Last time: • finish PCT proof; example of using PCT
• dual Perceptron, kernelization

End of our "concrete examples" for OLMB of specific $\mathcal{C}$'s.

Today: Start generic bounds: Algorithms & lower bounds that apply to any (finite) concept class. [Give up on comput. eff...]

pos results:
- Halving algorithm: $\log_2 |\mathcal{C}|$ mistake bound
- Randomized halving alg.: $\mathbb{E}[\#\text{mistakes}] \le \ln|\mathcal{C}| + O(1)$ (under "oblivious adversary" assumption)

neg. result:
- Start discussion of VC dimension of $\mathcal{C}$

Note: PS1 due on Tues,
PS2 out " "

Questions?

Halving Algorithm (HA)

Fix finite $\mathcal{C} = \{c_1, \ldots, c_N\}$ over X.

- Define CONS = all $c \in \mathcal{C}$ that are consistent w/ seq. of lab. ex. so far.
  Initially CONS = $\mathcal{C}$.
- Given example x, HA predicts according to maj of $c \in$ CONS.
  Ex: $|CONS| = 2000$
  $c(x) = 1$ for 1100 of them,
  $c(x) = 0$ " 900 " ".
  HA would output 1 on this x.
- On mistake: update CONS based on true $c(x)$.

If $c(x) = 0$, mistake;
new $|CONS| = 900$.

Thm: for any finite $\mathcal{C}$, HA's mistake bd is $\leq \log_2 |\mathcal{C}|$.

Pf: Every mistake causes $|\mathcal{C}| \to p|\mathcal{C}|$ some $p \leq \frac{1}{2}$; target c never leaves CONS so $|CONS| \geq 1$; so # mist. $\leq \log_2 |\mathcal{C}|$.

Discussion:

- Drawbacks:
  - On each ex x, must compute $c(x)$ for each $c \in CONS$; initially $|CONS| = |\mathcal{C}|$, so can be slow...

  - Weird/complex hyp: $h(x) = \underset{c \in CONS}{MAJ}(c(x))$
    (we'll fix this w/ RHA...)
  - Brittle if noise (we'll fix this with "WM")

- Ex: back to length-r OL's over $\{0,1\}^n$.
  $|\mathcal{C}| \approx (4n)^r$ $\Rightarrow$ HA's mist. bd. is $\leq \log|\mathcal{C}| = O(r \cdot \log n)$.

Alas, runtime per trial $\approx n^r$.

SOTA for length-r OL:

	Our first alg	Winnow 2	HA	??? Dream
Runtime per trial	$O(n)$ time ☺	$O(n)$ ☺	$\approx n^r$ ☹	$O(n)$ or $\text{poly}(n)$ ☺
Mistake bd	$O(n \cdot r)$ ~	$2^{O(r)} \cdot \log n$ ☺	$O(r \cdot \log n)$	$O(r \cdot \log n)$ ☺

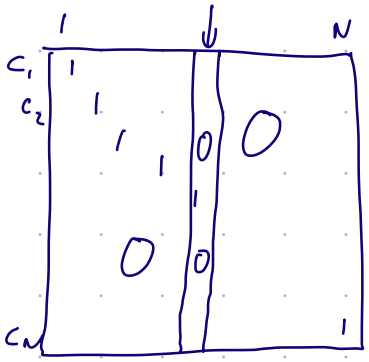
- For some $\mathcal{C}$'s, HA is optimal!

  Ex: $X = \{1, \dots, N\}$
  $\mathcal{C}$ = all 2^N fns $X \to \{0,1\}$.
  HA will make N mistakes,
  no alg can do better.

- Sometimes HA makes $\ll \log|\mathcal{C}|$ mistakes.

  Ex: $X = \{1, \dots, N\}$ $\mathcal{C} = \{\{1\}, \{2\}, \dots, \{N\}\}$

$c_i(x) = \begin{cases} 1 & x = i \\ 0 & x \neq i \end{cases}$



HA predicts 0 until makes 1st mistake; then $|\text{CONS}| = 1$ ☺.

Randomized Halving Alg. (RHA)

MB criterion: very worst-case.

↳ ≡ to assuming ex seq + lab. of ex. are generated by "omniscient online adversary": at trial i, "knows" everything abt learner, can use any $c \in \mathcal{C}$ as target, provided c is cons. w/ all $i-1$ previous labeled examples.

Too Paranoid?

Let's relax, + assume

- target c, +
- seq. of examples

} fixed once + for all before start of seq. of trials.

"Oblivious adversary".

↳ Learner can hope to use randomness.

RHA: for obliv. adv. setting.

- Updates its h only on mistake
- Its h = a uniform random concept in CONS.

(Update CONS at each mistake as before.)

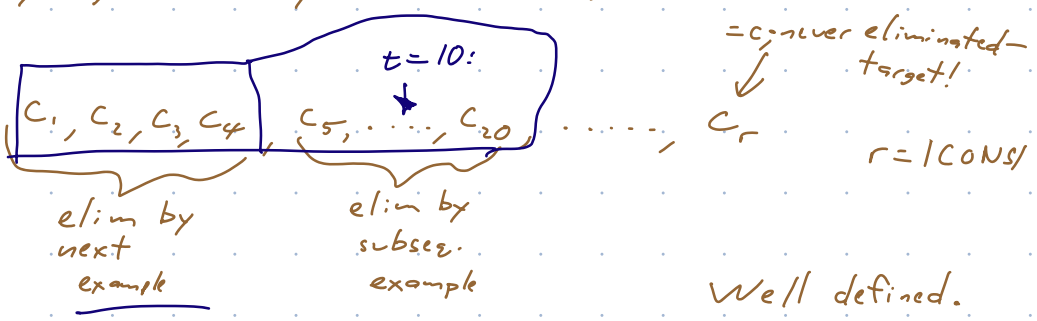
Thm: For any fixed target $c \in \mathcal{C}$, (obliv. adv. setup),
any fixed seq of examples,

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

Pf: Have fixed c, seq of ex $x^1, x^2, \ldots$.

Suppose at a pt when $|CONS| = r$.
Let $M_r := \mathbb{E}[\#\text{mist. RHA will make from this moment on}]$.
So $M_{|\mathcal{C}|} =$; want to bd $M_{|\mathcal{C}|}$.

Thought experiment: order elts of CONS by when they'll get elim. by c & remaining examples



Well defined.

When Alg chooses its rand $h \in \{c_1, \ldots, c_r\}$: what happens?
$h = c_i$ all =ly likely.

- if $h = c_r$ ($\frac{1}{r}$ prob): 0 no more mist! ☺
- if $h = c_t$ some $t < r$ ($\frac{r-1}{r}$): 1 mist. to elim. $c_1, \ldots, c_t$ (& even rest of c_t's block); ($\frac{1}{r}$ chance for each $c_1, \ldots, c_{r-1}$)
  at most $c_{t+1}, \ldots, c_r$ left, so $\mathbb{E}[\#\text{mist. from then on}] \le M_{r-t}$

So have: $M_r \le \frac{1}{r} \cdot 0 + \sum_{t=1}^{r-1} \frac{1}{r}\left[1 + M_{r-t}\right]$

Worst case: $M_r = \sum_{t=1}^{r-1} \frac{1}{r}(1 + M_{r-t})$

$$M_r = \frac{r-1}{r} + \frac{M_1 + M_2 + \dots + M_{r-1}}{r}$$

$$r \cdot M_r = r-1 + \boxed{M_1 + \dots + M_{r-2}} + M_{r-1}$$

Do $r-1$ case:

$$(r-1) M_{r-1} = r-2 + \boxed{M_1 + \dots + M_{r-2}}$$

Subtract:

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

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

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

$M_1 = 0$ " if $r=1$

$M_2 = \frac{1}{2}$ $\quad M_3 = \frac{1}{3} + \frac{1}{2}$ etc

$$M_r = \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{r} \leq \ln r + O(1)$$

$$M_{|e|} \leq \ln |e| + O(1). \qquad \blacksquare$$

Adv of RHA:

+ simpler hyp: c_i
+ $\ln |e|$ better than $\log_2 |e|$

Disadv:

- needed obliv. adv. assump.
- still hard to run in general.

So far... pos results.

Now:

Lower bound on OCMB learning based on "VC Dimension"

Vapnik Chervonenkis

VC Dim: • a # assoc. w/ a concept class $\mathcal{C}$.

$VC(\mathcal{C})$ $VCDIM(\mathcal{C})$

• way of measuring "how complex" (expressive) the the concept class.

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

S is shattered by $\mathcal{C}$ means:

(subset POV):

$\forall U \subseteq S$ $\exists$ some $c \in \mathcal{C}$ (subset of X) s.t. $c \cap S = U$.

(labeling POV)

For every labeling of pts in S by 0/1, some $c \in \mathcal{C}$ induces that labeling.

Ex: $X = \{1, \ldots 5\}$

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

$c_1 = \{1,2,3\}$ 2_ $c_1 \cap S = \{2\}$

$c_2 = \{2,4,5\}$ 24 $c_2 \cap S = \{2,4\}$

$c_3 = \{3,4\}$ _4 $c_3 \cap S = \{4\}$

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

$c_5 = \{1,3,5\}$ __ $c_5 \cap S = \emptyset$

$c_6 = \{5\}$.

This $\mathcal{C}$ shatters $S = \{2,4\}$.

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