Approximation Stability and Boosting

Wei Gao and Zhi-Hua Zhou

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
{gaow,zhouzh}@lamda.nju.edu.cn

Abstract. Stability has been explored to study the performance of learning algorithms in recent years and it has been shown that stability is sufficient for generalization and is sufficient and necessary for consistency of ERM in the general learning setting. Previous studies showed that AdaBoost has almost-everywhere uniform stability if the base learner has L_1 stability. The L_1 stability, however, is too restrictive and we show that AdaBoost becomes constant learner if the base learner is not real-valued learner. Considering that AdaBoost is mostly successful as a classification algorithm, stability analysis for AdaBoost when the base learner is not real-valued learner is an important yet unsolved problem. In this paper, we introduce the *approximation stability* and prove that approximation stability is sufficient for generalization, and sufficient and necessary for learnability of AERM in the general learning setting. We prove that AdaBoost has approximation stability and thus has good generalization, and an exponential bound for AdaBoost is provided.

1 Introduction

Stability has been considered as an important tool for studying the performance of learning algorithms in recent years. Intuitively, the stability of a learning algorithm can be referred as perturbation sensitivity in the training sample. It was first introduced in [5] for estimating leave-one-out error and further used to bound empirical risk of regression [10], which discovered a connection between finite VC dimension and stability. Bousquet and Elisseeff [3] obtained an exponential bound for uniform stability and proved that the Tikhonov regularized algorithms hold uniform stability property. Kutin and Niyogi [12] generalized the uniform stability to almost-everywhere algorithmic stability and derived generalization error bounds with extensions of McDiarmid's inequality. Stability has also been employed to bound the bias and variance of estimators for ERM (empirical risk minimization) or general algorithm [17]. An influential work of Mukherjee et al. [16] showed that stability is sufficient for generalization and sufficient and necessary for consistency of ERM in supervised regression and classification. Later, this result was extended to general learning setting by Shalev-Shwartz et al. [22].

AdaBoost [6, 7] is one of the most influential learning algorithms during the past decades. Many theoretical efforts have been devoted to studying the mysteries behind the great success of AdaBoost. There are different interpretations

from different aspects, and they have shed important insights for understanding the behaviors of AdaBoost. However, debates are still lasting to date, for example, on margin-based interpretation [4, 21, 19, 25] and statistical-view-based interpretation [2, 9, 15].

Considering the recent advances in stability, it is interesting to study the stability issues of AdaBoost. Kutin and Niyogi [11] proved that AdaBoost has almost-everywhere uniform stability if the base learner is L_1 stable. To the best of our knowledge, this is the only stability result for AdaBoost. The requirement of L_1 stability, however, is too restrictive, and as we will show in Section 4, AdaBoost becomes constant learner when the base learner is not real-valued learner. Note that as Freund and Schapire [8] indicated, AdaBoost is a *classification* algorithm, and so it is important to study the situation when the base learner is not real-valued learner.

In this paper, we introduce the notion of *approximation stability*, and prove that the approximation stability is sufficient for generalization, and is sufficient and necessary for learnability of AERM (asymptotical empirical risk minimization) in the general learning setting. Then, we prove that AdaBoost has approximation stability and thus has good generalization, and an exponential bound for AdaBoost is provided. All bounds obtained in this paper do not rely on any space complexity measure, but rather on the way the algorithm searches the space, and thus can be used even when the VC dimension is infinite.

In the rest of this paper we begin by introducing some notations and background knowledge in Section 2. Then, we give our results in Sections 3 and 4, and finally present the detail proofs in Section 5.

2 Preliminaries

2.1 Notations

Let $\mathcal{Z}$ denote an instance space and $\mathcal{D}$ denote an unknown probability distribution over $\mathcal{Z}$. We use $\Pr_{\mathcal{D}}[\cdot]$ to refer to the probability with respect to $\mathcal{D}$ and $\Pr_S[\cdot]$ to denote the probability with respect to a uniform distribution over the training sample S. Similarly, we use $E_D[\cdot]$ and $E_S[\cdot]$ to denote the expected values, respectively. For a positive number n, we denote by $[n]$ the set $\{1, 2, \cdots, n\}$. Given two distributions p and q with finite support, $||p - q||$ is defined to be the L_1-norm of $p - q$, i.e., $||p - q|| = \sum_{z \in \mathcal{Z}} |p(z) - q(z)|$. For a given sample $S = \{z_1, z_2, \cdots, z_n\}$ drawn i.i.d. according to distribution $\mathcal{D}$, let $S^i = S \setminus z_i$ be the sample with the i-th example z_i removed from S. For any $u \in \mathcal{Z}$, we denote by $S^{i,u} = S^i \cup \{u\}$ the sample with the i-th example z_i replaced by u in S.

A learning algorithm is a function $\mathbb{A}$ which maps a distribution p over $\mathcal{Z}$ onto a function $\mathbb{A}_p \in \mathcal{H}$, where $\mathcal{H}$ is a specific hypothesis class. Note that $\mathbb{A}_S$ means $\mathbb{A}_p$ where p is the uniform distribution on the training sample S. Throughout this paper, we consider symmetric algorithms, i.e., algorithms depend upon the given sample but not on the order of examples in the sample.

To measure the performance, we introduce a cost function $c\colon \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$. We assume such cost function is bounded by some constant B, i.e., $|c(h, z)| \leq B$

for every $h \in \mathcal{H}$ and $z \in \mathcal{Z}$. Given a sample S with size n and function $h \in \mathcal{H}$, define the empirical risk and expect risk, respectively, as

$$R_S(h) = E_{z\sim S}[c(h,z)] = \frac{1}{n}\sum_{z\in S} c(h,z) \text{ and } R(h) = E_{z\sim\mathcal{D}}[c(h,z)].$$

The general learning setting [24] is to minimize the expect risk, i.e., $\min_{h\in\mathcal{H}} R(h)$. Such setting comprises density estimation, stochastic optimization and supervised classification and regression. For instance, in supervised learning, $z = (x, y)$ is an instance-label pair and $c(h, z) = c(h(x), y)$ is the prediction loss for $h \in \mathcal{H}$.

Classical learning theory focuses on ERM, that is,

$$R_S(\mathbb{A}_S) = R_S(\hat{h}_S) = \min_{h\in\mathcal{H}} R_S(h),$$

where we denote by $\hat{h}_S = \arg\min_{h\in\mathcal{H}} R_S(h)$. A learning algorithm $\mathbb{A}$ is said to be AERM with rate $\epsilon_{\text{erm}}(n)$ under distribution $\mathcal{D}$ if

$$E_{S\sim\mathcal{D}^n}[R_S(\mathbb{A}_S) - R_S(\hat{h}_S)] \le \epsilon_{\text{erm}}(n).$$

We focus on AERM learning problems in this paper and ERM can be resolved in a similar way.

We say a learning algorithm $\mathbb{A}$ is consistent with rate $\epsilon_{\text{con}}(n)$ under distribution $\mathcal{D}$ if

$$E_{S\sim\mathcal{D}^n}[R(\mathbb{A}_S) - R(h^*)] \le \epsilon_{\text{con}}(n) \text{ for all } n,$$

where $h^* = \arg\min_{h\in\mathcal{H}} R(h)$. An algorithm $\mathbb{A}$ is universally consistent with rate $\epsilon_{\text{con}}(n)$ if it is consistent with rate $\epsilon_{\text{con}}(n)$ under all distributions $\mathcal{D}$ over $\mathcal{Z}$. A problem is learnable if there exists a universal consistent algorithm. The most influential result in classical learning theory for supervised classification and regression is that a problem is learnable if and only if the empirical risk $R_S(h)$ converges to the expect risk $R(h)$ [24]. This equivalence, however, does not always hold in the general learning setting [1, 23].

We say a learning algorithm $\mathbb{A}$ generalizes with rate $\epsilon_{\text{gen}}(n)$ under distribution $\mathcal{D}$ if

$$E_{S\sim\mathcal{D}^n}[|R(\mathbb{A}_S) - R_S(\mathbb{A}_S)|] \le \epsilon_{\text{gen}}(n) \text{ for all } n.$$

An algorithm $\mathbb{A}$ universally generalizes with $\epsilon_{\text{gen}}(n)$ if it generalizes with rate $\epsilon_{\text{gen}}(n)$ under all distributions $\mathcal{D}$ over $\mathcal{Z}$. In this paper we require $\epsilon_{\text{erm}}(n)$, $\epsilon_{\text{con}}(n)$, $\epsilon_{\text{gen}}(n) \to 0$ as $n \to \infty$.

2.2 Stability

Stability has been explored as an alternative for learnability. Definitions 1 and 2 show the CV stability [12, 17] and uniform stability [3], respectively.

Definition 1. *A learning algorithm $\mathbb{A}$ has CV stability $\eta(n)$ under distribution $\mathcal{D}$ if*

$$\forall\ i \in [n], \quad E_{S,u\sim\mathcal{D}^{n+1}}[|c(\mathbb{A}_S, u) - c(\mathbb{A}_{S^{i,u}}, u)|] \le \eta(n).$$

Algorithm 1 AdaBoost

Input: Sample $S = \{z_1 = (x_1, y_1), z_2 = (x_2, y_2), \cdots, z_n = (x_n, y_n)\} \in \mathcal{Z}^n$, base learner $\mathbb{A}$ and iteration rounds T.

Initialization: $P_S^1(z_i) = 1/n$ for each $z_i \in S$.

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

1. Call $\mathbb{A}$ with respect to distribution P_S^t to obtain a hypothesis $\mathbb{A}_{P_S^t}$.
2. Choose $\alpha_S^t = \frac{1}{2} \ln \frac{1-\mathrm{err}_S^t}{\mathrm{err}_S^t}$ with $\mathrm{err}_S^t = E_{z \sim P_S^t}[c(\mathbb{A}_{P_S^t}, z)]$, where $c(\mathbb{A}_{P_S^t}, z) = I[\mathbb{A}_{P_S^t}(x) \neq y]$.
3. Update $P_S^{t+1}(z_i) = \frac{1}{Z_t} P_S^t(z_i) \exp(-\alpha_S^t y_i \mathbb{A}_{P_S^t}(x_i))$, where Z_t is a normalization factor (such that P_S^{t+1} is a distribution).

end for

Output: The learner $\mathrm{sgn}(\mathbb{H}_S(x))$ where $\mathbb{H}_S(x) = \sum_{t=1}^T \alpha_S^t \mathbb{A}_{P_S^t}(x)$.

Definition 2. *A learning algorithm $\mathbb{A}$ has uniform stability $\beta(n)$ if*

$$\forall\ S \in \mathcal{Z}^n,\ \forall\ i \in [n]\ \textit{and}\ \forall\ z, u \in \mathcal{Z},\quad |c(\mathbb{A}_S, z) - c(\mathbb{A}_{S^{i,u}}, z)| \leq \beta(n).$$

A relevant concept, on-average-LOO stability [22], is defined as follows:

Definition 3. *A learning algorithm $\mathbb{A}$ has on-average-LOO stability $\beta(n)$ if*

$$\left| \frac{1}{n} \sum\nolimits_{i=1}^n E_{S \sim D^n}[c(\mathbb{A}_{S^i}, z_i) - c(\mathbb{A}_S, z_i)] \right| \leq \beta(n).$$

Here and whenever talking about "stability" $\beta(n)$ and $\eta(n)$, we require $\beta(n), \eta(n) \to 0$ as $n \to \infty$.

In this paper we will introduce *approximation stability* which is a kind of replacement version stability. As Shalev-Shwartz et al. [22] indicated, previously many researchers defined stability with respect to the deletion rather than replacement of an example. For instance, the deletion version uniform stability [5], the hypothesis stability [3], the cross-validation-(deletion) stability [17], the $\mathrm{CV}_{\mathrm{loo}}$ stability [16], etc. It is worth noting, however, that the deletion version stability implies the replacement version stability but not vice versa[1]; this is the reason why we focus on replacement version stability in this paper.

2.3 AdaBoost

Algorithm 1 shows a commonly used description of AdaBoost [6]. Kutin and Niyogi [11] studied the stability of AdaBoost and proved that when the base learner has L_1 stability and is real-valued with loss function $c(h, z) = |h(x) - y|$, AdaBoost has almost-everywhere uniform stability. The L_1 stability is given in Definition 4, which is equivalent to uniform stability, and the main result of [11] is shown in Theorem 1 using our notations.

[1] An example can be found in [16]: Let $\mathcal{Z} = \mathcal{X} \times \{+1, -1\}$ with $\mathcal{X}$ being uniform on $[0, 1]$. Suppose the target function is $t(x) = 1$ with 0/1 loss function. Given a sample S_n of size n, a non-AERM algorithm $\mathbb{A}_{S_n}(x) = (-1)^n$. Note that $\mathbb{A}_S$ does not have any deletion version stability but has replacement version uniform stability.

Definition 4. *A learning algorithm $\mathbb{A}$ has L_1 stability λ (constant) if $|c(\mathbb{A}_p, z) - c(\mathbb{A}_q, z)| \leq \lambda ||p - q||$ for any $z \in \mathcal{Z}$ and any given distributions p and q on $\mathcal{Z}$ with finite support.*

Theorem 1. *Suppose the base learner $\mathbb{A}$ has L_1 stability λ, and let*

$$\epsilon_\star = \frac{1}{2} \lim_{n\to\infty} \inf E_{S\sim\mathcal{D}^n}\Big[\inf_{\hat{S}\in\mathcal{Z}^m, m\leq n} R_S(\mathbb{A}_{\hat{S}})\Big] > 0.$$

Then, for all sufficiently large n and for all T, it holds for AdaBoost that

$$\Pr_{S\sim\mathcal{D}^n}[\forall\ i \in [n], \forall\ u, z \in \mathcal{Z}, |c(\mathbb{H}_S, z) - c(\mathbb{H}_{S^{i,u}}, z)| \leq \beta(n)] \geq 1 - \delta(n),$$

where $\beta(n) = \frac{2}{n}\sum_{t=1}^{T} 2^{t^2+1}(\lambda+1)^t/\epsilon_\star^{2t-1}$ and $\delta(n) = \exp(-n\epsilon_\star^2/2)$.

To the best of our knowledge, this is the only stability result for AdaBoost. It is worth noting, however, that it was obtained based on real-valued learner with the loss function $c(h, z) = |h(x) - y|$. We will show in Section 4 that for this result, AdaBoost becomes constant learner when the base learner is not real-valued learner. As Freund and Schapire [8] indicated, AdaBoost is a *classification* algorithm and therefore, it is important to study the stability of AdaBoost when the base learner is not real-valued learner, with the loss function $c(h, z) = I[h(x) \neq y]$ that is popularly used by classification algorithms; this remains an open problem and we will try to tackle it in the following sections.

3 Approximation Stability

We first introduce the empirical stability, expected empirical stability, validation stability and expected validation stability:

Definition 5. *A learning algorithm $\mathbb{A}$ has empirical stability $\beta(n)$ if*

$$\forall\ S \in \mathcal{Z}^n,\ \forall\ i \in [n]\ \text{and}\ \forall\ u \in \mathcal{Z},\quad |R_S(\mathbb{A}_S) - R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})| \leq \beta(n).$$

A learning algorithm $\mathbb{A}$ has expected empirical stability $\beta(n)$ under distribution $\mathcal{D}$ if

$$\forall\ i \in [n],\quad E_{S,u\sim\mathcal{D}^{n+1}}[|R_S(\mathbb{A}_S) - R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})|] \leq \beta(n).$$

Definition 6. *A learning algorithm $\mathbb{A}$ has validation stability $\beta(n)$ under distribution $\mathcal{D}$ if*

$$\forall\ S \in \mathcal{Z}^n\ \text{and}\ \forall\ i \in [n],\quad |R(\mathbb{A}_S) - E_{u\sim\mathcal{D}}[c(\mathbb{A}_{S^{i,u}}, u)]| \leq \beta(n).$$

A learning algorithm $\mathbb{A}$ has expected validation stability $\beta(n)$ under distribution $\mathcal{D}$ if

$$\forall\ i \in [n],\quad E_{S\sim\mathcal{D}^n}[|R(\mathbb{A}_S) - E_{u\sim\mathcal{D}}[c(\mathbb{A}_{S^{i,u}}, u)]|] \leq \beta(n).$$

An algorithm $\mathbb{A}$ has universally expected validation stability $\beta(n)$ if the stability holds with $\beta(n)$ for all distributions $\mathcal{D}$ over $\mathcal{Z}$. Combining the expected empirical stability and expected validation stability gives approximation stability:

Definition 7. *A learning algorithm $\mathbb{A}$ has approximation stability $(\beta_1(n), \beta_2(n))$ under distribution $\mathcal{D}$ if it exhibits both expected empirical stability $\beta_1(n)$ and expected validation stability $\beta_2(n)$.*

We prove that approximation stability is sufficient for generalization in the following theorem:

Theorem 2. *If an algorithm $\mathbb{A}$ has approximation stability $(\beta_1(n), \beta_2(n))$, then $\mathbb{A}$ generalizes with rate $\epsilon_{gen}(n) = B/\sqrt{n} + \sqrt{3\beta_1(n)B/2 + 4\beta_2(n)B + 3B^2/\sqrt{n}}$, that is,*

$$E_{S\sim\mathcal{D}^n}[|R(\mathbb{A}_S) - R_S(\mathbb{A}_S)|] \leq B/\sqrt{n} + \sqrt{3\beta_1(n)B/2 + 4\beta_2(n)B + 3B^2/\sqrt{n}}.$$

Note that the CLT (central limit theorem) guarantees that the average of i.i.d. random variables converges to expectation. However, $\mathbb{A}_S$ is dependent on S and thus the CLT is not applicable. The proof in Section 5 shows that the combination of expected validation stability and expected empirical stability implies generalization, though neither the expected validation stability nor the expected empirical stability is sufficient.

Next, we study the relationship between the approximation stability and the learnability of AERM in the general learning setting. Lemma 1 shows that AERM implies expected empirical stability. Hence we only need to study the relationship between the expected validation stability and the learnability of AERM. Theorem 3 establishes the equivalence between them.

Lemma 1 (AERM $\Rightarrow$ Expected empirical stability). *If a learning algorithm $\mathbb{A}$ is AERM with rate $\epsilon_{erm}(n)$ under distribution $\mathcal{D}$, then $\mathbb{A}$ has expected empirical stability $\beta(n) = 2\epsilon_{erm}(n) + 2B/n$.*

Proof. For any $i \in [n]$ and any $u \in \mathcal{Z}$, we have

$$\begin{aligned} E_{S\sim\mathcal{D}^n}[|R_S(\mathbb{A}_S) - R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})|] &\leq E_{S\sim\mathcal{D}^n}[|R_S(\mathbb{A}_S) - R_S(\hat{h}_S)|] \\ + E_{S\sim\mathcal{D}^n}[|R_S(\hat{h}_S) - R_{S^{i,u}}(\hat{h}_{S^{i,u}})|] &+ E_{S\sim\mathcal{D}^n}[|R_{S^{i,u}}(\hat{h}_{S^{i,u}}) - R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})|] \\ &\leq 2\epsilon_{\text{erm}}(n) + 2B/n, \end{aligned}$$

since $|R_S(\hat{h}_S) - R_{S^{i,u}}(\hat{h}_{S^{i,u}})| \leq 2B/n$ from the definition of ERM. □

Theorem 3. *The following are equivalent for an AERM:*

- *Universal expected validation stability;*
- *Universal consistency;*
- *Universal generalization.*

The equivalence between the on-average-LOO stability and learnability has been established in [22]. We thus work by establishing the equivalence between the expected validation stability and on-average-LOO stability under AERM, though this equivalence does not always hold in general and we use other techniques in such case in the proof.

It is worth noting that the uniform stability [3], which could be strictly stronger than any other stability, is not necessary for learnability [22]. Mukherjee et al. [16] suggested that LOO stability implies generalization, and is necessary and sufficient for consistency of ERM via uniform convergence of $R_S(h)$ to $R(h)$. It is well-known that uniform convergence is not equivalent to ERM consistency [1, 23] and thus their work is specific to supervised learning. In the general learning setting, the equivalence between on-average-LOO stability and learnability has been established for AERM [22]. However, out of the AERM framework there are also many useful learning algorithms, on which the on-average-LOO stability could not be applied. For instance, we could not guarantee that AdaBoost is AERM, and thus our approximation stability is meaningful for its analysis. Overall, comparing to previous stabilities, our approximation stability does not only promise generalization for general algorithm, but also guarantee sufficiency and necessity of learnability of AERM in the general setting.

Finally, we derive a bound for learning algorithm $\mathbb{A}$ which has both empirical stability and validation stability. The following theorem shows that the empirical risk converges to expect risk with high probability when $\beta_1(n) = o(n^{-\frac{1}{2}})$ and $\beta_2(n) = o(n^{-\frac{1}{2}})$ where $o(n)$ represents $\frac{o(n)}{n} \to 0$ as $n \to \infty$.

Theorem 4. *If a learning algorithm $\mathbb{A}$ has both empirical stability $\beta_1(n)$ and validation stability $\beta_2(n)$ under distribution $\mathcal{D}$, then, for all $n \geq 1$ and $\epsilon > 0$,*

$$\Pr_{S\sim\mathcal{D}^n}[|R(\mathbb{A}_S) - R_S(\mathbb{A}_S)| \geq \epsilon + \beta_2(n)] \leq 2\exp\left(\frac{-2\epsilon^2}{n(\beta_1(n) + 2\beta_2(n))^2}\right).$$

Uniform stability is sufficient for exponential generalization bound [3], however, it can only be used for regression or classification with real-valued learners. Note that the uniform stability implies empirical stability and validation stability, but not vice versa. Thus we get an exponential bound though our assumption is weaker than that used by [3] for the uniform stability bound.

Proof. Let $F(S) = R(\mathbb{A}_S) - R_S(\mathbb{A}_S)$. For any $i \in [n]$, we have

$$\begin{aligned}&|E_{S\sim\mathcal{D}^n}[F(S)]| \leq |E_{S,u\sim\mathcal{D}^{n+1}}[R(\mathbb{A}_S) - R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})]| \\ &+ |E_{S,u\sim\mathcal{D}^{n+1}}[R_{S^{i,u}}(\mathbb{A}_{S^{i,u}}) - R_S(\mathbb{A}_S)]| = |E_{S,u\sim\mathcal{D}^{n+1}}[R(\mathbb{A}_S) - R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})]|,\end{aligned}$$

by using $E_{S,u\sim\mathcal{D}^{n+1}}[R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})] = E_{S\sim\mathcal{D}^n}[R_S(\mathbb{A}_S)]$. From symmetry and i.i.d assumption, $E_{S,u\sim\mathcal{D}^{n+1}}[R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})] = E_{S,u\sim\mathcal{D}^{n+1}}[c(\mathbb{A}_{S^{i,u}}, u)]$, which leads to

$$\begin{aligned}&|E_{S,u\sim\mathcal{D}^{n+1}}[R(\mathbb{A}_S) - R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})]| \\ &\qquad = |E_{S\sim\mathcal{D}^n}[R(\mathbb{A}_S) - E_{u\sim\mathcal{D}}[c(\mathbb{A}_{S^{i,u}}, u)]]| \leq \beta_2(n).\end{aligned}$$

Thus we bound $|E_{S\sim\mathcal{D}^n}[F(S)]| \leq \beta_2(n)$. Meanwhile, it holds

$$\begin{aligned}|F(S) - F(S^{i,u})| &\leq |R(\mathbb{A}_S) - R(\mathbb{A}_{S^{i,u}})| + |R_S(\mathbb{A}_S) - R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})| \\ &\leq |R(\mathbb{A}_S) - E_{z\sim\mathcal{D}}[c(\mathbb{A}_{S^{i,z}}, z)] + E_{z\sim\mathcal{D}}[c(\mathbb{A}_{S^{i,z}}, z)] - R(\mathbb{A}_{S^{i,u}})| \\ &\qquad + \beta_1(n) \leq \beta_1(n) + 2\beta_2(n).\end{aligned}$$

This theorem follows by applying McDiarmid formula [14] to $F(S)$. □

4 Approximation Stability for AdaBoost

The following lemma shows that the L_1 stability is too restrictive for non real-valued learners with cost function $c(h,z) = I[h(x) \neq y]$.

Lemma 2. *If a learning algorithm $\mathbb{A}$ has L_1 stability λ, then $\mathbb{A}_{S_n}$ is a constant algorithm for $n > 2\lambda$.*

Proof. From the definition of L_1 stability, we have

$$\forall\ S \in \mathcal{Z}^n, \forall\ i \in [n] \text{ and } \forall\ z, u \in \mathcal{Z}, \quad |c(\mathbb{A}_S, z) - c(\mathbb{A}_{S^{i,u}}, z)| \leq 2\lambda/n.$$

It follows $\mathbb{A}_S(z) = \mathbb{A}_{S^{i,u}}(z)$ since $c(h,z) \in \{0,1\}$ and $|c(\mathbb{A}_S, z) - c(\mathbb{A}_{S^{i,u}}, z)| < 1$ for $n > 2\lambda$. We can also prove $\mathbb{A}_S(z) = \mathbb{A}_{S^i}(z)$ for all $z \in \mathcal{Z}$ in a similar way. □

If the base learner in AdaBoost is not real-valued learner for large-size sample, then the base learner is a constant learner according to the above lemma. This follows that AdaBoost becomes a constant learner. Thus, Theorem 1, the only stability result for AdaBoost, does not completely explain the stability of AdaBoost for general base learner.

Since AdaBoost is mostly successful as a classification algorithm, in contrast to considering real-valued base learner with loss function $c(h,z) = |h(x) - y|$, it may be more interesting to consider non real-valued base learner with loss function $c(h,z) = I[h(x) \neq y]$ which is adopted by classifiers such as decision trees and decision stumps that are popularly used with AdaBoost in practice.

Below we will discuss the stability of AdaBoost. Observing that

$$\Pr_S[y \neq \text{sgn}(\mathbb{H}_S(x))] = E_S[I[y\mathbb{H}_S(x) \leq 0]] \leq E_S[\exp(-y\mathbb{H}_S(x))],$$

we choose the cost function for $\mathbb{H}_S(x)$ as $c(\mathbb{H}_S, z) = \exp(-y\mathbb{H}_S(x))$. This is also in accordance with the theory that AdaBoost can be regarded as a coordinate descent algorithm [4, 9, 13, 18] for minimizing $R_S(\mathbb{H}_S)$. For base learner, we set the cost function $c(\mathbb{A}_S^t, z) = I[\mathbb{A}_S^t(x) \neq y]$ described in Algorithm 1.

We assume the iteration number T for AdaBoost is given in advance, and thus T is a constant since stability could not be used to analyze AdaBoost for unfixed or infinite T. We also assume $\gamma \leq \text{err}_S^t \leq 1 - \gamma$ for some small $\gamma > 0$, because $c(\mathbb{H}_S, z)$ may approach to infinity if $\text{err}_S^t \to 0$ or $\text{err}_S^t \to 1$, which goes beyond our discussion (bounded cost function). Such assumption can be viewed as a variation of "bounded edges" in [20]. A bound for $c(\mathbb{H}_S, z)$ is given as follows.

Lemma 3. *For constant $T \geq 1$ and any $S \in \mathcal{Z}^n$, if the base learner in each iteration satisfies $\gamma \leq err_S^t \leq 1 - \gamma$ with $\gamma > 0$, then $c(\mathbb{H}_S, z) \leq \big((1-\gamma)/\gamma\big)^{T/2}$.*

Proof. Since $\alpha_S^t = \frac{1}{2}\ln((1-\text{err}_S^t)/\text{err}_S^t)$ and $\gamma \leq \text{err}_S^t \leq 1-\gamma$, we have $\frac{1}{2}\ln(\gamma/(1-\gamma)) \leq \alpha_S^t \leq \frac{1}{2}\ln((1-\gamma)/\gamma)$. It follows $c(\mathbb{H}_S, z) = \exp\big(-y\sum_{t=1}^T \alpha_S^t \mathbb{A}_{P_S^t}(x)\big) \leq \exp\big(\sum_{t=1}^T |\alpha_S^t|\big) \leq \big((1-\gamma)/\gamma\big)^{T/2}$ as desired. □

Denote by B the bound of $c(\mathbb{H}_S, z)$ for notational simplicity. We have the following theorem on the approximation stability of AdaBoost:

Theorem 5. *AdaBoost has approximation stability $(\beta_1(n), \beta_2(n))$ for constant $T \geq 1$, if the base learner in each round has CV stability $\eta(n)$, and for any $u \in \mathcal{Z}$, $i \in [n]$, $t \in [T]$ and small $\gamma > 0$, the following holds:*

$$E_{S,u\sim\mathcal{D}^{n+1}}[|err_S^t - err_{S^{i,u}}^t|] \leq \zeta(n) \quad \text{and} \quad \gamma \leq err_S^t,\ err_{S^{i,u}}^t \leq 1-\gamma.$$

Here

$$\beta_1(n) = \frac{\zeta(n)T}{\sqrt{\gamma(1-\gamma)}} \text{ and } \beta_2(n) = \frac{BT}{2}\left(\eta(n)\ln\frac{1-\gamma}{\gamma} + \frac{\zeta(n)}{\gamma(1-\gamma)}\right).$$

Also, we have

$$E_{S\sim\mathcal{D}^n}[|R(\mathbb{H}_S) - R_S(\mathbb{H}_S)|] \leq B/\sqrt{n} + \sqrt{3\beta_1(n)B/2 + 4\beta_2(n)B + 3B^2/\sqrt{n}}.$$

We can also have a tighter bound for AdaBoost by considering Theorem 4:

Theorem 6. *AdaBoost has empirical stability $\beta_1(n)$ and validation stability $\beta_2(n)$ for constant $T \geq 1$, if for any $u, S \in \mathcal{Z}^{n+1}$, $i \in [n]$, $t \in [T]$ and small $\gamma > 0$, the following holds:*

$$E_{u\sim\mathcal{D}}[|c(\mathbb{A}_S, u) - c(\mathbb{A}_{S^{i,u}}, u)|] \leq \eta(n), \quad |err_S^t - err_{S^{i,u}}^t| < \zeta(n),$$

and $\gamma \leq err_S^t, err_{S^{i,u}}^t \leq 1-\gamma$. Here

$$\beta_1(n) = \frac{\zeta(n)T}{\sqrt{\gamma(1-\gamma)}} \text{ and } \beta_2(n) = \frac{BT}{2}\left(\eta(n)\ln\frac{1-\gamma}{\gamma} + \frac{\zeta(n)}{\gamma(1-\gamma)}\right).$$

For $\epsilon > 0$, we have

$$\Pr_{S\sim\mathcal{D}^n}[|R(\mathbb{H}_S) - R_S(\mathbb{H}_S)|] \geq \epsilon + \beta_2(n)] \leq 2\exp\left(\frac{-2\epsilon^2}{n(\beta_1(n) + 2\beta_2(n))^2}\right).$$

5 Proofs

This section presents detail proofs of our main theorems. Before proceeding our proofs, we introduce some tools which will be used:

Proposition 1. *[22] Let $|X_i| \leq B$ and $X = \sum_{i=1}^n X_i/n$ for i.i.d. X_i. Then we have $E[|X - E[X]|] \leq B/\sqrt{n}$.*

Proposition 2. *[22] If X, Y are random variables s.t. $X \leq Y$ almost surely, then $E[|X|] \leq |E[X]| + 2E[|Y|]$.*

Proposition 3. *If a learning algorithm $\mathbb{A}$ has expected validation stability $\beta(n)$ under distribution $\mathcal{D}$, then $E_{S,u\sim\mathcal{D}^{n+1}}[|R(\mathbb{A}_S) - R(\mathbb{A}_{S^{i,u}})|] \leq 2\beta(n)$ for all $i \in [n]$.*

The last proposition follows from the fact $E_{S,u\sim\mathcal{D}^{n+1}}[|R(\mathbb{A}_S) - R(\mathbb{A}_{S^{i,u}})|] = E_{S,u\sim\mathcal{D}^{n+1}}[|E_{z\sim\mathcal{D}}[c(\mathbb{A}_S, z) - c(\mathbb{A}_{S^{i,z}}, z) + c(\mathbb{A}_{S^{i,z}}, z) - c(\mathbb{A}_{S^{i,u}}, z)]|]$.

5.1 Proof of Theorem 2

We start by introducing a ghost sample $\hat{S} = \{\hat{z}_1, \hat{z}_2, \cdots, \hat{z}_n\}$ drawn i.i.d according to distribution $\mathcal{D}$ and denote $R_{\hat{S}}(\mathbb{A}_S) = \sum_{i=1}^{n} c(\mathbb{A}_S, \hat{z}_i)/n$. It follows

$$E_{S\sim\mathcal{D}^n}[|R_S(\mathbb{A}_S) - R(\mathbb{A}_S)|] \leq E_{S,\hat{S}\sim\mathcal{D}^{2n}}[|R(\mathbb{A}_S) - R_{\hat{S}}(\mathbb{A}_S)|] + E_{S,\hat{S}\sim\mathcal{D}^{2n}}[|R_S(\mathbb{A}_S) - R_{\hat{S}}(\mathbb{A}_S)|].$$

We bound the first term by $E_{S,\hat{S}\sim\mathcal{D}^{2n}}[|R(\mathbb{A}_S) - R_{\hat{S}}(\mathbb{A}_S)|] \leq B/\sqrt{n}$ from Proposition 1 since $\mathbb{A}_S$ is independent of $\hat{S}$. For the second term, using the Jensen's inequality,

$$E_{S,\hat{S}\sim\mathcal{D}^{2n}}[|R_S(\mathbb{A}_S) - R_{\hat{S}}(\mathbb{A}_S)|] \leq \sqrt{E_{S,\hat{S}\sim\mathcal{D}^{2n}}[(R_S(\mathbb{A}_S) - R_{\hat{S}}(\mathbb{A}_S))^2]}.$$

To bound this expression, we introduce a random permutation which swaps elements between S and $\hat{S}$, i.e., a permutation σ on $\{1, 2, \cdots, n, \hat{1}, \hat{2}, \cdots, \hat{n}\}$ s.t. $\{\sigma(i), \sigma(\hat{i})\} = \{i, \hat{i}\}$. Denote by S^σ and $\hat{S}^\sigma$ the permuted samples of S and $\hat{S}$, respectively, and define z_i^σ and $\hat{z}_i^\sigma$ in an obvious way. Since S and $\hat{S}$ are chosen i.i.d according to distribution $\mathcal{D}$, $E_{S,\hat{S}\sim\mathcal{D}^{2n}}[(R_S(\mathbb{A}_S) - R_{\hat{S}}(\mathbb{A}_S))^2]$ equals to

$$\begin{aligned}
&E_{S,\hat{S}\sim\mathcal{D}^{2n}}\Big[\sum\nolimits_\sigma \big(R_{S^\sigma}(\mathbb{A}_{S^\sigma}) - R_{\hat{S}^\sigma}(\mathbb{A}_{S^\sigma})\big)^2/2^n\Big] \\
&= \frac{1}{n^2 2^n} E_{S,\hat{S}\sim\mathcal{D}^{2n}}\Big[\sum_{i,j,\sigma} \big(c(\mathbb{A}_{S^\sigma}, z_i^\sigma) - c(\mathbb{A}_{S^\sigma}, \hat{z}_i^\sigma)\big)\big(c(\mathbb{A}_{S^\sigma}, z_j^\sigma) - c(\mathbb{A}_{S^\sigma}, \hat{z}_j^\sigma)\big)\Big].
\end{aligned}$$

Given σ and i, we define two permutations σ_1 and σ_2 as follows: $\sigma_1(i) = i$, $\sigma_1(\hat{i}) = \hat{i}$, $\sigma_2(i) = \hat{i}$, $\sigma_2(\hat{i}) = i$ and $\sigma_1(k) = \sigma_2(k) = \sigma(k)$ for $k \neq i, \hat{i}$. It holds

$$\begin{aligned}
&\sum\nolimits_\sigma \big(c(\mathbb{A}_{S^\sigma}, z_i^\sigma) - c(\mathbb{A}_{S^\sigma}, \hat{z}_i^\sigma)\big)\big(c(\mathbb{A}_{S^\sigma}, z_j^\sigma) - c(\mathbb{A}_{S^\sigma}, \hat{z}_j^\sigma\big) \\
&= \sum\nolimits_{\sigma_1} \big(c(\mathbb{A}_{S^{\sigma_1}}, z_i^{\sigma_1}) - c(\mathbb{A}_{S^{\sigma_1}}, \hat{z}_i^{\sigma_1})\big)\big(c(\mathbb{A}_{S^{\sigma_1}}, z_j^{\sigma_1}) - c(\mathbb{A}_{S^{\sigma_1}}, \hat{z}_j^{\sigma_1})\big)/2 \\
&\quad + \sum\nolimits_{\sigma_2} \big(c(\mathbb{A}_{S^{\sigma_2}}, z_i^{\sigma_2}) - c(\mathbb{A}_{S^{\sigma_2}}, \hat{z}_i^{\sigma_2})\big)\big(c(\mathbb{A}_{S^{\sigma_2}}, z_j^{\sigma_2}) - c(\mathbb{A}_{S^{\sigma_2}}, \hat{z}_j^{\sigma_2})\big)/2 \\
&= \sum(\Theta_{ij} + \Delta_{ij})/2
\end{aligned}$$

where $\Theta_{ij} = \chi_3(\chi_4 - \chi_2)$ and $\Delta_{ij} = \chi_2(\chi_1 + \chi_3)$ with $\chi_1 = c(\mathbb{A}_{S^{\sigma_1}}, z_i^{\sigma_1}) - c(\mathbb{A}_{S^{\sigma_1}}, \hat{z}_i^{\sigma_1})$, $\chi_2 = c(\mathbb{A}_{S^{\sigma_1}}, z_j^{\sigma_1}) - c(\mathbb{A}_{S^{\sigma_1}}, \hat{z}_j^{\sigma_1})$, $\chi_3 = c(\mathbb{A}_{S^{\sigma_2}}, z_i^{\sigma_2}) - c(\mathbb{A}_{S^{\sigma_2}}, \hat{z}_i^{\sigma_2})$ and $\chi_4 = c(\mathbb{A}_{S^{\sigma_2}}, z_j^{\sigma_2}) - c(\mathbb{A}_{S^{\sigma_2}}, \hat{z}_j^{\sigma_2})$. Noting σ_1 and σ_2 are independent of j,

$$\begin{aligned}
E_{S,\hat{S}}\Big[\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{n} \Theta_{ij}/n^2\Big] = E_{S,\hat{S}}\Big[\sum\nolimits_{i=1}^{n} &\big(c(\mathbb{A}_{S^{\sigma_2}}, z_i^{\sigma_2}) - c(\mathbb{A}_{S^{\sigma_2}}, \hat{z}_i^{\sigma_2})\big) \\
&\times \big(R_{S^{\sigma_2}}(\mathbb{A}_{S^{\sigma_2}}) - R_{S^{\sigma_1}}(\mathbb{A}_{S^{\sigma_1}}) + R_{\hat{S}^{\sigma_1}}(\mathbb{A}_{S^{\sigma_1}}) - R_{\hat{S}^{\sigma_2}}(\mathbb{A}_{S^{\sigma_2}})\big)/n\Big].
\end{aligned}$$

Since $|c(\mathbb{A}_S, z)| \leq B$ and $\mathbb{A}$ has approximation stability $(\beta_1(n), \beta_2(n))$, we obtain $E_{S,\hat{S}\sim\mathcal{D}^{2n}}[|R_{S^{\sigma_1}}(\mathbb{A}_{S^{\sigma_1}}) - R_{S^{\sigma_2}}(\mathbb{A}_{S^{\sigma_2}})|] \leq \beta_1(n)$ and

$$\begin{aligned} E_{S,\hat{S}\sim\mathcal{D}^{2n}}[|R_{\hat{S}^{\sigma_2}}(\mathbb{A}_{S^{\sigma_2}}) - R_{\hat{S}^{\sigma_1}}(\mathbb{A}_{S^{\sigma_1}})|] &\leq E_{S,\hat{S}\sim\mathcal{D}^{2n}}[|R_{\hat{S}^{\sigma_2}}(\mathbb{A}_{S^{\sigma_2}}) - R(\mathbb{A}_{S^{\sigma_2}})|] \\ + E_{S,\hat{S}\sim\mathcal{D}^{2n}}[|R_{\hat{S}^{\sigma_1}}(\mathbb{A}_{S^{\sigma_1}}) - R(\mathbb{A}_{S^{\sigma_1}})|] &+ E_{S,\hat{S}\sim\mathcal{D}^{2n}}[|R(\mathbb{A}_{S^{\sigma_1}}) - R(\mathbb{A}_{S^{\sigma_2}})|] \\ &\leq 2\beta_2(n) + 2B/\sqrt{n}, \quad (1) \end{aligned}$$

from Proposition 1 and Proposition 3. Thus we show

$$\Big|E_{S,\hat{S}}\Big[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Theta_{ij}\Big]\Big| \leq 2B\beta_1(n) + 4B\beta_2(n) + 4B^2/\sqrt{n}. \quad (2)$$

For $E_{S,\hat{S}}[\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_{ij}/n^2]$, we also have

$$\begin{aligned} E_{S,\hat{S}}\Big[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_{ij}\Big] &= E_{S,\hat{S}}\Big[\frac{1}{n}\sum_{i=1}^{n}\big(R_{S^{\sigma_1}}(\mathbb{A}_{S^{\sigma_1}}) - R_{\hat{S}^{\sigma_1}}(\mathbb{A}_{S^{\sigma_1}})\big) \\ &\times \big(c(\mathbb{A}_{S^{\sigma_1}}, z_i^{\sigma_1}) - c(\mathbb{A}_{S^{\sigma_1}}, \hat{z}_i^{\sigma_1}) + c(\mathbb{A}_{S^{\sigma_2}}, z_i^{\sigma_2}) - c(\mathbb{A}_{S^{\sigma_2}}, \hat{z}_i^{\sigma_2})\big)\Big]. \end{aligned}$$

This expression could not be summed directly since σ_1 and σ_2 are dependent on i. But from symmetry and i.i.d assumption, we have

$$\begin{aligned} E_{S,\hat{S}}\Big[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_{ij}\Big] &= E_{S,\hat{S}}[\big(R_{S^{\sigma_1^*}}(\mathbb{A}_{S^{\sigma_1^*}}) - R_{\hat{S}^{\sigma_1^*}}(\mathbb{A}_{S^{\sigma_1^*}})\big) \\ &\times \big(c(\mathbb{A}_{S^{\sigma_1^*}}, z_1) - c(\mathbb{A}_{S^{\sigma_1^*}}, \hat{z}_1) + c(\mathbb{A}_{S^{\sigma_2^*}}, \hat{z}_1) - c(\mathbb{A}_{S^{\sigma_2^*}}, z_1)\big)], \end{aligned}$$

where $\sigma_1^*(1) = 1$, $\sigma_1^*(\hat{1}) = \hat{1}$, $\sigma_2^*(1) = \hat{1}$, $\sigma_2^*(\hat{1}) = 1$ and $\sigma_1^*(k) = \sigma_2^*(k) = \sigma(k)$ for $k \neq 1, \hat{1}$. Let $z, \hat{z}$ be two new examples and set $S_1 = S^{1,z}$, $\hat{S}_1 = \hat{S}^{\hat{1},\hat{z}}$. In a similar way to prove Eq.(2), we have

$$\begin{aligned} E_{S,\hat{S},z,\hat{z}}[|R_{S^{\sigma_1^*}}(\mathbb{A}_{S^{\sigma_1^*}}) - R_{\hat{S}^{\sigma_1^*}}(\mathbb{A}_{S^{\sigma_1^*}}) - R_{S_1^{\sigma_1^*}}(\mathbb{A}_{S_1^{\sigma_1^*}}) + R_{\hat{S}_1{}^{\sigma_1^*}}(\mathbb{A}_{S_1^{\sigma_1^*}})|] \\ \leq \beta_1(n) + 2\beta_2(n) + 2B/\sqrt{n}. \quad (3) \end{aligned}$$

Since z_1 and $\hat{z}_1$ are independent to S_1 and $\hat{S}_1$, it holds

$$\begin{aligned} &\Big|E_{S,\hat{S},z,\hat{z}}\Big[\big(c(\mathbb{A}_{S^{\sigma_1^*}}, z_1) - c(\mathbb{A}_{S^{\sigma_1^*}}, \hat{z}_1) + c(\mathbb{A}_{S^{\sigma_2^*}}, \hat{z}_1) - c(\mathbb{A}_{S^{\sigma_2^*}}, z_1)\big)\times \\ &\big(R_{S_1^{\sigma_1^*}}(\mathbb{A}_{S_1^{\sigma_1^*}}) - R_{\hat{S}_1{}^{\sigma_1^*}}(\mathbb{A}_{S_1^{\sigma_1^*}})\big)\Big]\Big| = \Big|E_{S^1,\hat{S}^1,z,\hat{z}}\Big[\big(R_{S_1^{\sigma_1^*}}(\mathbb{A}_{S_1^{\sigma_1^*}}) - R_{\hat{S}_1{}^{\sigma_1^*}}(\mathbb{A}_{S_1^{\sigma_1^*}})\big) \\ &\times \big(E_{z_1\sim\mathcal{D}}[c(\mathbb{A}_{S^{\sigma_1^*}}, z_1) - c(\mathbb{A}_{S^{\sigma_2^*}}, z_1)] + E_{\hat{z}_1\sim\mathcal{D}}[c(\mathbb{A}_{S^{\sigma_2^*}}, \hat{z}_1) - c(\mathbb{A}_{S^{\sigma_1^*}}, \hat{z}_1)]\big)\Big]\Big| \\ &\leq 4B\beta_2(n). \end{aligned}$$

Thus we derive $\Big|E_{S,\hat{S}}\Big[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_{ij}\Big]\Big| \leq 4B(\beta_1(n)+3\beta_2(n)+2B/\sqrt{n})$ from Eq.(3), which yields $E_{S,\hat{S}\sim\mathcal{D}^{2n}}[(R_S(\mathbb{A}_S) - R_{\hat{S}}(\mathbb{A}_S))^2] \leq 3B\beta_1(n)/2 + 4B\beta_2(n) + 3B^2/\sqrt{n}$ from Eq.(2). This theorem follows. □

5.2 Proof of Theorem 3

The equivalence between the universal consistency and universal generalization has been established in [22]. Combining Lemma 1 and Theorem 2 proves that $\mathbb{A}$ generalizes with rate $\epsilon_{\text{gen}}(n) = \sqrt{3\epsilon_{\text{erm}}(n)B + 4\beta_2(n)B + 3B^2/\sqrt{n} + 3B^2/n} + B/\sqrt{n}$ if $\mathbb{A}$ is AERM with rate $\epsilon_{\text{erm}}(n)$ and has expected validation stability $\beta_2(n)$. Thus Theorem 3 follows from the following lemma.

Lemma 4 (AERM + generalization + consistency ⇒ Expected validation stability). *Suppose $\mathbb{A}$ is consistent with rate $\epsilon_{cons}(n)$, generalized with rate $\epsilon_{gen}(n)$, and AERM with rate $\epsilon_{erm}(n)$ such that $n\epsilon_{erm}(n) \to 0$ as $n \to \infty$. Then $\mathbb{A}$ has expected validation stability $\beta(n) = \epsilon_{gen} + 4\epsilon_{cons} + 2n\epsilon_{erm}(n)$.*

Proof. For $i \in [n]$, since $|E_{u\sim\mathcal{D}}[c(\mathbb{A}_S, u) - c(\mathbb{A}_{S^{i,u}}, u)]| \leq |E_{u\sim\mathcal{D}}[c(\mathbb{A}_S, u) - E_{z\sim\mathcal{D}}[c(\mathbb{A}_{S^{i,z}}, u)]]| + |E_{u\sim\mathcal{D}}[E_{z\sim\mathcal{D}}[c(\mathbb{A}_{S^{i,z}}, u)] - c(\mathbb{A}_{S^{i,u}}, u)]|$, we have

$$\begin{aligned} &E_{S\sim\mathcal{D}^n}[|E_{u\sim\mathcal{D}}[c(\mathbb{A}_S, u) - c(\mathbb{A}_{S^{i,u}}, u)]|] \\ &\qquad \leq E_{S,z\sim\mathcal{D}^{n+1}}[|E_{u\sim\mathcal{D}}[c(\mathbb{A}_S, u) - c(\mathbb{A}_{S^{i,z}}, u)]|] + \\ &\quad E_{S\sim\mathcal{D}^n}[|E_{u,z\sim\mathcal{D}^2}[c(\mathbb{A}_{S^{i,z}}, u) - c(\mathbb{A}_{S^{i,u}}, u) + c(\mathbb{A}_{S^{i,u}}, z) - c(\mathbb{A}_{S^{i,z}}, z)]|]/2. \end{aligned}$$

For the first term, we can easily upper bound

$$\begin{aligned} E_{S,z\sim\mathcal{D}^{n+1}}[|E_{u\sim\mathcal{D}}[c(\mathbb{A}_S, u) - c(\mathbb{A}_{S^{i,z}}, u)]|] &\leq E_{S,z\sim\mathcal{D}^{n+1}}[|R(\mathbb{A}_S) - R(h^*)|] \\ &+ E_{S,z\sim\mathcal{D}^{n+1}}[|R(h^*) - R(\mathbb{A}_{S^{i,z}})|] \leq 2\epsilon_{\text{cons}}(n) \quad (4) \end{aligned}$$

from the consistency of $\mathbb{A}_S$. For the second term, it holds

$$\begin{aligned} &E_{S\sim\mathcal{D}^n}[|E_{u,z\sim\mathcal{D}^2}[c(\mathbb{A}_{S^{i,z}}, u) - c(\mathbb{A}_{S^{i,u}}, u) + c(\mathbb{A}_{S^{i,u}}, z) - c(\mathbb{A}_{S^{i,z}}, z)]|] = \\ &nE_{S\sim\mathcal{D}^n}[|E_{u,z\sim\mathcal{D}^2}[R_{S^{i,u}}(\mathbb{A}_{S^{i,z}}) - R_{S^{i,u}}(\mathbb{A}_{S^{i,u}}) + R_{S^{i,z}}(\mathbb{A}_{S^{i,u}}) - R_{S^{i,z}}(\mathbb{A}_{S^{i,z}})]|] \\ &\qquad = 2nE_{S\sim\mathcal{D}^n}[|E_{u,z\sim\mathcal{D}^2}[R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})] - R_{S^{i,u}}(\mathbb{A}_{S^{i,z}})|]. \end{aligned}$$

We will use Proposition 2 to bound the above expression. It holds

$$\begin{aligned} &|E_{S\sim\mathcal{D}^n}[E_{u,z\sim\mathcal{D}^2}[R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})] - R_{S^{i,u}}(\mathbb{A}_{S^{i,z}})]| \\ &\qquad = |E_{S,u,z\sim\mathcal{D}^{n+2}}[c(\mathbb{A}_{S^{i,u}}, u) - c(\mathbb{A}_{S^{i,z}}, u)]|/n. \end{aligned}$$

Meanwhile, it is easy to obtain $E_{S,u\sim\mathcal{D}^{n+1}}[c(\mathbb{A}_{S^{i,u}}, u)] = E_{S,u\sim\mathcal{D}^{n+1}}[R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})]$ and $E_{S,u,z\sim\mathcal{D}^{n+2}}[c(\mathbb{A}_{S^{i,z}}, u)] = E_{S,z\sim\mathcal{D}^{n+1}}[R(\mathbb{A}_{S^{i,z}})]$ from symmetry and i.i.d assumption. This leads to

$$|E_{S\sim\mathcal{D}^n}[E_{u,z\sim\mathcal{D}^2}[R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})] - R_{S^{i,u}}(\mathbb{A}_{S^{i,z}})]| \leq \epsilon_{\text{gen}}/n + 2\epsilon_{\text{cons}}/n.$$

For ERM, $R_{S^{i,u}}(\mathbb{A}_{S^{i,u}}) - R_{S^{i,u}}(\mathbb{A}_{S^{i,z}}) \leq R_{S^{i,u}}(\mathbb{A}_{S^{i,u}}) - R_{S^{i,u}}(\hat{h}_{S^{i,u}})$ and

$$E_{S,u\sim\mathcal{D}^{n+1}}[|R_{S^{i,u}}(\mathbb{A}_{S^{i,u}}) - R_{S^{i,u}}(\hat{h}_{S^{i,u}})|] \leq \epsilon_{\text{erm}}(n).$$

By Proposition 2, we have

$$nE_{S\sim\mathcal{D}^n}[|E_{u,z\sim\mathcal{D}^2}[R_{S^{i,u}}(\mathbb{A}_{S^{i,u}})] - R_{S^{i,u}}(\mathbb{A}_{S^{i,z}})|] \leq \epsilon_{\text{gen}} + 2\epsilon_{\text{cons}} + 2n\epsilon_{\text{erm}}(n),$$

which concludes this lemma by combining with Eq.(4). □

5.3 Proofs of Theorems 5 and 6

The two proofs are relatively similar, and thus we only give the detail proof of Theorem 6. Set $\mathrm{err}(t) = \mathrm{err}^t_S$, $\mathrm{err}'(t) = \mathrm{err}^t_{S^{i,u}}$, $\alpha(t) = \alpha^t_S$, $\alpha'(t) = \alpha^t_{S^{i,u}}$, $h_t(x) = \mathbb{A}_{P^t_S}(x)$ and $h'_t(x) = \mathbb{A}_{P^t_{S^{i,u}}}(x)$ for short in this subsection. The following lemma establishes AdaBoost's empirical stability.

Lemma 5. *For any $u, S \in \mathcal{Z}^{n+1}$, any $i \in [n]$, any $t \in [T]$ and small $\gamma > 0$, if base learner satisfies $\gamma \le err^t_S$, $err^t_{S^{i,u}} \le 1-\gamma$ and $|err^t_S - err^t_{S^{i,u}}| \le \zeta(n)$, then the combined learner $\mathbb{H}_S(x)$ has empirical stability $\beta_1(n) = \zeta(n)T/\sqrt{\gamma(1-\gamma)}$.*

Proof. From [21] we derive

$$R_S(\mathbb{H}_S) = 2^T \prod_{t=1}^T \sqrt{\mathrm{err}(t)(1-\mathrm{err}(t))}$$
$$R_{S^{i,u}}(\mathbb{H}_{S^{i,u}}) = 2^T \prod_{t=1}^T \sqrt{\mathrm{err}'(t)(1-\mathrm{err}'(t))}.$$

Since $\gamma < \mathrm{err}(t) < 1-\gamma$, it is easy to get $\sqrt{\gamma(1-\gamma)} \le \sqrt{\mathrm{err}(t)(1-\mathrm{err}(t))} \le 1/2$, which leads to

$$\begin{aligned}&|\sqrt{\mathrm{err}(t)(1-\mathrm{err}(t))} - \sqrt{\mathrm{err}'(t)(1-\mathrm{err}'(t))}| \\ &\qquad = \frac{|\mathrm{err}(t) - \mathrm{err}'(t)| \times |1 - \mathrm{err}(t) - \mathrm{err}'(t)|}{\sqrt{\mathrm{err}(t)(1-\mathrm{err}(t))} + \sqrt{\mathrm{err}'(t)(1-\mathrm{err}'(t))}} \le \frac{\zeta(n)/2}{\sqrt{\gamma(1-\gamma)}},\end{aligned}$$

and $R_S(\mathbb{H}_S) < 1$. Thus $|R_S(\mathbb{H}_S) - R_{S^{i,u}}(\mathbb{H}_{S^{i,u}})|$ is bounded by

$$\begin{aligned}&2^{T-1}\Big|\prod_{t=1}^{T-1}\sqrt{\mathrm{err}(t)(1-\mathrm{err}(t))} - \prod_{t=1}^{T-1}\sqrt{\mathrm{err}'(t)(1-\mathrm{err}'(t))}\Big| \\ &+ 2^T\Big|\sqrt{\mathrm{err}(T)(1-\mathrm{err}(T))} - \sqrt{\mathrm{err}'(T)(1-\mathrm{err}'(T))}\Big|\prod_{t=1}^{T-1}\sqrt{\mathrm{err}(t)(1-\mathrm{err}(t))} \\ &\le 2^{T-1}\Big|\prod_{t=1}^{T-1}\sqrt{\mathrm{err}(t)(1-\mathrm{err}(t))} - \prod_{t=1}^{T-1}\sqrt{\mathrm{err}'(t)(1-\mathrm{err}'(t))}\Big| + \frac{\zeta(n)}{\sqrt{\gamma(1-\gamma)}}\end{aligned}$$

which leads to $|R_S(\mathbb{H}_S) - R_{S^{i,u}}(\mathbb{H}_{S^{i,u}})| \le \zeta(n)T/\sqrt{\gamma(1-\gamma)}$ as desired. □

Lemma 6. *If $h_t(x), h'_t(x)$ are two binary learners with cost function $c(h,z) = I[h(x) \ne y]$, then we have $E_{z\sim\mathcal{D}}[|h_t(x) - h'_t(x)|] = E_{z\sim\mathcal{D}}[|c(h_t,z) - c(h'_t,z)|]$.*

This lemma holds from the fact $|I[h_t(x) \ne y] - I[h'_t(x) \ne y]| = |h_t(x) - h'_t(x)|$. The following lemma establishes the validation stability of AdaBoost.

Lemma 7. *For any $u, S \in \mathcal{Z}^{n+1}$, any $i \in [n]$, any $t \in [T]$ and $0 < \gamma < 1/2$, if the base learner satisfies $\gamma \le err^t_S$, $err^t_{S^{i,u}} \le 1-\gamma$, $|err^t_S - err^t_{S^{i,u}}| \le \zeta(n)$ and $E_{u\sim\mathcal{D}}[|c(\mathbb{A}_S, u) - c(\mathbb{A}_{S^{i,u}}, u)|] \le \eta(n)$, then $\mathbb{H}_S(x)$ has validation stability $\beta_2(n) = TB\big(\frac{\eta(n)}{2}\ln\frac{1-\gamma}{\gamma} + \frac{\zeta(n)}{2\gamma(1-\gamma)}\big)$.*

Proof. We first set $u = (x, y)$. From mean value theorem and $\gamma \leq \mathrm{err}(t), \mathrm{err}'(t) \leq 1 - \gamma$, we have $|\alpha(t) - \alpha'(t)| \leq \zeta(n)/(2\gamma(1-\gamma))$. It follows from Lemma 6 that

$$E_{u\sim\mathcal{D}}[|\alpha(t)h_t(x) - \alpha'(t)h'_t(x)|] \leq E_{u\sim\mathcal{D}}[|\alpha(t)||h_t(x) - h'_t(x)|] + E_{u\sim\mathcal{D}}[|h'_t(x)||\alpha(t) - \alpha'(t)|] \leq \frac{\eta(n)}{2}\ln\frac{1-\gamma}{\gamma} + \frac{\zeta(n)}{2\gamma(1-\gamma)}, \quad (5)$$

Using mean value theorem again, we obtain

$$|\exp(-y\alpha(t)h_t(x)) - \exp(-y\alpha'(t)h'_t(x))| \leq \sqrt{(1-\gamma)/\gamma}\,|\alpha(t)h_t(x) - \alpha'(t)h'_t(x)|\,.$$

Combining with Eq.(5) gives

$$|R(\mathbb{H}_S) - E_{u\sim\mathcal{D}}[R(\mathbb{H}_{S^{i,u}})]| \leq E_{u\sim\mathcal{D}}\left[|\exp(-y\alpha'(T)h'_T(x)) \times \Gamma|\right] + E_{u\sim\mathcal{D}}\Big[\Big|\exp(-y\sum_{t=1}^{T-1}\alpha(t)h_t(x))\big(\exp(-y\alpha(T)h_T(x)) - \exp(-y\alpha'(T)h'_T(x))\big)\Big|\Big] \leq (B\eta(n)/2)\ln((1-\gamma)/\gamma) + B\zeta(n)/(2\gamma(1-\gamma)) + E_{u\sim\mathcal{D}}[|\Gamma|]\sqrt{(1-\gamma)/\gamma},$$

where $\Gamma = \exp\left(-y\sum_{t=1}^{T-1}\alpha(t)h_t(x)\right) - \exp\left(-y\sum_{t=1}^{T-1}\alpha'(t)h'_t(x)\right)$. This completes the proof by straight evaluation. □

By Lemmas 5 and 7 we get that AdaBoost has empirical stability and validation stability, respectively. Thus, by using Theorem 4, we get Theorem 6.

Acknowledgements

We thank the anonymous reviewers for helpful comments. The work was supported by the National Science Foundation of China (60635020, 60721002), the National Fundamental Research Program of China (2010CB327903) and the Jiangsu Science Foundation (BK2008018).

References

1. N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnablity. *J. ACM*, 44(4):615–631, 1997.
2. J. P. Bickel, Y. Ritov, and A. Zakai. Some theory for generalized boosting algorithms. *J. Mach. Learn. Res.*, 7:705–732, 2006.
3. O. Bousquet and A. Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002.
4. L. Breiman. Prediction games and arcing classifiers. *Neural Comput.*, 11(7):1493–1517, 1999.
5. L. P. Devroye and T. J. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Trans. Inform. Theory*, 25:601–604, 1979.

6. Y. Freund and R. E. Schapire. Game theory, on-line prediction and boosting. In *Proc. of 9th COLT*, pages 325–332, Desenzano sul Garda, Italy, 1996.
7. Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.
8. Y. Freund and R. E. Schapire. Response to "Evidence contrary to the statistical view of Boosting". *J. Mach. Learn. Res.*, 9:171–174, 2008.
9. J. Friedman, T. Hastie, and R. Tibshirani. Addtive logistic regression: A statistical view of boosting (with discussion). *Ann. Statist.*, 28:337–407, 2000.
10. M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Comput.*, 11:1427–1453, 1999.
11. S. Kutin and P. Niyogi. The interaction of stability and weakness in Adaboost. Technical Report 30, Department of Computer Science, University of Chicago, Chicago, IL, 2001.
12. S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proc. of 18th UAI*, pages 275–282, Edmonton, Canada, 2002.
13. L. Mason, J. Baxter, P. L. Bartlett, and M. R. Freman. Boosting algorithms as gradient descent. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in NIPS 12*, pages 512–518. MIT Press, Cambridge, MA, 1999.
14. C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, Cambridge, UK, 1989.
15. D. Mease and A. Wyner. Evidence contrary to the statistical view of boosting with discussion. *J. Mach. Learn. Res.*, 9:131–201, 2008.
16. S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimzation. *Adv. Comput. Math.*, 25:161–193, 2006.
17. A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Anal. Appl.*, 4:397–417, 2005.
18. G. Rätsch, T. Onoda, and K. R. Müller. Soft margins for Adaboost. *Mach. Learn.*, 42:287–320, 2001.
19. L. Reyzin and R. E. Schapire. How boosting the margin can also boost classifier complexity. In *Proc. of 23rd ICML*, pages 753–760, Pittsburgh, PA, 2006.
20. C. Rudin, R. E. Schapire, and I. Daubechies. Precise statements of convergence for adaboost and arc-gv. In *Proc. of AMS-IMS-SIAM Joint Summer Research Conference: Machine learning, Statistics and Discovery*, pages 131–145, Snowbird, Utah, 2007.
21. R. Schapire, Y. Freund, P. L. Bartlett, and W. Lee. Boosting the margin: A new explanation for the effectives of voting methods. *Ann. Statist.*, 26:1651–1686, 1998.
22. S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability and stability in the general learning setting. In *Proc. of 22nd COLT*, Montreal, Canada, 2009.
23. S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimation. In *Proc. of 22nd COLT*, Montreal, Canada, 2009.
24. V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.
25. L. Wang, M. Sugiyama, C. Yang, Z.-H. Zhou, and J. Feng. On the margin explanation of boosting algorithm. In *Proc. of 21st COLT*, pages 479–490, Helsinki, Finland, 2008.