Last time:
- example of using PCT (King + advisors, $\text{sign}(\pm x_1 \pm x_2 \cdots \pm x_n)$, n^2 mistake bound)
- dual Perceptron, feature expansions, kernel functions
  $\rightsquigarrow$ running Perceptron over high-dimensional feature spaces in a computationally efficient way

Today: Start generic bounds for OLMB learning: algs + LBs that apply to any finite $\mathcal{C}$

positive results
- Halving Algorithm: m.b. $\le \log_2 |\mathcal{C}|$
- Randomized Halving Alg: $\mathbb{E}[\text{mistakes}] \le \ln |\mathcal{C}| + O(1)$ under "oblivious adversary" assumption

negative result
- VC dimension

Reminder: PS 1 due on Wed

Questions?

$\mathcal{C}$ = any finite concept class

Halving Algorithm

Let CONSIST = set of all $c \in \mathcal{C}$ that are consistent w/ all lab. ex. $(x, c(x))$ seen so far.

(Initially CONSIST $= \mathcal{C}$).

$\longrightarrow$ Given ex x: HA's prediction is maj vote over all $c \in$ CONSIST.

E.g. if $|\text{CONSIST}| = \underline{2000}$ & 1150 say 1 on x
850 " 0 " x,

HA would predict 1.

On every example mistake: update CONSIST. [E.g. if $c(x) = 0$, update & now $|\text{CONSIST}| = 850$.]

Thm: For any finite $\mathcal{C}$, m.b. of HA $\le \log_2 |\mathcal{C}|$.

Pf: Each mist. cuts CONSIST by $\geq \frac{1}{2}$. (mult. size by some $0 < p \le \frac{1}{2}$).

Goes from $|\mathcal{C}|$ to ≥ 1. (target always is in CONSIST), so $\blacksquare$.

Ex: $\mathcal{C} = \{$1-DL of length r, over $\{0,1\}^n\}$

$$|\mathcal{C}| \le (4n)^r \cdot 2 \le 2^{O(r)} n^{O(r)}$$

so $\log |\mathcal{C}| \le O(r \cdot \log n)$.

Time per trial: $2^{\Theta(r)} n^{\Theta(r)}$.

Discuss HA:

:) Great MB!

:($\underline{\text{Not}}$ comput. efficient: need to spend time $\approx |\mathcal{C}|$ per trial to maintain CONSIST + det. maj vote.

:(Uses weirdish hyp: $MAJ(c_{i_1}, c_{i_2}, \ldots, c_{i_M})$

we'll fix this: RHA

$c_{i_j} \in \mathcal{C}$

:(Brittle if noise

we'll fix this too: Weighted Maj

More obs abt HA:

- Some $\mathcal{C}$: HA is best poss MB.

Let X = any finite set

$\mathcal{C}$ = $\underline{\text{all}}$ fns $X \to \{0,1\}$ $\qquad 2^{|X|}$ conc. in $\mathcal{C}$.

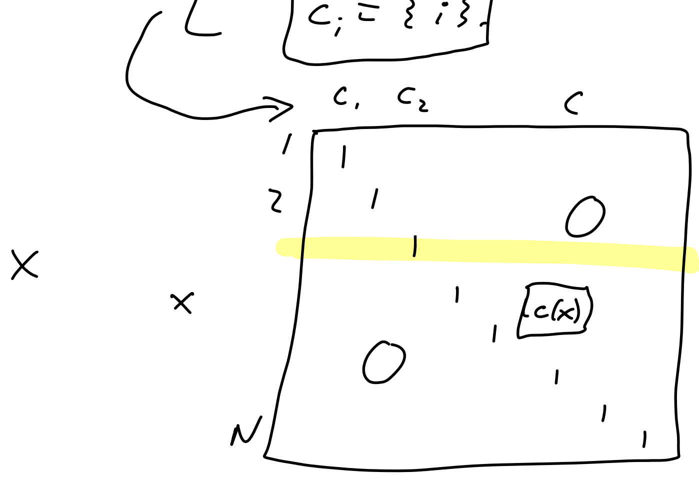
HA mb $= \log_2 |\mathcal{C}| = \log_2 2^{|X|} = |X|$,

+ nomatter what any alg says on each $x \in X$, could be wrong: no alg has m.b. $\leq |X| - 1$.

- Other $\mathcal{C}$'s: HA does $\underline{\text{better}}$ than $\log |\mathcal{C}|$.

mistakes.

$X = \{1, \dots, N\}$
C has a c_i for each $i \in [N]$: $\boxed{c_i = \{i\}}$



Initial pred. is 0;
one mist. elim. all but target C.

Randomized Halving Alg. (RHA)

Motiv.: Our OLMB def. is very "worst-case": equiv. to assuming seq of ex. $\underline{x^1, x^2, x^3, \dots \in X}$ + c (target concept) are chosen by malicious omniscient foe; can even change c as long as always there's some concept in class consistent w/ all lab. ex seen so far.

<u>Paranoid?</u>

Here's a relaxed assumption:

target $c \in \mathcal{C}$, ex seq $x^1, x^2, \ldots \in X$

chosen once + for all before learning process starts.

"oblivious adversary."

In this setting, learner can (try to) use randomness.

$\boxed{\text{Assume obliv. adv.}}$

In this setting, here's RHA (for any finite $\mathcal{C}$):

- Maintain CONSIST like before
- Each time we make a mistake $x \in X$, update hyp h to be a uniform random concept in CONSIST.

Thm: For $\overbrace{\text{any fixed } c \in \mathcal{C}\text{, seq } x^1, \ldots \text{ of ex}}^{\text{(obliv. adv. assump.)}}$,

$$\mathbb{E}[\#\text{mistakes RHA makes}] \leq \ln|\mathcal{C}| + O(1).$$

Pf: Fix c, seq $x^1, x^2, \ldots$

Consider pt in exec. of RHA when $|\text{CONSIST}| = r$.

Let $M_r = \mathbb{E}[\#$ mist. RHA makes from this point on, i.e. on rest of this ex. seq. + this $c]$.

Note $M_{|\mathcal{C}|} = \mathbb{E}[\#$ mistakes RHA makes$]$
= what we want to upper bound.

[Thought experiment (not in alg.):

Order concepts in CONSIST by when the remaining seq of ex. + c eliminate them.

$$\underbrace{\overbrace{c_1, c_2, c_3, c_4}^{\text{next ex. elim. these}}, \overbrace{c_5, \ldots, c_{13}}^{\text{subseq. ex. elim. them.}}, \ldots}\quad \boxed{c_r = \text{target } c}$$

Well-defined, by obliv. adv. assump.]

Consider pt in RHA's exec. when h (unif over $c_1, \ldots, c_r$) is chosen.

- If h chosen to be c_r: 0 more mist. :) ($\frac{1}{r}$ chance)
- 〃 〃 〃 〃 〃 c_t, some $t < r$:

One mistake eliminates $c_1, \ldots, c_t$ (+ maybe more)
Then (at most) $r-t$ concepts left;
$\mathbb{E}\{\#\text{ mist. then on}\} \le M_{r-t}$.

So

$$M_r \le \frac{1}{r}\cdot 0 + \frac{1}{r}\cdot\sum_{t=1}^{r-1}\left(1 + M_{r-t}\right).$$

Replace w/ $M_r = \frac{1}{r}\cdot 0 + \frac{1}{r}\cdot\sum_{t=1}^{r-1}\left(1 + M_{r-t}\right)$.

$$\boxed{M_r = \frac{r-1}{r} + \frac{1}{r}\left(M_1 + \ldots + M_{r-1}\right)}, \quad \boxed{M_1 = 0}$$

Solve:

$$r M_r = r-1 + \left(M_1 + \ldots + M_{r-1}\right) \qquad (\star)$$

Apply to $r-1$ instead of r:

$$(r-1)M_{r-1} = r-2 + \left(M_1 + \ldots + M_{r-2}\right) \qquad (\star)$$

Subtract:

$$r M_r - (r-1)M_{r-1} = (r-1)-(r-2) + M_{r-1}$$

$r(M_r - M_{r-1}) = 1$, i.e $r M_r = r M_{r-1} + 1$

$$M_r = M_{r-1} + \frac{1}{r}$$

$M_2 = \frac{1}{2}$, $M_3 = M_2 + \frac{1}{3} = \frac{1}{2} + \frac{1}{3}$, $M_4 = \frac{1}{2} + \frac{1}{3} + \frac{1}{4}$,

$$M_r = \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{r}$$

so $M_r \le \ln r + O(1)$

$+\ M_{|\mathcal{C}|} \le \ln|\mathcal{C}| + O(1).$ ∎

Discussion:

🙂 $\ln|\mathcal{C}| < \log_2 |\mathcal{C}|$

🙂 hyp. simpler (an elt of $\mathcal{C}$, not maj vote over subset of $\mathcal{C}$)

🙂 easier to predict $h(x)$ than would be for HA

🙁 we assumed obliv. adv.

🙁 still computationally inefficient to maintain CONSIST.

Lower bound for OLMB learning:

Vapnik-Chervonenkis dimension

(VC dimension)

VCDIM($\mathcal{C}$): a # ; (combin. parameter of $\mathcal{C}$).

a measure of "how complex" $\mathcal{C}$ is.

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.

Ex: $X = \{1,2,3,4,5\}$

Set $S = \{5\}$ is shattered:

$\mathcal{C} = \{c_1, \dots, c_6\}$

$c_1 = \{1,2,3\}$

$c_4 = \{1,2,5\}$ 2

$c_2 = \{2,4,5\}$ 2,4

$c_5 = \{1,3,5\}$ $\emptyset$

$c_3 = \{3,4\}$ 4

$c_6 = \{5\}$

Set $S = \{2,4\}$ also shattered:

Obs: no set of size 3 is shattered by this $\mathcal{C}$: $|\mathcal{C}| = 6 < 2^3$.

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