Journal of Machine Learning Research () Submitted ; Published

A Refined Margin Analysis for Boosting Algorithms via Equilibrium Margin

Liwei Wang WANGLW@CIS.PKU.EDU.CN
Key Laboratory of Machine Perception, MOE
School of Electronics Engineering and Computer Science
Peking University
Beijing, 100871, P.R.China

Masashi Sugiyama SUGI@CS.TITECH.AC.JP
Department of Computer Science
Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan

Zhaoxiang Jing JINGZX@CIS.PKU.EDU.CN
Key Laboratory of Machine Perception, MOE
School of Electronics Engineering and Computer Science
Peking University
Beijing, 100871, P.R.China

Cheng Yang YANGCHENUG@GMAIL.COM
Beijing Aerospace Control Center
Beijing, 100094, P.R.China

Zhi-Hua Zhou ZHOUZH@NJU.EDU.CN
National Key Laboratory for Novel Software Technology
Nanjing University
Nanjing 210093, P.R. China

Jufu Feng FJF@CIS.PKU.EDU.CN
Key Laboratory of Machine Perception, MOE
School of Electronics Engineering and Computer Science
Peking University
Beijing, 100871, P.R.China

Editor:

Abstract

Much attention has been paid to the theoretical explanation of the empirical success of AdaBoost. The most influential work is the margin theory, which is essentially an upper bound for the generalization error of any voting classifier in terms of the margin distribution over the training data. However, important questions were raised about the margin explanation. Breiman (1999) proved a bound in terms of the minimum margin, which is sharper than the margin distribution bound. He argued that the minimum margin would be better in predicting the generalization error. Grove and Schuurmans (1998) developed an algorithm called LP-AdaBoost which maximizes the minimum margin while keeping all other factors the same as AdaBoost. In experiments however, LP-AdaBoost usually performs worse than AdaBoost, putting the margin explanation into serious doubt. In this paper, we make a refined analysis of the margin theory. We prove a bound in terms of

a new margin measure called the *Equilibrium margin (Emargin)*. The Emargin bound is uniformly sharper than Breiman's minimum margin bound. Thus our result suggests that the minimum margin may be not crucial for the generalization error. We also show that a large Emargin and a small empirical error at Emargin imply a smaller bound of the generalization error. Experimental results on benchmark datasets demonstrate that AdaBoost usually has a larger Emargin and a smaller test error than LP-AdaBoost, which agrees well with our theory.

1. Introduction

The AdaBoost algorithm (Freund and Schapire, 1996, 1997) has achieved great success in the past ten years. It has demonstrated excellent experimental performance both on benchmark datasets and real applications (Bauer and Kohavi, 1999; Dietterich, 2000; Viola and Jones, 2001; Wang et al., 2007). According to a recent evaluation (Caruana and Niculescu-Mizil, 2006), boosting with decision trees as base learners is the leading classification algorithm. An important property of boosting is its relative (although not complete) resistance to overfitting. On many datasets it is observed that the test error keeps decreasing even after thousands of base classifiers have been combined (Breiman, 1998; Quinlan, 1996). This fact, at first sight, obviously violates Occam's razor.

Considerable efforts have been made to explain the "mystery" of boosting. Friedman et al. (2000) related boosting to fitting an additive logistic regression model. From this statistical view they developed the LogitBoost algorithm. Jiang (2004), Lugosi and Vayatis (2004), Zhang (2004), Bartlett et al. (2006) and others proved that boosting is Bayes consistent if it is properly regularized. These works provide deep understanding of boosting. However, these explanations each focused on some aspects of boosting. The consistency assures that boosting is asymptotically optimal, but it does not explain boosting's effectiveness on small sample problems. The statistical view led to many new algorithms, but left boosting's relative resistance to overfitting not well explained. Boosting algorithms involve several factors such as the type of base classifiers, regularization methods and loss functions to minimize. Recently, Mease and Wyner (2008) studied the effects of these factors. They provided a number of examples that are contrary to previous theoretical explanations.

Schapire et al. (1998) tried to give a comprehensive explanation in terms of the margins of the training examples. Roughly speaking, the margin of an example with respect to a classifier is a measure of the confidence of the classification result. They proved an upper bound for the generalization error of a voting classifier that does not depend on how many classifiers were combined, but only on the margin distribution over the training set, the number of the training examples and the size (the VC dimension for example) of the set of base classifiers. They also demonstrated that AdaBoost has the ability to produce a "good" margin distribution. This theory suggests that producing a good margin distribution is the key to the success of AdaBoost and explains well its relative resistance to overfitting.

Soon after that however, there were serious doubt cast on this margin explanation. First Breiman (1999) and Grove and Schuurmans (1998) developed algorithms that maximize the *minimum margin*. (Minimum margin is the smallest margin over all training examples, see Section 2 for the formal definition). Breiman (1999) then gave an upper bound for the generalization error of a voting classifier in terms of the minimum margin, as well as the number of training examples and the size of the set of base classifiers. This bound is sharper than

the bound based on the margin distribution given in Schapire et al. (1998). Breiman (1999) argued that if the bound of Schapire et al. implied that the margin distribution is important to the generalization error, his bound implied more strongly that the minimum margin is the key to the generalization error, and the minimum margin maximizing algorithms would achieve better performance than AdaBoost.

Grove and Schuurmans (1998) conducted a rigorous experimental comparison on the minimum margin. They developed an algorithm called LP-AdaBoost which first uses AdaBoost to train a series of base classifiers. Then by linear programming they obtained coefficients of the base classifiers, whose linear combination has the largest possible minimum margin. Thus LP-AdaBoost and AdaBoost have all relevant factors the same except the coefficients of the base classifiers. According to the minimum margin bound, LP-AdaBoost would have smaller generalization error than AdaBoost. In experiments, although LP-AdaBoost always achieves larger minimum margins, its test error is higher than AdaBoost on most datasets. This result puts the margin theory into serious doubt.

In this paper we provide a refined analysis of the margin theory. We propose a new upper bound for the generalization error of voting classifiers. This bound is uniformly sharper than Breiman's minimum margin bound. The key factor in this bound is a new margin notion which we refer to as the *Equilibrium margin (Emargin)*. The Emargin can be viewed as a measure of how good a margin distribution is. In fact, the Emargin depends, in a complicated way, on the margin distribution, and has little relation to the minimum margin. Experimental results show that AdaBoost usually produces a larger Emargin than LP-AdaBoost, which agrees with the Emargin explanation.

The margin theory has been studied and greatly improved by several authors. Especially Koltchinskii and Panchenko (2002, 2005) developed new tools for empirical processes and prove much sharper margin distribution bounds. However it is difficult to compare these bounds to the minimum margin bound of Breiman (1999), since they contain unspecified constants. Nevertheless, these results suggest that the margin distribution may be more important than the minimum margin for the generalization error of voting classifiers.

We also show that if a boosting algorithm returns a classifier that minimizes the Emargin bound or the margin distribution bound of Schapire et al. (1998) then the classifier learned converges to the best classifier in the hypothesis space as the number of training examples goes to infinity.

The rest of this paper is organized as follows: In Section 2 we briefly describe the background of the margin theory. Our main results—the Emargin bounds are given in Section 3. We provide further explanation of the main bound in Section 4 and the consistency results in Section 5. All the proofs can be found in Section 6. We provide experimental justification in Section 7 and conclude in Section 8.

2. Background

Consider binary classification problems. Examples are drawn independently according to an underlying distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$, where $\mathcal{X}$ is an instance space. Let $\mathcal{H}$ denote the space from which the base hypotheses are chosen. A base hypothesis $h \in \mathcal{H}$ is a

mapping from $\mathcal{X}$ to $\{-1, +1\}$. A voting classifier $f(x)$ is of the form

$$f(x) = \sum_i \alpha_i h_i(x), \quad h_i \in \mathcal{H},$$

where

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

An error occurs on an example (x, y) if and only if

$$yf(x) \leq 0.$$

We use $P_{\mathcal{D}}(A(x, y))$ to denote the probability of the event A when an example (x, y) is chosen randomly according to the distribution $\mathcal{D}$. Therefore, $P_{\mathcal{D}}(yf(x) \leq 0)$ is the generalization error of f which we want to bound. Let $\mathcal{S}$ be a training set containing n examples. We use $P_{\mathcal{S}}(A(x, y))$ to denote the probability with respect to choosing an example (x, y) uniformly at random from $\mathcal{S}$.

For an example (x, y), the value of $yf(x)$ reflects the confidence of the prediction. Since each base classifier outputs -1 or $+1$, one has

$$yf(x) = \sum_{i:y=h_i(x)} \alpha_i - \sum_{i:y\neq h_i(x)} \alpha_i.$$

Hence $yf(x)$ is the difference between the weights assigned to those base classifiers that correctly classify (x, y) and the weights assigned to those that misclassify the example. $yf(x)$ is called the *margin* for (x, y) with respect to f. If we consider the margins over the whole set of training examples, we can regard $P_{\mathcal{S}}(yf(x) \leq \theta)$ as a distribution over θ $(-1 \leq \theta \leq 1)$, since $P_{\mathcal{S}}(yf(x) \leq \theta)$ is the fraction of training examples whose margin is at most θ. This distribution is referred to as the *margin distribution*.

A description of AdaBoost is shown in Algorithm 1. In AdaBoost the linear coefficients α_t is set as

$$\alpha_t = \frac{1}{2} \log \frac{1 + \gamma_t}{1 - \gamma_t},$$

where γ_t is defined as:

$$\gamma_t = \sum_{i=1}^{n} D_t(i) y_i h_t(x_i).$$

γ_t is an affine transformation of the error rate of h_t with respect to the weight distribution D_t.

AdaBoost often does not overfit. Although it is known that boosting forever does overfit when there is high classification noise, on many datasets the performance of AdaBoost keeps improving even after a large number of rounds.

The first margin explanation (Schapire et al., 1998) of the AdaBoost algorithm is to upper bound the generalization error of voting classifiers in terms of the margin distribution, the number of training examples and the complexity of the set from which the base classifiers are chosen. The theory contains two bounds: one applies to the case that the base classifier set $\mathcal{H}$ is finite, and the other applies to the general case that $\mathcal{H}$ has a finite VC dimension.

Input: T, $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
where $x_i \in \mathcal{X}$, $y_i \in \{-1, 1\}$.
Initialization: $D_1(i) = 1/n$.
for $t = 1$ **to** T **do**

1. Train a base classifier $h_t \in \mathcal{H}$ using distribution D_t, where $h_t : \mathcal{X} \to \{-1, 1\}$.
2. Choose α_t.
3. Update:

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

where Z_t is the normalization factor chosen so that D_{t+1} will be a distribution.
end
Output: The final classifier

$$F(x) = \operatorname{sgn}\left(f(x)\right),$$

where

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x).$$

Algorithm 1: A description of AdaBoost.

Theorem 1 *(Schapire et al., 1998) For any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of the training set $\mathcal{S}$ of n examples, every voting classifier f satisfies the following bounds:*

$$P_{\mathcal{D}}\Big(yf(x) \le 0\Big) \le \inf_{\theta \in (0,1]} \left[P_{\mathcal{S}}\Big(yf(x) \le \theta\Big) + O\left(\frac{1}{\sqrt{n}} \left(\frac{\log n \log |\mathcal{H}|}{\theta^2} + \log \frac{1}{\delta}\right)^{1/2}\right)\right],$$

if $|\mathcal{H}| < \infty$. And

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

where d is the VC dimension of $\mathcal{H}$.

The theorem states that if the voting classifier generates a good margin distribution, that is, most training examples have large margins so that $P_{\mathcal{S}}(yf(x) \le \theta)$ is small for not too small θ, then the upper bound of the generalization error is also small. In Schapire et al. (1998) it has also been shown that for the AdaBoost algorithm, $P_{\mathcal{S}}(yf(x) \le \theta)$ decreases to zero exponentially fast with respect to the number of boosting iterations if θ is not too large. These results suggest that the excellent performance of AdaBoost is due to its good margin distribution.

Another important notion is the minimum margin which is the smallest margin achieved on the training set. Formally, the minimum margin, denoted by θ_0, of a voting classifier f on a training set $\mathcal{S}$ is defined as

$$\theta_0 = \min\left\{yf(x) : \ (x, y) \in \mathcal{S}\right\}. \tag{1}$$

Breiman (1999) proved an upper bound for the generalization error of voting classifiers which depends only on the minimum margin, not on the entire margin distribution.

Theorem 2 *(Breiman, 1999) Assume that $|\mathcal{H}| < \infty$. Let θ_0 be a real number that satisfies $\theta_0 > 4\sqrt{\frac{2}{|\mathcal{H}|}}$ and*

$$R = \frac{32\log(2|\mathcal{H}|)}{n\theta_0^2} \leq 2n.$$

Then for any $\delta > 0$, with probability at least $1-\delta$ over the random choice of the training set $\mathcal{S}$ of n examples, every voting classifier f whose minimum margin on $\mathcal{S}$ is at least θ_0 satisfies the following bound:

$$P_{\mathcal{D}}\Big(yf(x) \leq 0\Big) \leq R\left(\log(2n) + \log\frac{1}{R} + 1\right) + \frac{1}{n}\log\left(\frac{|\mathcal{H}|}{\delta}\right). \tag{2}$$

Breiman (1999) pointed out that his bound is sharper than the margin distribution bound of Schapire et al. If θ in Theorem 1 is taken to be the minimum margin θ_0, the bound in Theorem 2 is about the square of the bound in terms of the margin distribution, since the bound in Theorem 2 is $O\left(\frac{\log n}{n\theta_0^2}\right)$ and the bound in Theorem 1 is $O\Big(\sqrt{\frac{\log n}{n\theta_0^2}}\Big)$. Breiman then argued that compared to the margin distribution explanation, his bound implied more strongly that the minimum margin governs the generalization error.

Several authors developed algorithms to maximize the minimum margin. Among these, the most representative one is the LP-AdaBoost proposed by Grove and Schuurmans (1998). Let $h_1, \ldots, h_T$ be the base classifiers returned by AdaBoost on the training examples $\{(x_i, y_i),\ i = 1, \ldots, n\}$. Finding a voting classifier $g = \sum_{t=1}^{T}\beta_t h_t$ such that g maximizes the minimum margin can be formulated as a linear programming problem.

$$\begin{aligned} \max_{\beta, m}\ & m \\ s.t.\ & y_i\sum_{t=1}^{T}\beta_t h_t(x_i) \geq m, \quad i = 1, 2, \ldots, n \\ & \beta_t \geq 0, \quad \sum_{t=1}^{T}\beta_t = 1, \end{aligned}$$

where $\beta = (\beta_1, \cdots, \beta_T)$. Grove and Schuurmans called this algorithm LP-AdaBoost.

Comparing the performance of AdaBoost and LP-AdaBoost is a good test of significance of the minimum margin bound. Except the linear coefficients, the voting classifiers obtained by the two algorithms have all relevant factors the same. In experiments, although LP-AdaBoost always produces larger minimum margins, its test error is higher than AdaBoost more often than not. This result is different from what the minimum margin bound suggests and therefore puts the margin explanation into serious doubt.

Breiman (1999) and Meir and Rätsch (2003) developed arc-gv to maximize the minimum margin. Arc-gv can also be described by Algorithm 1. The only difference from AdaBoost is how to set α_t at each round. It can be shown that arc-gv converges to the maximum margin solution (Rätsch and Warmuth, 2005; Rudin et al., 2007) whereas AdaBoost does

not always do this (Rudin et al., 2004). However on some datasets AdaBoost has larger minimum margin than arc-gv after a finite number of rounds. Also note that arc-gv and AdaBoost generate different base classifiers. Recently Reyzin and Schapire (2006) gained an important discovery that when Breiman (1999) tried to maximize the minimum margin by arc-gv, he had not make a good control of the complexity of the base classifiers, while comparing the margin is only meaningful when the complexity of base learners are the same.

3. Emargin Bounds

In this section we propose upper bounds in terms of the Emargin. The bounds are sharper than the minimum margin bound.

First let us introduce some notions. Consider the Bernoulli relative entropy function $D(q||p)$ defined as

$$D(q||p) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}, \quad 0 \le p, q \le 1.$$

By convention, let $D(0||0) = 0$.

For a fixed q, $D(q||p)$ is a monotone increasing function of p for $q \le p \le 1$. It is easy to check that

$$D(q||p) = 0 \quad \text{when } p = q,$$

and

$$D(q||p) \to \infty \quad \text{as } p \to 1.$$

Thus one can define the inverse function of $D(q||p)$ for fixed q as $D^{-1}(q,u)$, such that

$$D\left(q||D^{-1}(q,u)\right) = u \quad \text{for all } u \ge 0 \text{ and } D^{-1}(q,u) \ge q.$$

See also Langford (2005).

The next theorem is our main result: the Emargin bound. Here we consider the case that the base classifier set $\mathcal{H}$ is finite. For the case that $\mathcal{H}$ is infinite but has a finite VC dimension, the bound is more complicated and will be given in Theorem 7. All the proofs can be found in Section 6.

Theorem 3 *If $8 < |\mathcal{H}| < \infty$, then for any $\delta > 0$, with probability at least $1-\delta$ over the random choice of the training set $\mathcal{S}$ of n examples ($n > 1$), every voting classifier f such that*

$$q_0 = P_{\mathcal{S}}\left(yf(x) \le \sqrt{\frac{8}{|\mathcal{H}|}}\right) < 1.$$

satisfies the following bound:

$$P_{\mathcal{D}}\Big(yf(x) \le 0\Big) \le \frac{\log|\mathcal{H}|}{n} + \inf_{q \in \{q_0, q_0+\frac{1}{n}, \ldots, \frac{n-1}{n}\}} D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right), \tag{3}$$

where

$$\hat{\theta}(q) = \sup\left\{\theta \in (0,1] : P_{\mathcal{S}}\Big(yf(x) \le \theta\Big) \le q\right\}, \tag{4}$$

$$u(\theta) = \frac{1}{n}\left(\frac{8}{\theta^2}\log\left(\frac{2n^2}{\log|\mathcal{H}|}\right)\log(2|\mathcal{H}|) + 2\log|\mathcal{H}| + \log\frac{n}{\delta}\right).$$

Note that the assumption $q_0 < 1$ in the theorem is very mild since it implies that at least one training example has a large margin (larger than $8/|\mathcal{H}|$), or equivalently the *largest* margin is not too small[1]. This contrasts with the fact that the minimum margin bound applies when the *minimum* margin is not too small.

Clearly the key factors in this bound are the optimal q and the corresponding $\hat{\theta}(q)$.

Definition 4 *Let q^* be the optimal q in Eq.(3), and denote*

$$\theta^* = \hat{\theta}(q^*).$$

*We call θ^* the Equilibrium margin (**Emargin**). It can be seen that q^* is the empirical error at margin θ^*, i.e.,*

$$q^* = P_{\mathcal{S}}(yf(x) < \theta^*).$$

q^* will be referred to as the Emargin error.

With Definition 4, the Emargin bound (3) can be simply written as

$$P_{\mathcal{D}}\Big(yf(x) \le 0\Big) \le \frac{\log|\mathcal{H}|}{n} + D^{-1}\Big(q^*, u\left(\theta^*\right)\Big). \tag{5}$$

Theorem 3 provides an upper bound of the generalization error of a voting classifier that depends on its Emargin and the Emargin error.

Our Emargin bound has a similar flavor to Theorem 1. Note that the Emargin depends, in a complicated way, on the whole margin distribution. Roughly, if most training examples have large margins, then θ^* is large and q^* is small. The minimum margin is only a special case of the Emargin. From (4) one can see that $\hat{\theta}(0)$ is the minimum margin. Hence the Emargin is equal to the minimum margin if and only if the optimal q^* is zero.

We next compare our Emargin bound to the minimum margin bound. We show that the Emargin bound is sharper than the minimum margin bound. Since the minimum margin bound applies only to the separable case, i.e., $\theta_0 > 0$, we assume that the conditions in Theorem 2 are satisfied.

Theorem 5 *Assume that the minimum margin θ_0 is larger than 0. Then the bound given in Theorem 3 is uniformly sharper than the minimum margin bound in Theorem 2. That is, if*

$$R = \frac{32\log(2|\mathcal{H}|)}{n\theta_0^2} \le 2n,$$

then

$$\frac{\log|\mathcal{H}|}{n} + D^{-1}\Big(q^*, u\left(\theta^*\right)\Big) \le R\left(\log(2n) + \log\frac{1}{R} + 1\right) + \frac{1}{n}\log\frac{|\mathcal{H}|}{\delta}.$$

1. This observation is due to a reviewer.

This theorem suggests that the Emargin and Emargin error may be more relevant to the generalization error than the minimum margin. The following theorem describes how the Emargin θ^* and the Emargin error q^* affect the upper bound of the generalization ability. It states that a larger Emargin and a smaller Emargin error result in a lower generalization error bound.

Theorem 6 *Let f_1, f_2 be two voting classifiers. Denote by θ_1, θ_2 the Emargins and by q_1, q_2 the Emargin errors of f_1, f_2 respectively. Thus*

$$q_i = P_{\mathcal{S}}\Big(yf_i(x) < \theta_i\Big), \qquad i = 1, 2.$$

Also denote by B_1, B_2 the Emargin upper bounds of the generalization error of f_1, f_2 (i.e., the right-hand side of (3)). Then

$$B_1 \leq B_2,$$

if

$$\theta_1 \geq \theta_2 \ \ \textit{and} \ \ q_1 \leq q_2.$$

Theorem 6 suggests that the Emargin and the Emargin error can be used as measures of the quality of a margin distribution. A large Emargin and a small Emargin error indicate a good margin distribution. Experimental results in Section 7 show that AdaBoost often has larger Emargins and smaller Emargin errors than LP-AdaBoost.

The last theorem of this section is the Emargin bound for the case that the set of base classifiers has a finite VC dimension.

Theorem 7 *Suppose the set of base classifiers $\mathcal{H}$ has VC dimension d. Then for any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of the training set $\mathcal{S}$ of n examples, every voting classifier f satisfies the following bound:*

$$P_{\mathcal{D}}\Big(yf(x) \leq 0\Big) \leq \frac{d^2+1}{n} + \inf_{q \in \{q_0, q_0 + \frac{1}{n}, \ldots, \frac{n-1}{n}\}} \frac{n-1}{n} \cdot D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right), \tag{6}$$

where

$$\hat{\theta}(q) = \sup\left\{\theta \in \big(0, 1\big] : P_{\mathcal{S}}\Big(yf(x) \leq \theta\Big) \leq q\right\}, \tag{7}$$

and

$$u\left(\theta\right) = \frac{1}{n}\left(\frac{16d}{\theta^2}\log\frac{n}{d}\log\frac{en^2}{d} + 3\log\left(\frac{16}{\theta^2}\log\frac{n}{d} + 1\right) + \log\frac{2n}{\delta}\right),$$

provided $q_0 = P_{\mathcal{S}}(yf(x) \leq 0) < 1$.

4. Explanation of the Emargin Bound

In Theorem 3, we adopted the partial inverse of the relative entropy to upper bound the generalization error. The key term in the Emargin bound is $\inf_q D^{-1}(q, u[\hat{\theta}(q)])$. To better understand the bound, we make use of three different upper bounds of $\inf_q D^{-1}(q, u[\hat{\theta}(q)])$ to obtain simpler forms and give explanations of the Emargin bound. We list in the following lemma the upper bounds of $\inf_q D^{-1}(q, u[\hat{\theta}(q)])$.

Lemma 8 *Let $u[\hat{\theta}(q)]$ be the one defined in Theorem 3. Let $\Gamma = \{q_0, q_0 + \frac{1}{n}, \ldots, \frac{n-1}{n}\}$, where q_0 was defined in Theorem 3. Then the following bounds hold. (In the first bound we assume that $q_0 = 0$.)*

$$\inf_{q\in\Gamma} D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right) \le D^{-1}\left(0, u\left[\hat{\theta}(0)\right]\right) \le u\left[\hat{\theta}(0)\right]. \tag{8}$$

$$\inf_{q\in\Gamma} D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right) \le \inf_{q\in\Gamma}\left(q + \left(\frac{u\left[\hat{\theta}(q)\right]}{2}\right)^{1/2}\right). \tag{9}$$

$$\inf_{q\in\Gamma} D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right) \le \inf_{q\in\Gamma,\ q\le Cu[\hat{\theta}(q)]} D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right) \le \inf_{q\in\Gamma,\ q\le Cu[\hat{\theta}(q)]} C'u[\hat{\theta}(q)], \tag{10}$$

where C is any constant such that there exists q such that $q \le Cu[\hat{\theta}(q)]$. Here $C' = \max(2C, 8)$.

Note from Theorem 3 that

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

and

$$q = P_{\mathcal{S}}\left(yf(x) < \hat{\theta}(q)\right).$$

Thus we can derive the following three bounds of the generalization error from the Emargin bound by using the three inequalities in Lemma 8 respectively.

Corollary 9 *If $8 < |\mathcal{H}| < \infty$, then for any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of the training set $\mathcal{S}$ of n examples ($n > 1$), every voting classifier $f \in C(\mathcal{H})$ such that $q_0 < 1$ satisfies the following bounds:*

1.

$$P_{\mathcal{D}}(yf(x) \le 0) \le O\left(\frac{1}{n}\left(\frac{\log n \log|\mathcal{H}|}{\theta_0^2} + \log\frac{1}{\delta}\right)\right). \tag{11}$$

Here we assume $\theta_0 > \sqrt{8/|\mathcal{H}|}$ is the minimum margin.

2.

$$P_{\mathcal{D}}\Big(yf(x) \le 0\Big) \le \inf_{\theta \in [\frac{8}{|\mathcal{H}|}, 1]} \left[P_{\mathcal{S}}\Big(yf(x) \le \theta\Big) + O\left(\frac{1}{\sqrt{n}} \left(\frac{\log n \log |\mathcal{H}|}{\theta^2} + \log \frac{1}{\delta} \right)^{1/2} \right) \right]. \tag{12}$$

3. For any constant C and $\theta \in [\sqrt{8/|\mathcal{H}|}, 1)$ such that

$$P_{\mathcal{S}}(yf(x) \le \theta) \le \frac{C}{n} \left(\frac{8}{\theta^2} \log \left(\frac{2n^2}{\log |\mathcal{H}|} \right) \log(2|\mathcal{H}| + \log |\mathcal{H}| + \log \frac{n}{\delta}) \right), \tag{13}$$

we have

$$P_{\mathcal{D}}(yf(x) \le \theta) \le \frac{\log |\mathcal{H}|}{n} + \frac{C'}{n} \left(\frac{8}{\theta^2} \log \left(\frac{2n^2}{\log |\mathcal{H}|} \right) \log(2|\mathcal{H}| + \log |\mathcal{H}| + \log \frac{n}{\delta}) \right), \tag{14}$$

where $C' = \max(2C, 8)$.

The first bound in the corollary has the same order as the minimum margin bound. The second bound is essentially the same as Theorem 1 except that θ cannot be too small. So previous bounds can be derived from the Emargin bound. The third bound states that the generalization error is $O\left(\frac{\log n \log |\mathcal{H}|}{n\theta^2}\right)$ even in the non-zero error case, provided the margin error $P_{\mathcal{S}}(yf(x) \le \theta)$ is small enough.

The third bound has a much simpler form than Theorem 3. If we use this bound to define Emargin, i.e., the optimal θ in the bound, it can be greatly simplified. It is easy to see that the optimal θ is just the largest θ satisfying (13). The price however is that this approximate bound is not uniformly sharper than the minimum margin bound.

5. Consistency

So far the results are finite sample generalization error bounds. In this section we point out that the Emargin bound and the margin distribution bound in Theorem 1 imply statistical consistency. In particular we show that if a boosting algorithm minimizes the bound, then the classifier learned converges to the optimal classifier in the hypothesis space, i.e., the convex hull of the base classifiers. Here we assume that the set of base classifiers $\mathcal{H}$ is symmetric. That is, if $h \in \mathcal{H}$ then $-h \in \mathcal{H}$. Therefore the best classifier in the convex hull of $\mathcal{H}$ is also the best classifier in the linear span of $\mathcal{H}$. An immediate consequence of this consistency is that margin bound optimization is Bayes consistent if the linear span of the base classifiers is dense in the space of all measurable functions. A typical example of such base classifiers is decision tree with the number of leaves larger than the dimension of the input space (Breiman, 2004).

Before stating the consistency theorem, we need some notions. Let $C(\mathcal{H})$ be the convex hull of the set of base classifiers. Also let

$$L^* = \inf_{f \in C(\mathcal{H})} P_{\mathcal{D}}\Big(yf(x) \le 0\Big).$$

That is, L^* is the minimal generalization error of the classifiers in $C(\mathcal{H})$.

We consider an algorithm that optimizes the Emargin: Given a training set $\mathcal{S}$ containing n examples, the learning algorithm returns a function $\hat{f}_n \in C(\mathcal{H})$ which minimizes the finite VC dimension Emargin bound (i.e., the right-hand side of (6)), or simply $D^{-1}\big(q^*, u(\theta^*)\big)$.

The next theorem states that margin bound optimization is consistent. With almost the same arguments one can show that minimizing the margin distribution bound in Theorem 1 is also consistent. But there is no such result for the minimum margin bound for the non-separable problems.

Theorem 10 *Let $C(\mathcal{H})$, L^* and $\hat{f}_n$ be defined as above. Then*

$$\lim_{n\to\infty} EP_{\mathcal{D}}\left(y\hat{f}_n(x) \le 0\right) = L^*,$$

where E is the expectation over the random draw of the training set $\mathcal{S}_n$.

6. Proofs

In this section, we give proofs of the theorems, lemmas and corollaries.

6.1 Proof of Theorem 3

The proof uses the tool developed in Schapire et al. (1998). The difference is that we do not bound the deviation of the generalization error from the empirical margin error directly, instead we consider the difference of the generalization error to a zero-one function of a certain empirical measure. This allows us to unify the zero-error and nonzero-error cases and it results in a sharper bound. For the sake of convenience, we follow the convention in Schapire et al. (1998).

Let $C(\mathcal{H})$ denote the convex hull of $\mathcal{H}$. Also let $C_N(\mathcal{H})$ denote the set of unweighted averages over N elements from the base classifier set $\mathcal{H}$. Formally,

$$C_N(\mathcal{H}) = \left\{ g : \ g = \frac{1}{N}\sum_{j=1}^{N} h_j, \ h_j \in \mathcal{H} \right\}.$$

Any voting classifier

$$f = \sum \beta_i h_i \in C(\mathcal{H}),$$

where

$$\sum \beta_i = 1, \ \ \beta_i \ge 0,$$

can be associated with a distribution over $\mathcal{H}$ by the coefficients $\{\beta_i\}$. We denote this distribution as $Q(f)$. By choosing N elements independently and randomly from $\mathcal{H}$ according to $Q(f)$, we can generate a classifier $g \in C_N(\mathcal{H})$. The distribution of g is denoted by $Q_N(f)$. For any fixed α $(0 < \alpha < 1)$

$$\begin{aligned} P_{\mathcal{D}}\Big(yf(x) \le 0\Big) &\le P_{\mathcal{D},g\sim Q_N(f)}\Big(yg(x) \le \alpha\Big) + P_{\mathcal{D},g\sim Q_N(f)}\Big(yg(x) > \alpha, \ yf(x) \le 0\Big) \\ &\le P_{\mathcal{D},g\sim Q_N(f)}\Big(yg(x) \le \alpha\Big) + \exp\left(-\frac{N\alpha^2}{2}\right). \end{aligned} \tag{15}$$

We next bound the first term on the right-hand side of the inequality. So far the argument is the same as Schapire et al. (1998). From now on we use some different techniques. For any fixed $g \in C_N(\mathcal{H})$, and for any positive number ε and nonnegative integer k such that $k \leq n\varepsilon$, we consider the probability (over the random draw of n training examples) that the training error at margin α is less than k/n, while the true error of g at margin α is larger than ε:

$$\Pr_{\mathcal{S}\sim\mathcal{D}^n}\left(P_{\mathcal{S}}\left(yg(x) \leq \alpha\right) \leq \frac{k}{n},\ P_{\mathcal{D}}\left(yg(x) \leq \alpha\right) > \epsilon\right). \tag{16}$$

Here $\Pr_{S\sim\mathcal{D}^n}$ denotes the probability over n training examples chosen independently at random according to $\mathcal{D}$. Note that the proof in Schapire et al. (1998) considers only the difference of $P_{\mathcal{D}}\left(yg(x) \leq \alpha\right)$ and $P_{\mathcal{S}}\left(yg(x) \leq \alpha\right)$, i.e., $P_{\mathcal{D}}\left(yg(x) \leq \alpha\right) - P_{\mathcal{S}}\left(yg(x) \leq \alpha\right)$; While here we consider the values of $P_{\mathcal{D}}\left(yg(x) \leq \alpha\right)$ and $P_{\mathcal{S}}\left(yg(x) \leq \alpha\right)$ themselves. The benefit is that this allows us to use the tightest version of Chernoff bound—the relative entropy Chernoff bound—rather than the relatively looser additive Chernoff bound. To derive the bound, we write (16) in the following equivalent form.

$$\Pr_{\mathcal{S}\sim\mathcal{D}^n}\left(P_{\mathcal{D}}(yg(x) \leq \alpha) > I\left[P_{\mathcal{S}}(yg(x) \leq \alpha) > \frac{k}{n}\right] + \varepsilon\right), \tag{17}$$

where I is the indicator function. (17) is important in our proof. It bounds the difference of the true and empirical margin distributions as α and k vary over their ranges. But k and α can take essentially finite number of values, so we can use union bounds. It's easy to see that no matter $P_{\mathcal{D}}(yg(x) \leq \alpha) > \varepsilon$ or $P_{\mathcal{D}}(yg(x) \leq \alpha) \leq \varepsilon$, we have the following inequality (In the former case, it is the tail bound for Bernoulli trials; and in the latter case the probability is actually zero).

$$\Pr_{\mathcal{S}\sim\mathcal{D}^n}\left(P_{\mathcal{D}}\Big(yg(x) \leq \alpha\Big) > I\left[P_{\mathcal{S}}\Big(yg(x) \leq \alpha\Big) > \frac{k}{n}\right] + \varepsilon\right) \leq \quad \sum_{r=0}^{k}\binom{n}{r}\varepsilon^r(1-\varepsilon)^{n-r}.$$

Then applying the relative entropy Chernoff bound (Hoeffding, 1963) to the Bernoulli trials, we further have

$$\sum_{r=0}^{k}\binom{n}{r}\varepsilon^r(1-\varepsilon)^{n-r} \leq \exp\left(-nD\left(\frac{k}{n}\middle\|\varepsilon\right)\right).$$

We thus obtain

$$\Pr_{\mathcal{S}\sim\mathcal{D}^n}\left(P_{\mathcal{D}}\Big(yg(x) \leq \alpha\Big) > I\left[P_{\mathcal{S}}\Big(yg(x) \leq \alpha\Big) > \frac{k}{n}\right] + \varepsilon\right) \leq \exp\left(-nD\left(\frac{k}{n}\middle\|\varepsilon\right)\right). \tag{18}$$

We only consider α at the values in the set

$$U = \left\{\frac{1}{|\mathcal{H}|}, \frac{2}{|\mathcal{H}|}, \ldots, 1\right\}.$$

There are no more than $|\mathcal{H}|^N$ elements in $C_N(\mathcal{H})$. Using the union bound we get

$$\begin{aligned}\Pr_{\mathcal{S}\sim\mathcal{D}^n}\Bigg(&\exists g \in C_N(\mathcal{H}),\ \exists\alpha \in U,\ P_{\mathcal{D}}\Big(yg(x) \leq \alpha\Big) > I\left[P_{\mathcal{S}}\Big(yg(x) \leq \alpha\Big) > \frac{k}{n}\right] + \varepsilon\Bigg) \\ &\leq |\mathcal{H}|^{(N+1)}\exp\left(-nD\left(\frac{k}{n}\middle\|\varepsilon\right)\right).\end{aligned} \tag{19}$$

The above formula upper bounds the probability that "$\exists g \in C_N(\mathcal{H})$" certain inequality of g holds. The bound also applies to "$\exists$ a distribution of g over $C_N(\mathcal{H})$" such that the inequality of the *expectation* over g holds, since the latter implies the former. Note that

$$E_{g\sim Q_N(f)} P_{\mathcal{D}}\Big(yg(x) \le \alpha\Big) = P_{\mathcal{D},g\sim Q_N(f)}\Big(yg(x) \le \alpha\Big),$$
$$E_{g\sim Q_N(f)} I\left[P_{\mathcal{S}}\Big(yg(x) \le \alpha\Big) > \frac{k}{n}\right] = P_{g\sim Q_N(f)}\left(P_{\mathcal{S}}\Big(yg(x) \le \alpha\Big) > \frac{k}{n}\right).$$

We thus have

$$\Pr_{S\sim D^n}\left(\exists f \in C(\mathcal{H}), \exists \alpha \in U, P_{D,g\sim Q_N(f)}\Big(yg(x) \le \alpha\Big) > P_{g\sim Q_N(f)}\left(P_{\mathcal{S}}(yg(x) \le \alpha) > \frac{k}{n}\right) + \varepsilon\right)$$
$$\le |\mathcal{H}|^{(N+1)} \exp\left(-nD\left(\frac{k}{n}\middle\|\varepsilon\right)\right).$$

Let

$$\delta = |\mathcal{H}|^{(N+1)} \exp\left(-nD\left(\frac{k}{n}\middle\|\varepsilon\right)\right),$$

then

$$\varepsilon = D^{-1}\left(\frac{k}{n}, \frac{1}{n}\left[(N+1)\log|\mathcal{H}| + \log\frac{1}{\delta}\right]\right).$$

We obtain that with probability at least $1-\delta$ over the draw of the training examples, for all $f \in C(\mathcal{H})$, all $\alpha \in U$, but fixed k,

$$P_{\mathcal{D},g\sim Q_N(f)}\Big(yg(x) \le \alpha\Big) \le P_{g\sim Q_N(f)}\left(P_{\mathcal{S}}\Big(yg(x) \le \alpha\Big) > \frac{k}{n}\right) + D^{-1}\left(\frac{k}{n}, \frac{1}{n}\left[(N+1)\log|\mathcal{H}| + \log\frac{1}{\delta}\right]\right). \tag{20}$$

We next bound the first term in the right-hand side of (20). Using the same argument for deriving (15), we have for any fixed $f, \mathcal{S}, \alpha$, k, any $\theta > \alpha$

$$P_{g\sim Q_N(f)}\left(P_{\mathcal{S}}\Big(yg(x) \le \alpha\Big) > \frac{k}{n}\right) \le I\left[P_{\mathcal{S}}\Big(yf(x) < \theta\Big) > \frac{k}{n}\right] + P_{g\sim Q_N(f)}\left(P_{\mathcal{S}}\Big(yg(x) \le \alpha\Big) > \frac{k}{n}, P_{\mathcal{S}}\Big(yf(x) < \theta\Big) \le \frac{k}{n}\right). \tag{21}$$

Note that the last term in (21) can be written in the following equivalent form and further bounded by

$$P_{g\sim Q_N(f)}\left(\exists (x_i, y_i) \in \mathcal{S} : \; y_i g(x_i) \le \alpha, \quad y_i f(x_i) \ge \theta\right) \le n \exp\left(-\frac{N(\theta-\alpha)^2}{2}\right). \tag{22}$$

Combining (15), (20), (21) and (22), we have that with probability at least $1-\delta$ over the draw of training examples, for all $f \in C(\mathcal{H})$, all $\alpha \in U$, all $\theta > \alpha$, but fixed k and N

$$P_{\mathcal{D}}\Big(yf(x) \le 0\Big) \le \exp\left(-\frac{N\alpha^2}{2}\right) + n\exp\left(-\frac{N(\theta-\alpha)^2}{2}\right) + I\left[P_{\mathcal{S}}\Big(yf(x) < \theta\Big) > \frac{k}{n}\right] + D^{-1}\left(\frac{k}{n}, \frac{1}{n}\left[(N+1)\log|\mathcal{H}| + \log\frac{1}{\delta}\right]\right).$$

Since θ is arbitrary, we set $\theta = \hat{\theta}(\frac{k}{n})$. Now we construct α by rounding $\theta/2$ to the nearest neighbor of $1/|\mathcal{H}|$. Let

$$\alpha = \frac{\theta}{2} - \frac{\eta}{|\mathcal{H}|} \in U,$$

where $0 \le \eta < 1$. The goal is to let α takes only a finite number of values. (Recall that $U = \{\frac{1}{|\mathcal{H}|}, \cdots, 1\}$.) It is easy to check that the sum of the first two terms on the right-hand side of the above inequality can be bounded by the following.

$$\begin{aligned} & \exp\left(-\frac{N\alpha^2}{2}\right) + n\exp\left(-\frac{N(\theta-\alpha)^2}{2}\right) \\ \le\ & \exp\left(-\frac{N\theta^2}{8}\right)\exp\left(-\frac{N\eta^2}{2|\mathcal{H}|^2}\right)\left[\exp\left(\frac{N\theta\eta}{2|\mathcal{H}|}\right) + n\exp\left(-\frac{N\theta\eta}{2|\mathcal{H}|}\right)\right] \\ \le\ & \max\left(2n, \exp\left(\frac{N}{2|\mathcal{H}|}\right)+1\right)\exp\left(-\frac{N\theta^2}{8}\right). \end{aligned}$$

The last inequality holds since $0 \le \theta, \eta \le 1$. Replacing δ by $\delta \cdot 2^{-N}$. we can get a union bound over all N by replacing $\log(\frac{n}{\delta})$ in all previous equations by $\log(\frac{n}{\delta \cdot 2^{-N}}) = N\log 2 + \log(\frac{n}{\delta})$. Put

$$N = \left\lceil \frac{8}{\theta^2}\log\left(\frac{2n^2}{\log|\mathcal{H}|}\right)\right\rceil.$$

Now for any sample $\mathcal{S}$ we only consider $f \in C(\mathcal{H})$ and k that satisfy $q_0 < 1$ and

$$\frac{k}{n} \ge q_0. \tag{23}$$

Note that by (23) and the assumption that $|\mathcal{H}| > 8$, we have

$$\theta > \sqrt{\frac{8}{|\mathcal{H}|}}.$$

So by some numerical calculations one can show

$$2n > \exp\left(\frac{N}{2|\mathcal{H}|}\right) + 1, \quad (n > 1).$$

Recall that $\theta = \hat{\theta}(k/n)$, so $P_{\mathcal{S}}\Big(yf(x) < \theta\Big) \le \frac{k}{n}$. We thus obtain that for fixed k, with probability at least $1-\delta$ over the random choice of the training set $\mathcal{S}$ of n examples, every $f \in C(\mathcal{H})$ with $q_0 < 1$ satisfies

$$P_{\mathcal{D}}\Big(yf(x) \le 0\Big) \le \frac{\log|\mathcal{H}|}{n} + D^{-1}\left(\frac{k}{n}, u\right),$$

where

$$u = \frac{1}{n}\left(\frac{8}{\theta^2}\log\left(\frac{2n^2}{\log|\mathcal{H}|}\right)\log(2|\mathcal{H}|) + 2\log|\mathcal{H}| + \log\frac{1}{\delta}\right).$$

Finally using the union bound over $k \in \{nq_0, \ldots, n-1\}$ and replacing δ by δ/n, we have with probability at least $1-\delta$ over the random choice of the training set $\mathcal{S}$ of n examples, every $f \in C(\mathcal{H})$ with $q_0 < 1$ satisfies

$$P_{\mathcal{D}}\Big(yf(x) \le 0\Big) \le \frac{\log|\mathcal{H}|}{n} + \inf_{k\in\{nq_0,\ldots,n-1\}} D^{-1}\left(\frac{k}{n}, u'\right),$$

where

$$u' = \frac{1}{n}\left(\frac{8}{\theta^2}\log\left(\frac{2n^2}{\log|\mathcal{H}|}\right)\log(2|\mathcal{H}|) + 2\log|\mathcal{H}| + \log\frac{n}{\delta}\right).$$

The theorem follows. ■

6.2 Proof of Theorem 5

The following lemma will be used to prove Theorem 5.

Lemma 11 *$D^{-1}(0,p) \le p$ for $p \ge 0$.*

Proof of Lemma 11. We only need to show

$$D(0||p) \ge p,$$

since $D(q||p)$ is a monotonic increasing function of p for $p \ge q$. By Taylor expansion

$$D(0||p) = -\log(1-p) = p + \frac{p^2}{2} + \frac{p^3}{3} + \cdots \ge p.$$

■

Proof of Theorem 5. By the assumption of Theorem 2 we have $\theta_0 > 4\sqrt{\frac{2}{|\mathcal{H}|}}$. Then it is easy to see that the right-hand side of the Emargin bound (3) is the minimum over all $q \in \left\{0, \ldots, \frac{n-1}{n}\right\}$. Take $q = 0$, it is clear that $\hat{\theta}(0)$ is the minimum margin. By Lemma 11, the Emargin bound can be relaxed to

$$\begin{aligned} P_{\mathcal{D}}\Big(yf(x) \le 0\Big) &\le \frac{1}{n}\left(\frac{8}{\theta_0^2}\log\left(\frac{2n^2}{\log|\mathcal{H}|}\right)\log(2|\mathcal{H}|) + 3\log|\mathcal{H}| + \log\frac{n}{\delta}\right) \\ &\le \frac{16\log(2n)\log(2|\mathcal{H}|)}{n\theta_0^2} + \frac{\log n + 2\log|\mathcal{H}|}{n} + \frac{1}{n}\log\left(\frac{|\mathcal{H}|}{\delta}\right). \end{aligned} \tag{24}$$

We only need to show that this relaxed bound is sharper than Theorem 2. For the minimum margin bound, we only consider the case that $R \le 1$, since otherwise the bound is larger than one. Simple calculations show that the right-hand side of (24) is smaller than the minimum margin bound. ■

6.3 Proof of Theorem 6

Remember that $q_i = P_{\mathcal{S}}(yf_i(x) < \theta_i)$ is the optimal q^* in the Emargin bound. Thus we only need to show

$$D^{-1}\Big(q_1, u(\theta_1)\Big) \leq D^{-1}\Big(q_2, u(\theta_2)\Big).$$

Note that if $\theta_1 \geq \theta_2$, then $u(\theta_1) \leq u(\theta_2)$. So

$$D^{-1}\Big(q_2, u(\theta_2)\Big) \geq D^{-1}\Big(q_2, u(\theta_1)\Big),$$

since $D^{-1}(q,u)$ is an increasing function of u for fixed q. Also $D^{-1}(q,u)$ is an increasing function of q for fixed u, we have

$$D^{-1}\Big(q_2, u(\theta_1)\Big) \geq D^{-1}\Big(q_1, u(\theta_1)\Big)$$

since $q_1 \leq q_2$. This completes the proof. ■

6.4 Proof of Theorem 7

The next lemma is a modified version of the uniform convergence result (Vapnik and Chervonenkis, 1971; Vapnik, 1998) and its refinement (Devroye, 1982). It will be used for proving Theorem 7.

Lemma 12 *Let $\mathcal{A}$ be a class of subsets of a space Z. Let $z_i \in Z$, $i = 1, \ldots, n$. Let $N^{\mathcal{A}}(z_1, z_2, \ldots, z_n)$ be the number of different sets in*

$$\Big\{\{z_1, z_2, \ldots, z_n\} \bigcap A : \ A \in \mathcal{A}\Big\}.$$

Define

$$s(\mathcal{A}, n) = \max_{(z_1, z_2, \ldots, z_n) \in Z^n} N^{\mathcal{A}}(z_1, z_2, \ldots, z_n).$$

Assume $\varepsilon \geq \frac{1}{n}$. Let $\varepsilon' = \frac{n}{n-1}\varepsilon - \frac{1}{n}$. Then for any distribution $\mathcal{D}$ over Z and any nonnegative integer k such that $\frac{k}{n} \leq \varepsilon'$

$$\Pr_{\mathcal{S}\sim\mathcal{D}^n}\left(\exists A \in \mathcal{A} : \ P_{\mathcal{D}}(A) > I\left[P_{\mathcal{S}}(A) > \frac{k}{n}\right] + \varepsilon\right) \leq 2 \cdot s(\mathcal{A}, n^2) \exp\left(-nD\left(\frac{k}{n} \middle\| \varepsilon'\right)\right).$$

Proof of Lemma 12. The proof is the standard argument. We first show that for any $0 < \alpha < 1$, $\varepsilon > 0$, and any integer n'

$$\begin{aligned}
&\Pr_{\mathcal{S}\sim\mathcal{D}^n}\left(\exists A \in \mathcal{A} : \ P_{\mathcal{D}}(A) > I\left[P_{\mathcal{S}}(A) > \frac{k}{n}\right] + \varepsilon\right) \\
&\quad \leq \left(\frac{1}{1 - e^{-2n'\alpha^2\varepsilon^2}}\right) \Pr_{S\sim\mathcal{D}^n,\ S'\sim\mathcal{D}^{n'}}\left(\exists A \in \mathcal{A} : \ P_{S'}(A) > I\left[P_{\mathcal{S}}(A) > \frac{k}{n}\right] + (1-\alpha)\varepsilon\right).
\end{aligned}$$

Or equivalently,

$$\Pr_{\mathcal{S}\sim\mathcal{D}^n}\left(\sup_{A\in\mathcal{A}}\left(P_{\mathcal{D}}(A)-I\left[P_{\mathcal{S}}(A)>\frac{k}{n}\right]\right)>\varepsilon\right)$$
$$\leq\left(\frac{1}{1-e^{-2n'\alpha^2\varepsilon^2}}\right)\Pr_{\mathcal{S}\sim\mathcal{D}^n,\ \mathcal{S}'\sim\mathcal{D}^{n'}}\left(\sup_{A\in\mathcal{A}}\left(P_{\mathcal{S}'}(A)-I\left[P_{\mathcal{S}}(A)>\frac{k}{n}\right]\right)>(1-\alpha)\varepsilon\right). \tag{25}$$

Let V denote the event

$$\sup_{A\in\mathcal{A}}\left(P_{\mathcal{D}}(A)-I\left[P_{\mathcal{S}}(A)>\frac{k}{n}\right]\right)>\varepsilon.$$

If the above event occurs, let A^* be any $A\in\mathcal{A}$ so that $P_{\mathcal{D}}(A)-I\left[P_{\mathcal{S}}(A)>\frac{k}{n}\right]>\varepsilon$. Otherwise let A^* be any $A\in\mathcal{A}$. Note that the following two events

$$P_{\mathcal{S}'}(A^*)\geq P_{\mathcal{D}}(A^*)-\alpha\varepsilon$$

and

$$P_{\mathcal{D}}(A^*)-I\left[P_{\mathcal{S}}(A^*)>\frac{k}{n}\right]>\varepsilon$$

imply that

$$P_{\mathcal{S}'}(A^*)-I\left[P_{\mathcal{S}}(A^*)>\frac{k}{n}\right]>(1-\alpha)\varepsilon.$$

Then

$$\begin{aligned}
&\Pr_{\mathcal{S}\sim\mathcal{D}^n,\ \mathcal{S}'\sim\mathcal{D}^{n'}}\left(\sup_{A\in\mathcal{A}}\left(P_{\mathcal{S}'}(A)-I\left[P_{\mathcal{S}}(A)>\frac{k}{n}\right]\right)>(1-\alpha)\varepsilon\right)\\
&\quad=\int d\mathcal{D}^n\int I\left[\sup_{A\in\mathcal{A}}\left(P_{\mathcal{S}'}(A)-I\left[P_{\mathcal{S}}(A)>\frac{k}{n}\right]\right)>(1-\alpha)\varepsilon\right]dD^{n'}\\
&\quad\geq\int_V d\mathcal{D}^n\int I\left[\sup_{A\in\mathcal{A}}\left(P_{\mathcal{S}'}(A)-I\left[P_{\mathcal{S}}(A)>\frac{k}{n}\right]\right)>(1-\alpha)\varepsilon\right]d\mathcal{D}^{n'}\\
&\quad\geq\int_V d\mathcal{D}^n\int I\left[P_{\mathcal{S}'}(A^*)-I\left[P_{\mathcal{S}}(A^*)>\frac{k}{n}\right]>(1-\alpha)\varepsilon\right]d\mathcal{D}^{n'}\\
&\quad\geq\int_V d\mathcal{D}^n\int I\left[P_{\mathcal{S}'}(A^*)\geq P_{\mathcal{D}}(A^*)-\alpha\varepsilon\right]d\mathcal{D}^{n'}\\
&\quad\geq\left(1-e^{-2n'\alpha^2\varepsilon^2}\right)\int_V d\mathcal{D}^n\\
&\quad=\left(1-e^{-2n'\alpha^2\varepsilon^2}\right)\Pr_{\mathcal{S}\sim\mathcal{D}^n}\left(\sup_{A\in\mathcal{A}}\left(P_{\mathcal{D}}(A)-I\left[P_{\mathcal{S}}(A)>\frac{k}{n}\right]\right)>\varepsilon\right).
\end{aligned}$$

This completes the proof of (25).

Take

$$n' = n^2 - n,$$
$$\alpha = \frac{1}{(n-1)\varepsilon},$$

we have

$$\Pr_{\mathcal{S}\sim\mathcal{D}^n}\left(\exists A \in \mathcal{A}:\ P_{\mathcal{D}}(A) > I\left[P_{\mathcal{S}}(A) > \frac{k}{n}\right] + \varepsilon\right)$$
$$\leq 2 \Pr_{\mathcal{S}\sim\mathcal{D}^n,\ \mathcal{S}'\sim\mathcal{D}^{n'}}\left(\exists A \in \mathcal{A}:\ P_{\mathcal{S}'}(A) > I\left[P_{\mathcal{S}}(A) > \frac{k}{n}\right] + (\varepsilon - \frac{1}{n-1})\right).$$

Proceeding as Devroye (1982) and using the relative entropy Hoeffding inequality, the lemma follows. ■

Proof of Theorem 7. The proof is the same as Theorem 3 until we have (18). Let $\alpha = \frac{\theta}{2}$, we need to bound

$$\Pr_{\mathcal{S}\sim\mathcal{D}^n}\left(\exists g \in C_N(\mathcal{H}),\ \exists\theta > 0,\ P_{\mathcal{D}}\Big(yg(x) \leq \frac{\theta}{2}\Big) > I\left[P_{\mathcal{S}}\Big(yg(x) \leq \frac{\theta}{2}\Big) > \frac{k}{n}\right] + \varepsilon\right).$$

Note that for fixed N, in order to derive a bound uniformly over all $0 < \theta \leq 1$ it suffices to show the bound holds for $\theta = \frac{1}{N}, \frac{2}{N}, \ldots, 1$. Let

$$A(g) = \left\{(x,y) \in \mathcal{X}\times\{-1,1\}:\ yg(x) \leq \frac{\theta}{2}\right\},$$

and

$$\mathcal{A} = \{A(g):\ g \in C_N(\mathcal{H})\}.$$

By Sauer's lemma (Sauer, 1972) it is easy to see that

$$s(\mathcal{A}, n) \leq \left(\frac{en}{d}\right)^{Nd},$$

where d is the VC dimension of $\mathcal{H}$. By Lemma 12, we have

$$\Pr_{S\sim D^n}\left(\exists g \in C_N(\mathcal{H}),\ \exists\theta > 0,\ P_{\mathcal{D}}\Big(yg(x) \leq \frac{\theta}{2}\Big) > I\left[P_{\mathcal{S}}\Big(yg(x) \leq \frac{\theta}{2}\Big) > \frac{k}{n}\right] + \varepsilon\right)$$
$$\leq 2(N+1)\left(\frac{en^2}{d}\right)^{Nd}\exp\left(-nD\left(\frac{k}{n}\middle\|\varepsilon'\right)\right),$$

where

$$\varepsilon' = \frac{n}{n-1}\varepsilon - \frac{1}{n}.$$

Proceeding as the proof of Theorem 3, we have that with probability at least $1-\delta$ the following holds for every $f \in C(\mathcal{H})$, every $\theta > 0$ but fixed k, where $0 \le k \le n\varepsilon$.

$$P_{\mathcal{D},g\sim Q_N(f)}\left(yg(x) \le \frac{\theta}{2}\right) \le P_{g\sim Q_N(f)}\left(P_{\mathcal{S}}\left(yg(x) \le \frac{\theta}{2}\right) > \frac{k}{n}\right) + \frac{1}{n} + \frac{n-1}{n} D^{-1}\left(\frac{k}{n}, \tau\right), \tag{26}$$

where

$$\tau = \frac{1}{n}\left[Nd\left(\log\frac{n^2}{d}+1\right) + \log(2(N+1)) + \log\frac{1}{\delta}\right].$$

Similar to the proof of Theorem 3, we can bound the first term of (26) as

$$\begin{aligned} P_{g\sim Q_N(f)}\left(P_{\mathcal{S}}\left(yg(x) \le \frac{\theta}{2}\right) > \frac{k}{n}\right) &\le I\left[P_{\mathcal{S}}\left(yf(x) < \theta\right) > \frac{k}{n}\right] \\ &\quad + P_{g\sim Q_N(f)}\left(P_{\mathcal{S}}\left(yg(x) \le \frac{\theta}{2}\right) > \frac{k}{n},\ P_{\mathcal{S}}\left(yf(x) < \theta\right) \le \frac{k}{n}\right) \\ &\le I\left[P_{\mathcal{S}}\left(yf(x) < \theta\right) > \frac{k}{n}\right] + n\exp\left(-\frac{N\theta^2}{8}\right). \end{aligned} \tag{27}$$

Setting $\theta = \hat{\theta}(\frac{k}{n})$ and combining (26), (27) and (15); recalling $\alpha = \theta/2$ we have with probability at least $1-\delta$ for all $f \in C(\mathcal{H})$, all $0 < \theta \le 1$, but fixed k and N

$$P_{\mathcal{D}}(yf(x) \le 0) \le \frac{1}{n} + (n+1)\exp\left(-\frac{N\theta^2}{8}\right) + \frac{n-1}{n} D^{-1}\left(\frac{k}{n}, \tau\right).$$

Use the union bound over N; put $N = \frac{16}{\theta^2}\log\frac{n}{d}$ and use the union bound over k as in the proof of Theorem 3 we obtain the theorem. ■

6.5 Proof of Lemma 8

The first inequality has already been proved in Lemma 11.

For the second inequality, we only need to show

$$D^{-1}(q,u) \le q + \sqrt{u/2},$$

or equivalently

$$D(q, q+\sqrt{u/2}) \ge u,$$

since D is an increasing function in the second parameter. But this is immediate by a well known result (Hoeffding, 1963):

$$D(q, q+\delta) \ge 2\delta^2.$$

For the third inequality we first show that for all $0 < q < 1$

$$D^{-1}\left(\frac{q}{2}, \frac{q}{8}\right) \le q, \tag{28}$$

which is equivalent to

$$D\left(\frac{q}{2}\middle\|q\right) \ge \frac{q}{8}.$$

For fixed q, let $\phi(x) = D(qx||q)$, $0 < x \leq 1$. Note that

$$\phi(1) = \phi'(1) = 0,$$

and

$$\phi''(x) = \frac{q}{x(1-qx)} \geq q,$$

we have

$$D\left(\frac{q}{2}\middle\|q\right) = \phi\left(\frac{1}{2}\right) \geq \frac{q}{8}.$$

This completes the proof of (28).

Now if $q \leq Cu[\hat{\theta}(q)]$, recall that $C' = \max(2C, 8)$, and note D^{-1} is increasing function on its first and second parameter respectively. If $C'u[\hat{\theta}(q)] < 1$ we have

$$\begin{aligned} D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right) &\leq D^{-1}\left(\frac{C'}{2}u\left[\hat{\theta}(q)\right], u\left[\hat{\theta}(q)\right]\right) \\ &\leq D^{-1}\left(\frac{C'}{2}u\left[\hat{\theta}(q)\right], \frac{C'}{8}u\left[\hat{\theta}(q)\right]\right) \\ &\leq C'u\left[\hat{\theta}(q)\right]. \end{aligned}$$

The lemma follows. ■

6.6 Proof of Corollary 9

The first and third bounds are straightforward from lemma 8. We only prove the second bound.

Let $\Phi(\theta)$ be the right hand side of the bound (without taking the infimum) we want to prove, i.e.,

$$\Phi(\theta) = P_{\mathcal{S}}\Big(yf(x) \leq \theta\Big) + O\left(\frac{1}{\sqrt{n}}\left(\frac{\log n \log|\mathcal{H}|}{\theta^2} + \log\frac{1}{\delta}\right)^{1/2}\right).$$

It is not difficult to see that there is no θ that can achieve $\inf_{\theta\in[8/|\mathcal{H}|,1]}\Phi(\theta)$. To see this, first note that for any θ, either $P_{\mathcal{S}}(yf(x) < \theta) = P_{\mathcal{S}}(yf(x) \leq \theta)$ (a continuous point), or $P_{\mathcal{S}}(yf(x) < \theta) < P_{\mathcal{S}}(yf(x) \leq \theta)$ (a jump point). In the former case, increasing θ decreases $\Phi(\theta)$ since $P_S(yf(x) \leq \theta)$ does not change but $u(\theta)$ is decreasing. In the latter case, decreasing θ also decreases $\Phi(\theta)$, since $P_S(yf(x) \leq \theta)$ decreases discontinuously while $u(\theta)$ increases continuously.

Let $\theta_1, \theta_2, \ldots$, be a sequence so that $\Phi(\theta_i)$ converges to $\inf_\theta \Phi(\theta)$. Let $\overline{\theta}$ be the limiting point of $\theta_1, \theta_2, \ldots$. It is not difficult to see from the above argument that for sufficiently large i, $\theta_i < \overline{\theta}$, since there is a jump of $\Phi(\theta)$ at those θ such that $P_{\mathcal{S}}(yf(x) \leq \theta)$ is discontinuous. Take any θ_i that is sufficiently close to $\overline{\theta}$. Let $q_i = P_{\mathcal{S}}(yf(x) \leq \theta_i)$, we must have $\hat{\theta}(q_i) = \overline{\theta}$ (recall that $\hat{\theta}(q_i) = \sup\{\theta \in (0,1] : P_{\mathcal{S}}(yf(x) \leq \theta) \leq q_i\}$). Therefore $u[\hat{\theta}(q_i)] < u(\theta_i)$ and hence $q_i + (u[\hat{\theta}(q_i)])^{1/2} < \Phi(\theta_i)$. Thus $\inf_q(q + (u[\hat{\theta}(q)])^{1/2}) \leq \inf_\theta \Phi(\theta)$. The corollary follows. ■

6.7 Proof of Theorem 10

We first give a simple lemma.

Lemma 13 *Let ξ be a random variable and κ a positive constant. If for any $t > 0$ we have $P(\xi > \kappa t) < \exp(-t^2)$, then $E\xi \leq \frac{\sqrt{\pi}}{2}\kappa$.*

Proof of Lemma 13.

$$E\xi = \int_{-\infty}^{\infty} u \; d(-P(\xi > u)) \leq \int_{0}^{\infty} u \; d(-P(\xi > u)).$$

By the assumption, we have

$$E\xi \leq \int_{0}^{\infty} \kappa t \; d(-e^{-t^2}) = \frac{\sqrt{\pi}}{2}\kappa.$$

■

Proof of Theorem 10.

Let $B(f)$ be the right-hand-side of the Emargin bound in Theorem 7. Then for any training set $\mathcal{S}$, $\hat{f}_n$ is the function f in $C(\mathcal{H})$ so that $B(f)$ is minimized, i.e., $\hat{f}_n = \arg\min_{f \in C(\mathcal{H})} B(f)$. According to the Emargin bound, with probability $1 - \delta$

$$P_{\mathcal{D}}(y\hat{f}_n(x) \leq 0) \leq B(\hat{f}_n).$$

Since $\hat{f}_n = \arg\min_{f \in C(\mathcal{H})} B(f)$, then for any $f \in C(\mathcal{H})$, we have $B(\hat{f}_n) \leq B(f)$. Therefore for all $f \in C(\mathcal{H})$, with probability $1 - \delta$

$$P_{\mathcal{D}}(y\hat{f}_n(x) \leq 0) \leq B(f) = \frac{d^2+1}{n} + \inf_{q \in \{q_0, \ldots, \frac{n-1}{n}\}} \frac{n-1}{n} D^{-1}(q, u[\hat{\theta}(q)]).$$

For any fixed $f \in C(\mathcal{H})$, let $q = P_{\mathcal{S}}(yf(x) \leq n^{-1/4})$. It is easy to see that $\hat{\theta}(q) \geq n^{-1/4}$ and $u[\hat{\theta}(q)] \leq u[n^{-1/4}]$, where $u[\hat{\theta}(q)]$ is defined in Theorem 7. By the second inequality of D^{-1} in lemma 8, we have

$$\begin{aligned} P_{\mathcal{D}}(y\hat{f}_n(x) \leq 0) &\leq \frac{d^2+1}{n} + \frac{n-1}{n}\left(q + (u\left[\hat{\theta}(q)\right])^{1/2}\right), \\ &\leq \frac{d^2+1}{n} + \frac{n-1}{n}\left(P_{\mathcal{S}}(yf(x) \leq n^{-1/4}) + (u\left[n^{-1/4}\right])^{1/2}\right). \end{aligned}$$

It is easy to see that there is a constant c (independent of f) such that the right-hand-side of the above inequality can be further bounded by

$$\frac{n-1}{n} P_{\mathcal{S}}(yf(x) \leq n^{-1/4}) + c\frac{d\log\frac{n}{d}}{n^{1/4}} + c\sqrt{\frac{\log n}{n}\log(\frac{1}{\delta})}.$$

Let $t = \sqrt{\log(\frac{1}{\delta})}$, we have that for any $t > 0$ with probability at most $\exp(-t^2)$

$$P_{\mathcal{D}}(y\hat{f}_n(x) \leq 0) - \frac{n-1}{n} P_{\mathcal{S}}(yf(x) \leq n^{-1/4}) - c\frac{d\log\frac{n}{d}}{n^{1/4}} > c\sqrt{\frac{\log n}{n}}t.$$

According to lemma 13, we obtain

$$EP_{\mathcal{D}}(y\hat{f}_n(x) \le 0) - \frac{n-1}{n} EP_{\mathcal{S}}(yf(x) \le n^{-1/4}) - c\frac{d\log\frac{n}{d}}{n^{1/4}} \le \frac{c\sqrt{\pi}}{2}\sqrt{\frac{\log n}{n}},$$

where the expectation is over the random choice of the training set. Note that

$$EP_{\mathcal{S}}(yf(x) \le n^{-1/4}) = P_{\mathcal{D}}(yf(x) \le n^{-1/4}),$$

we have

$$EP_{\mathcal{D}}(y\hat{f}_n(x) \le 0) \le \frac{n-1}{n} P_{\mathcal{D}}(yf(x) \le n^{-1/4}) + c\frac{d\log\frac{n}{d}}{n^{1/4}} + \frac{c\sqrt{\pi}}{2}\sqrt{\frac{\log n}{n}}.$$

Let $n \to \infty$, we obtain

$$\lim_{n\to\infty} EP_{\mathcal{D}}(y\hat{f}_n(x) \le 0) \le \lim_{n\to\infty} P_{\mathcal{D}}(yf(x) \le n^{-1/4}) = P_{\mathcal{D}}(yf(x) \le 0).$$

The last equality holds because $P_{\mathcal{D}}(yf(x) \le \theta)$ is a right continuous function of θ. Since the above inequality is true for every $f \in C(\mathcal{H})$, we have

$$\lim_{n\to\infty} EP_{\mathcal{D}}(y\hat{f}_n(x) \le 0) \le \inf_{f\in C(\mathcal{H})} P_{\mathcal{D}}(yf(x) \le 0) = L^*.$$

■

7. Experiments

In this section we provide experimental results to verify our theory. We compare AdaBoost and LP-AdaBoost in terms of their Emargin, Emargin error and the generalization error. Theorem 6 suggests that if a voting classifier f_1 has a larger Emargin and a smaller Emargin error than another classifier f_2, then f_1 has a smaller bound of the generalization error than f_2. Thus we expect f_1 will have better performance on the test data. The goal of the experiment is to see whether the empirical results agree with the theoretical prediction.

The experiments are conducted on 17 benchmark datasets all from the UCI repository (Asuncion and Newman, 2007). The datasets are grouped into two categories. Table 1 lists 12 "large" datasets, each containing at least 1000 data points. Table 2 lists 5 "small" datasets, each has at most 1000 examples. (We distinguish large and small datasets because we found they demonstrate somewhat different results, see below for discussions.) If the data is multiclass, we group them into two classes since we study binary classification problems. For instance, the "letter" dataset has 26 classes, we use the first 13 as the positive and the others as the negative. In the preprocessing stage, each feature is normalized to $[0, 1]$. For all datasets we use 5-fold cross validation, and average the results over 10 runs (for a total of 50 runs on each dataset).

In order to study the effect of the margins, we need to control and calculate the complexity of the base classifiers. We conduct two sets of experiments using different base classifiers. For one set of experiments, we use decision stumps. For the other, we use three-layer eight-leaf (complete) binary decision trees (Therefore the shape of the trees are fixed). We consider a finite set of base classifiers. Specifically, for each feature we consider 100

Table 1: Description of the large datasets

Dataset	# Examples	# Features	Dataset	# Examples	# Features
Image	2310	16	Page-block	5473	10
Isolet	7797	617	Pendigits	10992	16
Letter	20000	16	Satimage	6435	36
Magic04	19022	10	Shuttle	58000	9
Mfeat-fac	2000	216	Spambase	4601	57
Optdigits	5620	64	Waveform	5000	30

Table 2: Description of the small datasets

Dataset	# Examples	# Features
Breast	683	9
Diabetes	768	8
German	1000	24
Vehicle	845	18
Wdbc	569	30

thresholds uniformly distributed on $[0, 1]$. Therefore the size of the set of decision stumps is $2 \times 100 \times k$, and for the three-layer eight-leaf trees is $(2 \times 100 \times k)^7$, where k denotes the number of features.

We run AdaBoost 100 rounds, and use the obtained base classifiers to train the LP-AdaBoost voting classifier. We then calculate the Emargin, Emargin error, test error as well as the minimum margin of them respectively. The calculation of the Emargin involves solving the inverse relative entropy $D^{-1}(q, u)$. Since D is a monotone function on the second parameter, one can adopt the Newton method to find the root of $D(q||\cdot) - u = 0$ on $[q, 1]$. Another simple way to solve $D^{-1}(q, u)$ is just applying binary search on $[q, 1]$: Let $p_1 = q$, $p_2 = 1$. We have $D(q, p_1) = 0 \leq u$ and $D(q, p_2) = \infty > u$. Then let $p_3 = \frac{p_1+p_2}{2}$, compute $D(q, p_3)$ and see if $D(q, p_3) > u$ or not, etc.

The results are described in Tables 3, 4, 5 and 6 respectively according to the type of base classifiers used and the size of the datasets. To highlight the results we use boldface in the following manner: By a t-test with significant level 0.01, **larger Emargin**, **smaller Emargin error**, and **smaller test error** are denoted in boldface. If on a dataset, the empirical result agrees with the theory, the **name of the dataset** is marked in boldface. For example, if one algorithm has larger Emargin, smaller or equal Emargin error, and smaller test error, then the dataset is marked in boldface. Similarly, if one algorithm has smaller Emargin error, larger or equal Emargin, and smaller test error, then the dataset is marked in boldface. Also if the two algorithms have (statistically) the same Emargin, Emargin error and test error, it agrees with the theory.

In Table 3 we use decision stump base classifiers on large datasets. We see that only one dataset is not marked in boldface. On this "Shuttle" dataset, LP-AdaBoost has a larger

Table 3: Margin measures and performances of AdaBoost and LP-AdaBoost on the **large** datasets and using the **stump** base classifiers.

		Emargin	Emargin Error	Test Error	Min margin
Image	Ada	0.461 ± 0.024	0.799 ± 0.016	0.032 ± 0.009	-0.076 ± 0.010
	LP	**0.751 ± 0.238**	**0.664 ± 0.075**	**0.029 ± 0.009**	0.000 ± 0.001
Isolet	Ada	0.172 ± 0.057	**0.714 ± 0.040**	**0.163 ± 0.045**	-0.195 ± 0.063
	LP	0.145 ± 0.031	0.763 ± 0.021	0.180 ± 0.053	-0.069 ± 0.015
Letter	Ada	**0.199 ± 0.010**	**0.804 ± 0.017**	**0.190 ± 0.005**	-0.309 ± 0.009
	LP	0.000 ± 0.000	0.905 ± 0.021	0.202 ± 0.012	0.000 ± 0.000
Magic04	Ada	**0.190 ± 0.007**	**0.716 ± 0.017**	**0.230 ± 0.006**	-0.412 ± 0.034
	LP	0.000 ± 0.000	0.859 ± 0.063	0.265 ± 0.017	0.000 ± 0.000
Mfeat-fac	Ada	**0.184 ± 0.008**	**0.538 ± 0.033**	**0.040 ± 0.009**	-0.018 ± 0.007
	LP	0.171 ± 0.009	0.558 ± 0.038	0.045 ± 0.010	0.033 ± 0.003
Optdigits	Ada	**0.173 ± 0.009**	**0.654 ± 0.022**	**0.111 ± 0.013**	-0.231 ± 0.016
	LP	0.017 ± 0.046	0.708 ± 0.027	0.127 ± 0.019	-0.010 ± 0.027
Page-block	Ada	0.278 ±0.014	**0.458 ± 0.037**	**0.048 ± 0.005**	-0.213 ± 0.023
	LP	0.232 ± 0.374	0.686 ± 0.218	0.055 ± 0.008	0.000 ± 0.000
Pendigits	Ada	**0.176 ± 0.006**	**0.634 ± 0.020**	**0.091 ± 0.006**	-0.243 ± 0.015
	LP	0.135 ± 0.046	0.711 ± 0.028	0.131 ± 0.010	-0.085 ± 0.029
Satimage	Ada	**0.262 ± 0.008**	**0594 ± 0.018**	**0.057 ± 0.005**	-0.161 ± 0.014
	LP	0.092 ± 0.280	0.771 ± 0.036	0.066 ± 0.007	0.000 ± 0.000
Shuttle	Ada	0.173 ± 0.017	**0.062 ± 0.038**	0.001 ± 0.000	-0.087 ± 0.026
	LP	**0.204 ± 0.032**	0.251 ± 0.065	0.001 ± 0.000	0.000 ± 0.000
Spambase	Ada	**0.315 ± 0.217**	**0.591 ± 0.201**	**0.055 ± 0.020**	-0.126 ± 0.365
	LP	0.116 ± 0.316	0.737 ± 0.257	0.080 ± 0.028	0.096 ± 0.291
Waveform	Ada	**0.371 ± 0.014**	**0.721 ± 0.013**	**0.096 ± 0.008**	-0.185 ± 0.014
	LP	0.000 ± 0.000	0.780 ±0.014	0.104 ± 0.011	0.000 ± 0.000

Table 4: Margin measures and performances of AdaBoost and LP-AdaBoost on the **small** datasets and using the **stump** base classifiers.

		Emargin	Emargin Error	Test Error	Min margin
Breast	Ada	0.312 ± 0.045	**0.425 ± 0.082**	**0.044 ± 0.016**	-0.048 ± 0.017
	LP	0.299 ± 0.068	0.556 ± 0.135	0.053 ± 0.017	0.022 ± 0.012
Diabetes	Ada	0.216 ± 0.017	**0.753 ± 0.033**	**0.228 ± 0.026**	-0.199 ± 0.018
	LP	0.149 ± 0.294	0.821 ± 0.071	0.271 ± 0.040	-0.008 ± 0.015
German	Ada	**0.221 ± 0.015**	**0.769 ± 0.029**	**0.240 ± 0.026**	-0.246 ± 0.018
	LP	0.059 ± 0.173	0.818 ± 0.073	0.272 ± 0.030	0.000 ± 0.000
Vehicle	Ada	0.196 ± 0.012	**0.688 ± 0.035**	0.223 ± 0.026	-0.102 ± 0.011
	LP	0.273 ± 0.285	0.790 ± 0.075	0.231 ± 0.029	-0.018 ± 0.008
Wdbc	Ada	**0.400 ± 0.032**	0.537 ± 0.048	**0.028 ± 0.014**	0.096 ± 0.012
	LP	0.376 ± 0.032	0.546 ± 0.050	0.033 ± 0.015	0.139 ± 0.008

Table 5: Margin measures and performances of AdaBoost and LP-AdaBoost on the **large** datasets and using the **Tree** base classifiers.

		Emargin	Emargin Error	Test Error	Min margin
Image	Ada	0.370 ± 0.016	0.375 ± 0.034	0.010 ± 0.004	0.184 ± 0.008
	LP	0.374 ± 0.023	0.374 ± 0.054	0.010 ± 0.004	0.232 ± 0.007
Isolet	Ada	0.252 ± 0.076	0.589 ± 0.028	0.074 ± 0.067	0.020 ± 0.144
	LP	0.240 ± 0.010	0.591 ± 0.040	0.074 ± 0.056	0.063 ± 0.071
Letter	Ada	**0.246 ± 0.017**	**0.714 ± 0.034**	**0.077 ± 0.006**	-0.144 ± 0.012
	LP	0.236 ± 0.019	0.775 ± 0.031	0.086 ± 0.006	0.061 ± 0.004
Magic04	Ada	**0.312 ± 0.018**	**0.805 ± 0.018**	**0.156 ± 0.006**	-0.212 ± 0.012
	LP	0.282 ± 0.038	0.879 ± 0.028	0.225 ± 0.013	-0.085 ± 0.003
Mfeat-fac	Ada	**0.377 ± 0.029**	0.293 ± 0.104	0.017 ± 0.005	0.285 ± 0.006
	LP	0.350 ± 0.044	**0.146 ± 0.174**	0.018 ± 0.006	0.314 ± 0.005
Optdigits	Ada	0.288 ± 0.009	0.460 ± 0.025	0.018 ± 0.003	0.090 ± 0.006
	LP	0.288 ± 0.010	0.466 ± 0.022	0.018 ± 0.003	0.124 ± 0.004
Page-block	Ada	0.392 ± 0.024	**0.465 ± 0.038**	**0.030 ± 0.005**	-0.068 ± 0.009
	LP	**0.508 ± 0.041**	0.518 ± 0.057	0.033 ± 0.005	0.000 ± 0.000
Pendigits	Ada	**0.305 ± 0.008**	0.337 ± 0.017	0.005 ± 0.001	0.101 ± 0.008
	LP	0.301 ± 0.010	0.345 ± 0.022	0.005 ± 0.001	0.137 ± 0.005
Satimage	Ada	**0.319 ± 0.013**	0.484 ± 0.026	**0.044 ± 0.006**	0.012 ± 0.008
	LP	0.284 ± 0.014	0.496 ± 0.039	0.046 ± 0.006	0.055 ± 0.004
Shuttle	Ada	0.503 ± 0.037	**0.034 ± 0.020**	0.001 ± 0.000	-0.049 ± 0.013
	LP	**0.541 ± 0.066**	0.071 ± 0.042	0.001 ± 0.000	0.000 ± 0.000
Spambase	Ada	0.294 ± 0.014	**0.601 ± 0.034**	**0.052 ± 0.006**	-0.092 ± 0.008
	LP	0.309 ± 0.181	0.681 ± 0.077	0.067 ± 0.008	-0.002 ± 0.002
Waveform	Ada	**0.494 ± 0.023**	0.709 ± 0.011	**0.100 ± 0.009**	0.001 ± 0.006
	LP	0.473 ± 0.033	0.714 ± 0.018	0.103 ± 0.008	0.041 ± 0.003

Table 6: Margin measures and performances of AdaBoost and LP-AdaBoost on the **small** datasets and using the **tree** base classifiers.

		Emargin	Emargin Error	Test Error	Min margin
Breast	Ada	0.591 ± 0.057	0.392 ± 0.051	**0.030 ± 0.014**	0.317 ± 0.030
	LP	**0.667 ± 0.059**	0.404 ± 0.053	0.033 ± 0.014	0.385 ± 0.033
Diabetes	Ada	0.230 ± 0.032	0.706 ± 0.062	**0.272 ± 0.027**	0.035 ± 0.007
	LP	0.222 ± 0.026	0.709 ± 0.058	0.284 ± 0.030	0.082 ± 0.004
German	Ada	**0.202 ± 0.015**	0.704 ± 0.041	**0.242 ± 0.027**	-0.010 ± 0.010
	LP	0.192 ± 0.017	0.703 ± 0.050	0.259 ± 0.028	0.046 ± 0.004
Vehicle	Ada	**0.271 ± 0.018**	0.644 ± 0.038	0.216 ± 0.029	0.087 ± 0.007
	LP	0.256 ± 0.020	0.633 ± 0.046	0.216 ± 0.027	0.127 ± 0.004
Wdbc	Ada	0.539 ± 0.018	0.015 ± 0.010	0.028 ± 0.013	0.527 ± 0.019
	LP	**0.582 ± 0.020**	**0.002 ± 0.000**	0.030 ± 0.014	0.582 ± 0.020

Emargin and also a larger Emargin error. In this case, the comparison theorem (Theorem 6) does not apply. We mark such datasets by italic font. Note that AdaBoost does not always have larger Emargin than LP-AdaBoost. On the "Image" dataset, LP-AdaBoost achieves larger Emargin, smaller Emargin error and, as the bound predicts, a smaller test error.

In Table 4 we use decision stump base classifiers on small datasets. Four datasets agree with the theory. On the "Vehicle" dataset, although the bound predicts that AdaBoost would have a smaller generalization error, the test error of AdaBoost is not significantly smaller than LP-AdaBoost.

In Table 5 we use eight-leave decision tree base classifiers on large datasets. Eight datasets agree with the theory. For the "Mfeat-fac", "Page-block" and "Shuttle" datasets, our comparison theorem does not apply. Only the "Pendigits" dataset differs from the theoretical prediction: The test errors are the same while the theory predicts AdaBoost would perform better.

The last set of experiments, listed in Table 6, in which we use eight-leave decision tree base classifiers on small datasets, behaves different from all the previous results. Only one dataset agrees with the theory. On the "Breast" dataset, the test error is contrary to what the bound predicts.

To summarize, on large datasets, the Emargin theory usually agrees with empirical observations. AdaBoost has better performances because it has a larger Emargin and a smaller Emargin error. Note there are also cases that LP-AdaBoost achieves a larger Emargin and a smaller Emargin error and a smaller test error. However, on small datasets and with more complex base classifiers, the theory does not often give the correct predictions. We think the reason is that the bound is still loose, especially when the dataset contains only a few hundred of points. Also the number of classifiers is a loose bound for the complexity of complex decision trees.

Finally we plot in Figure 1 some margin distribution graphs and the corresponding Emargin and Emargin errors to give an illustration. AdaBoost often has intuitively "better" margin distributions.

8. Conclusions

In this paper we provided a refined analysis on the margin theory for boosting algorithms, which extended our preliminary study (Wang et al., 2008). We proposed a bound in terms of a new margin measure called the Emargin, which depends on the whole margin distribution. This bound is uniformly sharper than the minimum margin bound whose prediction is different from the empirical observations. Our theory suggests that a boosting classifier may not be necessarily achieve better performance even though it generates a larger minimum margin.

Our bound suggests that the Emargin and the Emargin error play important roles to guarantee a smaller bound of the generalization error of a voting classifier—a larger Emargin and a smaller Emargin error result in better generalization ability. Experimental results on (not-too-small) benchmark datasets agree well with our theory.

From a practical point of view, the Emargin bound is still too loose to give useful quantitative predictions. For most datasets, the bound is larger than 1/2. On the other

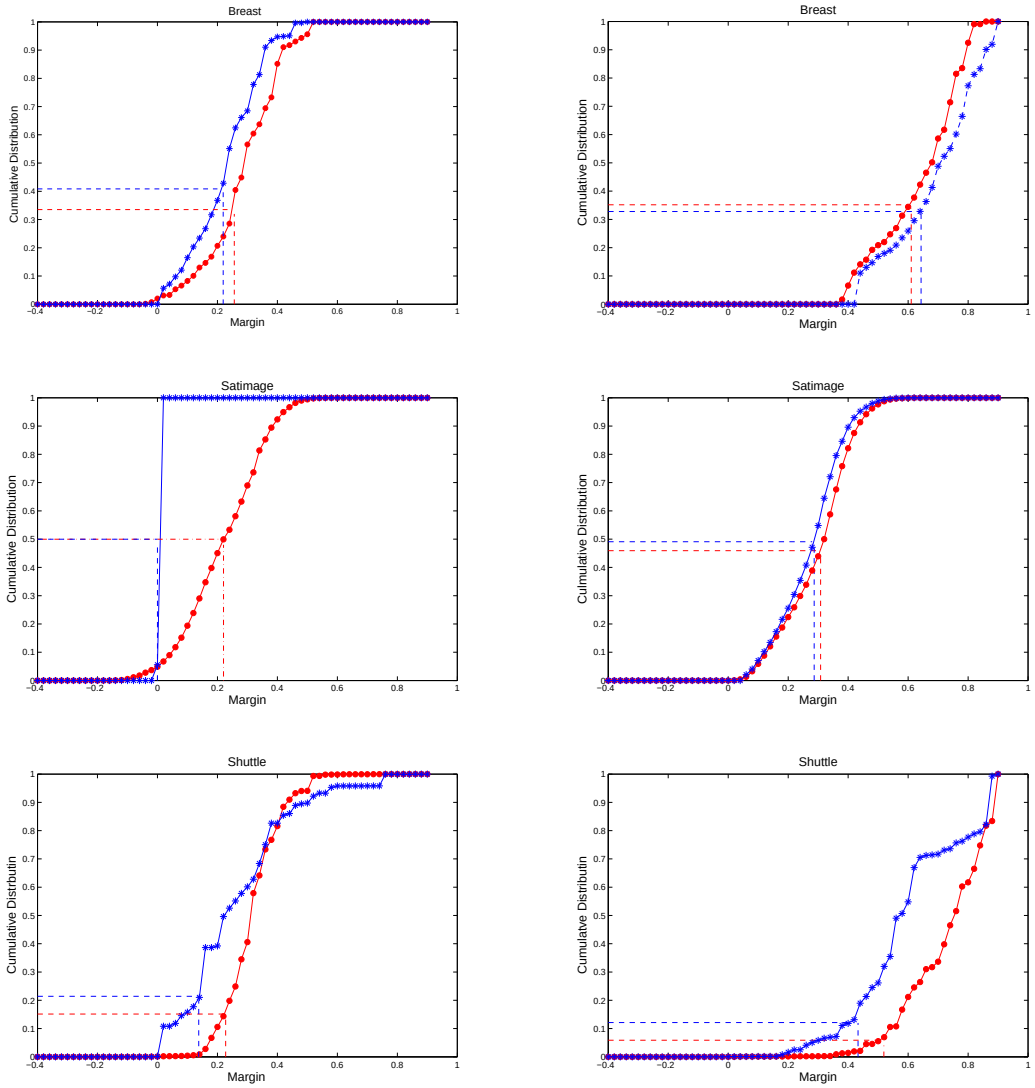

Figure 1: Margin distribution graphs with Emargin and Emargin errors. The lines marked with stars are the margin distributions of LP-AdaBoost. The lines marked with circles are of AdaBoost. Emargin and Emargin errors are plotted by lines parallel to the axes. The left column uses decision stump base classifiers, the right column uses decision tree classifiers. The three rows are from the datasets of Breast, Satimage and Shuttle respectively.

hand we can employ the bound to "compare" voting classifiers with the help of Emargin and Emargin error. This provides some guidance to choose classifiers. To calculate the Emargin, one needs to know the complexity (e.g., VC dimension) of the base classifiers. This can be difficult for some base learners like C4.5 decision trees.

A future work is to develop algorithms that generate voting classifiers with good margin distributions, i.e., large Emargin and small Emargin error. Directly optimizing Emargin and Emargin error would be computationally difficult. On the other hand, given a voting classifier $\sum \alpha_t h_t$, it might be possible to improve its margin distribution. One way is to solve the following linear optimization problem to obtain $\sum \beta_t h_t$.

$$\begin{aligned} \max_{\beta,\xi} \quad & \sum \xi_i \\ s.t. \quad & y_i \sum \beta_t h_t(x_i) \geq y_i \sum \alpha_t h_t(x_i) + \xi_i, \quad i = 1, 2, \dots \\ & \beta_t \geq 0, \quad \sum \beta_t = 1, \\ & \xi_i \geq 0, \end{aligned} \tag{29}$$

where $\alpha = (\alpha_1, \dots, \alpha_T)$, $\beta = (\beta_1, \dots, \beta_T)$, $\xi = (\xi_1, \dots, \xi_n)$. If there is a nontrivial solution (i.e., $\beta \neq \alpha$), $\sum \beta_t h_t$ would have a uniformly better margin distribution than $\sum \alpha_t h_t$ and therefore we expect it has a smaller generalization error. However, there is usually no nontrivial solutions when $\sum \alpha_t h_t$ is an AdaBoost classifier—it already has a good margin distribution. An open problem is to modify and relax (29) and obtain a solution with larger Emargin and smaller Emargin error. Then it would be a good test to see if such a classifier achieves better performance as our theory predicts.

Acknowledgement

We thank the referees for their useful and insightful comments. It greatly improves the quality of the paper. The first author would like to thank Phil Long, Gilles Blanchard and Lev Reyzin for helpful discussions. This work was supported by NSFC(61075003, 61073097), Global COE Program of Tokyo Institute of Technology and JiangsuSF(BK2008018) and the National Fundamental Research Program of China(2010CB327903). Part of the work was done when the first author was visiting Tokyo Institute of Technology.

References

A. Asuncion and D. J. Newman. UCI machine learning repository, 2007. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

P. Bartlett, M. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting and variants. *Machine Learning*, 36:105–139, 1999.

L. Breiman. Population theory for boosting ensembles. *Annals of Statistics*, 32:1–11, 2004.

L. Breiman. Prediction games and arcing algorithms. *Neural Computation*, 11:1493–1517, 1999.

L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26:801–849, 1998.

R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *23th International Conference on Machine Learning*, 2006.

L. Devroye. Bounds for the uniform deviation of empirical measures. *Journal of Multivariate Analysis*, 12:72–79, 1982.

T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning*, 40:139–157, 2000.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, 1996.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–407, 2000.

A. J. Grove and D. Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In *National Conference on Artificial Intelligence*, 1998.

W. Hoeffding. Probability inequalities for sum of bounded random variables. *Journal of American Statistical Society*, 58:13–30, 1963.

W. Jiang. Process consistency for adaboost. *The Annals of Statistics*, 32:13–29, 2004.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30:1–50, 2002.

V. Koltchinskii and D. Panchenko. Complexities of convex combinations and bounding the generalization error in classification. *Annals of Statistics*, 33:1455–1496, 2005.

J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

G. Lugosi and Nicolas Vayatis. On the bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, 32:30–55, 2004.

D. Mease and A. Wyner. Evidence contrary to the statistical view of boosting. *Journal of Machine Learning Research*, 9:131–156, 2008.

R. Meir and G. Rätsch. An introduction to boosting and leveraging. In *Advanced Lectures on Machine Learning*, pages 118–183, 2003.

J. R. Quinlan. Bagging, boosting, and c4.5. In *13th International Conference on Artificial Intelligence*, 1996.

G. Rätsch and M. Warmuth. Efficient margin maximization with boosting. *Journal of Machine Learning Research*, 6:2131–2152, 2005.

L. Reyzin and R. E. Schapire. How boosting the margin can also boost classifier complexity. In *International Conference on Machine Learning*, 2006.

C. Rudin, I. Daubechies, and R. Schapire. The dynamics of AdaBoost: Cyclic behavior and convergence of margins. *Journal of Machine Learning Research*, 5:1557–1595, Dec 2004.

C. Rudin, I. Daubechies, and R. Schapire. Analysis of boosting algorithms using the smooth margin function. *Annals of Statistics*, 35:2723–2768, 2007.

N. Sauer. On the density of family of sets. *Journal of Combinatorial Theory, Series A*, 13: 145–147, 1972.

R. Schapire, Y. Freund, P. Bartlett, and W. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26:1651–1686, 1998.

V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons Inc., 1998.

V. N. Vapnik and A. YA. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16:264–280, 1971.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.

L. Wang, C. Yang, and J. Feng. On learning with dissimilarity functions. In *24th International Conference on Machine Learning*, 2007.

L. Wang, M. Sugiyama, C. Yang, Z. Zhou, and J. Feng. On the margin explanation of boosting algorithms. In *21th Annual Conference on Learning Theory*, 2008.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–85, 2004.