Last time: started generic bounds for online learning

☺ algs
- HA: $\log_2 |\mathcal{C}|$
- RHA: $\mathbb{E} \leq \ln |\mathcal{C}| + O(1)$, for oblivious adversary
- started to define VCDIM($\mathcal{C}$)

Today: ☹
- VCDIM($\mathcal{C}$) as lower bound for any OCMB alg.
- Predicting from Expert Advice
  - Weighted Majority alg
  - Randomized Weighted Majority alg

rebranded noise-tolerant HA, RHA

Reminder: PS1 due today, PS2 out today

Questions?

Back to VCDIM($\mathcal{C}$):

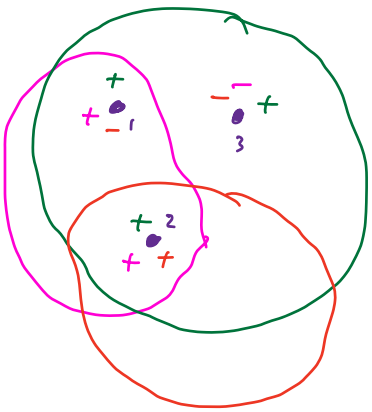
Def: Fix $\mathcal{C}$ over domain X.
Let $S \subseteq X$.

We say S is shattered by $\mathcal{C}$ if

- (subset POV) $\forall\, T \subseteq S$, some $c \in \mathcal{C}$ is s.t. $c \cap S = T$.

Equivalent def

- (labeling POV): for every Boolean labeling of S, i.e. every $f: S \to \{0,1\}$, some $c \in \mathcal{C}$ labels S that way.



all 8 such labellings are achieved by concepts in $\mathcal{C}$

Def: $VCDIM(\mathcal{C})$ = size of largest $S \subseteq X$ that is shattered by $\mathcal{C}$.

Equiv. def: $VCDIM(\mathcal{C})$ = smallest value d s.t. no set of $d+1$ ex. in X is shattered by $\mathcal{C}$.

If $\forall d \; \exists$ set S, $|S|=d$, s.t. $\mathcal{C}$ shatters S, then $VCDIM(\mathcal{C}) = \infty$.

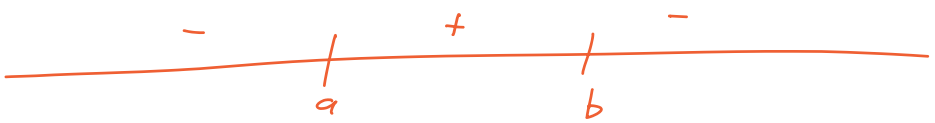
To show $VCDIM(\mathcal{C}) = d$, we

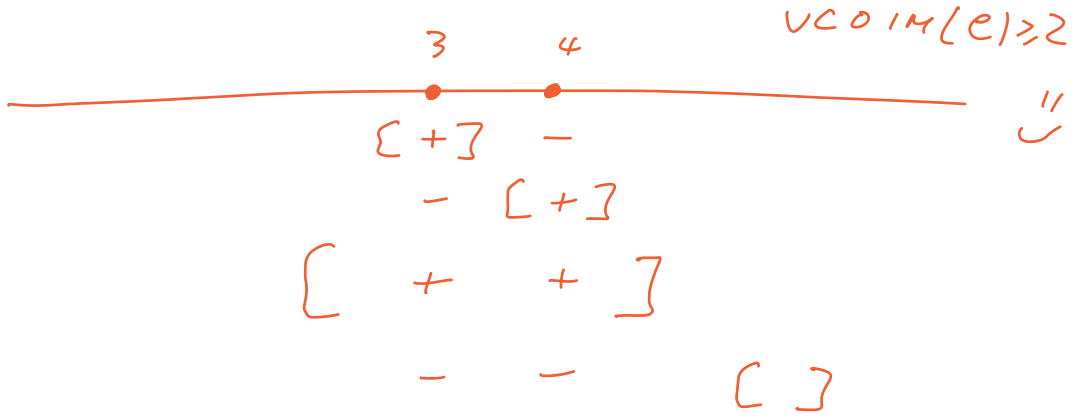
- exhibit a $S \subseteq X$, $|S|=d$, & argue that $\mathcal{C}$ shatters S (shows $VCDIM(\mathcal{C}) \geq d$);

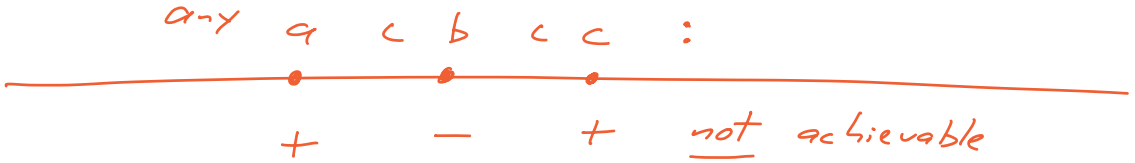
+

- argue that no $S \subseteq X$ with $|S| = d+1$ can be shattered by $\mathcal{C}$. (shows $VCDIM(\mathcal{C}) \leq d$.)

Ex1: $X = \mathbb{R}$, $\mathcal{C}$ = all closed intervals $[a,b]$.



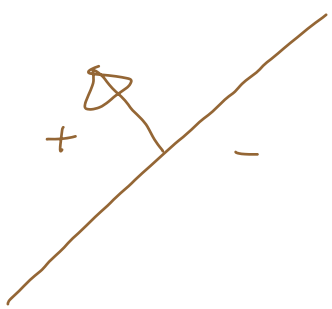




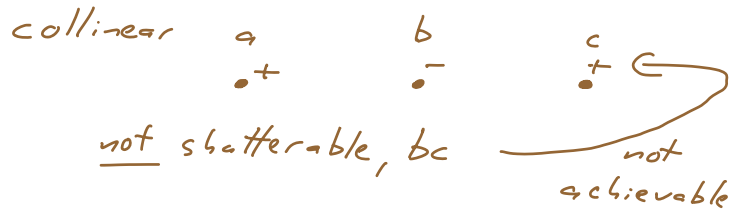
No set of 3 pts shatterable:

VCDIM(C) = 2.

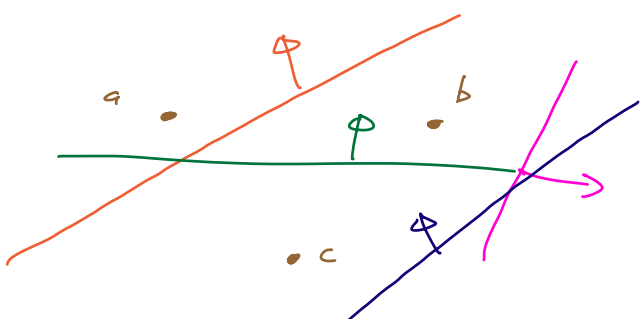
Ex 2: $X = \mathbb{R}^2$, C = all halfspaces (LTFs)



VCDIM(C) ≥ 3:



But

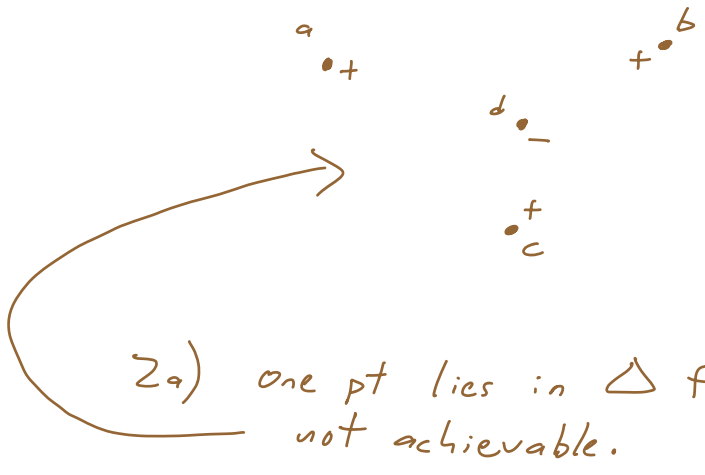


is shattered by LTFs.

Claim $VCDIM(\mathcal{C}) \leq 3$: must show no set of 4 points is shattered.

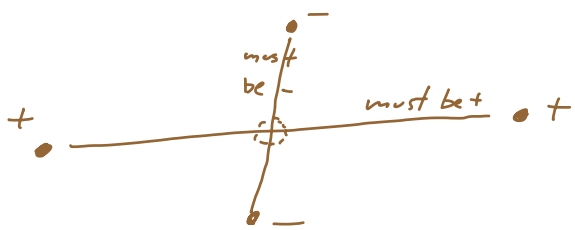
1) If 3 of the 4 pts collinear: not shattered.

2) no 3 pts collinear:



2a) one pt lies in △ formed by other 3: not achievable.

2b) the 4 points form a quadrilateral



contradiction: point of intersec. of diag's both + & −.

So $VCDIM(\mathcal{C}) = 3$.

Ex 3: $\mathcal{C}$ = all mon $\overset{disj}{conj}$ over $\{0,1\}^n$.

$VCDIM(\mathcal{C}) \geq n$: pf by ex.

$n=5$ The set $+-(0,1,1,1,1)+$
$++(1,0,1,1,1)-$ is shatt. by $\mathcal{C}$.
$++(1,1,0,1,1)-$
$++(1,1,1,0,1)+$
$++(1,1,1,1,0)+$

x_1 ϕ (empty conj)
$x_2 \wedge x_3$

etc.

$VCDIM(\mathcal{C}) \leq n$: $|\mathcal{C}| = 2^n$, +

$\forall\ \mathcal{C}$, have $VCDIM(\mathcal{C}) \leq \log_2 |\mathcal{C}|$.

(need 2^k concepts to shatter k pts.)

Back to OCMB learning:

Claim: if $VCDIM(\mathcal{C}) = d$, then any OCMB alg for $\mathcal{C}$ must have (worst-case) MB $\geq d$.

Pf: Let $S = \{x^1, \ldots, x^d\}$ be a size-d shatt. set.
On ex. seq. $x^1, x^2, \ldots, x^d$, no matter what bit $h_i(x^i)$ learner predicts, there is a $c \in \mathcal{C}$ st

$$c(x^i) \neq h_i(x^i) \quad \forall\ i = 1, \ldots, d.$$

So this means d mistakes. ▨

Note: • elim alg for mon conj / disj is optimal!

What about obliv. adv.?

Get l.b. of $\geq \frac{d}{2}$ expected mistakes:

adv. chooses ex seq to be $x^1, \ldots, x^d$, chooses c unif. at rand. from the 2^d concepts in $\mathcal{C}$ that shatter.

In this setting, learner's task is $\equiv$ to predicting a seq of d $\$$'s.
$\mathbb{E}[\#\text{mist}] = d/2$.

Prediction from Expert Advice:

Weighted Maj Alg (WMA)

Consider following scenario: You go to racetrack with group of friends.

Sequence of races.

Each race: each friend makes a bet.

You don't know horses: want to combine their predictions for each race to make your bet.

Your goal: do well!

- In hindsight: could have done as well as best friend.
- If all friends do badly, I can't expect to do well: no absolute guarantees. (best we can hope for.)

Turns out that using WMA, you can do almost as well as most successful friend does, without knowing in advance who will do well!

Setup for WMA:

- Have pool of N experts, sequence of trials

- Each expert makes 0/1 pred. at each trial

(Alg. has parameter $0 \leq \beta < 1$.)

- Each expert has a weight w_i; initially

all $w_i = 1$.

- At each trial, expert i predicts $z_i \in \{0,1\}$

WMA computes $q_0 := \sum_{i : z_i = 0} w_i$

$q_1 := \sum_{i : z_i = 1} w_i$

WMA predicts 0 if $q_0 \geq q_1$
1 if $q_0 < q_1$.

(☆) Update rule: given result of trial (true outcome), for each i s.t. z_i was wrong, set
$w_i \leftarrow \beta \cdot w_i$.

Observe: if $\beta = 0$ then this is HA:

expert i $\longleftrightarrow$ i^{th} concept in $\mathcal{C}$, c_i, $|\mathcal{C}| = N$

pred. of expert i on trial j $\longleftrightarrow$ $c_i(x^j)$ (j^{th} ex in ex seq)

Halving alg flops with noise - this doesn't!

Thm: For any seq of trials, suppose best expert in pool makes m mistakes. Then WMA makes

$$M \le \frac{\log N + m \cdot \log \frac{1}{\beta}}{\log \frac{2}{1+\beta}}$$

many mistakes.

$\beta = \frac{1}{2}$: $2.41(m + \ln N)$

$\beta = .75$: $2.2m + 5.2 \ln N$

$\beta \to 1$: $\simeq 2m + \frac{2}{\varepsilon} \cdot \ln N$

$\| \; 1-\varepsilon, \varepsilon \to 0$

Pf: Let $W = q_0 + q_1$ = tot wt of all experts.

Initially $W = N$.

At each mistake: at least half the total wt W predicts wrong & is $\cdot \beta$, so after a mistake, tot wt goes from

$$W \quad \text{to} \quad \le \left(\frac{W}{2}\right) + \left(\frac{W}{2}\right) \cdot \beta = \left(\frac{1+\beta}{2}\right) \cdot W$$

So after M mist., have

$$W \le N \cdot \left(\frac{1+\beta}{2}\right)^M.$$

$\frac{1+\beta}{2} < 1$

OTOH, best expert makes m mist.,

so her wt is $\ge \beta^m$ always. So

$$W \ge \text{her wt} \ge \beta^m.$$

So $\beta^m \le W \le N\left(\frac{1+\beta}{2}\right)^M$:

solve for m, & get

$$m \log \beta \le \log N + M \cdot \log \frac{1+\beta}{2}$$

$$m \log \tfrac{1}{\beta} \ge -\log N + M \log \frac{2}{1+\beta}$$

$$\frac{m \log \frac{1}{\beta} + \log N}{\log\left(\frac{2}{1+\beta}\right)} \ge M. \quad \blacksquare$$

$\frac{2}{1+\beta} > 1$, $\log \frac{2}{1+\beta} > 0$

Next time: randomized WMA

new unit : PAC Learning.