Last time:
- finish proof that can efficiently improperly learn 3-term DNF
- " can't eff. properly learn 3-term DNF (unless NP ⊆ RP).

Today: - finish pf that (prove L2, completing pf about the transformation f
- start VC dimension unit (KV Chap. 3) / unit on sample cxity of PAC learning
  - lower bound on : $\Omega(VCDIM(C)/\varepsilon)$
  - start upper bound on : (roughly) $O(VCDIM(C)/\varepsilon)$

Reminder: in-class closed-book midterm on Mon.

Questions?

To prove: $G \rightarrow [f] \rightarrow S^+, S^-$

01111 + 10111 + vert. of G

00111 - etc

(L2) If ∃ 3-term DNF cons. w/ S^+ & S^-, then G is 3-colorable.

Pf: Let $T_R \vee T_Y \vee T_B$ be any 3-term DNF cons. w/ S^+ & S^-.

↳ every + ex is acc. by some term

Consider the col. which colors vertex i of G

R
Y
or B

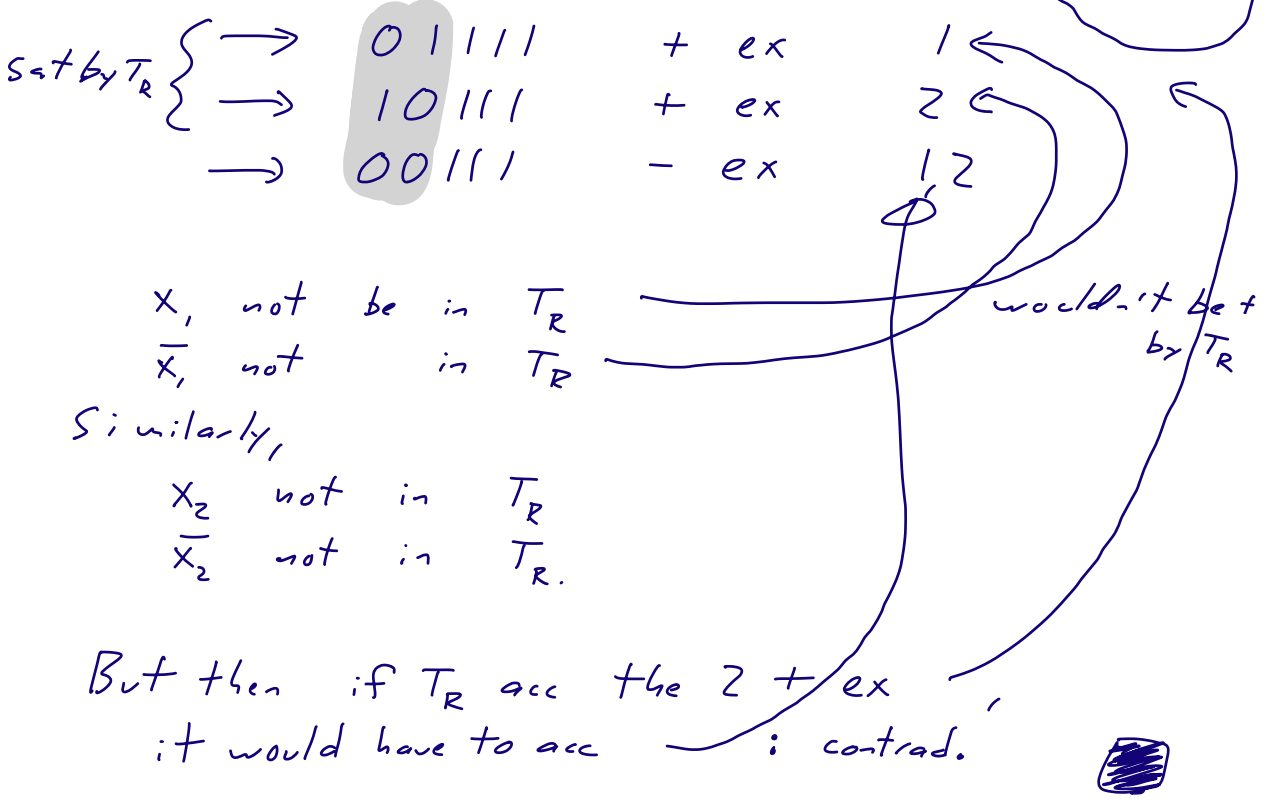
dep. on which of T_R, T_Y, T_B acc. $1 \ldots 1 \underset{i}{0} 1 \ldots 1$

(break ties arbitrarily).

Is this col. legit?

Claim: Let $\{i,j\}$ be any edge. Then above col. doesn't color both i & j R. (same for B, Y).

Pf: Consider $i,j = 1,2$: show not both sat by T_R



SAMPLE CXITY OF PAC LEARNING

(For this whole unit, we only worry about info theoretic, not comput. effiency/runtime.)

Motiv.: Given $\mathcal{C}$, basic questions:

- is $\mathcal{C}$ PAC learnable (from a finite # of samples)?
- if so, how many ex. needed?

Sps $\mathcal{C}$ finite.

CHF gives $\frac{1}{2}$ answer: $\frac{1}{\varepsilon}(\ln|\mathcal{C}| + \ln\frac{1}{\delta})$ ex.

is enough. (CHF for $\mathcal{C}$ usin $\mathcal{H}=\mathcal{C}$)

- lower bound?
- what if $|\mathcal{C}|=\infty$?

Fix $\delta = \frac{1}{10}$.

VCDIM($\mathcal{C}$) : whole story!

Main results of this unit: (pretend $\delta = \frac{1}{10}$)

Fix any $\mathcal{C}$. Let $d := \text{VCDIM}(\mathcal{C})$

- Any PAC learning alg for $\mathcal{C}$ must use $\Omega(\frac{d}{\varepsilon})$ ex.

(so if $d=\infty$, no finite # of ex. is enough to PAC learn)

- If $\mathcal{C}$ finite: then $\mathcal{C}$ is PAC learnable (CHF for $\mathcal{C}$ using $\mathcal{C}$) with $\approx O(\frac{d}{\varepsilon})$ ex.

  → like CHF result: VCDIM, not $\ln|\mathcal{C}|$, is "right" measure.

Lower bd on SC of PAC learning:

Thm Fix any $\mathcal{C}$, let $d = \text{VCDIM}(\mathcal{C})$.
Any PAC learner for $\mathcal{C}$ that learns to error ε & conf. $\delta = \frac{1}{10}$ must use

$\Omega(\frac{d}{\varepsilon})$ ex. $\left(\frac{d-1}{32\varepsilon}\right)$

Pf: First: give a $\mathcal{D}$ s.t. $\Omega(d)$ ex needed to achieve $\varepsilon = 1/8$, $\delta = 1/8$.

Let $S = \{x^1, \ldots, x^d\}$ be shatt. by $\mathcal{C}$.
Let $\mathcal{D}$ = unif over S. (0 wt elsewhere).

Suppose alg A only gets $\frac{d}{2}$ calls to $EX(c, \mathcal{D})$.
After calls, A has seen labels of $\leq d/2$ ex
not " " " $\geq \frac{d}{2}$ ".

Let target $c \in \mathcal{C}$ be chosen by picking it unif. from the 2^d concepts that shatter S.

So each of the 2^d poss. lab. of S = ly likely.

Hence
$$\mathbb{E}[\text{error of } A\text{'s hyp}] \geq \frac{1}{2} \cdot d/2 \cdot \frac{1}{d}$$
($\frac{1}{2}$ ← error on each unseen pt; $d/2$ ← $\geq \frac{d}{2}$ unseen pts; $\frac{1}{d}$ ← prob. of each unseen pt)
$$= 1/4.$$

Let $acc = 1 - error$; $\mathbb{E}[acc] \leq 3/4$

$\frac{3}{4} \cdot \frac{7}{6} = k$

Markov: $\Pr[acc \geq 7/8] \leq 6/7$.

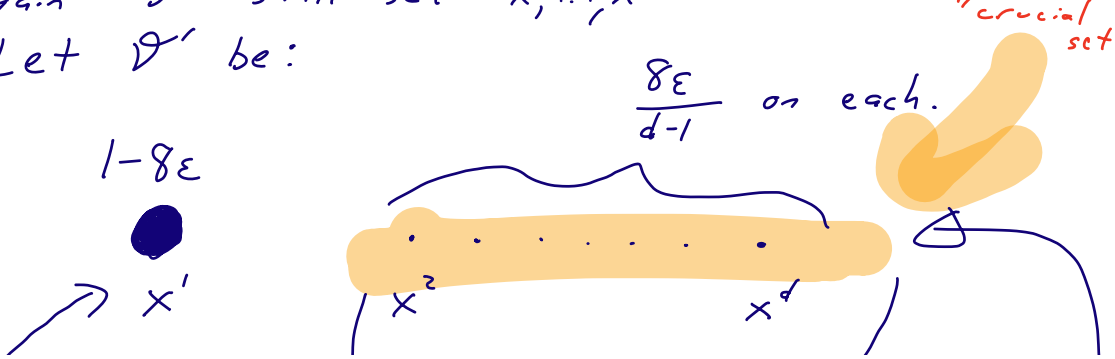
So $\Pr[error \geq 1/8] \geq 1/7$. Failed to $(\varepsilon = 1/8, \delta = 1/8)$-PAC learn.

Markov: if $X \geq 0$ is a r.v., & $\mu = \mathbb{E}[X]$, then for any $k \geq 1$, $\Pr[X \geq k \cdot \mu] \leq \frac{1}{k}$.

Let's prove whole thing: give a dist $\mathcal{D}'$ s.t. $\Omega(d/\varepsilon)$ ex. to get error ε & conf. $1-\delta = \frac{9}{10}$.

Again S = shatt set $x^1, \ldots, x^d$

Let $\mathcal{D}'$ be:



8ε total mass

Idea: $1-8\varepsilon$ acc. very easy to achieve; getting $7/8$ of last 8ε is slow: on avg, $\frac{1}{8\varepsilon}$ draws needed to get 1 ex from so "slowdown" of $\Omega(1/\varepsilon)$ on top of earlier arg.

Sps learning alg A only gets $\frac{d-1}{2}$ ex from crucial set $x^2, \ldots, x^d$.

Let $c \in \mathcal{C}$ (target) be rand. chosen s.t. each of the 2^{d-1} lab. of crucial set $=_y$ likely.

As above, w.p. $\geq \frac{1}{8}$, alg A has error rate $\geq \frac{1}{8}$ on crucial set. These have prob. 8ε under $\mathcal{D}'$, so if A gets $\leq \frac{d-1}{2}$ ex from crucial set, then w.p. $\geq \frac{1}{8}$, A has error $\geq \varepsilon \left(= \frac{1}{8} \cdot 8\varepsilon\right)$.

Each call to $EX(c, \mathcal{D})$ hits crucial set w.p. 8ε.

Sps A makes only $\frac{d-1}{32\varepsilon}$ calls to $EX(c, \mathcal{D})$.

Then $\mathbb{E}[\#\text{times hit crucial set}] = \frac{d-1}{32\varepsilon} \cdot 8\varepsilon = \frac{d-1}{4}$.

Mult. CB ($\gamma = 1$, $p_m = \frac{d-1}{4}$):

$$\Pr\left[X \geq (1+\gamma) p_m\right] \leq e^{-\frac{d-1}{12}}.$$

Can assume w.l.o.g. $d \geq 100$: then $\downarrow \leq \frac{1}{100}$.

So w.p. $\geq \frac{99}{100} \cdot \frac{1}{8} \geq \frac{1}{9}$,

error $\geq \varepsilon$.

Upper Bound on Sample Cxity of PAC Learning

Main result: $\forall \mathcal{D}$, any hyp $h \in \mathcal{C}$ which is consistent with $d = VCDIM(\mathcal{C})$

$$\approx \frac{d}{\varepsilon} \cdot \log \frac{1}{\varepsilon} + \frac{1}{\varepsilon} \log \frac{1}{\delta}$$

many rand ex from $EX(c, \mathcal{D})$ is, w.p. $\geq 1-\delta$, ε-accurate, i.e. $er_{\mathcal{D}}(h, c) \leq \varepsilon$.

↳ CHF for $\mathcal{C}$ using $\mathcal{H} = \mathcal{C}$ works, if you take ex.

Hi level outline of pf:

① Setup: Ponder $VCDIM(\mathcal{C})$. ← a #

We'll consider a function ↳ to get more info. "growth function of $\mathcal{C}$"

---

② Combinatorics arg: AMAZING THEOREM about how the growth fn behaves, for any $\mathcal{C}$.

---

③ Learning argument: a lot like CHF thm, but using AMAZING THM to control # ways conc in $\mathcal{C}$ can label a data set.

---

① Setup

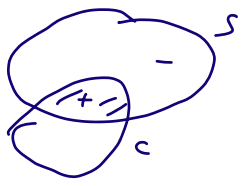
Recall $VCDIM(\mathcal{C})$ = size of largest shatt. set.

Another POV on $VCDIM(\mathcal{C})$:

Say $S \subseteq X$.

Define $\Pi_{\mathcal{C}}(S)$ = all labelings of S induced by concepts in $\mathcal{C}$.

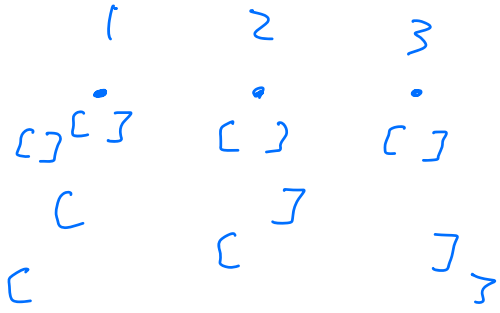
Equiv, $\Pi_{\mathcal{C}}(S) = \{c \cap S : c \in \mathcal{C}\}$



Ex: $S = \{1, 2, 3\}$

$\mathcal{C}$ = intervals of $\mathbb{R}$

$VCDIM(\mathcal{C}) = 2$.



$\Pi_{\mathcal{C}}(S) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}, \{1,2,3\}\}$ 7 subsets

Note: $\mathcal{C}$ shatters S $\Leftrightarrow$ $|\Pi_{\mathcal{C}}(S)| = 2^{|S|}$.

Def: $\Pi_{\mathcal{C}}(m)$ = growth function of $\mathcal{C}$

$$= \max_{\substack{S \subseteq X, \\ |S| = m}} |\Pi_{\mathcal{C}}(S)|$$

For $\mathcal{C}$ = intervals, $\Pi_{\mathcal{C}}(3) = 7$
