# Boosting

Chicheng Zhang

CSC 588, University of Arizona

# AdaBoost [1]: recap

Input: training examples $(x_1, y_1), \ldots, (x_m, y_m)$, $\gamma$-weak learner $\mathrm{WL}$

Initial distribution $(D_1(i) = \frac{1}{m})_{i=1}^m$

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

- Weak classifier $h_t \leftarrow \mathrm{WL}$ trained on weighted examples $(x_i, y_i, D_t(i))_{i=1}^m$
- Weighted error $\epsilon_t = \mathbb{P}_{(x,y)\sim D_t}(h_t(x) \neq y_t) \leq \frac{1}{2} - \gamma$
- Classifier weight $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$
- Update weight on training examples:

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

  where $Z_t > 0$ is a normalization factor.

Output final classifier $H_T(x) := \mathsf{sign}(f_T(x))$, where $f_T(x) := \sum_{t=1}^T \alpha_t h_t(x)$

# AdaBoost in action

(See Prof. Rob Schapire's slides)

# AdaBoost: Training error analysis

**Theorem**
*Suppose for every t,* $\epsilon_t \leq \frac{1}{2} - \gamma$, *then*

$$\mathrm{err}(H_T, S) \leq \exp(-2T\gamma^2).$$

**Proof.**

Step 1 : relaxing 0-1 error to exponential loss:

$$\mathrm{err}(H_T, S) \leq \frac{1}{m}\sum_{i=1}^{m} e^{-y_i f_T(x_i)} =: L_T$$

Step 2 : bounding $L_T$ using the normalization factors: $\frac{L_t}{L_{t-1}} = Z_t$

Reason: there exists some $N_t > 0$, such that $D_t(i) = e^{-y_i f_{t-1}(x_i)} \cdot N_t$.

Therefore,

$$\frac{L_t}{L_{t-1}} = \frac{\frac{1}{m}\sum_{i=1}^{m} e^{-y_i f_t(x_i)}}{\frac{1}{m}\sum_{i=1}^{m} e^{-y_i f_{t-1}(x_i)}} = \frac{N_t \cdot \sum_{i=1}^{m} D_t(i) e^{-y_i \alpha_t h_t(x_i)}}{N_t \cdot \sum_{i=1}^{m} D_t(i)} = Z_t$$

# AdaBoost: Training error analysis(cont'd)

**Proof.**

Step 2 : bounding $L_T$ using the normalization factors:
Note: $L_0 = \sum_i e^{-y_i f_0(x_i)}$, where $f_0 \equiv 0$. Therefore, $L_0 = 1$.
Consequently,

$$L_T = L_0 \cdot \frac{L_1}{L_0} \cdot \ldots \frac{L_T}{L_{T-1}} = \prod_{t=1}^{T} Z_t.$$

Step 3 : bounding the normalization factors:

$$\begin{aligned}
Z_t =& \sum_{i=1}^{m} D_t(i) e^{-\alpha_t y_i h_t(x_i)} \\
=& \sum_i D_t(i) e^{-\alpha_t} I(y_i = h_t(x_i)) + \sum_i D_t(i) e^{\alpha_t} I(y_i \neq h_t(x_i)) \\
=& e^{-\alpha_t}(1-\epsilon_t) + e^{\alpha_t}\epsilon_t \\
=& 2\sqrt{\epsilon_t(1-\epsilon_t)} \leq \sqrt{1-4\gamma^2} \leq \exp(-2\gamma^2).
\end{aligned}$$

$\square$

# AdaBoost: generalization error analysis

Question

*How should we choose $T$ in AdaBoost to optimize for generalization error?*

A plausible answer:

- $H_T$ is chosen from hypothesis class

$$\mathcal{H}_T = \left\{ \mathsf{sign}(\sum_{t=1}^{T} \alpha_t h_t(x)) : \alpha \in \mathbb{R}^T, h_1, \ldots, h_T \in \mathcal{B} \right\},$$

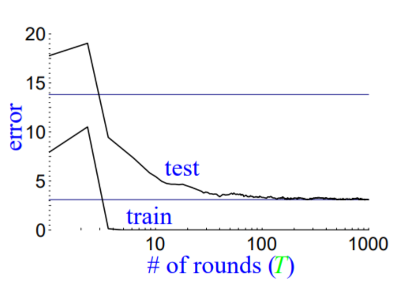
  where $\mathcal{B}$ is the class $\mathrm{WL}$ uses to choose weak classifiers from
- By VC theory:

$$\mathrm{err}(H_T, D) \leq \mathrm{err}(H_T, S) + O\left(\sqrt{\frac{\mathrm{VC}(\mathcal{H}_T)}{m}}\right),$$

  where $\mathrm{err}(H_T, S)$ decreases and $\mathrm{VC}(\mathcal{H}_T)$ increases in $T$
- So there is some tradeoff in the choice of $T$

# Theory vs. Practice

A typical learning curve of AdaBoost [2]:




How to explain this discrepancy between theory and practice?

# Margin-based generalization bounds for boosting [2]

**Theorem**
*Suppose base class $\mathcal{B}$ is finite, $\mathcal{C}(\mathcal{B}) = \left\{\sum_{h \in \mathcal{B}} \alpha_h h(x) : \sum_{h \in \mathcal{B}} |\alpha_h| \leq 1\right\}$ is the set of voting classifiers over $\mathcal{B}$. Fix margin $\theta \in [0, 1]$. Then, with probability $1 - \delta$, for all $f \in \mathcal{C}(\mathcal{B})$,*

$$\mathbb{P}_D(yf(x) \leq 0) \leq \underbrace{\mathbb{P}_S(yf(x) \leq \theta)}_{\text{“Margin error” of } f} + O\left(\frac{1}{\theta}\sqrt{\frac{\ln |\mathcal{B}|/\delta}{m}}\right)$$

Application to AdaBoost:

1. Let $\bar{f}_T(x) = \frac{\sum_{t=1}^{T} \alpha_t h_t(x)}{\sum_{t=1}^{T} \alpha_t} \in \mathcal{C}(\mathcal{B})$, and $\theta = \frac{\gamma}{2}$
2. $\mathbb{P}_S(y f_T(x) \leq \frac{\gamma}{2}) \leq \exp(-T\gamma^2)$
3. The “complexity term” $O\left(\frac{1}{\gamma}\sqrt{\frac{\ln |\mathcal{B}|/\delta}{m}}\right)$ is independent of $T$.

# What have we learned?

- Boosting: generic procedure that converts weak PAC learners to strong PAC learners
- AdaBoost's training error analysis
- AdaBoost's generalization error analysis: VC-based vs. margin-based

# References

[1] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[2] Robert E Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.