

Learning Theory

Nishant Mehta

Lecture 12 - Part *II* and 13

Generalization

Suppose a learning algorithm returns a hypothesis with low training error.

When can we guarantee that the hypothesis's true error is also low?

Main question: How can we use the training error of a learning algorithm to estimate the algorithm's true error?

Generalization

Main question: How can we use the training error of a learning algorithm to estimate the algorithm's true error?

We will focus on *binary classification*

Recall:

Training error: $\hat{R}(\hat{h}, D) = \frac{1}{n} \sum_{i=1}^n 1 [\hat{h}(X_i) \neq Y_i]$

True error:
(Risk) $R(\hat{h}) = E_{(X,Y) \sim P} [1 [\hat{h}(X) \neq Y]]$
 $= \Pr_{(X,Y) \sim P} (\hat{h}(X) \neq Y)$

Realizable setting

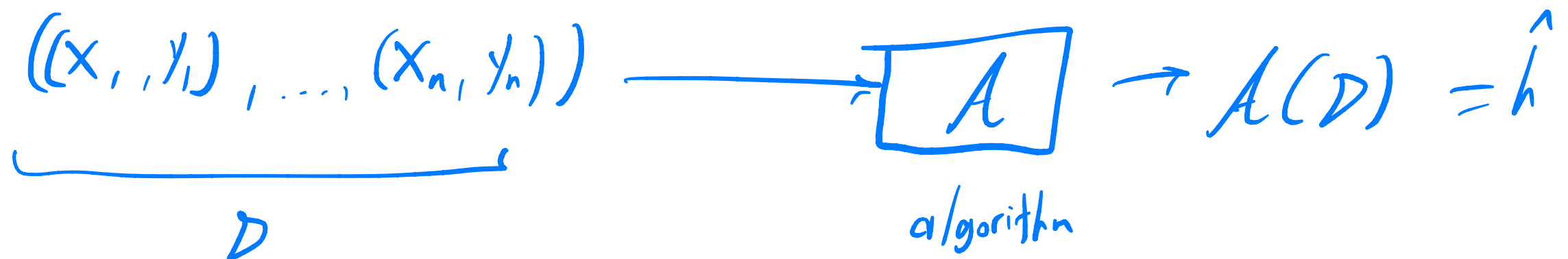
Suppose that each example is in input space $\mathcal{X}$

In the *realizable setting*:

There is a known *concept class* $\mathcal{C}$, a set of concepts, where each concept c is a rule mapping from $\mathcal{X}$ to $\{0, 1\}$

There is a concept $c \in \mathcal{C}$ such that, for any input X , the label is $Y = c(X)$

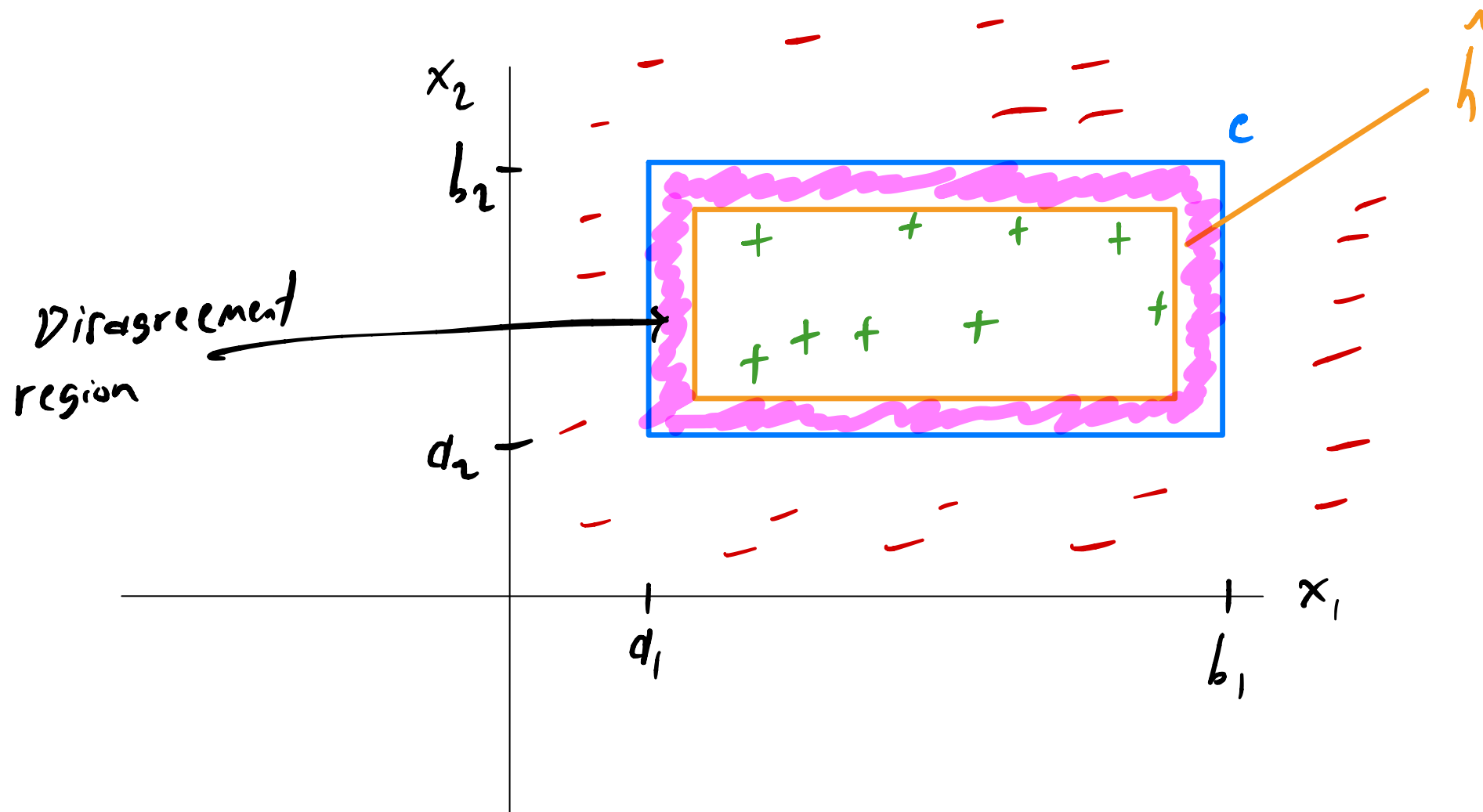
Given training data, learning algorithm selects hypothesis $\hat{h} \in \mathcal{H}$



Realizable setting: Example 1

Learning rectangles

$$\mathcal{X} = \mathbb{R}^2$$

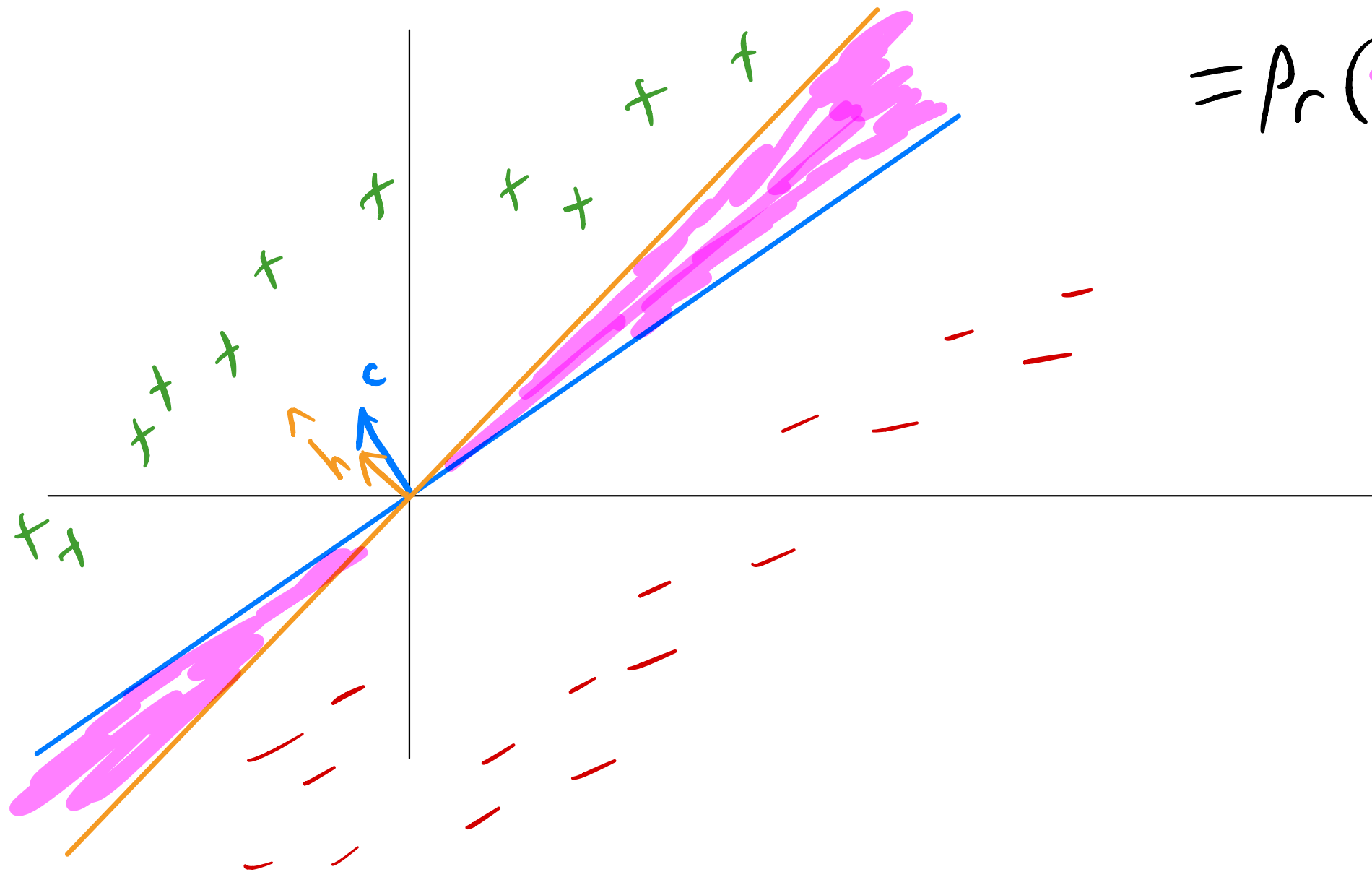


$$R(\hat{h}) = \Pr(X \in \text{disagreement region})$$

$$c(x) = \mathbb{1}[a_1 \leq x_1 \leq b_1] \cdot \mathbb{1}[a_2 \leq x_2 \leq b_2]$$

Realizable setting: Example 2

Homogeneous linear separators in $\mathbb{R}^2$



$$R(\hat{h}) = P_r(\text{magenta region})$$

Back to generalization

In the realizable setting, we have:

Training error: $\hat{R}(\hat{h}, D) = \frac{1}{n} \sum_{i=1}^n 1 \left[\hat{h}(X_i) \neq c(X_i) \right]$

Risk: $R(\hat{h}) = \Pr_{X \sim P} \left(\hat{h}(X) \neq c(X) \right)$

Main question:

How can we use training error $\hat{R}(\hat{h}, D)$ to upper bound risk $R(\hat{h})$,
no matter what distribution the data comes from?

A bad learning algorithm

memorization algorithm

$$\hat{h}(X) = \begin{cases} y_i & \text{if } \exists i \in [n] \text{ } X = X_i \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{array}{l} R(\hat{h}) \\ = \Pr_{(X,Y) \sim P} (Y \neq 1) \end{array}$$

Can we guarantee zero risk?

What type of guarantee can we hope to achieve?

What if we try to seek a hypothesis which gets zero risk? Is this possible?

Example: linear separators

No!

But we will try to guarantee

that $n \rightarrow \infty$, $R(\hat{h}) \rightarrow 0$

PAC Learning

A *Probably approximately correct (PAC) guarantee* is one of the form:

Suppose a learning algorithm outputs a hypothesis $\hat{h}$.

[Then with probability at least $1 - \delta$ (over the training sample),] "Probably"

[the risk $R(\hat{h})$ is at most ϵ .]

↑
"Approximately correct"

Towards achieving a PAC guarantee

In the realizable setting, we assume $\mathcal{C}$ contains a perfect classifier.

So, let's ensure that any hypothesis we select is consistent with the training sample.

We say that hypothesis $h \in \mathcal{H}$ is *consistent* (with training sample D) if it correctly classifies all the training examples, so $\hat{R}(h, D) = 0$

Towards achieving a PAC guarantee

In the realizable setting, we assume $\mathcal{C}$ contains a perfect classifier.

So, let's ensure that any hypothesis we select is consistent with the training sample.

We say that hypothesis $h \in \mathcal{H}$ is *consistent* (with training sample D) if it correctly classifies all the training examples, so $\hat{R}(h, D) = 0$

Version space: $\hat{V} = \left\{ h \in \mathcal{H} : \hat{R}(h, D) = 0 \right\}$

(set of hypotheses in $\mathcal{H}$ that are consistent with the training sample)

A PAC guarantee

Version space: $\hat{V} = \{h \in \mathcal{H} : \hat{R}(h, D) = 0\}$

(set of hypotheses in $\mathcal{H}$ that are consistent with the training sample)

Theorem

If $|\mathcal{H}| < \infty$, the probability that there is a hypothesis $h \in \hat{V}$ with risk $R(h) > \varepsilon$ is at most $|\mathcal{H}|e^{-n\varepsilon}$.

↑
"ε-bound"

$n = 1000$

$\mathcal{H}$ is a set of
 $|\mathcal{H}| = 100$

recognizers for some virus
 $\varepsilon = 2\% = 0.02$

↑
 $100 \cdot e^{-20} \approx 0$

A PAC guarantee

Version space: $\hat{V} = \left\{ h \in \mathcal{H} : \hat{R}(h, D) = 0 \right\}$

(set of hypotheses in $\mathcal{H}$ that are consistent with the training sample)

Theorem

If $|\mathcal{H}| < \infty$, the probability that there is a hypothesis $h \in \hat{V}$ with risk $R(h) > \varepsilon$ is at most $|\mathcal{H}|e^{-n\varepsilon} = \delta$

Inversion

$$\begin{aligned} |\mathcal{H}| e^{-n\varepsilon} &= \delta \\ \Leftrightarrow \frac{|\mathcal{H}|}{\delta} &= e^{n\varepsilon} \Leftrightarrow \frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{n} = \varepsilon \\ \text{w.p. at least } 1 - \delta, \text{ all hypotheses } h \text{ with } R(h) > \varepsilon &\text{ satisfy } \hat{R}(h, D) > 0. \end{aligned}$$

Proof

Lemma

Let $h \in \mathcal{H}$ be fixed hypothesis s.t. $R(h) > \varepsilon$.

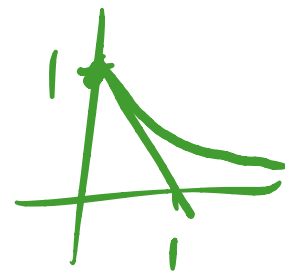
$$\Pr(\hat{R}(h, \mathcal{D}) = 0) \leq (1 - \varepsilon)^n \leq e^{-n\varepsilon}$$

$\swarrow^n$ examples

$1 - x \leq e^{-x}$

Proof

$$\text{Let } M_i = \begin{cases} 1 & \text{if } h(X_i) \neq Y_i \\ 0 & \text{if } h(X_i) = Y_i \end{cases}$$



$$\Pr(M_i = 1) = R(h)$$

$\nwarrow$ Bernoulli ($R(h)$)

$$\Pr(\hat{R}(h, \mathcal{D}) = 0) = \Pr(M_1 = 0, M_2 = 0, \dots, M_n = 0)$$

$= \prod_{i=1}^n \Pr(M_i = 0)$
 $= (1 - R(h))^n$
 $< (1 - \varepsilon)^n$

Proof of Theorem

$$\Pr(\exists h \in \mathcal{H} : \hat{R}(h, \mathcal{D}) = 0 \text{ and } R(h) > \varepsilon)$$

$$= \Pr\left(\bigcup_{h \in \mathcal{H}} \{ (x_1, \dots, x_n) \in \mathcal{X}^n : \hat{R}(h, \mathcal{D}) = 0 \text{ and } R(h) > \varepsilon \}\right)$$

$$\leq \sum_{h \in \mathcal{H}} \Pr(\hat{R}(h, \mathcal{D}) = 0 \text{ and } R(h) > \varepsilon)$$

$$= \sum_{\substack{h \in \mathcal{H} \\ h: R(h) > \varepsilon}} \underbrace{\Pr(\hat{R}(h, \mathcal{D}) = 0)}_{\leq e^{-n\varepsilon}}$$

$$\leq |\mathcal{H}| e^{-n\varepsilon}$$

PAC Learnability

We say that $\mathcal{C}$ is *PAC learnable* if there exists an algorithm $\mathcal{A}$ which, for all concepts $c \in \mathcal{C}$, for all distributions P over an input space $\mathcal{X}$ of dimension d , and for all $\varepsilon > 0$ and $\delta \in (0, 1)$, satisfies:

If $\mathcal{A}$ is given access to examples drawn from P and labeled according to c , then with probability at least $1 - \delta$, we have that the risk $\Pr(\hat{h}(X) \neq c(X)) \leq \varepsilon$.

We say $\mathcal{C}$ is *efficiently PAC learnable* if, in addition, $\mathcal{A}$ uses a number of examples polynomial in d , $1/\varepsilon$, and $1/\delta$.

Agnostic Learning

joint distribution over labeled examples (X, Y)

Assume $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P$

What if no hypothesis in $\mathcal{H}$ has zero risk?

What if no hypotheses (among all rules!) have zero risk?

Let's give up on learning $h \in \mathcal{H}$ with zero training error.

Instead, try to show that $R(h)$ isn't much larger than $\hat{R}(h, D)$

risk

empirical risk

Bounding the risk for a fixed hypothesis h

Hoeffding's inequality

Let $Z, Z_1, Z_2, \dots, Z_n \stackrel{\text{iid}}{\sim} P$, where $0 \leq Z \leq 1$
 $0 \leq Z_i \leq 1$ for $i = 1, \dots, n$

$$\text{Then } \Pr \left(\underbrace{\mathbb{E}[Z]}_{R(h)} \geq \underbrace{\frac{1}{n} \sum_{i=1}^n Z_i}_{\hat{R}(h, D)} + \varepsilon \right) \leq e^{-2n\varepsilon^2}$$

Lemma. Let h be a fixed hypothesis. Then

$$\Pr \left(R(h) \geq \hat{R}(h, D) + \varepsilon \right) \leq e^{-2n\varepsilon^2}$$

Equivalently, with probability at least $1 - \delta$

$$R(h) \leq \hat{R}(h, D) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

set δ
and solve
for ε

$$Z = \begin{cases} 1 & \text{if } h(x) \neq y \\ 0 & \text{o.w.} \end{cases}$$

$$Z_i = \begin{cases} 1 & \text{if } h(x_i) \neq y_i \\ 0 & \text{o.w.} \end{cases}$$

$$\mathbb{E}[Z] = R(h)$$

Bounding the risk of $\hat{h}$ selected from finite class $\mathcal{H}$

Suppose, given training data $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P$,

we learn a rule $\hat{h} \in \mathcal{H}$.

Then with probability at least $1 - \delta$,

$$R(\hat{h}) \leq \hat{R}(\hat{h}, D) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2n}}$$

Effective size

When $|\mathcal{H}|$ is finite, our notion of size was $|\mathcal{H}|$

What if $|\mathcal{H}|$ is infinite?

New measure of the size of $\mathcal{H}$: “effective size of $\mathcal{H}$ ”

Effective size of $\mathcal{H}$ relative to training sample $S = (x_1, x_2, \dots, x_n)$
is defined as:

unlabeled sample

$$|\mathcal{H}|_S = \left| \left\{ h \in \mathcal{H} : \begin{pmatrix} h(X_1) \\ h(X_2) \\ \vdots \\ h(X_n) \end{pmatrix} \right\} \right|$$

of distinct ways we can label
sample S using hypotheses in $\mathcal{H}$

Effective size

When $|\mathcal{H}|$ is finite, our notion of size was $|\mathcal{H}|$

What if $|\mathcal{H}|$ is infinite?

New measure of the size of $\mathcal{H}$: “effective size of $\mathcal{H}$ ”

Effective size of $\mathcal{H}$ relative to training sample $S = (x_1, x_2, \dots, x_n)$ is defined as:

$$|\mathcal{H}|_S = \left| \left\{ h \in \mathcal{H} : \begin{pmatrix} h(X_1) \\ h(X_2) \\ \vdots \\ h(X_n) \end{pmatrix} \right\} \right|$$

of distinct ways we can label
sample S using hypotheses in $\mathcal{H}$

Example

	h_1	h_2	h_3	h_4	h_5
x_1	1	1	1	1	0
x_2	0	0	0	0	0
x_3	0	1	1	0	1

Effective size

When $|\mathcal{H}|$ is finite, our notion of size was $|\mathcal{H}|$

What if $|\mathcal{H}|$ is infinite?

New measure of the size of $\mathcal{H}$: “effective size of $\mathcal{H}$ ”

Effective size of $\mathcal{H}$ relative to training sample $S = (x_1, x_2, \dots, x_n)$ is defined as:

$$|\mathcal{H}|_S = \left| \left\{ h \in \mathcal{H} : \begin{pmatrix} h(X_1) \\ h(X_2) \\ \vdots \\ h(X_n) \end{pmatrix} \right\} \right|$$

of distinct ways we can label
sample S using hypotheses in $\mathcal{H}$

Example

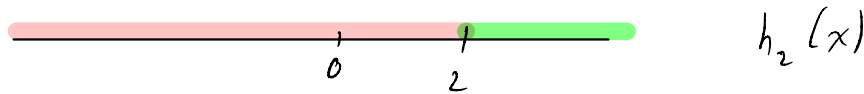
	h_1	h_2	h_3	h_4	h_5
x_1	1	1	1	1	0
x_2	0	0	0	0	0
x_3	0	1	1	0	1

effective size = 3

$|\mathcal{H}_S|$ = " # of distinct ways that we can label
 S (label $x_1, x_2, \dots, x_n$) using hypotheses in $\mathcal{H}$ "

Example (threshold functions)

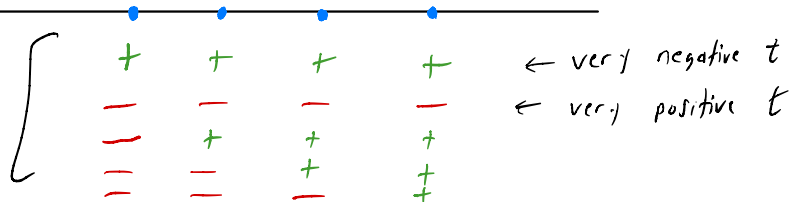
$X = \mathbb{R}$ $\mathcal{H} = \{h_t : t \in \mathbb{R}\}$ $h_t(x) = \begin{cases} 1 & \text{if } x \geq t \\ 0 & \text{if } x < t \end{cases}$



For any sample of
size n ,

$$|\mathcal{H}_S| \leq n+1$$

5 different labelings



In general, how can we upper bound $|\mathcal{H}_{|S}|$ — "effective size of $\mathcal{H}$ w.r.t. S "
↑
"examples"

Key Tool: VC dimension

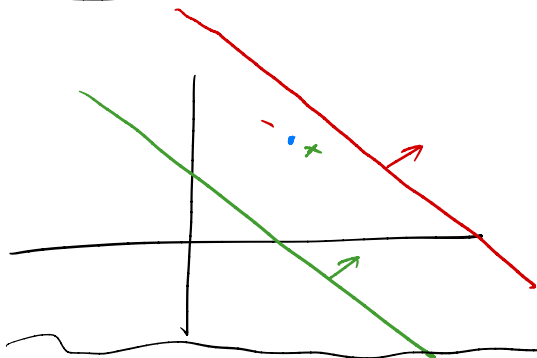
Definition: Shattering

finite unlabeled training sample ($|S| = k$)
↓
we say that $\mathcal{H}$ shatters a set $S \subseteq \mathcal{X}$ if,
for every possible labeling of S (2^k of these!),
there is $h \in \mathcal{H}$ that is consistent with labeling.

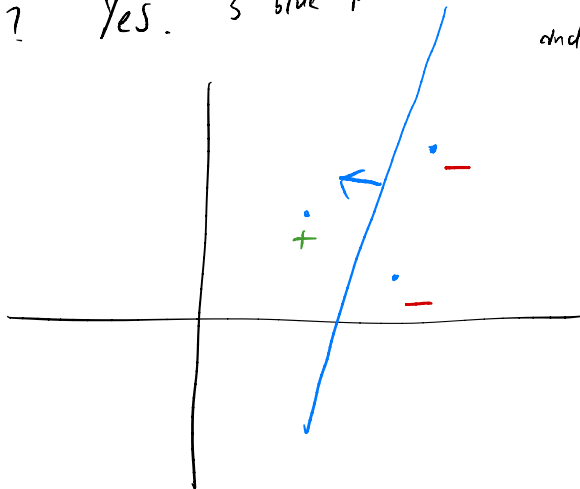
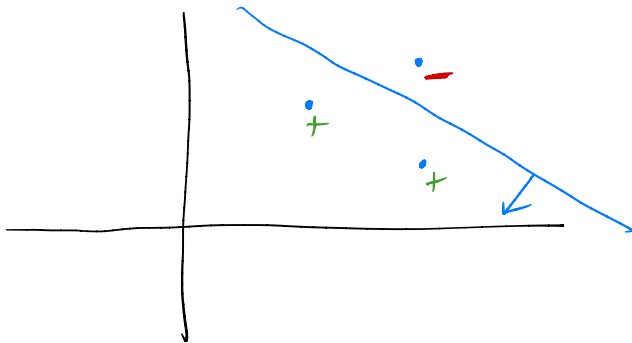
(implication: $|\mathcal{H}_{|S}| = 2^k$)

Example (Linear separators in $\mathbb{R}^2$) $\leftarrow \mathcal{H}$

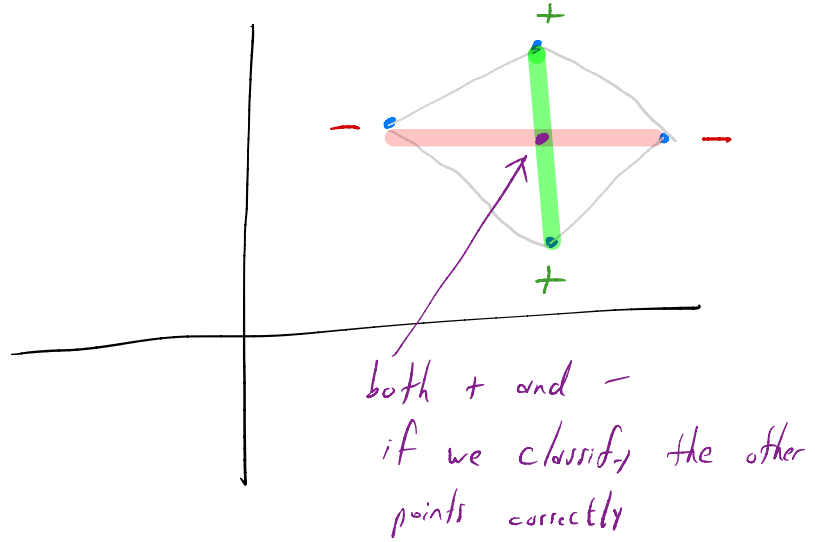
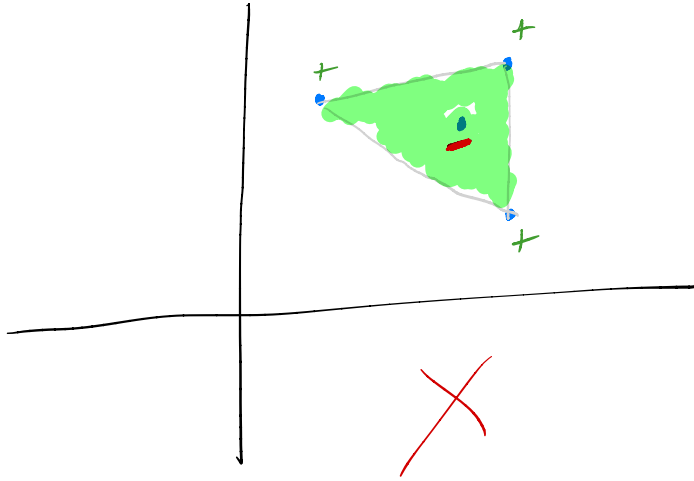
we can shatter $\mathcal{S}$ with one example



Can we shatter some $\mathcal{S}$ with 3 examples? Yes. 3 blue points represent a set s.t. $|\mathcal{S}|=3$ and $\mathcal{H}$ shatters $\mathcal{S}$.



Question: Does there exist S with $|S|=4$ s.t.
it shatters S ? No!



The largest S that we can shatter is of cardinality 3.

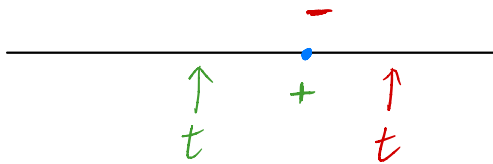
Definition (Vapnik - Chervonenkis Dimension)

The VC dimension of $\mathcal{H}$, $VC(\mathcal{H})$, is the cardinality of the largest finite $S \subseteq \mathcal{X}$ that is shattered by $\mathcal{H}$.

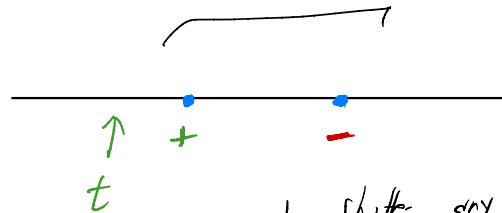
If $\mathcal{H}$ can shatter arbitrarily large sets S ,
then $VC(\mathcal{H}) = \infty$

Example (Threshold functions)

can shatter on S
with $|S|=1$



we cannot achieve this labeling



cannot shatter any S
with $|S|=2$

$\Rightarrow$
 $\Downarrow$
 $\Leftarrow$

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

Bounding the risk when $\mathcal{H}$ has finite VC dimension

Suppose, given training data $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P$,

we learn a rule $\hat{h} \in \mathcal{H}$.

Then with probability at least $1 - \delta$,

$$R(\hat{h}) \leq \hat{R}(\hat{h}, D) + O\left(\sqrt{\frac{\text{VC}(H) + \log \frac{1}{\delta}}{n}}\right)$$