Last time: • finished proof of our CHF theorem
• Application: efficiently PAC learning LTFs over $\mathbb{R}^n$
• We now understand sample cxity of PAC learning"

Today: Boosting!
→ • confidence boosting (easy)
→ • start accuracy boosting:
- setup, boosting framework
- proof of concept: boosting a 40% error learner to a 35.2% error learner (three stage process)

Questions?

Recall PAC learning def: "strong PAC learning"
• Alg. A PAC learns $\mathcal{C}$ if $\forall c \in \mathcal{C}$, $\forall$ dist $\mathcal{D}$,
• $\forall \varepsilon > 0$,
• $\forall \delta > 0$,
given ε, δ & $EX(c, \mathcal{D})$, A produces h s.t. $err_{\mathcal{D}}[h, c] \leq \varepsilon$ w.p. $\geq 1-\delta$.

What happens if we relax the def.? Turns out...
nothing -- if $\mathcal{C}$ ~~efficiently~~ learnable under relaxed def, then " " " " original def!

Relax δ (confidence) requirement:

Say $\mathcal{C}$ is "low-confidence PAC learnable" if:

- $\forall c \in \mathcal{C}$, $\forall$ dist $\mathcal{D}$,
- $\forall \varepsilon > 0$,
- ~~$\forall \delta > 0$,~~ for $\delta = 0.9$,

given ε, δ & $EX(c, \mathcal{D})$, A produces h s.t. $err_{\mathcal{D}}[h, c] \le \varepsilon$ w.p. $\ge 1 - \delta = 0.1$

Fact: if $\mathcal{C}$ eff. LC-PAC learnable with $m(\varepsilon)$ examples, then $\mathcal{C}$ is also eff. "strongly PAC learnable", in fact using $O(\log(1/\delta) \cdot m(\varepsilon/2)) + O(\log(1/\delta)/\varepsilon^2)$ ex.

Pf sketch:

The alg has 2 stages:

1) Run orig. LC-PAC learner with error param $\varepsilon/2$, $C_1 = O(\log(1/\delta))$ times.

$\Pr[$none of C_1 runs yields an $\varepsilon/2$-error hyp$] \le (1 - 0.1)^{C_1} < \delta/2$.

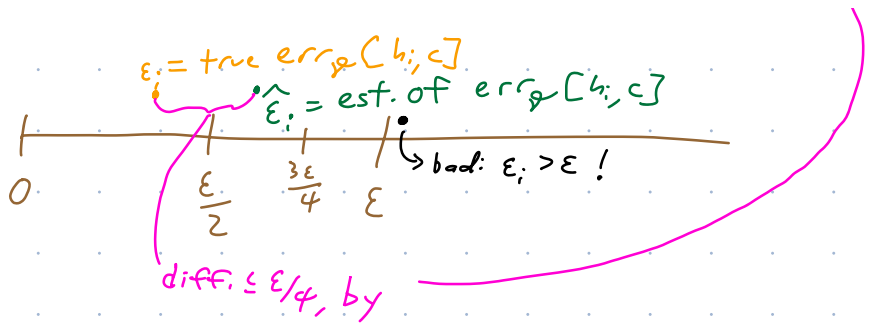
So have hyp's $h_1, \ldots, h_{C_1}$, & w.p. $\ge 1 - \delta/2$, at least one h_i has $err_{\mathcal{D}}[h_i, c] \le \varepsilon/2$.

2) Hyp. testing: Draw $O\left(\frac{\log(1/\delta^2)}{\varepsilon^2}\right)$ fresh ex., measure observed error of each h_i on them, output the h_i that does best.

CB: for any fixed $j \in [C_1]$, w.p. $\ge 1 - \frac{\delta}{2C_1} \ge 1 - \delta^2$, my est. of $err_{\mathcal{D}}[h_j, c]$ is $\pm \varepsilon/4$ accurate.

So UB $\Rightarrow$ w.p. $\ge 1 - \frac{\delta}{2}$, every est. is $\pm \varepsilon/4$-acc.

(*)



Any Bad h_i will appear to have error $> \frac{3\varepsilon}{4}$

Any $\frac{\varepsilon}{2}$-acc. h_i) / / / / $< \frac{3\varepsilon}{4}$

So tot fail prob. (of ~~not~~ outputting an h_i with $err_{\mathcal{D}}[h_i, c] \leq \varepsilon$) is $\leq \frac{\delta}{2} + \frac{\delta}{2} = \delta$ ☺

We'll usually ignore δ going forward ☺

What about accuracy?

Def: Alg A is a weak PAC learner for $\mathcal{C}$ with advantage γ if $\forall c \in \mathcal{C}$, $\forall$ dist $\mathcal{D}$,

- ~~$\forall \varepsilon \in (0,$~~
- $\forall \delta > 0$,

given δ & $EX(c, \mathcal{D})$, A produces h s.t. $err_{\mathcal{D}}[h, c] \leq \frac{1}{2} - \gamma$ w.p. $\geq 1 - \delta$.

(i.e. achieve acc. $\frac{1}{2} + \gamma$).

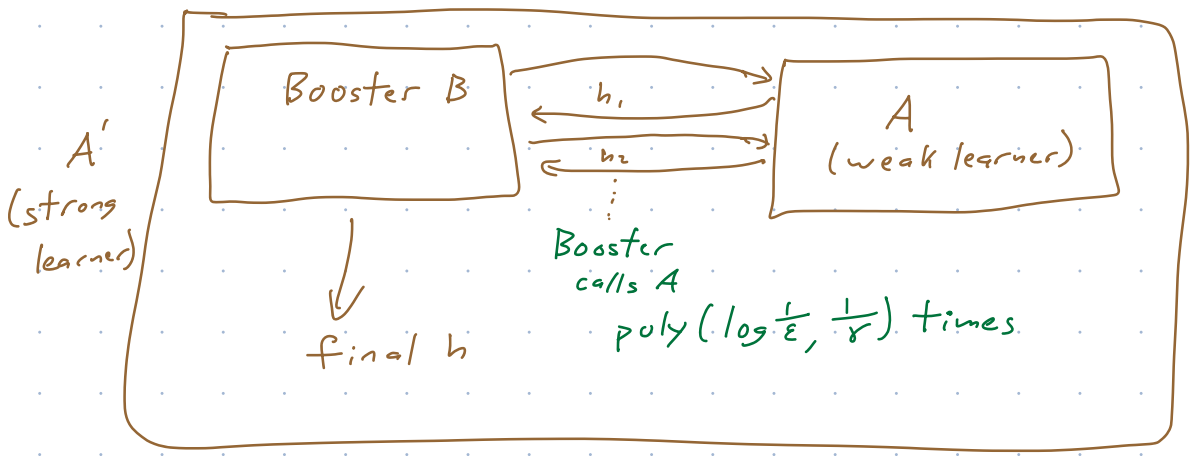
Turns out: given any weak PAC learner, can "boost" acc. to achieve strong PAC learning!

Thm: Let $\mathcal{C}$ be any conc. class.

Sps A is an eff. weak PAC learner for $\mathcal{C}$ with advantage γ.

Then there is a $\text{poly}(1/\gamma, \frac{1}{\varepsilon}, \log\frac{1}{\delta}, T)$-time strong PAC learner A' for $\mathcal{C}$ (where T = runtime for A to achieve confidence $9/10$, i.e. $\delta = 0.1$)

Done via efficient, explicit boosting alg. B.



Basic idea of a booster: to learn c given $EX(c, \mathcal{D})$ + weak learner A:

- booster runs A multiple times using a seq. of <u>different distributions</u> $EX(c, \mathcal{D}_1)$, $EX(c, \mathcal{D}_2)$, ... to get hyp's h_1, h_2, ...
- booster combines $h_1, h_2, \ldots$ to create final h.

To describe a boosting alg, must describe:

- what are $\mathcal{D}_1, \mathcal{D}_2, \ldots$? (weak learner)

  Cleverly designed dist's to force A to

"give some new info".

- How can you run A on $EX(c, \mathcal{D}_i)$ when we only have $EX(c, \mathcal{D})$?

  Few ways:
  - filter examples; or
  - work over explicit sample.

- how to combine h_i's? Maj vote. (why)
- Does it work? Does it give final h s.t. $\Pr_{x \sim \mathcal{D}}[h(x) \neq c(x)] \leq \varepsilon$?

Yes - takes a pf.

Let's do an example pf of concept: 3-stage boosting.

Let A be a weak PAC learner for some $\mathcal{C}$ with $\gamma = 0.1$ (acc of h is 60%; error $\leq$ 40%)

We'll run A 3 times, & get a hyp. w/error $\leq 35.2\%$.

2 simplifying assumptions: (ignore δ)

assume every run of A gives a hyp. with error exactly 40% on whatever dist. was used.

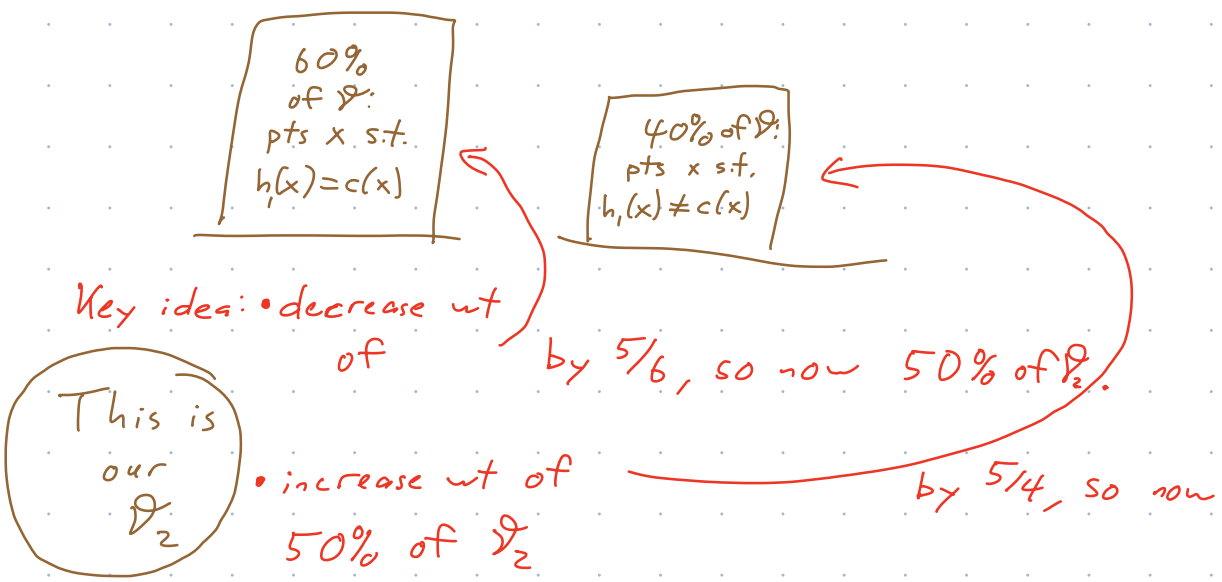
Notation:
write "$\mathcal{D}[E]$" for $Pr_{\mathcal{D}}[E]$, e.g.

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

Simple 3-stage booster: [Have w.l. A (40% error); $EX(c,\mathcal{D})$]

1) Run A on $EX(c,\mathcal{D})$ to get h_1. ($\mathcal{D}_1 = \mathcal{D}$)
Have $\mathcal{D}_1[h_1(x)\neq c(x)] = 0.4$

2) One idea: let $\mathcal{D}_2$ be $\mathcal{D}_1$ restricted to those x s.t. $h_1(x)\neq c(x)$.
Problem: the hyp $\overline{h_1(x)}$ has 100% acc on this $\mathcal{D}_2$, so w.l. could give you that... no new info "



Q: how to sim. access to $EX(c,\mathcal{D}_2)$ given h_1 & $EX(c,\mathcal{D})$?

1 way: when $EX(c, \vartheta)$ has $h_1(x) = c(x)$,
keep w.p. $2/3$ & try again w.p. $1/3$
↳ $p = 2/3$

Another way: toss fair coin to decide between
$h_1 = c$ or $h_1 \neq c$
↳ $p = .6$ ↳ $p = 0.4$

& call $EX(c, \vartheta)$ till you get an example of the type you want.

Both ways efficient:

given $\$ \sim \Pr[\$ = H] = p$,

$\mathbb{E}[\#\text{ tosses till get } H] = \frac{1}{p}$.
