On the Margin Explanation of Boosting Algorithms

(Wang, et al., COLT 2008)

Presented by Ameet Talwalkar, 10/7/08

Motivation

- "Finding many rough rules of thumb can be a lot easier than finding a single, highly accurate prediction rule" [Schapire, 2002]

- Produce highly accurate prediction rule by combining several moderately accurate rules of thumb

Motivation

- "Finding many rough rules of thumb can be a lot easier than finding a single, highly accurate prediction rule" [Schapire, 2002]

- Produce highly accurate prediction rule by combining several moderately accurate rules of thumb

- Goal: Explain why boosting works?

What is Boosting?

- Binary Classification
- Base classifier/weak learner
  - “moderately accurate rule of thumb”
  - Trained on (weighted) training data
- Boosting: combination of weak learners
  - Weighted majority vote

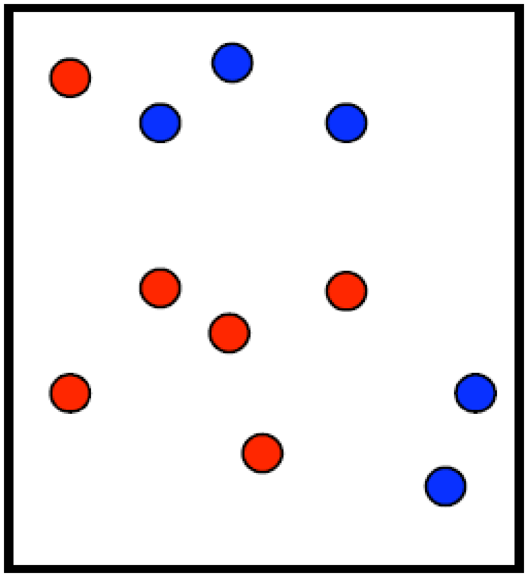
Outline

- Adaboost
  - Algorithm, Example, Analysis and Experiments
- Initial Margin bounds
  - Margin distribution and Min margin
- New Margin Bounds
  - E-margin

Outline

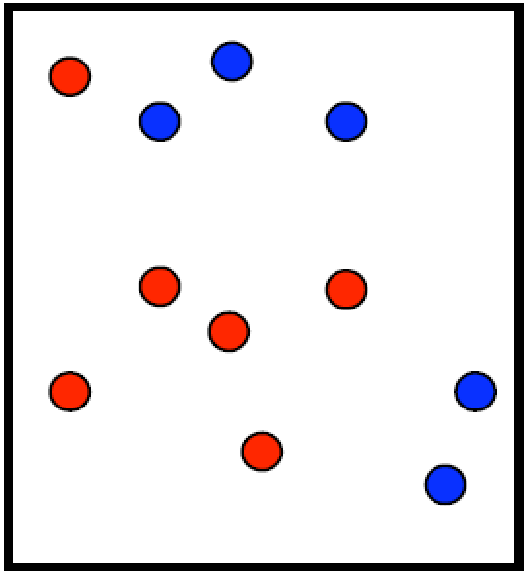
- Adaboost
  - Algorithm, Example, Analysis and Experiments
- Initial Margin bounds
  - Margin distribution and Min margin
- New Margin Bounds
  - E-margin

Adaboost [Freund & Schapire, 1997]



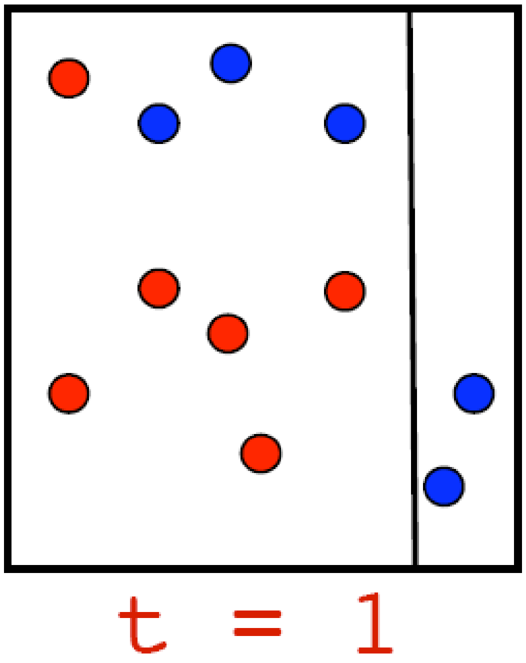
- Focus on “hard” training points
  - Increase mass of points misclassified by previous WL
- Weighted majority classification
  - More weight for accurate WLs

Adaboost [Freund & Schapire, 1997]

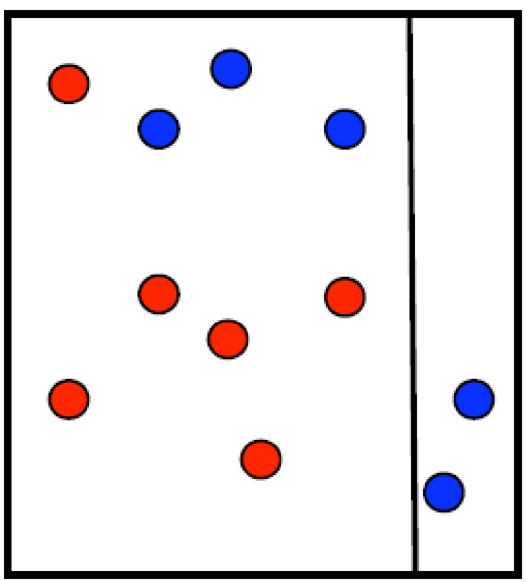


Animation from
[Mohri, FML lecture 8]

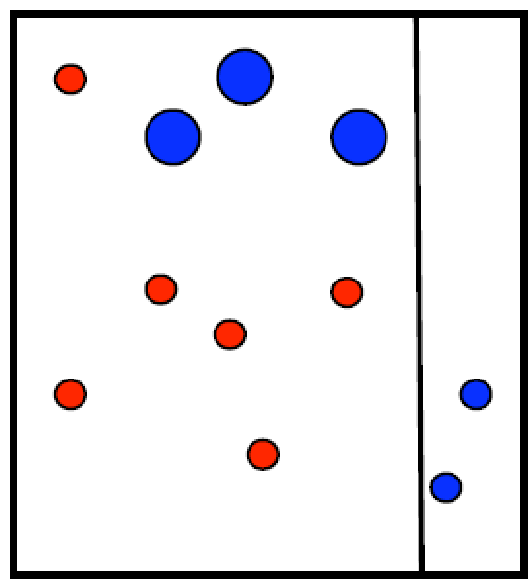
Adaboost [Freund & Schapire, 1997]



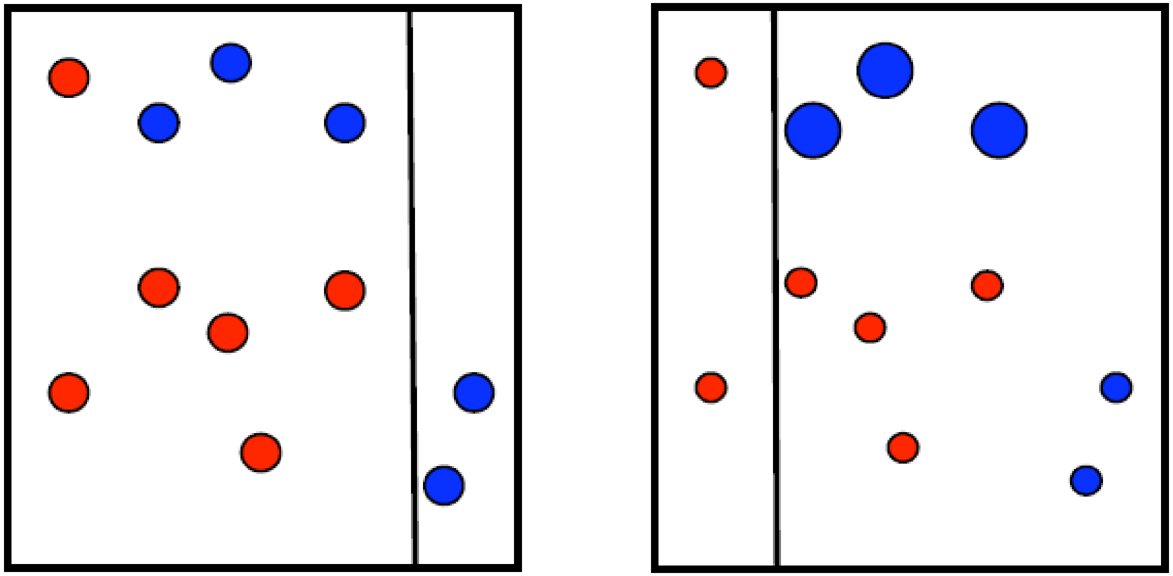
Adaboost [Freund & Schapire, 1997]



t = 1

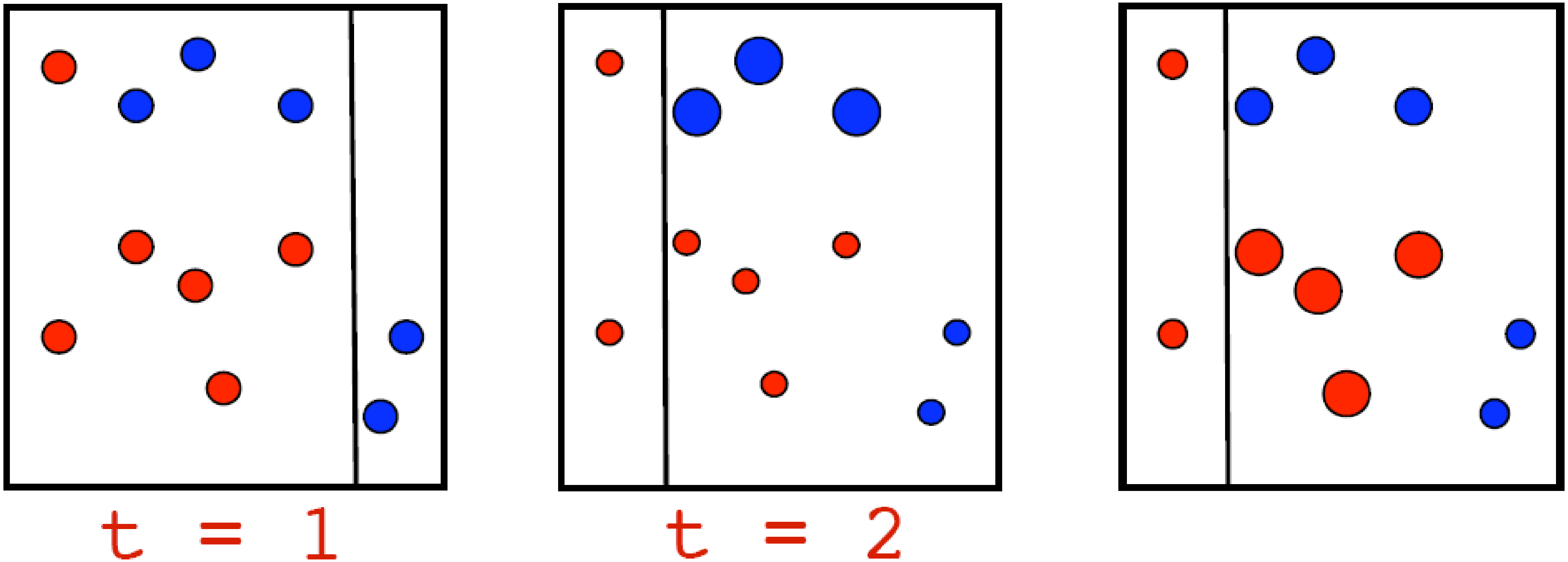


Adaboost [Freund & Schapire, 1997]

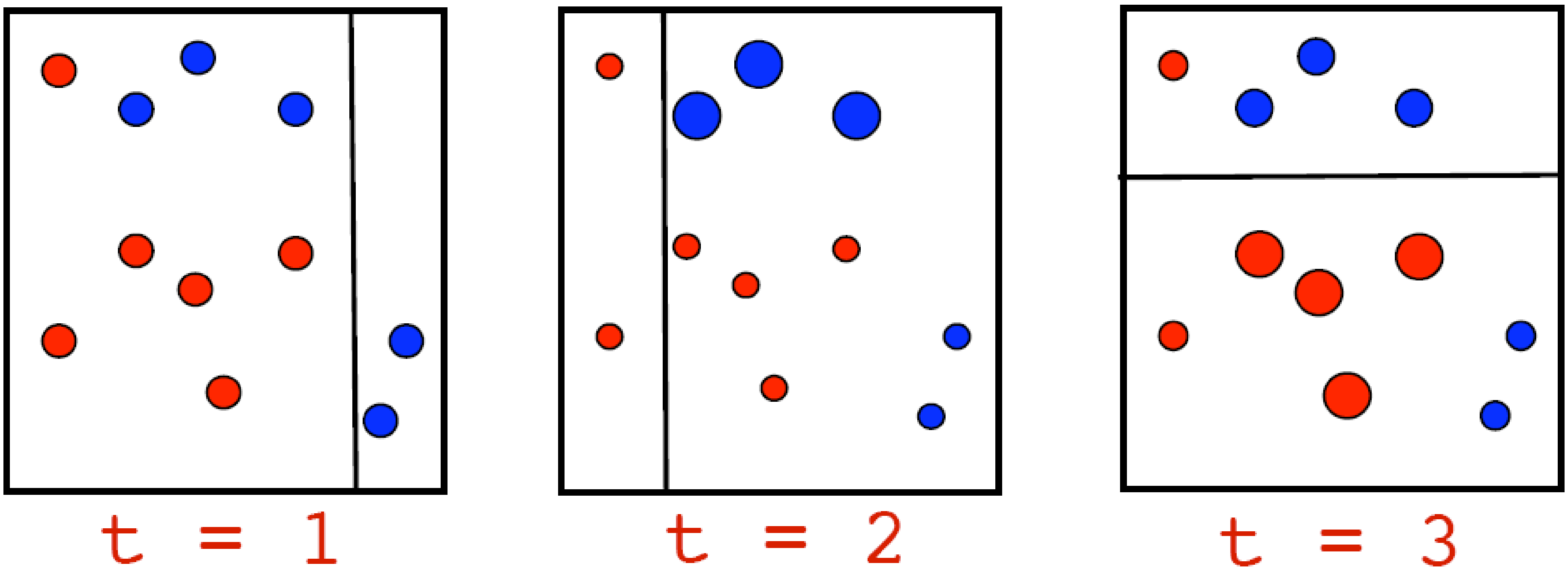


t = 1   t = 2

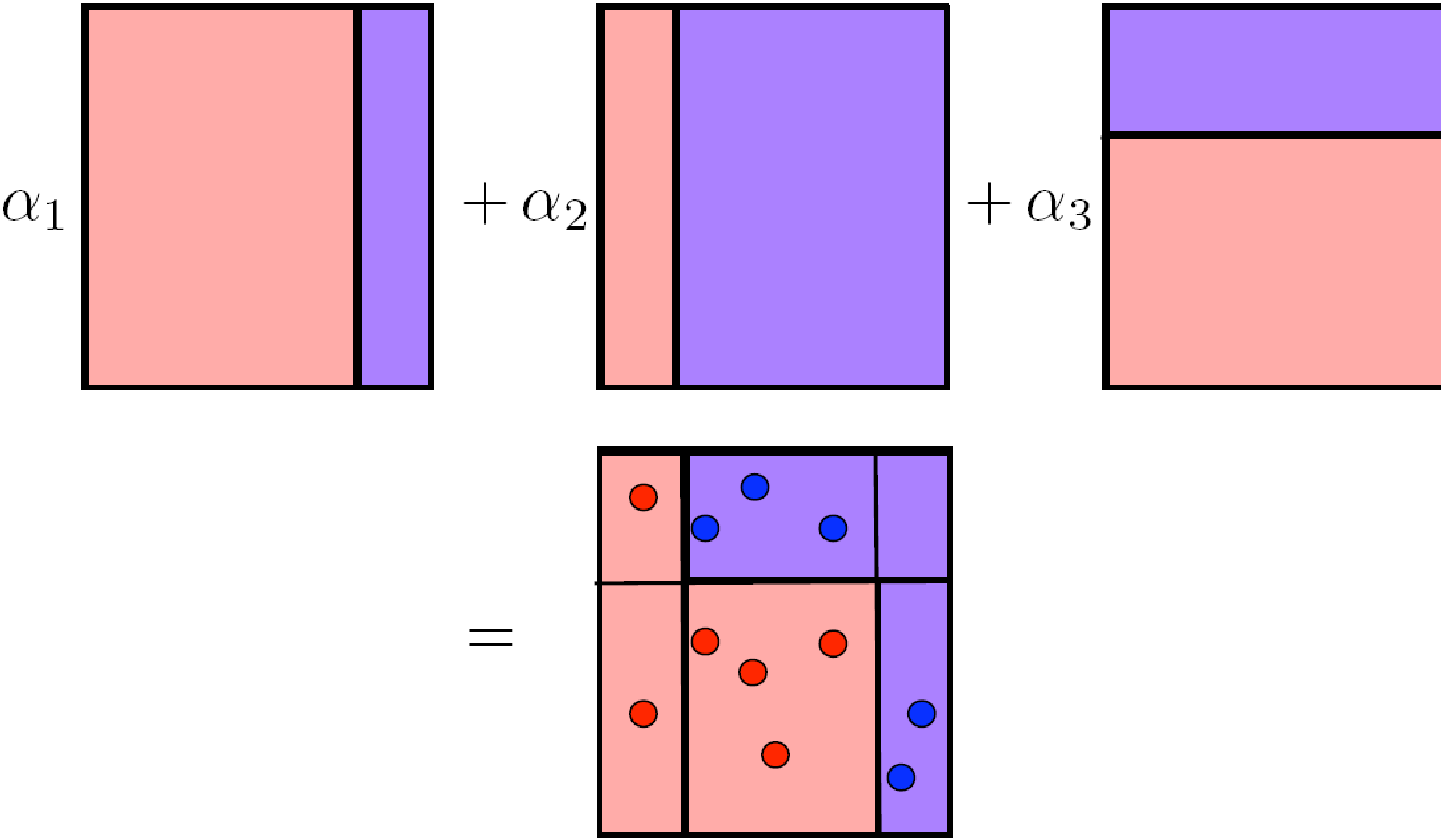
Adaboost [Freund & Schapire, 1997]



Adaboost [Freund & Schapire, 1997]



Adaboost [Freund & Schapire, 1997]



Adaboost [Freund & Schapire, 1997]

$$\textbf{Input}:\ S = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$
$$\text{where } x_i \in X,\ y_i \in \{-1, 1\}.$$

Adaboost [Freund & Schapire, 1997]

Input: $S = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
where $x_i \in X$, $y_i \in \{-1, 1\}$.

Create WLs: $h_t : \bar{X} \rightarrow \{-1, 1\}$ (next slide)

Adaboost [Freund & Schapire, 1997]

Input: $S = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
where $x_i \in X$, $y_i \in \{-1, 1\}$.

Create WLs: $h_t : \bar{X} \to \{-1, 1\}$ (next slide)

Output: $f(x) = \operatorname{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

where $\sum \alpha_i = 1, \quad \alpha_i \geq 0$

Adaboost [Freund & Schapire, 1997]

Initialization: $D_1(i) = 1/n.$

Adaboost [Freund & Schapire, 1997]

Initialization: $D_1(i) = 1/n.$

for $t = 1$ **to** T **do**

1. Train base learner using $D_t \left[h_t : \bar{X} \rightarrow \{-1, 1\} \right]$

end

Adaboost [Freund & Schapire, 1997]

Initialization: $D_1(i) = 1/n$.

for $t = 1$ **to** T **do**

1. Train base learner using $D_t \left[h_t : X \to \{-1, 1\} \right]$
2. Choose $\alpha_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$

error: $\Pr_{D_t}[h_t(x_i) \neq y_i]$

end

Adaboost [Freund & Schapire, 1997]

Initialization: $D_1(i) = 1/n$.

for $t = 1$ **to** T **do**

1. Train base learner using $D_t \left[h_t : X \rightarrow \{-1, 1\} \right]$
2. Choose $\alpha_t = \dfrac{1}{2} \log \dfrac{1-\epsilon_t}{\epsilon_t}$

   error: $\Pr_{D_t}[h_t(x_i) \neq y_i]$

3. Update:

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

Normalization constant

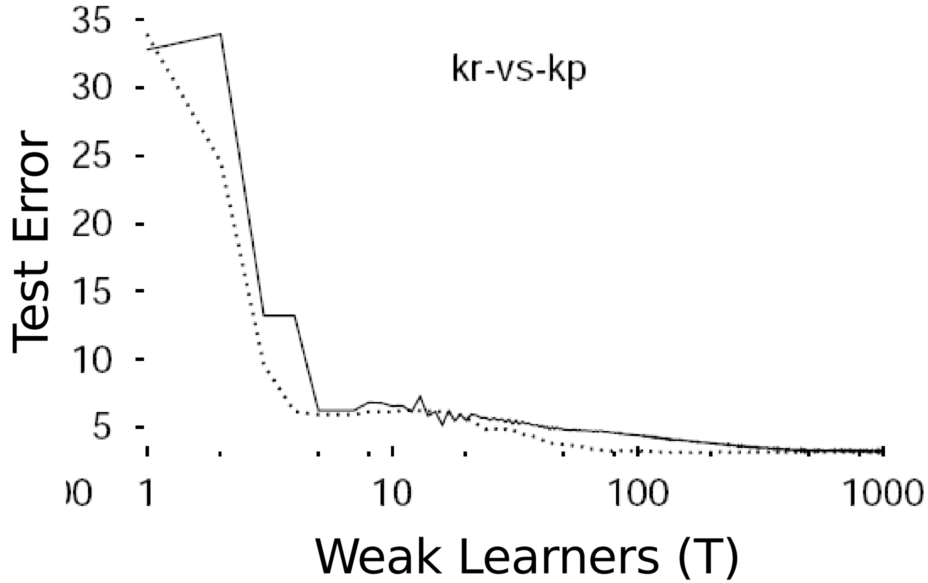
end

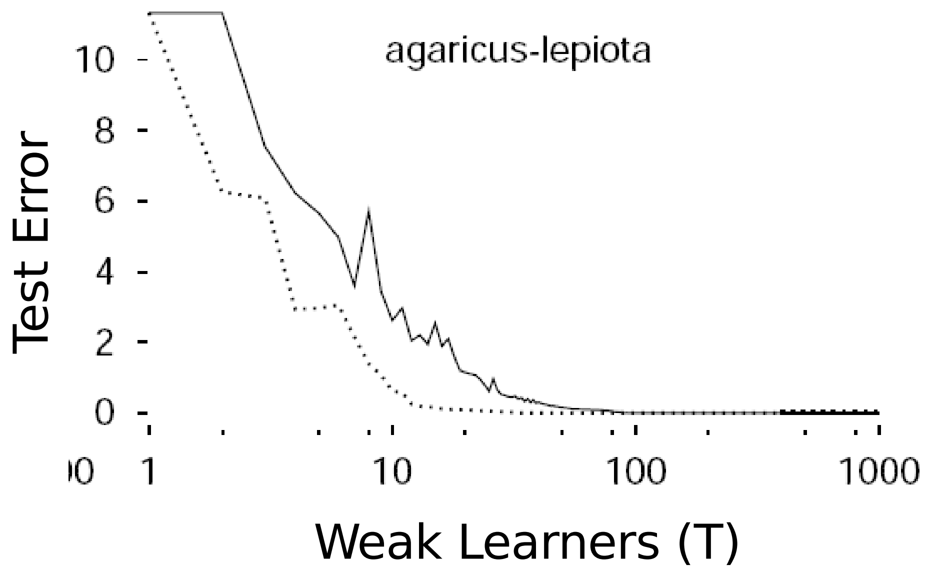
Boosting Example [Mohri, FML lecture 8] [Schapire & Singer, 1999]

- n training points, m features
- Weak learners = Decision stumps (trees of length 1)
- Algorithm:
  - Associate stump with each feature
  - Pre-sort each feature: $O(mn \log n)$
  - Find best feature/threshold at each round: $O(mnT)$

Boosting Example [Mohri, FML lecture 8] [Schapire & Singer, 1999]

- n training points, m features
- Weak learners = Decision stumps (trees of length 1)
- Algorithm:
  - Associate stump with each feature
  - Pre-sort each feature: $O(mn \log n)$
  - Find best feature/threshold at each round: $O(mnT)$





Initial Adaboost Theory

- Complexity (VC dim) of combined classifiers increases as T increases [Freund & Schapire, 1997]

- Initial analysis showed typical tradeoff between training error and complexity hypothesis class
  - overfitting as T increases

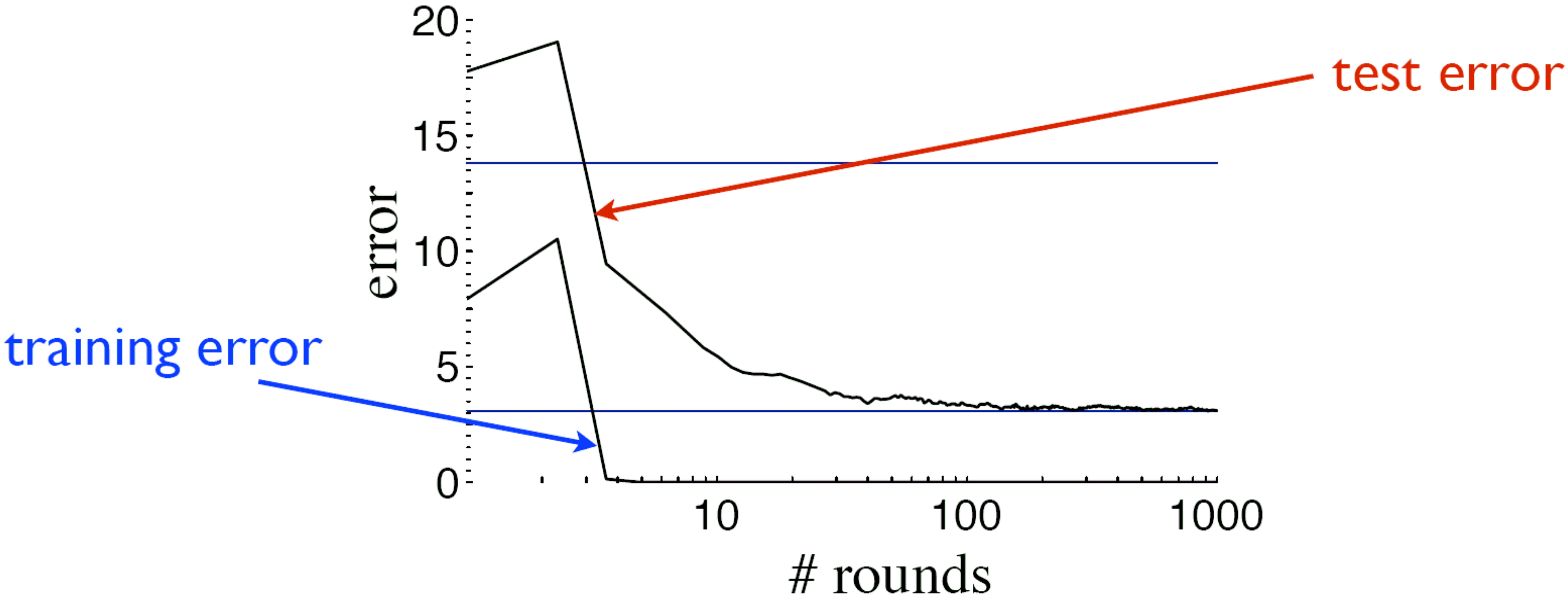
Adaboost Experiments

- Theory implied overfitting with increased T

Adaboost Experiments

- Theory implied overfitting with increased T

- Empirical evidence does NOT follow
  - E.g., Boosting C4.5 trees [Freund & Schapire, 1998] [Mohri, FML lecture 8]



Outline

- Adaboost
  - Algorithm, Example, Analysis and Experiments
- Initial Margin bounds
  - Margin distribution and Min margin
- New Margin Bounds
  - E-margin

Why margin?

- Key idea: Look at *confidence of training classification (margin)* instead of number of incorrect classifications

Why margin?

- Key idea: Look at *confidence of training classification (margin)* instead of number of incorrect classifications

- Goal: bound that depend on margin and complexity of WLs but NOT on # of WLs

Why margin?

- Key idea: Look at *confidence of training classification (margin)* instead of number of incorrect classifications

- Goal: bound that depend on margin and complexity of WLs but NOT on # of WLs

- Results [SFBL, 1998]
  - margin bound on error for *any voting classifier*
  - bound on margin distribution for Adaboost

Margin definitions

- margin of (x, y): confidence of prediction
  - ranges from -1 to 1
  - $yf(x) = \sum_{i:y=h_i(x)} \alpha_i - \sum_{i:y\neq h_i(x)} \alpha_i$

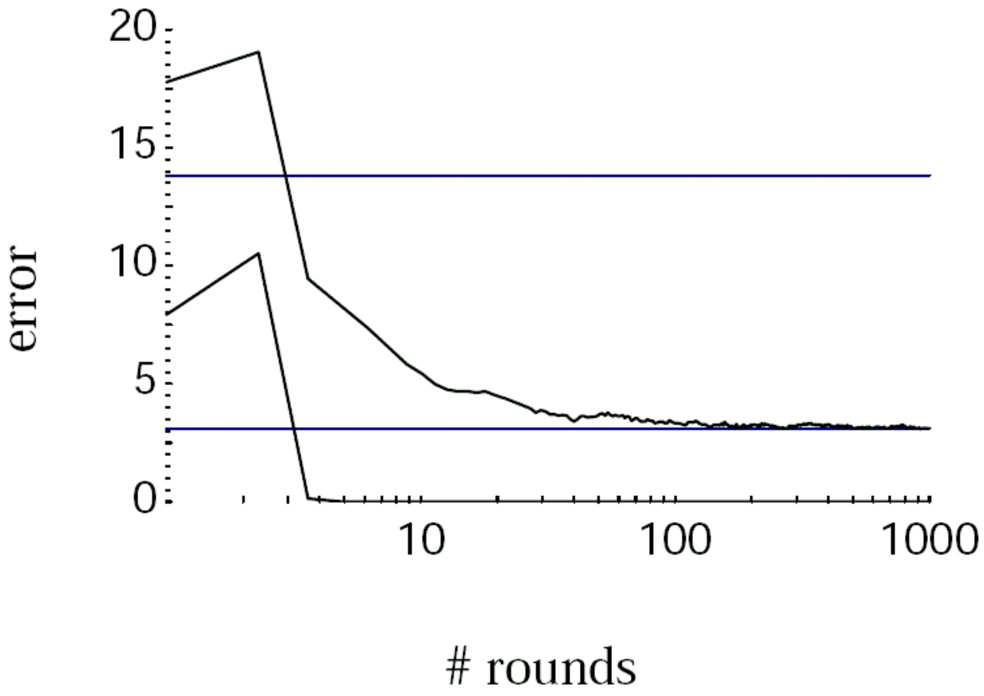
Margin definitions

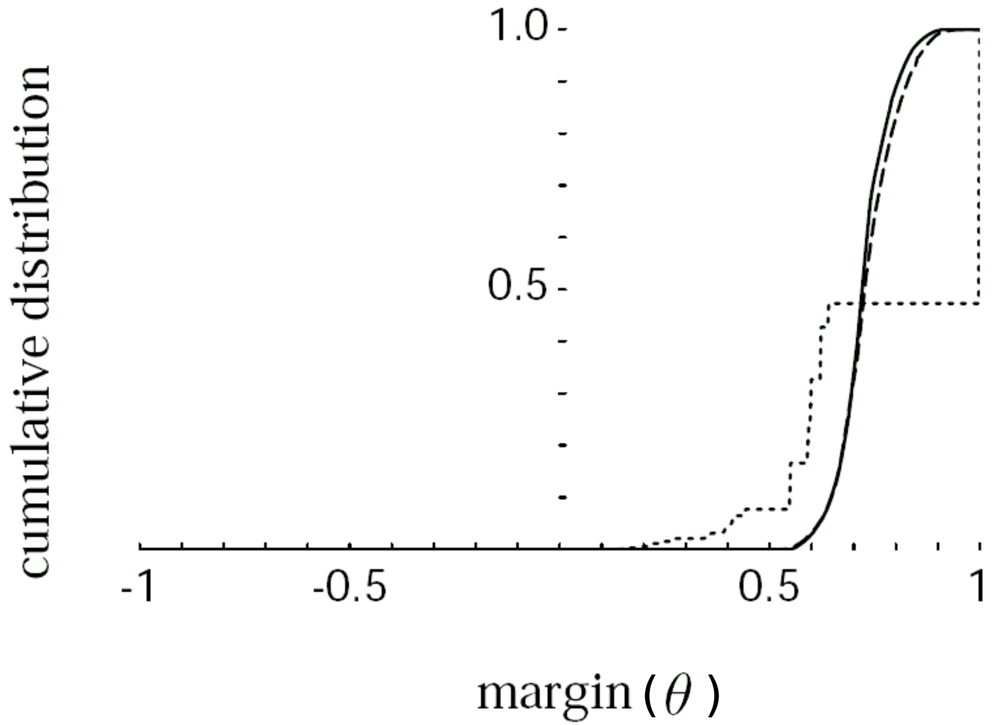
- margin of (x, y): confidence of prediction
  - ranges from -1 to 1
  - $yf(x) = \sum_{i:y=h_i(x)} \alpha_i - \sum_{i:y\neq h_i(x)} \alpha_i$
- margin distribution over training set (*S*)
  - fraction with margin at most θ , $-1 \leq \theta \leq 1$
  - $P_S(yf(x) \leq \theta)$

Margin definitions

- margin of (x, y): confidence of prediction
  - ranges from -1 to 1
  - $$yf(x) = \sum_{i:y=h_i(x)} \alpha_i - \sum_{i:y\neq h_i(x)} \alpha_i$$
- margin distribution over training set (*S*)
  - fraction with margin at most θ , $-1 \leq \theta \leq 1$
  - $P_S(yf(x) \leq \theta)$
- min margin: smallest margin over S
  - max θ s.t. $P_S(yf(x) \leq \theta) = 0$

Margin Distribution

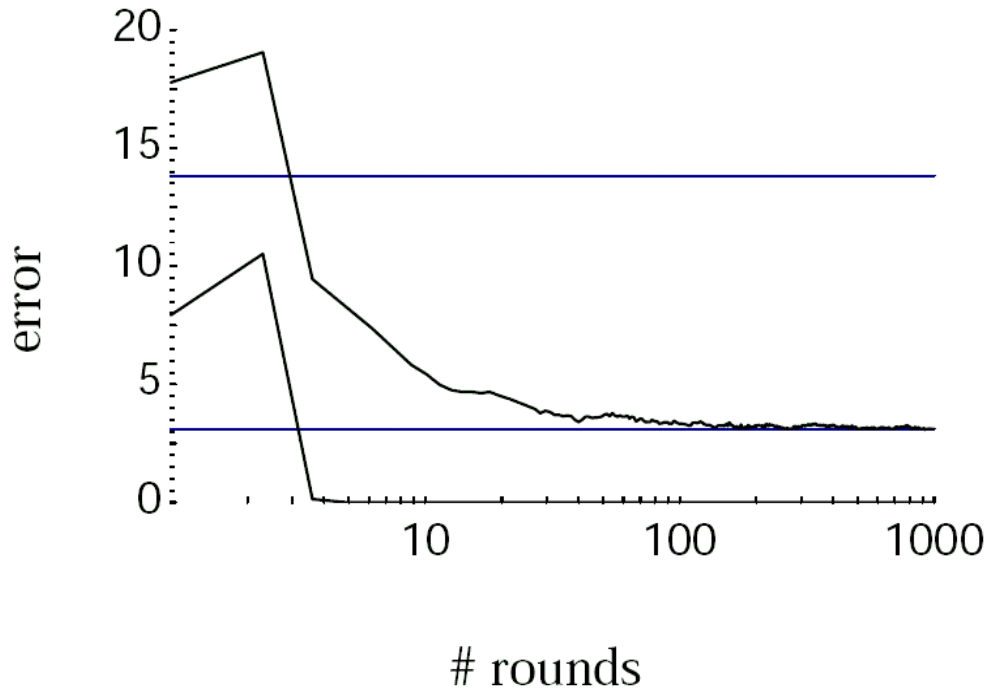


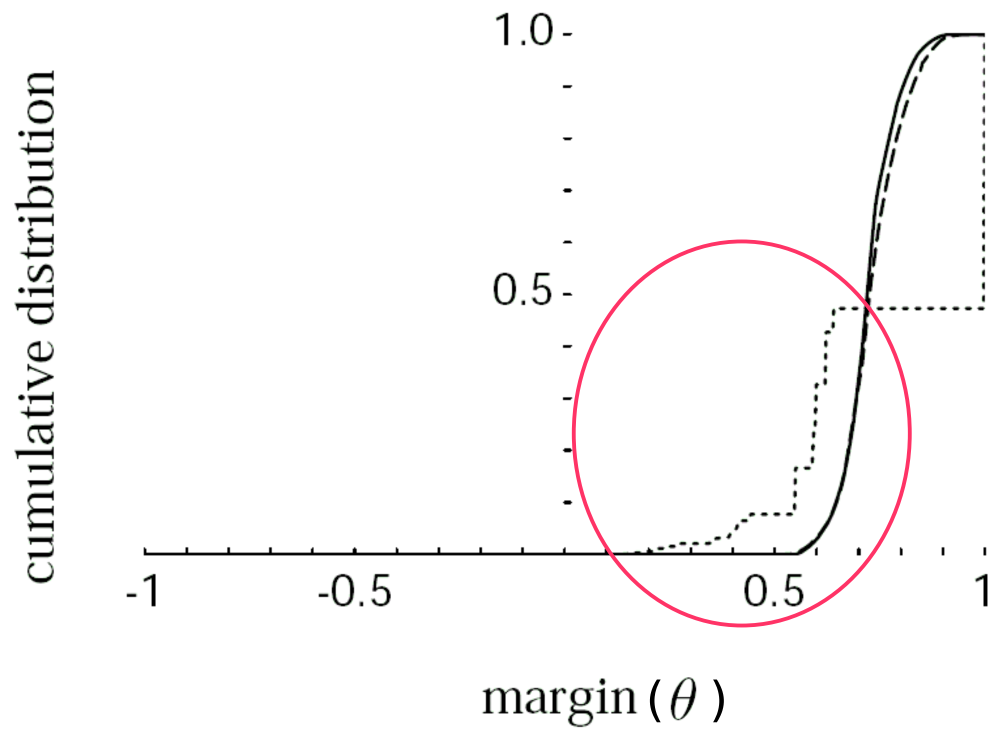


LEGEND: (small dash, large dash, solid) lines equal (5, 100, 1000) rounds of boosting

- Adaboost with C4.5 trees [Freund & Schapire, 1998]

Margin Distribution





LEGEND: (small dash, large dash, solid) lines equal (5, 100, 1000) rounds of boosting

- Adaboost with C4.5 trees [Freund & Schapire, 1998]
- Margin distribution "improves" with more rounds ($P_S(yf(x) \leq \theta)$)

Margin Distribution Bound

Theorem 1 *[SFBL98] For any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of the training set S of n examples, every voting classifier f satisfies the following bounds:*

$$P_D\Big(yf(x) \leq 0\Big) \leq \inf_{\theta \in (0,1]} \Bigg[P_S\Big(yf(x) \leq \theta\Big) + O\left(\frac{1}{\sqrt{n}}\left(\frac{d \log^2(n/d)}{\theta^2} + \log\frac{1}{\delta}\right)^{1/2}\right)\Bigg],$$

where d is the VC dimension of H.

Margin Distribution Bound

Theorem 1 *[SFBL98] For any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of the training set S of n examples, every voting classifier f satisfies the following bounds:*

$$P_D\Big(yf(x) \le 0\Big) \le \inf_{\theta \in (0,1]} \Bigg[P_S\Big(yf(x) \le \theta\Big) + O\left(\frac{1}{\sqrt{n}}\left(\frac{d \log^2(n/d)}{\theta^2} + \log\frac{1}{\delta}\right)^{1/2}\right)\Bigg],$$

where d is the VC dimension of H.

Margin Distribution Bound

Theorem 1 *[SFBL98] For any $\delta > 0$, with probability at least $1-\delta$ over the random choice of the training set S of n examples, every voting classifier f satisfies the following bounds:*

$$P_D\left(\boxed{yf(x) \le 0}\right) \le \inf_{\theta \in (0,1]} \left[P_S\left(yf(x) \le \theta\right) + O\left(\boxed{\frac{1}{\sqrt{n}}}\left(\frac{\boxed{d \log^2(n/d)}}{\theta^2} + \boxed{\log\frac{1}{\delta}}\right)^{1/2}\right)\right],$$

Generalization Error

where d is the VC dimension of H.

Dependence on size of training set (*n*), VC dim of WLs (*d*) and confidence parameter

Margin Distribution Bound

Theorem 1 *[SFBL98] For any $\delta > 0$, with probability at least $1-\delta$ over the random choice of the training set S of n examples, every voting classifier f satisfies the following bounds:*

$$P_D\Big(yf(x) \leq 0\Big) \leq \inf_{\theta \in (0,1]} \left[\boxed{P_S\Big(yf(x) \leq \theta\Big)} + O\left(\frac{1}{\sqrt{n}} \left(\frac{d \log^2(n/d)}{\boxed{\theta^2}} + \log \frac{1}{\delta} \right)^{1/2} \right) \right],$$

where d is the VC dimension of H.

Want most examples to have large margin so that we have small $P_S(yf(x) \leq \theta)$ for not too small θ

Adaboost's Margin

THEOREM 5. *Suppose the base learning algorithm, when called by AdaBoost, generates classifiers with weighted training errors* $\varepsilon_1, \ldots, \varepsilon_T$. *Then for any* θ, *we have that*

$$\mathbf{P}_{(x,y)\sim S}\left[yf(x) \leq \theta \right] \leq 2^T \prod_{t=1}^{T} \sqrt{\varepsilon_t^{1-\theta} (1-\varepsilon_t)^{1+\theta}}$$

[SFBL, 1998]

Adaboost's Margin

THEOREM 5. *Suppose the base learning algorithm, when called by AdaBoost, generates classifiers with weighted training errors $\varepsilon_1, \ldots, \varepsilon_T$. Then for any θ, we have that*

$$\mathbf{P}_{(x,y)\sim S}[yf(x) \leq \theta] \leq 2^T \prod_{t=1}^{T} \sqrt{\varepsilon_t^{1-\theta}(1-\varepsilon_t)^{1+\theta}}$$

[SFBL, 1998]

- If $\varepsilon_t \leq 1/2 - \gamma$ for all *t* and $\theta < \gamma$, then we have exponential convergence in *T*
  - Exponential convergence for small enough θ

Min-Margin Bound

- Tighter bounds exist for min-margin

  [Breiman, 1999]

- Adaboost does not maximize min margin

  [Rudin, 2004]

Min-Margin Bound

- Tighter bounds exist for min-margin [Breiman, 1999]
- Adaboost does not maximize min margin [Rudin, 2004]
- Arc-GV [Breiman, 1999]
  - converges to min-margin
  - larger min-margin than Adaboost in practice

Min-Margin Bound

- Tighter bounds exist for min-margin [Breiman, 1999]
- Adaboost does not maximize min margin [Rudin, 2004]
- Arc-GV [Breiman, 1999]
  - converges to min-margin
  - larger min-margin than Adaboost in practice
  - worse generalization error!

Min-Margin Bound

- Tighter bounds exist for min-margin
[Breiman, 1999]

- Adaboost does not maximize min margin
[Rudin, 2004]

- Arc-GV [Breiman, 1999]
  - converges to min-margin
  - larger min-margin than Adaboost in practice
  - worse generalization error!

- So what really drives performance?

Outline

- Adaboost
  - Algorithm, Example, Analysis and Experiments
- Initial Margin bounds
  - Margin distribution and Min margin
- New Margin Bounds
  - E-margin

E-Margin Bound

- Goal: Bounds that explain empirical results
  - clarify Adaboost vs Arc-GV issue

E-Margin Bound

- Goal: Bounds that explain empirical results
  - clarify Adaboost vs Arc-GV issue

- Bernoulli Relative Entropy

$$D(q||p) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}, \qquad 0 \leq p, q \leq 1$$

E-Margin Bound

- Goal: Bounds that explain empirical results
  - clarify Adaboost vs Arc-GV issue

- Bernoulli Relative Entropy

$$D(q||p) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}, \qquad 0 \leq p, q \leq 1$$

- Inverse for fixed *q* and $q \leq p \leq 1$
  - exists b/c $D(q||p)$ is a monotone increasing fct
  - not defined in paper

E-Margin Bound

Warning: This is a monster!!!

E-Margin Bound

Theorem 3 *If* $|H| < \infty$, *then for any* $\delta > 0$, *with probability at least* $1-\delta$ *over the random choice of the training set* S *of* n *examples, every voting classifier* f *satisfies the following bound:*

$$P_D\Big(yf(x) \le 0\Big) \le \frac{\log|H|}{n} + \inf_{q \in \{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}} D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right), \qquad (3)$$

where

$$u\left[\hat{\theta}(q)\right] = \frac{1}{n}\left(\frac{8}{\hat{\theta}^2(q)} \log\left(\frac{2n^2}{\log|H|}\right)\log|H| + \log|H| + \log\frac{n}{\delta}\right),$$

and $\hat{\theta}(q)$ *is given by*

$$\hat{\theta}(q) = \sup\left\{\theta \in \left(\sqrt{8/|H|}, 1\right] : P_S\Big(yf(x) \le \theta\Big) \le q\right\}.$$

E-Margin Bound

Theorem 3 *If $|H| < \infty$, then for any $\delta > 0$, with probability at least $1-\delta$ over the random choice of the training set S of n examples, every voting classifier f satisfies the following bound:*

$$P_D\Big(yf(x) \le 0\Big) \le \frac{\log|H|}{n} + \inf_{q\in\{0,\frac{1}{n},\frac{2}{n},\ldots,1\}} D^{-1}\left(q, u\left[\hat\theta(q)\right]\right), \qquad (3)$$

where

$$u\left[\hat\theta(q)\right] = \frac{1}{n}\left(\frac{8}{\hat\theta^2(q)}\log\left(\frac{2n^2}{\log|H|}\right)\log|H| + \log|H| + \log\frac{n}{\delta}\right),$$

and $\hat\theta(q)$ is given by

$$\hat\theta(q) = \sup\left\{\theta \in \left(\sqrt{8/|H|}, 1\right] : P_S\Big(yf(x) \le \theta\Big) \le q\right\}.$$

E-margin

$$\theta^* = \hat\theta(q^*)$$

E-margin error

Analysis of E-margin Bound

- Reduction to Margin Distribution bound (and proof relies on very similar techniques)

Reduction to Previous Bound

Lemma 9

$$\inf_q D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right) \leq \inf_q \left(\boxed{q} + \left(\frac{u\left[\hat{\theta}(q)\right]}{2}\right)^{1/2}\right)$$

$$\Downarrow$$

Margin Distribution Bound [SFBL, 1998]

$$P_D\Big(yf(x) \leq 0\Big) \leq \inf_{\theta \in (0,1]} \left[\boxed{P_S\Big(yf(x) \leq \theta\Big)} + O\left(\frac{1}{\sqrt{n}}\left(\frac{\log n \log |H|}{\theta^2} + \log\frac{1}{\delta}\right)^{1/2}\right)\right]$$

Analysis of E-margin Bound

- Reduction to Margin Distribution bound (and proof relies on very similar techniques)

- Tighter than min-margin bound (Theorem 6)
  - suggests that E-margin (not min-margin) dictates generalization error!

Analysis of E-margin Bound

- Reduction to Margin Distribution bound (and proof relies on very similar techniques)

- Tighter than min-margin bound (Theorem 6)
  - suggests that E-margin (not min-margin) dictates generalization error!

- Theorem 7 – Compare voting classifiers: If voting classifier *f1* has larger E-margin and smaller E-margin error than another voting classifier (*f2*), then *f1* will have smaller generalization error

Analysis of E-margin Bound

- Reduction to Margin Distribution bound (and proof relies on very similar techniques)

- Tighter than min-margin bound (Theorem 6)
  - suggests that E-margin (not min-margin) dictates generalization error!

- Theorem 7 – Compare voting classifiers: If voting classifier *f1* has larger E-margin and smaller E-margin error than another voting classifier (*f2*), then *f1* will have smaller generalization error
  - Is this true empirically?

E-margin Experiments

		Emargin	Emargin Error
Breast	AdaBoost	**0.313**	**0.803**
	arc-gv	0.281	0.909
Diabetes	AdaBoost	**0.110**	**0.748**
	arc-gv	0.049	0.759
German	AdaBoost	0.157	0.824
	arc-gv	0.034	0.780
Image	AdaBoost	**0.196**	**0.610**
	arc-gv	0.195	0.705
Ionosphere	AdaBoost	0.323	0.800
	arc-gv	0.131	0.577
Letter	AdaBoost	**0.078**	**0.645**
	arc-gv	0.063	0.958
Satimage	AdaBoost	**0.133**	**0.521**
	arc-gv	0.133	0.956
USPS	AdaBoost	**0.108**	**0.972**
	arc-gv	0.053	0.990
Vehicle	AdaBoost	**0.129**	**0.737**
	arc-gv	0.052	0.794
Wdbc	AdaBoost	**0.350**	**0.581**
	arc-gv	0.350	0.710

E-margin Experiments

		Emargin	Emargin Error	Test Error
Breast	AdaBoost	**0.313**	**0.803**	**0.052**
	arc-gv	0.281	0.909	0.057
Diabetes	AdaBoost	**0.110**	**0.748**	**0.255**
	arc-gv	0.049	0.759	0.256
German	AdaBoost	0.157	0.824	0.258
	arc-gv	0.034	0.780	0.261
Image	AdaBoost	**0.196**	**0.610**	0.023
	arc-gv	0.195	0.705	**0.021**
Ionosphere	AdaBoost	0.323	0.800	0.100
	arc-gv	0.131	0.577	0.106
Letter	AdaBoost	**0.078**	**0.645**	**0.174**
	arc-gv	0.063	0.958	0.178
Satimage	AdaBoost	**0.133**	**0.521**	**0.053**
	arc-gv	0.133	0.956	0.057
USPS	AdaBoost	**0.108**	**0.972**	**0.450**
	arc-gv	0.053	0.990	0.460
Vehicle	AdaBoost	**0.129**	**0.737**	**0.297**
	arc-gv	0.052	0.794	0.304
Wdbc	AdaBoost	**0.350**	**0.581**	**0.035**
	arc-gv	0.350	0.710	**0.035**

Summary

- Adaboost
  - Combine weak learners => excellent classifier
  - Does not seem to overfit

- Initial Margin bounds
  - based on margin distribution/min margin
  - make intuitive sense, but inconsistent with empirical results

- New Margin Bounds
  - complicated, but consistent with experiments

Initial Margin Bound

THEOREM 1. *Let $\mathscr{D}$ be a distribution over $X \times \{-1, 1\}$, and let S be a sample of m examples chosen independently at random according to $\mathscr{D}$. Assume that the base-classifier space $\mathscr{H}$ is finite, and let $\delta > 0$. Then with probability at least $1 - \delta$ over the random choice of the training set S, every weighted average function $f \in \mathscr{C}$ satisfies the following bound for all $\theta > 0$:*

$$\mathbf{P}_{\mathscr{D}}[yf(x) \leq 0] \leq \mathbf{P}_S[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}}\left(\frac{\log m \log|\mathscr{H}|}{\theta^2} + \log(1/\delta)\right)^{1/2}\right).$$

Brief Proof Sketch (finite)

- $\mathscr{C}$: set of majority vote classifiers (MVC)
- Uniform MVC: $\mathscr{C}_N \doteq \left\{ f\colon\, x \mapsto \frac{1}{N} \sum_{i=1}^{N} h_i(x) \,\middle|\, h_i \in \mathscr{H} \right\}$
- $f \in \mathscr{C}$ induces distribution over $\mathscr{H}$
  – select $h \in \mathscr{H}$ from distribution to get $g \in \mathscr{C}_N$
  – $\mathscr{Q}$ is distribution over $\mathscr{C}_N$ induced by $f \in \mathscr{C}$

Brief Proof Sketch (finite)

- $\mathscr{C}$: set of majority vote classifiers (MVC)
- Uniform MVC: $\mathscr{C}_N \doteq \left\{ f\colon\ x \mapsto \frac{1}{N} \sum_{i=1}^{N} h_i(x) \,\middle|\, h_i \in \mathscr{H} \right\}$
- $f \in \mathscr{C}$ induces distribution over $\mathscr{H}$
  - select $h \in \mathscr{H}$ from distribution to get $g \in \mathscr{C}_N$
  - $\mathscr{Q}$ is distribution over $\mathscr{C}_N$ induced by $f \in \mathscr{C}$
- $\mathbf{P}[A] = \mathbf{P}[B \cap A] + \mathbf{P}[\overline{B} \cap A] \leq \mathbf{P}[B] + \mathbf{P}[\overline{B} \cap A]$

$$\mathbf{P}_{\mathscr{D}}[yf(x) \leq 0] \leq \mathbf{P}_{\mathscr{D}}[yg(x) \leq \theta/2] + \mathbf{P}_{\mathscr{D}}[yg(x) > \theta/2, yf(x) \leq 0]$$

Brief Proof Sketch (finite)

- $\mathbf{P}_{\mathscr{D}}[\, yf(x) \le 0] \le \mathbf{P}_{\mathscr{D}}[\, yg(x) \le \theta/2]$
  $+ \mathbf{P}_{\mathscr{D}}[\, yg(x) > \theta/2,\, yf(x) \le 0]$
- This holds for all $g \in \mathscr{C}_N$ so we can take expectation and bound each term separately
- $\mathbf{P}_{\mathscr{D},\, g \sim \mathscr{Q}}[\, yg(x) \le \theta/2]$
  - bound difference of empirical and test error for particular g and θ (Chernoff bound)
  - union bound over (finite!)
  - bound $\mathbf{P}_{S,\, g \sim \mathscr{Q}}[\, yf(x) \le \theta]$ by $\mathbf{P}_S[\, yf(x) \le \theta]$

Brief Proof Sketch (finite)

- $$\mathbf{P}_{\mathscr{D}}[yf(x) \le 0] \le \mathbf{P}_{\mathscr{D}}[yg(x) \le \theta/2] + \mathbf{P}_{\mathscr{D}}[yg(x) > \theta/2, yf(x) \le 0]$$
- This holds for all $g \in \mathscr{C}_N$ so we can take expectation and bound each term separately
- $\mathbf{P}_{\mathscr{D}, g \sim \mathscr{Q}}[yg(x) > \theta/2, yf(x) \le 0]$
  - Use fact that $f(x) = \mathbf{E}_{g \sim \mathscr{Q}}[g(x)]$ and apply Chernoff bound