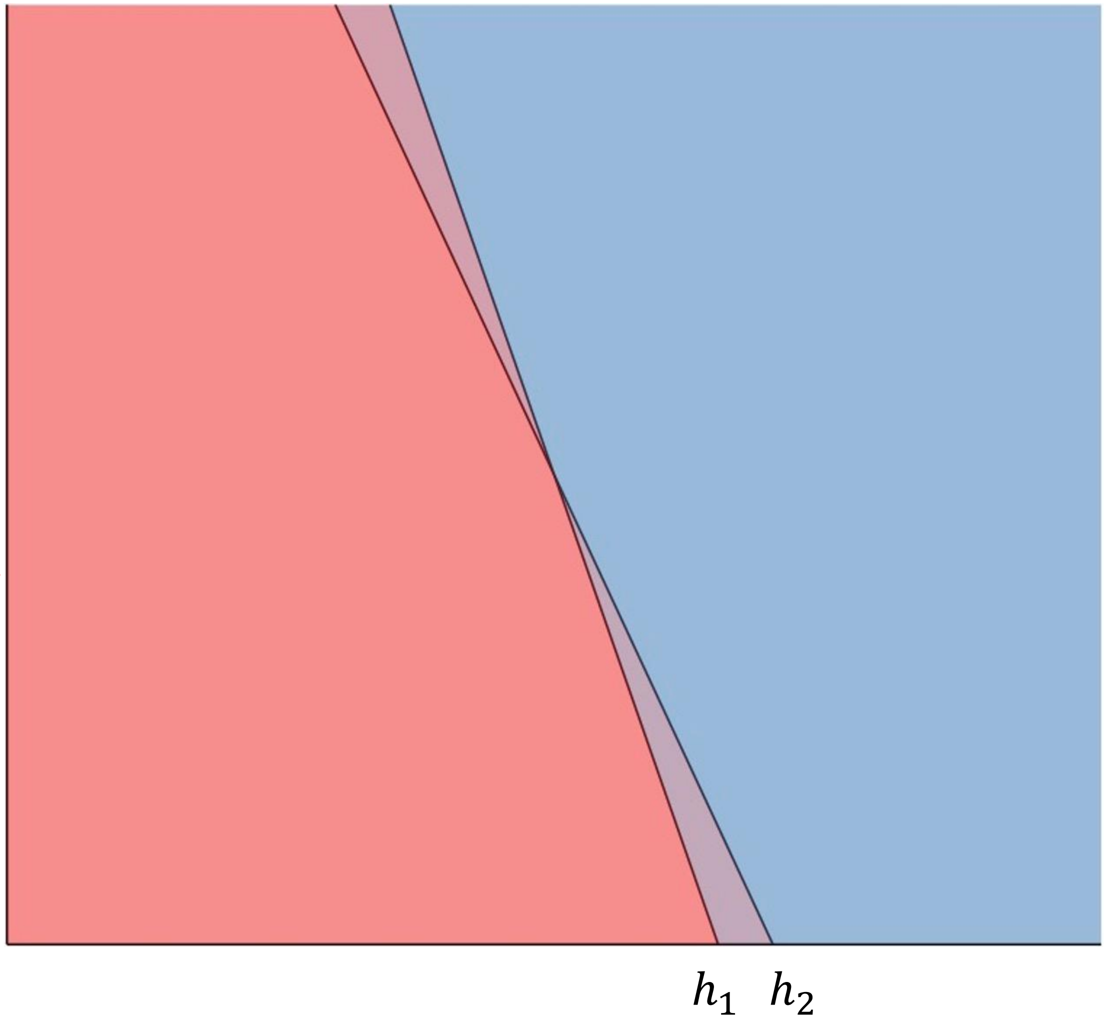
- $\hat{R}(h_K) = 0$

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" $\hat{R}(h_1) = 0$
- "h_2 is consistent with the first m training data points" $\hat{R}(h_2) = 0$

will overlap a lot!

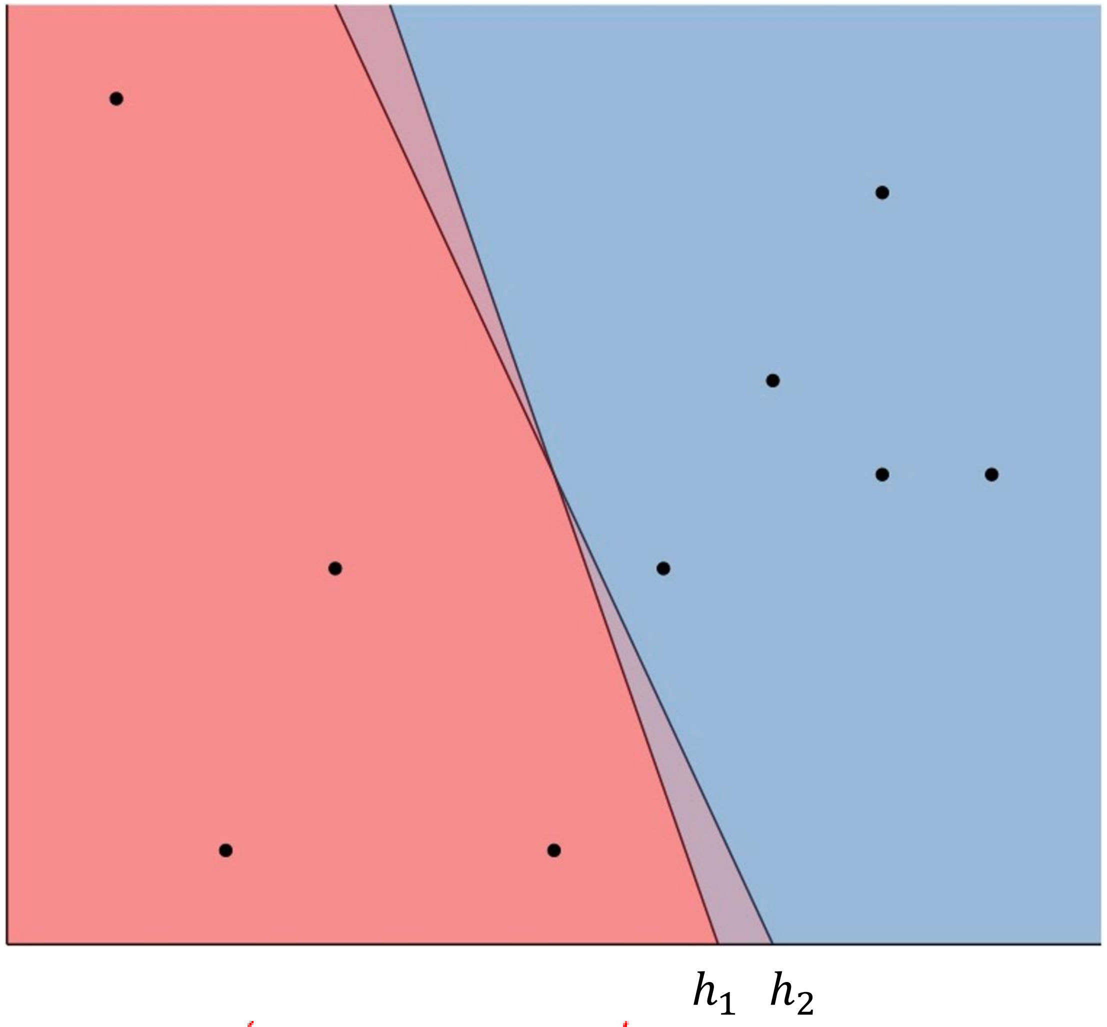


Intuition

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!



N samples $\rightarrow$ $\leq 2^N$ labelings

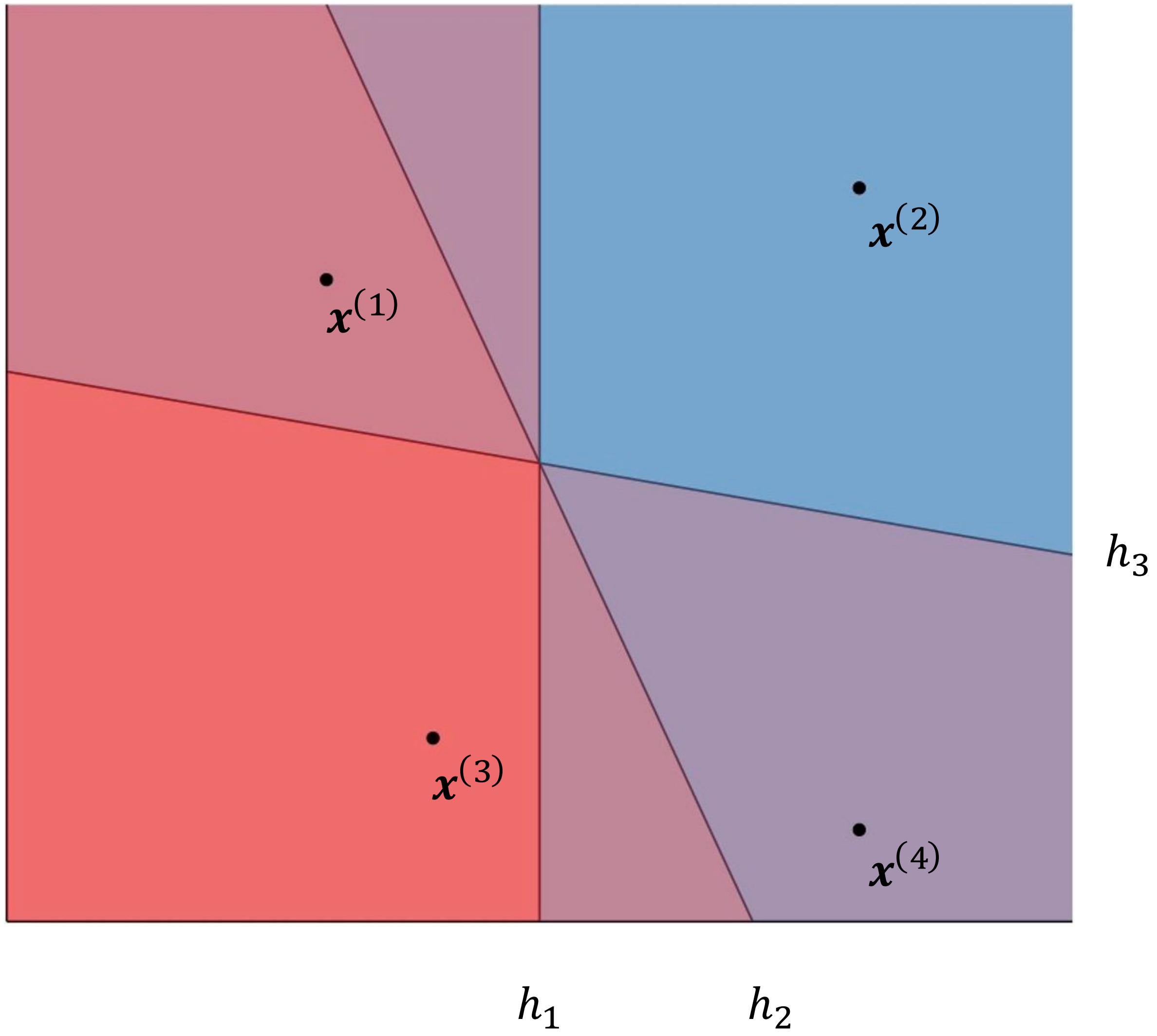
Labellings

- 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
- Insight: 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

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

Example: Labellings

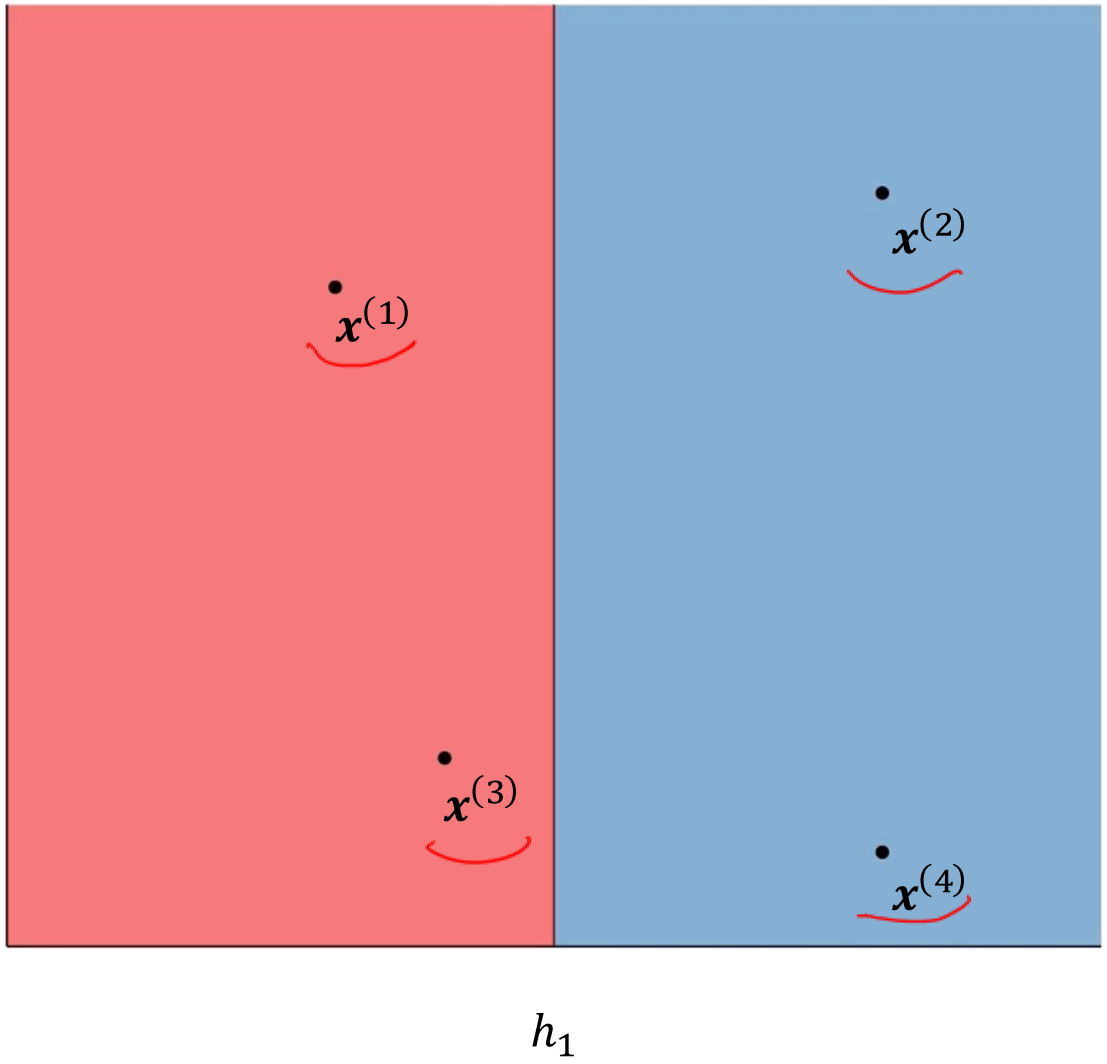
$\mathcal{H} = \{h_1, h_2, h_3\}$



Example: Labellings

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

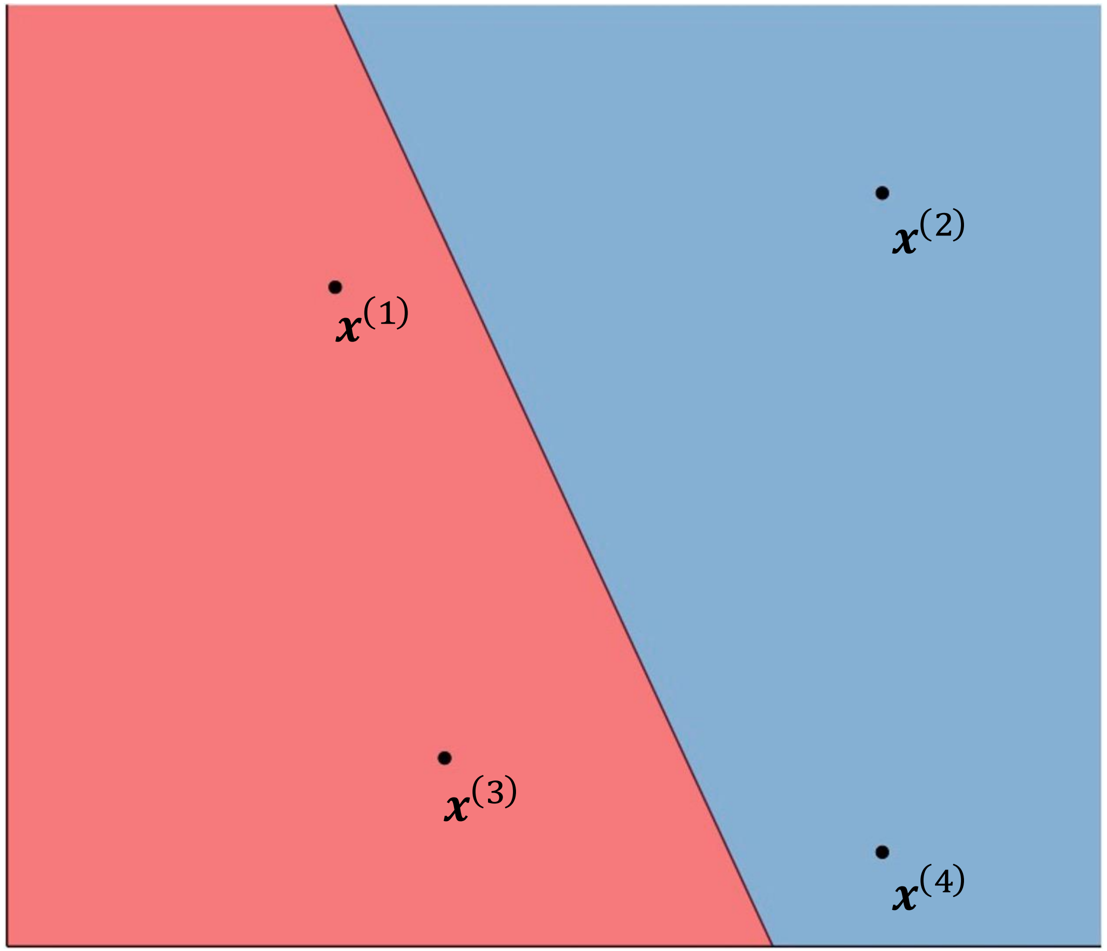
$$\left(h_1(\boldsymbol{x}^{(1)}), h_1(\boldsymbol{x}^{(2)}), h_1(\boldsymbol{x}^{(3)}), h_1(\boldsymbol{x}^{(4)})\right)$$
$$= \underbrace{(-1, +1, -1, +1)}$$



Example: Labellings

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

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

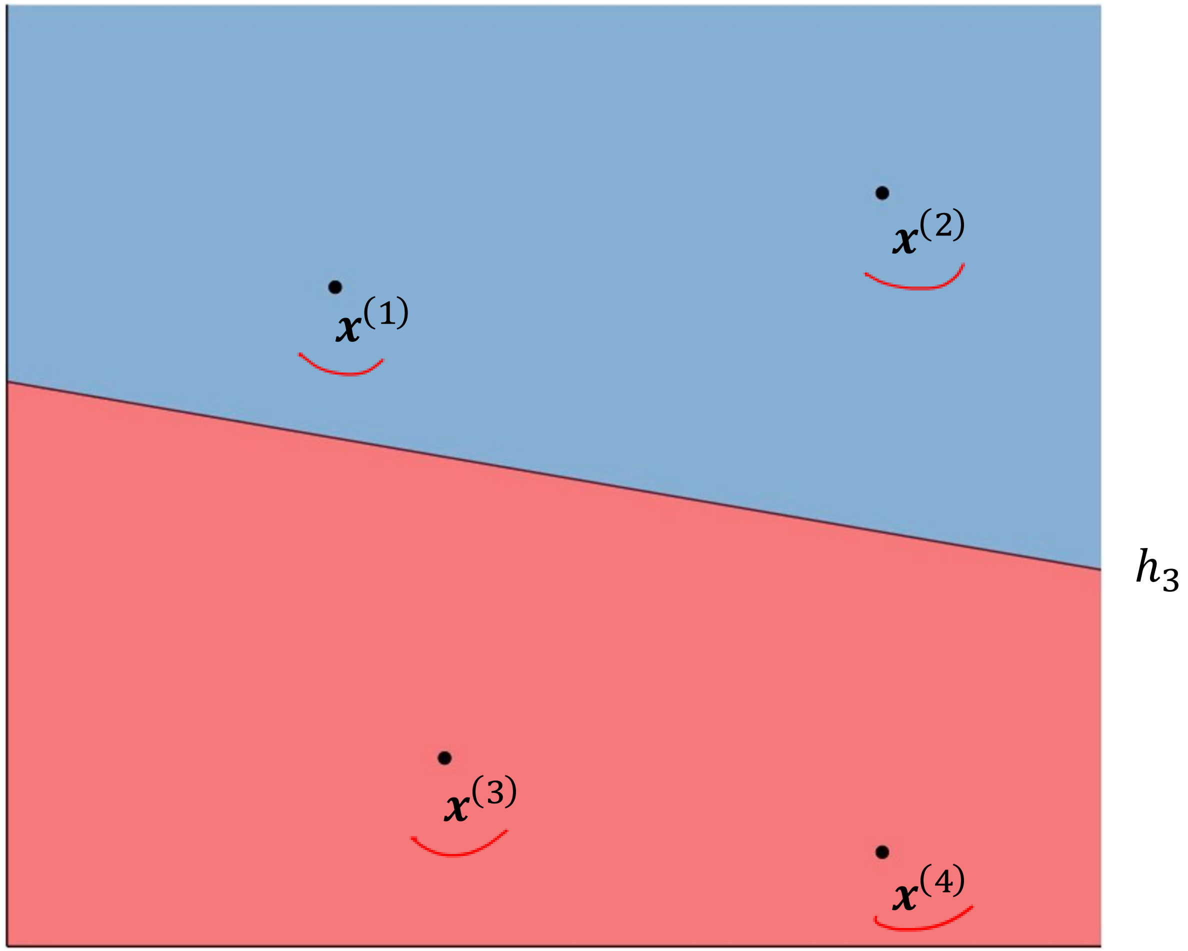


h_2

Example: Labellings

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

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

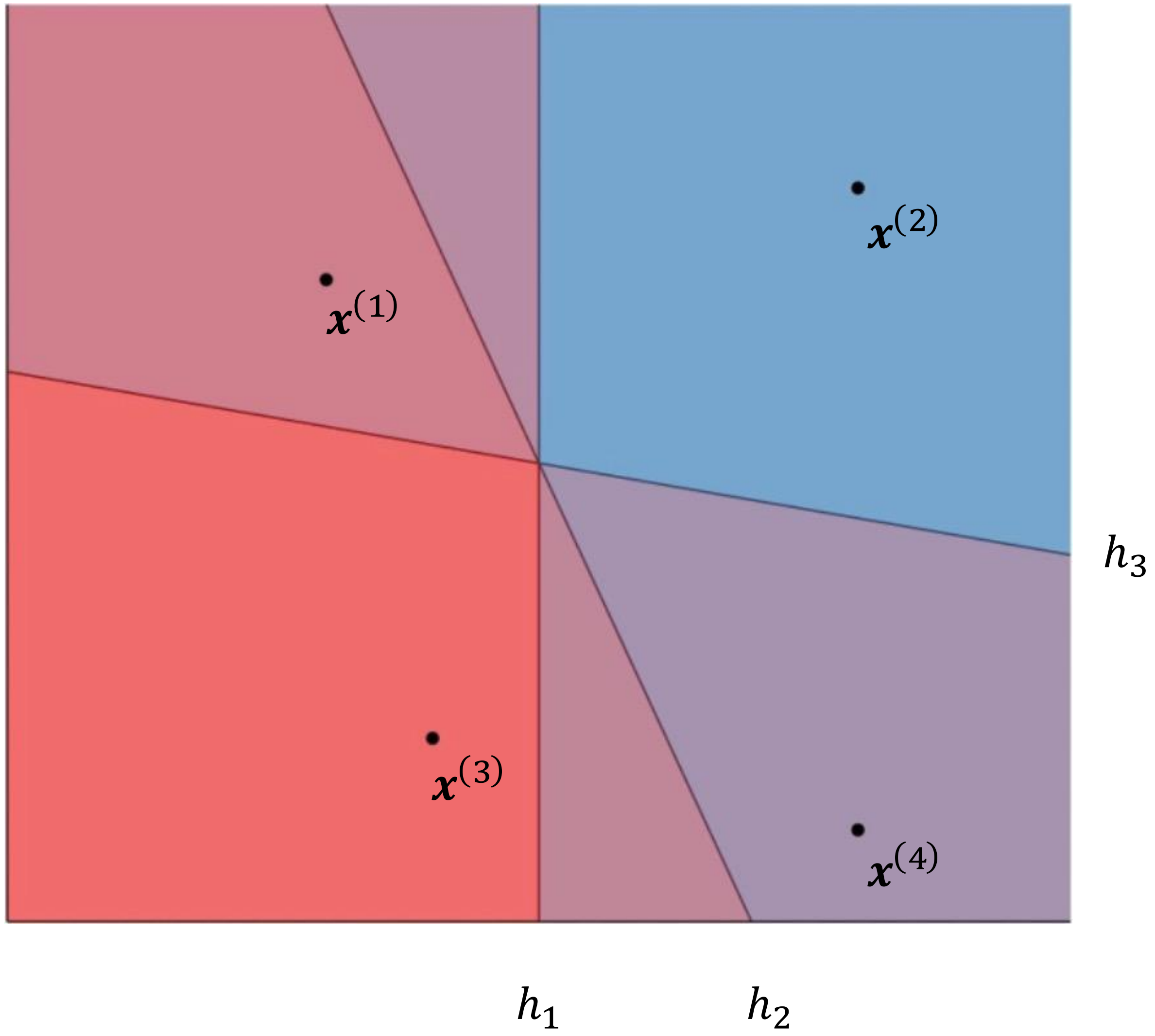


Example: Labellings

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

$L_{\mathcal{H}}(S)$
$= \{\underbrace{(+1, +1, -1, -1)}_{h_1,\, h_2}, \underbrace{(-1, +1, -1, +1)}_{h_3}\}$

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

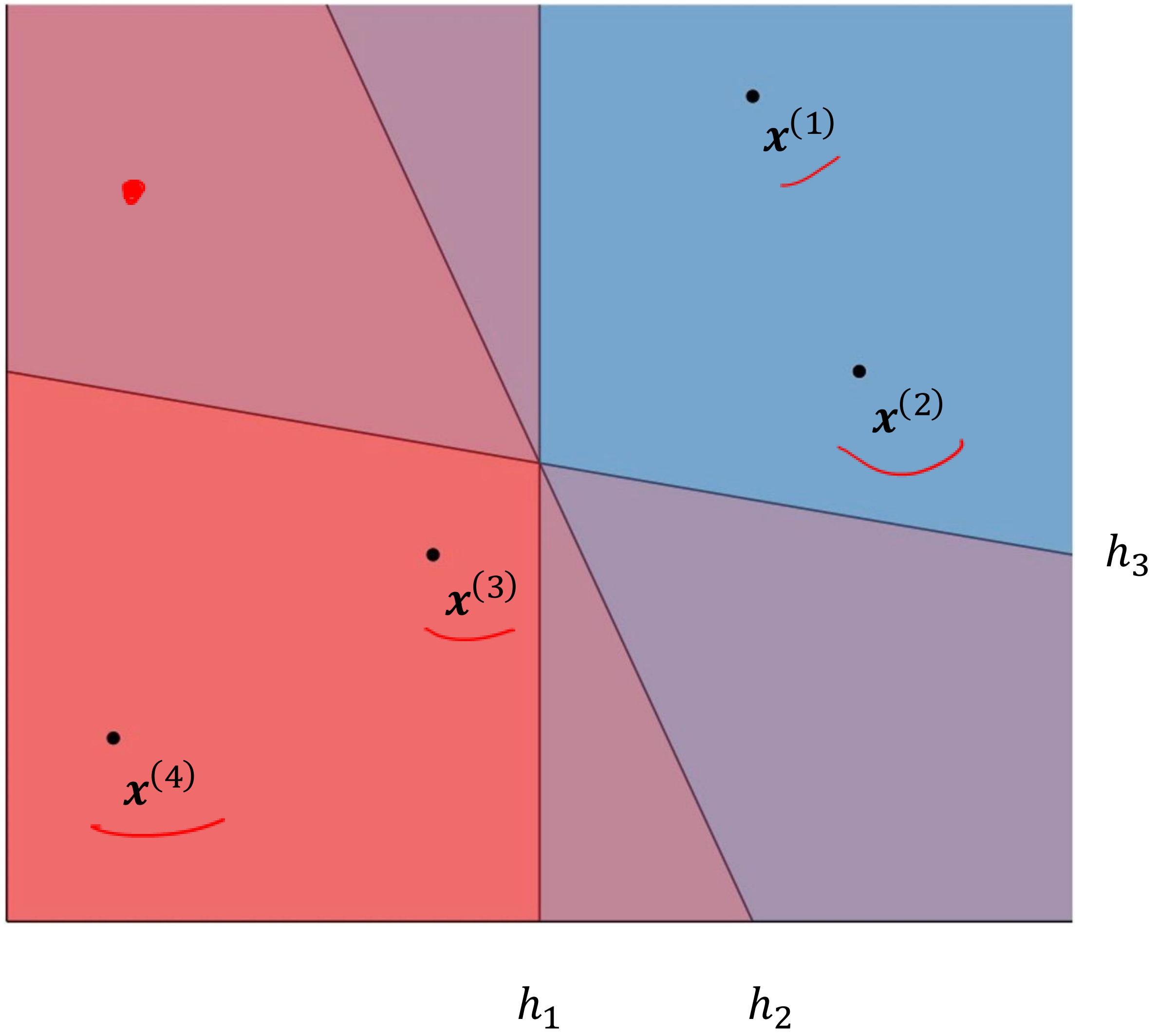


Example: Labellings

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

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

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



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:

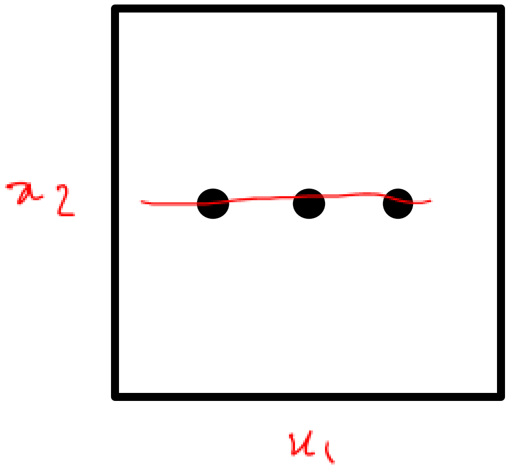
$$g_{\mathcal{H}}(M) = \boxed{\max_{S\,:\,|S|=M}} |L_{\mathcal{H}}(S)|$$

- $g_{\mathcal{H}}(M) \leq 2^M \; \forall \, \mathcal{H}$ and M (assuming we're in binary classification)
- $\mathcal{H}$ **shatters** S if $|L_{\mathcal{H}}(S)| = 2^M$
- If $\exists \, S$ s.t. $|S| = M$ and $\mathcal{H}$ shatters S, then $g_{\mathcal{H}}(M) = 2^M$

Growth Function: Example

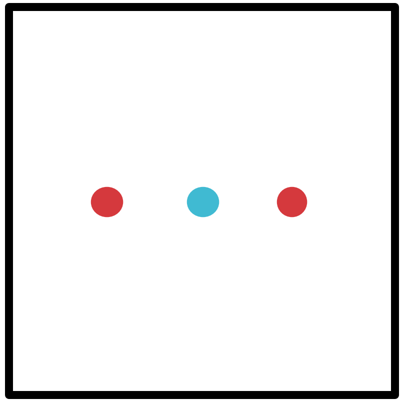
- $\boldsymbol{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators

- What is $g_{\mathcal{H}}(\overset{m}{3})$? $2^3 = 8$



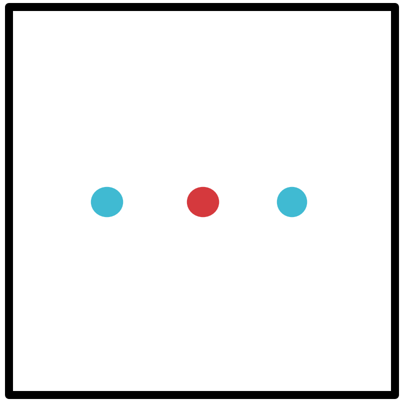
Growth Function: Example

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



Growth Function: Example

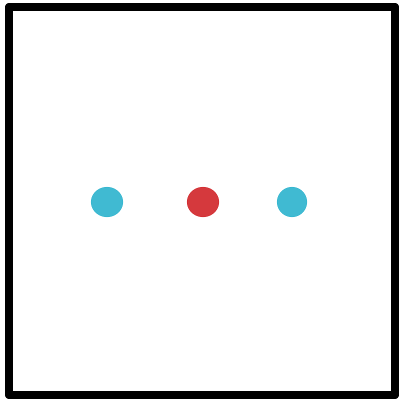
- $\boldsymbol{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(3)$?



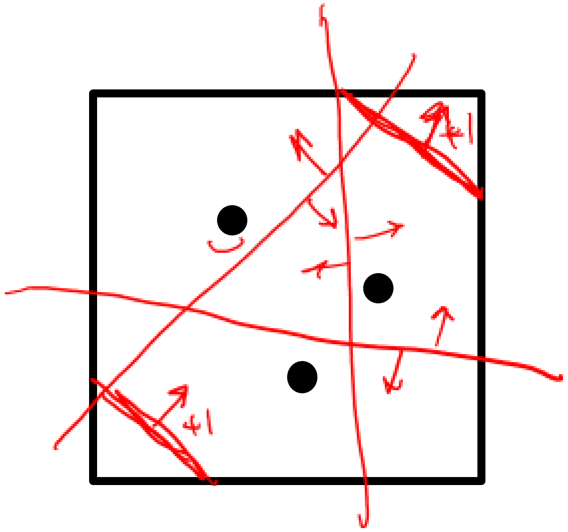
Growth Function: Example

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

- What is $g_{\mathcal{H}}(3)$? $= 2^3 = 8$



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

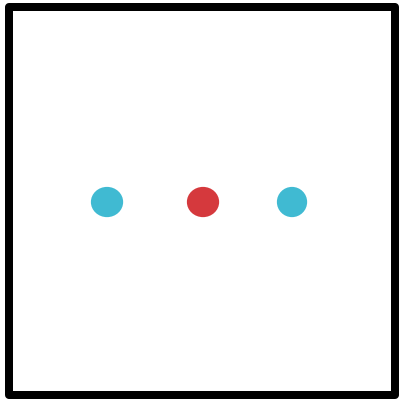


$|\mathcal{H}(S_2)| = 8$

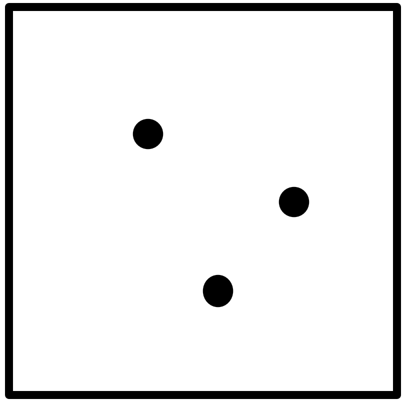
Growth Function: Example

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

- $g_{\mathcal{H}}(3) = 8 = 2^3$



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

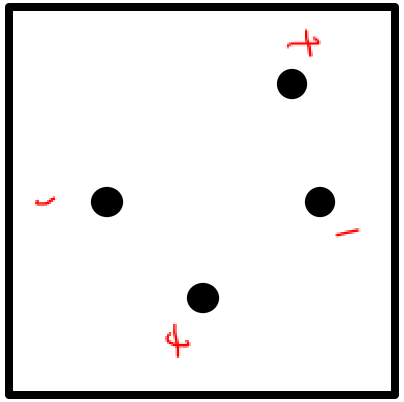


$|\mathcal{H}(S_2)| = 8$

Growth Function: Example

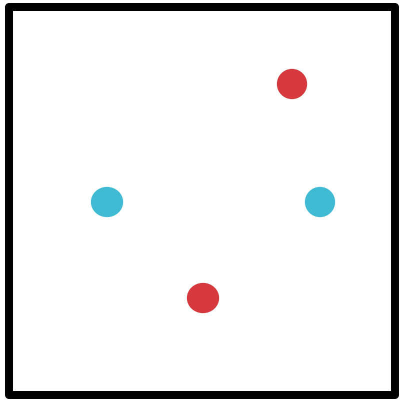
- $\boldsymbol{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators

- What is $g_{\mathcal{H}}(4)$? $\leq 2^4 = 16$



Growth Function: Example

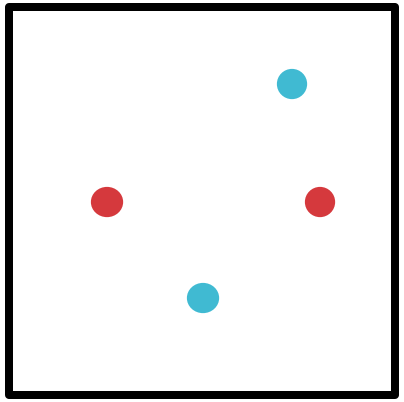
- $\boldsymbol{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(4)$?



Growth Function: Example

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

- What is $g_{\mathcal{H}}(4)$?

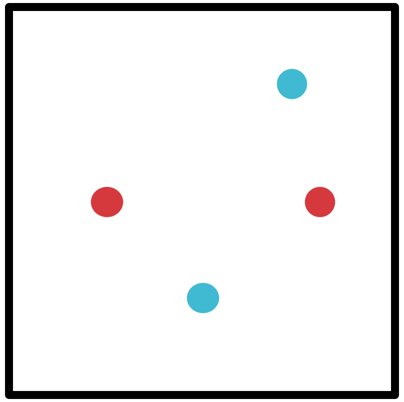


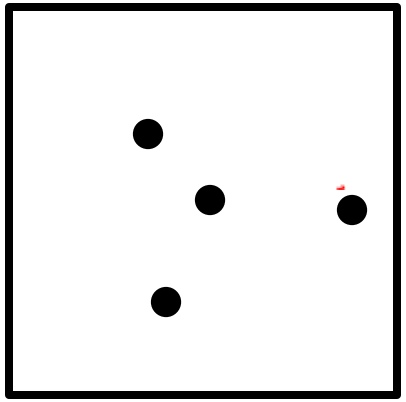
$|\mathcal{H}(S_1)| = 14$

Growth Function: Example

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

- What is $g_{\mathcal{H}}(4)$?



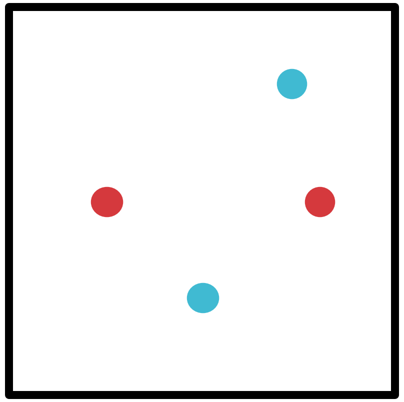


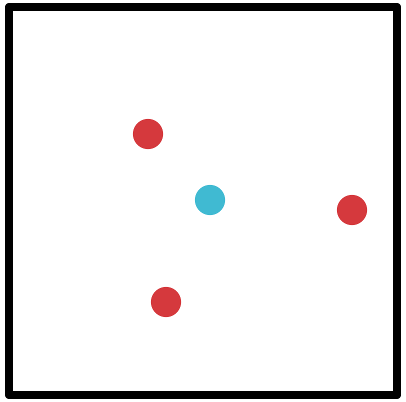
$$|\mathcal{H}(S_1)| = 14$$

Growth Function: Example

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

- What is $g_{\mathcal{H}}(4)$?



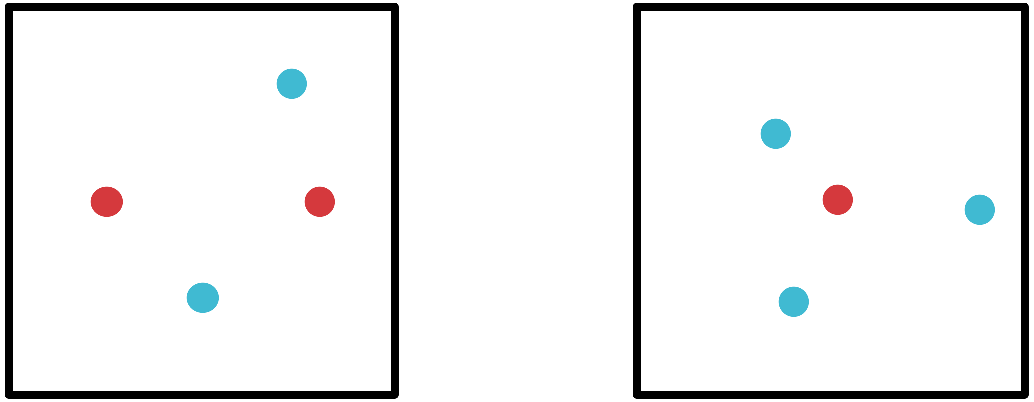


$|\mathcal{H}(S_1)| = 14$

Growth Function: Example

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

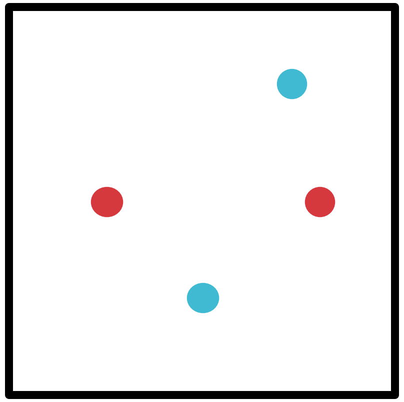
- What is $g_{\mathcal{H}}(4)$?



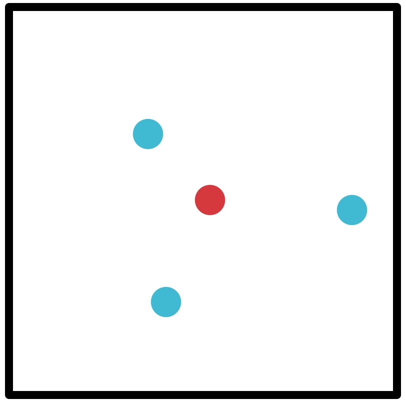
$|\mathcal{H}(S_1)| = 14$

Growth Function: Example

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



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



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

Growth Function: Example

$g_H(3) = 2^3$

$g_H(4) = 14 < 2^4$

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

- What is $\underbrace{g_{\mathcal{H}}(5)}$?

$g_H(5) \stackrel{?}{=} 2^5$

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

$|\mathcal{H}|$

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

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

M appears on both sides of the inequality…

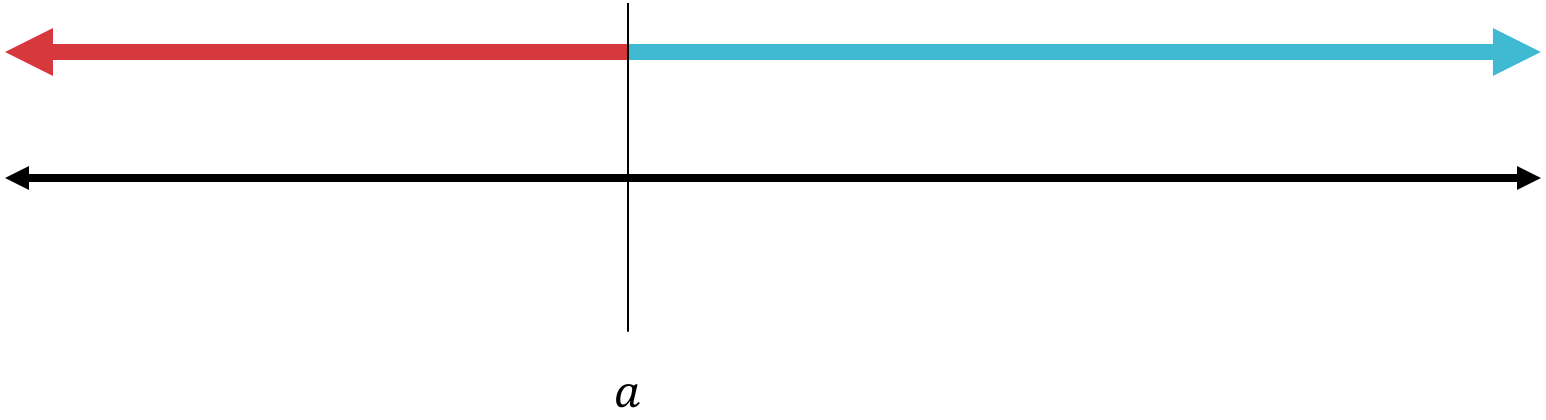
Theorem 3: Vapnik-Chervonenkis (VC)-Dimension

- $d_{VC}(\mathcal{H})$ = 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 $d_{VC}(\mathcal{H}) = \infty$
  - $g_{\mathcal{H}}(M) = O\left(M^{d_{VC}(\mathcal{H})}\right)$ (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

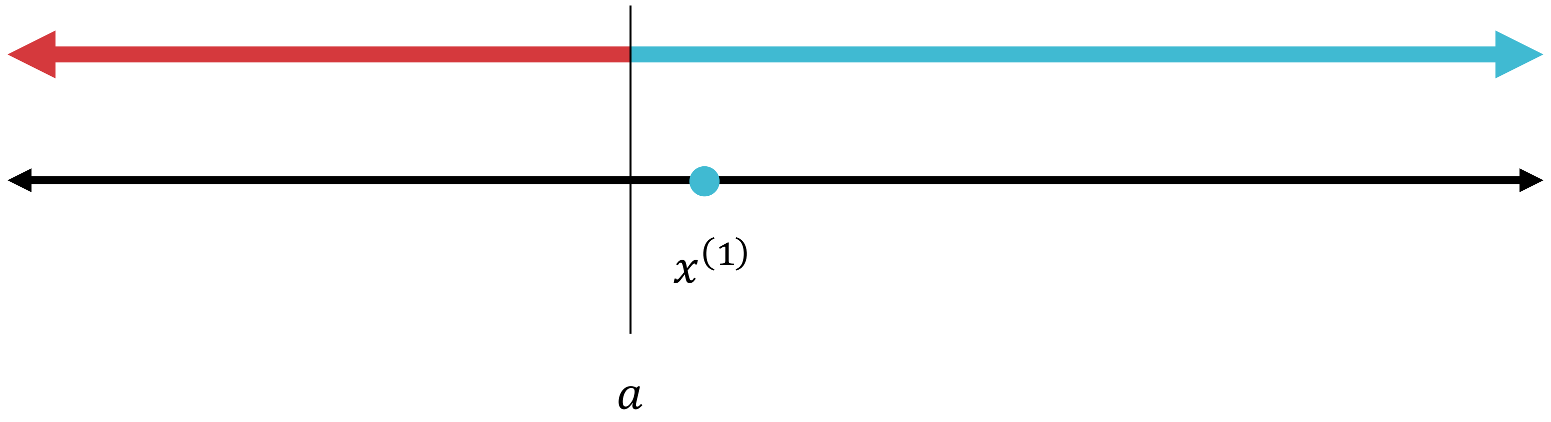
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

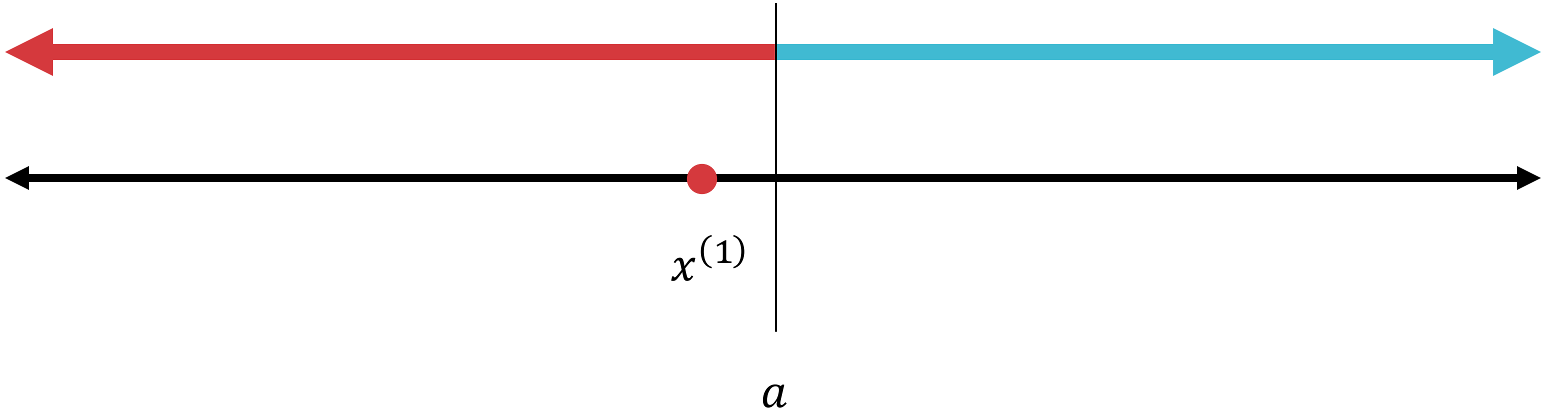
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

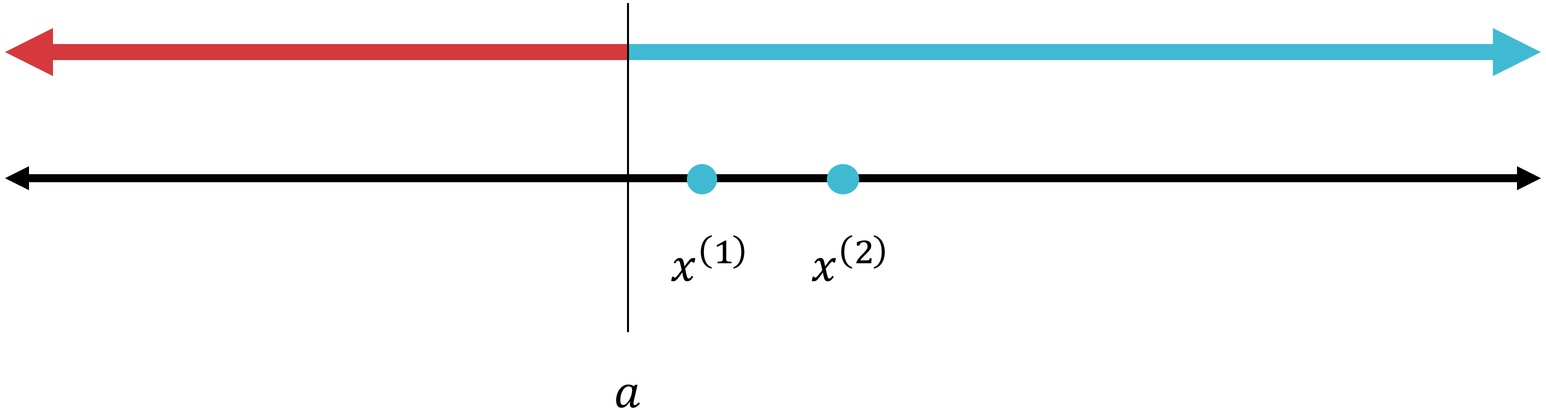
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

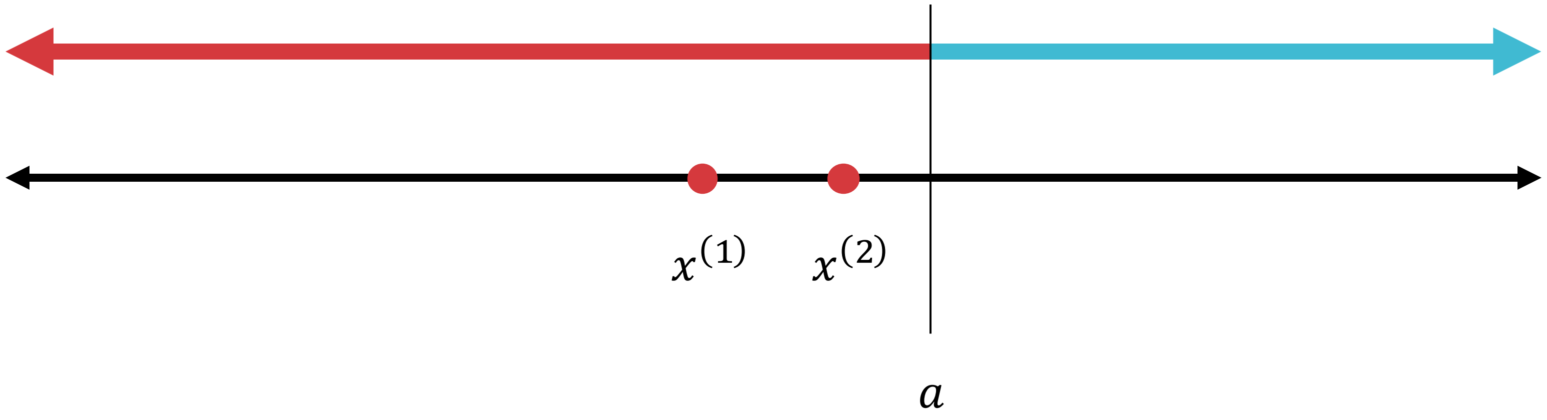
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

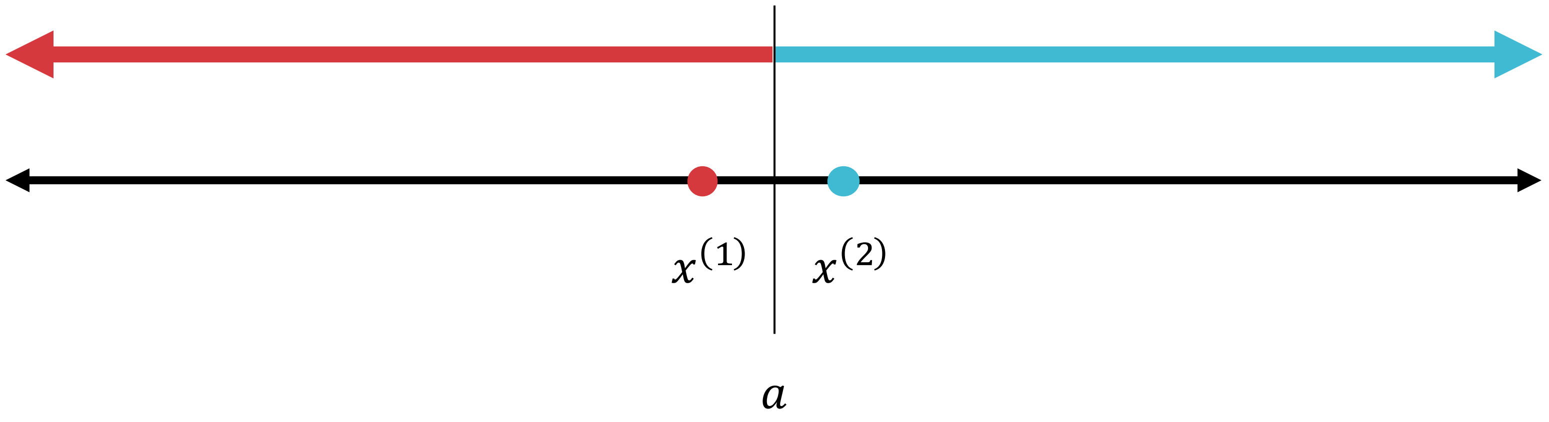
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

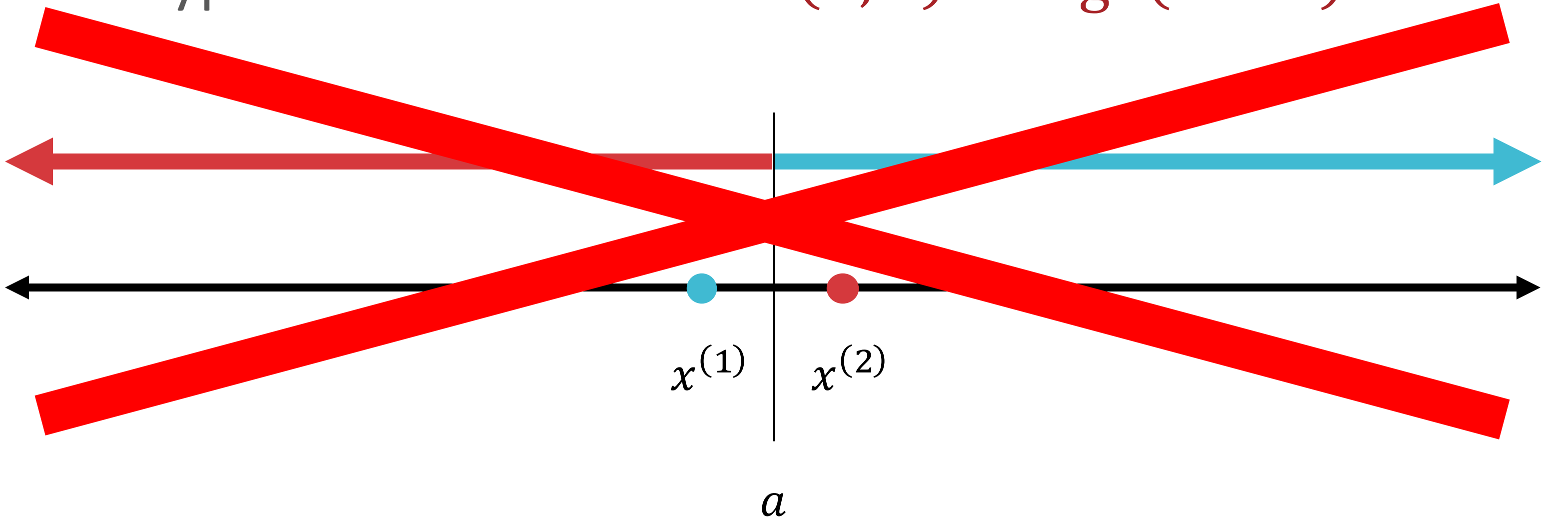
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

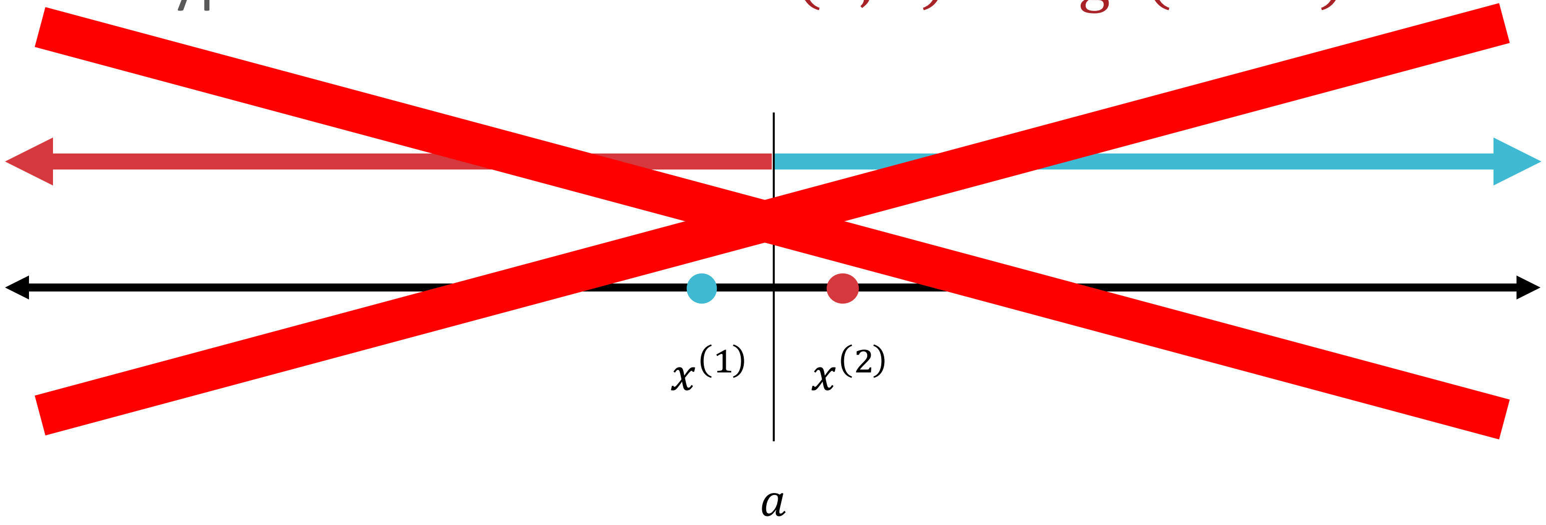
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

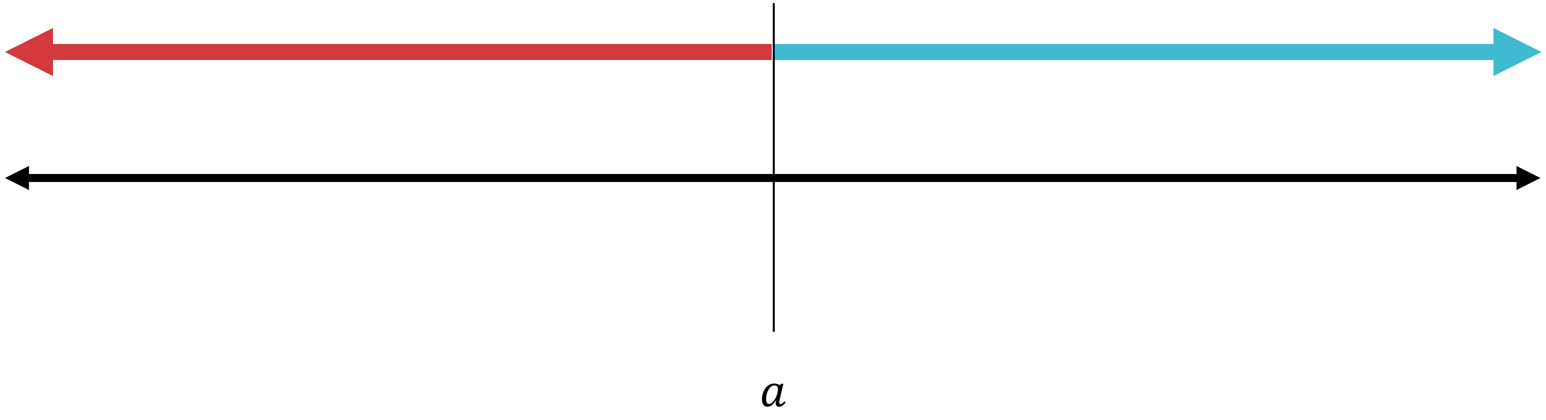
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

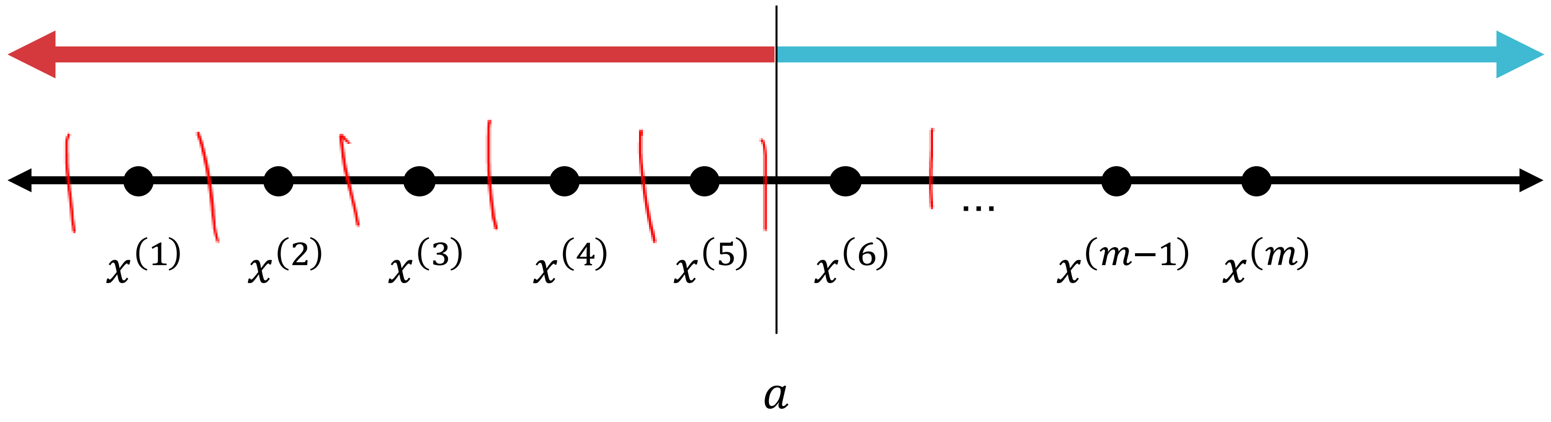
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

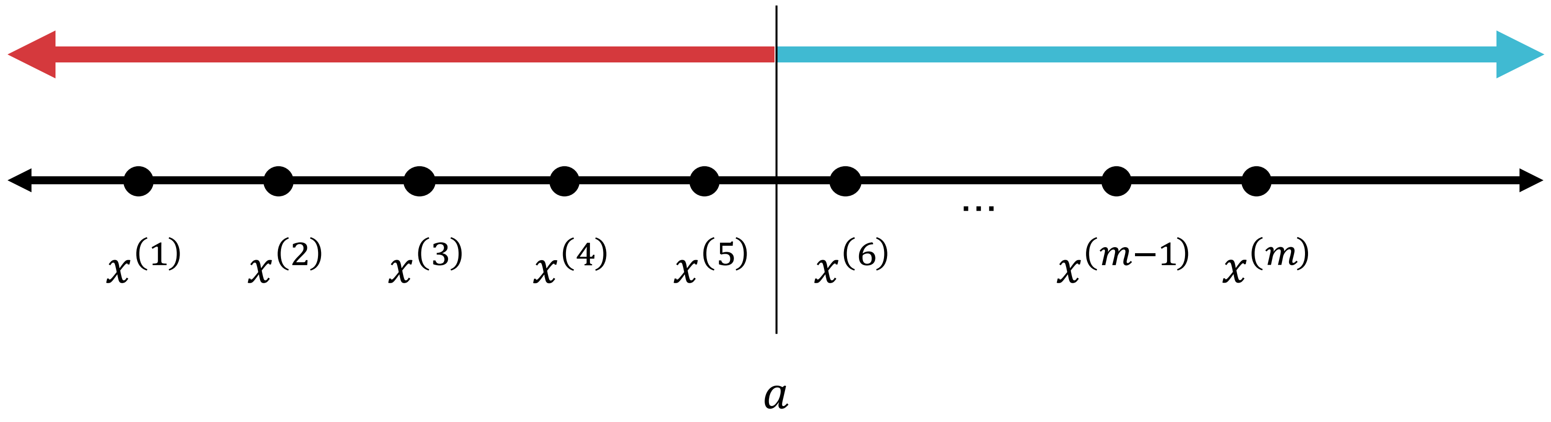
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

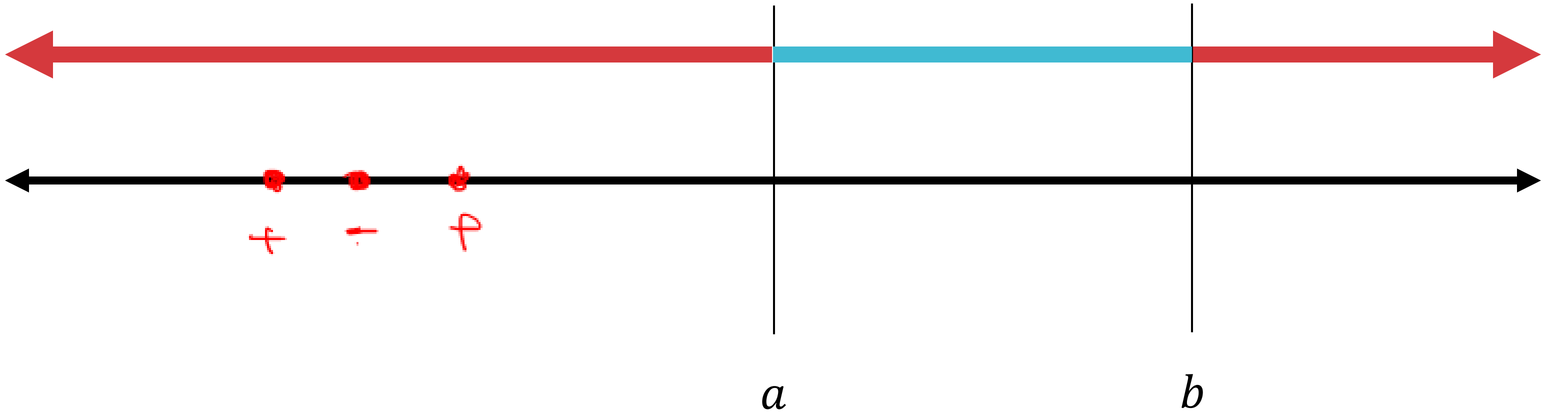
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



- $g_{\mathcal{H}}(m) = m + 1 = O(m^1)$

VC-Dimension: In-class Poll

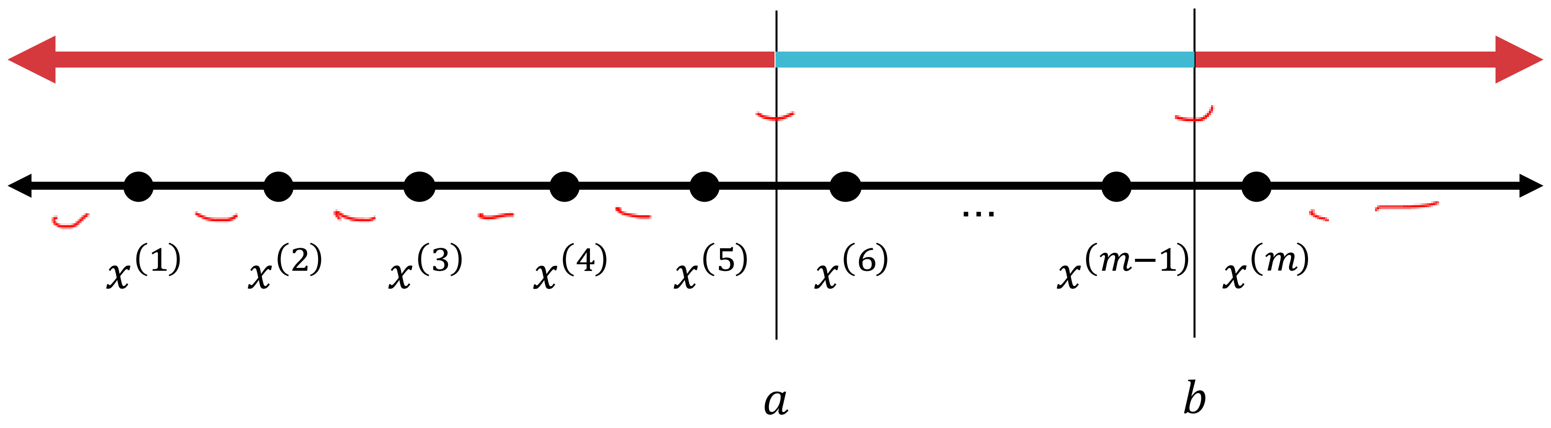
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive intervals



- What are $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(m)$?
  - 1 and m+1
  - 2 and m+1
  - 2 and 1/2(m^2 + m + 4)
  - 2 and 1/2(m^2 + m + 2)

VC-Dimension: Example

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

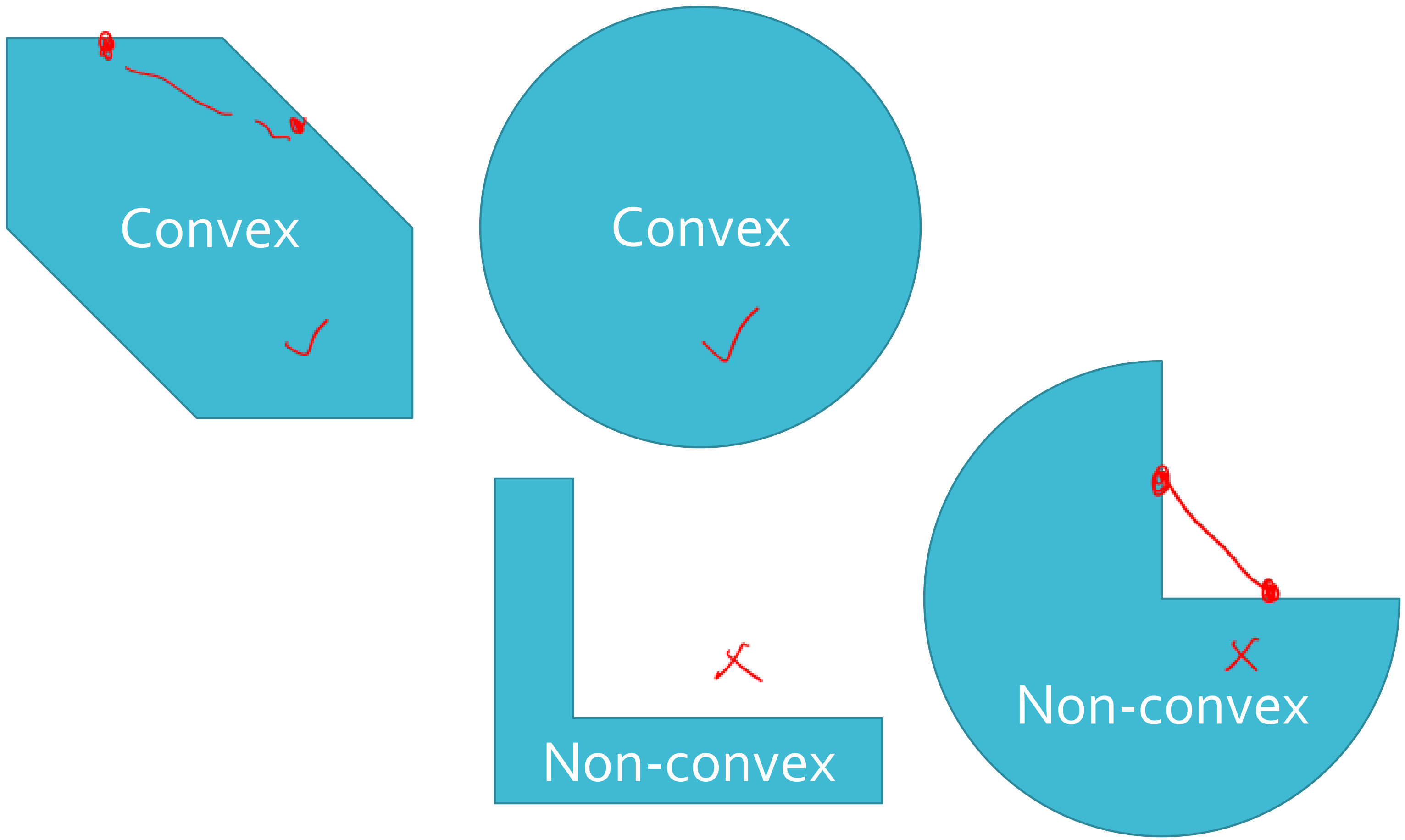


- What are $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(m)$?

$$\binom{m+1}{2} + 2 = \frac{(m+1)m}{2} + 2 = \frac{m^2+m+4}{2}$$

Growth Function: Example

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



Growth Function: Example

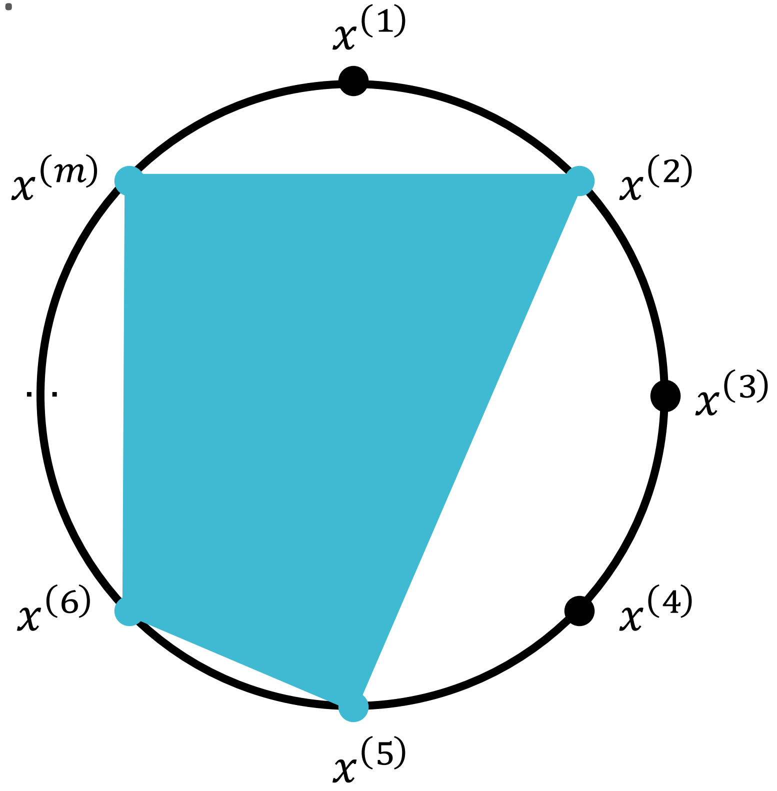
- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional positive convex sets

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

Growth Function: Example

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

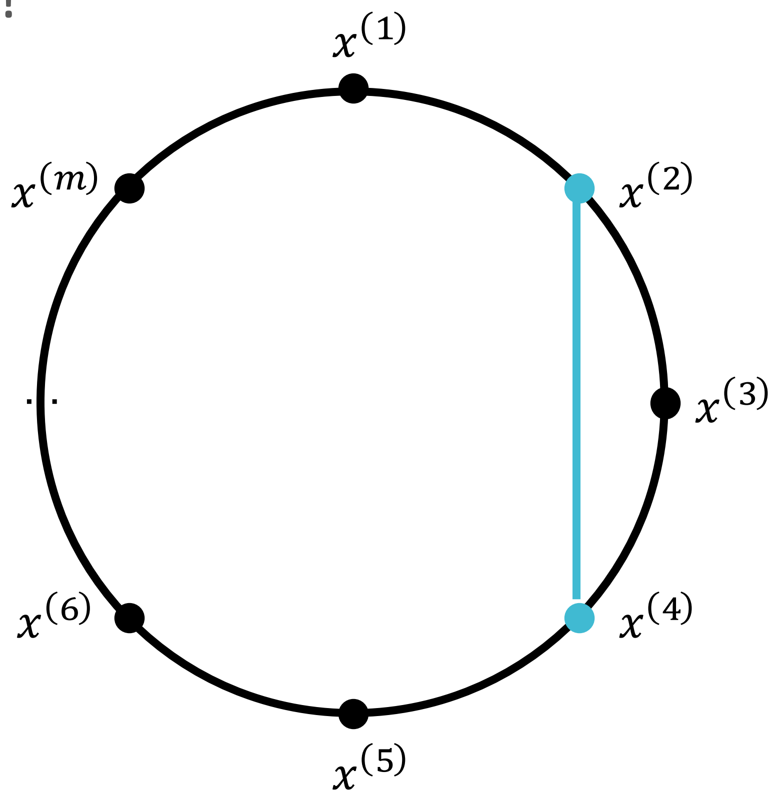
- What are $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(M)$?



Growth Function: Example

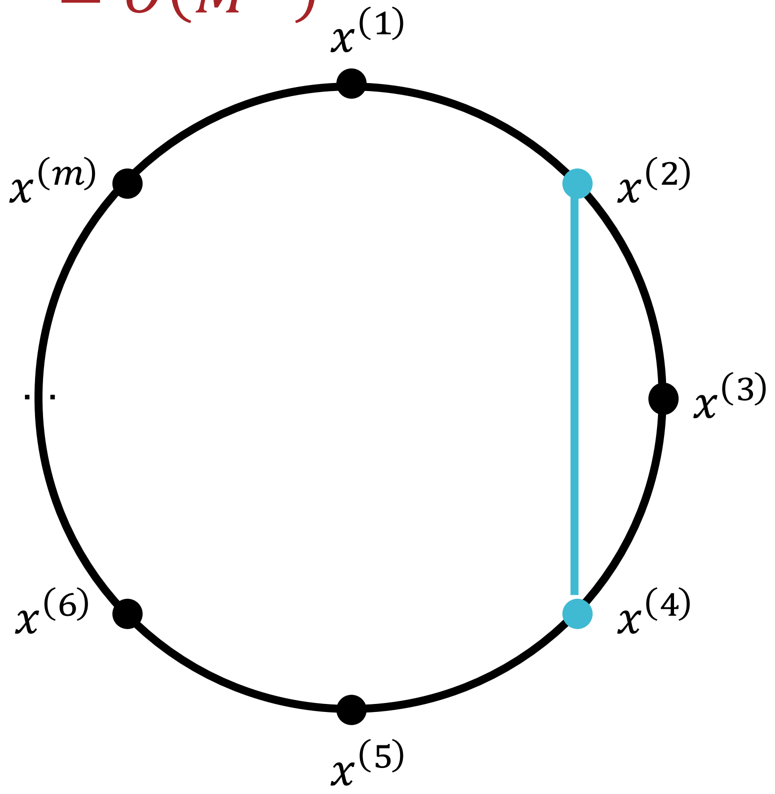
- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional positive convex sets

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



Growth Function: Example

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



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

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

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

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

Approximation Generalization Tradeoff

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

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 $d_{VC}(\mathcal{H})$