Boosting

The Weak PAC Model

- Exactly the same as PAC, but no ε
- Ask that $\forall c \in \mathcal{C}, P, \delta$ algo outputs $h \in \mathcal{H}$ s.t. $\varepsilon(h) \leq \frac{1}{2} - \alpha \triangleq \varepsilon_0$
- Here α is a fixed constant (e.g. $\alpha = 0.01$) or even $\alpha = \frac{1}{poly(n)}$
- Algo runs in time $poly(n, size(c), \delta)$

Once upon a time...

- Every C known to be **weakly** learnable was known to be **strongly** learnable
- Every C known to be **hard** for **strong** learning was also hard for **weak** learning
- **Q**: Could weak learning $\equiv$ strong learning?

Note: If so, needs distribution-free property of PAC.

classes C and specific P s.t.

- C weakly learnable w.r.t. P
- C not strongly learnable w.r.t. P

Boosting

- Let L be an algo for weakly learning C by $\mathcal{H}$
- Idea: Use L as a subroutine in an algo L' for strongly learning C by $\mathcal{H}'$

Attempt #1

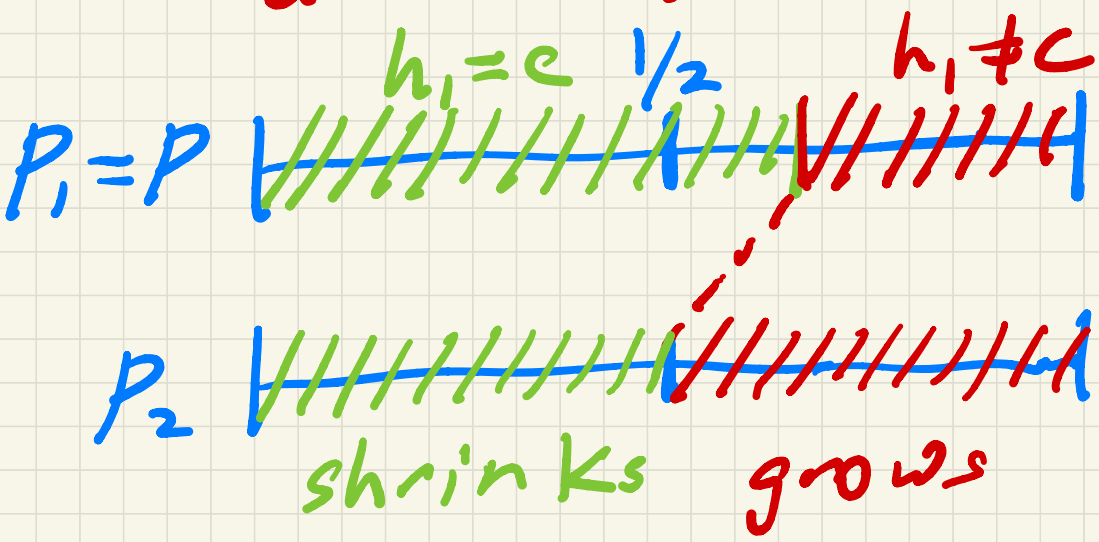
- Run weak L l times on $c, P \rightarrow$ get $h_1, h_2, \ldots h_l$
- Let

$$h(x) = \underbrace{MAJ}_{\hookrightarrow \text{majority vote}}(h_1(x), \ldots h_l(x))$$

- Good idea?

Attempt #2

- Run L on $c, P \to$ get h_1
- Run L again, but force it to do something different:



More formally, if

$$\varepsilon(h_1) = \varepsilon_0 < 1/2$$

then define P_2:

$$\forall x \in X:$$

if $h_1(x) = c(x)$

$$P_2(x) = \frac{1}{2(1-\varepsilon_0)} P_1(x)$$

if $h_1(x) \neq c(x)$

$$P_2(x) = \frac{1}{2\varepsilon_0} P_1(x)$$

- By design,
$$\varepsilon_{P_2}(h_1) = 1/2$$
- We can simulate P_2 from P_1, h_1 (how?)
- Run L on $P_2 \Rightarrow$ get $h_2 \neq h_1$
- Now what?

Breaking ties

- Define P_3:

If $h_1(x) \neq h_2(x)$

$$P_3(x) = P_1(x)/Z$$

$$Z = P_1[h_1(x) \neq h_2(x)]$$

If $h_1(x) = h_2(x)$

$$P_3(x) = 0.$$

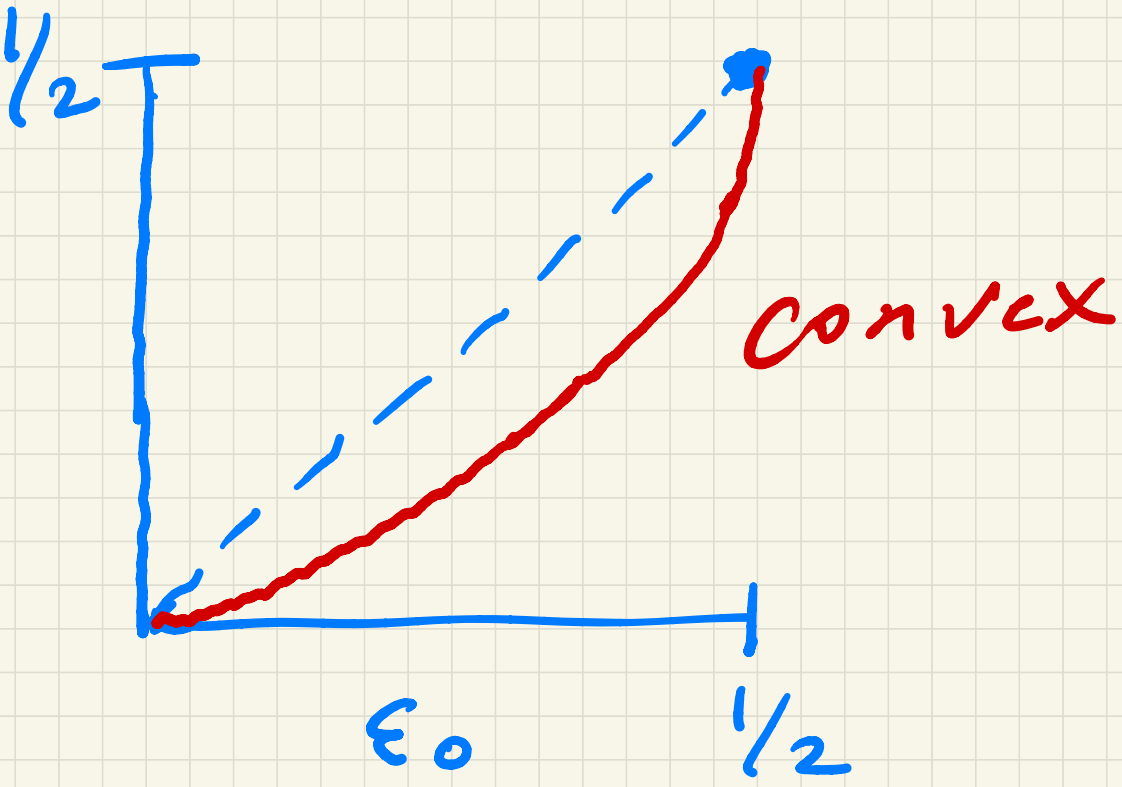
- We can simulate P_3 from P_1, h_1, h_2 (how?)
- Run L on P_3
  $\Rightarrow$ get h_3
- Now let

$$h(x) = MAJ(h_1(x), h_2(x), h_3(x))$$

- Now what?

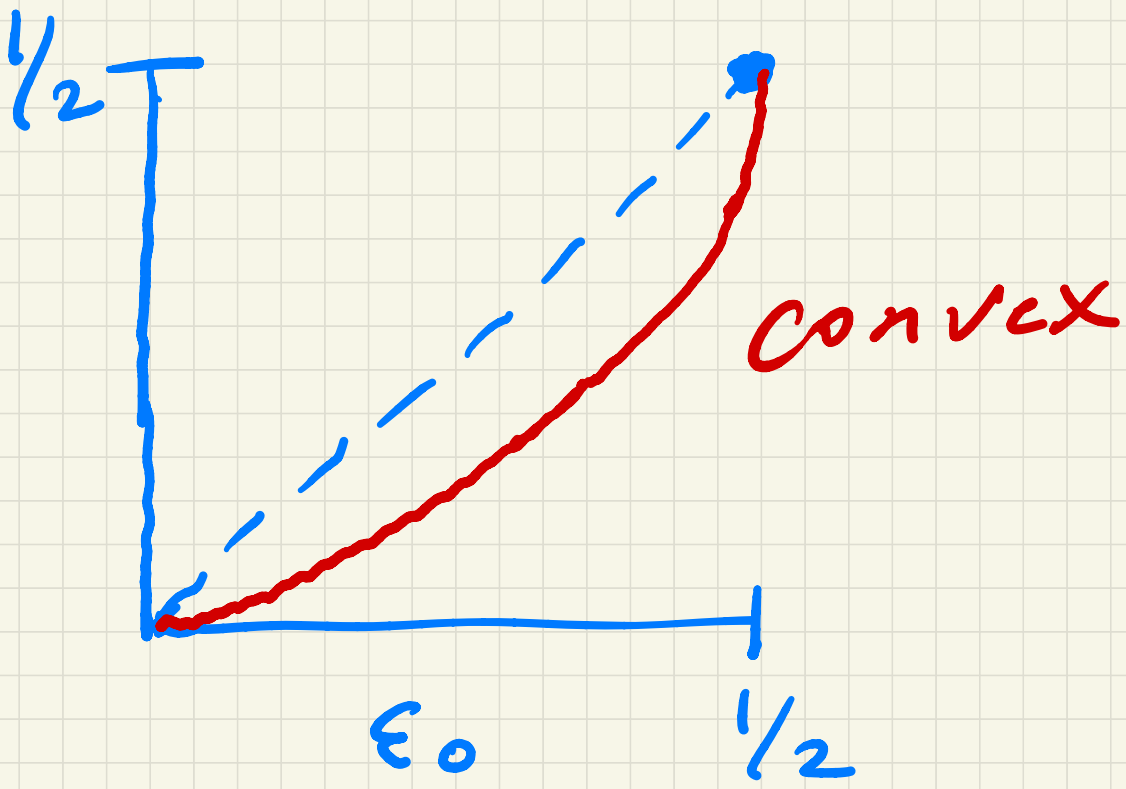
Lemma

$$\varepsilon_{p_t}(h) \leq 3\varepsilon_0^2 - 2\varepsilon_0^3$$

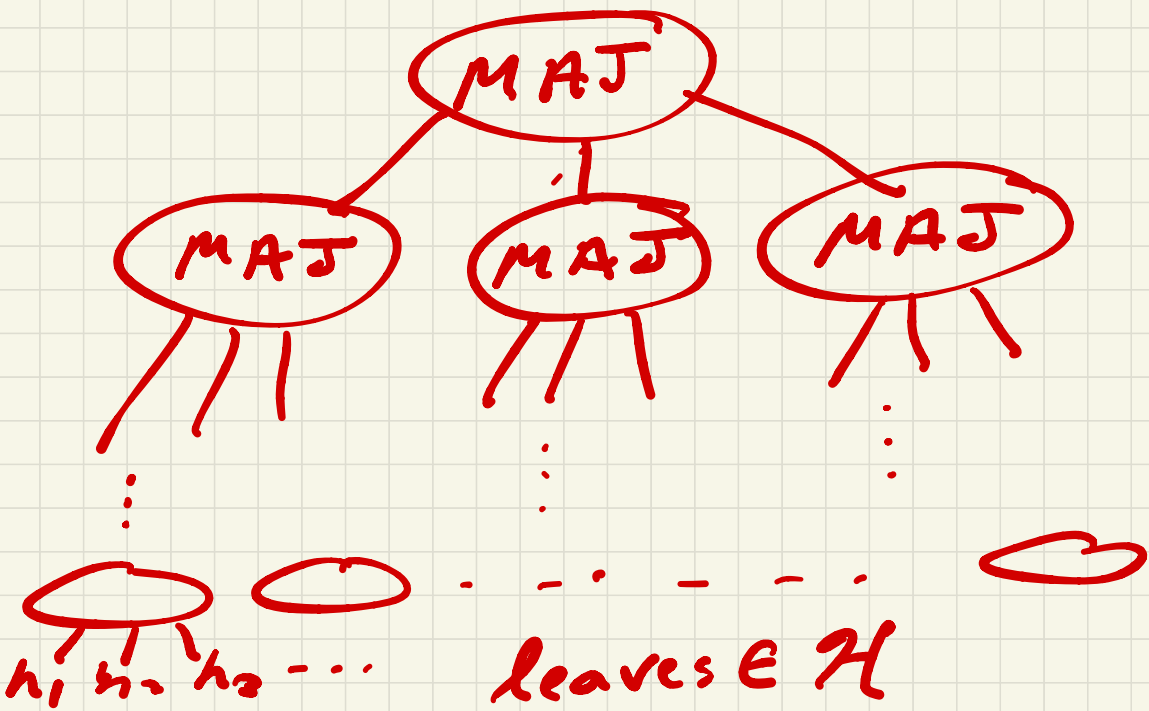


So now have a better weak learner!

So now let's recurse:



- Need $\sim \log\log(1/\varepsilon)$ levels to go from $\varepsilon_0 \to \varepsilon$
- Final hypothesis:



A More Practical Formulation: Adaboost

Setup

- Given data

$$S = \langle x_1, y_1 \rangle, \ldots \langle x_m, y_m \rangle$$

$$y_i \in \{+1, -1\}$$

- Goal: drive error on $S \to 0$ using weak learner
- Initial distribution p_1: uniform on S
- Subsequent p_t will be reweightings on S

Adaboost

For $t = 1, 2, \ldots T$:

- run weak algo on P_t
  $\Rightarrow$ get $h_t \in \mathcal{H}$
- choose $\alpha_t > 0$
- define $\forall i$:

$$P_{t+1}(i) = \frac{P_t(i)\, e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$$

Z_t = normalization for P_{t+1}

Final classifier:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

So $\mathcal{H}'$ = linear combos over $\mathcal{H}$.

Some Notation

- Define
$$\varepsilon_t = P_t[h_t(x_i) \neq y_i]$$
- Allowing ε_t to vary (including $\varepsilon_t = 1/2$)
- Can easily measure
- Define $\gamma_t = 1/2 - \varepsilon_t$
- Advantage over $1/2$

Adaboost: Analysis

$$y_i \neq H(x_i) = sign\left(\sum_t \alpha_t h_t(x_i)\right)$$

$$\Rightarrow y_i \sum_t \alpha_t h_t(x_i) \leq 0$$

$$\Rightarrow e^{-y_i \sum_t \alpha_t h_t(x_i)} \geq 1$$

So

$$\frac{1}{m}\sum_i I[H(x_i) \neq y_i] \quad \leftarrow \text{error of } H \text{ on } S$$

$$\leq \frac{1}{m}\sum_i e^{-y_i \sum_t \alpha_t h_t(x_i)}$$

focus on bounding this (*)

For any fixed i, recall:

$$P_{t+1}(i) = \frac{P_t(i) e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$$

Rewrite:

$$Z_t = \frac{P_t(i) e^{-\alpha_t y_i h_t(x_i)}}{P_{t+1}(i)}$$

Let's define P_i:

$$P_{T+1}(i) = 1/m = P_1(i)$$

(not used anyway)

Next: show that

$$(*) = \prod_{t=1}^{T} Z_t$$

$$\prod Z_t = Z_T \cdot Z_{T-1} \cdots Z_1 =$$

$$p_T(i)\, e^{-\alpha_T y_i h_T(x_i)} / p_{T+1}(i)$$

$$\times p_{T-1}(i)\, e^{-\alpha_{T-1} y_i h_{T-1}(x_i)} / p_T(i)$$

$$\times p_{T-2}(i)\, e^{-\alpha_{T-2} y_i h_{T-2}(x_i)} / p_{T-1}(i)$$

$$\vdots$$

$$\times p_1(i)\, e^{-\alpha_1 y_i h_1(x_i)} / p_2(i)$$

$$= e^{-y_i \sum_t \alpha_t h_t(x_i)}$$

This holds $\forall i$, so get same if we average:

$$\prod_t Z_t = \frac{1}{m} \sum_i e^{-y_i \sum_t \alpha_t h_t(x_i)}$$

$$= (*) \text{ from before!}$$

$$\geq \text{error on } S$$

Next: choose α_t's to make Z_t's as small as possible.

Adaboost

For $t = 1, 2, \ldots T$:

- run weak algo on P_t
  $\Rightarrow$ get $h_t \in \mathcal{H}$
- choose $\alpha_t > 0$
- define $\forall i$:

$$P_{t+1}(i) = \frac{P_t(i)\, e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$$

Z_t = normalization for P_{t+1}

$$Z_t = \sum_i P_t(i)\, e^{-\alpha_t y_i h_t(x_i)}$$

$$= \text{normalization for } P_{t+1}$$

$$= \sum_{i:\, h_t(x_i) = y_i} P_t(i)\, e^{-\alpha_t}$$

$$+ \sum_{i:\, h_t(x_i) \neq y_i} P_t(i)\, e^{\alpha_t}$$

$$= (1-\varepsilon_t)\, e^{-\alpha_t}$$

$$+ \varepsilon_t e^{\alpha_t}$$

If we choose

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

then

$$Z_t = (1-\varepsilon_t)\sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} + \varepsilon_t \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}$$

$$= 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$$

So error on $S \leq$

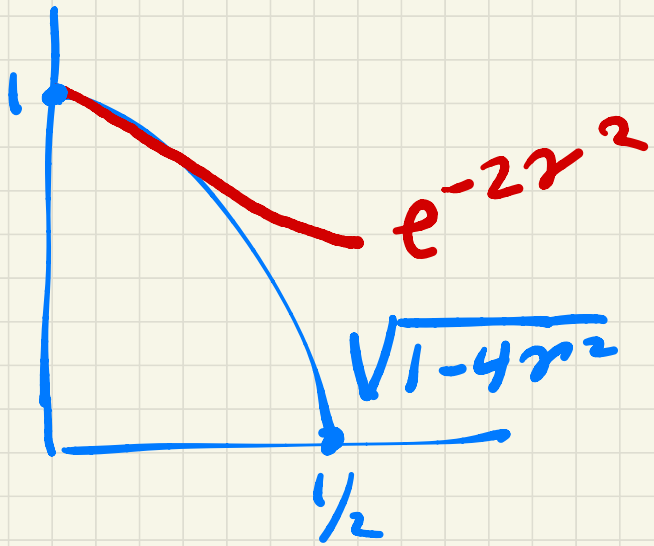
$$\prod_t Z_t = \prod_t \underbrace{2\sqrt{\varepsilon_t(1-\varepsilon_t)}}_{\leq 1}$$

$$\sqrt{4\varepsilon(1-\varepsilon)} =$$

$$\sqrt{4(1/2-\gamma)(1/2+\gamma)} =$$

$$\sqrt{4(1/4-\gamma^2)} =$$

$$\sqrt{1-4\gamma^2} \leq e^{-2\gamma^2}$$



So $\prod_t z_t \le \prod_t e^{-2\gamma_t^2}$

$$= e^{-2\sum_t \gamma_t^2}$$

- exponential decay in $\sum_t \gamma_t^2$
- adaptive in γ_t's

- If all $\gamma_t \geq \gamma > 0$:
  error on $S \leq e^{-2T\gamma^2}$
- Set $< 1/m$ (consistency):
  $$T > \frac{1}{2\gamma^2} \ln(m)$$
- $VCD = VCD(\mathcal{H}) \cdot T = dT$
- So $\boxed{T > \frac{1}{2\gamma^2} \ln(dT) \text{ suffices.}}$

Weak = Strong via Adaboost

- Suppose $\mathcal{C}$ weakly PAC by $\mathcal{H}$ with $\varepsilon_0 = 1/2 - \gamma$
- If we make

$$e^{-2T\gamma^2} < 1/m$$

we get a consistent

$$H(x) = \text{sign}\left(\sum_t \alpha_t h_t(x)\right)$$

- If $VCD(\mathcal{H}) = d$, then VCD of these $\leq Td$

So:

- Choose $m = \frac{Td}{\varepsilon} \log(1/\delta)$
- Then solve

$$e^{-2T\gamma^2} < 1/m$$

$$-2T\gamma^2 < \ln(1/m)$$

$$T > \frac{1}{2\gamma^2} \ln(m)$$

$$T > \frac{1}{2\gamma^2} \ln\left(\frac{Td}{\varepsilon} \ln(1/\delta)\right)$$

With these choices for m and T, with prob. $\geq 1-\delta$, $\varepsilon(H) \leq \varepsilon$.

$\therefore$ Weak PAC $\Rightarrow$ PAC.