Learning Theory

Nishant Mehta

Lecture 12 - Part II

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

True error: (Risk)

$$R(\hat{h}) = \mathrm{E}_{(X,Y)\sim P}\left[1\left[\hat{h}(X) \neq Y\right]\right]$$
$$= \Pr_{(X,Y)\sim P}\left(\hat{h}(X) \neq Y\right)$$

Realizable setting

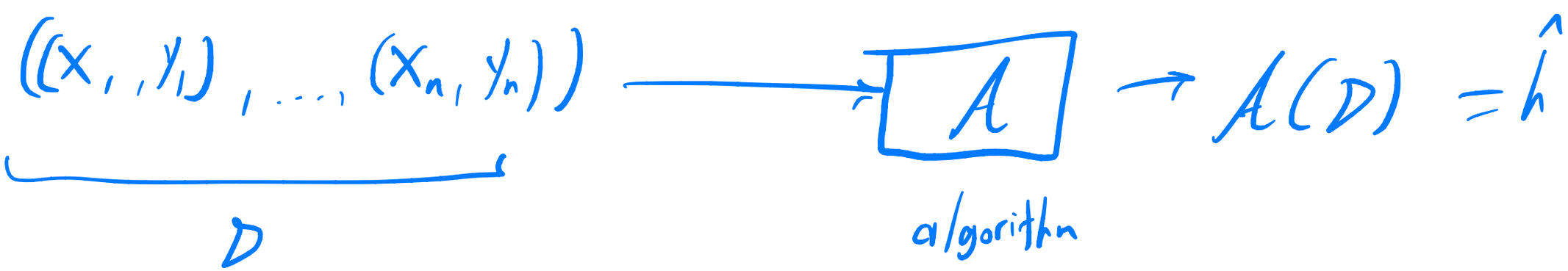
Suppose that each example is in input space $\mathcal{X}$

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 $Y = c(X)$

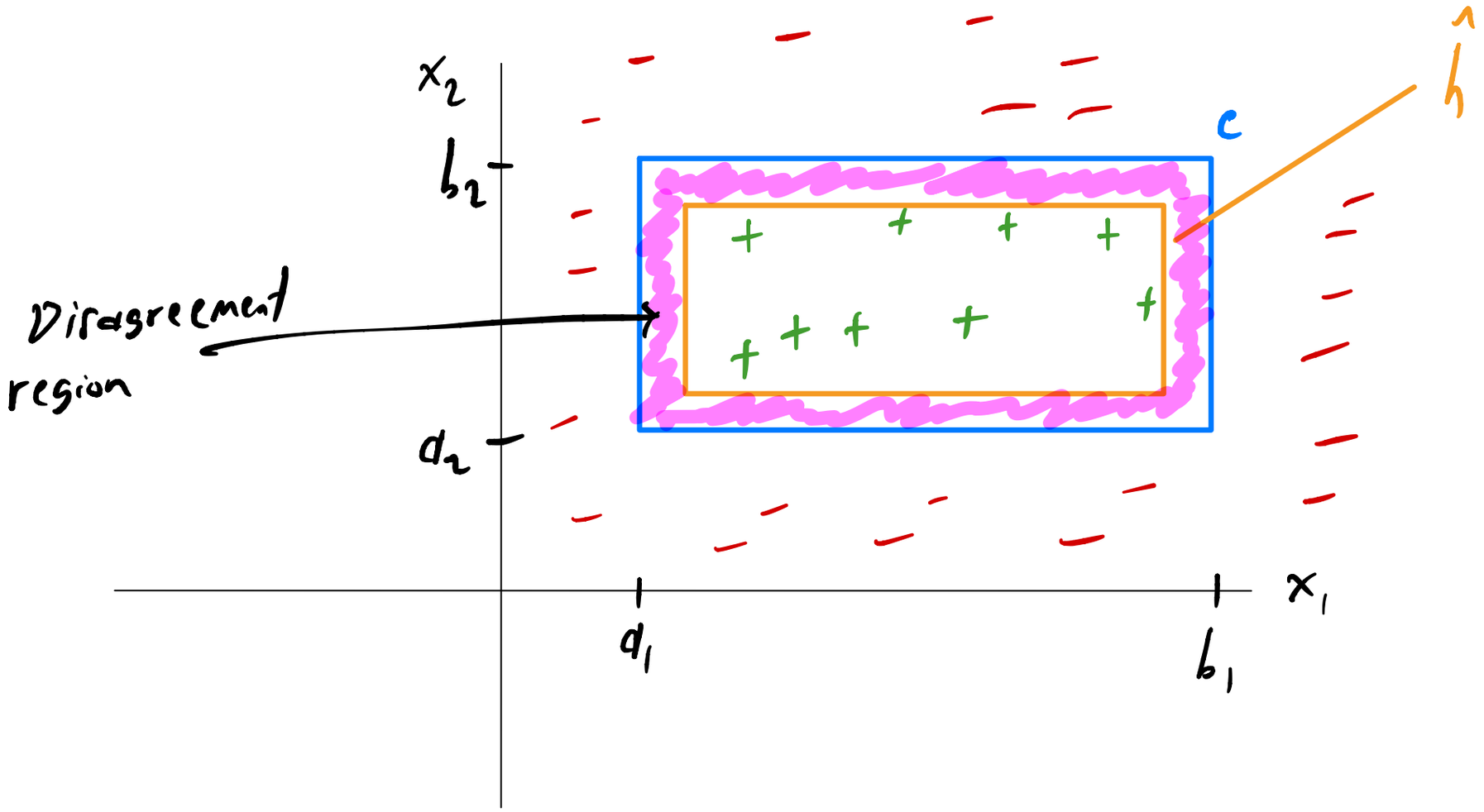
Given training data, learning algorithm selects hypothesis $\hat{h} \in \mathcal{H}$



Realizable setting: Example 1

Learning rectangles

$X = \mathbb{R}^2$



$R(\hat{h}) = \Pr(X \in \square)$

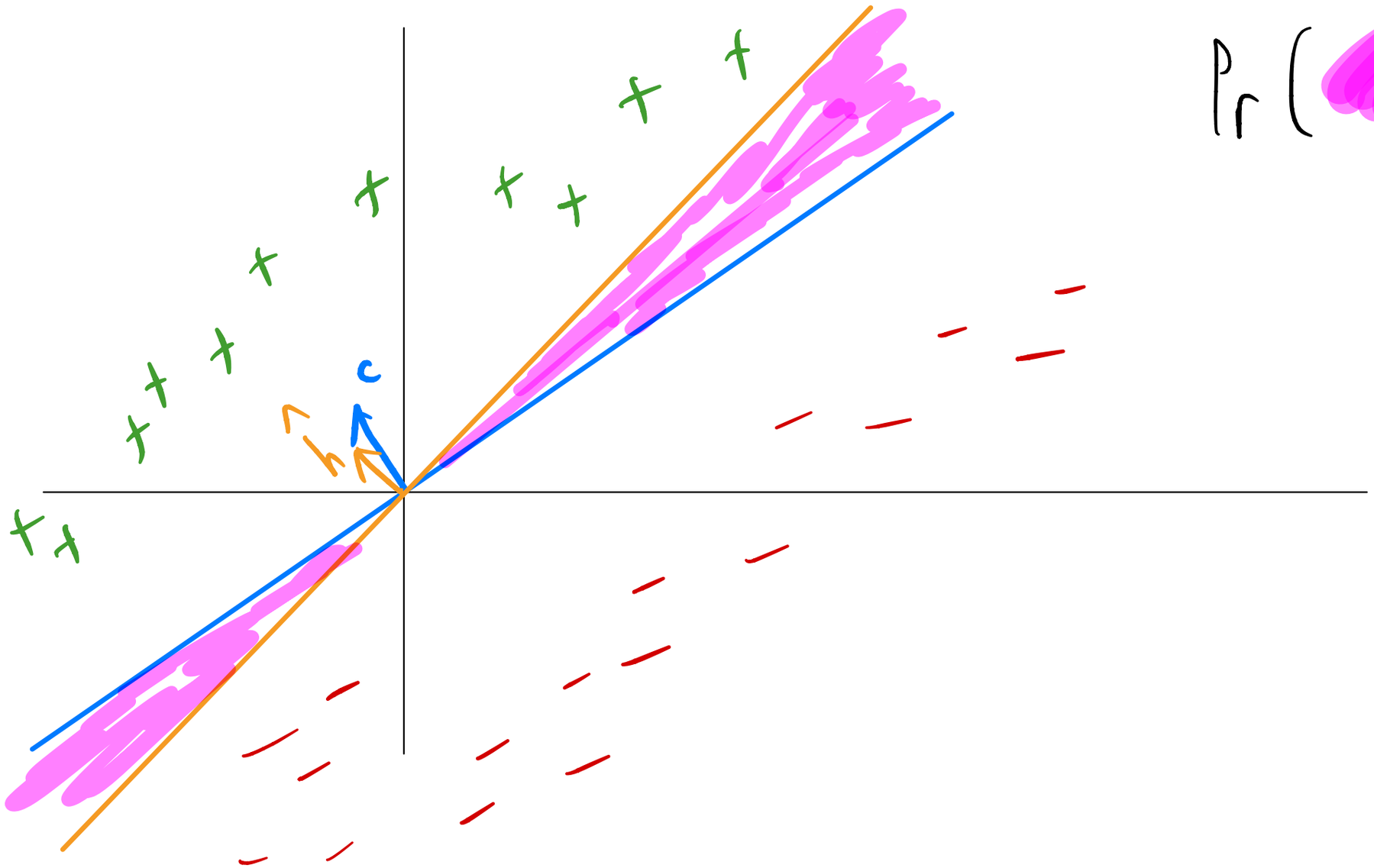
$c(x) = 1[a_1 \le x_1 \le b_1] \cdot 1[a_2 \le x_2 \le b_2]$

Realizable setting: Example 2

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

$R(\hat{h}) =$

$P_r($  $)$



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

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

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