Admin: video lectures next Tues, Thurs (Rocco OH 9-11 tomorrow)

PS4 due Tues.

Last time: Boosting: 3-stage boosting, Schapire; 40% error \( \rightarrow \) 35.2% error.

\( \mathcal{Y}_1 = \mathcal{Y} \); \( \mathcal{Y}_2 : \mathcal{Y}_1 \) balanced s.t. 4, useless

\( \mathcal{D}_3 \) focused on "tiebreakers"

Today: Boosting over fixed sample; AdaBoost; analysis.

Reading: Schapire stuff.

Rec. analysis of Schapire booster: say \( \delta = \frac{1}{10} \). Start from \( 3 \beta^2 - 2 \beta \), rec. applied to itself: to reach \( \epsilon \), need \( O(\log \frac{1}{\epsilon}) \) many rec. calls. Hence overall tree of rec. calls has poly \( (\frac{1}{\epsilon}) \) size.
Schap. booster we saw: “boosting by filtering.” Have \( E[X(c, \theta)] \); filter draws from to gen. new distr. Costly in practice.

In practice: have data set \((x', y'), \ldots, (x'', y'')\). **Fixed sample.** “Boosting over a fixed sample.”

- Have \( S \); only pts ever used.
  - \( \Theta \): weighting of pts in \( S \).
  - Init dist \( \Theta(x) = \frac{1}{2} \forall i \in \{1, \ldots, n\} \sum \Theta(x) = 1 \).

New dist \( \Theta' \): diff weighting, again \(-/\Theta'(x) > 0\), \( \sum \Theta'(x) = 1 \).

\[ \sum w_i \leq \frac{1}{2} - \gamma, \quad \text{orig. } \Theta(x), \]

\[ \sum w_i \leq \frac{\gamma}{\gamma'} \quad \text{wtet. } \Theta(x), \]

Want final \( h \) s.t.

\[ (\exists \frac{1}{\gamma} : \text{zero error on orig.}) \]

\[ \frac{1}{m} < \frac{\gamma}{\gamma'} \quad \text{if } h(x) \neq y \]

\[ \Rightarrow \sum \frac{1}{\gamma'} \leq \frac{\gamma}{\gamma'} \]

AdaBoost: simple, powerful, adaptive
BBS alg. \( \sqrt{\text{takes adv. of "extra-good" weak hyp's}} \)

boosting by sampling

One more bit of intuition: \( \alpha_t \) of \( h_t \) in final hyp. is
\[
\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \epsilon_t}{\epsilon_t} \right).
\]

\( \epsilon_t = 0.4: \quad \alpha_t = 0.202. \quad \epsilon_t = 0.01: \quad \alpha_t = 2.29. \)

**AdaBoost update rule:** multi & \( \alpha_t \) on \( x_i \) that were just wrong, \( \alpha_t \) on \( x_i \) just right (assuming adv. \( \epsilon_t > 0 \)).

Like weighted Maj update rule:

**WM**  
**AdaBoost**

\[
\begin{align*}
\text{expert } i = 1, \ldots, m \quad \rightarrow & \quad \text{example } x'_1, \ldots, x'_m \\
\text{pred. of expert } i \text{ at start } \rightarrow & \quad h_t(x_i) \\
 t^{th} \text{ trial } \quad \leftarrow & \quad t^{th} \text{ run of WL} \\
\text{wt of expert } i \text{ at trial } \rightarrow & \quad \beta_t(x_i).
\end{align*}
\]

Now: mistake increases \( \text{wt on } x_i \).

Justify claim that AdaBoost sets up its \( \beta_t \) so as to make \( h_t \) "useless" under \( \beta_{t+1} \) (50% error).
Consider some \( t \), some \( i \). Suppose \( h_t(x^i) = y^i \) \((h_t \) right on \( x^i)\). 

Have \( h_t(x^i) \cdot y^i = 1 \). So 

\[
\exp(-\alpha_t h_t(x^i)y^i) = \exp(-\alpha_t) \exp(\alpha_t) = \exp(-\alpha_t) = \exp(\alpha_t) = \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} = \sqrt{\frac{\epsilon_t}{1-\epsilon_t}}
\]

So ignoring \( \epsilon_t \),

\( D_{t+1}(i) \) is obt from \( D_t(i) \) by 

\[
\sqrt{\frac{\epsilon_t}{1-\epsilon_t}}
\]

\[
\Rightarrow E_x \text{ like this had } 1 - \epsilon_t \text{ wt under } D_t; \text{ new total wt under } D_{t+1} \text{ is } (1 - \epsilon_t) \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} = \sqrt{\epsilon_t(1-\epsilon_t)}.
\]

So new wt of (ex on which \( h_t \) was wrong)

\[
\frac{1 - \epsilon_t}{\epsilon_t} = \text{ new wt of (ex on which } h_t \text{ was right).}
\]

Main thm abt Ada Boost: if \( h_t \)'s are decently good, get hi-acc.

Final hyp.
Thm: Sps run AdaBoost for T stages. (Recall \( \epsilon_t = \Pr_{x \sim \mathcal{D}_t}[h_t(x) \neq y_i] \), \( \epsilon_t = \frac{1}{t} - \epsilon_t \).) The AdaBoost final hyp. \( H(x) \) makes errors on at most a

\[
\frac{1}{T} \sum_{t=1}^{T} \sqrt{1 - 4 \epsilon_t^2} \leq e^{-2 \sum_{t=1}^{T} \epsilon_t^2}
\]

frac. of pts in \( S \).

Pf: 3 claims.

\[1 - 4x \leq e^{-4x}\]
\[\sqrt{1 - 4x} \leq e^{-2x}\]