Totally Corrective Boosting Algorithms that Maximize the Margin

Manfred K. Warmuth[1]
Jun Liao[1]
Gunnar Rätsch[2]

[1]University of California, Santa Cruz
[2]Friedrich Miescher Laboratory, Tübingen, Germany

Last update: March 2, 2007

1. Basics
2. Large margins
3. Games
4. Convergence
5. Illustrative Experiments

Outline

1. Basics
2. Large margins
3. Games
4. Convergence
5. Illustrative Experiments

Protocol of Boosting

- Maintain a distribution $\mathbf{d}^t$ on the examples
- At iteration $t = 1, \ldots, T$:
  1. Receive a "weak" hypothesis h_t
  2. Update $\mathbf{d}^t$ to $\mathbf{d}^{t+1}$, put more weights on "hard" examples
- Output a convex combination of the weak hypotheses

$$f_{\alpha}(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})$$

[Freund & Schapire, 1995]

Protocol of Boosting

- Maintain a distribution $\mathbf{d}^t$ on the examples
- At iteration $t = 1, \ldots, T$:
  1. Receive a “weak” hypothesis h_t
  2. Update $\mathbf{d}^t$ to $\mathbf{d}^{t+1}$, put more weights on “hard” examples
- Output a convex combination of the weak hypotheses

$$f_\alpha(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})$$

[Freund & Schapire, 1995]

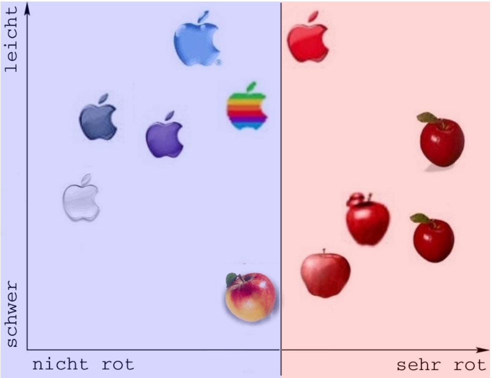
Protocol of Boosting

- Maintain a distribution $\mathbf{d}^t$ on the examples
- At iteration $t = 1, \ldots, T$:
  1. Receive a "weak" hypothesis h_t
  2. Update $\mathbf{d}^t$ to $\mathbf{d}^{t+1}$, put more weights on "hard" examples
- Output a convex combination of the weak hypotheses

$$f_{\boldsymbol{\alpha}}(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})$$

[Freund & Schapire, 1995]

Boosting: 1st Iteration



First hypothesis:

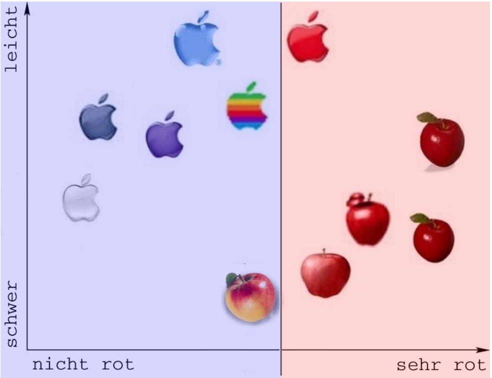
- Error rate: $\frac{2}{11}$

$$\epsilon_t = \sum_{n=1}^{N} d_n^t \mathbf{I}(h_t(\mathbf{x}_n) = y_n)$$

- Edge: $\frac{9}{22}$

$$\gamma_t = \sum_{n=1}^{N} d_n^t y_n h_t(\mathbf{x}_n)$$
$$= 1 - 2\epsilon_t$$

Boosting: 1st Iteration



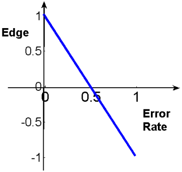
First hypothesis:

- Error rate: $\frac{2}{11}$

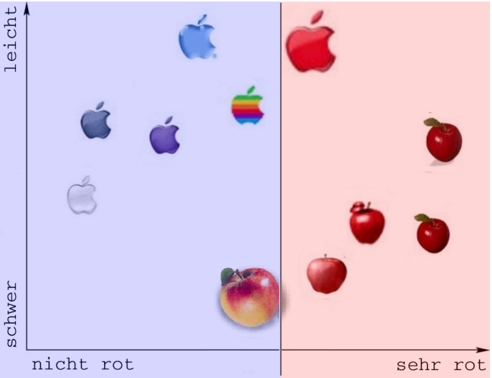
$$\epsilon_t = \sum_{n=1}^{N} d_n^t \mathbf{I}(h_t(\mathbf{x}_n) = y_n)$$

- Edge: $\frac{9}{22}$

$$\gamma_t = \sum_{n=1}^{N} d_n^t y_n h_t(\mathbf{x}_n)$$
$$= 1 - 2\epsilon_t$$



Update Distribution

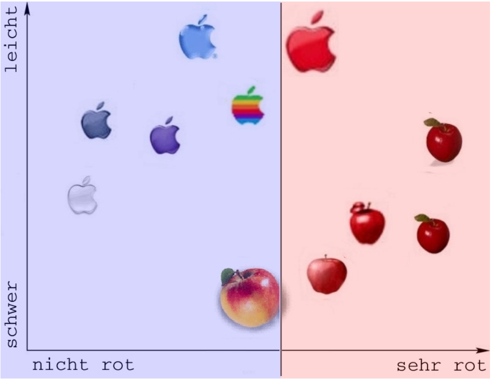


Misclassified examples
⇒ Increased weights

After update:

- Error rate:
  $\epsilon(h_t, \mathbf{d}^{t+1}) = \frac{1}{2}$
- Edge:
  $\gamma(h_t, \mathbf{d}^{t+1}) = 0$

Update Distribution



Misclassified examples
⇒ Increased weights

After update:

- Error rate: $\epsilon(h_t, \mathbf{d}^{t+1}) = \frac{1}{2}$
- Edge: $\gamma(h_t, \mathbf{d}^{t+1}) = 0$

Ada-Boost as Entropy Projection

Minimize relative entropy to last distribution subject to constraint

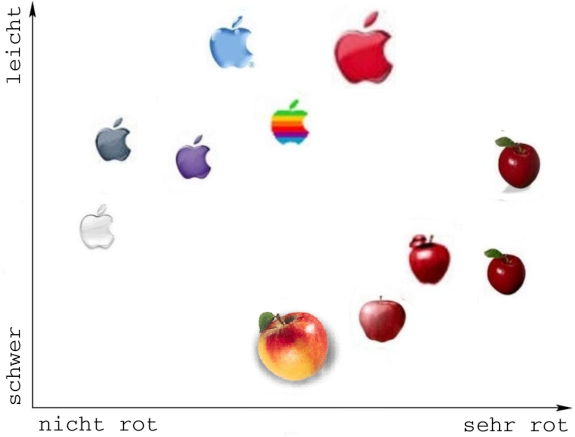
$$\begin{aligned} \min_{\mathbf{d}} \quad & \Delta(\mathbf{d}, \mathbf{d}^t) \\ \text{s.t.} \quad & \sum_{n=1}^{N} d_n y_n h_t(\mathbf{x}_n) = 0 \\ & \mathbf{d} \in \mathcal{P}^N \end{aligned}$$

where

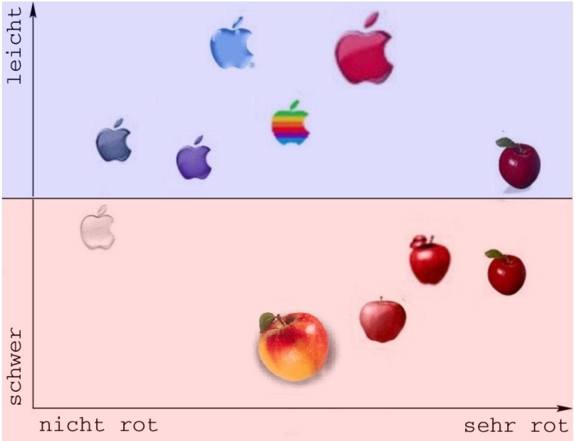
- $\Delta(\mathbf{d}, \mathbf{d}^t) = \sum_{n=1}^{N} d_n \ln \frac{d_n}{d_n^t}$ and
- $\mathcal{P}^N$ is the N dimensional probability simplex

[Lafferty, 1999; Kivinen & Warmuth, 1999]

Before 2nd Iteration

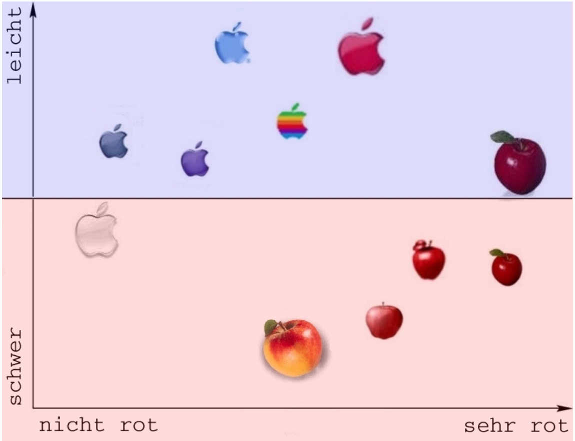


Boosting: 2nd Hypothesis



AdaBoost assumption:
Edge $\gamma > \nu$

Update Distribution



Edge $\gamma = 0$

AdaBoost update sets edge of last hypothesis to 0

Number of iterations:

$$\leq \frac{2 \ln N}{\nu^2}$$

Which constraints?

Corrective - Ada-Boost: Single constraint

$$\min_{\mathbf{d}\in\mathcal{P}^N} \quad \Delta(\mathbf{d}, \mathbf{d}^t)$$

$$\text{s.t.} \quad \sum_{n=1}^{N} d_n y_n h_t(\mathbf{x}_n) = 0 \qquad \text{for } n = 1, \ldots, t$$

Totally corrective: One constraint per past weak hypothesis

$$\min_{\mathbf{d}\in\mathcal{P}^N} \quad \Delta(\mathbf{d}, \mathbf{d}^1)$$

$$\text{s.t.} \quad \sum_{n=1}^{N} d_n y_n h_q(\mathbf{x}_n) = 0 \qquad \text{for } q = 1, \ldots, t$$

[Lafferty, 1999; Kivinen & Warmuth, 1999]

Which constraints?

Corrective - Ada-Boost: Single constraint

$$\min_{\mathbf{d}\in\mathcal{P}^N} \quad \Delta(\mathbf{d},\mathbf{d}^t)$$

$$\text{s.t.} \quad \sum_{n=1}^{N} d_n y_n h_{t}(\mathbf{x}_n) = 0 \qquad \text{for } n = 1, \dots, t$$

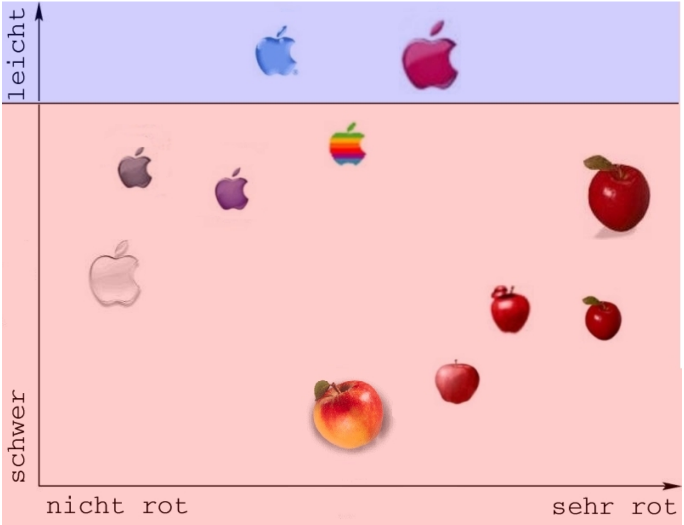
Totally corrective: One constraint per past weak hypothesis

$$\min_{\mathbf{d}\in\mathcal{P}^N} \quad \Delta(\mathbf{d},\mathbf{d}^1)$$

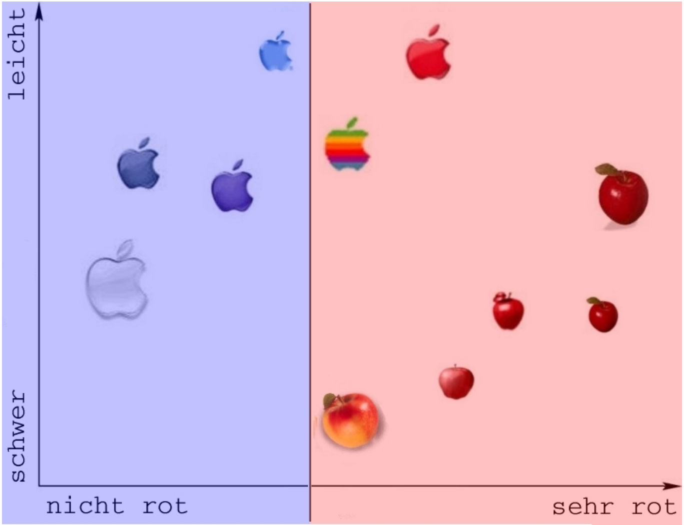
$$\text{s.t.} \quad \sum_{n=1}^{N} d_n y_n h_{q}(\mathbf{x}_n) = 0 \qquad \text{for } q = 1, \dots, t$$

[Lafferty, 1999; Kivinen & Warmuth, 1999]

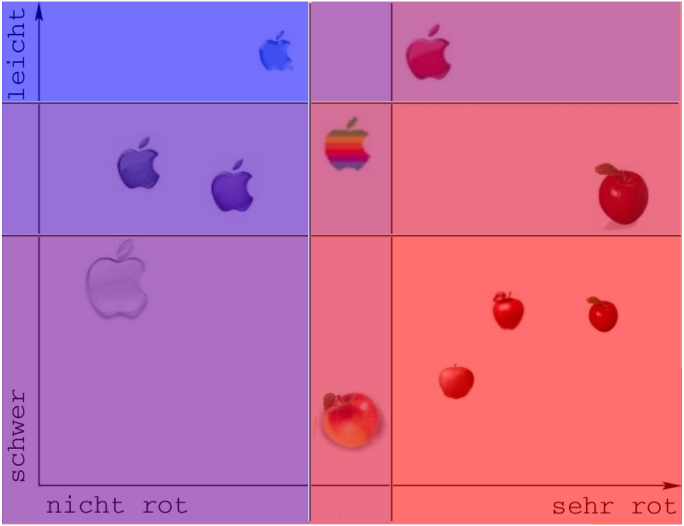
Boosting: 3nd Hypothesis



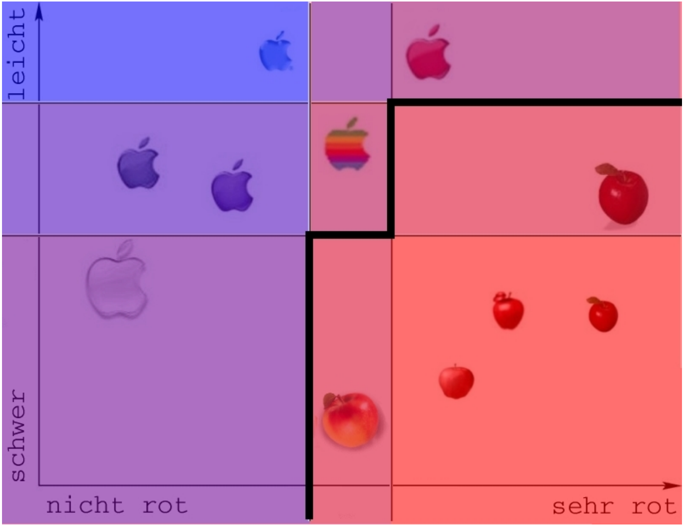
Boosting: 4th Hypothesis



All Hypotheses



Decision: $f_\alpha(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x}) > 0$?



Outline

1 Basics

2 Large margins

3 Games

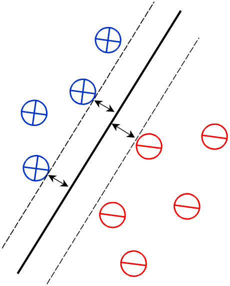
4 Convergence

5 Illustrative Experiments

Large margins in addition to correct classification

- Margin of the combined hypothesis $f_{\boldsymbol{\alpha}}$ for example $(\mathbf{x}_n, y_n)$

$$\begin{aligned}\rho_n(\boldsymbol{\alpha}) &= y_n f_{\boldsymbol{\alpha}}(\mathbf{x}_n)\\ &= y_n \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x}_n) \qquad (\boldsymbol{\alpha} \in \mathcal{P}^T)\end{aligned}$$

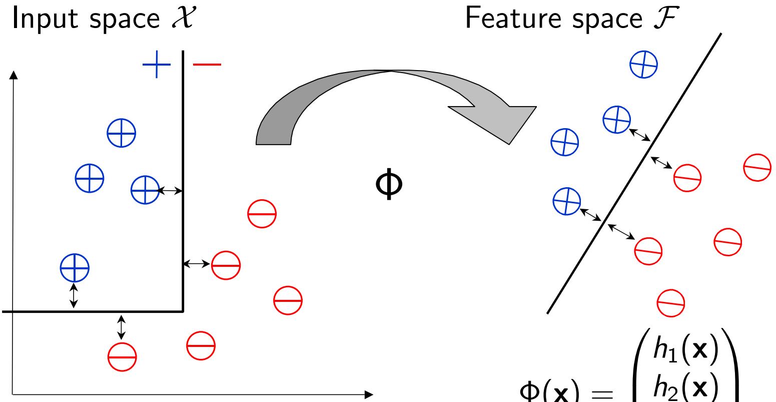


Margin of set of examples is minimum over examples

$$\rho(\boldsymbol{\alpha}) := \min_n \rho_n(\boldsymbol{\alpha})$$

[Freund, Schapire, Bartlett & Lee, 1998]

Large Margin and Linear Separation



$$\Phi(\mathbf{x}) = \begin{pmatrix} h_1(\mathbf{x}) \\ h_2(\mathbf{x}) \\ \vdots \end{pmatrix}$$

$$\mathcal{H} = \{h_1, h_2, \ldots\}$$

Linear separation in $\mathcal{F}$ is
nonlinear separation in $\mathcal{X}$

[Mangasarian, 1999; G.R., Mika, Schölkopf & Müller, 2002]

Margin vs. edge

Margin

- Measure for "confidence" in prediction for a hypothesis weighting
- Margin of example n for current hypothesis weighting $\boldsymbol{\alpha}$

$$\rho_n(\boldsymbol{\alpha}) \;=\; y_n f_{\boldsymbol{\alpha}}(\mathbf{x}_n) = y_n \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x}_n) \qquad \boldsymbol{\alpha} \in \mathcal{P}^T$$

Edge

- Measurement of "goodness" of a hypothesis w.r.t. a distribution
- Edge of a hypothesis h for a distribution $\mathbf{d}$ on the examples

$$\gamma_h(\mathbf{d}) = \sum_{n=1}^{N} d_n \; y_n h(\mathbf{x}_n) \qquad \mathbf{d} \in \mathcal{P}^N$$

What is the connection?

[Breiman, 1999]

Margin vs. edge

Margin

- Measure for “confidence” in prediction for a hypothesis weighting
- Margin of example *n* for current hypothesis weighting $\boldsymbol{\alpha}$

$$\rho_n(\boldsymbol{\alpha}) \;=\; y_n f_{\boldsymbol{\alpha}}(\mathbf{x}_n) = y_n \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x}_n) \qquad \boldsymbol{\alpha} \in \mathcal{P}^T$$

Edge

- Measurement of “goodness” of a hypothesis w.r.t. a distribution
- Edge of a hypothesis *h* for a distribution **d** on the examples

$$\gamma_h(\mathbf{d}) = \sum_{n=1}^{N} d_n \; y_n h(\mathbf{x}_n) \qquad \mathbf{d} \in \mathcal{P}^N$$

What is the connection?

[Breiman, 1999]

Margin vs. edge

Margin

- Measure for "confidence" in prediction for a hypothesis weighting
- Margin of example *n* for current hypothesis weighting $\boldsymbol{\alpha}$

$$\rho_n(\boldsymbol{\alpha}) \;=\; y_n f_{\boldsymbol{\alpha}}(\mathbf{x}_n) = y_n \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x}_n) \qquad \boldsymbol{\alpha} \in \mathcal{P}^T$$

Edge

- Measurement of "goodness" of a hypothesis w.r.t. a distribution
- Edge of a hypothesis *h* for a distribution **d** on the examples

$$\gamma_h(\mathbf{d}) = \sum_{n=1}^{N} d_n \; y_n h(\mathbf{x}_n) \qquad \mathbf{d} \in \mathcal{P}^N$$

What is the connection?

[Breiman, 1999]

von Neumann's Minimax-Theorem

Set of examples $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$
and hypotheses set $\mathcal{H}^t = \{h_1, \ldots, h_t\}$,

$$\text{minimum edge: } \gamma_t^* = \min_{\mathbf{d}\in\mathcal{P}^N} \underbrace{\max_{h\in\mathcal{H}^t} \overbrace{\gamma_h(\mathbf{d})}^{\text{edge of } h}}_{\text{edge of } \mathcal{H}^t}$$

$$\text{maximum margin: } \rho_t^* = \max_{\boldsymbol{\alpha}\in\mathcal{P}^t} \underbrace{\min_{n} \overbrace{y_n f_{\boldsymbol{\alpha}}(\mathbf{x}_n)}^{\text{margin of } (\mathbf{x}_n, y_n)}}_{\text{margin of S}}$$

$$\text{Duality: } \gamma_t^* = \rho_t^*$$

[von Neumann, 1928]

von Neumann's Minimax-Theorem

Set of examples $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$
and hypotheses set $\mathcal{H}^t = \{h_1, \ldots, h_t\}$,

$$\text{minimum edge: } \gamma_t^* = \min_{\mathbf{d} \in \mathcal{P}^N} \underbrace{\max_{q=1}^{t} \overbrace{\sum_{n=1}^{N} d_n y_n h_q(\mathbf{x}_n)}^{\text{edge of } h_q}}_{\text{edge of } \mathcal{H}^t}$$

$$\text{maximum margin: } \rho_t^* = \max_{\boldsymbol{\alpha} \in \mathcal{P}^t} \underbrace{\min_{n} \overbrace{y_n \sum_{q=1}^{t} \alpha_t h_q(\mathbf{x}_n)}^{\text{margin of } (\mathbf{x}_n, y_n)}}_{\text{margin of S}}$$

$$\text{Duality: } \gamma_t^* = \rho_t^*$$

Outline

1. Basics
2. Large margins
3. Games
4. Convergence
5. Illustrative Experiments

Two-player Zero Sum Game

Rock, Paper, Scissors game

column player

$$\begin{array}{llccc} & & \text{R} & \text{P} & \text{S} \\ & & \alpha_1 & \alpha_2 & \alpha_3 \\ & \text{R} \quad d_1 & 0 & 1 & -1 \\ \text{row player} & \text{P} \quad d_2 & -1 & 0 & 1 \\ & \text{S} \quad d_3 & 1 & -1 & 0 \end{array}$$

gain matrix

Single row is pure strategy of row player and $\mathbf{d}$ is mixed strategy

Single column is pure strategy of column player and $\boldsymbol{\alpha}$ is mixed strategy

Row player minimizes
Column player maximizes

$$\begin{aligned} \text{payoff} &= \mathbf{d}^T M \boldsymbol{\alpha} \\ &= \textstyle\sum_{i,j} d_i M_{i,j} \alpha_j \end{aligned}$$

[Freund and Schapire, 1997]

Optimum Strategy

$$\begin{array}{ccc|ccc}
 & & & R & P & S \\
 & & & \alpha_1 & \alpha_2 & \alpha_3 \\
 & & & .33 & .33 & .33 \\
R & d_1 & .33 & 0 & 1 & -1 \\
P & d_2 & .33 & -1 & 0 & 1 \\
S & d_3 & .33 & 1 & -1 & 0
\end{array}$$

- Min-max theorem:

$$\begin{aligned}
&\min_{d} \max_{\boldsymbol{\alpha}} \mathbf{d}^T M \boldsymbol{\alpha} = \min_{d} \max_{j} \mathbf{d}^T \; M e_j \\
= \quad & \max_{\boldsymbol{\alpha}} \min_{d} \; \mathbf{d}^T M \boldsymbol{\alpha} = \max_{\boldsymbol{\alpha}} \min_{i} \; e_i M \boldsymbol{\alpha} \\
= \quad & \text{value of the game (0 in example)}
\end{aligned}$$

e_i is pure strategy

Connection to Boosting?

- Rows are the examples
- Columns the weak hypothesis
- $M_{i,j} = h_j(\mathbf{x}_i)y_i$
- Row sum: margin of example
- Column sum: edge of weak hypothesis
- Value of game: $\gamma^* = \rho^*$

New column added: boosting

			R	P	S		
			α_1	α_2	α_3	α_4	margin
			.44	0	.22	.33	
R	d_1	.22	0	1	-1	1	.11
P	d_2	.33	-1	0	1	1	.11
S	d_3	.44	1	-1	0	-1	.11
	edge		.11	-.22	.11	.11	

Value of game **increases** from 0 to .11

Row added: on-line learning

			R	P	S	
			α_1	α_2	α_3	margin
			.33	.44	.22	
R	d_1	0	0	1	-1	.22
P	d_2	.22	-1	0	1	-.11
S	d_3	.44	1	-1	0	-.11
	d_4	.33	-1	1	-1	-.11
	edge		-.11	-.11	-.11	

Value of game **decreases** from 0 to -.11

Boosting: maximize margin incrementally

$$\begin{array}{c|c} & \alpha_1^1 \\ d_1^1 & 0 \\ d_2^1 & 1 \\ d_3^1 & -1 \end{array}$$

iteration 1

$$\begin{array}{c|cc} & \alpha_1^2 & \alpha_2^2 \\ d_1^2 & 0 & -1 \\ d_2^2 & 1 & 0 \\ d_3^2 & -1 & 1 \end{array}$$

iteration 2

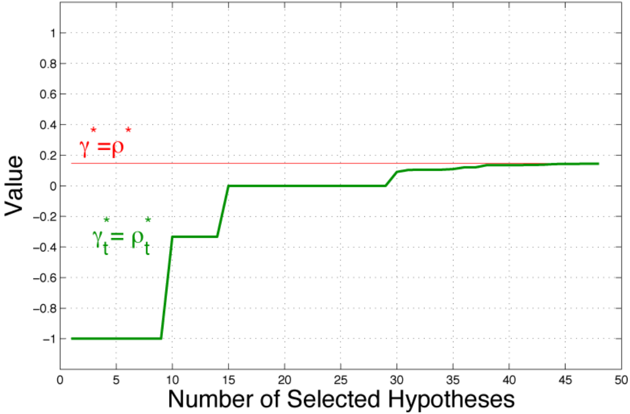
$$\begin{array}{c|ccc} & \alpha_1^3 & \alpha_2^3 & \alpha_3^3 \\ d_1^3 & 0 & -1 & 1 \\ d_2^3 & 1 & 0 & -1 \\ d_3^3 & -1 & 1 & 0 \end{array}$$

iteration 3

- In each iteration solve optimization problem to update d
- Column player adds new column - weak hypothesis
- Some assumptions will be needed on the edge of the added hypothesis

Value non-decreasing

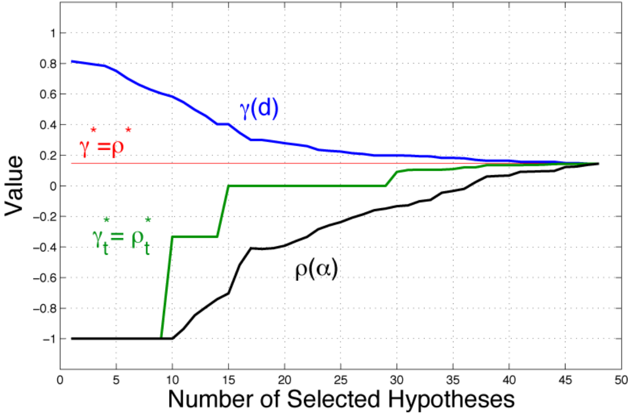
γ^*, ρ^*: edge/margin for all hypotheses



Duality gap

For any non-optimal $\mathbf{d} \in \mathcal{P}^N$ and $\boldsymbol{\alpha} \in \mathcal{P}^t$,

$$\gamma(\mathbf{d}) \geq \gamma_t^* = \rho_t^* \geq \rho(\boldsymbol{\alpha})$$



How Large is the Maximal Margin?

Assumptions on Weak learner

For any distribution $\mathbf{d}$ on the examples, the weak learner returns a hypothesis h with edge $\gamma_h(\mathbf{d})$ at least g.
Best case: $g = \rho^* = \gamma^*$

[Breiman, 1999; Bennett et al.; G.R. et al., 2001; Rudin et al., 2004]

How Large is the Maximal Margin?

Assumptions on Weak learner

For any distribution $\mathbf{d}$ on the examples, the weak learner returns a hypothesis h with edge $\gamma_h(\mathbf{d})$ at least g.
Best case: $g = \rho^* = \gamma^*$

Implication from Minimax Theorem

There exists $\boldsymbol{\alpha} \in \mathcal{P}^N$, such that $\rho(\boldsymbol{\alpha}) \geq g$

[Breiman, 1999; Bennett et al.; G.R. et al., 2001; Rudin et al., 2004]

How Large is the Maximal Margin?

Assumptions on Weak learner

For any distribution $\mathbf{d}$ on the examples, the weak learner returns a hypothesis h with edge $\gamma_h(\mathbf{d})$ at least g.
Best case: $g = \rho^* = \gamma^*$

Implication from Minimax Theorem

There exists $\boldsymbol{\alpha} \in \mathcal{P}^N$, such that $\rho(\boldsymbol{\alpha}) \geq g$

Idea to iteratively solve LP: LPBoost

Add "best" hypothesis $h = \operatorname{argmax} \gamma_h(\mathbf{d}^t)$ to $\mathcal{H}^t$ and resolve

$$\mathbf{d}^{t+1} = \operatorname*{argmin}_{d \in \mathcal{P}^N} \max_{h \in \mathcal{H}^t} \gamma_h(\mathbf{d})$$

[Breiman, 1999; Bennett et al.; G.R. et al., 2001; Rudin et al., 2004]

Convergence?

LPBoost?

- No iteration bounds known

AdaBoost?

- May “oscillate”
- Does not find maximizing $\boldsymbol{\alpha}$ (counter examples)
- But there some guarantees:
  - $\rho(\boldsymbol{\alpha}^t) \geq 0$ after $2 \ln N/g^2$ iterations
  - $\rho(\boldsymbol{\alpha}^t) \geq g/2$ in the limit

[Breiman, 1999; Bennett et al.; G.R. et al., 2001; Rudin et al., 2004]

How to maximize the margin?

Modify AdaBoost for maximizing margin

- Arc-GV asymptotically maximizes the margin
  - quite slow, no convergence rates
- LPBoost uses a Linear Programming solver
  - Often very fast in practice, but no converge rates
- AdaBoost$^*_\nu$ requires $\frac{2\log(N)}{\nu^2}$ iterations to get $\rho^t \in [\rho^* - \nu, \rho^*]$
  - Slow in practice, i.e. not faster than theory predicts
- TotalBoost$_\nu$ requires $\frac{2\log(N)}{\nu^2}$ iterations to get $\rho^t \in [\rho^* - \nu, \rho^*]$
  - Fast in practice
  - Combination of benefits

[G.R. & Warmuth, 2004; Warmuth et al., 2006]

How to maximize the margin?

Modify AdaBoost for maximizing margin

- Arc-GV asymptotically maximizes the margin
  - quite slow, no convergence rates
- LPBoost uses a Linear Programming solver
  - Often very fast in practice, but no converge rates
- AdaBoost$^*_\nu$ requires $\frac{2\log(N)}{\nu^2}$ iterations to get $\rho^t \in [\rho^* - \nu, \rho^*]$
  - Slow in practice, i.e. not faster than theory predicts
- TotalBoost$_\nu$ requires $\frac{2\log(N)}{\nu^2}$ iterations to get $\rho^t \in [\rho^* - \nu, \rho^*]$
  - Fast in practice
  - Combination of benefits

[G.R. & Warmuth, 2004; Warmuth et al., 2006]

How to maximize the margin?

Modify AdaBoost for maximizing margin

- Arc-GV asymptotically maximizes the margin
  - quite slow, no convergence rates
- LPBoost uses a Linear Programming solver
  - Often very fast in practice, but no converge rates
- AdaBoost$^*_\nu$ requires $\frac{2\log(N)}{\nu^2}$ iterations to get $\rho^t \in [\rho^* - \nu, \rho^*]$
  - Slow in practice, i.e. not faster than theory predicts
- TotalBoost$_\nu$ requires $\frac{2\log(N)}{\nu^2}$ iterations to get $\rho^t \in [\rho^* - \nu, \rho^*]$
  - Fast in practice
  - Combination of benefits

[G.R. & Warmuth, 2004; Warmuth et al., 2006]

How to maximize the margin?

Modify AdaBoost for maximizing margin

- Arc-GV asymptotically maximizes the margin
    - quite slow, no convergence rates
- LPBoost uses a Linear Programming solver
    - Often very fast in practice, but no converge rates
- AdaBoost$^*_\nu$ requires $\frac{2\log(N)}{\nu^2}$ iterations to get $\rho^t \in [\rho^* - \nu, \rho^*]$
    - Slow in practice, i.e. not faster than theory predicts
- TotalBoost$_\nu$ requires $\frac{2\log(N)}{\nu^2}$ iterations to get $\rho^t \in [\rho^* - \nu, \rho^*]$
    - Fast in practice
    - Combination of benefits

[G.R. & Warmuth, 2004; Warmuth et al., 2006]

Outline

1 Basics

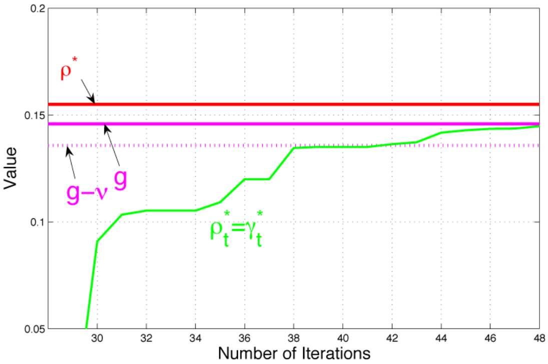
2 Large margins

3 Games

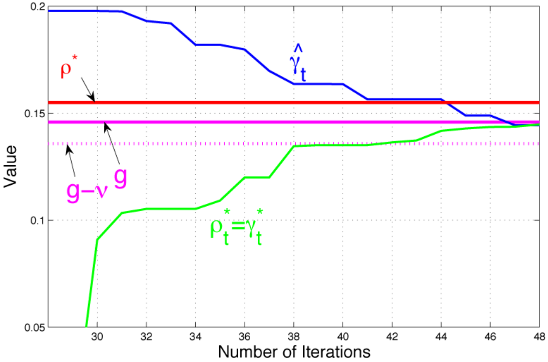
4 Convergence

5 Illustrative Experiments

Want margin $\geq g - \nu$



Want margin $\geq g - \nu$



- Assumption: $\gamma_t \geq g$
- Estimate of target: $\widehat{\gamma}_t = (\min_{q=1}^{t} \gamma_q) - \nu$

Idea: Projections to $\hat{\gamma}_t$ instead of 0

Corrective: Single constraint

$$\min_{\mathbf{d} \in \mathcal{P}^N} \quad \Delta(\mathbf{d}, \mathbf{d}^t)$$

AdaBoost$^*_\nu$

$$\text{s.t.} \quad \sum_{n=1}^{N} d_n y_n h_t(\mathbf{x}_n) \leq \hat{\gamma}_t \qquad \text{for } r = 1, \ldots, t$$

Totally corrective: One constraint per past weak hypothesis

$$\min_{\mathbf{d} \in \mathcal{P}^N} \quad \Delta(\mathbf{d}, \mathbf{d}^1)$$

TotalBoost$_\nu$

$$\text{s.t.} \quad \sum_{n=1}^{N} d_n y_n h_q(\mathbf{x}_n) \leq \hat{\gamma}_t \qquad \text{for } q = 1, \ldots, t$$

TotalBoost$_\nu$

1. **Input:** $S = \langle(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\rangle$, desired accuracy ν
2. **Initialize:** $d_n^1 = 1/N$ for all $n = 1 \ldots N$
3. **Do for** $t = 1, \ldots$
    1. Train classifier on $\{S, \mathbf{d}^t\}$ and obtain hypothesis $h_t : \mathbf{x} \mapsto [-1, 1]$ and let $u_i^t = y_i h_t(\mathbf{x}_i)$
    2. Calculate the edge γ_t of h_t: $\gamma_t = \mathbf{d}^t \cdot \mathbf{u}^t$
    3. Set $\widehat{\gamma}_t = (\min_{q=1}^{t} \gamma_q) - \nu$ and solve
       $$\mathbf{d}^{t+1} = \operatorname*{argmin}_{\{\mathbf{d} \in \mathcal{P}^N \mid \mathbf{d} \cdot \mathbf{u}^q \le \widehat{\gamma}_t, \text{ for } 1 \le q \le t\} = \mathcal{C}_t} \Delta(\mathbf{d}, \mathbf{d}^1)$$
    4. **If** above infeasible or $\mathbf{d}^{t+1}$ contains a zero then $T = t$ and break
4. **Output:** $f_\alpha(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})$, where the coefficients α_t maximize margin over hypotheses set $\{h_1, \ldots, h_T\}$.

TotalBoost$_\nu$

1. **Input:** $S = \langle(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\rangle$, desired accuracy ν
2. **Initialize:** $d_n^1 = 1/N$ for all $n = 1 \ldots N$
3. **Do for** $t = 1, \ldots$
   1. Train classifier on $\{S, \mathbf{d}^t\}$ and obtain hypothesis $h_t : \mathbf{x} \mapsto [-1, 1]$ and let $u_i^t = y_i h_t(\mathbf{x}_i)$
   2. Calculate the edge γ_t of h_t: $\gamma_t = \mathbf{d}^t \cdot \mathbf{u}^t$
   3. Set $\widehat{\gamma}_t = (\min_{q=1}^{t} \gamma_q) - \nu$ and solve

$$\mathbf{d}^{t+1} = \operatorname*{argmin}_{\{\mathbf{d} \in \mathcal{P}^N \mid \mathbf{d} \cdot \mathbf{u}^q \leq \widehat{\gamma}_t, \text{ for } 1 \leq q \leq t\} = \mathcal{C}_t} \Delta(\mathbf{d}, \mathbf{d}^1)$$

   4. If above infeasible or $\mathbf{d}^{t+1}$ contains a zero then $T = t$ and break
4. **Output:** $f_\alpha(\mathbf{x}) = \sum_{t=1}^T \alpha_t h_t(\mathbf{x})$, where the coefficients α_t maximize margin over hypotheses set $\{h_1, \ldots, h_T\}$.

TotalBoost$_\nu$

1. **Input:** $S = \langle(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\rangle$, desired accuracy ν
2. **Initialize:** $d^1_n = 1/N$ for all $n = 1 \ldots N$
3. **Do for** $t = 1, \ldots$
    1. Train classifier on $\{S, \mathbf{d}^t\}$ and obtain hypothesis $h_t : \mathbf{x} \mapsto [-1, 1]$ and let $u^t_i = y_i h_t(\mathbf{x}_i)$
    2. Calculate the edge γ_t of h_t: $\gamma_t = \mathbf{d}^t \cdot \mathbf{u}^t$
    3. Set $\widehat{\gamma}_t = (\min\limits_{q=1}^{t} \gamma_q) - \nu$ and solve
    4. **Optimization Problem**

       $$\begin{aligned} \mathbf{d}^{t+1} &= \operatorname*{argmin}_{\mathbf{d} \in \mathcal{C}_t} \Delta(\mathbf{d}, \mathbf{d}^1) \\ \text{with} \quad & \mathcal{C}_t := \{\mathbf{d} \in \mathcal{P}^N \mid \mathbf{d} \cdot \mathbf{u}^q \leq \widehat{\gamma}_t, \text{ for } 1 \leq q \leq t\} \end{aligned}$$
4. **Output:** $f_\alpha(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})$, where the coefficients α_t maximize margin over hypotheses set $\{h_1, \ldots, h_T\}$.

Effect of entropy regularization

- High weight given to hard examples

$$d_n^{t+1} \sim exp(- \quad \underbrace{y_n f_{\boldsymbol{\alpha}^{t+1}}(\mathbf{x}_n)}_{\text{margin of ex. } n} \quad)$$

- Softmin of margins
- α_{t+1} are current coefficients of hypotheses

Lower Bounding the Progress

Theorem

For $\mathbf{d}^t, \mathbf{d}^{t+1} \in \mathcal{P}^N$ *and* $\mathbf{u}^t \in [-1,1]^N$,
if $\Delta(\mathbf{d}^{t+1}, \mathbf{d}^t)$ *finite and* $\mathbf{d}^{t+1} \cdot \mathbf{u}^t \neq \mathbf{d}^t \cdot \mathbf{u}^t$ *then*

$$\Delta(\mathbf{d}^{t+1}, \mathbf{d}^t) > \frac{(\mathbf{d}^{t+1} \cdot \mathbf{u}^t - \mathbf{d}^t \cdot \mathbf{u}^t)^2}{2}$$

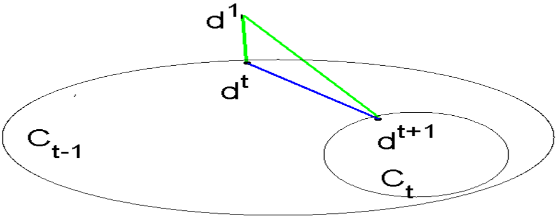
- $\mathbf{d}^t \cdot \mathbf{u}^t = \gamma_t$
- $\mathbf{d}^{t+1} \cdot \mathbf{u}^t \leq \widehat{\gamma}_t \leq \gamma_t - \nu$
- Thus $\mathbf{d}^t \cdot \mathbf{u}^t - \mathbf{d}^{t+1} \cdot \mathbf{u}^t \geq \nu$ and $\Delta(\mathbf{d}^{t+1}, \mathbf{d}^t) > \frac{\nu^2}{2}$

Generalized Pythagorean Theorem

$\mathcal{C}_t = \{\mathbf{d} \in \mathcal{P}^N | \mathbf{d} \cdot \mathbf{u}^q \leq \widehat{\gamma}_t, 1 \leq q \leq t\}, \mathcal{C}_0 = \mathcal{P}^N, \mathcal{C}_t \subseteq \mathcal{C}_{t-1}$
$\mathbf{d}^t$ is projection of $\mathbf{d}^1$ onto $\mathcal{C}_{t-1}$ at iteration $t-1$

$$\mathbf{d}^t = \operatorname*{argmin}_{\mathbf{d} \in \mathcal{C}_{t-1}} \Delta(\mathbf{d}, \mathbf{d}^1)$$

$$\Delta(\mathbf{d}^{t+1}, \mathbf{d}^1) \;\geq\; \Delta(\mathbf{d}^t, \mathbf{d}^1) + \Delta(\mathbf{d}^{t+1}, \mathbf{d}^t)$$



[Herbster & Warmuth, 2001]

Sketch of Proof

$$\begin{array}{rcccc}
1: & \Delta(\mathbf{d}^2, \mathbf{d}^1) - \Delta(\mathbf{d}^1, \mathbf{d}^1) & \geq & \Delta(\mathbf{d}^2, \mathbf{d}^1) & > & \frac{\nu^2}{2} \\
2: & \Delta(\mathbf{d}^3, \mathbf{d}^1) - \Delta(\mathbf{d}^2, \mathbf{d}^1) & \geq & \Delta(\mathbf{d}^3, \mathbf{d}^2) & > & \frac{\nu^2}{2} \\
 & \cdots & & \cdots & & \\
t: & \Delta(\mathbf{d}^{t+1}, \mathbf{d}^1) - \Delta(\mathbf{d}^t, \mathbf{d}^1) & \geq & \Delta(\mathbf{d}^{t+1}, \mathbf{d}^t) & > & \frac{\nu^2}{2} \\
 & \cdots & & \cdots & & \\
T-2: & \Delta(\mathbf{d}^{T-1}, \mathbf{d}^1) - \Delta(\mathbf{d}^{T-2}, \mathbf{d}^1) & \geq & \Delta(\mathbf{d}^{T-1}, \mathbf{d}^{T-2}) & > & \frac{\nu^2}{2} \\
T-1: & \Delta(\mathbf{d}^T, \mathbf{d}^1) - \Delta(\mathbf{d}^{T-1}, \mathbf{d}^1) & \geq & \Delta(\mathbf{d}^T, \mathbf{d}^{T-1}) & > & \frac{\nu^2}{2}
\end{array}$$

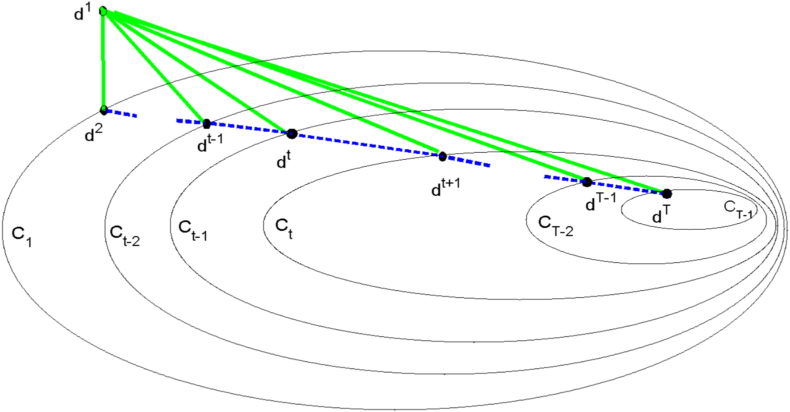
[Warmuth, Liao & G.R., 2006]

Cancellation

$$\begin{array}{rlcccc}
1: & \cancel{\Delta(\mathbf{d}^2,\mathbf{d}^1)} - \overbrace{\Delta(\mathbf{d}^1,\mathbf{d}^1)}^{0} & \geq & \Delta(\mathbf{d}^2,\mathbf{d}^1) & > & \frac{\nu^2}{2} \\
2: & \cancel{\Delta(\mathbf{d}^3,\mathbf{d}^1)} - \cancel{\Delta(\mathbf{d}^2,\mathbf{d}^1)} & \geq & \Delta(\mathbf{d}^3,\mathbf{d}^2) & > & \frac{\nu^2}{2} \\
 & \cdots & & \cdots & & \\
t: & \cancel{\Delta(\mathbf{d}^{t+1},\mathbf{d}^1)} - \cancel{\Delta(\mathbf{d}^t,\mathbf{d}^1)} & \geq & \Delta(\mathbf{d}^{t+1},\mathbf{d}^t) & > & \frac{\nu^2}{2} \\
 & \cdots & & \cdots & & \\
T-2: & \cancel{\Delta(\mathbf{d}^{T-1},\mathbf{d}^1)} - \cancel{\Delta(\mathbf{d}^{T-2},\mathbf{d}^1)} & \geq & \Delta(\mathbf{d}^{T-1},\mathbf{d}^{T-2}) & > & \frac{\nu^2}{2} \\
T-1: & \underbrace{\Delta(\mathbf{d}^T,\mathbf{d}^1)}_{\ln N} - \cancel{\Delta(\mathbf{d}^{T-1},\mathbf{d}^1)} & \geq & \Delta(\mathbf{d}^T,\mathbf{d}^{T-1}) & > & \frac{\nu^2}{2}
\end{array}$$

Therefore, $T \leq \lceil \frac{2 \ln N}{\nu^2} \rceil$

Overview



[Long & Wu, 2002]

Iteration Bounds for Other Variants

Using the same techniques, we can prove iteration bound for:

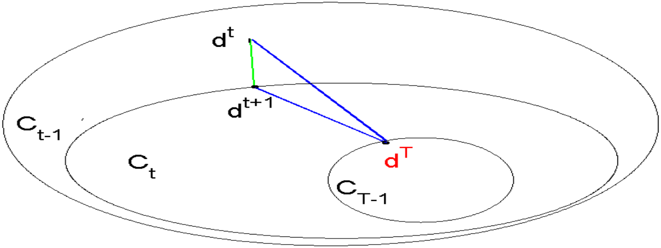
- TotalBoost$_\nu$ which optimizes the divergence $\Delta(\mathbf{d}, \mathbf{d}^t)$ to the last distribution $\mathbf{d}^t$
- TotalBoost$_\nu$ which uses the binary relative entropy $\Delta_2(\mathbf{d}, \mathbf{d}^1)$ or $\Delta_2(\mathbf{d}, \mathbf{d}^t)$ as the divergence
- The variant of AdaBoost$_\nu^*$ which terminates when $\widehat{\gamma}_t < \gamma_t^*$

TotalBoost$_\nu$ which optimizes $\Delta(\mathbf{d}, \mathbf{d}^t)$

$\mathbf{d}^{t+1}$ is projection of $\mathbf{d}^t$ onto $\mathcal{C}_t$ at iteration t

$$\mathbf{d}^{t+1} = \operatorname*{argmin}_{\mathbf{d}\in\mathcal{C}_t} \Delta(\mathbf{d}, \mathbf{d}^t)$$

$$\Delta(\mathbf{d}^T, \mathbf{d}^t) \;\geq\; \Delta(\mathbf{d}^T, \mathbf{d}^{t+1}) + \Delta(\mathbf{d}^{t+1}, \mathbf{d}^t)$$



Sketch of Proof

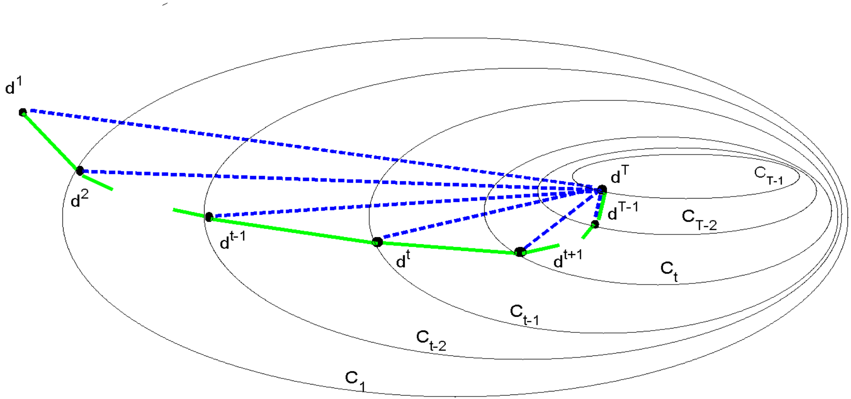
$$
\begin{array}{rcccccc}
1: & \Delta(\mathbf{d}^T, \mathbf{d}^1) - \Delta(\mathbf{d}^T, \mathbf{d}^2) & \geq & \Delta(\mathbf{d}^2, \mathbf{d}^1) & > & \frac{\nu^2}{2} \\
2: & \Delta(\mathbf{d}^T, \mathbf{d}^2) - \Delta(\mathbf{d}^T, \mathbf{d}^3) & \geq & \Delta(\mathbf{d}^3, \mathbf{d}^2) & > & \frac{\nu^2}{2} \\
 & \cdots & & \cdots & & \\
t: & \Delta(\mathbf{d}^T, \mathbf{d}^t) - \Delta(\mathbf{d}^T, \mathbf{d}^{t+1}) & \geq & \Delta(\mathbf{d}^{t+1}, \mathbf{d}^t) & > & \frac{\nu^2}{2} \\
 & \cdots & & \cdots & & \\
T-2: & \Delta(\mathbf{d}^T, \mathbf{d}^{T-2}) - \Delta(\mathbf{d}^T, \mathbf{d}^{T-1}) & \geq & \Delta(\mathbf{d}^{T-1}, \mathbf{d}^{T-2}) & > & \frac{\nu^2}{2} \\
T-1: & \Delta(\mathbf{d}^T, \mathbf{d}^{T-1}) - \Delta(\mathbf{d}^T, \mathbf{d}^T) & \geq & \Delta(\mathbf{d}^T, \mathbf{d}^{T-1}) & > & \frac{\nu^2}{2}
\end{array}
$$

Cancellation

$$\begin{array}{rcccccc}
1: & \overbrace{\Delta(\mathbf{d}^T,\mathbf{d}^1)}^{\ln N} - \cancel{\Delta(\mathbf{d}^T,\mathbf{d}^2)} & \geq & \Delta(\mathbf{d}^2,\mathbf{d}^1) & > & \frac{\nu^2}{2} \\
2: & \cancel{\Delta(\mathbf{d}^T,\mathbf{d}^2)} - \cancel{\Delta(\mathbf{d}^T,\mathbf{d}^3)} & \geq & \Delta(\mathbf{d}^3,\mathbf{d}^2) & > & \frac{\nu^2}{2} \\
 & \dots & & \dots & & \\
t: & \cancel{\Delta(\mathbf{d}^T,\mathbf{d}^t)} - \cancel{\Delta(\mathbf{d}^T,\mathbf{d}^{t+1})} & \geq & \Delta(\mathbf{d}^{t+1},\mathbf{d}^t) & > & \frac{\nu^2}{2} \\
 & \dots & & \dots & & \\
T-2: & \cancel{\Delta(\mathbf{d}^T,\mathbf{d}^{T-2})} - \cancel{\Delta(\mathbf{d}^T,\mathbf{d}^{T-1})} & \geq & \Delta(\mathbf{d}^{T-1},\mathbf{d}^{T-2}) & > & \frac{\nu^2}{2} \\
T-1: & \cancel{\Delta(\mathbf{d}^T,\mathbf{d}^{T-1})} - \underbrace{\Delta(\mathbf{d}^T,\mathbf{d}^T)}_{0} & \geq & \Delta(\mathbf{d}^T,\mathbf{d}^{T-1}) & > & \frac{\nu^2}{2}
\end{array}$$

Therefore, $T \leq \lceil \frac{2 \ln N}{\nu^2} \rceil$

TotalBoost$_\nu$ which optimizes $\Delta(\mathbf{d}, \mathbf{d}^t)$



Outline

1 Basics

2 Large margins

3 Games

4 Convergence

5 Illustrative Experiments

[Warmuth, Liao & G.R., 2006]

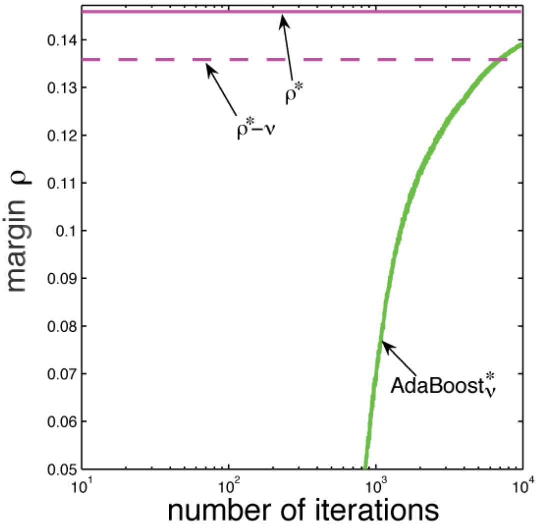
Illustrative Experiments

Cox-1 dataset from Telik Inc.

- Relatively small drug-design data set
    - 125 binary labeled examples
    - 3888 binary features
- Compare convergence of margin versus number of iterations

[Warmuth, Liao & G.R., 2006]

Illustrative Experiments



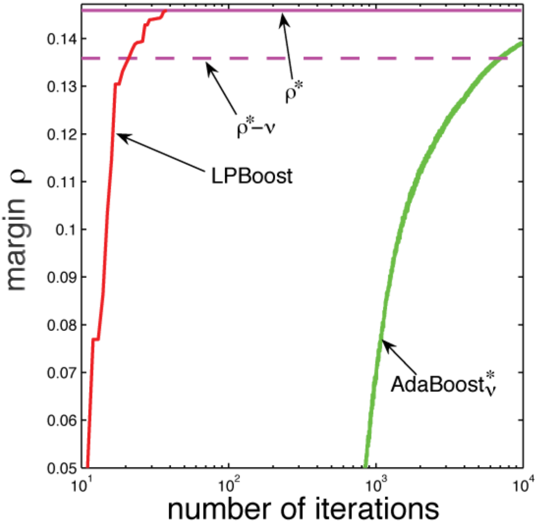
Cox-1 ($\nu = 0.01$)

- AdaBoost^*_ν
- LPBoost
- TotalBoost_ν

Results

- Corrective algorithms very slow
- LPBoost & TotalBoost_ν need few iterations
- Initial speed crucially depends on ν

Illustrative Experiments



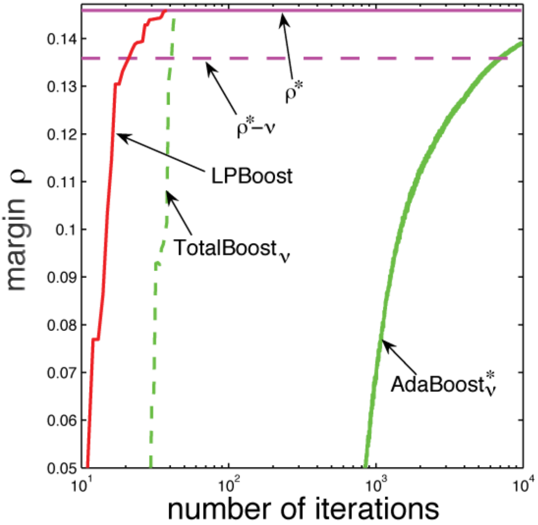
Cox-1 ($\nu = 0.01$)

- $\mathrm{AdaBoost}^*_\nu$
- LPBoost
- $\mathrm{TotalBoost}_\nu$

Results

- Corrective algorithms very slow
- LPBoost & $\mathrm{TotalBoost}_\nu$ need few iterations
- Initial speed crucially depends on ν

Illustrative Experiments



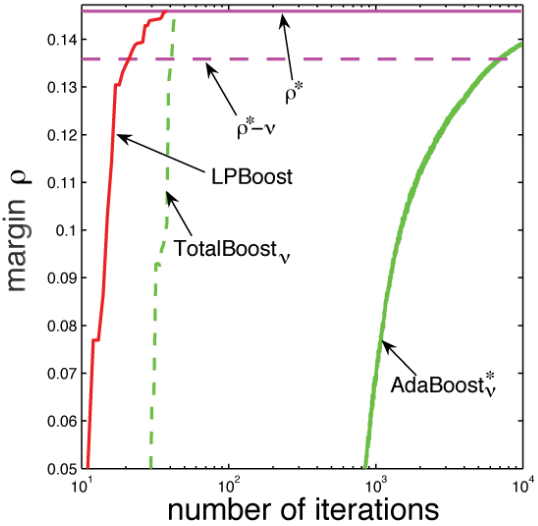
Cox-1 ($\nu = 0.01$)

- AdaBoost^*_ν
- LPBoost
- TotalBoost_ν

Results

- Corrective algorithms very slow
- LPBoost & TotalBoost_ν need few iterations
- Initial speed crucially depends on ν

Illustrative Experiments



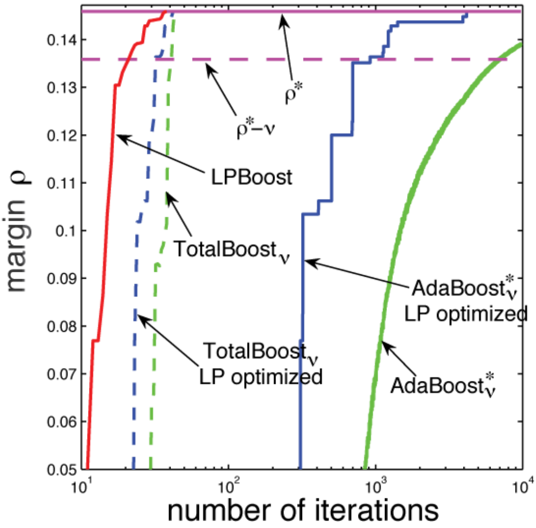
Cox-1 ($\nu = 0.01$)

- AdaBoost$^*_\nu$
- LPBoost
- TotalBoost$_\nu$

Results

- Corrective algorithms very slow
- LPBoost & TotalBoost$_\nu$ need few iterations
- Initial speed crucially depends on ν

Illustrative Experiments



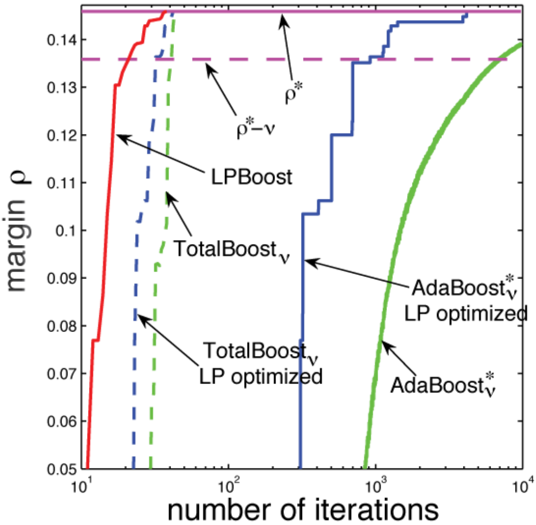
Cox-1 ($\nu = 0.01$)

- AdaBoost$^*_\nu$
- LPBoost
- TotalBoost$_\nu$

Results

- Corrective algorithms very slow
- LPBoost & TotalBoost$_\nu$ need few iterations
- Initial speed crucially depends on ν

Illustrative Experiments



Cox-1 ($\nu = 0.01$)

- AdaBoost$^*_\nu$
- LPBoost
- TotalBoost$_\nu$

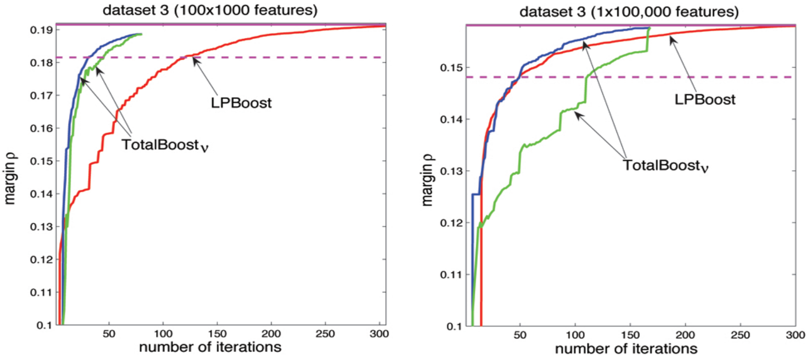
Results

- Corrective algorithms very slow
- LPBoost & TotalBoost$_\nu$ need few iterations
- Initial speed crucially depends on ν

LPBoost May Perform Much Worse Than TotalBoost

- Identified cases where LPBoost converges considerably slower than TotalBoost$_\nu$
- Dataset is a series of artificial datasets of 1000 examples with varying number of features created as follows:
    - First generated N_1 random ± 1-valued features $x_1, \ldots, x_{N_1}$ and set the label of the examples as $y = \mathrm{sign}(x_1 + x_2 + x_3 + x_4 + x_5)$
    - Then duplicated each features N_2 times, perturbed the features by Gaussian noise with $\sigma = 0.1$, and clipped the feature values so that they lie in the interval [-1,1]
    - Considered different N_1, N_2, the total number of features is $N_2 \times N_1$

LPBoost performs worse for high dimensional data with many redundant features



LPBoost vs. TotalBoost$_\nu$ on two 100,000 dimensional datasets: [*left*] many redundant features ($N_1 = 1,000, N_2 = 100$) and [*right*] independent features ($N_1 = 100,000, N_2 = 1$). Show margin vs. number of iterations

Bound Not True for LPBoost

A counter example:

	Hypothesis No.					
	1	2	3	4	5	6
Examples	+	-	-	-	-	-
	+	-	-	-	-	-
	+	-	-	-	-	-
	+	-	-	-	-	-
	+	-	-	-	-	-
	-	+	-	-	-	+
	-	-	+	-	-	+
	-	-	-	+	-	+
	-	-	-	-	+	+

+: correct prediction   -: incorrect prediction

TotalBoost averages hypothesis 1 and 6 (i.e. 2 iterations) to achieve maximum margin 0

Bound Not True for LPBoost

						Distribuition of Examples				
Hypothesis No.						Iteration No.				
1	2	3	4	5	6	1	2	3	4	5
+	-	-	-	-	-	1/9	0	0	0	0
+	-	-	-	-	-	1/9	0	0	0	0
+	-	-	-	-	-	1/9	0	0	0	0
+	-	-	-	-	-	1/9	0	0	0	0
+	-	-	-	-	-	1/9	0	0	0	0
-	+	-	-	-	+	1/9	1	0	0	0
-	-	+	-	-	+	1/9	0	1	0	0
-	-	-	+	-	+	1/9	0	0	1	0
-	-	-	-	+	+	1/9	0	0	0	1

Selected Hypothesis No.

1 2 3 4 5

Simplex-based LPBoost uses 5 iterations ((N+1)/2 iterations) to achieve margin 0

Regularized LPBoost

- LPBoost makes edge constraints as tight as possible
- Picking solutions at the corners can lead to slow convergence. Interior points methods avoid corners
- Regularized LPBoost: pick solution that minimizes relative entropy to initial distribution. It is identical to TotalBoost$_\nu$, but the latter algorithm uses a higher edge bound
- Open problem: find an example where all versions of LPBoost need O(N) iterations

Algorithms for Feature Selection

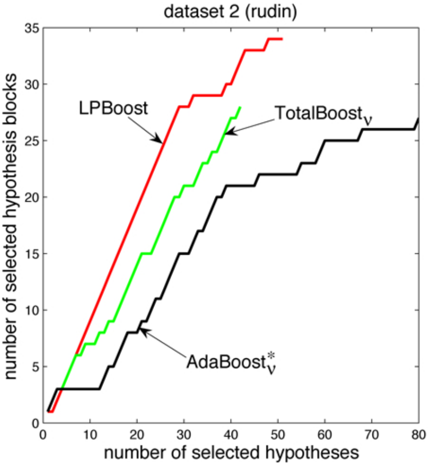
We test for each algorithm:

- Whether selected hypotheses are redundant
- Size of final hypothesis

Redundancy of Selected Hypotheses

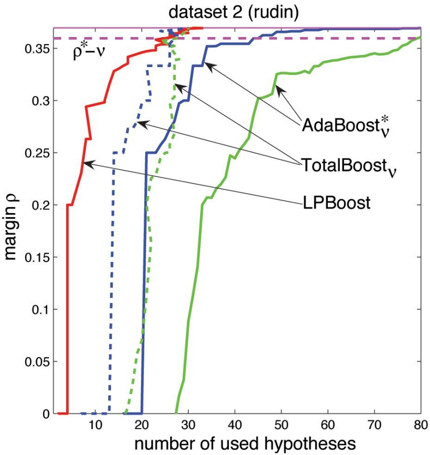
- Test which algorithm selects redundant hypotheses
- Dataset: Dataset 2 was created by expanding Rudin's dataset. Dataset 2 has 100 blocks of features. Each block contains one original feature and 99 mutations of this feature. Each mutation feature inverts the feature value of one randomly chosen example (with replacement). The mutation features are considered redundant hypotheses.
- A good algorithm would avoid repeatedly selecting hypotheses from the same block.

Experiment on Redundancy of Selected Hypotheses



- Show number of selected blocks v.s. number of selected hypotheses
- Redundancy of selected hypotheses: AdaBoost $>$ TotalBoost$_\nu$(ν=0.01) $>$ LPBoost
- If ρ^* is known, TotalBoost$^g_\nu$ selects one hypothesis per block (not shown).

Size of Final Hypothesis

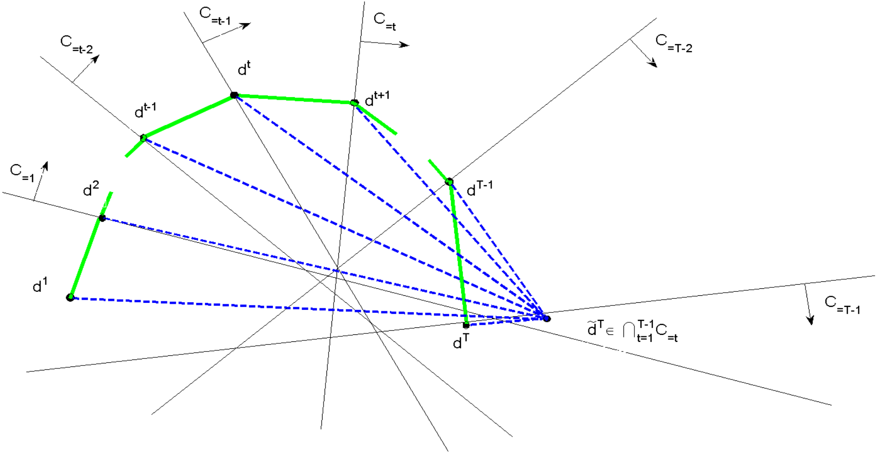


- Show margins v.s. number of used hypotheses (nonzero weight)
- TotalBoost$_\nu$(ν=0.01) and LPBoost use a small number of hypotheses in final hypothesis

Summary

- AdaBoost can be viewed as entropy projection
- TotalBoost projects based on all previous hypotheses
- Provably maximizes the margin
  - Theory: as fast as AdaBoost$^*_\nu$
  - Practice: much faster ($\approx$ LPBoost)
- Versatile techniques for proving iteration bounds
- Experiments corroborate our theory
  - Good for feature selection
  - LPBoost may have problems of maximizing the margin
- Future: extension to the soft margin case

Iteration Bound for Variant of AdaBoost$^*_\nu$



$C_{=t-1}$
$C_{=t}$
$C_{=t-2}$
$C_{=T-2}$
d^t
d^{t-1}
d^{t+1}
$C_{=1}$
d^2
d^{T-1}
d^1
$C_{=T-1}$
$\tilde{d}^T \in \bigcap_{t=1}^{T-1} C_{=t}$
d^T