Entropy Regularized LPBoost

Manfred K. Warmuth
Karen Glocer
S.V.N. Vishwanathan
(pretty slides from Gunnar Rätsch)

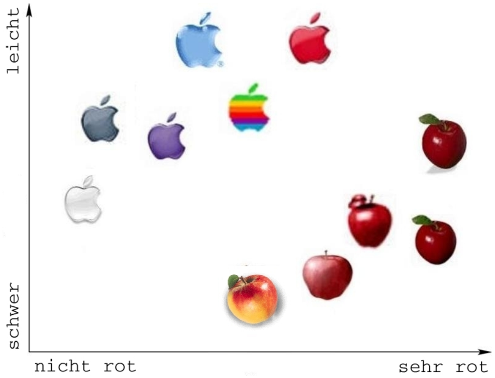
Updated: October 13, 2008

Protocol of Boosting [FS97]

- Maintain distribution on N ± 1 labeled examples
- At iteration $t = 1, \ldots, T$:
  - Receive a "weak" hypothesis h^t
  - Update $\mathbf{d}^{t-1}$ to $\mathbf{d}^t$ - put more weights on "hard" examples
- Output a convex combination of the weak hypotheses $\sum_{t=1}^{T} w_t h^t(x)$

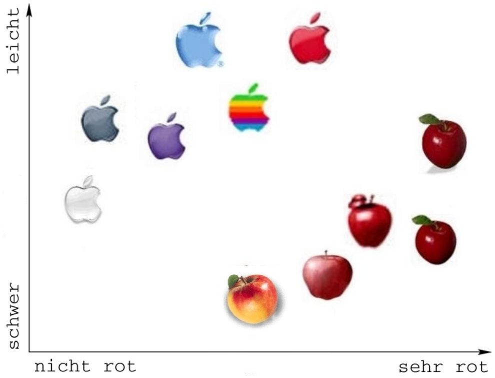
Two sets of weights:
- distribution on $\mathbf{d}$ on examples
- distribution on $\mathbf{w}$ on hypotheses

Setup for Boosting



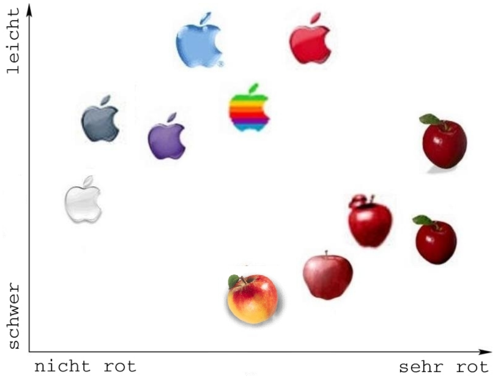
- 11 apples (examples)
- labeled +1 if natural and -1 if artificial
- want to classify the apples

Setup for Boosting



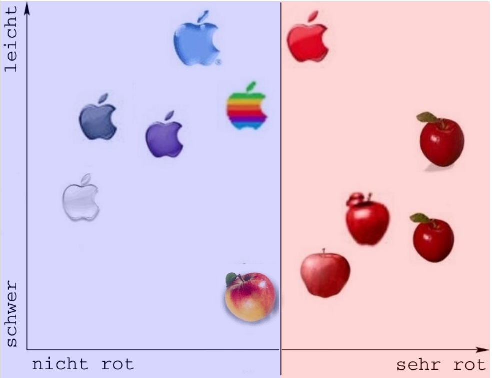
- 11 apples (examples)
- labeled +1 if natural and -1 if artificial
- want to classify the apples

Setup for Boosting



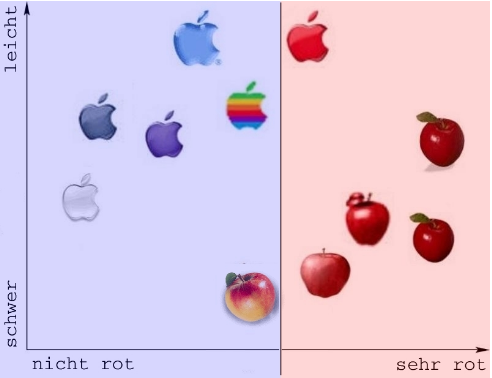
- 11 apples (examples)
- labeled $+1$ if natural and -1 if artificial
- want to classify the apples

Weak hypothesis



- weak hypotheses are decision stumps along the two features
- examples = rows
- weak hypotheses = possible columns

Weak hypothesis

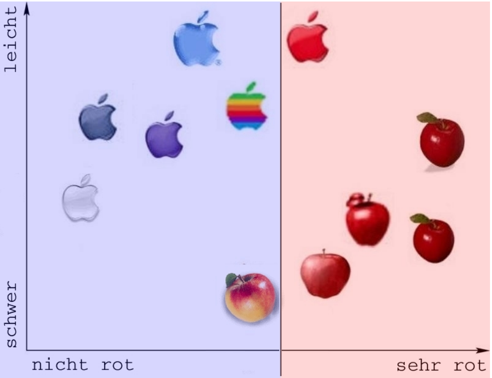


- weak hypotheses are decision stumps along the two features
- examples = rows
- weak hypotheses = possible columns

Examples and Hypotheses

Examples	Labels	h_1: redness	
	-1	-1	
	-1	-1	
	-1	-1	
	-1	1	mistake
	1	1	
	1	1	
	1	1	
	1	-1	mistake

Boosting: 1st Iteration



First hypothesis:

- error: $\frac{2}{11}$

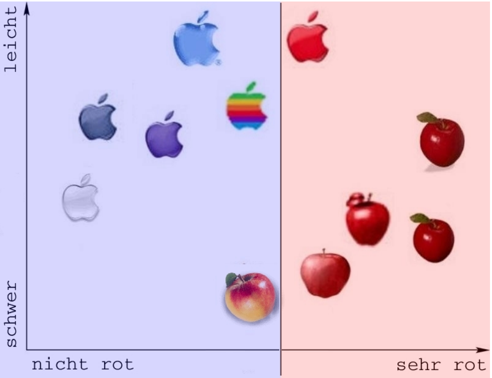
$$\sum_{n=1}^{N} d_n^0 \, \mathbf{I}(h^1(\mathbf{x}_n) \neq y_n)$$

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

$$\underbrace{\sum_{n=1}^{N} \underbrace{y_n h(\mathbf{x}_n)}_{\text{goodness on ex. } n} \quad d_n}_{\text{average goodness}}$$

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

Boosting: 1st Iteration



First hypothesis:

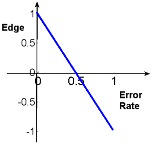
- error: $\frac{2}{11}$

$$\sum_{n=1}^{N} d_n^0 \, \mathbf{I}(h^1(\mathbf{x}_n) \neq y_n)$$

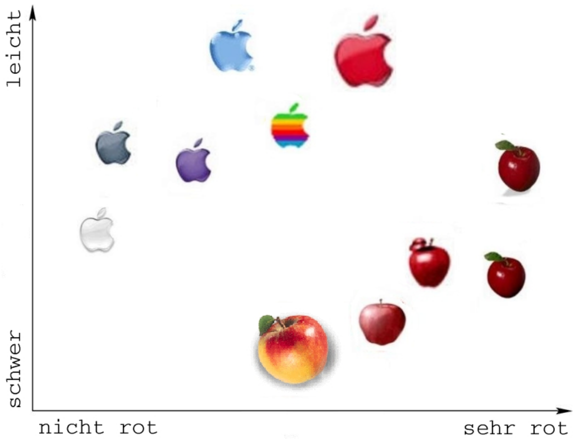
- edge: $\frac{9}{22}$

$$\underbrace{\sum_{n=1}^{N} \underbrace{y_n h(\mathbf{x}_n)}_{\text{goodness on ex. } n} \quad d_n}_{\text{average goodness}}$$

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



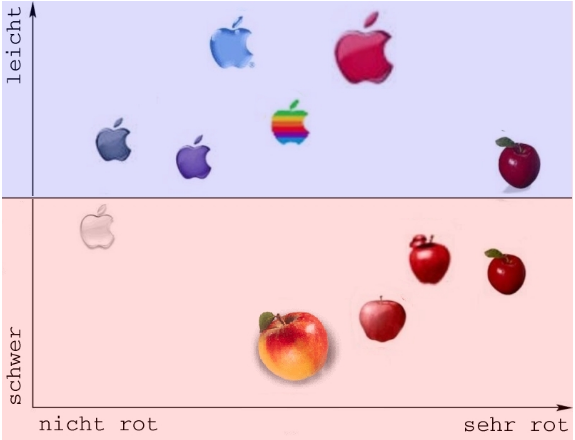
Before 2nd Iteration



Hard examples

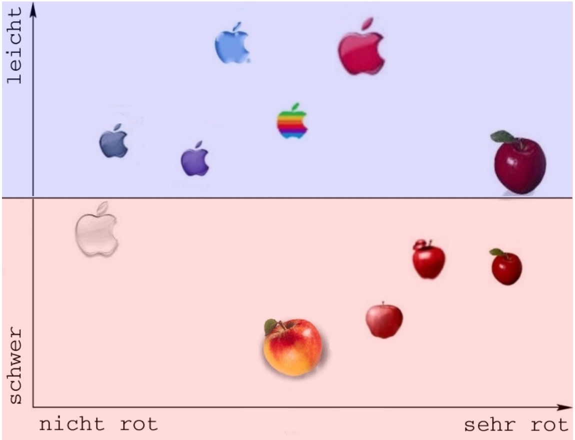
- high weight

Boosting: 2nd Hypothesis



Pick hypotheses
with high edge

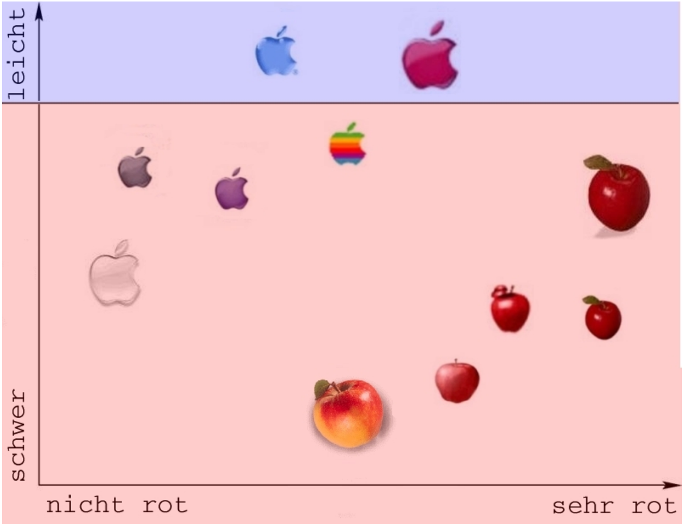
Update Distribution



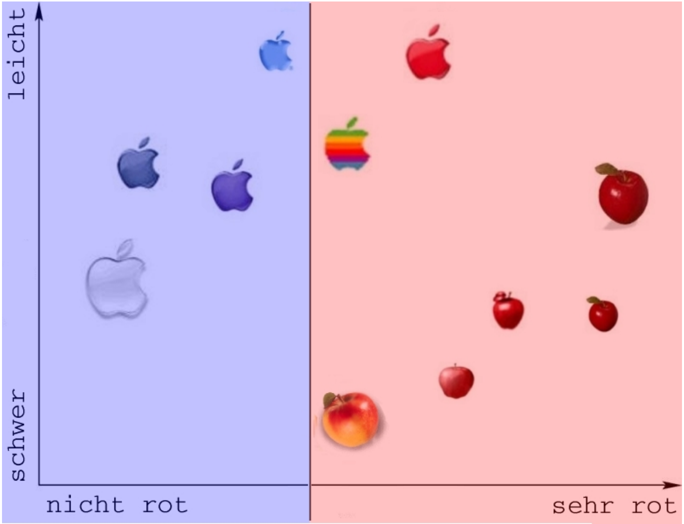
After update:

edges of all
chosen hypotheses
should be small

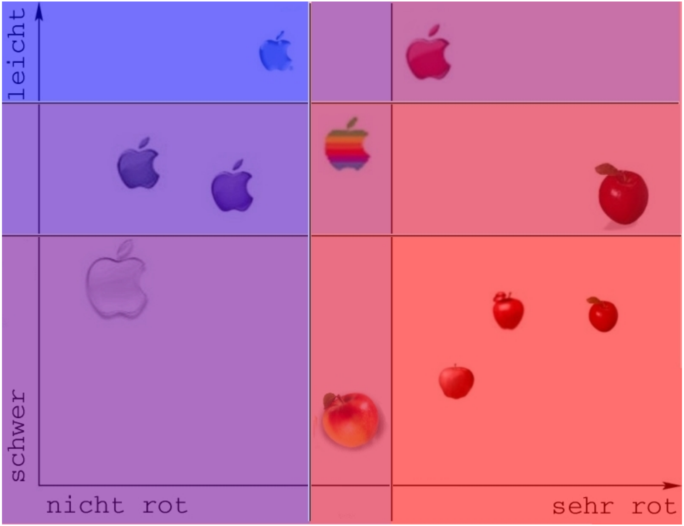
Boosting: 3nd Hypothesis



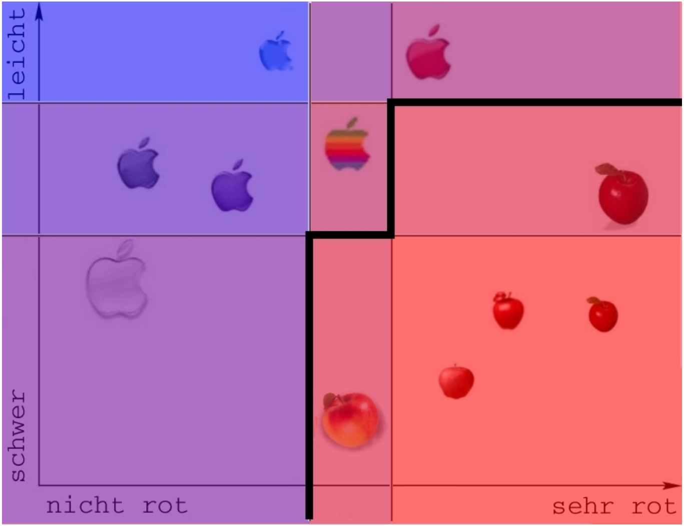
Boosting: 4th Hypothesis



All Hypotheses



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



Edge vs. margin [Br99]

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

$$\underbrace{\sum_{n=1}^{N} \underbrace{y_n h(\mathbf{x}_n)}_{\text{goodness on ex. } n} \quad d_n}_{\text{average goodness}} \qquad \mathbf{d} \in \mathcal{P}^N$$

Margin

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

$$y_n \sum_{t=1}^{T} h^t(\mathbf{x}_n)\, w_t \qquad \mathbf{w} \in \mathcal{P}^T$$

Edge vs. margin [Br99]

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

$$\underbrace{\sum_{n=1}^{N} \underbrace{y_n h(\mathbf{x}_n)}_{\text{goodness on ex. } n} \quad d_n}_{\text{average goodness}} \qquad \mathbf{d} \in \mathcal{P}^N$$

Margin

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

$$y_n \sum_{t=1}^{T} h^t(\mathbf{x}_n)\; w_t \qquad \mathbf{w} \in \mathcal{P}^T$$

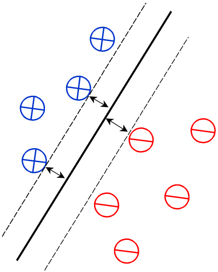
Objectives

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?
SVN	2-norm
Boosting	1-norm

Connection between objectives?

Edge vs. margin

$$\min_{\mathbf{d}\in\mathcal{S}^N} \max_{q=1,2,\ldots,t-1} \underbrace{\sum_{n=1}^{N} y_n h^q(x_n) d_n}_{\text{edge of hypothesis q}}$$

$$= \max_{\mathbf{w}\in\mathcal{S}^{t-1}} \min_{n=1,2,\ldots,N} \underbrace{\sum_{q=1}^{t-1} y_n h^q(x_n)\, w_q}_{\text{margin of example } n}$$

Linear Programming duality

Min max thm for the inseparable case

Slack variables in **w** domain = capping in **d** domain

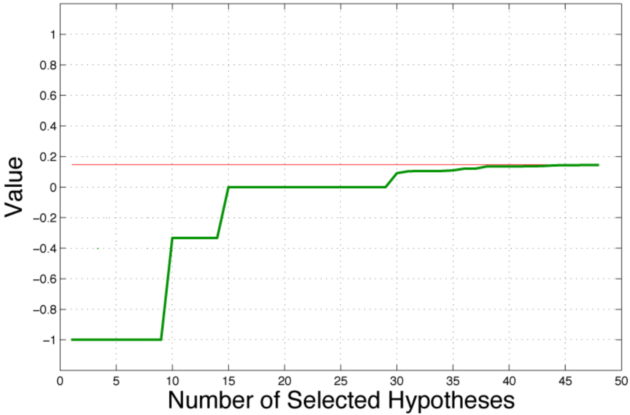
$$\max_{\mathbf{w}\in\mathcal{S}^t,\boldsymbol{\psi}\geq\mathbf{0}} \min_{n=1,2,\ldots,N} \underbrace{\left(\sum_{q=1}^{t} u_n^q\, w_q + \psi_n\right)}_{\text{margin of example } n} - \frac{1}{\nu}\sum_{n=1}^{N}\psi_n$$

$$= \min_{\mathbf{d}\in\mathcal{S}^N,\mathbf{d}\leq\frac{1}{\nu}\mathbf{1}} \max_{q=1,2,\ldots,t} \underbrace{\mathbf{u}^q\cdot\mathbf{d}}_{\text{edge of hypothesis q}}$$

Notation: $u_n^q = y_n h^q(x_n)$

Boosting = greedy method for increasing margin

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



Assumption on next weak hypothesis

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

Goal

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

LPBoost [GS98,RSS+00,DBST02]

Choose distribution that minimizes the maximum edge via LP

$$\min_{\sum_n d_n = 1, \mathbf{d} \leq \frac{1}{\nu}\mathbf{1}} \; \max_{q=1,2,\ldots,t} \; \mathbf{u}^q \cdot \mathbf{d}$$

- Good practical boosting algorithm
- All weight is put on examples with minimum soft margin
- Brittle: iteration bound can be linear in N
  on malign artificial data sets [WGR07]

Entropy Regularized LPBoost

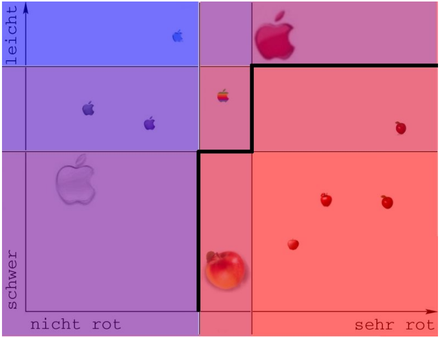
$$\min_{\sum_n d_n=1, \mathbf{d} \leq \frac{1}{\nu}\mathbf{1}} \max_{q=1,2,\ldots,t} \mathbf{u}^q \cdot \mathbf{d} + \frac{1}{\eta}\Delta(\mathbf{d}, \mathbf{d}^0)$$

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

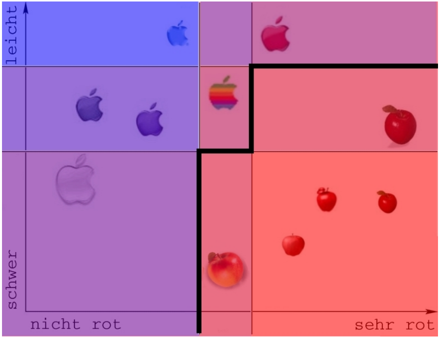
- Within ϵ of maximum soft margin in $O(\frac{\log n/\nu}{\epsilon^2})$ iterations
- Above form of weights first appeared in ν-Arc algorithm [RSS+00]

The effect of entropy regularization

Different distribution on the examples



LPBoost: lots of zeros



ERLPBoost: smoother distribution

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

- Easy to implement
- Gets within half of the optimal hard margin [RSD07]
  but only in the limit

Corrective versus totally corrective

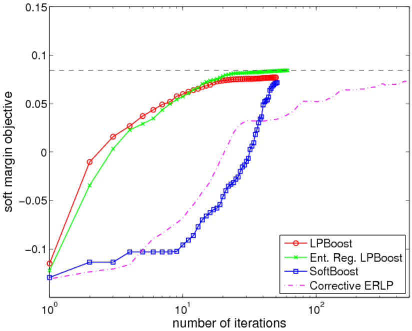
Processing last hypothesis versus all past hypotheses

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

Myths about boosting

- LPBoost does the trick in practice most of the time
- For safety, add relative entropy regularization
- Corrective algs
  - Sometimes easy to code
  - Fast per iteration
- Totally corrective algs
  - Smaller number of iterations
  - Nevertheless faster overall time
- Weak versus strong oracle makes a big difference

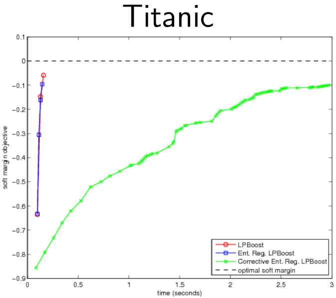
  weak: return hypothesis of edge larger than some guarantee g
  strong: return hypothesis of maximum edge

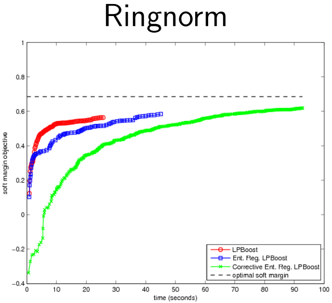


Soft margin objective vs. the number of iterations on a single run for the Banana data set with $\epsilon = 0.01$ and $\nu/N = 0.1$. For ERLPBoost, $\eta = \frac{2}{\epsilon} \log \frac{N}{\nu}$.

- LPBoost indistinguishable from ERLPBoost
- SoftBoost's margin begins increasing much later than the others
- Corrective alg. converges more slowly than totally corrective

Corrective vs. Totally Corrective





- Results for a single run of each algorithm
- Margin vs. time
- Titanic is the smallest dataset we used
- Ringnorm is the largest dataset we used

Conclusion

- Adding relative entropy regularization of LPBoost leads to good boosting alg.
- Boosting is instantiation of MaxEnt and MinxEnt principles [Jaines 57,Kullback 59]
- Is sparsity neccessary for good generalization or is relative entropy reguarization sufficient?

From AdaBoost to SoftBoost

AdaBoost (as interpreted in [KW99,La99])

Primal:

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

Dual:

$$\max_{\mathbf{w}} \quad -ln \sum_n d_n^{t-1} \exp(u_n^{t-1} w_{t-1})$$
$$\text{s.t.} \quad \mathbf{w} \geq 0$$

Achieves half of optimum hard margin in the limit

AdaBoost* [RW05]

Primal:

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

Dual:

$$\max_{\mathbf{w}} \ -ln \sum_n d_n^{t-1} \exp(u_n^{t-1} w_{t-1})$$
$$-\gamma_{t-1} ||\mathbf{w}||_1$$
$$\text{s.t.} \quad \mathbf{w} \geq 0$$

where edgebound γ_t is adjusted downward by a heuristic

Good iteration bound for reaching optimum hard margin

SoftBoost [WGR07]

Primal:

$$\min_{\mathbf{d}} \quad \Delta(\mathbf{d}, \mathbf{d}^0)$$
$$\text{s.t.} \quad \|\mathbf{d}\|_1 = 1,\ \mathbf{d} \le \tfrac{1}{\nu}\mathbf{1}$$
$$\mathbf{d} \cdot \mathbf{u}^q \le \gamma_{t-1},$$
$$1 \le q \le t-1$$

Dual:

$$\min_{\mathbf{w},\boldsymbol{\psi}} \quad -\ln \sum_n \mathbf{d}_n^0 \exp(-\eta \sum_{q=1}^{t-1} u_n^q w_q - \eta\psi_n) - \tfrac{1}{\nu}\|\boldsymbol{\psi}\|_1 - \gamma_{t-1}\|\mathbf{w}\|_1$$
$$\text{s.t.} \quad \mathbf{w} \ge 0,\ \boldsymbol{\psi} \ge 0$$

where edgebound γ_{t-1} is adjusted downward by a heuristic

ERLPBoost [WGV08]

Primal:

$$\min_{\mathbf{d},\gamma} \quad \gamma + \tfrac{1}{\eta}\Delta(\mathbf{d}, \mathbf{d}^0)$$
$$\text{s.t.} \quad \|\mathbf{d}\|_1 = 1,\ \mathbf{d} \le \tfrac{1}{\nu}\mathbf{1}$$
$$\mathbf{d} \cdot \mathbf{u}^q \le \gamma,$$
$$1 \le q \le t-1$$

Dual:

$$\min_{\mathbf{w},\boldsymbol{\psi}} \quad -\tfrac{1}{\eta} \ln \sum_n \mathbf{d}_n^0 \exp(-\eta \sum_{q=1}^{t-1} u_n^q w_q - \eta\psi_n) - \tfrac{1}{\nu}\|\boldsymbol{\psi}\|_1$$
$$\text{s.t.} \quad \mathbf{w} \ge 0,\ \|\mathbf{w}\|_1 = 1,\ \boldsymbol{\psi} \ge 0$$

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