Multiclass Boosting and the Cost of Weak Learning

Nataly Brukhim
Princeton University
Google AI Princeton
nbrukhim@princeton.edu

Elad Hazan
Princeton University
Google AI Princeton
ehazan@princeton.edu

Shay Moran
Technion
Google Research
smoran@technion.ac.il

Indraneel Mukherjee
indraneel.mukherjee@protonmail.com

Robert E. Schapire
Microsoft Research
schapire@microsoft.com

Abstract

Boosting is an algorithmic approach which is based on the idea of combining weak and moderately inaccurate hypotheses to a strong and accurate one. In this work we study multiclass boosting with a possibly large number of classes or categories. Multiclass boosting can be formulated in various ways. Here, we focus on an especially natural formulation in which the weak hypotheses are assumed to belong to an "easy-to-learn" base class, and the weak learner is an agnostic PAC learner for that class with respect to the standard classification loss. This is in contrast with other, more complicated losses as have often been considered in the past. The goal of the overall boosting algorithm is then to learn a combination of weak hypotheses by repeatedly calling the weak learner.

We study the resources required for boosting, especially how they depend on the number of classes k, for both the booster and weak learner. We find that the boosting algorithm itself only requires $O(\log k)$ samples, as we show by analyzing a variant of AdaBoost for our setting. In stark contrast, assuming typical limits on the number of weak-learner calls, we prove that the number of samples required by a weak learner is at least polynomial in k, exponentially more than the number of samples needed by the booster. Alternatively, we prove that the weak learner's accuracy parameter must be smaller than an inverse polynomial in k, showing that the returned weak hypotheses must be nearly the best in their class when k is large. We also prove a trade-off between number of oracle calls and the resources required of the weak learner, meaning that the fewer calls to the weak learner the more that is demanded on each call.

1 Introduction

Boosting [16] is a fundamental primitive in machine learning, most widely studied and applied in the context of binary supervised learning. Yet many important problems require classification into a great many target classes. For example, in image object recognition and building language models, the number of classes scales as the number of possible objects, or the dictionary size, respectively. It is thus of practical as well as theoretical interest to generalize boosting to the multiclass setting.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

In contrast with binary classification, there are several versions of multiclass boosting since there is more than one way to extend the weak learnability assumption. With only two labels, we simply require that the weak learner find a hypothesis with error slightly better than $1/2$, that is, a bit better than random guessing. So when the number of labels k is more than 2, perhaps the most natural extension requires that the weak learner outputs hypotheses with classification error slightly better than a random guess among the k labels, that is, with error a little better than $1 - 1/k$. However, this turns out to be too weak for boosting to be possible. Instead requiring classification error less than $1/2$ is sufficient for boosting, but far stronger than mere random guessing. Other approaches are based on reducing the multiclass problem to multiple binary problems, or constructing loss functions to be minimized by the weak learner that are significantly different than the simple classification error. See [13, 16] and references therein for a detailed discussion.

A Weak Learning Assumption. In this work we return to the classical formulation of boosting in which the weak learner aims to minimize ordinary classification loss, though in a way that sidesteps some of the hurdles just discussed. Specifically, we assume the weak learner is an agnostic PAC learner for some base class, meaning that it can find the approximately best hypothesis in that class. This requirement differs subtly, but importantly, from the kind we have been discussing in which, say, the weak hypotheses must have error not greater than some value; instead, the criterion is relative to this particular base class. Further, we take this class to consist of *simple* hypotheses, in line with the common supposition that weak hypotheses are rules of thumb from an "easy-to-learn" class [16, 18]. Thus, the weakness of the weak learner is manifested in the simplicity of the hypotheses in the base class, and the goal of the overall boosting algorithm is to learn target concepts *outside* the base class by aggregating the weak hypotheses provided by the weak learner. A similar setting was recently considered by [3] who studied binary-labeled classification.

Which target concepts can be approximated by boosting weak hypotheses from a given base-class? In this work we assume that the target concept can be represented by *weighted plurality votes* over the base class. To motivate this assumption, note that: (i) Most boosting algorithms (in both binary and multiclass settings) output a weighted vote of weak hypotheses [16, 10, 3, 8, 1]. Thus, if one focuses on such algorithms, it becomes a *necessary* condition that the target concept can be approximated arbitrarily well by such votes, and hence yields our assumption. (ii) In addition, the work by [16] conducted a thorough study of a variety of natural weak learning assumptions for multiclass boosting, and showed that each of those assumptions *implies* the plurality vote assumption. (iii) Finally, weighted votes are extremely expressive: for example, in the binary-labelled case, the work by [3] showed that under extremely mild assumptions on the base-class, one can approximate arbitrarily complex concepts by weighted votes. Relatedly, [19] proved that assuming a non-trivial base-class, one can obtain Bayes-consistent learning algorithms by boosting using weighted votes.

1.1 Contributions

We study the resources required for boosting, especially how they depend on k, the number of classes. For the booster, these include the number of examples needed for learning and the number of oracle calls to the weak learner. For the weak learner, we measure the required accuracy, that is, how close the returned hypothesis must be to the best in the base class, or alternatively, the number of unweighted samples the boosting algorithm feeds the weak learner in each round.

A Multiclass Boosting Algorithm. Our starting point (and first contribution) is a natural variant of AdaBoost (Algorithm 1), which demonstrates that it is possible to replace previously considered weak learners by one which aims to minimize the classification loss. Indeed, we prove that Algorithm 1 is able to learn any task which satisfies the general multiclass weak learning assumption of [13], which they proved to be necessary and sufficient (see Section 2).

However, a closer inspection of our algorithm reveals a surprising phenomenon: while the sample complexity of the booster exhibits a standard logarithmic dependence on k (the number of labels), the sample-complexity of the weak learner depends quadratically on k, which is *exponentially* worse. In more detail, our algorithm requires that at each round t, the weak learner returns a hypothesis h_t whose excess loss relative to the best hypothesis in the base class is less than $1/k$, which means, since k is assumed to be large, that h_t is nearly optimal. Stated differently, using standard statistical arguments, in each round t, our boosting algorithm provides the weak learner with $O(k^2)$ unweighted training examples (ignoring parameters other than k). This demonstrates a striking discrepancy

between the amount of information (in the form of labeled examples) the boosting algorithm uses from the source distribution and between the amount of information exchanged between the boosting algorithm and the weak learner.

Thus, the central question of study in this paper is whether a large dependence on the number of labels k is inherent in the interaction between the weak learner and the boosting algorithm. This question has a direct and significant relationship with the overall complexity of multiclass boosting. Indeed, each call to the weak learner amounts to solving a learning problem over the base class whose complexity is controlled by the prescribed excess loss. Thus, when the number of labels is large, it would be highly beneficial to improve the dependence exhibited by our Algorithm 1.

An Impossibility Result. Interestingly (and somewhat disappointingly), we show that the discrepancy between the sample complexities of the booster and weak learner is inevitable: in Section 5 we prove that, assuming typical limits on the number of weak-learner calls, the number of samples required by the weak learner is at least polynomial in k, which is exponentially more than the number of samples needed by the booster. Alternatively, we prove that the weak learner's accuracy parameter must be smaller than an inverse polynomial in k, meaning that the returned weak hypotheses must be nearly the best in the base class when k is large. This lower bound follows from a more general trade-off that we prove between the number of oracle calls (to the weak learner) and the sample complexity of the weak learner, meaning that the fewer calls to the weak learner the more that is demanded on each call. Specifically, we show that at least one of these quantities must be $\Omega(\sqrt{k})$.

From a technical perspective, the main challenge in proving this trade-off is the construction of the base class; specifically, the requirement that the base class is easy-to-learn rules out naive randomized constructions (e.g. as in the oracle-complexity lower bound in [16]), and requires a subtler treatment.

Summarizing, for the model of multiclass boosting that we study, we prove that there is a striking discrepancy between the amount of information (in the form of labeled examples) the boosting algorithm requires from the source distribution and the amount of information exchanged in the channel between the weak learner and the booster.

1.2 Related work

The boosting methodology is a fairly mature method, originally formulated for binary classification (e.g. AdaBoost and similar variants) [16]. In contrast with binary classification, there are several versions of multiclass boosting, since there is more than one way to extend the weak learnability assumption. The earlier extensions of boosting to the multiclass setting include AdaBoost.MH and AdaBoost.MR [17], as well as [2]. These works often reduce the k-class problem into a single binary problem with a k-fold augmented dataset, or to multiple one-versus-all binary problems using Error-Correcting Output Codes (ECOC). The binary reduction can have various problems, including increased complexity, and lack of guarantees of an optimal joint predictor.

When requiring weak learners with classification error less than $1/2$, a reduction to binary problems is unnecessary. Then, AdaBoost.M1 is a straightforward extension of its binary counterpart. We refer the reader to the book by [16] (Chapter 10) for a through survey of the various extensions of AdaBoost to the multiclass setting.

More recent methods [20, 12, 11, 4, 5, 15] were designed from a practical perspective, and provide empirical results of their approaches. Thus, while these works achieve improved performance across various applications, they do not provide a comprehensive theoretical framework for the multiclass boosting problem, and are often based on earlier formulations (e.g., a one-versus-all reduction to the binary setting, or multi-dimensional predictors and codewords).

Notably, previous work by [13] established a theoretical framework for multiclass boosting, which generalizes previous learning conditions. Moreover, as previous boosting methods were shown to require weak learning conditions that are either too strong or too weak, [13] outline an appropriate weak learning condition that is both necessary and sufficient for multiclass boosting. However, this requires the assumption that the weak learner minimizes a tailored loss function that is significantly different from simple classification error. See the next section for a detailed discussion.

2 Background

We study boosting algorithms that aim to learn a combination of hypotheses from a fixed base class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. The data domain $\mathcal{X}$ is finite or countable, and $\mathcal{Y}$ is the finite set of labels $[k] := \{1, ..., k\}$, for $k \geq 2$. In particular, we consider boosting as the task of learning a target concept that can be presented as a *plurality-vote* of hypotheses from $\mathcal{H}$ with margin bounded away from 0. A plurality-vote over the hypotheses class is characterized by a distribution λ over $\mathcal{H}$, and the corresponding classifier $\bar{h} = \bar{h}(\lambda)$ is defined by

$$\bar{h}(x) := \arg\max_{\ell \in [k]} \mathbb{P}_{h \sim \lambda}[h(x) = \ell]. \tag{1}$$

The boosting algorithm is given oracle access to $\mathcal{W}$, a weak learner that is an agnostic PAC learner for $\mathcal{H}$. The weakness of $\mathcal{W}$ is manifested in the simplicity of $\mathcal{H}$.[1] Concretely, $\mathcal{W}$ is defined as follows; for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, and $\epsilon^w, \delta^w > 0$, when given a sample of $m^w(\varepsilon^w, \delta^w)$ examples drawn i.i.d from $\mathcal{D}$, $\mathcal{W}$ outputs $h \in \mathcal{H}$ such that with probability $1 - \delta^w$,

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) = y] \geq \sup_{h^* \in \mathcal{H}} \mathbb{P}_{(x,y)\sim\mathcal{D}}[h^*(x) = y] - \epsilon^w. \tag{2}$$

The weak learner $\mathcal{W}$ considered in this work is defined with respect to the standard classification loss, rather than more complex formulations given in previous work on multiclass boosting. Next, we discuss an alternative weak learning condition that was shown to be necessary and sufficient for multiclass boosting [13].

2.1 Optimal weak learnability

We briefly review the formulation introduced by [13] which considers a weak learning condition with respect to a particular gain function, rather than the standard classification objective. Albeit it requires solving a seemingly unnatural classification sub-problem, it was shown to be, in a certain sense, optimal. Consider the following formulation. For any hypothesis $h \in \mathcal{H}$, any labeled example $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and any incorrect label $\ell \in \mathcal{Y}$, define:

$$\sigma_h(x, y, \ell) = \mathbb{1}[h(x) = y] - \mathbb{1}[h(x) = \ell] = \begin{cases} +1, & \text{if } h(x) = y, \\ -1, & \text{if } h(x) = \ell, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Consider the following dual perspective shown by [13], that is associated with the above gain function:

γ-margin. The *γ-margin* assumption with respect to a sample S, holds if there exists a distribution λ over $\mathcal{H}$ (corresponding to a plurality-vote), that has a large margin for every tuple (x, y, ℓ), of a labeled example from S, and an incorrect label. Concretely,

$$\exists \lambda \in \Delta_{\mathcal{H}}, \quad \forall (x, y) \in S, \ell \neq y, \qquad \mathbb{E}_{h\sim\lambda}\big[\sigma_h(x, y, \ell)\big] \geq \gamma. \tag{4}$$

γ-edge of $\mathcal{H}$. The edge is the gap between the performance of a certain hypothesis over random prediction. Define $\tilde{S} = \{(x, \ell) | (x, y) \in S, \ \ell \neq y\}$ the set of all incorrectly-labeled examples (with respect to S). The *empirical γ-weak learnability* assumption of $\mathcal{H}$ holds, if for any distribution $\tilde{\mathcal{D}}$ over $\tilde{S}$, there is a hypothesis $h \in \mathcal{H}$ that has an *edge* of γ (as is typically referred to in the binary boosting setting[2]), with respect to $\tilde{\mathcal{D}}$, over random prediction. Concretely,

$$\forall \tilde{\mathcal{D}} \in \Delta_{\tilde{S}}, \qquad \exists h \in \mathcal{H}, \qquad \mathbb{E}_{(x,\ell)\sim\tilde{\mathcal{D}}}\big[\sigma_h(x, y, \ell)\big] \geq \gamma. \tag{5}$$

Previous work by [13] has shown via a minimax argument that, for a given training set S, for the realizability of S by some γ-margin plurality-vote $\bar{h}$ to hold, i.e., for equation (4) to hold, it is both a sufficient and a necessary condition that equation (5) is satisfied. That is, (4) holds if and only if (5) holds. Note that these observations are only concerning a fixed training set S. Next, we extend these ideas using the following notion of a γ-realizable distribution.

[1] The simplicity of the base class is measured via its Natarajan dimension, an extension of the VC-dimension to the non-binary setting. See Section 4 for the formal definition and discussion.

[2] Note that for a uniformly-random predictor f, and any distribution $\mathcal{D}$, $\mathop{\mathbb{E}}_{(x,y,\ell)\sim\mathcal{D}}[\sigma_f(x, y, \ell)] = 0$.

Definition 1 (***γ-realizable sample/distribution***). *Let $\gamma \in (0,1)$, $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ a hypotheses class. A sample $S \subseteq \mathcal{X} \times \mathcal{Y}$ is γ-realizable with respect to $\mathcal{H}$ if Equation* (4) *holds. A distribution* $\mathbf{D}$ *over $\mathcal{X} \times \mathcal{Y}$ is γ-realizable with respect to $\mathcal{H}$ if any independent sample from it is γ-realizable.*[3]

The end goal of the boosting algorithm is to approximate a target γ-realizable distribution $\mathbf{D}$ with respect to classification loss. The booster is provided with oracle access to a learner for the base class $\mathcal{H}$, as well as $\epsilon^B, \delta^B > 0$ and $m(\epsilon^B, \delta^B)$ examples drawn i.i.d. from $\mathcal{D}$. The booster makes T iterative oracle calls to the learner and outputs a plurality-vote $\bar{h}_T$ such that with probability $1 - \delta^B$,

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}[\bar{h}_T(x) \neq y] \leq \epsilon^B. \tag{6}$$

3 Multiclass Adaboost: algorithm and analysis

We give an efficient adaptive algorithm for multiclass boosting, which converts an agnostic-PAC learner for the base class $\mathcal{H}$, to a learner for any target distribution that is γ-realizable with respect to $\mathcal{H}$. Towards that end, we first give a reduction to the framework provided by [13].

Specifically, we show the that the σ-gain weak-learning condition (defined in Equation (5)), can be captured by a standard classification loss, agnostic PAC learner for $\mathcal{H}$ (Equation (2)).

Lemma 2. *Let $S \subseteq \mathcal{X} \times \mathcal{Y}$ be a training set, and $\tilde{S} = \{(x,\ell) | (x,y) \in S,\ \ell \neq y\}$ the corresponding set of incorrectly-labelled examples. Let $\mathcal{W}$ be an agnostic-PAC learner for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. Let $\epsilon^w, \delta^w > 0$, and edge $\gamma > 0$, and assume that S is a γ-realizable sample w.r.t. $\mathcal{H}$. Then, for any distribution $\tilde{\mathcal{D}}$ over $\tilde{S}$, there is a distribution $\mathcal{D}$ over $S \cup \tilde{S}$, such that when given a sample of $m_w \geq m_w(\varepsilon^w, \delta^w)$ examples drawn i.i.d from $\mathcal{D}$, $\mathcal{W}$ outputs $h \in \mathcal{H}$ s.t. with probability $1 - \delta^w$,*

$$\mathbb{E}_{(i,\ell)\sim\tilde{\mathcal{D}}}\big[\sigma_h(x_i, y_i, \ell)\big] \geq \gamma - \varepsilon^w k.$$

Observe that our conversion of a classification loss learner to a σ-gain learner comes at a cost, as it incurs a multiplicative overhead of factor k in the excess error. The implications of that fact on the sample complexity of the learner are further described in Section 4.

Next, we present our multiclass boosting approach, described in Algorithm 1. We remark that AdaBoost.MR [17] a previous boosting method that is a version of AdaBoost for ranking loss, applies a similar technique as described in Algorithm 1. However, their algorithm requires access to a variant of the σ-gain learner for $\mathcal{H}$, whereas our method only assumes a standard classification loss learner.

[3] Our results remain valid for a relaxed definition of a γ-realizable distribution for which a γ-realizable samples are only drawn with high probability. However, we consider the above for simplicity of presentation.

Algorithm 1 Multiclass Adaboost

1: Input: $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.
2: Initialize: $\tilde{\mathcal{D}}_1(i, \ell) = \frac{1}{m(k-1)}$ if $\ell \neq y_i$, else set 0, for $\ell \in [k], i \in [m]$.
3: **for** $t = 1, \ldots, T$ **do**
4: Obtain $\mathcal{D}_t$ from $\tilde{\mathcal{D}}_t$ (see Lemma 2), and pass m_w examples to $\mathcal{W}$ drawn i.i.d from $\mathcal{D}_t$.
5: Let h_t be the weak hypothesis returned by $\mathcal{W}$, compute its empirical edge:

$$\gamma_t = \mathbb{E}_{(i,\ell)\sim\tilde{\mathcal{D}}_t}\left[\sigma_{h_t}(x_i, y_i, \ell)\right].$$

6: Set $\alpha_t = \frac{1}{2}\ln\left(\frac{1+\gamma_t}{1-\gamma_t}\right)$. Update, for all $i \in [m], y_i \neq \ell \in [k]$:

$$\begin{aligned}\tilde{\mathcal{D}}_{t+1}(i, \ell) &= \frac{\tilde{\mathcal{D}}_t(i, \ell)}{Z_t} \times \begin{cases} e^{-\alpha_t}, & \text{if } h_t(x_i) = y_i, \\ e^{\alpha_t}, & \text{if } h_t(x_i) = \ell, \\ 1, & \text{otherwise.} \end{cases} \\ &= \frac{\tilde{\mathcal{D}}_t(i, \ell)\exp(-\alpha_t \cdot \sigma_{h_t}(x_i, y_i, \ell))}{Z_t},\end{aligned}$$

where Z_t is a normalization factor (chosen so that $\tilde{\mathcal{D}}_{t+1}$ will be a distribution).
7: **end for**
8: Output the weighted plurality-vote,

$$\bar{h}_T(x) := \operatorname*{argmax}_{\ell\in[k]}\left(\sum_{t=1}^{T} \alpha_t \cdot \mathbb{1}[h_t(x) = \ell]\right).$$

Theorem 3. ***(Upper bound - Consistency)*** *Let $\mathcal{W}$ be an agnostic-PAC learner for the base class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. Let $\gamma, \delta > 0$, and training set S of m labelled examples. If S is γ-realizable with respect to $\mathcal{H}$ (see Definition 1), then applying Algorithm 1 with $T \geq \frac{8(\log(m)+\log(k))}{\gamma^2}$, and $m_w = m_w(\gamma/k, \delta/T)$, outputs with probability at least $1 - \delta$, an $\bar{h}_T$ that is consistent with S.*

Runtime analysis. The number of iterations required for Algorithm 1 to succeed has only a logarithmic dependence on the number of labels k. Nonetheless, since the algorithm maintains a distribution $\tilde{\mathcal{D}}$ over the data, of size $m \times k$, it might seem as if the cost of each round is linearly dependent on k. However, observe that the updates to distribution $\tilde{\mathcal{D}}$ only pertain to incorrect labels ℓ which at least one of $h_1, \ldots, h_t$ mistakenly predicted to be correct. Thus, at most mT such values are being updated. Moreover, note that performing a random sample from $\mathcal{D}_t$ (step 4 of Algorithm 1) can be done in time $O(\log(mk))$ [7], as $\tilde{\mathcal{D}}, \mathcal{D}$ maintained by the algorithm can be kept sorted throughout the iterations. Therefore, the running time of each round t is $O(m\log(mk))$, and the overall running time only has poly-logarithmic dependence on k.

4 Sample complexity upper bounds

In the previous section we have focused on the consistency model, and have shown that a multiclass boosting algorithm with a running time of only log dependence on k, can achieve zero training error. In this section we examine the generalization aspects of the algorithm. The results described below have two main implications, concerning the sample-complexity required for the boosting algorithm as a black-box (subsection 4.2), and the sample-complexity of the weak-learner required in each iteration, within the boosting framework (subsection 4.1).

We show that on the one hand, the sample complexity required for the booster defined in Algorithm 1 to learn a γ-realizable distribution $\mathbf{D}$ over $\mathcal{X} \times \mathcal{Y}$, is upper bounded by $m = O(\log(k))$. However, the upper bound we give on sample complexity of the weak learning oracle $\mathcal{W}$ is $m_w = O(k^2)$, omitting other terms. This striking contrast naturally raises the question of whether a different boosting approach exists, such that the large dependence on k is significantly improved to match the

logarithmic upper bound of the algorithm. In Section 5 we provide a lower bound showing that these gaps are inherent to the problem of multiclass boosting.

Natarajan dimension of $\mathcal{H}$. In the binary case, the VC-dimension characterizes the sample complexity of hypothesis classes. In the multiclass case, [14] introduced an extension of the VC-dimension to non-binary functions, defined as follows.

Definition 4. *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ a hypothesis class and let $S \subseteq \mathcal{X}$. We say that $\mathcal{H}$* N-shatters *S if $\exists\, \pi_1, \pi_2 : S \to \mathcal{Y}$ such that $\forall x \in S,\ \pi_1(x) \neq \pi_2(x)$, and for every $T \subseteq S$ there is $h \in \mathcal{H}$ such that,*

$$\forall x \in T,\ h(x) = \pi_1(x),\ \textit{and}\ \forall x \in S \setminus T,\ h(x) = \pi_2(x).$$

The Natarajan dimension *$d_N(\mathcal{H})$, is the maximal cardinality of a set that is N-shattered by $\mathcal{H}$.*

The Natarajan dimension coincides with the VC-dimension for $k = 2$. The Natarajan dimension of a hypothesis class, $d_N(\mathcal{H})$, characterize the sample complexity required for PAC-learning the class $\mathcal{H}$, up to a multiplicative factor of $O(\log(k))$. Specifically, by [6, 9] the sample complexity required to agnostically-PAC learn a class $\mathcal{H}$ is upper bounded by,

$$m_{\mathcal{H}}(\epsilon, \delta) = O\left(\frac{d_N(\mathcal{H})\log(k) + \ln(1/\delta)}{\epsilon^2}\right). \tag{7}$$

4.1 The sample complexity of weak-learning

In section 3 we have discussed a reduction of a classification-loss learner to a σ-gain learner. However, this transformation required a multiplicative overhead of factor k in the excess error. Observe that by plugging in any $\epsilon^w < \gamma/k$, we get that the hypothesis returned by $\mathcal{W}$ has a strictly positive edge, which allows boosting. The implications of that fact on the sample complexity of $\mathcal{W}$ are summarized in Remark 5, which immediately follows from Theorem 3, Lemma 2, and Equation (7).

Remark 5. *Given oracle access to an agnostic-PAC learner $\mathcal{W}$ for $\mathcal{H}$, and sample set S of size m, applying Algorithm 1 with $T = O(\log(mk)/\gamma^2)$ oracle calls to $\mathcal{W}$, setting $\epsilon_t^w < \gamma/k$, and passing $m_w = O\big(\frac{d_N(\mathcal{H})\log(k)}{\gamma^2} \cdot k^2\big)$ examples to $\mathcal{W}$, yields an output $\bar{h}_T$ is consistent with S.*

4.2 The sample complexity of Boosting

Next, we show that the sample complexity required for multiclass boosting to learn a γ-realizable distribution $\mathbf{D}$ over $\mathcal{X} \times \mathcal{Y}$, does only require a logarithmic dependence on k. This bound on the sample complexity is given by considering the output predictor in Algorithm 1 and bound the sample complexity that is required for it to generalize.

In the following result, formally stated in Theorem 6, we bound the generalization error of a predictor outputted by the boosting algorithm, given in Algorithm 1. In particular, we use a margins-based analysis of the generalization error (see margins-based analysis for the binary case in [16], chapter 5.2), utilizing two key facts; (i) the predictor $\bar{h}_T$ obtained by Algorithm 1 has a large empirical margin, and (ii) uniform convergence applies to the margin of $\bar{h}_T$. Moreover, the large large margin property both holds empirically, and generalizes, while requiring only a logarithmic dependence on k. These facts are formally stated and proved in the appendix.

Theorem 6. ***(Upper bound - Generalization)*** *Let $\mathbf{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ that is γ-realizable by a hypotheses class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, and $\gamma > 0$. Let $\epsilon, \delta > 0$, and the training set S be a random sample of $m = \Theta\left(\frac{d_N(\mathcal{H})\log(k)\log(\frac{1}{\epsilon})}{\gamma^2\epsilon^2} + \frac{\log(1/\delta)}{\epsilon^2}\right)$ examples drawn i.i.d from $\mathbf{D}$. Then, with probability $1-\delta$, over the random choice of the training set S, applying Algorithm 1 with $T = \Theta(\frac{\log(mk)}{\gamma^2})$, outputs a predictor $\bar{h}_T$ such that,*

$$\mathbb{P}_{(x,y)\sim\mathbf{D}}\left[\bar{h}_T(x) \neq y\right] \leq \epsilon.$$

Note that although Theorem 6 pertains to the complexity required for learning via a *boosting* algorithm; a similar statement can even be verified to hold for any ERM learner for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, which can learn any γ-realizable for $\mathcal{H}$, with a similar sample complexity. This follows from the definition of γ-realizability, the fact that a sparse plurality-vote can be sampled from the realizable one (a proof technique also used in Theorem 6), and the sample complexity bound given (7). Overall, we get that the task of multiclass learning, either directly or via boosting, has a logarithmic dependence on k.

5 Matching Lower Bounds

In this section, we give lower bounds on the sample and oracle complexities required for a boosting algorithm to attain a small generalization error, for a large number of labels $|\mathcal{Y}| = k$. Alternatively, such bounds can be inverted to give a a lower bound on the generalization error that can be achieved by a boosting algorithm for a fixed sample complexity and running time. Thus, these bounds characterize the efficiency of any multiclass boosting algorithm, in terms of how the number of rounds and samples must depend on the desired accuracy, with respect to k.

5.1 Model definition

Recall that in previous sections, a weak learning algorithm $\mathcal{W}$ was defined to be an agnostic-PAC learner for a hypothesis class $\mathcal{H}$, which satisfies Equation (2). Towards the goal of providing a lower bound for a general boosting algorithm, we consider two alternative, standard models of interaction between a booster and weak learner, that only differ in the input the weak learner expects to receive.

1. **Boosting-by-resampling** - in this form of interaction, the oracle $\mathcal{W}$ is a learning algorithm for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that when given a set of $m_w(\epsilon_w, \delta_w = 1/3)$ labelled examples drawn i.i.d. from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, it returns an ϵ_w-optimal hypothesis in $\mathcal{H}$, with respect to the classification loss over $\mathcal{D}$. In this setting, we consider the learner $\mathcal{W}$ to simply be an ERM (empirical risk minimizer) for a given sample.
2. **Boosting-by-reweighting** - in this model, the learner $\mathcal{W}$ is allowed to directly utilize the distribution over examples $\mathcal{D}$, so that the weighted training error is approximately minimized explicitly, but $\mathcal{W}$ may still return any ϵ_w-optimal hypothesis in $\mathcal{H}$ over $\mathcal{D}$.

In addition, we now define what is a general boosting algorithm here, for either interaction form. A boosting algorithm B is one which, when provided with oracle access a weak learner $\mathcal{W}$ for $\mathcal{H}$ (in either model of interaction; resampling or reweighting), as well as $\epsilon^B, \delta^B > 0$, and a training set S of $m = m(\epsilon^B, \delta^B)$ examples drawn i.i.d. from a target distribution $\mathbf{D}$ over $\mathcal{X} \times \mathcal{Y}$ that is γ-realizable (see Definition 1), will with probability at least $1 - \delta^B$, output a combined hypothesis that is a plurality-vote of hypotheses $h \in \mathcal{H}$ obtained via the learner $\mathcal{W}$, such that its classification error with respect to $\mathbf{D}$ is at most ϵ^B. In our construction, we simply fix $\epsilon^B = 2\gamma$, and $\delta^B = 1/2$.

Although the boosting algorithm may be randomized, we can regard its randomization source, denoted $\mathcal{R}$, as itself an input to the algorithm. Thus, we view B as a fixed and deterministic function of the training set S, the internal randomization $\mathcal{R}$, and the responses obtained via the learning oracle $\mathcal{W}$.

Note that we assume B does not know the hypothesis class $\mathcal{H}$, and its only access to it is through its access to the learning oracle $\mathcal{W}$. However, it is allowed to have other information about the weak learner, such as m^w the required sample size of $\mathcal{W}$, the associated edge γ, etc. Note that we also restricted the booster such that it may only output a plurality-vote of hypotheses $h \in \mathcal{H}$ obtained via the learner $\mathcal{W}$. Aside from the above, the boosting algorithm B is entirely unrestricted.

5.2 Results

We derive lower bounds using a simple construction of a class $\mathcal{H}$, and a base learner $\mathcal{W}$, applied in both models described above; by-resampling and by-reweighting. The main idea of the construction is of designing a simple class $\mathcal{H}$, such that its hypotheses will make diverse mistakes which the booster cannot know in advance. This forces the booster to pick a distribution $\mathcal{D}_t$ with a large support of $\Omega(\sqrt{k})$ elements, and require high accuracy $\epsilon_t \approx 1/k$. Otherwise, it can be shown no matter how the obtained hypotheses are combined in a plurality-vote, there will always be a large mistake, unless a large number of rounds $T = \Omega(\sqrt{k})$ is used. This trade-off between oracle and sample complexities provides the lower bound we seek, showing that any booster incurs a cost of $\Omega(\sqrt{k})$.

Theorem 7. ***(Lower bound)*** *Let $\gamma \in (0, 1/2)$ denote the margin parameter, $k > 2$ number of labels, and assume that $k > 2/\gamma$. Let $T > 0$ denote the number of iterations. For any boosting algorithm B, by-reweighting or by-resampling, there exists a class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, an agnostic-PAC learner $\mathcal{W}$ for $\mathcal{H}$, and a target $\frac{\gamma}{2}$-realizable distribution $\mathcal{D}$, such that; If B is applied with oracle access to $\mathcal{W}$ for T iterations, and for all $t \leq T$, B picks error rates $\epsilon_t^w > \theta := \frac{2T}{k - 1/\gamma - 1}$ (reweighting model), or picks*

sample size $m_t^w < 1/\theta$ (resampling model), and then outputs a final hypothesis $\bar{h}_B$, then,

$$\mathbb{P}_{S,\mathcal{R}}\left[\mathbb{P}_{\mathcal{D}}\left[\bar{h}_B(x) \neq y\right] > 2\gamma\right] \;\geq\; 1/2,$$

where S is the initial random training sample drawn from $\mathcal{D}$, and $\mathcal{R}$ denotes the internal randomization of the boosting algorithm.

The trade-off between the oracle and sample complexity bounds is demonstrated in Corollary 8 next.

Corollary 8. ***(Lower bound - complexity trade-off)*** *By applying Theorem 7 with different values of T, we get that for any boosting algorithm B to achieve error rate at most 2γ,*

1. *There exists a class $\mathcal{H}$ and learner $\mathcal{W}$, s.t. if B makes $T = O(\frac{1}{\gamma})$ oracle calls to $\mathcal{W}$, then there is some $t \leq T$ in which $\mathcal{W}$ must incur a sample complexity of $m_t^w = \Omega(k)$.*
2. *There exists a class $\mathcal{H}$ and learner $\mathcal{W}$, s.t. either B makes $T = \Omega(\sqrt{k})$ oracle calls to $\mathcal{W}$, or there is some $t \leq T$ in which $\mathcal{W}$ must incur a sample complexity of $m_t^w = \Omega(\sqrt{k})$.*
3. *There exists a class $\mathcal{H}$ and learner $\mathcal{W}$, s.t. if for all $t \leq T$, the sample complexity of $\mathcal{W}$ is $m_t^w = O(1)$, then B makes $T = \Omega(\sqrt{k})$ oracle calls to $\mathcal{W}$.*

Recall that k is thought of as the very large, dominating term, compared to $1/\gamma$. However, the results shown in Corollary 8 only rely on the assumption $k > 1/\gamma$. We remark that if $k \gg 1/\gamma$, the interplay between the time and sample complexity bounds changes accordingly. For example, if $k > 1/\gamma^2$, it can be similarly shown that the first item of Corollary 8 holds for $T = O(\frac{1}{\gamma^2})$ as well.

5.3 Proof overview

Below is an informal, high-level, overview of the proof. We refer the reader to the appendix for the formal statements and proofs. We begin with describing the construction of the class $\mathcal{H}$. Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be the domain, with $n = \frac{1}{2\gamma}$ data points, and assume that $n \ll k$. We fix the constant function $f(x) = 1$ to be the target labelling. Then, the labelled data is $\{(x_1, 1)...(x_n, 1)\}$, and let the target distribution $\mathcal{D}$ be the uniform distribution over it. We note that our construction and the lower bound argument hold even for boosting algorithms that have full knowledge of the data and distribution $\mathcal{D}$. However, recall that the boosting algorithm is restricted to only output a weighted plurality-vote of hypotheses $h^1, ...h^T \in \mathcal{H}$ obtained via $\mathcal{W}$. Next, consider hypotheses of the following form; for all $j \in [n]$ and $\ell \in \{2, ..., k\}$, define

$$h_{j,\ell}(x) = \begin{cases} 1 & \text{if } x = x_j, \\ \ell & \text{otherwise.} \end{cases} \tag{8}$$

We denote the set of all such hypotheses as $\mathcal{H}_{all}$. The base-class $\mathcal{H}$ will be a carefully chosen random subset of $\mathcal{H}_{all}$. The key properties that guide the construction of $\mathcal{H}$ are: (i) a uniform plurality vote over all hypotheses in $\mathcal{H}$ yields the target-function f with margin $\Omega(\gamma)$, and (ii) every weighted plurality vote of a <u>proper</u> subset of $\mathcal{H}$ is far-away from f.

We then use the probabilistic method to randomly construct $\mathcal{H}$. Specifically, the class $\mathcal{H}$ contains $2n - 2$ carefully chosen hypotheses that are <u>fixed</u>, and a single random hypothesis, $h_{j,L}$, where L is drawn uniformly at random from the set of labels $2n, ..., k$ (recall that $k \gg n$). We stress that the $2n - 2$ hypotheses, and the index j, are chosen such that for each possible L, the obtained class satisfies the two properties noted above.

Observe that the class $\mathcal{H}_{all}$ is an easy-to-learn class, and it can be verified that it has a constant Natarajan dimension $= 2$, and that it is learnable by any ERM with sample complexity independent of k. Since $\mathcal{H} \subset \mathcal{H}_{all}$ for any L, it exhibits theses traits as well. Moreover, by property (i) discussed above, it holds that there exists a plurality-vote over $\mathcal{H}$ that is consistent with the data, with margin $\gamma/2$. That is, the distribution $\mathcal{D}$ which is uniform over $\mathcal{X} \times \{1\}$ is $\gamma/2$-realizable, with respect to $\mathcal{H}$.

We refer to the randomly chosen hypothesis $h_{j,L}$ as the *hidden* hypothesis. Note that if the booster cannot "discover" the hidden hypothesis throughout its run (i.e., obtain it via the weak learner), then by property (ii) of $\mathcal{H}$, any weighted plurality it can output <u>must</u> have large error. Thus, it suffices to exhibit an agnostic PAC learner for $\mathcal{H}$, for which the booster will fail to obtain the hidden hypothesis unless the accuracy/sample-complexity are high.

The "finder-chooser" game. The above construction forces any successful boosting algorithm to find the hidden hypothesis. This is the crucial observation needed for our proof and it enables us to design a weak learner which reduces the task of finding the hidden hypothesis to the following simple "finder-chooser" game. This is an iterative game in which a *"chooser"* player picks a "hidden" element L out of a set of elements K, of size $|K| = k$, and a *"finder"* player iteratively attempts to discover it. Specifically, the finder is only required to find any small subset of elements from K that the hidden element L belongs to. Towards that end, in each round the finder queries the chooser, and the chooser replies with a binary response, which indicates if the chooser has indeed picked the hidden element. We consider two different query-models in which the finder is allowed to interact with the chooser; via weights, or via examples (which correspond to the alternative model of boosting, discussed in subsection 5.1). In both cases, the finder is given a binary response in each round, of whether or not the hidden element is in the subset it picked. If the finder failed, it continues to the next round, otherwise it succeeded to find a subset which contains L, and the game ends.

It is clear that there is a trade-off between the size of the picked subsets and the number of rounds needed to find the containing subset. We formalize this trade-off, and give bounds on the probability that the finder succeeds, when the chooser is uniform over K. This in turn, yields the desired bounds by reduction to the boosting setting.

6 Conclusion

We study the problem of multiclass boosting, where we focus on the case where the number of labels $k \gg 2$ is large, and study how k affects complexity aspects of boosting. Our first result is a natural extension of Adaboost to the multiclass case, using a simple weak learning oracle. We then show that there is a prominent gap between the sample complexities of the weak learner $O(k^2)$, and of the boosting algorithm $O(\log(k))$, that are needed for learning. Finally, we give a lower bound which shows that indeed, even for a trivial (constant-time) learning tasks of both the base class and the target, learning via boosting is a more challenging task. That is, a large dependence on k is in fact an inherent cost required for multiclass boosting.

Acknowledgments and Disclosure of Funding

S. Moran is a Robert J. Shillman Fellow and is supported by the ISF, grant no. 1225/20, by an Azrieli Faculty Fellowship, and by BSF grant 2018385. Part of this work was done when S. Moran was at Google AI Princeton. N. Brukhim and E. Hazan acknowledge the support of NSF grant 1704860.

References

[1] Naman Agarwal, Nataly Brukhim, Elad Hazan, and Zhou Lu. Boosting for control of dynamical systems. In *International Conference on Machine Learning*, pages 96–103. PMLR, 2020.

[2] Erin L Allwein, Robert E Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of machine learning research*, 1(Dec):113–141, 2000.

[3] Noga Alon, Alon Gonen, Elad Hazan, and Shay Moran. Boosting simple learners. *STOC*, 2021.

[4] Ron Appel and Pietro Perona. A simple multi-class boosting framework with theoretical guarantees and empirical proficiency. In *International Conference on Machine Learning*, pages 186–194. PMLR, 2017.

[5] Oscar Beijbom, Mohammad Saberian, David Kriegman, and Nuno Vasconcelos. Guess-averse loss functions for cost-sensitive multiclass boosting. In *International Conference on Machine Learning*, pages 586–594. PMLR, 2014.

[6] Shai Ben-David, Nicolo Cesabianchi, David Haussler, and Philip M Long. Characterizations of learnability for classes of (0,..., n)-valued functions. *Journal of Computer and System Sciences*, 50(1):74–86, 1995.

[7] Karl Bringmann and Konstantinos Panagiotou. Efficient sampling methods for discrete distributions. In *International Colloquium on Automata, Languages, and Programming*, pages 133–144. Springer, 2012.

[8] Nataly Brukhim, Xinyi Chen, Elad Hazan, and Shay Moran. Online agnostic boosting via regret minimization. In *Advances in Neural Information Processing Systems*, 2020.

[9] Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. *The Journal of Machine Learning Research*, 16(1):2377–2404, 2015.

[10] Varun Kanade and Adam Kalai. Potential-based agnostic boosting. In *Advances in neural information processing systems*, 2009.

[11] Balázs Kégl. The return of adaboost. mh: Multi-class hamming trees. *ICLR*, 2014.

[12] Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. Multi-class deep boosting. In *Advances in Neural Information Processing Systems*, pages 2501–2509, 2014.

[13] Indraneel Mukherjee and Robert E Schapire. A theory of multiclass boosting. *Journal of Machine Learning Research*, 14:437–497, 2011.

[14] Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

[15] Mohammad J Saberian and Nuno Vasconcelos. Multiclass boosting: Theory and algorithms. In *Advances in Neural Information Processing Systems*, pages 2124–2132, 2011.

[16] Robert E Schapire and Yoav Freund. *Boosting: Foundations and algorithms*. Cambridge university press, 2012.

[17] Robert E Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.

[18] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014.

[19] Peter L Bartlett Mikhail Traskin. Adaboost is consistent. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, volume 19, page 105. MIT Press, 2007.

[20] Shaodan Zhai, Tian Xia, and Shaojun Wang. A multi-class boosting method with direct optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 273–282, 2014.

A Omitted proofs of Section 3

A.1 Proof of Lemma 2

Proof. Assume $|S| = m$, for some $m > 0$. Then, corresponding set of incorrectly-labelled examples is of size $|\tilde{S}| = m(k-1)$. Let $\tilde{\mathcal{D}} \in \Delta_{\tilde{S}}$, be any distribution over $m \times (k-1)$. We show that we can construct a distribution $\mathcal{D} \in \Delta_{\tilde{S} \cup S}$ over $m \times k$, for which the guarantee in the Lemma holds.

First, observe that there must exist $a \in \Delta_m$ and $b_1, ..., b_m \in \Delta_{(k-1)}$, such that for all i, j, we have $\tilde{\mathcal{D}}(i,j) = a(i)b_i(j)$. For every $i \in [m]$, let $b'_i \in \Delta_k$ be a distribution constructed as follows:

$$b'_i(j) = \begin{cases} \frac{2}{k} & \text{if } j = y_i \\ \frac{1}{k} - \frac{b_i(j)}{k} & \text{otherwise} \end{cases} \tag{9}$$

Then, we set $\mathcal{D}$ such that for all i, j, we have $\mathcal{D}(i,j) := a(i)b'_i(j)$. Then, observe that for any $h \in \mathcal{H}$,

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}}[h(x) = y] &= \sum_{i=1}^{m} \sum_{j=1}^{k} a(i) b'_i(j) \mathbb{1}[h(x_i) = j] && \text{(By substituting } \mathcal{D} \text{ for } a, b') \\
&= \sum_{i=1}^{m} a(i) \left(\frac{2}{k} \mathbb{1}[h(x_i) = y_i] + \sum_{j \neq y_i} \frac{1 - b_i(j)}{k} \mathbb{1}[h(x_i) = j] \right) && \text{(By (9))} \\
&= \frac{2}{k} \mathbb{P}_{\tilde{\mathcal{D}}}[h(x) = y] + \frac{1}{k}(1 - \mathbb{P}_{\tilde{\mathcal{D}}}[h(x) = y]) - \frac{1}{k} \sum_{i=1}^{m} \sum_{j \neq y_i} \tilde{\mathcal{D}}(i,j) \mathbb{1}[h(x_i) = j] && (10) \\
&= \frac{1}{k} \mathbb{P}_{\tilde{\mathcal{D}}}[h(x) = y] + \frac{1}{k} - \frac{1}{k} \sum_{i=1}^{m} \sum_{j \neq y_i} \tilde{\mathcal{D}}(i,j) \mathbb{1}[h(x_i) = j] && (11) \\
&= \frac{1}{k} + \frac{1}{k} \left(\mathbb{P}_{\tilde{\mathcal{D}}}[h(x) = y] - \sum_{i=1}^{m} \sum_{j \neq y_i} \tilde{\mathcal{D}}(i,j) \mathbb{1}[h(x_i) = j] \right) && (12) \\
&= \frac{1}{k}(1 + \mathbb{E}_{(i,j) \sim \tilde{\mathcal{D}}} \big[\sigma_h(x_i, y_i, j) \big]). && (13)
\end{aligned}$$

By running the learner $\mathcal{W}$ over m^w i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypothesis $h \in \mathcal{H}$ such that, with probability of at least $1 - \delta^w$ (over the sampling of the m^w examples from $\mathcal{D}$), we have,

$$\mathbb{P}_{\mathcal{D}}[h(x) = y] \geq \max_{h^* \in \mathcal{H}} \mathbb{P}_{\mathcal{D}}[h^*(x) = y] - \epsilon^w. \tag{14}$$

Then, by applying the transformation given in (13) to (14) on both size, and re-arrange terms, we get,

$$\mathbb{E}_{(i,\ell) \sim \tilde{\mathcal{D}}} \big[\sigma_h(x_i, y_i, \ell) \big] \geq \max_{h^* \in \mathcal{H}} \mathbb{E}_{(i,\ell) \sim \tilde{\mathcal{D}}} \big[\sigma_{h^*}(x_i, y_i, \ell) \big] - \epsilon^w k \geq \gamma - \epsilon^w k,$$

where the last inequality follows by the assumption that S is γ-realizable with respect to $\mathcal{H}$ (see Definition 1). □

A.2 Proof of Theorem 3

Proof. For any $x \in \mathcal{X}, \ell \in [k]$, we define:

$$F(x, \ell) = \sum_{t=1}^{T} \alpha_t \mathbb{1}[h_t(x) = \ell]. \tag{15}$$

Unraveling the recurrence in Algorithm 1 that defines $\tilde{\mathcal{D}}_{t+1}$ in terms of $\tilde{\mathcal{D}}_t$ gives,

$$\begin{aligned}\tilde{\mathcal{D}}_{T+1}(i,\ell) &= \tilde{\mathcal{D}}_1(i,\ell) \times \frac{e^{-\alpha_1 \cdot \sigma_{h_1}(x_i,y_i,\ell)}}{Z_1} \times ... \times \frac{e^{-\alpha_T \cdot \sigma_{h_T}(x_i,y_i,\ell)}}{Z_T} \\ &= \tilde{\mathcal{D}}_1(i,\ell) \times \frac{e^{-\sum_{t=1}^T \alpha_t \cdot \sigma_{h_t}(x_i,y_i,\ell)}}{\prod_{t=1}^T Z_t} \\ &= \tilde{\mathcal{D}}_1(i,\ell) \times \frac{e^{-\sum_{t=1}^T \alpha_t \cdot \left(\mathbb{1}[h_t(x_i)=y_i] - \mathbb{1}[h_t(x_i)=\ell]\right)}}{\prod_{t=1}^T Z_t} \\ &= \tilde{\mathcal{D}}_1(i,\ell) \times \frac{e^{-F(x_i,y_i)+F(x_i,\ell)}}{\prod_{t=1}^T Z_t}.\end{aligned} \tag{16}$$

Next, note that $\bar{h}_T(x) = \text{argmax}_{\ell \in [k]} F(x,\ell)$. Therefore, for any $i \in [m]$ and $\ell \neq y_i$, we get that if $\bar{h}_T(x_i) = \ell$, then $F(x_i,\ell) \geq F(x_i,y_i)$. This implies that $e^{F(x_i,\ell)-F(x_i,y_i)} \geq 1$. Hence, we have, $\mathbb{1}[\bar{h}_T(x_i) = \ell] \leq e^{F(x_i,\ell)-F(x_i,y_i)}$. We can now bound the label-weighted error:

$$\sum_{i=1}^m \sum_{\ell \neq y_i} \tilde{\mathcal{D}}_1(i,\ell) \cdot \mathbb{1}[\bar{h}_T(x_i) = \ell] \leq \sum_{i=1}^m \sum_{\ell \neq y_i} \tilde{\mathcal{D}}_1(i,\ell) \cdot e^{F(x_i,\ell)-F(x_i,y_i)} \tag{17}$$

$$= \sum_{i=1}^m \sum_{\ell \neq y_i} \tilde{\mathcal{D}}_{T+1}(i,\ell) \prod_{t=1}^T Z_t \tag{18}$$

$$= \prod_{t=1}^T Z_t, \tag{19}$$

where equation (18) uses equation (16), and equation (19) uses the fact that $\tilde{\mathcal{D}}_{T+1}$ is a distribution which sums to 1 (over all examples and their *incorrect* labels). Denote $q_t = \sum_{i=1}^m \sum_{\ell \neq y_i} \tilde{\mathcal{D}}_t(i,\ell) \cdot \mathbb{1}[h_t(x_i) = \ell]$. Then, bounding the normalization factor,

$$Z_t = \sum_{i=1}^m \sum_{\ell \neq y_i} \tilde{\mathcal{D}}_t(i,\ell) \cdot e^{-\alpha_t \cdot \sigma_{h_t}(x_i,y_i,\ell)} \tag{20}$$

$$= \sum_{i=1}^m \sum_{\ell \neq y_i} \tilde{\mathcal{D}}_t(i,\ell) \cdot \Big(\mathbb{1}[h_t(x_i) = y_i] \cdot e^{-\alpha_t} + \mathbb{1}[h_t(x_i) = \ell] \cdot e^{\alpha_t} + \mathbb{1}[h_t(x_i) \notin \{y_i,\ell\}]\Big) \tag{21}$$

$$= (q_t + \gamma_t) \cdot e^{-\alpha_t} + q_t \cdot e^{\alpha_t} + (1 - 2q_t - \gamma_t), \tag{22}$$

where equation (22) simply follows by the definition of q_t and γ_t (step 5 of Algorithm 1). Observe that by plugging our choice of α_t in to equation (22) and re-arranging terms, we get that the coefficient of q_t is: $\frac{2}{\sqrt{1-\gamma_t^2}} - 2$, which is a positive term for any $\gamma_t \in (0,1)$. Therefore, equation (22) is a monotonic increasing function of q_t. Moreover, since $(1 - 2q_t - \gamma_t) \geq 0$, we have $q_t \leq \frac{1}{2} - \frac{\gamma_t}{2}$. Thus, we get,

$$Z_t \leq e^{-\alpha_t} \cdot \left(\frac{1}{2} + \frac{\gamma_t}{2}\right) + e^{\alpha_t} \cdot \left(\frac{1}{2} - \frac{\gamma_t}{2}\right) = \sqrt{1-\gamma_t^2}. \tag{23}$$

Lastly, plugging into equation (19) gives,

$$\sum_{i=1}^m \sum_{\ell \neq y_i} \tilde{\mathcal{D}}_1(i,\ell) \mathbb{1}[\bar{h}_T(x_i) = \ell] \leq \prod_{t=1}^T \sqrt{1-\gamma_t^2} \tag{24}$$

$$\leq e^{-1/2 \cdot \sum_{t=1}^T \gamma_t^2} \tag{25}$$

$$\leq e^{-T\gamma^2/8} \tag{26}$$

$$\leq \frac{1}{m \cdot k}, \tag{27}$$

where equation (25) follows by applying the approximation $1 + r \leq e^r$ for all real r, equation (26) follows by Lemma 2, where $\epsilon = \gamma/(2k)$, and thus $\gamma/2 \leq \gamma_t$ for all t. Equation (27) follows by plugging T. Since $\tilde{\mathcal{D}}_1$ is uniform over all examples and incorrect labels, then if the label-weighted training error of the combined classifier $\bar{h}_T$, which is always an integer multiple of $1/m(k-1)$, is at most $1/(m \cdot k)$, then the training error must in fact be zero. □

B Omitted proofs of Section 4.2

Lemma 9. *Let $\mathcal{W}$ be an agnostic-PAC learner for the base hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. Let $\gamma > 0$, and training set S of m labelled examples. If $\mathcal{H}$ is empirically γ-weak learnable with respect to S, then applying Algorithm 1 with $T = \frac{8(\log(m)+\log(k))}{\gamma^2}$, outputs $\bar{h}_T$ that is consistent with S. Furthermore, the empirical margin of $\bar{h}_T$ is at least $\gamma/8$.*

Proof. In this proof we use the notation introduced in the proof of Theorem 3. The consistency statement of the Lemma follows by Theorem 3. Next, we show that the empirical probability of $\bar{h}_T$ to have a low margin is zero. Specifically we show that,

$$F(x,y) - \max_{\ell \neq y} F(x,\ell) \geq \bar{\alpha} \cdot \gamma/8,$$

holds for all $(x,y) \in S$, where $\bar{\alpha} = \sum_{t=1}^{T} \alpha_t$ denotes the normalizing factor of F. Observe that the converse event occurs if and only if, for some $(x,y) \in S$,

$$\exp\Big(- F(x,y) + \max_{\ell \neq y} F(x,\ell) + \bar{\alpha} \cdot \gamma/8 \Big) \geq 1,$$

which in turn occurs if,

$$1 \leq e^{\bar{\alpha}\gamma/8} \max_{\ell \neq y} e^{F(x,\ell)-F(x,y)} \leq e^{\bar{\alpha}\gamma/8} \sum_{\ell \neq y} e^{F(x,\ell)-F(x,y)} = e^{\bar{\alpha}\gamma/8} \sum_{\ell \neq y} e^{F(x,\ell)-F(x,y)}. \tag{28}$$

Then, we get that,

$$\mathbb{P}_{(x,y)\sim S}\Big[F(x,y) \text{ has margin} \leq \gamma/8\Big] = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\Big[F(x,y) - \max_{\ell \neq y} F(x,\ell) \leq \bar{\alpha} \cdot \gamma/8\Big] \tag{29}$$

$$\leq \frac{e^{\bar{\alpha}\gamma/8}}{m} \sum_{i=1}^{m} \sum_{\ell \neq y} e^{F(x,\ell)-F(x,y)} \tag{30}$$

$$= e^{\bar{\alpha}\gamma/8}(k-1)\frac{1}{m(k-1)} \sum_{i=1}^{m} \sum_{\ell \neq y} e^{F(x,\ell)-F(x,y)} \tag{31}$$

$$= e^{\bar{\alpha}\gamma/8}(k-1) \prod_{t=1}^{T} Z_t, \tag{32}$$

$$\tag{33}$$

where the first inequality follows by (28), and the last equality uses Equation (18) (and summing $\tilde{\mathcal{D}}_{T+1}$ to 1 as in (19)). Recall that, as in the proof of Theorem 3, applying Lemma 2 with a choice of $\epsilon^w = \gamma/(2k)$, we get that $\gamma/2 \leq \gamma_t$ for all t. Thus, by that, and by the definition of $\bar{\alpha}$, and the bound on Z_t (see (23)), we get,

$$\mathbb{P}_{(x,y)\sim S}\Big[F(x,y) \text{ has margin} \leq \gamma/8\Big] \leq (k-1)\prod_{t=1}^{T} e^{\alpha_t \gamma_t/4} Z_t, \tag{34}$$

$$= (k-1)\prod_{t=1}^{T} (1-\gamma_t)^{\frac{1}{2}-\frac{\gamma_t}{4}} (1+\gamma_t)^{\frac{1}{2}+\frac{\gamma_t}{4}} \tag{35}$$

$$\leq (k-1)\Big[(1-\gamma/2)^{1-\frac{\gamma}{4}}(1+\gamma/2)^{1+\frac{\gamma}{4}}\Big]^{T/2} \tag{36}$$

$$\leq (k-1)\, e^{-T\gamma^2/8}, \tag{37}$$

where the first inequality holds since the inner expression of the product is a decreasing function of γ_t, and (36) holds since $\gamma/2 \leq \gamma_t$. Equation (37) follows by the the fact that the following inequality holds for all $\gamma \in (0,1)$,

$$-\frac{\gamma^2}{2\big((1-\frac{\gamma}{4})\ln(1-\frac{\gamma}{2}) + (1+\frac{\gamma}{4})\ln(1+\frac{\gamma}{2})\big)} < 8$$

Lastly, by setting T as in the Lemma and plugging into (37), we get,

$$\mathbb{P}_{(x,y)\sim S}\left[F(x,y) \text{ has margin} \leq \gamma/8\right] \leq \frac{k-1}{m\cdot k}, \tag{38}$$

and since the empirical probability must be a multiple of $\frac{1}{m}$, we get that it must in fact be zero. □

B.1 Definitions needed for the following proofs

For the sake of the next proofs, we define the following notions of combined hypotheses. Define $\overline{\mathcal{H}} \subseteq [0,1]^{\mathcal{X}\times\mathcal{Y}}$ be the set of all weighted hypotheses from $\mathcal{H}$, i.e.,

$$\overline{\mathcal{H}} = \Big\{\bar{h} : (x,\ell) \mapsto \sum_{h\in\mathcal{H}} \lambda(h)\cdot \mathbb{1}[h(x)=\ell] \,\Big|\, \lambda \in \Delta_{\mathcal{H}}\Big\}. \tag{39}$$

Observe that taking $\arg\max_\ell \bar{h}(x,\ell)$ corresponds to the plurality-vote prediction (see Equation (1)). Furthermore, observe that there exists a combined hypotheses $\bar{h} \in \overline{\mathcal{H}}$ which corresponds to the weighted average $\sum_{t=1}^{T} \alpha_t \cdot \mathbb{1}[h_t(x)=\ell]$, obtained by the booster, in Algorithm 1 (and by taking its plurality-vote prediction we get the final output predictor $\bar{h}_T(x)$). Assume that $\alpha = (\alpha_1, ..., \alpha_T)$ is the normalized weight vector of the weights obtained by the algorithm. We overload notation by denoting this combined, weighted, hypothesis by $\bar{h}_T(x,\ell)$. Next, we define $\overline{\mathcal{H}}_n \subseteq [0,1]^{\mathcal{X}\times\mathcal{Y}}$ as the set of all <u>unweighted</u> combined hypotheses of at most n elements,

$$\overline{\mathcal{H}}_n = \Big\{\bar{h} : (x,\ell) \mapsto \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}[h_j(x)=\ell] \,\Big|\, h_1, ..., h_n \in \mathcal{H}\Big\}. \tag{40}$$

Note that the same $h \in \mathcal{H}$ may appear multiple times in such a combined hypothesis. The main idea of the proof is to approximate the (weighted) combined hypotheses $\bar{h}_T$, by randomly polling its constituents from its corresponding distribution α. Towards that end, we prove a useful variation of the uniform convergence property applied to the margin of the combined hypotheses. Denote the margin of a combined hypotheses f, for a fixed $(x,y) \in \mathcal{X}\times\mathcal{Y}$, as follows,

$$\sigma(f;x,y) = f(x,y) - \max_{\ell\neq y} f(x,\ell). \tag{41}$$

The σ uniform convergence bound given in Lemma 11 is based on the following combinatorial result.

Lemma 10. *[14] For every hypothesis class $F \subseteq \mathcal{Y}^{\mathcal{X}}$, $|F| \leq |\mathcal{X}|^{d_N(F)}|\mathcal{Y}|^{2d_N(F)}$.*

Lemma 11 (σ uniform convergence)**.** *Let $\mathbf{D}$ be a distribution over $\mathcal{X}\times\mathcal{Y}$ that is γ-realizable for $\mathcal{H}$ and $\gamma > 0$, and let S be a training set of m i.i.d. samples from $\mathbf{D}$. Let $\delta > 0$. Then, with probability $1-\delta$ over the random choice of S, for all $n \geq 1$, $\bar{h} \in \overline{\mathcal{H}}_n$,*

$$\mathbb{P}_{\mathbf{D}}\left[\sigma(\bar{h};x,y) \leq \frac{\gamma}{2}\right] \leq \mathbb{P}_S\left[\sigma(\bar{h};x,y) \leq \frac{\gamma}{2}\right] + \epsilon_n,$$

where $\epsilon_n = \sqrt{32\big(4dn\ln(mk) + \ln(8n(n+1)/\delta)\big)/m}$, and $d = d_N(\mathcal{H})$.

Proof. Let $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$, and for any $\bar{h} \in \overline{\mathcal{H}}_n$, define a *low-margin* subset by,

$$\mathcal{Z}_{\bar{h}} = \Big\{(x,y) \in \mathcal{Z} : \sigma(\bar{h};x,y) \leq \gamma/2\Big\}.$$

Observe that any $\bar{h} \in \overline{\mathcal{H}}_n$ is determined by some $h_1, \ldots, h_n \in \mathcal{H}$, and that the restriction $\bar{h}|_Z$ of $\bar{h}$ to subset Z is determined by the restrictions $h_1|_Z, \ldots, h_n|_Z$. Denote the collection of all low-margin subsets of $\mathcal{Z}$ by, $F_n = \left\{\mathcal{Z}_{\bar{h}} : \bar{h} \in \overline{\mathcal{H}}_n\right\}$. Define the *growth function* of F_n as,

$$\Pi_{F_n}(m) = \sup\left\{|\{Z \cap Z' : Z' \in F_n\}| : Z \in \mathcal{Z}^m\right\}.$$

That is, $\Pi_{F_n}(m)$ is the maximal number of "in-out" behaviors (dichotomies) realizable by low-margin subsets (sets in F_n) on a finite set of m points. Observe that the number of such dichotomies is determined by the σ margin behaviors of all $\bar{h} \in \overline{\mathcal{H}}_n$, which in turn can be captured by 2 values for each $\bar{h}$ per example, (i.e., $\bar{h}(x, y)$ and $\max_{\ell \neq y} \bar{h}(x, \ell)$). Each $\bar{h}$ is in turn determined by some choice of $h_1, ..., h_n \in \mathcal{H}$. Using these observations we bound the growth function,

$$\begin{aligned}
\Pi_{F_n}(m) &= \sup_{Z \in \mathcal{Z}^m} \left|\left\{\langle \mathbf{1}[z_1 \in Z'], ..., \mathbf{1}[z_m \in Z']\rangle : Z' \in F_n\right\}\right| \\
&= \sup_{Z \in \mathcal{Z}^m} \left|\left\{\langle ..., \mathbf{1}[\sigma(\bar{h}; x_i, y_i) \leq \gamma/2], ...\rangle : \bar{h} \in \overline{\mathcal{H}}_n\right\}\right| \\
&\leq \sup_{Z \in \mathcal{Z}^m} \left|\left\{\langle \bar{h}(x_1, y_1), ..., \bar{h}(x_m, y_m)\rangle : \bar{h} \in \overline{\mathcal{H}}_n\right\}\right|^2 \\
&\leq \sup_{Z \in \mathcal{Z}^m} \left|\left\{\langle \mathbf{1}[h(x_1) = y_1], ..., \mathbf{1}[h(x_m) = y_m]\rangle : h \in \mathcal{H}\right\}\right|^{2n} \\
&= \sup_{Z \in \mathcal{Z}^m} |\mathcal{H}|_Z|^{2n} \;\leq\; m^{2dn} k^{4dn},
\end{aligned}$$

where the last inequality follows by Lemma 10. Next, we employ a bound given by a generalized form of uniform-convergence, as in Theorem 2.6, [16], which rather than showing that the training error generalizes to the population loss, shows the characterization of how any property, captured by subsets of the space $\mathcal{Z}$, generalizes. In particular, we get that for $n \geq 1$,

$$\mathbb{P}\left[\exists \mathcal{Z}_{\bar{h}} \in F_n,\ \mathbb{P}_{z \sim \mathcal{D}}[z \in \mathcal{Z}_{\bar{h}}] \geq \mathbb{P}_{z \sim S}[z \in \mathcal{Z}_{\bar{h}}] + \epsilon_n\right] \leq 8\Pi_{F_n}(m) e^{-m\epsilon_n^2/32}.$$

Therefore, by plugging in ϵ_n, we get that the statement in the Lemma occurs with probability at least $1 - \delta/(n(n+1))$, for all $\bar{h} \in \overline{\mathcal{H}}_n$. By the union bound, this same statement holds for all $n \geq 1$ simultaneously with probability at least $1 - \delta$, proving the lemma. □

B.2 Proof of Theorem 6

Proof. In this proof we consider the notation introduced above in Section B.1. First, consider a *fixed* training set S, and let $\bar{h}_T$ be the predictor outputted by Algorithm 1 applied to S. We approximate $\bar{h}_T$ by sampling $\tilde{h}_j \sim \alpha$ (i.e., set $\tilde{h}_j := h_t$ w.p. α_t), i.i.d. for each $j = 1...n$, and set,

$$\tilde{h}(x, \ell) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}[\tilde{h}_j(x) = \ell].$$

Observe that $\tilde{h} \in \overline{\mathcal{H}}_n$ (defined in Equation (40)). In this proof, we will use $\tilde{h}$ to approximate $\bar{h}_T$. Towards that end, we first show that for a fixed $x \in \mathcal{X}$, $\ell \in [k]$, $\gamma' > 0$, $n \geq 1$, we have:

$$\mathbb{P}_{\tilde{h}}\left[\left|\tilde{h}(x, \ell) - \bar{h}_T(x, \ell)\right| \geq \gamma'/2\right] \leq 2e^{-n\gamma'^2/2}. \tag{42}$$

This holds since by the definition of $\tilde{h}$ we have $\mathbb{E}_{\tilde{h}}[\tilde{h}(x, \ell)] = \bar{h}_T(x, \ell)$, and by applying Hoeffding's inequality we obtain Equation (42). Next, we show further that the *margins* of $\bar{h}_T$ and of $\tilde{h}$ are close. Specifically, we consider the σ notation in Equation (41), and show that for any fixed x, y, and for any $\gamma' > 0$, $n \geq 1$, we have,

$$\mathbb{P}_{\tilde{h}}\left[\left|\sigma(\tilde{h}; x, y) - \sigma(\bar{h}_T; x, y)\right| \geq \gamma'/2\right] \tag{43}$$

$$= \mathbb{P}_{\tilde{h}}\left[\left|\left(\tilde{h}(x, y) - \max_{\ell \neq y} \tilde{h}(x, \ell)\right) - \left(\bar{h}_T(x, y) - \max_{\ell \neq y} \bar{h}_T(x, \ell)\right)\right| \geq \gamma'/2\right] \tag{44}$$

$$\leq \mathbb{P}_{\tilde{h}}\left[\left|\tilde{h}(x,y)-\bar{h}_T(x,y)\right|+\max_{\ell\neq y}\left|\tilde{h}(x,\ell)-\bar{h}_T(x,\ell)\right|\geq \gamma'/2\right] \tag{45}$$

$$\leq 2\,\mathbb{P}_{\tilde{h}}\left[\max_{\ell\neq y}\left|\tilde{h}(x,\ell)-\bar{h}_T(x,\ell)\right|\geq \gamma'/4\right] \tag{46}$$

$$\leq 2\sum_{\ell\neq y}\mathbb{P}_{\tilde{h}}\left[\left|\tilde{h}(x,\ell)-\bar{h}_T(x,\ell)\right|\geq \gamma'/4\right]. \tag{47}$$

Then, by combining the above with (42), we get that for any D distribution over $\mathcal{X}\times\mathcal{Y}$, and for any $\gamma'>0$, $n\geq 1$, we have,

$$\mathbb{E}_D\left[\mathbb{P}_{\tilde{h}}\left[\left|\sigma(\tilde{h};x,y)-\sigma(\bar{h}_T;x,y)\right|\geq\gamma'/2\right]\right]\leq \mathbb{E}_D\left[4(k-1)e^{-n\gamma'^2/2}\right]=4(k-1)e^{-n\gamma'^2/2}. \tag{48}$$

Next, we apply the bound in (48) to distribution $\mathbf{D}$, and get that,

$$\mathbb{P}_{\mathbf{D}}\left[\sigma(\bar{h}_T;x,y)\leq 0\right]=\mathbb{P}_{\mathbf{D},\tilde{h}}\left[\sigma(\bar{h}_T;x,y)\leq 0\right] \tag{49}$$

$$\leq \mathbb{P}_{\mathbf{D},\tilde{h}}\left[\sigma(\tilde{h};x,y)\leq\frac{\gamma'}{2}\right]+\mathbb{P}_{\mathbf{D},\tilde{h}}\left[\sigma(\bar{h}_T;x,y)\leq 0\bigwedge\sigma(\tilde{h};x,y)>\frac{\gamma'}{2}\right] \tag{50}$$

$$\leq \mathbb{P}_{\mathbf{D},\tilde{h}}\left[\sigma(\tilde{h};x,y)\leq\frac{\gamma'}{2}\right]+\mathbb{P}_{\mathbf{D},\tilde{h}}\left[\left|\sigma(\tilde{h};x,y)-\sigma(\bar{h}_T;x,y)\right|>\frac{\gamma'}{2}\right] \tag{51}$$

$$\leq \mathbb{P}_{\mathbf{D},\tilde{h}}\left[\sigma(\tilde{h};x,y)\leq\frac{\gamma'}{2}\right]+4(k-1)e^{-n\gamma'^2/2}, \tag{52}$$

where the first inequality follows from the the simple fact that for any two events a and b, $\mathbb{P}[a]=\mathbb{P}[a\wedge b]+\mathbb{P}[a\wedge\neg b]\leq\mathbb{P}[b]+\mathbb{P}[a\wedge\neg b]$. Using a similar derivation, applied to the empirical distribution that is uniform over S, we obtain,

$$\mathbb{P}_{S,\tilde{h}}\left[\sigma(\tilde{h};x,y)\leq\frac{\gamma'}{2}\right]\leq\mathbb{P}_{S,\tilde{h}}\left[\sigma(\bar{h}_T;x,y)\leq\gamma'\right]+\mathbb{P}_{S,\tilde{h}}\left[\sigma(\tilde{h};x,y)\leq\frac{\gamma'}{2}\bigwedge\sigma(\bar{h}_T;x,y)>\gamma'\right] \tag{53}$$

$$\leq \mathbb{P}_{S,\tilde{h}}\left[\sigma(\bar{h}_T;x,y)\leq\gamma'\right]+\mathbb{P}_{S,\tilde{h}}\left[\left|\sigma(\tilde{h};x,y)-\sigma(\bar{h}_T;x,y)\right|>\frac{\gamma'}{2}\right] \tag{54}$$

$$\leq \mathbb{P}_{S,\tilde{h}}\left[\sigma(\bar{h}_T;x,y)\leq\gamma'\right]+4(k-1)e^{-n\gamma'^2/2}. \tag{55}$$

Lastly, by applying Lemma 11 and using the σ uniform-convergence property, we get that with probability at least $1-\delta$ over the random choice of S, for any $n\geq 1$, and $\bar{h}_n\in\overline{\mathcal{H}}_n$,

$$\mathbb{P}_{\mathbf{D}}\left[\sigma(\bar{h}_n;x,y)\leq\frac{\gamma'}{2}\right]\leq\mathbb{P}_S\left[\sigma(\bar{h}_n;x,y)\leq\frac{\gamma'}{2}\right]+\epsilon_n, \tag{56}$$

where $\epsilon_n=O\Big(\sqrt{\frac{dn\log(mk)+\log(1/\delta)}{m}}\Big)$, with $d=d_N(\mathcal{H})$ denoting the Natarajan dimension of $\mathcal{H}$. Observe that by using marginalization, we get that the above bound holds with respect to a random $\tilde{h}$, rather than a fixed $\bar{h}_n$. Then, combining the above Equations (52), then (56) (using marginalization), then (55), we get that with probability $1-\delta$,

$$\mathbb{P}_{\mathbf{D}}\left[\sigma(\bar{h}_T;x,y)\leq 0\right]\leq\mathbb{P}_{\mathbf{D},\tilde{h}}\left[\sigma(\tilde{h};x,y)\leq\frac{\gamma'}{2}\right]+4e^{-n\gamma'^2/2} \tag{57}$$

$$\leq \mathbb{P}_{S,\tilde{h}}\left[\sigma(\tilde{h};x,y)\leq\frac{\gamma'}{2}\right]+4e^{-n\gamma'^2/2}+\epsilon_n \tag{58}$$

$$\leq \mathbb{P}_{S,\tilde{h}}\left[\sigma(\bar{h}_T; x, y) \leq \gamma'\right] + 8e^{-n\gamma'^2/2} + \epsilon_n. \tag{59}$$

By setting $n = \frac{2}{\gamma^2}\ln(\frac{16}{\epsilon})$, we get that $8e^{-n\gamma'^2/2} = \epsilon/2$. Then, by setting m as stated in the Theorem, we get that with probability $1-\delta$, the overall bound is,

$$\mathbb{P}_{\mathbf{D}}\left[\sigma(\bar{h}_T; x, y) \leq 0\right] \leq \mathbb{P}_S\left[\sigma(\bar{h}_T; x, y) \leq \gamma'\right] + \epsilon. \tag{60}$$

Lastly, by Lemma 9, we get that the empirical margin probability is zero, for $\gamma' = \gamma/8$, which completes the proof. □

C Lower Bound: Omitted proofs of Section 5

The proof of Theorem 7 will occupy this section. Although the intuitive idea outlined in Section 5 is simple, there are many subtle but technical details that will need to be worked out to ensure that all of the formal requirements of the learning model are satisfied. We consider the model definition as detailed in Section 5, and describe the construction of the base class $\mathcal{H}$ below.

C.1 The construction of $\mathcal{H}$

Let $\mathcal{X}$ be the finite domain, with $n = \frac{1}{2\gamma}$ data points, and assume that $4n < k$. We fix the constant function $f(x) = 1$ to be the target labelling. Then, the labelled data is $\{(x_1, 1)...(x_n, 1)\}$, and let the target distribution $\mathcal{D}$ be the uniform distribution over it. We note that our construction and the lower bound argument hold even for the case that the boosting algorithm has full access to the data and distribution $\mathcal{D}$. However, recall that it is restricted to only output a plurality-vote of hypotheses $h^1, ...h^T \in \mathcal{H}$ obtained via $\mathcal{W}$. Towards defining the base class $\mathcal{H}$, consider hypotheses of the following form; for all $j \in [n]$ and $\ell \in \{2, ..., k\}$, define

$$h_{j,\ell}(x) = \begin{cases} 1 & \text{if } x = x_j, \\ \ell & \text{otherwise.} \end{cases} \tag{61}$$

We denote the set of all such hypotheses as $\mathcal{H}_{all}$. Next, we describe the desired properties of the constructed base class. For any hypothesis class $\mathcal{H} \subset \mathcal{H}_{all}$ of size $|\mathcal{H}| = 2n-1$, consider the following two conditions:

(I) For each index $j = 1, ..., n-1$ there are exactly 2 distinct labels $\ell_1, \ell_2 \in [k]$ such that $h_{j,\ell_1}, h_{j,\ell_2} \in \mathcal{H}$. For index $j = n$, there exists exactly 1 label ℓ such that, $h_{n,\ell} \in \mathcal{H}$.

(II) For each label $\ell = 2, ..., k$, there is at most 1 index $j \in [n]$ such that $h_{j,\ell}$ in $\mathcal{H}$.

Lemma 12. *Let $\mathcal{H} \subset H_{all}$ be a hypothesis class of size $2n-1$, such that both (I), (II) hold. Then, the distribution $\mathcal{D}$, uniform over $\mathcal{X} \times \{1\}$, is $\gamma/2$-realizable w.r.t. $\mathcal{H}$ (see Definition 1).*

Moreover, if only condition (I) fails such that there are q indices $j \in [n]$ for each of which there is at most 1 hypothesis of the form $h_{j,\ell}$ in $\mathcal{H}$, then any plurality-vote over $\mathcal{H}$ incurs error of at least $2\gamma(q-1) = \frac{q-1}{n}$.

We use the probabilistic method to construct the base hypothesis class $\mathcal{H}$, a subset of H_{all} of size $2n-1$. Specifically, we let the class $\mathcal{H}$ be entirely fixed, apart from a single random hypothesis, $h_{n-1,L}$, where L denotes a random variable drawn uniformly at random from the set of labels $2n, ..., k$. In particular, the class $\mathcal{H} \subset \mathcal{H}_{all}$ is constructed as follows. For each index $j < n-1$, we have exactly 2 fixed hypotheses $h_{j,2j}$ and $h_{j,2j+1}$ in $\mathcal{H}$; for $j = n-1$, we have one fixed hypotheses $h_{n-1,2n-2}$, and the other is the random hypotheses $h_{n-1,L}$. Lastly, for $j = n$ we have a single fixed

hypotheses $h_{n,2n-1}$. The overall class is:

$$\begin{aligned}\mathcal{H} = \big\{ & h_{1,2}, h_{1,3}, \\ & h_{2,4}, h_{2,5}, \\ & h_{3,6}, h_{3,7}, \\ & \quad \vdots \\ & h_{n-3,2n-6}, h_{n-3,2n-5} \\ & h_{n-2,2n-4}, h_{n-2,2n-3} \\ & h_{n-1,2n-2}, h_{n-1,L} \\ & h_{n,2n-1} \big\}.\end{aligned} \tag{62}$$

Observe that the class $\mathcal{H}_{all}$ is an easy-to-learn class, and it be easily verified that it has a constant Natarajan dimension of 2. Since $\mathcal{H} \subset \mathcal{H}_{all}$ for any L, we get that $\mathcal{H}$ has a constant Natarajan dimension as well. By Lemma 12, there exists a plurality-vote over $\mathcal{H}$ that is consistent with the data, with margin $\gamma/2$. That is, the distribution $\mathcal{D}$ which is uniform over $\mathcal{X} \times \{1\}$ is $\gamma/2$-realizable, with respect to $\mathcal{H}$.

We refer to the randomly chosen hypothesis $h_{n-1,L}$ as the *hidden* hypothesis. Note that if the booster cannot "discover" the hidden hypothesis throughout its run (i.e., obtain it via the weak learner), then by Lemma 12, it will fail to achieve a low generalization error (with error of at least 2γ). Next, we show that indeed it will fail to obtain the hidden hypothesis.

C.2 The "finder-chooser" game

Intuitively, applying any boosting algorithm on the random class $\mathcal{H}$ constructed above requires the booster to find the hidden hypothesis, which is a challenging task. This is the crucial observation needed for proving the bound, and is demonstrated in the simpler setting of a "finder-chooser" game, described next. Consider a setting of an iterative game in which a *"chooser"* player picks a "hidden" element L out of a set of elements K, of size $|K| = k$, and a *"finder"* player iteratively attempts to discover it. Specifically, the finder is only required to find any small <u>subset</u> of elements from K that the hidden element L belongs to. Towards that end, in each round the finder queries the chooser, and the chooser replies with a binary response, which indicates if the chooser has indeed picked the hidden element. Concretely, we consider 2 different query-models in which the finder is allowed to interact with the chooser; via weights, or via examples. In either case, the finder is given a binary response in each round, of whether or not the hidden element is in the subset it picked. If the finder failed, it continues to the next round, otherwise it succeeded to find a subset which contains L, and the game ends. It is clear that there is a trade-off between the size of the picked subsets and the number of rounds needed to find the containing subset. The Lemmas below formalize this trade-off in both models, and give bounds on the probability that the finder succeeds, when the chooser is uniform over K.

C.2.1 The "finder-chooser" game: Weighting model

Let $k, T > 0$, and fix a threshold $\theta \in (0, 1)$. The model is defined as follows. First, the element L is sampled uniformly over the set $[k] = \{1, ..., k\}$. Next, for each round $t = 1...T$, let $D_t \in \Delta_k$ be some distribution of $[k]$ determined iteratively as the computation output of a "finder" algorithm F, defined as follows. Formally, the finder F is a fixed and deterministic function of its previous picks $D_{t'}$, and previous binary responses $I_{t'}$, for all $t' < t$, and randomness source $\mathcal{R}$, where $I_t = \mathbf{1}\Big[D_t(L) \geq \theta\Big]$.

Lemma 13 (**Weights**). *For any finder F algorithm in the the weighting model, it holds that,*

$$\mathbb{P}_{L,\mathcal{R}}\left[\exists\, t \leq T,\;\; D_t(\ell) \geq \theta\right] \leq \frac{T}{\theta k}.$$

C.2.2 The "finder-chooser" game: Example model

We now consider a slight modification of the setting defined above. Let $T, m, k > 0$. The model is defined as follows. As in the previous model, the element L is sampled uniformly over the set

$[k] = \{1, ..., k\}$. Next, for each round $t = 1...T$, let $\mathcal{L}_t \subset K$, be a subset of at most m elements, that is determined iteratively via a "finder" algorithm F, defined as follows. Formally, the finder F is a fixed and deterministic function of its previous picks $\mathcal{L}_{t'}$, and previous binary responses $I_{t'}$, for all $t' < t$, and randomness source $\mathcal{R}$, where $I_t = \mathbf{1}\Big[L \in \mathcal{L}_t\Big]$.

Lemma 14 (**Examples**). *For any finder F algorithm in the the example model, it holds that,*

$$\mathbb{P}_{L,\mathcal{R}}\left[\exists\, t \leq T,\ \ L \in \mathcal{L}_t\right] \leq \frac{Tm}{k}.$$

Next, we are ready to prove the main result, stated in Theorem 7.

C.3 Proof of Theorem 7

Proof. Consider the randomly-constructed class $\mathcal{H}$ in Equation (62), and the distribution $\mathcal{D}$ that is uniform over $\mathcal{X} \times \{1\}$, which as shown above is $\gamma/2$-realizable for $\mathcal{H}$ (see Definition 1). We use the probabilistic method to complete the proof, for any boosting algorithm B (see the model definition in subsection 5.1).

There are several sources of randomness that are part of either the learning process or of our construction, namely the class $\mathcal{H}$ (in particular, the label L), the initial training set S, the internal randomness of the booster $\mathcal{R}$.

We will show that with respect to all of the sources of randomness, B's error is probable to be large. That is, we show that,

$$\mathbb{P}_{\mathcal{H},S,\mathcal{R}}\left[\mathbb{P}_{\mathcal{D}}\left[\bar{h}_B(x) \neq y\right] \leq \epsilon^B\right] < 1 - \delta^B. \tag{63}$$

This is sufficient for the proof since it implies, $\mathbb{P}_{\mathcal{H},S,\mathcal{R}}\left[\mathbb{P}_{\mathcal{D}}\left[\bar{h}_B(x) \neq y\right] > \epsilon^B\right] \geq \delta^B$, which is equivalent by marginalization to, $\mathbb{E}_{\mathcal{H}}\Big[\mathbb{P}_{S,\mathcal{R}}\left[\mathbb{P}_{\mathcal{D}}\left[\bar{h}_B(x) \neq y\right] > \epsilon^B \Big| \mathcal{H}\right]\Big] \geq \delta^B$. This in turn implies that there exists a *particular* class $\mathcal{H}$ for which, $\mathbb{P}_{S,\mathcal{R}}\left[\mathbb{P}_{\mathcal{D}}\left[\bar{h}_B(x) \neq y\right] > \epsilon^B \Big| \mathcal{H}\right] \geq \delta^B$, as claimed. Therefore, it is sufficient to prove Equation (63). Next, we give a weak learner $\mathcal{W}$ for both models of Boosting-by-reweighting and Boosting-by-resampling, for which (63) holds.

Boosting-by-reweighting Let the learner $\mathcal{W}$ be defined as follows: for any t, if $\epsilon_t^w > \theta$, and there exists any ϵ_t^w-optimal hypothesis with respect to $\mathcal{D}_t$ that is <u>not</u> the hidden hypothesis, then $\mathcal{W}$ returns it to B. Otherwise, $\mathcal{W}$ returns the hidden hypothesis. Denote $\mathcal{H}_T$ as the set of all hypotheses observed by the booster up to round T. Denote $OPT_t \subset \mathcal{H}$ as the set of all ϵ_t^w-optimal hypotheses with respect to $\mathcal{D}_t$. Then, we upper bound the probability that B succeeds to have low error,

$$\begin{aligned}
\mathbb{P}_{\mathcal{H},S,\mathcal{R}}\left[\mathbb{P}_{\mathcal{D}}\left[\bar{h}_B(x) \neq y\right] \leq \epsilon^B\right] &\leq \mathbb{P}_{L,S,\mathcal{R}}\left[h_{n-1,L} \in \mathcal{H}_T\right] && \text{(by Lemma 12)}\\
&\leq \mathbb{P}_{L,S,\mathcal{R}}\left[\exists\, t \leq T,\ OPT_t = \{h_{n-1,L}\}\right] && \text{(by definition of } \mathcal{W})\\
&\leq \mathbb{P}_{L,S,\mathcal{R}}\left[\exists\, t \leq T,\ D_t(L) > \epsilon_t^w\right] && \text{(by Lemma 15)}\\
&\leq \mathbb{P}_{L,S,\mathcal{R}}\left[\exists\, t \leq T,\ D_t(L) > \theta\right], && \text{(by assumption on } \epsilon_t^w)
\end{aligned}$$

where D_t denotes the marginal distribution over labels of $\mathcal{D}_t$, i.e., $D_t(\ell) = \mathbb{P}_{(x,y)\sim\mathcal{D}_t}[y = \ell]$.

We next describe a reduction to the "finder-chooser" setting defined in Lemma 13. In more detail, we will show how to convert the given boosting algorithm B, to a finder's strategy as defined in C.2.1. Assume we are given a boosting algorithm B as above, and consider an instance of the finder-chooser game which we are going to solve using access to B. Towards this end, define the base class $\mathcal{H}$ as in Equation (62), where the label L is determined by the secret label held by the chooser. Observe that the probability in the last inequality above, is equivalent to the probability that the finder succeeds to "win" the game. The reduction is obtained by substituting the sequence weights generated by the

finder, with the marginal distributions $D_1, ..., D_T$ generated by the booster, where the set K and size k in Lemma 13 corresponds to the set of labels $2n...k$, and size $k-2n-1$.

In particular, whenever the boosting algorithm submits a distribution $\mathcal{D}_t$ to the weak learner, our finder submits the corresponding marginal distribution D_t to the chooser. Observe that the binary responses of the chooser correspond to the binary alternative cases for the booster $I_t = [D_t(L) \geq \theta]$.

Note that for the task of finding the hidden label L (i.e., the probability that the booster finds the hidden hypothesis $h_{n-1,L}$), can by upper bounded by the setting in which it is only answered with binary responses, since no additional information of L is gained by the booster for different responses of non-hidden hypotheses [4]. Then, we get,

$$\begin{aligned}
\mathbb{P}_{\mathcal{H},S,\mathcal{R}}\left[\mathbb{P}_{\mathcal{D}}\left[\bar{h}_B(x) \neq y\right] \leq \epsilon^B\right] &\leq \mathbb{P}_{L,S,\mathcal{R}}\left[\exists\, t \leq T,\ D_t(L) > \theta\right] && \text{(by last inequality)} \\
&\leq T \cdot \frac{1}{\theta} \cdot \frac{1}{k-2n-1} && \text{(by Lemma 13)} \\
&< 1/2 \leq 1 - \delta^B,
\end{aligned}$$

where the last inequality follows by the definition of θ, and by the assumption that $\delta^B \leq 1/2$.

Boosting-by-resampling Let the learner $\mathcal{W}$ be an ERM learner such that for any t, when given $m_t^w < 1/\theta$ labelled examples, all of which are <u>not</u> labelled with the hidden label L, return any optimal hypothesis that minimizes the number of sample errors, other than the hidden hypothesis. Otherwise, return any optimal hypothesis. Note that the sample complexity required for the ERM $\mathcal{W}$ to agnostically-PAC learn $\mathcal{H}$ is at most $m_t^w(\epsilon, \delta) \leq O\left(\frac{\log(k)+\log(1/\delta)}{\epsilon^2}\right)$ (see Equation 7, and recall that $\mathcal{H}$ has a constant dimension). However, below we show that when m_t^w is too small ($m_t^w < 1/\theta \approx O(k/T)$), then the booster B fails. Denote $\mathcal{H}_T$ as the set of all hypotheses observed by the booster up to round T. Let $\mathcal{L}_t$ denote the set of all labels for which a labelled example was is fed to $\mathcal{W}$ at time t. Then, we upper bound the probability that B succeeds to have low error,

$$\begin{aligned}
\mathbb{P}_{\mathcal{H},S,\mathcal{R}}\left[\mathbb{P}_{\mathcal{D}}\left[\bar{h}_B(x) \neq y\right] \leq \epsilon^B\right] &\leq \mathbb{P}_{L,S,\mathcal{R}}\left[h_{n-1,L} \in \mathcal{H}_T\right] && \text{(by Lemma 12)} \\
&\leq \mathbb{P}_{L,S,\mathcal{R}}\left[\exists\, t \leq T,\ L \in \mathcal{L}_t\right]. && \text{(by definition of } \mathcal{W})
\end{aligned}$$

Recall that the "hidden label" $L \sim \mathcal{U}(2n, ..., k)$ is chosen uniformly at random. As in the previous model, we apply the reduction to the "finder-chooser" setting defined in Lemma 14, by replacing the sequence of subsets $\mathcal{L}_1, ..., \mathcal{L}_T$ picked by the booster, with the subset of elements generated by the finder, where the set K and size k in Lemma 14 corresponds to the set of labels $2n...k$, and size $k-2n-1$. Then, we get,

$$\mathbb{P}_{\mathcal{H},S,\mathcal{R}}\left[\mathbb{P}_{\mathcal{D}}\left[\bar{h}_B(x) \neq y\right] \leq \epsilon^B\right] \leq T \cdot m \cdot \frac{1}{k-2n-1} < \frac{T/\theta}{k-2n-1} < 1/2 \leq 1 - \delta^B,$$

where the last inequalities follow by the assumption that $m_t^w < 1/\theta$ for all t, by the definition of θ, and by the assumption that $\delta^B \leq 1/2$.

□

Lemma 15. *Let $\mathcal{H}$ be defined as in Equation* (62)*, for some fixed label $L \in \{2n, ..., k\}$. Let $\mathcal{W}$ be an agnostic PAC learner for $\mathcal{H}$. Let $\mathcal{D}$ be any distribution over $\mathcal{X} \times \mathcal{Y}$, and $\epsilon^w > 0$. Then, if the only ϵ-optimal hypothesis in $\mathcal{H}$ with respect to $\mathcal{D}$ is the hypothesis $h_{n-1,L}$, i.e., if it holds that,*

$$\max_{\substack{j \leq n \\ \ell < 2n}} \mathbb{P}_{\mathcal{D}}[h_{j,\ell}(x) = y] < \max_{h^* \in \mathcal{H}} \mathbb{P}_{\mathcal{D}}[h^*(x) = y] - \epsilon^w,$$

then, for the corresponding marginal distribution over labels D, it holds that,

$$D(L) = \sum_{j=1}^{n} \mathcal{D}(x_j, L) > \epsilon^w.$$

[4] That is, for different $\ell', \ell'' < 2n$, and any t_1, t_0 such that $t_1 > t_0$, it holds that $\mathbb{P}[D_{t_1}(L) \geq \theta | h_{t_0} = h_{j,\ell'}] = \mathbb{P}[D_{t_1}(L) \geq \theta | h_{t_0} = h_{j,\ell''}]$.

Proof. Observe that by the assumption, $h_{n-1,L} = \arg\max_{h^* \in \mathcal{H}} \mathbb{P}_{\mathcal{D}}[h^*(x) = y]$. Then, we have,

$$\begin{aligned}
\mathcal{D}(x_{n-1}, 1) &\leq \max_{j \leq n} \mathcal{D}(x_j, 1) \\
&\leq \max_{\substack{j \leq n \\ \ell < 2n}} \left(\mathcal{D}(x_j, 1) + \sum_{j' \neq j} \mathcal{D}(x_{j'}, \ell) \right) \\
&= \mathbb{P}_{\mathcal{D}}[h_{j,\ell}(x) = y] \\
&< \mathbb{P}_{\mathcal{D}}[h_{n-1,L}(x) = y] - \epsilon^w \\
&= \mathcal{D}(x_{n-1}, 1) + \sum_{j' \neq n-1} \mathcal{D}(x_{j'}, L) - \epsilon^w,
\end{aligned}$$

where the last inequality follows from the assumption. By re-arranging terms we get that,

$$\epsilon^w < \sum_{j' \neq n-1} \mathcal{D}(x_{j'}, L) \leq \sum_{j=1}^{n} \mathcal{D}(x_j, L) = D(L),$$

as claimed.

□

C.4 Proof of Lemma 12

Proof. We define λ a distribution over the $2n-1$ hypotheses of a class $\mathcal{H}$ satisfying the conditions of the lemma, and show that it induces a consistent plurality-vote with margin $\gamma/2$. Denote each hypotheses in $\mathcal{H}$ by an index i, and assume w.l.o.g that hypothesis $i = 2n-1$ corresponds to the single hypotheses $h_{n,\ell}$ as defined in condition (I). Set $\xi = \frac{1}{4n}$. For all $i < 2n-1$, set $\lambda(i) = \frac{1-\xi}{2n-1}$, and set $\lambda(2n-1) = \frac{1}{2n-1} + \frac{2n-2}{2n-1}\xi$. Then, a point x_j, for any $j \leq n-1$, corresponds to exactly 2 hypotheses with indices i_1, i_2 that predict its label correctly, and the corresponding weight of that prediction is $\lambda(i_1) + \lambda(i_2)$, and any incorrect label will have weight of at most $\lambda(2n-1)$. For a point x_n, the corresponding correct and incorrect weights are $\lambda(2n-1)$ and $\lambda(i)$ for some $i < 2n-1$. Overall, the plurality-vote induced by λ has a margin of,

$$\begin{aligned}
\min\{2\lambda(i) - \lambda(2n-1), \lambda(2n-1) - \lambda(i)\} &= \min\left\{\frac{1-2n\xi}{2n-1}, \xi\right\} \\
&= \min\left\{\frac{1}{2} \cdot \frac{1}{2n-1}, \xi\right\} \\
&\geq \frac{1}{4n} = \gamma/2.
\end{aligned}$$

To prove the second statement in the Lemma, assume that for a class $\mathcal{H}$ of size $2n-q$, condition (II) holds, but condition (I) fails, i.e., there are $j_1, ..., j_q \in [n]$ distinct indices such that for each such j, there is exactly 1 hypotheses of the form $h_{j,\ell} \in \mathcal{H}$, for some unique ℓ, denoted ℓ_j. Denote each hypothesis in $\mathcal{H}$ by an index $i = 1, ..., (2n-q)$, and assume w.l.o.g that for all $i \leq q$, the i-th hypothesis h_i corresponds to $h_{j_i,\ell_{j_i}}$. Let λ be any distribution over $\mathcal{H}$.

Then, let $i^* \in [2n-q]$ be the index of a hypothesis in $\mathcal{H}$, such that for all other $i \leq q$, $i \neq i^*$, $\lambda(i^*) > \lambda(i)$ (if the inequality is not strict, plurality-vote breaks ties arbitrarily, pick i^* maximal by it). Since for each x_i with $i \leq q$, the correct label 1 will get $\lambda(i)$ of the votes, whereas some incorrect label corresponding to hypothesis i^* will get $\lambda(i^*)$ of the votes. If $i^* > q$, then for all $i \leq q$, the plurality-vote induced by λ will err on x_i. Otherwise, if $i^* \leq q$, then it will err on exactly $q-1$ such points.

Overall, the plurality-vote will err for at least $q-1$ such i's. Each such error corresponds to a portion of $1/n = 2\gamma$ of the population loss. Therefore, the plurality-vote induced by λ over $\mathcal{H}$ incurs error of at least $2\gamma(q-1)$. □

C.5 Proof of Lemma 13

Proof. Let $\mathcal{I}_t = \mathbf{1}\left[\exists t' \leq t,\ I_{t'} = 1\right]$. Define $\mathcal{I}_0 = 0$. Fix an algorithm F and its randomness, and observe that when simulated with any fixed sequence of binary responses, $I_1^F, ..., I_{t-1}^F$, it outputs

a sequence $v_1, ..., v_t$ that is deterministic. This follows by definition, as the output sequence v_t is a deterministic function of the random bits of F and the binary responses it observes $I_1^F, ..., I_{t-1}^F$, and if those are fixed, then so is the output. next, for any t, simulate F with a response sequence of all zeros, $I_1^F = 0, ..., I_{t-1}^F = 0$, which then outputs v_t, and let $\mathcal{L}_t^{\mathbf{0}} = \{\ell \in [k] \mid D_t(\ell) \geq \theta\}$. First, observe that $|\mathcal{L}_t^{\mathbf{0}}| \leq 1/\theta$, and $|\cup_{t=1}^T \mathcal{L}_t^{\mathbf{0}}| \leq T/\theta$. Moreover, notice that for a fixed F, the set $\cup_{t=1}^T \mathcal{L}_t^{\mathbf{0}}$ is deterministic. Thus,

$$\mathbb{P}_{L,\mathcal{R}}\left[\exists\, t \leq T,\ \ D_t(L) \geq \theta\right] = \mathbb{P}_{L,\mathcal{R}}\left[\bigvee_{t=1}^{T} \mathcal{I}_{t-1} = 0 \wedge I_t = 1\right] \tag{64}$$

$$= \mathbb{E}_{\mathcal{R}}\left[\mathbb{P}_L\left[\bigvee_{t=1}^{T} \mathcal{I}_{t-1} = 0 \wedge I_t = 1 \middle| \mathcal{R}\right]\right] \tag{65}$$

$$= \mathbb{E}_{\mathcal{R}}\left[\mathbb{P}_L\left[\bigvee_{t=1}^{T} L \in \mathcal{L}_t^{\mathbf{0}} \middle| \mathcal{R}\right]\right] \tag{66}$$

$$= \mathbb{E}_{\mathcal{R}}\left[\mathbb{P}_L\left[L \in \bigcup_{t=1}^{T} \mathcal{L}_t^{\mathbf{0}} \middle| \mathcal{R}\right]\right] \tag{67}$$

$$\leq \frac{T}{\theta k}. \tag{68}$$

□

C.6 Proof of Lemma 14

Proof. Let $\mathcal{I}_t = \mathbf{1}\big[\exists t' \leq t,\ I_{t'} = 1\big]$. Define $\mathcal{I}_0 = 0$. Fix $\mathcal{R}$, and for any t, simulate F with a response sequence of all zeros, $I_1 = 0, ..., I_{t-1} = 0$, which then outputs v_t, and let $\mathcal{L}_t^{\mathbf{0}}$ denote the set $\mathcal{L}_t$ obtained by the simulation. note that $|\cup_{t=1}^T \mathcal{L}_t^{\mathbf{0}}| \leq Tm$. Thus,

$$\mathbb{P}_{L,\mathcal{R}}\left[\exists\, t \leq T,\ \ \ell \in \mathcal{L}_t\right] = \mathbb{P}_{L,\mathcal{R}}\left[\bigvee_{t=1}^{T} \mathcal{I}_{t-1} = 0 \wedge I_t = 1\right] \tag{69}$$

$$= \mathbb{E}_{\mathcal{R}}\left[\mathbb{P}_L\left[\bigvee_{t=1}^{T} \mathcal{I}_{t-1} = 0 \wedge I_t = 1 \middle| \mathcal{R}\right]\right] \tag{70}$$

$$= \mathbb{E}_{\mathcal{R}}\left[\mathbb{P}_L\left[\bigvee_{t=1}^{T} \ell \in \mathcal{L}_t^{\mathbf{0}} \middle| \mathcal{R}\right]\right] \tag{71}$$

$$= \mathbb{E}_{\mathcal{R}}\left[\mathbb{P}_L\left[\ell \in \bigcup_{t=1}^{T} \mathcal{L}_t^{\mathbf{0}} \middle| \mathcal{R}\right]\right] \tag{72}$$

$$\leq \frac{Tm}{k}. \tag{73}$$

□