Lecture 17

Boosting

Weak Learning

def. algorithm $\mathcal{A}$ weakly PAC learns concept class $\mathcal{C}$ if $\exists\, \gamma > 0$ s.t.

$\forall\, c \in \mathcal{C}$ & $\forall$ dists $\mathcal{D}$,

given examples of c according to $\mathcal{D}$

$\mathcal{A}$ outputs h s.t. $\Pr_{\mathcal{D}}[h(x) \neq c(x)] \leq \frac{1}{2} - \frac{\gamma}{2}$

↑ advantage

Thm if $\mathcal{C}$ can be weakly PAC learned (on any $\mathcal{D}$) then

$\mathcal{C}$ can be (strongly) PAC learned.

Weak vs. Strong Learning

<u>Def.</u> Algorithm $\mathcal{A}$ weakly "PAC learns" concept class $\mathcal{C}$

if $\forall\ c \in \mathcal{C}$ & $\forall$ dists $\mathcal{D}$ $\quad \exists\ \gamma > 0$

$\forall\ \cancel{\varepsilon}, \delta > 0$ $\quad \left(\delta = \frac{1}{4} \text{ or } \frac{1}{n^2} \text{ doesn't affect}\right)$

with prob $\geq 1-\delta$

given examples of c

$\mathcal{A}$ outputs h s.t. $\Pr_{\mathcal{D}}[h(x) \neq c(x)] \leq \cancel{\varepsilon}\ \frac{1}{2} - \frac{\gamma}{2}$

$\uparrow$ advantage

It was conjectured that distribution free weak learning was really weaker but surprise!

Can "boost" a weak learner

<u>Thm</u> if $\mathcal{C}$ can be weakly learned on <u>any</u> dist $\mathcal{D}$ then $\mathcal{C}$ can be (strongly) learned.

Applications

1) "theoretical"

- Unif dist Algorithms for poly term DNF
  weight w - poly threshhold fctns
  { low degree alg doesn't work well

  (Boosting + KM)

- Ave case vs. worst case

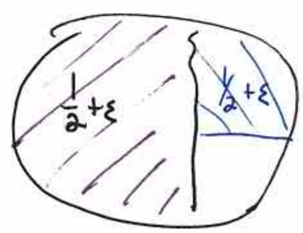
2) practical - Boosting
   Freund - Schapire

Good & Bad Ideas

1) simulate weak learner several times on same distribution & take majority answer
   -or-
   best answer

   gives better confidence
   but doesn't reduce error, what if always get same answer?

2) filter out examples on which current hypothesis does well & run weak learner on part where you do badly.



Problem: given a new example, how do you know which section it is in?

3) Keep some samples on which you are ok
alway use majority vote on all previous hypotheses
to predict value of new samples

history: Schapire, Freund-Schapire, Impagliazzo-Servedio. Klivans

Filtering Procedures

- decide which samples to keep, which to throw out
- samples on which so far you guess correctly ← need for checking future hypotheses
  incorrectly ← need to improve on these

The setting

- Given labelleled examples

  $$(x_1, f(x_1)), (x_2, f(x_2)), \ldots$$

  $$x_i \in_R \mathcal{D}$$
  $$f \in \mathcal{C}$$

- Given weak learning alg WL which weakly learns (advantage $\frac{\gamma}{2}$) on any dist $\mathcal{D}'$

Boosting Algorithm

- Stage 0 (Initialize)

$\mathcal{D}_0 \leftarrow \mathcal{D}$

run WL on $\mathcal{D}_0$ to generate (whp)

$$C_1 \text{ s.t. } \Pr_{\mathcal{D}_0}[f(x) = C_1(x)] \geq \frac{1}{2} + \gamma/2$$

- For $i = 1 \ldots T = O(\frac{1}{\gamma^2 \varepsilon^2})$ stages, stage i: (can stop if Majority$(C_1 \ldots C_i)$ correct on $\geq 1-\varepsilon$ inputs)

  (1) Construct $\mathcal{D}_i$ via "filtering procedure":

  { favor pts on which maj of $C_1 \cdots C_i$ don't do well

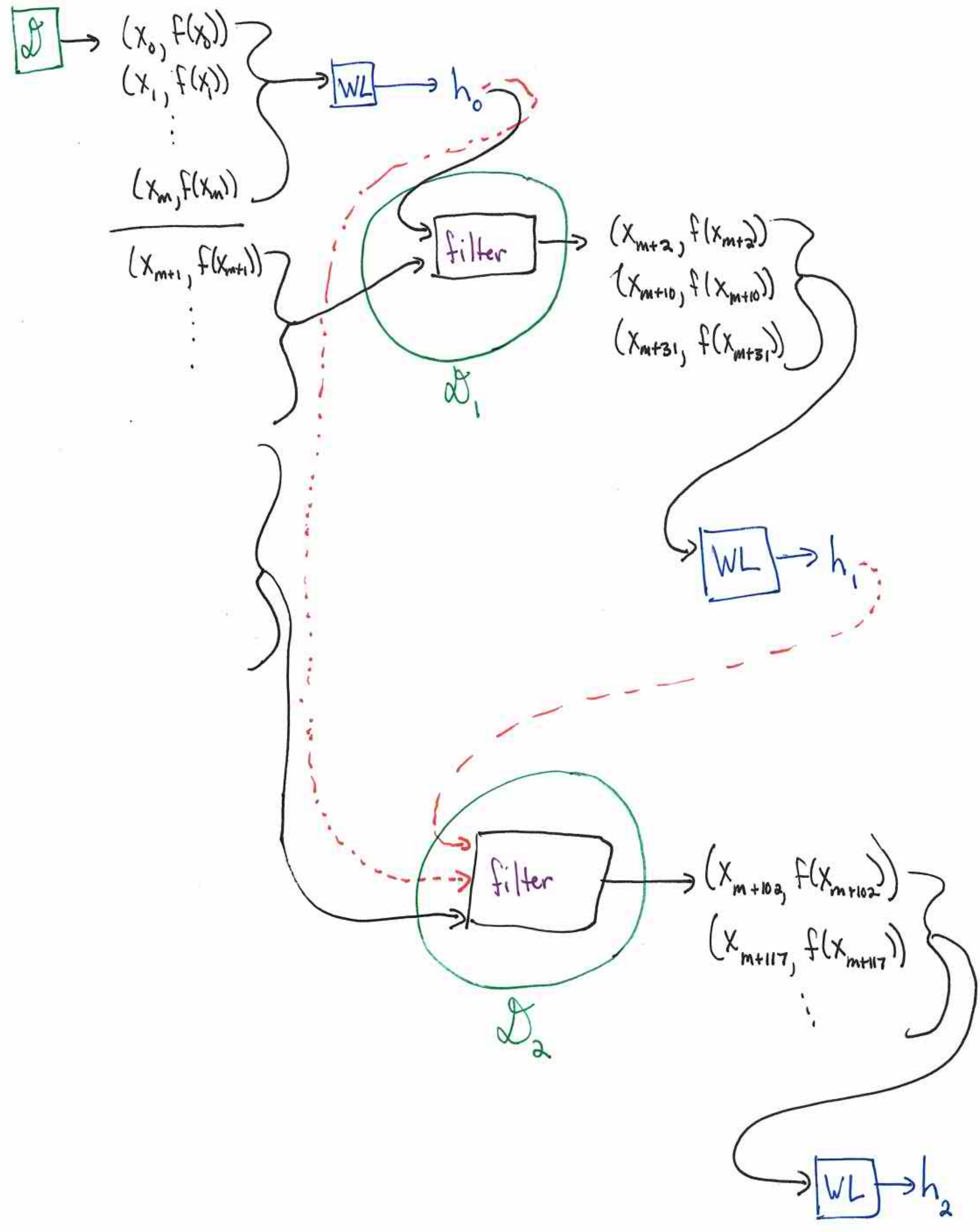
  but also keep some other points }

  Will specify soon

  (2) run WL on examples from $\mathcal{D}_i$ to output

  $$C_{i+1} \text{ s.t. } \Pr_{\mathcal{D}_i}[f(x) = C_{i+1}(x)] \geq \frac{1}{2} + \frac{\gamma}{2}$$

- output $C = MAJ(C_1 \ldots C_T)$


$\mathcal{D}$
$(x_0, f(x_0))$
$(x_1, f(x_1))$
$(x_m, f(x_m))$
WL
h_0
$(x_{m+1}, f(x_{m+1}))$
filter
$\mathcal{D}_1$
$(x_{m+2}, f(x_{m+2}))$
$(x_{m+10}, f(x_{m+10}))$
$(x_{m+31}, f(x_{m+31}))$
WL
h_1
filter
$\mathcal{D}_2$
$(x_{m+102}, f(x_{m+102}))$
$(x_{m+117}, f(x_{m+117}))$
WL
h_2

Filtering procedure

Given new example $x, f(x)$ from example oracle

- if majority of $c_1 \dots c_i$ wrong, Keep it

  ie. $\geq \frac{i}{2}$

- if large majority right, then discard

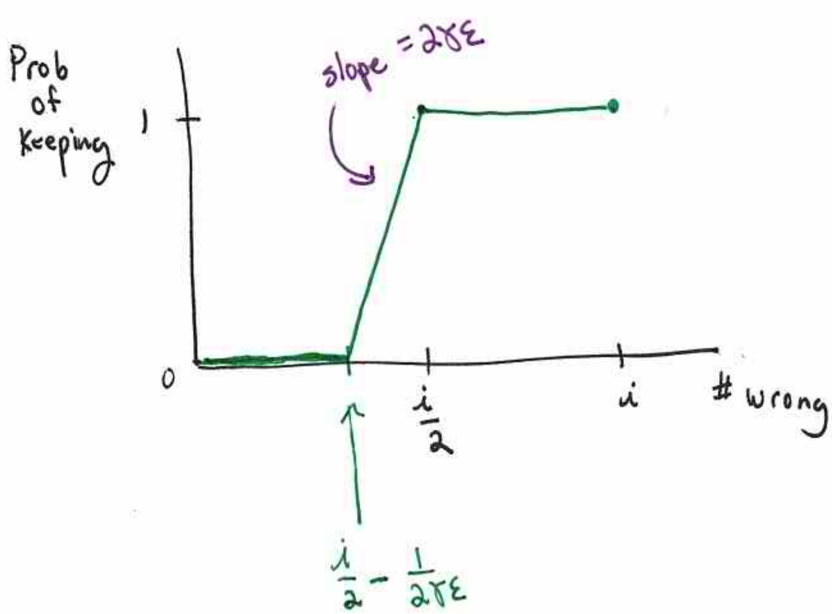
  ie. # right − # wrong $> \frac{1}{\gamma \varepsilon}$

  or # wrong $\leq \frac{i}{2} - \frac{1}{2\gamma\varepsilon}$

- else # right − # wrong $= \frac{\alpha}{\gamma\varepsilon}$ for $0 < \alpha < 1$

  #wrong − #right $= \frac{-\alpha}{\gamma\varepsilon}$

  so Keep with prob $= 1 - \alpha$



Need to show:

1) Output ~~is~~ has nontrivial agreement with f

2) # samples needed not too bad

why could it be bad? if throw out lots of samples, might need to wait a long time before WL can give an output.

<u>but</u> if throw out too many samples then you already have a good hypothesis!

will stop if $Maj(c_1 \cdots c_i)$ correct on $\geq 1-\varepsilon$ fraction of inputs

o.w. $Maj(c_1 \cdots c_i)$ incorrect on $> \varepsilon$ fraction

so filtering procedure outputs sample with prob $\geq \varepsilon$

(+ in expectation, every $1/\varepsilon$ samples of $\mathcal{D}$ at least one makes it thru the filtering system)

$\Rightarrow$ filtering slows down sample collection by $\leq O(1/\varepsilon)$

So lets focus on 1

Notation

- $R_c(x) = \begin{cases} +1 & \text{if } f(x) = c(x) \\ -1 & \text{if } f(x) \neq c(x) \end{cases}$ "is c correct on x?"

- $N_i(x) = \sum_{1 \le j \le i} R_{c_i}(x)$ after iteration i, how many c's correct? (#right − #wrong)

- $M_i(x) = \begin{cases} 1 & \text{if } N_i(x) \le 0 \\ 0 & \text{if } N_i(x) \ge \frac{1}{\varepsilon \gamma} \\ 1 - \varepsilon \cdot \gamma \cdot N_i(x) & \text{o.w.} \end{cases}$ prob of keeping x in filtering (after stage i)

note – all "wrong" x included in M
also some "right" x included

Note that new distribution on samples is proportional to M_i:

- $D_{M_i}(x) \equiv \dfrac{M_i(x)}{\sum_x M_i(x)}$ distribution induced by M

<u>note</u> $D_{M_i}(x) = \mathcal{D}_i$

$\sum_x M_i(x)$ includes all "wrong" x but also x for which maj that isn't overwhelming are correct } upper bounds # wrong x

How correct are we wrt. D_{M_i}?

- $Adv_c(M_i) = \sum_x R_c(x) M_i(x)$ "Advantage" of c on M
$\sim Pr[\text{correct}] - Pr[\text{incorrect}]$
$= 2 \cdot Pr[\text{correct}] - 1$

- $Pr_{x \in D_{M_i}}[c(x) = f(x)] \equiv \frac{1}{2} + \underbrace{\dfrac{Adv_c(M_i)}{2 \cdot \sum_x M_i(x)}}_{\gamma/2}$

Boost (8)

Sp2014

Note:

if $\sum M_i(x) \geq \varepsilon 2^n$

$$Adv_c(M_i) \geq \gamma \cdot \varepsilon \cdot 2^n$$

convert claim about WL ⇒ claim about advantage

ie. if have γ advantage on output of WL

+ still almost wrong on lots of inputs

then new advantage is pretty good

if not, then you are done

Begin Proof

For input x

let $A_i(x) \Leftarrow \sum_{0 \leq j \leq i-1} R_{c_{j+1}}(x) M_j(x)$

strange – indices don't match

$c_1 \cdots c_j$ define $\mathcal{D}_j$

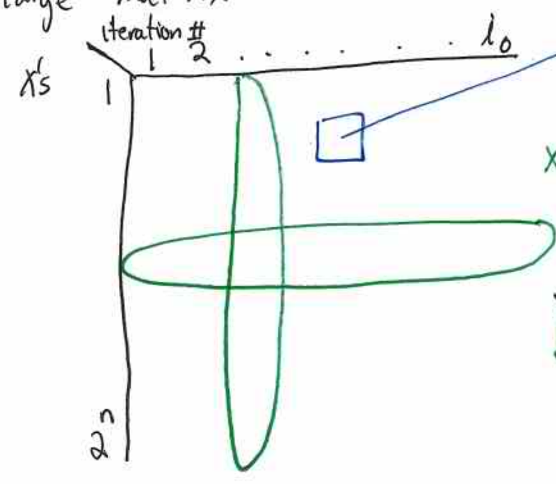
but c_{j+1} learned from WL on $\mathcal{D}_j$

Claim $A_i(x) \leq \frac{1}{\varepsilon\gamma} + \frac{\varepsilon\gamma}{2} \cdot i$

- bounds advantage per input
- only helps after $\frac{1}{\varepsilon\gamma}$ rounds

Plan for use of claim:

Consider large matrix:



$(k,j)^{th}$ Entry: $R_{c_{j+1}}(x_k) M_j(x_k)$

x's row sum $= \sum_{0 \leq j \leq i_0} R_{c_{j+1}}(x) M_j(x) = A_{i_0+1}(x)$

j^{th} col sum $= \sum_x R_{c_{j+1}}(x) M_j(x)$

$= Adv_{c_{j+1}}(M_j) \Leftarrow \geq \gamma \varepsilon \sum_x M_j(x)$

else algorithm stops