Last time:
  - Start generic bounds for online learning
    - Halving alg positive results
    - Randomized HA for oblivious adversaries

Today:
  - VC Dimension negative results
  - Predicting from Expert Advice
    - Weighted Majority alg
    - Randomized Weighted Majority alg
    (Next time: Probably Approximately Correct (PAC) learning)

Reminder: PS1 due Thurs, PS2 out Thurs

Questions?

Recall

**Def**: Fix C over X. Let S \subseteq X.

S is shattered by C if \forall U \subseteq S, \exists c \in C s.t.

\[ c \cap S = U \] (subset P.O.V.)

equivalently,

\[ c(x) = 1 \quad x \in U \]
\[ c(x) = 0 \quad x \in (S \setminus U) \]

"No matter how you want to label the pts in S, some c \in C does the job"

Ex: \[ X = \{1, 2, 3, 4, 5\} \]
$E = \{c_1, \ldots, c_6\}$:

\[ c_1 = \{1, 2, 3\} \]
\[ c_2 = \{2, 4, 5\} \]
\[ c_3 = \{3, 4\} \]
\[ c_4 = \{1, 2, 5\} \]
\[ c_5 = \{1, 3, 5\} \]
\[ c_6 = \{5\} \]

This $E$ shatters \{5\}.
Also shatt. \{3, 4, 5\}
\[ \text{vc01m}(E) = 2 \]

**Def:**

\[ \text{vc01m}(E) = \text{size of largest } S \subseteq X \text{ that is shattered by } E. \]

**Eqv:**

\[ \text{vc01m}(E) = \text{smallest } d \text{ s.t. no set of } d+1 \text{ pts is shattered by } E. \]

\[ If \forall d \exists S \text{ s.t. } |S| = d + 1 \text{ and } S \text{ is shattered by } E:\]

\[ \text{then } \text{vc01m}(E) = \infty. \]

To show $\text{vc01m}(E) = d$:

- exhibit set of $d$ pts that are shatt by $E$;
- argue no set of $d+1$ pts is shattered by $E$.

**Ex:** $X = \mathbb{R}$. $E = \text{all closed intervals } [a, b]$. $c_a, c_b(x) = 1$ if $a \leq x \leq b$.
\[ [a, b] \not\subset 0 \text{ o/w.} \]

\[
\begin{cases}
\text{VC01M}(c) \geq 2: & \text{\underline{\underline{\underline{\underline{\underline{\underline{\underline{3}}}}}}}} \quad \underline{\underline{\underline{\underline{\underline{\underline{\underline{7}}}}}}}
\
\text{VC01M} = 2
\
\text{VC01M}(c) < 3:
\end{cases}
\]

no interval contains \( x \) or \( z \) but not \( y \)

\[ \text{Ex: } X = \mathbb{R}^2, \; C = \text{LTFs over } X. \]

\[
\begin{array}{ccccc}
+ & \uparrow & + & - & -
\end{array}
\]

\[
\begin{array}{cccc}
\text{VC01M}(c) \geq 3: & \text{any 3 non-collinear pts are shattered.}
\end{array}
\]

\[
\begin{array}{cccc}
\text{VC01M}(c) < 4: & \text{pick any 4 pts } P_1, P_2, P_3, P_4 \in \mathbb{R}^2.
\end{array}
\]

\[
\begin{array}{cccc}
\text{case 1: } & 3 \text{ pts are collinear.} & 3 \text{ pts are collinear.} & \text{unachievable}
\end{array}
\]

\[
\begin{array}{ccc}
& P_1 & P_2 \quad P_3 \quad P_4
\end{array}
\]

\[
\begin{array}{cc}
\text{case 2: } & 7 \text{ no 3 pts collinear.} & \text{\triangle by } P_1, P_2, P_3
\end{array}
\]

\[
\begin{array}{c}
P_1
\end{array}
\]

\[
\begin{array}{c}
Z_9: \text{ in } \triangle
\end{array}
\]
Zb: $p_4$ outside $\Delta$. not achievable.

So $\text{VCDIM} = 3$.

Fact: $X = \mathbb{R}^n$, $C = \text{LTF}_3$ over $X$:
$\text{VCDIM}(C) = n+1$

Ex: $C = \text{mon. conj.}$ $X = \{0,1\}^n$
$\text{VCDIM}(C) \geq n$:
$n = 5$: can shatter

$c(x) = x \land x_2$ achieves
$c(x) = \text{empty conj} \equiv 1$

$\text{VCDIM}(C) < n+1$:

FACT: $\text{VCDIM}(C) \leq \log_2 |C|$. 
(Need $2^k$ concepts in $C$ to shatter set of $k$ pts.)

$|E| = 2^n$ so $VOCOMB(E) \leq n$.

Connection to OCMB learning:

**Lemma:** If $VOCOMB(E) = d$, then any OCMB alg for $E$ must have (worst-case) mist. bound $\geq d$.

**Pf:** Let \( \{x', \ldots, x^d\} \) be set shatt. by $E$.

Adv. gives example seq $x', x^2, \ldots, x^d$

Learner says $b \in \{0, 1\}$ as its pred. for $c(x^i)$:

Adv. says “wrong; $c(x^i) = \overline{b}$.”

*Since shattered, this is a valid labelling for some concept in $E$.*

(So, clin alg has best poss MB for non-cov.)

What about obliv. adv.? He can give shatt. set as ex seq, $d$ coin tosses to choose labels “once for all” in advance; learner’s job is predicting $d$ coin tosses; $E[\# mistakes] = d/2$.

"Predicting from Expert Advice" rebranded online learning "without the xi"
Weighed Maj. Algorithm ("noise-tolerant H.A.")

Setting: go to racetrack with group of friends
Seq. of horse races (oppty's for prediction)
Friend have bets on each race
Q: how to best aggregate friends' predictions to make your predictions?

- No absolute guarantee possible (experts do badly)

- But: can guarantee you do almost as well, at end of day, as most successful expert -- without knowing in advance who it is!

Weighted Maj. Alg. (WMA):

Scenario: pool of N experts
  seq. of trials
  Each expert predicts 0/1 in each trial.

Algorithm has parameter \( \beta < 1 \).

Each expert i has weight \( w_i \).

• Initialize: \( w_i = 1 \ \forall i \in [N] \).
At each trial, expert $i$ predicts $z_i \in \{0, 1\}$.

Let $g_0 = \sum_{i: z_i = 0} w_i$, $g_1 = \sum_{i: z_i = 1} w_i$.

WMA predicts 0 if $g_0 \geq g_1$, 1 otherwise.

Given outcome of trial, for each $i$ s.t. $z_i$ was wrong, set $w_i \leftarrow w_i \cdot \beta$ (penalty for being wrong).

Note: if $\beta = 0$, this is H.A.: remove on expert as soon as a mistake.

$\text{exp. } i \iff \text{i}^{th} \text{ concept in C}$

$pred. z_i \text{ on } t^{th} \text{ trial} \iff c_i(x), x = t^{th} \text{ example.}$

Like OLMG learning with $1 \leq N, \forall$ no concept in $C$ is expected to be perfect.

Thus: For any seq. of trials: Suppose best expert in pool makes $m$ mistakes.

Then WMA makes $M \leq \frac{\log N + m \cdot \log(1/\beta)}{\log(\frac{2}{1+\beta})}$ many mistakes.
\[
\begin{align*}
\text{(Ex: if } \beta &= \frac{1}{e} : \ L & \approx 2.41 \cdot (m + 1nN) \\
\beta &= \frac{3}{4} : \ 2.2m + 5.2hN \\
\beta &= 1- \nu, \ \nu \to 0 : \approx Zm + \frac{2}{e} \cdot hN
\end{align*}
\]

\text{Pf of thm:}

Let \( W \) = tot wt of all experts.

Initially \( W = N \).

At a given trial, \( W = g_0 + g \).

At each mistake, at least half of tot wt predicted wrong and gets \( x \beta \).

So after a mistake, tot wt went from \( W \) to at most \( \frac{W}{2} + \frac{W}{2} \beta = \left( \frac{1+\beta}{2} \right) W \).

So after \( M \) mistakes (if weighted majority made \( M \) mistakes), have

\[
W \leq N \cdot \left( \frac{1+\beta}{2} \right)^M
\]

On other hand, since best expert only makes \( m \) mistakes, \( W \) always \( \geq \beta^m \). So tot wt \( W \) always \( \geq \beta^m \).

\[
\beta^m \leq W \leq N \cdot \left( \frac{1+\beta}{2} \right)^M
\]
take logs:

\[ m \log B \leq \log N + M \log \frac{1 + \beta}{2} \]

\[ m \log \frac{1}{3} \geq -\log N + M \frac{2}{\log 1 + \beta} \]

\[ \frac{m \log \frac{1}{3} + \log N}{\log \frac{2}{1 + \beta}} > M. \]

Next time: Rand WMA

end of Chebyshev unit.

start PAC Learning unit.