Fantastic Bounds for Adaptive Boosting and Where to Find Them

Final Project, STAT 535

Songqian Chen

Start: Nov 25, 2016

Contents

Preface

That was a long and short journey.

I learnt adaptive boosting for the first time three years ago, when I was still a junior undergraduate student. At that time this algorithm is so hard to learn for me because there are many strange rules in this algorithm, like the way it updates the weights for each samples, and, especially, its choice of the weights for each classifiers $\alpha_t = \frac{1}{2}\log(\frac{1-\epsilon_t}{\epsilon_t})$, $t = 1, ..., T$, where ϵ_t is the error rate for each classifiers h_t among the test sample set. A good explanation our professor provided at that time is that this choice of α_t can let the classifiers with low error rate ϵ_t have relative larger weights, while those with high error rate would have smaller weights. However, this explanation just explain the $\frac{1-\epsilon_t}{\epsilon_t}$ part, but tells nothing about the logarithm. Therefore, a question left for me at that time:

- Why do we choose $\alpha_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$?

Two years passed, as I got into graduate study, I took another machine learning course and learnt adaptive boosting algorithm again. At that course the professor began to talk something about overfitting, and stated generally in practice, adaptive boosting will not overfit. He carried out the concept of bounds for adaptive boosting at that time, but did not talk much about it. Thus, another question left in my mind:

- Why in general will adaptive boosting not be overfitting? And what does the "bounds" mean?

All these questions for adaptive boosting were connected in this quarter. It was not until I learnt the the concept of probably approximately correct (PAC) learning model, VC dimensions, Rademacher complexity, Saucer's lamma and generalization error bound that I finally realized there are such many strict and beautiful mathematics behind the machine learning algorithms. They are powerful and fantastic, bring me to a new perspective to view all the machine learning knowledge I have learnt, and giving me more rigorous explanations about bias and variance for machine learning algorithm, the terms everyone is talking about but few knows what exactly they mean.

That is the reason I love mathematics! It is the logistic that makes our real world clear and reasonable!

Therefore I decided to use these new-learnt tools to solve the questions I kept in mind for a long time for adaptive boosting, and I began to read. At the beginning I read the slides from other universities, and almost all of them would just state and explain the generalization error bound without proving it. Fortunately, some of them led me to the original paper written by Freund and Schapire [3], and thus led me to the further discussion about adaptive boosting [5][6][7] (by the way, some literatures led me to the concept of "data compression" [2][4], which I found are not for our case after trying to understand them for one or two days). They built my basic understanding of theories behind adaptive boosting, and the roof were finally built by a book also written by Freund and Schapire [9] in 2012. I really appreciate the help from this book:

Boosting: Foundations and Algorithms
(Schapire, R. E., & Freund, Y. (2012) [9]. *MIT press*)

That was a fantastic book that make everything for boosting clear, especially for chapter 2 to chapter 5, although some parts I think are not rigorous enough.

Above is the long explanation for the "long" journey in the first sentence, similarly there is also a short explanation for the "short" journey: since the quarter is short, there was just around 10 days left to finish this project. Therefore, some parts are not as good written as I expected. However, I am still proud of them. Finally, extra credits are given to the book and the movie that name my project:

Fantastic Beast and Where to Find Them
(Rowling, J. K. (2009) [8] and Scamander, N., & Rowling, J. K. (2015) [12])

Songqian Chen
December, 2016
Two days after the first snow in Seattle

Chapter 1 Adaptive Boosting and Its Basic Bounds

Section 1.1 What is Adaptive Boosting

Originally raised by Yoav Freund and Robert Schapire [3] in 1995, Adaptive Boosting is a fantastic and powerful idea that ensemble a bunch of weak classifiers to form a strong classifier. Specifically, suppose we have a set of sample points $(x_1, y_1), ..., (x_n, y_n)$, with $y_i \in \{-1, 1\}$ for all i, and initially we would give equal weights to each of the sample point, $\mathcal{D}_0(i) = \frac{1}{n}$, $i = 1, ..., n$. Then, with a weak learning algorithm $\mathcal{A}$, we would run this algorithm on our sample for T rounds, and on each run t we would redistribute the weights $\mathcal{D}_t(i)$, $i = 1, ..., n$, so that the weights would pay more attention to the samples we classified wrong on the last round.

Denote h_t as the classifier produced by the weak learning algorithm $\mathcal{A}$ on the weighted samples on t round, the error rate ϵ_t on the t round would be calculated as $\epsilon_t = \sum_{i=1}^n \mathcal{D}_t(i)\mathbb{1}(h_t(x_i) \neq y_i)$. After T rounds, now we would have a bunch of weak classifiers h_t, $t = 1, ..., T$ and the method to ensemble them is by using a similar idea of weighted average:

$$F(x) = \sum_{t=1}^{T} \alpha_t h_t(x),$$

where α_t are the weights assigned to h_t and are supposed to be large when h_t has a low error rate ϵ_t on our samples while being small when h_t has a high error rate. That is, for our ensemble classifier, we would give more "voting power" to the those h_t who classify our sample better. Finally, we could construct our final classifier as

$$H(x) = \operatorname{sign}\big(F(x)\big).$$

That is just a general idea, since how we redistribute the weights $\mathcal{D}_t(i)$ for each samples and how the "voting" weights α_t are correlated with the error rate ϵ_t are still unclear. Therefore, to be more detailed, now we would have the adaptive boosting algorithm stated as follows:

Algorithm 1 Adaptive Boosting Algorithm

```
procedure AdaBoost
    Initialize $\mathcal{D}_1(i) \leftarrow \frac{1}{n}$, $i = 1, ..., n$
    for round $t = 1, ..., T$ do
        - Draw a sufficiently large samples from $\{(x_1, y_1), ..., (x_n, y_n)\}$ based on weights $\mathcal{D}_t(i)$
        - Train algorithm $\mathcal{A}$ on the samples to generate classifier $h_t$
        - Calculate weighted error rate $\epsilon_t$:
              $$\epsilon_t = \sum_{i=1}^{n} \mathcal{D}_t(i)\mathbb{1}(h_t(x_i) \neq y_i)$$
        - Calculate $\alpha_t$:
              $$\alpha_t = \frac{1}{2}\log\Big(\frac{1-\epsilon_t}{\epsilon_t}\Big)$$
        - Update the weights $\mathcal{D}_{t+1}(i)$, $i = 1, ..., n$:
              $$\mathcal{D}_{t+1}(i) = \frac{\mathcal{D}_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$
          where $Z_t = \sum_{i=1}^n \mathcal{D}_t(i)\exp(-\alpha_t y_i h_t(x_i))$ is the normalization factor.
    return final classifier
          $$H(x) = \operatorname{sign}\big(\sum_{t=1}^{T} \alpha_t h_t(x)\big)$$
```

Section 1.2 Traning Errors Bound and Generalization Error Bound

Now we begin to find the fantastic bounds for this algorithm! To begin with, we would first derive a bound for the training error,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i \neq H(x_i)).$$

The bound is proposed by Yoav Freund and Robert Schapire [3] with in their original paper for adaptive boosting, and we would state it as follows:

Theorem 1. *Let H be the ensemble classifier generate by adaptive boosting after T rounds on samples $\{(x_1, y_1), ..., (x_n, y_n)\}$, and also define γ_t as*

$$\gamma_t = \frac{1}{2} - \epsilon_t,$$

where ϵ_t is the weighted error rate on each round t. Then we would have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i \neq H(x_i)) \leq e^{-2\sum_{t=1}^{T}\gamma_t^2}.$$

Therefore if $\gamma_t \geq \gamma$ for all t, then

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i \neq H(x_i)) \leq e^{-2T\gamma^2}.$$

Proof. The proof of Theorem 1 is threefold, and most of them are just simple algebra. However, the proof answers an important and mysterious question for adaptive boosting algorithm:

- Why do we choose $\alpha_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$?

Now let's start to unravel this question by the proof.

Step I Firstly, we would prove that $\mathcal{D}_{T+1}(i) = \frac{1}{n}\left(\frac{\exp(-y_iF(x_i))}{\prod_{t=1}^{T}Z_t}\right)$:

$$\begin{aligned}
\mathcal{D}_{T+1}(i) =& \frac{\exp(-y_i\alpha_T h_T(x_i))}{Z_T} \times \mathcal{D}_T \\
=& \frac{\exp(-y_i\alpha_T h_T(x_i))}{Z_T} \times \frac{\exp(-y_i\alpha_{T-1} h_{T-1}(x_i))}{Z_{T-1}} \times \mathcal{D}_{T-1} \\
& \cdots \\
=& \frac{\exp[-y_i\sum_{t=1}^{T}\alpha_t h_t(x_i)]}{\prod_{t=1}^{T}Z_t} \times \mathcal{D}_0 \\
=& \frac{1}{n}\frac{\exp[-y_iF(x_i)]}{\prod_{t=1}^{T}Z_t}.
\end{aligned}$$

Step II In Step II we would prove $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i \neq H(x_i)) \leq \prod_{t=1}^{T}Z_t$. This step need a small trick, that is, for $\forall x \in \mathbb{R}$ and $y \in \{-1, 1\}$,

$$\mathbb{1}\big(y \neq H(x)\big) \leq \exp\big(-yF(x)\big),$$

where $H(x) = \text{sign}\big(F(x)\big)$. The reason is that when $yF(x) > 0$, then y and $F(x)$ must have the same sign, thus $\mathbb{1}\big(y \neq H(x)\big) = 0 \leq \exp\big(-yF(x)\big)$, while when $yF(x) < 0$, then y and $F(x)$ mush have different signs thus $\mathbb{1}\big(y \neq H(x)\big) = 1 \leq \exp\big(-yF(x)\big)$. Therefore, we have the above trick proved. Thus, we could have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i \neq H(x_i)) \leq & \frac{1}{n}\sum_{i=1}^{n}\exp\big(-y_iF(x_i)\big)\\ =&\frac{1}{n}\sum_{i=1}^{n}\Big[n\mathcal{D}_{T+1}(i)\prod_{t=1}^{T}Z_t\Big] \quad \text{(according to Step I)}\\ =&\prod_{t=1}^{T}Z_t\sum_{i=1}^{n}\mathcal{D}_{T+1}(i)\\ =&\prod_{t=1}^{T}Z_t.\end{aligned}$$

Step III Step III is the most important step to answer our question about α_t raised at the beginning of our proof. We would start with departing each Z_t into two part: a sum over all points labeled correctly and a sum over all points labeled incorrectly:

$$Z_t = \sum_{i=1}^{n}\mathcal{D}_t(i)e^{-\alpha_t y_i h_t(x_i)} = \sum_{i:\ y_i=h_t(x_i)}\mathcal{D}_t(i)e^{-\alpha_t} + \sum_{i:\ y_i\neq h_t(x_i)}\mathcal{D}_t(i)e^{\alpha_t}.$$

Notice that since $\epsilon_t = \sum_{i=1}^{n}\mathcal{D}_t(i)\mathbb{1}(h_t(x_i) \neq y_i)$, therefore $\sum_{i:\ y_i=h_t(x_i)}\mathcal{D}_t(i) = 1-\epsilon_t$ and $\sum_{i:\ y_i\neq h_t(x_i)}\mathcal{D}_t(i) = \epsilon_t$. Thus, we would further have

$$Z_t = (1-\epsilon_t)e^{-\alpha_t} + \epsilon_t e^{\alpha_t}.$$

That is the key point! In order to minimize Z_t so as to minimize the training error bound in Step II, we thus would take

$$\alpha_t = \frac{1}{2}\log\Big(\frac{1-\epsilon_t}{\epsilon_t}\Big) \quad , t = 1, ..., T.$$

That is exactly the reason why we would choose α_t to be such a strange form! Therefore, by choosing these specific α_t, we now have

$$\begin{aligned}\prod_{t=1}^{T}Z_t =& \prod_{t=1}^{T}2\sqrt{\epsilon_t(1-\epsilon_t)}\\ =&\prod_{t=1}^{T}2\sqrt{\Big(\frac{1}{2}-\gamma_t\Big)\Big(\frac{1}{2}+\gamma_t\Big)} \quad \text{(since } \gamma_t = \frac{1}{2}-\epsilon_t)\\ =&\prod_{t=1}^{T}\sqrt{1-4\gamma_t^2}\\ \leq& e^{-2\sum_{t=1}^{T}\gamma_t^2} \quad \text{(since } (1+x) \leq e^x).\end{aligned}$$

Therefore, we finally have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i \neq H(x_i)) \leq \prod_{t=1}^{T}Z_t \leq e^{-2\sum_{t=1}^{T}\gamma_t^2},$$

which complete the proof of Theorem 1.

□

From Theorem 1 we know that the training error for adaptive boosting will decrease to 0 exponentially fast, however, the thing that we really care about is the performance of adaptive boosting on the testing set, which is measured by testing error. Could we get a bound for the testing error just based on our already know information on the training set? The answer is YES! Recall all possible classifier produced from the adaptive boosting is of the form

$$H(x) = \text{sign}(F(x)),$$

where $F(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$ with $h_1, ..., h_T \in \mathcal{H}$, the hypothesis space of base classifiers, and $\alpha_1, ..., \alpha_T > 0$. Then, by denoting $\mathcal{C}_T$ be the function space of all possible $H(x)$ that can be generated by adaptive boosting, we could state the generalization error bound theorem for adaptive boosting as follows:

Theorem 2. *For $H \in \mathcal{C}_T$ that is generated after running adaptive boosting for T rounds on n random samples, with base classifier $h_1, ..., h_T$ from hypothesis space $\mathcal{H}$ that has finite VC-dimension $\nu(\mathcal{H})$, we would have with probability at least $1 - \delta$ with respect to random set of samples,*

$$P(Y \neq H(X)) \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Y_i \neq H(X_i)) + \sqrt{\frac{32\Big[T\big(\log(\frac{en}{T}) + \nu(\mathcal{H}) \log(\frac{en}{\nu(\mathcal{H})})\big) + \log(\frac{8}{\delta})\Big]}{n}}.$$

The proof is a simple application of generalized Glivenko-Cantelli theorem, that is Lemma 2 in the next chapter, with a trick of using saucer lemma to derive a bound for the growth function of $\mathcal{C}_T$. Since we will prove a stronger statement of the bound in the next chapter, we would not prove this theorem here (the actual reason is that there was not sufficient time to finish the proof). Generally, we would simply the description of Theorem 2 as with probability at least $1 - \delta$, we would have

$$P(Y \neq H(X)) \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Y_i \neq H(X_i)) + O\Big(\sqrt{\frac{T\nu(\mathcal{H})}{n}}\Big).$$

That is, we would have the testing error increase if we run a larger number of T rounds, which is usually called overfitting. However, in practice, the testing error would continue decreasing with T increase even as the training error has already comes to 0, as shown in the following figures:

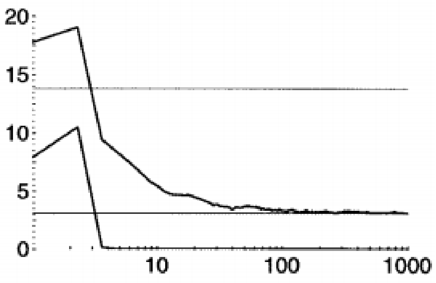

Figure 1: Testing and training error versus number of classifiers. Adapted from "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods," by Schapire, R. E., Freund, Y., Bartlett, P.,& Lee, W. S. (1998) [5]. *Annals of statistics*, 1651-1686.

Therefore, the generalization error bound provided by Theorem 2 is not enough for further analysis. We need another perspective to view this question and thus derive improved bounds. That is the reason that Schapire and Freund [5] risked the concept of margin in 1998, which we would discuss in the following chapter.

Chapter 2 Improved Bounds for Illustrating Non-overfitting

Section 2.1 Margin

The concept of margin is actually raised from the idea to measure how confident does our final classifier $H(x)$ is in classifying the data. Recall that our final classifier is $H(x) = \text{sign}\big(F(x)\big)$ where $F(x) = \sum_{i=1}^{T} \alpha_t h_t(x)$. Now we first normalize $F(x)$ as

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

so that $f(x) \in [-1, 1]$ since $h_t(x) \in [-1, 1]$ for $t = 1, ..., T$. Therefore, we then could define *margin* as

$$margin = yf(x) \in [-1, 1].$$

When most of h_t could correctly classify the sample point x, we would have $yf(x)$ close to 1, while when most of h_t classify the sample point x wrong, we would have $yf(x)$ close to -1. In between, when some h_t classify x correctly while some does not, then the value of $yf(x)$ would be in $[-1, 1]$ and the more neural the "vote", the closer $yf(x)$ is to 0. Therefore, the margin is actually measuring how confident the final classifier $H(x) = \text{sign}(F(x)) = \text{sign}(f(x))$ is in classifying the data, as the following figure illustrate:

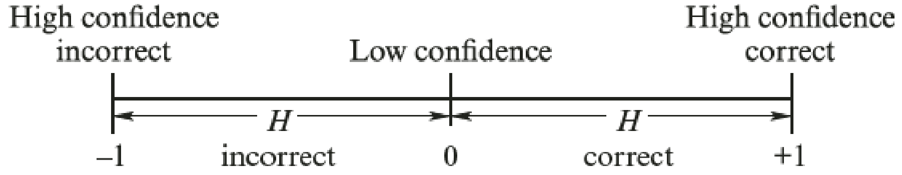

Figure 2: Explanation of margin. Adapted from "Boosting: Foundations and algorithms" by Schapire, R. E., & Freund, Y. (2012) [9]. *MIT press.*

Observe that the margin $yf(x)$ is a function of the sample (x, y), and since we have a generative distribution for the sample (x, y), we would also have a distribution of all values of $yf(x)$ based on the generative distribution of (x, y), which we might call *generative distribution of margin.* However, since we would never know what is exactly the generative distribution of our sample (x, y), we also could not know that of the margin. However, using the concept of empirical cumulative distribution, for any set of samples, we could calculate the empirical cumulative distribution of the margin. That is, for $\forall\theta \in [-1, 1]$, we can calculate the fraction of the n training samples that has a margin equal or less than θ:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i f(x_i) \leq \theta).$$

Then, the empirical distribution of margin could be drawed like Figure 3. An important observation is that as the number of rounds the adaptive boosting run on a specific training data increase, the fraction of the samples with small margin is decrease, that is, more and more sample would have a larger margin. This is a key property of adaptive boosting, and we would illustrate it later.

Now we come to the key step to connect the concept of generalization and training error with margin. Recall that when our final classifier $H(x) = \text{sign}(f(x)) \in \{-1, 1\}$ assign an incorrect label of x, we would have $yf(x) \leq 0$, therefore, we could write the generalization error and training error as

$$P(Y \neq H(X)) = P(Yf(X) \leq 0) \quad \text{and} \quad \frac{1}{n}\mathbb{1}(Y_i \neq H(X_i)) = \frac{1}{n}\mathbb{1}(Y_i f(X_i) \leq 0),$$

which lead us to view the problem of how to bound the generalization error from a new perspective of margin!

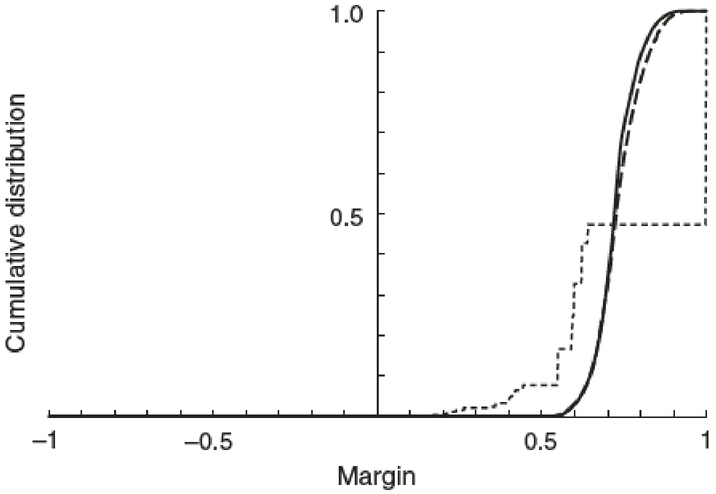

Figure 3: Empirical distribution of margin after 5, 100, 1000 rounds of boosting, indicated by short-dashed, long-dashed and solid curves respectively. Adapted from "Boosting: Foundations and algorithms" by Schapire, R. E., & Freund, Y. (2012) [9]. *MIT press.*

Therefore, before the further proof, we could now state the general idea of getting a bound for generalization error from the view of margin to illustrate non-overfitting of adaptive boosting as follows:

- Firstly, for $\theta \in [-1, 1]$, derive a bound for generalization error depends only on the fraction of random samples with margin smaller than θ, $\frac{1}{n}\mathbb{1}(Y_i f(X_i) \leq \theta)$, and thus and bound is independent of the number of rounds T (Section 2.2).
- Then, prove that for any set of samples $\{(x_1, y_1), ..., (x_n, y_n)\}$ and $\theta \in [-1, 1]$ satisfying some conditions, $\frac{1}{n}\mathbb{1}(y_i f(x_i) \leq \theta)$ would decrease exponentially fast to 0 as T increase (Section 2.3).

Section 2.2 Improved Bound Based on Generalized Glivenko-Cantelli Theorem

We would first state the theorem:

Theorem 3. *Suppose the base-classifier space $\mathcal{H}$ has VC-dimension $\nu(\mathcal{H})$. Then with probability at least $1 - \delta$, universally over all weighted average function $f \in co(\mathcal{H})$ and $\forall \theta > 0$, we have*

$$P\big(Yf(X) \leq 0\big) \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_i f(X_i) \leq \theta\big) + O\left(\frac{1}{\sqrt{n}}\sqrt{\frac{\nu(\mathcal{H})\log(\frac{n}{\nu(\mathcal{H})})\log(\frac{n\theta}{\nu(\mathcal{H})})}{\theta^2} + \log(\frac{1}{\delta})}\right).$$

The key idea for proving this theorem is that for $\forall f = \sum_{t=1}^{T}\alpha_t h_t \in co(\mathcal{H})$, we want to approximate f with a $\tilde{f}_m$ that only consist of m random base classifiers and m is independent of the number of rounds T. By this way we could derive a bound based on $\tilde{f}_m$ which would be also independent of T. Finally, by connect the bound for $\tilde{f}_m$ to that for f, and then choosing the optimal m, we would finish the proof.

Proof. The proof is threefold: in the first two steps, we would construct the $\tilde{f}_m$ and prove that $\tilde{f}_m$ is a good approximation for f, therefore we could prove that the probability of event $\{Y\tilde{f}_m(X) \leq \frac{\theta}{2}\}$ approximate that of event $\{Yf(X) \leq 0\}$ in terms of the generative distribution of (X, Y) and the sample D_n respectively, where the approximation error is a bound related to the value of θ. Then in the third step we would gain a bound based on $\tilde{f}_m$ for the generalized error so that the bound will be independent of T. Finally we would connect all steps together and thus derive a generalization bound for f and thus finish the proof. Now, let's start our proof.

Step I. To begin with, we would firstly construct an independent set of random base classifiers $\tilde{h}_j$, $j = 1, ..., m$, in a way such that the probability of $\tilde{h}_j$ equal to h_t, $t = 1, ..., T$, is exactly α_t. That is,

$$\begin{cases} P(\tilde{h}_j = h_1) & = \alpha_1, \\ P(\tilde{h}_j = h_2) & = \alpha_2, \\ & \vdots \\ P(\tilde{h}_j = h_T) & = \alpha_T \end{cases}, \quad j = 1, ..., m.$$

Recall that $f = \sum_{t=1}^{T} \alpha_t h_t$ with $\sum_{t=1}^{T} \alpha_t = 1$, therefore, like the concept of random variables, we could find out that $\tilde{h}_j$ is actually a "random function" that obey a multivariate distribution with all possible values are also functions $h_1, ..., h_T$. Therefore, $E[\tilde{h}_j] = \sum_{t=1}^{T} \alpha_t h_t = f$, and we know that a good approximation of the expectation $f = E[\tilde{h}_j]$, $j = 1, ..., m$, is actually their mean. That is,

$$\tilde{f}_m = \frac{1}{m} \sum_{j=1}^{m} \tilde{h}_j.$$

Thus $\tilde{f}_m$ will be an approximation, or estimate, for f with only m random base classifiers $\tilde{h}_j$, $j = 1, ..., m$. With this $\tilde{f}_m$, we need the following lemma before further analysis:

Lemma 1. *For $\forall x \in \mathcal{X}$, $\theta > 0$ and $m \geq 1$, we have*

$$P_{\tilde{f}_m}\Big(\Big|\tilde{f}_m(x) - f(x)\Big| \geq \frac{\theta}{2}\Big) \leq 2e^{-m\theta^2/8},$$

where $P_{\tilde{f}_m}$ means we take probability with respect to the randomness of $\tilde{f}_m$, which is the same as taking the randomness of $\tilde{h}_j$.

Proof. With x fixed, we know $h_t(x) \in \{-1, 1\}$, $t = 1, ..., T$, therefore we have $\tilde{h}_j(x) \in \{-1, 1\}$, $j = 1, ..., m$. Thus according to Hoeffding inequality, we would have

$$P_{\tilde{f}_m}\Big(\Big|\tilde{f}_m(x) - f(x)\Big| \geq \frac{\theta}{2}\Big) = P_{\tilde{f}_m}\Big(\Big|\frac{1}{m}\sum_{j=1}^{m} \tilde{h}_j(x) - \frac{1}{m}\sum_{j=1}^{m} E[\tilde{h}_j(x)]\Big| \geq \frac{\theta}{2}\Big) \leq 2e^{-m\theta^2/8}.$$

Thus we finish the proof of Lemma 1. □

Lemma 1 says that $\tilde{f}_m$ is a good approximation of f in term of any fixed $x \in \mathcal{X}$ respectively, therefore we could further examine that $P\big(Y\tilde{f}_m(X) \leq \frac{\theta}{2}\big)$ is an approximation of $P\big(Yf(X) \leq 0\big)$ with the approximation error related to θ for $\forall\theta > 0$:

$$\begin{aligned} P\Big(Yf(X) \leq 0\Big) =& P\Big(Yf(X) \leq 0, Y\tilde{f}_m(X) \leq \frac{\theta}{2}\Big) + P\Big(Yf(X) \leq 0, Y\tilde{f}_m(X) > \frac{\theta}{2}\Big) \\ \leq& P\Big(Y\tilde{f}_m(X) \leq \frac{\theta}{2}\Big) + P\Big(Yf(X) \leq 0, Y\tilde{f}_m(X) > \frac{\theta}{2}\Big) \\ \leq& P\Big(Y\tilde{f}_m(X) \leq \frac{\theta}{2}\Big) + P\Big(\Big|Yf(X) - Y\tilde{f}_m(X)\Big| > \frac{\theta}{2}\Big) \\ =& P\Big(Y\tilde{f}_m(X) \leq \frac{\theta}{2}\Big) + P\Big(\Big|f(X) - \tilde{f}_m(X)\Big| > \frac{\theta}{2}\Big) \quad \text{(since } Y \in \{-1, 1\}\text{)} \\ =& P\Big(Y\tilde{f}_m(X) \leq \frac{\theta}{2}\Big) + E_X\Big[E_{\tilde{f}_m}\Big[\mathbb{1}\Big(\big|f(X) - \tilde{f}_m(X)\big| > \frac{\theta}{2}\Big)\big|X\Big]\Big] \\ =& P\Big(Y\tilde{f}_m(X) \leq \frac{\theta}{2}\Big) + E_X\Big[P_{\tilde{f}_m}\Big(\big|f(X) - \tilde{f}_m(X)\big| > \frac{\theta}{2}\big|X\Big)\Big] \\ \leq& P\Big(Y\tilde{f}_m(X) \leq \frac{\theta}{2}\Big) + E_X\big[2e^{-m\theta^2/8}\big] \quad \text{(according to Lemma 1)} \\ =& P\Big(Y\tilde{f}_m(X) \leq \frac{\theta}{2}\Big) + 2e^{-m\theta^2/8}. \end{aligned}$$

After illustrating Step I, now we could state our whole line of proof more clearly:

$$P\Big(Yf(X) \le \underbrace{0\Big) \lesssim P\Big(Y}_{\textbf{Step I.}} \tilde{f}_m(X) \le \underbrace{\frac{\theta}{2}\Big) \lesssim \frac{1}{n}\sum_{i=1}^{n}}_{\textbf{Step III.}} P_{\tilde{f}_m}\Big(Y_i\tilde{f}_m(X_i) \le \underbrace{\frac{\theta}{2}\Big) \lesssim \frac{1}{n}\sum_{i=1}^{n}}_{\textbf{Step II.}} \mathbb{1}\Big(Y_if(X_i) \le \theta\Big),$$

where $\lesssim$ means that with a large probability (Step III) or almost surely (Step I&II), the term on the left is smaller than that on the right plus some small term. By finishing proving this chain, we could find a bound for $P\Big(Yf(X) \le 0\Big)$ based on $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\Big(Y_i\tilde{f}_m(X_i) \le \theta\Big)$ and another term that is independent of T as Theorem 3 states.

Step II. As stated above, we now want to prove $\frac{1}{n}\sum_{i=1}^{n}P_{\tilde{f}_m}\Big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}\Big)$ is smaller than $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\Big(Y_if(X_i) \le \theta\Big)$ plus a small term, where it should be noticed that $P_{\tilde{f}_m}(\cdot)$ only consider the randomness of $\tilde{f}_m$, therefore $P_{\tilde{f}_m}\Big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}\Big)$, $i = 1, ...n$, are still a function of random samples X_i and Y_i. The proof is quite similar to that in Step I, although some parts need to be modified since we are now dealing with the identity function $\mathbb{1}(\cdot)$. To start with, for $\forall i = 1, ..., n$, we would have

$$\begin{aligned}\mathbb{1}\Big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}\Big) \le &\mathbb{1}\Big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}, Y_if(X_i) \le \theta\Big) + \mathbb{1}\Big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}, Y_if(X_i) > \theta\Big) \\ \le &\mathbb{1}\Big(Y_if(X_i) \le \theta\Big) + \mathbb{1}\Big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}, Y_if(X_i) > \theta\Big) \\ \le &\mathbb{1}\Big(Y_if(X_i) \le \theta\Big) + \mathbb{1}\Big(\Big|Y_i\tilde{f}_m(X_i) - Y_if(X_i)\Big| > \frac{\theta}{2}\Big) \\ = &\mathbb{1}\Big(Y_if(X_i) \le \theta\Big) + \mathbb{1}\Big(\Big|\tilde{f}_m(X_i) - f(X_i)\Big| > \frac{\theta}{2}\Big). \quad \text{(since } Y_i \in \{-1, 1\}\text{)}\end{aligned}$$

Take the expectation of both side on with respect to $\tilde{f}_m$, thus we would have

$$\begin{aligned}P_{\tilde{f}_m}\Big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}\Big) \le &\mathbb{1}\Big(Y_if(X_i) \le \theta\Big) + P_{\tilde{f}_m}\Big(\Big|\tilde{f}_m(X_i) - f(X_i)\Big| > \frac{\theta}{2}|X_i\Big) \\ \le &\mathbb{1}\Big(Y_if(X_i) \le \theta\Big) + 2e^{-m\theta^2/8},\end{aligned}$$

where it need to emphasized that $P_{\tilde{f}_m}\Big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}\Big)$ is still a function of random variable X_i and Y_i.

Therefore, taking the average over n sample, we thus have

$$\frac{1}{n}\sum_{i=1}^{n}P_{\tilde{f}_m}\Big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}\Big) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\Big(Y_if(X_i) \le \theta\Big) + 2e^{-m\theta^2/8}.$$

Step III. Before we start connecting Step I & II with Step III, we would first introduce the following important lemma mentioned by Devroye, Gyorfi, Lugosi [10] (Theorem 12.5) and was originally proposed by Vapnik and Chervonenkis [1] in 1971:

Lemma 2 (Generalized Glivenko-Cantelli Theorem)**.** *Let Ψ be a class of bounded functions satisfying $0 \le \psi(x) \le M$ for all $x \in \mathbb{R}^d$. Define the collection of sets:*

$$\mathcal{A} = \{A_{\psi,t} : \ \psi \in \Psi, \ t \in [0, M]\},$$

where for every $\psi \in \Psi$ and $t \in [0, M]$, the set $A_{\psi,t} \in \mathbb{R}^d$ is defined as

$$A_{\psi,t} = \{z : \ \psi(z) > t\}.$$

Then for $\forall n, \epsilon > 0$,

$$P\Big(\sup_{\psi\in\Psi}\Big|\frac{1}{n}\sum_{i=1}^{n}\psi(X_i) - E\psi(X_i)\Big| > \epsilon\Big) \le 8\Pi_{\mathcal{A}}(n)e^{-n\epsilon^2/(32M^2)}.$$

With Lemma 2 and notations mentioned before, we now could state and prove the following lemma that is the main goal of Step III:

Lemma 3. *With probability at least* $1-\delta$, *universally over* $\forall\theta > 0$ *and* $\forall m \geq 1$, *we have*

$$P\Big(Y\tilde{f}_m(X) \leq \frac{\theta}{2}\Big) \leq \frac{1}{n}\sum_{i=1}^{n} P_{\tilde{f}_m}\Big(Y_i\tilde{f}_m(X_i) \leq \frac{\theta}{2}\Big) + \sqrt{\frac{32\Big[\log\big(m(m+1)^2\big) + \nu(\mathcal{H})m\log(\frac{en}{\nu(\mathcal{H})}) + \log(\frac{8}{\delta})\Big]}{n}}.$$

Proof. To begin with, for a fixed m, we would denote $\mathcal{G}_m$ as the set of unweighted averages over m elements from $\mathcal{H}$:

$$\mathcal{G}_m = \Big\{g: \; x \mapsto \frac{1}{m}\sum_{i=1}^{m} h_j(x) \;\Big|\; h_1, ..., h_m \in \mathcal{H}\Big\}.$$

Thus we know that $\tilde{f}_m$ is a random function taking values from $\mathcal{G}_m$, that is, all possible values g of this random function $\tilde{f}_m$, which are functions, are elements of $\mathcal{G}_m$. Therefore, before considering the randomness in $\tilde{f}_m$, we would first start with studying $g \in \mathcal{G}_m$.

In order to leverage Lemma 2, for an arbitrary fixed $g \in \mathcal{G}_m$ and an arbitrary fixed $\theta > 0$, we construct

$$\psi_{g,\theta}(x,y) = \mathbb{1}\Big(yg(x) \leq \frac{\theta}{2}\Big),$$

thus we have $0 \leq \psi_{g(x,y),\theta} \leq 1 = M$ for all $(x,y) \in \mathcal{X} \times \{-1,1\}$. Also, we would denote $\Psi_\theta = \{\psi_{g,\theta} : \; g \in \mathcal{G}_m\}$. Now for any $t \in [0,1]$, construct

$$A_{\psi,t,\theta} = \{(x,y) : \; \psi_{g,\theta}(x,y) > t\} \qquad \text{and} \qquad \mathcal{A}_\theta = \{A_{\psi,t,\theta} : \; \psi_{g,\theta} \in \Psi_\theta, t \in [0,1]\}.$$

According to the definition of $\psi_{g,\theta}$, we have $A_{\psi,t,\theta} = \{(x,y) : \; \psi_{g,\theta}(x,y) > t\} = \{(x,y) : \; \mathbb{1}\Big(yg(x) \leq \frac{\theta}{2}\Big) > t\}$. Since the identity function $\mathbb{1}(\cdot)$ will only take values $\{0,1\}$, therefore, for $\forall t \in [0,1)$, (x,y) satisfies $\mathbb{1}\Big(yg(x) \leq \frac{\theta}{2}\Big) > t$ is the same as satisfying $yg(x) \leq \frac{\theta}{2}$. Therefore, we have an equivalent form of $A_{\psi,t,\theta}$ as

$$A_{\psi,t,\theta} = \{(x,y) : \; \psi_{g,\theta}(x,y) > t\} = \{(x,y) : \; yg(x) \leq \frac{\theta}{2}\}.$$

With this equivalent form of $A_{\psi,t,\theta}$, we could start to derive a bound the growth function of $\mathcal{A}_\theta$, that is, $\Pi_{\mathcal{A}_\theta}(n)$. First, let $\{(x_1,y_1), ..., (x_n,y_n)\}$ be arbitrary n samples from $\mathcal{X} \times \{-1,1\}$, according to Sauer's lemma, we have

$$\Big|\Big\{\big(h(x_1), ..., h(x_n)\big)\Big\} : \; h \in \mathcal{H}\Big| \leq \Big(\frac{en}{\nu(\mathcal{H})}\Big)^{\nu(\mathcal{H})}.$$

Then, since each $g \in \mathcal{G}_m$ is a combination of m functions in $\mathcal{H}$, therefore we then have

$$\Big|\Big\{\big(y_1g(x_1), ..., y_ng(x_n)\big)\Big\} : \; g \in \mathcal{G}_m\Big| \leq \Big(\frac{en}{\nu(\mathcal{H})}\Big)^{\nu(\mathcal{H})m}.$$

Finally, for any fixed θ, since $A_{\psi,t,\theta} = \{(x,y) : \; yg(x) \leq \frac{\theta}{2}\}$ and $\mathcal{A}_\theta = \{A_{\psi,t,\theta} : \; \psi_{g,\theta} \in \Psi_\theta, t \in [0,1]\} = \{A_{\psi,t,\theta} : \; g \in \mathcal{G}_m\}$, we thus have

$$\Pi_{\mathcal{A}_\theta}(n) \leq \Big(\frac{en}{\nu(\mathcal{H})}\Big)^{\nu(\mathcal{H})m}.$$

Now we could make use of Lemma 2. That is, for any fixed $\theta > 0$, based on Lemma 2, we would have that

$$P\Big(\sup_{g \in \mathcal{G}_m}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_ig(X_i) \leq \frac{\theta}{2}\big) - P\big(Yg(X) \leq \frac{\theta}{2}\big)\Big| > \epsilon\Big) \leq 8\Big(\frac{en}{\nu(\mathcal{H})}\Big)^{\nu(\mathcal{H})m} e^{-n\epsilon^2/32}, \tag{1}$$

and thus

$$P\Big(\sup_{g\in\mathcal{G}_m}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_ig(X_i)\le\frac{\theta}{2}\big)-P\big(Yg(X)\le\frac{\theta}{2}\big)\Big|\le\epsilon\Big)\ge 1-8\Big(\frac{en}{\nu(\mathcal{H})}\Big)^{\nu(\mathcal{H})m}e^{-n\epsilon^2/32}.$$

Notice that so far this bound only holds for any fixed $\theta>0$, but not uniformly over all $\theta>0$. In order to make the latter statement true, we just need one more step: denote $\Theta=\{\frac{2j}{m}: j=0,1,...,m\}$, then we would state that

$$\begin{aligned}&P\Big(\bigcup_{\theta>0}\sup_{g\in\mathcal{G}_m}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_ig(X_i)\le\frac{\theta}{2}\big)-P\big(Yg(X)\le\frac{\theta}{2}\big)\Big|>\epsilon\Big)\\ =&P\Big(\bigcup_{\theta\in\Theta}\sup_{g\in\mathcal{G}_m}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_ig(X_i)\le\frac{\theta}{2}\big)-P\big(Yg(X)\le\frac{\theta}{2}\big)\Big|>\epsilon\Big).\end{aligned}\tag{2}$$

The reason is that since Y_i only takes values from $\{-1,1\}$ and $\forall g(\cdot)\in\mathcal{G}_m$ also only take values from $\{\frac{k}{m}: k=0,1,...,m\}$, therefore all possible values of $Y_ig(X_i)$ will only be $\{\frac{k}{m}: k=-m,-m+1,...,0,1,...,m\}$. Hence, consider $\forall\theta',\theta''\in[\frac{2j}{m},\frac{2(j+1)}{m})$, $\theta'\ne\theta''$, for a specific $j\in\{0,...,m\}$, then we would have

$$\begin{aligned}&P\Big(\bigcup_{\theta\in\{\theta',\theta''\}}\sup_{g\in\mathcal{G}_m}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_ig(X_i)\le\frac{\theta}{2}\big)-P\big(Yg(X)\le\frac{\theta}{2}\big)\Big|>\epsilon\Big)\\ =&P\Big(\sup_{g\in\mathcal{G}_m}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_ig(X_i)\le\frac{\theta'}{2}\big)-P\big(Yg(X)\le\frac{\theta'}{2}\big)\Big|>\epsilon\Big),\end{aligned}$$

because these two events inside the probability measure will happen or not happen simultaneously. Therefore, denote $\theta_j=\frac{2j}{m}$, $j=0,1,...,m$, we thus further have

$$\begin{aligned}&P\Big(\bigcup_{\theta\in[\frac{2j}{m},\frac{2(j+1)}{m})}\sup_{g\in\mathcal{G}_m}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_ig(X_i)\le\frac{\theta}{2}\big)-P\big(Yg(X)\le\frac{\theta}{2}\big)\Big|>\epsilon\Big)\\ =&P\Big(\sup_{g\in\mathcal{G}_m}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_ig(X_i)\le\frac{\theta_j}{2}\big)-P\big(Yg(X)\le\frac{\theta_j}{2}\big)\Big|>\epsilon\Big),\quad\forall j=0,1,...,m,\end{aligned}$$

which proves the (2) we state before (we don't need to consider the case when $\theta>2$ and $\theta<0$ because in both cases the probability $P\big(\sup_{g\in\mathcal{G}_m}\big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_ig(X_i)\le\frac{\theta}{2}\big)-P\big(Yg(X)\le\frac{\theta}{2}\big)\big|>\epsilon\big)$ is 0 for $\forall\epsilon>0$).

Therefore, according to equation (2), we thus have

$$\begin{aligned}&P\Big(\bigcup_{\theta>0}\sup_{g\in\mathcal{G}_m}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_ig(X_i)\le\frac{\theta}{2}\big)-P\big(Yg(X)\le\frac{\theta}{2}\big)\Big|>\epsilon\Big)\\ =&P\Big(\bigcup_{\theta\in\Theta}\sup_{g\in\mathcal{G}_m}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_ig(X_i)\le\frac{\theta}{2}\big)-P\big(Yg(X)\le\frac{\theta}{2}\big)\Big|>\epsilon\Big)\\ \le&\sum_{\theta\in\Theta}P\Big(\sup_{g\in\mathcal{G}_m}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_ig(X_i)\le\frac{\theta}{2}\big)-P\big(Yg(X)\le\frac{\theta}{2}\big)\Big|>\epsilon\Big)\\ \le&8(m+1)\Big(\frac{en}{\nu(\mathcal{H})}\Big)^{\nu(\mathcal{H})m}e^{-n\epsilon^2/32},\quad\text{(according to equation (1)).}\end{aligned}\tag{3}$$

and also

$$P\Big(\bigcap_{\theta>0}\sup_{g\in\mathcal{G}_m}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_ig(X_i)\le\frac{\theta}{2}\big)-P\big(Yg(X)\le\frac{\theta}{2}\big)\Big|\le\epsilon\Big)\ge 1-8(m+1)\Big(\frac{en}{\nu(\mathcal{H})}\Big)^{\nu(\mathcal{H})m}e^{-n\epsilon^2/32}.\tag{4}$$

Let's interpret equation (4): with probability at least $1-8(m+1)\Big(\frac{en}{\nu(\mathcal{H})}\Big)^{\nu(\mathcal{H})m} e^{-n\epsilon^2/32}$, for $\forall\theta > 0$ and $\forall g \in \mathcal{G}_m$, we have

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_i g(X_i) \le \frac{\theta}{2}\big) - P\big(Yg(X) \le \frac{\theta}{2}\big)\Big| \le \epsilon,$$

which obviously leads to that with probability at least $1-8(m+1)\Big(\frac{en}{\nu(\mathcal{H})}\Big)^{\nu(\mathcal{H})m} e^{-n\epsilon^2/32}$, for $\forall\theta > 0$ and $\forall g \in \mathcal{G}_m$, we have

$$P\big(Yg(X) \le \frac{\theta}{2}\big) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_i g(X_i) \le \frac{\theta}{2}\big) + \epsilon.$$

Since the random function $\tilde{f}_m$ only take values from $\mathcal{G}_m$, and the probability $1-8(m+1)\Big(\frac{en}{\nu(\mathcal{H})}\Big)^{\nu(\mathcal{H})m} e^{-n\epsilon^2/32}$ is with respect to the randomness of drawing a random set of samples set from the generative distribution and whenever a set of sample $\{(x_1, y_1), ..., (x_n, y_n)\}$ let the $P\big(Yg(X) \le \frac{\theta}{2}\big) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(y_i g(x_i) \le \frac{\theta}{2}\big) + \epsilon$ holds for $\forall\theta > 0$ and $\forall g \in \mathcal{G}_m$, we would also have $P_{X,Y}\big(Y\tilde{f}_m(X) \le \frac{\theta}{2}\big) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(y_i\tilde{f}_m(x_i) \le \frac{\theta}{2}\big) + \epsilon$ holds, therefore, we would have with probability $1-8(m+1)\Big(\frac{en}{\nu(\mathcal{H})}\Big)^{\nu(\mathcal{H})m} e^{-n\epsilon^2/32}$ with respect to the random set of samples, for $\forall\theta > 0$,

$$P_{X,Y}\big(Y\tilde{f}_m(X) \le \frac{\theta}{2}\big) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}\big) + \epsilon.$$

Since the randomness of $\tilde{f}_m$ is independent of the randomness of samples, thus by taking expectation with respect to $\tilde{f}_m$ on both sides, we thus have that with probability $1-8(m+1)\Big(\frac{en}{\nu(\mathcal{H})}\Big)^{\nu(\mathcal{H})m} e^{-n\epsilon^2/32}$ with respect to the random set of samples, for $\forall\theta > 0$,

$$P\big(Y\tilde{f}_m(X) \le \frac{\theta}{2}\big) \le \frac{1}{n}\sum_{i=1}^{n}P_{\tilde{f}_m}\big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}\big) + \epsilon. \tag{5}$$

Therefore, by letting

$$\epsilon = \sqrt{\frac{32\Big[\log\big(m(m+1)^2\big) + \nu(\mathcal{H})m\log(\frac{en}{\nu(\mathcal{H})}) + \log(\frac{8}{\delta})\Big]}{n}},$$

then for the fixed m we would have that with probability at least $1 - \frac{\delta}{m(m+1)}$ with respect to random sets of samples, for $\forall\theta > 0$,

$$P\big(Y\tilde{f}_m(X) \le \frac{\theta}{2}\big) > \frac{1}{n}\sum_{i=1}^{n}P_{\tilde{f}_m}\big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}\big) + \sqrt{\frac{32\Big[\log\big(m(m+1)^2\big) + \nu(\mathcal{H})m\log(\frac{en}{\nu(\mathcal{H})}) + \log(\frac{8}{\delta})\Big]}{n}},$$

which is equivalent as stating that for a fixed m, with probability at most $\frac{\delta}{m(m+1)}$ with respect to random sets of samples, $\exists\theta > 0$ such that

$$P\big(Y\tilde{f}_m(X) \le \frac{\theta}{2}\big) \le \frac{1}{n}\sum_{i=1}^{n}P_{\tilde{f}_m}\big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}\big) + \sqrt{\frac{32\Big[\log\big(m(m+1)^2\big) + \nu(\mathcal{H})m\log(\frac{en}{\nu(\mathcal{H})}) + \log(\frac{8}{\delta})\Big]}{n}}. \tag{6}$$

Since $\sum_{m=1}^{\infty}\frac{\delta}{m(m+1)} = \delta\sum_{m=1}^{\infty}(\frac{1}{m} - \frac{1}{m+1}) = \delta$, therefore (6) for all $m \ge 1$, by a similar steps as statement (3), we would have that with probability at most $\sum_{m=1}^{\infty}\frac{\delta}{m(m+1)} = \delta$, $\theta > 0$ and $\exists m \ge 1$ ($m \in \mathbb{Z}^+$, which is the set of positive integer) such that

$$P\big(Y\tilde{f}_m(X) \le \frac{\theta}{2}\big) > \frac{1}{n}\sum_{i=1}^{n}P_{\tilde{f}_m}\big(Y_i\tilde{f}_m(X_i) \le \frac{\theta}{2}\big) + \sqrt{\frac{32\Big[\log\big(m(m+1)^2\big) + \nu(\mathcal{H})m\log(\frac{en}{\nu(\mathcal{H})}) + \log(\frac{8}{\delta})\Big]}{n}},$$

which finally leads to our statement in the lemma that with probability at least $1-\delta$ with respect to random sets of samples, universally over $\forall\theta > 0$ and $\forall m \geq 1$, we have

$$P\big(Y\tilde{f}_m(X) \leq \frac{\theta}{2}\big) > \frac{1}{n}\sum_{i=1}^{n} P_{\tilde{f}_m}\big(Y_i\tilde{f}_m(X_i) \leq \frac{\theta}{2}\big) + \sqrt{\frac{32\Big[\log\big(m(m+1)^2\big) + \nu(\mathcal{H})m\log(\frac{en}{\nu(\mathcal{H})}) + \log(\frac{8}{\delta})\Big]}{n}}.$$

Thus we finished the proof for Lemma 3.

□

Now we could use Lemma 3 to connect all things together. That is, as the idea we stated before in Step I, according to Lemma 3 we would have that with probability at least $1-\delta$, universally over $\forall\theta > 0$ and $\forall m \geq 1$,

$$\begin{aligned} P\Big(Yf(X) \leq 0\Big) \leq& P\Big(Y\tilde{f}_m(X) \leq \frac{\theta}{2}\Big) + 2e^{-m\theta^2/8} \quad \text{(Step I)} \\ \leq& \frac{1}{n}\sum_{i=1}^{n} P_{\tilde{f}_m}\Big(Y_i\tilde{f}_m(X_i) \leq \frac{\theta}{2}\Big) + 2e^{-n\theta/8} \\ &+ \sqrt{\frac{32\Big[\log\big(m(m+1)^2\big) + \nu(\mathcal{H})m\log(\frac{en}{\nu(\mathcal{H})}) + \log(\frac{8}{\delta})\Big]}{n}} \quad \text{(Lemma 3)} \\ \leq& \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\Big(Y_if(X_i) \leq \frac{\theta}{2}\Big) + 4e^{-m\theta^2/8} \\ &+ \sqrt{\frac{32\Big[\log\big(m(m+1)^2\big) + \nu(\mathcal{H})m\log(\frac{en}{\nu(\mathcal{H})}) + \log(\frac{8}{\delta})\Big]}{n}} \quad \text{(Step II).} \end{aligned} \tag{7}$$

In order to let the latter two terms on the right side of (7) to the same magnitude, for $\forall\theta > 0$ and any fixed $n \geq 1$, choosing

$$m = \Big\lceil\frac{4}{\theta^2}\log\Big(\frac{n\theta^2}{8\nu(\mathcal{H})\log\big(\frac{en}{\nu(\mathcal{H})}\big)}\Big)\Big\rceil,$$

we therefore could have the bound stated in Theorem 3. Thus we finished the proof.

□

Section 2.3 Bounding the Fraction of Samples with Margin Less than θ

Finally, as stated in the last part of Section 2.1, we would prove that for any set of samples $\{(x_1, y_1), ..., (x_n, y_n)\}$ and $\theta \in [-1, 1]$ satisfying some conditions, $\frac{1}{n}\mathbb{1}(y_if(x_i) \leq \theta)$ would decrease exponentially fast to 0 as T increase, which is actually an improved theorem of Theorem 1:

Theorem 4. *Let H be the ensemble classifier generate by adaptive boosting after T rounds on samples $\{(x_1, y_1), ..., (x_n, y_n)\}$, and also define γ_t as $\gamma_t = \frac{1}{2} - \epsilon_t$. Then we would have*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_if(x_i) \leq \theta) \leq \prod_{i=1}^{T}\sqrt{(1+2\gamma_t)^{1+\theta}(1-2\gamma_t)^{1-\theta}}.$$

Proof. The proof is very similar to Step II of the proof of Theorem 1. Recall that

$$f(x) = \frac{\sum_{i=1}^{T}\alpha_th_t(x)}{\sum_{i=1}^{T}\alpha_t},$$

therefore we have

$$y_i f(x_i) \leq \theta \quad \Leftrightarrow \quad y\frac{\sum_{i=1}^{T}\alpha_t h_t(x_i)}{\sum_{i=1}^{T}\alpha_t} \leq \theta \quad \Leftrightarrow \quad y\sum_{i=1}^{T}\alpha_t h_t(x_i) \leq \theta\sum_{i=1}^{T}\alpha_t \quad \Leftrightarrow \quad 1 \leq \exp\big(-y\sum_{i=1}^{T}\alpha_t h_t(x_i)+\theta\sum_{i=1}^{T}\alpha_t\big),$$

which leads to that

$$\mathbb{1}(y_i f(x_i) \leq \theta) \leq \exp\big(-y\sum_{i=1}^{T}\alpha_t h_t(x_i)+\theta\sum_{i=1}^{T}\alpha_t\big).$$

Thus, we could have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i f(x_i) \leq \theta) \leq & \frac{1}{n}\sum_{i=1}^{n}\exp\big(-y\sum_{i=1}^{T}\alpha_t h_t(x_i)+\theta\sum_{i=1}^{T}\alpha_t\big) \\
= & \frac{\exp\big(\theta\sum_{i=1}^{T}\alpha_t\big)}{n}\sum_{i=1}^{n}\exp\big(-y\sum_{i=1}^{T}\alpha_t h_t(x_i)\big) \\
= & \frac{\exp\big(\theta\sum_{i=1}^{T}\alpha_t\big)}{n}\sum_{i=1}^{n}\big[n\mathcal{D}_{T+1}(i)\prod_{t=1}^{T}Z_t\big] \\
= & \exp\big(\theta\sum_{i=1}^{T}\alpha_t\big)\prod_{t=1}^{T}Z_t\sum_{i=1}^{n}\mathcal{D}_{T+1}(i) \\
= & \exp\big(\theta\sum_{i=1}^{T}\alpha_t\big)\prod_{t=1}^{T}Z_t.
\end{aligned}$$

Lastly, by plugging the values we derived in the Step III in the proof of Theorem 1:

$$\alpha_t = \frac{1}{2}\log\Big(\frac{1-\epsilon_t}{\epsilon_t}\Big) = \frac{1}{2}\log\Big(\frac{\frac{1}{2}+\gamma_t}{\frac{1}{2}-\gamma_t}\Big),\ t = 1, ..., T \quad \text{and} \quad \prod_{t=1}^{T}Z_t = \prod_{t=1}^{T}\sqrt{(1-2\gamma_t)(1+2\gamma_t)},$$

we therefore could have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i f(x_i) \leq \theta) \leq \prod_{i=1}^{T}\sqrt{(1+2\gamma_t)^{1+\theta}(1-2\gamma_t)^{1-\theta}},$$

and thus finish the proof of Theorem 4. □

Notice that if for $\forall t = 1, ..., T$ we have $\gamma_t \geq \gamma$ and

$$\sqrt{(1+2\gamma)^{1+\theta}(1-2\gamma)^{1-\theta}} < 1,$$

or equivalently

$$\theta < \frac{-\log(1-4\gamma^2)}{\log(1+2\gamma)-\log(1-2\gamma)},$$

then we would have Theorem 4 state that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i f(x_i) \leq \theta) \leq \Big(\sqrt{(1+2\gamma)^{1+\theta}(1-2\gamma)^{1-\theta}}\Big)^T,$$

which means that the $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i f(x_i) \leq \theta)$ will decrease to 0 exponentially fast as T increase.

Recall that in section 2.2, we have proved that with a large probability, the generalization error is bound by $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i f(x_i) \leq \theta)$ plus a relative small constant with respect to $\nu(\mathcal{H}), \theta$ and sample size n, both of which are independent of T. Therefore, when the above condition is satisfied, we would have the $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i f(x_i) \leq \theta)$ term decrease to 0 exponential fast as T increase, and thus what is left in the bound of generalization error is the relative constant. This thus illustrate the non-overfitting of adaptive boosting algorithm, and is marked as the end of our whole story.

References

[1] Vapnik, V. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264-280.

[2] Littlestone, N., & Warmuth, M. (1986). Relating data compression and learnability (Vol. 23). *Technical report*, University of California, Santa Cruz.

[3] Freund, Y., & Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory* (pp. 23-37). Springer Berlin Heidelberg.

[4] Floyd, S., & Warmuth, M. (1995). Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine learning*, 21(3), 269-304.

[5] Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of statistics*, 1651-1686.

[6] Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3), 297-336.

[7] Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification* (pp. 149-171). Springer New York.

[8] Rowling, J. K. (2009). Fantastic beasts and where to find them. *Bloomsbury Publishing.*

[9] Schapire, R. E., & Freund, Y. (2012). Boosting: Foundations and algorithms. *MIT press.*

[10] Devroye, L., Gyorfi, L., & Lugosi, G. (2013). A probabilistic theory of pattern recognition (Vol. 31). *Springer Science* & *Business Media.*

[11] Wyner, A. J., Olson, M., Bleich, J., & Mease, D. (2015). Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers. arXiv preprint arXiv:1504.07676.

[12] Scamander, N., & Rowling, J. K. (2015). Fantastic Beasts and Where to Find Them.