Last time: • finished pf of UB on sample cxity of PAC learning: if $\mathrm{VCDIM}(\mathcal{C}) = d$,
then any CHF for $\mathcal{C}$ using $\mathcal{C}$ as hyp class, run on $m \approx \frac{1}{\varepsilon}\left(d \log \frac{1}{\varepsilon} + \log 1/\delta\right)$ examples from $EX(c, \mathcal{D})$, is an (ε, δ)-PAC learner.

> **Key Thm:** Fix any (may be ∞!) c.c. $\mathcal{C}$, any target $c \in \mathcal{C}$, any dist $\mathcal{D}$.
> Given m samples from $EX(c, \mathcal{D})$, where
>
> $$m \geq \frac{2}{\varepsilon}\left(\log\left(\Pi_{\mathcal{C}}(2m)\right) + \log 2/\delta\right), \quad m \geq 8/\varepsilon \quad \star$$
>
> w.p. $\geq 1-\delta$ all bad $h \in \mathcal{C}$ are inconsistent w/ the m samples.

proved this, + combined w/ Amazing Thm $\Pi_{\mathcal{C}}(m) \leq \left(\frac{em}{d}\right)^d$, to get

• Application: $\mathcal{C}$ = LTFs over $\mathbb{R}^n$ is efficiently PAC learnable

Today: start unit on boosting confidence ✓ + accuracy
• setup, framework of boosting
• proof-of-concept: 3-stage boosting
40% error ⟶ 35% error

Questions?

Motivation for boosting: Def PAC learning:

Def: Alg A PAC learns conc. class $\mathcal{C}$ if:
• $\forall$ dist $\mathcal{D}$,
• $\forall c \in \mathcal{C}$,
• $\forall \varepsilon > 0$, $\forall \delta > 0$
given ε, δ, $EX(c, \mathcal{D})$, w.p. $\geq 1-\delta$ A outputs h st

strong PAC learning

$\Pr_{x\sim\mathcal{D}}[h(x)\neq c(x)] \le \varepsilon.$

What happens if we weaken def?

Weaken δ req? New def:

Def: Alg A' "low-confidence" PAC learns conc. class $\mathcal{C}$ if:

- $\forall$ dist $\mathcal{D}$,
- $\forall c \in \mathcal{C}$,
- $\forall \varepsilon > 0$, for $\delta = \frac{1}{10}$

given ε, $EX(c,\mathcal{D})$, w.p. $\ge 1-\delta = 0.9$ A' outputs h st

$$\Pr_{x\sim\mathcal{D}}[h(x)\neq c(x)] \le \varepsilon.$$

If $\mathcal{C}$ has a low-conf PAC learner ($\delta = \frac{1}{10}$)
then $\mathcal{C}$ also " " regular/strong PAC learner.

Sketch: like HW:

$1-\frac{\delta}{2}$ fail $\left\{\right.$ • run A' repeatedly, k times: w.p. $1-(0.1)^k$, some run gives good hyp;

$1-\frac{\delta}{2}$ fail $\left\{\right.$ • test the k hyp's on fresh ex, output one that does best.

(B/c of this: we'll sometimes ignore δ in our analyses.)

What about ε?

Def: Alg A is a weak PAC learner for $\mathcal{C}$ with advantage γ if

- $\forall$ dist $\mathcal{D}$,
- $\forall c \in \mathcal{C}$,
- $\forall \delta > 0$,

w.p. $\geq 1-\delta$, A outputs an h s.t. given $EX(c, \mathcal{D})$

$$\Pr_{x\sim\mathcal{D}}[h(x) \neq c(x)] \leq \tfrac{1}{2} - \gamma.$$

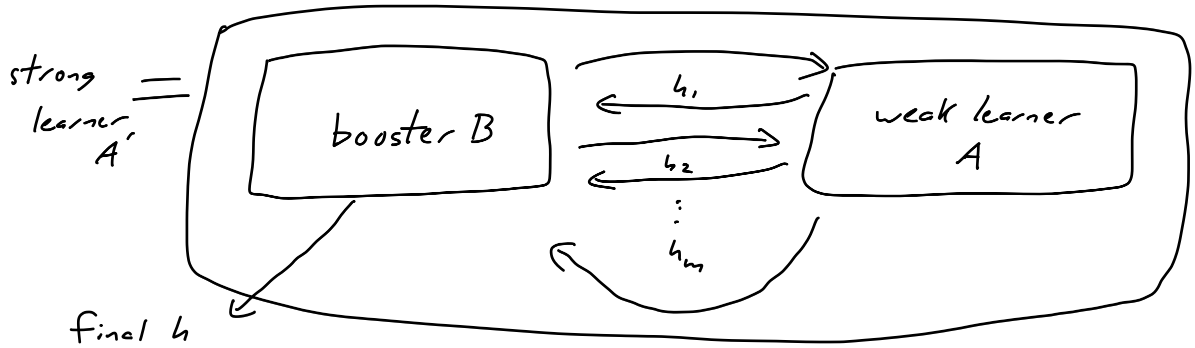
Surprisingly, not really weaker than strong PAC!

runtime of A is T, to achieve $\delta = \frac{1}{10}$ with adv. γ,

Thm: Let $\mathcal{C}$ be any concept class.
If there's an (efficient) weak PAC learner A for $\mathcal{C}$,
then there is an (eff) strong PAC learner A' for $\mathcal{C}$.
where A' runs in time

$$\mathrm{poly}(1/\gamma, \tfrac{1}{\varepsilon}, \log 1/\delta, T)$$

Pf is by giving explicit, eff. "boosting alg" B.
B runs weak learner repeatedly, $\mathrm{poly}(\log\frac{1}{\varepsilon}, \frac{1}{\gamma})$ times.



Hi level idea of boosting

to learn c $\xrightarrow{\text{to acc } \varepsilon}$ given $EX(c, \mathcal{D})$: ($\neq$ have w.l. A with adv. γ)

booster

① runs A repeatedly using different dist's on examples: i.e. using $EX(c, \underset{\mathcal{D}_1}{\underset{\shortparallel}{\mathcal{D}}})$, $EX(c, \mathcal{D}_2)$, $EX(c, \mathcal{D}_3)$, ...

② combines $h_1, \ldots, h_m$ to get final hyp h.

Questions / requirements to fully describe a booster:

ⓐ what's idea behind $\mathcal{D}_1, \mathcal{D}_2, \ldots$?

Cook them up so as to "force weak learner A to give some new info"

ⓑ How can we run A on $EX(c, \mathcal{D})$, $EX(c, \mathcal{D}_2)$, $EX(c, \mathcal{D}_3)$, ... when we only have access to $EX(c, \mathcal{D})$?

Two poss. approaches:

★ 1) "filter" examples to change $\mathcal{D}$ into $\mathcal{D}_2, \mathcal{D}_3, \ldots$

or

2) draw a fixed sample + reweight them as you please -- explicitly maintain dist over

ⓒ how to combine $h_1, h_2, \ldots h_m$ to get final h?

Typically by maj vote, or weighted maj vote.

(d) Why does it all work - why is

$$\Pr_{x \sim \mathcal{D}}[h(x) \neq c(x)] \leq \varepsilon \, ?$$

takes work/pf :)

Concrete simple ex. of boosting in action: Schapire's orig. booster.

recursive; we'll do one step.

Setup:

Let A be weak PAC learner for $\mathcal{C}$ w/ $\gamma = \frac{1}{10}$:
for any dist $\mathcal{D}$, achieves hyp w/ error ≤ 0.4.

We'll show how to eff boost acc. to achieve error ≤ 0.352, by doing 3 runs of A to get h_1, h_2, h_3, + taking final $h = MAJ(h_1, h_2, h_3)$

---

2 simplif. assump:

- at every run, A gives a hyp. w/ error $= 0.4$; +
- assume this happens w.p. 1 (like $\delta = 0$).

<u>Notation:</u> for a dist $\mathcal{D}$ + event E,

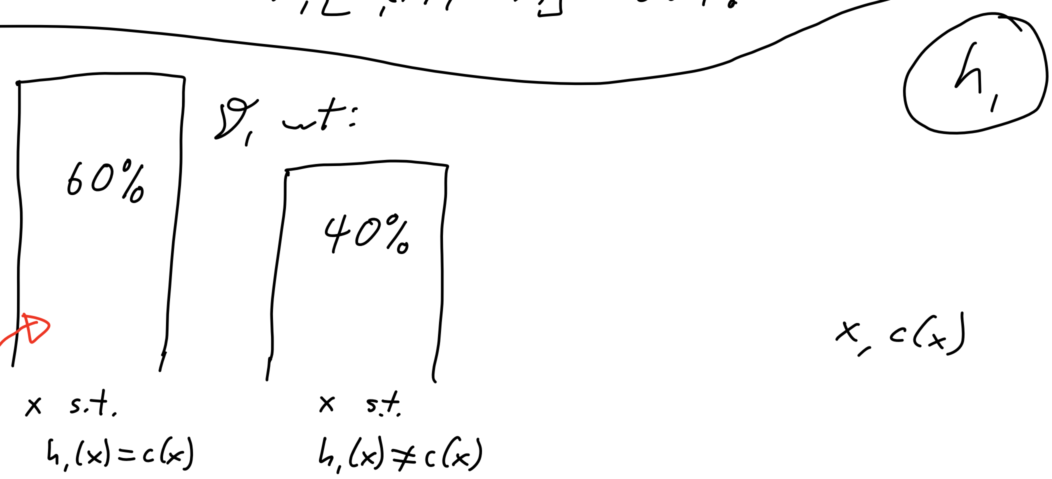
write "$\mathcal{D}(E)$" for "$\Pr_{\mathcal{D}}[E \text{ holds}]$"

"$\mathcal{D}(h(x)=c(x))$" for "$\Pr_{x\sim\mathcal{D}}[h(x)=c(x)]$".

Booster does this

① Run A on $EX(c, \mathcal{D})$ (with $\mathcal{D} = \mathcal{D}_1$) to get hyp h_1: have

$$\mathcal{D}_1[h_1(x) \neq c(x)] = 0.4.$$



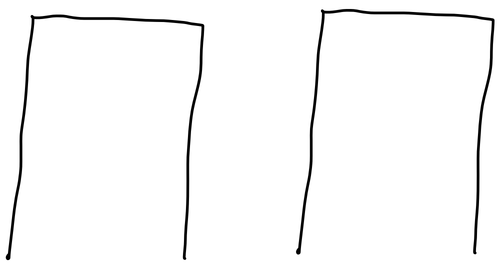
② What's a new dist $\mathcal{D}_2$ s.t. running A on $EX(c, \mathcal{D}_2)$ would "tell me new that h_1 didn't tell me"?

Not best approach: filter by discarding all x in : then $\overline{h_1}$ has perfect acc, + WL can give it as hyp, but doesn't help (nothing new).

Let $\mathcal{D}_2$ be dist which • "scales up" wt of each $x \in X$ st $h_1(x) \neq c(x)$ by $5/4$ $\quad 40\% \to 50\%$

• "scales down" wt of each $x \in X$ st $h_1(x) = c(x)$ by $5/6$ $\quad 60\% \to 50\%$

$\mathcal{D}_2$ wt:



x s.t. $h_1(x) = c(x)$ $\qquad$ x s.t. $h_1(x) \neq c(x)$

With this $\mathcal{D}_2$, $\Pr_{x \sim \mathcal{D}_2}[h_1(x) \neq c(x)] = \frac{1}{2}$, we'll get a <u>different</u> h_2.

Can sample from $EX(c, \mathcal{D}_2)$ as follows:

- toss fair coin

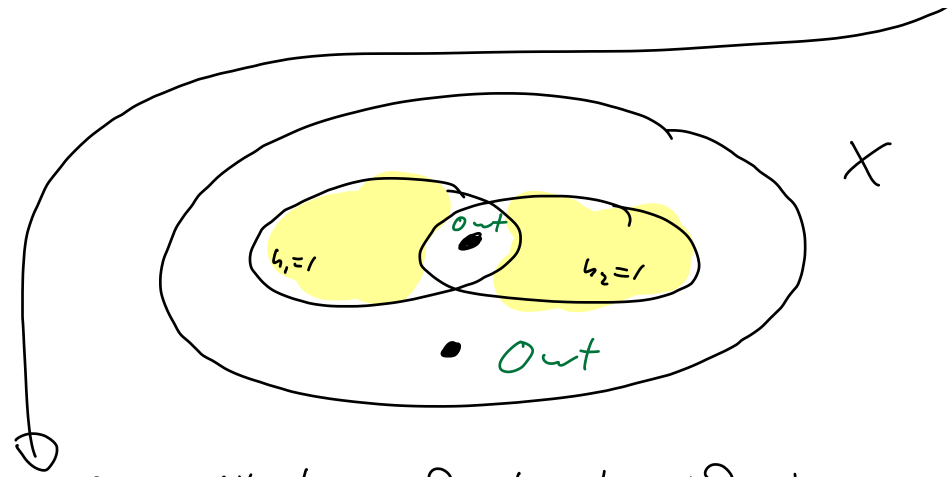
  H: draw from $EX(c, \mathcal{D})$ till get an x st $h_1(x) = c(x)$

  T: draw from $EX(c, \mathcal{D})$ till get an x st $h_1(x) \neq c(x)$.

$\mathbb{E}[\#\text{ rep.}]$ for $\checkmark = \frac{1}{0.4} = 2.5$: efficient.

What's $\mathcal{D}_3$? $\qquad$ (recall, Final $h = MAJ(h_1, h_2, h_3)$...)

h_3 only matters on x st $h_1(x) \neq h_2(x)$:



so we'll take $\mathcal{D}_3$ to be $\mathcal{D}_1$ but conditioned on $h_1(x) \neq h_2(x)$.

To $\underbrace{\text{do this}}^{\text{simulate } EX(c,\mathcal{D}_3)}$: draw from $EX(c,\mathcal{D}_1)$ until get x st $h_1(x) \neq h_2(x)$; use that $(x, c(x))$.

Run A on $EX(c,\mathcal{D}_3)$ to get h_3,

use $h = MAJ(h_1, h_2, h_3)$ as final hyp.

Q: what if $\Pr_{x \sim \mathcal{D}_1}[h_1(x) \neq h_2(x)] := \tau$ is tiny?

Runtime overhead to simulate 1 draw from $EX(c,\mathcal{D}_3)$ is $1/\tau$ in expectation...

inefficient!

If τ tiny, though, then w.p. $1-\tau$, $h_1(x) = h_2(x) = h(x)$

regardless of h_3.

So if you see τ is tiny, use anything for h_3, & only changes

acc $\Pr_{x \sim \mathcal{D}}[h(x) \neq c(x)]$, by $\pm\tau$ (tiny).

Next time: analyze this. 35.2% error.
