Optimal Weak to Strong Learning

Kasper Green Larsen
Department of Computer Science
Aarhus University
Aarhus, Denmark
larsen@cs.au.dk

Martin Ritzert
Department of Computer Science
Aarhus University
Aarhus, Denmark
ritzert@cs.au.dk

Abstract

The classic algorithm AdaBoost allows to convert a weak learner, that is an algorithm that produces a hypothesis which is slightly better than chance, into a strong learner, achieving arbitrarily high accuracy when given enough training data. We present a new algorithm that constructs a strong learner from a weak learner but uses less training data than AdaBoost and all other weak to strong learners to achieve the same generalization bounds. A sample complexity lower bound shows that our new algorithm uses the minimum possible amount of training data and is thus optimal. Hence, this work settles the sample complexity of the classic problem of constructing a strong learner from a weak learner.

1 Introduction

The field of boosting has been started from a classic question in learning theory asking whether classifiers that are just slightly better than random guessing can be used to create a classifier with arbitrarily high accuracy when given enough training data. This question was initially asked by Kearns and Valiant [14, 15] and ignited the line of research that eventually lead to the development of AdaBoost [6], the prototype boosting algorithm to date. AdaBoost carefully combines the predictions of several inaccurate classifiers trained with a focus on different parts of the training data to come up with a voting classifier that performs well everywhere.

We quantify the performance of an inaccurate learner by its *advantage* γ over random guessing. Said loosely, a γ*-weak learner* will correctly classify new data points with probability at least $1/2 + \gamma$. In contrast, given $0 < \varepsilon, \delta < 1$ and enough training data a *strong learner* outputs with probability $1 - \delta$ over the choice of the training data and possible random choices of the algorithm a hypothesis that correctly classifies new data points with probability at least $1 - \varepsilon$. The number of samples $m(\varepsilon, \delta)$ such that the learning algorithm achieves the desired accuracy and confidence levels is called the *sample complexity*. The sample complexity is the key metric for the performance of a strong learner and depends on the weak learner's advantage γ, the weak learner's flexibility measured in terms of the VC-dimension, as well as ε and δ. Essentially, a construction with low sample complexity makes the most out of the available training data.

AdaBoost [6] is the classic algorithm for constructing a strong learner from a γ-weak learner. If the weak learner outputs a hypothesis from a base set of hypotheses $\mathcal{H}$, then AdaBoost constructs a strong learner by taking a weighted majority vote among several hypotheses $h_1, \ldots, h_t$ from $\mathcal{H}$. Each of these hypotheses is obtained by invoking the γ-weak learning algorithm on differently weighted versions of a set of training samples S. The number of samples required by AdaBoost for strong learning depends both on the advantage γ of the weak learner and the complexity of the hypothesis set $\mathcal{H}$. If we let d denote the VC-dimension of $\mathcal{H}$, i.e. the cardinality of the largest set of data points $x_1, \ldots, x_d$ such that every classification of $x_1, \ldots, x_d$ can be realized by a hypothesis $h \in \mathcal{H}$, then it

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

is known that AdaBoost is a strong learner, which for error ε and failure probability δ, requires

$$O\left(\frac{d\ln(1/(\varepsilon\gamma))\ln(d/(\varepsilon\gamma))}{\varepsilon\gamma^2}+\frac{\ln(1/\delta)}{\varepsilon}\right), \tag{1}$$

samples. This sample complexity is state-of-the-art for producing a strong learner from a γ-weak learner. However, is this the best possible sample complexity? This is the main question we ask and answer in this work.

First, we present a new algorithm for constructing a strong learner from a weak learner and prove that it requires only

$$O\left(\frac{d}{\varepsilon\gamma^2}+\frac{\ln(1/\delta)}{\varepsilon}\right)$$

samples. In addition to improving over AdaBoost by two logarithmic factors, we complement our new algorithm by a lower bound, showing that any algorithm for converting a γ-weak learner to a strong learner requires

$$\Omega\left(\frac{d}{\varepsilon\gamma^2}+\frac{\ln(1/\delta)}{\varepsilon}\right)$$

samples. Combining these two results, we have a tight bound on the sample complexity of weak to strong learning. In the remainder of the section, we give a more formal introduction to weak and strong learning as well as present our main results and survey previous work.

1.1 Weak and strong learning

Consider a binary classification task in which there is an unknown concept $c : \mathcal{X} \to \{-1, 1\}$ assigning labels to a ground set $\mathcal{X}$. The goal is to learn or approximate c to high accuracy. Formally, we assume that there is an unknown but fixed data distribution $\mathcal{D}$ over $\mathcal{X}$. A learning algorithm then receives a set S of i.i.d. samples $x_1, \ldots, x_m$ from $\mathcal{D}$ together with their labels $c(x_1), \ldots, c(x_m)$ and produces a hypothesis h with $h \approx c$ based on S and the labels. To measure how well h approximates c, it is assumed that a new data point x is drawn from the same unknown distribution $\mathcal{D}$, and the goal is to minimize the probability of mispredicting the label of x. We say that a learning algorithm is a weak learner if it satisfies the following:

Definition 1. *Let $C \subseteq \mathcal{X} \to \{-1, 1\}$ be a set of concepts and $\mathcal{A}$ a learning algorithm. We say that $\mathcal{A}$ is a γ-*weak learner *for C, if there is a constant $\delta_0 < 1$ and an integer $m_0 \in \mathbb{N}$, such that for every distribution $\mathcal{D}$ over $\mathcal{X}$ and every concept $c \in C$, when given m_0 i.i.d. samples $S = x_1, \ldots, x_{m_0}$ from $\mathcal{D}$ together with their labels $c(x_1), \ldots, c(x_{m_0})$, it holds with probability at least $1 - \delta_0$ over the sample S and the randomness of $\mathcal{A}$, that $\mathcal{A}$ outputs a hypothesis $h : \mathcal{X} \to \{-1, 1\}$ such that*

$$\mathcal{L}_{\mathcal{D}}(h) \;=\; \Pr_{x\sim\mathcal{D}}\left[h(x) \neq c(x)\right] \;\leq\; 1/2 - \gamma.$$

A γ-weak learner thus achieves an advantage of γ over random guessing when given m_0 samples. Note that $\mathcal{A}$ knows neither the distribution $\mathcal{D}$, nor the concrete concept $c \in C$ but achieves the advantage γ for all $\mathcal{D}$ and c. We remark that in several textbooks (e.g. Mohri et al. [17]) a weak learner needs to work for any arbitrary $\delta > 0$ while Definition 1 only requires the existence of some δ_0. Thus, every weak learner satisfying the definition of Mohri et al. also satisfies Definition 1, making our results more general.

In contrast to a weak learner, a strong learner can obtain arbitrarily high accuracy:

Definition 2. *Let $C \subseteq \mathcal{X} \to \{-1, 1\}$ be a set of concepts and $\mathcal{A}$ a learning algorithm. We say that $\mathcal{A}$ is a* strong learner *for C, if for all $0 < \varepsilon, \delta < 1$, there is some number $m(\varepsilon, \delta)$ such that for every distribution $\mathcal{D}$ over $\mathcal{X}$ and every concept $c \in C$, when given $m = m(\varepsilon, \delta)$ i.i.d. samples $S = x_1, \ldots, x_m$ from $\mathcal{D}$ together with their labels $c(x_1), \ldots, c(x_m)$, it holds with probability at least $1 - \delta$ over the sample S and the randomness of $\mathcal{A}$, that $\mathcal{A}$ outputs a hypothesis $h : \mathcal{X} \to \{-1, 1\}$ such that*

$$\mathcal{L}_{\mathcal{D}}(h) \;=\; \Pr_{x\sim\mathcal{D}}\left[h(x) \neq c(x)\right] \;\leq\; \varepsilon.$$

The definition of a strong learner is essentially identical to the classic notion of (ε, δ)-PAC learning in the realizable setting. Unlike the γ-weak learner, we here require the learner to output a classifier with arbitrarily high accuracy (ε small) and confidence (δ small) when given enough samples S.

Kearns and Valiant [14, 15] asked whether one can always obtain a strong learner when given access only to a γ-weak learner for a $\gamma > 0$. This was answered affirmatively by Schapire [20] and is the motivation behind the design of AdaBoost [6]. If we let $\mathcal{H}$ denote the set of hypotheses that a γ-weak learner may output from, then AdaBoost returns a *voting* classifier $f(x) = \text{sign}(\sum_{i=1}^{t} \alpha_i h_i(x))$ where each $h_i \in \mathcal{H}$ is the output of the γ-weak learner when trained on some carefully weighted version of the training set S and each α_i is a real-valued weight. In terms of sample complexity $m(\varepsilon, \delta)$, the number of samples stated in Eq. (1) is sufficient for AdaBoost. There are several ways to prove this. For instance, it can be argued that when given m samples, AdaBoost combines only $t = O(\gamma^{-2} \ln m)$ hypotheses $h_1, \ldots, h_t$ from $\mathcal{H}$ in order to produce an f that perfectly classifies all the training data S, i.e. $f(x_i) = c(x_i)$ for all $x_i \in S$. Using that the class $\mathcal{H}^t$ can generate at most $O(\binom{m}{d}^t)$ distinct classifications of m points (i.e. its growth function is bounded by this), one can intuitively invoke classic generalization bounds for PAC-learning in the realizable case to conclude that the hypothesis f satisfies

$$\mathcal{L}_{\mathcal{D}}(f) \;\leq\; O\left(\frac{td\ln(m/d) + \ln(1/\delta)}{m}\right) \;=\; O\left(\frac{d\ln(m/d)\ln m}{\gamma^2 m} + \frac{\ln(1/\delta)}{m}\right) \tag{2}$$

with probability at least $1 - \delta$ over S (and potentially the randomness of the weak learner). Using $\mathcal{L}_{\mathcal{D}}(f) = \varepsilon$ and solving Eq. (2) for m gives the sample complexity stated in Eq. (1). This is the best sample complexity bound of any weak to strong learner prior to this work.

Our main upper bound result is a new algorithm with better sample complexity than AdaBoost and other weak to strong learners. It guarantees the following:

Theorem 1. *Assume we are given access to a γ-weak learner for some $0 < \gamma < 1/2$, using a base hypothesis set $\mathcal{H} \subseteq \mathcal{X} \to \{-1, 1\}$ of VC-dimension d. Then there is a universal constant $\alpha > 0$ and an algorithm $\mathcal{A}$, such that $\mathcal{A}$ is a strong learner with sample complexity $m(\varepsilon, \delta)$ satisfying*

$$m(\varepsilon, \delta) \;\leq\; \alpha \cdot \left(\frac{d\gamma^{-2}}{\varepsilon} + \frac{\ln(1/\delta)}{\varepsilon}\right).$$

We remark that it is often required that a strong learner runs in polynomial time given a polynomial-time weak learner. This is indeed the case for our new algorithm.

Next, we complement our algorithm from Theorem 1 by the following lower bound:

Theorem 2. *There is a universal constant $\alpha > 0$ such that for all integers $d \in \mathbb{N}$ and every $0 < \gamma < 1/80$, there is a finite set $\mathcal{X}$, a concept class $\mathcal{C} \subset \mathcal{X} \to \{-1, 1\}$ and a hypothesis set $\mathcal{H} \subseteq \mathcal{X} \to \{-1, 1\}$ of VC-dimension at most d, such that for every (ε, δ) with $0 < \varepsilon < 1$ and $0 < \delta < 1/3$, there is a distribution $\mathcal{D}$ over $\mathcal{X}$ such that the following holds:*

1. *For every $c \in \mathcal{C}$ and every distribution $\mathcal{D}'$ over $\mathcal{X}$, there is an $h \in \mathcal{H}$ with $\Pr_{x \sim \mathcal{D}'}[h(x) \neq c(x)] \leq 1/2 - \gamma$.*
2. *For any algorithm $\mathcal{A}$, there is a concept $c \in \mathcal{C}$ such that $\mathcal{A}$ requires at least*

$$m \;\geq\; \alpha \cdot \left(\frac{d\gamma^{-2}}{\varepsilon} + \frac{\ln(1/\delta)}{\varepsilon}\right)$$

samples S and labels $c(S)$ to guarantee $\mathcal{L}_{\mathcal{D}}(h_S) \leq \varepsilon$ with probability at least $1 - \delta$ over S, where h_S is the hypothesis produced by $\mathcal{A}$ on S and $c(S)$.

The first statement of Theorem 2 says that the concept class $\mathcal{C}$ can be γ-weakly learned. The second point then states that any learner requires as many samples as our new algorithm. Moreover, the lower bound does not require the algorithm to even use a γ-weak learner, nor does it need to run in polynomial time for the lower bound to apply. Furthermore, the algorithm is even allowed to use the full knowledge of the set $\mathcal{C}$ and the distribution $\mathcal{D}$. The only thing it does not know is which concept $c \in \mathcal{C}$ provides the labels $c(S)$ to the training samples.

In the next section, we present the overall ideas in our new algorithm, as well as a new generalization bound for voting classifiers that is key to our algorithm. Finally, we sketch the main ideas in the lower bound. Due to the space requirements, several proofs have been moved to the supplementary material.

1.2 Main ideas and voting classifiers

One of the key building blocks in our new algorithm is voting classifiers. To formally introduce voting classifiers, define from a hypothesis set $\mathcal{H} \subseteq \mathcal{X} \to \{-1, 1\}$ the set of all convex combinations $\Delta(\mathcal{H})$ of hypotheses in $\mathcal{H}$. That is, $\Delta(\mathcal{H})$ contains all functions f of the form $f(x) = \sum_{i=1}^{t} \alpha_i h_i(x)$ with $\alpha_i > 0$ and $\sum_i \alpha_i = 1$. AdaBoost can be thought of as producing a voting classifier $g(x) = \text{sign}(f(x))$ for an $f \in \Delta(\mathcal{H})$ by appropriate normalization of the weights it uses.

Classic work on understanding the surprisingly high accuracy of AdaBoost introduced the notion of *margins* [1]. For a function $f \in \Delta(\mathcal{H})$, and a sample x with label y, the margin of f on (x, y) is $yf(x)$. Notice that the margin is positive if and only if $\text{sign}(f(x))$ correctly predicts the label y of x. It was empirically observed that AdaBoost produces voting classifiers $g(x) = \text{sign}(f(x))$ where f has large margins. This inspired multiple generalization bounds based on the margins of a voting classifier, considering both the minimum and the k-th margin [11, 3, 2, 18, 19, 16]. The simplest bound when all margins are assumed to be at least γ, is Breiman's min margin bound:

Theorem 3 (Breiman [3])**.** *Let $c \in \mathcal{X} \to \{-1, 1\}$ be an unknown concept, $\mathcal{H} \subseteq \mathcal{X} \to \{-1, 1\}$ a hypothesis set of VC-dimension d and $\mathcal{D}$ an arbitrary distribution over $\mathcal{X}$. With probability at least $1 - \delta$ over a set of m samples $S \sim \mathcal{D}^m$, it holds for every voting classifier $g(x) = \text{sign}(f(x))$ with $f \in \Delta(\mathcal{H})$ satisfying $c(x)f(x) \geq \gamma$ on all $x \in S$, that:*

$$\mathcal{L}_{\mathcal{D}}(g) = O\left(\frac{d \ln(m/d) \ln m}{\gamma^2 m}\right)$$

The resemblance to the generalization performance of AdaBoost in Eq. (2) is no coincidence. Indeed, a small twist to AdaBoost, presented in the AdaBoost$^*_\nu$ algorithm [19], ensures that the voting classifier produced by AdaBoost$^*_\nu$ from a γ-weak learner has all margins at least $\gamma/2$. This gives an alternative way of obtaining the previous best sample complexity in Eq. (1). We remark that more refined generalization bounds based on margins exist, such as the k-th margin bound by Gao and Zhou [7] which is known to be near-tight [9]. These bounds take the whole sequence of margins $c(x_i)f(x_i)$ of all samples $x_i \in S$ into account, not only the smallest. However, none of these bounds leads to better generalization from a γ-weak learner.

We note that the notion of margins has not only been considered in the context of boosting algorithms but also plays a key role in understanding the generalization performance of Support Vector Machines, see e.g. the recent works [13, 10] giving tight SVM generalization bounds in terms of margins.

In our new algorithm, we make use of a voting classifier with good margins as a subroutine. Concretely, we invoke AdaBoost$^*_\nu$ to obtain margins of at least $\gamma/2$ on all training samples. At first sight, this seems to incur logarithmic losses, at least if the analysis by Breiman is tight. Moreover, Grønlund et al. [8] proved a generalization lower bound showing that there are voting classifiers with margins γ on all training samples, but where at least one of the logarithmic factors in the generalization bound must occur. To circumvent this, we first notice that the lower bound only applies when m is sufficiently larger than $d\gamma^{-2}$. We carefully exploit this loophole and prove a new generalization bound for voting classifiers:

Theorem 4. *Let $c \in \mathcal{X} \to \{-1, 1\}$ be an unknown concept, $\mathcal{H} \subseteq \mathcal{X} \to \{-1, 1\}$ a hypothesis set of VC-dimension d and $\mathcal{D}$ an arbitrary distribution over $\mathcal{X}$. There is a universal constant $\alpha > 0$, such that with probability at least $1 - \delta$ over a set of $m \geq \alpha(d\gamma^{-2} + \ln(1/\delta))$ samples $S \sim \mathcal{D}^m$, every voting classifier $g(x) = \text{sign}(f(x))$ with $f \in \Delta(\mathcal{H})$ satisfying $c(x)f(x) \geq \gamma$ on all $x \in S$ achieves*

$$\mathcal{L}_{\mathcal{D}}(g) \leq \tfrac{1}{200}.$$

The value $1/200$ is arbitrary and chosen to match the requirements in the proof of Theorem 1. Notice how our new generalization bound avoids the logarithmic factors when aiming merely at generalization error $1/200$. Breiman's bound would only guarantee that $d\gamma^{-2} \ln(1/\gamma) \ln(d/\gamma)$ samples suffice for such a generalization error. While the focus of previous work on generalization bounds was not on the constant error case, we remark that any obvious approaches to modify the previous proofs could perhaps remove the $\ln m$ factor but not the $\ln(m/d)$ factor. The $\ln(m/d)$ factor turns into $\Theta(\ln(1/\gamma))$ when solving for m in $d \ln(m/d)/(\gamma^2 m) = 1/200$ and is insufficient for our purpose.

With the new generalization bound on hand, we can now construct our algorithm for producing a strong learner from a γ-weak learner. Here we use as template the sample optimal algorithm by

Hanneke [12] for PAC learning in the realizable case (which improved over a previous near-tight result by Simon [21]). Given a training set S, his algorithm carefully constructs a number of sub-samples $S_1, S_2, \ldots, S_k$ of S and trains a hypothesis h_i on each S_i using empirical risk minimization. As the final classifier, he returns the voter $g(x) = \text{sign}\big((1/k)\sum_{i=1}^{k} h_i(x)\big)$.

For our new algorithm, we use Hanneke's approach to construct sub-samples $S_1, \ldots, S_k$ of a training set S. We then run AdaBoost$^*_\nu$ on each S_i to produce a voting classifier $g_i(x) = \text{sign}(f_i(x))$ for an $f_i \in \Delta(\mathcal{H})$ with margins $\gamma/2$ on all samples in S_i. We finally return the voter $h(x) = \text{sign}((1/k)\sum_{i=1}^{k} g_i(x))$. Our algorithm thus returns a majority of majorities.

To prove that our algorithm achieves the desired sample complexity $m(\varepsilon, \delta)$ claimed in Theorem 1, we then revisit Hanneke's proof and show that it suffices for his argument that the base learning algorithm (in his case empirical risk minimization, in our case AdaBoost$^*_\nu$) achieves an error of at most $1/200$ when given τ samples. If this is the case, then his proof can be modified to show that the final error of the output voter drops to $O(\tau/m)$. Plugging in the $\tau = \alpha(d\gamma^{-2} + \ln(1/\delta))$ from our new generalization bound in Theorem 4 completes the proof.

Let us remark that a lower bound by Grønlund et al. [8] shows the existence of a voting classifier with simultaneously large margins and a generalization error with an additional log-factor. It is thus conceivable that a simple majority vote is not sufficient and a majority of majorities is indeed necessary, although the lower bound only guarantees the *existence* of a 'bad' voter with good margins and not that *all* such voters are 'bad'.

In the following, we start by proving our new generalization bound (Theorem 4) in Section 2. We then proceed in Section 3 to present our new algorithm and sketch the proof that it gives the guarantees in Theorem 1. Finally, in Section 4 we give the proof ideas of the lower bound in Theorem 2.

2 New margin-based generalization bounds for voting classifiers

In this section, we prove the new generalization bound stated in Theorem 4. For ease of notation, we write that $\mathcal{D}$ is a distribution over $\mathcal{X} \times \{-1, 1\}$ (and not just a distribution over $\mathcal{X}$) and implicitly assume that the label of each $x \in \mathcal{X}$ is $c(x)$ for the unknown concept c. Moreover, for a voting classifier $g(x) = \text{sign}(f(x))$ with $f \in \Delta(\mathcal{H})$, we simply refer to f as the voting classifier and just remark that one needs to take the sign to make a prediction. Finally, we think of the sample S as a set of pairs (x_i, y_i) with $x_i \in \mathcal{X}$ and $y_i = c(x_i) \in \{-1, 1\}$.

The key step in the proof of Theorem 4 is to analyze the generalization performance for a voting classifier obtained by combining randomly drawn hypotheses among the hypotheses $h_1, \ldots, h_t$ making up a voting classifier $f = \sum_i \alpha_i h_i$ from $\Delta(\mathcal{H})$. We then relate that to the generalization performance of f itself. Formally, we define a distribution $\mathcal{D}_{f,t}$ for every f and look at a random hypothesis from $\mathcal{D}_{f,t}$. We start by defining this distribution.

Let $f(x) = \sum_h \alpha_h h(x) \in \Delta(\mathcal{H})$ be a voting classifier. Let $\mathcal{D}_f$ be the distribution over $\mathcal{H}$ (the base hypotheses used in f) where h has probability α_h. Consider drawing t i.i.d. hypotheses $h'_1, \ldots, h'_t$ from $\mathcal{D}_f$ and then throwing away each h'_i independently with probability $1/2$. Let t' be the number of remaining hypotheses, denote them $h_1, \ldots, h_{t'}$, and let $g = \frac{1}{t'}\sum_{i=1}^{t'} h_i$. One can think of g as a sub-sampled version of f with replacement. Denote by $\mathcal{D}_{f,t}$ the distribution over g.

Key properties of $\mathcal{D}_{f,t}$. In the following, we analyze how a random g from $\mathcal{D}_{f,t}$ behaves and show that while it behaves similar to f it produces with good probability predictions that are big in absolute value (even if $f(x) \approx 0$). The proofs of the following three lemmas are given in the supplementary material. First, we note that predictions made by a random g are often close to those made by f and then observe that g rarely makes predictions $g(x)$ that are small in absolute value.

Lemma 1. *For any $x \in \mathcal{X}$, any $f \in \Delta(\mathcal{H})$, and any $\mu > 0$:* $\Pr_{g\sim\mathcal{D}_{f,t}}[|f(x) - g(x)| \geq \mu] < 5e^{-\mu^2 t/32}$.

Lemma 2. *For any $x \in \mathcal{X}$, any $f \in \Delta(\mathcal{H})$, and any $\mu \geq 1/t$:* $\Pr_{g\sim\mathcal{D}_{f,t}}[|g(x)| \leq \mu] \;\leq\; 2\mu\sqrt{t}$.

Lemma 2 states that even if $f(x) \approx 0$ for an unseen sample x, $g(x)$ will still be large with good probability. Thus we can think of g as having large margins (perhaps negative) also on unseen data. This is crucial for bounding the generalization error. The proof follows from an invocation of Erdős' improved Littlewood-Offord lemma [5].

As the last property, we look at the out-of-sample and in-sample error of a random g and start with relating the generalization error of f to that of a random g. To formalize this, define for any distribution $\mathcal{D}$, the loss $\mathcal{L}^t_{\mathcal{D}}(f) := \Pr_{(x,y)\sim\mathcal{D},\, g\sim\mathcal{D}_{f,t}}\left[yg(x) \le 0\right]$ and when writing $\mathcal{L}^t_S(f)$ we implicitly assume S to also denote the uniform distribution over all $(x, y) \in S$. We then have:

Lemma 3. *For any distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}$, any $t \ge 36$ and any voting classifier $f \in \Delta(\mathcal{H})$ for a hypothesis set $\mathcal{H} \subset \mathcal{X} \to \{-1, 1\}$, we have $\mathcal{L}_{\mathcal{D}}(f) \le 3\mathcal{L}^t_{\mathcal{D}}(f)$.*

Moreover, if f has margins γ on all training samples $(x, y) \in S$, then g is correct on most of S provided that we set t big enough:

Lemma 4. *Let S be a set of m samples in $\mathcal{X} \times \{-1, 1\}$ and assume f is a voting classifier with $yf(x) \ge \gamma$ for all $(x, y) \in S$. For $t \ge 1024\gamma^{-2}$, we have $\mathcal{L}^t_S(f) \le 1/1200$.*

Proof. By Lemma 1, it holds for all $(x, y) \in S$, that $|f(x) - g(x)| \ge \gamma$ with probability at most $5\exp\left(-\gamma^2 t/32\right) \le 5e^{-32} \ll 1/1200$. Since $yf(x) \ge \gamma$, this implies $\operatorname{sign}\big(g(x)\big) = \operatorname{sign}\big(f(x)\big) = y$. □

The last ingredient for the proof of Theorem 4 is to relate $\mathcal{L}^t_S(f)$ and $\mathcal{L}^t_{\mathcal{D}}(f)$. For the proof we use Lemma 2 to infer that with good probability $|g(x)| = \Omega(\gamma)$, i.e. has large absolute value. We use this to argue that $\operatorname{sign}(g)$ often belongs to a class with small VC-dimension and then apply a growth-function argument to relate $\mathcal{L}^t_S(f)$ and $\mathcal{L}^t_{\mathcal{D}}(f)$. Formally, we prove the following lemma.

Lemma 5. *Let $\mathcal{D}$ be an arbitrary distribution over $\mathcal{X} \times \{-1, 1\}$ and let $\mathcal{H} \subset \mathcal{X} \to \{-1, 1\}$ be a hypothesis set of VC-dimension d. There is a universal constant $\alpha > 0$ such that for any $t \in \mathbb{N}$ and any $m \ge \alpha t d$, it holds that:*

$$\Pr_S\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{\mathcal{D}}(f)| > \tfrac{1}{1200}\Big] \;\le\; \alpha \cdot \exp(-m/\alpha).$$

Before we prove Lemma 5, we show how to use it to prove Theorem 4. Since we are only aiming to prove the generalization of voting classifiers f with $yf(x) \ge \gamma$ for all samples $(x, y) \in S$, Lemma 4 tells us that such f have small $\mathcal{L}^t_S(f)$ when $t \ge 1024\gamma^{-2}$. We thus fix $t = 1024\gamma^{-2}$ and get that $\mathcal{L}^t_S(f) \le 1/1200$ from Lemma 4. By Lemma 5, with probability at least $1 - \alpha\exp(-m/\alpha)$ over the sample S, we have for all $f \in \Delta(\mathcal{H})$ that $|\mathcal{L}^t_S(f) - \mathcal{L}^t_{\mathcal{D}}(f)| \le 1/1200$ and thus $\mathcal{L}^t_{\mathcal{D}}(f) \le 1/600$. Finally, Lemma 3 gives us that $\mathcal{L}_{\mathcal{D}}(f) \le 3\mathcal{L}^t_{\mathcal{D}}(f)$ for all $f \in \Delta(\mathcal{H})$. Together we thus have $\mathcal{L}^t_{\mathcal{D}}(f) \le 1/600 \Rightarrow \mathcal{L}_{\mathcal{D}}(f) \le 1/200$ for any $m \ge \alpha t d \ge \alpha' d\gamma^{-2}$ where $\alpha' > 0$ is a universal constant. By observing that $\alpha\exp(-m/\alpha) < \delta$ for $m \ge \alpha\ln(\alpha/\delta)$, this completes the proof of Theorem 4. What remains is thus to prove Lemma 5 which we do in the remainder of this section.

2.1 Relating $\mathcal{L}^t_S(f)$ and $\mathcal{L}^t_{\mathcal{D}}(f)$

The last remaining step to show Theorem 4 is thus to relate $\mathcal{L}^t_S(f)$ to $\mathcal{L}^t_{\mathcal{D}}(f)$, i.e. to prove Lemma 5. In the proof, we rely on the classic approach for showing generalization for classes $\mathcal{H}$ of bounded VC-dimension and introduce a *ghost* set that only exists for the sake of analysis. In addition to the sample S, we thus consider a *ghost* set S' of another m i.i.d. samples from $\mathcal{D}$. This allows us to prove:

Lemma 6. *For $m \ge 2400^2$ any t and any f, it holds that:*

$$\Pr_S\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{\mathcal{D}}(f)| > \tfrac{1}{1200}\Big] \;\le\; 2\cdot\Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{S'}(f)| > \tfrac{1}{2400}\Big].$$

As the proof is standard, it can be found in the supplementary material.

We thus only need to bound $\Pr_{S,S'}[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{S'}(f)| > 1/2400]$. To do this, consider drawing a data set P of $2m$ i.i.d. samples from $\mathcal{D}$ and then drawing S as a set of m uniform samples from P without replacement and letting S' be the remaining samples. Then S and S' have the same distribution as if they were drawn as two independent sets of m i.i.d. samples each. From here on, we thus think of S and S' as being sampled via P.

Now consider a fixed set P in the support of $\mathcal{D}^{2m}$ and define $\Delta^{\mu}_{\delta}(\mathcal{H}, P)$ as the set of voting classifiers $f \in \Delta(\mathcal{H})$ for which $\Pr_{(x,y)\sim P}[|f(x)| \ge \mu] \ge 1 - \delta$. These are the voting classifiers that make

predictions of large absolute value on most of *both* S and S' (if $\delta \ll 1/2$). The crucial point, and the whole reason for introducing g, is that regardless of what f is, a random $g \sim \mathcal{D}_{f,t}$ often lies in the set $\Delta^{\mu}_{\delta}(\mathcal{H}, P)$:

Lemma 7. *For any data set P, parameters $0 < \delta < 1$ and t, and every $\mu \leq \delta/(9600\sqrt{t})$, we have* $\Pr_{g\sim\mathcal{D}_{f,t}}[g \notin \Delta^{\mu}_{\delta}(\mathcal{H}, P)] \leq 1/4800$.

Proof. Define an indicator X_i for each $(x_i, y_i) \in P$ taking the value 1 if $|g(x_i)| \leq \mu$. By Lemma 2, we have $\mathbb{E}[\sum_i X_i] \leq |P|\, 2\mu\sqrt{t} \leq |P|\, \delta/4800$. By Markov's inequality $\Pr[\sum_i X_i \geq \delta|P|] \leq 1/4800$. □

If we had just considered f, we had no way of arguing that f makes predictions of large absolute value on S', since the only promise we are given is that it does so on S. That g makes predictions of large absolute value even outside of S is crucial for bounding the generalization error in the following.

Let us now define $\hat{\Delta}^{\mu}_{\delta}(P) = \operatorname{sign}\big(\Delta^{\mu}_{\delta}(\mathcal{H}, P)\big)$ which means that $\hat{\Delta}^{\mu}_{\delta}(P)$ contains all the hypotheses that are obtained by voting classifiers in $\Delta^{\mu}_{\delta}(\mathcal{H}, P)$ when taking the sign. Since g is in $\Delta^{\mu}_{\delta}(\mathcal{H}, P)$ except with probability $1/4800$ by Lemma 7, we can prove:

Lemma 8. *For any $0 < \delta < 1$, every t, and every $\mu \leq \delta/(9600\sqrt{t})$, we have*

$$\Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} \big|\mathcal{L}^t_S(f) - \mathcal{L}^t_{S'}(f)\big| > \tfrac{1}{2400}\Big] \;\leq\; \sup_P 2\,\big|\hat{\Delta}^{\mu}_{\delta}(P)\big| \exp\big(-2m/9600^2\big).$$

Again, the proof of the lemma can be found in the supplementary material. What Lemma 8 gives us, is that it relates the generalization error to the growth function $|\hat{\Delta}^{\mu}_{\delta}(P)|$. The key point is that $\hat{\Delta}^{\mu}_{\delta}(P)$ was obtained from voting classifiers with predictions of large absolute value on all but a δ fraction of points in P. This implies that we can bound the VC-dimension of $\hat{\Delta}^{\mu}_{\delta}(P)$ when restricted to the point set P using Rademacher complexity:

Lemma 9. *Let $\mathcal{H}$ be a hypothesis set of VC-dimension d. For any $\delta, \mu > 0$ and point set P, we have that the largest subset P' of P that $\hat{\Delta}^{\mu}_{\delta}(P) = \operatorname{sign}\big(\Delta^{\mu}_{\delta}(\mathcal{H}, P)\big)$ can shatter, has size at most $|P'| = d' < \max\{2\delta|P|,\, 4\alpha^2\mu^{-2}d\}$, where $\alpha > 0$ is a universal constant.*

Proof. Recall that the VC-dimension of $\mathcal{H}$ is d. Thus the Rademacher complexity of $\mathcal{H}$ for any point set P' is:

$$\mathop{\mathbb{E}}_{\sigma\in P'\to\{-1,1\}}\left[\frac{1}{|P'|}\sup_{h\in\mathcal{H}}\left|\sum_{x\in P'} h(x)\sigma(x)\right|\right] \;<\; \alpha\sqrt{\frac{d}{|P'|}}$$

for a universal constant $\alpha > 0$ (see e.g. [22]). Assume $P' \subseteq P$ with $|P'| = d'$ can be shattered. Fix any labeling $\sigma \in P' \to \{-1, 1\}$. Let $h_\sigma \in \hat{\Delta}^{\mu}_{\delta}(P)$ be the hypothesis generating the dichotomy σ (which exists since P' is shattered). Since $h_\sigma \in \hat{\Delta}^{\mu}_{\delta}(P)$, there must be some $g \in \Delta^{\mu}_{\delta}(\mathcal{H}, P)$ such that $h_\sigma = \operatorname{sign}(g)$ on the point set P'. If $|P'| \geq 2\delta|P|$, then by definition of $\Delta^{\mu}_{\delta}(\mathcal{H}, P)$, there are at least $|P'| - \delta|P| \geq |P'|/2$ points $x \in P'$ for which $|g(x)| \geq \mu$. This means that $(1/|P'|)\sum_{x\in P'} g(x)\sigma(x) \geq (1/2)\mu$. But $g(x)$ is a convex combination of hypotheses from $\mathcal{H}$, hence there is also a hypothesis $h \in \mathcal{H}$ for which $(1/|P'|)\sum_{x\in P'} h(x)\sigma(x) \geq (1/2)\mu$. Since this holds for all σ, by the bound on the Rademacher complexity, we conclude $\alpha\sqrt{d/|P'|} > (1/2)\mu \implies |P'| < 4\alpha^2\mu^{-2}d$. We thus conclude that the largest set that $\hat{\Delta}^{\mu}_{\delta}(P)$ can shatter, has size less than $\max\big\{2\delta|P|,\; 4\alpha^2\mu^{-2}d\big\}$. □

We remark that it was crucial to introduce the random hypothesis g, since all we are promised about the original hypothesis f is that it has large margins on S, i.e. on only half the points in P. That case would correspond to $\delta = 1/2$ in Lemma 9 and would mean that we could potentially shatter all of P. In order for the bound to be useful, we thus need $\delta \ll 1/2$ and thus large margins on much more than half of P (which we get by using g).

For a $0 < \delta < 1$ to be determined, let us now fix $\mu = \delta/(9600\sqrt{t})$ and assume that the number of samples m satisfies $m \geq \max\{\alpha^2\mu^{-2}d/\delta,\; 2400^2\}$ where α is the constant from Lemma 9. By Lemma 8, we have

$$\Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{S'}(f)| > \tfrac{1}{2400}\Big] \;\leq\; \sup_P 2|\hat{\Delta}^{\mu}_{\delta}(P)| \exp(-2m/9600^2).$$

Algorithm 1: *Sub-Sample*(A, B) (Hanneke [12])

Input: Sets A and B

if $|A| \leq 3$ **then** // stop when A is too small to recurse

  return $A \cup B$

else

  Let A_0 denote the first $|A| - 3\lfloor |A|/4 \rfloor$ elements of A, // split A evenly
    A_1 the next $\lfloor |A|/4 \rfloor$ elements,
    A_2 the next $\lfloor |A|/4 \rfloor$ elements, and
    A_3 the remaining $\lfloor |A|/4 \rfloor$ elements.

  return *Sub-Sample*$(A_0, A_2 \cup A_3 \cup B) \cup$ // recurse in leave-one-out fashion
    Sub-Sample$(A_0, A_1 \cup A_3 \cup B) \cup$
    Sub-Sample$(A_0, A_1 \cup A_2 \cup B)$

Algorithm 2: Optimal weak-to-strong learner

Input: Set S of m samples.

$\{C_1, \ldots, C_k\} =$ *Sub-Sample*$(S, \emptyset)$ // create highly overlapping subsamples of S

for $i = 1, \ldots, k$ **do**

  $h_i = \mathcal{A}^*_\nu(C_i)$ // run AdaBoost$^*_\nu$ on all those sub-samples

return $h(x) = \text{sign}\left(\sum_{i=1}^k h_i(x)\right)$. // return unweighted majority vote

Lemma 9 gives us that the largest subset $P' \subseteq P$ that $\hat{\Delta}^\mu_\delta(P)$ shatters has size at most $d' < \max\{2\delta|P|, 4\alpha^2\mu^{-2}d\}$. By our assumption on m, the term $2\delta|P| = 4\delta m$ is at least $4c^2\mu^{-2}d$ and thus $2\delta|P| = 4\delta m$ takes the maximum in the bound on d'. By the Sauer-Shelah lemma, we have that $|\hat{\Delta}^\mu_\delta(P)| \leq \sum_{i=0}^{4\delta m-1} \binom{2m}{i}$. For $\delta \leq 1/4$, this is at most $\binom{2m}{4\delta m} \leq (e\delta^{-1}/2)^{4\delta m} = \exp\left(4\delta m \ln(e\delta^{-1}/2)\right)$.

As conclusion we have:

$$\Pr_{S,S'}\left[\sup_{f \in \Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{S'}(f)| > \tfrac{1}{2400}\right] \leq 2\exp(4\delta m \ln(e\delta^{-1}/2))\exp(-2m/9600^2).$$

Let us now fix $\delta = 10^{-10}$. We then have

$$\begin{aligned} & 2\exp\left(4\delta m \ln(e\delta^{-1}/2)\right)\exp(-2m/9600^2) \\ = \; & 2\exp\left(m(4\delta \ln(e\delta^{-1}/2) - 2/9600^2)\right) \\ \leq \; & 2\exp\left(-m/10^8\right) \end{aligned}$$

where the last step is a numerical calculation. By Lemma 6, this in turn implies:

$$\Pr_S\left[\sup_{f \in \Delta(\mathcal{H})} \left|\mathcal{L}^t_S(f) - \mathcal{L}^t_{\mathcal{D}}(f)\right| > \tfrac{1}{1200}\right] \leq 4\exp(-m/10^8).$$

Since we only required $m \geq \max\{\alpha^2\mu^{-2}d/\delta, 2400^2\}$ and we had $\mu = \delta/(9600\sqrt{t})$, this is satisfied for $m \geq \alpha' t d$ for a large enough constant $\alpha' > 0$. This completes the proof of Lemma 5 and thus also finishes the proof of Theorem 4.

3 Weak to strong learning

In this section, we give our algorithm for obtaining a strong learner from a γ-weak learner with optimal sample complexity and sketch the main proof idea.

The algorithm obtaining the guarantees of Theorem 1 is as follows: Let $\mathcal{A}^*_\nu$ be an algorithm that on a sample S outputs a classifier $g = \text{sign}(f)$, where f is a voting classifier with margins at least $\gamma/2$ on all samples in S such as AdaBoost$^*_\nu$ [19]. Given a set S of m i.i.d. samples from an unknown distribution $\mathcal{D}$, we run $\mathcal{A}^*_\nu$ on a number of samples $C_1, C_2, \ldots, C_k \subset S$ obtaining hypotheses $h_1, h_2, \ldots, h_k$. We then return the (unweighted) majority vote among $h_1, \ldots, h_k$ as our final hypothesis h^*. The subsets C_i are chosen by the algorithm *Sub-Sample* (shown in Algorithm 1) as in the optimal PAC learning algorithm by Hanneke [12]. The final algorithm (Algorithm 2) calls $\mathcal{A}^*_\nu$ on all subsets

returned by Algorithm 1 and returns the majority vote. Note that the final hypothesis returned by Algorithm 2 is a majority of majorities since $\mathcal{A}_\nu^*$ already returns a voting classifier. Let us also remark that in Theorem 4 we have a failure probability $\delta_0 > 0$, while the analysis of AdaBoost$_\nu^*$ assumes $\delta_0 = 0$, i.e. that the weak learner always achieves an advantage of at least γ. If one knows γ in advance, this is not an issue as AdaBoost$_\nu^*$ only calls the weak learner on distributions over the training data S and one can thus compute the advantage from the training data. After in expectation $1/(1-\delta_0)$ invocations of the weak learner, we thus get a hypothesis with advantage γ.

In total there are $k = 3^{\lceil \log_4(m) \rceil} \approx m^{0.79}$ calls to the weak learner, each with a sub-sample of linear size. Since AdaBoost$_\nu^*$ runs in polynomial time on its input, given that the weak learner is polynomial, Algorithm 2 is polynomial under the same condition.

The formal proof that Algorithm 2 has the guarantees of Theorem 1 is given in the supplementary material due to space constraints. It follows the proof of Hanneke [12] pretty much uneventfully, although carefully using that a generalization error of $1/200$ suffices for each call of $\mathcal{A}_\nu^*$. The key observation is that each of the recursively generated sub-samples in Algorithm 1 leaves out a subset A_i of the training data, whereas the two other recursive calls always include all of A_i in their sub-samples. If one considers a hypothesis h trained on the data leaving out A_i, then A_i serves as an independent sample from $\mathcal{D}$. This implies that if h has large error probability over $\mathcal{D}$, then many of the samples in A_i will be classified incorrectly by h. Now, since the two other recursive calls always include A_i, any hypothesis h' trained on a sub-sample from those calls will have margin at least $\gamma/2$ on all points misclassified by h in A_i. But the generalization bound in Theorem 2 then implies that h' makes a mistake only with probability $1/200$ on *the conditional distribution* $\mathcal{D}(\cdot \mid h \text{ errs})$. Thus, the probability that they both err at the same time is at most the probability that h errs, times $1/200$. Applying this reasoning inductively gives the conclusion that it is very unlikely that the majority of all trained hypotheses err at the same time which then finishes the proof.

4 Lower bound

In this section, we sketch the proof of Theorem 2 which gives a lower bound that matches the sample complexity from Theorem 1. The full proof is given in the supplementary material. Informally, Theorem 2 says that there exists a weakly learnable concept class C such that the hypothesis $\mathcal{A}(S)$ of any learning algorithm $\mathcal{A}$ satisfies

$$\mathcal{L}_{\mathcal{D}}(\mathcal{A}(S)) \geq \alpha \left(\frac{d}{\gamma^2 m} + \frac{\ln(1/\delta)}{m} \right).$$

Here, the term $\ln(1/\delta)/m$ follows from previous work. In particular, we could let $C = \mathcal{H}$ and invoke the tight lower bounds for PAC-learning in the realizable setting [4].

Thus, we let $\delta = 1/3$ and only prove that the loss of $\mathcal{A}(S)$ is at least $\alpha d/(\gamma^2 m)$ with probability $1/3$ over S when $|S| = m$ for some weakly learnable concept class C. This proof uses a construction from Grønlund et al. [8] to obtain a hypothesis set $\mathcal{H}$ over a domain $\mathcal{X} = \{x_1, \ldots, x_u\}$ of cardinality $u = \alpha d \gamma^{-2}$ such that a constant fraction of all concepts in $\mathcal{X} \to \{-1, 1\}$ can be γ-weakly learned from $\mathcal{H}$. We then create a distribution $\mathcal{D}$ where the first point x_1 is sampled with probability $1 - u/(4m)$ and with the remaining probability, we receive a uniform sample among $x_2, \ldots, x_u$. The key point is that we only expect to see $1 + m \cdot u/(4m) \approx u/4$ distinct points from $\mathcal{X}$ in a sample S of cardinality m. Thus, if we consider a random concept that can be γ-weakly learned, the labels it assigns to points not in the sample are almost uniform random and independent. This in turn implies that the best any algorithm $\mathcal{A}$ can do is to guess the labels of points in $\mathcal{X} \setminus S$. In that way, $\mathcal{A}$ fails with constant probability if we condition on receiving a sample other than x_1. This happens with probability $u/(4m) = 4\alpha d \gamma^{-2}/m$ and the lower bound follows.

To formally carry out the intuitive argument above, we first argue that for a random concept $c \in C$, the Shannon entropy of c is high, even conditioned on S and the labels $c(S)$. Secondly, we argue that if $\mathcal{A}(S)$ has a small error probability under $\mathcal{D}$, then it must be the case that the hypothesis $\mathcal{A}(S)$ reveals a lot of information about c, i.e. the entropy of c is small conditioned on $\mathcal{A}(S)$. Since $\mathcal{A}(S)$ is a function of S and $c(S)$, the same holds if we condition on S and $c(S)$. This contradicts that c has high entropy and thus we conclude that $\mathcal{A}(S)$ cannot have a small error probability.

5 Conclusion

Overall, we presented a new weak to strong learner with a sample complexity that removes two logarithmic factors from the best-known bound. By accompanying the algorithm with a matching lower bound, we showed that the achieved sample complexity of our algorithm is indeed optimal. Our algorithm uses the same sub-sampling technique as Hanneke [12] and computes a voting classifier with large margins for each sample for example with AdaBoost$^*_\nu$ [19]. The analysis of our algorithm uses a new generalization bound for voting classifiers with large margins.

Although we determined the exact sample complexity of weak to strong learning (up to multiplicative constants), there are a few connected open problems. Currently, our construction uses $3^{\log_4(m)} \approx m^{0.79}$ many sub-samples of linear size as input to AdaBoost$^*_\nu$. For very large datasets, it would be great to reduce the number and size of these calls. We conjecture that the most promising way to do so is to revisit Hanneke's optimal PAC learner and improve the sub-sampling strategy there. This could lead to an improvement for the realizable case as well as to faster weak-to-strong learners.

Next, the output of our algorithm is a majority vote over majority voters. It is unclear whether a simple voter could achieve the same bounds. We believe that a majority of majorities is actually necessary. This is supported by a lower bound showing that there are voters with large margin and poor generalization (paying a logarithmic factor) and thus the learning algorithm has to avoid this 'bad' voter. We currently see no indication of how a variant of AdaBoost could do that.

We have a new generalization bound for large-margin classifiers, which is better than the k-th margin bound (Gao and Zhou [7]) for constant error. Can the k-th margin bound in general be improved, perhaps by one logarithmic factor? One of our key new ideas is the application of the Littlewood-Offord lemma which might also be helpful for the more general case of non-constant error.

Acknowledgments and Disclosure of Funding

This work was supported by the Digital Research Centre Denmark (DIREC) and by the Independent Research Fund Denmark (DFF) Sapere Aude Research Leader grant No 9064-00068B.

References

[1] Peter Bartlett, Yoav Freund, Wee Sun Lee, and Robert E Schapire. 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics* 26, 5 (1998), 1651–1686.

[2] Kristin P Bennett, Ayhan Demiriz, and John Shawe-Taylor. 2000. A column generation algorithm for boosting. In *ICML*. Citeseer, 65–72.

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

[4] Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. 1989. A general lower bound on the number of examples needed for learning. *Information and Computation* 82, 3 (1989), 247–261.

[5] Paul Erdős. 1945. On a lemma of Littlewood and Offord. *Bull. Amer. Math. Soc.* 51, 12 (1945), 898 – 902.

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

[7] Wei Gao and Zhi-Hua Zhou. 2013. On the doubt about margin explanation of boosting. *Artificial Intelligence* 203 (2013), 1–18.

[8] Allan Grønlund, Lior Kamma, Kasper Green Larsen, Alexander Mathiasen, and Jelani Nelson. 2019. Margin-based generalization lower bounds for boosted classifiers. *Advances in Neural Information Processing Systems* 32 (2019).

[9] Allan Grønlund, Lior Kamma, and Kasper Green Larsen. 2020. Margins are Insufficient for Explaining Gradient Boosting. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

[10] Allan Grønlund, Lior Kamma, and Kasper Green Larsen. 2020. Near-Tight Margin-Based Generalization Bounds for Support Vector Machines. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020 (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 3779–3788.

[11] Adam J Grove and Dale Schuurmans. 1998. Boosting in the limit: Maximizing the margin of learned ensembles. In *AAAI/IAAI*. 692–699.

[12] Steve Hanneke. 2016. The optimal sample complexity of PAC learning. *The Journal of Machine Learning Research* 17, 1 (2016), 1319–1333.

[13] Steve Hanneke and Aryeh Kontorovich. 2021. Stable Sample Compression Schemes: New Applications and an Optimal SVM Margin Bound. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (Proceedings of Machine Learning Research, Vol. 132)*. PMLR, 697–721.

[14] Michael Kearns. 1988. Learning Boolean formulae or finite automata is as hard as factoring. *Technical Report TR-14-88 Harvard University Aikem Computation Laboratory* (1988).

[15] Michael Kearns and Leslie Valiant. 1994. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM (JACM)* 41, 1 (1994), 67–95.

[16] Alexander Mathiasen, Kasper Green Larsen, and Allan Grønlund. 2019. Optimal minimal margin maximization with boosting. In *International Conference on Machine Learning*. PMLR, 4392–4401.

[17] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. *Foundations of machine learning*. MIT press.

[18] Gunnar Rätsch and Manfred K Warmuth. 2002. Maximizing the margin with boosting. In *International Conference on Computational Learning Theory*. Springer, 334–350.

[19] Gunnar Rätsch, Manfred K Warmuth, and John Shawe-Taylor. 2005. Efficient Margin Maximizing with Boosting. *Journal of Machine Learning Research* 6, 12 (2005).

[20] Robert E Schapire. 1990. The strength of weak learnability. *Machine learning* 5, 2 (1990), 197–227.

[21] Hans U. Simon. 2015. An Almost Optimal PAC Algorithm. In *Proceedings of The 28th Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 40)*. PMLR, Paris, France, 1552–1563.

[22] Jon Wellner et al. 2013. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] *The main limitations are the points we stated in the open questions paragraph of the conclusion*
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] *Our work is purely theoretic (set within learning theory) and is thus not expected to impact society outside of the scientific community*
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] *Due to space constraints, a large part of the proofs had to be moved to the supplementary material. We still managed to present the largest part of the proof of our new generalization bound for classifiers with large margins within the main body of the paper.*
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Weak to strong learning

The following theorem is essentially a restatement of Theorem 1 from the main paper.

Theorem A.1. *Assume we are given access to a γ-weak learner for a $0 < \gamma < 1/2$, using base hypothesis set $\mathcal{H} \subseteq \mathcal{X} \to \{-1, 1\}$ of VC-dimension d. Then there is a universal constant $\alpha > 0$ and an algorithm $\mathcal{A}$, such that for every $0 < \delta < 1$ and every distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}$, it holds with probability at least $1 - \delta$ over a set of m samples $S \sim \mathcal{D}^m$, that $\mathcal{A}$ on S outputs a classifier $h_S = \mathcal{A}(S) \in \mathcal{X} \to \{-1, 1\}$ with*

$$\mathcal{L}_{\mathcal{D}}(h_S) \;\leq\; \alpha \cdot \frac{d\gamma^{-2} + \ln(1/\delta)}{m}.$$

Theorem 1 of the main paper follows by setting $\varepsilon = \mathcal{L}_{\mathcal{D}}(h_S)$ and solving for m and letting the label in the distribution $\mathcal{D}$ be $c(x)$ for every $x \in \mathcal{X}$. The algorithm that obtains the guarantees has been described in the main paper. We thus only present (again) the two algorithms (Algorithm 1 and Algorithm 2), as well as AdaBoost$^*_\nu$ (Algorithm A.1) by Rätsch et al. [19] that achieves almost optimal margins and is used in Algorithm 2.

In the remainder of the section, we prove that Algorithm 2 has the guarantees of Theorem A.1. The proof follows that of Hanneke [12] pretty much uneventfully, although carefully using that a generalization error of $1/200$ suffices. For simplicity, we assume m is a power of 4. This can easily be ensured by rounding m down to the nearest power of 4 and ignoring all excess samples. This only affects the generalization bound by a constant factor. With m being a power of 4 we can observe from Algorithm 1 that the cardinalities of all recursively generated sets A_0 (which are the input to the next level of the recursion) are also powers of 4. Hence we can ignore all roundings.

A.1 Proof of Optimal Strong Learning

Let $C \subseteq \mathcal{X} \to \{-1, 1\}$ be a concept class and assume there is a γ-weak learner for C using hypothesis set $\mathcal{H}$ of VC-dimension d. Let $\mathcal{A}^*_\nu$ be an algorithm that on a sample S consistent with a concept $c \in C$, computes a voting classifier $f \in \Delta(\mathcal{H})$ with $yf(x) \geq \gamma/2$ for all $(x, y) \in S$ and returns as its output hypothesis $g(x) = \text{sign}(f(x))$. We could e.g. let $\mathcal{A}^*_\nu$ be AdaBoost$^*_\nu$. For a sample S, we use the notation $\mathcal{M}_\gamma(S)$ to denote the set of hypotheses $g(x) = \text{sign}(f(x))$ for an $f \in \Delta(\mathcal{H})$ satisfying $yf(x) \geq \gamma$ for all $(x, y) \in S$. The set $\mathcal{M}_\gamma(S)$ is thus the set of all voting classifiers obtained by taking the sign of a voter that has margins at least γ on all samples in S. By definition, the output hypothesis g of $\mathcal{A}^*_\nu$ on a set of samples S always lies in $\mathcal{M}_{\gamma/2}(S)$.

Let $c \in C$ be an unknown concept in C and let $\mathcal{D}$ be an arbitrary distribution over $\mathcal{X}$. Let $S = \{(x_i, c(x_i))\}_{i=1}^m \in (\mathcal{X} \times \{-1, 1\})^m$ be a set of m samples with each x_i an i.i.d. sample from $\mathcal{D}$. Let $S_{1:k}$ denote the first k samples of S. Let $c' \geq 4$ be a constant to be determined later. We will prove by induction that for every $m' \in \mathbb{N}$ that is a power of 4, for every $\delta' \in (0, 1)$, and every finite sequence B' of samples in $\mathcal{X} \times \{-1, 1\}$ with $y_i = c(x_i)$ for each $(x_i, y_i) \in B'$, with probability at least $1 - \delta'$, the classifier

$$\hat{h}_{m',B'} \;=\; \text{sign}\left(\sum_{C_i \in \textit{Sub-Sample}(S_{1:m'}, B')} \mathcal{A}^*_\nu(C_i) \right)$$

satisfies

$$\mathcal{L}_{\mathcal{D}}(\hat{h}_{m',B'}) \;\leq\; \frac{c'}{m'} \left(d\gamma^{-2} + \ln(1/\delta') \right). \tag{3}$$

The conclusion of Theorem A.1 follows by letting $B' = \emptyset$ and $m' = m$ (and recalling that we assume m is a power of 4). Thus what remains is to give the inductive proof.

As the base case, consider any $m' \in \mathbb{N}$ with $m' \leq c'$ and m' a power of 4. In this case, the bound $c'(d\gamma^{-2} + \ln(1/\delta'))/m'$ is at least $d\gamma^{-2} \geq 1$ and $\mathcal{L}_{\mathcal{D}}(\hat{h}_{m',B'}) \leq 1$ obviously holds.

For the inductive step, take as inductive hypothesis that, for some $m \in \mathbb{N}$ with $m > c'$ and m a power of 4, it holds for all $m' \in \mathbb{N}$ with $m' < m$ and m' a power of 4, that for every $\delta' \in (0, 1)$ and every

Algorithm A.1: AdaBoost$^*_\nu$ [19]

Input: training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$
number of rounds T
desired accuracy ν

Result: An ensemble hypothesis H_{out} with almost optimal margins

$\mathcal{D}^{(1)} \leftarrow \left(\frac{1}{m}, \ldots \frac{1}{m}\right)$ // uniform initialization of $\mathcal{D}$
for $t \in \{1, \ldots, T\}$ **do**
  $h_t \leftarrow \text{WL}(\mathcal{D}^{(t)}, S)$ // invoke weak learner
  $\gamma_t \leftarrow \sum_{i=1}^m \mathcal{D}_i^{(t)} y_i h_t(x_i)$ // average margin of h_t
  if $|\gamma_t| = 1$ **then**
    /* h_t is consistent $\Rightarrow$ taking only h_t as 'ensemble' maximizes the margin */
    $w_1 \leftarrow \text{sign}(\gamma_t), \quad h_1 \leftarrow h_t, \quad T \leftarrow 1$
    break
  $\gamma_t^{\min} \leftarrow \min_{r \in [t]} \gamma_r,$ // update assumed advantage
  $\rho_t^{\min} \leftarrow \gamma_t^{\min} - \nu$
  $w_t = \frac{1}{2} \ln \frac{1+\gamma_t}{1-\gamma_t} - \frac{1}{2} \ln \frac{1+\rho_t}{1-\rho_t}$ // weight for the current hypothesis
  for $i \in \{1, \ldots, m\}$ **do**
    $\mathcal{D}_i^{(t+1)} \leftarrow \frac{\mathcal{D}_i^{(t)} \exp\left(-w_t y_i h_t(x_i)\right)}{\sum_{j=1}^m \mathcal{D}_i^{(t)} \exp\left(-w_t y_j h_t(x_j)\right)}$ // update $\mathcal{D}$
return $f_{out}(x) = \frac{1}{\sum_{i=1}^T |w_i|} \sum_{t=1}^T w_t h_t(x)$ // (normalized) weighted majority vote

finite sequence B' of samples in $\mathcal{X} \times \{-1, 1\}$ with $y_i = c(x_i)$ for each $(x_i, y_i) \in B'$, with probability at least $1 - \delta'$, Eq. (3) holds. We need to prove that the inductive hypothesis also holds for $m' = m$.

Fix a $\delta \in (0, 1)$ and any finite sequence B of points in $\mathcal{X} \times \{-1, 1\}$ with $y_i = c(x_i)$ for each (x_i, y_i) in B. Since $m > c' \geq 4$, we have that *Sub-Sample*$(S_{1:m}, B)$ returns in Step 5 of Algorithm 1. Let A_0, A_1, A_2, A_3 be as defined in Step 4 of Algorithm 1. Also define $B_1 = A_2 \cup A_3 \cup B$, $B_2 = A_1 \cup A_3 \cup B$, $B_3 = A_1 \cup A_2 \cup B$, and for each $i \in \{1, 2, 3\}$, denote

$$h_i = \text{sign}\left(\sum_{C_i \in \textit{Sub-Sample}(A_0, B_i)} \mathcal{A}^*_\nu(C_i)\right).$$

Note that the h_i's correspond to the majority vote classifiers trained on the sub-samples of the three recursive calls in Algorithm 1. Moreover, notice that $h_i = \hat{h}_{m/4, B_i}$. Therefore, the inductive hypothesis may be used on h_1, h_2, h_3 to conclude that for each $i \in \{1, 2, 3\}$, there is an event E_i of probability at least $1 - \delta/9$, on which

$$\mathcal{L}_{\mathcal{D}}(h_i) \leq \frac{c'}{|A_0|}\left(d\gamma^{-2} + \ln(9/\delta)\right) \leq \frac{4c'}{m}\left(d\gamma^{-2} + \ln(1/\delta) + 3\right) \leq \frac{12c'}{m}\left(d\gamma^{-2} + \ln(1/\delta)\right). \quad (4)$$

Here we chose the probability $1 - \delta/9$ in order to perform a union bound in the end of the induction step which is possible since the inductive hypothesis holds for every δ'. Next, define $\text{Err}(h_i)$ as the set of points $x \in \mathcal{X}$ for which $h_i(x) \neq c(x)$. Now fix an $i \in \{1, 2, 3\}$ and denote by $\{(Z_{i,1}, c(Z_{i,1})), \ldots, (Z_{i,N_i}, c(Z_{i,N_i})\} = A_i \cap (\text{Err}(h_i) \times \{-1, 1\})$, where $N_i = |A_i \cap (\text{Err}(h_i) \times \{-1, 1\})|$. Said in words, the set $\{(Z_{i,j}, c(Z_{i,j})\}_{j=1}^{N_i}$ is the subset of samples in A_i on which h_i makes a mistake. Notice that h_i is not trained on any samples from A_i (B_i excludes A_i), hence h_i and A_i are independent. Therefore, given h_i and N_i, the samples $Z_{i,1}, \ldots, Z_{i,N_i}$ are conditionally independent samples with distribution $\mathcal{D}(\cdot \mid \text{Err}(h_i))$ (provided $N_i > 0$). From Theorem 6 in the main paper, we get that there is an event E_i' of probability at least $1 - \delta/9$, such that if $N_i \geq c''\big(d\gamma^{-2} + \ln(1/\delta)\big)$, then every $h \in \mathcal{M}_{\gamma/2}\big(\{(Z_{i,j}, c(Z_{i,j}))\}_{j=1}^{N_i}\big)$ satisfies

$$\mathcal{L}_{\mathcal{D}(\cdot \mid \text{Err}(h_i))}(h) \leq \tfrac{1}{200}.$$

Note that this is a key step where our proof differs from Hanneke's original proof since we exploit that a bound of $\frac{1}{200}$ on the generalization error suffices for the rest of the proof. We continue by observing

that for each $j \in \{1,2,3\} \setminus \{i\}$, the set B_j contains A_i and this remains the case in all recursive calls of *Sub-Sample*(A_0, B_i). Thus for $\{C_1, \ldots, C_k\} =$ *Sub-Sample*(A_0, B_j), it holds for all C_k that $\mathcal{A}^*_\nu(C_k) \in \mathcal{M}_{\gamma/2}(B_j) \Rightarrow \mathcal{A}^*_\nu(C_k) \in \mathcal{M}_{\gamma/2}(A_0) \Rightarrow \mathcal{A}^*_\nu(C_k) \in \mathcal{M}_{\gamma/2}\big(\{(Z_{i,j}, c(Z_{i,j}))\}_{j=1}^{N_i}\big)$. Thus on the event E'_i defined above, if $N_i > c''\big(d\gamma^{-2} + \ln(1/\delta)\big)$, then it holds for all $j \in \{1,2,3\} \setminus \{i\}$ and all $C_k \in$ *Sub-Sample*(A_0, B_j), that the hypothesis $h = \mathcal{A}^*_\nu(C_k)$ satisfies

$$\begin{aligned}\Pr_{x\sim\mathcal{D}}\left[h_i(x) \neq c(x) \wedge h(x) \neq c(x)\right] &= \mathcal{L}_{\mathcal{D}}(h_i) \cdot \mathcal{L}_{\mathcal{D}(\cdot|\mathrm{Err}(h_i))}(h) \\ &\leq \tfrac{1}{200}\,\mathcal{L}_{\mathcal{D}}(h_i).\end{aligned}$$

Assume now that $\mathcal{L}_{\mathcal{D}}(h_i) \geq \big((10/7)c''(d\gamma^{-2} + \ln(1/\delta)) + 23\ln(9/\delta)\big)/(m/4) \geq 23\ln(9/\delta)/|A_i|$. Using that h_i and A_i are independent, it follows by a Chernoff bound that

$$\begin{aligned}\Pr\left[N_i \geq (7/10)\mathcal{L}_{\mathcal{D}}(h_i)|A_i|\right] &\geq 1 - \exp\left(-(3/10)^2\mathcal{L}_{\mathcal{D}}(h_i)|A_i|/2\right) \\ &\geq 1 - \exp\left(-(3/10)^2 \cdot 23\ln(9/\delta)/2\right) \\ &> 1 - \delta/9.\end{aligned}$$

Thus there is an event E''_i of probability at least $1 - \delta/9$, on which, if

$$\mathcal{L}_{\mathcal{D}}(h_i) \;\geq\; \frac{(10/7)c''(d\gamma^{-2} + \ln(1/\delta)) + 23\ln(9/\delta)}{m/4}$$

then

$$\begin{aligned}N_i &\geq (7/10)\mathcal{L}_{\mathcal{D}}(h_i)\,|A_i| \\ &= (7/10)\mathcal{L}_{\mathcal{D}}(h_i)\,m/4 \\ &\geq c''\big(d\gamma^{-2} + \ln(1/\delta)\big).\end{aligned}$$

Combining it all, we have that on the event $E_i \cap E'_i \cap E''_i$, which occurs with probability at least $1 - \delta/3$, if $\mathcal{L}_{\mathcal{D}}(h_i) \geq \big((10/7)c''(d\gamma^{-2} + \ln(1/\delta)) + 23\ln(9/\delta)\big)/(m/4)$, then every $h = \mathcal{A}^*_\nu(C_k)$ for a $C_k \in$ *Sub-Sample*(A_0, B_j) with $j \neq i$ has:

$$\Pr_{x\sim\mathcal{D}}\left[h_i(x) \neq c(x) \wedge h(x) \neq c(x)\right] \;\leq\; \tfrac{1}{200}\mathcal{L}_{\mathcal{D}}(h_i)$$

By Eq. (4), this is at most

$$\begin{aligned}\Pr_{x\sim\mathcal{D}}\left[h_i(x) \neq c(x) \wedge h(x) \neq c(x)\right] &\leq \frac{1}{200} \cdot \frac{12c'}{m}\left(d\gamma^{-2} + \ln(1/\delta)\right) \\ &\leq \frac{c'}{16m}\left(d\gamma^{-2} + \ln(1/\delta)\right).\end{aligned}$$

On the other hand, if $\mathcal{L}_{\mathcal{D}}(h_i) < \big(c''(d\gamma^{-2} + \ln(1/\delta)) + 23\ln(9/\delta)\big)/(m/4)$, then

$$\begin{aligned}\Pr_{x\sim\mathcal{D}}\left[h_i(x) \neq c(x) \wedge h(x) \neq c(x)\right] &\leq \mathcal{L}_{\mathcal{D}}(h_i) \\ &\leq \big(c''(d\gamma^{-2} + \ln(1/\delta)) + 23\ln(9/\delta)\big)/(m/4) \\ &\leq 4c''(d\gamma^{-2} + 24\ln(1/\delta) + 23\ln 9)/m\end{aligned}$$

Using that $23 \cdot \ln 9 < 51 \leq 51d\gamma^{-2}$, the above is at most $204c''(d\gamma^{-2} + \ln(1/\delta))/m$. Fixing the constant c' to $c' \geq (16 \cdot 204)c''$, this is at most

$$\frac{c'}{16m}\left(d\gamma^{-2} + \ln(1/\delta)\right).$$

We conclude that on the event $\bigcap_{i=1,2,3}\{E_i \cap E'_i \cap E''_i\}$, which occurs with probability at least $1 - \delta$ by a union bound, it holds for all i and all $C_k \in$ Sub-Sample(A_0, B_j) with $j \neq i$ that the hypothesis $h = \mathcal{A}^*_\nu(C_k)$ satisfies:

$$\Pr_{x\sim\mathcal{D}}\left[h_i(x) \neq c(x) \wedge h(x) \neq c(x)\right] \;\leq\; \frac{c'}{16m}\left(d\gamma^{-2} + \ln(1/\delta)\right).$$

Now consider an x on which $\hat{h}_{m,B}$ errs. On such an x, the majority among the classifiers

$$\bigcup_{C_i \in \textit{Sub-Sample}(S_{1:m},B)} \left\{\mathcal{A}^*_\nu(C_i)\right\} \;=\; \bigcup_{i=1,2,3} \;\; \bigcup_{C_k \in \textit{Sub-Sample}(S_{1:m/4},B_i)} \left\{\mathcal{A}^*_\nu(C_k)\right\}$$

errs. For the majority to err, there must be an $i \in \{1,2,3\}$ for which the majority of

$$\bigcup_{C_k \in \textit{Sub-Sample}(S_{1:m/4},B_i)} \left\{\mathcal{A}^*_\nu(C_k)\right\}$$

errs. This is equivalent to $h_i(x) \neq c(x)$. Furthermore, even when all of the classifiers in

$$\bigcup_{C_k \in \textit{Sub-Sample}(S_{1:m/4},B_i)} \left\{\mathcal{A}^*_\nu(C_k)\right\}$$

err, there still must be another $(1/6)$-fraction of all the classifiers

$$\bigcup_{i=1,2,3} \;\; \bigcup_{C_k \in \textit{Sub-Sample}(S_{1:m/4},B_i)} \left\{\mathcal{A}^*_\nu(C_k)\right\}$$

that err. This follows since each of the three recursive calls in *Sub-Sample* generated equally many classifiers/samples. It follows that if we pick a uniform random $i \in \{1,2,3\}$ and a uniform random hypothesis h in

$$\bigcup_{j \in \{1,2,3\}\setminus\{i\}} \;\; \bigcup_{C_k \in \textit{Sub-Sample}(S_{1:m/4},B_j)} \left\{\mathcal{A}^*_\nu(C_k)\right\}$$

then with probability at least $(1/3)(1/6)(3/2) = 1/12$, we have that $h_i(x) \neq c(x) \wedge h(x) \neq c(x)$. It follows by linearity of expectation that on the event $\bigcap_{i=1,2,3}\{E_i \cap E'_i \cap E''_i\}$, we have:

$$\mathcal{L}_{\mathcal{D}}(\hat{h}_{m,B}) \;\leq\; 12 \cdot \frac{c'}{16m}\left(d\gamma^{-2} + \ln(1/\delta)\right) \;<\; \frac{c'}{m}\left(d\gamma^{-2} + \ln(1/\delta)\right).$$

This completes the inductive proof and shows Theorem A.1.

B Lower bound

In this section, we prove the following lower bound which directly implies Theorem 2 from the main paper:

Theorem B.1. *There is a universal constant $\alpha > 0$ such that for all integers $d \in \mathbb{N}$ and every $0 < \gamma < 1/80$, there is a finite set $\mathcal{X}$, a concept class $\mathcal{C} \subset \mathcal{X} \to \{-1,1\}$ and a hypothesis set $\mathcal{H} \subseteq \mathcal{X} \to \{-1,1\}$ of VC-dimension at most d, such that for every integer $m \in \mathbb{N}$ and $0 < \delta < 1/3$, there is a distribution $\mathcal{D}$ over $\mathcal{X}$ such that the following holds:*

1. *For every $c \in \mathcal{C}$ and every distribution $\mathcal{D}'$ over $\mathcal{X}$, there is an $h \in \mathcal{H}$ with*
$$\Pr_{x \sim \mathcal{D}'}\left[h(x) \neq c(x)\right] \;\leq\; 1/2 - \gamma.$$
2. *For any algorithm $\mathcal{A}$, there is a concept $c \in \mathcal{C}$ such that with probability at least δ over a set of m samples $S \sim \mathcal{D}^m$, the classifier $\mathcal{A}(S) \in \mathcal{X} \to \{-1,1\}$ produced by $\mathcal{A}$ on S and $c(S)$ must have*
$$\mathcal{L}_{\mathcal{D}}(\mathcal{A}(S)) \;\geq\; \alpha \cdot \frac{d\gamma^{-2} + \ln(1/\delta)}{m}.$$

Theorem B.1 immediately implies Theorem 2 by solving the equation in the second statement for $\varepsilon = \mathcal{L}_{\mathcal{D}}(\mathcal{A}(S))$.

The proof of the term $\ln(1/\delta)/m$ in the lower bound follows from previous work. In particular, we could let $\mathcal{C} = \mathcal{H}$ and invoke the tight lower bounds for PAC-learning in the realizable setting [4]. Thus, we focus on $\delta = 1/3$ and only need to prove that the loss of $\mathcal{A}(S)$ is at least $\alpha d\gamma^{-2}/m$ with probability $1/3$ over S.

For the proof, we make use of the following lemma by Grønlund et al. [8] to construct the 'hard' hypothesis set $\mathcal{H}$ and concept class $\mathcal{C}$:

Lemma B.1 (Grønlund et al. [8]). *For every $\gamma \in (0, 1/40), \delta \in (0, 1)$ and integers $k \leq u$, there exists a distribution $\mu = \mu(u, d, \gamma, \delta)$ over a hypothesis set $\mathcal{H} \subset \mathcal{X} \to \{-1, 1\}$, where $\mathcal{X}$ is a set of size u, such that the following holds.*

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

$$\Pr_{\mathcal{H}\sim\mu}\left[\exists f \in \Delta(\mathcal{H}) : \forall x \in \mathcal{X} : \ell(x)f(x) \geq \gamma\right] \;\geq\; 1 - \delta.$$

where $N = \Theta\big(\gamma^{-2} \ln u \ln(\gamma^{-2} \ln u\delta^{-1}) e^{\Theta(\gamma^2 k)}\big)$.

To prove Theorem B.1 for a given $\gamma \in (0, 1/80)$ and $m, d \in \mathbb{N}$, let $u = k$ for a u to be determined. Invoke Lemma B.1 with $\delta = 1/2$ and $\gamma' = 2\gamma$ to conclude that there exists a hypothesis set $\mathcal{H}$ such that among all labelings $\ell \in \{-1, 1\}^u$, at least half of them satisfy:

$$\exists f \in \Delta(\mathcal{H}) : \forall x \in \mathcal{X} : \; \ell(x)f(x) \geq 2\gamma.$$

Moreover, we have $N = |\mathcal{H}| = \Theta\big(\gamma^{-2} \ln u \ln(\gamma^{-2} \ln u) e^{\Theta(\gamma^2 u)}\big)$. Let the concept class $\mathcal{C}$ be the set of such labelings.

For the given VC-dimension d, we need to bound the VC-dimension of $\mathcal{H}$ by d. For this, note that the VC-dimension is bounded by $\lg |\mathcal{H}| = \Theta(\gamma^2 u)$. We thus choose $u = \Theta(\gamma^{-2} d)$ which implies the claimed VC-dimension of $\mathcal{H}$.

Next, we have to argue that any concept $c \in \mathcal{C}$ can be γ-weakly learned from $\mathcal{H}$. That is, the first statement of Theorem B.1 holds for $\mathcal{H}, \mathcal{C}$. To see this, we must show that for every distribution $\mathcal{D}$ over $\mathcal{X}$, there is a hypothesis $h \in \mathcal{H}$ such that $\Pr_{x\sim\mathcal{D}}[h(x) = c(x)] \geq 1/2 + \gamma$. To argue that this is indeed the case, let $f \in \Delta(\mathcal{H})$ satisfy $\forall x \in \mathcal{X} : \; c(x)f(x) \geq 2\gamma$. Such an f exists by definition of $\mathcal{C}$. Then, $\mathbb{E}_{x\sim\mathcal{D}}[c(x)f(x)] \geq 2\gamma$. Since $f(x)$ is a convex combination of hypotheses from $\mathcal{H}$, it follows that there is a hypothesis $h \in \mathcal{H}$ also satisfying $\mathbb{E}_{x\sim\mathcal{D}}[c(x)h(x)] \geq 2\gamma$. But

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[c(x)h(x)] \;&=\; \sum_{x\in\mathcal{X}} \mathcal{D}(x)c(x)h(x) \\
&=\; \sum_{x\in\mathcal{X}:\, c(x)=h(x)} \mathcal{D}(x) \;-\; \sum_{x\in\mathcal{X}:\, c(x)\neq h(x)} \mathcal{D}(x) \\
&=\; \Pr_{x\sim\mathcal{D}}[c(x) = h(x)] - \Pr_{x\sim\mathcal{D}}[c(x) \neq h(x)] \\
&=\; \Pr_{x\sim\mathcal{D}}[c(x) = h(x)] - (1 - \Pr_{x\sim\mathcal{D}}[c(x) = h(x)]) \\
&=\; 2 \Pr_{x\sim\mathcal{D}}[c(x) = h(x)] - 1.
\end{aligned}$$

Hence, $2 \cdot \Pr_{x\sim\mathcal{D}}[c(x) = h(x)] - 1 \geq 2\gamma \implies \Pr_{x\sim\mathcal{D}}[c(x) = h(x)] \geq 1/2 + \gamma$ as claimed.

We have thus constructed $\mathcal{H}$ and $\mathcal{C}$ satisfying the first statement of Theorem B.1, where $\mathcal{C}$ contains at least half of all possible labelings of the points $\mathcal{X} = \{x_1, \ldots, x_u\}$ with $u = \Theta(\gamma^{-2} d)$. For the remainder of the proof, we assume u is at least some large constant, which is true for γ small enough.

What remains is to establish the second statement of Theorem B.1. For this, we first define the hard distribution $\mathcal{D}$ over $\mathcal{X}$. The distribution $\mathcal{D}$ returns the point x_1 with probability $1 - (u - 1)/4m$ and with the remaining probability $(u - 1)/4m$ it returns a uniform random sample x_i among $x_2, \ldots, x_u$. Also, let c be a uniform random concept drawn from $\mathcal{C}$.

Let $\mathcal{A}$ be any (possibly randomized) learning algorithm that on a set of samples S from $\mathcal{X}$ and a labeling $\ell(S)$ of S that is consistent with at least one concept $c \in \mathcal{C}$ (i.e. $\ell(S) = c(S)$), outputs a hypothesis $h_{S,\ell(S)}$ in $\mathcal{X} \to \{-1, 1\}$. The algorithm $\mathcal{A}$ is not constrained to output a hypothesis from $\Delta(\mathcal{H})$ or $\mathcal{H}$, but instead may output any desirable hypothesis in $\mathcal{X} \to \{-1, 1\}$, using the full knowledge of $\mathcal{C}, \ell(S), \mathcal{H}$ and the promise that $c \in \mathcal{C}$. Our goal is to show that

$$\mathop{\mathbb{E}}_{c\sim\mathcal{C}}\left[\Pr_{S\sim\mathcal{D}^m}\left[\Pr_{x\sim\mathcal{D}}[h_{S,c(S)}(x) \neq c(x)] \geq \alpha' \frac{d\gamma^{-2}}{m}\right]\right] \geq 1/3 \tag{5}$$

where $c \sim C$ denotes the uniform random choice of c. Notice that if this is the case, there must exist a concept c for which

$$\Pr_{S\sim\mathcal{D}^m}\left[\Pr_{x\sim\mathcal{D}}[h_{S,c(S)}(x) \neq c(x)] \geq \alpha'\frac{d\gamma^{-2}}{m}\right] \geq 1/3.$$

To establish Eq. (5), we start by observing that for any randomized algorithm $\mathcal{A}$, there is a deterministic algorithm $\mathcal{A}'$ obtaining a smaller than or equal value of the left hand side of Eq. (5) (by Yao's principle). Thus, we assume from here on that $\mathcal{A}$ is deterministic.

The main idea in our proof is to first show that conditioned on the set S and label $c(S)$, the concept c is still largely unknown. We formally measure this by arguing that the binary Shannon entropy of c is large conditioned on S and $c(S)$. Next, we argue that if a learning algorithm often manages to produce an accurate hypothesis from S and $c(S)$, then that reveals a lot of information about c, i.e. the entropy of c is small conditioned on S and $c(S)$. This contradicts the first statement and thus the algorithm cannot produce an accurate hypothesis. We now proceed with the two steps.

Large Conditional Entropy. Consider the binary Shannon entropy of the uniform random c conditioned on S and $c(S)$, denoted $H(c \mid S, c(S))$. We know that $H(c) = \lg|C| \geq \lg(2^u/2) = u-1$. The random variable c is independent of S, hence $H(c \mid S) = H(c)$. We therefore have $H(c \mid S, c(S)) \geq H(c \mid S) - H(c(S) \mid S) = u - 1 - H(c(S) \mid S)$. For a fixed $s \in \mathcal{X}^m$, let $p_s = \Pr_{S\sim\mathcal{D}^m}[S = s]$. Then $H(c(S) \mid S) = \sum_{s\in\mathcal{X}^m} p_s H(c(S) \mid S = s) \leq \sum_{s\in\mathcal{X}^m} p_s|s|$, where the last step follows from the fact that, conditioned on s, the labeling $c(s)$ consists of $|s|$ signs. Note that the size of the set $|s|$ is possibly smaller than m due to repetitions.

Now notice that $\Pr[|S| > u/3]$ is exponentially small in u since each of the m samples from $\mathcal{D}$ is among $x_2, \ldots, x_u$ with probability only $(u-1)/(4m)$. Therefore, we get $H(c(S) \mid S) \leq u/3 + \exp(-\Omega(u))u \leq u/2 - 1$. It follows that

$$H(c \mid S, c(S)) \;\geq\; u - 1 - (u/2 - 1) \;=\; u/2. \tag{6}$$

Accuracy Implies Low Entropy. Now assume that $h_{S,c(S)}$ is such that $\Pr_{x\sim\mathcal{D}}[h_{S,c(S)} \neq c(x)] < \alpha' d\gamma^{-2}/m$ for a sufficiently small constant α'. Any point x_i where $c(x_i)$ disagrees with $h_{S,c(S)}(x_i)$ adds at least $1/(4m)$ to $\Pr_{x\sim\mathcal{D}}[h_{S,c(S)} \neq c(x)]$ (the point x_1 would add more), hence $h_{S,c(S)}$ makes a mistake on at most $\alpha' d\gamma^{-2}/m \cdot (4m) = 4\alpha' d\gamma^{-2}$ points. Recalling that $u = \Theta(d\gamma^{-2})$, we get that for α' small enough, this is less than $u/100$. Thus, conditioned on $\Pr_{x\sim\mathcal{D}}[h_{S,c(S)} \neq c(x)] < \alpha' d\gamma^{-2}/m$ and $h_{S,c(S)}$, we get that the entropy of the concept c is no more than $\lg\left(\sum_{i=0}^{u/100}\binom{u}{i}\right)$ since c is within a Hamming ball of radius $u/100$ from $h_{S,c(S)}$. Now $\sum_{i=0}^{u/100}\binom{u}{i} \leq 2^{H_b(1/100)u}$, where H_b is the binary entropy of a Bernoulli random variable with success probability $1/100$. Numerical calculations give $H_b(1/100) = (1/100)\lg_2(100) + (99/100)\lg_2(100/99) < 0.09$. Thus

$$H\Big(c \;\Big|\; h_{S,c(S)},\; \Pr_{x\sim\mathcal{D}}[h_{S,c(S)} \neq c(x)] < \alpha' d\gamma^{-2}/m\Big) \;\leq\; 0.09u. \tag{7}$$

Now let $X_{S,c}$ be an indicator random variable for the event that $\Pr_{x\sim\mathcal{D}}[h_{S,c(S)} \neq c(x)] < \alpha' d\gamma^{-2}/m$. Then $H(c \mid S, c(S)) \leq H(c \mid S, c(S), h_{S,c(S)}, X_{S,c}) + H(X_{S,c})$. Here we remark that we add $h_{S,c(S)}$ in the conditioning for free since it depends only on S and $c(S)$. Adding $X_{S,c}$ costs at most its entropy which satisfies $H(X_{S,c}) \leq 1$. Since removing variables that we condition on only increases entropy, we get $H(c \mid S, c(S)) \leq H(c \mid h_{S,c(S)}, X_{S,c}) + 1$. Now observe that $H(c \mid h_{S,c(S)}, X_{S,c}) = \Pr[X_{S,c} = 1]H(c \mid h_{S,c(S)}, X_{S,c} = 1) + \Pr[X_{S,c} = 0]H(c \mid h_{S,c(S)}, X_{S,c} = 0)$. The latter entropy we simply bound by u and the former is bounded by $0.09u$ by Eq. (7). Thus $H(c \mid S, c(S)) \leq 1 + \Pr[X_{S,c} = 1]0.09u + (1 - \Pr[X_{S,c} = 1])u$.

Combining the Bounds. Combining the above with Eq. (6) we conclude that

$$1 + \Pr[X_{S,c} = 1]0.09u + (1 - \Pr[X_{S,c} = 1])u \;\geq\; u/2.$$

It follows that $\Pr[X_{S,c} = 1] \leq 2/3$. This completes the proof since

$$\mathop{\mathbb{E}}_{c\sim C}\left[\Pr_{S\sim\mathcal{D}^m}\left[\Pr_{x\sim\mathcal{D}}[h_{S,c(S)}(x) \neq c(x)] \geq \alpha'\frac{d\gamma^{-2}}{m}\right]\right] = \mathop{\mathbb{E}}_{c\sim C}\Big[\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[(1 - X_{S,c})]\Big] \;=\; 1 - \Pr[X_{S,c} = 1]$$

and thus

$$\mathop{\mathbb{E}}_{c\sim\mathcal{C}}\left[\Pr_{S\sim\mathcal{D}^m}\left[\Pr_{x\sim\mathcal{D}}[h_{S,c(S)}(x)\neq c(x)]\geq \alpha'\frac{d\gamma^{-2}}{m}\right]\right] \;\geq\; \frac{1}{3}.$$

This finishes the proof of Theorem B.1.

C Proofs for the margin-based generalization bound for voting classifiers

This section covers the proofs of all lemmas needed to show the generalization bound for voting classifiers with large margins (Theorem 4 in the main paper) that did not fit into the main text.

C.1 Proofs of key properties of $\mathcal{D}_{f,t}$

First, we present the proofs of Lemma 1, 2, and 3 from the main paper covering different properties of the distribution $\mathcal{D}_{f,t}$.

Restatement of Lemma 1. *For any $x\in\mathcal{X}$, any $f\in\Delta(\mathcal{H})$ and any $\mu>0$:*

$$\Pr_{g\sim\mathcal{D}_{f,t}}\left[|f(x)-g(x)|\geq\mu\right] \;<\; 5\exp(-\mu^2 t/32).$$

Proof. This lemma follows using standard concentration inequalities: In the first step of sampling g from $\mathcal{D}_{f,t}$, where we draw t i.i.d. hypotheses, it follows from Hoeffding's inequality that the hypothesis $g'(x)=(1/t)\sum_{i=1}^t h_i'(x)$ satisfies

$$\Pr_{g'}\left[|f(x)-g'(x)|\geq\mu/2\right] \;\leq\; 2\exp\big(-2(\mu/2)^2t^2/(4t)\big) \;=\; 2\exp(-\mu^2t/8).$$

In the second step, we first get by a Chernoff bound that $\Pr[t'<t/4]<\exp(-t/16)$. Secondly, let us condition on any fixed value of t' that is at least $t/4$. Then $h_1,\ldots,h_{t'}$ is a uniform sample without replacement from $h_1',\ldots,h_t'$. It follows by a Hoeffding bound without replacement that

$$\Pr\left[|g(x)-g'(x)|\geq\mu/2\right] \;\leq\; 2\exp\big(-2(\mu/2)^2(t')^2/(4t')\big) \;<\; 2\exp(-\mu^2t/32).$$

In total, we conclude that

$$\Pr\left[|f(x)-g(x)|\geq\mu\right] \;<\; 2\exp(-\mu^2t/8)+\exp(-t/16)+2\exp(-\mu^2t/32) \;<\; 5\exp(-\mu^2t/32). \quad\square$$

Restatement of Lemma 2. *For any $x\in\mathcal{X}$, any $f\in\Delta(\mathcal{H})$ and any $\mu\geq 1/t$:*

$$\Pr_{g\sim\mathcal{D}_{f,t}}\left[|g(x)|\leq\mu\right] \;\leq\; 2\mu\sqrt{t}.$$

Proof. Let $h_1',\ldots,h_t'$ be the hypotheses sampled in the first step of drawing g. Define σ_i to be 1 if h_i' is sampled in g and -1 otherwise. That is, we have

$$g(x)=\frac{1}{|\{i:\sigma_i=1\}|}\sum_{i:\sigma_i=1}h_i'(x).$$

Let $\Gamma=\sum_{i=1}^t h_i'(x)$. Then

$$\begin{aligned}
&\Gamma+\sum_{i=1}^t\sigma_i h_i'(x)\\
=\;&\sum_{i:\sigma_i=1}h_i'(x)+\sum_{i:\sigma_i=-1}h_i'(x)+\sum_{i=1}^t\sigma_i h_i'(x)\\
=\;&2\sum_{i:\sigma_i=1}h_i'(x)=2t'g(x).
\end{aligned}$$

Therefore, $|g(x)|\leq\mu$ if and only if

$$\left|\frac{\Gamma+\sum_i\sigma_i h_i'(x)}{2t'}\right|\leq\mu.$$

Since $t' \leq t$, this implies

$$\left|\frac{\Gamma + \sum_i \sigma_i h'_i(x)}{2t}\right| \leq \mu.$$

Hence, we have $\Pr[|g(x)| \leq \mu] \leq \Pr[\sum_i \sigma_i h'_i(x) \in -\Gamma \pm 2t\mu]$. By Erdös' improved Littlewood-Offord lemma, as long as $2t\mu \geq 2$, this happens with probability at most $2t\mu\binom{t}{\lfloor t/2 \rfloor}2^{-t}$. The central binomial coefficient satisfies $\binom{t}{\lfloor t/2 \rfloor} \leq 2^t/\sqrt{\pi t/2} \leq 2^t/\sqrt{t}$ and thus the probability is at most $2t\mu/\sqrt{t} = 2\mu\sqrt{t}$. □

Finally, we prove Lemma 3 from the main paper:

Restatement of Lemma 3. *For any distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}$, any $t \geq 36$ and any voting classifier $f \in \Delta(\mathcal{H})$ for a hypothesis set $\mathcal{H} \subset \mathcal{X} \to \{-1, 1\}$, we have:*

$$\mathcal{L}_{\mathcal{D}}(f) \;\leq\; 3\mathcal{L}^t_{\mathcal{D}}(f).$$

For the proof, we first need the following auxiliary lemma:

Lemma C.1. *For any $x \in \mathcal{X}$ and any $f \in \Delta(\mathcal{H})$, if $f(x) \neq 0$, then*

$$\Pr_{g \sim \mathcal{D}_{f,t}}\big[\operatorname{sign}(f(x)) = \operatorname{sign}(g(x))\big] \;\geq\; 1/2 - 1/\sqrt{t}.$$

Proof. If we condition on t', then $h_1, \ldots, h_{t'}$ are i.i.d samples from $\mathcal{D}_f$ and thus $\Pr[\operatorname{sign}(g(x)) = \operatorname{sign}(f(x))] \geq \Pr[\operatorname{sign}(g(x)) = -\operatorname{sign}(f(x))]$. We therefore have $\Pr[\operatorname{sign}(f(x)) = \operatorname{sign}(g(x))] \geq \Pr[g(x) \neq 0]/2$, regardless of t'. We thus only need to bound $\Pr[g(x) \neq 0]$. For this, Lemma 2 with $\mu = 1/t$ implies $\Pr[g(x) = 0] \leq 2/\sqrt{t}$. □

Using this lemma, we can prove Lemma 3:

Proof of Lemma 3 from the main paper. Consider any example $(x, y) \in \mathcal{X} \times \{-1, 1\}$ for which $\Pr_{g \sim \mathcal{D}_{f,t}}[yg(x) \leq 0] < 1/2 - 1/\sqrt{t}$. By Lemma C.1, it must be the case that $\operatorname{sign}(f(x)) = y$. We therefore have by Markov's inequality:

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}(f) \;&\leq\; \Pr_{(x,y)\sim\mathcal{D}}\big[\Pr_{g\sim\mathcal{D}_{f,t}}[yg(x) \leq 0] \geq 1/2 - 1/\sqrt{t}\big] \\
&\leq\; \frac{\mathbb{E}_{(x,y)\sim\mathcal{D}}[\Pr_{g\sim\mathcal{D}_{f,t}}[yg(x) \leq 0]]}{1/2 - 1/\sqrt{t}} \\
&=\; \mathcal{L}^t_{\mathcal{D}}(f)/(1/2 - 1/\sqrt{t}) \\
&\leq\; 3\mathcal{L}^t_{\mathcal{D}}(f).
\end{aligned}$$

□

C.2 Relating generalization error to the ghost set

In the following, we give the proof of Lemma 6 from the main paper:

Restatement of Lemma 6. *For $m \geq 2400^2$ any t and any f, it holds that:*

$$\Pr_S\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{\mathcal{D}}(f)| > \tfrac{1}{1200}\Big] \;\leq\; 2 \cdot \Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{S'}(f)| > \tfrac{1}{2400}\Big].$$

Proof. The proof uses standard techniques uneventfully. We can assume $\Pr_S[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{\mathcal{D}}(f)| > 1/1200] > 0$, otherwise we are done. We have:

$$\begin{aligned}
&\Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{S'}(f)| > \tfrac{1}{2400}\Big] \\
\geq\; &\Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{S'}(f)| > \tfrac{1}{2400} \wedge \sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{\mathcal{D}}(f)| > \tfrac{1}{1200}\Big] \\
=\; &\Pr_{S}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{\mathcal{D}}(f)| > \tfrac{1}{1200}\Big] \times \\
&\Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{S'}(f)| > \tfrac{1}{2400} \mid \sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}^t_S(f) - \mathcal{L}^t_{\mathcal{D}}(f)| > \tfrac{1}{1200}\Big].
\end{aligned}$$

Fix a data set S in the non-empty event $\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{\mathcal{D}}^t(f)| > 1/1200$. Let $f^* \in \mathcal{H}$ be any hypothesis on which $|\mathcal{L}_S^t(f^*) - \mathcal{L}_{\mathcal{D}}^t(f^*)| > 1/1200$. The hypothesis f^* does not depend on S' but only on S. We now condition on S as well and get:

$$\begin{aligned}
&\Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{S'}^t(f)| > \tfrac{1}{2400} \,\Big|\, S;\ \sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{\mathcal{D}}^t(f)| > \tfrac{1}{1200}\Big] \\
\geq\ &\Pr_{S'}\Big[|\mathcal{L}_S^t(f^*) - \mathcal{L}_{S'}^t(f^*)| > \tfrac{1}{2400} \,\Big|\, S;\ \sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{\mathcal{D}}^t(f)| > \tfrac{1}{1200}\Big] \\
\geq\ &\Pr_{S'}\Big[|\mathcal{L}_{S'}^t(f^*) - \mathcal{L}_{\mathcal{D}}^t(f^*)| \leq \tfrac{1}{2400} \,\Big|\, S;\ \sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{\mathcal{D}}^t(f)| > \tfrac{1}{1200}\Big].
\end{aligned}$$

Here the last inequality follows because the events $|\mathcal{L}_{S'}^t(f^*) - \mathcal{L}_{\mathcal{D}}^t(f^*)| \leq 1/2400$ and $|\mathcal{L}_S^t(f^*) - \mathcal{L}_{\mathcal{D}}^t(f^*)| > 1/1200$ (which holds by definition of f^*) implies $|\mathcal{L}_S^t(f^*) - \mathcal{L}_{S'}^t(f^*)| > 1/2400$. Since f^* is fixed and independent of S', we may now use Hoeffding's inequality to conclude

$$\Pr_{S'}\Big[|\mathcal{L}_{S'}^t(f^*) - \mathcal{L}_{\mathcal{D}}^t(f^*)| \leq 1/2400 \,\Big|\, S;\ \sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{\mathcal{D}}^t(f)| > 1/1200\Big] \ \geq\ 1 - 2e^{-2(1/2400)^2 m}.$$

For $m \geq 2400^2$, this is at least $1 - 2e^{-2} \geq 1/2$.

Multiplying with $\Pr[S \mid \sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{\mathcal{D}}^t(f)| > 1/1200]$ and integrating over S, we get

$$\begin{aligned}
&\int_S \Big(\Pr_{S'}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{S'}^t(f)| > \tfrac{1}{2400} \,\Big|\, S;\ \sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{\mathcal{D}}^t(f)| > \tfrac{1}{1200}\Big] \\
&\qquad \times\ \Pr\Big[S \,\Big|\, \sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{\mathcal{D}}^t(f)| > \tfrac{1}{1200}\Big]\Big) \\
\geq\ &\int_S \tfrac{1}{2} \Pr\Big[S \,\Big|\, \sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{\mathcal{D}}^t(f)| > \tfrac{1}{1200}\Big].
\end{aligned}$$

The right hand side is simply $1/2$ and the left hand side is $\Pr_{S,S'}[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{S'}^t(f)| > 1/2400 \mid \sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{\mathcal{D}}^t(f)| > 1/1200]$. We finally conclude that for $m \geq 2400^2$, we have:

$$\Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{S'}^t(f)| > \tfrac{1}{2400}\Big] \ \geq\ \tfrac{1}{2}\Pr_{S}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{\mathcal{D}}^t(f)| > \tfrac{1}{1200}\Big]. \qquad \square$$

C.3 Relation to the growth function

Last, we prove Lemma 8 from the main paper, which is restated here for convenience:

Restatement of Lemma 8. *For any $0 < \delta < 1$, every t, and every $\mu \leq \delta/(9600\sqrt{t})$, we have*

$$\Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} \big|\mathcal{L}_S^t(f) - \mathcal{L}_{S'}^t(f)\big| > \tfrac{1}{2400}\Big] \ \leq\ \sup_P 2\,\big|\hat{\Delta}_\delta^\mu(P)\big| \exp\big(-2m/9600^2\big).$$

Proof. Let $\mu \leq \delta/(9600\sqrt{t})$. We have that:

$$\begin{aligned}
&\Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{S'}^t(f)| > \tfrac{1}{2400}\Big] \\
=\ &\int_P \Pr[P] \Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{S'}^t(f)| > \tfrac{1}{2400} \,\Big|\, P\Big] \\
\leq\ &\sup_P \Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} |\mathcal{L}_S^t(f) - \mathcal{L}_{S'}^t(f)| > \tfrac{1}{2400} \,\Big|\, P\Big] \\
=\ &\sup_P \Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} \Big|\Pr_{(x,y)\sim S, g\sim\mathcal{D}_{f,t}}[yg(x) \leq 0] - \Pr_{(x,y)\sim S', g\sim\mathcal{D}_{f,t}}[yg(x) \leq 0]\Big| > \tfrac{1}{2400} \,\Big|\, P\Big] \\
=\ &\sup_P \Pr_{S,S'}\Big[\sup_{f\in\Delta(\mathcal{H})} \Big|\int_g \Pr[g]\Big(\Pr_{(x,y)\sim S}[yg(x) \leq 0] - \Pr_{(x,y)\sim S'}[yg(x) \leq 0]\Big)\Big| > \tfrac{1}{2400} \,\Big|\, P\Big].
\end{aligned}$$

We always have $\left(\Pr_{(x,y)\sim S}[yg(x) \le 0] - \Pr_{(x,y)\sim S'}[yg(t) \le 0]\right) \le 1$, and by Lemma 13, we have $\Pr[g \notin \Delta^{\mu}_{\delta}(\mathcal{H}, P)] \le 1/4800$, hence

$$\begin{aligned}
&\left|\int_g \Pr[g]\left(\Pr_{(x,y)\sim S}[yg(x) \le 0] - \Pr_{(x,y)\sim S'}[yg(x) \le 0]\right)\right| \\
\le\ & \Pr_{g\sim\mathcal{D}_{f,g}}\left[g \notin \Delta^{\mu}_{\delta}(\mathcal{H}, P)\right] + \sup_{g\in\Delta^{\mu}_{\delta}(\mathcal{H},P)}\left|\Pr_{(x,y)\sim S}[yg(x) \le 0] - \Pr_{(x,y)\sim S'}[yg(x) \le 0]\right| \\
\le\ & \tfrac{1}{4800} + \sup_{g\in\Delta^{\mu}_{\delta}(\mathcal{H},P)}\left|\Pr_{(x,y)\sim S}[yg(x) \le 0] - \Pr_{(x,y)\sim S'}[yg(x) \le 0]\right|.
\end{aligned}$$

We thus have

$$\begin{aligned}
&\Pr_{S,S'}\left[\sup_{f\in\Delta(\mathcal{H})}|\mathcal{L}^t_S(f) - \mathcal{L}^t_{S'}(f)| > \tfrac{1}{2400}\right] \\
\le\ & \sup_P \Pr_{S,S'}\left[\tfrac{1}{4800} + \sup_{g\in\Delta^{\mu}_{\delta}(\mathcal{H},P)}\left|\Pr_{(x,y)\sim S}[yg(x) \le 0] - \Pr_{(x,y)\sim S'}[yg(x) \le 0]\right| > \tfrac{1}{2400}\ \middle|\ P\right] \\
=\ & \sup_P \Pr_{S,S'}\left[\sup_{g\in\Delta^{\mu}_{\delta}(\mathcal{H},P)}\left|\Pr_{(x,y)\sim S}[yg(x) \le 0] - \Pr_{(x,y)\sim S'}[yg(x) \le 0]\right| > \tfrac{1}{4800}\ \middle|\ P\right].
\end{aligned}$$

To bound this, let $\hat{\Delta}^{\mu}_{\delta}(P) = \mathrm{sign}(\Delta^{\mu}_{\delta}(\mathcal{H}, P))$. Then the above equals:

$$\sup_P \Pr_{S,S'}\left[\sup_{h\in\hat{\Delta}^{\mu}_{\delta}(P)}\left|\Pr_{(x,y)\sim S}[h(x) \ne y] - \Pr_{(x,y)\sim S'}[h(x) \ne y]\right| > \tfrac{1}{4800}\ \middle|\ P\right].$$

Since we have restricted to the fixed set P, the set $\hat{\Delta}^{\mu}_{\delta}(P)$ is finite. Hence we may use the union bound to bound the above by

$$\sup_P |\hat{\Delta}^{\mu}_{\delta}(P)| \sup_{h\in\hat{\Delta}^{\mu}_{\delta}(P)} \Pr_{S,S'}\left[\left|\Pr_{(x,y)\sim S}[h(x) \ne y] - \Pr_{(x,y)\sim S'}[h(x) \ne y]\right| > \tfrac{1}{4800}\ \middle|\ P\right].$$

For a set P and hypothesis $h \in \hat{\Delta}^{\mu}_{\delta}(P)$, let p denote the fraction of samples $(x, y) \in P$ for which $h(x) \ne y$. Recall that S and the ghost set S' are obtained from P by letting S be a uniform set of m samples from P without replacement, and S' are the remaining m samples. For shorthand, define $p_S = \Pr_{(x,y)\sim S}[h(x) \ne y \mid P]$ and $p_{S'}$ symmetrically. Then $p = (1/2)(p_S + p_{S'})$. By Hoeffding's inequality for sampling without replacement, we have $\Pr_{S,S'}[|p_S - p| > \varepsilon \mid P] = \Pr_S[|p_S - p| > \varepsilon \mid P] < 2\exp(-2\varepsilon^2 m)$. Setting $\varepsilon = 1/9600$, we get that for $|p - p_S| \le 1/9600$, it must be the case that $p'_S = 2p - p_S \in p \pm 1/9600$. Hence $|p_S - p_{S'}| \le 1/4800$ and we conclude $\Pr_{S,S'}[|p_S - p_{S'}| > 1/4800 \mid P] < 2\exp(-2m/9600^2)$. Thus, we end up with the bound

$$\Pr_{S,S'}\left[\sup_{f\in\Delta(\mathcal{H})}|\mathcal{L}^t_S(f) - \mathcal{L}^t_{S'}(f)| > \tfrac{1}{2400}\right] \le \sup_P 2|\hat{\Delta}^{\mu}_{\delta}(P)|\exp(-2m/9600^2). \qquad \square$$