Artificial Intelligence 203 (2013) 1–18

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

On the doubt about margin explanation of boosting

Wei Gao, Zhi-Hua Zhou *

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

ARTICLE INFO

Article history:
Received 8 May 2012
Received in revised form 4 July 2013
Accepted 13 July 2013
Available online 24 July 2013

Keywords:
Classification
Boosting
Ensemble methods
Margin theory

ABSTRACT

Margin theory provides one of the most popular explanations to the success of AdaBoost, where the central point lies in the recognition that *margin* is the key for characterizing the performance of AdaBoost. This theory has been very influential, e.g., it has been used to argue that AdaBoost usually does not overfit since it tends to enlarge the margin even after the training error reaches zero. Previously the *minimum margin bound* was established for AdaBoost, however, Breiman (1999) [9] pointed out that maximizing the minimum margin does not necessarily lead to a better generalization. Later, Reyzin and Schapire (2006) [37] emphasized that the margin distribution rather than minimum margin is crucial to the performance of AdaBoost. In this paper, we first present the *kth margin bound* and further study on its relationship to previous work such as the minimum margin bound and Emargin bound. Then, we improve the previous empirical Bernstein bounds (Audibert et al. 2009; Maurer and Pontil, 2009) [2,30], and based on such findings, we defend the margin-based explanation against Breiman's doubts by proving a new generalization error bound that considers exactly the same factors as Schapire et al. (1998) [39] but is sharper than Breiman's (1999) [9] minimum margin bound. By incorporating factors such as average margin and variance, we present a generalization error bound that is heavily related to the whole margin distribution. We also provide margin distribution bounds for generalization error of voting classifiers in finite VC-dimension space.

1. Introduction

The AdaBoost algorithm [18,19], which aims to construct a "strong" classifier by combining some "weak" learners (slightly better than random guess), is a representative of ensemble methods [47] and has been one of the most influential classification algorithms [13,45], and it has exhibited excellent performance both on benchmark datasets and real applications [5,16].

Many studies are devoted to understanding the mysteries behind the success of AdaBoost, among which the margin theory proposed by Schapire et al. [39] has been very influential. For example, AdaBoost often tends to be empirically resistant (but not completely) to overfitting [8,17,35], i.e., the generalization error of the combined learner keeps decreasing as its size becomes very large and even after the training error has reached zero; it seems violating the Occam's razor [7], i.e., the principle that less complex classifiers should perform better. This remains one of the most famous mysteries of AdaBoost. The margin theory provides the most intuitive and popular explanation to this mystery, that is: AdaBoost tends to improve the margin even after the error on training sample reaches zero.

However, Breiman [9] raised serious doubt on the margin theory by designing arc-gv, a boosting-style algorithm. This algorithm is able to maximize the *minimum margin*, i.e., the smallest margin over the training data (the formal definition

* Corresponding author.
E-mail address: zhouzh@lamda.nju.edu.cn (Z.-H. Zhou).

0004-3702/$ – see front matter
http://dx.doi.org/10.1016/j.artint.2013.07.002

will be given in Eq. (2)), but its generalization error is high on empirical datasets, and similar experimental evidence has also been observed in [22]. Thus, Breiman [9] concluded that the margin theory for AdaBoost failed. Breiman's argument was backed up with a minimum margin bound, which is sharper than the generalization bound given by Schapire et al. [39], and a lot of experiments. Garg and Roth [21] presented a margin-distribution algorithm based on a data-dependent complexity measure. Later, Reyzin and Schapire [37] found that there were flaws in the design of experiments: Breiman used CART trees [11] as base learners and fixed the number of leaves for controlling the complexity of base learners. However, Reyzin and Schapire [37] found that the trees produced by arc-gv were usually much deeper than those produced by AdaBoost. Generally, for two trees with the same number of leaves, the deeper one is with a larger complexity because more judgments are needed for making a prediction. Therefore, Reyzin and Schapire [37] concluded that Breiman's observation was biased due to the poor control of model complexity. They repeated the experiments by using decision stumps for base learners, considering that decision stump has exactly two leaves and thus with a fixed complexity, and observed that though arc-gv produced a larger minimum margin, its margin distribution was quite poor. Nowadays, it is well-accepted that the margin distribution is crucial to relate margin to the generalization performance of AdaBoost. To support the margin theory, Wang et al. [44] presented a sharper bound in term of *Emargin* (the formal definition will be given in Theorem 3), which was believed to be relevant to margin distribution.

In this paper, we first present the *kth margin* bound and further study its relationship to previous work such as the minimum margin bound and Emargin bound. Then, by using empirical Bernstein bounds, we present a new generalization error bound for voting classifier, which considers exactly the same factors as Schapire et al. [39], but is sharper than the bounds of Schapire et al. [39] and Breiman [9]. Therefore, we defend the margin-based explanation against Breiman's doubt. Moreover, we provide a generalization error bound, by incorporating other factors such as average margin and variance, which are heavily relevant to the whole margin distribution. We also give a margin distribution bound for generalization error of voting classifiers in finite VC-dimension space. It is also worth mentioning that our new empirical Bernstein bounds improve the main results of [2,30], with a simpler proof, and we present empirical Bernstein bounds for finite VC-dimension space; these results can be interesting, independently to the main purpose of the paper, to the machine learning community.

The rest of this paper is organized as follows. We begin with some notations and background in Sections 2 and 3, respectively. Then, we prove the kth margin bound and discuss on its relation to previous bounds in Section 4. Our main results are presented in Section 5, and detailed proofs are provided in Section 6. We conclude in Section 7.

2. Notations

Let $\mathcal{X}$ and $\mathcal{Y}$ denote an input space and output space, respectively. In this paper, we focus on binary classification problems, i.e., $\mathcal{Y} = \{+1, -1\}$. Denote by D an (unknown) underlying probability distribution over the product space $\mathcal{X} \times \mathcal{Y}$. A training sample of size m

$$S = \big\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\big\}$$

is drawn independently and identically (i.i.d.) according to the distribution D. We use $\Pr_D[\cdot]$ to refer as the probability with respect to D, and $\Pr_S[\cdot]$ to denote the probability with respect to uniform distribution over the sample S. Similarly, we use $E_D[\cdot]$ and $E_S[\cdot]$ to denote the expected values, respectively. For an integer $m > 0$, we set $[m] = \{1, 2, \ldots, m\}$.

The Bernoulli Kullback–Leibler (or KL) divergence is defined as

$$KL(q\|p) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} \quad \text{for } 0 \leqslant p, q \leqslant 1.$$

For a fixed q, we can easily find that $KL(q\|p)$ is a monotone increasing function for $q \leqslant p < 1$, and thus, the inverse of $KL(q\|p)$ for the fixed q is given by

$$KL^{-1}(q; u) = \inf_w \big\{w\colon\ w \geqslant q \text{ and } KL(q\|w) \geqslant u\big\}.$$

Let $\mathcal{H}$ be a hypothesis space. A base learner is a function which maps a distribution over $\mathcal{X} \times \mathcal{Y}$ onto a function $h : \mathcal{X} \to \mathcal{Y}$. In this paper, we only focus on binary base classifiers, i.e., the outputs are in $\{-1, 1\}$. Let $\mathcal{C}(\mathcal{H})$ denote the convex hull of H, i.e., a voting classifier $f \in \mathcal{C}(\mathcal{H})$ is of the following form

$$f = \sum \alpha_i h_i \quad \text{with } \sum \alpha_i = 1 \text{ and } \alpha_i \geqslant 0.$$

For $N \geqslant 1$, denote by $\mathcal{C}_N(\mathcal{H})$ the set of unweighted averages over N elements from $\mathcal{H}$, that is

$$\mathcal{C}_N(\mathcal{H}) = \left\{g\colon\ g = \sum_{j=1}^{N} \frac{h_j}{N},\ h_j \in \mathcal{H}\right\}. \tag{1}$$

For voting classifier $f \in \mathcal{C}(\mathcal{H})$, we can associate with a distribution over $\mathcal{H}$ by using the coefficients $\{\alpha_i\}$, denoted by $\mathcal{Q}(f)$. For convenience, $g \in \mathcal{C}_N(\mathcal{H}) \sim \mathcal{Q}(f)$ implies $g = \sum_{j=1}^{N} h_j/N$ where $h_j \sim \mathcal{Q}(f)$.

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

Input: Sample $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ and the number of iterations T.

Initialization: $D_1(i) = 1/m$.

for $t = 1$ to T **do**

1. Construct base learner $h_t : \mathcal{X} \to \mathcal{Y}$ using the distribution D_t.
2. Choose α_t.
3. Update

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

where Z_t is a normalization factor (such that D_{t+1} is a distribution).

end for

Output: The final classifier $\operatorname{sgn}[f(x)]$, where

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

For an example (x, y), the *margin* with respect to the voting classifier $f = \sum \alpha_i h_i(x)$ is defined as $yf(x)$; in other words,

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

which shows the difference between the weights of base learners that classify (x, y) correctly and the weights of base learners that misclassify (x, y). Therefore, margin can be viewed as a measure of the confidence of the classification. Given a sample $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, we denote by $\hat{y}_1 f(\hat{x}_1)$ the *minimum margin* and $E_S[yf(x)]$ the *average margin*, which are defined respectively as follows:

$$\hat{y}_1 f(\hat{x}_1) = \min_{i\in[m]}\big\{y_i f(x_i)\big\} \quad \text{and} \quad E_S\big[yf(x)\big] = \sum_{i=1}^{m} \frac{y_i f(x_i)}{m}. \tag{2}$$

3. Background

In the statistics community, great efforts have been devoted to understanding how and why `AdaBoost` works. Friedman et al. [20] made an important stride by viewing `AdaBoost` as a stagewise optimization and relating it to fitting an additive logistic regression model. Various new boosting-style algorithms were developed by performing a gradient decent optimization of some potential loss functions [12,28,36]. Based on this optimization view, some boosting-style algorithms and their variants have been shown to be Bayes's consistent under different settings [3,4,6,10,24,27,34,46], i.e., those studies theoretically ensure that boosting is asymptotically convergent to the Bayes's classifiers. However, such theories cannot be used to explain the resistance of `AdaBoost` to overfitting for small sample problems, and some statistical views have been questioned by Mease and Wyner [33] with empirical evidences. In this paper, we focus on margin theory.

Algorithm 1 provides a unified description of `AdaBoost` and `arc-gv`. The only difference between them lies in the choice of α_t. In `AdaBoost`, α_t is chosen by

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

where $\gamma_t = \sum_{i=1}^{m} D_t(i) y_i h_t(x_i)$ is called the *edge* of h_t, which is an affine transformation of the error rate of $h_t(x)$. However, `Arc-gv` sets α_t in a different way. Denote by ρ_t the minimum margin of the voting classifier of round $t-1$, that is,

$$\rho_t = \hat{y}_1 f_t(\hat{x}_1) \quad \text{with } \rho_1 = 0$$

where

$$f_t = \sum_{s=1}^{t-1} \frac{\alpha_s}{\sum_{s=1}^{t-1}\alpha_s} h_s(x).$$

Then, `Arc-gv` sets α_t as to be

$$\alpha_t = \frac{1}{2}\ln\frac{1+\gamma_t}{1-\gamma_t} - \frac{1}{2}\ln\frac{1+\rho_t}{1-\rho_t}.$$

Schapire et al. [39] proposed the first margin theory for `AdaBoost` and upper bounded the generalization error as follows:

Theorem 1. *(See [39].) For any $\delta > 0$ and $\theta > 0$, with probability at least $1 - \delta$ over the random choice of sample S with size m, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound*:

$$\Pr_D\big[yf(x) < 0\big] \leqslant \Pr_S\big[yf(x) \leqslant \theta\big] + O\left(\frac{1}{\sqrt{m}}\left(\frac{\ln m \ln|\mathcal{H}|}{\theta^2} + \ln\frac{1}{\delta}\right)^{1/2}\right).$$

Breiman [9] provided the minimum margin bound for `arc-gv` by Theorem 2 with our notations.

Theorem 2. *(See [9].) If*

$$\theta = \hat{y}_1 f(\hat{x}_1) > 4\sqrt{\frac{2}{|\mathcal{H}|}} \quad \text{and} \quad R = \frac{32\ln 2|\mathcal{H}|}{m\theta^2} \leqslant 2m,$$

then, for any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of sample S with size m, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound:

$$\Pr_D\big[yf(x) < 0\big] \leqslant R\left(\ln(2m) + \ln\frac{1}{R} + 1\right) + \frac{1}{m}\ln\frac{|\mathcal{H}|}{\delta}.$$

Empirical results show that `arc-gv` probably generates a larger minimum margin but with higher generalization error, and Breiman's minimum bound is $O(\ln m/m)$, sharper than $O(\sqrt{\ln m/m})$ in Theorem 1. Thus, Breiman cast serious doubt on margin theory. To support the margin theory, Wang et al. [44] presented a sharper bound in term of *Emargin* by Theorem 3, which was believed to be related to margin distribution. Notice that the factors considered by Wang et al. [44] are different from that considered by Schapire et al. [39] and Breiman [9].

Theorem 3. *(See [44].) If $8 < |\mathcal{H}| < \infty$, then for any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of the training set S of size $m > 1$, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ such that*

$$q_0 = \Pr_S\big[yf(x) \leqslant \sqrt{8/|\mathcal{H}|}\,\big] < 1 \tag{3}$$

satisfies the following bound:

$$\Pr_D\big[yf(x) < 0\big] \leqslant \frac{\ln|\mathcal{H}|}{m} + \inf_{q\in\{q_0, q_0+\frac{1}{m},\ldots,1\}} KL^{-1}\big(q; u\big[\hat{\theta}(q)\big]\big),$$

where

$$u\big[\hat{\theta}(q)\big] = \frac{1}{m}\left(\frac{8\ln|\mathcal{H}|}{\hat{\theta}^2(q)}\ln\frac{2m^2}{\ln|\mathcal{H}|} + \ln|\mathcal{H}| + \ln\frac{m}{\delta}\right)$$

and $\hat{\theta}(q) = \sup\{\theta \in (\sqrt{8/|\mathcal{H}|}, 1]\colon \Pr_S[yf(x) \leqslant \theta] \leqslant q\}$. Also, the Emargin is given by $\theta^* \in \arg\inf_{q\in\{q_0, q_0+\frac{1}{m},\ldots,1\}} KL^{-1}(q; u[\hat{\theta}(q)])$.

Instead of the whole function space, much work developed margin-based data-dependent bounds for generalization error, e.g., empirical cover number [40], empirical fat-shattering dimension [1], Rademacher and Gaussian complexities [25,26], etc. Some of these bounds are proven to be sharper than Theorem 1, but it is hard to show that these bounds are sharper than the bounds of Theorems 2 and 3, and fail to explain the resistance of `AdaBoost` to overfitting.

4. The *k*th margin bounds

Given a sample S of size m, we define the *kth margin* $\hat{y}_k f(\hat{x}_k)$ as the kth smallest margin over sample S, i.e., the kth smallest value in $\{y_i f(x_i),\ i \in [m]\}$. The following theorem shows that the kth margin can be used to measure the performance of a voting classifier, whose proof is deferred in Section 6.1.

Theorem 4. *For any $\delta > 0$ and $k \in [m]$, if $\theta = \hat{y}_k f(\hat{x}_k) > \sqrt{8/|\mathcal{H}|}$, then with probability at least $1 - \delta$ over the random choice of sample with size m, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound*:

$$\Pr_D\big[yf(x) < 0\big] \leqslant \frac{\ln|\mathcal{H}|}{m} + KL^{-1}\left(\frac{k-1}{m}; \frac{q}{m}\right), \tag{4}$$

where

$$q = \frac{8\ln(2|\mathcal{H}|)}{\theta^2}\ln\frac{2m^2}{\ln|\mathcal{H}|} + \ln|\mathcal{H}| + \ln\frac{m}{\delta}.$$

Particularly, when k is constant with $m > 4k$, we have

$$\Pr_D\big[yf(x) < 0\big] \leqslant \frac{\ln|\mathcal{H}|}{m} + \frac{2}{m}\left(\frac{8\ln(2|\mathcal{H}|)}{\theta^2}\ln\frac{2m^2}{\ln|\mathcal{H}|} + \ln|\mathcal{H}| + \ln\frac{km^{k-1}}{\delta}\right). \tag{5}$$

Here, we present the kth margin bound to link previous results on margin bounds, and it is interesting to study the relation between Theorem 4 and previous results, especially Theorems 2 and 3. It is straightforward to get a result similar to Breiman's minimum margin bound in Theorem 2, by setting $k = 1$ in Eq. (5):

Corollary 1. *For any $\delta > 0$, if $\theta = \hat{y}_1 f(\hat{x}_1) > \sqrt{8/|\mathcal{H}|}$, then with probability at least $1-\delta$ over the random choice of sample S with size m, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound*:

$$\Pr_D\big[yf(x) < 0\big] \leqslant \frac{\ln|\mathcal{H}|}{m} + \frac{2}{m}\left(\frac{8\ln(2|\mathcal{H}|)}{\theta^2}\ln\frac{2m^2}{\ln|\mathcal{H}|} + \ln\frac{|\mathcal{H}|}{\delta}\right).$$

Notice that when k is a constant, the bound in Eq. (5) is $O(\ln m/m)$ and the only difference lies in the coefficient. Thus, there is no essential difference to select constant kth margin (such as the 2nd margin, the 3rd margin, etc.) to measure the confidence of classification for large-size sample.

Based on Theorem 4, it is not difficult to get a result similar to the Emargin bound in Theorem 3 as follows:

Corollary 2. *For any $\delta > 0$, if $\theta_k = \hat{y}_k f(\hat{x}_k) > \sqrt{8/|\mathcal{H}|}$, then with probability at least $1-\delta$ over the random choice of the sample S with size m, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound*:

$$\Pr_D\big[yf(x) < 0\big] \leqslant \frac{\ln|\mathcal{H}|}{m} + \inf_{k\in[m]} KL^{-1}\left(\frac{k-1}{m}; \frac{q}{m}\right),$$

where

$$q = \frac{8\ln(2|\mathcal{H}|)}{\theta_k^2}\ln\frac{2m^2}{\ln|\mathcal{H}|} + \ln|\mathcal{H}| + \ln\frac{m}{\delta}.$$

From this corollary, we can easily understand that the Emargin bound ought to be tighter than the minimum margin bound because the former takes the infimum over all $k \in [m]$ while the latter only focuses on the minimum margin. Intuitively, the bound of Corollary 2 might be sharper than that of Corollary 1 if the minimum margin is very small whereas some kth margin is very large. We also notice that, as shown by Eq. (2), the minimum margin can also be expressed as taking the infimum over all margin, whereas it is well accepted that the minimum margin bound is a single-margin bound.

5. Main results

We begin with the standard deviation bounds as follows:

Theorem 5. *For independent random variables $X_1, X_2, \ldots, X_m$ ($m \geqslant 5$) with values in $[0,1]$, and for $\delta \in (0,1)$, we have*

$$\Pr\left[\sqrt{E[\hat{V}_m]} < \sqrt{\hat{V}_m} - \sqrt{\frac{\ln 1/\delta}{4m}}\right] \leqslant \delta, \tag{6}$$

$$\Pr\left[\sqrt{E[\hat{V}_m]} > \sqrt{\hat{V}_m} + \sqrt{\frac{2\ln 1/\delta}{m}}\right] \leqslant \delta, \tag{7}$$

where the sample variance $\hat{V}_m = \sum_{i\neq j}(X_i - X_j)^2/2m(m-1)$.

The detailed proof is presented in Section 6.2. This theorem improves the results of [30, Theorem 10], especially for Eq. (6). Based on this result, we can derive the following empirical Bernstein bounds, with proof deferred to Section 6.3.

Theorem 6. *For independent random variables $X_1, X_2, \ldots, X_m$ ($m \geqslant 5$) with values in $[0,1]$, and for $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\frac{1}{m}\sum_{i=1}^m E[X_i] - \frac{1}{m}\sum_{i=1}^m X_i \leqslant \sqrt{\frac{2\hat{V}_m\ln(2/\delta)}{m}} + \frac{7\ln(2/\delta)}{3m}, \tag{8}$$

$$\frac{1}{m}\sum_{i=1}^m E[X_i] - \frac{1}{m}\sum_{i=1}^m X_i \geqslant -\sqrt{\frac{2\hat{V}_m\ln(2/\delta)}{m}} - \frac{7\ln(2/\delta)}{3m}, \tag{9}$$

where $\hat{V}_m = \sum_{i\neq j}(X_i - X_j)^2/2m(m-1)$.

For identical and independent distribution (i.i.d.) variables, we have

Corollary 3. *For i.i.d. random variables $X, X_1, X_2, \ldots, X_m$ ($m \geqslant 5$) with values in $[0,1]$, and for $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$E[X] - \frac{1}{m}\sum_{i=1}^{m} X_i \leqslant \sqrt{\frac{2\hat{V}_m \ln(2/\delta)}{m}} + \frac{7\ln(2/\delta)}{3m},$$

$$E[X] - \frac{1}{m}\sum_{i=1}^{m} X_i \geqslant -\sqrt{\frac{2\hat{V}_m \ln(2/\delta)}{m}} - \frac{7\ln(2/\delta)}{3m},$$

where $\hat{V}_m = \sum_{i\neq j}(X_i - X_j)^2/2m(m-1)$.

There are two results [2,30] closely related to Theorem 6 (or Corollary 3). Audibert et al. [2] presented the first empirical Bernstein bound and applied to analyze multi-armed bandit algorithms. Soon after, Maurer and Pontil [30] improved the constants and explored the sample variance penalization methods. Comparing with these results, our bounds in Eqs. (8) and (9) are with better constants and the technique of proof is simpler.

Based on this Corollary 3, we can derive the following corollary for the finite function space:

Corollary 4. *Let $S = \{X_1, \ldots, X_m\}$ ($m \geqslant 5$) be drawn i.i.d. from a distribution $\mathcal{D}$ over $\mathcal{X}$, and let $\mathcal{H} = \{h : \mathcal{X} \to [0,1]\}$ be a finite function space. For any $\delta \in (0,1)$, every $h \in \mathcal{H}$ satisfies the following bound with probability at least $1-\delta$:*

$$E_{\mathcal{D}}\big[h(X)\big] - \frac{1}{m}\sum_{i=1}^{m} h(X_i) \leqslant \sqrt{\frac{2\hat{V}_m(h)\ln(2|\mathcal{H}|/\delta)}{m}} + \frac{7\ln(2|\mathcal{H}|/\delta)}{3m}$$

where $\hat{V}_m(h) = \sum_{i\neq j}(h(X_i) - h(X_j))^2/2m(m-1)$.

Then, we get a new generalization bound for infinite hypothesis space with finite VC-dimension, with proof deferred to Section 6.4.

Theorem 7. *Let $S = \{X_1, \ldots, X_m\}$ ($m \geqslant 5$) be drawn i.i.d. from a distribution $\mathcal{D}$ over $\mathcal{X}$, and let $\mathcal{H} = \{h : \mathcal{X} \to \{0,1\}\}$ be a hypothesis space with finite VC-dimension d. For any $\delta \in (0,1)$, every $h \in \mathcal{H}$ satisfies the following bound with probability at least $1-\delta$:*

$$E_{\mathcal{D}}\big[h(X)\big] - \sum_{i=1}^{m}\frac{h(X_i)}{m} \leqslant \sqrt{\frac{2\hat{V}_m(h)}{m}\left(d\ln\frac{2m}{d} + \ln\frac{8}{\delta}\right)} + \frac{19}{3m}\left(d\ln\frac{2m}{d} + \ln\frac{8}{\delta}\right)$$

where $\hat{V}_m(h) = \sum_{i\neq j}(h(X_i) - h(X_j))^2/2m(m-1)$.

We now present our first margin bound for AdaBoost as follows:

Theorem 8. *For any $\delta > 0$, with probability at least $1-\delta$ over the random choice of sample S with size $m \geqslant 5$, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound*:

$$\Pr_{D}\big[yf(x) < 0\big] \leqslant \frac{2}{m} + \inf_{\theta\in(0,1]}\left[\Pr_{S}\big[yf(x) < \theta\big] + \frac{7\mu + 3\sqrt{3\mu}}{3m} + \sqrt{\frac{3\mu}{m}\Pr_{S}\big[yf(x) < \theta\big]}\right]$$

where

$$\mu = \frac{8}{\theta^2}\ln m \ln(2|\mathcal{H}|) + \ln\frac{2|\mathcal{H}|}{\delta}.$$

This proof is based on the techniques developed by Schapire et al. [39], and the main difference is that we utilize the empirical Bernstein bound of Eq. (8) in Theorem 6 for the derivation of generalization error. The detailed proof is deferred to Section 6.5.

It is noteworthy that Theorem 8 shows that the generalization error can be bounded in term of the empirical margin distribution $\Pr_S[yf(x) \leqslant \theta]$, the training sample size and the hypothesis complexity; in other words, this bound considers exactly the same factors as Schapire et al. [39] in Theorem 1. However, the following corollary shows that, the bound in Theorem 8 is sharper than the bound of Schapire et al. [39] in Theorem 1, as well as the minimum margin bound of Breiman [9] in Theorem 2.

Corollary 5. *For any $\delta > 0$, if the minimum margin $\theta_1 = \hat{y}_1 f(\hat{x}_1) > 0$ and $m \geqslant 5$, then we have*

$$\inf_{\theta\in(0,1]}\left[\Pr_S\big[yf(x) < \theta\big] + \frac{7\mu + 3\sqrt{3\mu}}{3m} + \sqrt{\frac{3\mu}{m}\Pr_S\big[yf(x) < \theta\big]}\right] \leqslant 7\mu_1/3m + \sqrt{3\mu_1}/m, \tag{10}$$

where $\mu = 8\ln m \ln(2|\mathcal{H}|)/\theta^2 + \ln(2|\mathcal{H}|/\delta)$ and $\mu_1 = 8\ln m \ln(2|\mathcal{H}|)/\theta_1^2 + \ln(2|\mathcal{H}|/\delta)$; moreover, if

$$\theta_1 = \hat{y}_1 f(\hat{x}_1) > 4\sqrt{\frac{2}{|\mathcal{H}|}}, \tag{11}$$

$$R = \frac{32\ln 2|\mathcal{H}|}{m\theta_1^2} \leqslant 2m, \tag{12}$$

$$m \geqslant \max\left\{4, \exp\left(\frac{\theta_1^2}{4\ln(2|\mathcal{H}|)}\ln\frac{|\mathcal{H}|}{\delta}\right)\right\}, \tag{13}$$

then we have

$$\frac{2}{m} + \inf_{\theta\in(0,1]}\left[\Pr_S\big[yf(x) < \theta\big] + \frac{7\mu + 3\sqrt{3\mu}}{3m} + \sqrt{\frac{3\mu}{m}\Pr_S\big[yf(x) < \theta\big]}\right] \leqslant R\left(\ln(2m) + \ln\frac{1}{R} + 1\right) + \frac{1}{m}\ln\frac{|\mathcal{H}|}{\delta}. \tag{14}$$

This proof is deferred to Section 6.6. From Eq. (10), we can see clearly that the bound of Theorem 8 is $O(\ln m/m)$, sharper than the bound of Schapire et al. [39] $O(\sqrt{\ln m/m})$ in Theorem 1. In fact, we could also guarantee that bound of Theorem 8 is $O(\ln m/m)$ even under weaker assumption that $\hat{y}_k f(\hat{x}_k) > 0$ for some $k \leqslant O(\ln m)$.

It is also noteworthy Eqs. (11) and (12) are the conditions of Theorem 2, and the term $\exp(\frac{\theta_1^2}{4\ln(2|\mathcal{H}|)}\ln\frac{|\mathcal{H}|}{\delta}) \leqslant (\frac{e}{\delta})^{\frac{1}{4}}$ in Eq. (13), which is small for many real applications, e.g., it is less than 13 even if $\delta = 0.0001$. Eq. (14) shows that the bound of Theorem 8 is sharper than Breiman's minimum margin bound of Theorem 2.

Breiman [9] doubted the margin theory because of two recognitions: (i) the minimum margin bound of Breiman [9] is sharper than the margin distribution bound of Schapire et al. [39], and therefore, the minimum margin is more essential than margin distribution to characterize the generalization performance; (ii) `arc-gv` maximizes the minimum margin, but demonstrates worse performance than `AdaBoost` empirically. However, our result shows that the margin distribution bound in Theorem 1 can be greatly improved such that it is even sharper than the minimum margin bound, and therefore, it is natural that `AdaBoost` outperforms `arc-gv` empirically on some datasets; in a word, our results provide a complete answer to Breiman's doubt on margin theory.

The Emargin bounds of Wang et al. [44] are also proven to be sharper than those of Schapire et al. [39] and Breiman [9]. The main difference between Theorem 8 and the Emargin bounds lies in the consideration of different factors for margin theory, e.g., Theorem 8 considers exactly the same factors as Schapire et al. [39], whereas Wang et al. [44] considered the Emargin as the key factor. Moreover, Theorem 8 is advantageous in that its margin interval is wider than that of Emargin.[1] Note that it is not easy to directly compare Theorem 8 and the Emargin bounds because it is difficult to get a closed-form for the $D^{-1}(p\|q)$ term contained in the Emargin bounds, whereas Theorem 8 is relatively easier to estimate.

It is well-accepted that the margin distribution is crucial to relate margin to the generalization performance of `AdaBoost`, whereas it is unclear how to measure the "goodness" of a margin distribution. The first-order and second-order statistics, i.e., the average margin and variance, are natural and intuitive measures. Indeed, Reyzin and Schapire [37] have recommended to take the average margin for a characterization of the margin distribution. However, there is no theory, to the best of our knowledge, to support that a larger average margin or a smaller variance implies a smaller generalization error. The following theorem fills the gap for such theory:

Theorem 9. *For any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of sample S with size $m \geqslant 5$, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound*:

$$\Pr_D\big[yf(x) < 0\big] \leqslant \frac{1}{m^{50}} + \inf_{\theta\in(0,1]}\left[\Pr_S\big[yf(x) < \theta\big] + m^{-2/(1-E_S^2[yf(x)]+\theta/9)} + \frac{3\sqrt{\mu}}{m^{3/2}} + \frac{7\mu}{3m} + \sqrt{\frac{3\mu}{m}\hat{\mathcal{I}}(\theta)}\right]$$

where

$$\mu = 144\ln m\ln\big(2|\mathcal{H}|\big)/\theta^2 + \ln\big(2|\mathcal{H}|/\delta\big),$$
$$\hat{\mathcal{I}}(\theta) = \Pr_S\big[yf(x) < \theta\big]\Pr_S\big[yf(x) \geqslant 2\theta/3\big].$$

[1] This observation owes to a reviewer.

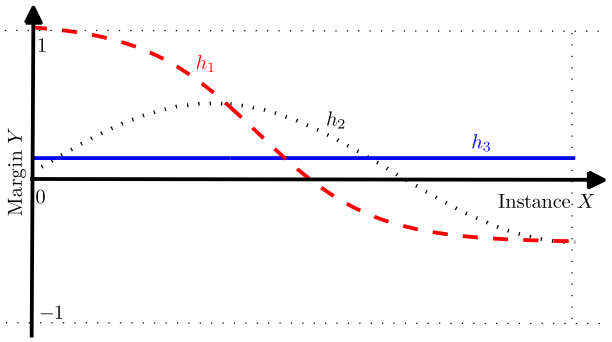

Fig. 1. Each curve represents a voting classifier. The X-axis and Y-axis denote example and margin, respectively, and uniform distribution is assumed on the example space. The voting classifiers h_1, h_2 and h_3 have the same average margin but with different generalization error rates: 1/2, 1/3 and 0.

The detailed proof is deferred to Section 6.7. It is easy to find in almost all boosting experiments that the average margin $E_S[yf(x)]$ is positive. Thus, the bound of Theorem 9 can be sharper for larger average margin. The statistics $\hat{\mathcal{I}}(\cdot)$ reflects the margin variance in some sense, and the term including $\hat{\mathcal{I}}(\cdot)$ can be small or even vanished except for a small interval when the variance is small. This new generalization error bound depends not only on the sample size and the complexity of base classifiers, but also on the average margin, variance, and empirical margin distribution; this implying that, completely explaining AdaBoost's resistance to overfitting is more difficult than what has been expected and disclosed by previous theoretical results.

Theorem 9 also provides a theoretical support to the suggestion of Reyzin and Schapire [37]; that is, the average margin can be used to measure the performance. It is noteworthy that, however, merely considering the average margin is insufficient to bound the generalization error tightly, as shown by the simple example in Fig. 1. Indeed, as this theorem discloses, "average" and "variance" are two important statistics to capture a distribution, and it is reasonable that both the average margin and margin variance are considered.

We have the following corollary with proof presented in Section 6.8.

Corollary 6. *If the minimum margin $\theta_1 = \hat{y}_1 f(\hat{x}_1) > 0$, then, for any $\delta > 0$, with probability at least $1-\delta$ over the random choice of sample S with size $m \geqslant 5$, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound*:

$$\frac{1}{m^{50}} + \inf_{\theta\in(0,1]}\left[\Pr_S\left[yf(x) < \theta\right] + m^{-2/(1-E_S^2[yf(x)]+\theta/9)} + \frac{3\sqrt{\mu}}{m^{3/2}} + \frac{7\mu}{3m} + \sqrt{\frac{3\mu}{m}\hat{\mathcal{I}}(\theta)}\right] \leqslant \frac{1}{m^{50}} + \frac{1}{m^2} + \frac{3\sqrt{\mu_1}}{m^{3/2}} + \frac{7\mu_1}{3m}$$

where $\mu_1 = 144\ln m\ln(2|\mathcal{H}|)/\theta_1^2 + \ln(2|\mathcal{H}|/\delta)$, μ and $\hat{\mathcal{I}}(\theta)$ are given in Theorem 9.

This corollary shows that the bounds of Theorem 9 are $O(\ln m/m)$, comparable to the Emargin bounds [44] and the bounds of Theorem 8, but with different constants. The main difference lies in the consideration of different factors, as we have considered the average margin and variance, that are better for the characterization of margin distribution. It is noteworthy that the best bounds for `AdaBoost` and `arc-gv` are both $O(\ln m/m)$ whereas `AdaBoost` outperforms `arc-gv` empirically because `AdaBoost` tends to improve the margin distribution; this provides an example showing that it is very important to consider factors that are heavily relevant to the whole distribution. We also notice that a recent study in [41] provides empirical evidence to support our theoretical result. Indeed, designing new Boosting algorithms that maximize average margin but minimize variance simultaneously is an interesting direction, and [42] may shed some light.

Finally, we generalize our main margin bounds to the case when the space of base classifiers has finite VC-dimension. The detailed proofs are presented in Section 6.9.

Theorem 10. *If the base classifiers space $\mathcal{H}$ has finite VC-dimension d, then for any $\delta > 0$, with probability at least $1-\delta$ over the random choice of sample S with size $m \geqslant 5$, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound*:

$$\Pr_D\left[yf(x) < 0\right] \leqslant \frac{2}{m} + \inf_{\theta\in(0,1]}\left[\Pr_S\left[yf(x) < \theta\right] + \frac{19\mu + 3\sqrt{3\mu}}{3m} + \sqrt{\frac{3\mu}{m}\Pr_S\left[yf(x) < \theta\right]}\right]$$

where $\mu = \frac{8\ln m}{\theta^2}(\ln 2 + d\ln(2em/d)) + \ln(\frac{8}{\delta}(1 + \frac{8\ln m}{\theta^2}))$.

Theorem 11. *If the base classifiers space $\mathcal{H}$ has finite VC-dimension d, then for any $\delta > 0$, with probability at least $1-\delta$ over the random choice of sample S with size $m \geqslant 5$, every voting classifier $f \in \mathcal{C}(\mathcal{H})$ satisfies the following bound*:

$$\Pr_D\left[yf(x) < 0\right] \leqslant \frac{1}{m^{50}} + \inf_{\theta\in(0,1]}\left[\Pr_S\left[yf(x) < \theta\right] + m^{-2/(1-E_S^2[yf(x)]+\theta/9)} + \frac{3\sqrt{\mu}}{m^{3/2}} + \frac{19\mu}{3m} + \sqrt{\frac{3\mu}{m}\hat{\mathcal{I}}(\theta)}\right]$$

where

$$\mu = 144\big(\ln 2 + d\ln(2em/d)\big)\ln m/\theta^2 + \ln\big(\big(8 + 576\ln m/\theta^2\big)/\delta\big),$$
$$\hat{\mathcal{I}}(\theta) = \Pr_S\big[yf(x) < \theta\big]\Pr_S\big[yf(x) \geqslant 2\theta/3\big].$$

6. Proofs

In this section, we provide the detailed proofs for the main theorems and corollaries. First, we present a series of useful lemmas as follows:

Lemma 1 *(Chernoff bound). (See [14].) Let $X, X_1, X_2, \ldots, X_m$ be $m+1$ i.i.d. random variables with $X \in [0,1]$. Then, for any $\epsilon > 0$, we have*

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i \geqslant E[X] + \epsilon\right] \leqslant \exp\left(-\frac{m\epsilon^2}{2}\right),$$
$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i \leqslant E[X] - \epsilon\right] \leqslant \exp\left(-\frac{m\epsilon^2}{2}\right).$$

Lemma 2 *(Relative entropy Chernoff bound). (See [23].) For $0 < \epsilon < 1$, we have*

$$\sum_{i=0}^{k-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i} \leqslant \exp\left(-mKL\left(\frac{k-1}{m}\middle\|\epsilon\right)\right).$$

Lemma 3 *(Bennett's inequalities). (See [32].) For independent random variables $X, X_1, X_2, \ldots, X_m$ with $X_i \in [0,1]$, and for any $\delta > 0$, the followings hold with probability at least $1-\delta$*

$$\frac{1}{m}\sum_{i=1}^{m} E[X_i] - \frac{1}{m}\sum_{i=1}^{m} X_i \leqslant \sqrt{\frac{2V(X)\ln 1/\delta}{m}} + \frac{\ln 1/\delta}{3m}, \tag{15}$$
$$\frac{1}{m}\sum_{i=1}^{m} E[X_i] - \frac{1}{m}\sum_{i=1}^{m} X_i \geqslant -\sqrt{\frac{2V(X)\ln 1/\delta}{m}} - \frac{\ln 1/\delta}{3m}, \tag{16}$$

where $V(X)$ denotes the variance $\sum_{i=1}^{m} E[(X_i - E[X_i])^2]/m$.

6.1. Proof of Theorem 4

We begin with a lemma as follows:

Lemma 4. *For $f \in \mathcal{C}(\mathcal{H})$, let $g \in \mathcal{C}_N(\mathcal{H})$ be drawn i.i.d. according to distribution $\mathcal{Q}(f)$. If $\hat{y}_k f(\hat{x}_k) \geqslant \theta$ and $\hat{y}_k g(\hat{x}_k) \leqslant \alpha$ with $\theta > \alpha$, then there is an example (x_i, y_i) in S such that $y_i f(x_i) \geqslant \theta$ and $y_i g(x_i) \leqslant \alpha$.*

Proof. There exists a bijection between $\{y_j f(x_j):\ j \in [m]\}$ and $\{y_j g(x_j):\ j \in [m]\}$ according to the original position in S. Suppose $\hat{y}_k f(\hat{x}_k)$ corresponds to $\hat{y}_l g(\hat{x}_l)$ for some l. If $l \leqslant k$ then the example $(\hat{x}_k, \hat{y}_k)$ of $\hat{y}_k f(\hat{x}_k)$ is desired; otherwise, except for $(\hat{x}_k, \hat{y}_k)$ of $\hat{y}_k f(\hat{x}_k)$ in S, there are at least $m-k$ elements larger than or equal to θ in $\{y_j f(x_j):\ j \in [m] \setminus \{k\}\}$ but at most $m-k-1$ elements larger than α in $\{y_j g(x_j):\ j \in [m] \setminus \{l\}\}$. This completes the proof from the bijection. □

Proof of Theorem 4. For finite $\mathcal{H}$, we denote by $\mathcal{A} = \{i/|\mathcal{H}|:\ i \in [|\mathcal{H}|]\}$. For every $f \in \mathcal{C}(\mathcal{H})$, we can construct a $g \in \mathcal{C}_N(\mathcal{H})$ by choosing N elements i.i.d. according to distribution $\mathcal{Q}(f)$, and thus $E_{g\sim\mathcal{Q}(f)}[g] = f$. For $\alpha > 0$, the Chernoff's bound in Lemma 1 gives

$$\begin{aligned}\Pr_D\big[yf(x) < 0\big] &= \Pr_{D,\mathcal{Q}(f)}\big[yf(x) < 0,\ yg(x) \geqslant \alpha\big] + \Pr_{D,\mathcal{Q}(f)}\big[yf(x) < 0,\ yg(x) < \alpha\big] \\ &\leqslant \exp\big(-N\alpha^2/2\big) + \Pr_{D,\mathcal{Q}(f)}\big[yg(x) < \alpha\big].\end{aligned} \tag{17}$$

For any $\epsilon_N > 0$, we consider the following probability:

$$\Pr_{S\sim D^m}\Big[\Pr_D\big[yg(x)<\alpha\big]>I\big[\hat{y}_k g(\hat{x}_k)\leqslant\alpha\big]+\epsilon_N\Big]\leqslant \Pr_{S\sim D^m}\Big[\hat{y}_k g(\hat{x}_k)>\alpha\Big|\Pr_D\big[yg(x)<\alpha\big]>\epsilon_N\Big] \leqslant \sum_{i=0}^{k-1}\binom{m}{i}\epsilon_N^i(1-\epsilon_N)^{m-i} \tag{18}$$

where $\hat{y}_k g(\hat{x}_k)$ denotes the kth margin with respect to g. For any k, Eq. (18) can be bounded by $\exp(-mKL(\frac{k-1}{m}\|\epsilon_N))$ from Lemma 2; for constant k with $m>4k$, we have

$$\sum_{i=0}^{k-1}\binom{m}{i}\epsilon_N^i(1-\epsilon_N)^{m-i}\leqslant k(1-\epsilon_N)^{m/2}\binom{m}{k-1}\leqslant km^{k-1}(1-\epsilon_N)^{m/2}\leqslant km^{k-1}e^{-\epsilon_N m/2}.$$

By using the union bound and $|\mathcal{C}_N(\mathcal{H})|\leqslant|\mathcal{H}|^N$, we have, for any $k\in[m]$,

$$\Pr_{S\sim D^m,g\sim\mathcal{Q}(f)}\Big[\exists g\in\mathcal{C}_N(\mathcal{H}),\ \exists\alpha\in\mathcal{A},\Pr_D\big[yg(x)<\alpha\big]>I\big[\hat{y}_k g(\hat{x}_k)\leqslant\alpha\big]+\epsilon_N\Big]\leqslant|\mathcal{H}|^{N+1}\exp\left(-mKL\left(\frac{k-1}{m}\middle\|\epsilon_N\right)\right).$$

Setting $\delta_N=|\mathcal{H}|^{N+1}\exp(-mKL(\frac{k-1}{m}\|\epsilon_N))$ gives

$$\epsilon_N=KL^{-1}\left(\frac{k-1}{m};\frac{1}{m}\ln\frac{|\mathcal{H}|^{N+1}}{\delta_N}\right).$$

Thus, with probability at least $1-\delta_N$ over sample S, for all $f\in\mathcal{C}(\mathcal{H})$ and all $\alpha\in\mathcal{A}$, we have

$$\Pr_D\big[yg(x)<\alpha\big]\leqslant I\big[\hat{y}_k g(\hat{x}_k)\leqslant\alpha\big]+KL^{-1}\left(\frac{k-1}{m};\frac{1}{m}\ln\frac{|\mathcal{H}|^{N+1}}{\delta_N}\right). \tag{19}$$

Similarly, for constant k, with probability at least $1-\delta_N$ over sample S, it holds that

$$\Pr_D\big[yg(x)<\alpha\big]\leqslant I\big[\hat{y}_k g(\hat{x}_k)\leqslant\alpha\big]+\frac{2}{m}\ln\frac{km^{k-1}|\mathcal{H}|^{N+1}}{\delta_N}. \tag{20}$$

From $E_{g\sim\mathcal{Q}(f)}[I[\hat{y}_k g(\hat{x}_k)\leqslant\alpha]]=\Pr_{g\sim\mathcal{Q}(f)}[\hat{y}_k g(\hat{x}_k)\leqslant\alpha]$, we have, for any $\theta>\alpha$,

$$\Pr_{g\sim\mathcal{Q}(f)}\big[\hat{y}_k g(\hat{x}_k)\leqslant\alpha\big]\leqslant I\big[\hat{y}_k f(\hat{x}_k)<\theta\big]+\Pr_{g\sim\mathcal{Q}(f)}\big[\hat{y}_k f(\hat{x}_k)\geqslant\theta,\hat{y}_k g(\hat{x}_k)\leqslant\alpha\big]. \tag{21}$$

Notice that the example $(\hat{x}_k,\hat{y}_k)$ in $\{\hat{y}_i f(\hat{x}_i)\}$ may be different from example $(\hat{x}_k,\hat{y}_k)$ in $\{\hat{y}_i g(\hat{x}_i)\}$; therefore, we cannot bound the last term on the right-hand side of Eq. (21) as done in [44], whereas it can be bounded by using Lemma 4

$$\Pr_{g\sim\mathcal{Q}(f)}\big[\exists(x_i,y_i)\in S\colon\ y_i f(x_i)\geqslant\theta,\ y_i g(x_i)\leqslant\alpha\big]\leqslant m\exp\big(-N(\theta-\alpha)^2/2\big). \tag{22}$$

Combining Eqs. (17), (19), (21) and (22), we have that with probability at least $1-\delta_N$ over the sample S, for all $f\in\mathcal{C}(\mathcal{H})$, all $\theta>\alpha$, all $k\in[m]$ but fixed N:

$$\Pr_D\big[yf(x)<0\big]\leqslant I\big[\hat{y}_k f(\hat{x}_k)\leqslant\theta\big]+m\exp\big(-N(\theta-\alpha)^2/2\big)+\exp\big(-N\alpha^2/2\big)+KL^{-1}\left(\frac{k-1}{m};\frac{1}{m}\ln\frac{|\mathcal{H}|^{N+1}m}{\delta_N}\right). \tag{23}$$

To obtain the probability of failure for any N at most δ, we select $\delta_N=\delta/2^N$. Setting $\alpha=\frac{\theta}{2}-\frac{\eta}{|\mathcal{H}|}\in\mathcal{A}$ and $N=\lceil\frac{8}{\theta^2}\ln\frac{2m^2}{\ln|\mathcal{H}|}\rceil$ with $0\leqslant\eta<1$, we have

$$\exp\big(-N\alpha^2/2\big)+m\exp\big(-N(\theta-\alpha)^2/2\big)\leqslant 2m\exp\big(-N\theta^2/8\big)\leqslant\ln|\mathcal{H}|/m$$

from the fact $2m>\exp(N/(2|\mathcal{H}|))$ for $\theta>\sqrt{8/|\mathcal{H}|}$. Finally we obtain

$$\Pr\big[yf(x)<0\big]\leqslant I\big[\hat{y}_k f(\hat{x}_k)<\theta\big]+\frac{\ln|\mathcal{H}|}{m}+KL^{-1}\left(\frac{k-1}{m}\middle\|\frac{q}{m}\right)$$

where $q=\frac{8\ln(2|\mathcal{H}|)}{\theta^2}\ln\frac{2m^2}{\ln|\mathcal{H}|}+\ln|\mathcal{H}|+\ln\frac{m}{\delta}$. This completes the proof of Eq. (4). In a similar manner, we have

$$\Pr\big[yf(x)<0\big]\leqslant I\big[\hat{y}_k f(\hat{x}_k)<\theta\big]+\ln|\mathcal{H}|/m+\frac{2}{m}\left(\frac{8\ln(2|\mathcal{H}|)}{\theta^2}\ln\frac{2m^2}{\ln|\mathcal{H}|}+\ln|\mathcal{H}|+\ln\frac{km^{k-1}}{\delta}\right),$$

for constant $k<m/4$. This completes the proof of Eq. (5) as desired. □

6.2. Proof of Theorem 5

For notational simplicity, we denote by $\bar{X} = (X_1, X_2, \ldots, X_m)$ a vector of m i.i.d. random variables, and further set

$$\bar{X}^{k,Y} = (X_1, \ldots, X_{k-1}, Y, X_{k+1}, \ldots, X_m),$$

i.e., the vector with the kth variable X_k in $\bar{X}$ replaced by variable Y. We first introduce some lemmas as follows:

Lemma 5 *(McDiarmid formula). (See [31].) Suppose that $\bar{X} = (X_1, X_2, \ldots, X_m)$ is a vector of m i.i.d. random variables taking values in a set $\mathcal{A}$. If $|F(\bar{X}) - F(\bar{X}^{k,Y})| \leqslant c_k$ for $k \in [m]$ and $Y \in \mathcal{A}$, then the following holds for any $t > 0$,*

$$\Pr\big[F(\bar{X}) - E\big[F(\bar{X})\big] \geqslant t\big] \leqslant \exp\left(\frac{-2t^2}{\sum_{k=1}^m c_k^2}\right).$$

Lemma 6. *(See [29, Theorem 13].) Let $\bar{X} = (X_1, X_2, \ldots, X_m)$ be a vector of m independent random variables tanking values in a set $\mathcal{A}$. If $F : \mathcal{A}^m \to \mathbb{R}$ satisfies that*

$$F(\bar{X}) - \inf_{Y\in\mathcal{A}} F\big(\bar{X}^{k,Y}\big) \leqslant 1 \quad \text{and} \quad \sum_{k=1}^m \Big(F(\bar{X}) - \inf_{Y\in\mathcal{A}} F\big(\bar{X}^{k,Y}\big)\Big)^2 \leqslant F(\bar{X}),$$

then for any $t > 0$, we have

$$\Pr\big[E\big[F(\bar{X})\big] - F(\bar{X}) > t\big] \leqslant \exp\big(-t^2/2E\big[F(\bar{X})\big]\big).$$

Lemma 7. *For two i.i.d. random variables X and Y, we have*

$$E\big[(X-Y)^2\big] = 2E\big[\big(X - E[X]\big)^2\big] = 2V(X).$$

Proof. This lemma follows from the obvious fact $E[(X-Y)^2] = E(X^2+Y^2-2XY) = 2E[X^2] - 2E^2[X] = 2E[(X-E[X])^2]$. □

Proof of Theorem 5. We will utilize Lemmas 5 and 6 to prove Eqs. (6) and (7), respectively. For Eq. (6), we first observe that, for any $k \in [m]$,

$$\Big|\sqrt{\hat{V}_m(\bar{X})} - \sqrt{\hat{V}_m\big(\bar{X}^{k,Y}\big)}\Big| = \left|\frac{\hat{V}_m(\bar{X}) - \hat{V}_m(\bar{X}^{k,Y})}{\sqrt{\hat{V}_m(\bar{X})} + \sqrt{\hat{V}_m(\bar{X}^{k,Y})}}\right| \leqslant \frac{1}{\sqrt{2}m},$$

where we use $\hat{V}_m(\bar{X}), \hat{V}_m(\bar{X}^{k,Y}) \leqslant 1/2$ from $X_i \in [0,1]$. By using the Jenson's inequality, we have $E[\sqrt{\hat{V}_m(\bar{X})}] \leqslant \sqrt{E[\hat{V}_m(\bar{X})]}$ and thus,

$$\Pr\Big[\sqrt{E\big[\hat{V}_m(\bar{X})\big]} < \sqrt{\hat{V}_m(\bar{X})} - \epsilon\Big] \leqslant \Pr\Big[E\big[\sqrt{\hat{V}_m(\bar{X})}\big] < \sqrt{\hat{V}_m(\bar{X})} - \epsilon\Big] \leqslant \exp\big(-4m\epsilon^2\big).$$

where the last inequality holds by applying McDiarmid formula in Lemma 5 to $\sqrt{\hat{V}_m}$. Therefore, we complete the proof of Eq. (6) by setting $\delta = \exp(-4m\epsilon^2)$.

To prove Eq. (7), we set $\xi_m(\bar{X}) = m\hat{V}_m(\bar{X})$. For $X_i \in [0,1]$ and $\xi_m(\bar{X}^{k,Y})$, it is easy to obtain the optimal solution by simple calculation

$$Y^* = \arg\inf_{Y\in[0,1]}\big[\xi_m\big(\bar{X}^{k,Y}\big)\big] = \sum\nolimits_{i\neq k} \frac{X_i}{m-1},$$

which yields that

$$\xi_m(\bar{X}) - \inf_{Y\in[0,1]}\big[\xi_m\big(\bar{X}^{k,Y}\big)\big] = \frac{1}{m-1}\sum_{i\neq k}(X_i - X_k)^2 - \big(Y^* - X_i\big)^2 = \left(X_k - \sum_{i\neq k}\frac{X_i}{m-1}\right)^2.$$

For $X_i \in [0,1]$, it is obvious that

$$\xi_m(\bar{X}) - \inf_{Y\in[0,1]}\big[\xi_m\big(\bar{X}^{k,Y}\big)\big] \leqslant 1,$$

and from Lemma 7, we have

$$\frac{1}{m}\sum_{k=1}^{m}\left(X_k-\sum_{i=1}^{m}\frac{X_i}{m}\right)^2\leqslant\frac{1}{2m^2}\sum_{i,k}(X_i-X_k)^2=\frac{1}{2m^2}\sum_{i\neq k}(X_i-X_k)^2,$$

which yields that, for $m\geqslant 5$,

$$\sum_{k=1}^{m}\Big(\xi_m(\bar{X})-\inf_{Y\in[0,1]}\big[\xi_m\big(\bar{X}^{k,Y}\big)\big]\Big)^2\leqslant\frac{m^3}{4(m-1)^4}\sum_{i\neq k}(X_i-X_k)^2\leqslant\xi_m(\bar{X}).$$

Therefore, for any $t>0$, the following holds by using Lemma 6 to $\xi_m(\bar{X})$,

$$\Pr\big[E\big[\hat{V}_m(\bar{X})\big]-\hat{V}_m(\bar{X})>t\big]=\Pr\big[E\big[\xi_m(\bar{X})\big]-\xi_m(\bar{X})>mt\big]\leqslant\exp\big(-mt^2/2E\big[\hat{V}_m(\bar{X})\big]\big).$$

Setting $\delta=\exp(-mt^2/2E[\hat{V}_m(\bar{X})])$ gives

$$\Pr\Big[E\big[\hat{V}_m(\bar{X})\big]-\hat{V}_m(\bar{X})>\sqrt{2E\big[\hat{V}_m(\bar{X})\big]\ln(1/\delta)/m}\Big]\leqslant\delta$$

which completes the proof of Eq. (7) by using the square-root's inequality and $\sqrt{a+b}\leqslant\sqrt{a}+\sqrt{b}$ for $a,b\geqslant 0$. □

6.3. Proof of Theorem 6

For independent random variables $\bar{X}=(X_1,X_2,\ldots,X_m)$, we set $\hat{V}_m(\bar{X})=\sum_{i\neq j}(X_i-X_j)^2/2m(m-1)$, and observe that

$$\begin{aligned}E\big[\hat{V}_m(\bar{X})\big]&=\frac{1}{2m(m-1)}\sum_{i\neq j}E\big[(X_i-X_j)^2\big]\\&=\frac{1}{2m(m-1)}\sum_{i\neq j}\big(E\big[\big(X_i-E[X_i]\big)^2\big]+E[\big(X_j-E[X_j]\big)^2+\big(E[X_i]-E[X_j]\big)^2\big)\\&\geqslant\frac{1}{m}\sum_{i}E\big(X_i-E[X_i]\big)^2=V,\end{aligned}$$

where we denote by $V=\sum_i E(X_i-E[X_i])^2/m$ and the second equality holds from $(a+b+c)^2=a^2+b^2+c^2+2ab+2ac+2bc$. For any $\delta>0$, the following holds with probability at least $1-\delta$ from Eq. (15),

$$\frac{1}{m}\sum_{i=1}^{m}\big(E[X_i]-X_i\big)\leqslant\sqrt{\frac{2V\ln 1/\delta}{m}}+\frac{\ln 1/\delta}{3m}\leqslant\sqrt{\frac{2E[\hat{V}_m(\bar{X})]\ln 1/\delta}{m}}+\frac{\ln 1/\delta}{3m}$$

which completes the proof of Eq. (8) by combining with Eq. (7) in a union bound and simple calculations. Similar proof could be made for Eq. (9). □

6.4. Proof of Theorem 7

We will use classical double sample method [15,43] to prove Theorem 7. Let $\mathscr{A}$ be a subsets of space $\mathcal{Z}$, and we define

$$s(\mathscr{A},m)=\max\big\{\big|\{A\cap S\colon\ A\in\mathscr{A}\}\big|\colon\ S\subseteq\mathcal{Z}\text{ and }|S|=m\big\}.$$

We first introduce a useful lemma as follows:

Lemma 8. *For space $\mathscr{A}$ of subsets of $\mathcal{Z}$, and for sample $S=(z_1,z_2,\ldots,z_m)$ drawn i.i.d. from distribution $\mathcal{D}$ over $\mathcal{Z}$, we have, for $t>\ln 4$*

$$\Pr_{S\sim\mathcal{D}^m}\left[\exists A\in\mathscr{A}\colon\ \Pr_{\mathcal{D}}[A]>\Pr_{S}[A]+\sqrt{\frac{2t}{m}\hat{V}_S(A)}+\frac{19t}{3m}\right]\leqslant 8s(\mathscr{A},2m)e^{-t}$$

where $\Pr_{\mathcal{D}}[A]=\Pr_{z\sim\mathcal{D}}[z\in A]$, $\Pr_S[A]=\Pr_{z\sim S}[z\in A]$ and $\hat{V}_S(A)=\sum_{i\neq j}(I[z_i\in A]-I[z_j\in A])^2/2m(m-1)$.

Proof. We begin with another sample $\hat{S}=(\hat{z}_1,\hat{z}_2,\ldots,\hat{z}_m)$ drawn identically and independently from distribution $\mathcal{D}$, and denote by

$$\Psi_S(A)=\Pr_S[A]+\sqrt{2\hat{V}_S(A)t/m}+7t/3m.$$

From Corollary 3, we have $\Pr_{\hat{S}\sim\mathcal{D}^m}[\Pr_{\mathcal{D}}[A] \leqslant \Psi_{\hat{S}}(A)] \geqslant 1/2$ for $h \in \mathcal{H}$ and $t > \ln 4$. This follows for any $\epsilon > 0$

$$\begin{aligned}\Pr_{S\sim\mathcal{D}^m}\big[\exists A \in \mathscr{A}\colon \Pr_{\mathcal{D}}[A] > \Psi_S(A) + \epsilon\big] &= E_{S\sim\mathcal{D}^m} \sup_{A\in\mathscr{A}} I\big[\Pr_{\mathcal{D}}[A] > \Psi_S(A) + \epsilon\big] \\ &\leqslant 2E_{S\sim\mathcal{D}^m} \sup_{A\in\mathscr{A}} I\big[\Pr_{\mathcal{D}}[A] > \Psi_S(A) + \epsilon\big] E_{\hat{S}\sim\mathcal{D}^m} I\big[\Pr_{\mathcal{D}}[A] \leqslant \Psi_{\hat{S}}(A)\big] \\ &\leqslant 2 \Pr_{S\sim\mathcal{D}^m,\hat{S}\sim\mathcal{D}^m}\big[\exists A \in \mathscr{A}\colon \Psi_{\hat{S}}(A) > \Psi_S(A) + \epsilon\big].\end{aligned}$$

Now, we introduce the sign random variable vector $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_m)$ with probability $\Pr[\sigma_i = 1] = \Pr[\sigma_i = -1] = 1/2$ for $i \in [m]$, and denote by $S^\sigma = (z_i^\sigma)_{i=1}^m$ and $\hat{S}^\sigma = (\hat{z}_i^\sigma)_{i=1}^m$

$$z_i^\sigma = z_i, \quad \hat{z}_i^\sigma = \hat{z}_i \quad \text{if } \sigma = 1; \quad \text{otherwise}, \quad z_i^\sigma = \hat{z}_i, \quad \hat{z}_i^{\sigma_i} = z_i.$$

Given S and S', z_i^σ $(i \in [m])$ are not identically distributed but independent. Conditioned on S and S', we have

$$\begin{aligned}\Pr_\sigma\big[\exists A \in \mathscr{A}\colon \Psi_{\hat{S}^\sigma}(A) > \Psi_{S^\sigma}(A) + \epsilon | S, S'\big] &\leqslant s(\mathscr{A}, 2m) \sup_{A\in\mathscr{A}} \Pr_\sigma\big[\Psi_{\hat{S}^\sigma}(A) > \Psi_{S^\sigma}(A) + \epsilon | S, S'\big] \\ &= s(\mathscr{A}, 2m) \Pr_\sigma\big[\Psi_{\hat{S}^\sigma}(A^*) > \Psi_{S^\sigma}(A^*) + \epsilon | S, S'\big] \\ &= s(\mathscr{A}, 2m) \Pr_\sigma\big[\Psi_{\hat{S}^\sigma}(A^*) > \Pr_\sigma[A^*] + \epsilon/2 | S, S'\big] \\ &\quad + s(\mathscr{A}, 2m) \Pr_\sigma\big[\Pr_\sigma[A^*] > \Psi_{S^\sigma}(A^*) + \epsilon/2 | S, S'\big],\end{aligned}$$

where we denote by $A^* \in \arg\sup_{A\in\mathscr{A}} \Pr_\sigma[\Psi_{\hat{S}^\sigma}(A) > \Psi_{S^\sigma}(A) + \epsilon | S, S']$ and $\Pr_\sigma[A^*] = E_\sigma[\Pr_{S^\sigma}[A^*] | S, \hat{S}] = E_\sigma[\Pr_{\hat{S}^\sigma}[A^*] | S, \hat{S}]$. Further, we denote by

$$V_\sigma(A^*) = E_{S^\sigma}\big[\hat{V}_{S^\sigma}(A^*) | S, \hat{S}\big] = E_{\hat{S}^\sigma}\big[\hat{V}_{\hat{S}^\sigma}(A^*) \big| S, \hat{S}\big].$$

Thus, we have

$$\begin{aligned}\Pr_\sigma\big[\Psi_{\hat{S}^\sigma}(A^*) > \Pr_\sigma[A^*] + \epsilon/2 | S, S'\big] &= \Pr_\sigma\Big[\Pr_{\hat{S}^\sigma}[A^*] + \sqrt{2\hat{V}_{\hat{S}^\sigma}(A^*)t/m} + 7t/3m > \Pr_\sigma[A^*] + \epsilon/2 | S, S'\Big] \\ &\leqslant \Pr_\sigma\Big[\Pr_{\hat{S}^\sigma}[A^*] + \sqrt{2V_\sigma(A^*)t/m} + 7t/3m > \Pr_\sigma[A^*] | S, S'\Big] \\ &\quad + \Pr_\sigma\Big[\sqrt{2\hat{V}_{\hat{S}^\sigma}(A^*)t/m} > \sqrt{2V_\sigma(A^*)t/m} + \epsilon/2 | S, S'\Big].\end{aligned}$$

The first term in the above can be bounded by e^{-t} from Bennett's inequality (Lemma 3), and the second term can be bound by e^{-t} by setting $\epsilon = 4t/m$ and using Theorem 5. Similarly, we can prove

$$\Pr_\sigma\big[\Pr_\sigma[A^*] > \Psi_{S^\sigma}(A^*) + \epsilon/2 | S, S'\big] \leqslant 2e^{-t}$$

by setting $\epsilon = 4t/m$. This completes the proof as desired. □

Proof of Theorem 7. Let

$$\mathscr{A} = \big\{A(h)\colon h \in \mathcal{H}\big\}$$

where $A(h) = \{(X, h(X) \in \mathcal{X} \times \{-1, +1\})\}$. For space $\mathcal{H}$ with finite VC-dimension d, Sauer's lemma [38] gives

$$s(\mathscr{A}, 2m) \leqslant (2em/d)^d.$$

Combining with Lemma 8, we have, for $t \geqslant \ln 4$

$$\Pr_{S\sim\mathcal{D}^m}\left[\exists h \in \mathcal{H}\colon E_{\mathcal{D}}[h(X)] > \sum_{i=1}^m \frac{h(X_i)}{m} + \sqrt{\frac{2t\hat{V}_S(h)}{m}} + \frac{19t}{3m}\right] \leqslant 8\left(\frac{2m}{d}\right)^d e^{-t}.$$

Setting $\delta = 8(2m/d)^d e^{-t}$, we have

$$t = d\ln(2m/d) + \ln(8/\delta) \geqslant \ln 4 \quad \text{for } \delta \in (0, 1),$$

which completes the proof. □

6.5. Proof of Theorem 8

Similarly to the proof of Theorem 4, we have

$$\Pr_{D}\big[yf(x)<0\big]\leqslant\exp\big(-N\alpha^2/2\big)+\Pr_{D,\mathcal{Q}(f)}\big[yg(x)<\alpha\big], \tag{24}$$

for any given $\alpha>0$, $f\in\mathcal{C}(\mathcal{H})$ and $g\in\mathcal{C}_N(\mathcal{H})$ drawn i.i.d. according to $\mathcal{Q}(f)$. Recall that $|\mathcal{C}_N(\mathcal{H})|\leqslant|\mathcal{H}|^N$. Therefore, for any $\delta_N>0$, combining union bound with Eq. (8) in Theorem 3 guarantees that the following holds with probability at least $1-\delta_N$ over sample S, for any $g\in\mathcal{C}_N(\mathcal{H})$ and $\alpha\in\mathcal{A}$,

$$\Pr_{D}\big[yg(x)<\alpha\big]\leqslant\Pr_{S}\big[yg(x)<\alpha\big]+\sqrt{\frac{2}{m}\hat{V}_m\ln\left(\frac{2}{\delta_N}|\mathcal{H}|^{N+1}\right)}+\frac{7}{3m}\ln\left(\frac{2}{\delta_N}|\mathcal{H}|^{N+1}\right), \tag{25}$$

where

$$\hat{V}_m=\sum_{i\neq j}\frac{(I[y_ig(x_i)<\alpha]-I[y_jg(x_j)<\alpha])^2}{2m(m-1)}.$$

Furthermore, we have

$$\sum_{i\neq j}\big(I\big[y_ig(x_i)<\alpha\big]-I\big[y_jg(x_j)<\alpha\big]\big)^2=2m^2\Pr_{S}\big[yg(x)<\alpha\big]\Pr_{S}\big[yg(x)\geqslant\alpha\big],$$

which yields that

$$\hat{V}_m=\frac{m}{m-1}\Pr_{S}\big[yg(x)<\alpha\big]\Pr_{S}\big[yg(x)\geqslant\alpha\big]\leqslant\frac{3}{2}\Pr_{S}\big[yg(x)<\alpha\big], \tag{26}$$

for $m\geqslant5$. By using Lemma 1 again, the following holds for any $\theta_1>0$,

$$\Pr_{S}\big[yg(x)<\alpha\big]\leqslant\exp\big(-N\theta_1^2/2\big)+\Pr_{S}\big[yf(x)<\alpha+\theta_1\big]. \tag{27}$$

Setting $\theta_1=\alpha=\theta/2$ and combining Eqs. (24), (25), (26) and (27), we have

$$\Pr_{D}\big[yf(x)<0\big]\leqslant\Pr_{S}\big[yf(x)<\theta\big]+2\exp\big(-N\theta^2/8\big)+\frac{7\mu}{3m}+\sqrt{\frac{3\mu}{m}\left(\Pr_{S}\big[yf(x)<\theta\big]+\exp\left(-\frac{N\theta^2}{8}\right)\right)},$$

where $\mu=\ln(2|\mathcal{H}|^{N+1}/\delta_N)$. By utilizing the fact $\sqrt{a+b}\leqslant\sqrt{a}+\sqrt{b}$ for $a\geqslant0$ and $b\geqslant0$, we further have

$$\sqrt{\frac{3\mu}{m}\left(\Pr_{S}\big[yf(x)<\theta\big]+\exp\left(-\frac{N\theta^2}{8}\right)\right)}\leqslant\sqrt{\frac{3\mu}{m}\Pr_{S}\big[yf(x)<\theta\big]}+\sqrt{\frac{3\mu}{m}\exp\left(-\frac{N\theta^2}{8}\right)}.$$

Finally, we set $\delta_N=\delta/2^N$ so that the probability of failure for any N will be no more than δ. This theorem follows by setting $N=\lceil8\ln m/\theta^2\rceil$. □

6.6. Proof of Corollary 5

If the minimum margin $\theta_1=\hat{y}_1f(\hat{x}_1)>0$, then we have $\Pr_S[yf(x)<\theta_1]=0$ and further get

$$\begin{aligned}&\inf_{\theta\in(0,1]}\left[\Pr_{S}\big[yf(x)<\theta\big]+\frac{7\mu+3\sqrt{3\mu}}{3m}+\sqrt{\frac{3\mu}{m}\Pr_{S}\big[yf(x)<\theta\big]}\right]\\&\leqslant\Pr_{S}\big[yf(x)<\theta_1\big]+\frac{7\mu_1+3\sqrt{3\mu_1}}{3m}+\sqrt{\frac{3\mu_1}{m}\Pr_{S}\big[yf(x)<\theta_1\big]}\\&=\frac{7\mu_1+3\sqrt{3\mu_1}}{3m},\end{aligned} \tag{28}$$

where $\mu_1=8\ln m\ln(2|\mathcal{H}|)/\theta_1^2+\ln(2|\mathcal{H}|/\delta)$. This gives the proof of Eq. (10). If $m\geqslant5$, then we have

$$\mu_1\geqslant\frac{8}{\theta_1^2}\ln m\ln(2|\mathcal{H}|)\geqslant8\quad\text{leading to}\quad\sqrt{3\mu_1}\leqslant2\mu_1/3.$$

Therefore, the following holds by combining Eq. (28) and the above facts,

$$\begin{aligned}&\frac{2}{m}+\inf_{\theta\in(0,1]}\left[\Pr_S\big[yf(x)<\theta\big]+\frac{7\mu+3\sqrt{3\mu}}{3m}+\sqrt{\frac{3\mu}{m}\Pr_S\big[yf(x)<\theta\big]}\right]\\&\leqslant\frac{2}{m}+\frac{7\mu_1+3\sqrt{2\mu_1}}{3m}\leqslant\frac{2}{m}+\frac{3\mu_1}{m}=\frac{2}{m}+\frac{24\ln m}{m\theta_1^2}\ln(2|\mathcal{H}|)+\frac{3}{m}\ln\frac{2|\mathcal{H}|}{\delta}\\&\leqslant\frac{8}{m}+\frac{24\ln m}{m\theta_1^2}\ln(2|\mathcal{H}|)+\frac{3}{m}\ln\frac{|\mathcal{H}|}{\delta}\leqslant R\left(\ln(2m)+\ln\frac{1}{R}+1\right)+\frac{1}{m}\ln\frac{|\mathcal{H}|}{\delta}\end{aligned}$$

where the last inequality holds from the conditions of Eq. (13) and $8/m<R$. This completes the proof of Eq. (14). □

6.7. Proof of Theorem 9

Our proof is based on a new Bernstein-type bound as follows:

Lemma 9. *For $f\in\mathcal{C}(\mathcal{H})$ and $g\in\mathcal{C}_N(\mathcal{H})$ drawn i.i.d. according to distribution $\mathcal{Q}(f)$, we have*

$$\Pr_{S,g\sim\mathcal{Q}(f)}\big[yg(x)-yf(x)\geqslant t\big]\leqslant\exp\left(\frac{-Nt^2}{2-2E_S^2[yf(x)]+4t/3}\right).$$

Proof. For $\lambda>0$, we utilize the Markov's inequality to have

$$\begin{aligned}\Pr_{S,g\sim\mathcal{Q}(f)}\big[yg(x)-yf(x)\geqslant t\big]&=\Pr_{S,g\sim\mathcal{Q}(f)}\big[\big(yg(x)-yf(x)\big)N\lambda/2\geqslant N\lambda t/2\big]\\&\leqslant\exp\left(-\frac{\lambda Nt}{2}\right)E_{S,g\sim\mathcal{Q}(f)}\left[\exp\left(\frac{\lambda}{2}\sum_{j=1}^{N}yh_j(x)-yf(x)\right)\right]\\&=\exp(-\lambda Nt/2)\prod_{j=1}^{N}E_{S,h_j\sim\mathcal{Q}(f)}\big[\exp\big(\lambda\big(yh_j(x)-yf(x)\big)/2\big)\big],\end{aligned}$$

where the last inequality holds from the independence of h_j. Notice that $|yh_j(x)-yf(x)|\leqslant 2$ from $\mathcal{H}\subseteq\{h:\mathcal{X}\to\{-1,+1\}\}$. By using Taylor's expansion, we further get

$$\begin{aligned}E_{S,h_j\sim\mathcal{Q}(f)}\big[\exp\big(\lambda\big(yh_j(x)-yf(x)\big)/2\big)\big]&\leqslant 1+E_{S,h_j\sim\mathcal{Q}(f)}\big[\big(yh_j(x)-yf(x)\big)^2\big]\big(e^{\lambda}-1-\lambda\big)/4\\&=1+E_S\big[1-\big(yf(x)\big)^2\big]\big(e^{\lambda}-1-\lambda\big)/4\\&\leqslant\exp\big(\big(1-E_S^2\big[yf(x)\big]\big)\big(e^{\lambda}-1-\lambda\big)/4\big),\end{aligned}$$

where the last inequality holds from Jensen's inequality and $1+x\leqslant e^x$. Therefore, it holds that

$$\Pr_{S,g\sim\mathcal{Q}(f)}\big[yg(x)-yf(x)\geqslant t\big]\leqslant\exp\big(N\big(e^{\lambda}-1-\lambda\big)\big(1-E_S^2\big[yf(x)\big]\big)/4-\lambda Nt/2\big).$$

If $0<\lambda<3$, then we could use Taylor's expansion again to have

$$e^{\lambda}-\lambda-1=\sum_{i=2}^{\infty}\frac{\lambda^i}{i!}\leqslant\frac{\lambda^2}{2}\sum_{i=0}^{\infty}\frac{\lambda^m}{3^m}=\frac{\lambda^2}{2(1-\lambda/3)}.$$

Now by picking $\lambda=t/(1/2-E_S^2[yf(x)]/2+t/3)$, we have

$$-\frac{\lambda t}{2}+\frac{\lambda^2(1-E_S^2[yf(x)])}{8(1-\lambda/3)}\leqslant\frac{-t^2}{2-2E_S^2[yf(x)]+4t/3},$$

which completes the proof as desired. □

Proof of Theorem 9. This proof is rather similar to the proof of Theorem 8, and we just give main steps. For any $\alpha>0$ and $\delta_N>0$, the following holds with probability at least $1-\delta_N$ over sample S_m ($m\geqslant 5$),

$$\Pr_D\big[yf(x)<0\big]\leqslant\Pr_S\big[yg(x)<\alpha\big]+\exp\big(-N\alpha^2/2\big)+\sqrt{\frac{3\hat{V}_m^*\ln(\frac{2}{\delta_N}|\mathcal{H}|^{N+1})}{m}}+\frac{7}{3m}\ln\left(\frac{2}{\delta_N}|\mathcal{H}|^{N+1}\right),$$

where $\hat{V}_m^*=\Pr_S[yg(x)<\alpha]\Pr_S[yg(x)\geqslant\alpha]$. For any $\theta_1>0$, we use Lemma 1 to obtain

$$\hat{V}_m^* = \Pr_S[yg(x) < \alpha]\Pr_S[yg(x) \geqslant \alpha] \leqslant 3\exp(-N\theta_1^2/2) + \Pr_S[yf(x) < \alpha + \theta_1]\Pr_S[yf(x) > \alpha - \theta_1].$$

From Lemma 9, it holds that

$$\Pr_S[yg(x) < \alpha] \leqslant \Pr_S[yf(x) < \alpha + \theta_1] + \exp\left(\frac{-N\theta_1^2}{2 - 2E_S^2[yf(x)] + 4\theta_1/3}\right).$$

Let $\theta_1 = \theta/6$, $\alpha = 5\theta/6$, and set $\delta_N = \delta/2^N$ so that the probability of failure for any N will be no more than δ. We complete the proof by setting $N = \lceil 144\ln m/\theta^2 \rceil$ and simple calculation. □

6.8. Proof of Corollary 6

If the minimum margin $\theta_1 = \hat{y}_1 f(\hat{x}_1) > 0$, then we have $\Pr_S[yf(x) < \theta_1] = 0$ and $\hat{\mathcal{I}}(\theta_1) = \Pr_S[yf(x) < \theta_1]\Pr_S[yf(x) \geqslant 2\theta_1/3] = 0$. Further, we have

$$\begin{aligned}
&\inf_{\theta\in(0,1]}\left[\Pr_S[yf(x) < \theta] + \frac{\sqrt{6\mu}}{m^{3/2}} + \frac{7\mu}{3m} + \sqrt{\frac{3\mu}{m}\hat{\mathcal{I}}(\theta)} + m^{-2/(1-E_S^2[yf(x)+\theta/9)}\right] \\
&\quad\leqslant \frac{\sqrt{6\mu_1}}{m^{3/2}} + \frac{7\mu_1}{3m} + m^{-2/(1-E_S^2[yf(x)]+\theta_1/9)} \\
&\quad\leqslant \frac{\sqrt{6\mu_1}}{m^{3/2}} + \frac{7\mu_1}{3m} + \frac{1}{m^2}
\end{aligned}$$

where $\mu_1 = 144\ln m\ln(2|\mathcal{H}|)/\theta_1^2 + \ln(2|\mathcal{H}|/\delta)$. This completes the proof. □

6.9. Proof of Theorems 10 and 11

For finite VC-dimension space $\mathcal{H}$, we denote by $\mathcal{A} = \{i/N\colon\ i \in [N]\}$. Similarly to the proof of Theorem 4, we have

$$\Pr_D[yf(x) < 0] \leqslant \exp(-N\alpha^2/2) + \Pr_{D,\mathcal{Q}(f)}[yg(x) < \alpha], \tag{29}$$

for $\alpha \in \mathcal{A}$, $f \in \mathcal{C}(\mathcal{H})$ and $g \in \mathcal{C}_N(\mathcal{H})$ chosen i.i.d. according to $\mathcal{Q}(f)$. Define

$$\mathscr{A} = \left\{\left\{(x, y) \in \mathcal{X} \times \{+1, -1\}\colon\ yg(x) < \alpha\right\}\colon\ g \in \mathcal{C}_N(\mathcal{H}),\ \alpha \in \mathcal{A}\right\},$$

and by using Sauer's lemma [38], we have

$$s(\mathscr{A}, m) \leqslant (N+1)(em/d)^{Nd} \tag{30}$$

for $m > d$. By setting $4s(\mathscr{A}, 2m)e^{-t} = \delta_N > 0$ in Lemma 8, the following holds with probability at least $1 - \delta_N$ over sample S, for any $g \in \mathcal{C}_N(\mathcal{H})$ and $\alpha \in \mathcal{A}$,

$$\Pr_D[yg(x) < \alpha] \leqslant \Pr_S[yg(x) < \alpha] + \sqrt{\frac{3}{m}\hat{V}_m\ln\left(\frac{8s(\mathscr{A}, 2m)}{\delta_N}\right)} + \frac{19}{3m}\ln\left(\frac{8s(\mathscr{A}, 2m)}{\delta_N}\right), \tag{31}$$

where $\hat{V}_m^* = \Pr_S[yg(x) < \alpha]\Pr_S[yg(x) \geqslant \alpha]$.

To prove Theorem 10, we proceed as the proof of Theorem 8. Setting $\alpha = \theta/2$, we have

$$\Pr_D[yf(x) < 0] \leqslant \Pr_S[yf(x) < \theta] + 2\exp(-N\theta^2/8) + \frac{19\mu}{3m} + \sqrt{\frac{3\mu}{m}\left(\Pr_S[yf(x) < \theta] + \exp\left(-\frac{N\theta^2}{8}\right)\right)},$$

where $\mu = \ln(8s(\mathscr{A}, 2m)/\delta_N)$. This completes the proof by using $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ and setting $\delta_N = \delta/2^N$ and $N = \lceil 8\ln m/\theta^2 \rceil$.

To prove Theorem 11, we proceed as the proof of Theorem 9. Setting $\alpha = 5\theta/6$, we have

$$\begin{aligned}
\Pr_D[yf(x) < 0] &\leqslant \Pr_S[yf(x) < \theta] + \exp(-25N\theta^2/72) + 19\mu/3m \\
&\quad + \exp\left(\frac{-N\theta^2/36}{2 - 2E_S^2[yf(x)] + 2\theta/9}\right) + \sqrt{\frac{3\mu}{m}\left(\hat{\mathcal{I}}(\theta) + 3\exp\left(-\frac{N\theta^2}{72}\right)\right)},
\end{aligned}$$

where $\mu = \ln(8s(\mathscr{A}, 2m)/\delta_N)$ and $\hat{\mathcal{I}}(\theta) = \Pr_S[yf(x) < \theta]\Pr_S[yf(x) \geqslant 2\theta/3]$. This completes the proof by using $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ and setting $\delta_N = \delta/2^N$ and $N = \lceil 144\ln m/\theta^2 \rceil$. □

7. Conclusion

The margin theory provides one of the most intuitive and popular theoretical explanations to AdaBoost. It is well-accepted that the margin distribution is crucial for characterizing the performance of AdaBoost, and it is desirable to theoretically establish generalization bounds based on margin distribution.

In this paper, we first present the *kth margin bound* and further study on its relationship to previous work such as the minimum margin bound and Emargin bound. Then, we improve the empirical Bernstein bound with different skills. As our main results, we prove a new generalization bound which considers exactly the same factors as Schapire et al. [39] but is sharper than the bounds of Schapire et al. [39] and Breiman [9], and thus provide a complete answer to Breiman's doubt on the margin theory. By incorporating other factors such as average margin and variance, we present another generalization error bound which is heavily related to the whole margin distribution. In addition, we provide margin bounds for generalization error of voting classifiers in finite VC-dimension space. An interesting future issue is to develop new algorithms based on our theory.

Acknowledgements

We want to thank the editor and reviewers for helpful comments and suggestions. This work was supported by the National Fundamental Research Program of China (2010CB327903), the National Science Foundation of China (61073097, 61021062), the Jiangsu Province Graduate Students Innovative Research Project (CXZZ11_0046) and the Nanjing University PhD Students Promoting Program (201301A07).

References

[1] A. Antos, B. Kégl, T. Linder, G. Lugosi, Data-dependent margin-based generalization bounds for classification, Journal of Machine Learning Research 3 (2002) 73–98.
[2] J.Y. Audibert, R. Munos, C. Szepesvári, Exploration-exploitation tradeoff using variance estimates in multi-armed bandits, Theoretical Computer Science 410 (19) (2009) 1876–1902.
[3] P.L. Bartlett, M.I. Jordan, J.D. McAuliffe, Convexity, classification, and risk bounds, Journal of the American Statistical Association 101 (473) (2006) 138–156.
[4] P.L. Bartlett, M. Traskin, Adaboost is consistent, Journal of Machine Learning Research 8 (2007) 2347–2368.
[5] E. Bauer, R. Kohavi, An empirical comparison of voting classification algorithms: Bagging, boosting and variants, Machine Learning 36 (1) (1999) 105–139.
[6] J.P. Bickel, Y. Ritov, A. Zakai, Some theory for generalized boosting algorithms, Journal of Machine Learning Research 7 (2006) 705–732.
[7] A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth, Occam's razor, Information Processing Letter 24 (6) (1987) 377–380.
[8] L. Breiman, Arcing algorithm, Annuals of Statistics 26 (7) (1998) 801–849.
[9] L. Breiman, Prediction games and arcing classifiers, Neural Computation 11 (7) (1999) 1493–1517.
[10] L. Breiman, Some infinity theory for predictor ensembles, Tech. Rep. 577, Statistics Department, University of California, Berkeley, CA, 2000.
[11] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Chapman & Hall/CRC, Wadsworth, 1984.
[12] P. Buhlmann, B. Yu, Boosting with l_2 loss: Regression and classification, Journal of the American Statistical Association 98 (462) (2003) 324–339.
[13] R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in: Proceeding of 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, 2006, pp. 161–168.
[14] H. Chernoff, A measure of asymptotic efficiency of tests of a hypothesis based upon the sum of the observations, Annals of Mathematical Statistics 24 (4) (1952) 493–507.
[15] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, 1996.
[16] T. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization, Machine Learning 40 (2) (2000) 139–157.
[17] H. Drucker, C. Cortes, Boosting decision trees, in: D.S. Touretzky, M. Mozer, M.E. Hasselmo (Eds.), Advances in Neural Information Processing Systems, vol. 8, MIT Press, Cambridge, MA, 1996, pp. 479–485.
[18] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, in: Proceeding of 13th International Conference on Machine Learning, Bari, Italy, 1996, pp. 148–156.
[19] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55 (1) (1997) 119–139.
[20] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: A statistical view of boosting. with discussions, Annuals of Statistics 28 (2) (2000) 337–407.
[21] A. Garg, D. Roth, Margin distribution and learning, in: Proceedings of the 20th International Conference on Machine Learning, Washington, DC, 2003, pp. 210–217.
[22] A.J. Grove, D. Schuurmans, Boosting in the limit: Maximizing the margin of learned ensembles, in: Proceedings of the 15th National Conference on Artificial Intelligence, Menlo Park, CA, 1998, pp. 692–699.
[23] W. Hoeffding, Probability inequalities for sum of bounded random variables, Journal of American Statistical Society 58 (301) (1963) 13–30.
[24] W. Jiang, Process consistency for AdaBoost, Annuals of Statistics 32 (1) (2004) 13–29.
[25] L. Koltchinskii, D. Panchanko, Empirical margin distributions and bounding the generalization error of combined classifiers, Annuals of Statistics 30 (1) (2002) 1–50.
[26] L. Koltchinskii, D. Panchanko, Complexities of convex combinations and bounding the generalization error in classification, Annuals of Statistics 33 (4) (2005) 1455–1496.
[27] G. Lugosi, N. Vayatis, On the bayes-risk consistency of regularized boosting methods, Annuals of Statistics 32 (1) (2004) 30–55.
[28] L. Mason, J. Baxter, P.L. Bartlett, M.R. Frean, Boosting algorithms as gradient descent, in: S.A. Solla, T.K. Leen, K.-R. Müller (Eds.), Advances in Neural Information Processing Systems, vol. 12, MIT Press, Cambridge, MA, 1999, pp. 512–518.
[29] A. Maurer, Concentration inequalities for functions of independent variables, Random Structures and Algorithms 29 (2) (2006) 121–138.
[30] A. Maurer, M. Pontil, Empirical Bernstein bounds and sample-variance penalization, in: Proceedings of the 22nd Annual Conference on Learning Theory, Montreal, Canada, 2009.

[31] C. McDiarmid, On the method of bounded differences, in: Surveys in Combinatorics, Cambridge University Press, Cambridge, UK, 1989, pp. 148–188.
[32] C. McDiarmid, Concentration, in: Probabilistic Methods for Algorithmic Discrete Mathematics, Springer, 1998, pp. 195–248.
[33] D. Mease, A. Wyner, Evidence contrary to the statistical view of boosting with discussion, Journal of Machine Learning Research 9 (2008) 131–201.
[34] I. Mukherjee, C. Rudin, R. Schapire, The rate of convergence of Adaboost, in: Proceedings of the 24th Annual Conference on Learning Theory, Budapest, Hungary, 2011.
[35] J.R. Quinlan, Bagging, boosting, and C4.5, in: Proceeding of 13th National Conference on Artificial Intelligence, Portland, OR, 1996, pp. 725–730.
[36] G. Rätsch, T. Onoda, K.R. Müller, Soft margins for Adaboost, Machine Learning 42 (3) (2001) 287–320.
[37] L. Reyzin, R.E. Schapire, How boosting the margin can also boost classifier complexity, in: Proceeding of 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006, pp. 753–760.
[38] N. Sauer, On the density of families of sets, Journal of Combinatorial Theory, Series A 13 (1) (1972) 145–147.
[39] R. Schapire, Y. Freund, P.L. Bartlett, W. Lee, Boosting the margin: A new explanation for the effectives of voting methods, Annuals of Statistics 26 (5) (1998) 1651–1686.
[40] J. Shawe-Taylor, R.C. Williamson, Generalization performance of classifiers in terms of observed covering numbers, in: H.U.S.P. Fischer (Ed.), Proceedings of the 14th European Computational Learning Theory Conference, Springer, Berlin, 1999, pp. 153–167.
[41] C. Shen, H. Li, Boosting through optimization of margin distributions, IEEE Transactions on Neural Networks 21 (4) (2010) 659–666.
[42] P.K. Shivaswamy, T. Jebara, Variance penalizing adaboost, in: J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F.C.N. Pereira, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 24, MIT Press, Cambridge, MA, 2011, pp. 1908–1916.
[43] V.N. Vapnik, Statistical Learning Theory, John Wiley & Sons, New York, 1998.
[44] L.W. Wang, M. Sugiyama, C. Yang, Z.-H. Zhou, J. Feng, A refined margin analysis for boosting algorithms via equilibrium margin, Journal of Machine Learning Research 12 (2011) 1835–1863.
[45] X. Wu, V. Kumar, The Top Ten Algorithms in Data Mining, Chapman and Hall/CRC, 2009.
[46] T. Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, Annals of Statistics 32 (1) (2004) 56–85.
[47] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman & Hall/CRC, Boca Raton, FL, 2012.