

## Lecture 13

~~RECAP~~

Definition (SQ dimension) The SQ-dimension of a class  $\mathcal{C}$  wrt a distribution  $D$  over  $X$  is the size of the largest subset  $\mathcal{C}' \subseteq \mathcal{C}$  st. for all  $f, g \in \mathcal{C}'$ ,

$$\left| \Pr_{x \sim D} [f(x) = g(x)] - \frac{1}{2} \right| < \frac{1}{|\mathcal{C}'|}.$$

Thm 2 : If  $\text{SQ-DIM}_D(\mathcal{C}) > \text{poly}(d)$  then you cannot efficiently learn  $\mathcal{C}$  over  $D$  by SQ-algorithms (even "weak-learning" to error  $\leq \frac{1}{2} - \frac{1}{\text{poly}(d)}$  is impossible).

Parity function :

$$X^d = \{0, 1\}^d$$

$$Y = \{0, 1\}$$

$$\mathcal{C} = \{w(x) = \langle w, x \rangle \bmod 2 : w \in \{0, 1\}^d\}$$

Thm : PARTIES are efficiently PAC learnable.

Let  $U$  be the uniform distribution over  $\{0,1\}^d$ .

Claim: Any two parity functions  $(w_1(x))$  &  $(w_2(x))$  (where  $w_1 \neq w_2$ ) are uncorrelated :

$$\Pr_{x \sim U} [w_1(x) = w_2(x)] = \frac{1}{2}.$$

TODAY

It will be convenient to define PARITIES as a function on  $\{-1, +1\}^d \rightarrow \{-1, +1\}$

$$x^d = \{\pm 1\}^d$$

$$y = \pm 1$$

$$\ell = \left\{ \begin{array}{l} (s(x) = \prod_{i \in s} x_i : s \subseteq \{1, \dots, d\} \end{array} \right\}$$

New Claim ; If  $s \neq T$

$$\Pr_{x \sim U} [s(x) = \ell_T(x)] = \frac{1}{2}.$$

Proof

$$\mathbb{E}_{x \sim U} [s(x) \cdot \ell_T(x)] = \mathbb{E}_{x \sim U} \left[ \prod_{i \in s} x_i \cdot \prod_{i \in T} x_i \right]$$

$$= \mathbb{E}_{x \sim U} \left[ \prod_{i \in S \Delta T} x_i \right]$$

$$(S \Delta T = (S - T) \cup (T - S))$$

$$= 0 \text{ if } S \neq T \text{ (unif dist.)}$$

$$\Pr[(s(n) = c_\tau(n))] + \Pr[(s(x) \neq c_\tau(x))] = 1$$

$$\mathbb{E}_{x \sim u} [(s(n) \cdot c_\tau(x))] = \Pr[(s(x) = c_\tau(n))] - \Pr[(s(x) \neq c_\tau(x))]$$

$$\therefore \mathbb{E}((s(x) \cdot c_\tau(x))) = 0$$

$$\Rightarrow \Pr[(s(x) = c_\tau(x))] = \frac{1}{2}.$$

Corollary of Thm 2: It is not possible to efficiently learn PARITIES in the SQ model over the uniform distribution

Proof:  $\text{SQ-DIM}_u(\epsilon) = 2^d$ .

We'll now prove Thm 2 for the special case of PARITIES.

Thm (hardness of parities in SQ): Any SQ algorithm for learning PARITIES over  $D = U$ , which makes queries of tolerance  $T \geq T_{\min}$  must make  $\Omega(T_{\min}^2 2^d)$  queries to  $\text{STAT}(c_0)$ .

Proof

Correlation SQ (CSQ) oracle:

for any query function  $\Psi : \mathcal{X} \rightarrow \{-1, +1\}$ , & tolerance  $\tau$

$$\text{let } P_\Psi = \mathbb{E} [\Psi(x). c(x)]$$

Oracle returns  $\hat{P}_\Psi \in [P_\Psi - \tau, P_\Psi + \tau]$

Lemma: If learner knows target distribution  $D$ , can simulate SQ oracle with CSQ oracle.

Proof: We can decompose any SQ  $\phi$  into:

$$\begin{aligned} \mathbb{E}_{x \sim D} [\phi(x, c(x))] &= \mathbb{E}_{x \sim D} [\phi(x, 1) \cdot \mathbb{1}(c(x) = 1)] \\ &\quad + \mathbb{E}_{x \sim D} [\phi(x, -1) \cdot \mathbb{1}(c(x) = -1)] \\ &= \mathbb{E}_{x \sim D} [\phi(x, 1) \left(1 + \frac{c(x)}{2}\right)] \\ &\quad + \mathbb{E}_{x \sim D} [\phi(x, -1) \cdot \left(1 - \frac{c(x)}{2}\right)] \end{aligned}$$

"target-independent query"  $\leftarrow \frac{1}{2} \left( \mathbb{E}_{x \sim D} [\phi(x, 1)] + \mathbb{E}_{x \sim D} [\phi(x, -1)] \right)$

"correlational query"  $\leftarrow \frac{1}{2} \left( \mathbb{E} (\phi(x, 1) c(x)) - \mathbb{E} (\phi(x, -1) \cdot c(x)) \right)$



Since we consider  $D = U$  (fixed), suffices to show hardness for CSQ oracle.

### Basics of Boolean function Analysis

Think of any function  $f: \{-1\}^d \rightarrow \{-1\}$  as a vector  $\vec{f}$  of  $2^d$  entries.

$$\left( \frac{1}{2^{d/2}} f(-1, -1, -1), \frac{1}{2^{d/2}} f(-1, -1, 1), \dots, \frac{1}{2^{d/2}} f(1, 1, 1) \right)$$

Note that  $\langle \vec{f}, \vec{g} \rangle = \mathbb{E}_{x \sim U} [f(x) \cdot g(x)]$

$$\langle \vec{f}, \vec{g} \rangle = \sum_{i=1}^{2^d} \frac{1}{2^{d/2}} f(x_i) \cdot \frac{1}{2^{d/2}} g(x_i)$$

$$= \mathbb{E}_{x \sim U} [f(x) \cdot g(x)].$$

\*  $\langle \vec{f}, \vec{f} \rangle = 1$

Fourier analysis: change "basis" to understand  $f$ .

Recall that an orthonormal basis for a vector space is a set of orthogonal unit vectors that span the space.

If  $v_1, v_2$  are orthonormal basis for  $\mathbb{R}^2$ , we can write any vector

$$w = \langle w, v_1 \rangle v_1 + \langle w, v_2 \rangle v_2.$$

Claim PARITIES form an orthonormal basis for our vector space.

Proof Note that for  $S \neq T$

$$\begin{aligned} \langle \vec{c}_S(x), \vec{c}_T(x) \rangle &= \mathbb{E} (c_S(x) \cdot c_T(x)) \\ &= 0. \end{aligned}$$

$$\langle \vec{c}_S(x), \vec{c}_S(x) \rangle = 1. \quad \forall S$$

■

For any CSQ  $\psi : \mathbb{R}^d \rightarrow \{-1, 1\}$

$$\vec{\psi} = \sum \alpha_S \vec{c}_S$$

$$S: S \subseteq \mathbb{R}^d \quad \alpha_S$$

where  $\vec{c}_S$  is the parity function over  $S$ .

$$\text{Note that } \alpha_S = \mathbb{E}_{x \sim U} [\psi(x) \cdot c_S(x)]$$

Expected response to CSQ  $\psi$  if the target function is  $c_S(x)$ .

Since  $\langle \vec{\Psi}, \vec{\Psi} \rangle = 1$

$$\sum d_{S^*}^2 = 1$$

i.e. There can be atmost  $\frac{1}{\tau^2}$  S.s.t.

$$|d_S| > \tau.$$

Note that if target is  $S^*$ , then CSQ oracle can just answer 0 to this query  $\Psi$  if  $|d_{S^*}| < \tau$ .

We draw the target parity function (the set  $S^*$ ) uniformly at random from all possible subsets.

Claim: If alg. makes  $\leq \tau^2 2^d$  queries, whp over choice of  $S^*$ , CSQ can answer 0 to all these queries.

Proof: There are  $2^d$  options

□ □ □ □ □ □ □ □ □

Any query  $\Psi$  is non-zero on atmost  $\frac{1}{\tau^2}$  options.

If  $s^*$  is not among these, just answer 0.

Since  $s^*$  is random, whp cannot "find" it in  $O(\tau^2 2^d)$  queries. [Exercise]

This finishes proof.

Thm Over the unif dist.  $D$ , in the presence of  $R \in N$  with noise level  $\eta$ , PARITIES are learnable  $O\left(\frac{d}{(1-2\eta)^2}\right)$  samples, whp.  
(information-theoretically)

Proof Let  $\varepsilon = \frac{1}{2} - \eta$

We keep label w.p.  $\frac{1}{2} + \varepsilon$   
" flip "  $\frac{1}{2} - \varepsilon$

Get  $m = O\left(\frac{d}{\varepsilon^2}\right)$  samples from  $Ex^{\eta}(c, D)$ .

(where  $c(\pi) = c_{S^*}(\pi) = \prod_{i \in S^*} \pi_i$ )

Claim: With high prob,

- ①  $S^*$  is consistent with  $\geq \left(\frac{1}{2} + \frac{\varepsilon}{2}\right)$  fraction of examples
- ② Any  $S \neq S^*$  is consistent with  $\leq \frac{1}{2}$  fraction.

Proof Exercise. Use Chernoff/Hoeffding & property that parity functions are uncorrelated for SFST.

■

Best known algorithm for learning parity

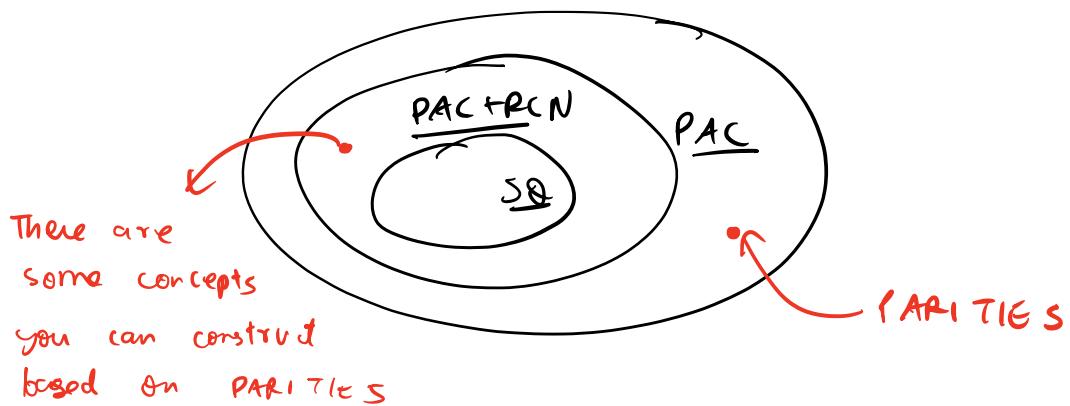
with noise:  $2^{d/\log d}$  time

(Blum-Wasserman-Kalai '03)

Slightly less than exponential, non-SQ algorithm.

L PNC (Learning Parity w/ Noise) is believed to be hard.

Efficient also.



$2^{d/\log d}$  alg.  $\Rightarrow$  we can learn parities over  $O(\log d \log \log d)$  co-ordinates with RCN in poly-time

$$SQ \subset PAC + RCN \subseteq PAC$$

↑

conditioned on hardness of  
LPN, this is proper.

With exception of Gaussian elimination,  
almost all known algo. can be run in  
SQ model.

SQ is sort of frontier of algorithmic knowledge.

∴ To show that some learning problem is  
hard, many recent papers show hardness in  
SQ models.

Thanks to SQ-dim, we have an information  
theoretic way to show hardness in SQ.

## Boosting

Recall our definition of weak-learning.

Weak-learning: An algorithm  $A$  is a weak learner with edge/advantage  $\gamma$  for class  $C$  if: for any dist.  $D$  & any target  $c \in C$ , given access to  $\text{Ex}(c, D)$ , w.p.  $(1-\delta)$   $A$  produces a hypothesis with  $\text{error}(h, c, D) \leq \frac{1}{2} - \gamma$ .

If  $A$  runs in time  $\text{poly}(d, \frac{1}{\gamma})$  &  $\frac{1}{\gamma} > \frac{1}{\text{poly}(d)}$   
then  $C$  is efficiently weakly-PAC learnable

Thm If  $C$  is weakly-PAC learnable (efficiently),  
then  $C$  is PAC-learnable (efficiently).

Proof AdaBoost (Freund & Schapire)  
(Nobel prize '03, very influential  
in practice)