



# Speed and Sparsity of Regularized Boosting

Yongxin Taylor Xi, Zhen James Xiang, Peter J. Ramadge, Robert Schapire  
 Dept. of Electrical Engineering, Dept. of Computer Science, Princeton University, NJ 08544

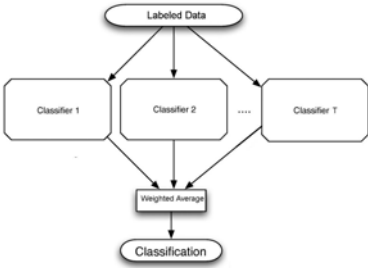


## Abstract

Boosting algorithms with L1-regularization are of interest because L1 regularization leads to sparser composite classifiers. Moreover, Rosset et al. have shown that for separable data, standard Lp regularized loss minimization results in a margin maximizing classifier in the limit as regularization is relaxed. For the L1 case, we extend these results by obtaining explicit convergence bounds on the regularization required to yield a margin within prescribed accuracy of the maximum achievable margin. We also introduce a new hybrid algorithm, AdaBoost+L1, that combines the virtues of AdaBoost with the sparsity of L1 regularization in a computationally efficient fashion. We prove that the algorithm is margin maximizing and empirically examine its performance on UCI data sets.

## Introduction

Boosting as a way of aggregating classifiers:



Example: Adaboost

Pseudocode for AdaBoost:

```

given  $(x_1, y_1), \dots, (x_m, y_m)$   $y_i \in \{-1, +1\}$ 
for  $t = 1, \dots, T$ 
  construct  $d_t$ 
  train  $h_t$  on  $d_t$ 
   $\epsilon_t = \text{err}_{d_t}(h_t) \leq \frac{1}{2} - \theta$ 
output
 $H(x) = \text{sign}\left(\sum_t \alpha_t h_t(x)\right)$ 
    
```

**Weak learnable assumption:** Given any distribution of the data, we can always find at least one classifier better than random guessing.

**Weak learnability = Linear separability (positive margin)**

$$\mu^* = \max_{\alpha \in \Delta_N} \min_{i=1, \dots, m} \sum_{j=1}^N \alpha_j y_i h_j(x_i)$$

$$= \min_{d \in \Delta_m} \max_{j=1, \dots, N} \sum_{i=1}^m d_i y_i h_j(x_i) = \theta^*$$

**Motivation: sparse classifier with large margin**

1. Large margin improves generalization [Schapire 98]:

With probability at least  $1 - \delta$ ,  $\forall f \in \text{co}(\mathcal{H}), \forall \theta > 0$ :

$$\Pr_{\mathcal{D}}[yf(x) \leq 0] \leq \Pr_{\mathcal{S}}[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}} \sqrt{\ln m \cdot \ln |\mathcal{H}| + \ln \frac{1}{\delta}}\right)$$

2. Sparsity facilitates consistency [Zhang & Yu 05]:

Boosting is shown to be consistent if early stopping or certain regularization is applied.

**Margin maximization through RLMP**

**Definition**

Regularized Loss Minimization Problem (RLMP) with loss function  $L$  and parameter  $r > 0$  is:

$$\min_{\alpha} \mathcal{L}(\alpha) = \sum_{i=1}^m L\left(y_i \sum_{j=1}^N \alpha_j h_j(x_i)\right)$$

$$\text{such that } \sum_{j=1}^N \alpha_j \leq r, \quad \alpha_j \geq 0, \quad j = 1, 2, \dots, N.$$

**Existing theory [Rosset 2004]**

For margin maximizing loss functions, the solution to RLMP converges to a margin maximizing solution as regularization vanishes.

**Our contribution**

[Y. T. Xi et. al., submitted to AISTATS 2009]

Prove a convergence rate for the margin maximizing process.

**Theorem 1.** Assume  $L$  is convex, differentiable and  $L'(z) < 0$  for all  $z$ . Let  $\alpha^{(r)}$  be a solution of RLMP and  $h_{\alpha^{(r)}}$  have margin  $\mu(\alpha^{(r)})$ . Then

$$\frac{L'(r\theta^*)}{L'(r(\theta^* - \epsilon))} < \frac{\epsilon}{r(1 - \theta^*)} \Rightarrow \mu(\alpha^{(r)}) \geq \theta^* - \epsilon.$$

Thus, if we further assume that  $\forall \epsilon > 0$ ,  $\lim_{z \rightarrow \infty} L'(z)/L'(z(1 - \epsilon)) = 0$ , then for any sequence  $\{r_k\}$  that converges to  $\infty$ ,  $\lim_{k \rightarrow \infty} \mu(\alpha^{(r_k)}) = \theta^*$ .

Note that Theorem 1 provides an explicit rate of convergence of the margin to  $\theta^*$ .

For instance, for exponential loss, it shows that

$$\text{if } r > (1/\epsilon) \ln(m(1 - \theta^*)/\epsilon) \text{ then } \mu(\alpha^{(r)}) \geq \theta^* - \epsilon.$$

	AdaBoost	AdaBoost+L1	Rel. Improvement
Ringnorm	25.5%	25.3%	0.8%
Diabetes	26.7%	26.6%	0.4%
German	26.9%	26.7%	0.8%
Spam	11.3%	11.3%	0.2%
Ionosphere	12.5%	12.6%	-0.6%

Table 1: Comparison of error rates

	AdaBoost	AdaBoost+L1	Rel. Improvement
Ringnorm	121.8	55.0	54.8%
Diabetes	7.6	11.9	-56.6%
German	22.1	16.5	25.2%
Spam	31.3	26.5	15.3%
Ionosphere	26.7	19.6	26.8%

Table 2: Comparison of numbers of classifiers

## New algorithm: Adaboost + L1

Although RLMP is conceptually easy, implementing is time consuming.

Eg: Epsilon Boost, a variant of Adaboost

**AdaBoost+L1: combine the virtue of Adaboost with sparsity of L1 regularization**

**AdaBoost+L1**

1. Initialize: select  $\nu \in (0, 1]$ , set  $r_0 = 0$ ,  $\alpha_0 = \mathbf{0} \in \mathbb{R}^N$ ,  $U_0 = \emptyset$ , and  $d_0(i) = \frac{1}{m}, i = 1, \dots, m$ .

For  $t = 1, 2, \dots$

2. Find  $h_k \in \mathcal{H}$  such that  $e(h_k, \mathbf{d}_{t-1}) \geq \theta^*$ .

3. Update  $U$  and  $r$ :

$$U_t = U_{t-1} \cup \{k\}, \quad r_t = r_{t-1} + \frac{\nu}{2} \ln \frac{1 + e(h_k, \mathbf{d}_{t-1})}{1 - e(h_k, \mathbf{d}_{t-1})}$$

4. Solve the (small) convex minimization problem over  $\{\alpha_j\}_{j \in U_t}$ :

$$\min_{\alpha} \sum_{i=1}^m \exp\left(-y_i \sum_{j \in U_t} \alpha_j h_j(x_i)\right)$$

$$\text{s.t. } \sum_{j \in U_t} \alpha_j \leq r_t, \quad \alpha_j \geq 0, \quad \forall j \in U_t$$

5. Update the coefficients:

$$\alpha_t(j) = \begin{cases} \alpha_j & \text{if } j \in U_t; \\ 0 & \text{otherwise} \end{cases}$$

6. Update the distribution:

$$d_t(i) = \frac{\exp(-y_i \sum_{j=1}^N \alpha_t(j) h_j(x_i))}{\sum_{i=1}^m \exp(-y_i \sum_{j=1}^N \alpha_t(j) h_j(x_i))}, \quad i = 1, \dots, m.$$

Figure 1: The AdaBoost+L1 algorithm.

**Each iteration of Adaboost+L1:**

Step 2 and 3: perform first part of Standard Adaboost update

Step 4 (key): solves a small convex problem (L1 regularized loss minimization)

Step 5 and 6: update the coefficients and the distribution

**Theorem 2.** Let  $\alpha_t$  be the solution of AdaBoost+L1 after round  $t$  and set  $\tilde{\alpha}_t = \alpha_t / \|\alpha_t\|_1$ . Then  $\lim_{t \rightarrow \infty} \mu(\tilde{\alpha}_t) = \theta^*$ , and every limit point of  $\tilde{\alpha}_t$  is margin maximizing.

**Note:** Boost+L1 framework can be generalized to other loss functions such as Logistic loss, to achieve a maximum margin solution.

## Experiment Results

**Comparison between Adaboost and Adaboost+L1 on five datasets:**

- Breiman Ringnorm 20 dimension
- Diabetes (detecting diabetes based on medical information)
- German (determining credit based on financial histories)
- Spam (identifying spam email messages based on word frequencies)
- Ionosphere (classifying radar returns from the ionosphere)

We run experiments for 20 times with different training and testing examples, and average the results, shown in Figure 2. We use simple stumps as the basic classifiers.

From Table 1, we find Adaboost+L1 unnecessary if the only objective is low classification error. On the other hand, from Table 2, Adaboost+L1 reduces the number of active classifiers by a considerable percentage. Therefore, this algorithm is preferable if a sparse set of features is advantageous for other purposes.

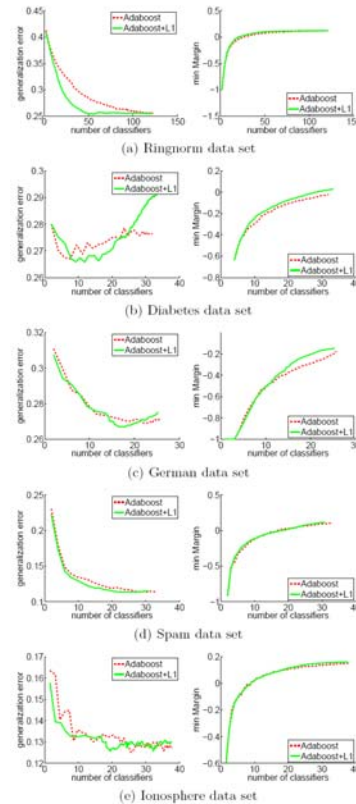


Figure 2: Experiment results on 5 data sets. (The left column shows the error rate on testing data, the right column shows the minimal margin on training data)

## Conclusion

1. For RLMP, we have obtained a quantitative relationship between the regularization parameter and the achievable margin.

2. We derive a new efficient algorithm Adaboost+L1, which provably achieves the maximum margin in the limit. The algorithm is able to yield a much sparser composite classifier with same or better generalization performance.

Figure 3: The frequencies of different features being selected, Spam Data Set

