Last time:
- Randomized Weighted Majority alg.
- Start PAC Learning unit
  - motivation
  - basic defs
  - learning \( C = \) intervals of \( \mathbb{R} \)

Today:
- Finish
  - \( OLM \rightarrow PAC \) conversion
  - Revisit PAC learning def.
    - "size" of concepts
    - effiency of evol. \( h \)
    - "Chernoff bounds" for hypoth. testing

Questions?

Ex: PAC Learning Intervals

\[ X = [0, 1] \cdot C = \text{all closed intervals } [a, b] \]

PAC learning: there's some \( D \) over \( X \), unknown to us.

Target interval \( C = [a, b] \).

\[ m = \frac{2}{3} \cdot \frac{1}{\gamma (\frac{3}{2})} \]
Alg A:  
- draw in examples

Let $a' = \text{smallest pos ex}$
$\quad b' = \text{largest pos ex.}$

A outputs hyp $h = [a', b'].$

Total error of $h$ on $c$ w.r.t. $y = \text{area of}$

A only makes false neg. errors.
To show: w.h.p., $h$ "doesn't miss much mass".

Intuition: unlikely to miss "big" region;
if don't miss "big" region, error is "small".

Let $a, b$ s.t. $Pr \{c \in [a, b]\} = \frac{\epsilon}{2}$, $b, b'$ be s.t.

If alg's sample contains $\geq 1$ ex. in $L$
$\quad > 1$ ex in $R$

$\Rightarrow \text{total error of hyp } h \leq \epsilon$.
Goal: \[ \text{u.b. } \Pr[\text{all n ex. miss L}], \]
\[ \Pr[\text{all n ex. miss R}]. \]

For a single \( x \sim \mathcal{D} \), \( \Pr[\text{x miss L}] = 1 - \frac{\epsilon}{2} \).

Independence \( \Rightarrow \) \( \Pr[\text{u ex. all miss L}] = (1 - \frac{\epsilon}{2})^m \).

Union bd:
\[ \Pr[\text{u ex. all miss L or all miss R}] \leq 2(1 - \frac{\epsilon}{2})^m. \]

What \( m \) is s.t. \( 2(1 - \frac{\epsilon}{2})^m \leq \delta \)?

Recall:
\[ 1 - x \leq e^{-x} \]
So
\[ 2(1 - \frac{\epsilon}{2})^m \leq 2e^{-m/2} \]
So
\[ m \geq \frac{2}{\epsilon} \cdot \ln \left( \frac{2}{\delta} \right) \text{ is enough.} \]

Note: \( \delta \) is "confidence" param.
\( \epsilon \) is prob. over random sample used to train learner.
Protects against bad/typical samples.

PAC Learning vs OLMDB Learning.
Thm: Let A be OCMC alg for c.c. $E$ w. m.b. M. Then there is a PAC alg for $E$ which uses at most

$$m = M + \frac{M+1}{\epsilon} \log \left( \frac{M+1}{\delta} \right)$$

examples to $(\epsilon, \delta)$-PAC learn.

**Pf:** WLOG, assume alg $A$ is lazy (only changes hyp if makes a mistake.)

**PAC alg:**
- Run $A$ on set of ex from $EX(c, \theta)$.
- If hyp $h$ of $A$ ever goes for many consec. ex. from $\gamma$ making mistake, stop & output $h$.

Indeed this alg. uses $\leq m$ ex. from $EX(c, \theta)$:

$$\checkmark \checkmark \ldots \checkmark \times \checkmark \checkmark \ldots \checkmark \times \checkmark \ldots$$

$h$ was right $\uparrow$ right $\uparrow$ wrong

- $\leq M$ many $\times$'s
- every run of $\checkmark$ is $\leq \frac{1}{\epsilon} \ln \left( \frac{M+1}{\delta} \right)$ (b/c alg stops then)
- $\leq M+1$ runs of $\checkmark \ldots$
So \( \text{tot } \# \text{ex} \leq M + (M+1) \cdot \frac{1}{\epsilon} \ln \left( \frac{M+1}{\delta} \right) \).

Why is this \((\epsilon, \delta)\)-PAC?

**Def:** Say \( h \) is bad if \( \epsilon_{\rho}(h, c) > \epsilon \).

Fix any bad \( h \). For such an \( h \),

\[
\Pr \left[ \text{it gets } k \text{ consec. ex. all right} \right] < (1 - \epsilon)^k.
\]

\[
k = \frac{1}{\epsilon} \ln \left( \frac{M+1}{\delta} \right); \quad (1 - \epsilon)^k \leq \frac{\delta}{M+1}.
\]

There are \( \leq M+1 \) bad \( h \)'s ever in the mix for the alg (\( \leq M+1 \) h's overall are ever used), so

\[
\Pr \left[ \text{any of } \leq M+1 \text{ bad h's gets output} \right] \leq (M+1) \cdot \frac{\delta}{M+1} = \delta,
\]

\( \text{u.b.} \)

**Ex:** now know a PAC alg for \( E = \text{disj. over } S_q, B^m \) with sample complexity \( M=n+1 \)

\[
(n+1) + \frac{(n+2)}{\epsilon} \cdot \ln \left( \frac{n+2}{\delta} \right)
\]

\[
= O \left( \frac{n}{\epsilon} \cdot \log \left( \frac{n}{\delta} \right) \right).
\]
So... $C$ OLCMB learnable

\[ \iff \quad \exists \text{ ? No} \]

$C$ PAC learnable.

- over $\mathbb{R}$ or $\{0, 1\}$
- $C =$ intervals is *very efficiently* PAC learnable $\lor O(\frac{1}{\epsilon} \cdot \ln \frac{1}{\delta})$ many ex.
- but not OLCMB learnable $\lor$ any finite m.b.

Revisit def of PAC learning.
See book for details.

2 new issues:

1) "size" of concept $c \in C$.
   For some $C$'s, there is a notion of the "size" of $c \in C$.

**Ex 2:** DNF formulas. $X = \{0, 1\}$
- $C =$ all DNF formulas, is really $C =$ all $f : \{0, 1\}^n \rightarrow \{0, 1\}$.
But some $c$ are very hard to write as DNFs.
For any $c: \{0,1\}^n \to \{0,1\}$ define

$$\text{DNF}_{\text{size}}(c) := \min \text{ # of terms in any DNF for } c.$$ 

A fn like $x_1$ is a 1-term DNF; also expressible as, e.g.

$$x_1, x_2 \bar{x}_3 \lor \bar{x}_1 \bar{x}_2 \bar{x}_3 \lor \bar{x}_1 \bar{x}_3 \lor x_1 \bar{x}_2 \bar{x}_3.$$ 

**Ex 2:** $C = \text{all decision trees over } \{0,1\}^n$

Every $f: \{0,1\}^n \to \{0,1\}$ has some DNF rep,

$$\text{DNF}_{\text{size}}(c) = \min \text{ # internal nodes of all DNFs}$$

computing $c$

When PAC learning $C$ with a notion of size: if target $c$ has $\text{size}(c) = s$,
we allow PAC learning alg's runtime, sample complexity to depend on $s$.

- Efficient alg: $\text{poly}\left(s \frac{1}{\epsilon}, \frac{1}{\delta}\right)$ (if $\text{size}(c)$ or $\text{R}$)

- $A$ gets $s = \text{size}(c), \epsilon, \delta$ as input params.
2) complexity of hyp $h$.

What hyp classes $\mathcal{H}$ are "reasonable"?

(Recall $h \in \mathcal{H}$ is a program.)

**Definition:**

A hyp. class $\mathcal{H}$ over $X = \mathbb{R}^n$ or $\{0,1\}^n$ is **polynomially evaluable** if we can evaluate $h$ on $x$ (run the program) in $\text{poly}(n)$ time.

An efficient PAC alg. must use a polynomially evaluable $\mathcal{H}$.

$h(x)$: "if $\geq 2^{2^n}$ the $h$ digit of $\pi$ is $\geq 4$, output 1, else output 0."

---

**Chernoff Bounds** (hyp. testing)

**Motivation:**

\[
\begin{cases} 
\text{have hyp } h, \text{ in PAC world} \\
\text{have } EX(c, D) \\
\end{cases}
\]

is my $h$ good or not?
Obvious approach: * draw \( m \) ex from \( EX(c, D) \)
  * eval. \( h \) on each
  * count frac \( \frac{\# \text{ of ex. where } h(x) \neq c(x)}{m} \)

How good is \( \hat{E} \) as estimate of \( E_{\rho}(h, c) \)?

Next time: tools to answer this.