AdaBoost is not an Optimal Weak to Strong Learner

Mikael Møller Høgsgaard, Kasper Green Larsen, Martin Ritzert

hogsgaard@cs.au.dk, larsen@cs.au.dk, ritzert@informatik.uni-goettingen.de

Abstract

AdaBoost is a classic boosting algorithm for combining multiple inaccurate classifiers produced by a *weak learner*, to produce a *strong learner* with arbitrarily high accuracy when given enough training data. Determining the optimal number of samples necessary to obtain a given accuracy of the strong learner, is a basic learning theoretic question. Larsen and Ritzert (NeurIPS'22) recently presented the first provably optimal weak-to-strong learner. However, their algorithm is somewhat complicated and it remains an intriguing question whether the prototypical boosting algorithm AdaBoost also makes optimal use of training samples. In this work, we answer this question in the negative. Concretely, we show that the sample complexity of AdaBoost, and other classic variations thereof, are sub-optimal by at least one logarithmic factor in the desired accuracy of the strong learner.

1 Introduction

The algorithm AdaBoost [2] is the textbook example of a boosting algorithm. Boosting algorithms in general make use of a *weak learner*, i.e. a learning algorithm that produces classifiers with accuracy slightly better than chance, and produces from it a so-called *strong learner*, achieving arbitrarily high accuracy when given enough training samples. The question whether one can always produce a strong learner from a weak learner was initially asked by Kearns and Valiant [7, 8] and initiated the field of boosting.

Given a weak learner $\mathcal{W}$, AdaBoost uses $\mathcal{W}$ to train multiple inaccurate classifiers/hypotheses that focus on different parts of the training data and combines them using a weighted majority vote. In more detail, it runs for some T iterations, each time invoking $\mathcal{W}$ to produce a hypothesis h_t. It then computes weights w and outputs the final voting classifier $f(x) = \text{sign}(\sum_t w_t h_t(x))$. For the calls of $\mathcal{W}$, AdaBoost maintains a distribution $\mathcal{D}_t$ over the training samples that puts a large weight on training samples misclassified by most of $h_1, \ldots, h_{t-1}$ and a smaller weight on samples classified correctly. Using this distribution, in iteration t AdaBoost invokes the weak learner to produce a hypothesis h_t performing better than chance under $\mathcal{D}_t$. This way, h_t focuses on training examples which are hard for the voting classifier so far.

In this paper, we study the sample complexity of AdaBoost, answering the question whether AdaBoost is able to make optimal use of its training data. To formally answer this question, we need to introduce a few parameters. A γ-weak learner is a learning algorithm that, given some constant number of training samples from an unknown data distribution $\mathcal{D}$, produces a hypothesis h that correctly predicts the label of a new sample from $\mathcal{D}$ with probability at least $1/2 + \gamma$. We let $\mathcal{H}$ denote the set of possible hypotheses that the weak learner may output. A strong learner, on the other hand, is a learning algorithm that for any $0 < \varepsilon, \delta < 1$, with probability at least $1 - \delta$ over a set of $m(\varepsilon, \delta)$ training samples from an unknown distribution $\mathcal{D}$, outputs a hypothesis that correctly predicts the label of a new sample from $\mathcal{D}$ with probability at least $1 - \varepsilon$. The function $m(\varepsilon, \delta)$ is referred to as the sample complexity. A strong learner can thus obtain arbitrarily high accuracy $1 - \varepsilon$ when given enough training samples $m(\varepsilon, \delta)$. See Section 1.1 for a formal definition of weak and strong learning.

Recently, Larsen & Ritzert [10] showed that the optimal sample complexity of weak-to-strong learning is given by

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

where d is the VC-dimension of the hypothesis set $\mathcal{H}$ of the weak learner. The paper provides both a learning algorithm achieving this sample complexity as well as an asymptotically matching lower bound. Their algorithm is based on a majority vote among hypotheses produced by a version of AdaBoost. It is thus a majority of majorities. Is this necessary for optimal weak-to-strong learning? Or does it suffice to use a classic algorithm like AdaBoost? The current best upper bound on the sample complexity of AdaBoost (for constant δ) is [14]:

$$m_{\text{Ada}}(\varepsilon) = O\left(\frac{d \ln \frac{1}{\varepsilon\gamma} \ln \frac{d}{\varepsilon\gamma}}{\gamma^2 \varepsilon}\right) \tag{2}$$

However, this is just an upper bound, and until now, it remained completely plausible that a better analysis could remove the two logarithmic factors.

The main contribution of this work is to show that AdaBoost is *not always optimal*. Concretely, we show that there exists a weak learner $\mathcal{W}$, such that if AdaBoost is run with $\mathcal{W}$ as its weak learner, its sample complexity is sub-optimal by at least one logarithmic factor. This is stated in the following theorem:

Theorem 1.1. *For any $0 < \gamma < C$ for $C > 0$ sufficiently small, any $d = \Omega(\ln(1/\gamma))$, and any $\exp(-\exp(\Omega(d))) \leq \varepsilon \leq C$, there exists a γ-weak learner $\mathcal{W}$ using a hypothesis set $\mathcal{H}$ of VC-dimension d and a distribution $\mathcal{D}$, such that AdaBoost run with $\mathcal{W}$ is sub-optimal and needs*

$$m_{Ada}(\varepsilon) = \Omega\left(\frac{d \ln(1/\varepsilon)}{\gamma^2 \varepsilon}\right)$$

samples from $\mathcal{D}$ to output with constant probability, a hypothesis with error at most ε under $\mathcal{D}$.

This lower bound does not only apply to AdaBoost but extends to many of its variants such as AdaBoost$_\nu$ [12], AdaBoost$^*_\nu$ [13], and DualLPboost [4]. The key property those algorithms share and that we manage to exploit is that they run the weak learner $\mathcal{W}$ on the full training data set. This allows $\mathcal{W}$ to adversarially return hypotheses that accumulate mistakes outside of the training data, leading to poor generalization performance.

The rest of the paper is structured as follows. In the remainder of this section, we describe some preliminaries and give an overview of the related work. In Section 2, we present the high-level ideas of our proof and in Section 3 we sketch the formal details of the proof. The proofs of the main lemmas and parts of the formal proof of Theorem 1.1 are deferred to the appendix.

1.1 Preliminaries and Notation

We now formally define our setup. Weak and strong learning are studied in the general framework of *probably approximately correct* (PAC) learning, see e.g. [14] for an introduction. In the PAC learning framework, one assumes that training samples are chosen i.i.d. from an underlying distribution $\mathcal{D}$ over elements of some universe $\mathcal{X}$. Furthermore, we assume an underlying but unknown 'correct' labeling function $c\colon \mathcal{X} \to \{-1, 1\}$ called the *concept*, which assigns every element from the universe $\mathcal{X}$ its 'true' label. The concept is assumed to belong to a concept class $\mathcal{C} \subseteq \mathcal{X} \to \{-1, 1\}$.

A learning algorithm $\mathcal{A}$ is a *γ-weak learner* for $\mathcal{C}$, if for every distribution $\mathcal{D}$ over $\mathcal{X}$ and every concept $c \in \mathcal{C}$, there is a constant number of samples m_0 and a constant $\delta_0 < 1$, such that with probability at least $1 - \delta_0$ over m_0 i.i.d. samples $x_1, \ldots, x_{m_0}$ from $\mathcal{D}$ and their corresponding labels $c(x_1), \ldots, c(x_{m_0})$, $\mathcal{A}$ outputs a hypothesis h with error

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

Algorithm 1: AdaBoost

Input: training set $S = \{(x_1, c(x_1)), \ldots, (x_m, c(x_m))\}$,
number of rounds T

Result: A majority hypothesis h_{out}

$\mathcal{D}^{(1)} \leftarrow \left(\frac{1}{m}, \ldots \frac{1}{m}\right)$ // uniform init of $\mathcal{D}$

for $t = 1, \ldots, T$ **do**

  $h_t \leftarrow \mathcal{W}(\mathcal{D}^{(t)}, S)$ // invoke weak learner $\mathcal{W}$

  $\gamma_t \leftarrow \sum_{i=1}^{m} \mathcal{D}^{(t)} \operatorname{sign}(c(x_i)h_t(x_i))$ // error

  $w_t = \frac{1}{2} \ln\left(\frac{1-\gamma_t}{\gamma_t}\right)$ // weight for h_t

  for $i \in \{1, \ldots, m\}$ **do**

    /* update $\mathcal{D}$ based on success of h_t */

    $\mathcal{D}_i^{(t+1)} \leftarrow \frac{\mathcal{D}_i^{(t)} \exp\left(-w_t c(x_i) h_t(x_i)\right)}{\sum_{j=1}^{m} \mathcal{D}_j^{(t)} \exp\left(-w_t c(x_j) h_t(x_j)\right)}$

return $h_{out}(x) = \operatorname{sign}\left(\sum_{t=1}^{T} w_t h_t(x)\right)$

We refer to γ as the *advantage* of the weak learner. We let $\mathcal{H}$ denote the hypothesis set used by the weak learner, i.e. we assume that $h \in \mathcal{H}$ and that $\mathcal{H}$ has a finite VC-dimension d.

A learning algorithm $\mathcal{A}$ is a *strong learner* for $\mathcal{C}$, if for every $0 < \varepsilon, \delta < 1$, there exists some number of samples $m(\varepsilon, \delta)$, such that with probability at least $1 - \delta$ over $m(\varepsilon, \delta)$ i.i.d. samples from $\mathcal{D}$ and their corresponding labels, $\mathcal{A}$ outputs a hypothesis with error $\mathcal{L}_{\mathcal{D}}(h) \leq \varepsilon$.

AdaBoost is the classic algorithm for constructing a strong learner from a γ-weak learner. For completeness, we have included the full algorithm as Algorithm 1.

Related Work

In terms of sample complexity, most previous works prove generalization bounds for *voting classifiers* in general. A voting classifier over a hypothesis set $\mathcal{H}$, is a majority vote $f(x) = \operatorname{sign}\left(\sum_{h \in \mathcal{H}} \alpha_h h(x)\right)$ for coefficients $\alpha_h > 0$ such that $\sum_h \alpha_h = 1$. AdaBoost can be seen to output a voting classifier by appropriate normalization of the coefficients w_t chosen in Algorithm 1. The generalization bounds for voting classifiers are typically *data-dependent* in the sense that they depend on the so-called *margin* of the voting classifier. For a voting classifier $f(x) = \operatorname{sign}\left(\sum_{h \in \mathcal{H}} \alpha_h h(x)\right)$ and a sample $(x, c(x))$, the margin of f on $(x, c(x))$ is defined as $c(x) \sum_{h \in \mathcal{H}} \alpha_h h(x)$. The margin is thus a number between -1 and 1 and is positive if and only if $f(x) = c(x)$. Intuitively, large margins correspond to high certainty/agreement among the hypotheses. In terms of upper bounds, Breiman [1] showed that with probability $1 - \delta$ over a training set S of m samples, all voting classifiers f with margin at least γ on all samples in S have

$$\mathcal{L}_{\mathcal{D}}(f) = O\left(\frac{d \ln(m/d) \ln m}{\gamma^2 m}\right). \tag{3}$$

A small tweak to AdaBoost, known as AdaBoost$_\nu^*$ [13], guarantees that the output hypothesis f has margins $\Omega(\gamma)$ on all samples when AdaBoost$_\nu^*$ is run with a γ-weak learner. Solving for $\varepsilon = \mathcal{L}_{\mathcal{D}}(f)$ in Equation (3) matches the sample complexity bound for AdaBoost from Equation (2).

In terms of sample complexity lower bounds for boosting, or for AdaBoost in particular, there are some relevant works. First, as mentioned earlier and stated in (1), it is known that any weak-to-strong learner must have a sample complexity of $\Omega\big(d/(\gamma^2\varepsilon) + \ln(1/\delta)/\varepsilon\big)$ [10]. While not directly comparable, work by [3] showed that there are data distributions, such that with constant probability over a set of $m = (d/\gamma^2)^{1+\Omega(1)}$ samples, there *exists* a voting classifier f with margin at least γ on all samples, yet its generalization error is at least $\Omega\big(d \ln(m)/(\gamma^2 m)\big)$. This lower bound is in some sense similar to our work, as it manages to squeeze out a logarithmic factor. However, the voting classifier f is only shown to exist and as such might not correspond to the output of any reasonable learning algorithm, certainly not AdaBoost.

At this point, we would like to compare AdaBoost to the optimal weak-to-strong learning algorithm given by Larsen & Ritzert [10]. First, their learning algorithm is more complicated. It runs AdaBoost$^*_\nu$ on various sub-samples of the training data to obtain voting classifiers $f_1, \ldots, f_T$ which it then combines in a majority vote $g(x) = \text{sign}(\sum_i f_i(x))$. It thus outputs a majority of majorities. Moreover, the number of sub-samples is a rather large $T = m^{\lg_4 3} \approx m^{0.79}$ and their size is linear in the overall number of training samples m, thus resulting in somewhat slow training time. The sub-samples are constructed with a very careful overlap as pioneered by Hanneke [5] in his optimal algorithm for PAC learning in the realizable setting. A recent manuscript [9] shows that one may replace the T sub-samples by just $O(\lg(m/\delta))$ bootstrap samples (sub-samples each consisting of m samples with replacement from the training data) in the algorithm from Larsen & Ritzert [10]. While reducing the number of sub-samples, it still remains a majority of majorities. It would thus have been desirable if one could show that AdaBoost also had an optimal sample complexity. Sadly, as already stated in Theorem 1.1, this is not true.

2 Proof Overview

In this section, we give an overview of the main ideas in our proof that AdaBoost is not always an optimal weak-to-strong learner. Concretely, for any γ, m, and $d = \Omega(\ln(1/\gamma))$ we show that there exists an input domain $\mathcal{X}$, a distribution $\mathcal{D}$ over $\mathcal{X}$, a concept $c : \mathcal{X} \to \{-1, 1\}$, a hypothesis set $\mathcal{H}$ of VC-dimension at most d, and a γ-weak learner $\mathcal{W}$ for c that outputs hypotheses from $\mathcal{H}$, such that with constant probability over a set of m samples $S \sim \mathcal{D}^m$ and their corresponding labels $c(S)$, AdaBoost run with the weak learner $\mathcal{W}$ produces a voting classifier f with $\mathcal{L}_\mathcal{D}(f) = \Omega\big(d \ln(\gamma^2 m/d)/(\gamma^2 m)\big)$. Solving for $\varepsilon = \mathcal{L}_\mathcal{D}(f)$ gives $m = \Omega\big((d \ln(1/\varepsilon))/(\gamma^2 \varepsilon)\big)$ as claimed in Theorem 1.1.

When proving the lower bound for AdaBoost, we consider just one fixed concept c, namely the concept that assigns the label 1 to all elements of $\mathcal{X}$. AdaBoost of course does not know this but executes precisely as in Algorithm 1. As distribution $\mathcal{D}$ we consider the uniform distribution $\mathcal{U}$ over the input domain $\mathcal{X}$. Thus, if f is the output of AdaBoost and $\mathcal{X} = [u]$, then $\mathcal{L}_\mathcal{U}(f)$ is precisely equal to the fraction of elements $i \in [u]$ for which $f(i) = -1$. Our goal is thus to show that AdaBoost will produce a voting classifier f with a negative prediction on many $i \in [u]$.

To prove the above, we need to construct a weak learner $\mathcal{W}$ that somehow returns hypotheses that result in AdaBoost making many negative predictions. Although the formal definition of a γ-weak learner given in Section 1.1 allows $\mathcal{W}$ to sometimes (with probability δ_0) return a hypothesis with advantage less than γ, we will *not* do so in our construction. Thus, our adversarial weak learner always returns hypotheses with advantage at least γ which only makes our lower bound stronger.

To define our adversarial weak learner $\mathcal{W}$, we carefully examine the "interface" it must support. Concretely, the way AdaBoost accesses a weak learner is to feed it the training data $S = \{(x_i, c(x_i))\}_{i=1}^m$ and a distribution $\mathcal{D}_t$ over S. From this, AdaBoost expects that $\mathcal{W}$ returns a hypothesis h_t with advantage at least γ under the distribution $\mathcal{D}_t$ which is supported only on S. Our adversarial weak learner $\mathcal{W}$ will support this interface. In fact, it will completely ignore the set S and return a hypothesis that is solely a function of $\mathcal{D}_t$. Our weak learner thus needs to be a function, that for any probability distribution $\mathcal{D}$ over $\mathcal{X}$ returns a hypothesis h with advantage at least γ under $\mathcal{D}$ (for the all-1 concept c).

Our main challenge is now to design a weak learner that always has advantage γ under the distributions fed to it by AdaBoost, yet under the uniform distribution $\mathcal{U}$ over $\mathcal{X} = [u]$, the voting classifier produced by AdaBoost must often make negative predictions. Here, our first observation is that if the universe size u is $cm/\ln(\gamma^2 m/d)$ for a sufficiently small constant $c > 0$, then by a coupon collector argument, with constant probability there are $\Omega(d/\gamma^2)$ elements $i \in [u]$ that are not sampled into the training set S. Our basic idea is to force that the final voting classifier f produced by AdaBoost makes negative predictions on a constant fraction of these non-sampled elements. This would imply $\mathcal{L}_\mathcal{U}(f) = \Omega\big((d/\gamma^2)/u\big) = \Omega\big(d \ln(\gamma^2 m/d)/(\gamma^2 m)\big)$ as claimed.

Our next key observation is that all distributions $\mathcal{D}_t$ fed to $\mathcal{W}$ by AdaBoost put a non-zero probability on *every* element in the training data set. Crucially, this implies that the weak learner knows the complete training set and can thus compute the $\Omega(d/\gamma^2)$ points $\bar{S}$ that were not sampled. Our adversarial weak

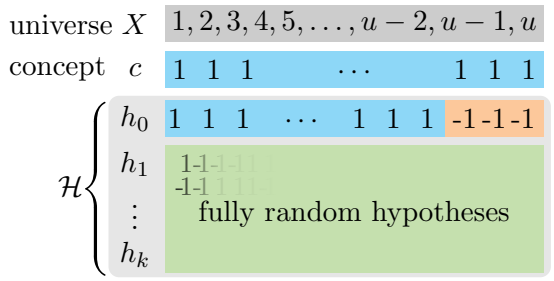

Figure 1: Illustration of our hypothesis set $\mathcal{H}$

learner does precisely this and chooses an arbitrary subset $\bar{S}' \subseteq \bar{S}$ of size $O(d/\gamma^2)$ (the same deterministic choice for a given $\bar{S}$). It then returns a hypothesis h that has advantage γ under $\mathcal{D}_t$ but at the same time under the uniform distribution over $\bar{S}'$ is *wrong* with probability $1/2 + \gamma$. Notice that it is wrong on $\bar{S}'$ with probability more than half which we call a *negative advantage* of $-\gamma$. Intuitively, since this holds for every h returned by $\mathcal{W}$ on an execution of AdaBoost (for the same $\bar{S}'$), the output f of AdaBoost will be mistaken on about half the points in $\bar{S}'$ which is sufficient for the lower bound.

To carry out the above argument, we need to construct a hypothesis set $\mathcal{H}$ that contains hypotheses with advantage γ on S under D_t and negative advantage over $\bar{S}'$. Then the weak learner can essentially just return such a hypothesis. For this construction, we use a probabilistic argument and show that by sampling a *random* hypothesis set $\mathcal{H}$ in an appropriate manner and defining an associated weak learner $\mathcal{W}_{\mathcal{H}}$, there is a constant probability that the weak learner satisfies all of the above. Hence, a weak learner must exist. The point of considering a random $\mathcal{H}$ is that it allows us to give simple probabilistic arguments that show that all the hypotheses that $\mathcal{W}_{\mathcal{H}}$ needs to return on an execution of AdaBoost indeed exist in $\mathcal{H}$. We illustrate $\mathcal{H}$ in Figure 1.

For the random construction of $\mathcal{H}$, we sample at most 2^{d-1} hypotheses $h : \mathcal{X} \to \{-1, 1\}$ independently and uniformly at random. This clearly implies that the VC-dimension of $\mathcal{H}$ is less than d. We now have to argue that we can use $\mathcal{H}$ to design a γ-weak learner for the all-1 concept. Here, we distinguish two cases. First, consider any distribution $\mathcal{D}$ over $[u]$ where most of the probability mass is concentrated on some r entries. Anti-concentration results imply that a random hypothesis has an advantage of $\Omega(\sqrt{\ln(1/\delta)/r})$ with probability at least δ. We need the advantage to be at least γ and we have $\exp(\Omega(d))$ hypotheses to choose from. Thus, if we plug in $\delta = \exp(-\Omega(d))$, we see that for $r = O(d/\gamma^2)$ we expect that the random $\mathcal{H}$ contains a hypothesis with advantage γ under $\mathcal{D}$. Thus, for distributions with small support, we can get a high advantage. A similar argument shows that we can at the same time get a *negative* advantage of $-\gamma$ on $\bar{S}'$ as required earlier. However, AdaBoost might feed $\mathcal{W}$ a distribution $\mathcal{D}_t$ that is not concentrated on some $O(d/\gamma^2)$ entries. In this second case, we would intuitively like to add the all-ones hypothesis $h_0^\star$ to $\mathcal{H}$ to achieve an advantage on such $\mathcal{D}_t$. Then $\mathcal{W}_{\mathcal{H}}$ can always return $h_0^\star$ when being fed a distribution that is far from concentrated on a few entries. This is problematic for our lower bound since now AdaBoost could put a large weight on $h_0^\star$ which would cancel out any mistakes/negative advantage we accumulated in $\bar{S}'$.

To remedy this, we introduce the hypothesis h_0 which resembles $h_0^\star$ on most elements (returning 1 there) but returns -1 on cd/γ^2 elements of $\mathcal{X}$ for some constant $c > 0$. Then, similar to $h_0^\star$, the hypothesis h_0 has a γ advantage under all $\mathcal{D}$ that are "spread out", i.e. do not have most of its mass on $O(d/\gamma^2)$ entries. Thus, we can let $\mathcal{W}_{\mathcal{H}}$ return h_0 for such $\mathcal{D}$. If on the other hand $\mathcal{D}$ is concentrated on few entries, we can find one of the random h that has advantage at least γ under $\mathcal{D}$ and at most $-\gamma$ for a uniform element in $\bar{S}'$. But $\bar{S}'$ might be (mostly) among the coordinates where h_0 returns 1. Thus, if AdaBoost puts too large a weight on h_0, then the negative advantage we accumulated on $\bar{S}'$ is still canceled out by h_0. This is where we use that h_0 has many -1's. Concretely, we show that if h_0 receives a weight of more than some $O(\gamma)$, then there is no way to cancel out the -1's that h_0 produces. In summary, if AdaBoost assigns a large weight to h_0 in its output classifier f, then f makes negative predictions where h_0 is negative. If AdaBoost assigns a small weight to h_0, then f makes negative predictions in $\bar{S}'$. In both

cases, we have $\Omega(d/\gamma^2)$ negative predictions. We illustrate this in Figure 2.

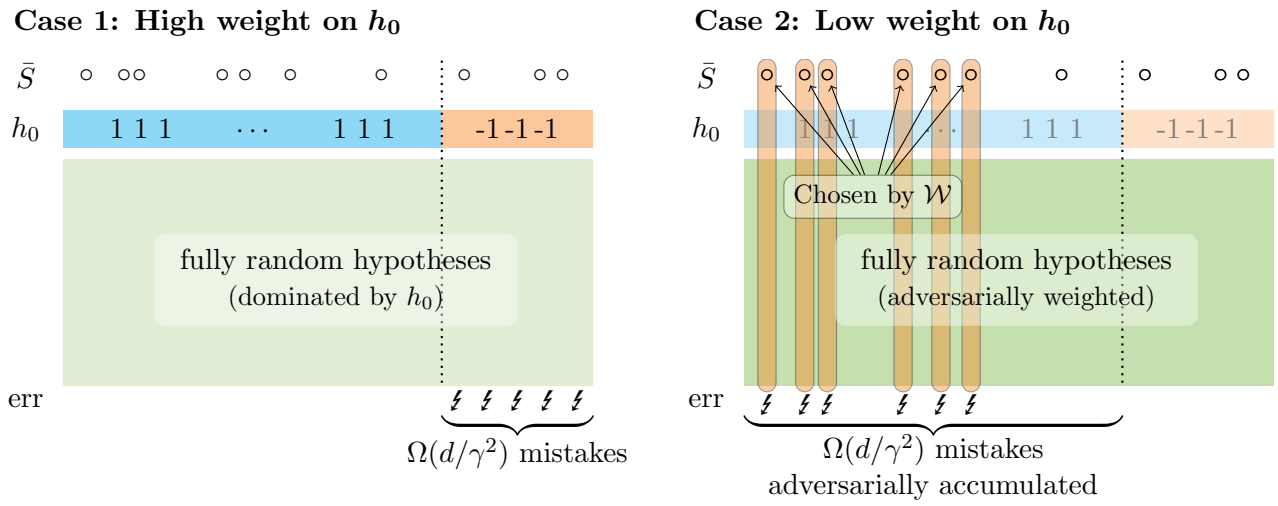

Figure 2: Illustration of where errors will occur

Finally, let us summarize precisely what properties of AdaBoost we exploited above. As mentioned earlier, the key point is that the adversarial weak learner can determine the elements $\bar{S}$ of $\mathcal{X}$ that are not part of the training set S. It can thus return hypotheses that have a negative advantage of $-\gamma$ on some $\Omega(d/\gamma^2)$ elements of $\bar{S}$. This negative advantage is enough that it is not canceled out by any weight that AdaBoost assigns to a nearly all-1 hypothesis h_0. Note though that it is crucial that the negative advantage achieved by $\mathcal{W}$ is $-\gamma$ and not just negative as AdaBoost may use h_0 "a little bit", i.e. with a weight of up to some small constant times γ. If AdaBoost would put more weight on h_0, this would induce negative predictions where h_0 is negative.

Let us also remark that it is vital for our argument that every distribution $\mathcal{D}_t$ fed to $\mathcal{W}$ by AdaBoost is non-zero on *all* of the training data. Assume for instance that $\mathcal{D}_t$ was only non-zero on a random constant fraction of S. Then the weak learner could only identify some random superset of $\bar{S}$ having linear size in u. But the weak learner needs to force a negative advantage of $-\gamma$ on some $\Omega(d/\gamma^2)$ points to cancel out the positive contributions by h_0. Concentration results show that this can only be done on $O(d/\gamma^2)$ points and thus the adversarial weak learner would have to pick $O(d/\gamma^2)$ points among the random $\Omega(u)$ with zero mass under $\mathcal{D}_t$. If these are in the training data S, which is the most likely case as $\bar{S}$ has cardinality only $\Theta(d/\gamma^2)$, then these $O(d/\gamma^2)$ points will have non-zero mass in most other $\mathcal{D}_{t'}$, allowing a boosting algorithm to correct the negative predictions.

The above proof outline can be seen to work for any boosting algorithm producing voting classifiers and that always invokes the weak learner with a probability distribution that is strictly positive on all of the training data. For this reason, our lower bound argument also applies to many other classic boosting algorithms as mentioned in Section 1. In addition to showing that these algorithms are sub-optimal, we believe our lower bound may help inspire new boosting algorithms. Concretely, as just sketched above, if the weak learner was invoked with probability distributions that have mass on only a constant fraction of the training data, our argument breaks down. In fact, the optimal weak-to-strong learner by Larsen & Ritzert [10] precisely samples subsets of the training data and runs AdaBoost$^*_\nu$ on such subsets. Perhaps a similar sub-sampling could be used without the two-level majority. We leave this as an exciting direction for future research.

3 AdaBoost is not Optimal

In this section, we prove our main result that AdaBoost is not an optimal weak-to-strong learner.

In the following, we let $\mathcal{X} = [u] = \{1, \ldots, u\}$ be the universe where u is the universe size. Further we

let $\Delta_{\mathcal{X}}$ be the set of probability distributions over $\mathcal{X}$. In our construction, we use the all ones hypothesis, i.e. $h_0^\star(x) = 1$ for all $x \in \mathcal{X}$, as the underlying concept that is to be learned. Since we do not consider any other concept, the error of a hypothesis f under a distribution $\mathcal{D} \in \Delta_{\mathcal{X}}$ is given as

$$\mathcal{L}_{\mathcal{D}}(f) = P_{x\sim\mathcal{D}}\left[f(x) \neq 1\right].$$

This is equivalent to $\mathcal{L}_{\mathcal{D}}(f) = \sum_{i=1}^{u} \mathcal{D}(i)(1 - f(i))/2$ such that we can write the error requirement of a γ-weak learner as

$$\sum_{i=1}^{u} \mathcal{D}(i) f(i) \geq 2\gamma$$

which we will use in the analysis.

In our construction, we will need the hypothesis h_0, which is "close" to the all ones hypothesis $h_0^\star$. Let h_0 be the hypothesis from $\mathcal{X}$ into $\{-1, 1\}$ such that $h_0(i) = 1$ for $i = 1, \ldots, u - r_1$ and $h_0(i) = -1$ for $i = u - r_1 + 1, \ldots, u$, for r_1 to be defined later (think of r_1 as small compared to u). Let $\mathcal{A}$ be any learning algorithm which takes as input a sample S and a weak learner $\mathcal{W}$, and satisfies the following:

Properties 1.

1. $\mathcal{A}$ outputs a weighted majority classifier, i.e. a classifier of the form $\operatorname{sign}(\sum_i w_i h_i)$ where w_i are non-negative weights with $\sum_i w_i = 1$ and h_i are hypotheses obtained from the weak learner $\mathcal{W}$. The weights w_i only depend on the performance of the h_i's on S (i.e. w_i may depend on $h_j(S)$ for $j \neq i$ but not on any $h_j(x)$ for $x \notin S$).
2. In every query to the weak learner $\mathcal{W}$, the algorithm $\mathcal{A}$ provides a distribution $\mathcal{D} \in \Delta_{\mathcal{X}}$ with $\operatorname{supp}(\mathcal{D}) = S$ ($\mathcal{D}_i > 0$ for $i \in S$ and 0 otherwise).
3. The learning algorithm $\mathcal{A}$ provides the true labels to the items in the sample in its query to $\mathcal{W}$.

The conditions above are necessary and sufficient for our construction of the adversarial weak learner. 1) ensures that the learning algorithm actually uses the weak learner to compute the majority classifier with weights based only on the samples in S (and not $\bar{S}$). 2) gives away the sample to the adversarial weak learner such that it can accumulate errors outside the sample i.e. on points in $\bar{S} = \mathcal{X}\backslash S$. And 3) ensures that the weak learner is always asked to learn the all ones hypothesis, so we only need to guarantee an advantage of γ on that. Under those conditions, $\mathcal{D}$ already encodes S such that we view the weak learner as a function of a distribution $\mathcal{D} \in \Delta_{\mathcal{X}}$, instead of a function of $\mathcal{D}$ and the sample S. Furthermore, we write $\mathcal{W}_{\mathcal{H}}$ to make the hypothesis set $\mathcal{H}$ that is used by a weak learner explicit.

The following lower bound is a more general version of Theorem 1.1. Since AdaBoost satisfies the above properties, the lower bound applies to AdaBoost as well.

Theorem 3.1. *There exist a universal constant $c \leq 1$ such that for $\gamma \leq c$, $d \geq \ln(1/\gamma)$, and $d\gamma^{-2}/16 \leq m \leq \exp(\exp(d))$ there exist a universe $\mathcal{X}$, a distribution $\mathcal{D} \in \Delta_{\mathcal{X}}$, a hypothesis set $\mathcal{H}$ of VC-dimension $O(d)$, and a weak learner $\mathcal{W}_{\mathcal{H}}$ on $\mathcal{H}$ for the all one hypothesis i.e.*

$$\forall \mathcal{D} \in \Delta_{\mathcal{X}} : \quad \sum_{i\in[u]} \mathcal{D}(i)\, \mathcal{W}_{\mathcal{H}}(\mathcal{D})(i) \geq 2\gamma,$$

such that for any learning algorithm $\mathcal{A}$ satisfying Properties 1, we have with constant probability over $S \sim \mathcal{D}^m$:

$$\mathcal{L}_{\mathcal{D}}(\mathcal{A}(S, \mathcal{W}_{\mathcal{H}})) = \Omega\left(\frac{d \ln\left(m\gamma^2/d\right)\right)}{m\gamma^2}\right)$$

Formally, Theorem 1.1 follows from Theorem 3.1 by invoking it with $d' = O(d)$ (implying $m = \exp(\exp(O(d)))$ and solving the loss $\mathcal{L}_{\mathcal{D}}(\text{AdaBoost}(S, \mathcal{W}_{\mathcal{H}})) = \varepsilon = C\,\frac{d\ln\left(m\gamma^2/d\right)}{m\gamma^2}$ for m.

To prove Theorem 3.1 we use the following three lemmas whose proofs are deferred to Section 4. The first lemma is a concentration inequality for linear combinations of independent, negatively biased $\{-1, 1\}$-variables. Notationwise, we denote a fixed hypothesis set by $\mathcal{H}$ and a random one by $\mathbf{H}$. Similarly, a concrete hypothesis (which can be encoded by a vector) is denoted by h and a random hypothesis by $\mathbf{h}$.

Lemma 3.2. *Let $w \in \mathbb{R}^d$ such that $\|w\|_1 = 1$ and let $\tilde{\alpha} \geq 1$. Let further $\mathbf{h}$ be a random vector in $\{-1,1\}^d$ with i.i.d. entries such that $\mathbb{P}[\mathbf{h}(i) = 1] = 1/2 - \tilde{\alpha}\beta$ and $\mathbb{P}[\mathbf{h}(i) = -1] = 1/2 + \tilde{\alpha}\beta$ where $\beta < 1/(2\tilde{\alpha})$. We then have for $\alpha' < \tilde{\alpha}$ that*

$$\mathbb{P}\left[\sum_{i=1}^{d} w_i \mathbf{h}(i) \leq -\alpha'\beta\right] \geq \min\left(\frac{1}{4}, \frac{1}{2} - \frac{4\tilde{\alpha}\alpha'}{(2\tilde{\alpha} - \alpha')^2}\right).$$

The lemma will be used to get the $-\gamma$ advantages outside the sample S as described in the proof overview. The second lemma is of a coupon collector style.

Lemma 3.3. *Let $\zeta m/\ln(m/r)$ be the number of coupons where $m \geq 4r$, $r \geq 1$, and $\zeta \geq 8$. Let X denote the number of samples with replacement from the coupons before seeing $\zeta m/\ln(m/r) - 2r$ distinct coupons, then $\mathbb{P}[X \leq m] \leq \frac{1}{2}$*

In the proof we virtually split the universe into a main part and the last r_1 points and are interested in the probability of sampling a training set $S \in \mathcal{S}_{\text{part1}} := \{S : |\bar{S} \cap [u - r_1]| \geq r\}$ (for some r and $r \leq r_1$) capturing the case that there are "enough" unsampled points in the main part of the universe. We will use Lemma 3.3 and carefully chosen constants to show that this probability is at least constant.

The third lemma describes properties of two functions which we combine to get the random adversarial weak learner $\mathcal{W}_{\mathbf{H}}$.

Lemma 3.4. *Let $c_0, c_1 \leq 1$, and $c_2 \geq 1$ denote universal constants. For a universe $\mathcal{X}$ of size u, integers r, r_1 with $r_1 = \alpha^2 r$ for $\alpha \geq 1$, and $\gamma \leq c_0/(2\alpha)$ there exist two independent random hypothesis sets $\mathbf{H}^1$ and $\mathbf{H}^2$ such that*

- *For $\mathbf{H} \coloneqq \mathbf{H}^1 \cup \mathbf{H}^2$ and $k = \ln(u)\gamma^{-2}$,*

$$|\mathbf{H}| \leq 4c_1^{-2} k \ln(k/\delta) \exp(8c_2\gamma^2 r_1) + 1 \tag{4}$$

- *There exists a mapping $g_{\mathbf{H}^1} : \Delta_{\mathcal{X}} \to \mathbf{H}^1$ such that for $r_1 \geq 40\lg(|\mathbf{H}^1|)$ and $S \in \mathcal{S}_{part1} := \{S : |\bar{S} \cap [u - r_1]| \geq r\}$, the mapping $g_{\mathbf{H}^1}$ and the hypothesis set $\mathbf{H}^1$ satisfy the following four properties with probability at least $1 - \delta - 2^{-0.01 r_1}$ (over the outcome of $\mathbf{H}^1$):*
  1. *For any distribution $\mathcal{D} \in \mathcal{D}_S := \{\mathcal{D} : \mathcal{D}(i) > 0 \text{ for } i \in S \text{ else } \mathcal{D}(i) = 0, \|\mathcal{D}\|_1 = 1\}$ supported on S, $\sum_{i \in S} \mathcal{D}(i) g_{\mathbf{H}^1}(\mathcal{D})(i) \geq \gamma/4$.*
  2. *Let $F_{r,S}$ denote the first r points from $\bar{S} \cap [u - r_1]$ and recall that $supp(\mathcal{D}) = S$. If for $\mathcal{D} \in \mathcal{D}_S$, $g_{\mathbf{H}^1}(\mathcal{D}) \neq h_0$, then the hypothesis $g_{\mathbf{H}^1}(\mathcal{D})$ has $(1/2 + \alpha\gamma/2)r$ minus signs in $F_{r,S}$. Further, the outcome of $g_{\mathbf{H}^1}(\mathcal{D})$ on $F_{r,S}$ is uniformly distributed among all vectors in $\{-1,1\}^r$ which have at least $(1/2 + \alpha\gamma/2)r$ minus signs.*
  3. *The randomness over $F_{r,S}$ in Item 2 is independent for all hypotheses in $\{g_{\mathbf{H}^1}(\mathcal{D}) \text{ for } \mathcal{D} \in \Delta_{\mathcal{X}}\}$. Further, the outcome of $g_{\mathbf{H}^1}$ on $F_{r,S}$ is independent of $g_{\mathbf{H}^1}$ on $\overline{F_{r,S}}$.*
  4. *For any weight vector $w \in \Delta_{\mathbf{H}^1 \setminus h_0} := \{w \in \mathbb{R}^{|\mathbf{H}^1|} : 0 \leq w_i, w_0 = 0, \sum_{i \in |\mathbf{H}^1|} w_i = 1\}$ weighing the hypotheses in $\mathbf{H}^1$, we have for at least $r_1/10$ of the i's in $\{u - r_1 + 1, \ldots, u\}$, that $\sum_{j \in |\mathbf{H}^1|} w_j h_j(i) \leq 14\sqrt{\lg(|\mathbf{H}^1|)/r_1}$.*
- *There exists a mapping $t_{\mathbf{H}^2} : \mathcal{D} \to \mathbf{H}^2$ such that with probability at least $1 - \delta$ over $\mathbf{H}^2$, it holds for all $\mathcal{D} \in \Delta_{\mathcal{X}}$ that $\sum_{i \in [u]} \mathcal{D}(i) t_{\mathbf{H}^2}(\mathcal{D})(i) \geq \gamma/4$.*

Let us carefully go over the statements in Lemma 3.4. The first bullet bounds the size of the hypothesis set $\mathbf{H}$, ensuring that its VC-dimension is at most $O(d)$. The second and third bullet consider the functions $g_{\mathbf{H}^1}$ and $t_{\mathbf{H}^2}$ from which we construct the weak learner $\mathcal{W}_{\mathbf{H}}$. These functions, as well as $\mathcal{W}_{\mathbf{H}}$, take as input a distribution and output a hypothesis from $\mathbf{H}$. The key idea is that whenever $g_{\mathbf{H}^1}$ outputs a hypothesis with sufficient advantage on $\mathcal{D}$, $\mathcal{W}_{\mathbf{H}}$ will use that hypothesis (and therefore $g_{\mathbf{H}^1}$ to compute it), and otherwise $\mathcal{W}_{\mathbf{H}}$ will use the hypothesis computed by $t_{\mathbf{H}^2}$. We thus think of $t_{\mathbf{H}^2}$ as a safety mechanism

that ensures that we can always get the required advantage $\Omega(\gamma)$ which is guaranteed by the last bullet of Lemma 3.4. We will call the lemma with 8γ instead of γ to achieve an advantage of 2γ.

With this in mind, consider the second bullet of Lemma 3.4 and consider some $S \in \mathcal{S}_{\text{part1}}$. Let us denote by E_S the event that the four properties in the bullet hold for S and $\mathbf{H}^1$. Now assume a weak-to-strong learning algorithm $\mathcal{A}$ that satisfies 1), 2), and 3) from Properties 1 and that receives an $S \in \mathcal{S}_{\text{part1}}$, i.e. at least r of the unsampled points receive a positive label under the hypothesis h_0. Assume further that E_S occurs. Then our weak learner $\mathcal{W}_{\mathbf{H}}$ has the following interesting properties.

First, in this case, the weak learner $\mathcal{W}_{\mathbf{H}}$ always returns a hypothesis produced by $g_{\mathbf{H}^1}$. This holds as Item 1 of the second bullet guarantees a sufficient advantage regardless of what distribution $\mathcal{A}$ queries the weak learner with.

From the second bullet's Item 2 and Item 3, we get that $\mathcal{A}$ can not put too much mass on the hypotheses provided by $g_{\mathbf{H}^1}$ (those different from h_0), without making at least $\Omega(r)$ mistakes on the r unsampled points $F_{r,S}$. These mistakes would imply an error of at least $\Omega\left((d\ln(m\gamma^2/d))/(m\gamma^2)\right)$.

Finally, Item 4 gives us that $\mathcal{A}$ can neither put too much mass on h_0, without making $\Omega(r)$ mistakes on the last r_1 points of $\mathcal{X}$. Combining this with the previous point gives the desired lower bound. We now give the proof of Theorem 3.1.

Proof of Theorem 3.1. Let γ, d, and m be as in Theorem 3.1. Let the concept that $\mathcal{A}$ is trying to learn be the all ones hypothesis $h_0^\star$. We now show the existence of a universe $\mathcal{X}$, a hypothesis set $\mathcal{H}$ of VC-dimension at most d, and a γ-weak learner $\mathcal{W}_{\mathcal{H}} : \Delta_{\mathcal{X}} \to \mathcal{H}$ for $h_0^\star$ (mapping distributions over $\mathcal{X}$ to hypotheses from $\mathcal{H}$), such that when $\mathcal{A}$ uses hypotheses from $\mathcal{W}_{\mathcal{H}}$ and receives samples from the uniform distribution $\mathcal{U}$ on $\mathcal{X}$, then with constant probability over the sample $\mathbf{S} \sim \mathcal{U}^m$, it has an error of $\mathcal{L}_{\mathcal{U}}(\mathcal{A}(S, \mathcal{W}_{\mathcal{H}})) = \Omega\left((d\ln\left(m\gamma^2/d\right))/(m\gamma^2)\right)$.

To show the existence of such a hypothesis set and weak learner, we show for a random hypothesis set $\mathbf{H}$ (with VC-dimension $O(d)$) that we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{H}}\left[\mathbb{P}_{\mathbf{S}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(\mathbf{S}, \mathcal{W}_{\mathbf{H}})) \geq C\,\frac{d\ln\left(m\gamma^2/d\right)}{m\gamma^2},\ \forall \mathcal{D} \in \Delta_{\mathcal{X}} : \sum_{i\in[u]} \mathcal{D}(i)\,\mathcal{W}_{\mathbf{H}}(\mathcal{D})(i) \geq 2\gamma\right]\right] \\
&= \mathbb{E}_{\mathbf{S}}\left[\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(\mathbf{S}, \mathcal{W}_{\mathbf{H}})) \geq C\,\frac{d\ln\left(m\gamma^2/d\right)}{m\gamma^2},\ \forall \mathcal{D} \in \Delta_{\mathcal{X}} : \sum_{i\in[u]} \mathcal{D}(i)\,\mathcal{W}_{\mathbf{H}}(\mathcal{D})(i) \geq 2\gamma\right]\right] \geq \frac{1}{64} \qquad (5)
\end{aligned}$$

for some universal constant C. Here, the first part states that $\mathcal{A}$ has a large error while the second part ensures that $\mathcal{W}_{\mathbf{H}}$ is indeed a weak learner. As the event of $\mathcal{W}_{\mathbf{H}}$ being a weak learner is independent of $\mathbf{S}$, the expectation implies that there exists a concrete hypothesis set $\mathcal{H}$ such that $\mathcal{W}_{\mathcal{H}}$ is a weak learner and with constant probability over the sample $\mathbf{S}$, the algorithm $\mathcal{A}$ has error probability $\Omega\left(d\ln\left(m\gamma^2/d\right)\right)/(m\gamma^2)\right)$ when using $\mathcal{W}_{\mathcal{H}}$ as its weak learner. The equality uses that a probability can be written as the expectation of an indicator variable.

Establishing Equation (5). Our adversarial weak learner accumulates errors on r elements in $\bar{S}$, such that the overall error is connected to the fraction r/u. Next, we show that we can invoke Lemma 3.4 with parameters such that $r/u \geq C(d\ln\left(m\gamma^2/d\right))/(m\gamma^2)$ for some universal constant C, and where $t_{\mathbf{H}^2}$ is a weak learner with probability at least $1-\delta$ for $\delta = 1/4$. Using this, we can phrase Equation (5) as

$$\mathbb{E}_{\mathbf{S}}\left[\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(\mathbf{S}, \mathcal{W}_{\mathbf{H}})) \geq \frac{r}{10u},\ \forall \mathcal{D} \in \Delta_{\mathcal{X}} : \sum_{i\in[u]} \mathcal{D}(i)\,\mathcal{W}_{\mathbf{H}}(\mathcal{D})(i) \geq 2\gamma\right]\right] > \frac{1}{64}. \qquad (6)$$

We now show that such a choice of parameters is indeed possible.

Preliminary Setting of Parameters. Let $m \geq 8$ be the sample size and $\gamma' = 8\gamma$ where γ is the (sufficiently small) advantage needed for the weak learner. This choice implies that the weak learner constructed in Lemma 3.4 has a 2γ advantage.

Now, let $u = 8\alpha^2 m/\ln(m/r)$ be the universe size where $r := d\gamma'^{-2}$ and where the value of α will be chosen larger than 1. From the assumption $m \geq d\gamma^{-2}/16$ in the theorem we get that $m/r \geq 4$ and thus $\ln(m/r)$ is non-negative. We now choose $r_1 = \alpha^2 r = \alpha^2 d\gamma'^{-2}$. In the definition of h_0 the last r_1 positions return -1, thereby splitting the universe in a "first" and "second" part. Note that $u \geq 8\alpha^2 m/\ln(m/r) \geq 8\alpha^2 r \geq 8r_1$ (using that $x/\ln x > 1$ for $x > 1$ in the second inequality), thus we may assume that the set of samples $\mathcal{S}_{\text{part1}} := \{S : |\bar{S} \cap [u - r_1]| \geq r\}$ is not $\emptyset$.

We wish to invoke Lemma 3.4 with u, $\gamma = \gamma'$, r, α, and $\delta = 1/4$ as above. First, Equation (4) guarantees that the size of $\mathbf{H}$ in Lemma 3.4 is upper bounded by $4c_0^{-2}k\ln(k/\delta)\exp(8c_2\gamma'^2 r_1) + 1 \leq 5c_0^{-2}k\ln(k/\delta)\exp(8c_2\gamma'^2 r_1)$. Lemma 3.4 only holds when $\gamma' \leq c_0/(2\alpha)$. We guarantee this with the constraint in Theorem 3.1 saying that $\gamma \leq c$, where c is less than $c_0/(16\alpha)$.

We now decide on the choice of α. Later in the proof, we will need that $\ln(|\mathbf{H}|)/r_1 \leq c_3\gamma^2$ where c_3 is a universal constant that will determine the concrete value of α. To upper bound $\ln(|\mathbf{H}|)/r_1$, we first notice that since $m \leq \exp(\exp(d))$ we get that $\ln(\ln(u)) \leq \ln(\ln(8\alpha^2 m)) \leq \ln(\ln(8\alpha^2)) + d$. Further, since $d \geq \ln(1/\gamma)$ (one of the conditions in Theorem 3.1) we get that $\ln(k) = \ln(\ln(u)\,\gamma'^{-2}) \leq \ln(\ln(8\alpha^2)) + 3d$. By these two inequalities as well as $\delta = 1/4$ and $r_1 = \alpha^2 d\gamma'^{-2}$ we get that

$$\ln(|\mathbf{H}|) \leq 8c_2\gamma'^2 r_1 + \ln(5c_0^{-2}) + \ln(k) + \ln(\ln(k/\delta)) \leq \ln(5c_0^{-2}) + 5(\ln(\ln(8\alpha^2)) + (8c_2\alpha^2 + 3)d. \quad (7)$$

implying that for any $\alpha, d \geq 1$ if we choose $c_3 = \big(\ln(5c_0^{-2}) + 5\ln(\ln(8)) + (8c_2 + 3)\big)8^2$

$$\frac{\ln(|\mathbf{H}|)}{r_1} \leq \left(\ln(5c_0^{-2}) + \frac{5\ln(\ln(8\alpha^2))}{\alpha^2 d} + 8c_2 + \frac{3}{\alpha^2}\right)8^2\gamma^2 \leq c_3\gamma^2 \quad (8)$$

since the middle expression in Equation (8) is decreasing in $\alpha, d \geq 1$. This allows us to fix $\alpha = 5 \cdot 28\sqrt{c_3}$.

Further notice that Equation (8), the before mentioned constraint $\gamma \leq c_0/(16\alpha)$, $c_0 \leq 1$ implied by Lemma 3.4, and the now fixed $\alpha = 5 \cdot 28\sqrt{c_3}$, $c_3 \geq 1$ implies that $r_1 \geq \ln(|\mathbf{H}|)/(c_3\gamma^2) \geq 40\lg(|\mathbf{H}^1|)$. This a condition for the second bullet of Lemma 3.4 to hold. We thus have that we can invoke Lemma 3.4 as claimed.

Bounded VC-Dimension. Using the parameters we have chosen above, we can now bound the VC-dimension of $\mathbf{H}$. Here we use that the VC-dimension of $\mathbf{H}$ is trivially bounded by $\ln|\mathbf{H}|/\ln(2)$. Together with the size bound on $\mathbf{H}$ from Equation (7) we get that the VC-dimension of $\mathbf{H}$ is $O(d)$ as claimed.

We now construct our weak learner $\mathcal{W}$ in the following way using $g_{\mathbf{H}^1}$ and $t_{\mathbf{H}^2}$ from Lemma 3.4.

$$\begin{aligned}\mathcal{W}_{\mathbf{H}}(\mathcal{D}) = {} & \mathbb{1}_{\sum_{i=1}^{u}\mathcal{D}(i)g_{\mathbf{H}^1}(\mathcal{D})(i)\geq 2\gamma}\, g_{\mathbf{H}^1} \\ & + \mathbb{1}_{\sum_{i=1}^{u}\mathcal{D}(i)g_{\mathbf{H}^1}(\mathcal{D})(i)<2\gamma}\, t_{\mathbf{H}^2}(\mathcal{D}) \quad \forall\mathcal{D} \in \Delta_{\mathcal{X}},\end{aligned}$$

Said in words, $\mathcal{W}_{\mathbf{H}}$ is $g_{\mathbf{H}^1}$ when $g_{\mathbf{H}^1}$ achieves an advantage of 2γ and it defaults back to $t_{\mathbf{H}^2}$ otherwise.

First, we notice that if $t_{\mathbf{H}^2}$ is a weak learner, then $\mathcal{W}_{\mathbf{H}}$ is also a weak learner. Thus we can replace the weak learning requirement on $\mathcal{W}_{\mathbf{H}}$ by a similar requirement on $t_{\mathbf{H}^2}$, implying

$$\begin{aligned}& \mathbb{E}_{\mathbf{S}}\left[\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(\mathbf{S},\mathcal{W}_{\mathbf{H}})) \geq \frac{r}{10u},\ \forall\mathcal{D} \in \Delta_{\mathcal{X}} : \sum_{i\in[u]}\mathcal{D}(i)\,\mathcal{W}_{\mathbf{H}}(\mathcal{D})(i) \geq 2\gamma\right]\right] \\ \geq\ & \mathbb{E}_{\mathbf{S}}\left[\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(\mathbf{S},\mathcal{W}_{\mathbf{H}})) \geq \frac{r}{10u},\ \forall\mathcal{D} \in \Delta_{\mathcal{X}} : \sum_{i\in[u]}\mathcal{D}(i)\,t_{\mathbf{H}^2}(\mathcal{D})(i) \geq 2\gamma\right]\right].\end{aligned}$$

Further notice that if we have a sample S, then $\mathcal{A}$ would by Item 2) in Properties 1 only give inputs $\mathcal{D}$ in $\mathcal{D}_S := \{\mathcal{D} : \mathcal{D}(i) > 0 \text{ for } i \in S \text{ else } \mathcal{D}(i) = 0, \|\mathcal{D}\|_1 = 1,\}$ to the weak learner $\mathcal{W}_{\mathbf{H}}$. Thus, we have for a fixed sample S and the definition of $\mathcal{W}_{\mathbf{H}}$ that

$$\begin{aligned}& \left\{\mathcal{H} = (\mathcal{H}^1 \cup \mathcal{H}^2) : \mathcal{L}_{\mathcal{U}}(\mathcal{A}(S, g_{\mathcal{H}^1})) \geq \frac{r}{10u},\ \forall\mathcal{D} \in \mathcal{D}_S \sum_{i\in[u]}\mathcal{D}(i)\,g_{\mathcal{H}^1}(\mathcal{D})(i) \geq 2\gamma\right\} \\ \subseteq & \left\{\mathcal{H} : \mathcal{L}_{\mathcal{U}}(\mathcal{A}(S,\mathcal{W}_{\mathcal{H}})) \geq \frac{r}{10u}\right\}\end{aligned}$$

where we use that $\mathcal{W}_{\mathbf{H}}$ becomes $g_{\mathbf{H}^1}$ when $g_{\mathbf{H}^1}$ produces large margins. Thus, we conclude that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S}}\left[\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(\mathbf{S},\mathcal{W}_{\mathbf{H}})) \geq \frac{r}{10u},\ \forall \mathcal{D} \in \Delta_{\mathcal{X}} : \sum_{i\in[u]} \mathcal{D}(i)\, t_{\mathbf{H}^2}(\mathcal{D})(i) \geq 2\gamma\right]\right] \\
&\geq \mathbb{E}_{\mathbf{S}}\left[\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(\mathbf{S},g_{\mathbf{H}^1})) \geq \frac{r}{10u},\ \forall \mathcal{D} \in \mathcal{D}_{\mathbf{S}} : \sum_{i\in[u]} \mathcal{D}(i)\, g_{\mathbf{H}^1}(\mathcal{D})(i) \geq 2\gamma,\right.\right. \\
&\qquad\qquad \left.\left.\forall \mathcal{D} \in \Delta_{\mathcal{X}} : \sum_{i\in[u]} \mathcal{D}(i)\, t_{\mathbf{H}^2}(\mathcal{D})(i) \geq 2\gamma\right]\right] \\
&\geq \mathbb{E}_{\mathbf{S}}\left[\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(\mathbf{S},g_{\mathbf{H}^1})) \geq \frac{r}{10u},\ \forall \mathcal{D} \in \mathcal{D}_{\mathbf{S}} \sum_{i\in[u]} \mathcal{D}(i)\, g_{\mathbf{H}^1}(\mathcal{D})(i) \geq 2\gamma\right]\right](1-\delta) \qquad (9)
\end{aligned}$$

where the last inequality follows from the last point of Lemma 3.4, which says that $t_{\mathbf{H}^2}$ is a weak learner with probability at least $1-\delta$ and $t_{\mathbf{H}^2}$ is independent of $g_{\mathbf{H}^1}$.

We will now show that

$$\mathbb{P}_{\mathbf{S}}[\mathbf{S} \in \mathcal{S}_{\text{part1}}] \geq 1/4, \qquad (10)$$

and for any sample S in the set $\mathcal{S}_{\text{part1}} := \{S : |\bar{S} \cap [u-r_1]| \geq r\}$ (from Lemma 3.4) we have that

$$\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(S,g_{\mathbf{H}^1})) \geq \frac{r}{10u},\ \forall \mathcal{D} \in \mathcal{D}_S : \sum_{i\in[u]} \mathcal{D}(i)\, g_{\mathbf{H}^1}(\mathcal{D})(i) \geq 2\gamma\right] \geq \frac{1}{12}. \qquad (11)$$

Now combining Equation (9), Equation (10), Equation (11), and $\delta = 1/4$ we get

$$\mathbb{E}_{\mathbf{S}}\left[\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(\mathbf{S},\mathcal{W}_{\mathbf{H}})) \geq \frac{r}{10u},\ \forall \mathcal{D} \in \Delta_{\mathcal{X}} : \sum_{i\in[u]} \mathcal{D}(i)\, \mathcal{W}_{\mathbf{H}}(\mathcal{D})(i) \geq 2\gamma\right]\right] \geq \frac{1}{64}$$

as desired. Thus, if we can show Equation (10) and Equation (11) we are done. Essentially, Equation (10) makes sure that we (often enough) have space in $\bar{S}$ to accumulate errors using $g_{\mathbf{H}^1}$. Equation (11) gives us that if there is space to accumulate errors, many of the random hypothesis sets $\mathbf{H}^1$ allow us to actually do so. Equation (9) accounts for the behavior of the weak learner, i.e. its decision rule between the adversarial function $g_{\mathbf{H}^1}$ and the 'normal' weak learner $t_{\mathbf{H}^2}$.

Establishing Equation (10): Recall that we chose the universe size to be $u = 8\alpha^2 m/\ln(m/r)$ and the sample distribution to be uniform on $\mathcal{X} = [u]$ (corresponding to drawing with replacement from $\mathcal{X}$). Further we had $r = d\gamma'^{-2}$ which by the assumption $m \geq d\gamma^{-2}/16$, implied that $m/r \geq 4$. Using this, we get from Lemma 3.3 with $\zeta = 8\alpha^2 \geq 8$ that with probability at least $1/2$ there are $2r$ points in u that are not sampled into $\mathbf{S}$. Further, by the choice of $r_1 = \alpha^2 r$ we noticed that $u = 8\alpha^2 m/\ln(m/r) \geq 8r_1$ thus the universe has at least 8 times the size of r_1. Using this together with $r \leq r_1$ (since $\alpha \geq 1$) and the sampling distribution being uniform/with replacement, we conclude that at least half of the samples where $2r$ points were not sampled in $\mathcal{X}$ have r entries outside of $\{u-r_1+1,\ldots,u\}$ implying $r \leq |\bar{S} \cap [u-r_1]|$, i.e. $S \in \mathcal{S}_{\text{part1}}$. Thus, we conclude that $\mathbb{P}_{\mathbf{S}}[\mathbf{S} \in \mathcal{S}_{\text{part1}}] \geq \mathbb{P}_{\mathbf{S}}[|\mathbf{S}| \leq u-2r]/2 \geq 1/4$ which shows Equation (10).

Establishing Equation (11): For Equation (11) let S be in $\mathcal{S}_{\text{part1}}$ and notice that by Lemma 3.4 we have with probability at least $1-\delta-2^{-0.01r_1}$ over $\mathbf{H}$ that all the 4 items regarding $g_{\mathbf{H}^1}$ in Lemma 3.4 hold. Let E_S denote the corresponding event that those 4 properties regarding $g_{\mathbf{H}^1}$ in Lemma 3.4 hold. In particular, Item 1 says that $g_{\mathbf{H}^1}$ is indeed a weak learner on $\mathcal{D}_S$. Using this event E_S we get that

$$\begin{aligned}
&\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(S,g_{\mathbf{H}^1})) \geq \frac{r}{10u}, \quad \forall \mathcal{D} \in \mathcal{D}_S : \sum_{i\in[u]} \mathcal{D}(i)\, g_{\mathbf{H}^1}(\mathcal{D})(i) \geq 2\gamma\right] \\
&\geq \mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(S,g_{\mathbf{H}^1})) \geq \frac{r}{10u} \,\middle|\, E_S\right](1-\delta-2^{-0.01r_1}). \qquad (12)
\end{aligned}$$

We now show that conditioned on E_S, with probability at least $1/6$ the algorithm $\mathcal{A}$ has an out-of-sample error of at least $r/(10u)$ when using $g_{\mathbf{H}^1}$ as the weak learner, formally $\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(S, g_{\mathbf{H}^1})) \geq \frac{r}{10u} | E_S\right] \geq 1/6$. We further show that $1 - \delta - 2^{-0.01 r_1} \geq 1/2$ which combined with Equation (12) implies Equation (11).

By the definition of the event E_S we know we know that in the event E_S the random hypothesis set $\mathbf{H}$ satisfies the 4 items of the second bullet of Lemma 3.4 (the ones about $g_{\mathbf{H}^1}$). Thus, $g_{\mathbf{H}^1}$ is a weak learner on S by Item 1 and $\mathcal{A}$ terminates using only hypotheses given by $g_{\mathbf{H}^1}$, which satisfy the conditions given in the 4 items. Let $w^{\mathcal{A}} = (w_0^{\mathcal{A}}, \ldots, w_{|\mathbf{H}^1|}^{\mathcal{A}})$ be the weights that $\mathcal{A}$ calculates, where $w_0^{\mathcal{A}}$ is the weight put on h_0. Notice that the weights are random as they depend on the outputs of $g_{\mathbf{H}^1}$ which themselves depend on the random hypothesis set $\mathbf{H}$. From the first item of the second bullet in Lemma 3.4 we know that the weights $w^{\mathcal{A}}$ depend only on $g_{\mathbf{H}^1}(\cdot)(i)$ for $i \in S$. Thus, we get by Item 2 and Item 3 in Lemma 3.4 that the minus signs of $g_{\mathbf{H}^1}$ in the first r points of $\bar{S} \cap [u - r_1]$, which we denoted as $F_{r,S}$, are independent of the weights $w^{\mathcal{A}}$. We will use this property below in the second case. In the following let $\{\mathbf{h}_i\}_{i=1,\ldots,|\mathbf{H}^1|}$ be the hypotheses in $\mathbf{H}^1$. Note that whenever a hypothesis $\mathbf{h}_i$ has a positive weight $w_i > 0$, there must be a distribution $\mathcal{D} \in \Delta_{\mathcal{X}}$ such that $g_{\mathbf{H}^1}(\mathcal{D}) = \mathbf{h}_i$. We now consider two cases for the weight $w_0^{\mathcal{A}}$ of the all-one hypothesis h_0. For this let E_{small} be the event that $w_0^{\mathcal{A}} < 14\sqrt{c_3\gamma^2}/(1 + 14\sqrt{c_3\gamma^2})$.

Case 1: $w_0^{\mathcal{A}} \geq 14\sqrt{c_3\gamma^2}/(1 + 14\sqrt{c_3\gamma^2})$ **($\overline{E_{\text{small}}}$).** Consider the r_1 last points in the universe $\mathcal{X} = [u]$, i.e. the points where h_0 is -1. Thus, for $i \in \{u - r_1 + 1, \ldots, u\}$ we have that the prediction of $\mathcal{A}$ is $w_0^{\mathcal{A}} h_0(i) + \sum_{j=1}^{|\mathbf{H}^1|} w_j^{\mathcal{A}} \mathbf{h}_j(i) = -w_0^{\mathcal{A}} + (1 - w_0^{\mathcal{A}}) \sum_{j=1}^{|\mathbf{H}^1|} w_j^{\mathcal{A}}/(1 - w_0^{\mathcal{A}}) \mathbf{h}_j(i)$, where we have used that $h_0(i) = -1$ for $i \in \{u - r_1 + 1, \ldots, u\}$. Now, conditioned on E_S we know by Item 4 in Lemma 3.4 that for any weighted combination of $(\mathbf{h}_j)_{1,\ldots,|\mathbf{H}^1|}$ there are at least $r_1/10$, i's in $\{u - r_1 + 1, \ldots, u\}$ where the linear combination is at most $14\sqrt{\lg(|\mathbf{H}^1|)/r_1}$ i.e. for such i's we have $\sum_{j=1}^{\mathbf{H}^1} w_j \mathbf{h}_j(i) \leq 14\sqrt{\lg(|\mathbf{H}^1|)/r_1}$. By Equation (8) and $|\mathbf{H}^1| < |\mathbf{H}|$ we know that $14\sqrt{\lg(|\mathbf{H}^1|)/r_1}$ is strictly less than $14\sqrt{c_3\gamma^2}$. Thus, we get for such elements i that $-w_0^{\mathcal{A}} + (1 - w_0^{\mathcal{A}}) \sum_{j=1}^{|\mathbf{H}|} w_j^{\mathcal{A}}/(1 - w_0^{\mathcal{A}}) \mathbf{h}_j(i) < -w_0^{\mathcal{A}} + (1 - w_0^{\mathcal{A}}) 14\sqrt{c_3\gamma^2}$, which for $w_0^{\mathcal{A}} \geq 14\sqrt{c_3\gamma^2}/(1 + 14\sqrt{c_3\gamma^2})$ is less than zero. Thus, conditioned on E_S, if $\mathcal{A}$ puts more than $14\sqrt{c_3\gamma^2}/(1 + 14\sqrt{c_3\gamma^2})$ mass on $w_0^{\mathcal{A}}$, then $\mathcal{A}$ gets at least $r_1/10 \geq r/10$ points misclassified, resulting in an out of sample error of at least $r/(10u)$. Thus, we conclude that

$$\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(S, g_{\mathbf{H}^1})) \geq \frac{r}{10u}, \overline{E_{\text{small}}} \,\middle|\, E_S\right] = \mathbb{P}_{\mathbf{H}}\left[\, \overline{E_{\text{small}}} \,\middle|\, E_S \,\right]. \tag{13}$$

Case 2: $w_0^{\mathcal{A}} < 14\sqrt{c_3\gamma^2}/(1 + 14\sqrt{c_3\gamma^2})$ **(E_{small}).** Let R be the set of all indices of hypotheses with nonzero weights in $w^{\mathcal{A}}$ except the index of h_0. Notice that R depends on the vector if weights $w^{\mathcal{A}}$ which depends on the random hypothesis set $\mathbf{H}$, making R random too. Further, by the comments before Case 1, $i \in R$ implies that $\exists \mathcal{D} \in \Delta_{\mathcal{X}}$ such that $g_{\mathbf{H}^1}(\mathcal{D}) = \mathbf{h}_i$. Thus, we have by Item 2 of Lemma 3.4 that for every $j \in R$ the vector $(\mathbf{h}_j(i))_{i \in F_{r,S}}$ corresponds to a random vector of length r with at least $(1/2 + 8\alpha\gamma/2)r$ minus signs and the vector is uniformly distributed between all permutations of $\{-1, 1\}^r$ with at least $(1/2 + 8\alpha\gamma/2)r$ minus signs (where we used $\gamma' = 8\gamma$). Further, Item 3 of Lemma 3.4 states that these vectors (one for each hypothesis $j \in R$) are independent of each other and of $(\mathbf{h}_j(i))_{j \in R, i \in \overline{F_{r,S}}}$, which the weights w_i are a function of. Therefore, the vectors $(\mathbf{h}_j(i))_{i \in F_{r,S}}$ for $j \in R$ are also independent of the weights. If we now let $\tilde{w}_j^{\mathcal{A}} := w_j^{\mathcal{A}}/(1 - w_0^{\mathcal{A}})$ for $j \in R$ and use that for every $i \in F_{r,S}$ we know $h_0(i) = 1$, we get for $i \in F_{r,S}$ that

$$\begin{aligned}
&\mathbb{P}_{\mathbf{H}}\left[w_0^{\mathcal{A}} h_0(i) + \sum_{j=1}^{|\mathbf{H}^1|} w_j^{\mathcal{A}} \mathbf{h}_j(i) < 0, E_{\text{small}} \,\middle|\, E_S\right] \\
=&\mathbb{P}_{\mathbf{H}}\left[\sum_{j \in R} \tilde{w}_j^{\mathcal{A}} \mathbf{h}_j(i) < -w_0^{\mathcal{A}}/(1 - w_0^{\mathcal{A}}), E_{\text{small}} \,\middle|\, E_S\right]
\end{aligned}$$

Since $-x/(1-x)$ is decreasing for $0 \leq x \leq 1$ and we have $w_0^{\mathcal{A}} < 14\sqrt{c_3\gamma^2}/(1+14\sqrt{c_3\gamma^2})$, which implies $-w_0^{\mathcal{A}}/(1-w_0^{\mathcal{A}}) > -14\sqrt{c_3\gamma^2}$ and we get that

$$\geq \mathbb{P}_{\mathbf{H}}\left[\sum_{j\in R} \tilde{w}_j^{\mathcal{A}} \mathbf{h}_j(i) \leq -14\sqrt{c_3}\gamma, E_{\text{small}} \;\middle|\; E_S\right]$$

Now using the law of total probability gives us

$$\begin{aligned}
&= \int \mathbb{P}_{\mathbf{H}}\left[\sum_{j\in R} \tilde{w}_j^{\mathcal{A}} \mathbf{h}_j(i) \leq -14\sqrt{c_3}\gamma, E_{\text{small}} \;\middle|\; E_S, \tilde{w}^{\mathcal{A}} = z\right] \\
&\qquad d\mathbb{P}_{\mathbf{H}}\left[\tilde{w}^{\mathcal{A}} = z \mid E_S\right] \\
&= \int_{E_{\text{small}}} \mathbb{P}_{\mathbf{H}}\left[\sum_{j\in R} \tilde{w}_j^{\mathcal{A}} \mathbf{h}_j(i) \leq -14\sqrt{c_3}\gamma \;\middle|\; E_S, \tilde{w}^{\mathcal{A}} = z\right] \\
&\qquad d\mathbb{P}_{\mathbf{H}}\left[\tilde{w}^{\mathcal{A}} = z \mid E_S\right] \qquad (14)
\end{aligned}$$

We will now work towards lower bounding $\mathbb{P}_{\mathbf{H}}\left[\sum_{j\in R} \tilde{w}_j^{\mathcal{A}} \mathbf{h}_j(i) \leq -14\sqrt{c_3}\gamma \;\middle|\; E_S, \tilde{w}^{\mathcal{A}} = z\right]$ by $1/4$ for any $z \in E_{\text{small}}$. As noted above, we have for $i \in F_{r,S}$ that $(\mathbf{h}_j(i))_{j\in R}$ are -1 with probability at least $1/2+4\alpha\gamma$, independent of each other and independent of the weights $w^{\mathcal{A}}$. Thus, using that we chose $\alpha = 5 \cdot 28\sqrt{c_3}$ and by invoking Lemma 3.2 with $\tilde{\alpha} = 4\alpha = 4 \cdot 5 \cdot 28\sqrt{c_3}$ and $\alpha' = 14\sqrt{c_3}$, we get that

$$\frac{4\tilde{\alpha}\alpha'}{(2\tilde{\alpha}-\alpha')^2} = \frac{4(4\cdot 5\cdot 28\sqrt{c_3})\cdot(14\sqrt{c_3})}{\left(2\cdot 4\cdot 5\cdot 28\sqrt{c_3} - 14\sqrt{c_3}\right)^2} = \frac{4^2\cdot 5\cdot 2}{(2\cdot 4\cdot 5\cdot 2-1)^2} \leq \frac{1}{4}.$$

Thus, $\min\left(\frac{1}{4}, \frac{4\tilde{\alpha}\alpha'}{(2\tilde{\alpha}-\alpha')^2}\right)$ in Lemma 3.2 is realized by $\frac{1}{4}$ and we get

$$\mathbb{P}_{\mathbf{H}}\left[\sum_{j\in R} \tilde{w}_j^{\mathcal{A}} \mathbf{h}_j(i) \leq -14\sqrt{c_3}\gamma \;\middle|\; E_S, \tilde{w}^{\mathcal{A}} = z\right] \geq \frac{1}{4}. \qquad (15)$$

Notice that the condition $\gamma \leq 1/(2\tilde{\alpha}) = 1/(8\alpha)$ of Lemma 3.2 is already satisfied since we already imposed the condition $\gamma \leq c_0/(16\alpha)$ with $c_0 \leq 1$ in the main theorem in order to apply Lemma 3.4.

We now consider the error of the points in $F_{r,S}$, or more specifically, the part of the total error that is induced by points from $F_{r,S}$. We get the following upper bound by observing that there are r points in $F_{r,S}$:

$$\begin{aligned}
E_{F_{r,S}} &= (1/u) \sum_{i\in F_{r,S}} \mathbb{1}_{\text{sign}\left(\sum_{j=0}^{|\mathbf{H}^1|} w_j^{\mathcal{A}} \mathbf{h}_j(i)\right)\neq 1} \\
&= (1/u) \sum_{i\in F_{r,S}} \mathbb{1}_{\sum_{j=0}^{|\mathbf{H}^1|} w_j^{\mathcal{A}} \mathbf{h}_j(i) < 0} \\
&\leq r/u.
\end{aligned}$$

By Equation (15) we get that $\mathbb{E}_{\mathbf{H}}\left[E_{F_{r,S}} \mid E_S, \tilde{w}^{\mathcal{A}} = z\right] \geq r/(4u)$. This allows us to use a reverse Chernoff bound from which we get that

$$\mathbb{P}_{\mathbf{H}}\left[E_{F_{r,S}} \geq r/(10u) \;\middle|\; E_S, \tilde{w}^{\mathcal{A}} = z\right] \geq \frac{r/(4u) - r/(10u)}{r/u - r/(10u)} = \frac{1/4 - 1/10}{1 - 1/10} = 1/6. \qquad (16)$$

Using that $\mathcal{L}_{\mathcal{U}} \geq E_{F_{r,S}}$, Equation (16), and following calculations as in Equation (14) we conclude that

$$\begin{aligned}
&\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(S, g_{\mathbf{H}^1})) \geq \frac{r}{10u}, E_{\text{small}} \;\middle|\; E_S\right] \\
&\geq \mathbb{P}_{\mathbf{H}}\left[E_{F_{r,S}} \geq \frac{r}{10u}, E_{\text{small}} \;\middle|\; E_S\right] \\
&= \int_{E_{\text{small}}} \mathbb{P}_{\mathbf{H}}\left[E_{F_{r,S}} \geq r/(10u) \mid E_S, \tilde{w}^{\mathcal{A}} = z\right] \\
&\qquad\qquad d\mathbb{P}_{\mathbf{H}}\left[\tilde{w}^{\mathcal{A}} = z \mid E_S\right] \\
&\geq \mathbb{P}_{\mathbf{H}}\left[E_{\text{small}} \mid E_S\right]/6 \qquad (17)
\end{aligned}$$

Combining the two cases: Now using Equation (13) and Equation (17) we get that

$$\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(S, g_{\mathbf{H}^1})) \geq \frac{r}{10u} \;\middle|\; E_S\right] \geq 1/6. \qquad (18)$$

Combining this with Equation (12) we conclude that

$$\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(S, g_{\mathbf{H}^1})) \geq \frac{r}{10u}, \forall \mathcal{D} \in \mathcal{D}_S : \sum_{i\in[u]} \mathcal{D}(i)\, g_{\mathbf{H}^1}(\mathcal{D})(i) \geq 2\gamma\right] \;\geq\; \frac{1}{6}(1 - \delta - 2^{-0.01 r_1}), \qquad (19)$$

that is, for any $S \in \mathcal{S}_{\text{part1}}$, the function $g_{\mathbf{H}^1}$ is a weak learner on $\mathcal{D}_S$ and $\mathcal{A}$ using $g_{\mathbf{H}^1}$ makes at least $r/(10u)$ errors with probability at least $(1 - \delta - 2^{-0.01 r_1})/6$ over the random hypothesis set $\mathbf{H}$. Now, since $\gamma \leq 1/(16\alpha)$, $c_3 \geq 1$, and $\alpha = 5 \cdot 28\sqrt{c_3}$ we get that $r_1 = \alpha \ln(m)/(8\gamma)^2 \geq 4\alpha^3 \ln(m) \geq 4(5 \cdot 28)^3 \ln(m)$. Using $m \geq 2$ we get that $2^{-0.01 r_1} \leq 1/4$ and since we chose $\delta = 1/4$ we get that

$$\mathbb{P}_{\mathbf{H}}\left[\mathcal{L}_{\mathcal{U}}(\mathcal{A}(S, g_{\mathbf{H}^1})) \geq \frac{r}{10u}, \forall \mathcal{D} \in \mathcal{D}_S : \sum_{i\in[u]} \mathcal{D}(i)\, g_{\mathbf{H}^1}(\mathcal{D})(i) \geq 2\gamma\right] \geq \frac{1}{12},$$

which shows Equation (11) and concludes the proof.

□

4 Proof of Lemmas

In this section, we restate the lemmas from Section 3 and give their proofs. A main part of the proof in Section 3 makes use of the functions $g_{\mathbf{H}^1}$ and $t_{\mathbf{H}^2}$ which on the random hypothesis set $\mathbf{H}$ have "nice" properties (Lemma 3.4). As $g_{\mathbf{H}^1}$ and $t_{\mathbf{H}^2}$ played the main role in Section 3 we start off by proving Lemma 3.4. To prove the lemma, we need the following algorithm which we use to show the existence of a the hypotheses $g_{\mathbf{H}^1}$ and $t_{\mathbf{H}^2}$ will output.

Algorithm 2: Majority Voter

Input: $(\mathcal{H}_1, \ldots, \mathcal{H}_k)$, $S \subset \mathcal{X}$
Output: f adversarial weak learner on S

1 $\eta \leftarrow \ln\left((1+2\gamma)/(1-2\gamma)\right)/2$
2 $f_0(i) \leftarrow 0$ for all $i \in u$
3 $D_1(i) \leftarrow \frac{1}{S}$ for all $i \in S$
4 **for** $j \in \{1, \ldots, k\}$ **do**
5 | **if** $\sum_{i=1, i\in S}^{u-r_1} D_j(i) > 1/2 + \gamma$ **then**
6 | | set $h_j = h_0$ (notice that if this is the case then $\sum_{i\in S} D_j(i)h_j(i) \geq 2\gamma$)
7 | **else if** *there is a hypothesis* $h_j \in H_j$ *such that* $\sum_{i\in S} D_j(i)h_j(i) \geq 2\gamma$ *and* h_j *has* $(1/2 + \alpha\gamma/2)r$ *minus signs on the first* r *elements in* $\bar{S} \cap [u - r_1]$ **then**
8 | | choose this hypothesis
9 | **else**
10 | | **return** Fail
11 | $f_j \leftarrow f_{j-1} + h_j$
12 | $Z_j \leftarrow \sum_{i\in S} D_j(i)\exp\left(-\eta h_j(i)\right)$
13 **for** $i \in S$ **do**
14 | $D_{j+1}(i) \leftarrow D_j(i)\exp\left(-\eta h_j(i)\right)/Z_j$
15 **return** $f = f_k/k$

In the following proof of Lemma 3.4 we will run the above algorithm on a sequence of random hypothesis sets whose union will be $\mathbf{H}$. Running the above algorithm will then create a voting classifier with a γ advantage which implies that one of the hypotheses also has this advantage. Thus, $\mathbf{H}$ contains a hypothesis with a γ advantage that $g_{\mathbf{H}^1}$ or $t_{\mathbf{H}^2}$ can output. In the case of $g_{\mathbf{H}^1}$ we will also make these hypotheses adversarial by using the minus signs in Line 8. For the above argument to go through we need that the random hypothesis set $\mathbf{H}$ contains at least one hypothesis that has a γ advantage given a distribution $\mathcal{D}$ over the universe $\mathcal{X}$ (for all distributions $\mathcal{D}$ that the algorithm computes). This is captured in the following lemma, which we will prove later in this section.

Lemma 4.1. *Let* $c_0, c_1 \leq 1$, *and* $c_2 \geq 1$ *denote some universal constants. Let* $\mathcal{X}$ *be a universe of size* u *and* $\mathcal{D} \in \Delta_{\mathcal{X}}$ *a distribution over* $\mathcal{X}$. *Further let* r *and* r_1 *be non-negative numbers such that* $r_1 = \alpha^2 r$ *for* $\alpha \geq 1$ *and* $r_1 \leq u$. *Let* $0 < \delta \leq 1$, $\gamma \leq c_0/(2\alpha)$, *and* $k = \ln(u)\gamma^{-2}$. *Let* $\mathbf{H}_i$ *be a random hypothesis set consisting of* h_0 *and independent random vectors in* $\{-1, 1\}^u$ *with i.i.d. uniform random entries. Further let the size of* $\mathbf{H}_i$ *be* N/k *without counting* h_0, *where* $N = 2c_1^{-2}k\ln(k/\delta)\exp(8c_2\gamma^2 r_1)$. *With the above, we have with probability at least* $1 - \delta/k$ *over* $\mathbf{H}_i$ *that:*

1. *There exists a hypothesis* $\mathbf{h} \in \mathbf{H}_i$ *such that*

$$\sum_{i\in supp(D)} D_i\mathbf{h}(i) \geq 2\gamma$$

where $\mathbf{h} = h_0$ *if* $\sum_{i=1, i\in supp(D)}^{u-r_1} D_i > 1/2 + \gamma$ *else* $\mathbf{h}$ *is random.*

Further, if $\sum_{i=1, i\in supp(D)}^{u-r_1} D_i \leq 1/2 + \gamma$ *and* $r \leq |\overline{supp(\mathcal{D})} \cap [u - r_1]|$

2. $\mathbf{h}$ *in Item 1 is such that the first* r *entries of* $\{\mathbf{h}(i)\}_{i\in\overline{supp(\mathcal{D})}\cap[u-r_1]}$ *has at least* $(1/2 + \alpha\gamma/2)r$ *minus signs.*

Recall that $supp(D)$ in AdaBoost is just the training set S (without the labels which are all 1 in our setting). Intuitively, the first item states that there is a hypothesis with a sufficient advantage on the training set. In the case that there is not much weight on the first part (where h_0 is positive, i.e. $\mathcal{D}$ focuses on the second part) and there are at least r points in the first part that are not part of the training set, then Item 2 states that we can even find a hypothesis with many minus signs in this first part (outside of

the training data). Since we are trying to learn the all ones hypothesis, those minus signs will induce a large error later on.

Further, we need the following lemma in the proof of Lemma 3.4, to say that for any linear combination over hypotheses in $\mathbf{H}^1$ can not achieve a large advantage on too many points within the last r_1 points of $\mathcal{X}$. Thus, it is impossible to achieve a large advantage where h_0 is -1.

Lemma 4.2. *Let $\mathbf{A}$ be uniform random in $\{-1,1\}^{r\times n}$ and assume that $r \geq 40\lg(n)$. With probability at least $1-2^{-0.01r}$, it holds for all $w \in \mathbb{R}^n$ with $\|w\|_1 = 1$ that $\mathbf{A}w$ has at least $r/10$ entries i with $(\mathbf{A}w)_i < 14\sqrt{\lg(n)/r}$.*

We will prove Lemma 4.2 later in this section. We now restate Lemma 3.4 and give the proof under the assumption that Lemma 4.1 and Lemma 4.2 hold.

Lemma 3.4. *Let $c_0, c_1 \leq 1$, and $c_2 \geq 1$ denote universal constants. For a universe $\mathcal{X}$ of size u, integers r, r_1 with $r_1 = \alpha^2 r$ for $\alpha \geq 1$, and $\gamma \leq c_0/(2\alpha)$ there exist two independent random hypothesis sets $\mathbf{H}^1$ and $\mathbf{H}^2$ such that*

- *For $\mathbf{H} \coloneqq \mathbf{H}^1 \cup \mathbf{H}^2$ and $k = \ln(u)\gamma^{-2}$,*

$$|\mathbf{H}| \leq 4c_1^{-2}k\ln(k/\delta)\exp(8c_2\gamma^2 r_1) + 1 \tag{4}$$

- *There exists a mapping $g_{\mathbf{H}^1} : \Delta_{\mathcal{X}} \to \mathbf{H}^1$ such that for $r_1 \geq 40\lg(|\mathbf{H}^1|)$ and $S \in \mathcal{S}_{part1} := \{S : |\bar{S} \cap [u-r_1]| \geq r\}$, the mapping $g_{\mathbf{H}^1}$ and the hypothesis set $\mathbf{H}^1$ satisfy the following four properties with probability at least $1-\delta-2^{-0.01r_1}$ (over the outcome of $\mathbf{H}^1$):*
  1. *For any distribution $\mathcal{D} \in \mathcal{D}_S := \{\mathcal{D} : \mathcal{D}(i) > 0$ for $i \in S$ else $\mathcal{D}(i) = 0, \|\mathcal{D}\|_1 = 1\}$ supported on S, $\sum_{i\in S}\mathcal{D}(i)g_{\mathbf{H}^1}(\mathcal{D})(i) \geq \gamma/4$.*
  2. *Let $F_{r,S}$ denote the first r points from $\bar{S}\cap[u-r_1]$ and recall that $supp(\mathcal{D}) = S$. If for $\mathcal{D} \in \mathcal{D}_S$, $g_{\mathbf{H}^1}(\mathcal{D}) \neq h_0$, then the hypothesis $g_{\mathbf{H}^1}(\mathcal{D})$ has $(1/2+\alpha\gamma/2)r$ minus signs in $F_{r,S}$. Further, the outcome of $g_{\mathbf{H}^1}(\mathcal{D})$ on $F_{r,S}$ is uniformly distributed among all vectors in $\{-1,1\}^r$ which have at least $(1/2+\alpha\gamma/2)r$ minus signs.*
  3. *The randomness over $F_{r,S}$ in Item 2 is independent for all hypotheses in $\{g_{\mathbf{H}^1}(\mathcal{D})$ for $\mathcal{D} \in \Delta_{\mathcal{X}}\}$. Further, the outcome of $g_{\mathbf{H}^1}$ on $F_{r,S}$ is independent of $g_{\mathbf{H}^1}$ on $\overline{F_{r,S}}$.*
  4. *For any weight vector $w \in \Delta_{\mathbf{H}^1\setminus h_0} := \{w \in \mathbb{R}^{|\mathbf{H}^1|} : 0 \leq w_i, w_0 = 0, \sum_{i\in|\mathbf{H}^1|} w_i = 1\}$ weighing the hypotheses in $\mathbf{H}^1$, we have for at least $r_1/10$ of the i's in $\{u-r_1+1,\ldots,u\}$, that $\sum_{j\in|\mathbf{H}^1|} w_j h_j(i) \leq 14\sqrt{\lg(|\mathbf{H}^1|)/r_1}$.*
- *There exists a mapping $t_{\mathbf{H}^2} : \mathcal{D} \to \mathbf{H}^2$ such that with probability at least $1-\delta$ over $\mathbf{H}^2$, it holds for all $\mathcal{D} \in \Delta_{\mathcal{X}}$ that $\sum_{i\in[u]}\mathcal{D}(i)t_{\mathbf{H}^2}(\mathcal{D})(i) \geq \gamma/4$.*

Proof. Let $\mathbf{H}^1 = \cup_{i=1}^{k}\mathbf{H}_i$ and $\mathbf{H}^2 = \cup_{i=k+1}^{2k}\mathbf{H}_i$ for independent outcomes of $\mathbf{H}_i$ from Lemma 4.1. In the proof, we consider the three bullets of the lemma separately.

The first bullet, i.e. the bound on the size of $\mathbf{H}$ follows immediately from Lemma 4.1 and the bound on $|\mathbf{H}_i|$ of N/k, and the fact that we use $2k$ hypothesis sets $\mathbf{H}_i$ in $\mathbf{H}$. We thus end up with at most

$$2N = 4c_0^{-2}k\ln(k/\delta)\exp(8c_2\gamma^2 r_1)$$

random hypothesis in $\mathbf{H}$ adding h_0 gives the desired bound on $\mathbf{H}$'s size. Thus, what remains to be shown is the second and third bullet of Lemma 4.1.

Second bullet / Properties of the event E_S: We now show the second bullet, which intuitively states that $g_{\mathbf{H}^1}$ outputs hypothesès with a $\gamma/4$ advantage on S, many minus signs in $F_{r,S}$, and linear combinations of them on the last r_1 points can not all have large margins (the part where h_0 is -1).

Let the function $g_{\mathbf{H}^1}$ that searches for the first hypothesis in $\mathbf{H}_1, \ldots, \mathbf{H}_k$ which has a $\gamma/4$ advantage (i.e. fulfils Item 1 in Lemma 3.4) for a given distribution $\mathcal{D} \in \Delta_{\mathcal{X}}$ and additionally has at least $(1/2 + \alpha\gamma/2)r$ minus signs in the first r points of $\overline{supp(\mathcal{D})} \cap [u - r_1] = \bar{S} \cap [u - r_1]$, matching Algorithm 2. If there is no such hypothesis, $g_{\mathbf{H}^1}$ chooses the hypothesis h_0. Let further $S \in \mathcal{S}_{\text{part1}}$ and define E_S^1 to be the event (over the outcome $\mathcal{H}$ of $\mathbf{H}$) that

$$E_S^1 \coloneqq \Big\{ \mathcal{H} : \forall \mathcal{D} \in \mathcal{D}_S \, \exists h \in \mathcal{H} \text{ such that: } \sum_{i \in S} \mathcal{D}(i) h(i) \geq \gamma/4 \text{ and } h(F_{r,S}) \text{ has } (1/2 + \alpha\gamma/2)r \text{ minus signs} \\ \text{or } h_0 \in \mathcal{H} \text{ and } \sum_{i \in S} \mathcal{D}(i) h_0(i) \geq \gamma/4 \Big\}. \tag{20}$$

E_S^1 will be one part of E_S (E_S will be a union of two events) and used in arguing for Item 1, Item 2, and Item 3. We now argue that E_S^1 happens with probability at least $1 - \delta$ over $\mathbf{H}^1$. For this we run Algorithm 2 on input $S \in \mathcal{S}_{\text{part1}}$ and $\mathbf{H}_1, \ldots, \mathbf{H}_k$. Using Lemma 3.4, we show that a run of Algorithm 2 finishes on input S and $\mathbf{H}_1, \ldots, \mathbf{H}_k$ with probability at least $1 - \delta$ and that this implies that $\mathbf{H}^1$ is in the event E_S^1. To see this we show that whenever Algorithm 2 finishes, it produces an f such that $f(i) \geq \gamma/4$ for any $i \in S$ (large margin on S) and that the hypotheses that f is made of (when they are not h_0) have at least $(1/2 + \alpha\gamma/2)r$ minus signs in the first r points of $\bar{S} \cap [u - r_1]$. We then notice that $f(i) \geq \gamma/4$ for any $i \in S$ implies that for any $\mathcal{D} \in \mathcal{D}_S$ one of the hypotheses f is made of must have a $\gamma/4$ advantage on the all-ones label. This follows from $supp(\mathcal{D}) = S$ for $\mathcal{D} \in \mathcal{D}_S$, $\mathcal{D}$ being a probability distribution, $f = (1/k) \sum_{j=1}^{k} h_j$, and

$$\gamma/4 \leq \sum_{i \in S} D_S(i) f(i) = \sum_{j=1}^{k} 1/k \sum_{i \in S} D_S(i) h_j(i). \tag{21}$$

We therefore conclude that the event that Algorithm 2 finishes is contained in E_S^1. Thus if we can show that Algorithm 2 with input $S \in \mathcal{S}_{\text{part1}}$ and $\mathbf{H}_1, \ldots, \mathbf{H}_k$ finish with probability at least $1 - \delta$, then $\mathbf{H}^1$ is in E_S^1 with probability at least $1 - \delta$ over $\mathbf{H}^1$. We show that Algorithm 2 finishes with probability $1 - \delta$ in the end of this section and has the promised guarantees.

To handle Item 4, we define the event E^2 as

$$E^2 \coloneqq \left\{ \mathcal{H} : \forall w \in \Delta_{\mathcal{H} \setminus h_0} \text{ at least } r_1/10 \text{ } i\text{'s in } \{u - r_1 + 1, \ldots, u\} \text{ satisfies: } \sum_{j \in |\mathcal{H}|} w_j h_j(i) \leq 14 \sqrt{\lg\left(|\mathcal{H}|\right)/r_1} \right\}. \tag{22}$$

We show that $\mathbf{H}^1$ is in E^2 with probability at least $1 - 2^{-0.01 r_1}$ over $\mathbf{H}^1$. To see this, we form a matrix of all hypotheses created by $\mathbf{H}_1, \ldots, \mathbf{H}_k$ excluding h_0 (the hypotheses as columns). Now using $r_1 \geq 40 \lg(|\mathbf{H}^1|)$ by the assumption in the bullet of the lemma, Lemma 4.2 invoked on the lower $r_1 \times |\mathbf{H}^1|$ part of this matrix, gives us that $\mathbf{H}^1$ is in E^2 with probability at least $1 - 2^{-0.01 r_1}$. Now setting $E_S = E_S^1 \cap E^2$ and using a union bound we get that that $\mathbf{H}^1$ is in E_S with probability at least $1 - \delta - 2^{-0.01 r_1}$.

First notice that conditioned on E_S, we get by the E^2 part of E_S that Item 4 of the second bullet follows. From the definition of $g_{\mathbf{H}^1}$ choosing a hypothesis with $\gamma/4$ advantage with at least $(1/2 + \alpha\gamma/2)r$ minus signs in $F_{r,S}$ or else h_0 it follows from the E_S^1 part of E_S that Item 1 holds and the guarantee about at least $(1/2 + \alpha\gamma/2)r$ minus signs in $F_{r,S}$ of Item 2. Further, the part of Item 2 claiming that the minus signs in $F_{r,S}$ of $g_{\mathbf{H}^1}$ are uniformly distributed between any permutation in $\{-1, 1\}^r$ with at least $(1/2 + \alpha\gamma/2)r$ minus signs follows from the hypothesis in $\mathbf{H}^1 \setminus h_0$ being random vectors in $\{-1, 1\}^u$ with i.i.d. uniform entries, i.e. all outcomes of $\{-1, 1\}^r$ with at least $(1/2 + \alpha\gamma/2)r$ minus signs are equally likely. That the entries of $\mathbf{H}^1 \setminus h_0$ are i.i.d. and the constrains different from, $g_{\mathbf{H}^1}$ having at least

$(1/2 + \alpha\gamma/2)r$ minus signs in $F_{r,S}$, imposed in E_S^1 and E^2 only depend on points in $\overline{F_{r,S}}$ gives the claims of independence in Item 3 for $g_{\mathbf{H}^1}$ on $F_{r,S}$.

What is left to show is that Algorithm 2 with input $\mathbf{H}_1, \ldots, \mathbf{H}_k$ and S finishes with probability at least $1 - \delta$ and that on the event that Algorithm 2 finishes it produces an f such that $f(i) \geq \gamma/4$ for any $i \in S$ and that the hypotheses that f is made of (when they are not h_0) have at least $(1/2 + \alpha\gamma/2)r$ minus signs in the first r points of $F_{r,S} = \bar{S} \cap [u - r_1]$. By Lemma 4.1, Algorithm 2 with S and $\mathbf{H}_1, \ldots, \mathbf{H}_k$ as input finishes with probability at least $(1 - \delta/k)^k \geq 1 - \delta$, where we have used the independence of the hypothesis sets $\mathbf{H}_1, \ldots, \mathbf{H}_k$. The claim that the f produced when Algorithm 2 finishes consists of hypotheses (when they are not h_0) with at least $(1/2 + \alpha\gamma/2)r$ minus signs in $F_{r,S}$ follows from Line 6, Line 8, and Line 10 of Algorithm 2.

Thus, we still need to show that $f(i) \geq \gamma/4$ for all $i \in S$ when Algorithm 2 finishes. In this case, we know that the hypotheses $h_1, \ldots, h_k$ chosen by Algorithm 2 fulfill Line 6 and Line 8 in Algorithm 2 which ensures that hypothesis chosen in the i'th round h_i for the distribution in the i'th round D_i has a 2γ advantage. Let $\eta = \frac{1}{2}\ln\frac{1+2\gamma}{1-2\gamma}$ and $f_k = k \cdot f = \sum_{i=1}^k h_i$. We now follow a standard AdaBoost argument to show that $\exp(-\eta f_k(i)) \leq \exp\left(\ln(|S|) - 2k\gamma^2\right)$, for any $i \in S$ when Algorithm 2 finishes.

Showing $\exp(-\eta f_k(i)) \leq \exp\left(\ln(|S|) - 2k\gamma^2\right)$, for any $i \in S$ implies that $f(i) \geq \left(2k\gamma^2 - \ln(|S|)\right)/(k\eta)$ and since for $\gamma < 1/4$, it holds that

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

we get

$$f(i) \;\geq\; \frac{2k\gamma^2 - \ln(|S|)}{4k\gamma} \;\geq\; \frac{\gamma}{2} - \frac{\ln(|S|)}{4k\gamma}$$

and using that $k = \ln(u)\gamma^{-2}$ and $S \subseteq [u]$ it follows that $f(i) \geq \gamma/4$. Thus, if we show $\exp(-\eta f_k(i)) \leq \exp\left(\ln(|S|) - 2k\gamma^2\right)$ for all $i \in S$ we are done. Let Z_l be the normalization factor for the multiplicative weight update step in Algorithm 2. We now argue that $\exp(-\eta f_j(i)) = |S| D_{j+1}(i) \prod_{l\in[j]} Z_l$ for all $j \in [k]$ and $i \in [u]$ and that $\prod_{l\in[k]} Z_l \leq (1 - 2\gamma^2)^k$. Showing these two relations implies that

$$\exp(-\eta f_k(i)) \;\leq\; |S| \prod_{l\in[k]} Z_l \;\leq\; |S|(1-2\gamma^2)^k \;\leq\; \exp\left(\ln(|S|) - 2k\gamma^2\right) \tag{23}$$

where the first inequality uses $D_{k+1} \leq 1$ and the last inequality follows from $\lg(1 + x) \leq x$ for $x > -1$.

We show that $\exp(-\eta f_j(i)) = |S| D_{j+1}(i) \prod_{l\in[j]} Z_l$ for all $j \in [k]$ and $i \in [u]$ by induction. For the induction base $j = 1$ we have $\exp(-\eta f_1(i)) = \exp(-\eta h_1(i))$ and $|S| D_2(i) Z_1 = |S| D_1(i) \exp(-\eta h_1) = \exp(-\eta h_1)$, where we have used that $D_2(i) = D_1(i)\exp(-\eta h_1(i))/Z_1$ and $D_1(i) = 1/|S|$. For the induction step we have

$$\exp(-\eta f_{j+1}(i)) = \exp(-\eta(f_j(i) + h_{j+1}(i))) = |S| D_{j+1}(i) \prod_{l\in[j]} Z_l \exp(-\eta h_{j+1}(i)) = |S| D_{j+2}(i) \prod_{l\in[j+1]} Z_l$$

where the second equality follows from the induction hypothesis for j and the last by $D_{j+2}(i) = D_{j+1}(i)\exp(\eta h_{j+1}(i))/Z_{j+1}$ (see Algorithm 2).

To show $\prod_{l\in[k]} Z_l \leq (1 - 2\gamma^2)^k$, i.e. the second inequality in Equation (23), we show $Z_l \leq (1 - 2\gamma^2)$ for

$l = 1, \ldots, k$. Using that $\exp(\eta) = \left(\frac{1+2\gamma}{1-2\gamma}\right)^{1/2}$ we notice that

$$\begin{aligned}
Z_l &= \sum_{i \in S} D_l(i) \exp\left(-\eta h_l\left(i\right)\right) \\
&= \sum_{\substack{i \in S: \\ h_l(i)=1}} D_l(i) \exp\left(-\eta\right) + \sum_{\substack{i \in S: \\ h_l(i)=-1}} D_l(i) \exp\left(\eta\right) \\
&= \sum_{\substack{i \in S: \\ h_l(i)=1}} D_l(i) \sqrt{\frac{1-2\gamma}{1+2\gamma}} + \left(1 - \sum_{\substack{i \in S: \\ h_l(i)=1}} D_l(i)\right) \sqrt{\frac{1+2\gamma}{1-2\gamma}} \\
&= \left(\sum_{\substack{i \in S: \\ h_l(i)=1}} D_l(i) \frac{1}{1+2\gamma} + \left(1 - \sum_{\substack{i \in S: \\ h_l(i)=1}} D_l(i)\right) \frac{1}{1-2\gamma}\right) \sqrt{(1+2\gamma)\left(1-2\gamma\right)}. \qquad (24)
\end{aligned}$$

Using that we noticed that Line 6, Line 8, and Line 10 in Algorithm 2 together with Algorithm 2 finishing implied $\sum_{i \in S} D_j(i) h_j(i) \geq 2\gamma$ for any $j \in k$ we get that

$$\sum_{\substack{i \in S \\ h_l(i)=1}} D_l(i) = \sum_{i \in S} D_l(i) \frac{1 + h_l(i)}{2} \geq 1/2 + \gamma,$$

and using this together with $\frac{x}{1+2\gamma} + \frac{1-x}{1-2\gamma}$ being decreasing we get

$$\left(\sum_{\substack{i \in S \\ h_l(i)=1}} D_l(i) \frac{1}{1+2\gamma} + \left(1 - \sum_{\substack{i \in S \\ h_l(i)=1}} D_l(i)\right) \frac{1}{1-2\gamma}\right) \leq 1.$$

Further using that $(1-2x)(1+2x) = 1-4x^2 \leq (1-2x^2)^2$ we conclude by Equation (24) that $Z_l \leq (1-2\gamma^2)$ as claimed.

Third bullet / Properties of $t_{\mathbf{H}^2}$: Let $t_{\mathbf{H}^2}$ be such that given a $\mathcal{D} \in \Delta_{\mathcal{X}}$ it returns the first hypothesis in $\mathbf{H}^2$ that has a $\gamma/4$ advantage on $\mathcal{D}$ otherwise report fail. Note that $t_{\mathbf{H}^2}$ does not include any adversarial behavior, it is a simple and straightforward γ-weak learner. We now show with probability at least $1-\delta$ over $\mathbf{H}^2$ that $t_{\mathbf{H}^2}$ succeeds simultaneously for all $\mathcal{D} \in \Delta_{\mathcal{X}}$. Here, we use a slightly different argument compared to the case for $g_{\mathbf{H}^1}$ above and run Algorithm 2 in a slightly modified version. The slight modification is that in Line 8 we have no constraints on the number of minus signs in the first r positions of $\bar{S} \cap [u - r_1]$ and that we run the algorithm with the input $\mathcal{X}$ and $\mathbf{H}_{k+1}, \ldots, \mathbf{H}_{2k}$ (instead of $\mathbf{H}_1, \ldots, \mathbf{H}_k$). We then show that this variant of Algorithm 2 succeeds with probability at least $1-\delta$ and that the produced f satisfies $f(i) \geq \gamma/4$ for all $i \in [u]$. By the same argument as above for Equation (21), it follows that $f(i) \geq \gamma/4$ for all $i \in u$ implies that for any $\mathcal{D}$ there exist an $\mathbf{h} \in \mathbf{H}_{k+1}, \ldots, \mathbf{H}_{2k}$ with a $\gamma/4$ advantage on $\mathcal{D}$. Thus, the event that this slightly modified version of Algorithm 2 succeeds on $\mathcal{X}$ and $\mathbf{H}_{k+1}, \ldots, \mathbf{H}_{2k}$ is contained in the event

$$\left\{\mathcal{H} : \forall \mathcal{D} \in \Delta_{\mathcal{X}}\ \exists h \in \mathcal{H} \text{ such that: } \sum_{i \in [u]} \mathcal{D}(i) h(i) \geq \gamma/4\right\}.$$

Hence, with probability at least $1-\delta$ for any $\mathcal{D} \in \Delta_{\mathcal{X}}$, $t_{\mathbf{H}^2}$ finds a hypothesis in $\mathbf{H}^2$ with $\gamma/4$ advantage (choosing the first it finds) and outputs this as the weak learner for the distribution $\mathcal{D}$.

The claim that Algorithm 2 with $\mathcal{X}$ and $\mathbf{H}_{k+1}, \ldots, \mathbf{H}_{2k}$ succeeds with probability at least $1 - \delta$ over $\mathbf{H}^2$ follows as in the $g_{\mathbf{H}^1}$-case from Lemma 4.1 and $\mathbf{H}_{k+1}, \ldots, \mathbf{H}_{2k}$ being independent.

We now notice that when we argued that the non-modified version of Algorithm 2 finishing would produce an f such that $f(i) \geq \gamma/4$ for $i \in S$, we never used the constraint on the minus signs, and only that $|S| \leq u$. Thus, reusing the above arguments but now for the modified version of Algorithm 2 finishing, with $S = \mathcal{X}$, again yields that the produced f satisfies $f(i) \geq \gamma/4$ for $i \in \mathcal{X}$, which concludes the proof of Lemma 3.4.

□

Having established the proof of Lemma 3.4 using Lemma 4.2 and Lemma 4.1 we now move on to the proof of those. We start by restating and giving the proof of Lemma 4.2.

Lemma 4.2. *Let* $\mathbf{A}$ *be uniform random in* $\{-1, 1\}^{r \times n}$ *and assume that* $r \geq 40 \lg(n)$. *With probability at least* $1 - 2^{-0.01r}$, *it holds for all* $w \in \mathbb{R}^n$ *with* $\|w\|_1 = 1$ *that* $\mathbf{A}w$ *has at least* $r/10$ *entries* i *with* $(\mathbf{A}w)_i < 14\sqrt{\lg(n)/r}$.

Proof. The following proof proceeds by bounding the probability of the complementary event of the above, i.e. we will show that the probability of there existing a $w \in \mathbf{R}^n$, $\|w\| = 1$ such that $\mathbf{A}w$ has strictly less than $r/10$ entries such that $(\mathbf{A}w)_i < 14\sqrt{\lg(n)/r}$ happens with probability at most $2^{-0.01r}$. For this we first discretize the set of all unit vectors, call this set $\mathcal{W}$. We then show that if there exists a unit vector with the above property, then there exists a vector $\tilde{w}$ in $\mathcal{W}$ such that $\mathbf{A}\tilde{w}$ has at least $(13/20)r$ strictly positive entries. Now using that $\mathbf{A}$ has i.i.d. uniform $\{-1, 1\}$-random variables as entries, $(\mathbf{A}\tilde{w})_i$ is strictly positive with a probability at most $1/2$, i.e. in expectation we see at most $(1/2)r$ strictly positive entries. The result then follows by applying Hoeffding's inequality and union bounding over $\mathcal{W}$.

Consider the set $\mathcal{W}$ containing all w whose coordinates w_i are of the form $j_i 40 \lg(n)/r$ for integers $j_i \in \{-r/(40 \lg n), \ldots, r/(40 \lg n)\}$ and $\|w\|_1 = 1$. We now want to bound $|\mathcal{W}|$. For this, consider throwing $r/(40 \lg(n))$ balls with a sign and absolute value $40 \lg(n)/r$ into n buckets. There are $(2n)^{r/(40 \lg n)} \leq 2^{r/20}$ outcomes of this experiment. We now map each $w \in \mathcal{W}$ to an outcome of the above experiment. For this, notice that $\sum_{i=1}^n j_i = r/(40 \lg(n))$ since $w \in \mathcal{W}$ has unit length. Now for a $w \in \mathcal{W}$ consider any outcome of the experiment where for $i = 1, \ldots, n$: j_i balls fell into the i'th bucket, and all the balls signs coincide with $\text{sign}(w_i)$. In this case the value of the i'th bucket is the same value as w_i. Thus, we conclude that $|\mathcal{W}| \leq 2^{r/20}$.

Now consider an outcome A of the random matrix $\mathbf{A}$ and assume there exists $w \in \mathbb{R}^n$ with $\|w\|_1 = 1$ such that Aw has strictly less than $r/10$ entries i with $(Aw)_i < 14\sqrt{\lg(n)/r}$. We now show that this implies that there exists a vector $\tilde{w} \in \mathcal{W}$ such that $A\tilde{w}$ has at least $(13/20)r$ strictly positive entries. For $t = 1, \ldots, r/(40 \lg n)$ sample independently an index $\mathbf{j}(\mathbf{t})$ from w such that the i'th index is sampled with probability $|w_i|/\|w\|_1$. Let $\tilde{\mathbf{w}}$ be the vector whose i'th coordinate is $\mathbf{j_i} \,\text{sign}(w_i) 40 \lg(n)/r$. Here $\mathbf{j_i}$ denotes the number of times index i was sampled.

Consider any coordinate $(A\tilde{\mathbf{w}})_i$. Using i.i.d. random variables X_t taking the value $a_{i,\mathbf{j}(\mathbf{t})} \,\text{sign}(w_{\mathbf{j}(\mathbf{t})}) 40 \lg(n)/r$, we can write $(A\tilde{\mathbf{w}})_i$ as $\sum_{t=1}^{r/(40 \lg n)} X_t$. Note that $\mathbb{E}[X_t] = \sum_{i=1}^n a_{i,j} w_i 40 \lg(n)/r = (Aw)_i 40 \lg(n)/r$. Thus, we see that $\mathbb{E}[(A\tilde{\mathbf{w}})_i] = (r/(40 \lg n)) \,\mathbb{E}[X_1] = (Aw)_i$. Notice that since X_t takes values in $\{-40 \lg(n)/r, 40 \lg(n)/r\}$, its variance is at most $(40 \lg(n)/r)^2$. Further, by the independence of the X_t's, we have that $(A\tilde{\mathbf{w}})_i$ has variance at most $(r/(40 \lg n))(40 \lg(n)/r)^2 = 40 \lg(n)/r$. Thus, Chebyshev's inequality implies that $\Pr[\,|(A\tilde{\mathbf{w}})_i - (Aw)_i| > 2\sqrt{40 \lg(n)/r}] \leq 1/4$. Now noticing that $\tilde{\mathbf{w}} \in \mathcal{W}$ and using the linearity of expectation, we conclude that there must be some vector $\tilde{w} \in \mathcal{W}$ for which there are less than $r/4$ entries i such that $|(A\tilde{\mathbf{w}})_i - (Aw)_i| > 2\sqrt{40 \lg(n)/r}$. This, combined with the assumption of $(Aw)_i < 14\sqrt{\lg(n)/r}$ for strictly less than $r/10$ entries, implies that $A\tilde{w}$ has at least $r - r/10 - r/4 = (13/20)r$ entries i such that $(Aw)_i \geq 14\sqrt{\lg(n)/r}$ and $|(A\tilde{w})_i - (Aw)_i| > 2\sqrt{40 \lg(n)/r}$. Thus, we conclude that at least $(13/20)r$ entries i satisfy $(A\tilde{w}_i) \geq 14\sqrt{\lg(n)/r} - 2\sqrt{40 \lg(n)/r} > 0$, i.e. if there exists $w \in \mathbb{R}^n$ with $\|w\|_1 = 1$ such that $\mathbf{A}w$ has strictly less than $r/10$ entries i then there also exists $\tilde{w} \in \mathcal{W}$ such that $\mathbf{A}\tilde{w}$ has at least $(13/20)r$ entries that are strictly positive.

Thus, what remains is to argue that $\mathcal{W}$ with small probability over $\mathbf{A}$ contains a vector w with at least $(13/20)r$ entries i such that $(\mathbf{A}w)_i > 0$. For this, consider any fixed $w \in \mathcal{W}$. The probability that

$(\mathbf{A}w)_i > 0$ is at most $1/2$ for all i. Now Hoeffding's inequality implies that the probability that there are $(13/20)r$ entries i with $(\mathbf{A}w)_i > 0$ is no more than $\exp(-2((3/10)r)^2/(4r)) = \exp(-(9/200)r)$. A union bound over all of $\mathcal{W}$ (recall $|\mathcal{W}| \leq 2^{r/20}$) shows that the probability that there exists a vector $w \in \mathcal{W}$ which has at least $(13/20)r$ strictly positive entries is at most $e^{-(9/200)r}2^{r/20} < 2^{-0.01r}$ over $\mathbf{A}$. Thus, we conclude that the probability of existence of a $w \in \mathbb{R}^n$ with $\|w\|_1 = 1$ such that $\mathbf{A}w$ has strictly less than $r/10$ entries i with $(\mathbf{A}w)_i < 14\sqrt{\lg(n)/r}$ is at most $2^{-0.01r}$ which concludes the proof. □

To show Lemma 4.1 we need the following corollary which follows from a use of the Montgomery-Smith inequality [11]. The corollary says that a linear combination of i.i.d. uniform $\{-1, 1\}$-variables where the coefficient's absolute values sums to at least $1/2 - \beta/2$ with some probability are greater than β. This will be used in Lemma 4.1 to say that $\mathbf{H}_i$ for a given $\mathcal{D} \in \Delta_{\mathcal{X}}$ contains a hypothesis $\mathbf{h}$ with an advantage of 2γ.

Corollary 4.3. *There exist universal constants $\tilde{c}_1, \tilde{c}_2 \leq 1$, and $\tilde{c}_3 \geq 1$ such that for $\beta \leq \tilde{c}_1/6$, $x \in \mathbb{R}^n$, $x_i \geq 0\ \forall i \in [n]$, and $\sum_{i=1}^n x_i \geq (1-\beta)/2$, we have for a random $\mathbf{h} \in \{-1, 1\}^n$ with i.i.d. uniform entries that*

$$\mathbb{P}\left[\sum_{i=1}^{n} \mathbf{h}(i)x_i \geq \beta\right] \geq \tilde{c}_2 \exp\left(-\tilde{c}_3 \frac{16\beta^2 n}{\tilde{c}_1^2}\right)$$

We will show Corollary 4.3 after the proof of Lemma 4.1. We now restate and give the proof of Lemma 4.1

Lemma 4.1. *Let $c_0, c_1 \leq 1$, and $c_2 \geq 1$ denote some universal constants. Let $\mathcal{X}$ be a universe of size u and $\mathcal{D} \in \Delta_{\mathcal{X}}$ a distribution over $\mathcal{X}$. Further let r and r_1 be non-negative numbers such that $r_1 = \alpha^2 r$ for $\alpha \geq 1$ and $r_1 \leq u$. Let $0 < \delta \leq 1$, $\gamma \leq c_0/(2\alpha)$, and $k = \ln(u)\gamma^{-2}$. Let $\mathbf{H}_i$ be a random hypothesis set consisting of h_0 and independent random vectors in $\{-1, 1\}^u$ with i.i.d. uniform random entries. Further let the size of $\mathbf{H}_i$ be N/k without counting h_0, where $N = 2c_1^{-2}k\ln(k/\delta)\exp(8c_2\gamma^2 r_1)$. With the above, we have with probability at least $1 - \delta/k$ over $\mathbf{H}_i$ that:*

1. *There exists a hypothesis $\mathbf{h} \in \mathbf{H}_i$ such that*

$$\sum_{i \in supp(D)} D_i \mathbf{h}(i) \geq 2\gamma$$

where $\mathbf{h} = h_0$ if $\sum_{i=1, i\in supp(D)}^{u-r_1} D_i > 1/2 + \gamma$ else $\mathbf{h}$ is random.

Further, if $\sum_{i=1, i\in supp(D)}^{u-r_1} D_i \leq 1/2 + \gamma$ and $r \leq |\overline{supp(\mathcal{D})} \cap [u - r_1]|$

2. *$\mathbf{h}$ in Item 1 is such that the first r entries of $\{\mathbf{h}(i)\}_{i \in \overline{supp(\mathcal{D})} \cap [u-r_1]}$ has at least $(1/2 + \alpha\gamma/2)r$ minus signs.*

Proof. If the distribution $\mathcal{D}$ has more than $1/2+\gamma$ mass on the points $1, \ldots, u-r_1$, i.e. $\sum_{i=1, i\in supp(D)}^{u-r_1} D_i > 1/2 + \gamma$, we have $\sum_{i=u-r_1+1, i\in supp(D)}^{u} D_i < 1/2 - \gamma$. Thus, we notice that h_0 satisfies

$$\sum_{i \in supp(D)} D_i h_0(i) = \sum_{\substack{i=1 \\ i \in supp(D)}}^{u-r_1} D_i - \sum_{\substack{i=u-r_1+1 \\ i \in supp(D)}}^{u} D_i \geq 2\gamma,$$

i.e. h_0 fulfills Item 1.

Now assume that $\sum_{i=1, i\in supp(D)}^{u-r_1} D_i \leq 1/2 + \gamma$. Then we have $1/2 - \gamma$ mass on the points $\{u - r_1 + 1, u\} \cap supp(D)$, i.e. $\sum_{i=u-r_1+1, i\in supp(D)}^{u} D_i \geq 1/2 - \gamma$. Since we know that the entries of any $\mathbf{h}$ in $\mathbf{H}_i$ for $\mathbf{h} \neq h_0$ are i.i.d. uniform $\{-1, 1\}$-variables, we get that $\sum_{i=1, i\in supp(D)}^{u-r_1} \mathbf{h}(i)D_i \geq 0$ with probability $1/2$. Thus, we give a lower bound on the probability of $\sum_{i=u-r_1+1, i\in supp(D)}^{u} \mathbf{h}(i)D_i \geq 2\gamma$. Using that

$\sum_{i=u-r_1+1, i\in supp(D)}^{u} D_i \geq 1/2-\gamma$ (by the assumption in this paragraph), Corollary 4.3 implies that for $2\gamma \leq c_0$

$$\mathbb{P}\left[\sum_{i=u-r_1+1, i\in supp(D)}^{u} D_i\mathbf{h}(i) \geq 2\gamma\right] \geq c_1 \exp(-4c_2\gamma^2 r_1)$$

so we conclude by the independence of the entries in $\mathbf{h}(i)$ that

$$\mathbb{P}\left[\sum_{i\in supp(D)} D_i\mathbf{h}(i) \geq 2\gamma\right] \geq \mathbb{P}\left[\sum_{\substack{i=1\\ i\in supp(D)}}^{u-r_1} D_i\mathbf{h}(i) \geq 0, \sum_{\substack{i=u-r_1+1\\ i\in supp(D)}}^{u} D_i\mathbf{h}(i) \geq 2\gamma\right] \geq c_1\exp(-4c_2\gamma^2 r_1)/2, \quad (25)$$

where c_0, c_1, and c_2 are universal constants (some of them are the product of universal constants in Corollary 4.3). Thus, Item 1 holds for every $\mathbf{h}$ in $\mathbf{H}_i\backslash h_0$ with at least the above probability. Now if $r \leq |\overline{supp(D)} \cap [u-r_1]|$ let F_r be the first r indices of $\overline{supp(D)} \cap \{1,\ldots,u-r_1\}$. Note that F_r has the same role as $F_{r,S}$ in other parts of the paper, but in this lemma we make no assumptions about the support of $\mathcal{D}$. Then by Corollary 4.3, we get that for $\alpha\gamma \leq c_0$

$$\mathbb{P}\left[\sum_{i\in F_r} \mathbf{h}(i)/r \leq -\alpha\gamma\right] = \mathbb{P}\left[\sum_{i\in F_r} \mathbf{h}(i)/r \geq \alpha\gamma\right] \geq c_1\exp(-c_2(\alpha\gamma)^2 r) \geq c_1\exp(-4c_2\gamma^2 r_1), \quad (26)$$

where the equality is due to the $\mathbf{h}(i)$ being i.d.d. uniform $\{-1,1\}$-variables and the last inequality follows from $r \leq r_1$. If we have $\sum_{i\in F_r}\mathbf{h}(i)/r \leq -\alpha\gamma$ then $\{\mathbf{h}(i)\}_{i\in F_r}$ must contain at least $(1/2+\alpha\gamma/2)r$ minus ones. Thus, we conclude by Equation (25) and Equation (26), and the independence of the entries of $\mathbf{h}$ that

$$\mathbb{P}\left[\sum_{i\in supp(D)} \mathbf{h}(i)D_i \geq 2\gamma,\ |\{i\in F_r \mid \mathbf{h}(i)=-1\}| \geq (1/2+\alpha\gamma/2)r\right] \geq c_1^2\exp(-8c_2\gamma^2 r_1)/2.$$

By the definition of $N = 2c_1^{-2}k\ln(k/\delta)\exp(8c_2\gamma^2 r_1)$ we get that we have that

$$c_1^2\exp(-8c_2\gamma^2 r_1)/2 = \frac{k\ln(k/\delta)}{N}.$$

Now define $f(h) = \mathbb{1}_{\{\sum_{i\in supp(D)}\mathbf{h}(i)D_i \geq 2\gamma, |\{i\in F_r|\mathbf{h}(i)=-1\}| \geq (1/2+\alpha\gamma/2)r\}}$. Using f, independence of the $\mathbf{h}$'s in $\mathbf{H}_i$, and that the size of $\mathbf{H}_i$ is N/k we get that

$$\begin{aligned}
\Pr[\exists \mathbf{h}\in\mathbf{H}_i \text{ s.t. } f(\mathbf{h})=1] &= 1-\Pr[\forall \mathbf{h}\in\mathbf{H}_i \text{ we have } f(\mathbf{h})=0]\\
&= 1-\Pr[f(\mathbf{h})=0]^{N/k}\\
&= 1-(1-\Pr[f(\mathbf{h})=1])^{N/k}\\
&\geq 1-\left(1-\frac{k\ln(k/\delta)}{N}\right)^{N/k}\\
&\geq 1-\exp(-\ln(k/\delta))\\
&= 1-\delta/k
\end{aligned}$$

where the last inequality follows from $(1+x/n)^n = \exp(n\ln(1+x/n)) \leq \exp(x)$ for $n\geq 1$ and $x\geq -1$, since $\ln(1+x)\leq x$ for $x\geq -1$. This shows Item 1 in the case $\sum_{i=1,i\in supp(D)}^{u-r_1} D_i \leq 1/2+\gamma$ and Item 2 if $r \leq |\overline{supp(\mathcal{D})}\cap[u-r_1]|$ which finishes the proof of Lemma 4.1. □

We now prove and restate Corollary 4.3.

Corollary 4.3. *There exist universal constants $\tilde{c}_1, \tilde{c}_2 \leq 1$, and $\tilde{c}_3 \geq 1$ such that for $\beta \leq \tilde{c}_1/6$, $x \in \mathbb{R}^n$, $x_i \geq 0\ \forall i \in [n]$, and $\sum_{i=1}^n x_i \geq (1-\beta)/2$, we have for a random $\mathbf{h} \in \{-1,1\}^n$ with i.i.d. uniform entries that*

$$\mathbb{P}\left[\sum_{i=1}^n \mathbf{h}(i)x_i \geq \beta\right] \geq \tilde{c}_2 \exp\left(-\tilde{c}_3 \frac{16\beta^2 n}{\tilde{c}_1^2}\right)$$

Proof. In the following we will assume that the x_i's are ordered by their absolute value, which we can assume without loss of generality since the $\mathbf{h}(i)$'s are i.d.d. uniform $\{-1,1\}$-variables. By [11] there exist universal constants $\tilde{c}_1$, $\tilde{c}_2$, and $\tilde{c}_3$ such that

$$f(x,t) := \sum_{i=1}^{\min(\lceil t^2 \rceil, n)} x_i + t\sqrt{\sum_{i=\lceil t^2 \rceil+1}^n x_i^2}, \tag{27}$$

and

$$\mathbb{P}\left[\sum_{i=1}^n \mathbf{h}(i)x_i \geq \tilde{c}_1 f(x,t)\right] \geq \tilde{c}_2 \exp(-\tilde{c}_3 t^2). \tag{28}$$

Notice that we may assume that $\tilde{c}_1 < 1$. If $\tilde{c}_1$ was greater than 1, we could lower it to 1 and the claim in Equation (28) would still hold. Similarly, we also assume $\tilde{c}_2 \leq 1$ and $\tilde{c}_3 \geq 1$.

Now consider $t = \frac{4\beta\sqrt{n}}{\tilde{c}_1}$ which implies that $t^2 \leq n/2$ since $\beta \leq \tilde{c}_1/6$. Thus the first sum of Equation (27) goes up to $\lceil t^2 \rceil$. Formally, if $\tilde{c}_1 f(x,t) \geq \tilde{c}_1 \sum_{i=1}^{\lceil t^2 \rceil} x_i \geq \beta$ we get by Equation (27) and Equation (28) that

$$\mathbb{P}\left[\sum_i^n \mathbf{h}(i)x_i \geq \beta\right] \geq \mathbb{P}\left[\sum_{i=1}^n \mathbf{h}(i)x_i \geq \tilde{c}_1 f(x,t)\right] \geq \tilde{c}_2 \exp(-\tilde{c}_3 t^2) = \tilde{c}_2 \exp\left(-\tilde{c}_3 \frac{16\beta^2 n}{\tilde{c}_1^2}\right).$$

For the other case, assume that $\tilde{c}_1 \sum_{i=1}^{\lceil t^2 \rceil} x_i \ \leq\ \beta$, which combined with $\sum_{i=1}^n x_i \geq 1/2 - \beta/2$ implies that

$$\tilde{c}_1 \sum_{i=\lceil t^2 \rceil+1}^n x_i = \tilde{c}_1 \left(\sum_{i=1}^n x_i - \sum_{i=1}^{\lceil t^2 \rceil} x_i\right) \geq \tilde{c}_1(1 - \beta - 2\beta/\tilde{c}_1)/2.$$

By Cauchy-Schwarz (in the second inequality below) and $\lceil t^2 \rceil \leq n$ we get that

$$\tilde{c}_1(1-\beta-2\beta/\tilde{c}_1)/2 \leq \tilde{c}_1 \sum_{i=\lceil t^2 \rceil+1}^n 1 \cdot x_i \leq \tilde{c}_1 \sqrt{|n - \lceil t^2 \rceil| \sum_{i=\lceil t^2 \rceil+1}^n x_i^2} \leq \tilde{c}_1 \sqrt{n \sum_{i=\lceil t^2 \rceil+1}^n x_i^2}.$$

$$\Rightarrow (1-\beta-2\beta/\tilde{c}_1)/2 \leq \sqrt{n \sum_{i=\lceil t^2 \rceil+1}^n x_i^2}. \tag{29}$$

We notice that $\beta \leq \tilde{c}_1/6$ implies $(1-\beta-2\beta/\tilde{c}_1) \geq 1/2$. From Equation (27) we get with Equation (29), $t = \frac{4\beta\sqrt{n}}{\tilde{c}_1}$, and $(1-\beta-2\beta/\tilde{c}_1) \geq 1/2$ that

$$\tilde{c}_1 f(x,t) \geq \tilde{c}_1 t \sqrt{\sum_{i=\lceil t^2 \rceil+1}^n x_i^2} = 4\beta \sqrt{n \sum_{i=\lceil t^2 \rceil+1}^n x_i^2} \geq 4\beta \frac{(1-\beta-2\beta/\tilde{c}_1)}{2} \geq \beta$$

Now using this and Equation (28) we get that

$$\mathbb{P}\left[\sum_i^n \mathbf{h}(i)x_i \geq \beta\right] \geq \mathbb{P}\left[\sum_{i=1}^n \mathbf{h}(i)x_i \geq \tilde{c}_1 f(x,t)\right] \geq \tilde{c}_2 \exp\left(-\tilde{c}_3 t^2\right) = \tilde{c}_2 \exp\left(-\tilde{c}_3 \frac{16\beta^2 n}{\tilde{c}_1^2}\right)$$

as in the other case which finishes the proof. □

We now have shown Lemma 3.4 and the two lemmas Lemma 4.2 and Lemma 4.1 that are used in the lemma. This leaves us to prove Lemma 3.2 and Lemma 3.3 which both appear in the proof of the main theorem. We start by restating Lemma 3.2.

Lemma 3.2. *Let $w \in \mathbb{R}^d$ such that $\|w\|_1 = 1$ and let $\tilde{\alpha} \geq 1$. Let further $\mathbf{h}$ be a random vector in $\{-1,1\}^d$ with i.i.d. entries such that $\mathbb{P}[\mathbf{h}(i) = 1] = 1/2 - \tilde{\alpha}\beta$ and $\mathbb{P}[\mathbf{h}(i) = -1] = 1/2 + \tilde{\alpha}\beta$ where $\beta < 1/(2\tilde{\alpha})$. We then have for $\alpha' < \tilde{\alpha}$ that*

$$\mathbb{P}\left[\sum_{i=1}^{d} w_i \mathbf{h}(i) \leq -\alpha'\beta\right] \geq \min\left(\frac{1}{4}, \frac{1}{2} - \frac{4\tilde{\alpha}\alpha'}{(2\tilde{\alpha} - \alpha')^2}\right).$$

Proof. First, if there is $j \in \{1, \ldots, d\}$ such that $w_j \geq \alpha'\beta$ (i.e. there is a hypothesis h_j with a large weight in the output of Algorithm 2) we get that

$$\mathbb{P}\left[\sum_{i=1}^{d} w_i \mathbf{h}(i) \leq -\alpha'\beta\right] \geq \mathbb{P}\left[\sum_{\substack{i=1 \\ i \neq j}}^{d} w_i \mathbf{h}(i) \leq 0, w_j r_j \leq -\alpha'\beta\right] \geq 1/4$$

which follows from the $\mathbf{h}(i)$'s being biased towards minus so if we changed them to i.i.d. uniform $\{-1,1\}$-variables the above probability would be lower and equal to $1/4$.

Thus, we may assume that $\|w\|_\infty \leq \alpha'\beta$, i.e. the largest entry in w is less than $\alpha'\beta$. We now introduce the random variables η_i and $\tilde{\mathbf{h}}(i)$ where $\tilde{\mathbf{h}}(i)$ are i.i.d. uniform $\{-1,1\}$-variables and the η_i's have the distribution $\mathbb{P}\left[\eta_i = 1 | \tilde{\mathbf{h}}(i) = -1\right] = 1$, $\mathbb{P}\left[\eta_i = -1 | \tilde{\mathbf{h}}(i) = 1\right] = 2\tilde{\alpha}\beta$ and $\mathbb{P}\left[\eta_i = 1 | \tilde{\mathbf{h}}(i) = 1\right] = 1 - 2\tilde{\alpha}\beta$. We immediately get

$$\begin{aligned}
\mathbb{P}\left[\eta_i \tilde{\mathbf{h}}(i) = -1\right] &= 1/2 + 1/2(2\tilde{\alpha}\beta) = 1/2 + \tilde{\alpha}\beta \qquad \text{and} \\
\mathbb{P}\left[\eta_i \tilde{\mathbf{h}}(i) = 1\right] &= 1/2(1 - (2\tilde{\alpha}\beta)) = 1/2 - \tilde{\alpha}\beta
\end{aligned}$$

thus $\eta_i \tilde{\mathbf{h}}(i)$ has the same distribution as $\mathbf{h}(i)$. Using this decomposition of the $\mathbf{h}(i)$'s we get that

$$\begin{aligned}
&\mathbb{P}\left[\sum_{i=1}^{d} w_i \mathbf{h}(i) \leq -\alpha'\beta\right] \\
&= \mathbb{P}\left[\sum_{i=1}^{d} w_i \eta_i \tilde{\mathbf{h}}(i) \leq -\alpha'\beta\right] \\
&= \mathbb{P}\left[\sum_{i=1}^{d} w_i \tilde{\mathbf{h}}(i) + \sum_{i=1}^{d} w_i (\eta_i - 1)\tilde{\mathbf{h}}(i) \leq -\alpha'\beta\right] \\
&\geq \mathbb{P}\left[\sum_{i=1}^{d} w_i \tilde{\mathbf{h}}(i) \leq 0, \sum_{i=1}^{d} w_i (\eta_i - 1)\tilde{\mathbf{h}}(i) \leq -\alpha'\beta\right] \\
&\geq 1 - \frac{1}{2} - \mathbb{P}\left[\sum_{i=1}^{d} w_i (\eta_i - 1)\tilde{\mathbf{h}}(i) > -\alpha'\beta\right] \qquad (30)
\end{aligned}$$

where the last inequality follows from $\mathbb{P}[A \cap B] \geq 1 - \mathbb{P}[A] - \mathbb{P}[B]$ and the $1/2$-term by a weighted sum of i.i.d. uniform $\{-1,1\}$-variables being symmetric around 0. We now notice that $(\eta_i - 1)\tilde{\mathbf{h}}(i)$ has the same distribution as a random variable $-2x_i$ where x_i follows $\mathbb{P}[x_i = 0] = 1 - \tilde{\alpha}\beta$ and $\mathbb{P}[x_i = 1] = \tilde{\alpha}\beta$. We also see that $\mathbb{E}\left[\sum_{i=1}^{n} -2w_i x_i\right] = -2\tilde{\alpha}\beta$ and by independence of the x_i's

$$\mathrm{Var}\left(\sum_{i=1}^{n} -2w_i x_i\right) = 4\sum_{i=1}^{d} w_i^2 \left(\mathbb{E}\left[x_i^2\right] - \mathbb{E}\left[x_i\right]^2\right) \leq 4\sum_{i=1}^{d} (\alpha'\beta\, w_i)\left(\tilde{\alpha}\beta - (\tilde{\alpha}\beta)^2\right) = 4\tilde{\alpha}\alpha'\beta^2(1 - \tilde{\alpha}\beta)$$

where the inequality follows from $\|w\|_\infty \leq \alpha'\beta$ and the last equality uses $\sum_{i=1}^d w_i = 1$. Using that the $\tilde{h}(i)$'s follow the same distribution as $-2x_i$ we get from Chebyshev's inequality, the above calculation of the expected value of $\sum_{i=1}^n -2w_i x_i$, and the upper bounds on its variance that

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^d w_i(\eta_i - 1)\tilde{\mathbf{h}}(i) > -\alpha'\beta\right] &= \mathbb{P}\left[\sum_{i=1}^d w_i(-2x_i) > -\alpha'\beta\right] = \mathbb{P}\left[\sum_{i=1}^d 2w_i(-x_i + \tilde{\alpha}\beta) > (2\tilde{\alpha} - \alpha')\beta\right] \\ &\leq \frac{4\tilde{\alpha}\alpha'\beta^2(1-\tilde{\alpha}\beta)}{(2\tilde{\alpha}-\alpha')^2\beta^2} \leq \frac{4\tilde{\alpha}\alpha'}{(2\tilde{\alpha}-\alpha')^2}\end{aligned}$$

where the last inequality uses that $\beta < 1/(2\tilde{\alpha})$.

Thus, we conclude by the above and Equation (30) that in the case that $\|w\|_\infty \leq \alpha'\beta$ we have

$$\mathbb{P}\left[\sum_{i=1}^d w_i \mathbf{h}(i) \leq -\alpha'\beta\right] \geq \frac{1}{2} - \frac{4\tilde{\alpha}\alpha'}{(2\tilde{\alpha}-\alpha')^2}.$$

Together with the case that $\|w\|_\infty \geq \alpha'\beta$ the claim follows. □

We now restate and prove Lemma 3.3

Lemma 3.3. *Let $\zeta m/\ln(m/r)$ be the number of coupons where $m \geq 4r$, $r \geq 1$, and $\zeta \geq 8$. Let X denote the number of samples with replacement from the coupons before seeing $\zeta m/\ln(m/r) - 2r$ distinct coupons, then $\mathbb{P}[X \leq m] \leq \frac{1}{2}$*

Proof. First, notice that seeing a new item in the next sample after having seen i distinct items happens with probability

$$p_i = \frac{\zeta m/\ln(m/r) - i}{\zeta m/\ln(m/r)}.$$

Now if we use X_i to denote the number of samples between having seen i distinct items and $i+1$ distinct items, we can write X as $\sum_{i=0}^{\zeta m/\ln(m/r)-2r-1} X_i$, i.e. as sum of independent geometric random variables with success probability p_i. By Theorem 3.1 in [6] for $0 < \lambda \leq 1$ it holds that

$$\mathbb{P}[X \leq \lambda\mathbb{E}[X]] \leq \exp\left(-\min_{i=0,\ldots,\zeta m/\ln(m/r)-2r-1}(p_i)\mathbb{E}[X](\lambda - 1 - \ln(\lambda))\right). \tag{31}$$

We now notice that

$$\min_{i=0,\ldots,\zeta m/\ln(m/r)-2r-1}(p_i) = \frac{2r+1}{\zeta m/\ln(m/r)} \geq \frac{2r}{\zeta m/\ln(m/r)}$$

and that

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i=0}^{\zeta m/\ln(m/r)-2r-1} \frac{\zeta m/\ln(m/r)}{\zeta m/\ln(m/r) - i} \\ &= \zeta m/\ln(m/r) \sum_{i=2r+1}^{\zeta m/\ln(m/r)} \frac{1}{i} \\ &\geq \zeta m/\ln(m/r) \int_{2r+1}^{\zeta m/\ln(m/r)} \frac{1}{x}dx \\ &= \zeta m/\ln(m/r) \ln\left(\frac{\zeta m/\ln(m/r)}{2r+1}\right) \\ &\geq \zeta(m/\ln(m/r)) \ln\left(\frac{\zeta m/\ln(m/r)}{4r}\right)\end{aligned} \tag{32}$$

where the first inequality follows from $1/x$ being monotonically decreasing. Using that $x/\lg(x) \geq \sqrt{x}$ for $x \geq 1$ and $\zeta \geq 8$ we get that $\mathbb{E}[X] \geq \zeta(m/\ln(m/r))\ln\left(\zeta\sqrt{m/r}/4\right) \geq \zeta m/2$.

We can now combine all those ingredients. By choosing $\lambda = 2/\zeta$ and using $\zeta \geq 8$ we get that $\lambda - 1 - \ln(\lambda) \geq 1/2$. First notice that, this choice of λ with $\mathbb{E}[X] \geq \zeta m/2$ implies $\mathbb{P}[X \leq m] \leq \mathbb{P}[X \leq \lambda\mathbb{E}[X]]$. Together with the bound on the minimum of the p_i and the lower bound on $\mathbb{E}[X]$ from Equation (32) we get from Equation (31) that

$$\mathbb{P}[X \leq m] \leq \mathbb{P}[X \leq \lambda\mathbb{E}[X]] \leq \exp\left(-\frac{2r\mathbb{E}[X](\lambda - 1 - \ln(\lambda))}{\zeta m/\ln(m/r)}\right) \leq \exp\left(-r\ln\left(\frac{\zeta m/\ln(m/r)}{4r}\right)\right)$$

From $m \geq 4r$ we get that $(m/r)/\ln(m/r) \geq 1$. Together with $\zeta \geq 8$ we get that $\ln((\zeta m/\ln(m/r))/(4r)) \geq 1$ and since $r \geq 1$ we conclude that $\mathbb{P}[X \leq m] \leq 1/2$ as claimed which concludes the proof. □

5 Conclusion

We have presented a lower bound on the sample complexity of AdaBoost, establishing that AdaBoost is sub-optimal by at least one logarithmic factor. In the proof, we make use of an adversarial weak learner that accumulates errors outside of the training set. Technically, this is achieved by relying on concentration and anti-concentration bounds to show that a random hypothesis set will be able to achieve both an advantage within the training set and a negative advantage on a small subset of points outside of it. In order to work, the weak learner needs to know the training set S, which happens to be the case in AdaBoost and many of its variants. This makes our lower bound applicable to a variety of boosting algorithms, showing that they are all sub-optimal.

In contrast, the optimal weak-to-strong learner from Larsen & Ritzert [10] precisely calls the weak learner on subsets of S, avoiding the lower bound. One key question here is whether a generalization of their idea allows to reach optimal generalization performance with a simple majority vote as in AdaBoost instead of their two-level majority scheme. Another interesting open question is the exact sample complexity of AdaBoost which currently has a logarithmic gap between our lower bound and the best known upper bound.

Acknowledgements

Supported by Independent Research Fund Denmark (DFF) Sapere Aude Research Leader grant No 9064-00068B.

References

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

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

[3] Grønlund, A., Kamma, L., Green Larsen, K., Mathiasen, A., and Nelson, J. Margin-based generalization lower bounds for boosted classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[5] Hanneke, S. The optimal sample complexity of pac learning. *The Journal of Machine Learning Research*, 17(1):1319–1333, 2016.

[6] Janson, S. Tail bounds for sums of geometric and exponential variables. *Statistics and Probability Letters*, 135:1–6, 2018.

[7] Kearns, M. Learning boolean formulae or finite automata is as hard as factoring. *Technical Report TR-14-88 Harvard University Aikem Computation Laboratory*, 1988.

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

[9] Larsen, K. G. Bagging is an optimal PAC learner. *arXiv preprint*, arXiv/2212.02264, 2022.

[10] Larsen, K. G. and Ritzert, M. Optimal weak to strong learning. *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 2022. To appear.

[11] Montgomery-Smith, S. J. The distribution of rademacher sums. *Proceedings of the American Mathematical Society*, 109(2):517–522, 1990.

[12] Rätsch, G. and Warmuth, M. K. Maximizing the margin with boosting. In *International Conference on Computational Learning Theory*, pp. 334–350. Springer, 2002.

[13] Rätsch, G., Warmuth, M. K., and Shawe-Taylor, J. Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, 6(12), 2005.

[14] Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.