THEORY OF MACHINE LEARNING

LECTURE 3 4

PAC MODEL, VC DIMENSION

RECAP – VALIANT'S THEORY OF (SUPERVISED) LEARNING

(concept / hyp)
labeling function

- **Learnability (from examples).** [Suppose D is fixed.] We say that a concept class is "learnable" if there exists an [efficient] algorithm ***A*** with the property: for all $\epsilon > 0$, there exists *m* (number of samples) such that when given *m* <u>*i.i.d.*</u> samples from D along with their labels, ***A*** produces a hypothesis *h* with risk less than ϵ, with prob. >= 0.9

$$R_{D,\ell}(h) = \Pr_{x \sim D}\left(h(x) \neq \ell(x)\right)$$

- (Recall, risk = expected error on sample from distribution)
- Beyond examples? (technically yes, e.g., teacher/student)

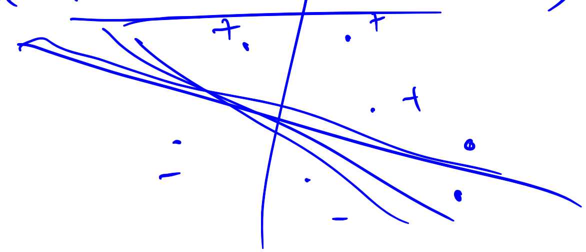
RECAP: NO FREE LUNCH THEOREM

inductive bias.

- Motivation: do we really need to restrict the hypothesis/concept class *before starting* learning? – yes!
- No free lunch: (informal) there is no "universal" learner, even if it's allowed to be inefficient (even for binary classification under a uniform distribution, unless it "sees most of the labels")
- Proof via a counting argument – too many hypotheses

$\mathcal{H}$: - class of hypotheses (all linear thresholds..)

- all 100 layer NNs with width 100k.



TODAY'S PLAN

- Definition. (Agnostic) PAC learning
- Finite classes are PAC learnable
- Dealing with infinite classes: 'growth function' and VC dimension

PAC LEARNING (REALIZABLE CASE)

(over some domain X) . $\mathcal{H}$: bunch of hypotheses over X .

- **Learnability of a concept class.** *A concept class H is PAC learnable* (over domain X) if there exists an algorithm **A** that for all $\epsilon, \delta > 0$ and distributions D, has the following property:

N $\quad \frac{d \log N}{\varepsilon^2} \quad \log\frac{1}{\delta}$

$2^{1/\varepsilon^2} \quad 2^{2^{1/\delta}}$

∃ a function $m(\varepsilon, \delta)$

  - given $m(\epsilon, \delta)$ samples (x, f(x)), where x ~ D and f is a (unknown) function in H, it outputs *h* with risk at most ϵ with probability at least $1 - \delta$.

0.05 $\quad$ 2/3

(label function is in $\mathcal{H}$.)

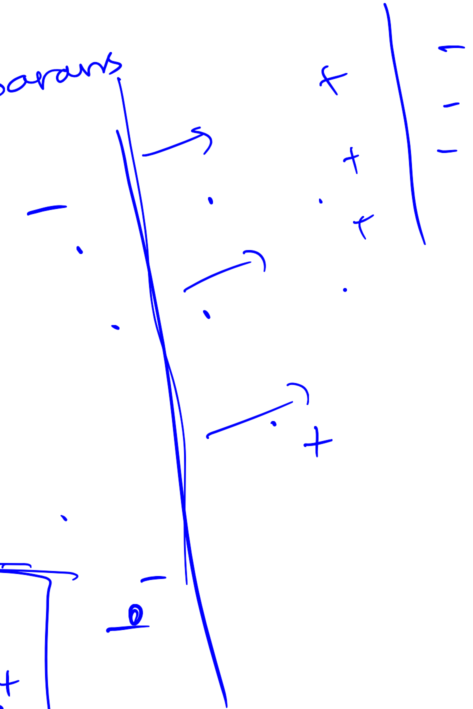
- (The sample size must not depend on D)
- As such *h* need not belong to H (improper learning)

- goal is to find true label function f
- declare success, if we find h that has risk $\leq \varepsilon$ $\quad h \rightsquigarrow f$.

Concept ≡ hypothesis.

Example of $\mathcal{H}$: $X = \mathbb{R}^2$

3 params

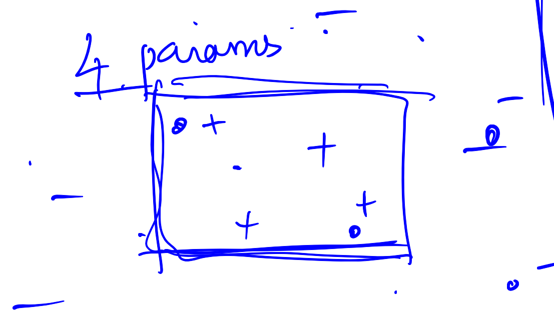


$\text{sign}(\cancel{ax+by+c})$ $= ax+by+c$ $\mathcal{H}$: all linear classifiers.

$\text{sign}(-(ax+by+c))$

$X = \mathbb{R}^2$

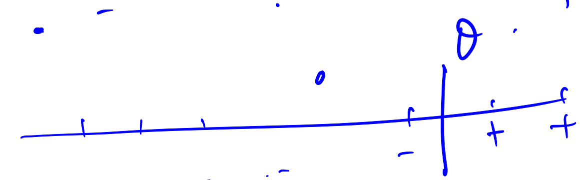
4 params



$\mathcal{H}$: all rectangles

$X : \mathbb{R}$.

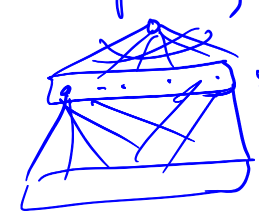
1 param



$\mathcal{H}$: all threshold functions

$X = \mathbb{R}^{100}$; $\mathcal{H}$: all depth 2 nns with 100 inputs,

every gate is ReLU.

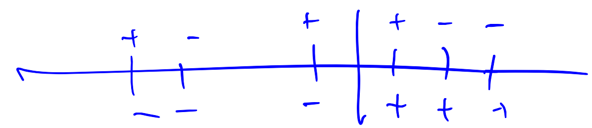


50

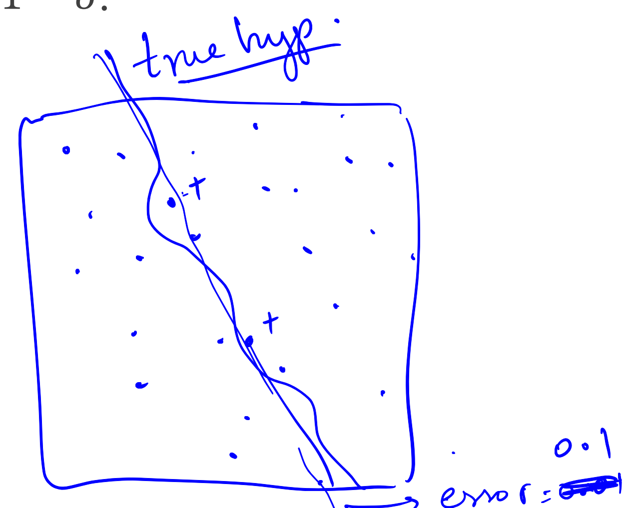
100

PAC LEARNING (NON-REALIZABLE CASE)

(agnostic)



- **Learnability of a concept class.** *A* concept class H is **agnostically** *PAC learnable* (over domain X) if there exists an algorithm ***A*** that for all $\epsilon, \delta > 0$ and distributions D, has the following property:
  - given $m(\epsilon, \delta)$ samples (x, f(x)), where x ~ D and f is a (unknown) function not necessarily in H, it outputs *h* with risk at most ϵ **more than the risk of the** h′ in H that is **"closest" to f,** with probability at least $1 - \delta$.
- (The sample size must not depend on D)
- Again, *h* need not belong to H (improper learning)



→ can find h with risk ≤ 0.1 + ε.

"weakens" the inductive bias.

EVERY FINITE CLASS IS PAC LEARNABLE (EVEN AGNOSTIC)

$\mathcal{H}: \{h_1, h_2, \ldots, h_N\}$; input space X (potentially infinite).

Theorem: $\mathcal{H}$ is PAC-learnable.

$(x, f(x))$

- Suppose H has only finitely many hypotheses

  (input space X may still be infinite)

example: $(x, f(x))$

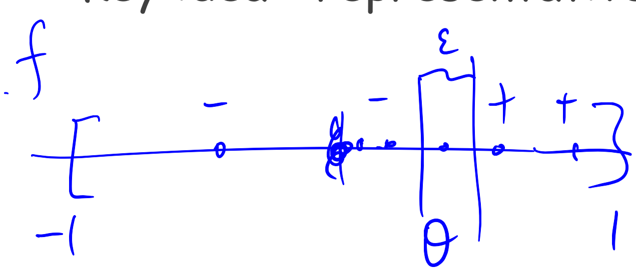
Sample S: collection of m these examples.

- **Generic algorithm:** empirical risk minimization (ERM)

  - get (m examples) (we'll fix m later)
  - find $h \in \mathcal{H}$ that is consistent with all these examples.
  - output (one such) h.

  - get m examples
  - find $h \in \mathcal{H}$ that minimizes "empirical risk"
  - output h.

- Key idea: "representative sample"



$\boxed{1/\varepsilon^2 \text{ samples}}$

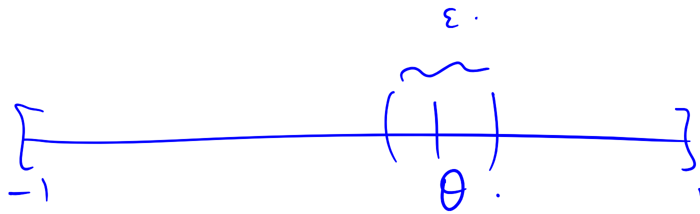
"empirical risk" of h & a sample S $= \frac{1}{|S|} \cdot \sum_{x \in S} \mathbb{1}_{h(x) \neq f(x)}$

(minimizing training error)

When is ERM bad?

- if there are too few examples (ie, m is too small)

- just got unlucky with the examples?



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

REPRESENTATIVE SAMPLE

- Let H be a hypothesis class and X be an input space with a distribution D on it, and let f be a target function. Sample $S \subseteq X$ is said to be ϵ –"representative" if **for all** h in H, we have:

$$\left| \frac{1}{|S|} \text{error}(S, h) - \text{risk}_D(h, f) \right| < \epsilon$$

empirical risk.

(risk on examples.

true risk w.r.t. $\mathcal{D}$.

$$\Pr_{x \sim \mathcal{D}} \left[h(x) \neq f(x) \right].$$

$$\frac{1}{|S|} \cdot \sum_{x \in S} \mathbb{1}[h(x) \neq f(x)]$$

RANDOM SAMPLE IS REPRESENTATIVE WHP!

- **Chernoff bound (Hoeffding).** Suppose $X_1, X_2, \ldots X_n$ are n iid samples from a distribution with mean μ and support [a, b]. Then we have

$$\Pr\left[\,\left|\frac{1}{n}(X_1 + \cdots + X_n) - \mu\right| > \epsilon\,\right] \le 2\exp\left(-\frac{\epsilon^2 n}{(a-b)^2}\right)$$

WHAT ABOUT INFINITE CLASSES?

- Note: as long as sample is representative, we are good!
- What if we can divide hypotheses into finitely many "classes"?
- Example of threshold functions on a line

GROWTH FUNCTION OF A CLASS

- For a class H and an input space X, we can define a notion of "growth function"