

# MLSS 2010 Boosting Tutorial

## Recent Advances in Boosting

Part I: Entropy Regularized LPBoost

Part II: Boosting from an Optimization  
Perspective

Manfred K. Warmuth - UCSC

prepared jointly with S.V.N. Vishwanathan - Purdue  
adapted from ICML 2009 tutorial

Updated September 25, 2010

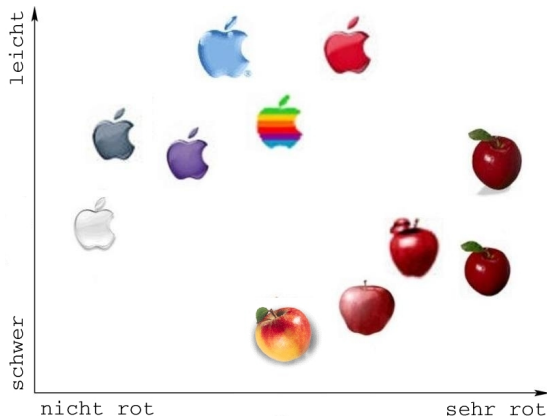
- 1 Introduction to Boosting
- 2 What is Boosting?
- 3 Entropy Regularized LPBoost
- 4 Overview of Boosting algorithms
- 5 Conclusion and Open Problems

# Outline

- 1 Introduction to Boosting
- 2 What is Boosting?
- 3 Entropy Regularized LPBoost
- 4 Overview of Boosting algorithms
- 5 Conclusion and Open Problems

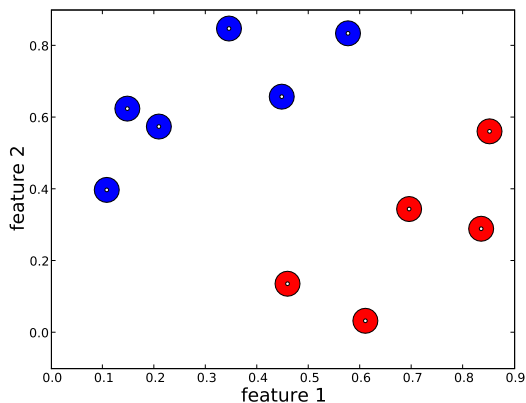
# Setup for Boosting

[Giants of field: Schapire, Freund]



- examples: 11 apples
- $+1$  if artificial  
 $-1$  if natural
- goal:  
classification

# Setup for Boosting

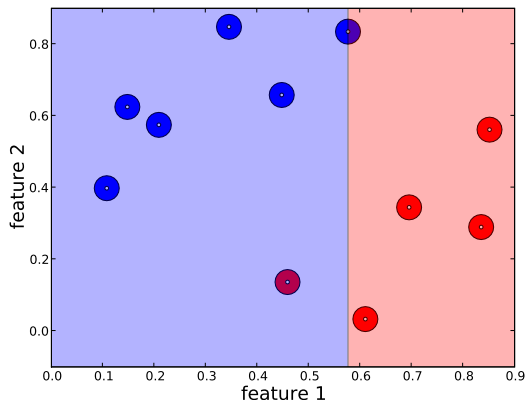


•  $+1/-1$  examples

• weight  $d_n \approx \text{size}$

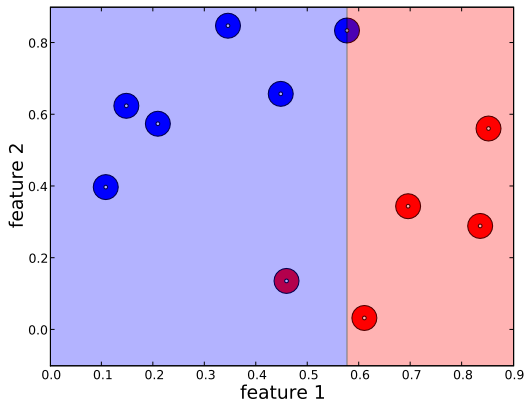
• separable

# Weak hypotheses



- weak hypotheses:  
decision stumps on two features  
**one can't do it**
- goal:  
find convex combination of weak hypotheses that classifies all

# Boosting: 1st iteration

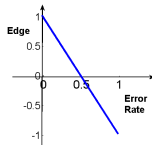


First hypothesis:

• error:  $\frac{1}{11}$

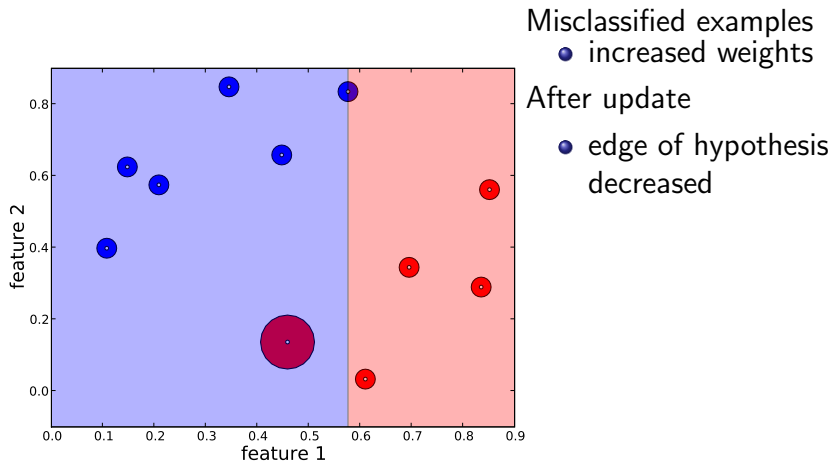
• edge:  $\frac{9}{11}$

low error = high edge



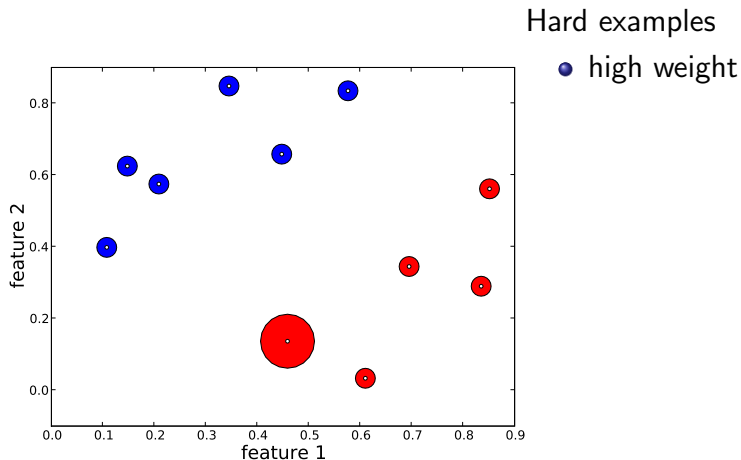
$$\text{edge} = 1 - 2 \text{ error}$$

# Update after 1st



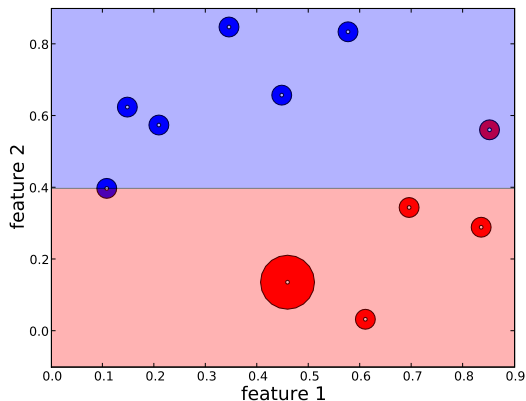


# Before 2nd iteration

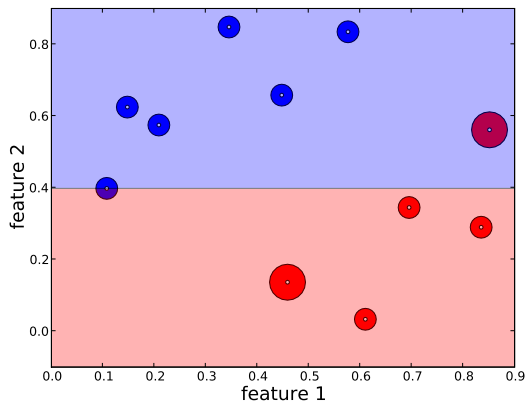


# Boosting: 2nd hypothesis

Pick hypotheses  
with high (weighted) edge



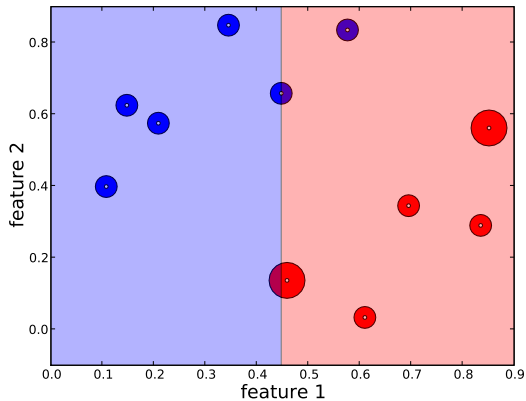
# Update after 2nd



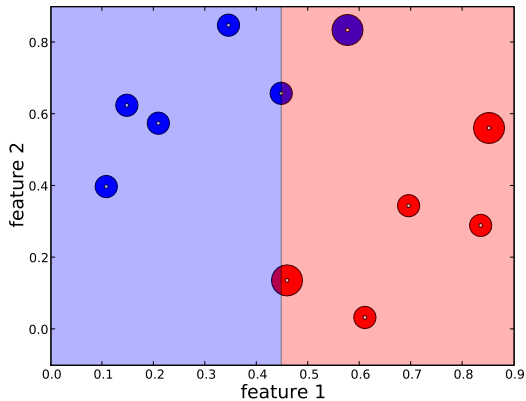
After update

- edges of all past hypotheses should be small

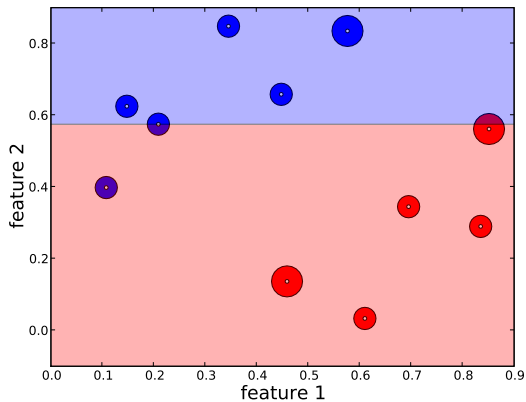
# 3rd hypothesis



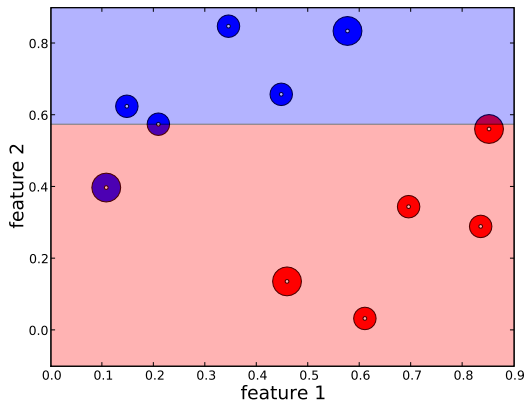
# Update after 3rd



# 4th hypothesis

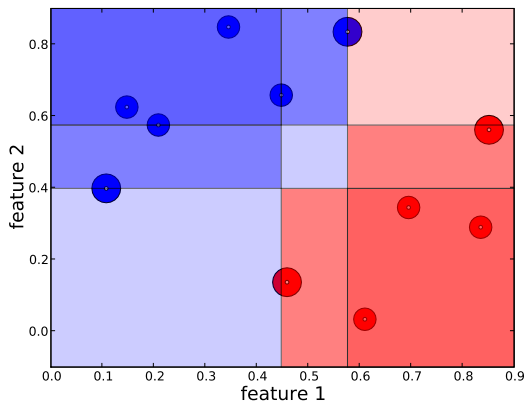


# Update after 4th



# Final convex combination of all hypotheses

Decision:  $\sum_{t=1}^T w_t h^t(\mathbf{x}) \geq 0$  ?



Positive total weight - Negative total weight



# Protocol of Boosting









[FS97]

- Maintain distribution on  $N$   $\pm 1$  labeled examples
- At iteration  $t = 1, \dots, T$ :
  - Receive “weak” hypothesis  $h^t$  of high edge
  - Update  $\mathbf{d}^{t-1}$  to  $\mathbf{d}^t$  **more weights on “hard” examples**
- Output convex combination of the weak hypotheses
 
$$\sum_{t=1}^T w_t h^t(x)$$

Two sets of weights:

- distribution on  $\mathbf{d}$  on examples
- distribution on  $\mathbf{w}$  on hypotheses

# Data representation

	$y_n h^t(x_n) := u_n^t$	examples $x_n$	labels $y_n$	$h^1(x_n)$	$u^1$
			-1	-1	<b>1</b>
			-1	-1	<b>1</b>
			-1	-1	<b>1</b>
perfect	+1		-1	1	<b>-1</b>
opposite	-1		1	1	<b>1</b>
neutral	0		1	1	<b>1</b>
			1	1	<b>1</b>
			1	-1	<b>-1</b>

## Edge vs. margin

[Br99]

Edge of a hypothesis  $h^t$  for a distribution  $\mathbf{d}$  on the examples

$$\underbrace{\sum_{n=1}^N \overbrace{u_n^t}^{\text{accuracy of example}} d_n}_{\text{weighted accuracy of hypothesis}} \quad \mathbf{d} \in \mathcal{P}^N$$

Margin of example  $n$  for current hypothesis weighting  $\mathbf{w}$

$$\underbrace{\sum_{t=1}^T \overbrace{u_n^t}^{\text{accuracy of example}} w_t}_{\text{weighted accuracy of example}} \quad \mathbf{w} \in \mathcal{P}^T$$

## Edge vs. margin

[Br99]

Edge of a hypothesis  $h^t$  for a distribution  $\mathbf{d}$  on the examples

$$\underbrace{\sum_{n=1}^N \overbrace{u_n^t}^{\text{accuracy of example}} d_n}_{\text{weighted accuracy of hypothesis}} \quad \mathbf{d} \in \mathcal{P}^N$$

Margin of example  $n$  for current hypothesis weighting  $\mathbf{w}$

$$\underbrace{\sum_{t=1}^T \overbrace{u_n^t}^{\text{accuracy of example}} w_t}_{\text{weighted accuracy of example}} \quad \mathbf{w} \in \mathcal{P}^T$$

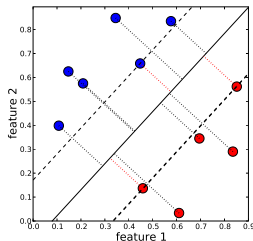
# Objectives

## Edge

- Edges of past hypotheses should be small after update
- Minimize maximum edge of past hypotheses

## Margin

- Choose convex combination of weak hypotheses that maximizes the minimum margin



	Which margin?
SVM	2-norm (weights on examples)
Boosting	1-norm (weights on base hypotheses)

## Connection between objectives?

# Edge vs. margin

$$\min \max \text{ edge} = \max \min \text{ margin}$$

$$\min_{\mathbf{d} \in \mathcal{S}^N} \max_{q=1,2,\dots,t-1} \underbrace{\mathbf{u}^q \cdot \mathbf{d}}_{\text{edge of hypothesis } q} = \max_{\mathbf{w} \in \mathcal{S}^{t-1}} \min_{n=1,2,\dots,N} \underbrace{\sum_{q=1}^{t-1} u_n^q w_q}_{\text{margin of example } n}$$

Linear Programming duality

# Boosting as zero-sum-game

[FS97]

Rock, Paper, Sciissors game

		column player		
		R	P	S
		$w_1$	$w_2$	$w_3$
row player	R	$d_1$ 0	1	-1
	P	$d_2$ -1	0	1
	S	$d_3$ 1	-1	0

gain matrix

Row player minimizes  
Column player maximizes

$$\begin{aligned} \text{payoff} &= \mathbf{d}^T \mathbf{U} \mathbf{w} \\ &= \sum_{i,j} d_i U_{i,j} w_j \end{aligned}$$

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

Single column is pure strategy of  
column player and  $\mathbf{w}$  is mixed strategy

# Optimum strategy

			R	P	S
			$w_1$	$w_2$	$w_3$
			.33	.33	.33
R	$d_1$	.33	0	1	-1
P	$d_2$	.33	-1	0	1
S	$d_3$	.33	1	-1	0

- Min-max theorem:

$$\begin{aligned}
 \min_d \max_w \mathbf{d}^T \mathbf{U} \mathbf{w} &= \min_d \max_j \mathbf{d}^T \mathbf{U} \mathbf{e}_j \\
 &= \max_w \min_d \mathbf{d}^T \mathbf{U} \mathbf{w} = \max_w \min_i e_i \mathbf{U} \mathbf{w} \\
 &= \text{value of the game (0 in example)}
 \end{aligned}$$

$\mathbf{e}_j$  is pure strategy



# Connection to Boosting?

- Rows are the examples
- Columns  $\mathbf{u}^q$  encode weak hypothesis  $h^q$
- Row sum: margin of example
- Column sum: edge of weak hypothesis
- Value of game:

$$\min \max \text{ edge} = \max \min \text{ margin}$$

Van Neumann's Minimax Theorem

## Edges/margins

			R	P	S		
			$w_1$	$w_2$	$w_3$	margin	
			.33	.33	.33		
R	$d_1$	.33	0	1	1	0	
P	$d_2$	.33	-1	0	1	0	min
S	$d_3$	.33	1	-1	-1	0	
	edge		0	0	0		
				max			

value of game 0

# New column added: boosting

			R	P	S		
			$w_1$	$w_2$	$w_3$	$w_4$	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	
			max				

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

# Row added: on-line learning

			R	P	S		
			$w_1$	$w_2$	$w_3$	margin	
			.33	.44	.22		
R	$d_1$	0	0	1	-1	.22	min
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		
			max				

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

# Boosting: maximize margin incrementally

	$w_1^1$		$w_1^2$	$w_2^2$		$w_1^3$	$w_2^3$	$w_3^3$
$d_1^1$	0	$d_1^2$	0	-1	$d_1^3$	0	-1	1
$d_2^1$	1	$d_2^2$	1	0	$d_2^3$	1	0	-1
$d_3^1$	-1	$d_3^2$	-1	1	$d_3^3$	-1	1	0
iteration 1		iteration 2			iteration 3			

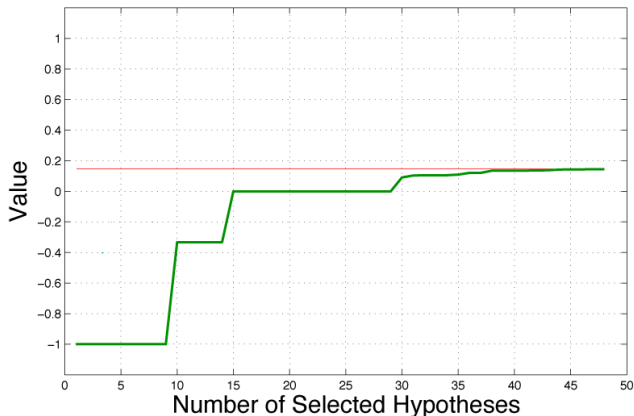
- In each iteration solve optimization problem to update  $\mathbf{d}$
- Column player / oracle provides new hypothesis
- Boosting is column generation method in  $\mathbf{d}$  domain and coordinate descent in  $\mathbf{w}$  domain

# Outline

- 1 Introduction to Boosting
- 2 What is Boosting?
- 3 Entropy Regularized LPBoost
- 4 Overview of Boosting algorithms
- 5 Conclusion and Open Problems

# Boosting = greedy method for increasing margin

Converges to optimum margin w.r.t. all hypotheses



Want small number of iterations

# Assumption on next weak hypothesis

For current weighting of examples,  
oracle returns hypothesis of edge  $\geq g$

## Goal

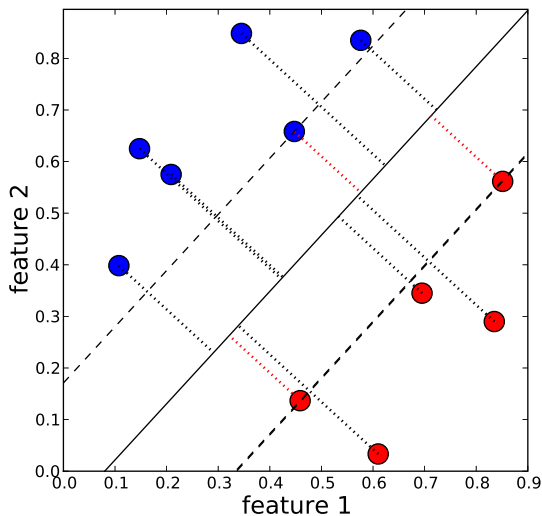
- For given  $\epsilon$ , produce convex combination of weak hypotheses with soft margin  $\geq g - \epsilon$
- Number of iterations  $O(\frac{\log N}{\epsilon^2})$



# Recall min max thm

$$\begin{aligned}
 & \min_{\mathbf{d} \in S^N} \max_{q=1,2,\dots,t} \underbrace{\mathbf{u}^q \cdot \mathbf{d}}_{\text{edge of hypothesis } q} \\
 &= \max_{\mathbf{w} \in S^t} \min_{n=1,2,\dots,N} \underbrace{\left( \sum_{q=1}^t u_n^q w_q \right)}_{\text{margin of example } n}
 \end{aligned}$$

# Visualizing the margin

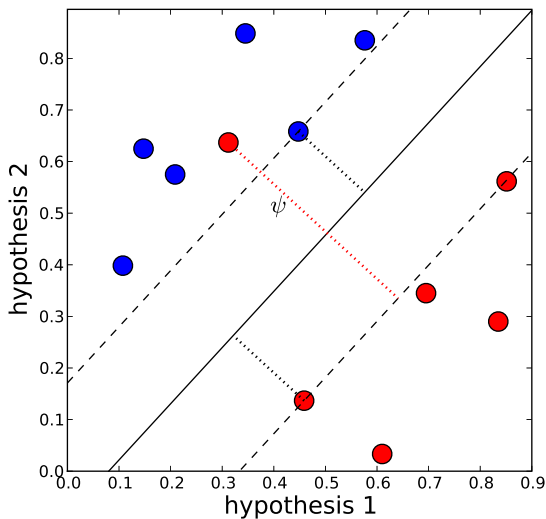


# Min max thm - inseparable case

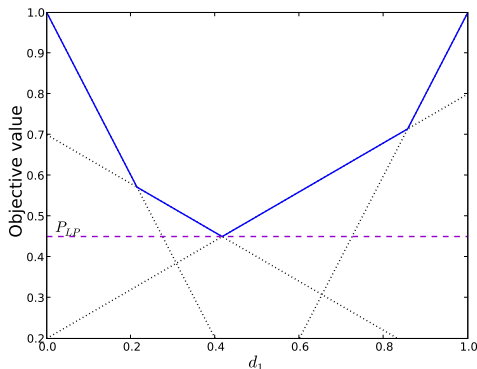
Slack variables in  $\mathbf{w}$  domain = capping in  $\mathbf{d}$  domain

$$\begin{aligned}
 & \min_{\mathbf{d} \in \mathcal{S}^N, \mathbf{d} \leq \frac{1}{\nu} \mathbf{1}} \max_{q=1,2,\dots,t} \underbrace{\mathbf{u}^q \cdot \mathbf{d}}_{\text{edge of hypothesis } q} \\
 &= \max_{\mathbf{w} \in \mathcal{S}^t, \psi \geq 0} \min_{n=1,2,\dots,N} \underbrace{\left( \sum_{q=1}^t u_n^q w_q + \psi_n \right)}_{\text{soft margin of example } n} - \frac{1}{\nu} \sum_{n=1}^N \psi_n
 \end{aligned}$$

# Visualizing the soft margin



# LPBoost



Choose distribution that minimizes the maximum edge of current hypotheses by solving:

$$\underbrace{\min_{\sum_n d_n = 1, \mathbf{d} \leq \frac{1}{\nu} \mathbf{1}} \max_{q=1,2,\dots,t} \mathbf{u}^q \cdot \mathbf{d}}_{P_{LP}^t}$$

All weight is put on examples with minimum soft margin

# Outline

- 1 Introduction to Boosting
- 2 What is Boosting?
- 3 Entropy Regularized LPBoost**
- 4 Overview of Boosting algorithms
- 5 Conclusion and Open Problems

# Entropy Regularized LPBoost

$$\min_{\sum_n d_n=1, \mathbf{d} \leq \frac{1}{\nu} \mathbf{1}} \max_{q=1,2,\dots,t} \mathbf{u}^q \cdot \mathbf{d} + \frac{1}{\eta} \Delta(\mathbf{d}, \mathbf{d}^0)$$

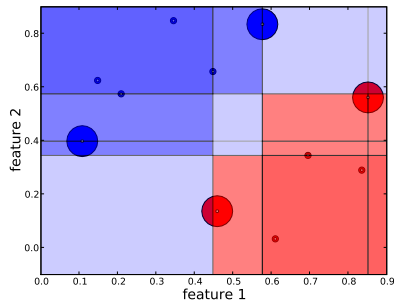


$$\mathbf{d}_n = \frac{\exp^{-\eta \text{ soft margin of example } n}}{Z} \quad \text{"soft min"}$$

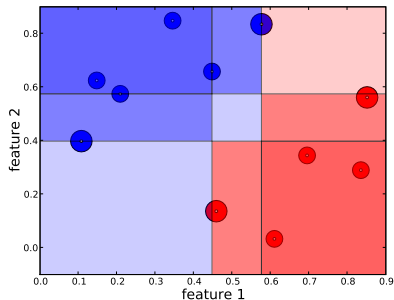
- Form of weights first in  $\nu$ -Arc algorithm [RSS+00]
- Regularization in  $\mathbf{d}$  domain makes problem strongly convex
- Gradient of dual Lipschitz continuous in  $\mathbf{w}$  [e.g. HL93,RW97]

# The effect of entropy regularization

Different distribution on the examples



LPBoost: lots of zeros / brittle



ERLPBoost: smoother



# Outline

- 1 Introduction to Boosting
- 2 What is Boosting?
- 3 Entropy Regularized LPBoost
- 4 Overview of Boosting algorithms**
- 5 Conclusion and Open Problems

## AdaBoost

[FS97]

$$d_n^t := \frac{d_n^{t-1} \exp(-w_t u_n^t)}{\sum_{n'} d_{n'}^{t-1} \exp(-w_t u_{n'}^t)},$$

where  $w_t$  s.t.  $\sum_{n'} d_{n'}^{t-1} \exp(-w u_{n'}^t)$  is minimized

$$\text{i.e. } \left. \frac{\partial \sum_{n'} d_{n'}^{t-1} \exp(-w u_{n'}^t)}{\partial w} \right|_{w=w_t} = \sum_n u_n^t \frac{d_n^{t-1} \exp(-w_t u_n^t)}{\sum_{n'} d_{n'}^{t-1} \exp(-w_t u_{n'}^t)} = \mathbf{u}^t \cdot \mathbf{d}^t = 0$$

- Easy to implement
- Adjusts distribution so that edge of **last** hypothesis is zero
- Gets within half of the optimal hard margin but only in the limit

[RSD07]

# Corrective versus totally corrective

Processing **last** hypothesis versus **all** past hypotheses

Corrective	Totally Corrective
AdaBoost	LPBoost
LogitBoost	TotalBoost
AdaBoost*	SoftBoost
SS,Colt08	ERLPBoost

# From AdaBoost to ERLPBoost

## AdaBoost

(as interpreted in [KW99,La99])

Primal:

Dual:

$$\begin{aligned} \min_{\mathbf{d}} \quad & \Delta(\mathbf{d}, \mathbf{d}^{t-1}) \\ \text{s.t.} \quad & \mathbf{d} \cdot \mathbf{u}^t = 0, \|\mathbf{d}\|_1 = 1 \end{aligned}$$

$$\begin{aligned} \max_{\mathbf{w}} \quad & -\ln \sum_n d_n^{t-1} \exp(-\eta u_n^t w_t) \\ \text{s.t.} \quad & \mathbf{w} \geq 0 \end{aligned}$$

Achieves half of optimum hard margin in the limit

## AdaBoost\*

[RW05]

Primal:

Dual:

$$\begin{aligned} \min_{\mathbf{d}} \quad & \Delta(\mathbf{d}, \mathbf{d}^{t-1}) \\ \text{s.t.} \quad & \mathbf{d} \cdot \mathbf{u}^t \leq \gamma_t, \\ & \|\mathbf{d}\|_1 = 1 \end{aligned}$$

$$\begin{aligned} \max_{\mathbf{w}} \quad & -\ln \sum_n d_n^{t-1} \exp(-\eta u_n^t w_t) \\ & -\gamma_t \|\mathbf{w}\|_1 \\ \text{s.t.} \quad & \mathbf{w} \geq 0 \end{aligned}$$

where edge bound  $\gamma_t$  is adjusted downward by a heuristic

Good iteration bound for reaching optimum hard margin

**SoftBoost**

[WGR07]

Primal:

$$\begin{aligned}
\min_{\mathbf{d}} \quad & \Delta(\mathbf{d}, \mathbf{d}^0) \\
\text{s.t.} \quad & \|\mathbf{d}\|_1 = 1, \mathbf{d} \leq \frac{1}{\nu} \mathbf{1} \\
& \mathbf{d} \cdot \mathbf{u}^q \leq \gamma_t, \\
& 1 \leq q \leq t
\end{aligned}$$

Dual:

$$\begin{aligned}
\min_{\mathbf{w}, \psi} \quad & -\ln \sum_n \mathbf{d}_n^0 \exp(-\eta \sum_{q=1}^t u_n^q w_q \\
& -\eta \psi_n) - \frac{1}{\nu} \|\psi\|_1 - \gamma_t \|\mathbf{w}\|_1 \\
\text{s.t.} \quad & \mathbf{w} \geq 0, \psi \geq 0
\end{aligned}$$

where edge bound  $\gamma_t$  is adjusted downward by a heuristic

Good iteration bound for reaching soft margin

**ERLPBoost**

[WGV08]

Primal:

$$\begin{aligned}
\min_{\mathbf{d}, \gamma} \quad & \gamma + \frac{1}{\eta} \Delta(\mathbf{d}, \mathbf{d}^0) \\
\text{s.t.} \quad & \|\mathbf{d}\|_1 = 1, \mathbf{d} \leq \frac{1}{\nu} \mathbf{1} \\
& \mathbf{d} \cdot \mathbf{u}^q \leq \gamma, \\
& 1 \leq q \leq t
\end{aligned}$$

Dual:

$$\begin{aligned}
\min_{\mathbf{w}, \psi} \quad & -\frac{1}{\eta} \ln \sum_n \mathbf{d}_n^0 \exp(-\eta \sum_{q=1}^t u_n^q w_q \\
& -\eta \psi_n) - \frac{1}{\nu} \|\psi\|_1 \\
\text{s.t.} \quad & \mathbf{w} \geq 0, \|\mathbf{w}\|_1 = 1, \psi \geq 0
\end{aligned}$$

where for the iteration bound  $\eta$  is fixed to  $\max(\frac{2}{\epsilon} \ln \frac{N}{\nu}, \frac{1}{2})$

Good iteration bound for reaching soft margin

# Corrective ERLPBoost

[SS08]

Primal:

$$\begin{aligned} \min_{\mathbf{d}} \quad & \sum_{q=1}^t w_q (\mathbf{u}^q \cdot \mathbf{d}) + \frac{1}{\eta} \Delta(\mathbf{d}, \mathbf{d}^0) \\ \text{s.t.} \quad & \|\mathbf{d}\|_1 = 1, \mathbf{d} \leq \frac{1}{\nu} \mathbf{1} \end{aligned}$$

Dual:

$$\begin{aligned} \min_{\psi} \quad & -\frac{1}{\eta} \ln \sum_n \mathbf{d}_n^0 \exp(-\eta \sum_{q=1}^t u_n^q w_q - \eta \psi_n) - \frac{1}{\nu} \|\psi\|_1 \\ \text{s.t.} \quad & \psi \geq 0 \end{aligned}$$

where for the iteration bound  $\eta$  is fixed to  $\max(\frac{2}{\epsilon} \ln \frac{N}{\nu}, \frac{1}{2})$

Good iteration bound for reaching soft margin

# Iteration bounds

Corrective	Totally Corrective
AdaBoost	LPBoost
LogitBoost	TotalBoost
AdaBoost*	SoftBoost
SS, Colt08	ERLPBoost

- Strong oracle: returns hypothesis with maximum edge
- Weak oracle: returns hypothesis with edge  $\geq g$

- In  $O(\frac{\log \frac{N}{\nu}}{\epsilon^2})$  iterations  
within  $\epsilon$  of maximum soft margin for strong oracle  
or within  $\epsilon$  of  $g$  for weak oracle
- Ditto for hard margin case
- In  $O(\frac{\log N}{g^2})$  iterations consistency with weak oracle

# LPBoost may require $\Omega(N)$ iterations

		$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	margin
		0	0	0	0	0	
$d_1$	.125	+1	-.95	-.93	-.91	-.99	—
$d_2$	.125	+1	-.95	-.93	-.91	-.99	—
$d_3$	.125	+1	-.95	-.93	-.91	-.99	—
$d_4$	.125	+1	-.95	-.93	-.91	-.99	—
$d_5$	.125	-.98	+1	-.93	-.91	+.99	—
$d_6$	.125	-.97	-.96	+1	-.91	+.99	—
$d_7$	.125	-.97	-.95	-.94	+1	+.99	—
$d_8$	.125	-.97	-.95	-.93	-.92	+.99	—
edge		.0137	-.7075	-.6900	-.6725	.0000	
value	-1						



# LPBoost may require $\Omega(N)$ iterations

		$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	margin
		1	0	0	0	0	
$d_1$	0	+1	-.95	-.93	-.91	-.99	1
$d_2$	0	+1	-.95	-.93	-.91	-.99	1
$d_3$	0	+1	-.95	-.93	-.91	-.99	1
$d_4$	0	+1	-.95	-.93	-.91	-.99	1
$d_5$	1	-.98	+1	-.93	-.91	+.99	-.98
$d_6$	0	-.97	-.96	+1	-.91	+.99	-.97
$d_7$	0	-.97	-.95	-.94	+1	+.99	-.97
$d_8$	0	-.97	-.95	-.93	-.92	+.99	-.97
edge		-.98	1	-.93	-.91	.99	
value	-1	-.98					

# LPBoost may require $\Omega(N)$ iterations

		$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	margin
		0	1	0	0	0	
$d_1$	0	+1	-.95	-.93	-.91	-.99	-.95
$d_2$	0	+1	-.95	-.93	-.91	-.99	-.95
$d_3$	0	+1	-.95	-.93	-.91	-.99	-.95
$d_4$	0	+1	-.95	-.93	-.91	-.99	-.95
$d_5$	0	-.98	+1	-.93	-.91	+.99	1
$d_6$	1	-.97	-.96	+1	-.91	+.99	-.96
$d_7$	0	-.97	-.95	-.94	+1	+.99	-.95
$d_8$	0	-.97	-.95	-.93	-.92	+.99	-.95
edge		-.97	-.96	1	-.91	.99	
value	-1	-.98	-.96				

# LPBoost may require $\Omega(N)$ iterations

		$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	margin
		0	0	1	0	0	
$d_1$	0	+1	-.95	-.93	-.91	-.99	-.93
$d_2$	0	+1	-.95	-.93	-.91	-.99	-.93
$d_3$	0	+1	-.95	-.93	-.91	-.99	-.93
$d_4$	0	+1	-.95	-.93	-.91	-.99	-.93
$d_5$	0	-.98	+1	-.93	-.91	+.99	-.93
$d_6$	0	-.97	-.96	+1	-.91	+.99	1
$d_7$	1	-.97	-.95	-.94	+1	+.99	-.94
$d_8$	0	-.97	-.95	-.93	-.92	+.99	-.93
edge		-.97	-.95	-.94	1	.99	
value	-1	-.98	-.96	-.94			

# LPBoost may require $\Omega(N)$ iterations

		$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	margin
		0	0	0	1	0	
$d_1$	0	+1	-.95	-.93	-.91	-.99	-.91
$d_2$	0	+1	-.95	-.93	-.91	-.99	-.91
$d_3$	0	+1	-.95	-.93	-.91	-.99	-.91
$d_4$	0	+1	-.95	-.93	-.91	-.99	-.91
$d_5$	0	-.98	+1	-.93	-.91	+.99	-.91
$d_6$	0	-.97	-.96	+1	-.91	+.99	-.91
$d_7$	0	-.97	-.95	-.94	+1	+.99	1
$d_8$	1	-.97	-.95	-.93	-.92	+.99	-.92
edge		-.97	-.95	-.94	-.92	.99	
value	-1	-.98	-.96	-.94	-.92		

# LPBoost may require $\Omega(N)$ iterations

		$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	margin
		.5	.0026	0	0	.4975	
$d_1$	.497	+1	-.95	-.93	-.91	-.99	.0051
$d_2$	0	+1	-.95	-.93	-.91	-.99	.0051
$d_3$	0	+1	-.95	-.93	-.91	-.99	.0051
$d_4$	0	+1	-.95	-.93	-.91	-.99	.0051
$d_5$	0	-.98	+1	-.93	-.91	+.99	.0051
$d_6$	.490	-.97	-.96	+1	-.91	+.99	.0051
$d_7$	0	-.97	-.95	-.94	+1	+.99	.0051
$d_8$	.013	-.97	-.95	-.93	-.92	+.99	.0051
edge		.0051	.0051	.9055	.9100	.0051	
value	-1	-.98	-.96	-.94	-.92	.0051	

**No ties!**

# LPBoost may return bad final hypothesis

How good is the master hypothesis returned by LPBoost compared to the best possible convex combination of hypotheses?

Any linearly separable dataset can be reduced to a dataset on which LPBoost misclassifies all examples by

- adding a bad example
- adding a bad hypothesis

# Adding a bad example

		$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	margin
		.5	.0026	0	0	.4975	
$d_1$	0	+1	-.95	-.93	-.91	-.99	.0051
$d_2$	0	+1	-.95	-.93	-.91	-.99	.0051
$d_3$	0	+1	-.95	-.93	-.91	-.99	.0051
$d_4$	0	+1	-.95	-.93	-.91	-.99	.0051
$d_5$	0	-.98	+1	-.93	-.91	+.99	.0051
$d_6$	0	-.97	-.96	+1	-.91	+.99	.0051
$d_7$	0	-.97	-.95	-.94	+1	+.99	.0051
$d_8$	0	-.97	-.95	-.93	-.92	+.99	.0051
$d_9$	1	<b>-.03</b>	<b>-.03</b>	<b>-.03</b>	<b>-.03</b>	<b>-.03</b>	<b>-.03</b>
edge		-.03	-.03	-.03	-.03	-.03	
value	-1	-.98	-.96	-.94	-.92	-.03	

# Adding a bad hypothesis

		$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	$w_6$	margin
		0	0	0	0	0	1	
$d_1$	0	+1	-.95	-.93	-.91	-.99	-.01	.0051
$d_2$	0	+1	-.95	-.93	-.91	-.99	-.01	.0051
$d_3$	0	+1	-.95	-.93	-.91	-.99	-.01	.0051
$d_4$	0	+1	-.95	-.93	-.91	-.99	-.01	.0051
$d_5$	0	-.98	+1	-.93	-.91	+.99	-.01	.0051
$d_6$	0	-.97	-.96	+1	-.91	+.99	-.01	.0051
$d_7$	0	-.97	-.95	-.94	+1	+.99	-.01	.0051
$d_8$	0	-.97	-.95	-.93	-.92	+.99	-.01	.0051
$d_9$	1	-.03	-.03	-.03	-.03	-.03	-.02	.0051
edge		-.03	-.03	-.03	-.03	-.03	-.02	
value	-1	-.98	-.96	-.94	-.92	-.03		



# Adding a bad hypothesis

		$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	$w_6$	margin
		0	0	0	0	0	1	
$d_1$	0	+1	-.95	-.93	-.91	-.99	<b>-.01</b>	-.01
$d_2$	0	+1	-.95	-.93	-.91	-.99	<b>-.01</b>	-.01
$d_3$	0	+1	-.95	-.93	-.91	-.99	<b>-.01</b>	-.01
$d_4$	0	+1	-.95	-.93	-.91	-.99	<b>-.01</b>	-.01
$d_5$	0	-.98	+1	-.93	-.91	+.99	<b>-.01</b>	-.01
$d_6$	0	-.97	-.96	+1	-.91	+.99	<b>-.01</b>	-.01
$d_7$	0	-.97	-.95	-.94	+1	+.99	<b>-.01</b>	-.01
$d_8$	0	-.97	-.95	-.93	-.92	+.99	<b>-.01</b>	-.01
$d_9$	1	-.03	-.03	-.03	-.03	-.03	<b>-.02</b>	-.02
edge		-.03	-.03	-.03	-.03	-.03	-.02	
value	-1	-.98	-.96	-.94	-.92	-.03	<b>-.02</b>	

# Adding a bad hypothesis

		$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	$w_6$	margin
		0	0	0	0	0	1	
$d_1$	0	+1	-.95	-.93	-.91	-.99	-.01	-.01
$d_2$	0	+1	-.95	-.93	-.91	-.99	-.01	-.01
$d_3$	0	+1	-.95	-.93	-.91	-.99	-.01	-.01
$d_4$	0	+1	-.95	-.93	-.91	-.99	-.01	-.01
$d_5$	0	-.98	+1	-.93	-.91	+.99	-.01	-.01
$d_6$	0	-.97	-.96	+1	-.91	+.99	-.01	-.01
$d_7$	0	-.97	-.95	-.94	+1	+.99	-.01	-.01
$d_8$	0	-.97	-.95	-.93	-.92	+.99	-.01	-.01
$d_9$	1	-.03	-.03	-.03	-.03	-.03	-.02	-.02
edge		-.03	-.03	-.03	-.03	-.03	-.02	
value	-1	-.98	-.96	-.94	-.92	-.03	-.02	

# Adding a bad hypothesis

		$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	$w_6$	margin
		.5	0	0	0	.5	0	
$d_1$	0	<b>+1</b>	-.95	-.93	-.91	<b>-.99</b>	-.01	<b>+.005</b>
$d_2$	0	<b>+1</b>	-.95	-.93	-.91	<b>-.99</b>	-.01	<b>+.005</b>
$d_3$	0	<b>+1</b>	-.95	-.93	-.91	<b>-.99</b>	-.01	<b>+.005</b>
$d_4$	0	<b>+1</b>	-.95	-.93	-.91	<b>-.99</b>	-.01	<b>+.005</b>
$d_5$	0	<b>-.98</b>	+1	-.93	-.91	<b>+.99</b>	-.01	<b>+.005</b>
$d_6$	0	<b>-.97</b>	-.96	+1	-.91	<b>+.99</b>	-.01	<b>+.01</b>
$d_7$	0	<b>-.97</b>	-.95	-.94	+1	<b>+.99</b>	-.01	<b>+.01</b>
$d_8$	0	<b>-.97</b>	-.95	-.93	-.92	<b>+.99</b>	-.01	<b>+.01</b>
$d_9$	1	<b>-.03</b>	-.03	-.03	-.03	<b>-.03</b>	-.02	<b>-.03</b>

# Synopsis

- LPBoost often unstable
- For safety, add relative entropy regularization
- Corrective algs
  - Sometimes easy to code
  - Fast per iteration
- Totally corrective algs
  - Smaller number of iterations
  - Faster overall time when  $\epsilon$  small
- **Weak** versus **strong** oracle makes a big difference in practice

# $O(\frac{\log N}{\epsilon^2})$ iteration bounds

## Good

- Bound is major design tool
- Any reasonable Boosting algorithm should have this bound

## Bad

- Bound is weak
 

	$\frac{\ln N}{\epsilon^2} \geq N$
$\epsilon = .01$	$N \leq 1.2 \cdot 10^5$
$\epsilon = .001$	$N \leq 1.7 \cdot 10^7$
- Why are totally corrective algorithms much better in practice?

# Lower bounds on the number of iterations

- Majority of  $\Omega(\frac{\log N}{g^2})$  hypotheses for achieving consistency with **weak oracle** of guarantee  $g$  [Fr95]
- Easy:  $\Omega(\frac{1}{\epsilon^2})$  iteration bound for getting within  $\epsilon$  of hard margin with **strong oracle**
- Harder:  $\Omega(\frac{\log N}{\epsilon^2})$  iteration bound for **strong** oracle [Ne83?]

# Outline

- 1 Introduction to Boosting
- 2 What is Boosting?
- 3 Entropy Regularized LPBoost
- 4 Overview of Boosting algorithms
- 5 Conclusion and Open Problems

# Conclusion

- Adding relative entropy regularization of LPBoost leads to good boosting alg.
- Boosting is instantiation of MaxEnt and MinxEnt principles  
[Jaines 57, Kullback 59]
- Relative entropy regularization smoothes one-norm regularization

## Open

- When hypotheses have one-sided error then  $O(\frac{\log N}{\epsilon})$  iterations suffice [As00, HW03]
- Does ERLPBoost have  $O(\frac{\log N}{\epsilon})$  bound when hypotheses one-sided?
- Replace geometric optimizers by entropic ones
- Compare ours with Freund's algorithms that don't just cap, but forget examples



# Acknowledgment

- Rob Schapire and Yoav Freund for pioneering Boosting
- Gunnar Rätsch for bringing in optimization
- Karen Glocer for helping with figures and plots