10-301/601: Introduction to Machine Learning Lecture 17 – Learning Theory (Infinite Case)

Henry Chai

7/5/23

Front Matter

- Announcements
  - PA4 released 6/15, due 7/13 at 11:59 PM
    - You still have one week from this Thursday!
  - Quiz 6: Deep Learning & Learning Theory on 7/11
- Recommended Readings
  - Mitchell, Chapter 7.4

Recall: Theorem 1: Finite, Realizable Case

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

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

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

- Solving for ϵ gives...

Statistical Learning Theory Corollary

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

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

with probability at least $1 - \delta$.

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

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

- Bound is inversely quadratic in ϵ, e.g., halving ϵ means we need four times as many labelled training data points

- Solving for ϵ gives…

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

as M increases, the stuff under the $\sqrt{}$ decreases

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

with probability at least $1 - \delta$.

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

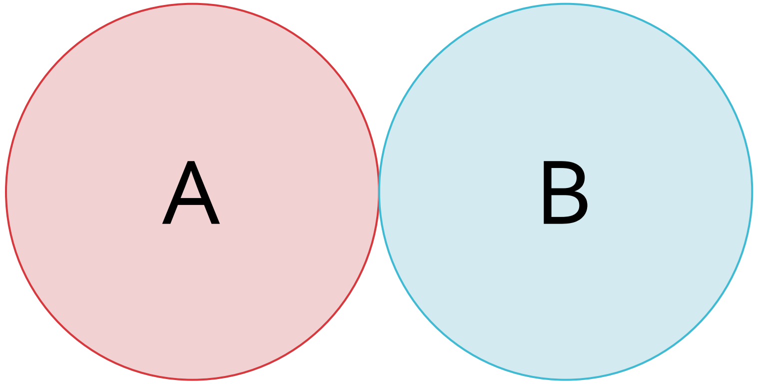
- For a finite hypothesis set $\mathcal{H}$ and arbitrary distribution p^*, given a training data set S s.t. $|S| = M$, all $h \in \mathcal{H}$ have

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

with probability at least $1 - \delta$.

The Union Bound...

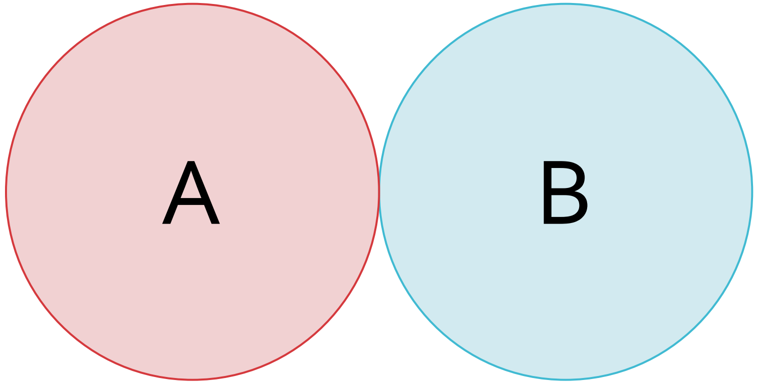
$$P\{A \cup B\} \leq P\{A\} + P\{B\}$$



The Union Bound is Bad!

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

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

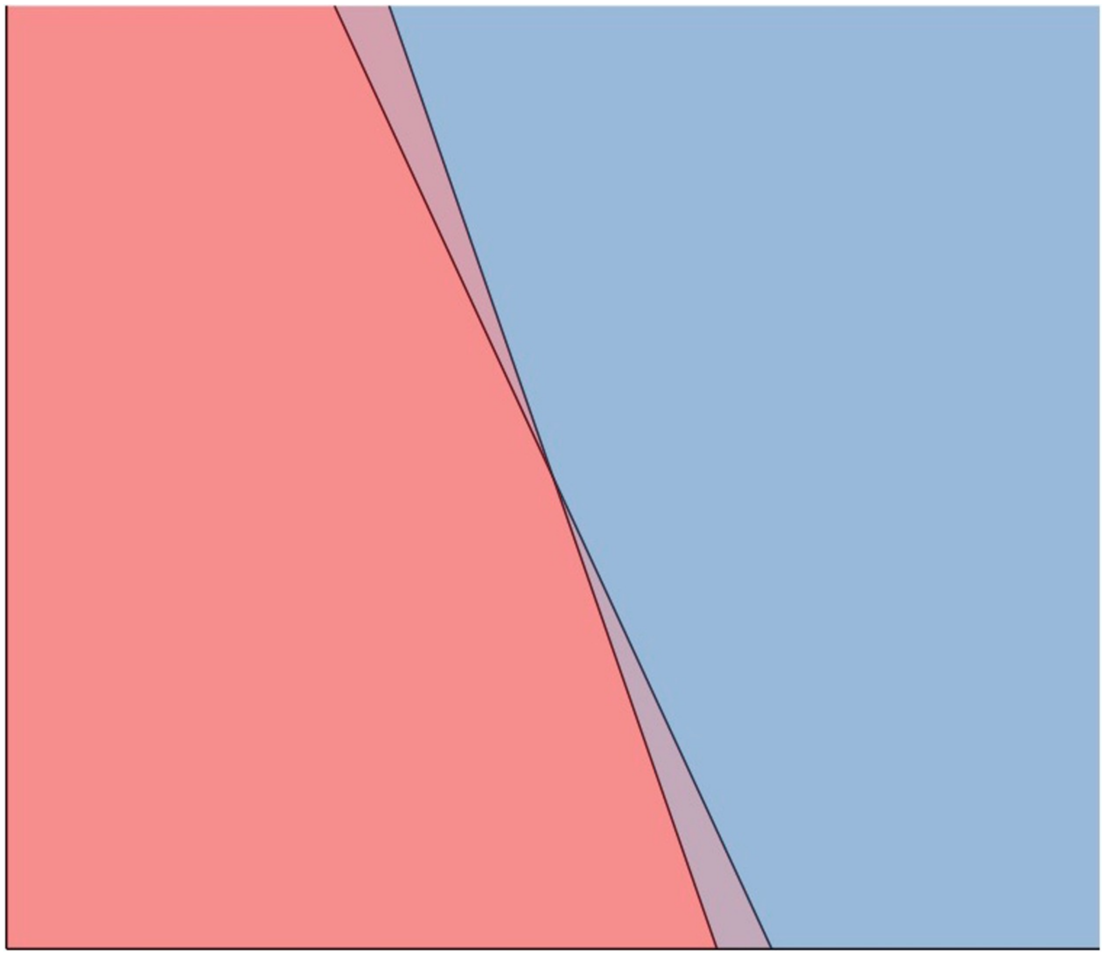


Intuition

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!

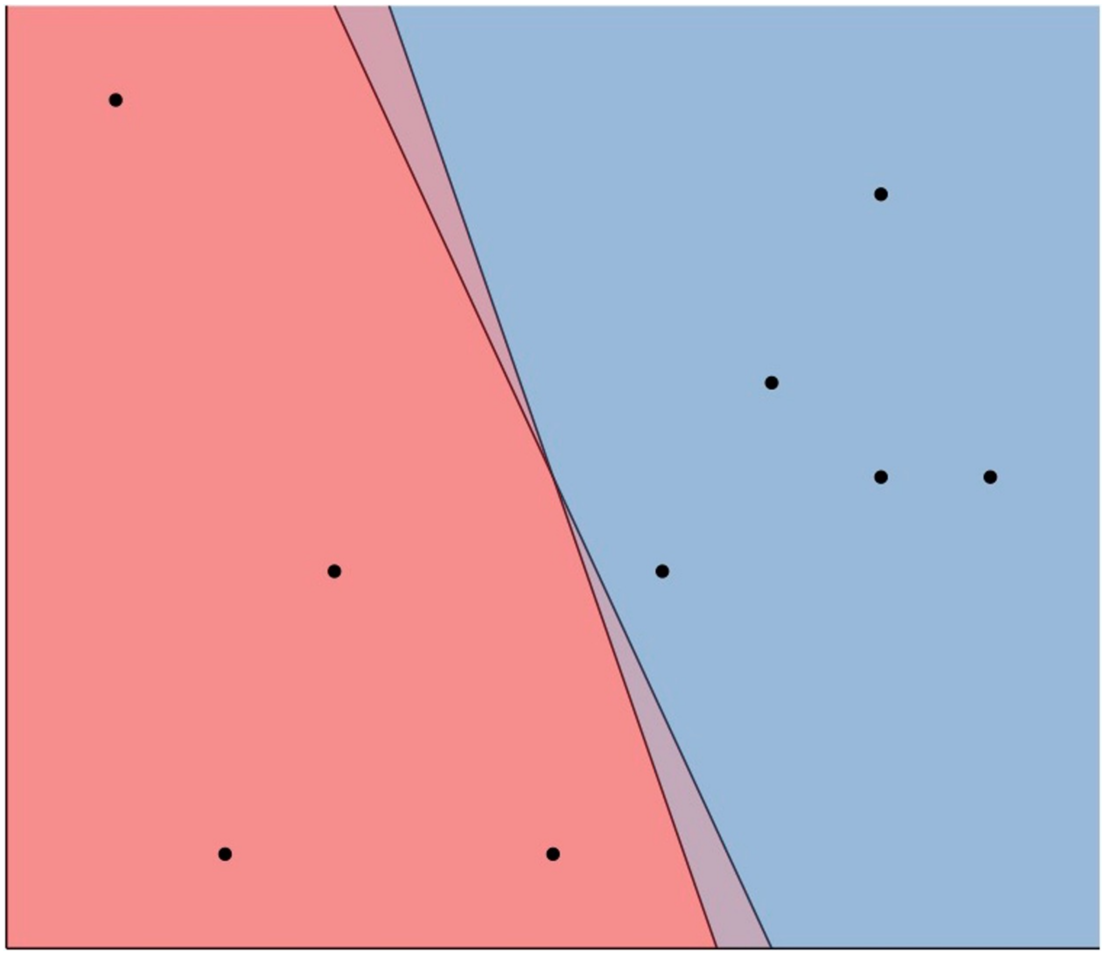


Intuition

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!



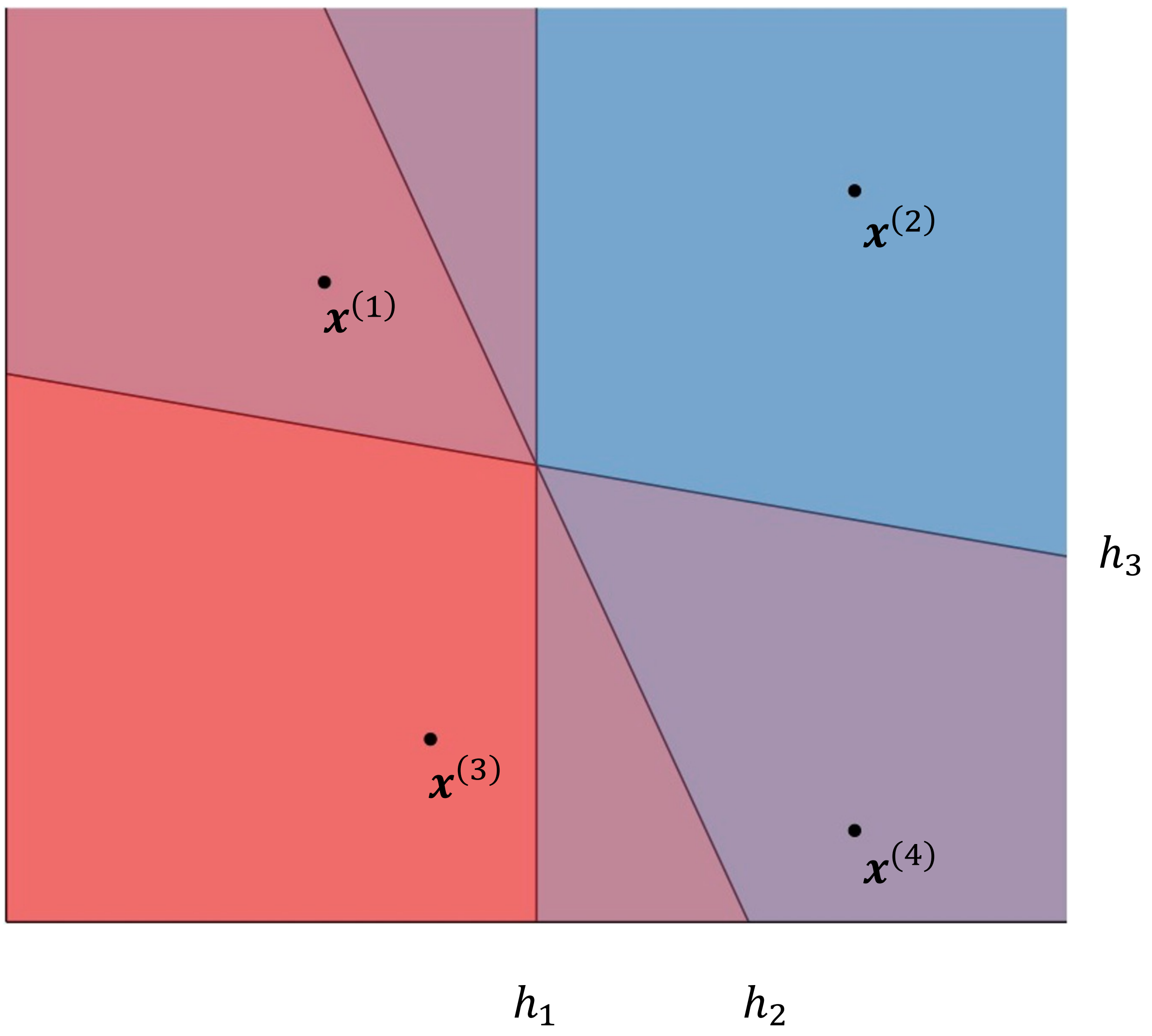
Labellings

- Given some finite set of data points $S = \left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}\right)$ 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

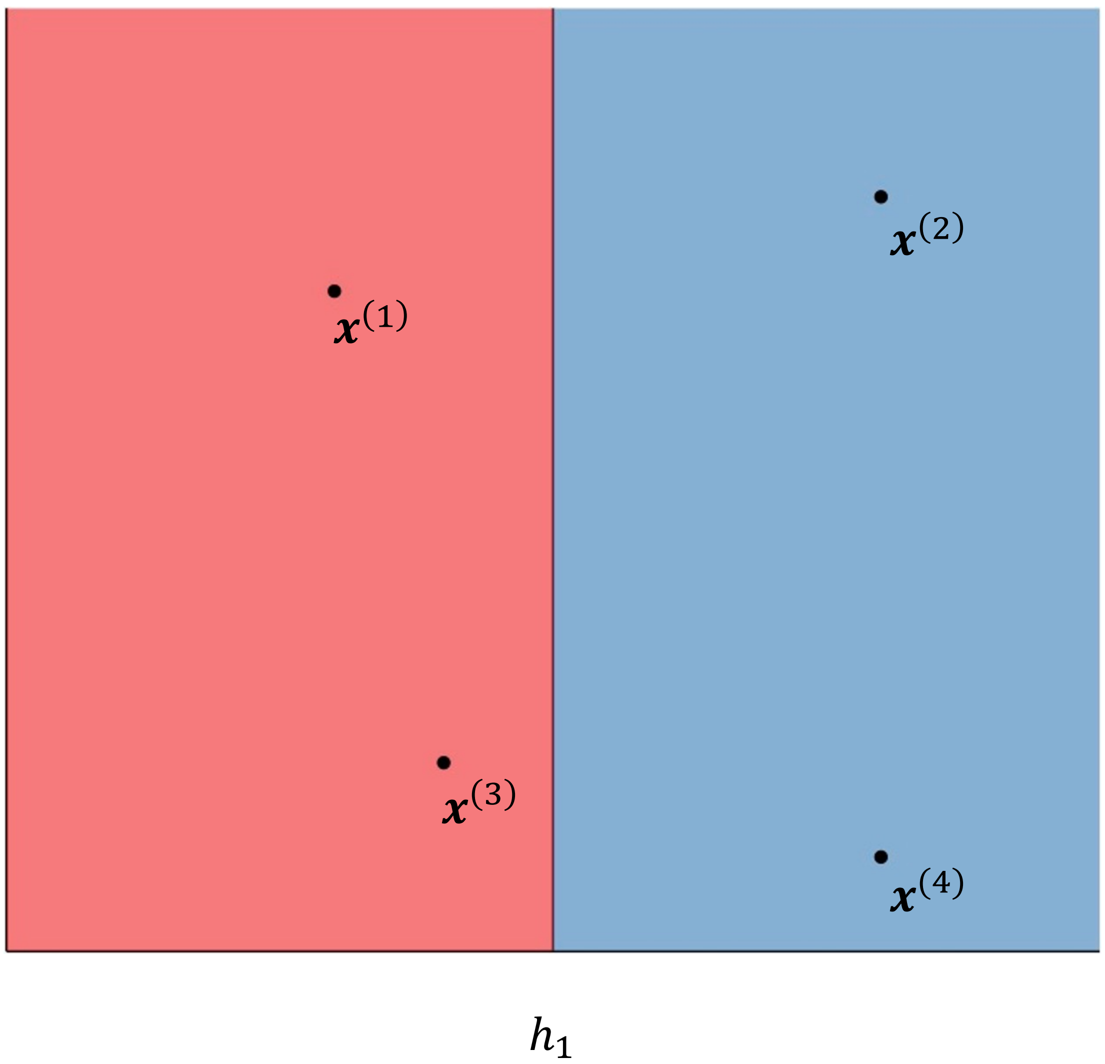
$\mathcal{H} = \{h_1, h_2, h_3\}$



Example: Labellings

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

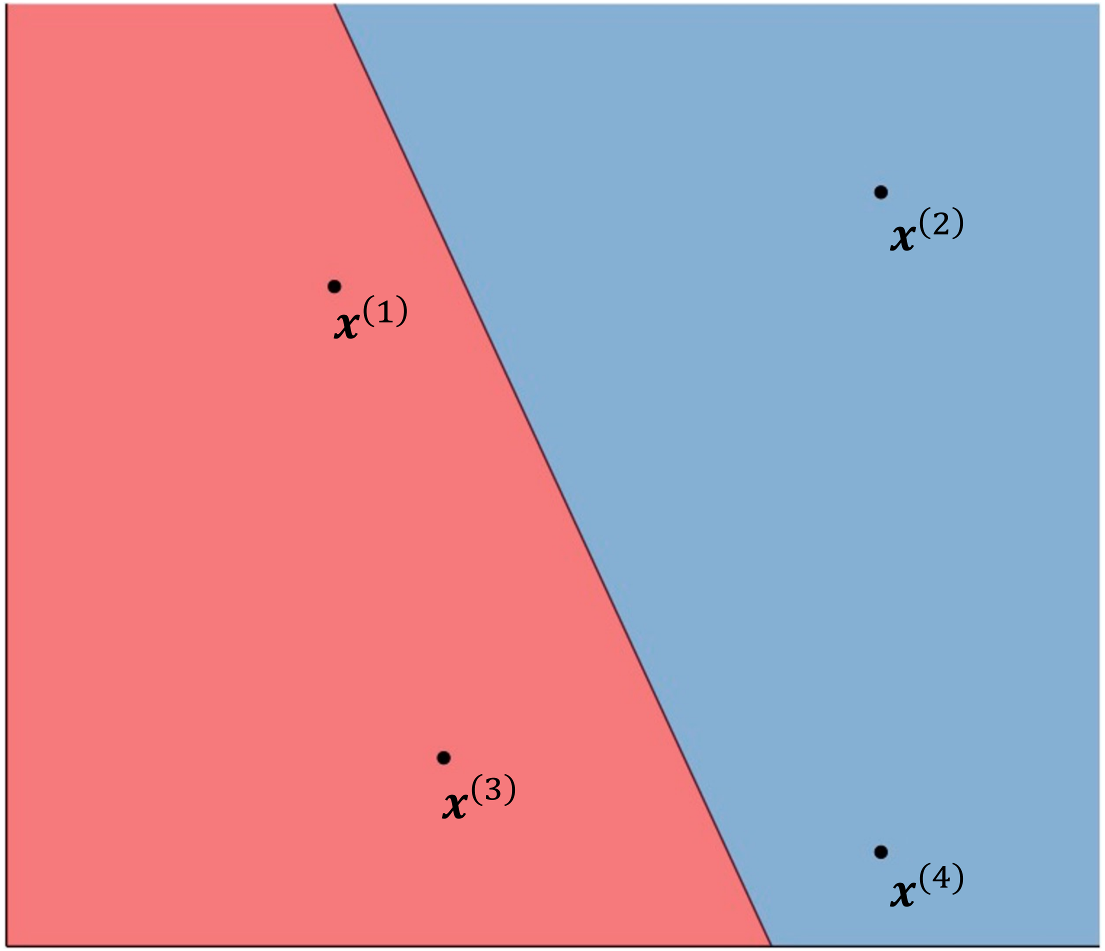
$$\left(h_1(\boldsymbol{x}^{(1)}), h_1(\boldsymbol{x}^{(2)}), h_1(\boldsymbol{x}^{(3)}), h_1(\boldsymbol{x}^{(4)})\right)$$
$$= (-1, +1, -1, +1)$$



Example: Labellings

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

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

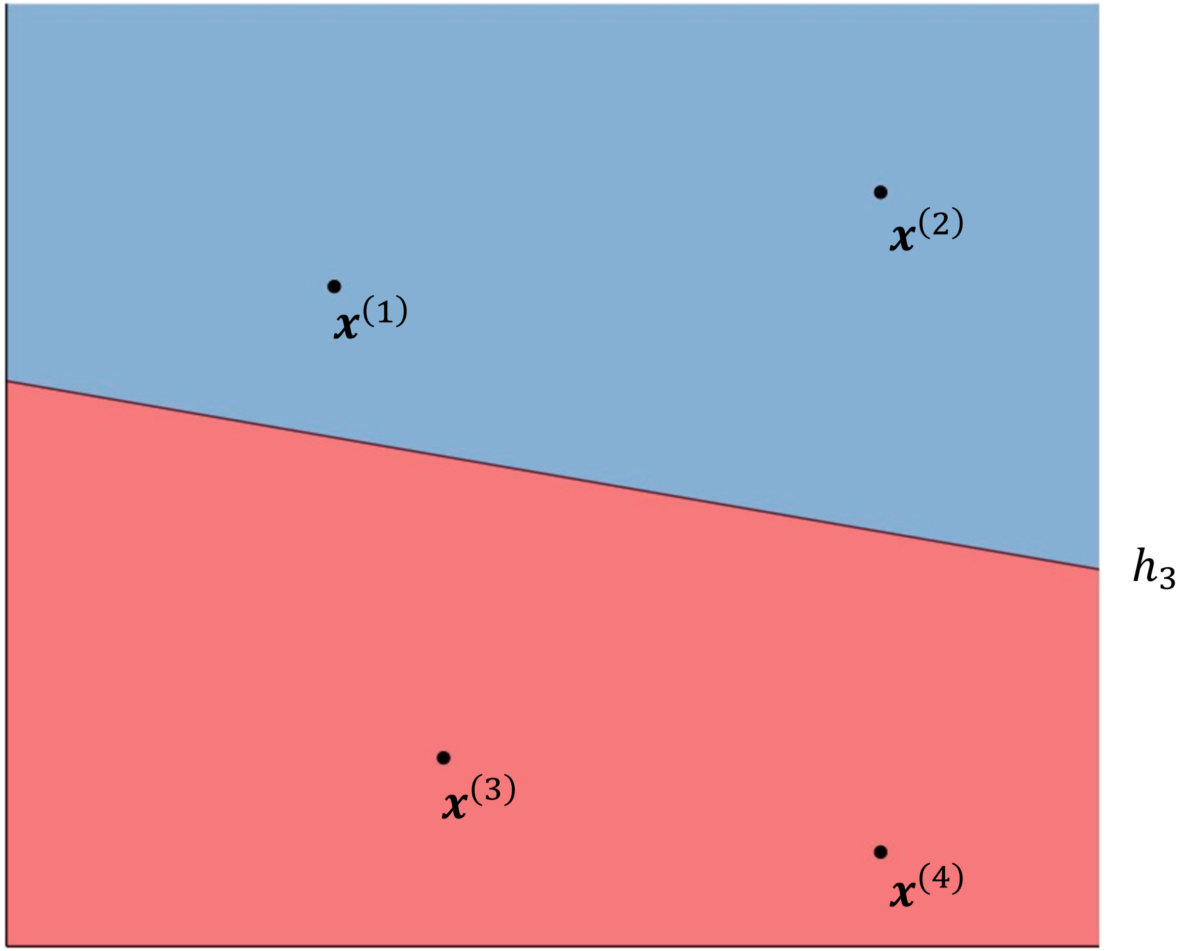


h_2

Example: Labellings

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

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

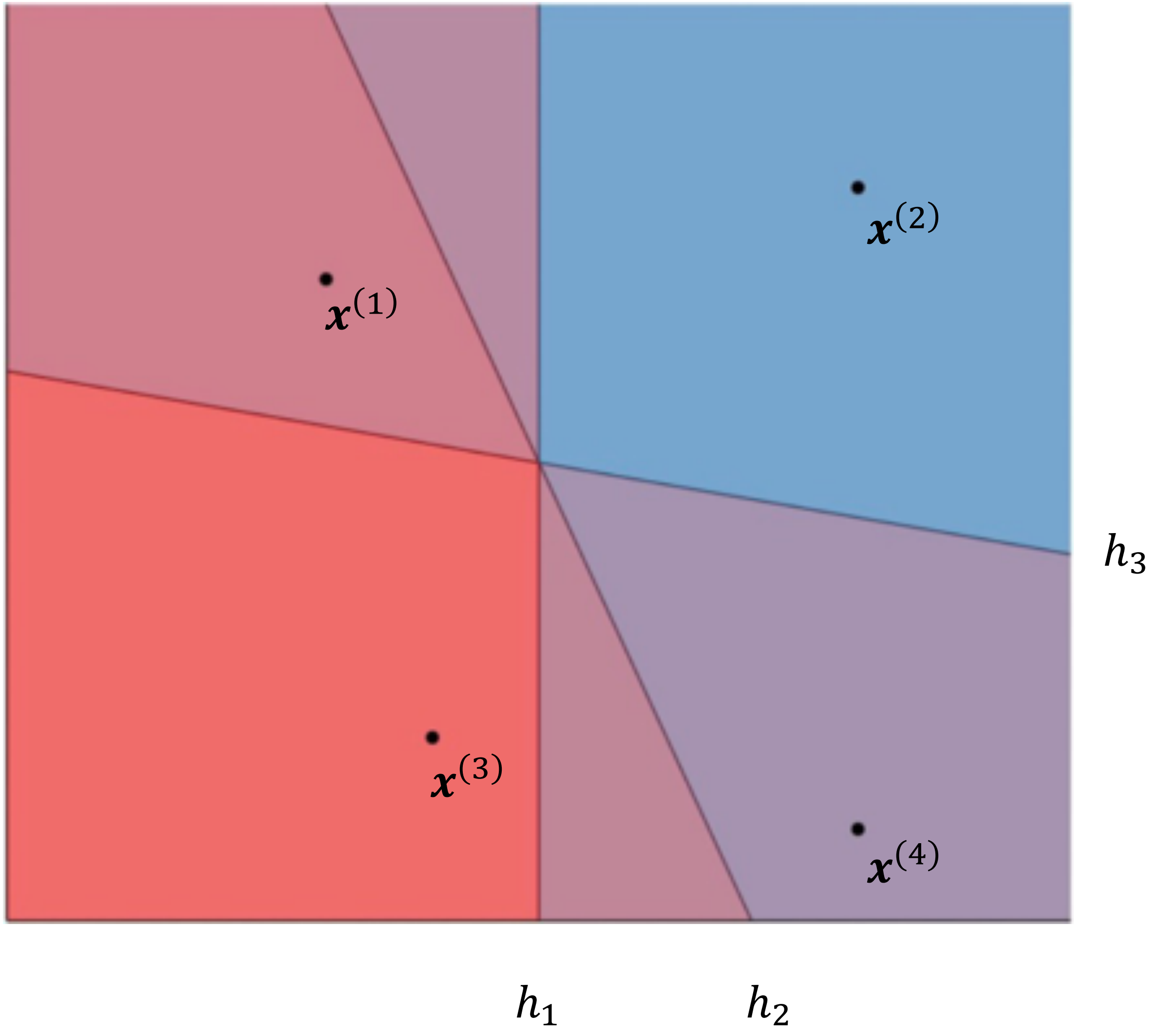


Example: Labellings

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

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

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

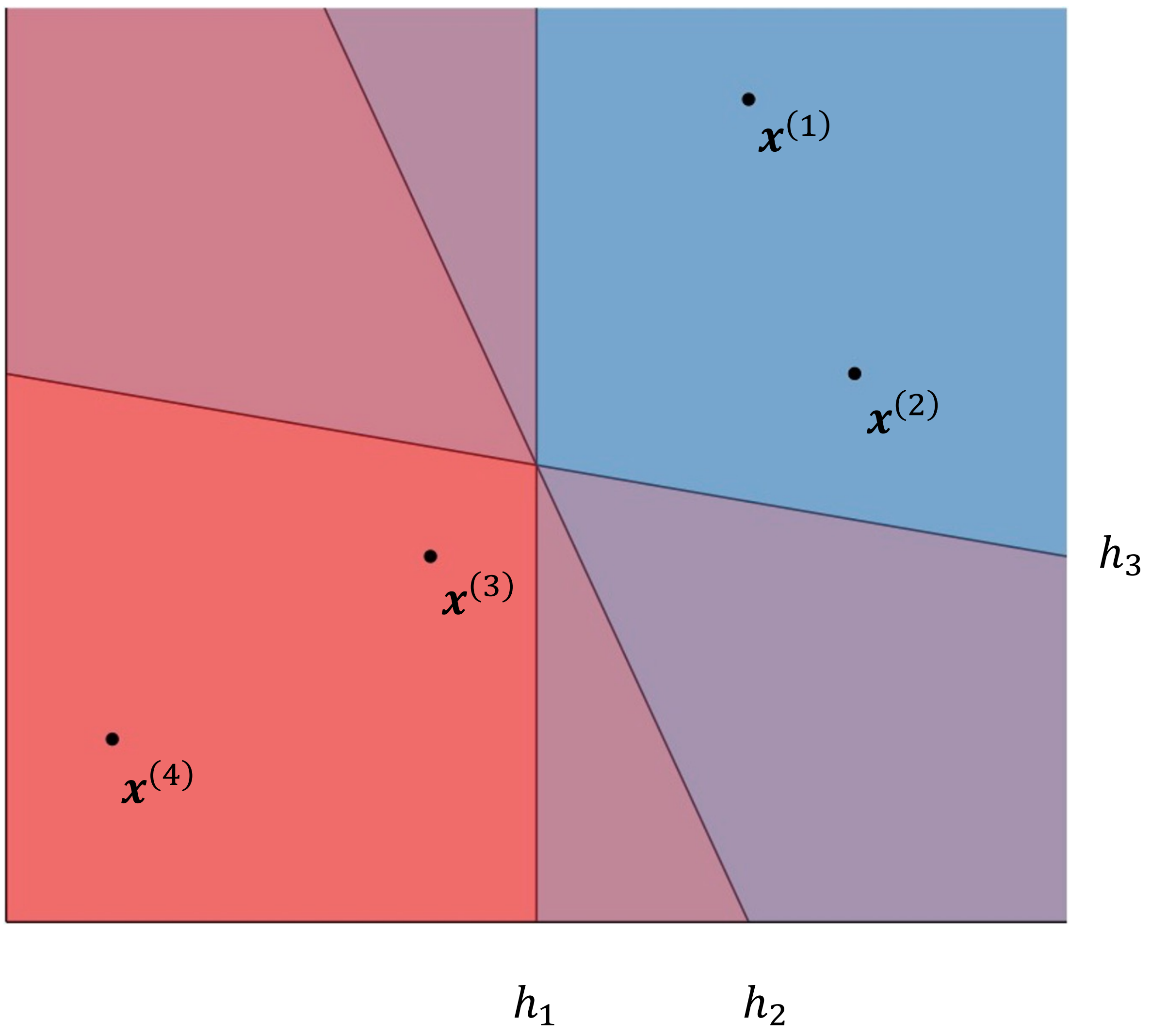


Example: Labellings

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

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

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



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 $g_{\mathcal{H}}(M) = 2^M$

if H shatters S and $|S| = M$

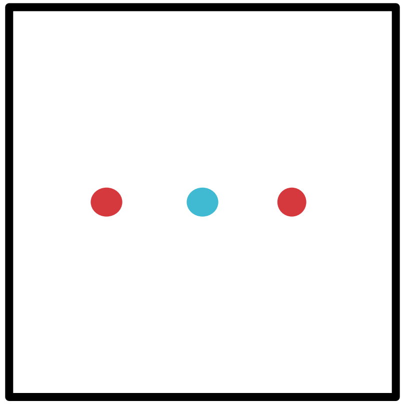
then $|H(S)| = 2^M$

Growth Function: Example

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

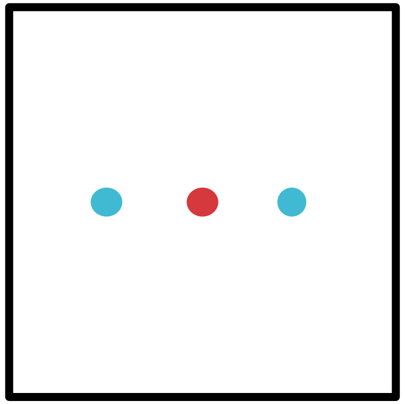
Growth Function: Example

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



Growth Function: Example

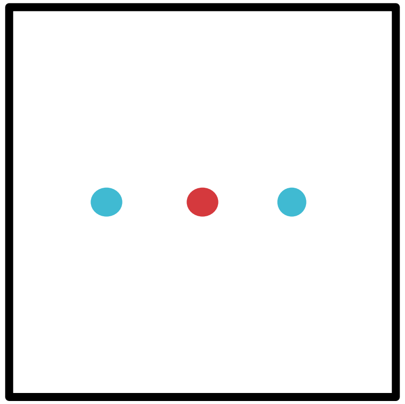
- $\boldsymbol{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(3)$?



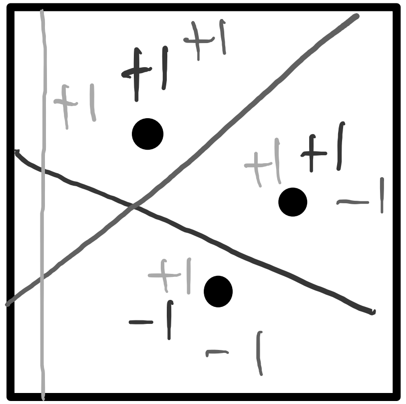
Growth Function: Example

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

- What is $g_{\mathcal{H}}(3)$?



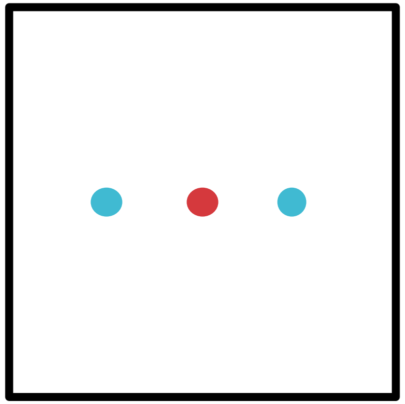
$|\mathcal{H}(S_1)| = 6$



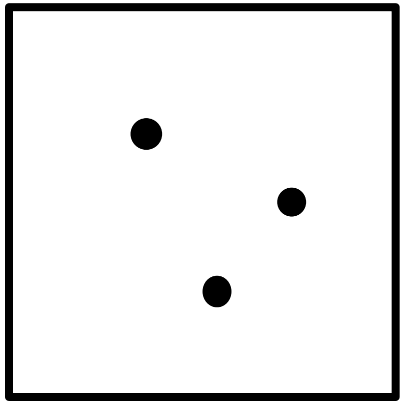
$|\mathcal{H}(S_2)| = 8$

Growth Function: Example

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



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



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

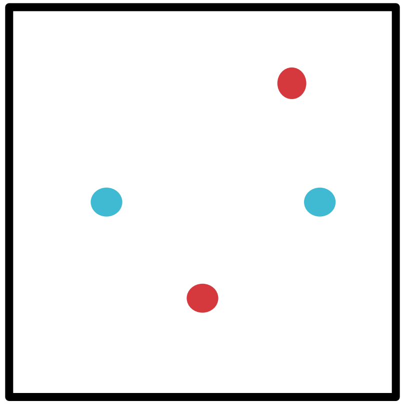
Growth Function: Example

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

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

Growth Function: Example

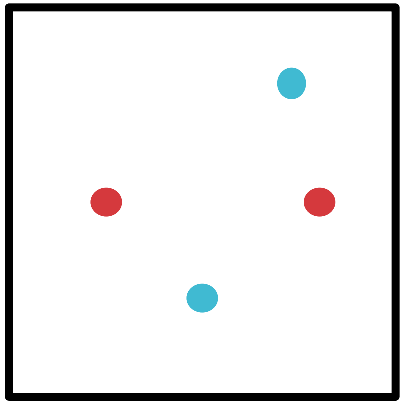
- $\boldsymbol{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(4)$?



Growth Function: Example

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

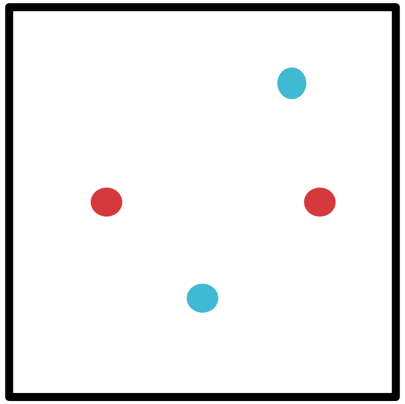
- What is $g_{\mathcal{H}}(4)$?

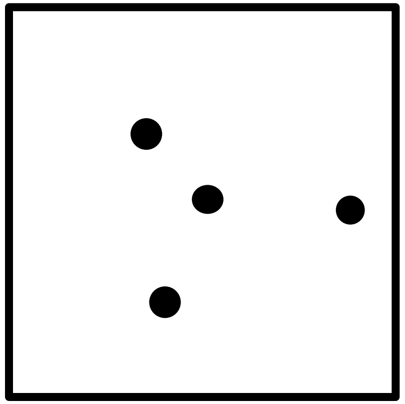


$|\mathcal{H}(S_1)| = 14$

Growth Function: Example

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



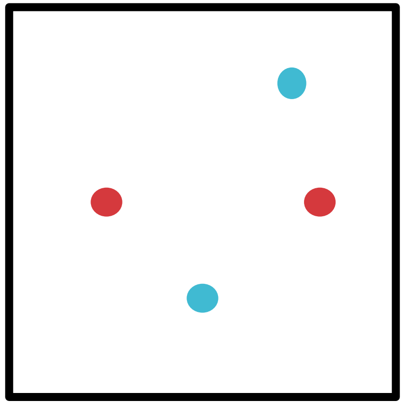


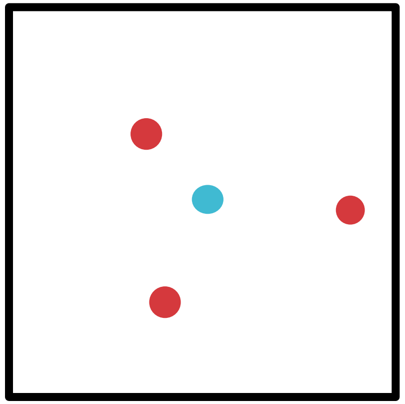
$|\mathcal{H}(S_1)| = 14$

Growth Function: Example

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

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

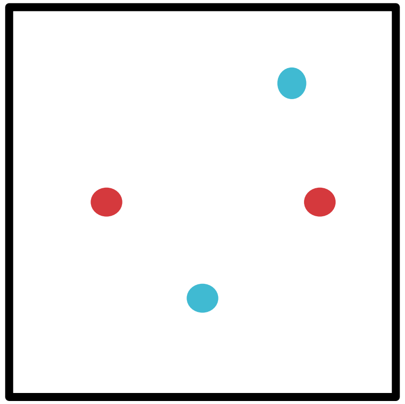


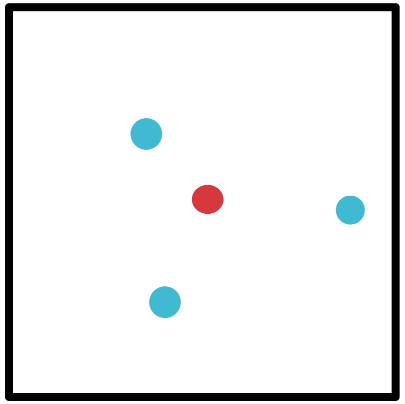


$|\mathcal{H}(S_1)| = 14$

Growth Function: Example

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

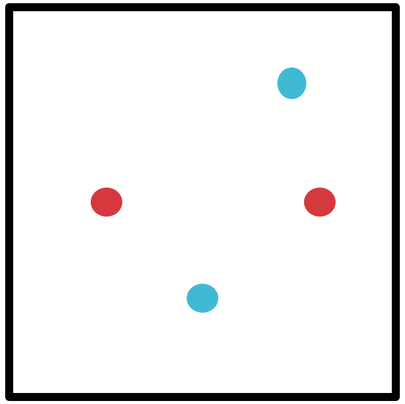




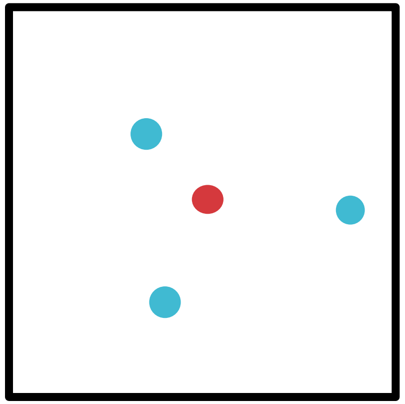
$|\mathcal{H}(S_1)| = 14$

Growth Function: Example

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



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



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

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

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

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

- M appears on both sides of the inequality…

Vapnik-Chervonenkis (VC)-Dimension

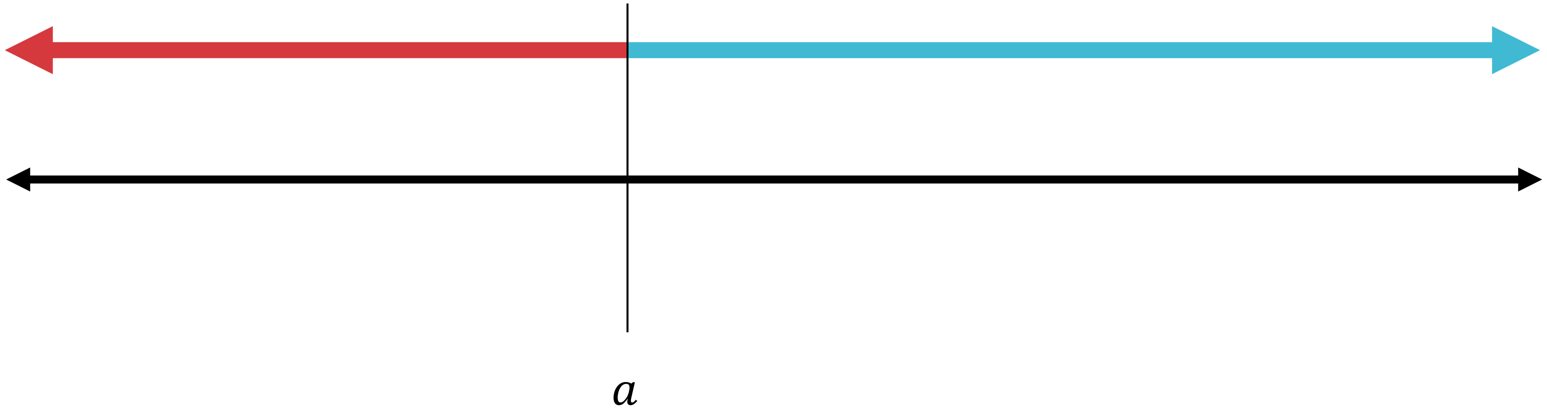
- $d_{VC}(\mathcal{H})$ = the largest value of M s.t. $g_{\mathcal{H}}(M) = 2^M$, i.e., the greatest number of data points that can be shattered by $\mathcal{H}$
  - If $\mathcal{H}$ can shatter arbitrarily large finite sets, then $d_{VC}(\mathcal{H}) = \infty$
  - $g_{\mathcal{H}}(M) = O\left(M^{d_{VC}(\mathcal{H})}\right)$ (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

As we saw d_{VC}(2-d linear separators) = 3
– d_{VC}(d-dimensional linear separators) = d+1

VC-Dimension: Example

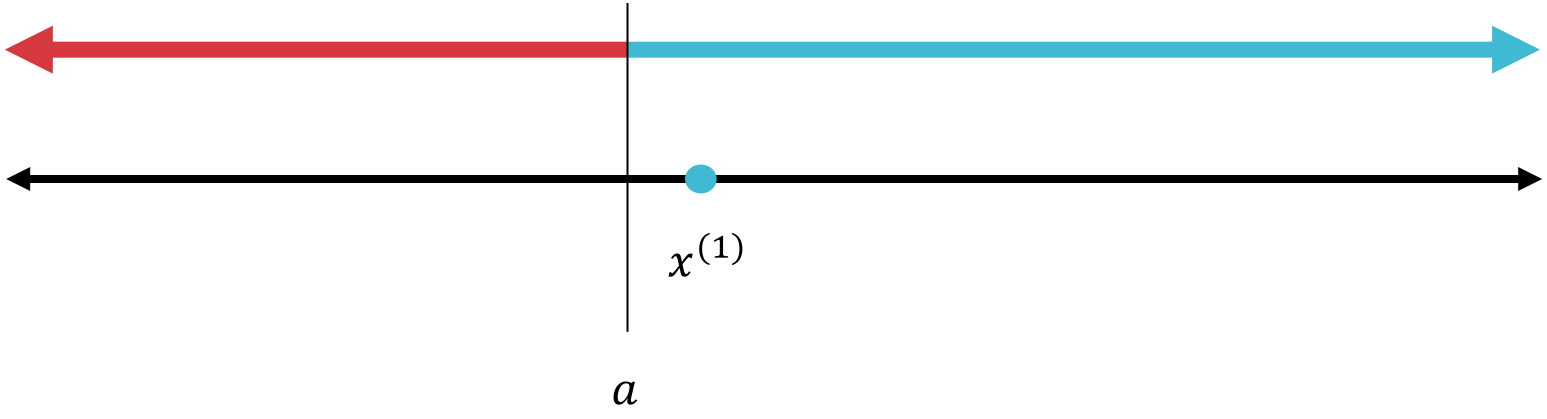
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

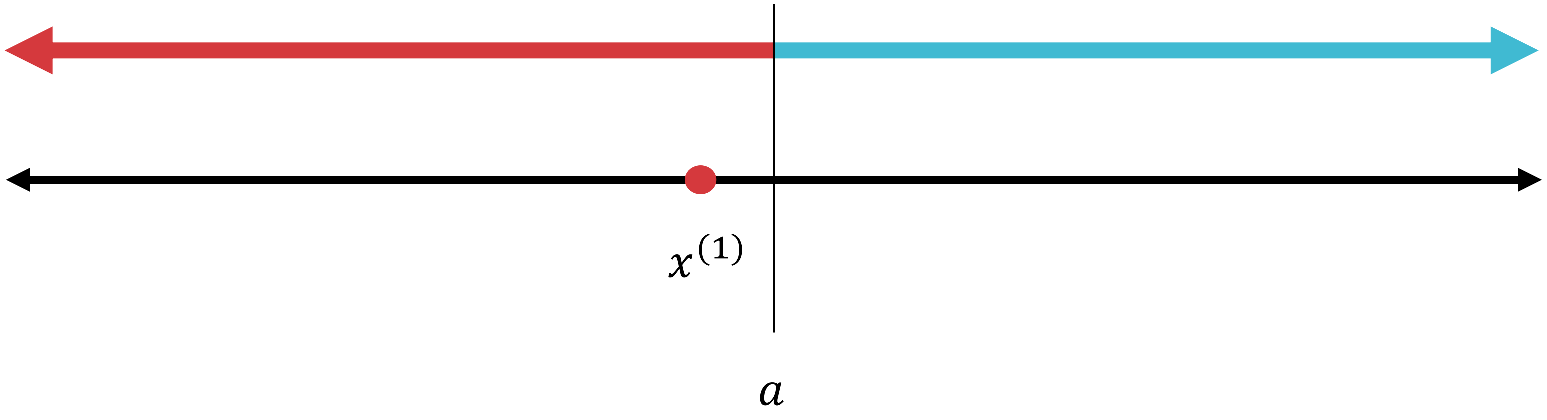
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

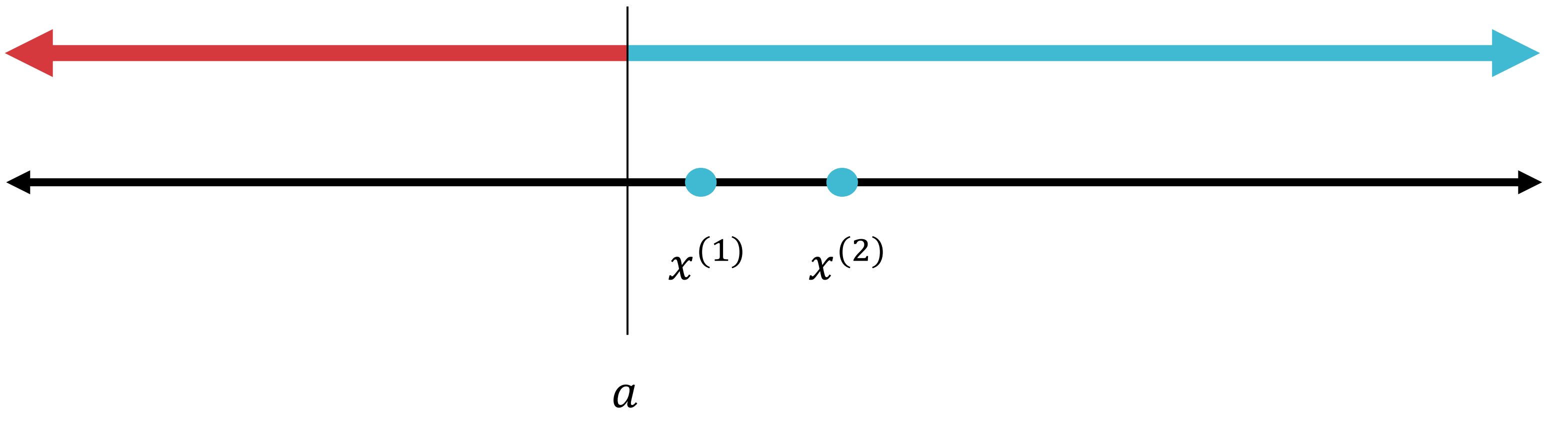
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

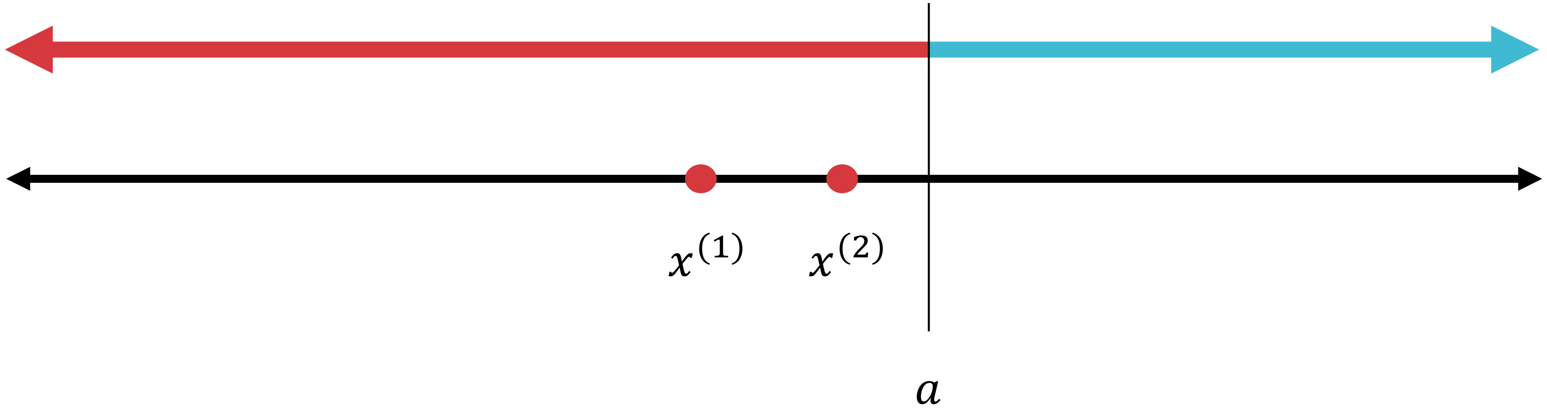
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

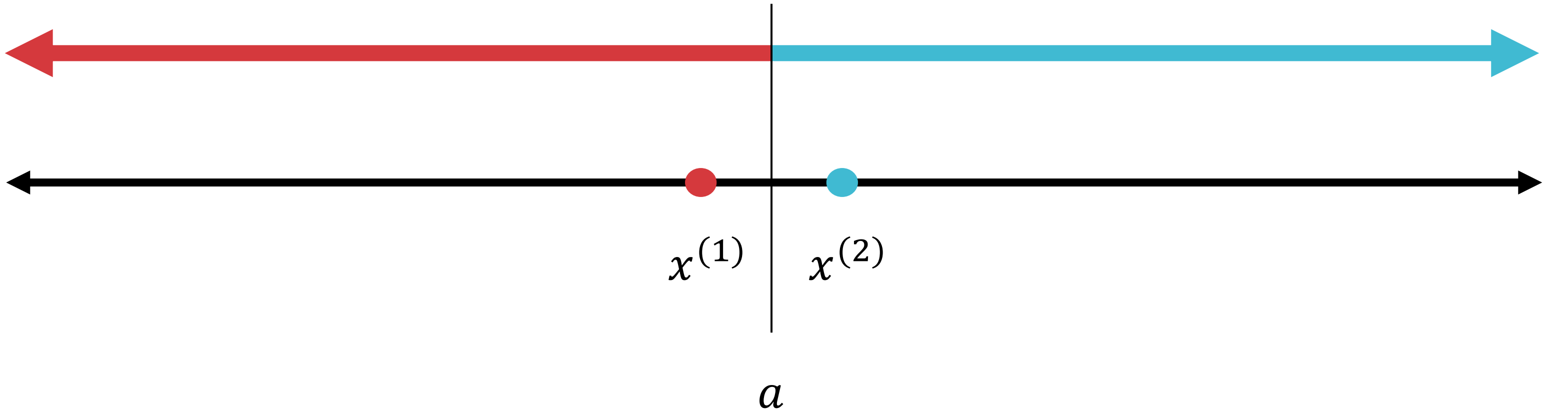
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

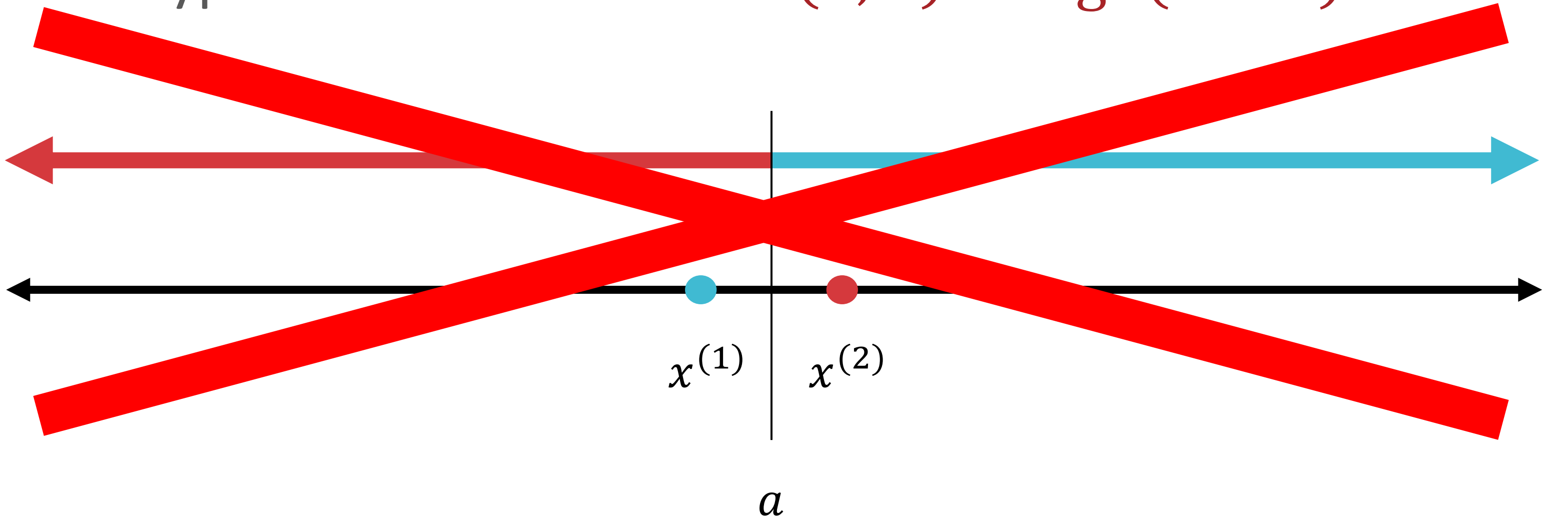
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

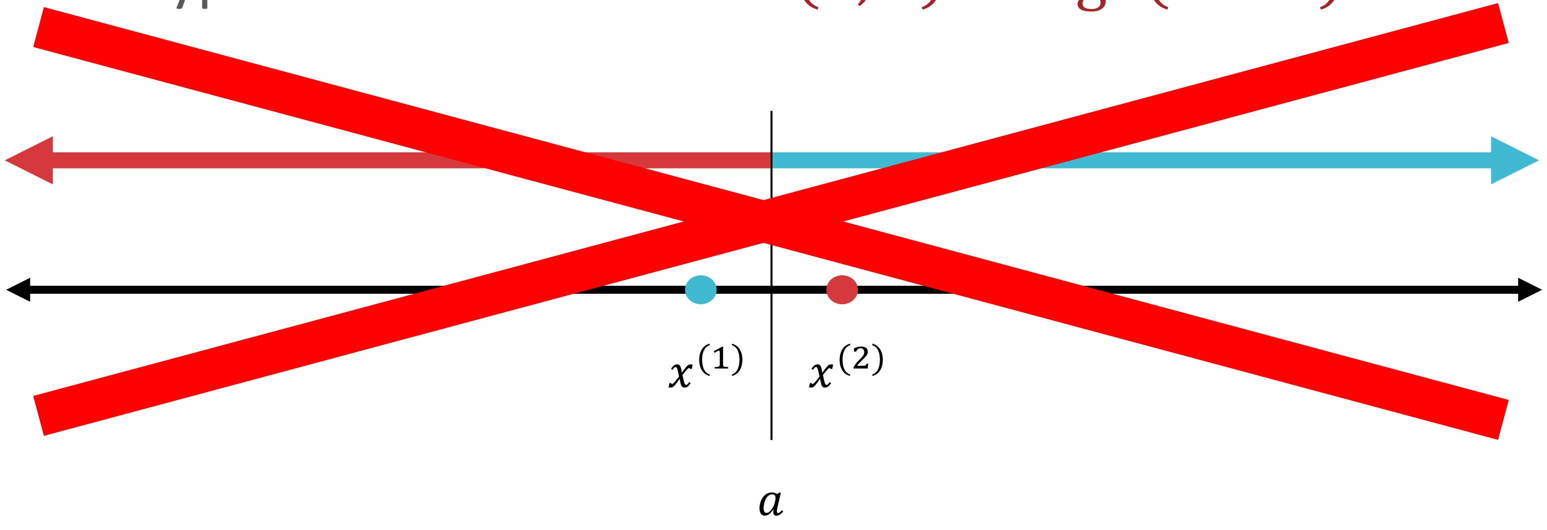
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

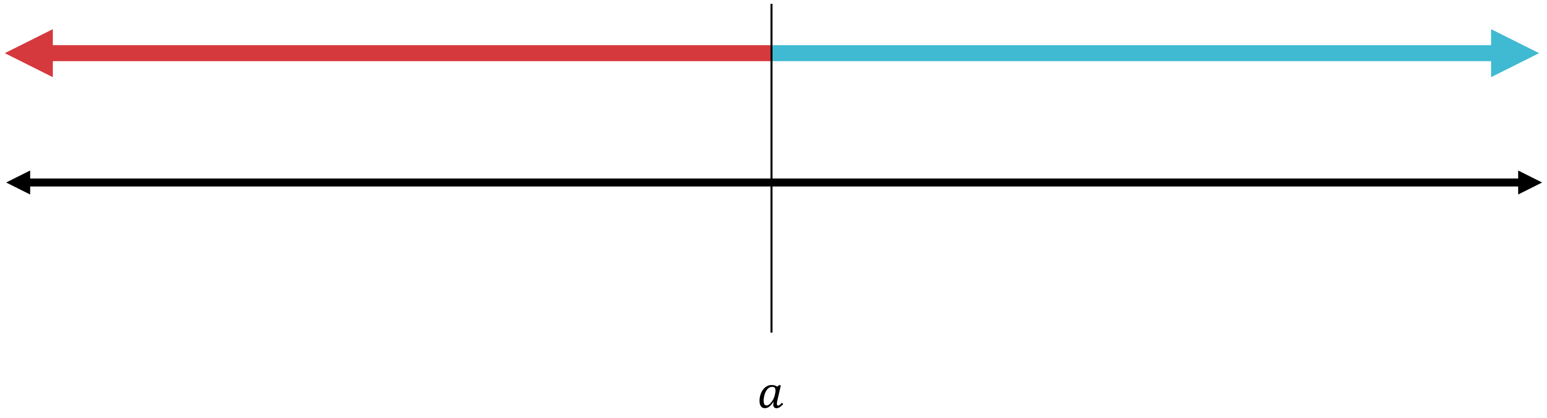
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

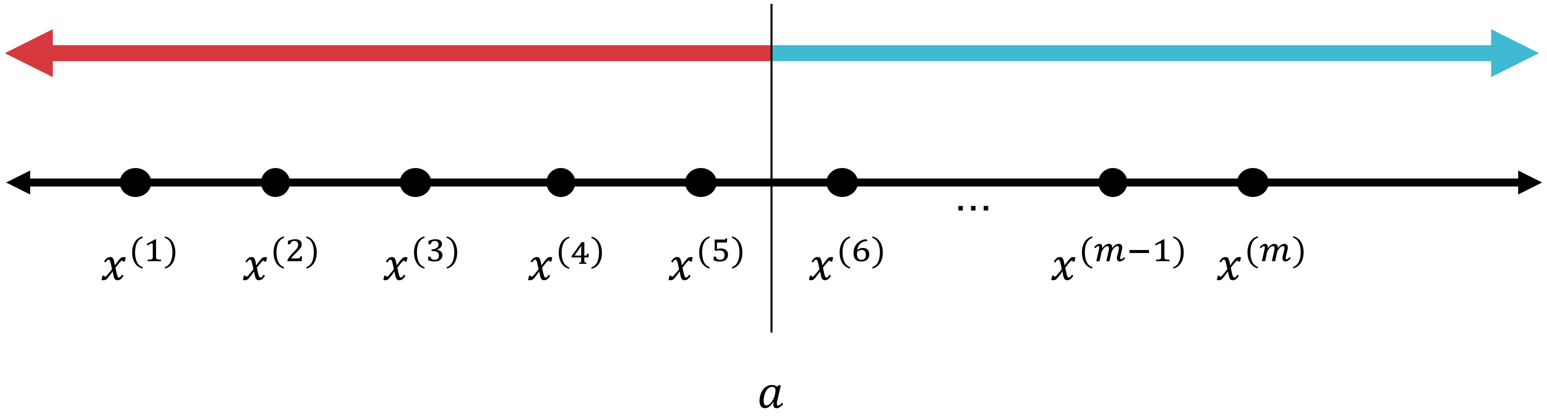
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

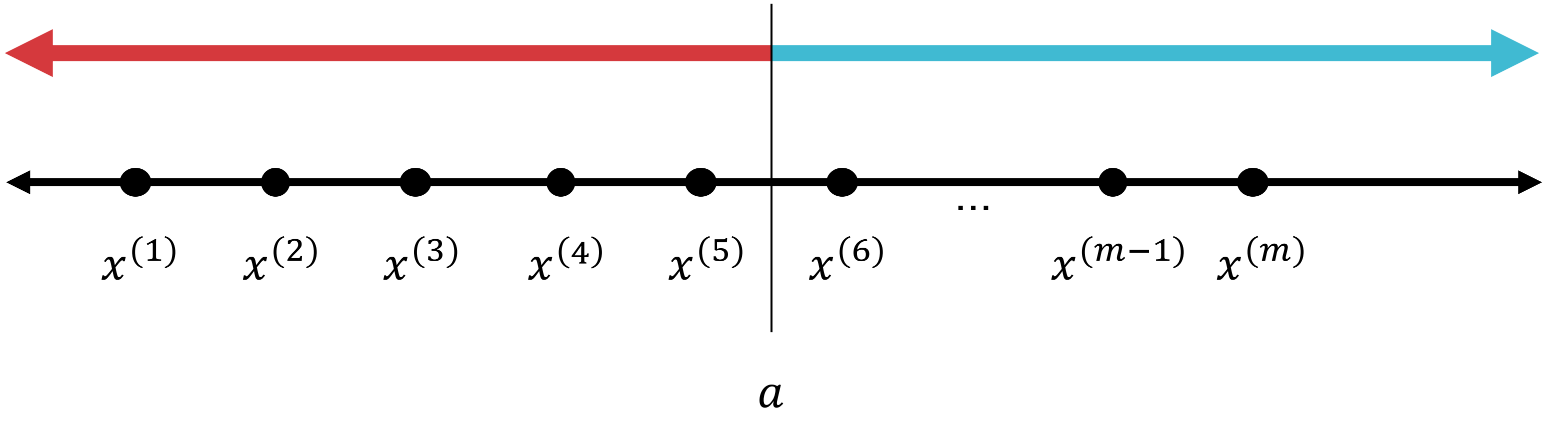
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

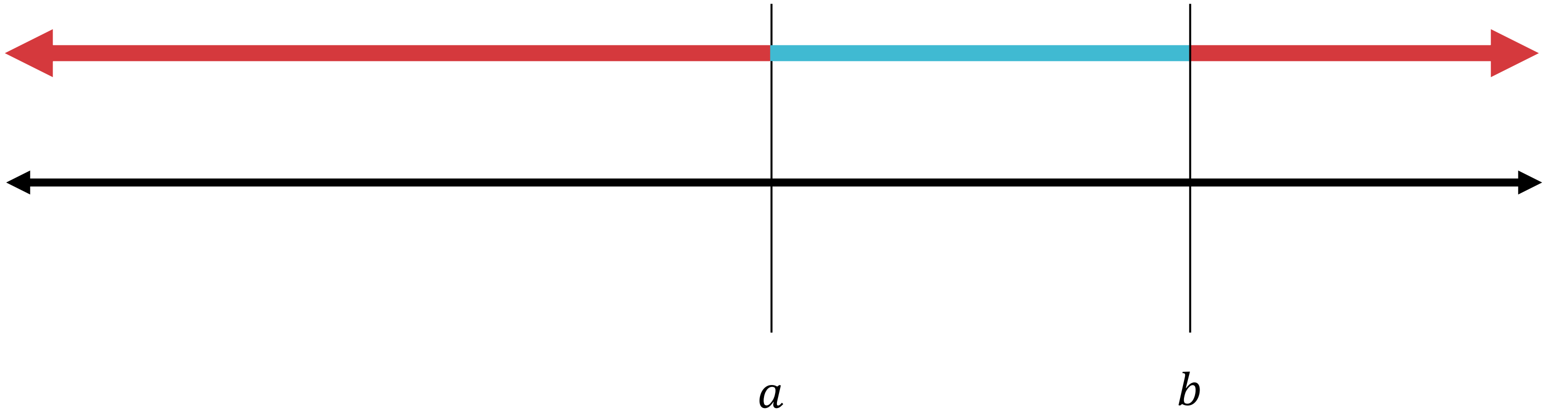
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

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



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

$1 \text{ and } m+1$

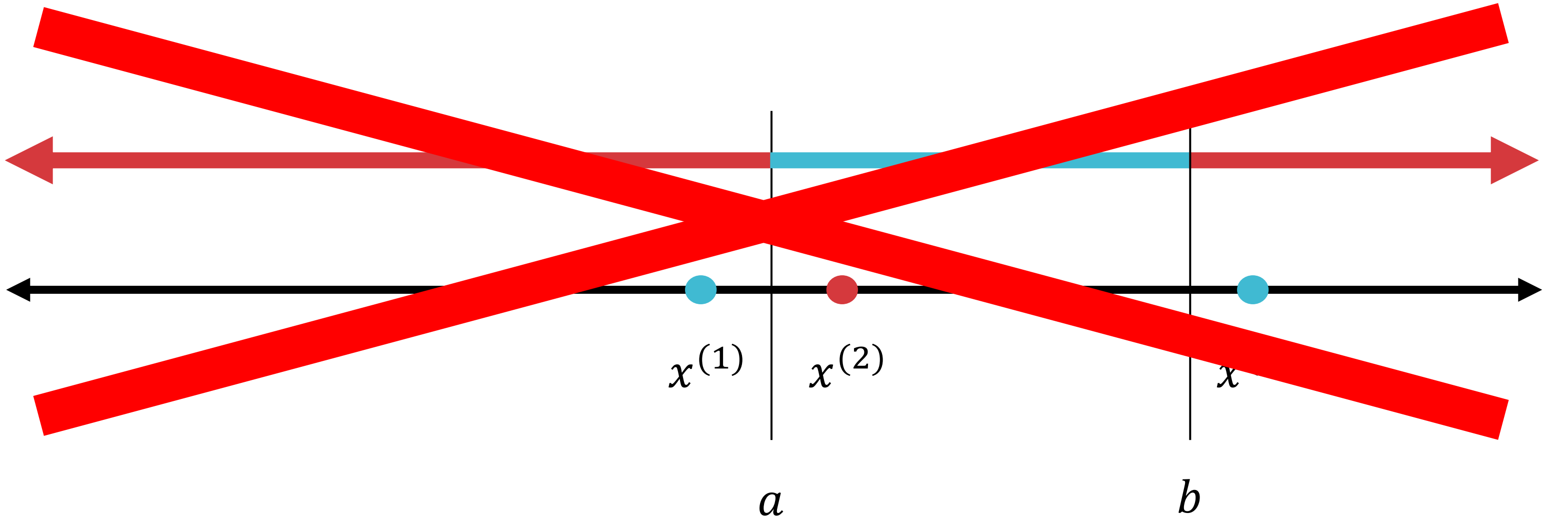
$2 \text{ and } m+1$

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

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

VC-Dimension: Example

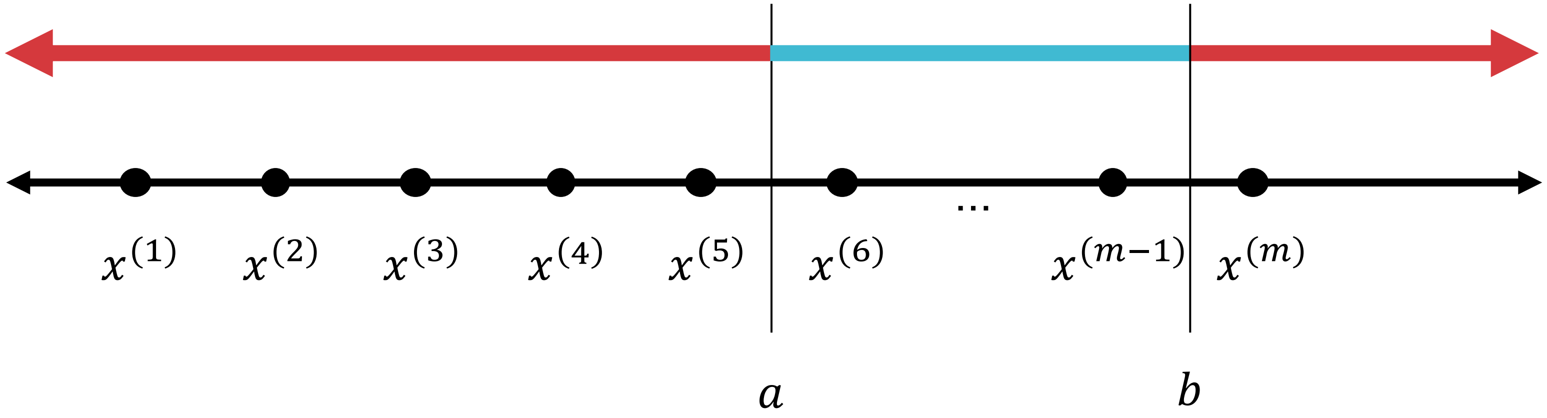
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive intervals



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

VC-Dimension: Example

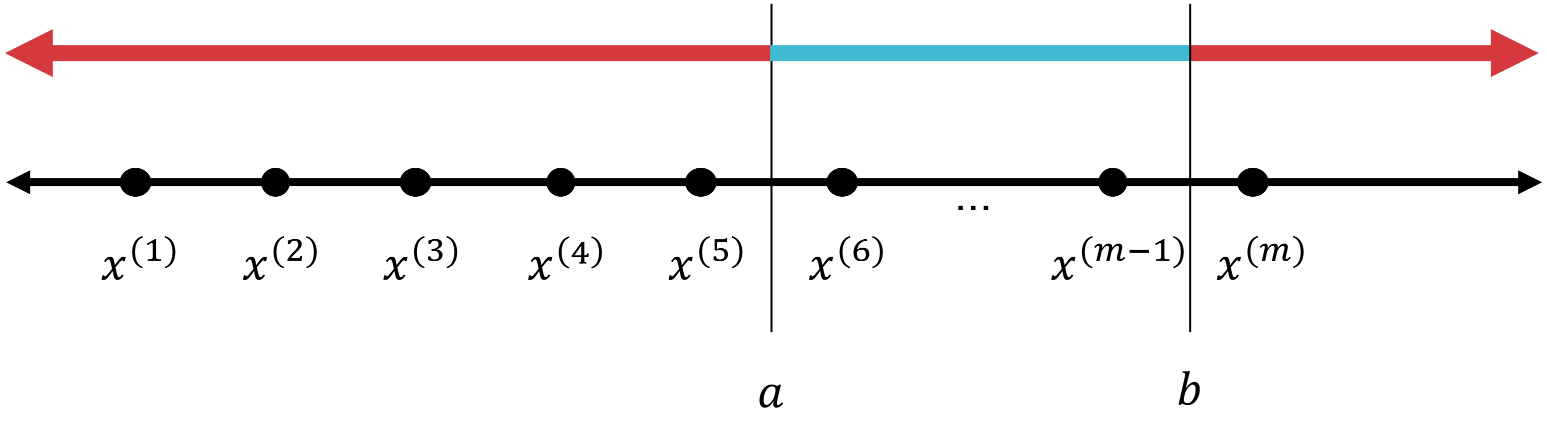
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive intervals



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

VC-Dimension: Example

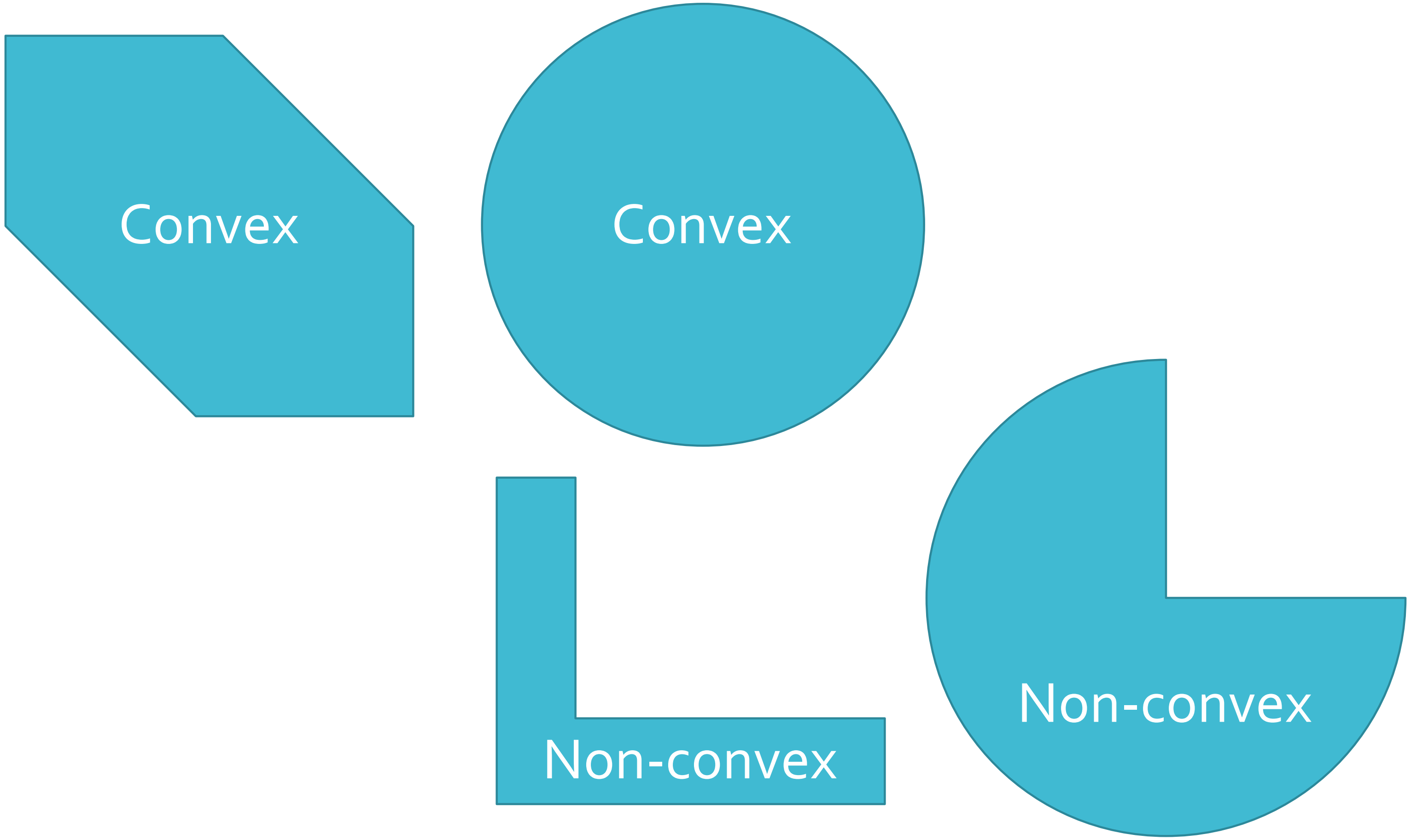
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive intervals



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

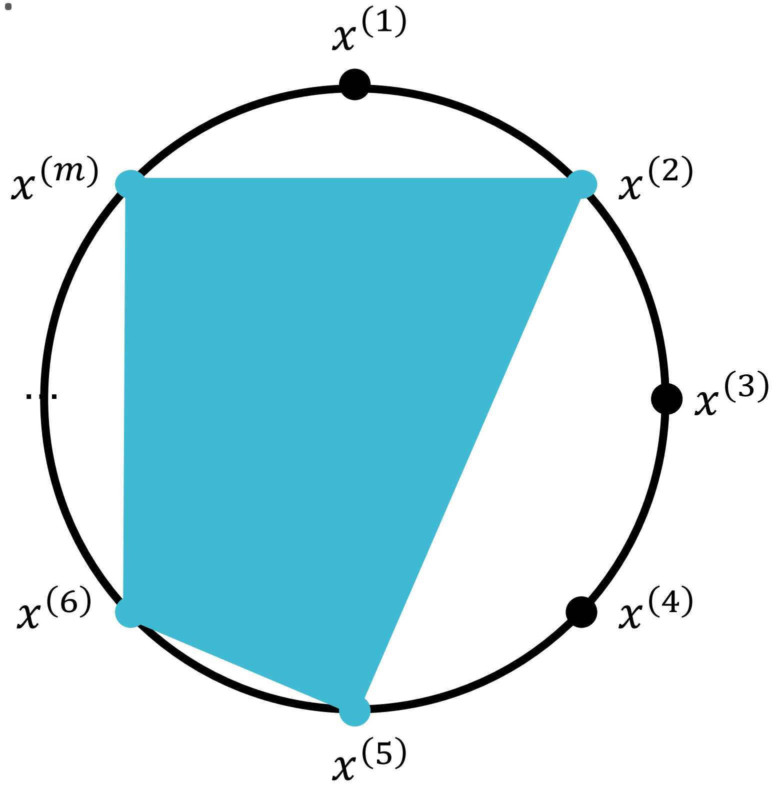
- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H}$ $=$ all 2-dimensional positive convex sets

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

Growth Function: Example

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

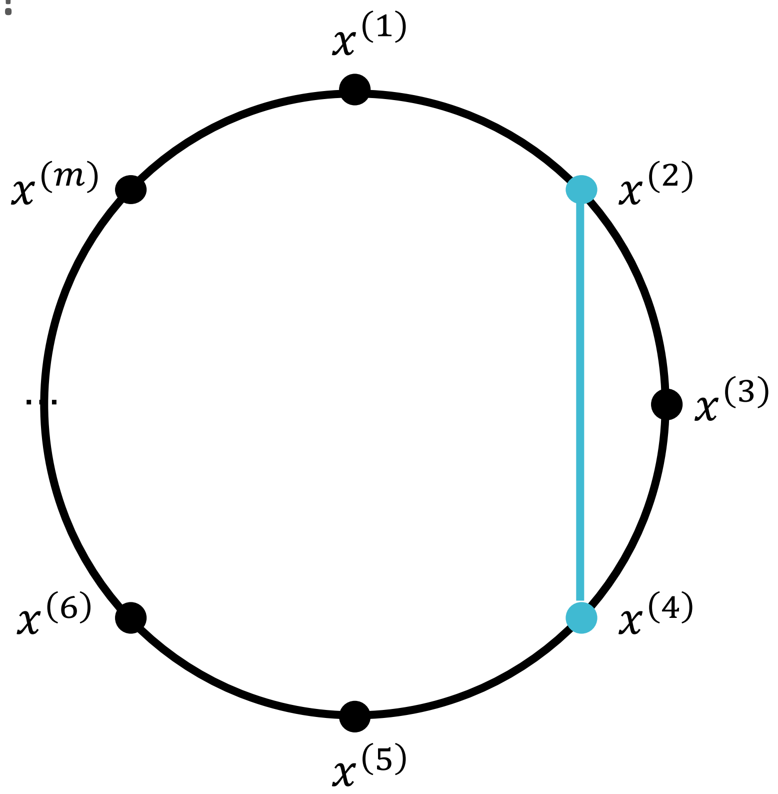
- What are $d_{VC}(\mathcal{H})$ and $g_{\mathcal{H}}(M)$?



Growth Function: Example

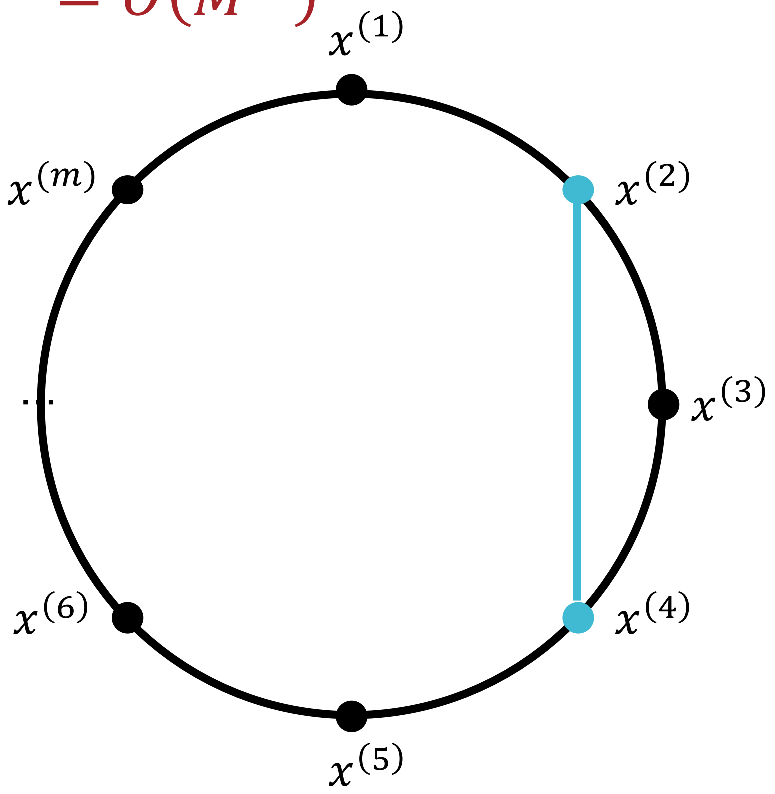
- $x^{(m)} \in \mathbb{R}^2$ and $\mathcal{H}$ = all 2-dimensional positive convex sets

- 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

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

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

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

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

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

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

Bias-Variance Tradeoff

- Assume a regression task with squared error and let $h_S \in \mathcal{H}$ = the hypothesis trained on training data S
- $err_D(h_S) = \mathbb{E}_{\boldsymbol{x} \sim D}\left[\left(h_S(\boldsymbol{x}) - c^*(\boldsymbol{x})\right)^2\right]$

Bias-Variance Tradeoff

- Assume a regression task with squared error and let $h_S \in \mathcal{H}$ = the hypothesis trained on training data S
- $err_D(h_S) = \mathbb{E}_{\boldsymbol{x}\sim D}\left[\left(h_S(\boldsymbol{x}) - c^*(\boldsymbol{x})\right)^2\right]$
- $\mathbb{E}_S[err_D(h_S)] = \mathbb{E}_{\boldsymbol{x}\sim D}\left[\mathbb{E}_S[h_S(\boldsymbol{x})^2] - 2\bar{h}(\boldsymbol{x})c^*(\boldsymbol{x}) + c^*(\boldsymbol{x})^2\right]$

Bias-Variance Tradeoff

How much does h change if the training data set changes?

$$\mathbb{E}_S[err_D(h_S)] = \mathbb{E}_{x \sim D}\left[\mathbb{E}_S[h_S(\boldsymbol{x})^2 - \bar{h}(\boldsymbol{x})^2] + \left(\bar{h}(\boldsymbol{x}) - c^*(\boldsymbol{x})\right)^2\right]$$

How well on average does h approximate c^*?

Bias-Variance Tradeoff

How well could h approximate anything?

$$\mathbb{E}_S[err_D(h_S)] = \mathbb{E}_{\boldsymbol{x}\sim D}\left[\mathbb{E}_S[h_S(\boldsymbol{x})^2 - \bar{h}(\boldsymbol{x})^2] + \left(\bar{h}(\boldsymbol{x}) - c^*(\boldsymbol{x})\right)^2\right]$$

How well on average does h approximate c^*?

Bias-Variance Tradeoff

Increases as $\mathcal{H}$ becomes more complex

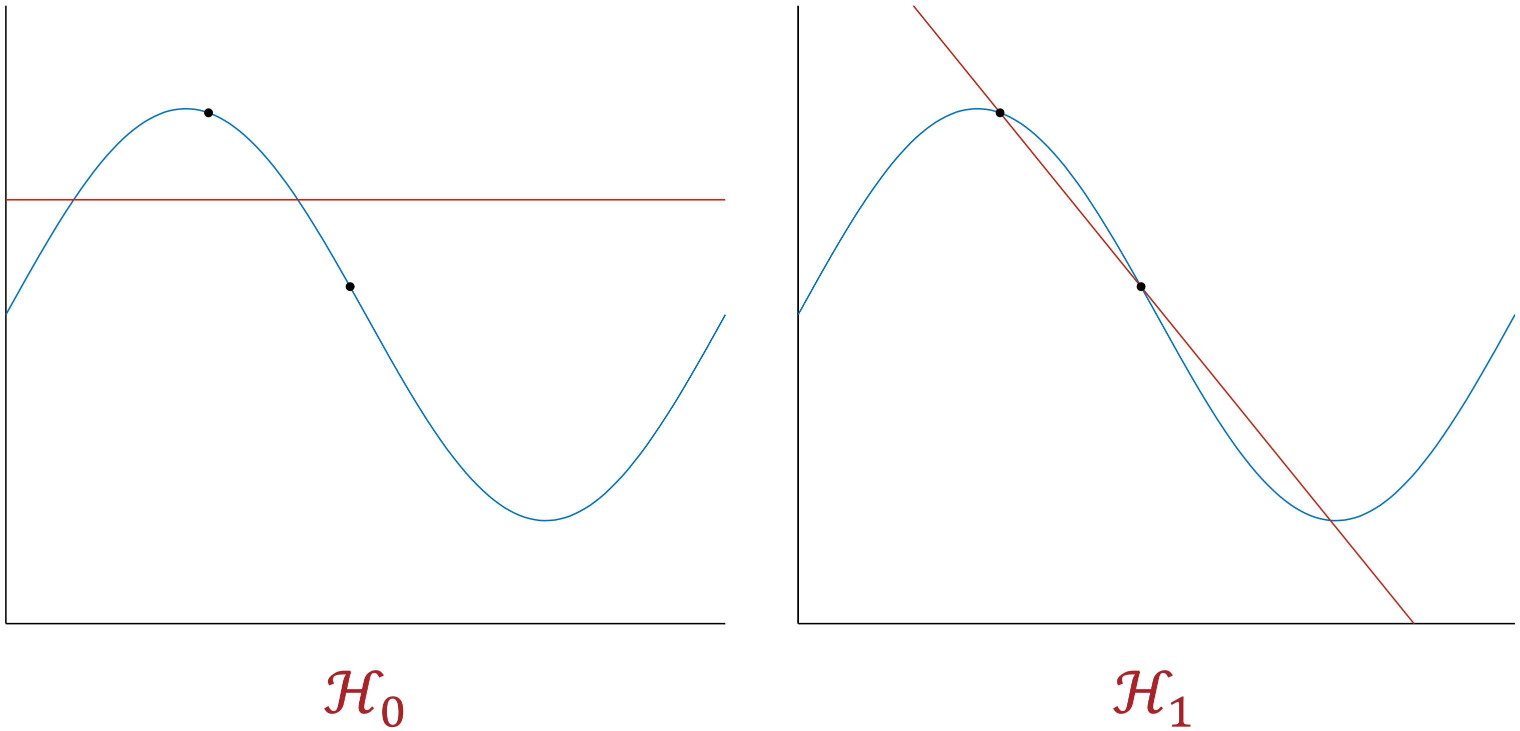
$$\mathbb{E}_S[err_D(h_S)] = \mathbb{E}_{\boldsymbol{x}\sim D}\left[\mathbb{E}_S\left[h_S(\boldsymbol{x})^2 - \bar{h}(\boldsymbol{x})^2\right] + \left(\bar{h}(\boldsymbol{x}) - c^*(\boldsymbol{x})\right)^2\right]$$

Decreases as $\mathcal{H}$ becomes more complex

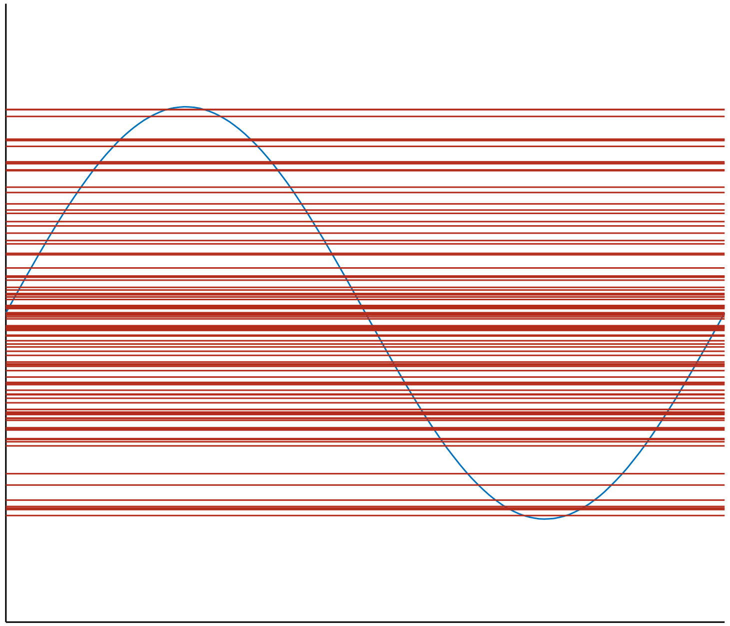
Bias-Variance Tradeoff: Example

- $x^{(i)} \in \mathbb{R}$ and $D = \text{Uniform}(0, 2\pi)$
- $c^* = \sin(\cdot)$, i.e., $y = \sin(x)$
- $N = 2 \rightarrow \mathcal{D} = \{(x^{(1)}, \sin(x^{(1)})), (x^{(2)}, \sin(x^{(2)}))\}$
- $\mathcal{H}_0 = \{h : h(x) = b\}$ and $\mathcal{H}_1 = \{h : h(x) = ax + b\}$

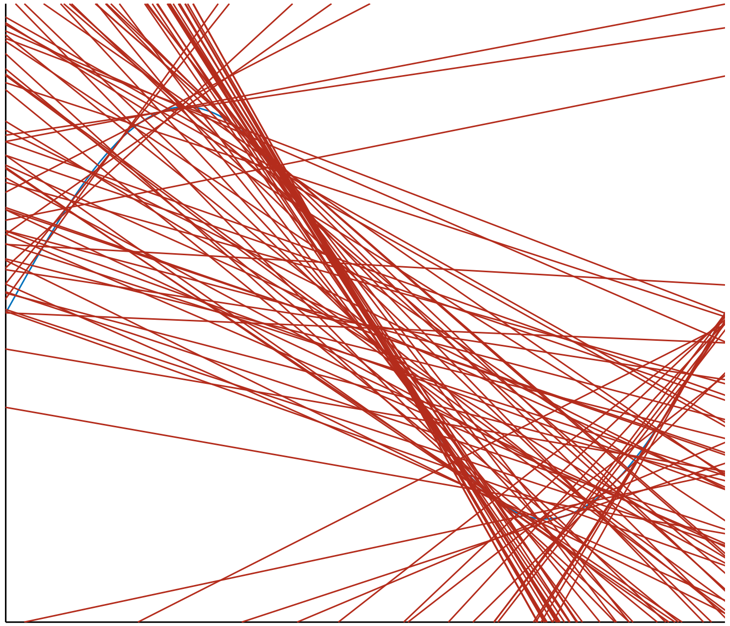
Bias-Variance Tradeoff: Example



Bias-Variance Tradeoff: Example

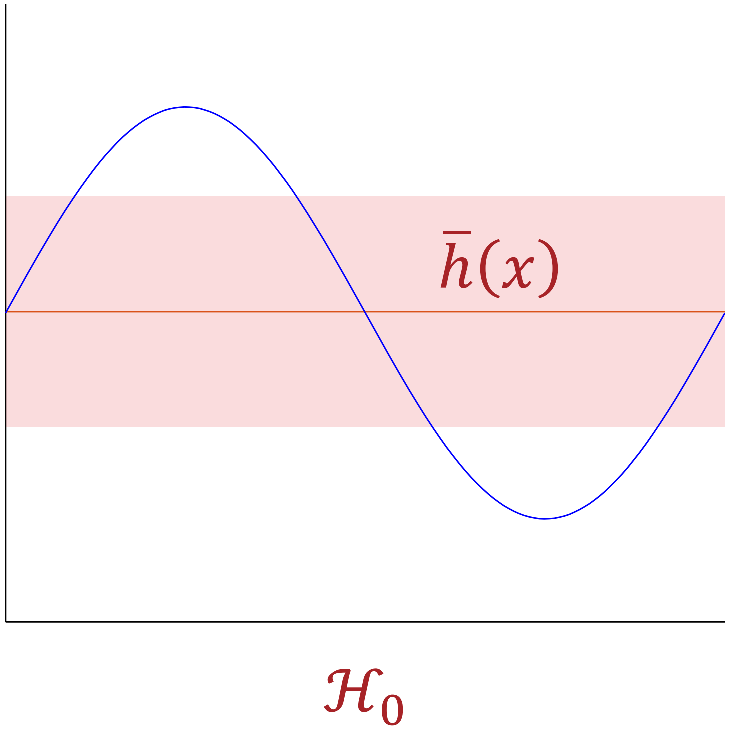


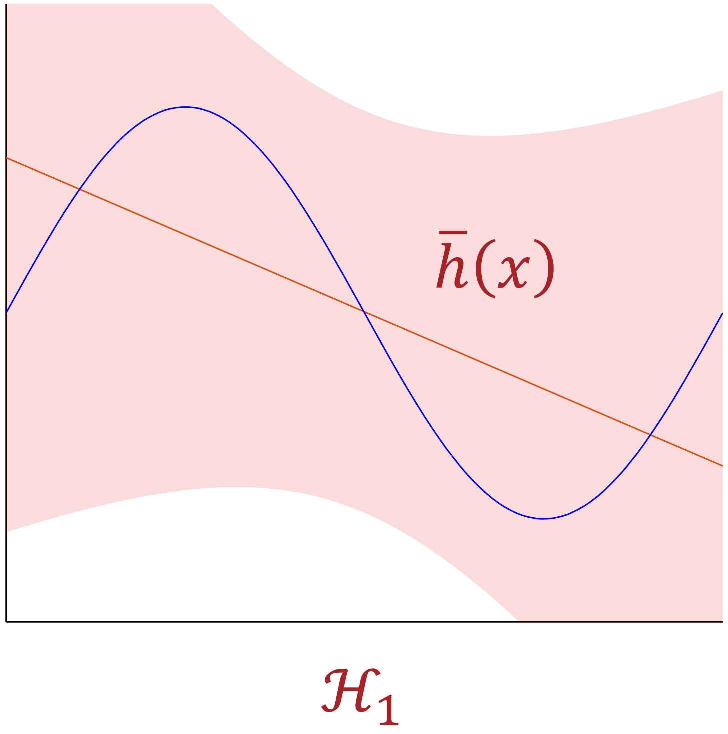
$\mathcal{H}_0$



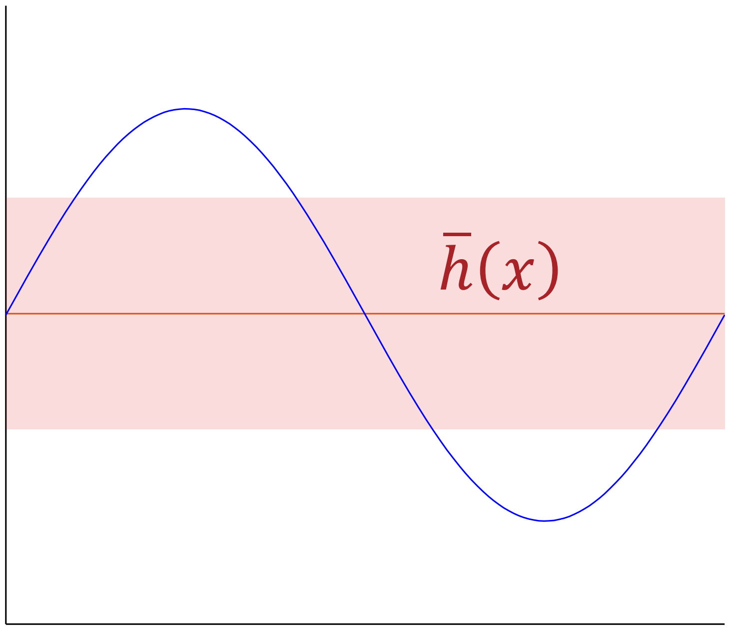
$\mathcal{H}_1$

Bias-Variance Tradeoff: Example





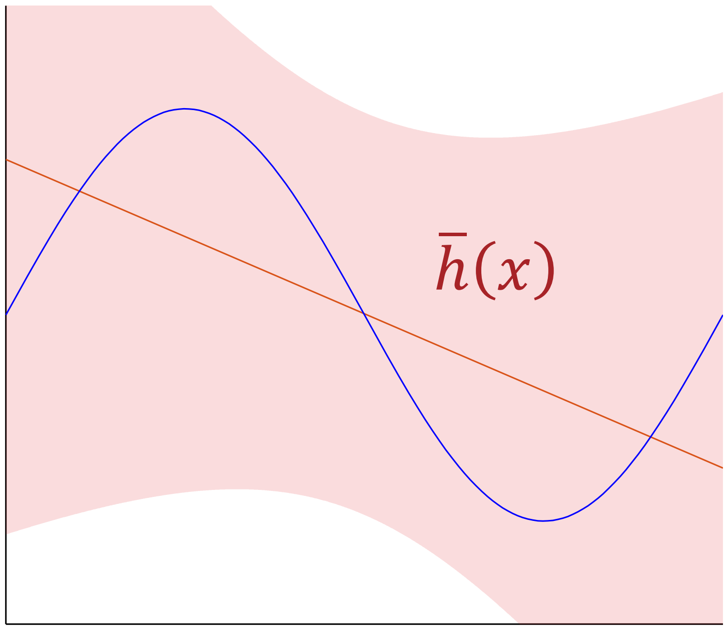
Bias-Variance Tradeoff: Example



Bias of $\bar{h}(x) \approx 0.50$

Variance of $h_S(x) \approx 0.25$

$\mathbb{E}_S[err_D(h_S)] \approx 0.75$

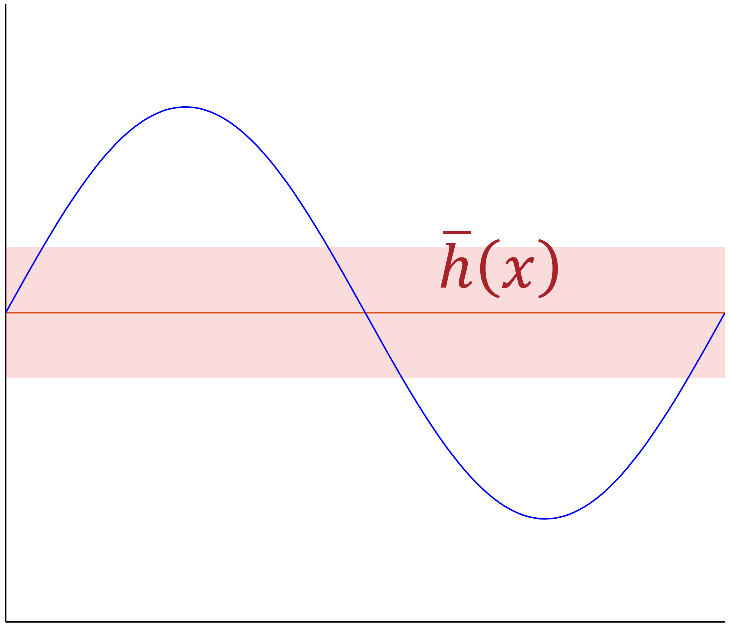


Bias of $\bar{h}(x) \approx 0.21$

Variance of $h_S(x) \approx 1.74$

$\mathbb{E}_S[err_D(h_S)] \approx 1.95$

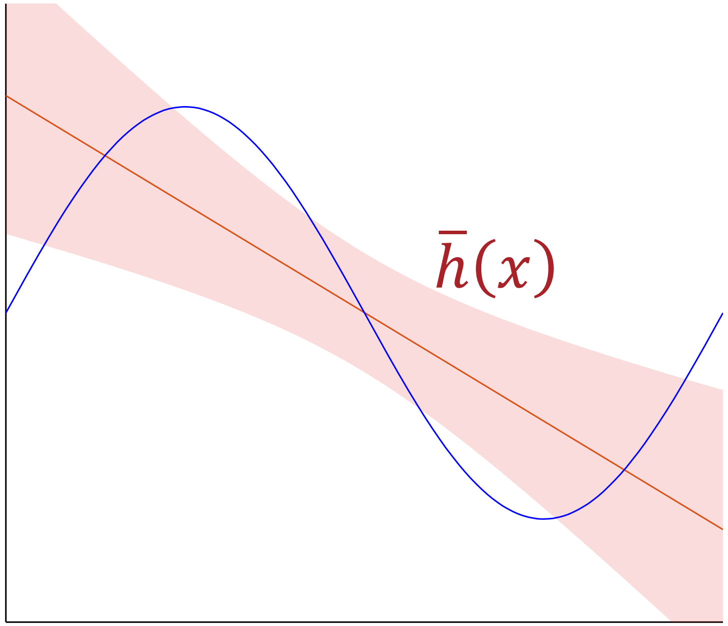
Bias-Variance Tradeoff: Example $(N = 5)$



Bias of $\bar{h}(x) \approx 0.50$

Variance of $h_S(x) \approx 0.10$

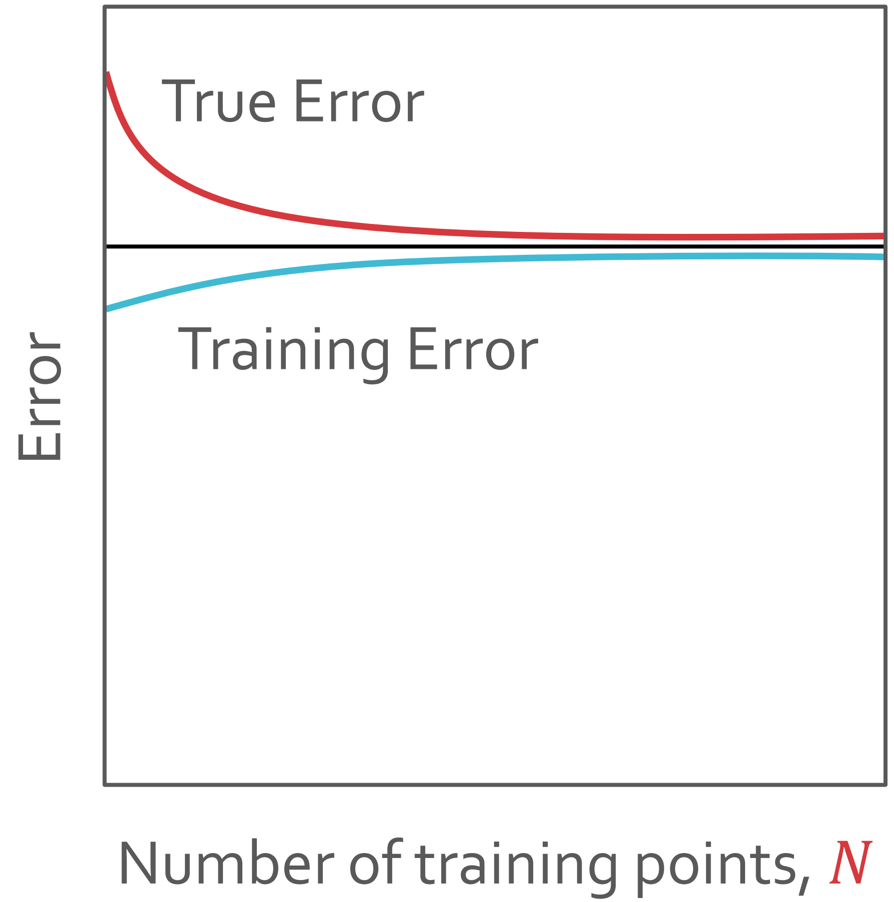
$\mathbb{E}_S[err_D(h_S)] \approx 0.60$

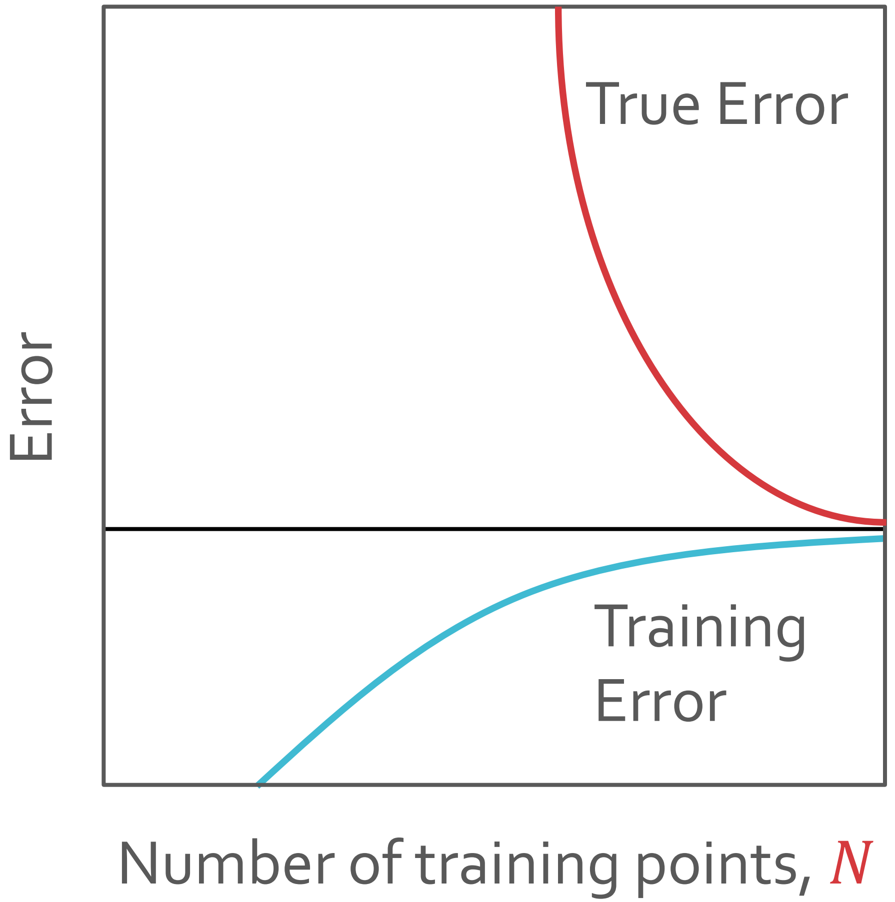


Bias of $\bar{h}(x) \approx 0.21$

Variance of $h_S(x) \approx 0.21$

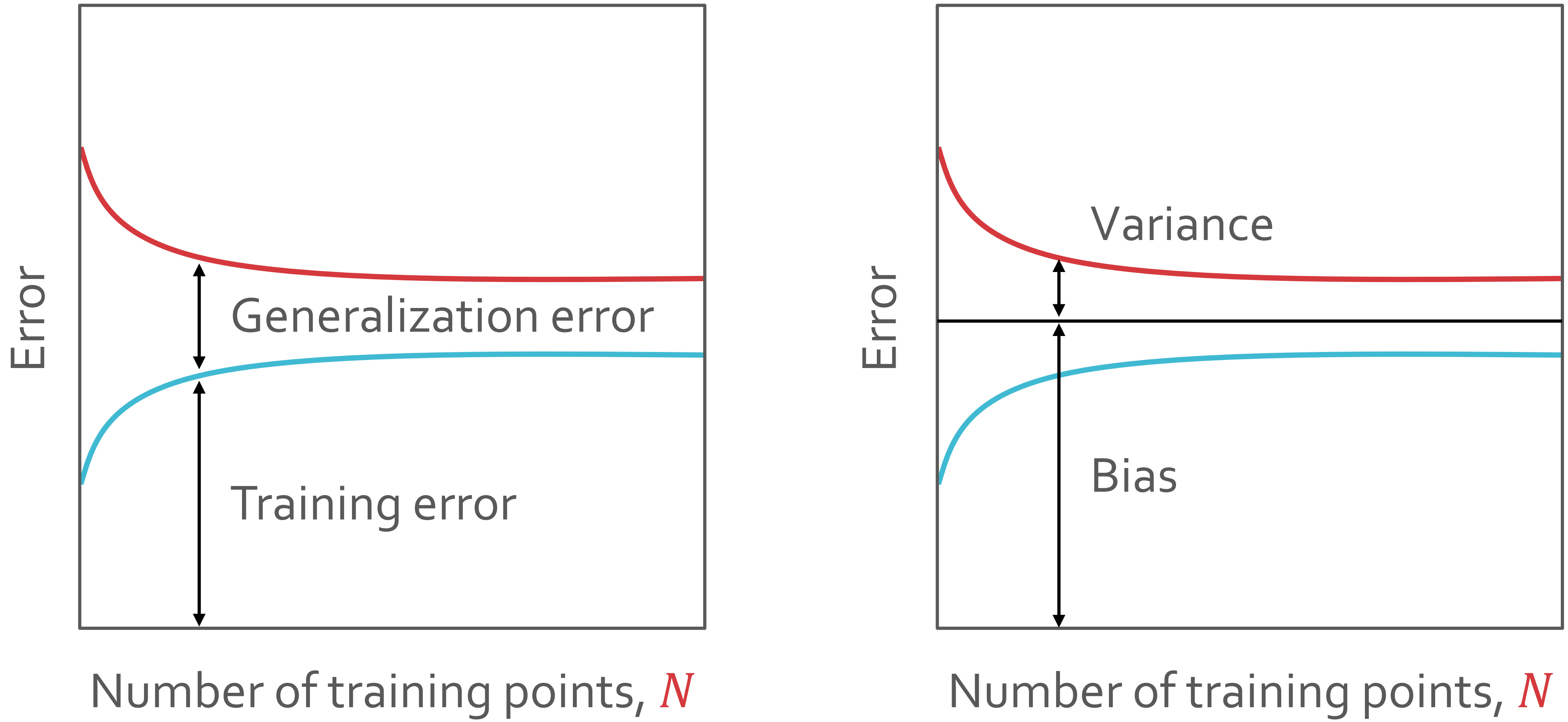
$\mathbb{E}_S[err_D(h_S)]$ $\boxed{\approx 0.42}$





Simple model

Complex model



VC analysis

Bias-Variance analysis