Margin-Based Generalization Lower Bounds for Boosted Classifiers

Anonymous Author(s)
Affiliation
Address
email

Abstract

Boosting is one of the most successful ideas in machine learning. The most well-accepted explanations for the low generalization error of boosting algorithms such as AdaBoost stem from margin theory. The study of margins in the context of boosting algorithms was initiated by Schapire, Freund, Bartlett and Lee (1998) and has inspired numerous boosting algorithms and generalization bounds. To date, the strongest known generalization (upper bound) is the kth margin bound of Gao and Zhou (2013). Despite the numerous generalization upper bounds that have been proved over the last two decades, nothing is known about the tightness of these bounds. In this paper, we give the first margin-based lower bounds on the generalization error of boosted classifiers. Our lower bounds nearly match the kth margin bound and thus almost settle the generalization performance of boosted classifiers in terms of margins.

1 Introduction

Boosting algorithms produce highly accurate classifiers by combining several less accurate classifiers and are amongst the most popular learning algorithms, obtaining state-of-the-art performance on several benchmark machine learning tasks [KMF$^+$17, CG16]. The most famous of these boosting algorithm is arguably AdaBoost [FS97]. For binary classification, AdaBoost takes a training set $S = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$ of m labeled samples as input, with $x_i \in \mathcal{X}$ and labels $y_i \in \{-1, 1\}$. It then produces a classifier f in iterations: in the jth iteration, a base classifier $h_j : \mathcal{X} \to \{-1, 1\}$ is trained on a reweighed version of S that emphasizes data points that f struggles with and this classifier is then added to f. The final classifier is obtained by taking the sign of $f(x) = \sum_j \alpha_j h_j(x)$, where the α_j's are non-negative coefficients carefully chosen by AdaBoost. The base classifiers h_j all come from a *hypothesis set* $\mathcal{H}$, e.g. $\mathcal{H}$ could be a set of small decision trees or similar. As AdaBoost's training progresses, more and more base classifiers are added to f, which in turn causes the training error of f to decrease. If $\mathcal{H}$ is rich enough, AdaBoost will eventually classify all the data points in the training set correctly [FS97].

Early experiments with AdaBoost report a surprising generalization phenomenon [SFB$^+$98]. Even after perfectly classifying the entire training set, further iterations keeps improving the test accuracy. This is contrary to what one would expect, as f gets more complicated with more iterations, and thus prone to overfitting. The most prominent explanation for this phenomena is margin theory, introduced by Schapire *et al.* [SFB$^+$98]. The margin of a training point (x_i, y_i) is a number in $[-1, 1]$, which can be interpreted, loosely speaking, as the classifier's confidence on that point. Formally, we say that $f(x) = \sum_j \alpha_j h_j(x)$ is a *voting classifier* if $\alpha_j \geq 0$ for all j. Note that one can additionally assume without loss of generality that $\sum_j \alpha_j = 1$ since normalizing each α_i by $\sum_j \alpha_j$ leaves the sign of $f(x_i)$ unchanged. The margin of a point (x_i, y_i) with respect to a voting classifier f is then defined

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

as

$$\text{margin}(x_i) := y_i f(x_i) \quad = \quad y_i \sum_j \alpha_j h_j(x_i) \; .$$

Thus $\text{margin}(x_i) \in [-1, 1]$, and if $\text{margin}(x_i) > 0$, then taking the sign of $f(x_i)$ correctly classifies (x_i, y_i). Informally speaking, margin theory guarantees that voting classifiers with large (positive) margins have a smaller generalization error. Experimentally AdaBoost has been found to continue to improve the margins even when training past the point of perfectly classifying the training set. Margin theory may therefore explain the surprising generalization phenomena of AdaBoost. Indeed, the original paper by Schapire *et al.* [SFB$^+$98] that introduced margin theory, proved the following margin-based generalization bound. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X} \times \{-1, 1\}$ and assume that the training data S is obtained by drawing m i.i.d. samples from $\mathcal{D}$. Then with high probability over S it holds that for every margin $\theta \in (0, 1]$, *every* voting classifier f satisfies

$$\Pr_{(x,y)\sim\mathcal{D}}[yf(x) \leq 0] \leq \Pr_{(x,y)\sim S}[yf(x) < \theta] + O\left(\sqrt{\frac{\ln|\mathcal{H}| \ln m}{\theta^2 m}}\right). \tag{1}$$

The left-hand side of the equation is the out-of-sample error of f (since $\text{sign}(f(x)) \neq y$ precisely when $yf(x) < 0$). On the right-hand side, we use $(x, y) \sim S$ to denote a uniform random point from S. Hence $\Pr_{(x,y)\sim S}[yf(x) < \theta]$ is the fraction of training points with margin less than θ. The last term is increasing in $|\mathcal{H}|$ and decreasing in θ and m. Here it is assumed $\mathcal{H}$ is finite. A similar bound can be proved for infinite $\mathcal{H}$ by replacing $|\mathcal{H}|$ by $d \lg m$, where d is the VC-dimension of $\mathcal{H}$. This holds for all the generalization bounds below as well. The generalization bound thus shows that f has low out-of-sample error if it attains large margins on most training points. This fits well with the observed behaviour of AdaBoost in practice.

The generalization bound above holds for every voting classifier f, i.e. regardless of how f was obtained. Hence a natural goal is to design boosting algorithms that produce voting classifiers with large margins on many points. This has been the focus of a long line of research and has resulted in numerous algorithms with various margin guarantees, see e.g. [GS98, Bre99, BDST00, RW02, RW05, GLM19]. One of the most well-known of these is Breimann's ArcGV [Bre99]. ArcGV produces a voting classifier maximizing the *minimal* margin, i.e. it produces a classifier f for which $\min_{(x,y)\in S} yf(x)$ is as large as possible. Breimann complemented the algorithm with a generalization bound stating that with high probability over the sample S, it holds that every voting classifier f satisfies:

$$\Pr_{(x,y)\sim\mathcal{D}}[yf(x) \leq 0] \leq O\left(\frac{\ln|\mathcal{H}| \ln m}{\hat{\theta}^2 m}\right), \tag{2}$$

where $\hat{\theta} = \min_{(x,y)\in S} yf(x)$ is the minimal margin over all training examples. Notice that if one chooses θ as the minimal margin in the generalization bound (1) of Schapire *et al.* [SFB$^+$98], then the term $\Pr_{(x,y)\sim S}[yf(x) < \theta]$ becomes 0 and one obtains the bound

$$\Pr_{(x,y)\sim\mathcal{D}}[yf(x) \leq 0] \leq O\left(\sqrt{\frac{\ln|\mathcal{H}| \ln m}{\hat{\theta}^2 m}}\right),$$

which is weaker than Breimann's bound and motivated his focus on maximizing the minimal margin. Minimal margin is however quite sensitive to outliers and work by Gao and Zhou [GZ13] proved a generalization bound which provides an interpolation between (1) and (2). Their bound is known as the kth margin bound, and states that with high probability over the sample S, it holds for every margin $\theta \in (0, 1]$ and every voting classifier f that:

$$\Pr_{(x,y)\sim\mathcal{D}}[yf(x) < 0] \leq \Pr_{(x,y)\sim S}[yf(x) < \theta] + O\left(\frac{\ln|\mathcal{H}| \ln m}{\theta^2 m} + \sqrt{\Pr_{(x,y)\sim S}[yf(x) < \theta] \frac{\ln|\mathcal{H}| \ln m}{\theta^2 m}}\right).$$

The kth margin bound remains the strongest margin-based generalization bound to date (see Section 1.2 for further details). The kth margin bound recovers Breimann's minimal margin bound by choosing θ as the minimal margin (making $\Pr_{(x,y)\sim S}[yf(x) < \theta] = 0$), and it is always at most the same as the bound (1) by Schapire *et al.* As with previous generalization bounds, it suggests that boosting algorithms should focus on obtaining a large margin on as large a fraction of training points as possible.

Despite the decades of progress on generalization *upper* bounds, we still do not know how tight these bounds are. That is, we do not have any margin-based generalization *lower* bounds. Generalization lower bounds are not only interesting from a theoretical point of view, but also from an algorithmic point of view: If one has a provably tight generalization bound, then a natural goal is to design a boosting algorithm minimizing a loss function that is equal to this generalization bound. This approach makes most sense with a matching lower bound as the algorithm might otherwise minimize a sub-optimal loss function. Furthermore, a lower bound may also inspire researchers to look for other parameters than margins when explaining the generalization performance of voting classifiers. Such new parameters may even prove useful in designing new algorithms, with even better generalization performance in practice.

1.1 Our Results

In this paper we prove the first margin-based generalization lower bounds for voting classifiers. Our lower bounds almost match the kth margin bound and thus essentially settles the generalization performance of voting classifiers in terms of margins.

To present our main theorems, we first introduce some notation. For a ground set $\mathcal{X}$ and hypothesis set $\mathcal{H}$, let $C(\mathcal{H})$ denote the family of all voting classifiers over $\mathcal{H}$, i.e. $C(\mathcal{H})$ contains all functions $f : \mathcal{X} \to [-1, 1]$ that can be written as $f(x) = \sum_{h\in\mathcal{H}} \alpha_h h(x)$ such that $\alpha_h \geq 0$ for all h and $\sum_h \alpha_h = 1$. For a (randomized) learning algorithm $\mathcal{A}$ and a sample S of m points, let $f_{\mathcal{A},S}$ denote the (possibly random) voting classifier produced by $\mathcal{A}$ when given the sample S as input. With this notation, our first main theorem is the following:

Theorem 1. *For every large enough integer N, every $\theta \in (1/N, 1/40)$ and every $\tau \in [0, 49/100]$ there exist a set $\mathcal{X}$ and a hypothesis set $\mathcal{H}$ over $\mathcal{X}$, such that $\ln|\mathcal{H}| = \Theta(\ln N)$ and for every $m = \Omega\left(\theta^{-2}\ln|\mathcal{H}|\right)$ and for every (randomized) learning algorithm $\mathcal{A}$, there exist a distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}$ and a voting classifier $f \in C(\mathcal{H})$ such that with probability at least $1/100$ over the choice of samples $S \sim \mathcal{D}^m$ and the random choices of $\mathcal{A}$*

1. $\Pr_{(x,y)\sim S}[yf(x) < \theta] \leq \tau$*; and*
2. $\Pr_{(x,y)\sim\mathcal{D}}[yf_{\mathcal{A},S}(x) < 0] \geq \tau + \Omega\left(\frac{\lg|\mathcal{H}|}{m\theta^2} + \sqrt{\tau\cdot\frac{\lg|\mathcal{H}|}{m\theta^2}}\right)$.

Theorem 1 states that for any algorithm $\mathcal{A}$, there is a distribution $\mathcal{D}$ for which the out-of-sample error of the voting classifier produced by $\mathcal{A}$ is at least that in the second point of the theorem. At the same time, one can find a voting classifier f obtaining a margin of at least θ on at least a $1 - \tau$ fraction of the sample points. Notice that we cannot hope to prove that the algorithm $\mathcal{A}$ constructs a voting classifier that has a margin of at least θ on a $1 - \tau$ fraction, since we make no assumptions on the algorithm. For example, if the constant hypothesis h_1 that always outputs 1 is in $\mathcal{H}$, then $\mathcal{A}$ could be the algorithm that simply outputs h_1. The interpretation is thus: It is always possible for an algorithm $\mathcal{A}$ to produce a voting classifier f with margin at least θ on a $1 - \tau$ fraction of samples, and regardless of which voting classifier $\mathcal{A}$ produces, it still has large out-of-sample error. Comparing Theorem 1 to the kth margin bound, we see that the parameter τ corresponds to $\Pr_{(x,y)\sim S}[yf(x) < \theta]$. The magnitude of the out-of-sample error in the second point in the theorem thus matches that of the kth margin bound, except for a factor $\lg m$ in the first term inside the $\Omega(\cdot)$ and a $\sqrt{\lg m}$ factor in the second term. If we consider the range of parameters $\theta, \tau, \ln|\mathcal{H}|$ and m for which the lower bound applies, then these ranges are almost as tight as possible. For τ, note that the theorem cannot generally be true for $\tau > 1/2$, as the algorithm $\mathcal{A}$ that outputs a uniform random choice of hypothesis among h_1 and h_{-1} (the constant hypothesis outputting -1), gives a (random) voting classifier $f_{\mathcal{A},S}$ with an out-of-sample error of $1/2$. This is less than the second point of the theorem would state if it was true for $\tau > 1/2$. For $\ln|\mathcal{H}|$, observe that our theorem holds for arbitrarily large values of $|\mathcal{H}|$. That is, the integer N can be as large as desired, making $\ln|\mathcal{H}| = \Theta(\ln N)$ as large as desired. Finally, for the constraint on m, notice again that the theorem simply cannot be true for smaller values of m as then the term $\lg|\mathcal{H}|/(m\theta^2)$ exceeds 1.

Our second main result gets even closer to the kth margin bound:

Theorem 2. *For every large enough integer N, every $\theta \in (1/N, 1/40)$, $\tau \in [0, 49/100]$ and every $m = \left(\theta^{-2}\ln N\right)^{1+\Omega(1)}$, there exist a set $\mathcal{X}$, a hypothesis set $\mathcal{H}$ over $\mathcal{X}$ and a distribution $\mathcal{D}$ over*

$\mathcal{X} \times \{-1,1\}$ *such that* $\ln|\mathcal{H}| = \Theta(\ln N)$ *and with probability at least* $1/100$ *over the choice of samples* $S \sim \mathcal{D}^m$ *there exists a voting classifier* $f_S \in C(\mathcal{H})$ *such that*

1. $\Pr_{(x,y)\sim S}[yf_S(x) < \theta] \leq \tau$*; and*
2. $\Pr_{(x,y)\sim \mathcal{D}}[yf_S(x) < 0] \geq \tau + \Omega\left(\frac{\lg|\mathcal{H}|\lg m}{m\theta^2} + \sqrt{\tau \cdot \frac{\lg|\mathcal{H}|}{m\theta^2}}\right)$.

Observe that the second point of Theorem 2 has an additional $\lg m$ factor on the first term in $\Omega(\cdot)$ compared to Theorem 1. It is thus only off from the kth margin bound by a $\sqrt{\lg m}$ factor in the second term and hence completely matches the kth margin bound for small values of τ. To obtain this strengthening, we replaced the guarantee in Theorem 1 saying that *all* algorithms $\mathcal{A}$ have such a large out-of-sample error. Instead, Theorem 2 demonstrates only the existence of a voting classifier f_S (that is chosen as a function of the sample S) that simultaneously achieves a margin of at least θ on a $1-\tau$ fraction of the sample points, and yet has out-of-sample error at least that in point 2. Since the kth margin bound holds with high probability *for all* voting classifiers, Theorem 2 rules out any strengthening of the kth margin bound, except for possibly a $\sqrt{\lg m}$ factor on the second additive term. Again, our lower bound holds for almost the full range of parameters of interest.

Finally, we mention that both our lower bounds are proved for a finite hypothesis set $\mathcal{H}$. This only makes the lower bounds stronger than if we proved it for an infinite $\mathcal{H}$ with bounded VC-dimension, since the VC-dimension of a finite $\mathcal{H}$, is no more than $\lg|\mathcal{H}|$.

1.2 Related Work

We mentioned above that the kth margin bound is the strongest margin-based generalization bound to date. Technically speaking, it is incomparable to the so-called *emargin* bound by Wang *et al.* [WSJ+11]. The kth margin bound by Gao and Zhou [GZ13], the minimum margin bound by Breimann [Bre99] and the bound by Schapire *et al.* [SFB+98] all have the form $\Pr_{(x,y)\sim\mathcal{D}}[yf(x) < 0] \leq \Pr_{(x,y)\sim S}[yf(x) < \theta] + \Gamma(\theta, m, |\mathcal{H}|, \Pr_{(x,y)\sim S}[yf(x) < \theta])$ for some function Γ. The emargin bound has a different (and quite involved) form, making it harder to interpret and compute. We will not discuss it in further detail here and just remark that our results show that for generalization bounds of the form studied in most previous work [SFB+98, Bre99, GZ13], one cannot hope for much stronger upper bounds than the kth margin bound.

2 Proof Overview

The main argument that lies in the heart of both proofs is a probabilistic method argument. With every labeling $\ell \in \{-1,1\}^u$ we associate a distribution $\mathcal{D}_\ell$ over $\mathcal{X} \times \{-1,1\}$. We then show that with some positive probability if we sample $\ell \in \{-1,1\}^u$, $\mathcal{D}_\ell$ satisfies the requirements of Theorem 1 (respectively Theorem 2). We thus conclude the existence of a suitable distribution. We next give a more detailed high-level description of the proof for Theorem 1. The proof of Theorem 2 follows similar lines.

Constructing a Family of Distributions. We start by first describing the construction of $\mathcal{D}_\ell$ for $\ell \in \{-1,1\}^u$. Our construction combines previously studied distribution patterns in a subtle manner.

Ehrenfeucht *et al.* [EHKV89] observed that if a distribution $\mathcal{D}$ assigns each point in $\mathcal{X}$ a fixed (yet unknown) label, then, loosely speaking, every classifier f, that is constructed using only information supplied by a sample S, cannot do better than random guessing the labels for the points in $\mathcal{X} \setminus S$. Intuitively, consider a distribution $\mathcal{D}_\ell$ that assigns very small probability, say $\frac{1}{10m}$, to each element $x \in \mathcal{X}$. With very high probability over a sample S of m points, many elements of $\mathcal{X}$ are not in S. Moreover, assume that $\mathcal{D}_\ell$ associates every $x \in \mathcal{X}$ with a unique "correct" label $\ell(x)$. Consider some (perhaps random) learning algorithm $\mathcal{A}$, and let $f_{\mathcal{A},S}$ be the classifier it produces given a sample S as input. If ℓ is chosen randomly, then, loosely speaking, for every point x not in the sample, $f_{\mathcal{A},S}(x)$ and $\ell(x)$ are independent, and thus $\mathcal{A}$ returns the wrong label with probability $1/2$. In turn, this implies that there exists a labeling ℓ such that $\mathcal{A}$ is wrong on a constant fraction of $\mathcal{X}$ when receiving a sample $S \sim \mathcal{D}_\ell^m$. We remark that the argument above can in fact be used to prove an arbitrarily large generalization error. However, assigning *every* point in $\mathcal{X}$ a probability of $1/10m$

requires $|\mathcal{X}| \geq 10m$. This conflicts with the first point in Theorem 1, that is, we have to argue that a voting classifier f with good margins exist for the sample S. If S consists of m distinct points, and each point in $\mathcal{X}$ can have an arbitrary label, then intuitively $\mathcal{H}$ needs to be very large to ensure the existence of f. In order to overcome this difficulty, we set $\mathcal{D}_\ell$ to assign very high probability to one designated point in $\mathcal{X}$, and the rest of the probability mass is then equally distributed between all other points. The argument above still applies for the subset of small-probability points. More precisely, if $\mathcal{D}_\ell$ assigns all but one point in $\mathcal{X}$ probability $\frac{1}{10m}$, then the expected generalization error (over the choice of ℓ) is still $\Omega\left(\frac{1}{10m}|\mathcal{X}|\right)$. Therefore there exists a labeling ℓ such that $\mathcal{A}$ is wrong on a constant fraction of $\mathcal{X}$ when receiving a sample $S \sim \mathcal{D}_\ell^m$. In the notations of the theorem, in order for a hypothesis set $\mathcal{H}$ to satisfy $\ln|\mathcal{H}| = \Theta(\ln N)$, and at the same time, have an $f \in C(\mathcal{H})$ obtaining margins of θ on most points in a sample, our proof (and specifically Lemma 3, described hereafter) requires $\mathcal{X}$ to be not significantly larger than $\frac{\ln N}{\theta^2}$, and therefore the generalization error we get is $\Omega\left(\frac{\ln|\mathcal{H}|}{\theta^2 m}\right)$. This accounts for the first term inside the Ω-notation in the second point of Theorem 1.

Anthony and Bartlett [AB09, Chapter 5] additionally observed that for a distribution $\mathcal{D}$ that assigns each point in $\mathcal{X}$ a random label, if S does not sample a point x enough times, any classifier f, that is constructed using only information supplied by S, cannot determine with good probability the Bayes label of x, that is, the label of x that minimizes the error probability. Intuitively, consider once more a distribution $\mathcal{D}_\ell$ that is uniform over $\mathcal{X}$. However, instead of associating every point $x \in \mathcal{X}$ with one correct label $\ell(x)$, $\mathcal{D}_\ell$ is now only slightly biased towards ℓ. That is, given that x is sampled, the label in the sample point is $\ell(x)$ with probability that is a little larger than $1/2$, say $(1+\alpha)/2$ for some small $\alpha \in (0,1)$. Note that every classifier f has an error probability of at least $(1-\alpha)/2$ on every given point in $\mathcal{X}$. Consider once again a learning algorithm $\mathcal{A}$ and the voting classifier $f_{\mathcal{A},S}$ it constructs. Loosely speaking, if S does not sample a point x enough times, then with good probability $f_{\mathcal{A},S}(x) \neq \ell(x)$. More formally, in order to correctly assign the Bayes label of x, an algorithm must see $\Omega(\alpha^{-2})$ samples of x. Therefore if we set the bias α to be $\sqrt{|\mathcal{X}|/(10m)}$, then with high probability the algorithm does not see a constant fraction of $\mathcal{X}$ enough times to correctly assign their label. In turn, this implies an expected generalization error of $(1-\alpha)/2 + \Omega(\sqrt{|\mathcal{X}|/m})$, where the expectation is over the choice of ℓ. By once again letting $|\mathcal{X}| = \frac{\ln N}{\theta^2}$ we conclude that there exists a labeling ℓ such that for $S \sim \mathcal{D}_\ell^m$, the expected generalization error of $f_{\mathcal{A},S}$ is $\frac{1-\alpha}{2} + \Omega\left(\sqrt{\frac{\ln|\mathcal{H}|}{\theta^2 m}}\right)$. This expression is almost the second term inside the Ω-notation in the theorem statement, though slightly larger. We note, however, for large values of m, the in-sample error is arbitrarily close to $1/2$. One challenge is therefore to reduce the in-sample-error, and moreover guarantee that we can find a voting classifier f where the $(m\tau)$'th smallest margin for f is at least θ, where τ, θ are the parameters provided by the theorem statement.

To this end, our proof subtly weaves the two ideas described above and constructs a family of distributions $\{\mathcal{D}_\ell\}_{\ell \in \{-1,1\}^u}$. Informally, we partition $\mathcal{X}$ into two disjoint sets, and conditioned on the sample point $x \in \mathcal{X}$ belonging to each of the subsets, $\mathcal{D}_\ell$ is defined similarly to be one of the two distribution patterns defined above. The main difficulty lies in delicately balancing all ingredients and ensuring that we can find an f with margins of at least θ on all but τm of the sample points, while still enforcing a large generalization error. Our proof refines the proof given by Ehrenfeucht *et al.* and Anthony and Bartlett and shows that not only does there exists a labeling ℓ such that $f_{\mathcal{A},S}$ has large generalization error with respect to $\mathcal{D}_\ell$ (with probability at least $1/100$ over the randomness of $\mathcal{A}, S$), but rather that a large (constant) fraction of labelings ℓ share this property. This distinction becomes crucial in the proof.

Small yet Rich Hypothesis Sets. The technical crux in our proofs is the construction of an appropriate hypothesis set. Loosely speaking, the size of $\mathcal{H}$ has to be small, and most importantly, independent of the size m of the sample set. On the other hand, the set of voting classifiers $C(\mathcal{H})$ is required to be rich enough to, intuitively, contain a classifier that with good probability has good in-sample margins for a sample $S \sim \mathcal{D}_\ell^m$ with a large fraction of labelings $\ell \in \{-1,1\}^u$. Our main technical lemma presents a distribution μ over small hypothesis sets $\mathcal{H} \subset \mathcal{X} \to \{-1,1\}$ such that for every *sparse* $\ell \in \{-1,1\}^u$, that is $\ell_i = -1$ for a small number of entries $i \in [u]$, with high probability over $\mathcal{H} \sim \mu$, there exists some voting classifier $f \in C(\mathcal{H})$ that has minimum margin θ

with ℓ over the entire set $\mathcal{X}$. In fact, the size of the hypothesis set does not depend on the size of $\mathcal{X}$, but only on the sparsity parameter d. More formally, we show the following.

Lemma 3. *For every $\theta \in (0, 1/40)$, $\delta \in (0,1)$ and integers $d \leq u$, there exists a distribution $\mu = \mu(u, d, \theta, \delta)$ over hypothesis sets $\mathcal{H} \subset \mathcal{X} \to \{-1, 1\}$, where $\mathcal{X}$ is a set of size u, such that the following holds.*

1. *For all $\mathcal{H} \in \mathrm{supp}(\mu)$, we have $|\mathcal{H}| = N$; and*
2. *For every labeling $\ell \in \{-1, +1\}^u$, if no more than d points $x \in \mathcal{X}$ satisfy $\ell(x) = -1$, then*

$$\Pr_{\mathcal{H} \sim \mu}[\exists f \in \mathcal{C}(\mathcal{H}) : \forall x \in \mathcal{X}.\ \ell(x) f(x) \geq \theta] \geq 1 - \delta\ ,$$

where $N = \Theta\left(\theta^{-2} \ln d \ln(\theta^{-2} d \delta^{-1}) e^{\Theta(\theta^2 d)}\right)$

In fact, we prove that if $\mathcal{H}$ is a random hypothesis set that also contains the hypothesis mapping all points to 1, then with good probability $\mathcal{H}$ satisfies the second requirement in the theorem.

To show the existence of a good voting classifier in $C(\mathcal{H})$ our proof actually employs a slight variant of the celebrated AdaBoost algorithm, and shows that with high probability (over the choice of the random hypothesis set $\mathcal{H}$), the voting classifier constructed by this algorithm attains minimum margin at least θ over the entire set $\mathcal{X}$.

Note that Lemma 3 speaks of a distribution over hypothesis sets. When using Lemma 3 in our proofs, we will invoke Yao's principle to conclude the existence of a suitable fixed hypothesis set $\mathcal{H}$.

Existential Lower Bound. Our proof of Theorem 2 uses many of the same ideas as the proof of Theorem 1. The difference between the generalization lower bound (second point) in Theorem 1 and 2 is an $\ln m$ factor in the first term inside the $\Omega(\cdot)$ notation. That is, Theorem 2 has an $\Omega(\frac{\ln |\mathcal{H}| \ln m}{\theta^2 m})$ where Theorem 1 has an $\Omega(\frac{\ln |\mathcal{H}|}{\theta^2 m})$. This term originated from having $\ln |\mathcal{H}|/\theta^2$ points with a probability mass of $1/10m$ in $\mathcal{D}_\ell$ and one point having the remaining probability mass. In the proof of Theorem 2, we first exploit that we are proving an existential lower bound by assigning all points the same label 1. That is, our hard distribution $\mathcal{D}$ assigns all points the label 1 (ignoring the second half of the distribution with the random and slightly biased labels). Since we are not proving a lower bound for every algorithm, this will not cause problems. We then change $|\mathcal{X}|$ to about $m/\ln m$ and assign each point the same probability mass $\ln m/m$ in distribution $\mathcal{D}$. The key observation is that on a random sample S of m points, by a coupon-collector argument, there will still be $m^{\Omega(1)}$ points from $\mathcal{X}$ that were not sampled. From Lemma 3, we can now find a voting classifier f, such that $\mathrm{sign}(f(x))$ is 1 on all points in $x \in S$, and -1 on a set of $d = \ln |\mathcal{H}|/\theta^2$ points in $\mathcal{X} \setminus S$. This means that f has out-of-sample error $\Omega(d \ln m/m) = \Omega(\frac{\ln |\mathcal{H}| \ln m}{\theta^2 m})$ under distribution $\mathcal{D}$ and obtains a margin of θ on all points in the sample S.

As in the proof Theorem 1, we can combine the above distribution $\mathcal{D}$ with the ideas of Anthony and Bartlett to add the terms depending on τ to the lower bound.

3 Margin-Based Generalization Lower Bounds

In this section we prove Theorems 1 and 2 assuming Lemma 3, whose proof is deferred to Section 4, and we start by describing the outlines of the proofs. To this end fix some integer N, and fix $\theta \in (1/N, 1/40)$. Let u be an integer, and let $\mathcal{X} = \{\xi_1, \ldots, \xi_u\}$ be some set with u elements. With every $\ell \in \{-1, 1\}^u$ we associate a distribution $\mathcal{D}_\ell$ over $\mathcal{X} \times \{-1, 1\}$, and show that with some constant probability over a random choice of ℓ, a voting classifier of interest has a high generalization probability with respect to $\mathcal{D}_\ell$. By a voting classifier of interest we mean one constructed by a learning algorithm in the proof of Theorem 1 and an adversarial classifier in the proof of Theorem 2. We additionally show existence of a hypothesis set $\hat{\mathcal{H}}$ such that with very high (constant) probability over a random choice of $\ell \in \{-1, 1\}^u$, $C(\hat{\mathcal{H}})$ contains a voting classifier that attains high margins with ℓ over the entire set $\mathcal{X}$. Finally, we conclude that with positive probability over a random choice of $\ell \in \{-1, 1\}^u$ both properties are satisfied, and therefore there exists at least one labeling ℓ that satisfies both properties.

We start by constructing the family $\{\mathcal{D}_\ell\}_{\ell\in\{-1,1\}^u}$ of distributions over $\mathcal{X}\times\{-1,1\}$. To this end, let $d\leq u$ be some constant to be fixed later, and let $\ell\in\{-1,1\}^u$. We define $\mathcal{D}_\ell$ separately for the first $u-d$ points and the last d points of $\mathcal{X}$. Intuitively, every point in $\{\xi_i\}_{i\in[u-d]}$ has a fixed label determined by ℓ, however all points but one have a very small probability of being sampled according to $\mathcal{D}_\ell$. Every point in $\{\xi_i\}_{i\in[u-d,u]}$, on the other hand, has an equal probability of being sampled, however its label is not fixed by ℓ rather than slightly biased towards ℓ. Formally, let $\alpha,\beta,\varepsilon\in[0,1]$ be constants to be fixed later. We construct $\mathcal{D}_\ell$ using the ideas described earlier in Section 2, by sewing them together over two parts of the set $\mathcal{X}$. We assign probability $1-\beta$ to $\{\xi_i\}_{i\in[u-d]}$ and β to $\{\xi_i\}_{i\in[u-d+1,u]}$. That is, for $(x,y)\sim\mathcal{D}_\ell$, the probability that $x\in\{\xi_i\}_{i\in[u-d]}$ is $1-\beta$. Next, conditioned on $x\in\{\xi_i\}_{i\in[u-d]}$, (ξ_1,ℓ_1) is assigned high probability $(1-\varepsilon)$ and the rest of the measure is distributed uniformly over $\{(\xi_i,\ell_i)\}_{i\in[2,u-d]}$. That is

$$\Pr_{\mathcal{D}_\ell}[(\xi_1,\ell_1)]=(1-\beta)(1-\varepsilon)\ ,\ and\ \forall j\in[2,u-d].\ \Pr_{\mathcal{D}_\ell}[(\xi_j,\ell_j)]=\frac{(1-\beta)\varepsilon}{u-d-1}\ .$$

Finally, conditioned on $x\in\{\xi_i\}_{i\in[u-d+1,u]}$, x distributes uniformly over $\{\xi_i\}_{i\in[u-d+1,u]}$, and conditioned on $x=\xi_i$, we have $y=\ell_i$ with probability $\frac{1+\alpha}{2}$. That is

$$\forall j\in[u-d+1,u].\ \Pr_{\mathcal{D}_\ell}[(\xi_j,\ell_j)]=\frac{(1+\alpha)\beta}{2d}\ ,and\ \Pr_{\mathcal{D}_\ell}[(\xi_j,-\ell_j)]=\frac{(1-\alpha)\beta}{2d}\ .$$

In order to give a lower bound on the generalization error for some classifier f of interest, we define new random variables such that their sum is upper bounded by $\Pr_{(x,y)\sim\mathcal{D}_\ell}[yf(x)<0]$, and give a lower bound on that sum. To this end, for every $\ell\in\{-1,1\}^u$ and $f:\mathcal{X}\to\mathbb{R}$, denote

$$\Psi_1(\ell,f)=\frac{(1-\varepsilon)\beta}{u-d-1}\sum_{i\in[2,u-d]}\mathbb{1}_{\ell_i f(\xi_i)<0}\quad;\quad\Psi_2(\ell,f)=\frac{\alpha\beta}{d}\sum_{i\in[u-d+1,u]}\mathbb{1}_{\ell_i f(\xi_i)<0}\ . \tag{3}$$

When f,ℓ are clear from the context we shall simply denote Ψ_1,Ψ_2. We show next that indeed proving a lower bound on $\Psi_1+\Psi_2$ implies a lower bound on the generalization error.

Claim 4. *For every ℓ,f we have* $\Pr_{(x,y)\sim\mathcal{D}_\ell}[yf(x)<0]\geq\frac{\beta(1-\alpha)}{2}+\Psi_1+\Psi_2$.

Before getting proving the claim, we explain why focusing on $\Psi_1+\Psi_2$, rather than bounding the generalization error directly is essential for the proof. The reason lies in the fact that we need a lower bound to hold *with constant probability* over the choice of ℓ and S (and in the case of Theorem 1 also the random choices made by the algorithm) and not only *in expectation*. While lower bounding $\mathbb{E}[\Pr_{(x,y)\sim\mathcal{D}_\ell}[yf(x)<0]]$ is clearly not harder than lower bounding $\mathbb{E}[\Psi_1+\Psi_2]$, showing that a lower bound holds with some constant probability is slightly more delicate. Our proof uses the fact that with probability 1, $\Psi_1+\Psi_2$ is not larger than a constant from its expectation, and therefore we can use Markov's inequality to lower bound $\Psi_1+\Psi_2$ with constant probability. We next turn to prove the claim.

Proof. We first observe that

$$\begin{aligned}\Pr_{(x,y)\sim\mathcal{D}_\ell}[yf(x)<0]&=\mathbb{E}_{(x,y)\sim\mathcal{D}_\ell}[\mathbb{1}_{yf(x)<0}]\\&=\sum_{i\in[u-d],y\in\{-1,1\}}\mathbb{1}_{yf(\xi_i)<0}\Pr_{\mathcal{D}_\ell}[(\xi_i,y)]+\sum_{i\in[u-d+1,u],y\in\{-1,1\}}\mathbb{1}_{yf(\xi_i)<0}\Pr_{\mathcal{D}_\ell}[(\xi_i,y)]\end{aligned} \tag{4}$$

For every $i\in[u-d]$ and $y\in\{-1,1\}$, if $y\neq\ell_i$ then $\Pr_{\mathcal{D}_y}[(\xi_j,y)]=0$. Moreover, if $i\geq 2$ and $y=\ell_i$ then $\Pr_{\mathcal{D}_y}[(\xi_i,y)]=\frac{(1-\beta)\varepsilon}{u-d-1}$. Therefore

$$\sum_{j\in[u-d],y\in\{-1,1\}}\mathbb{1}_{yf(\xi_j)<0}\Pr_{\mathcal{D}_y}[(\xi_j,y)]\geq\frac{(1-\beta)\varepsilon}{u-d-1}\sum_{j\in[2,u-d]}\mathbb{1}_{yf(\xi_j)<0}=\Psi_1\ . \tag{5}$$

Next, for every $i\in[u-d+1,u]$ we have that

$$\begin{aligned}\sum_{y\in\{-1,1\}}\mathbb{1}_{yf(\xi_i)<0}\Pr_{\mathcal{D}_\ell}[(\xi_i,y)]&=\mathbb{1}_{\ell_i f(\xi_i)<0}\Pr_{\mathcal{D}_\ell}[(\xi_i,\ell_i)]+\mathbb{1}_{\ell_i f(\xi_i)>0}\Pr_{\mathcal{D}_\ell}[(\xi_i,-\ell_i)]\\&=\frac{(1-\alpha)\beta}{2d}+\mathbb{1}_{\ell_i f(\xi_i)<0}\frac{\alpha\beta}{d}\ ,\end{aligned}$$

and therefore

$$\sum_{i\in[u-d+1,u],y\in\{-1,1\}} \mathbb{1}_{yf(\xi_i)<0} \Pr_{\mathcal{D}_\ell}[(\xi_i,y)] = \frac{(1-\alpha)\beta}{2} + \frac{\alpha\beta}{d} \sum_{i\in[u-d+1,u]} \mathbb{1}_{\ell_i f(\xi_i)<0} \,. \tag{6}$$

Plugging (5) and (6) into (4) we conclude the claim. □

To prove existence of a "rich" yet small enough hypothesis set $\hat{\mathcal{H}}$ we apply Lemma 3 together with Yao's minimax principle. In order to ensure that the hypothesis sets constructed using Lemma 3 is small enough, and specifically has size $N^{O(1)}$, we need to focus our attention on sparse labelings $\ell \in \{-1,1\}^u$ only. That is, the labelings cannot contain more than $\Theta\left(\frac{\ln N}{\theta^2}\right)$. To this end we will focus on $2d$-sparse vectors, and more specifically, a designated set of $2d$-sparse labelings. More formally, we define a set of labelings of interest $\mathcal{L}(u,d)$ as the set of all labelings $\ell \in \{-1,1\}^u$ such that the restriction to the first $u-d$ entries is d-sparse. That is

$$\mathcal{L}(u,d) := \{\ell \in \{-1,1\}^u : |\{i \in [u-d] : \ell_i = -1\}| \leq d\} \,. \tag{7}$$

We next show that there exists a small enough (with respect to N) hypothesis set $\hat{\mathcal{H}}$ that is rich enough. That is, with high probability over $\ell \in \mathcal{L}(u,d)$, there exists a voting classifier $f \in C(\hat{\mathcal{H}})$ that attains high minimum margin with ℓ over the entire set $\mathcal{X}$. Note that the following result, similarly to Lemma 3 does not depend on the size of $\mathcal{X}$, but only on the sparsity of the labelings in question.

Claim 5. *If $d \leq \frac{\ln N}{\theta^2}$ then there exists a hypothesis set $\hat{\mathcal{H}}$ such that $\ln|\hat{\mathcal{H}}| = \Theta(\ln N)$ and*

$$\Pr_{\ell\in_R\mathcal{L}(u,d)}[\exists f \in \mathcal{C}(\hat{\mathcal{H}}) : \forall i \in [u].\ \ell_i f(\xi_i) \geq \theta] \geq 1 - 1/N \,.$$

Proof. Let $\mu = \mu(u,d,\theta,1/N)$, be the distribution whose existence is guaranteed in Lemma 3. Then for every labeling $\ell \in \mathcal{L}(u,d)$, with probability at least $99/100$ over $\mathcal{H} \sim \mu$, there exists a voting classifier $f \in C(\mathcal{H})$ that has minimal margin of θ. That is, for every $i \in [u]$, $\ell_i f(\xi_i) \geq \theta$. By Yao's minimax principle, there exists a hypothesis set $\hat{\mathcal{H}} \in \mathrm{supp}(\mu)$ such that

$$\Pr_{\ell\in_R\mathcal{L}(u,d)}[\exists f \in \mathcal{C}(\hat{\mathcal{H}}) : \forall i \in [u].\ \ell_i f(x_i) \geq \theta] \geq 1 - 1/N \,.$$

Moreover, since $\hat{\mathcal{H}} \in \mathrm{supp}(\mu)$, then $|\hat{\mathcal{H}}| = \Theta\left(\theta^{-2}\ln d \cdot \ln(N\theta^{-2}\ln d)\cdot e^{\Theta(\theta^2 d)}\right)$. Since $\theta \geq 1/N$ and since $d = \frac{\ln N}{\theta^2}$ and thus $e^{\theta^2 d} = N$ we get that there exists some univeral constant $C > 0$ such that $|\hat{\mathcal{H}}| = \Theta(N^C)$, and thus $\ln|\hat{\mathcal{H}}| = \Theta(\ln N)$. □

3.1 Proof Algorithmic Lower Bound

This section is devoted to the proof of Theorem 1. That is, we show that for every algorithm $\mathcal{A}$, there exist some distribution $\mathcal{D} \in \{\mathcal{D}_\ell\}_{\ell\in\{-1,1\}^u}$ and some classifier $\hat{f} \in C(\hat{\mathcal{H}})$ such that with constant probability over $S \sim \mathcal{D}^m$, $\hat{f}$ has large margins on points in S, yet $f_{\mathcal{A},S}$ has large generalization error. To this end we now fix u to be $\frac{2\ln N}{\theta^2}$ and $d = \frac{u}{2} = \frac{\ln N}{\theta^2}$. For these values of u, d we get that $\mathcal{L}(u,d)$ is, in fact, the set of all possible labelings, i.e. $\mathcal{L}(u,d) = \{-1,1\}^u$. Next, fix $\mathcal{A}$ be a (perhaps randomized) learning algorithm. For every m-point sample S and recall that $f_{\mathcal{A},S}$ denotes the classifier returned by $\mathcal{A}$ when running on sample S.

The main challenge is to show that there exists a labeling $\hat{\ell} \in \{-1,1\}^u$ such that $C(\hat{\mathcal{H}})$ contains a good voting classifier for $\hat{\ell}$ and, in addition, $f_{\mathcal{A},S}$ has a large generalization error with respect to $\mathcal{D}_{\hat{\ell}}$. We will show that if α is small enough, then indeed such a labeling exists. Formally, we show the following.

Lemma 6. *If $\alpha \leq \sqrt{\frac{u}{40\beta m}}$, then there exists $\hat{\ell} \in \{-1,1\}^u$ such that*

1. *There exists $\hat{f} = \hat{f}_{\hat{\ell}} \in \mathcal{C}(\hat{\mathcal{H}})$ such that for every $i \in [u]$, $\hat{\ell}_i \hat{f}(\xi_i) \geq \theta$; and*
2. *with probability at least $1/25$ over $S \sim \mathcal{D}_{\hat{\ell}}^m$ and the randomness of $\mathcal{A}$ we have*

$$\Pr_{(x,y)\sim\mathcal{D}_{\hat{\ell}}}[y f_{\mathcal{A},S}(x) < 0] \geq \frac{(1-\alpha)\beta}{2} + \frac{(1-\beta)\varepsilon}{24} + \frac{\alpha\beta}{24} \,.$$

Before proving the lemma, we first show how it implies Theorem 1

Proof of Theorem 1. Fix some $\tau \in [0, 49/100]$. Assume first that $\tau \leq \frac{u}{300m}$, and let $\varepsilon = \frac{u}{10m}$ and $\beta = \alpha = 0$. Let $\hat{\ell}, \hat{f}$ be as in Lemma 6, then for every sample $S \sim \mathcal{D}_{\hat{\ell}}^m$, $\Pr_{(x,y)\sim S}[y\hat{f}(x) < \theta] = 0 \leq \tau$, and moreover with probability at least $1/25$ over S and the randomness of $\mathcal{A}$

$$\Pr_{(x,y)\sim\mathcal{D}_{\hat{\ell}}}[yf_{\mathcal{A},S}(x) < 0] \geq \frac{(1-\beta)\varepsilon}{24} \geq \tau + \Omega\left(\frac{u}{m}\right) = \tau + \Omega\left(\frac{\ln|\hat{\mathcal{H}}|}{m\theta^2} + \sqrt{\frac{\tau\ln|\hat{\mathcal{H}}|}{m\theta^2}}\right) .$$

where the last transition is due to the fact that $u = 2\theta^{-2}\ln N = \Theta(\theta^{-2}\lg|\hat{\mathcal{H}}|)$ and $\tau = O(u/m)$.

Otherwise, assume $\tau > \frac{u}{300m}$, and let $\varepsilon = \frac{u}{10m}$, $\alpha = \sqrt{\frac{u}{2560\tau m}}$ and $\beta = \frac{64\tau}{32-31\alpha}$. Since $\tau \geq \frac{u}{300m}$, then $\alpha \in [0,1]$. Moreover, if $m > Cu$ for large enough but universal constant $C > 0$, then $32 - 31\alpha \geq 64 \cdot \frac{49}{100} \geq 64\tau$, and hence $\beta \in [0,1]$. Moreover, since $\alpha \leq 1$ then $\beta \leq 64\tau$, and therefore $\alpha = \sqrt{\frac{u}{2560\tau m}} \leq \sqrt{\frac{u}{40\beta m}}$. Let therefore $\hat{\ell}, \hat{f}$ be a labeling and a classifier in $C(\hat{\mathcal{H}})$ whose existence is guaranteed in Lemma 6. Let $\langle(x_1,y_1),\ldots,(x_m,y_m)\rangle \sim \mathcal{D}_{\hat{y}}^m$ be a sample of m points drawn independently according to $\mathcal{D}_{\hat{\ell}}$. For every $j \in [m]$, we have $\mathbb{E}[\mathbb{1}_{y_j\hat{f}(x_j)<\theta}] = \frac{(1-\alpha)\beta}{2}$. Therefore by Chernoff we get that for large enough N,

$$\begin{aligned}\Pr_{S\sim\mathcal{D}_{\hat{\ell}}^m}\left[\Pr_{(x,y)\sim S}\left[y\hat{f}(x) < \theta\right] \geq \tau\right] &= \Pr_{S\sim\mathcal{D}_{\hat{\ell}}^m}\left[\frac{1}{m}\sum_{j\in[m]}\mathbb{1}_{\hat{y}_j\hat{f}(x_j)<\theta} \geq \frac{(1-31\alpha/32)\beta}{2}\right] \\ &\leq e^{-\Theta\left(\alpha^2\beta m\right)} \leq e^{-\Theta(u)} \leq 10^{-3} ,\end{aligned}$$

where the inequality before last is due to the fact that $\alpha^2\beta m = \frac{u\beta}{2560\tau} = \Omega(u)$, since $\beta \geq 2\tau$. Moreover, by Lemma 6 we get that with probability at least $1/25$ over S and $\mathcal{A}$ we get that

$$\begin{aligned}\Pr_{(x,y)\sim\mathcal{D}_{\hat{\ell}}}[yf_{\mathcal{A},S}(x) < 0] &\geq \frac{(1-\alpha)\beta}{2} + \frac{\alpha\beta}{32} = \frac{(1-31\alpha/32)\beta}{2} + \frac{\alpha\beta}{64} = \tau + \Omega\left(\sqrt{\frac{\tau u}{m}}\right) \\ &\geq \tau + \Omega\left(\frac{\ln|\hat{\mathcal{H}}|}{m\theta^2} + \sqrt{\frac{\tau\ln|\hat{\mathcal{H}}|}{m\theta^2}}\right) ,\end{aligned}$$

where the last transition is due to the fact that $\tau = \Omega(u/m)$. This completes the proof of Theorem 1. □

For the rest of the section we therefore prove Lemma 6. We start by lower bounding the expected value of $\Psi_1 + \Psi_2$, where the expectation is over the choice of labeling $\ell \in \{-1,1\}^u$, $S \sim \mathcal{D}_\ell^m$ and the random choices made by $\mathcal{A}$. Intuitively, as points in $\{\xi_2,\ldots,\xi_u\}$ are sampled with very small probability, it is very likely that the sample S does not contain many of them, and therefore the algorithm cannot do better than randomly guessing many of the labels. Moreover, if α is small enough, and S does not sample a point in $\{\xi_{u/2+1},\ldots,\xi_u\}$ enough times, there is a larger probability that $\mathcal{A}$ does not determine the bias correctly.

Claim 7. *If* $\alpha \leq \sqrt{\frac{u}{40\beta m}}$, *then* $\mathbb{E}_{\ell\in\{-1,1\}^u}\left[\mathbb{E}_{\mathcal{A},S}\left[\Psi_1(\ell, f_{\mathcal{A},S}) + \Psi_2(\ell, f_{\mathcal{A},S})\right]\right] \geq \frac{(1-\beta)\varepsilon}{6} + \frac{\alpha\beta}{6}$.

Proof. To lower bound the expectation, we lower bound the expectations of Ψ_1 and Ψ_2 separately. For every $i \in [2, u-d] \setminus \{1\}$, if $\xi_i \notin S$ then ℓ_i and $f_{\mathcal{A},S}(\xi_i)$ are independent, and therefore $\mathbb{E}_\ell[\mathbb{1}_{\ell_i f_{\mathcal{A},S}(\xi_i)<0}] = \frac{1}{2}$. Let $\mathcal{S}$ be the set of all samples for which $|S \cap \{\xi_2,\ldots,\xi_{u-d}\}| \leq \frac{u-d-1}{2}$, then for every $S \in \mathcal{S}$,

$$\mathbb{E}_\ell\left[\sum_{i\in[2,u-d-1]}\mathbb{1}_{\ell_i f_{\mathcal{A},S}(\xi_i)<0}\right] \geq \frac{u-d-1-|S\cap\{\xi_2,\ldots,\xi_{u-d}\}|}{2} \geq \frac{u-d-1}{4} ,$$

As this holds for every $S \in \mathcal{S}$, we conclude that

$$\mathbb{E}_{\mathcal{A},S}\left[\mathbb{E}_{\ell}\left[\Psi_1(\ell, f_{\mathcal{A},S})\right] \mid S \in \mathcal{S}\right] \geq \frac{(1-\beta)\varepsilon}{u-d-1} \cdot \frac{u-d-1}{4} = \frac{(1-\beta)\varepsilon}{4} .$$

Next, for large enough N a Chernoff bound gives $\Pr_{S\sim\mathcal{D}^m}[\mathcal{S}] \geq 1 - e^{-\Theta(u)} \geq 2/3$, and therefore $\mathbb{E}_{\mathcal{A},S}\left[\mathbb{E}_{\ell}\left[\Psi_1(\ell, f_{\mathcal{A},S})\right]\right] \geq \frac{(1-\beta)\varepsilon}{6}$, and by Fubini's theorem $\mathbb{E}_{\ell}\left[\mathbb{E}_{\mathcal{A},S}[\Psi_1(\ell, f_{\mathcal{A},S})]\right] \geq \frac{(1-\beta)\varepsilon}{6}$.

Next, let $i \in [u-d+1, u]$. Denote by $\sigma_i \in [m]$ the number of times ξ_i was sampled into S. Then

$$\mathbb{E}_{\ell}\left[\mathbb{E}_{\mathcal{A},S}\left[\mathbb{1}_{\ell_i f_{\mathcal{A},S}(\xi_i)<0}\right]\right] = \sum_{n=0}^{m} \mathbb{E}_{\ell}\left[\mathbb{E}_{\mathcal{A},S}\left[\mathbb{1}_{\ell_i f_{\mathcal{A},S}(\xi_i)<0} \middle| \sigma_i = n\right]\right] \cdot \Pr[\sigma_i = n] \tag{8}$$

For every $x > 0$ and $y \in (0,1)$, let $\Phi(x,y) = \frac{1}{4}\left(1 - \sqrt{1 - \exp\left(\frac{-xy^2}{1-y^2}\right)}\right)$, then a result by Anthony and Bartlett [AB09, Lemma 5.1] shows that

$$\mathbb{E}_{\ell}\left[\mathbb{E}_{\mathcal{A},S}\left[\mathbb{1}_{\ell_i f_{\mathcal{A},S}(\xi_i)<0} \middle| \sigma_i = n\right]\right] \geq \Phi(n+2, \alpha)$$

Plugging this into (8), by the convexity of $\Phi(\cdot, \alpha)$ and Jensen's inequality we get that

$$\mathbb{E}_{\ell}\left[\mathbb{E}_{\mathcal{A},S}\left[\mathbb{1}_{\ell_i f_{\mathcal{A},S}(\xi_i)<0}\right]\right] \geq \sum_{n=0}^{m} \Phi(n+2,\alpha) \cdot \Pr[\sigma_i = n] \geq \Phi(\mathbb{E}[\sigma_i] + 2, \alpha) .$$

Since $\mathbb{E}[\sigma_i] = \frac{2\beta m}{u}$, and Since $\Phi(\cdot, \alpha)$ is monotonically decreasing we get that

$$\mathbb{E}_{\ell}\left[\mathbb{E}_{\mathcal{A},S}\left[\mathbb{1}_{\ell_i f_{\mathcal{A},S}(\xi_i)<0}\right]\right] \geq \Phi\left(\frac{4\beta m}{u}, \alpha\right) .$$

Summing over all $i \in [u-d+1, u]$ we get that $\mathbb{E}_{\ell}\left[\mathbb{E}_{\mathcal{A},S}[\Psi_2(\ell, f_{\mathcal{A},S})]\right] \geq \alpha\beta\Phi\left(\frac{4\beta m}{u}, \alpha\right)$. The claim then follows from the fact that for every $\alpha \leq \sqrt{\frac{u}{40\beta m}}$ we have $\Phi(\frac{8\beta m}{u}, \alpha) \geq \frac{1}{6}$. □

We next show that for small values of α, a large fraction of labelings $\ell \in \{-1,1\}^u$ satisfy that $\Psi_1 + \Psi_2$ is large with some positive constant probability over the random choices of $\mathcal{A}$ and the choice of $S \in \mathcal{S}$.

Claim 8. *If $\alpha \leq \sqrt{\frac{u}{40\beta m}}$, then with probability at least $1/11$ over the choice of $\ell \in \{-1,1\}^u$ we have*

$$\Pr_{\mathcal{A},S}\left[\Psi_1(\ell, f_{\mathcal{A},S}) + \Psi_2(\ell, f_{\mathcal{A},S}) \geq \frac{(1-\beta)\varepsilon}{24} + \frac{\alpha\beta}{24}\right] \geq \frac{1}{25} .$$

Proof. First note that by substituting every indicator in (3) with 1 we get that with probability 1 over all samples S, labelings ℓ and random choices of $\mathcal{A}$ we have

$$\Psi_1 + \Psi_2 \leq (1-\beta)\varepsilon + \alpha\beta , \tag{9}$$

and therefore $\Pr_{\ell}\left[\mathbb{E}_{\mathcal{A},S}[\Psi_1 + \Psi_2] \leq (1-\beta)\varepsilon + \alpha\beta\right] = 1$. Furthermore, for every $\alpha \leq \sqrt{\frac{u}{40\beta m}}$ we get from Claim 7 that $\mathbb{E}_{\ell}\left[\mathbb{E}_{\mathcal{A},S}[\Psi_1 + \Psi_2]\right] \geq \frac{1}{6}\left((1-\beta)\varepsilon + \alpha\beta\right)$. Denote $X = \mathbb{E}_{\mathcal{A},S}[\Psi_1 + \Psi_2]$ and $a = (1-\beta)\varepsilon + \alpha\beta$. In these notations we have that (9) states that $\Pr_{\ell}[X \leq a] = 1$, and Claim 7 states that $\mathbb{E}_{\ell}[X] \geq a/6$. Therefore $a - X$ is a non-negative random variable, and from Markov's inequality we get that

$$\Pr_{\ell}[X \leq a/12] = \Pr_{\ell}[a - X \geq 11a/12] \leq \Pr_{\ell}[a - X \geq 1.1\mathbb{E}[a - X]] \leq 10/11$$

and therefore $\Pr_{\ell}[\mathbb{E}_{\mathcal{A},S}[\Psi_1 + \Psi_2] \geq \frac{1}{12}((1-\beta)\varepsilon + \alpha\beta)] \geq 1/11$.

Next, fix some $\ell \in \{-1,1\}^u$ for which $\mathbb{E}_{\mathcal{A},S}[\Psi_1 + \Psi_2] \geq \frac{1}{12}((1-\beta)\varepsilon + \alpha\beta)$. Once again, as $\Pr_{\mathcal{A},S}[\Psi_1 + \Psi_2 \leq 12\mathbb{E}_{\mathcal{A},S}[\Psi_1 + \Psi_2]] = 1$ we get from Markov's inequality that with probability at least $1/25$ we have

$$\Pr_{\mathcal{A},S}\left[\Psi_1 + \Psi_2 \geq \frac{(1-\varepsilon)\beta}{24} + \frac{\alpha\beta}{24}\right] \geq \Pr_{\mathcal{A},S}\left[\Psi_1 + \Psi_2 \geq \frac{1}{2}\mathbb{E}_{\mathcal{A},S}[\Psi_1 + \Psi_2]\right] \geq \frac{1}{25} .$$

□

To finish the proof of Lemma 6, observe that from Claims 5 and 8 we get that with positive probability over $\ell \in \{-1,1\}$ there exists a voting classifier $f \in C(\hat{\mathcal{H}})$ such that for every $i \in [u]$, $\ell_i f(x_i) \geq \theta$ and in addition $\Pr_{\mathcal{A},S}\left[\Psi_1 + \Psi_2 \geq \frac{(1-\varepsilon)\beta}{24} + \frac{\alpha\beta}{24}\right] \geq \frac{1}{25}$. As this occurs with positve probability, we conclude that there exists some labeling $\hat{\ell} \in \{-1,1\}^u$ satisfying both properties. Since for every set of random choices of $\mathcal{A}$, and every $S \sim \mathcal{D}_{\hat{\ell}}^m$, Claim 4 guarantees that

$$\Pr_{(x,y)\sim\mathcal{D}_{\hat{\ell}}}[yf_{\mathcal{A},S}(x)] \geq \frac{(1-\alpha)\beta}{2} + \Psi_1(\hat{\ell}, f_{\mathcal{A},S}) + \Psi_2(\hat{\ell}, f_{\mathcal{A},S}) \ ,$$

this concludes the proof of Lemma 6, and thus the proof of Theorem 1 is now complete.

3.2 Proof of Existential Lower Bound

This section is devoted to the proof of Theorem 2. That is, we show the existence of a distribution $\mathcal{D} \in \{\mathcal{D}_\ell\}_{\ell\in\{-1,1\}^u}$ such that with a constant probability over $S \sim \mathcal{D}^m$ there exists some voting classifier $f_S \in C(\hat{\mathcal{H}})$ such that f_S has large margins on points in S, but has large generalization probability with respect to $\mathcal{D}$. To this end, let m be such that $\frac{\ln N}{\theta^2} < \left(\frac{m}{\lg m}\right)^{9/10}$, and note that $m = \left(\frac{\ln N}{\theta^2}\right)^{1+\Omega(1)}$. Let $u = \frac{40m}{\lg m}$, and let $d = \frac{\ln N}{\theta^2}$.

Similarly to the proof of Theorem 1, the main challenge is to show the existence of a lebeling that satisfies all desired properties. We draw the reader's attention to the fact that unlike the previous proof, the distribution over labelings is not uniform over the entire set $\{-1,1\}^u$, but rather a designated subset of sparse labelings.

With every labeling $\ell \in \{-1,1\}^u$ and an m-point sample S, we associate a classifier $h_{\ell,S}$ as follows. Intuitively, $h_{\ell,S}$ "adverserially changes" at most d labels of points in $\{\xi_2,\ldots,\xi_{u-d}\}$ that were not picked by S, and chooses the majority label for points in $\{\xi_{u-d+1},\ldots,\xi_u\}$. Formally, let $\mathcal{I}_S \subseteq \{\xi_2,\ldots,\xi_{u-d}\} \setminus S$ be an arbitrary sets of size at most d, then for every $x \in \{\xi_1,\ldots,\xi_{u-d}\}$, $h_{\ell,S}(x) = -\ell(x)$ if and only if $x \in \mathcal{I}_S$, and for every $x \in \{\xi_{u-d+1},\ldots,\xi_u\}$, $h_{\ell,S}(x)$ is the majority of labels of x in S. That is $h_{\ell,S}(x) = 1$ if and only if $(x,1)$ appears in S more times than $(x,-1)$. Break ties arbitratily.

Lemma 9. *If $\alpha \leq \sqrt{\frac{d}{40\beta m}}$ then there exists $\hat{\ell} \in \{-1,1\}^u$ such that*

1. *For every $i \in [u-d]$, $\hat{\ell}_i = 1$;*
2. *With probability at least $99/100$ over the choice of sample $S \sim \mathcal{D}_{\hat{\ell}}^m$, there exists a voting classifier $f_S \in C(\hat{\mathcal{H}})$ such that $f_S(\xi_i)h_{\hat{\ell},S}(\xi_i) \geq \theta$ for all $i \in [u]$; and*
3. *with probability at least $1/25$ over $S \sim \mathcal{D}_{\hat{\ell}}^m$ we have*

$$\Pr_{(x,y)\sim\mathcal{D}_{\hat{\ell}}}[yh_{\hat{\ell},S}(x) < 0] \geq \frac{(1-\alpha)\beta}{2} + \frac{(1-\beta)\varepsilon d}{8(u-d-1)} + \frac{\alpha\beta}{24} \ .$$

We first show that the lemma implies Theorem 2.

Proof of Theorem 2. Fix some $\tau \in [0, 49/100]$. Assume first that $\tau \leq \frac{d}{50u}$, and let $\varepsilon = \frac{1}{2}$ and $\beta = \alpha = 0$. With probability $1/25$ over S we have

$$\Pr_{(x,y)\sim\mathcal{D}_{\hat{\ell}}}[yh_{\hat{\ell},S}(x) < 0] \geq \frac{(1-\beta)\varepsilon d}{8u} \geq \tau + \Omega\left(\frac{d}{u}\right) = \tau + \Omega\left(\frac{\ln|\hat{\mathcal{H}}|\ln m}{m\theta^2} + \sqrt{\frac{\tau\ln|\hat{\mathcal{H}}|\ln m}{m\theta^2}}\right) \ ,$$

where the last transition is due to the fact that $d = \theta^{-2}\ln N = \Theta(\theta^{-2}\ln|\hat{\mathcal{H}}|)$ and $\tau = O(d/u)$. Moreover, with probability $99/100$ over S there exists $f_S \in C(\hat{\mathcal{H}})$ such that $f_S(\xi_i)h_{\hat{\ell},S}(\xi_i) \geq \theta$ for all $i \in [u]$. We get that with probability at least $1/100$ over the sample S there exists $f_S \in C(\hat{\mathcal{H}})$ such that

$$\Pr_S[y_j f_S(x_j) < \theta] = \Pr_S[y_j h_{\hat{\ell},S}(x_j) < 0] = 0 \leq \tau \ ,$$

and moreover

$$\Pr_{(x,y)\sim\mathcal{D}_{\hat{\ell}}}[yf_S(x)<0] = \Pr_{(x,y)\sim\mathcal{D}_{\hat{\ell}}}[yh_{\hat{\ell},S}(x)<0] \geq \tau + \Omega\left(\frac{\ln|\hat{\mathcal{H}}|\ln m}{m\theta^2} + \sqrt{\frac{\tau\ln|\hat{\mathcal{H}}|\ln m}{m\theta^2}}\right) .$$

Otherwise, assume $\tau > \frac{d}{50u}$, and let $\varepsilon = \frac{1}{2}$, $\alpha = \sqrt{\frac{d}{2560\tau m}}$ and $\beta = \frac{64\tau}{32-31\alpha}$. Since $\tau \geq \frac{d}{50u}$, then $\alpha \in [0,1]$. Moreover, for large enough constant $C > 0$, if $m > Cd$, then $32 - 31\alpha \geq 64 \cdot \frac{499}{1000} \geq 64 \cdot \frac{101}{100}\tau$, and therefore $\beta \in [0, 100/101]$.

Next, let $\langle (x_1,y_1),\ldots,(x_m,y_m)\rangle \sim \mathcal{D}_{\hat{\ell}}^m$ be a sample of m points drawn independently according to $\mathcal{D}_{\hat{\ell}}$. For every $j \in [m]$, let $\mathcal{E}_j$ be the event that $(x_j,y_j) \in \{(\xi_i, -\hat{\ell}_i)\}_{i\in[u-d+1,u]}$, then we have $\mathbb{1}_{y_j f_S(x_j)<0} < \mathbb{1}_{\mathcal{E}_j}$. Moreover, $\mathbb{E}[\mathbb{1}_{\mathcal{E}_j}] = \frac{(1-\alpha)\beta}{2}$, and $\{\mathbb{1}_{\mathcal{E}_j}\}_{j\in[m]}$ are independent. Therefore by Chernoff we get that for large enough N,

$$\begin{aligned}\Pr_{S\sim\mathcal{D}_{\hat{\ell}}^m}\left[\Pr_{(x,y)\sim S}\left[yh_{\hat{\ell},S}(x)<0\right]\geq\tau\right] &\leq \Pr_{S\sim\mathcal{D}_{\hat{\ell}}^m}\left[\frac{1}{m}\sum_{j\in[m]}\mathbb{1}_{\mathcal{E}_j} \geq \frac{(1-31\alpha/32)\beta}{2}\right] \\ &\leq e^{-\Theta\left(\alpha^2\beta m\right)} = e^{-\Theta(d)} \leq 10^{-3} ,\end{aligned}$$

where the inequality before last is due to the fact that $\alpha^2\beta m = \frac{d\beta}{2560\tau} = \Omega(d)$, since $\beta \geq 2\tau$.

Moreover, since $\alpha \leq 1$ then $\beta \leq 64\tau$, and therefore $\alpha = \sqrt{\frac{d}{2560\tau m}} \leq \sqrt{\frac{d}{40\beta m}}$. Thus with probability at least $1/25$ over S we get that

$$\begin{aligned}\Pr_{(x,y)\sim\mathcal{D}_{\hat{\ell}}}[yh_{\hat{\ell},S}(x)<0] &\geq \frac{(1-\alpha)\beta}{2} + \frac{(1-\beta)\varepsilon d}{u-d-1} + \frac{\alpha\beta}{32} = \frac{(1-31\alpha/32)\beta}{2} + \frac{(1-\beta)\varepsilon d}{u-d-1} + \frac{\alpha\beta}{64} \\ &= \tau + \Omega\left(\frac{d}{u} + \sqrt{\frac{\tau d}{m}}\right) \geq \tau + \Omega\left(\frac{\ln|\hat{\mathcal{H}}|\ln m}{m\theta^2} + \sqrt{\frac{\tau\ln|\hat{\mathcal{H}}|}{m\theta^2}}\right) ,\end{aligned}$$

Therefore with probability at least $1/50$ over the sample S we get that $\Pr_{(x,y)\sim S}\left[yh_{\hat{\ell},S}(x)<0\right] \leq \tau$ and moreover

$$\Pr_{(x,y)\sim\mathcal{D}_{\hat{\ell}}}[yh_{\hat{\ell},S}(x)<0] \geq \tau + \Omega\left(\frac{\ln|\hat{\mathcal{H}}|\ln m}{m\theta^2} + \sqrt{\frac{\tau\ln|\hat{\mathcal{H}}|}{m\theta^2}}\right) .$$

Finally, from Lemma 9 and similarly to the first part of the proof, we get that with probability $1/100$ over the choice of S there exists $f_S \in C(\hat{\mathcal{H}})$ such that $h_{\hat{\ell},S}(\xi_i)f_S(\xi_i) \geq \theta$ for all $i \in [u]$. For all these samples S we get that $\Pr_{(x,y)\sim S}[yf_S(x)<\theta] = \Pr_{(x,y)\sim S}\left[yh_{\hat{\ell},S}(x)<0\right] \leq \tau$ and moreover

$$\Pr_{(x,y)\sim\mathcal{D}_{\hat{\ell}}}[yf_S(x)<0] = \Pr_{(x,y)\sim\mathcal{D}_{\hat{\ell}}}[yh_{\hat{\ell},S}(x)<0] \geq \tau + \Omega\left(\frac{\ln|\hat{\mathcal{H}}|\ln m}{m\theta^2} + \sqrt{\frac{\tau\ln|\hat{\mathcal{H}}|}{m\theta^2}}\right) .$$

□

For the rest of the section we therefore prove Lemma 9. As with the proof of Lemma 6, we start by lower bounding the expected value of $\Psi_1(\ell, h_{\ell,S}) + \Psi_2(\ell, h_{\ell,S})$ over a choice of a labeling ℓ and samples $S \in \mathcal{D}_\ell$. We consider next the subset $\mathcal{L}'$ of $\mathcal{L}(u,d)$ containing all labelings ℓ satisfying $\ell_i = 1$ for all $i \in [u]$. Intuitively, by a coupon-collector like argument we show that with very high probability over the sample S, there are at least d points in $\{\xi_i\}_{i\in[u-d]}$ not sampled into S. The argument lower bounding Ψ_2 is identical to the one in the proof of Lemma 9.

Claim 10. *If* $\alpha \leq \sqrt{\frac{d}{40\beta m}}$ *then*

$$\mathbb{E}_{\ell \in \mathcal{L}'}\left[\mathbb{E}_S\left[\Psi_1(\ell, h_{\ell,S}) + \Psi_2(\ell, h_{\ell,S})\right]\right] \geq \frac{(1-\varepsilon)\beta d}{2(u-d-1)} + \frac{\alpha\beta}{6} .$$

Proof. Let $\mathcal{S}$ be the set of all m-point samples S for which $|\{\xi_2, \ldots, \xi_{u-d}\} \setminus S| \geq d$. For every $S \in \mathcal{S}$ we have $|\mathcal{I}_S| = d$, and therefore

$$\sum_{i \in [2, u-d]} \mathbb{1}_{\ell_i f_S(\xi_i) < 0} = \sum_{i \in [2, u-d]} \mathbb{1}_{f_S(\xi_i) < 0} = |\mathcal{I}_S| = d .$$

Therefore $\mathbb{E}_\ell[\mathbb{E}_S[\Psi_1(\ell, f_S) | S \in \mathcal{S}]] = \frac{(1-\varepsilon)\beta d}{u-d-1}$. We will show next that $\Pr_S[\mathcal{S}] \geq 1/2$, and conclude that $\mathbb{E}_\ell[\mathbb{E}_S[\Psi_1(\ell, f_S)]] \geq \frac{(1-\varepsilon)\beta d}{2(u-d-1)}$. To see this, consider a random sampling $S \sim \mathcal{D}_\ell^m$. We will show by a coupon-collector argument that with high probability, no more than $(u-d-1)-d$ elements of $\{\xi_2, \ldots, \xi_{u-d}\}$ are sampled to S, and therefore $S \in \mathcal{S}$. Consider the set of elements of $\{\xi_2, \ldots, \xi_{u-d}\}$ sampled by S. For every $k \in [u-2d-1]$, let X_k be the number of samples between the time $(k-1)$th distinct element was sampled from $\{\xi_2, \ldots, \xi_{u-d}\}$ and the time the kth distinct element was sampled from $\{\xi_2, \ldots, \xi_{u-d}\}$. Then $X_k \sim Geom\,(p_k)$, where $p_k = (1-\beta)\varepsilon \cdot \frac{u-d-k}{u-d-1}$. Denote $X := \sum_{k \in [u-2d-1]} X_k$, then

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k \in [u-2d-1]} \frac{1}{p_k} = \sum_{k \in [u-2d-1]} \frac{u-d}{(1-\beta)\varepsilon(u-d-k)} = \frac{u-d-1}{(1-\beta)\varepsilon} \sum_{k=d+1}^{u-d-1} \frac{1}{k} \\
&\geq (u-d-1)[\ln(u-d-1) - \ln(d+1) - 1] \geq \frac{1}{2} u \ln\frac{u}{d} \geq \frac{1}{20} u \ln u \geq \frac{4}{3} m
\end{aligned}$$

Therefore by letting $\lambda = \frac{3}{4}$, and $p_* = \min_{k \in [u-2d-1]} p_k = (1-\beta)\varepsilon \cdot \frac{u-d-(u-2d-1)}{u-d-1} \geq \frac{d}{u}$ then known tail bounds on the sum of geometrically-distributed random variable (e.g. [Jan18, Theorem 3.1]) we get that for large enough values of m,

$$\Pr_{S \sim \mathcal{D}^m}[S \notin \mathcal{S}] = \Pr[X \leq m] \leq \Pr[X \leq \lambda \mathbb{E}[X]] \leq e^{-p_* \mathbb{E}[X](\lambda - 1 - \ln\lambda)} \leq e^{-\Omega(\ln u)} \leq 1/2 . \quad (10)$$

The lower bound on the expectation of Ψ_2 is proved identically to the proof in Claim 7. □

Similarly to Claim 8, we conclude the following.

Claim 11. *For* $\alpha \leq \sqrt{\frac{d}{40\beta m}}$*, then with probability at least* $1/11$ *over the choice of* $\ell \in \mathcal{L}'$ *we have*

$$\Pr_{S \sim \mathcal{D}_\ell^m}\left[\Psi_1(\ell, h_{\ell,S}) + \Psi_2(\ell, h_{\ell,S}) \geq \frac{(1-\beta)\varepsilon d}{4(u-d-1)} + \frac{\alpha\beta}{12}\right] \geq \frac{1}{25} .$$

We next want to show that there exists a labeling $\ell \in \mathcal{L}'$ such that with high probability over $S \sim \mathcal{D}_\ell^m$, there exists a voting classifier $f_S \in C(\hat{\mathcal{H}})$ attaining high margins with $h_{\ell,S}$. since the distribution induced on $\{\xi_i\})_{i \in [u-d+1,u]}$ by $\mathcal{D}_\ell$ is uniform, we conclude the following for a large enough value of N.

Claim 12. *With probability at least* $99/100$ *over the choice of a labeling* $\ell \in \mathcal{L}'$*,*

$$\Pr_{S \sim \mathcal{D}_\ell}\left[\exists f_S \in \mathcal{C}(\hat{\mathcal{H}}) : \forall i \in [i].\ h_{\ell,S}(\xi_i) f_S(\xi_i) \geq \theta\right] \geq \frac{99}{100} .$$

Proof. For two labelings $\ell \in \mathcal{L}(u,d)$ and $\ell' \in \mathcal{L}'$ we say that ℓ and ℓ' are similar, and denote $\ell \equiv \ell'$ if for all $i \in [u-d+1, u]$, $\ell_i = \ell'_i$. From Claim 5 we know that

$$\begin{aligned}
1 - 1/N &\leq \Pr_{\ell \in_R \mathcal{L}(u,d)}[\exists f \in \mathcal{C}(\hat{\mathcal{H}}) : \forall i \in [u].\ \ell_i f(\xi_i) \geq \theta] = \\
&= \sum_{\ell' \in \mathcal{L}} \Pr_{\ell \in_R \mathcal{L}(u,d)}[\exists f \in \mathcal{C}(\hat{\mathcal{H}}) : \forall i \in [u].\ \ell_i f(\xi_i) \geq \theta | \ell \equiv \ell'] \cdot \Pr_{\ell \in_R \mathcal{L}(u,d)}[\ell \equiv \ell'] \\
&= \sum_{\ell' \in \mathcal{L}} \Pr_{S \sim \mathcal{D}_{\ell'}^m}[\exists f_S \in \mathcal{C}(\hat{\mathcal{H}}) : \forall i \in [u].\ h_{\ell',S}(\xi_i) f(\xi_i) \geq \theta | \ell \equiv \ell'] \cdot \Pr_{\ell \in_R \mathcal{L}(u,d)}[\ell \equiv \ell']
\end{aligned}$$

For a large enough value of N we conclude that with probability at least $99/100$ over a choice of $\ell' \in \mathcal{L}'$, for at least a $99/100$ fraction of samples $S \sim \mathcal{D}_{\ell'}^m$ there exists a voting classifier $f_S \in C(\hat{\mathcal{H}})$ attaining high margins with $h_{\ell',S}$. □

Combining Claims 12 and 11 we conclude that if $\alpha \leq \sqrt{\frac{d}{40\beta m}}$ then there exists $\hat{\ell} \in \mathcal{L}'$ satisfying the guarantees in Lemma 9. The proof of the lemma, and therefore of Theorem 2 is now complete.

4 Existence of a Small Hypotheses Set

Fix some $\theta \in (0, 1/40)$, $\delta \in (0,1)$ and an integer $d \leq u$. Let $\gamma = 4\theta \in (0, 1/10)$ and let $N = 2\gamma^{-2} \ln d \cdot \ln \frac{\gamma^{-2} \ln d}{\delta} \cdot e^{O(\theta^2 d)}$.We define the distribution μ via the following procedure, that samples a hypothesis set $\mathcal{H} \sim \mu$. Let $\hat{h} : \mathcal{X} \to \{-1, 1\}$ be defined by $\hat{h}(x) = 1$ for all $x \in \mathcal{X}$. Sample independently and uniformly at random N hypotheses $h_1, \ldots, h_N \in_R \mathcal{X} \to \{-1, 1\}$, and define $\mathcal{H} := \{\hat{h}\} \cup \{h_j\}_{j \in [N]}$.

Clearly every $\mathcal{H} \in \mathrm{supp}(\mu)$ satisfies $|\mathcal{H}| = N + 1$. We therefore turn to prove the second property. To this end, let $k = \gamma^{-2} \ln d$. In order to show existence of a voting classifier, we conceptually change the procedure defining μ, and think of the random hypotheses as being sampled in k equally sized "batches", each of size N/k, and adding $\hat{h}$ to each of them. Denote the batches by $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_k$. We consider next the following procedure to construct a voting classifier $f \in C(\mathcal{H})$ given $\mathcal{H} \sim \mu$. We will use the main ideas from the AdaBoost algorithm. Recall that AdaBoost creates a voting classifier using a sample $S = ((x_1, y_1), \ldots, (x_u, y_u))$ in iterations. Staring with $f_0 = 0$, in iteration j, it computes a new voting classifier $f_j = f_{j-1} + \alpha_j h_j$ for some hypothesis $h_j \in \mathcal{H}$ and weight α_j. The heart of the algorithm lies in choosing h_j. In each iteration, AdaBoost computes a distribution D_j over S and chooses a hypothesis h_j minimizing

$$\varepsilon_j = \Pr_{i \sim D_j}[h_j(x_i) \neq y_i].$$

The weight it then assigns is $\alpha_j = (1/2)\ln((1-\varepsilon_j)/\varepsilon_j)$ and the next distribution D_{j+1} is

$$D_{j+1}(i) = \frac{D_j(i)\exp(-\alpha_j y_i h_j(x_i))}{Z_j}$$

where Z_j is a normalization factor, namely

$$Z_j = \sum_{i=1}^{d} D_j(i)\exp(-\alpha_j y_i h_j(x_i)).$$

The first distribution D_1 is the uniform distribution.

We alter the above slightly assigning uniform weights on the hypotheses, and setting $\alpha_j = \frac{1}{2}\ln\frac{1+2\gamma}{1-2\gamma}$ for all iterations j. The algorithm is formally described as Algorithm 1.

We will prove that the algorithm fails with probability at most δ (over the choice of $\mathcal{H}$), and that if the algorithm does not fail, then it returns a voting classifier with minimum margin at least θ. First note that if f is the classifier returned by the algorithm, then clearly $f = \frac{1}{k}\sum_{j \in [k]} h_j \in C(\mathcal{H})$ is a voting classifier.

Claim 13. *Algorithm 1 fails with probability at most δ.*

Proof. Since $\mathcal{H}_1, \ldots, \mathcal{H}_k$ are independent, it is enough to show that for every $j \in [k]$, for every $w \in \Delta_u$ with probability at least $1 - \delta/k$ there exsits $h_j \in \mathcal{H}_j$ such that

$$\sum_{i \in [u]} w_i \mathbb{1}_{y_i \neq h_j(x_i)} \leq \frac{1}{2} - \gamma\,, \tag{11}$$

where Δ_u is the u-dimensional simplex. First note that if $\sum_{i \in [u]: y_i = -1} w_i \leq \frac{1}{2} - \gamma$, then $\hat{h} \in \mathcal{H}_j$ satisfies (11). We can therefore assume $\sum_{i \in [u]: y_i = -1} w_i > \frac{1}{2} - \gamma$. Next, note that for every

Input: $(\mathcal{H}_1, \ldots, \mathcal{H}_k) \sim \mu$
Output: $f \in C\left(\bigcup_{j\in[k]} \mathcal{H}_j\right)$

let $\alpha = \frac{1}{2}\ln\frac{1+2\gamma}{1-2\gamma}$
let $f(x) = 0$ for all $x \in \mathcal{X}$
let $D_1(i) = \frac{1}{u}$ for all $i \in [u]$.
for $j = 1$ to k **do**
    Find a hypothesis $h_j \in \mathcal{H}_j$ satisfying $\sum_{i\in[u]} D_j(i)\mathbb{1}_{y_i \neq h_j(x_i)} \leq \frac{1}{2} - \gamma$.
    If there is no such hypothesis, **return** *fail*.
    $f_j \leftarrow f_{j-1} + h_j$.
    $Z_j \leftarrow \sum_{i\in[u]} D_j(i)\exp(-\alpha y_i h_j(x_i))$.
    for every $i \in [u]$ let $D_{j+1}(i) = \frac{1}{Z_j} D_j(i)\exp(-\alpha y_i h_j(x_i))$.
return $\frac{1}{k} f_k$.

Algorithm 1: Construct a Voting Classifier

$h : \mathcal{X} \to \{-1, 1\}$ we have

$$\sum_{i\in[u]} w_i \mathbb{1}_{y_i \neq h(x_i)} = \sum_{i\in[u]} \frac{1}{2}(w_i - w_i y_i h(x_i)) = \frac{1}{2}\left(\sum_{i\in[u]} w_i - \sum_{i\in[u]} w_i y_i h(x_i)\right) = \frac{1}{2} - \frac{1}{2}\sum_{i\in[u]} w_i y_i h(x_i)$$

Therefore $\sum_{i\in[u]} w_i \mathbb{1}_{y_i \neq h(x_i)} \geq \frac{1}{2} - \gamma$ if and only if $\sum_{i\in[u]} w_i y_i h(x_i) \geq 2\gamma$. We want to show that with probability at most $\frac{\delta}{k}$ every $h \in \mathcal{H}_j$ satisfies $\sum_{i\in[u]} w_i y_i h_j(x_i) \geq 2\gamma$. We claim that it is enough to show that

$$\Pr_{h\in_R \mathcal{X}\to\{-1,1\}}\left[\sum_{i\in[u]} w_i y_i h(x_i) \geq 2\gamma\right] \geq \frac{k\ln\frac{k}{\delta}}{N} = \frac{1}{2}e^{-\Theta(\gamma^2 d)} \tag{12}$$

To see why this is enough assume that (12) is true, then since sampling $\mathcal{H}_j$ means indepently and uniformly sampling N/k hypotheses $h \in_R \mathcal{X} \to \{-1, 1\}$, the probability that there exists $h \in \mathcal{H}_j$ such that (11) holds is at least

$$1 - (1 - \frac{k\ln\frac{k}{\delta}}{N})^{N/k} \geq 1 - \exp\left(-\frac{k\ln\frac{k}{\delta}}{N}\cdot\frac{N}{k}\right) = 1 - \frac{\delta}{k}\ .$$

We thus turn to prove that (12) holds. To this end, let $M := \{i \in [u] : \beta_i < 0\}$. Recall that $|M| \leq d$ and that we assumed $\sum_{i\in M} w_i = \sum_{i\in M} |y_i w_i| \geq \frac{1}{2} - \gamma$. From a known tail bound by Montgomery-Smith [MS90] on the sum of Rademacher random variables we have that since $\gamma \in (0, 1/10)$,

$$\Pr\left[\sum_{i\in[u]} w_i y_i h(x_i) \geq 2\gamma\right] \geq \Pr\left[\sum_{i\in M} w_i y_i h(x_i) \geq 2\gamma \text{ and } \sum_{i\in[u]\setminus M} w_i y_i h(x_i) \geq 0\right] \geq \frac{1}{2}e^{-\Theta(\gamma^2 d)}$$

□

Claim 14. *If Algorithm 1 does not fail, then for every $i \in [y]$, $y_i f(x_u) \geq \theta$.*

Proof. We first show by induction that for all $j \in [k]$ we have that for all $i \in [u]$

$$\exp(-\alpha y_i f_j(x_i)) = u \cdot D_{j+1}(i)\prod_{\ell\in[j]} Z_\ell\ .$$

To see this observe that for all $i \in [u]$, $D_2(i) = \frac{D_1(i)}{Z_1}\exp(-\alpha y_i h_1(x_i))$. Since $h_1 = f_1$ and by rearranging we get that $\exp(-\alpha y_i f_1(x_i)) = \frac{D_2(i)Z_1}{D_1(i)} = u \cdot D_2(i)Z_1$. For the induction step we have

that

$$\begin{aligned}\exp(-\alpha y_i f_j(x_i)) &= \exp(-\alpha y_i (f_{j-1}(x_i) + h_j(x_i))) = \exp(-\alpha y_i f_{j-1}(x_i)) \cdot \exp(-\alpha y_i h_j(x_i)) \\ &= u \cdot D_j(i) \prod_{\ell \in [j-1]} Z_\ell \cdot \frac{Z_j D_{j+1}(i)}{D_j(i)} \\ &= u \cdot D_{j+1}(i) \prod_{\ell \in [j]} Z_\ell\end{aligned}$$

Since $\sum_{i \in [u]} D_{k+1}(i) = 1$, we get that

$$\sum_{i \in [u]} \exp(-\alpha y_i f_k(x_i)) = u \prod_{\ell \in [k]} Z_\ell \ . \tag{13}$$

We turn therefore to bound Z_ℓ for $\ell \in [k]$. Denote $\varepsilon_\ell = \sum_{i \in [u]} D_\ell(i) \cdot \mathbb{1}_{h_\ell(x_i) \neq y_i}$. Then

$$\begin{aligned}Z_\ell &= \sum_{i \in [u]} D_\ell(i) \exp(-\alpha y_i h_\ell(x_i)) = \sum_{i \in [u]} D_\ell(i) \exp\left(-\frac{1}{2} \ln\left(\frac{1+2\gamma}{1-2\gamma}\right) y_i h_\ell(x_i)\right) \\ &= \sum_{i \in [u]} D_\ell(i) \left(\frac{1+2\gamma}{1-2\gamma}\right)^{-\frac{1}{2} y_i h_\ell(x_i)} = \varepsilon_\ell \left(\frac{1+2\gamma}{1-2\gamma}\right)^{\frac{1}{2}} + (1-\varepsilon_\ell)\left(\frac{1+2\gamma}{1-2\gamma}\right)^{-\frac{1}{2}} \\ &= \left(\frac{\varepsilon_\ell}{1-2\gamma} + \frac{1-\varepsilon_\ell}{1+2\gamma}\right) \sqrt{(1+2\gamma)(1-2\gamma)}\end{aligned}$$

By the condition in line 5 we know that $\varepsilon_\ell \leq \frac{1}{2} - \gamma$. Since $\left(\frac{\varepsilon_\ell}{1-2\gamma} + \frac{1-\varepsilon_\ell}{1+2\gamma}\right)$ is increasing as a function of ε_ℓ we therefore get that

$$Z_\ell \leq \left(\frac{\frac{1}{2}-\gamma}{1-2\gamma} + \frac{\frac{1}{2}+\gamma}{1+2\gamma}\right) \sqrt{(1+2\gamma)(1-2\gamma)} = \sqrt{(1+2\gamma)(1-2\gamma)} \leq 1 - 2\gamma^2 \ ,$$

where the last inequality follows from the fact that $1 - 4\gamma^2 \leq (1-2\gamma^2)^2$. Substituting in (13) we get that for every $i \in [u]$,

$$\exp(-\alpha y_i f_k(x_i) \leq \sum_{i \in [u]} \exp(-\alpha y_i f_k(x_i)) = u \prod_{\ell \in [k]} Z_\ell \leq u \cdot \left(1 - 2\gamma^2\right)^k \leq \exp(\ln d - 2k\gamma^2) \ ,$$

and therefore

$$y_i f(x_i) = \frac{1}{k} y_i f_k(x_i) \geq \frac{1}{k\alpha}(2k\gamma^2 - \ln d) \ . \tag{14}$$

Since $\ln(1+x) \leq x$ for all $x \geq 0$ we get that

$$\alpha = \frac{1}{2} \ln\left(\frac{1+2\gamma}{1-2\gamma}\right) = \frac{1}{2} \ln\left(1 + \frac{4\gamma}{1-2\gamma}\right) \leq \frac{2\gamma}{1-2\gamma} \leq 4\gamma \ ,$$

where the last inequality follows from the fact that $\gamma \in (0, 1/4)$. Substituting in (14) we get that

$$y_i f(x_i) \geq \frac{1}{4k\gamma}(2k\gamma^2 - \ln d) = \frac{\gamma}{2} - \frac{\ln d}{4k\gamma} \ .$$

Recall that $k = \gamma^{-2} \ln d$, and therefore $y_i f(x_i) \geq \gamma/4 = \theta$. □

5 Conclusions

In this work, we showed almost tight margin-based generalization lower bounds for voting classifiers. These new bounds essentially complete the theory of generalization for voting classifers based on margins alone. Closing the remaining gap between the upper and lower bounds is an intriguing open problem and we hope our techniques might inspire further improvements. Our results come in the form of two theorems, one showing generalization lower bounds for *any* algorithm producing a voting classifier, and a slightly stronger lower bound showing the *existence* of a voting classifier with poor generalization. This raises the important question of whether specific boosting algorithms can produce voting classifiers that avoid the $\lg m$ factor in the second lower bound via a careful analysis tailored to the algorithm. As a final important direction for future work, we suggest investigating whether natural parameters other than margins may be used to better explain the practical generalization error of voting classifiers. At least, we now have an almost tight understanding, if no further parameters are taken into consideration.

References

[AB09] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[BDST00] K. P. Bennett, A. Demiriz, and J. Shawe-Taylor. A column generation algorithm for boosting. In *ICML*, pages 65–72, 2000.

[Bre99] L. Breiman. Prediction games and arcing algorithms. *Neural computation*, 11(7):1493–1517, 1999.

[CG16] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[EHKV89] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247 – 261, 1989.

[FS97] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[GLM19] A. Grønlund, K. G. Larsen, and A. Mathiasen. Optimal minimal margin maximization with boosting. *arXiv preprint arXiv:1901.10789*, 2019.

[GS98] A. J. Grove and D. Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In *AAAI/IAAI*, pages 692–699, 1998.

[GZ13] W. Gao and Z.-H. Zhou. On the doubt about margin explanation of boosting. *Artificial Intelligence*, 203:1–18, 2013.

[Jan18] S. Janson. Tail bounds for sums of geometric and exponential variables. *Statistics & Probability Letters*, 135:1 – 6, 2018. doi:https://doi.org/10.1016/j.spl.2017.11.017.

[KMF+17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017.

[MS90] S. J. Montgomery-Smith. The distribution of Rademacher sums. *Proceedings of the American Mathematical Society*, 109(2):517–522, 1990. Available from: http://www.jstor.org/stable/2048015.

[RW02] G. Rätsch and M. K. Warmuth. Maximizing the margin with boosting. In *COLT*, volume 2375, pages 334–350. Springer, 2002.

[RW05] G. Rätsch and M. K. Warmuth. Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, 6(Dec):2131–2152, 2005.

[SFB+98] R. E. Schapire, Y. Freund, P. Bartlett, W. S. Lee, et al. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686, 1998.

[WSJ+11] L. Wang, M. Sugiyama, Z. Jing, C. Yang, Z.-H. Zhou, and J. Feng. A refined margin analysis for boosting algorithms via equilibrium margin. *Journal of Machine Learning Research*, 12(Jun):1835–1863, 2011.