On the Margin Explanation of Boosting Algorithms

Liwei Wang*[1], Masashi Sugiyama[2], Cheng Yang[1], Zhi-Hua Zhou[3], and Jufu Feng[1]
[1] Key Laboratory of Machine Perception, MOE, School of Electronics Engineering and Computer Science,
Peking University, Beijing, 100871, P.R.China. {wanglw,yangch,fjf}@cis.pku.edu.cn
[2] Department of Computer Science, Tokyo Institute of Technology,
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan. sugi@cs.titech.ac.jp
[3] National Key Laboratory for Novel Software Technology, Nanjing University
Nanjing 210093, P.R. China. zhouzh@nju.edu.cn

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, Breiman raised important questions about the margin explanation by developing a boosting algorithm *arc-gv* that provably generates a larger minimum margin than AdaBoost. He also gave a sharper bound in terms of the minimum margin, and argued that the minimum margin governs the generalization. In experiments however, arc-gv usually performs worse than AdaBoost, putting the margin explanation into serious doubts. In this paper, we try to give a complete answer to Breiman's critique by proving a bound in terms of a new margin measure called Equilibrium margin (Emargin). The Emargin bound is uniformly sharper than Breiman's minimum margin bound. This result suggests that the minimum margin is not crucial for the generalization error. We also show that a large Emargin implies good generalization. Experimental results on benchmark datasets demonstrate that AdaBoost usually has a larger Emargin and a smaller test error than arc-gv, which agrees well with our theory.

1 Introduction

The AdaBoost algorithm [FS96, FS97] has achieved great success in the past ten years. It has demonstrated excellent experimental performance both on benchmark datasets and real applications [BK99, Die00, VJ01]. It is observed in experiments that the test error of a combined voting classifier usually keeps decreasing as its size becomes very large and even after the training error is zero [Bre98, Qui96]. This fact, on the first sight, obviously violates Occam's razor.

Schapire et al. [SFBL98] tried to explain this phenomenon 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. Schapire et al. [SFBL98] 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 demonstrate that AdaBoost has the ability to produce a good margin distribution. This theory indicates that producing a good margin distribution is the key to the success of AdaBoost and explains well the surprising phenomenon observed in experiments.

Soon after that however, Breiman [Bre99] cast serious doubts on this margin explanation. He developed a boosting-type algorithm called arc-gv, which provably generates a larger minimum margin than AdaBoost[1] (Minimum margin is the smallest margin over all the training examples, see Section 2 for the formal definition). Then he 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 by Schapire et al.

Breiman argued that if the bound of Schapire et al. implied that the margin distribution is the key to the generalization error, his bound implied more strongly that the minimum margin is the key to the generalization error, and the arc-gv algorithm would achieve the best performance among all boosting-type algorithms. In experiments, even though arc-gv always produces larger minimum margins than AdaBoost, its test error is consistently higher. Breiman also investigated the margin distributions generated by AdaBoost and arc-gv, and found that arc-gv actually produced uniformly better margin distributions than AdaBoost. Thus he concluded that neither the minimum margin nor the margin distribution determined the generalization error and a new theoretical explanation is needed.

*This work was supported by NSFC(60775005, 60635030, 60721002) and Global COE Program of Tokyo Institute of Technology.

[1] Actually, the minimum margin of arc-gv converges to the largest possible value among all voting classifiers.

Breiman's argument seems convincing and put the margin explanation into serious doubts. Recently however, Reyzin and Schapire [RS06] gained important discovery after a careful study on Breiman's arc-gv algorithm. Note first that the bounds of both Breiman and Schapire et al. state that the generalization error also depends on the complexity of the set of base classifiers as well as the minimum margin or the margin distribution. To investigate how the margin affects the generalization error, one has to keep the complexity of the base classifiers fixed. In Breiman's experiments, he tried to control this by always using CART trees [BFOS84] of a fixed number of leaves as the base classifier. Reyzin and Schapire re-conducted Breiman's experiments and found that the trees produced by arc-gv were much deeper than those produced by AdaBoost. Since deeper trees are more complex even though the number of leaves is the same, arc-gv uses base classifiers of higher complexity than AdaBoost in Breiman's experiments. Thus it was not a fair comparison.

In order to study the margin explanation in a fair manner, a more controlled setting is needed. Reyzin and Schapire then compared arc-gv and AdaBoost by using the decision stump, whose complexity is fixed, as the base classifier. Experiments showed that arc-gv produced larger minimum margins yet still a higher error rate. But this time, the margin distribution generated by arc-gv is not as "good" as that AdaBoost generated (see Fig.7 in [RS06]). So they argued that according to the Schapire et al. bound in terms of the margin distribution, the empirical observation, i.e., the inferior performance of arc-gv, could be explained.

From a more critical point of view however, Breiman's doubt has not been fully answered by the above results. First of all, Breiman backed up his argument with a sharper bound in terms of the minimum margin. In Reyzin and Schapire's experiment with the decision stumps, arc-gv still produced larger minimum margin and had worse performance. Even though AdaBoost generates a "better" margin distribution than arc-gv, it would not disprove Breiman's critique unless we could show a bound in terms of the margin distribution and is uniformly sharper than Breiman's minimum margin bound. Another problem is how to measure the "goodness" of a margin distribution. The statement that AdaBoost generates "better" margin distributions than arc-gv is vague. Reyzin and Schapire used the average margin as a measure to compare margin distributions produced by AdaBoost and arc-gv. But the average margin does not explicitly appear in the bound of Schapire et al. Thus a larger average margin does not necessarily imply a smaller generalization error in theory.

In this paper, we try to give a complete answer to Breiman's doubt by solving the two problems mentioned above. We first propose a novel upper bound for the generalization error of voting classifiers. This bound is uniformly sharper than Breiman's 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 arc-gv when the complexity of the base classifier is well controlled. Our results thus explain the inferior performance of arc-gv and give Breiman's doubt a negative answer.

The rest of this paper is organized as follows: In Section 2 we briefly describe the margin theory of Schapire et al. and Breiman's argument. Our main results are given in Section 3. We provide further explanation of the main bound in Section 4. All the proofs can be found in Section 5. We provide experimental justification in Section 6 and conclude in Section 7.

2 Background and Related Work

In this section we briefly review the existing margin bounds and the two boosting algorithms.

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

$$f(x) = \sum \alpha_i h_i(x),$$

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 $P_D(A(x, y))$ to denote the probability of the event A when an example (x, y) is chosen randomly according to the distribution D. Therefore, $P_D(yf(x) \leq 0)$ is the generalization error which we want to bound. We also use $P_S(A(x, y))$ to denote the probability with respect to choosing an example (x, y) uniformly at random from the training set 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 $P_S(yf(x) \leq \theta)$ as a distribution over θ $(-1 \leq \theta \leq 1)$, since $P_S(yf(x) \leq \theta)$ is the fraction of training examples whose margin is at most θ. This distribution is referred to as the *margin distribution*. The *minimum margin* of f, which is the smallest margin over the training examples, then can

```
Input: $S = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
        where $x_i \in X$, $y_i \in \{-1, 1\}$.
Initialization: $D_1(i) = 1/n$.
for $t = 1$ to $T$ do
    1. Train base learner using distribution $D_t$.
    2. Get base classifier $h_t : X \to \{-1, 1\}$.
    3. Choose $\alpha_t$.
    4. Update:
        $$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t},$$
    where $Z_t$ is a normalization factor chosen so
    that $D_{t+1}$ will be a distribution.
end
Output: The final Classifier
    $$H(x) = \mathrm{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$
```

Algorithm 1: A unified description of AdaBoost and arc-gv.

be equivalently represented by the maximum value of θ such that $P_S(yf(x) \le \theta) = 0$.

A unified description of AdaBoost and arc-gv is shown in Algorithm 1. The only difference of the two algorithms is the choice of α_t. AdaBoost sets α_t as

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

where γ_t is the *edge* of the base classifier h_t, defined as:

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

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

Arc-gv chooses α_t in a different way. It takes into consideration of the minimum margin of the composite classifier up to the current round. Denote by ρ_t the minimum margin of the voting classifier of round $t-1$, that is,

$$\rho_t = \min_i \left(y_i \frac{\sum_{s=1}^{t-1} \alpha_s h_s(x_i)}{\sum_{s=1}^{t-1} \alpha_s} \right).$$

Let

$$\beta_t = \frac{1}{2}\log\frac{1+\gamma_t}{1-\gamma_t} - \frac{1}{2}\log\frac{1+\rho_t}{1-\rho_t}.$$

Arc-gv sets α_t as [Bre99]:

$$\alpha_t = \left\{ \begin{array}{ccl} 1 & : & \beta_t > 1, \\ \beta_t & : & 0 \le \beta_t \le 1, \\ 0 & : & \beta_t < 0. \end{array} \right.$$

The first margin explanation of the AdaBoost algorithm [SFBL98] 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 H is finite, and the other applies to the general case that H has a finite VC dimension.

Theorem 1 *[SFBL98] For any $\delta > 0$, with probability at least $1-\delta$ over the random choice of the training set S of n examples, every voting classifier f satisfies the following bounds:*

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

if $|H| < \infty$. And

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

where d is the VC dimension of H.

The theorem states that if the voting classifier generates a good margin distribution, that is, most training examples have large margins so that $P_S(yf(x) \le \theta)$ is small for not too small θ, then the generalization error is also small. In [SFBL98] it has also been shown that for the AdaBoost algorithm, $P_S(yf(x) \le \theta)$ decreases to zero exponentially fast with respect to the number of boosting iterations if θ is not too large. These results imply that the excellent performance of AdaBoost is due to its good margin distribution.

Breiman's doubts on the margin explanation came from the arc-gv algorithm. It can be shown that the minimum margin generated by arc-gv converges to the largest possible value among all voting classifiers. In practice, arc-gv has larger minimum margins than AdaBoost in most cases for a finite number of boosting iterations. Breiman also proved an upper bound for the generalization error of voting classifiers. This bound depends only on the minimum margin, not on the entire margin distribution.

Theorem 2 *[Bre99] Let θ_0 be the minimum margin defined as*

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

where S is the training set. If

$$|H| < \infty,$$

$$\theta_0 > 4\sqrt{\frac{2}{|H|}},$$

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

then for any $\delta > 0$, with probability at least $1-\delta$ over the random choice of the training set S of n examples, every voting classifier f satisfies the following bounds:

$$P_D\Big(yf(x) \le 0\Big) \le R\left(\log(2n) + \log\frac{1}{R} + 1\right) + \frac{1}{n}\log(\frac{|H|}{\delta}). \quad (2)$$

Breiman 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 $O(\log n/n)$ and the bound in Theorme 1 is $O(\sqrt{\log n/n})$. Breiman then argued that compared to the margin distribution explanation, his bound implied more strongly that the minimum margin governs the generalization error. However, arc-gv performs almost consistently worse than AdaBoost in experiments[2]. These empirical results contradict what the margin theory predicts and therefore put the margin explanation into serious doubts.

A lot of efforts have been made on providing better explanation of the boosting algorithms in recent years [MBG02, KP02, KP05, AKLL02]. Koltchinskii and Panchanko [KP02, KP05] proved a number of bounds in terms of the margin distribution which are sharper than Theorem 1. However, it is difficult to compare the minimum margin bound to these bounds since they contain unspecified constants. Nevertheless, these results imply that the margin distribution might be more important than the minimum margin for the generalization error of voting classifiers.

3 Main Results

In this section we propose upper bounds in terms of the Emargin. The bound is uniformly sharper than Breiman's minimum margin bound.

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

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

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

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

See also [Lan05].

The next theorem is our main result: the Emargin bound. Here we consider the case that the base classifier set H is finite. For the case that H is infinite but has a finite VC dimension, the bound is more complicated and will be given in Theorem 8. All the proofs can be found in Section 5.

Theorem 3 *If $|H| < \infty$, then for any $\delta > 0$, with probability at least $1-\delta$ over the random choice of the training set S of n examples, every voting classifier f satisfies the following bound:*

$$P_D\Big(yf(x) \le 0\Big) \le \frac{\log|H|}{n} + \inf_{q\in\{0,\frac{1}{n},\frac{2}{n},\ldots,1\}} D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right), \quad (3)$$

where

$$u\left[\hat{\theta}(q)\right] = \frac{1}{n}\left(\frac{8}{\hat{\theta}^2(q)}\log\left(\frac{2n^2}{\log|H|}\right)\log|H| + \log|H| + \log\frac{n}{\delta}\right),$$

and $\hat{\theta}(q)$ is given by

$$\hat{\theta}(q) = \sup\left\{\theta \in \left(\sqrt{8/|H|}, 1\right] : P_S\Big(yf(x) \le \theta\Big) \le q\right\}. \quad (4)$$

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**).*

The name *equilibrium* is due to the following fact.

Proposition 5 *q^* is the empirical error at the Emargin θ^*.*

$$P_S\Big(yf(x) < \theta^*\Big) = q^*. \quad (5)$$

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

$$P_D\Big(yf(x) \le 0\Big) \le \frac{\log|H|}{n} + D^{-1}\Big(q^*, u\left(\theta^*\right)\Big). \quad (6)$$

Theorem 3 then states that the generalization error of a voting classifier depends on its Emargin and the empirical error at the Emargin.

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 Eq.(4) one can see that $\hat{\theta}(0)$ is the minimum margin. Hence the Emargin is

[2] Actually, the inferior performance has also been observed when using other voting classifiers that maximize the minimum margin (see also [GS98, RW02]).

equal to the minimum margin if and only if the optimal q^* is zero.

We next compare our Emargin bound to Breiman's minimum margin bound. We show that the Emargin bound is uniformly sharper than the minimum margin bound.

Theorem 6 *The bound given in Theorem 3 is uniformly sharper than the minimum margin bound in Theorem 2. That is*

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

where

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

According to this theorem, the minimum margin is not crucial for the generalization error, i.e., a larger minimum margin does not necessarily imply a smaller test error. Thus arc-gv does not necessarily have better performance than AdaBoost. Our new bound implies that it is the Emargin θ^* and the empirical error q^* at θ^* that govern the performance of the classifier. The following theorem describes how the Emargin θ^* and the Emargin error q^* affect the generalization ability. It states that a larger Emargin and a smaller Emargin error result in a lower generalization error.

Theorem 7 *Let f_1, f_2 be two voting classifiers. Denote by θ_1, θ_2 the Emargin and by q_1, q_2 the empirical error at θ_1, θ_2 of f_1, f_2 respectively. That is*

$$q_i = P_S\Big(yf_i(x) < \theta_i\Big), \qquad i = 1, 2.$$

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

$$B_1 \leq B_2,$$

if

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

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

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

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

where

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

and $\hat{\theta}(q)$ is

$$\hat{\theta}(q) = \sup\left\{\theta \in \big(0, 1\big] : P_S\Big(yf(x) \leq \theta\Big) \leq q\right\}. \quad (7)$$

4 Explanation of the Emargin Bound

In Theorem 3, we adopt 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 $\inf_q D^{-1}(q, u)$ to obtain simpler forms of the Emargin bound. We list in the following lemma the upper bounds of $\inf_q D^{-1}(q, u[\hat{\theta}(q)])$.

Lemma 9 *The following bounds holds.*

1.

$$\inf_q D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right) \leq D^{-1}\left(0, u\left[\hat{\theta}(0)\right]\right) \leq u\left[\hat{\theta}(0)\right].$$

2.

$$\inf_q D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right) \leq \inf_q\left(q + \left(\frac{u\left[\hat{\theta}(q)\right]}{2}\right)^{1/2}\right).$$

3.

$$\inf_q D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right) \leq \inf_{q\leq Cu[\hat{\theta}(q)]} D^{-1}\left(q, u\left[\hat{\theta}(q)\right]\right) \leq \inf_{q\leq Cu[\hat{\theta}(q)]} C'u[\hat{\theta}(q)],$$

where $C > 0$ is any constant and $C' = \max(2C, 8)$.

Note from Theorem 3 that

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

and

$$q = P_S\left(yf(x) \leq \hat{\theta}(q)\right).$$

Thus we can derive the following three bounds from the Emargin bound by using the three inequalities in Lemma 9 respectively.

Corollary 10 *If $|H| < \infty$, then for any $\delta > 0$, with probability at least $1-\delta$ over the random choice of the training set S of n examples, every voting classifier f satisfies the following bounds:*

1.
$$P_D(yf(x) \le 0) \le O\left(\frac{1}{n}\left(\frac{\log n \log |H|}{\theta_0^2} + \log \frac{1}{\delta}\right)\right),$$
where θ_0 is the minimum margin.

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

3.
$$P_D(yf(x) \le 0) \le O\left(\frac{1}{n}\left(\frac{\log n \log |H|}{\theta^2} + \log \frac{1}{\delta}\right)\right),$$
for all θ such that
$$P_S(yf(x) \le \theta) \le O\left(\frac{1}{n}\left(\frac{\log n \log |H|}{\theta^2} + \log \frac{1}{\delta}\right)\right).$$

The first bound in the Corollary has the same order of magnitude as the minimum margin bound. The second bound is the same as Theorem 1. So essentially, previous bounds can be derived from the Emargin bound. The third bound in the Corollary is new. It states that the generalization error is $O(\frac{\log n \log |H|}{n\theta^2})$ even in the non-zero error case, provided the margin error $P_S(yf(x) \le \theta)$ is small enough.

5 Proofs

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

5.1 Proof of Theorem 3

The proof uses the tool developed in [SFBL98]. 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 [SFBL98].

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

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

For any voting classifier

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

where

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

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

$$\begin{aligned} &P_D\Big(yf(x) \le 0\Big) \\ &\le P_{D,g\sim Q(f)}\Big(yg(x) \le \alpha\Big) \\ &\quad + P_{D,g\sim Q(f)}\Big(yg(x) > \alpha, \ yf(x) \le 0\Big) \\ &\le P_{D,g\sim Q(f)}\Big(yg(x) \le \alpha\Big) + \exp\left(-\frac{N\alpha^2}{2}\right). \qquad (8) \end{aligned}$$

We next bound the first term on the right-hand side of the inequality. For any fixed $g \in C_N(H)$, and for any positive number ε and nonnegative integer k such that $k \le n\varepsilon$, 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. A compact representation of this probability is

$$\Pr_{S\sim D^n}\left(P_D(yg(x) \le \alpha) > I\left[P_S\left(yg(x) \le \alpha) > \frac{k}{n}\right)\right] + \varepsilon\right)$$

where $\Pr_{S\sim D^n}$ denotes the probability over n training samples chosen independently at random according to D, and I is the indicator function. Note that

$$\begin{aligned} &\Pr_{S\sim D^n}\Bigg(P_D\Big(yg(x) \le \alpha\Big) \\ &\qquad > I\left[P_S\Big(yg(x) \le \alpha\Big) > \frac{k}{n}\right] + \varepsilon\Bigg) \\ &\le \Pr_{S\sim D^n}\left(P_S\Big(yg(x) \le \alpha\Big) \le \frac{k}{n} \ \middle| \ P_D\Big(yg(x) \le \alpha\Big) > \varepsilon\right) \\ &\le \sum_{r=0}^{k}\binom{n}{r}\varepsilon^r(1-\varepsilon)^{n-r}. \end{aligned}$$

Then applying the relative entropy Chernoff bound to the Bernoulli trials, we further have

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

We thus obtain

$$\begin{aligned} &\Pr_{S\sim D^n}\Bigg(P_D\Big(yg(x) \le \alpha\Big) \\ &\qquad > I\left[P_S\Big(yg(x) \le \alpha\Big) > \frac{k}{n}\right] + \varepsilon\Bigg) \\ &\le \exp\left(-nD\left(\frac{k}{n}\Big\|\varepsilon\right)\right). \qquad (9) \end{aligned}$$

We only consider α at the values in the set

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

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

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

Note that

$$\begin{aligned}
&E_{g\sim Q(f)} P_D\Big(yg(x) \le \alpha\Big) \\
&\qquad = P_{D,g\sim Q(f)}\Big(yg(x) \le \alpha\Big), \\
&E_{g\sim Q(f)} I\left[P_S\Big(yg(x) \le \alpha\Big) > \frac{k}{n}\right] \\
&\qquad = P_{g\sim Q(f)}\left(P_S\Big(yg(x) \le \alpha\Big) > \frac{k}{n}\right).
\end{aligned}$$

We have

$$\begin{aligned}
&\Pr_{S\sim D^n}\Bigg(\exists f \in C(H), \exists \alpha \in U,\ P_{D,g\sim Q(f)}\Big(yg(x) \le \alpha\Big) \\
&\qquad > P_{g\sim Q(f)}\left(P_S(yg(x) \le \alpha) > \frac{k}{n}\right) + \varepsilon\Bigg) \\
&\le |H|^{(N+1)} \exp\left(-nD\left(\frac{k}{n}\Big\|\varepsilon\right)\right).
\end{aligned}$$

Let

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

then

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

We obtain that with probability at least $1-\delta$ over the draw of the training samples, for all $f \in C(H)$, all $\alpha \in U$,

$$\begin{aligned}
&P_{D,g\sim Q(f)}\Big(yg(x) \le \alpha\Big) \\
&\quad \le P_{g\sim Q(f)}\left(P_S\Big(yg(x) \le \alpha\Big) > \frac{k}{n}\right) \\
&\qquad + D^{-1}\left(\frac{k}{n}, \frac{1}{n}\left[(N+1)\log|H| + \log\frac{1}{\delta}\right]\right).
\end{aligned}$$

Using the union bound over $k = 0, 1, \dots, n$, then with probability at least $1-\delta$ over the draw of the training samples, for all $f \in C(H)$, all $\alpha \in U$, and all k

$$\begin{aligned}
&P_{D,g\sim Q(f)}\Big(yg(x) \le \alpha\Big) \\
&\quad \le P_{g\sim Q(f)}\left(P_S\Big(yg(x) \le \alpha\Big) > \frac{k}{n}\right) \\
&\qquad + D^{-1}\left(\frac{k}{n}, \frac{1}{n}\left[(N+1)\log|H| + \log\frac{n}{\delta}\right]\right). \qquad (10)
\end{aligned}$$

We next bound the first term in the right-hand side of Eq.(10). Using the same argument for deriving Eq.(8), we have for any $\theta > \alpha$

$$\begin{aligned}
&P_{g\sim Q(f)}\left(P_S\Big(yg(x) \le \alpha\Big) > \frac{k}{n}\right) \\
&\quad \le I\left[P_S\Big(yf(x) \le \theta\Big) > \frac{k}{n}\right] \\
&\qquad + P_{g\sim Q(f)}\Bigg(P_S\Big(yg(x) > \alpha\Big) > \frac{k}{n}, \\
&\qquad\qquad P_S\Big(yf(x) \le \theta\Big) \le \frac{k}{n}\Bigg). \qquad (11)
\end{aligned}$$

Note that the last term in Eq.(11) can be further bounded by

$$\begin{aligned}
&P_{g\sim Q(f)}\Big(\exists (x_i, y_i) \in S:\ y_i g(x_i) \le \alpha \text{ and } y_i f(x_i) > \theta\Big) \\
&\quad \le n \exp\left(-\frac{N(\theta-\alpha)^2}{2}\right). \qquad (12)
\end{aligned}$$

Combining (8), (10), (11) and (12), we have that with probability at least $1-\delta$ over the draw of training examples, for all $f \in C(H)$, all $\alpha \in U$, all $\theta > \alpha$, and all k, but fixed N

$$\begin{aligned}
&P_D\Big(yf(x) \le 0\Big) \\
&\quad \le \exp\left(-\frac{N\alpha^2}{2}\right) + n\exp\left(-\frac{N(\theta-\alpha)^2}{2}\right) \\
&\qquad + I\left[P_S\Big(yf(x) \le \theta\Big) > \frac{k}{n}\right] \\
&\qquad + D^{-1}\left(\frac{k}{n}, \frac{1}{n}\left[(N+1)\log|H| + \log\frac{n}{\delta}\right]\right).
\end{aligned}$$

Let

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

where $0 \le \eta < 1$. 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

$$\max\left(2n, \exp\left(\frac{N}{2|H|}\right)\right)\exp\left(-\frac{N\theta^2}{8}\right).$$

Let

$$\delta_N = \delta \cdot 2^{-N},$$

we can get a union bound over all N. Put

$$N = \frac{8}{\theta^2}\log\left(\frac{2n^2}{\log|H|}\right),$$

note that if

$$\theta > \sqrt{\frac{8}{|H|}},$$

then

$$2n > \exp\left(\frac{N}{2|H|}\right).$$

We obtain

$$P_D\Big(yf(x) \le 0\Big) \le \frac{\log|H|}{n} + \inf_{0\le k<n}\left(I\left[P_S\Big(yf(x)\le\theta\Big) > \frac{k}{n}\right] + D^{-1}\left(\frac{k}{n}, u\right)\right),$$

where

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

The theorem follows. ■

5.2 Proof of Proposition 5

Let M be the set defined as

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

Let q_0 be the minimal q in M. We will show that

$$q^* = q_0, \tag{13}$$

and

$$P_S\Big(yf(x) < \theta^*\Big) = q_0. \tag{14}$$

To show $q^* = q_0$, note that $D^{-1}(q,u)$ is an increasing function of q for fixed u. Since q^* is the optimal value such that $D^{-1}\left(q, u(\hat{\theta}(q))\right)$ achieves the minimum, one must have $q^* = q_0$.

To show

$$P_S\Big(yf(x) < \theta^*\Big) = q_0,$$

first note that

$$P_S\Big(yf(x) < \theta^*\Big) \in M.$$

For every $q \in M$, by the definition of $\hat{\theta}(q)$, one has

$$P_S\Big(yf(x) < \theta^*\Big) \le q.$$

This implies

$$P_S\Big(yf(x) < \theta^*\Big) = q_0.$$

This completes the proof. ■

5.3 Proof of Theorem 6

The following lemma will be used to prove Theorem 6.

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 the Taylor expansion

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

■

Proof of Theorem 6. The right-hand side of the Emargin bound (3) is the minimum over all $q \in \left\{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\right\}$. Take $q = 0$, it is clear that $\hat{\theta}(0)$ is the minimum margin. By Lemma 9, the Emargin bound can be relaxed to

$$P_D\Big(yf(x)\le 0\Big) \le \frac{1}{n}\left(\frac{8}{\theta_0^2}\log\left(\frac{2n^2}{\log|H|}\right)\log|H| + 2\log|H| + \log\frac{n}{\delta}\right). \tag{15}$$

We 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 (15) is smaller than the minimum margin bound. The theorem then follows. ■

5.4 Proof of of Theorem 7

According to Proposition 5, we have that $q_i = P_S(yf_i(x) < \theta_i)$ is also the optimal q^* in the Emargin bound. Thus we only need to show

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

Note that if $\theta_1 \ge \theta_2$, then $u(\theta_1) \le u(\theta_2)$. So

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

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

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

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

5.5 Proof of Theorem 8

The next lemma is a modified version of the uniform convergence result of [VC71, Vap98] and its refinement [Dev82]. It will be used for proving Theorem 8.

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

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

Define

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

Then for any fixed integer k

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

where

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

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'

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

Or equivalently,

$$\begin{aligned}
&\Pr_{S\sim D^n}\left(\sup_{A\in\mathcal{A}}\left(P_D(A) - I\left[P_S(A) > \frac{k}{n}\right]\right) > \varepsilon\right) \\
&\quad \le \left(\frac{1}{1-e^{-2n'\alpha^2\varepsilon^2}}\right) \Pr_{S\sim D^n,\ S'\sim D^{n'}}\Bigg(\sup_{A\in\mathcal{A}}\Bigg(P_{S'}(A) \\
&\qquad - I\left[P_S(A) > \frac{k}{n}\right]\Bigg) > (1-\alpha)\varepsilon\Bigg). \qquad (16)
\end{aligned}$$

Let V denote the event

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

Let A^* be (one of) the optimal A so that

$$P_D(A) - I\left[P_S(A) > \frac{k}{n}\right]$$

achieves the maximum. Note that the following two events

$$P_{S'}(A^*) \ge P_D(A^*) - \alpha\varepsilon$$

and

$$P_D(A^*) - I\left[P_S(A^*) > \frac{k}{n}\right] > \varepsilon$$

imply that

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

Then

$$\begin{aligned}
&\Pr_{S\sim D^n,\ S'\sim D^{n'}}\Bigg(\sup_{A\in\mathcal{A}}\left(P_{S'}(A) - I\left[P_S(A) > \frac{k}{n}\right]\right) \\
&\qquad\qquad > (1-\alpha)\varepsilon\Bigg) \\
&= \int dP \int I\Bigg[\sup_{A\in\mathcal{A}}\left(P_{S'}(A) - I\left[P_S(A) > \frac{k}{n}\right]\right) \\
&\qquad\qquad > (1-\alpha)\varepsilon\Bigg] dP' \\
&\ge \int_V dP \int I\Bigg[\sup_{A\in\mathcal{A}}\left(P_{S'}(A) - I\left[P_S(A) > \frac{k}{n}\right]\right) \\
&\qquad\qquad > (1-\alpha)\varepsilon\Bigg] dP' \\
&\ge \int_V dP \int I\Bigg[P_{S'}(A^*) - I\left[P_S(A^*) > \frac{k}{n}\right] \\
&\qquad\qquad > (1-\alpha)\varepsilon\Bigg] dP' \\
&\ge \int_V dP \int I\Big[P_{S'}(A^*) \ge P_D(A^*) - \alpha\varepsilon\Big] dP' \\
&\ge \left(1-e^{-2n'\alpha^2\varepsilon^2}\right)\int_V dP \\
&= \left(1-e^{-2n'\alpha^2\varepsilon^2}\right) \\
&\quad \times \Pr_{S\sim D^n}\left(\sup_{A\in\mathcal{A}}\left(P_D(A) - I\left[P_S(A) > \frac{k}{n}\right]\right) > \varepsilon\right).
\end{aligned}$$

This completes the proof of (16).

Take

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

we have

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

Proceeding as [Dev82] and using the relative entropy Hoeffding inequality, the theorem follows. ■

Proof of Theorem 8. The proof is the same as Theorem 3 until we have Eq.(9). Let $\alpha = \frac{\theta}{2}$, we need to bound

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

Note that we only need to consider $\theta = 0, \frac{1}{N}, \frac{2}{N}, \ldots, 1$. Let

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

and

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

By Sauer's lemma [Sau72] it is easy to see that

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

where d is the VC dimension of H. By Lemma 12, we have

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

where

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

Using the argument as Theorem 3, the theorem follows. ■

Proof of Lemma 9. 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 [Hoe63]:

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

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

$$D^{-1}(\frac{q}{2},\frac{q}{8})\le q, \qquad (17)$$

which is equivalent to

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

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

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

and

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

we have

$$D(\frac{q}{2}||q) = \phi(\frac{1}{2})\ge\frac{q}{8}.$$

This completes the proof of Eq. (17).

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

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

The lemma then follows. ■

6 Experiments

In this section we provide experimental results to verify our theory. We compare AdaBoost and arc-gv in terms of their Emargin, Emargin error and the generalization error. Theorem 7 indicates that if a voting classifier f_1 has a larger Emargin and a smaller Emargin error than another classifier f_2, then f_1 would 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 10 benchmark datasets described in Table 1. Except the USPS which contains handwritten digits, all datasets are from the UCI repository [AN07]. If the data is multiclass, we group them into two classes, since we study the binary classification problem. 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 $[0,1]$. All datasets are used in a five-fold cross validation manner. For the USPS which originally has a training set and a test set, we merge them and regenerate the cross validation data.

In all experiments, decision stumps are adopted as the base learner, so the complexity of the base classifiers is well controlled. We use a finite set of possible decision stumps. Specifically, for each feature we consider 100 thresholds uniformly distributed on $[0,1]$. Therefore the size of the base classifier set is $2\times 100\times k$, where k denotes the number of features.

We run AdaBoost and arc-gv for 500 rounds, then calculate the Emargin, Emargin error, test error as well as the minimum margin of them respectively. The results are described in Table 2. AdaBoost has a larger or equal Emargin and a smaller Emargin error than arc-gv on all the datasets except *German* and *Ionosphere*. According to our theory, it predicts that AdaBoost would have a lower generalization error. The experiments show that among these eight datasets, AdaBoost outperforms arc-gv on six datasets, ties on one dataset, and loses

Table 1: Description of the datasets

Dataset	# Examples	# Features	Dataset	# Examples	# Features
Breast	683	9	Letter	20000	16
Diabetes	768	8	Satimage	6435	36
German	1000	24	USPS	9298	256
Image	2310	16	Vehicle	846	20
Ionosphere	351	34	Wdbc	569	30

Table 2: Margin measures and performances of AdaBoost and arc-gv. For the datasets in bold-face, AdaBoost generates larger Emargins and smaller Emargin errors than arc-gv. AdaBoost outperforms arc-gv on all these datasets except the *Image* dataset.

		Emargin	Emargin Error	Test Error	Minimum margin
Breast	AdaBoost	**0.313**	**0.803**	**0.052**	0.005
	arc-gv	0.281	0.909	0.057	0.008
Diabetes	AdaBoost	**0.110**	**0.748**	**0.255**	-0.064
	arc-gv	0.049	0.759	0.256	-0.017
German	AdaBoost	0.157	0.824	0.258	-0.118
	arc-gv	0.034	0.780	0.261	-0.026
Image	AdaBoost	**0.196**	**0.610**	0.023	-0.009
	arc-gv	0.195	0.705	**0.021**	-0.003
Ionosphere	AdaBoost	0.323	0.800	0.100	0.084
	arc-gv	0.131	0.577	0.106	0.061
Letter	AdaBoost	**0.078**	**0.645**	**0.174**	-0.165
	arc-gv	0.063	0.958	0.178	-0.034
Satimage	AdaBoost	**0.133**	**0.521**	**0.053**	-0.054
	arc-gv	0.133	0.956	0.057	-0.019
USPS	AdaBoost	**0.108**	**0.972**	**0.450**	-0.142
	arc-gv	0.053	0.990	0.460	-0.024
Vehicle	AdaBoost	**0.129**	**0.737**	**0.297**	-0.117
	arc-gv	0.052	0.794	0.304	-0.033
Wdbc	AdaBoost	**0.350**	**0.581**	**0.035**	-0.130
	arc-gv	0.350	0.710	**0.035**	-0.100

only on one dataset. These results agree well with our theory.

Note also that on all the datasets except *Ionosphere*, arc-gv has a larger minimum margin than AdaBoost, but arc-gv has a lower test error than AdaBoost only on one dataset. This verifies that the minimum margin is not crucial for the generalization error.

7 Conclusions

In this paper we tried to give a complete answer to Breiman's doubt on the margin explanation of the AdaBoost algorithm. 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 used by Breiman to back up his argument. According to our theory, arc-gv does not necessarily outperform AdaBoost even though it generates larger minimum margins.

Our bounds also imply that the Emargin and the Emargin error are the key to the generalization error of a voting classifier—a larger Emargin and a smaller Emargin error result in better generalization ability. Experiments on benchmark datasets agree well with our theory.

A future work is to study why AdaBoost generates larger Emargins and smaller Emargin errors, i.e., better margin distributions, than arc-gv. Can we find a strategy that optimizes the margin distribution? If such an algorithm exists, it would be a good test of our theory to see whether it has better performance than AdaBoost as we predict.

References

[AKLL02] A. Antos, B. Kégl, T. Linder, and G. Lugosi. Data-dependent margin-based generalization bounds for classification. *Journal of Machine Learning Research*, 3:73–98, 2002.

[AN07] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007.

[BFOS84] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[BK99] E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting and variants. *Machine Learning*, 36:105–139, 1999.

[Bre98] L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26:801–849, 1998.

[Bre99] L. Breiman. Prediction games and arcing algorithms. *Neural Computation*, 11:1493–1517, 1999.

[Dev82] L. Devroye. Bounds for the uniform deviation of empirical measures. *Journal of Multivariate Analysis*, 12:72–79, 1982.

[Die00] T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning*, 40:139–157, 2000.

[FS96] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, 1996.

[FS97] 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.

[GS98] A. J. Grove and D. Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In *National Conference on Artificial Intelligence*, 1998.

[Hoe63] W. Hoeffding. Probability inequalities for sum of bounded random variables. *Journal of American Statistical Society*, 58:13–30, 1963.

[KP02] V. Koltchinskii and D. Panchanko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30:1–50, 2002.

[KP05] V. Koltchinskii and D. Panchanko. Complexities of convex combinations and bounding the generalization error in classification. *Annals of Statistics*, 33:1455–1496, 2005.

[Lan05] J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

[MBG02] L. Mason, P. Bartlett, and M. Golea. Generalization error of combined classifiers. *Journal of Computer and System Sciences*, 65:415–438, 2002.

[Qui96] J. R. Quinlan. Bagging, boosting, and c4.5. In *13th International Conference on Artificial Intelligence*, 1996.

[RS06] L. Reyzin and R. E. Schapire. How boosting the margin can also boost classifier complexity. In *International Conference on Machine Learning*, 2006.

[RW02] G. Rätsch and M. Warmuth. Maximizing the margin with boosting. In *15th Annual Conference on Computational Learning Theory*, 2002.

[Sau72] N. Sauer. On the density of family of sets. *Journal of Combinatorial Theory, Series A*, 13:145–147, 1972.

[SFBL98] 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.

[Vap98] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons Inc., 1998.

[VC71] 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.

[VJ01] 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.