10-701: Introduction to Machine Learning Lecture 20 – Learning Theory

Pradeep Ravikumar & Geoff Gordon

Spring 2026

Front Matter

- Announcements
  - Project check-in is due next Wednesday, April 8.
  - You should complete a second 15-30 minute meeting with your project mentor before that date (April 8)
    - It is recommended to bring a (hopefully complete) draft of check-in to that meeting.

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

Recall: Types of Error

- True error rate
  - Actual quantity of interest in machine learning
  - How well your hypothesis will perform on average across all possible data points
- Test error rate: used to evaluate hypothesis performance
  - Good estimate of the true error rate
- Validation error rate: used to set model hyperparameters
  - Slightly “optimistic” estimate of the true error rate
- Training error rate: used to set model parameters
  - Very “optimistic” estimate of the true error rate

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

- Expected risk of a hypothesis h (a.k.a. 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)

$$\begin{aligned}\hat{R}(h) &= P_{\boldsymbol{x} \sim \mathcal{D}}\left(c^*(\boldsymbol{x}) \neq h(\boldsymbol{x})\right) \\ &= \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left(c^*\left(\boldsymbol{x}^{(n)}\right) \neq h\left(\boldsymbol{x}^{(n)}\right)\right) \\ &= \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left(y^{(n)} \neq h\left(\boldsymbol{x}^{(n)}\right)\right)\end{aligned}$$

where $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)\right\}_{n=1}^{N}$ is the training data set and $\boldsymbol{x} \sim \mathcal{D}$ denotes a point sampled uniformly at random from $\mathcal{D}$

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?

A simple setting: finite possibilities

- Classification
  - n i.i.d. data points (X_i, Y_i), i = 1,…,n
  - **finite** number of possible hypotheses (e.g., decision trees of depth d)
- A learner finds a hypothesis *h*
- We are interested in:

$$\text{error}_{\text{train}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(h(X_i) \neq Y_i)$$

$$\text{error}_{\text{true}} = \mathbb{P}(h(X) \neq Y)$$

A simple setting

- Classification
  - n i.i.d. data points (X_i, Y_i), i = 1,…,n
  - **finite** number of possible hypotheses (e.g., decision trees of depth d)
- A learner finds a hypothesis *h* that is **consistent** with training data
  - Gets zero error in training, $error_{train}(h) = 0$
- What is the probability that *h* has more than ε true error?
  - $error_{true}(h) \geq \varepsilon$

Even if *h* makes zero errors in training data, may make errors in test

How likely is a bad hypothesis to get m data points right?

- Consider a bad hypothesis *h* i.e. $error_{true}(h) \geq \varepsilon$
- Probability that *h* gets one data point right (i.e. does not make an error) $\leq 1-\varepsilon$
- Probability that *h* gets *m* data points right

$$\leq (1-\varepsilon)^m$$

How likely is a learner to pick a bad hypothesis?

- Usually there are many (say k) bad hypotheses in the class

  $h_1, h_2, \ldots, h_k$ s.t. error(h_i) $\geq \varepsilon$ $\quad$ i = 1, …, k

- Probability that learner picks a bad hypothesis = Probability that some bad hypothesis is consistent with m data points

Prob(h_1 consistent with m data points OR
h_2 consistent with m data points OR … OR
h_k consistent with m data points)

≤ Prob(h_1 consistent with m data points) +
Prob(h_2 consistent with m data points) + □ +
Prob(h_k consistent with m data points)

≤ $k(1-\varepsilon)^m$

Union bound

Loose but works

How likely is a learner to pick a bad hypothesis?

- Usually there are many many (say k) bad hypotheses in the class

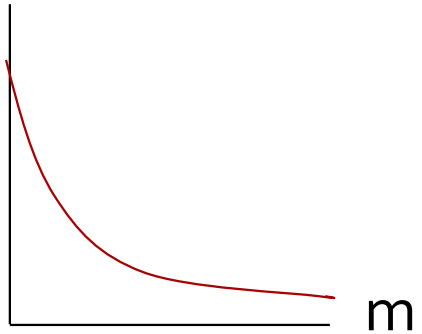
$$h_1, h_2, \ldots, h_k \quad \text{s.t. error}(h_i) \geq \varepsilon \qquad i = 1, \ldots, k$$

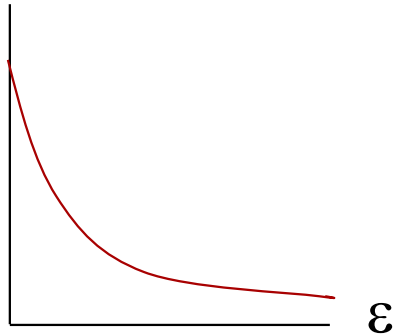
- Probability that learner picks a bad hypothesis

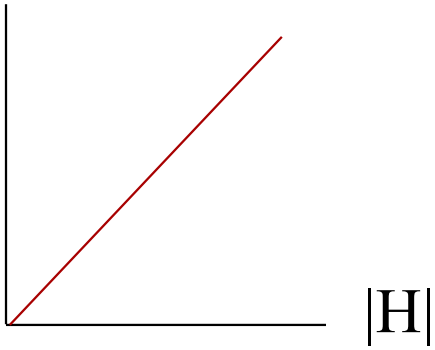
$$\leq \; |H| \, (1-\varepsilon)^m \leq \; |H| \, e^{-\varepsilon m}$$

$$\leq \; k \, (1-\varepsilon)^m$$

↳ Size of hypothesis class







Probability of Error

$|H|e^{-m\epsilon} \leq \delta$ .... Probability of error

- Given ε and δ, yields sample complexity

#training data $m \geq \frac{\ln |H| + \ln \frac{1}{\delta}}{\epsilon}$

PAC (Probably Approximately Correct) bound

- ***Theorem [Haussler'88]***: Hypothesis space *H* finite, dataset *D* with *m* i.i.d. samples, $0 < \varepsilon < 1$: for any learned hypothesis *h* that is consistent on the training data, for <u>sufficiently large</u> m:

$$P(\text{error}_{true}(h) > \epsilon) \leq |H|e^{-m\epsilon} \leq \delta$$

- Equivalently, with probability $\geq 1 - \delta$

$$\text{error}_{true}(h) \leq \epsilon$$

What if our classifier does not have zero error on the training data?

- Question: What about a learner with $error_{train}(h) \neq 0$ in training set?
- The error of a hypothesis is like estimating the parameter of a coin!

$$error_{true}(h) := P(h(X) \neq Y) \quad \equiv \quad P(H=1) =: \theta$$

$$error_{train}(h) := \frac{1}{m} \sum_i \mathbf{1}_{h(X_i) \neq Y_i} \quad \equiv \frac{1}{m} \sum_i Z_i =: \widehat{\theta}$$

Hoeffding's bound for a single hypothesis

- Consider *m* i.i.d. flips $x_1,\ldots,x_m$, where $x_i \in \{0,1\}$ of a coin with parameter θ. For $0<\varepsilon<1$:

$$P\left(\left|\theta - \frac{1}{m}\sum_i x_i\right| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

- For a single hypothesis h

$$P\left(|\text{error}_{true}(h) - \text{error}_{train}(h)| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

Hoeffding's bound for |H| hypotheses

- For each hypothesis h_i:

$$P\left(|\text{error}_{true}(h_i) - \text{error}_{train}(h_i)| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

- What if we are comparing |H| hypotheses?

  Union bound

- ***Theorem***: Hypothesis space *H* finite, dataset *D* with *m* i.i.d. samples, $0 < \varepsilon < 1$: for any learned hypothesis $h \in H$, with sufficiently large number of samples m:

$$P\left(|\text{error}_{true}(h) - \text{error}_{train}(h)| > \epsilon\right) \leq 2|H|e^{-2m\epsilon^2} \leq \delta$$

Summary of PAC bounds for finite hypothesis spaces

With probability $\geq 1-\delta$,

1) For all $h \in H$ s.t. $error_{train}(h) = 0$,

$$error_{true}(h) \leq \varepsilon = \frac{\ln |H| + \ln \frac{1}{\delta}}{m}$$

Haussler's bound

2) For all $h \in H$

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}}$$

Hoeffding's bound

PAC bound and Bias-Variance tradeoff

- With probability $\geq 1 - \delta$

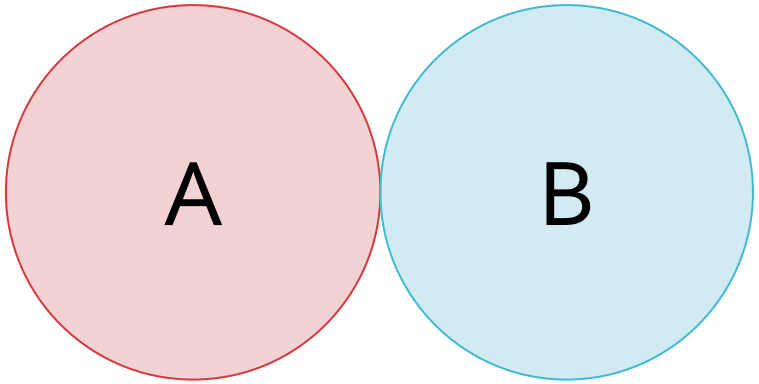
$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}$$

- Fixed m

hypothesis space		
complex	small	large
simple	large	small

The Union Bound…

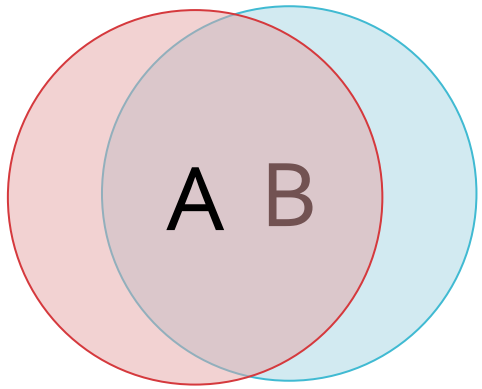
$$P\{A \cup B\} \leq P\{A\} + P\{B\}$$



The Union Bound is Bad!

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

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

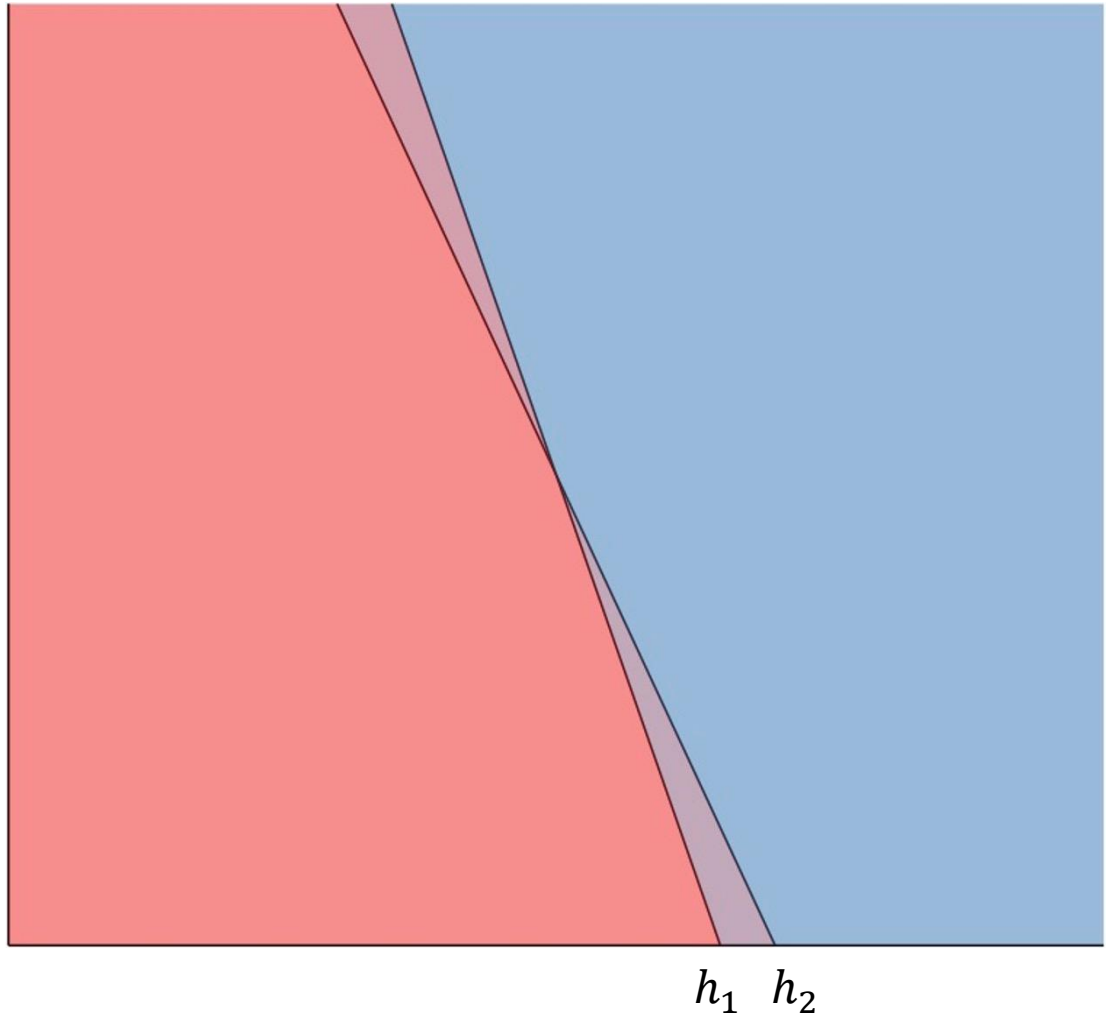


Intuition

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!

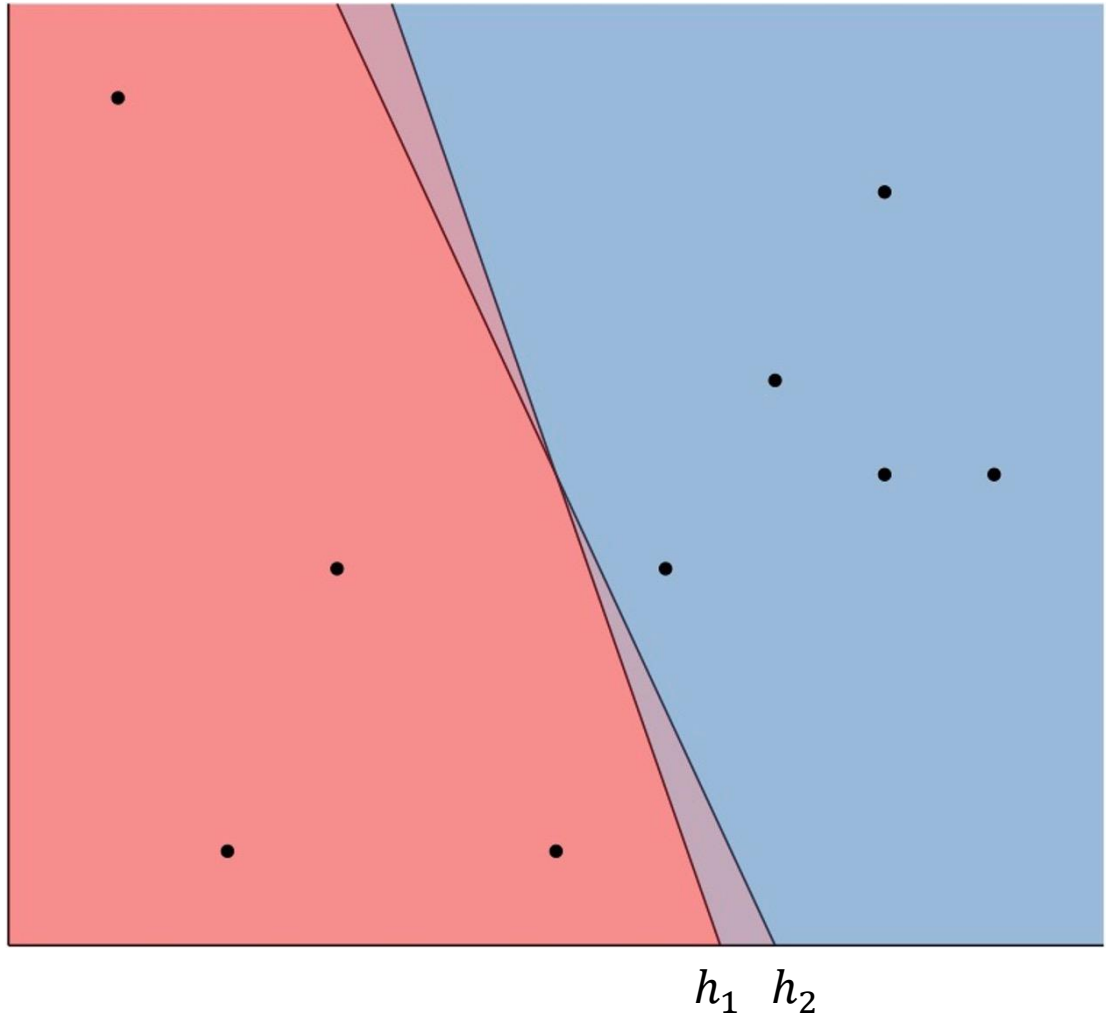


Intuition

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!



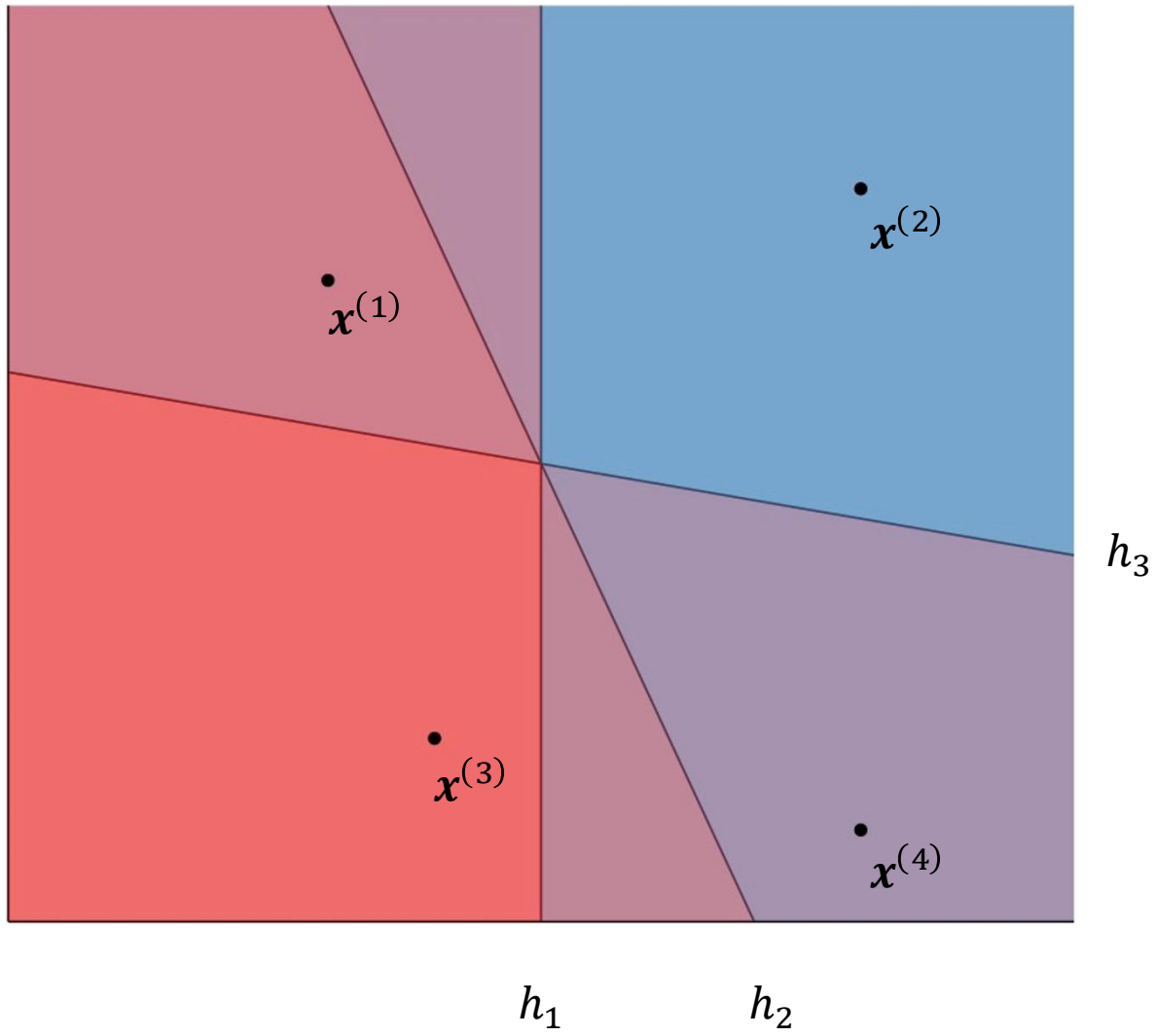
Labelings

- 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

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

Example: Labelings

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

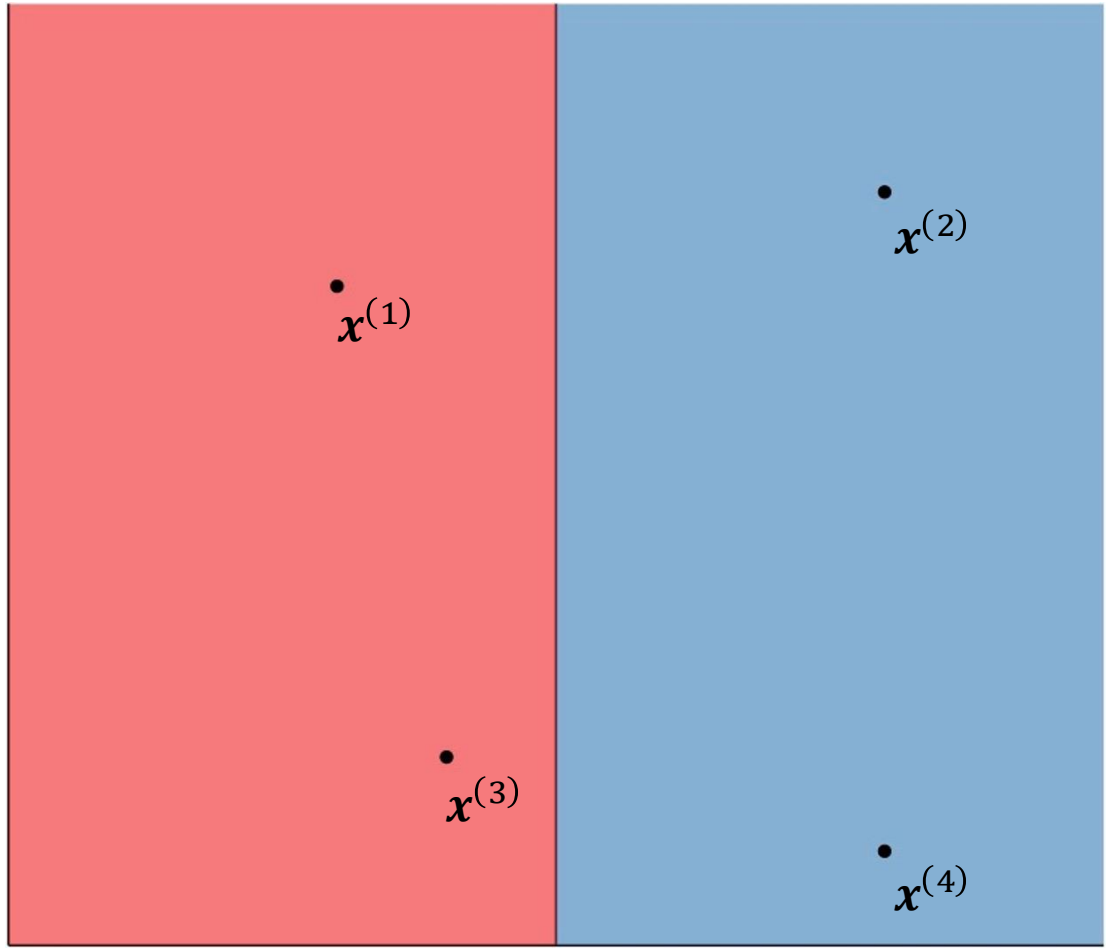


Example: Labelings

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

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

$= (-1, +1, -1, +1)$



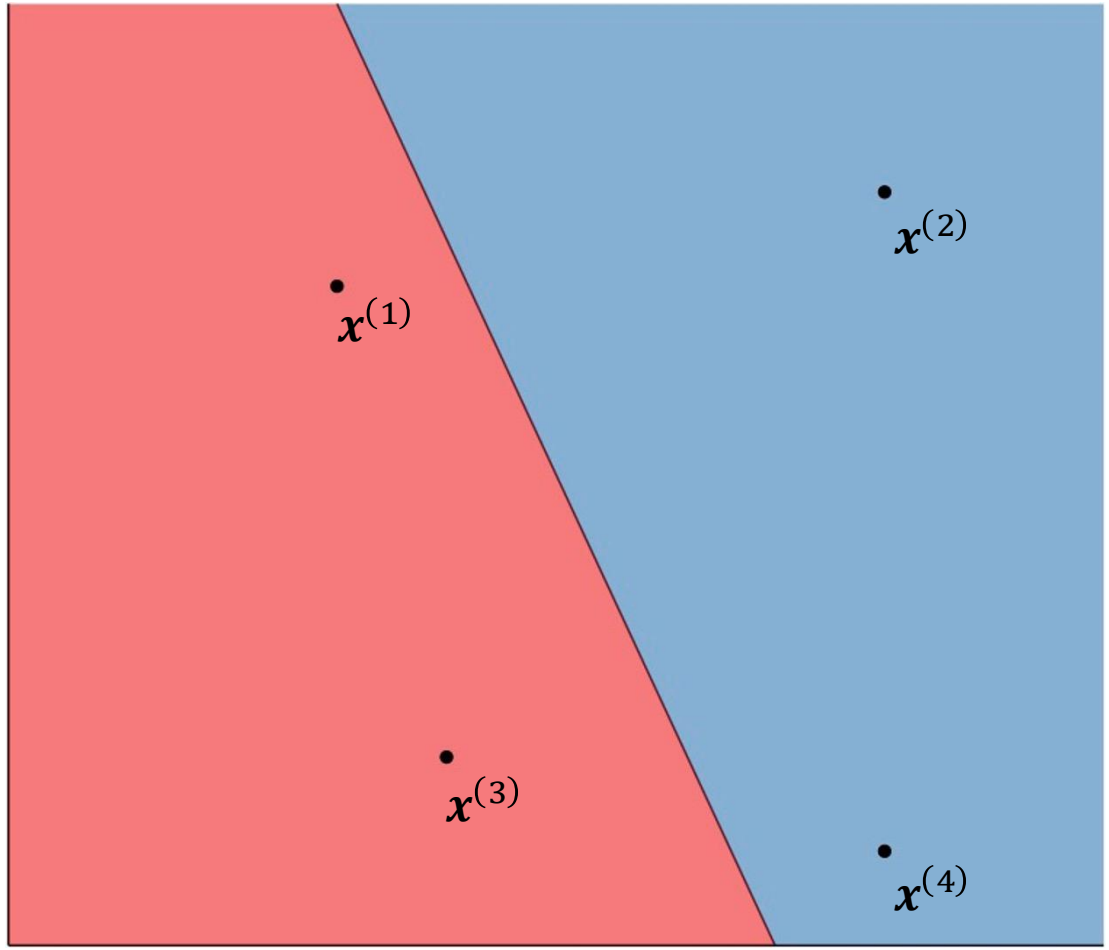
h_1

Example: Labelings

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

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

$= (-1, +1, -1, +1)$



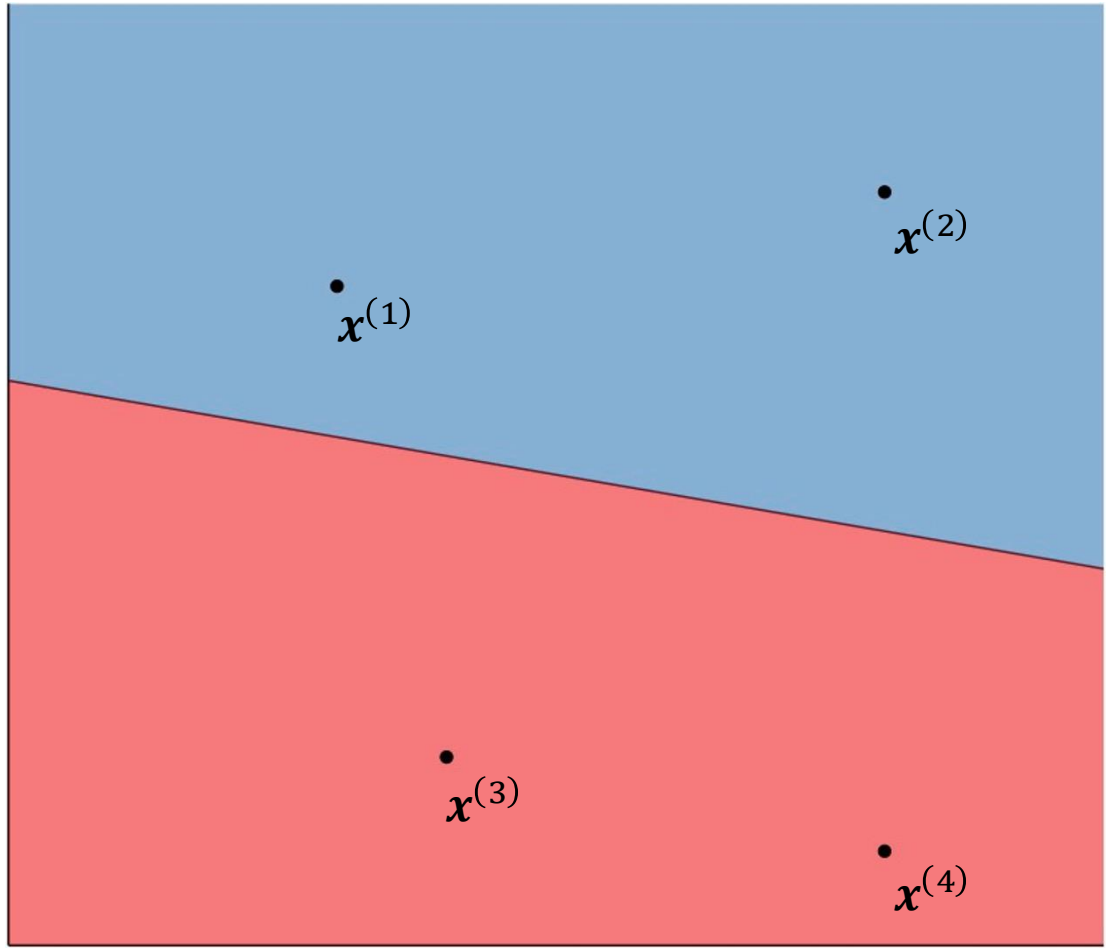
h_2

Example: Labelings

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

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

$= (+1, +1, -1, -1)$

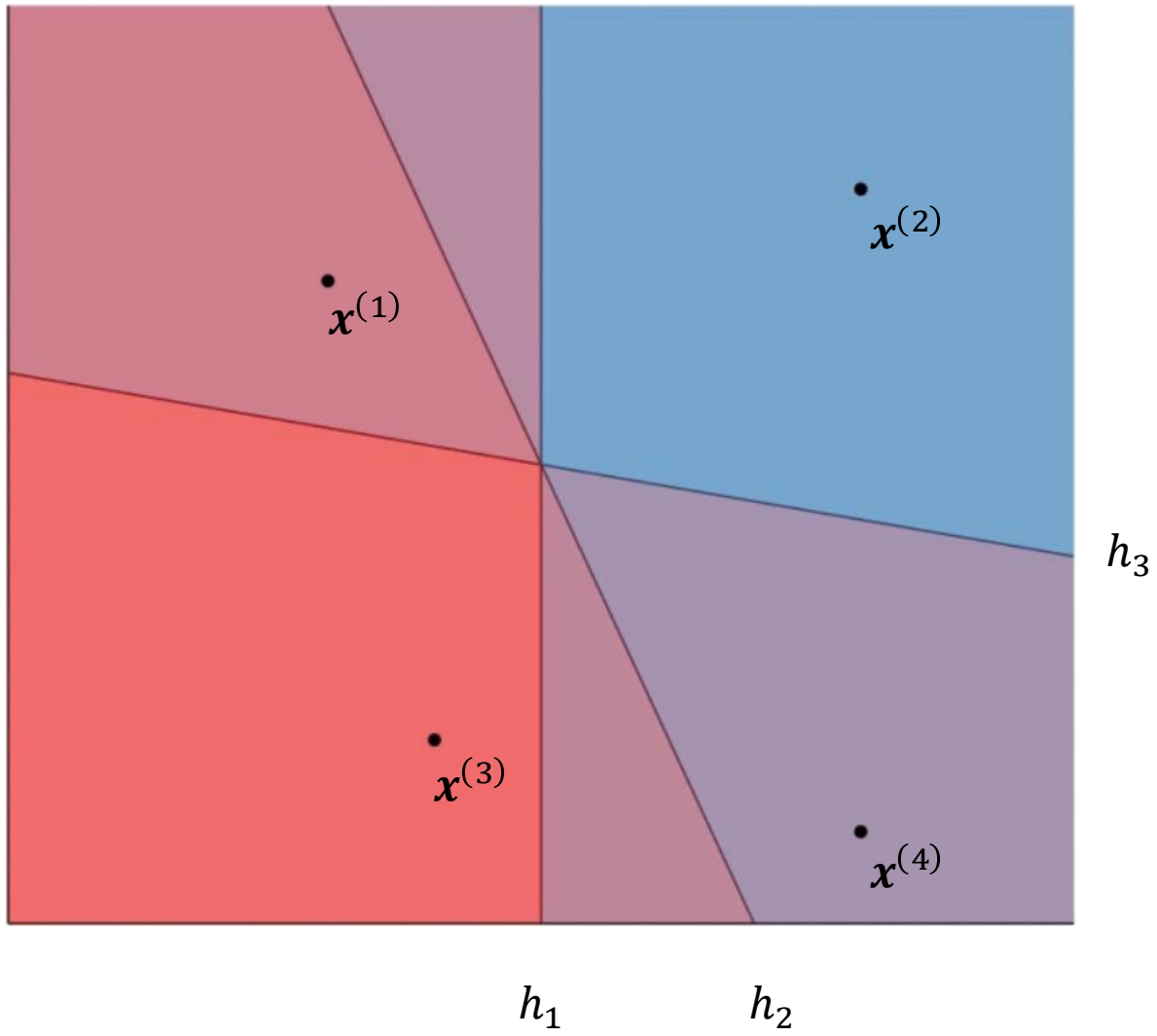


Example: Labelings

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

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

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

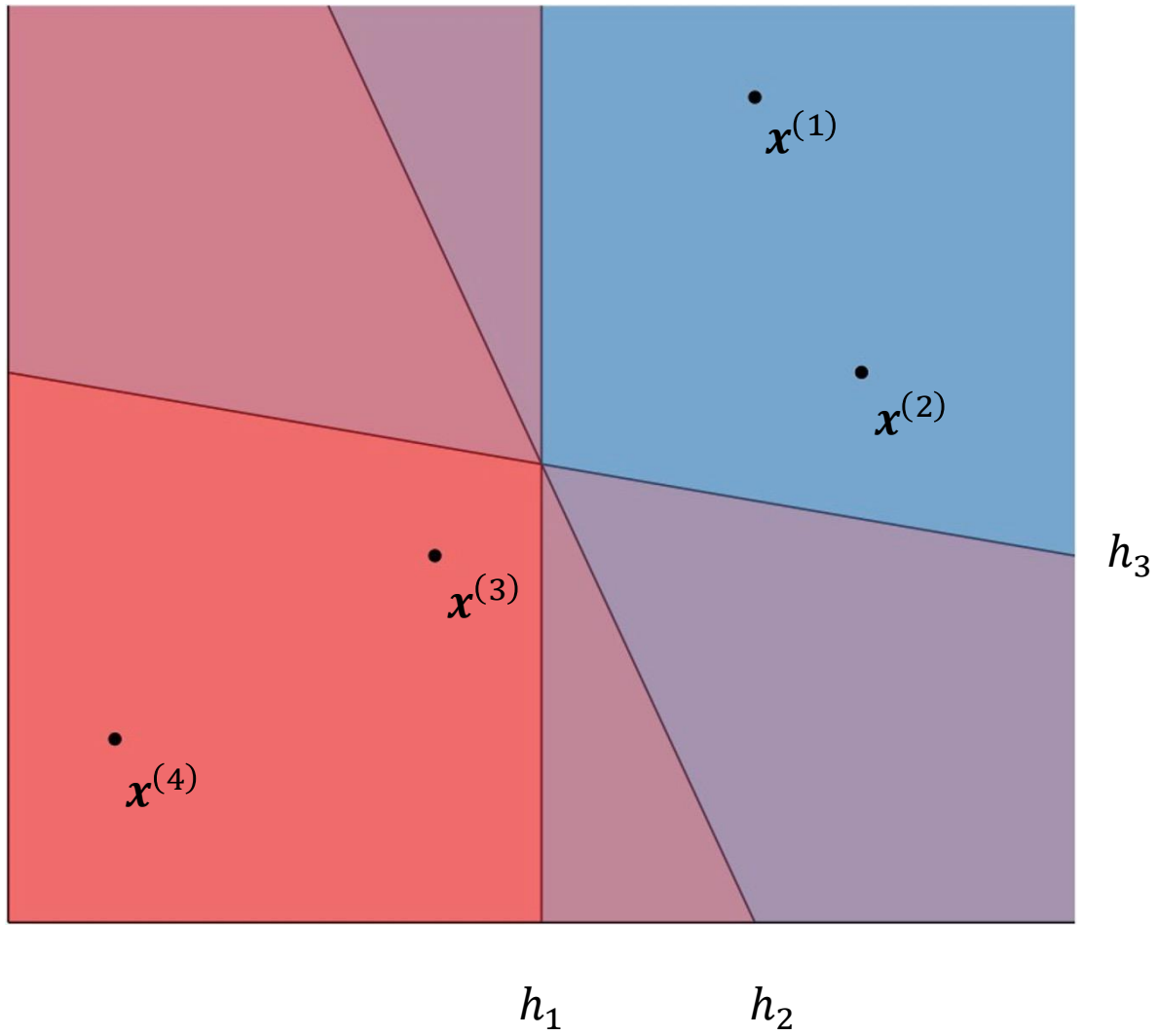


Example: Labelings

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

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

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



Key Takeaways

- Statistical learning theory model
- Expected vs. empirical risk of a hypothesis
- Four possible cases of interest
  - realizable vs. agnostic
  - finite vs. infinite
- Sample complexity bounds and statistical learning theory corollaries for finite hypothesis sets

Growth Function

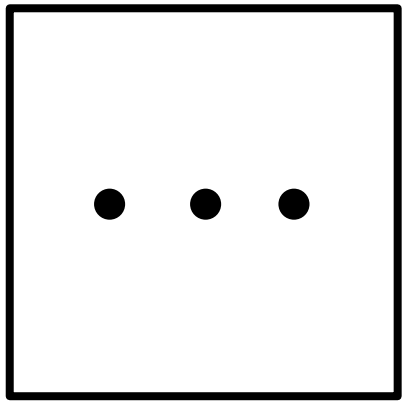
- The **growth function** of $\mathcal{H}$ is the maximum number of distinct labelings $\mathcal{H}$ can induce on ***any*** set of M data points:

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

- $g_{\mathcal{H}}(M) \leq 2^M \;\forall\; \mathcal{H}$ 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 $g_{\mathcal{H}}(M) = 2^M$

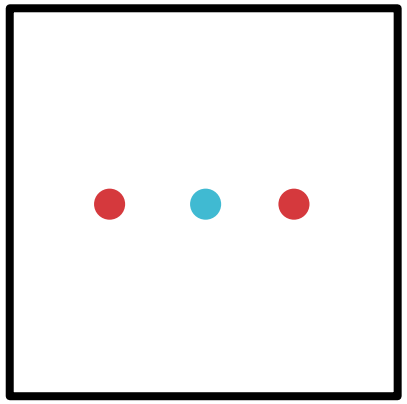
Growth Function: Example

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



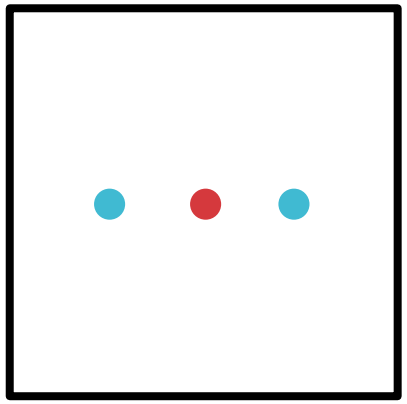
Growth Function: Example

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



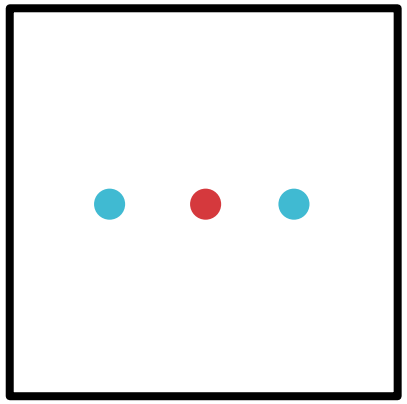
Growth Function: Example

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

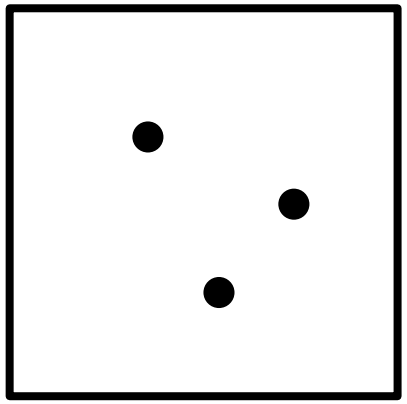


Growth Function: Example

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



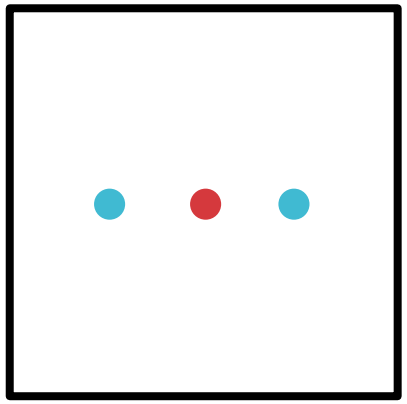
$|\mathcal{H}(S_1)| = 6$



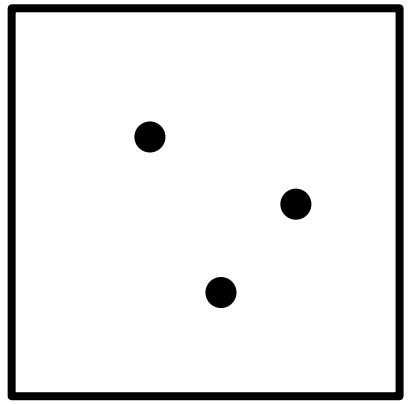
$|\mathcal{H}(S_2)| = 8$

Growth Function: Example

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



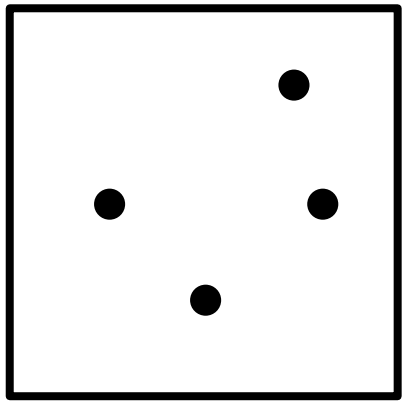
$|\mathcal{H}(S_1)| = 6$



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

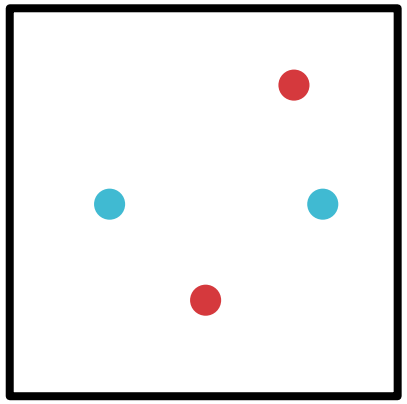
Growth Function: Example

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



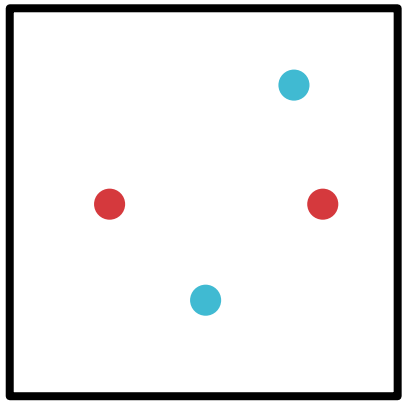
Growth Function: Example

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



Growth Function: Example

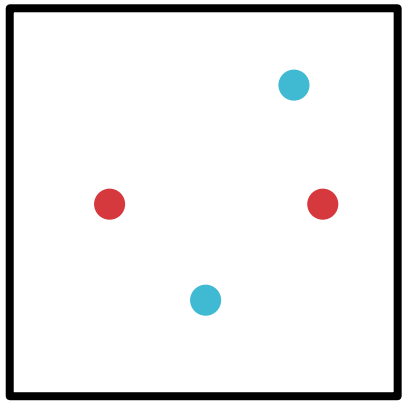
- $\boldsymbol{x}^{(m)} \in \mathbb{R}^2$ and $\mathcal{H} =$ all 2-dimensional linear separators
- What is $g_{\mathcal{H}}(4)$?

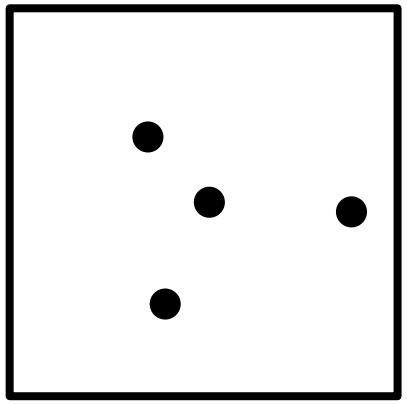


$$|\mathcal{H}(S_1)| = 14$$

Growth Function: Example

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

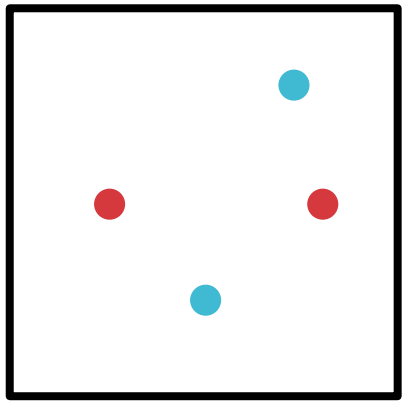


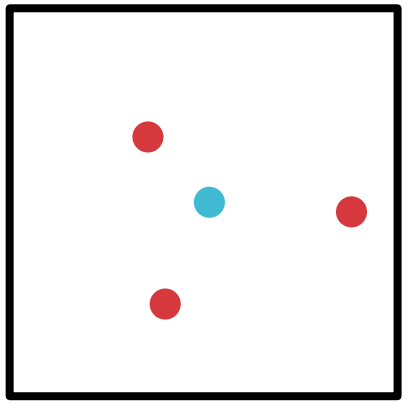


$$|\mathcal{H}(S_1)| = 14$$

Growth Function: Example

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

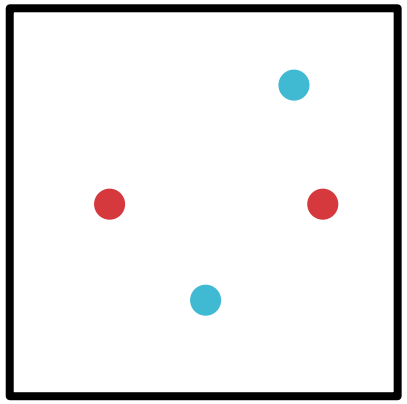


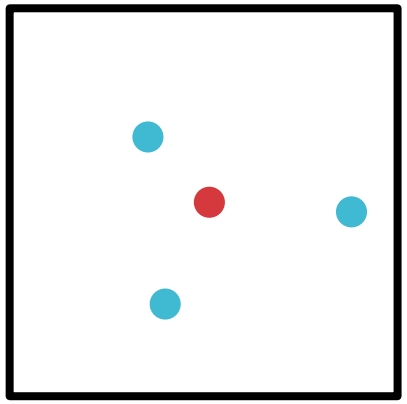


$$|\mathcal{H}(S_1)| = 14$$

Growth Function: Example

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

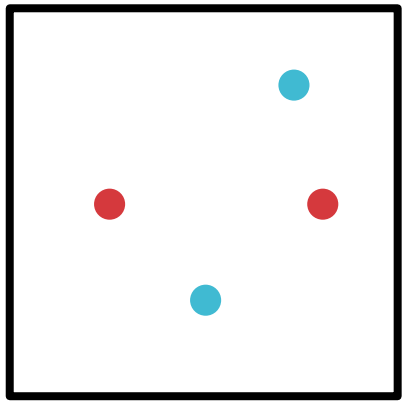




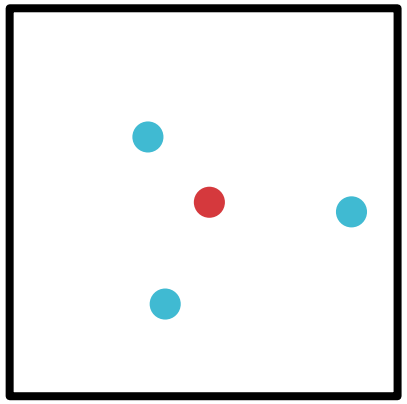
$$|\mathcal{H}(S_1)| = 14$$

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 $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$

M appears on both sides of the inequality…

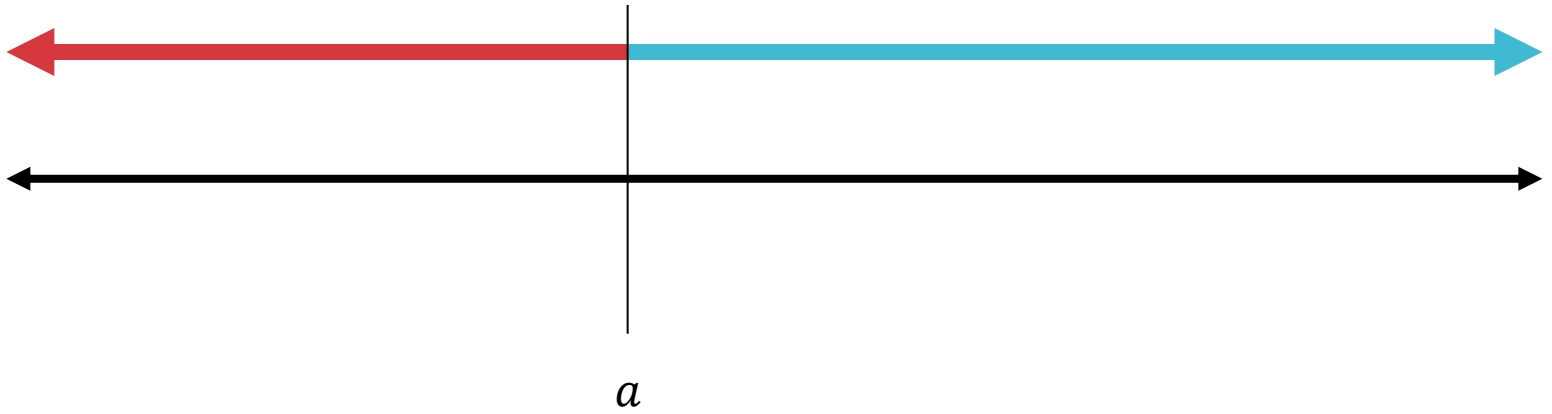
Theorem 3: Vapnik-Chervonenkis (VC)-Dimension

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

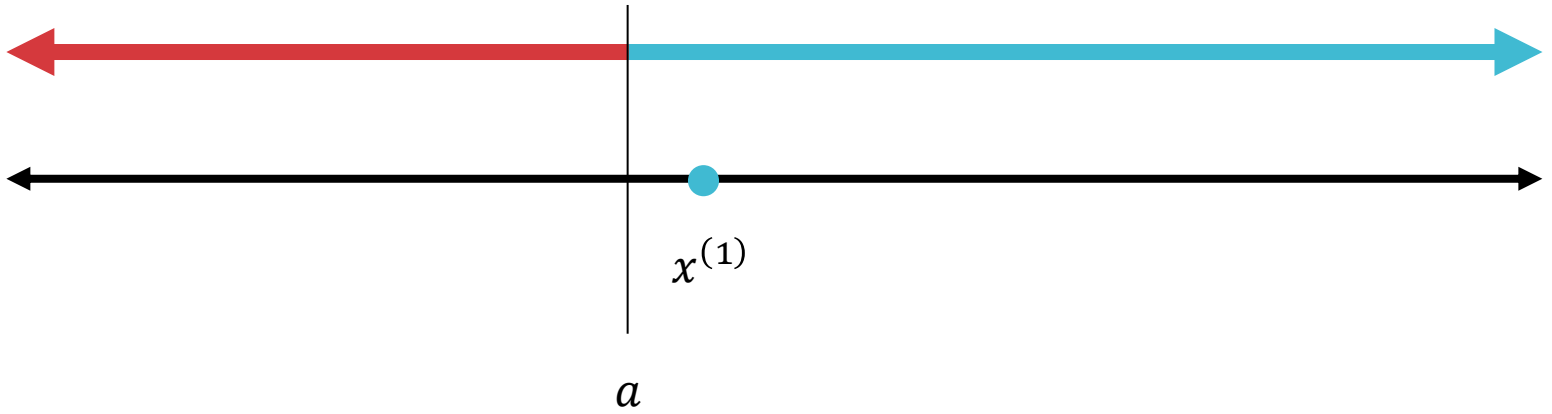
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

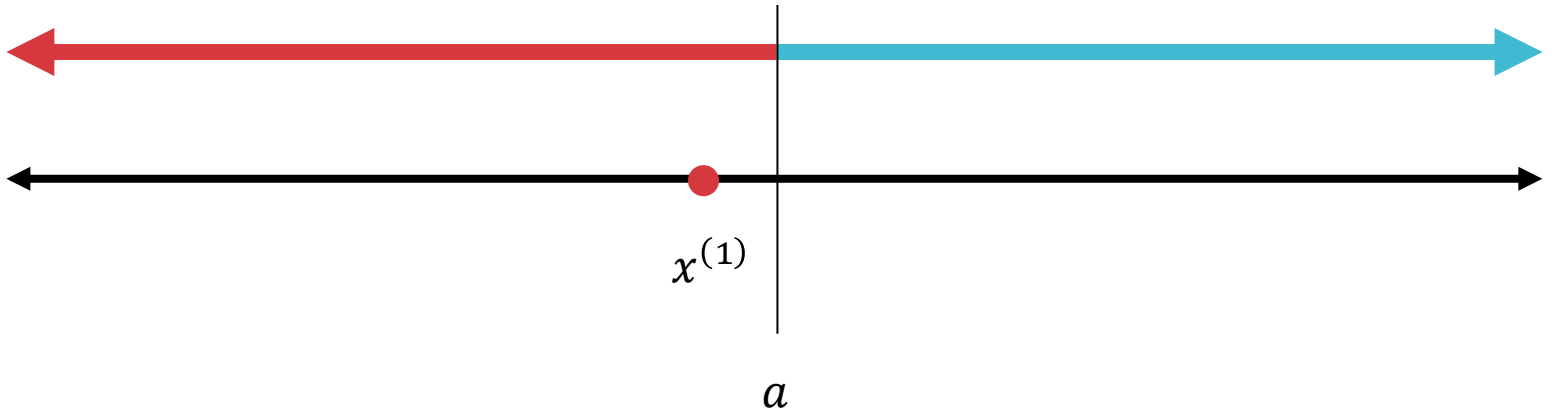
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

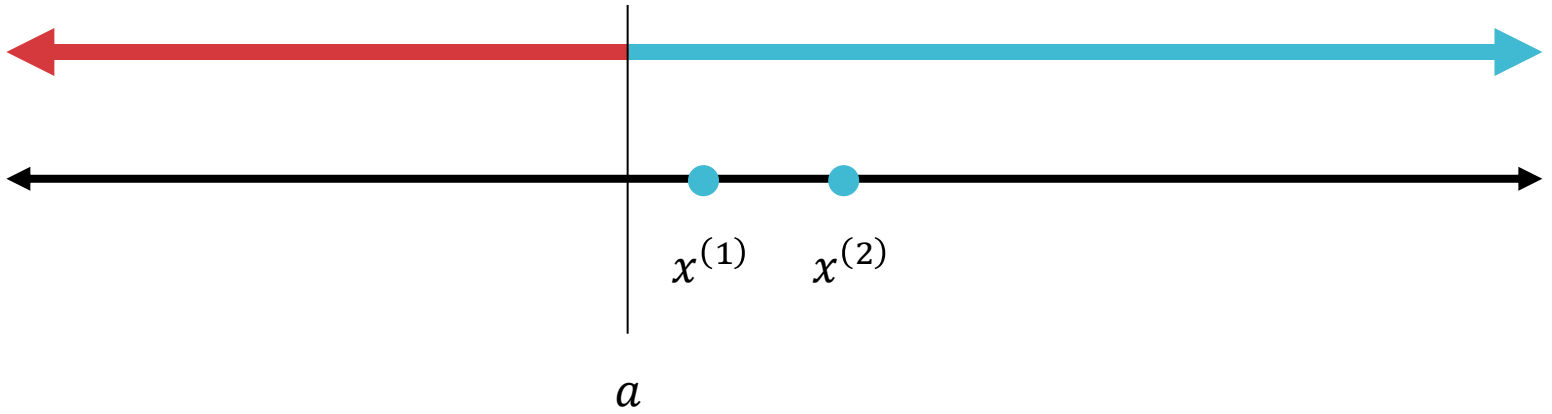
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

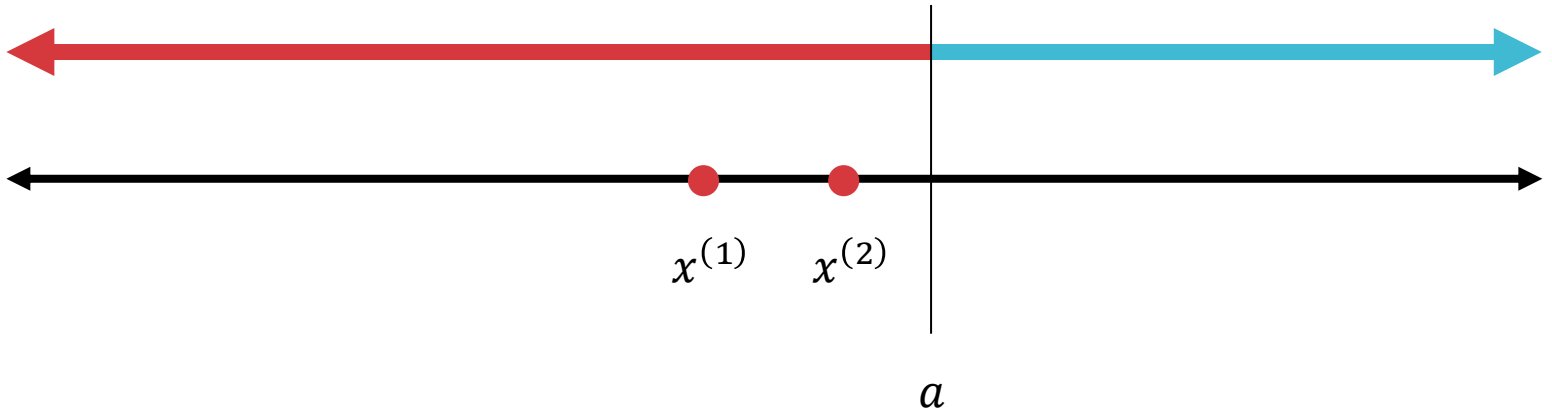
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

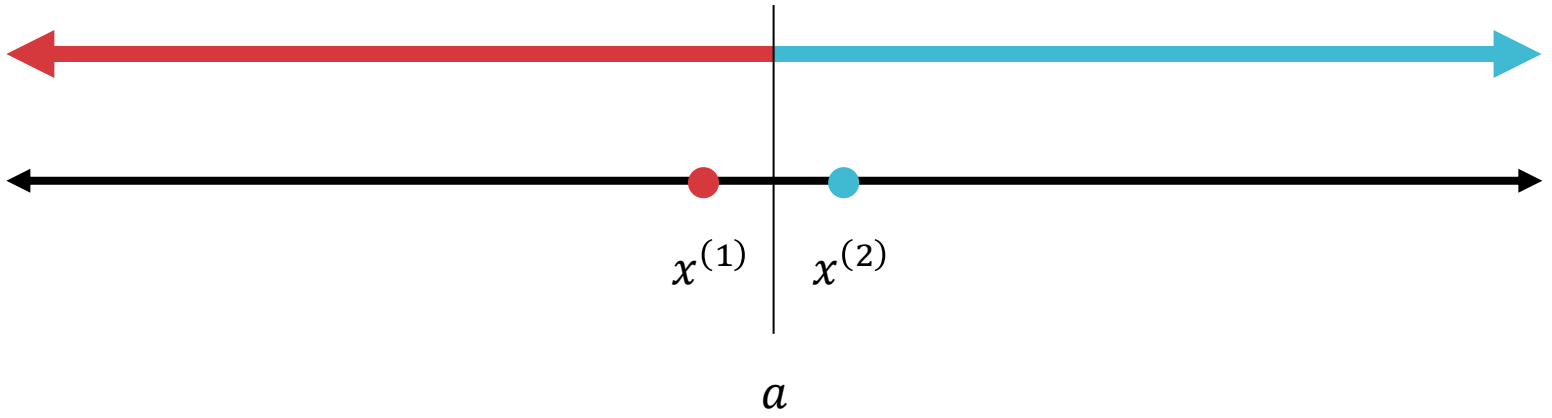
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

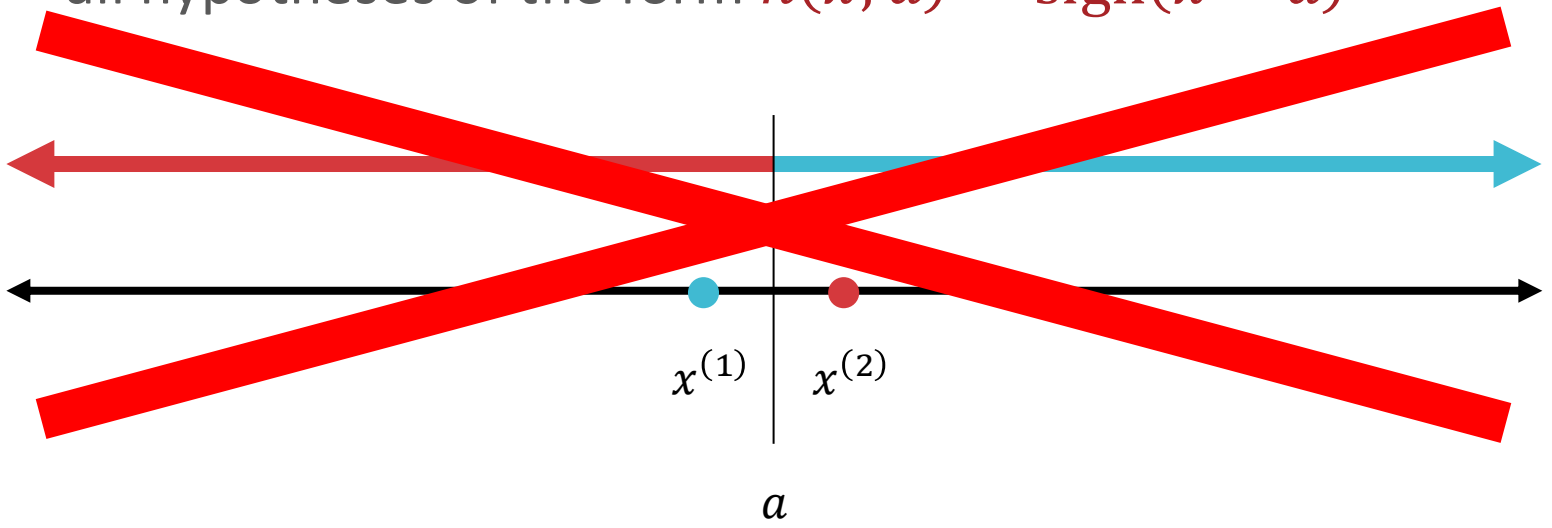
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

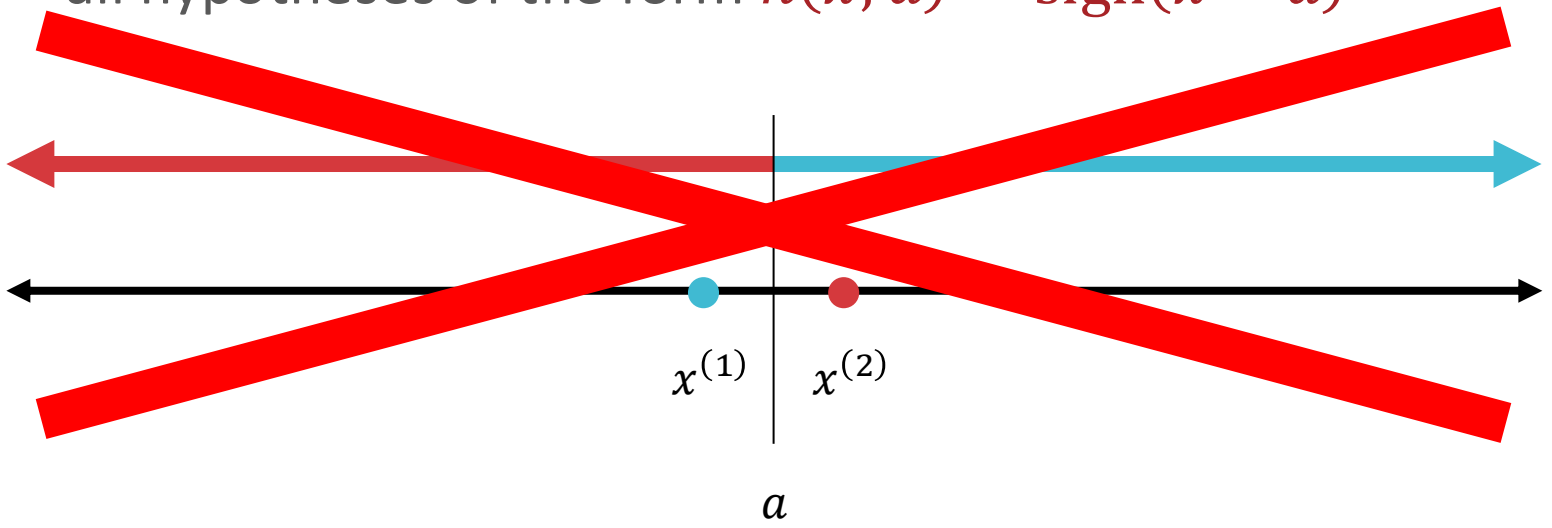
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

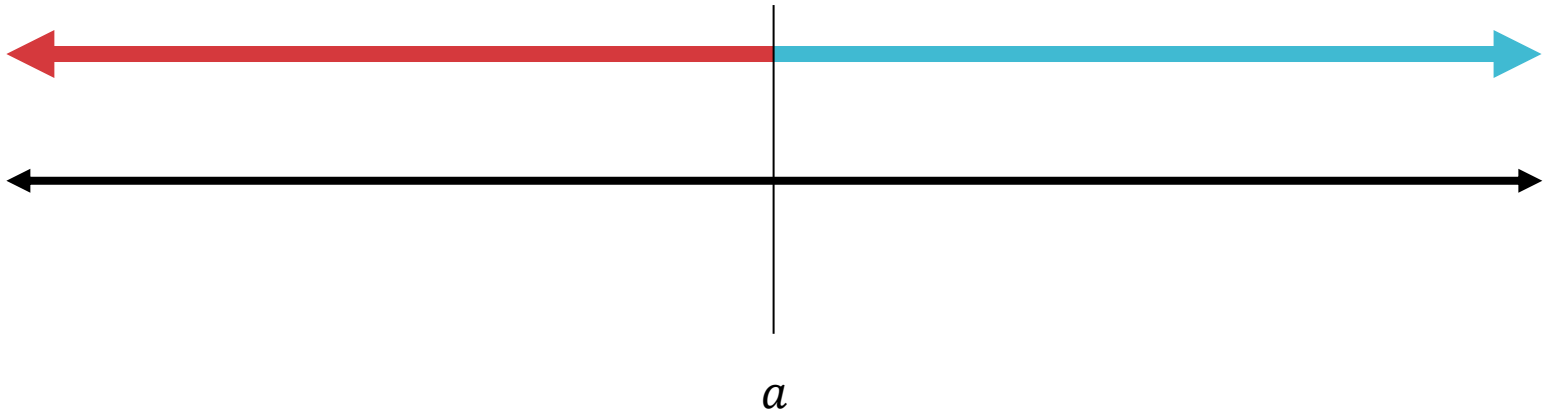
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

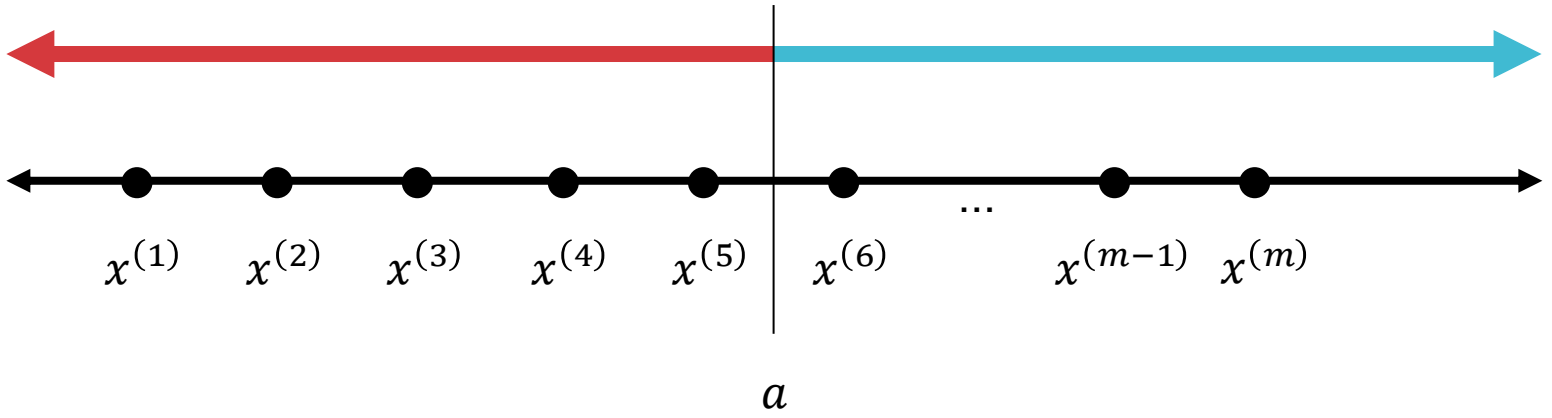
- $x^{(i)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

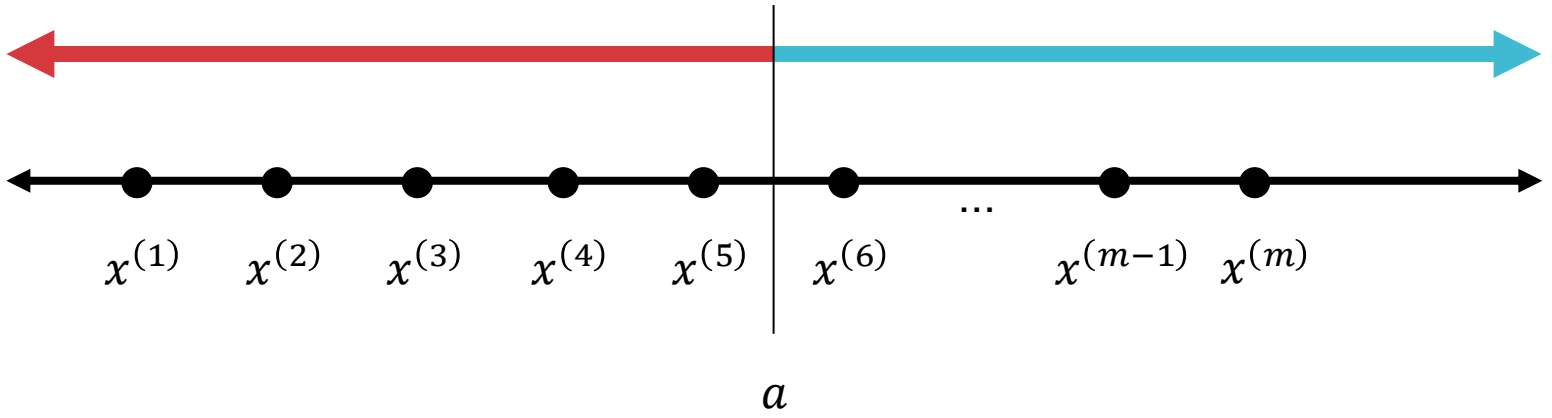
- $x^{(i)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

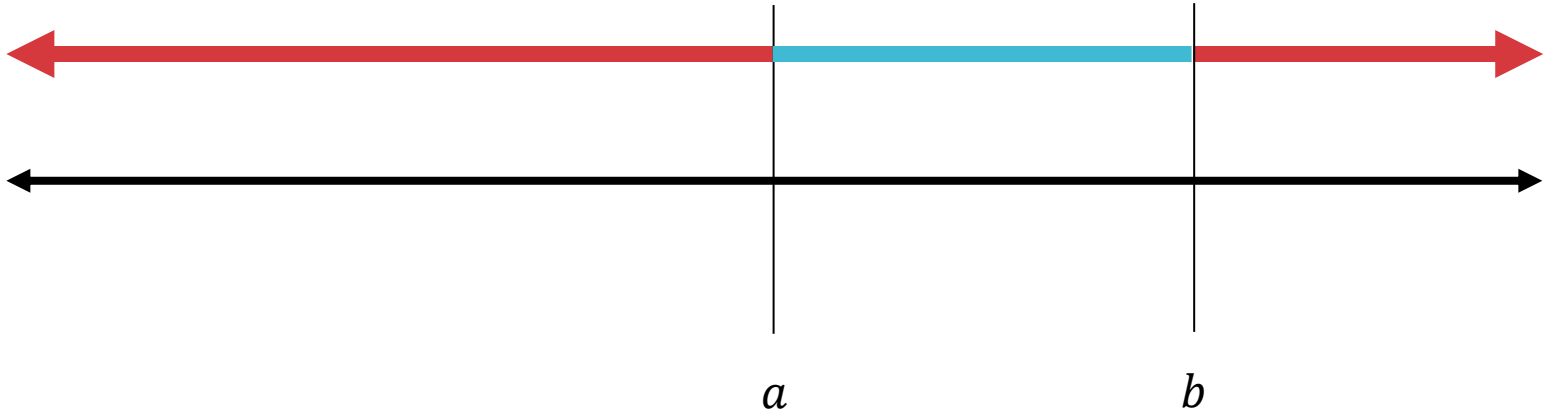
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H}$ = all 1-dimensional positive rays, i.e., all hypotheses of the form $h(x; a) = \text{sign}(x - a)$



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

VC-Dimension: Example

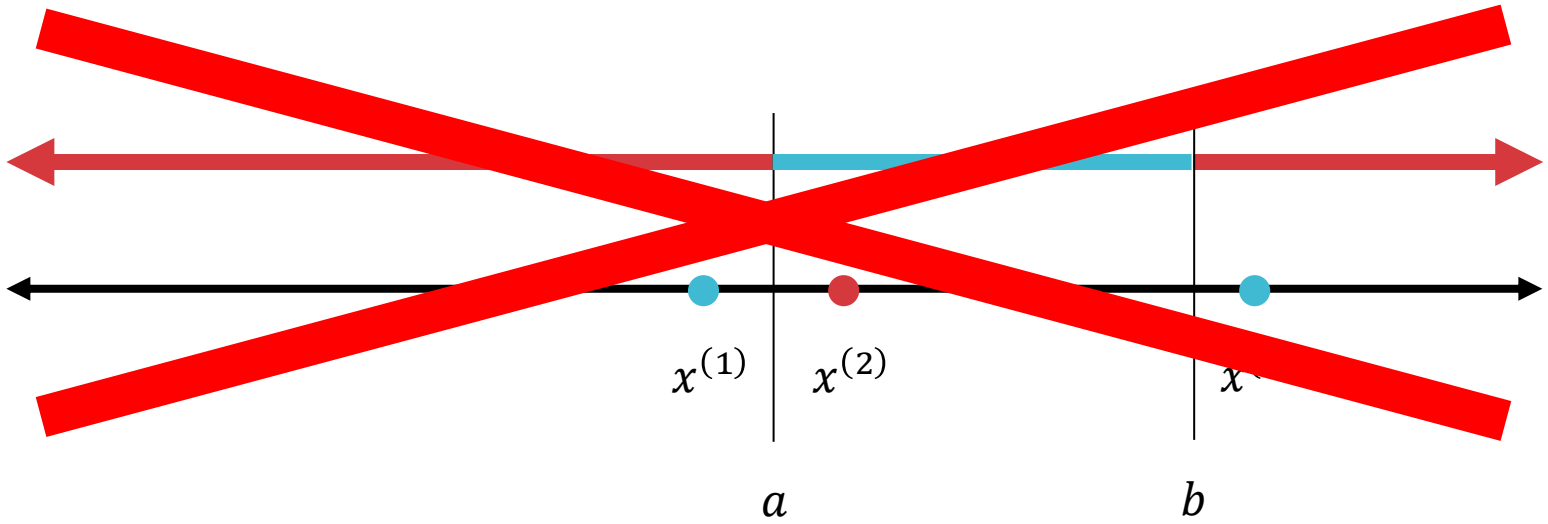
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive intervals



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

VC-Dimension: Example

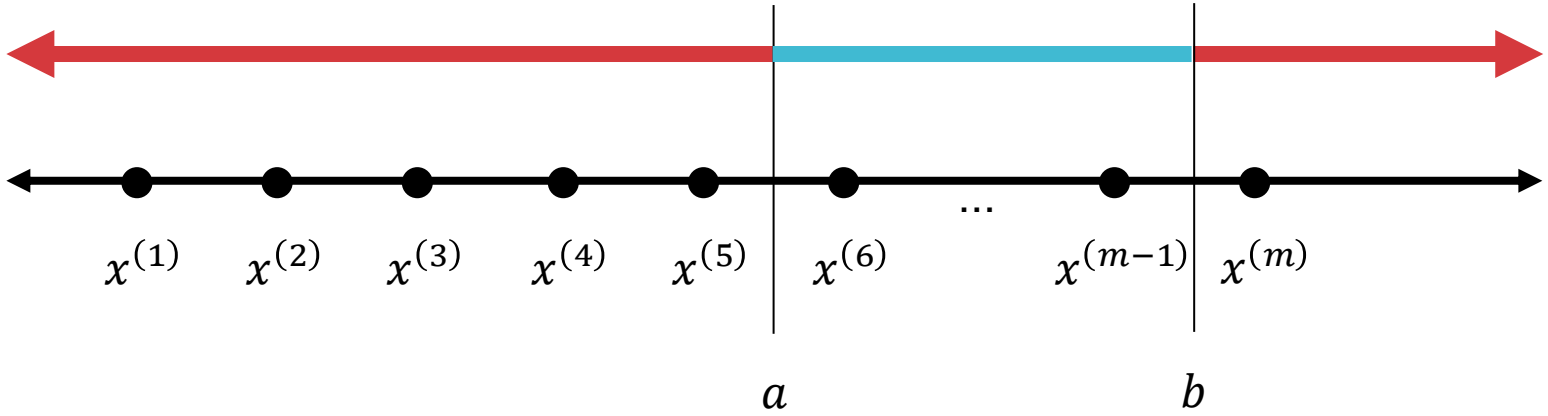
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive intervals



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

VC-Dimension: Example

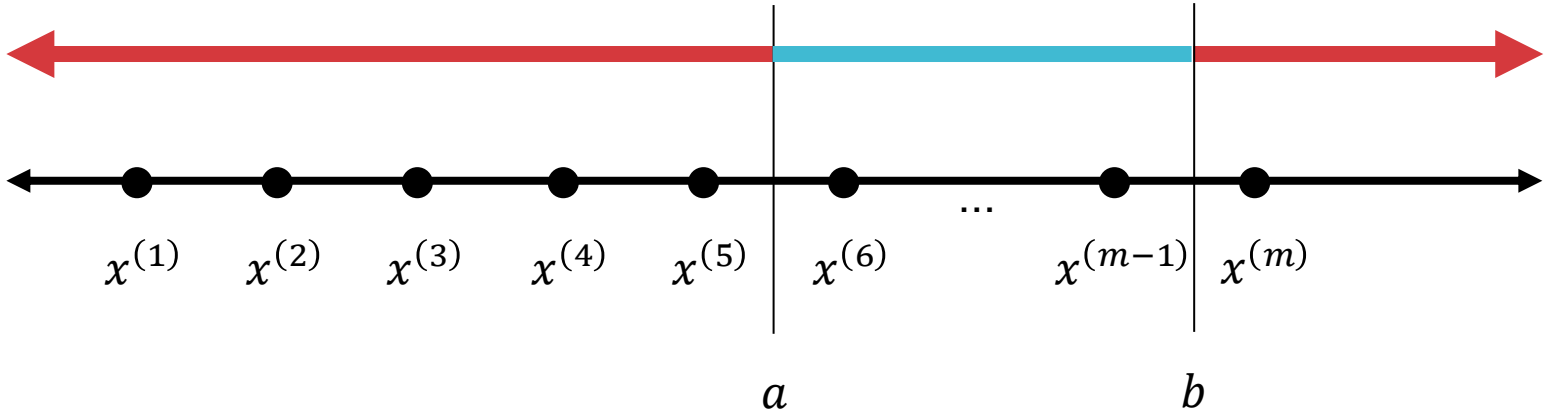
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive intervals



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

VC-Dimension: Example

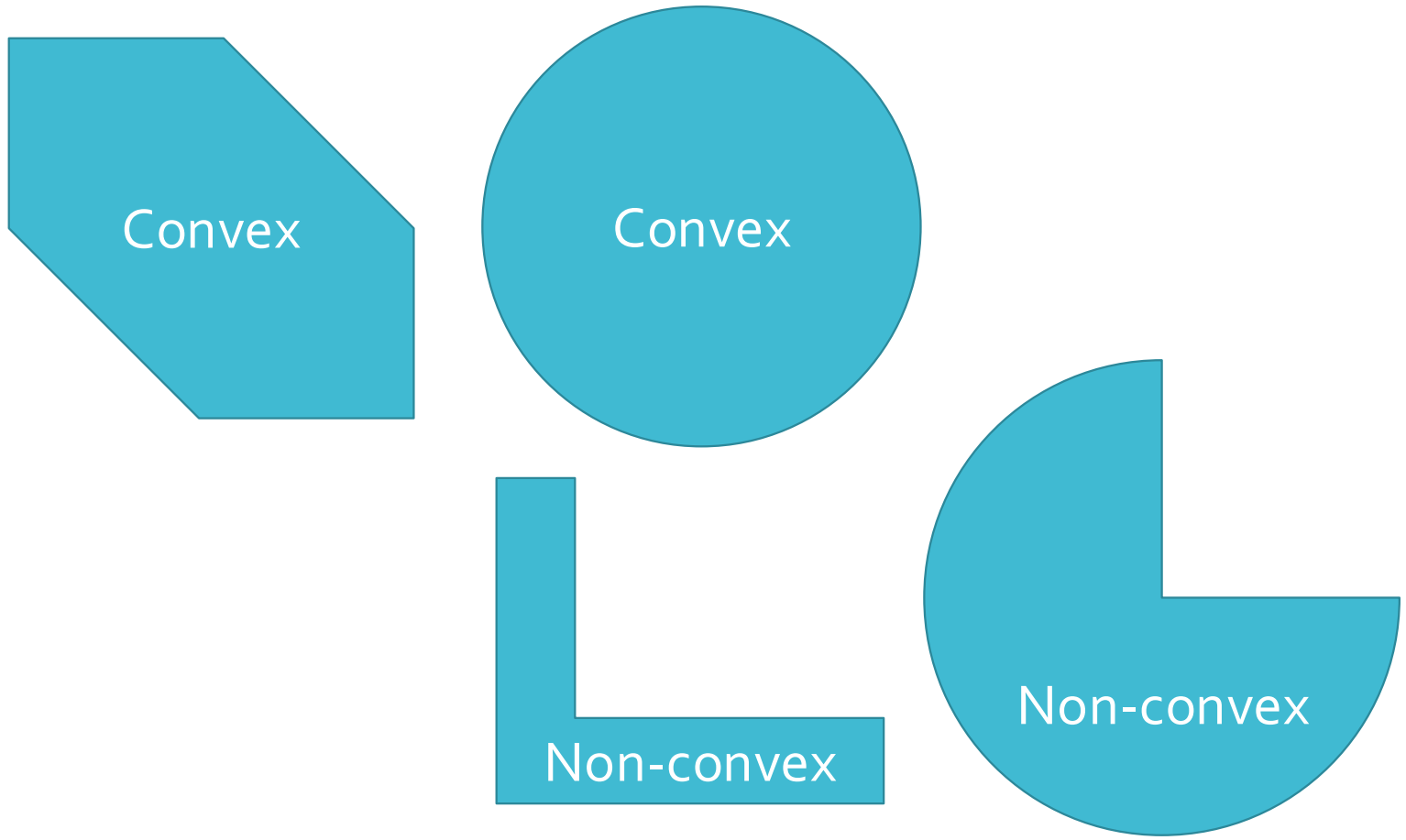
- $x^{(m)} \in \mathbb{R}$ and $\mathcal{H} =$ all 1-dimensional positive intervals



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

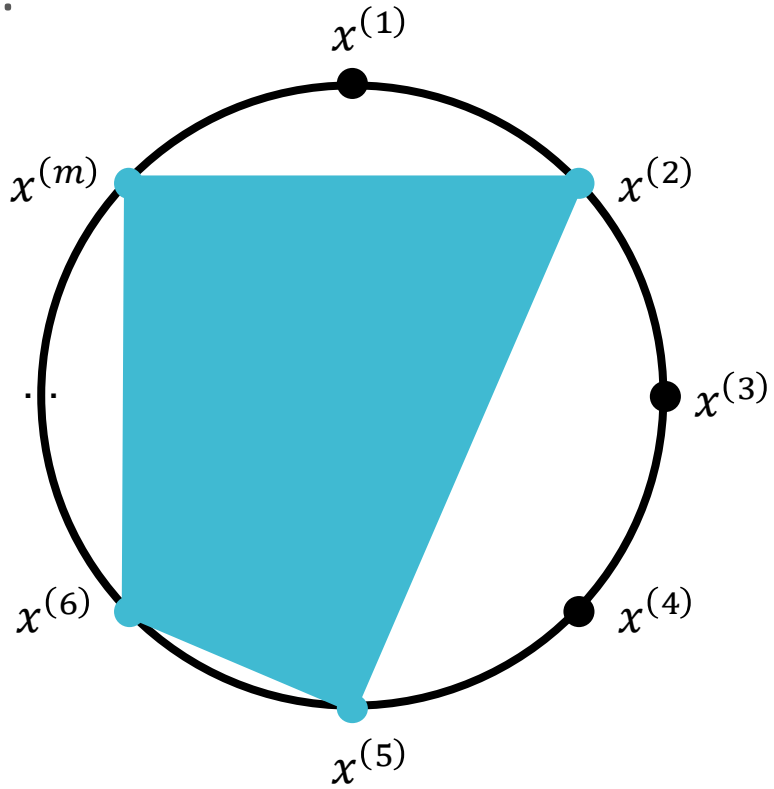


Growth Function: Example

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

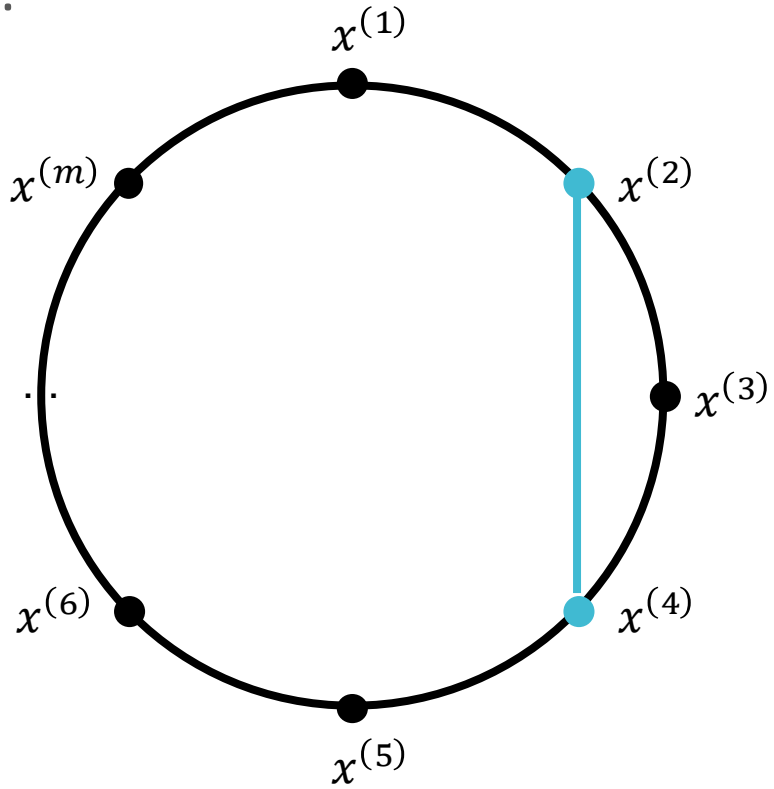
Growth Function: Example

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



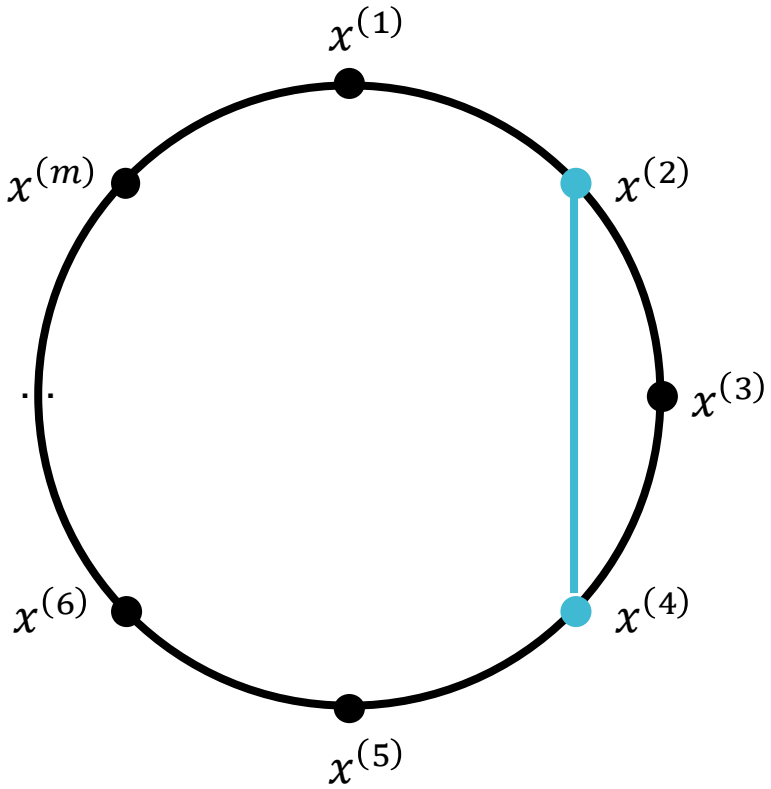
Growth Function: Example

- $x^{(m)} \in \mathbb{R}^2$ 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 $g_{\mathcal{H}}(M) = 2^M$



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^* on training data?

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

VC dimension and size of hypothesis space

- To be able to shatter m points, how many hypothesis do we need?

  2^m labelings ⇨ $|H| \geq 2^m$

- Given $|H|$ hypothesis, number of points we can shatter $m <= \log_2|H|$

  $VC(H) \leq \log_2|H|$

- So VC bound is tighter.

Limitation of VC dimension

- Hard to compute for many hypothesis spaces

  VC(H) ≥ lower bound (easy)

  VC(H) = … (HARD!)

  For all placements of VC(H)+1 points, there exists a labeling that can't be shattered

- Too loose for many hypothesis spaces

  linear SVMs, VC dim = d+1 (d features)

  kernel SVMs, VC dim = ??

  = ∞ (Gaussian kernels)

Suggests Gaussian kernels are really BAD!!

Quick Game

Quick game: name the case

The lecture organizes learning-theory guarantees into four cases. Match each scenario to the right quadrant.

Scenarios

- **A** Decision tree of depth d; one hypothesis fits the data perfectly
- **B** Decision tree of depth d; labels have noise, so zero train error may be impossible
- **C** Infinite hypothesis class H; assume H contains the target concept
- **D** Infinite H; want bounds when training error may remain nonzero

Quick Game

Quick game: name the case

The lecture organizes learning-theory guarantees into four cases. Match each scenario to the right quadrant.

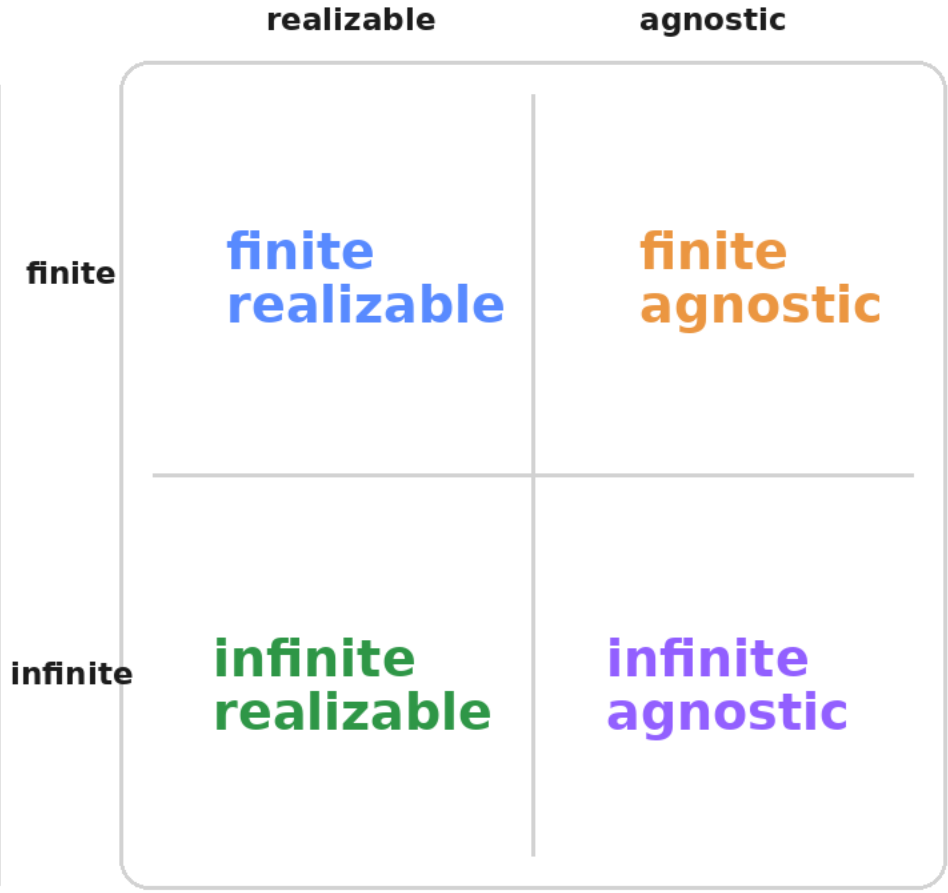
Scenarios

A Decision tree of depth d; one hypothesis fits the data perfectly

B Decision tree of depth d; labels have noise, so zero train error may be impossible

C Infinite hypothesis class H; assume H contains the target concept

D Infinite H; want bounds when training error may remain nonzero



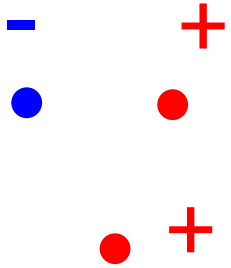
Reveal: A→finite realizable, B→finite agnostic, C→infinite realizable, D→infinite agnostic.

Rademacher Complexity

- Instead of all possible labelings, measure complexity by how accurately a hypothesis space can match a random labeling of the data.

For each data point i, draw random label

σ_i s.t. $P(\sigma_i = +1) = ½ = P(\sigma_i = -1)$



Then empirical Rademacher complexity of H is

$$\widehat{R}_m(H) = \mathbb{E}_\sigma \left[\sup_{h \in H} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i h(X_i) \right) \right]$$

Max correlation possible with random labels

Rademacher Bounds

- With probability ≥ 1-δ,

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \widehat{R}_m(H) + 3\sqrt{\frac{\log(2/\delta)}{m}}$$

where empirical Rademacher complexity of H

$$\widehat{R}_m(H) = \mathbb{E}_\sigma \left[\sup_{h \in H} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i h(X_i) \right) \right]$$

is purely data-dependent.

Summary of PAC bounds

With probability ≥ 1-δ,

1) for all h ∈ H s.t. $error_{train}(h) = 0$,

$$error_{true}(h) \leq \varepsilon = \frac{\ln |H| + \ln \frac{1}{\delta}}{m}$$

2) for all h ∈ H,

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}}$$

Finite hypothesis space

3) for all h ∈ H,

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = 8\sqrt{\frac{VC(H)\left(\ln \frac{m}{VC(H)} + 1\right) + \ln \frac{8}{\delta}}{2m}}$$

4) For all h ∈ H,

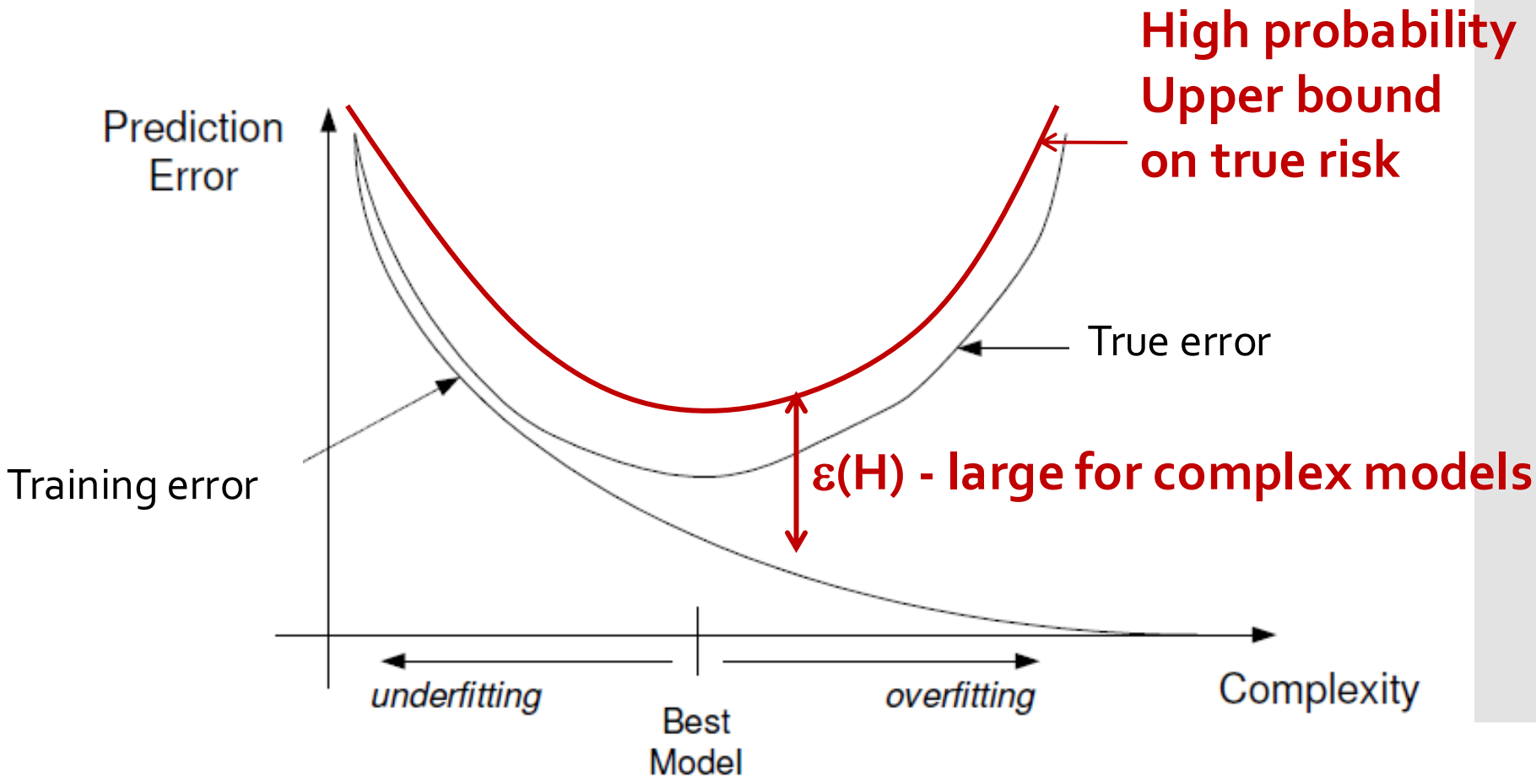
$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon = \widehat{R}_m(H) + 3\sqrt{\frac{\log(2/\delta)}{m}}$$

Infinite hypothesis space

PAC Bounds

With probability $\geq 1-\delta$, for all $h \in H$,

$$|error_{true}(h) - error_{train}(h)| \leq \varepsilon(H)$$



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

Poll:

Final Poll

Final poll (pick one)

Which statement best matches the lecture’s main takeaway about model complexity and generalization?

A. Once training error is zero, VC dimension no longer matters.

B. Increasing d_VC(H) makes both training error and generalization strictly better.

C. Increasing d_VC(H) can reduce training error, but it worsens the complexity term in generalization bounds.

D. Rademacher complexity ignores the data and depends only on the hypothesis class size.

E. Infinite VC dimension always means the model is unusable in practice.