Improved Algorithms for Collaborative PAC Learning

Huy Lê Nguyễn * Lydia Zakynthinou †

Abstract

We study a recent model of collaborative PAC learning where k players with k different tasks collaborate to learn a single classifier that works for all tasks. Previous work showed that when there is a classifier that has very small error on all tasks, there is a collaborative algorithm that finds a single classifier for all tasks and has $O((\ln(k))^2)$ times the worst-case sample complexity for learning a single task. In this work, we design new algorithms for both the realizable and the non-realizable setting, having sample complexity only $O(\ln(k))$ times the worst-case sample complexity for learning a single task. The sample complexity upper bounds of our algorithms match previous lower bounds and in some range of parameters are even better than previous algorithms that are allowed to output different classifiers for different tasks.

1 Introduction

There has been a lot of work in machine learning concerning learning multiple tasks simultaneously, ranging from multi-task learning [3, 4], to domain adaptation [10, 11], to distributed learning [2, 7, 14]. Another area in similar spirit to this work is meta-learning, where one leverages samples from many different tasks to train a single algorithm that adapts well to all tasks (see e.g. [8]).

In this work, we focus on a model of collaborative PAC learning, proposed by [5]. In the classic PAC learning setting introduced by [13], where PAC stands for *probably approximately correct*, the goal is to learn a task by drawing from a distribution of samples. The optimal classifier that achieves the lowest error on the task with respect to the given distribution is assumed to come from a concept class $\mathcal{F}$ of VC dimension d. The VC theorem [1] states that for any instance $m_{\epsilon,\delta} = O\left(\frac{1}{\epsilon}\left(d\ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right)$ labeled samples suffice to learn a classifier that achieves low error with probability at least $1-\delta$, where the error depends on ϵ.

In the collaborative model, there are k players attempting to learn their own tasks, each task involving a different distribution of samples. The goal is to learn a single classifier that also performs well on all the tasks. One example from [5], which motivates this problem, is having k hospitals with different patient demographics which want to predict the overall occurrence of a disease. In this case, it would be more fitting as well as cost efficient to develop and distribute a single classifier to all the hospitals. In addition, the requirement for a single classifier is imperative in settings where there are fairness concerns. For example, consider the case that the goal is to find a classifier that predicts loan defaults for a bank by gathering information from bank stores located in neighborhoods with diverse socioeconomic characteristics. In this setting, the samples provided by each bank store come from different distributions while it is desired to

*College of Computer and Information Science, Northeastern University. hu.nguyen@northeastern.edu. This work was supported by NSF CAREER 1750716.

†College of Computer and Information Science (CCIS), Northeastern University. zakynthinou.l@northeastern.edu. This work was supported by a Graduate Fellowship from CCIS.

guarantee low error rates for all the neighborhoods. Again, in this setting, the bank should employ a single classifier among all the neighborhoods.

If each player were to learn a classifier for their task without collaboration, they would each have to draw a sufficient number of samples from their distribution to train their classifier. Therefore, solving k tasks independently would require $k \cdot m_{\epsilon,\delta}$ samples in the worst case. Thus, we are interested in algorithms that utilize samples from all players and solve all k tasks with sample complexity $o\Big(\frac{k}{\epsilon}\Big(d\ln\Big(\frac{1}{\epsilon}\Big)+\ln\Big(\frac{1}{\delta}\Big)\Big)\Big)$.

Blum et al. [5] give an algorithm with sample complexity $O\Big(\frac{\ln^2(k)}{\epsilon}\Big((d+k)\ln\Big(\frac{1}{\epsilon}\Big)+k\ln\Big(\frac{1}{\delta}\Big)\Big)\Big)$ for the realizable setting, that is, assuming the existence of a single classifier with zero error on all the tasks. They also extend this result by proving that a slightly modified algorithm returns a classifier with error ϵ, under the relaxed assumption that there exists a classifier with error $\epsilon/100$ on all the tasks. In addition, they prove a lower bound showing that there is a concept class with $d=\Theta(k)$ where $\Omega\Big(\frac{k}{\epsilon}\ln\Big(\frac{k}{\delta}\Big)\Big)$ samples are necessary.

In this work, we give two new algorithms based on multiplicative weight updates which have sample complexities $O\Big(\frac{\ln(k)}{\epsilon}\Big(d\ln\Big(\frac{1}{\epsilon}\Big)+k\ln\Big(\frac{k}{\delta}\Big)\Big)\Big)$ and $O\Big(\frac{1}{\epsilon}\ln\Big(\frac{k}{\delta}\Big)\Big(d\ln\Big(\frac{1}{\epsilon}\Big)+k+\ln\Big(\frac{1}{\delta}\Big)\Big)\Big)$ for the realizable setting. Our first algorithm matches the sample complexity of [5] for the variant of the problem in which the algorithm is allowed to return different classifiers to the players and our second algorithm has the sample complexity almost matching the lower bound of [5] when $d=\Theta(k)$ and for typical values of δ. Both are presented in Section 3. Independently of our work, [6] use the multiplicative weight update approach and achieve the same bounds as we do in that section.

Moreover, in Section 4, we extend our results to the non-realizable setting, presenting two algorithms that generalize the algorithms for the realizable setting. These algorithms learn a classifier with error at most $(2+\alpha)\texttt{OPT}+\epsilon$ on all the tasks, where α is set to a constant value, and have sample complexities $O\Big(\frac{\ln(k)}{\alpha^4\epsilon}\Big(d\ln\Big(\frac{1}{\epsilon}\Big)+k\ln\Big(\frac{k}{\delta}\Big)\Big)\Big)$ and $O\Big(\frac{1}{\alpha^4\epsilon}\ln\Big(\frac{k}{\delta}\Big)\Big(d\ln\Big(\frac{1}{\epsilon}\Big)+k\ln\Big(\frac{1}{\alpha}\Big)+\ln\Big(\frac{1}{\delta}\Big)\Big)\Big)$. With constant α, these sample complexities are the same as in the realizable case. Finally, we give two algorithms with *randomized* classifiers whose error probability over the random choice of the example *and the classifier's randomness* is at most $(1+\alpha)\texttt{OPT}+\epsilon$ for all tasks. The sample complexities of these algorithms are $O\Big(\frac{\ln(k)}{\alpha^3\epsilon^2}\Big(d\ln\Big(\frac{1}{\epsilon}\Big)+k\ln\Big(\frac{k}{\delta}\Big)\Big)\Big)$ and $O\Big(\frac{1}{\alpha^3\epsilon^2}\ln\Big(\frac{k}{\delta}\Big)\Big((d+k)\ln\Big(\frac{1}{\epsilon}\Big)+\ln\Big(\frac{1}{\delta}\Big)\Big)\Big)$.

2 Model

In the traditional PAC learning model, there is a space of instances $\mathcal{X}$ and a set $\mathcal{Y}=\{0,1\}$ of possible labels for the elements of $\mathcal{X}$. A classifier $f:\mathcal{X}\to\mathcal{Y}$, which matches each element of $\mathcal{X}$ to a label, is called a *hypothesis*. The error of a hypothesis with respect to a distribution D on $\mathcal{X}\times\mathcal{Y}$ is defined as $\text{err}_D(f)=\Pr_{(x,y)\sim D}[f(x)\neq y]$. Let $\texttt{OPT}=\inf_{f\in\mathcal{F}}\text{err}_D(f)$, where $\mathcal{F}$ is a class of hypotheses. In the realizable setting we assume that there exists a target classifier with zero error, that is, there exists $f^*\in\mathcal{F}$ with $\text{err}_D(f^*)=\texttt{OPT}=0$ for all $i\in[k]$. Given parameters (ϵ,δ), the goal is to learn a classifier that has error at most ϵ, with probability at least $1-\delta$. In the non-realizable setting, the optimal classifier f^* is defined to have $\text{err}_D(f^*)\leq\texttt{OPT}+\varepsilon$ for any $\varepsilon>0$. Given parameters (ϵ,δ) and a new parameter α, which can be considered to be a constant, the goal is to learn a classifier that has error at most $(1+\alpha)\texttt{OPT}+\epsilon$, with probability at least $1-\delta$.

By the VC theorem and its known extension, the desired guarantee can be achieved in both settings by drawing a set of samples of size $m_{\epsilon,\delta}=O\Big(\frac{1}{\epsilon}\Big(d\ln\Big(\frac{1}{\epsilon}\Big)+\ln\Big(\frac{1}{\delta}\Big)\Big)\Big)$ and returning the classifier with minimum error on that sample. More precisely, in the non-realizable setting, $m_{\epsilon,\delta}=\frac{C}{\epsilon\alpha}\Big(d\ln\Big(\frac{1}{\epsilon}\Big)+\ln\Big(\frac{1}{\delta}\Big)\Big)$,

where C is also a constant. We consider an algorithm $\mathcal{O}_{\mathcal{F}}(S)$, where S is a set of samples drawn from an arbitrary distribution D over the domain $\mathcal{X} \times \{0,1\}$, that returns a hypothesis f_0 whose error on the sample set satisfies $\text{err}_S(f_0) \leq \inf_{f \in \mathcal{F}} \text{err}_S(f) + \varepsilon$ for any $\varepsilon > 0$, if such a hypothesis exists. The VC theorem guarantees that if $|S| = m_{\epsilon,\delta}$, then $\text{err}_D(f_0) \leq (1+\alpha)\text{err}_S(f_0) + \epsilon$.

In the collaborative model, there are k players with distributions $D_1, \ldots, D_k$. Similarly, $\texttt{OPT} = \inf_{f \in \mathcal{F}} \max_{i \in [k]} \text{err}_{D_i}(f)$ and the goal is to learn a single good classifier for all distributions. In [5], the authors consider two variants of the model for the realizable setting, the *personalized* and the *centralized*. In the former the algorithm can return a different classifier to each player, while in the latter it must return a single good classifier. For the personalized variant, Blum et al. give an algorithm with almost the same sample complexity as the lower bound they provide. We focus on the more restrictive centralized variant of the model, for which the algorithm that Blum et al. give does not match the lower bound. We note that the algorithms we present are *improper*, meaning that the classifier they return is not necessarily in the concept class $\mathcal{F}$.

3 Sample complexity upper bounds for the realizable setting

In this section, we present two algorithms and prove their sample complexity.

Both algorithms employ multiplicative weight updates, meaning that in each round they find a classifier with low error on the weighted mixture of the distributions and double the weights of the players for whom the classifier did not perform well. In this way, the next sample set drawn will include more samples from these players' distributions so that the next classifier will perform better on them. To identify the players for whom the classifier of the round did not perform well, the algorithms test the classifier on a small number of samples drawn from each player's distribution. If the error of the classifier on the sample is low, then the error on the player's distribution can not be too high and vise versa. In the end, both algorithms return the majority function over all the classifiers of the rounds, that is, for each point $x \in \mathcal{X}$, the label assigned to x is the label that the majority of the classifiers assign to x.

We note that for typical values of δ, Algorithm R2 is better than Algorithm R1. However, Algorithm R1 is always better than the algorithm of [5] for the centralized variant of the problem and matches their number of samples in the personalized variant, so we present both algorithms in this section. In the algorithms of [5], the players are divided into classes based on the number of rounds for which that player's task is not solved with low error. The number of classes could be as large as the number of rounds, which is $\Theta(\log(k))$, and their algorithm uses roughly $m_{\epsilon,\delta}$ samples from each class. On the other hand, Algorithm R1 uses only $m_{\epsilon,\delta}$ samples across all classes and saves a factor of $\Theta(\log(k))$ in the sample complexity. This requires analyzing the change in all classes together as opposed to class by class.

Algorithm R1

Initialize: $\forall i \in [k]\ w_i^{(0)} := 1$; $t := 5\lceil \log(k) \rceil$; $\epsilon' := \epsilon/6$; $\delta' := \delta/(3t)$;
for $r = 1$ **to** t **do**
  $\tilde{D}^{(r-1)} \leftarrow \frac{1}{\Phi^{(r-1)}} \sum_{i=1}^{k} \Big(w_i^{(r-1)} D_i \Big)$, where $\Phi^{(r-1)} = \sum_{i=1}^{k} w_i^{(r-1)}$;
  Draw a sample set $S^{(r)}$ of size $m_{\epsilon'/16,\delta'}$ from $\tilde{D}^{(r-1)}$;
  $f^{(r)} \leftarrow \mathcal{O}_{\mathcal{F}}(S^{(r)})$;
  $G_r \leftarrow \textsc{Test}(f^{(r)}, k, \epsilon', \delta')$;
  Update: $w_i^{(r)} = \begin{cases} 2w_i^{(r-1)}, & \text{if } i \notin G_r \\ w_i^{(r-1)}, & \text{otherwise} \end{cases}$;
end for
return $f_{\text{R1}} = \text{maj}(\{f^{(r)}\}_{r=1}^{t})$

Procedure $\textsc{Test}(f^{(r)}, k, \epsilon', \delta')$
for $i = 1$ **to** k **do**
  Draw a sample set T_i of size $O\Big(\frac{1}{\epsilon'} \ln\Big(\frac{k}{\delta'}\Big)\Big)$ from D_i;
end for
return $\{i \mid \text{err}_{T_i}(f^{(r)}) \leq \frac{3}{4}\epsilon'\}$;

Algorithm R1 runs for $t = \Theta(\log(k))$ rounds and learns a classifier $f^{(r)}$ in each round r that has low error on the weighted mixture of the distributions $\tilde{D}^{(r-1)}$. For each player at least $0.6t$ of the learned classifiers are "good", meaning that they have error at most $\epsilon' = \epsilon/6$ on the player's distribution. Since the algorithm returns the majority of the classifiers, in order for an instance to be mislabeled, at least $0.5t$ of the total number of classifiers should mislabel it. This implies that at least $0.1t$ of the "good" classifiers of that player should mislabel it, which amounts to $1/6$ of the "good" classifiers. Therefore, the error of the majority of the functions for that player is at most $6\epsilon' = \epsilon$.

To identify the players for whom the classifier of the round does not perform well, Algorithm R1 uses a procedure called TEST. This procedure draws $O\Big(\frac{1}{\epsilon'} \ln\Big(\frac{k}{\delta'}\Big)\Big)$ samples from each player's distribution and tests the classifier on these samples. If the error for a player's sample set is at most $3\epsilon'/4$ then TEST concludes that the classifier is good for that player and adds them to the returned set G_r. The samples that the TEST requires from each player suffice to make it capable of distinguishing between the players with error more than ϵ' and players with error at most $\epsilon'/2$ with respect to their distributions, with high probability.

Theorem 1. *For any $\epsilon, \delta \in (0,1)$, and hypothesis class $\mathcal{F}$ of VC dimension d, Algorithm R1 returns a classifier f_{R1} with $err_{D_i}(f_{R1}) \leq \epsilon\ \forall i \in [k]$ with probability at least $1-\delta$ using m samples, where*

$$m = O\Big(\frac{\ln(k)}{\epsilon}\Big(d \ln\Big(\frac{1}{\epsilon}\Big) + k \ln\Big(\frac{k}{\delta}\Big)\Big)\Big).$$

To prove the correctness and sample complexity of Algorithm R1, we need to prove Lemma 1.2, which describes the set G_r that the TEST returns. This proof uses the following multiplicative forms of the Chernoff bounds (proved as in Theorems 4.4 and 4.5 of [12]).

Lemma 1.1 (Chernoff Bounds)**.** *If X is the average of n independent random variables taking values in $\{0,1\}$, then*

$$\Pr[X \leq (1-s)\,\mathbb{E}[X]] \leq \exp\Big(-\frac{s^2\,\mathbb{E}[X]n}{2}\Big), \tag{1}$$

$$\Pr[X \geq (1+s)\,\mathbb{E}[X]] \leq \exp\Big(-\frac{s^2\,\mathbb{E}[X]n}{3}\Big), \tag{2}$$

$$\Pr[X \geq (1+s)\,\mathbb{E}[X]] \leq \exp\Big(-\frac{s\,\mathbb{E}[X]n}{3}\Big), \tag{3}$$

where the latter inequality holds for $s \geq 1$ and the first two hold for $s \in (0,1)$.

Lemma 1.2. TEST$(f^{(r)}, k, \epsilon', \delta')$ *is such that the following two properties hold, each with probability at least $1-\delta'$, for all $i \in [k]$ and for a given round $r \in [t]$.*

(a) If $\text{err}_{D_i}(f^{(r)}) > \epsilon'$, then $i \notin G_r$.

(b) If $\text{err}_{D_i}(f^{(r)}) \leq \frac{\epsilon'}{2}$, then $i \in G_r$.

Proof of Lemma 1.2. For this proof we assume that the number of samples $|T_i|$ for each $i \in [k]$ must be at least $\frac{32}{\epsilon'}\ln\Big(\frac{k}{\delta'}\Big) = O\Big(\frac{1}{\epsilon'}\ln\Big(\frac{k}{\delta'}\Big)\Big)$. For a given round $r \in [t]$:

(a) Assume $\text{err}_{D_i}(f^{(r)}) > \epsilon'$ for some $i \in [k]$. Then

$$\begin{aligned}
&\Pr\Big[i \in G_r\Big] \\
&= \Pr\Big[\text{err}_{T_i}(f^{(r)}) \leq \tfrac{3}{4}\epsilon'\Big] \\
&< \Pr\Big[\text{err}_{T_i}(f^{(r)}) \leq \Big(1-\tfrac{1}{4}\Big)\text{err}_{D_i}(f^{(r)})\Big] \\
&\overset{(1)}{\leq} \exp\Big(-\tfrac{1}{2}\Big(\tfrac{1}{4}\Big)^2\text{err}_{D_i}(f^{(r)})|T_i|\Big) \\
&< \exp\Big(-\tfrac{1}{32}\epsilon'|T_i|\Big) \\
&\leq \exp\Big(-\tfrac{1}{32}\epsilon'\tfrac{32}{\epsilon'}\ln\Big(\tfrac{k}{\delta'}\Big)\Big) \\
&\leq \tfrac{\delta'}{k}.
\end{aligned}$$

Hence, by union bound, $\text{err}_{D_i}(f^{(r)}) > \epsilon' \Rightarrow i \notin G_r$ holds for all $i \in [k]$ with probability at least $1-\delta'$.

(b) Assume $\text{err}_{D_i}(f^{(r)}) \leq \frac{\epsilon'}{2}$ for some $i \in [k]$. We consider two cases and we apply the Chernoff bounds with $s = \frac{\epsilon'}{4\text{err}_{D_i}(f^{(r)})}$. Note that if $\text{err}_{D_i}(f^{(r)}) = 0$ then $\text{err}_{T_i}(f^{(r)}) = 0$ and the property holds. So we only need to consider $\text{err}_{D_i}(f^{(r)}) \neq 0$. First, we need to prove that

$$\begin{aligned}
&\tfrac{3\epsilon'}{4} \geq (1+s)\text{err}_{D_i}(f^{(r)}) \\
&\Leftrightarrow \tfrac{3\epsilon'}{4\text{err}_{D_i}(f^{(r)})} \geq 1 + \tfrac{\epsilon'}{4\text{err}_{D_i}(f^{(r)})} \\
&\Leftrightarrow \tfrac{\epsilon'}{2\text{err}_{D_i}(f^{(r)})} \geq 1,
\end{aligned}$$

which is true.

Case 1. If $\text{err}_{D_i}(f^{(r)}) > \frac{\epsilon'}{4}$, which implies $s < 1$, then

$$\begin{aligned}
&\Pr\Big[i \notin G_r\Big] \\
&= \Pr\Big[\text{err}_{T_i}(f^{(r)}) > \tfrac{3}{4}\epsilon'\Big] \\
&\leq \Pr\Big[\text{err}_{T_i}(f^{(r)}) \geq \Big(1+s\Big)\text{err}_{D_i}(f^{(r)})\Big]
\end{aligned}$$

$$\overset{(2)}{\leq} \exp\Big(-\tfrac{1}{3}\Big(\tfrac{\epsilon'}{4\text{err}_{D_i}(f^{(r)})}\Big)^2 \text{err}_{D_i}(f^{(r)})|T_i|\Big)$$
$$= \exp\Big(-\tfrac{\epsilon'^2}{48\text{err}_{D_i}(f^{(r)})}|T_i|\Big)$$
$$\leq \exp\Big(-\tfrac{1}{48}2\epsilon'\tfrac{24}{\epsilon'}\ln\Big(\tfrac{k}{\delta'}\Big)\Big)$$
$$\leq \tfrac{\delta'}{k}.$$

Case 2. If $\text{err}_{D_i}(f^{(r)}) \leq \frac{\epsilon'}{4}$, which implies $s \geq 1$, then:

$$\Pr\Big[i \notin G_r\Big]$$
$$= \Pr\Big[\text{err}_{T_i}(f^{(r)}) > \tfrac{3}{4}\epsilon'\Big]$$
$$\leq \Pr\Big[\text{err}_{T_i}(f^{(r)}) \geq \Big(1+s\Big)\text{err}_{D_i}(f^{(r)})\Big]$$
$$\overset{(3)}{\leq} \exp\Big(-\tfrac{1}{3}\tfrac{\epsilon'}{4\text{err}_{D_i}(f^{(r)})}\text{err}_{D_i}(f^{(r)})|T_i|\Big)$$
$$= \exp\Big(-\tfrac{\epsilon'}{3}|T_i|\Big)$$
$$\leq \exp\Big(-\tfrac{\epsilon'}{3}\tfrac{3}{\epsilon'}\ln\Big(\tfrac{k}{\delta'}\Big)\Big)$$
$$\leq \tfrac{\delta'}{k}.$$

Hence, by union bound, $\text{err}_{D_i}(f^{(r)}) \leq \frac{\epsilon'}{2} \Rightarrow i \in G_r$ holds for all $i \in [k]$ with probability at least $1-\delta'$.

□

Having proven Lemma 1.2, we can now prove Theorem 1.

Proof of Theorem 1. First, we prove that Algorithm R1 indeed learns a good classifier, meaning that for every player $i \in [k]$ the returned classifier f_{R1} has error $\text{err}_{D_i}(f_{\text{R1}}) \leq \epsilon$ with probability at least $1-\delta$.

Let $e_i^{(r)}$ denote the number of rounds, up until and including round r, that i did not pass the TEST. More formally, $e_i^{(r)} = |\{r' \mid r' \in [r] \text{ and } i \notin G_{r'}\}|$.

Claim 1.1. *With probability at least* $1 - \frac{2\delta}{3}$, $e_i^{(t)} < 0.4t \;\forall i \in [k]$.

From Lemma 1.2(a) and union bound, with probability at least $1 - t\delta' = 1 - \frac{\delta}{3}$, the number of functions that have error more than ϵ' on D_i is the same as the number of rounds that i did not pass the TEST, for all $i \in [k]$. So, if the claim holds, with probability at least $1 - (\frac{2}{3}+\frac{1}{3})\delta = 1-\delta$, less than $0.4t$ functions have error more than ϵ' on D_i, for all $i \in [k]$. Equivalently, with probability at least $1-\delta$, more than $0.6t$ functions have error at most ϵ' on D_i, for all $i \in [k]$. As a result, with probability at least $1-\delta$, the error of the majority of the functions is $\text{err}_{D_i}(f_{\text{R1}}) \leq \frac{0.6}{0.1}\epsilon' = \epsilon$ for all $i \in [k]$.

Let us now prove the claim.

Proof of Claim 1.1. Recall that $\Phi^{(r)} = \sum_{i=1}^k w_i^{(r)}$ is the potential function in round r. By linearity of expectation, the following holds for the error on the mixture of distributions:

$$\begin{aligned} \text{err}_{\tilde{D}^{(r-1)}}(f^{(r)}) &= \tfrac{1}{\Phi^{(r-1)}}\sum_{i=1}^k \Big(w_i^{(r-1)}\text{err}_{D_i}(f^{(r)})\Big) \\ &\geq \tfrac{1}{\Phi^{(r-1)}}\sum_{i\notin G_r} \Big(w_i^{(r-1)}\text{err}_{D_i}(f^{(r)})\Big) \end{aligned} \tag{4}$$

From the VC theorem, it holds that, since $f^{(r)} = \mathcal{O}_{\mathcal{F}}(S^{(r)})$ and $|S^{(r)}| = m_{\epsilon'/16,\delta'}$, with probability at least $1-\delta'$, $\text{err}_{\tilde{D}^{(r-1)}}(f^{(r)}) \leq \frac{\epsilon'}{16}$. From Lemma 1.2($b$), with probability at least $1-\delta'$, $\text{err}_{D_i}(f^{(r)}) \geq \frac{\epsilon'}{2}$ for all $i \notin G_r$. So with probability at least $1-2\delta'$ the two hold simultaneously. Combining these inequalities with (4), we get that with probability at least $1-2\delta'$, $\frac{\epsilon'}{16} \geq \frac{1}{\Phi^{(r-1)}} \sum_{i=1}^{k} \left(w_i^{(r-1)} \frac{\epsilon'}{2} \right) \Leftrightarrow \sum_{i \notin G_r} w_i^{(r-1)} \leq \frac{1}{8}\Phi^{(r-1)}$.

Since only the weights of players $i \notin G_r$ are doubled, it holds that for a given round r

$$\Phi^{(r)} \leq \Phi^{(r-1)} + \sum_{i \notin G_r} w_i^{(r-1)} \leq \frac{9}{8}\Phi^{(r-1)}.$$

Therefore with probability at least $1 - 2t\delta' = 1 - \frac{2\delta}{3}$, the inequality holds for all rounds, by union bound. By induction:

$$\Phi^{(t)} \leq \left(\frac{9}{8}\right)^t \Phi^{(0)} = \left(\frac{9}{8}\right)^t k$$

Also, for every $i \in [k]$ it holds that $w_i^{(t)} = 2^{e_i^{(t)}}$, as each weight is only doubled every time i does not pass the TEST. Since the potential function is the sum of all weights, the following inequality is true.

$w_i^{(t)} \leq \Phi^{(t)}$

$\Rightarrow 2^{e_i^{(t)}} \leq \left(\frac{9}{8}\right)^t k$

$\Rightarrow e_i^{(t)} \leq t \log\left(\frac{9}{8}\right) + \log(k)$

$\Rightarrow e_i^{(t)} \leq 0.17t + 0.2t < 0.4t$

So with probability at least $1 - \frac{2\delta}{3}$, $e_i^{(t)} < 0.4t \ \forall i \in [k]$. ■

As for the total number of samples, it is the sum of TEST's samples and the $m_{\epsilon'/16,\delta'}$ samples for each round. Since TEST is called $t = 5\lceil \log(k) \rceil$ times and each time requests $O\left(\frac{1}{\epsilon'} \ln\left(\frac{k}{\delta'}\right)\right)$ samples from each of the k players, the total number of samples that it requests is $O\left(\log(k)\frac{k}{\epsilon'} \ln\left(\frac{k}{\delta'}\right)\right)$. Substituting $\epsilon' = \epsilon/6$ and $\delta' = \delta/(3t) = \delta/(15\lceil \log(k) \rceil)$, this yields

$$O\left(\frac{\log(k)}{\epsilon} k \ln\left(\frac{k\log(k)}{\delta}\right)\right) = O\left(\frac{\log(k)}{\epsilon} k \ln\left(\frac{k}{\delta}\right)\right)$$

samples in total.

In addition, the sum of the $m_{\epsilon'/16,\delta'}$ samples drawn in each round to learn the classifier for the mixture for $t = 5\lceil \log(k) \rceil$ rounds is $O\left(\frac{\log(k)}{\epsilon'}\left(d \ln\left(\frac{1}{\epsilon'}\right) + \ln\left(\frac{1}{\delta'}\right)\right)\right)$. Again, substituting ϵ' and δ', we get:

$$O\left(\frac{\log(k)}{\epsilon}\left(d \ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{\log(k)}{\delta}\right)\right)\right)$$

samples in total.

Hence, the overall bound is:

$$O\left(\frac{\log(k)}{\epsilon}\left(d \ln\left(\frac{1}{\epsilon}\right) + k \ln\left(\frac{k}{\delta}\right)\right)\right)$$

□

Algorithm R1 is the natural boosting alternative to the algorithm of [5] for the centralized variant of the model. Although it is discussed in [5] and mentioned to have the same sample complexity as their algorithm, it turns out that it is more efficient. Its sample complexity is slightly better (or the same, depending on the parameter regime) compared to the one of the algorithm for the personalized setting presented in [5], which is $O\Big(\frac{\log(k)}{\epsilon}\Big((d+k)\ln\Big(\frac{1}{\epsilon}\Big)+k\ln\Big(\frac{k}{\delta}\Big)\Big)\Big)$.

However, in the setting of the lower bound in [5] where $k=\Theta(d)$, there is a gap of $\log(k)$ multiplicatively between the sample complexity of Algorithm R1 and the lower bound. This difference stems from the fact that in every round, the algorithm uses roughly $\Theta(k)$ samples to find a classifier but roughly $\Theta(k\log(k))$ samples to test the classifier for k tasks. Motivated by this discrepancy, we develop Algorithm R2, which is similar to Algorithm R1 but uses fewer samples to test the performance of each classifier on the players' distributions. To achieve high success probability, Algorithm R2 uses a higher number of rounds.

Algorithm R2

Initialize: $\forall i\in[k]\ w_i^{(0)}:=1$; $t:=150\Big\lceil\log\Big(\frac{k}{\delta}\Big)\Big\rceil$; $\epsilon':=\epsilon/6$; $\delta':=\delta/(4t)$;
for $r=1$ **to** t **do**
  $\tilde{D}^{(r-1)}\leftarrow\frac{1}{\Phi^{(r-1)}}\sum_{i=1}^{k}\Big(w_i^{(r-1)}D_i\Big)$, where $\Phi^{(r-1)}=\sum_{i=1}^{k}w_i^{(r-1)}$;
  Draw a sample set $S^{(r)}$ of size $m_{\epsilon'/16,\delta'}$ from $\tilde{D}^{(r-1)}$;
  $f^{(r)}\leftarrow\mathcal{O}_{\mathcal{F}}(S^{(r)})$;
  $G_r\leftarrow$ FASTTEST$(f^{(r)},k,\epsilon',\delta')$;
  Update: $w_i^{(r)}=\begin{cases}2w_i^{(r-1)}, & \text{if } i\notin G_r\\ w_i^{(r-1)}, & \text{otherwise}\end{cases}$;
end for
return $f_{\text{R2}}=\text{maj}(\{f^{(r)}\}_{r=1}^{t})$;

Procedure FASTTEST$(f^{(r)},k,\epsilon',\delta')$
for $i=1$ **to** k **do**
  Draw a sample set T_i of size $O\Big(\frac{1}{\epsilon'}\Big)$ from D_i;
end for
return $\{i\mid \text{err}_{T_i}(f^{(r)})\le\frac{3}{4}\epsilon'\}$;

More specifically, Algorithm R2 runs for $t=150\lceil\log(\frac{k}{\delta})\rceil$ rounds. In addition, the test it uses to identify the players for whom the classifier of the round does not perform well requires $O\Big(\frac{1}{\epsilon'}\Big)$ samples from each player. This helps us save one logarithmic factor in the second term of the sample complexity of Algorithm R1. We call this new test FASTTEST. The fact that FASTTEST uses less samples causes it to be less successful at distinguishing the players for whom the classifier was "good" from the players for whom it was not, meaning that it has constant probability of making a mistake for a given player at a given round. There are two types of mistakes that FASTTEST can make: to return $i\notin G_r$ and double the weight of i when the classifier is good for i's distribution and to return $i\in G_r$ and not double the weight of i when the classifier is not good.

Theorem 2. *For any $\epsilon,\delta\in(0,1)$, and hypothesis class $\mathcal{F}$ of VC dimension d, Algorithm R2 returns a classifier f_{R2} with $err_{D_i}(f_{R2})\le\epsilon\ \forall i\in[k]$ with probability at least $1-\delta$ using m samples, where*

$$m=O\Big(\frac{1}{\epsilon}\ln\Big(\frac{k}{\delta}\Big)\Big(d\ln\Big(\frac{1}{\epsilon}\Big)+k+\ln\Big(\frac{1}{\delta}\Big)\Big)\Big).$$

To prove the correctness and sample complexity of Algorithm R2, we need Lemma 2.1, which describes the set G_r that the FASTTEST returns and is proven similarly to Lemma 1.2.

Lemma 2.1. FASTTEST$(f^{(r)}, k, \epsilon', \delta')$ *is such that the following two properties hold, each with probability at least* 0.99*, for given round* $r \in [t]$ *and player* $i \in [k]$.

(a) If $err_{D_i}(f^{(r)}) > \epsilon'$*, then* $i \notin G_r$.

(b) If $err_{D_i}(f^{(r)}) \leq \frac{\epsilon'}{2}$*, then* $i \in G_r$.

Proof of Lemma 2.1. For this proof, we assume that the number of samples $|T_i|$ for each $i \in [k]$ must be at least $\frac{148}{\epsilon'} = O\Big(\frac{1}{\epsilon'}\Big)$. For given $r \in [t]$ and $i \in [k]$:

(a) Assume $\text{err}_{D_i}(f^{(r)}) > \epsilon'$. Then

$$\begin{aligned}
&\Pr\Big[i \in G_r\Big] \\
&= \Pr\Big[\text{err}_{T_i}(f^{(r)}) \leq \tfrac{3}{4}\epsilon'\Big] \\
&< \Pr\Big[\text{err}_{T_i}(f^{(r)}) \leq \Big(1 - \tfrac{1}{4}\Big)\text{err}_{D_i}(f^{(r)})\Big] \\
&\overset{(1)}{\leq} \exp\Big(-\tfrac{1}{2}\Big(\tfrac{1}{4}\Big)^2 \text{err}_{D_i}(f^{(r)})|T_i|\Big) \\
&< \exp\Big(-\tfrac{1}{32}\epsilon'|T_i|\Big) \\
&\leq \exp\Big(-\tfrac{1}{32}\epsilon'\tfrac{148}{\epsilon'}\Big) \\
&< 0.01.
\end{aligned}$$

Hence, $\text{err}_{D_i}(f^{(r)}) > \epsilon' \Rightarrow i \notin G_r$ holds with probability at least 0.99.

(b) Assume $\text{err}_{D_i}(f^{(r)}) \leq \frac{\epsilon'}{2}$. We consider two cases and we apply the Chernoff bounds with $s = \frac{\epsilon'}{4\text{err}_{D_i}(f^{(r)})}$. Note that if $\text{err}_{D_i}(f^{(r)}) = 0$ then $\text{err}_{T_i}(f^{(r)}) = 0$ and the property holds. So we only need to consider $\text{err}_{D_i}(f^{(r)}) \neq 0$. First, we need to prove that

$$\begin{aligned}
&\tfrac{3\epsilon'}{4} \geq (1+s)\text{err}_{D_i}(f^{(r)}) \\
&\Leftrightarrow \tfrac{3\epsilon'}{4\text{err}_{D_i}(f^{(r)})} \geq 1 + \tfrac{\epsilon'}{4\text{err}_{D_i}(f^{(r)})} \\
&\Leftrightarrow \tfrac{\epsilon'}{2\text{err}_{D_i}(f^{(r)})} \geq 1,
\end{aligned}$$

which is true.

Case 1. If $\text{err}_{D_i}(f^{(r)}) > \frac{\epsilon'}{4}$, which implies $s < 1$, then

$$\begin{aligned}
&\Pr\Big[i \notin G_r\Big] \\
&= \Pr\Big[\text{err}_{T_i}(f^{(r)}) > \tfrac{3}{4}\epsilon'\Big] \\
&\leq \Pr\Big[\text{err}_{T_i}(f^{(r)}) \geq \Big(1+s\Big)\text{err}_{D_i}(f^{(r)})\Big] \\
&\overset{(2)}{\leq} \exp\Big(-\tfrac{1}{3}\Big(\tfrac{\epsilon'}{4\text{err}_{D_i}(f^{(r)})}\Big)^2 \text{err}_{D_i}(f^{(r)})|T_i|\Big) \\
&= \exp\Big(-\tfrac{\epsilon'^2}{48\text{err}_{D_i}(f^{(r)})}|T_i|\Big) \\
&\leq \exp\Big(-\tfrac{1}{48}2\epsilon'\tfrac{148}{\epsilon'}\Big) \\
&< 0.01.
\end{aligned}$$

Case 2. If $\text{err}_{D_i}(f^{(r)}) \leq \frac{\epsilon'}{4}$, which implies $s \geq 1$, then

$$\begin{aligned}
&\Pr\Big[i \notin G_r\Big] \\
&= \Pr\Big[\text{err}_{T_i}(f^{(r)}) > \tfrac{3}{4}\epsilon'\Big] \\
&\leq \Pr\Big[\text{err}_{T_i}(f^{(r)}) \geq \Big(1+s\Big)\text{err}_{D_i}(f^{(r)})\Big] \\
&\overset{(3)}{\leq} \exp\Big(-\tfrac{1}{3}\tfrac{\epsilon'}{4\text{err}_{D_i}(f^{(r)})}\text{err}_{D_i}(f^{(r)})|T_i|\Big) \\
&= \exp\Big(-\tfrac{\epsilon'}{12}|T_i|\Big) \\
&\leq \exp\Big(-\tfrac{\epsilon'}{12}\tfrac{148}{\epsilon'}\Big) \\
&< 0.01.
\end{aligned}$$

Hence, $\text{err}_{D_i}(f^{(r)}) \leq \frac{\epsilon'}{2} \Rightarrow i \in G_r$ holds with probability at least 0.99.

□

Proof of Theorem 2. First, we prove that Algorithm R2 indeed learns a good classifier, meaning that, with probability at least $1-\delta$, for every player $i \in [k]$ the returned classifier f_{R2} has error $\text{err}_{D_i}(f_{\text{R2}}) \leq \epsilon$. Let $e_i^{(t)}$ be the number of rounds for which the classifier's error on D_i was more than ϵ', i.e. $e_i^{(t)} = |\{r \mid r \in [t] \text{ and } \text{err}_{D_i}(f^{(r)}) > \epsilon'\}|$.

Claim 2.1. *With probability at least* $1-\delta$, $e_i^{(t)} < 0.4t \ \forall i \in [k]$.

If the claim holds, then with probability at least $1-\delta$, less than $0.4t$ functions have error more than ϵ' on $D_i, \forall i \in [k]$. Therefore, with probability at least $1-\delta$, $\text{err}_{D_i}(f_{\text{R2}}) \leq \frac{0.6}{0.1}\epsilon' \leq \epsilon$ for every $i \in [k]$.

Proof of Claim 2.1. Let us denote by $I^{(r)}$ the set of players having $\text{err}_{D_i}(f^{(r)}) > \frac{\epsilon'}{2}$ in round r, i.e., $I^{(r)} = \{i \in [k] \mid \text{err}_{D_i}(f^{(r)}) > \frac{\epsilon'}{2}\}$. We condition on the randomness in the first $r-1$ rounds and compute $\mathbb{E}[\Phi^{(r)} \mid \Phi^{(r-1)}]$. By linearity of expectation, the following hold for round r:

$$\text{err}_{\tilde{D}^{(r-1)}}(f^{(r)}) = \frac{1}{\Phi^{(r-1)}}\sum_{i=1}^{k}\Big(w_i^{(r-1)}\text{err}_{D_i}(f^{(r)})\Big) \geq \frac{1}{\Phi^{(r-1)}}\sum_{i \in I^{(r)}\setminus G_r}\Big(w_i^{(r-1)}\text{err}_{D_i}(f^{(r)})\Big) \tag{5}$$

By the definition of $I^{(r)}$, $\text{err}_{D_i}(f^{(r)}) > \frac{\epsilon'}{2}$ for $i \in I^{(r)}$. From the VC theorem, with probability at least $1-\delta'$, $\text{err}_{\tilde{D}^{(r-1)}}(f^{(r)}) \leq \frac{\epsilon'}{16}$. Using these two bounds and inequality (5), it follows that with probability at least $1-\delta'$,

$$\sum_{i \in I^{(r)}\setminus G_r} w_i^{(r-1)} \leq \frac{1}{8}\Phi^{(r-1)}. \tag{6}$$

For the rest of the analysis, we will condition our probability space to the event that inequality (6) holds for all t rounds. By the union bound, this event happens with probability $1 - t\delta' = 1 - \delta/4$.

Consider the set of players $i \notin I^{(r)} \cup G_r$. These are the players for whom the classifier of the round performed well but FASTTEST made a mistake and did not include them in the set G_r. By linearity of expectation:

$$\begin{aligned}
\mathbb{E}[\sum_{i \notin G_r} w_i^{(r-1)} \mid \Phi^{(r-1)}] &= \mathbb{E}\left[\sum_{i \in I^{(r)}\setminus G_r} w_i^{(r-1)} + \sum_{i \notin I^{(r)} \cup G_r} w_i^{(r-1)} \,\middle|\, \Phi^{(r-1)}\right] \\
&\overset{\text{(6), Lemma 2.1(b)}}{\leq} (0.125 + 0.01)\Phi^{(r-1)}
\end{aligned} \tag{7}$$

Thus, the expected value of the potential function in round r conditioned on its value in the previous round is bounded by

$$\mathbb{E}[\Phi^{(r)} \mid \Phi^{(r-1)}] = \mathbb{E}\left[\sum_{i=1}^{k} w_i^{(r-1)} + \sum_{i \notin G_r} w_i^{(r-1)} \middle| \Phi^{(r-1)}\right] \overset{(7)}{\leq} 1.135\Phi^{(r-1)}.$$

By the definition of the expected value, this implies that $\mathbb{E}[\Phi^{(r)}] \leq 1.135\,\mathbb{E}[\Phi^{(r-1)}]$. Conditioned on the fact that inequality (6) holds for all rounds, which is true with probability at least $1 - \frac{\delta}{4}$, we can conclude that $\mathbb{E}[\Phi^{(t)}] \leq k(1.135)^t$, by induction. Using Markov's inequality we can state that $\Pr\left[\Phi^{(t)} \geq \frac{\mathbb{E}[\Phi^{(t)}]}{\delta/2}\right] \leq \delta/2$. It follows that with probability at least $1 - \frac{\delta}{4} - \frac{\delta}{2} = 1 - \frac{3\delta}{4}$

$$\Phi^{(t)} \leq \frac{2k(1.135)^t}{\delta}. \tag{8}$$

We now need a lower bound for $w_i^{(t)}$. Let $m_i^{(r)}$ denote the number of rounds r', up until and including round r, for which the procedure FASTTEST made a mistake and returned $i \in G_{r'}$ although $\text{err}_{D_i}(f^{(r')}) > \epsilon'$. From Lemma 2.1($a$), it follows that $\mathbb{E}[m_i^{(r)} - m_i^{(r-1)}] \leq 0.01$ so for $M_i^{(r)} = m_i^{(r)} - 0.01r$ it holds that $\mathbb{E}[M_i^{(r)} \mid M_i^{(r-1)}] \leq M_i^{(r-1)}$. Therefore, the sequence $\{M_i^{(r)}\}_{r=0}^{t}$ is a super-martingale. In addition to this, since we can make at most one mistake in each round, it holds that $M_i^{(r)} - M_i^{(r-1)} < 1$. Using the Azuma-Hoeffding inequality with $M_i^{(0)} = m_i^{(0)} - 0.01 \cdot 0 = 0$ and the fact that $t \geq 150$ we calculate that

$$\Pr\left[m_i^{(t)} \geq 0.18t\right] \leq \exp\Big(-\frac{(0.17t)^2}{2t}\Big) \leq \frac{\delta}{4k}.$$

By union bound, $m_i^{(t)} < 0.18t$ holds $\forall i \in [k]$ with probability at least $1 - \frac{\delta}{4}$.

The number of times a weight is doubled throughout the algorithm is $\log(w_i^{(t)})$ and it is at least the number of times the error of the classifier was more than ϵ' minus the number of times the error was more than ϵ' but the FASTTEST made a mistake, which is exactly $e_i^{(t)} - m_i^{(t)}$. So $w_i^{(t)} \geq 2^{e_i^{(t)} - m_i^{(t)}} > 2^{e_i^{(t)} - 0.18t}$ holds for all $i \in [k]$ with probability at least $1 - \frac{\delta}{4}$. Combining this with the bound from inequality (8) we have that with probability at least $1 - \delta$:

$w_i^{(t)} \leq \Phi^{(t)} \Rightarrow 2^{e_i^{(t)} - 0.18t} < \frac{2k(1.135)^t}{\delta} \Rightarrow e_i^{(t)} - 0.18t < 1 + \log\left(\frac{k}{\delta}\right) + t\log(1.135)$

$\Rightarrow e_i^{(t)} < 0.18t + \frac{1}{150}t + \frac{1}{150}t + 0.183t < 0.4t$ ■

It remains to bound the number of samples. FASTTEST is called $t = 150\lceil\log(\frac{k}{\delta})\rceil$ times, so it requires $O\Big(\log\left(\frac{k}{\delta}\right)\frac{k}{\epsilon'}\Big) = O\Big(\frac{k}{\epsilon}\log\left(\frac{k}{\delta}\right)\Big)$ samples in total. The number of samples required to learn each round's classifier is $m_{\epsilon'/16,\delta'}$, so for all rounds there are required $O\Big(\log\left(\frac{k}{\delta}\right)\frac{1}{\epsilon'}\Big(d\ln\left(\frac{1}{\epsilon'}\right) + \ln\left(\frac{1}{\delta'}\right)\Big)\Big)$ samples. Substituting $\epsilon' = \epsilon/6$ and $\delta' = \delta/(4t) = \delta/\left(600\left\lceil\log\left(\frac{k}{\delta}\right)\right\rceil\right)$ we get $O\Big(\frac{1}{\epsilon}\log\left(\frac{k}{\delta}\right)\Big(d\ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{\log(k)}{\delta}\right)\Big)\Big)$ samples in total. From the addition of the two bounds above, the overall sample complexity bound is:

$$O\Big(\frac{1}{\epsilon}\ln\left(\frac{k}{\delta}\right)\Big(d\ln\left(\frac{1}{\epsilon}\right) + k + \ln\left(\frac{1}{\delta}\right)\Big)\Big)$$

□

4 Sample complexity upper bounds for the non-realizable setting

We design Algorithms NR1 and NR2 for the non-realizable setting, which generalize the results of Algorithms R1 and R2, respectively.

Theorem 3. *For any $\epsilon, \delta \in (0,1)$, $7\epsilon/6 < \alpha < 1$, and hypothesis class $\mathcal{F}$ of VC dimension d, Algorithm NR1 returns a classifier f_{NR1} such that $err_{D_i}(f_{NR1}) \leq (2+\alpha)OPT + \epsilon$ holds for all $i \in [k]$ with probability $1-\delta$ using m samples, where*

$$m = O\Big(\frac{\ln(k)}{\alpha^4 \epsilon}\Big(d \ln\Big(\frac{1}{\epsilon}\Big) + k \ln\Big(\frac{k}{\delta}\Big)\Big)\Big).$$

Theorem 4. *For any $\epsilon, \delta \in (0,1)$, $5\epsilon/4 < \alpha < 1$, and hypothesis class $\mathcal{F}$ of VC dimension d, Algorithm NR2 returns a classifier f_{NR2} such that $err_{D_i}(f_{NR2}) \leq (2+\alpha)OPT + \epsilon$ holds for all $i \in [k]$ with probability $1-\delta$ using m samples, where*

$$m = O\Big(\frac{1}{\alpha^4 \epsilon} \ln\Big(\frac{k}{\delta}\Big)\Big(d \ln\Big(\frac{1}{\epsilon}\Big) + k \ln\Big(\frac{1}{\alpha}\Big) + \ln\Big(\frac{1}{\delta}\Big)\Big)\Big).$$

Their main modification compared to the algorithms in the previous section is that these algorithms use a smoother update rule. Algorithms NR1 and NR2 are the following.

Algorithm NR1

Initialization: $\forall i \in [k]\ w_i^{(0)} := 1$; $\alpha' := \alpha/35$; $t := 2\lceil \ln(k)/\alpha'^3 \rceil$; $\epsilon' := \epsilon/60$; $\delta' := \delta/(4t)$;
for $r = 1, \ldots, t$ **do**
  $\tilde{D}^{(r-1)} \leftarrow \frac{1}{\Phi^{(r-1)}} \sum_{i=1}^k \Big(w_i^{(r-1)} D_i\Big)$, where $\Phi^{(r-1)} := \sum_{i=1}^k w_i^{(r-1)}$;
  Draw a sample set $S^{(r)}$ of size $O\Big(\frac{1}{\alpha'\epsilon'}\Big(d \ln\Big(\frac{1}{\epsilon'}\Big) + \ln\Big(\frac{1}{\delta'}\Big)\Big)\Big)$ from $\tilde{D}^{(r-1)}$;
  $f^{(r)} \leftarrow \mathcal{O}_{\mathcal{F}}(S^{(r)})$;
  for $i = 1, \ldots, k$ **do**
    Draw a sample set T_i of size $O\Big(\frac{1}{\alpha'\epsilon'} \ln\Big(\frac{k}{\delta'}\Big)\Big)$ from D_i;
    $s_i^{(r)} \leftarrow \min\left(\frac{\mathrm{err}_{T_i}(f^{(r)})\alpha'^2}{(1+3\alpha')\mathrm{err}_{S^{(r)}}(f^{(r)})+3\epsilon'}, \alpha'\right)$
    Update: $w_i^{(r)} \leftarrow w_i^{(r-1)}(1+s_i^{(r)})$
  end for
end for

return $f_{\mathrm{NR1}} = \mathrm{maj}(\{f^{(r)}\}_{r=1}^t)$;

Algorithm NR2

Initialization: $\forall i \in [k]\ w_i^{(0)} := 1$; $\alpha' := \alpha/40$; $t := 2\lceil \ln(4k/\delta)/\alpha'^3 \rceil$; $\epsilon' := \epsilon/64$; $\delta' := \delta/(4t)$;
for $r = 1, \ldots, t$ **do**
  $\tilde{D}^{(r-1)} \leftarrow \frac{1}{\Phi^{(r-1)}} \sum_{i=1}^{k} \left(w_i^{(r-1)} D_i \right)$, where $\Phi^{(r-1)} := \sum_{i=1}^{k} w_i^{(r-1)}$;
  Draw a sample set $S^{(r)}$ of size $O\left(\frac{1}{\alpha'\epsilon'}\left(d \ln\left(\frac{1}{\epsilon'}\right) + \ln\left(\frac{1}{\delta'}\right)\right)\right)$ from $\tilde{D}^{(r-1)}$;
  $f^{(r)} \leftarrow \mathcal{O}_{\mathcal{F}}(S^{(r)})$;
  for $i = 1, \ldots, k$ **do**
    Draw a sample set T_i of size $O\left(\frac{1}{\alpha'\epsilon'} \ln\left(\frac{1}{\alpha'}\right)\right)$ from D_i;
    $s_i^{(r)} \leftarrow \min\left(\frac{\mathrm{err}_{T_i}(f^{(r)})\alpha'^2}{(1+3\alpha')\mathrm{err}_{S^{(r)}}(f^{(r)})+3\epsilon'}, \alpha' \right)$
    Update: $w_i^{(r)} \leftarrow w_i^{(r-1)}(1 + s_i^{(r)})$
  end for
end for

return $f_{\text{NR2}} = \text{maj}(\{f^{(r)}\}_{r=1}^{t})$;

The algorithms of this section share many useful properties and the proofs of their corresponding theorems follow similar steps. We will first prove some of these shared properties.

Corollary (of Lemma 1.1)**.** *If X is the average of n independent random variables taking values in $\{0, 1\}$, then:*

$$\Pr[X \leq (1-\alpha)\,\mathbb{E}[X] - \epsilon] \leq \exp(-\alpha\epsilon n)\ \forall \alpha, \epsilon \in (0,1) \tag{9}$$

$$\Pr[X \geq (1+\alpha)\,\mathbb{E}[X] + \epsilon] \leq \exp\Big(-\frac{\alpha\epsilon n}{3}\Big)\ \forall \alpha, \epsilon \in (0,1) \tag{10}$$

Proof. We first prove inequality (9). Note that if $\mathbb{E}[X] \leq \epsilon$ then the inequality is trivially true so we only need to consider $\mathbb{E}[X] > \epsilon$. Let $s = \alpha + \frac{\epsilon}{\mathbb{E}[X]}$. Notice that $s^2 \geq \frac{2\alpha\epsilon}{\mathbb{E}[X]}$. Thus, by inequality (1),

$$\Pr[X \leq (1-\alpha)\,\mathbb{E}[X] - \epsilon] \leq \exp(-s^2\,\mathbb{E}[X]n/2) \leq \exp(-\alpha\epsilon n).$$

Next we prove inequality (10). Again let $s = \alpha + \frac{\epsilon}{\mathbb{E}[X]}$. If $s < 1$ then by inequality (2,

$$\Pr[X \geq (1+\alpha)\,\mathbb{E}[X] + \epsilon] \leq \exp(-s^2\,\mathbb{E}[X]n/3) \leq \exp(-2\alpha\epsilon n/3).$$

If $s \geq 1$ then by inequality (3),

$$\Pr[X \geq (1+\alpha)\,\mathbb{E}[X] + \epsilon] \leq \exp(-s\,\mathbb{E}[X]n/3) \leq \exp(-\epsilon n/3) \leq \exp(-\alpha\epsilon n/3).$$

□

Lemma 4.1 proves that the error of the classifier $f^{(r)}$ of each round on the weighted mixture of distributions is low. It holds due to a known extension of the VC Theorem and Chernoff bounds, but we prove it here for our parameters for completeness.

Lemma 4.1. *With probability at least $1-\delta/2$, for all rounds $r \in [t]$:*

(a) $(1+3\alpha')err_{S^{(r)}}(f^{(r)}) + 3\epsilon' \leq (1+7\alpha')\texttt{OPT} + 19\epsilon'$.

(b) $err_{\tilde{D}^{(r-1)}}(f^{(r)}) \leq (1+\alpha')err_{S^{(r)}}(f^{(r)}) + \epsilon'$.

Proof. Let $S^{(r)}$ be a set of samples of size $C \cdot \frac{1}{\alpha'\epsilon'}\left(d\ln\left(\frac{1}{\epsilon'}\right) + \ln\left(\frac{1}{\delta'}\right)\right)$ drawn from $\tilde{D}^{(r-1)}$, where C is a constant. We will prove that for large enough constant C the two statements hold simultaneously for all rounds, each with probability at least $1-t\delta'$. It suffices to prove that each statement in each round holds with probability at least $1-\delta'$. For a given round r:

(a) By f^*'s definition it holds that $\text{err}_{D_i}(f^*) \leq \texttt{OPT} + \epsilon' \; \forall i \in [k]$, so it must also hold that $\text{err}_{\tilde{D}^{(r-1)}}(f^*) \leq \texttt{OPT} + \epsilon'$, since $\tilde{D}^{(r-1)}$ is a weighted average of the distributions. From the Corollary it holds that $\Pr[\text{err}_{S^{(r)}}(f^*) \geq (1+\alpha')\text{err}_{\tilde{D}^{(r-1)}}(f^*) + \epsilon'] \leq \exp(-\alpha'\epsilon'|S^{(r)}|/3) \leq \delta'$ and since $\alpha' \leq 1$, it is easy to see that with probability at least $1-\delta'$,

$$\text{err}_{S^{(r)}}(f^*) \leq (1+\alpha')\texttt{OPT} + 3\epsilon' \tag{11}$$

Since $f^{(r)}$ is the error minimizing classifier for the sample $S^{(r)}$, it holds that $\text{err}_{S^{(r)}}(f^{(r)}) \leq \text{err}_{S^{(r)}}(f^*) + \epsilon'$. Therefore,

$$(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon' \leq (1+3\alpha')\text{err}_{S^{(r)}}(f^*) + 7\epsilon' \overset{(11)}{\leq} (1+7\alpha')\texttt{OPT} + 19\epsilon'.$$

(b) We prove the second statement for all $f \in \mathcal{F}$, using Theorem 5.7 from [1]. The theorem states that for every $h \in \mathcal{H}$, it holds that $\text{err}_D(h) \leq (1+\gamma)\text{err}_S(h) + \beta$ with probability at least $1 - 4\Pi_{\mathcal{H}}(2m)\exp\left(\frac{-\gamma\beta m}{4(\gamma+1)}\right)$, where S is a sample of size m drawn from a distribution D on $\mathcal{X} \times \{0,1\}$, $\gamma > 2\beta$, and $\Pi_{\mathcal{H}}(n) = \max\{|\mathcal{H}_{|S}| : S \subseteq \mathcal{X} \text{ and } |S| = n\}$ is the growth function of $\mathcal{H}$.

We apply Theorem 5.7 for $\gamma = \alpha'$, $\beta = \epsilon'$, $D = \tilde{D}^{(r-1)}$, $S = S^{(r)}$, $\mathcal{H} = \mathcal{F}$. Since the VC-dimension of $\mathcal{F}$ is d, from [[1], Theorem 3.7] it holds that $\Pi_{\mathcal{F}}(2m) \leq \left(\frac{2em}{d}\right)^d$. In our setting, the theorem states that, given round r, for every $f \in \mathcal{F}$, it holds that $\text{err}_{\tilde{D}^{(r-1)}}(f) \leq (1+\alpha')\text{err}_{S^{(r)}}(f) + \epsilon'$ with probability at least $1 - 4\left(\frac{2em}{d}\right)^d \exp\left(\frac{-\alpha'\epsilon' m}{4(\alpha'+1)}\right)$.

It remains to prove that, for large enough C, $m = C \cdot \frac{1}{\alpha'\epsilon'}\left(d\ln\left(\frac{1}{\epsilon'}\right) + \ln\left(\frac{1}{\delta'}\right)\right)$ samples suffice to guarantee that $4\left(\frac{2em}{d}\right)^d \exp\left(\frac{-\alpha'\epsilon' m}{4(\alpha'+1)}\right) \leq \delta'$ so that the statement holds with probability at least $1-\delta'$. It suffices to prove that for the given m:

$\ln(4) + d\ln(2e) + d\ln\left(\frac{m}{d}\right) - \frac{\alpha'}{8}\epsilon' m \leq -\ln\left(\frac{1}{\delta'}\right)$

$\Leftrightarrow \ln(4) + d\ln(2e) + d\ln\left(\frac{m}{d}\right) + \ln\left(\frac{1}{\delta'}\right) \leq \frac{C}{8}d\ln\left(\frac{1}{\epsilon'}\right) + \frac{C}{8}\ln\left(\frac{1}{\delta'}\right)$.

We consider two cases:

i. If $d\ln\left(\frac{1}{\epsilon'}\right) \geq \ln\left(\frac{1}{\delta'}\right)$, then $\frac{m}{d} \leq \frac{2C}{\alpha'\epsilon'}\ln\left(\frac{1}{\epsilon'}\right) < \frac{C}{\epsilon'^2}\ln\left(\frac{1}{\epsilon'}\right)$. So to prove the statement, it suffices to prove that

$$\ln(4) + d\ln(2e) + d\Big(\ln(C) + 2\ln\left(\frac{1}{\epsilon'}\right) + \ln\ln\left(\frac{1}{\epsilon'}\right)\Big) + \ln\left(\frac{1}{\delta'}\right) \leq \frac{C}{8}d\ln\left(\frac{1}{\epsilon'}\right) + \frac{C}{8}\ln\left(\frac{1}{\delta'}\right).$$

The latter inequality holds for large enough C.

ii. If $d\ln\left(\frac{1}{\epsilon'}\right) \leq \ln\left(\frac{1}{\delta'}\right)$, then $\frac{m}{d} \leq \frac{2C}{\alpha'\epsilon'}\frac{\ln(1/\delta')}{d} < \frac{C}{\epsilon'^2}\frac{\ln(1/\delta')}{d}$. So to prove the statement, it suffices to prove that

$$\ln(4)+d\ln(2e)+d\Big(\ln(C)+2\ln\left(\frac{1}{\epsilon'}\right)+\ln\left(\frac{\ln(1/\delta')}{d}\right)\Big)+\ln\left(\frac{1}{\delta'}\right) \leq \frac{C}{8}d\ln\left(\frac{1}{\epsilon'}\right)+\frac{C}{8}\ln\left(\frac{1}{\delta'}\right).$$

If we prove that $d\ln\left(\frac{\ln(1/\delta')}{d}\right) \leq \ln(1/\delta')$, then the inequality holds for large enough C. Indeed, it holds that $\ln\left(\frac{\ln(1/\delta')}{d}\right)/\frac{\ln(1/\delta')}{d} \leq \frac{1}{e}$, since $\max_{x\in\mathbb{R}}\{\ln(x)/x\} = \frac{1}{e}$.

Thus the second statement holds too with probability at least $1-\delta'$.

□

Lemmas 4.2 and 4.3 give us two inequalities that are useful for all the proofs of Section 4.

Lemma 4.2. *Let $L_r = \{i \in [k] \mid |err_{T_i}(f^{(r)}) - err_{D_i}(f^{(r)})| \leq \alpha' \cdot err_{D_i}(f^{(r)}) + \epsilon'\}$. With probability $1-\delta/2$, it holds that*

$$\sum_{i\in L_r}\left(w_i^{(r-1)} err_{T_i}(f^{(r)})\right) \leq [(1+3\alpha')err_{S^{(r)}}(f^{(r)}) + 3\epsilon']\Phi^{(r-1)} \leq [(1+7\alpha')OPT + 19\epsilon']\Phi^{(r-1)}.$$

Proof. By linearity of expectation,

$$\begin{aligned}
\text{err}_{\tilde{D}^{(r-1)}}(f^{(r)}) &= \frac{1}{\Phi^{(r-1)}}\sum_{i=1}^{k}\left(w_i^{(r-1)}\text{err}_{D_i}(f^{(r)})\right) \\
&\geq \frac{1}{\Phi^{(r-1)}}\sum_{i\in L_r}\left(w_i^{(r-1)}\text{err}_{D_i}(f^{(r)})\right) \\
&\geq \frac{1}{(1+\alpha')\Phi^{(r-1)}}\sum_{i\in L_r}\left(w_i^{(r-1)}\text{err}_{T_i}(f^{(r)})\right) - \frac{\epsilon'}{1+\alpha'}.
\end{aligned}$$

Therefore, $\sum_{i\in L_r}\left(w_i^{(r-1)}\text{err}_{T_i}(f^{(r)})\right) \leq [(1+\alpha')\text{err}_{\tilde{D}^{(r-1)}}(f^{(r)}) + \epsilon']\Phi^{(r-1)}$. By Lemma 4.1(b), it follows that with probability $1-\delta/2$,

$$\begin{aligned}
\sum_{i\in L_r}\left(w_i^{(r-1)}\text{err}_{T_i}(f^{(r)})\right) &\leq [(1+\alpha')(1+\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + (1+\alpha')\epsilon' + \epsilon']\Phi^{(r-1)} \\
&\leq [(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon']\Phi^{(r-1)} \\
&\overset{\text{Lemma 4.1(a)}}{\leq} [(1+7\alpha')\texttt{OPT} + 19\epsilon']\Phi^{(r-1)}.
\end{aligned}$$

□

Lemma 4.3. *For all $i \in [k]$ it holds that*

$$\sum_{r=1}^{t} s_i^{(r)} \leq \frac{\ln(\Phi^{(t)})}{1-\alpha'/2}.$$

Proof. In every round r, $w_i^{(r)} = w_i^{(r-1)}(1+s_i^{(r)})$. Therefore for any $i \in [k]$,

$$\begin{aligned} w_i^{(t)} &= \prod_{r=1}^{t}(1+s_i^{(r)}) \\ &\geq \prod_{r=1}^{t}\exp(s_i^{(r)} - (s_i^{(r)})^2/2) \\ &\overset{s_i^{(r)}\leq\alpha'}{\geq} \exp\left((1-\alpha'/2)\sum_{r=1}^{t}s_i^{(r)}\right), \end{aligned}$$

where the second to last inequality holds since $(1+x) \geq \exp(x - x^2/2)$ for $x \in \mathbb{R}_+$. The inequality follows since $w_i^{(t)} \leq \Phi^{(t)}$ for all $i \in [k]$. □

We will now give the proof of Theorem 3.

Proof of Theorem 3. By the Corollary, for a given round r and player i,

$$\Pr[|\text{err}_{T_i}(f^{(r)}) - \text{err}_{D_i}(f^{(r)})| \geq \alpha' \cdot \text{err}_{D_i}(f^{(r)}) + \epsilon'] \leq 2\exp(-\alpha'\epsilon'|T_i|/3).$$

If $|T_i| = \frac{3}{\epsilon'\alpha'}\ln\left(\frac{k}{\delta'}\right) = O\Big(\frac{1}{\epsilon'\alpha'}\ln\left(\frac{k}{\delta'}\right)\Big)$, the inequality

$$|\text{err}_{T_i}(f^{(r)}) - \text{err}_{D_i}(f^{(r)})| \leq \alpha' \cdot \text{err}_{D_i}(f^{(r)}) + \epsilon' \tag{12}$$

holds with probability at least $1 - 2\delta'/k$. By union bound, it follows that (12) holds for every i and every r with probability at least $1 - 2\delta' t = 1 - \delta/2$.

With probability at least $1-\delta$ inequality (12) and the inequality of Lemma 4.2 hold for all rounds and players. We restrict the rest of the proof to this event. It holds that,

$$\begin{aligned} \Phi^{(r)} &= \Phi^{(r-1)} + \sum_{i=1}^{k}\left(w_i^{(r-1)}\cdot s_i^{(r)}\right) \\ &\leq \Phi^{(r-1)} + \frac{\alpha'^2}{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'}\sum_{i=1}^{k}\left(w_i^{(r-1)}\text{err}_{T_i}(f^{(r)})\right) \\ &\overset{L_r=[k]}{\leq} \Phi^{(r-1)}\left(1 + \frac{\alpha'^2}{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'}[(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon']\right) \\ &= \Phi^{(r-1)}(1+\alpha'^2) \end{aligned}$$

By induction, $\Phi^{(t)} \leq \Phi^{(0)}(1+\alpha'^2)^t = k(1+\alpha'^2)^t \leq k\exp(t\alpha'^2)$. From Lemma 4.3 and $t = 2\lceil\ln(k)/\alpha'^3\rceil$, it follows that

$$\sum_{r=1}^{t}s_i^{(r)} \leq \frac{\ln(k) + t\alpha'^2}{1-\alpha'/2} \leq \frac{1+\alpha'}{1-\alpha'/2}t\alpha'^2. \tag{13}$$

Let G_i be the set of rounds r such that $s_i^{(r)} < \alpha'$. We consider these to be the "good" classifiers. Because of (13), we have $|[t] \setminus G_i| \leq \frac{1}{\alpha'}\sum_{r\in[t]\setminus G_i}\alpha' \leq \frac{1}{\alpha'}\sum_{r=1}^{t}s_i^{(r)} \leq \frac{1+\alpha'}{1-\alpha'/2}\alpha' t$. For the classifiers of the rounds

$r \in G_i$, it holds that

$$\sum_{r \in G_i} \frac{\text{err}_{T_i}(f^{(r)})\alpha'^2}{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'} = \sum_{r \in G_i} s_i^{(r)} \leq \sum_{r=1}^{t} s_i^{(r)} \overset{(13)}{\leq} \frac{1+\alpha'}{1-\alpha'/2}\alpha'^2 t.$$

Thus, $\sum_{r \in G_i} \text{err}_{T_i}(f^{(r)}) \overset{4.1(a)}{\leq} t\frac{1+\alpha'}{1-\alpha'/2}[(1+7\alpha')\texttt{OPT} + 19\epsilon']$. From inequality (12), it follows that:

$$\begin{aligned}
&(1-\alpha')\sum_{r \in G_i} \text{err}_{D_i}(f^{(r)}) - |G_i|\epsilon' \leq t\frac{1+\alpha'}{1-\alpha'/2}[(1+7\alpha')\texttt{OPT} + 19\epsilon'] \\
&\Rightarrow \sum_{r \in G_i} \text{err}_{D_i}(f^{(r)}) \leq t\frac{1+\alpha'}{(1-\alpha'/2)(1-\alpha')}[(1+7\alpha')\texttt{OPT} + 19\epsilon'] + \frac{t\epsilon'}{1-\alpha'} \\
&\Rightarrow \sum_{r \in G_i} \text{err}_{D_i}(f^{(r)}) \leq [(1+12\alpha')\texttt{OPT} + 25\epsilon']t,
\end{aligned}$$

which holds for $\alpha' < 1/12$.

For each example e that is a mistake for f_{NR1}, it must be a mistake for at least $t/2 - |[t] \setminus G_i|$ members of G_i. Thus the fraction of error of f_{NR1} is at most

$$\frac{\sum_{r \in G_i} \text{err}_{D_i}(f^{(r)})}{t/2 - |[t] \setminus G_i|} \leq \frac{(1+12\alpha')\texttt{OPT} + 25\epsilon'}{1/2 - (1+\alpha')\alpha'/(1-\alpha'/2)} \leq (2+35\alpha')\texttt{OPT} + 60\epsilon'.$$

Having set $\alpha' = \alpha/35$ and $\epsilon' = \epsilon/60$ we get that $\text{err}_{D_i}(f_{\text{NR1}}) \leq (2+\alpha)\texttt{OPT} + \epsilon$.

As for the total number of samples, it is the sum of $O(\frac{k}{\alpha'\epsilon'}\ln(k/\delta'))$ and $O\Big(\frac{1}{\alpha'\epsilon'}\Big(d\ln\Big(\frac{1}{\epsilon'}\Big) + \ln\Big(\frac{1}{\delta'}\Big)\Big)\Big)$ samples for each round. Because there are $O(\ln(k)/\alpha'^3)$ rounds, the total number of samples is

$$O\Big(\frac{\ln(k)}{\alpha'^4\epsilon'}\Big(k\ln\Big(\frac{k}{\delta'}\Big) + d\ln\Big(\frac{1}{\epsilon'}\Big)\Big)\Big) = O\Big(\frac{\ln(k)}{\alpha^4\epsilon}\Big(k\ln\Big(\frac{k}{\delta}\Big) + d\ln\Big(\frac{1}{\epsilon}\Big)\Big)\Big).$$

□

Algorithm NR2 faces a similar challenge as Algorithm R2. Given a player i, since the number of samples T_i used to estimate $\text{err}_{D_i}(f^{(r)})$ in each round is low, the estimation is not very accurate. Ideally, we would want the inequality

$$|\text{err}_{T_i}(f^{(r)}) - \text{err}_{D_i}(f^{(r)})| \leq \alpha' \cdot \text{err}_{D_i}(f^{(r)}) + \epsilon'$$

to hold for all players and all rounds with high probability. The "good" classifiers are now defined as the ones corresponding to rounds for which the inequality holds and $\text{err}_{T_i}(f^{(r)})$ is not very high (an indication of which is that $s_i^{(r)} < \alpha'$). The expected number of rounds that either one of these properties does not hold is a constant fraction of the rounds ($\approx t\alpha'$) and due to the high number of rounds it is concentrated around that value, as in Algorithm R2. The proof of Theorem 4 is the following.

Proof of Theorem 4. By the Corollary, for a given round r and player i,

$$\Pr[|\text{err}_{T_i}(f^{(r)}) - \text{err}_{D_i}(f^{(r)})| \geq \alpha' \cdot \text{err}_{D_i}(f^{(r)}) + \epsilon'] \leq 2\exp(-\alpha'\epsilon'|T_i|/3).$$

If $|T_i| = \frac{6}{\epsilon'\alpha'} \ln\left(\frac{\sqrt{2}}{\alpha'}\right) = O\Big(\frac{1}{\epsilon'\alpha'} \ln\left(\frac{1}{\alpha'}\right)\Big)$, then

$$\Pr[|\text{err}_{T_i}(f^{(r)}) - \text{err}_{D_i}(f^{(r)})| \geq \alpha' \cdot \text{err}_{D_i}(f^{(r)}) + \epsilon'] \leq \alpha'^2. \tag{14}$$

Assuming that the inequality of Lemma 4.2 holds, which is true with probability $1 - \delta/2$, it follows that

$$\begin{aligned}
&\mathbb{E}[\Phi^{(r)} \mid \Phi^{(r-1)}] \\
&\leq \mathbb{E}\left[\Phi^{(r-1)} + \frac{\alpha'^2}{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'} \sum_{i \in L_r} \left(w_i^{(r-1)} \text{err}_{T_i}(f^{(r)})\right) + \sum_{i \notin L_r} \left(w_i^{(r-1)} s_i^{(r-1)}\right) \middle| \Phi^{(r-1)}\right] \\
&\leq \mathbb{E}\left[\Phi^{(r-1)} + \frac{\alpha'^2}{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'} [(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon']\Phi^{(r-1)} + \alpha' \sum_{i \notin L_r} w_i^{(r-1)} \middle| \Phi^{(r-1)}\right] \\
&\overset{(14)}{\leq} \Phi^{(r-1)}(1 + \alpha'^2 + \alpha'^3)
\end{aligned}$$

By the definition of expectation, $\mathbb{E}[\Phi^{(r)}] \leq \mathbb{E}[\Phi^{(r-1)}](1 + \alpha'^2 + \alpha'^3)$. So by induction and the fact that $\Phi^{(0)} = k$, $\mathbb{E}[\Phi^{(t)}] \leq k \exp(t\alpha'^2(1+\alpha'))$. Markov's inequality states that $\Pr[\Phi^{(t)} \geq \frac{\mathbb{E}[\Phi^{(t)}]}{\delta/4}] \leq \delta/4$. So with overall probability $1 - \delta/4 - \delta/2 = 1 - 3\delta/4$ it holds that $\Phi^{(t)} \leq \frac{4k}{\delta} \exp(t\alpha'^2(1+\alpha'))$.

From Lemma 4.3 and $t = 2\lceil \ln(4k/\delta)/\alpha'^3 \rceil$, it follows that

$$\sum_{r=1}^{t} s_i^{(r)} \leq \frac{\ln(4k/\delta) + t\alpha'^2(1+\alpha')}{1 - \alpha'/2} \leq \frac{(1+2\alpha')}{1-\alpha'/2} t\alpha'^2. \tag{15}$$

For $G_i = \{r \in [t] \mid s_i^{(r)} < \alpha'\}$, we have $|[t] \setminus G_i| \leq \frac{1+2\alpha'}{1-\alpha'/2}\alpha' t$ because of (15).

Let $R_i = \{r \in [t] \mid |\text{err}_{T_i}(f^{(r)}) - \text{err}_{D_i}(f^{(r)})| \leq \alpha' \cdot \text{err}_{D_i}(f^{(r)}) + \epsilon'\}$. For the classifiers of the rounds $r \in G_i \cap R_i$:

$$\begin{aligned}
\sum_{r \in G_i \cap R_i} \text{err}_{D_i}(f^{(r)}) &\leq \sum_{r \in G_i \cap R_i} \frac{\text{err}_{T_i}(f^{(r)})}{1-\alpha'} + \frac{|G_i \cap R_i|\epsilon'}{1-\alpha'} \\
&\leq \sum_{r \in G_i \cap R_i} \frac{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'}{\alpha'^2} \frac{\text{err}_{T_i}(f^{(r)})\alpha'^2}{(1-\alpha')[(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon']} + \frac{t\epsilon'}{1-\alpha'} \\
&= \sum_{r \in G_i \cap R_i} \frac{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'}{(1-\alpha')\alpha'^2} s_i^{(r)} + \frac{t\epsilon'}{1-\alpha'} \\
&\overset{(15)}{\leq} \frac{(1+7\alpha')\texttt{OPT} + 19\epsilon'}{(1-\alpha')\alpha'^2} \frac{(1+2\alpha')}{1-\alpha'/2} t\alpha'^2 + \frac{t\epsilon'}{1-\alpha'} \\
&\leq [(1+15\alpha')\texttt{OPT} + 25\epsilon']t
\end{aligned}$$

which holds for $\alpha' < 1/15$.

We will now bound $|[t] \setminus R_i|$. For every round r, let $m^{(r)}$ be the indicator random variable of the set $[t] \setminus R_i$ and let $y^{(r)} = \alpha'^2$. It holds that for all rounds r, $|m^{(r)} - y^{(r)}| \leq 1$ and $m^{(r)}, y^{(r)} \geq 0$. In addition, from inequality (14) it follows that $\mathbb{E}[m^{(r)} - y^{(r)} \mid \sum_{r'<r} m^{(r')}, \sum_{r'<r} y^{(r')}] = \alpha'^2 - \alpha'^2 \leq 0$.

Using [[9], Lemma 10], with $\varepsilon = 1/2$ and $A = \alpha'^2$, we get that

$$\Pr\left[\sum_{r=1}^{t} m^{(r)} \geq 2\alpha'^2 t + 2\alpha'^2 t\right] \leq \exp(-\alpha'^2 t/2) \leq \delta/4k.$$

So $|[t] \setminus R_i| = \sum_{r=1}^{t} m^{(r)} \leq 4\alpha'^2 t$ for all i with probability at least $1 - \delta/4$, by union bound.

For each example e that is a mistake for f_{NR2}, it must be a mistake for at least $t/2 - |[t] \setminus (G_i \cap R_i)|$ members of $G_i \cap R_i$. Thus, with probability at least $1 - \delta$, the fraction of error of f_{NR2} is at most

$$\frac{\sum_{r \in G_i \cap R_i} \text{err}_{D_i}(f^{(r)})}{t/2 - |[t] \setminus (G_i \cap R_i)|} \leq \frac{(1 + 15\alpha')\texttt{OPT} + 25\epsilon'}{t/2 - 4\alpha'^2 t - (1+\alpha')\alpha' t/(1 - \alpha'/2)} \leq (2 + 40\alpha')\texttt{OPT} + 64\epsilon'.$$

Having set $\alpha' = \alpha/40$ and $\epsilon' = \epsilon/64$ we get that $\text{err}_{D_i}(f_{\text{NR2}}) \leq (2+\alpha)\texttt{OPT} + \epsilon$.

As for the total number of samples, it is the sum of $O(\frac{k}{\alpha'\epsilon'}\ln(1/\alpha'))$ samples and $O\Big(\frac{1}{\alpha'\epsilon'}\Big(d\ln\Big(\frac{1}{\epsilon'}\Big) + \ln\Big(\frac{1}{\delta'}\Big)\Big)\Big)$ samples for each round. Because there are $O(\ln(k/\delta)/\alpha'^3)$ rounds, the total number of samples is

$$O\Big(\frac{1}{\alpha^4\epsilon}\ln\Big(\frac{k}{\delta}\Big)\Big(k\ln\Big(\frac{1}{\alpha}\Big) + d\ln\Big(\frac{1}{\epsilon}\Big) + \ln\Big(\frac{1}{\delta}\Big)\Big)\Big).$$

□

We note that the classifiers returned by these algorithms have a multiplicative approximation factor of almost 2 on the error. A different approach would be to allow for randomized classifiers with low error probability over both the randomness of the example and the classifier. We design two algorithms, NR1-AVG and NR2-AVG that return a classifier which satisfies this form of guarantee on the error without the 2-approximation factor but use roughly $\frac{\alpha}{\epsilon}$ times more samples. The returned classifier is a randomized algorithm that, given an element x, chooses one of the classifiers of all rounds uniformly at random and returns the label that this classifier gives to x. For any distribution over examples, the error probability of this randomized classifier is exactly the average of the error probability of classifiers $f^{(1)}, f^{(2)}, \ldots, f^{(t)}$, hence the AVG in the names. The guarantees of the algorithms are stated in the next two theorems.

Theorem 5. *For any $\epsilon, \delta \in (0,1)$, $24\epsilon/25 < \alpha < 1$, and hypothesis class $\mathcal{F}$ of VC dimension d, Algorithm NR1-AVG returns a classifier $f_{\text{NR1-AVG}}$ such that for the expected error $\overline{err}_{D_i}(f_{\text{NR1-AVG}}) \leq (1+\alpha)\texttt{OPT} + \epsilon$ holds for all $i \in [k]$ with probability $1 - \delta$ using m samples, where*

$$m = O\Big(\frac{\ln(k)}{\alpha^3\epsilon^2}\Big(d\ln\Big(\frac{1}{\epsilon}\Big) + k\ln\Big(\frac{k}{\delta}\Big)\Big)\Big).$$

Theorem 6. *For any $\epsilon, \delta \in (0,1)$, $30\epsilon/29 < \alpha < 1$, and hypothesis class $\mathcal{F}$ of VC dimension d, Algorithm NR2-AVG returns a classifier $f_{\text{NR2-AVG}}$ such that for the expected error $\overline{err}_{D_i}(f_{\text{NR2-AVG}}) \leq (1+\alpha)\texttt{OPT} + \epsilon$ holds for all $i \in [k]$ with probability $1 - \delta$ using m samples, where*

$$m = O\Big(\frac{1}{\alpha^3\epsilon^2}\ln\Big(\frac{k}{\delta}\Big)\Big((d+k)\ln\Big(\frac{1}{\epsilon}\Big) + \ln\Big(\frac{1}{\delta}\Big)\Big)\Big).$$

Algorithms NR1-AVG and NR2-AVG are the following.

Algorithm NR1-AVG

Initialization: $\forall i \in [k]\ w_i^{(0)} := 1$; $\alpha' := \alpha/12$; $t := 2\lceil \ln(k)/(\epsilon'\alpha'^2)\rceil$; $\epsilon' := \epsilon/25$; $\delta' := \delta/(4t)$;
for $r = 1, \ldots, t$ **do**
  $\tilde{D}^{(r-1)} \leftarrow \frac{1}{\Phi^{(r-1)}} \sum_{i=1}^{k} \left(w_i^{(r-1)} D_i \right)$, where $\Phi^{(r-1)} := \sum_{i=1}^{k} w_i^{(r-1)}$;
  Draw a sample set $S^{(r)}$ of size $O\left(\frac{1}{\alpha'\epsilon'}\left(d \ln\left(\frac{1}{\epsilon'}\right) + \ln\left(\frac{1}{\delta'}\right)\right)\right)$ from $\tilde{D}^{(r-1)}$;
  $f^{(r)} \leftarrow \mathcal{O}_{\mathcal{F}}(S^{(r)})$;
  for $i = 1, \ldots, k$ **do**
    Draw a sample set T_i of size $O\left(\frac{1}{\alpha'\epsilon'} \ln\left(\frac{k}{\delta'}\right)\right)$ from D_i;
    $s_i^{(r)} \leftarrow \frac{\mathrm{err}_{T_i}(f^{(r)})\epsilon'\alpha'}{(1+3\alpha')\mathrm{err}_{S^{(r)}}(f^{(r)})+3\epsilon'}$
    Update: $w_i^{(r)} \leftarrow w_i^{(r-1)}(1 + s_i^{(r)})$
  end for
end for

return $f_{\text{NR1-AVG}}$, where $f_{\text{NR1-AVG}}(x) \overset{R}{\leftarrow} \{f^{(r)}(x)\}_{r=1}^{t}$;

Algorithm NR2-AVG

Initialization: $\forall i \in [k]\ w_i^{(0)} := 1$; $\alpha' := \alpha/15$; $t := 2\lceil \ln(4k/\delta)/(\epsilon'\alpha'^2)\rceil$; $\epsilon' := \epsilon/29$; $\delta' := \delta/(4t)$;
for $r = 1, \ldots, t$ **do**
  $\tilde{D}^{(r-1)} \leftarrow \frac{1}{\Phi^{(r-1)}} \sum_{i=1}^{k} \left(w_i^{(r-1)} D_i \right)$, where $\Phi^{(r-1)} := \sum_{i=1}^{k} w_i^{(r-1)}$;
  Draw a sample set $S^{(r)}$ of size $O\left(\frac{1}{\alpha'\epsilon'}\left(d \ln\left(\frac{1}{\epsilon'}\right) + \ln\left(\frac{1}{\delta'}\right)\right)\right)$ from $\tilde{D}^{(r-1)}$;
  $f^{(r)} \leftarrow \mathcal{O}_{\mathcal{F}}(S^{(r)})$;
  for $i = 1, \ldots, k$ **do**
    Draw a sample set T_i of size $O\left(\frac{1}{\alpha'\epsilon'} \ln\left(\frac{1}{\epsilon'}\right)\right)$ from D_i;
    $s_i^{(r)} \leftarrow \frac{\mathrm{err}_{T_i}(f^{(r)})\epsilon'\alpha'}{(1+3\alpha')\mathrm{err}_{S^{(r)}}(f^{(r)})+3\epsilon'}$
    Update: $w_i^{(r)} \leftarrow w_i^{(r-1)}(1 + s_i^{(r)})$
  end for
end for

return $f_{\text{NR2-AVG}}$, where $f_{\text{NR2-AVG}}(x) \overset{R}{\leftarrow} \{f^{(r)}(x)\}_{r=1}^{t}$;

We first prove the guarantee for Algorithm NR1-AVG.

Proof of Theorem 5. The expected error of the returned classifier $f_{\text{NR1-AVG}}$ on player i's distribution is $\overline{\mathrm{err}}_{D_i}(f_{\text{NR1-AVG}}) = \frac{1}{t}\sum_{r=1}^{t} \mathrm{err}_{D_i}(f^{(r)})$. We will prove that with probability at least $1-\delta$, $\overline{\mathrm{err}}_{D_i}(f_{\text{NR1-AVG}}) \leq (1+\alpha)\texttt{OPT} + \epsilon$ for all $i \in [k]$.

By the Corollary, for a given round r and player i,

$$\Pr[|\mathrm{err}_{T_i}(f^{(r)}) - \mathrm{err}_{D_i}(f^{(r)})| \geq \alpha' \cdot \mathrm{err}_{D_i}(f^{(r)}) + \epsilon'] \leq 2\exp(-\alpha'\epsilon'|T_i|/3).$$

If $|T_i| = \frac{3}{\epsilon'\alpha'} \ln\left(\frac{k}{\delta'}\right) = O\Big(\frac{1}{\epsilon'\alpha'} \ln\left(\frac{k}{\delta'}\right)\Big)$, the inequality holds with probability at least $1 - 2\delta'/k$. By union bound, it follows that it holds for every i and every r with probability at least $1 - 2\delta' t = 1 - \delta/2$.

With probability at least $1 - \delta$ the previous inequality as well as the inequality of Lemma 4.2 hold for all rounds and players. We restrict the rest of the proof to this event.

It holds that,

$$\begin{aligned}
\Phi^{(r)} &= \Phi^{(r-1)} + \sum_{i=1}^{k} \left(w_i^{(r-1)} s_i^{(r)}\right) \\
&\leq \Phi^{(r-1)} + \frac{\epsilon'\alpha'}{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'} \sum_{i=1}^{k} \left(w_i^{(r-1)} \text{err}_{T_i}(f^{(r)})\right) \\
&\overset{L_r=[k]}{\leq} \Phi^{(r-1)} \left(1 + \frac{\epsilon'\alpha'}{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'} [(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon']\right) \\
&\leq \Phi^{(r-1)}(1 + \epsilon'\alpha')
\end{aligned}$$

By induction, $\Phi^{(t)} \leq k \exp(t\epsilon'\alpha')$. From Lemma 4.3 and since $t = 2\lceil \ln(k)/(\epsilon'\alpha'^2)\rceil$, it follows that

$$\sum_{r=1}^{t} s_i^{(r)} \leq \frac{\ln(k) + t\epsilon'\alpha'}{1 - \alpha'/2} \leq \frac{1+\alpha'}{1-\alpha'/2} t\epsilon'\alpha'. \tag{16}$$

Therefore, the total error is:

$$\begin{aligned}
\sum_{r=1}^{t} \text{err}_{D_i}(f^{(r)}) &\leq \sum_{r=1}^{t} \frac{\text{err}_{T_i}(f^{(r)})}{1-\alpha'} + \frac{t\epsilon'}{1-\alpha'} \\
&\leq \sum_{r=1}^{t} \frac{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'}{\epsilon'\alpha'} \frac{\text{err}_{T_i}(f^{(r)})\epsilon'\alpha'}{(1-\alpha')[(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon']} + \frac{t\epsilon'}{1-\alpha'} \\
&= \sum_{r=1}^{t} \frac{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'}{(1-\alpha')\epsilon'\alpha'} s_i^{(r)} + \frac{t\epsilon'}{1-\alpha'} \\
&\overset{(16)}{\leq} \frac{(1+7\alpha')\texttt{OPT} + 19\epsilon'}{(1-\alpha')\epsilon'\alpha'} \frac{(1+\alpha')}{1-\alpha'/2} t\epsilon'\alpha' + \frac{t\epsilon'}{1-\alpha'} \\
&\leq [(1+12\alpha')\texttt{OPT} + 25\epsilon']t \\
&= [(1+\alpha)\texttt{OPT} + \epsilon]t,
\end{aligned}$$

where the last inequality holds for $\alpha' < 1/12$ and we have set $\alpha' = \alpha/12$ and $\epsilon' = \epsilon/25$.

As for the total number of samples, it is the sum of $O(\frac{k}{\alpha'\epsilon'} \ln(k/\delta'))$ samples and $O\Big(\frac{1}{\alpha'\epsilon'}\Big(d \ln\left(\frac{1}{\epsilon'}\right) + \ln\left(\frac{1}{\delta'}\right)\Big)\Big)$ samples for each round. Because there are $O(\ln(k)/(\epsilon'\alpha'^2))$ rounds, the total number of samples is

$$O\Big(\frac{\ln(k)}{\alpha^3\epsilon^2}\Big(k \ln\left(\frac{k}{\delta}\right) + d \ln\left(\frac{1}{\epsilon}\right)\Big)\Big).$$

□

Finally, we prove the guarantee of Algorithm NR2-AVG.

Proof of Theorem 6. The expected error of the returned classifier $f_{\text{NR2-AVG}}$ on player i's distribution is $\overline{\text{err}}_{D_i}(f_{\text{NR2-AVG}}) = \frac{1}{t}\sum_{r=1}^{t} \text{err}_{D_i}(f^{(r)})$. We will prove that with probability at least $1-\delta$, $\overline{\text{err}}_{D_i}(f_{\text{NR2-AVG}}) \leq (1+\alpha)\texttt{OPT} + \epsilon$ for all $i \in [k]$.

By the Corollary, for a given round r and player i,

$$\Pr[|\text{err}_{T_i}(f^{(r)}) - \text{err}_{D_i}(f^{(r)})| \geq \alpha' \cdot \text{err}_{D_i}(f^{(r)}) + \epsilon'] \leq 2\exp(-\alpha'\epsilon'|T_i|/3).$$

If $|T_i| = \frac{3}{\epsilon'\alpha'}\ln\left(\frac{2}{\epsilon'\alpha'}\right) \overset{\alpha' \geq 2\epsilon'}{=} O\left(\frac{1}{\epsilon'\alpha'}\ln\left(\frac{1}{\epsilon'}\right)\right)$, then

$$\Pr[|\text{err}_{T_i}(f^{(r)}) - \text{err}_{D_i}(f^{(r)})| \geq \alpha' \cdot \text{err}_{D_i}(f^{(r)}) + \epsilon'] \leq \epsilon'\alpha'. \tag{17}$$

Assuming that the inequality of Lemma 4.2 holds, which is true with probability $1-\delta/2$, it follows that

$$\begin{aligned}
&\mathbb{E}[\Phi^{(r)} \mid \Phi^{(r-1)}] \\
&= \mathbb{E}\left[\Phi^{(r-1)} + \frac{\epsilon'\alpha'}{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'}\sum_{i\in L_r}\left(w_i^{(r-1)}\text{err}_{T_i}(f^{(r)})\right) + \sum_{i\notin L_r}\left(w_i^{(r-1)}s_i^{(r-1)}\right)\middle|\Phi^{(r-1)}\right] \\
&\leq \mathbb{E}\left[\Phi^{(r-1)} + \frac{\epsilon'\alpha'}{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'}[(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon']\Phi^{(r-1)} + \alpha'\sum_{i\notin L_r} w_i^{(r-1)}\middle|\Phi^{(r-1)}\right] \\
&\overset{(17)}{\leq} \Phi^{(r-1)}(1+\epsilon'\alpha' + \epsilon'\alpha'^2)
\end{aligned}$$

By the definition of expectation, $\mathbb{E}[\Phi^{(r)}] \leq \mathbb{E}[\Phi^{(r-1)}](1+\epsilon'\alpha'+\epsilon'\alpha'^2)$. So by induction, $\mathbb{E}[\Phi^{(t)}] \leq k\exp(t\epsilon'\alpha'(1+\alpha'))$. Markov's inequality states that $\Pr[\Phi^{(t)} \geq \frac{\mathbb{E}[\Phi^{(t)}]}{\delta/4}] \leq \delta/4$. So with probability $1-\delta/4-\delta/2 = 1-3\delta/4$ it holds that $\Phi^{(t)} \leq \frac{4k}{\delta}\exp(t\epsilon'\alpha'(1+\alpha'))$.

From Lemma 4.3 and $t = 2\lceil\ln(4k/\delta)/(\epsilon'\alpha'^2)\rceil$, it follows that

$$\sum_{r=1}^{t} s_i^{(r)} \leq \frac{\ln(4k/\delta) + t\epsilon'\alpha'(1+\alpha')}{1-\alpha'/2} \leq \frac{(1+2\alpha')}{1-\alpha'/2}t\epsilon'\alpha'. \tag{18}$$

Let $R_i = \{r \in [t] \mid |\text{err}_{T_i}(f^{(r)}) - \text{err}_{D_i}(f^{(r)})| \leq \alpha' \cdot \text{err}_{D_i}(f^{(r)}) + \epsilon'\}$. For the classifiers of the rounds $r \in R_i$:

$$\begin{aligned}
\sum_{r\in R_i}\text{err}_{D_i}(f^{(r)}) &\leq \sum_{r\in R_i}\frac{\text{err}_{T_i}(f^{(r)})}{1-\alpha'} + \frac{|R_i|\epsilon'}{1-\alpha'} \\
&\leq \sum_{r\in R_i}\frac{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'}{\epsilon'\alpha'}\frac{\text{err}_{T_i}(f^{(r)})\epsilon'\alpha'}{(1-\alpha')[(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon']} + \frac{t\epsilon'}{1-\alpha'} \\
&= \sum_{r\in R_i}\frac{(1+3\alpha')\text{err}_{S^{(r)}}(f^{(r)}) + 3\epsilon'}{(1-\alpha')\epsilon'\alpha'}s_i^{(r)} + \frac{t\epsilon'}{1-\alpha'} \\
&\overset{(18)}{\leq} \frac{(1+7\alpha')\texttt{OPT} + 19\epsilon'}{(1-\alpha')\epsilon'\alpha'}\frac{(1+2\alpha')}{1-\alpha'/2}t\epsilon'\alpha' + \frac{t\epsilon'}{1-\alpha'} \\
&\leq [(1+15\alpha')\texttt{OPT} + 25\epsilon']t
\end{aligned}$$

which holds for $\alpha' < 1/15$.

We will now bound $|[t] \setminus R_i|$. For every round r, let $m^{(r)}$ be the indicator random variable of the set $[t] \setminus R_i$ and let $y^{(r)} = \epsilon'\alpha'$. It holds that for all rounds r, $|m^{(r)} - y^{(r)}| \leq 1$ and $m^{(r)}, y^{(r)} \geq 0$. In addition, from inequality (17) it follows that $\mathbb{E}[m^{(r)} - y^{(r)} \mid \sum_{r'<r} m^{(r')}, \sum_{r'<r} y^{(r')}] = \epsilon'\alpha' - \epsilon'\alpha' \leq 0$.

Using [[9], Lemma 10], with $\varepsilon = 1/2$ and $A = \epsilon'\alpha'$, we get that

$$\Pr\left[\sum_{r=1}^{t} m^{(r)} \geq 2\epsilon'\alpha' t + 2\epsilon'\alpha' t\right] \leq \exp(-\epsilon'\alpha' t/2) \leq \delta/4k.$$

So $|[t] \setminus R_i| = \sum_{r=1}^{t} m^{(r)} \leq 4\epsilon'\alpha' t$ for all i with probability at least $1 - \delta/4$.

Thus, for the expected error it holds that:

$$\begin{aligned}\frac{\sum_{r=1}^{t} \mathrm{err}_{D_i}(f^{(r)})}{t} &= \frac{\sum_{r\in R_i} \mathrm{err}_{D_i}(f^{(r)}) + \sum_{r\notin R_i} \mathrm{err}_{D_i}(f^{(r)})}{t} \\ &\leq (1+15\alpha')\texttt{OPT} + 25\epsilon' + 4\epsilon'\alpha' \leq (1+15\alpha')\texttt{OPT} + 29\epsilon'.\end{aligned}$$

Having set $\alpha' = \alpha/15$ and $\epsilon' = \epsilon/29$ we get that $\overline{\mathrm{err}}_{D_i}(f_{\text{NR2-AVG}}) \leq (1+\alpha)\texttt{OPT} + \epsilon$ with probability at least $1 - \delta$.

As for the total number of samples, it is the sum of $O(\frac{k}{\alpha'\epsilon'}\ln(1/\epsilon'))$ samples and $O\Big(\frac{1}{\alpha'\epsilon'}\Big(d\ln\Big(\frac{1}{\epsilon'}\Big) + \ln\Big(\frac{1}{\delta'}\Big)\Big)\Big)$ samples for each round. Because there are $O(\ln(k/\delta)/\epsilon'\alpha'^2)$ rounds, the total number of samples is

$$O\Big(\frac{1}{\alpha^3\epsilon^2}\ln\Big(\frac{k}{\delta}\Big)\Big((d+k)\ln\Big(\frac{1}{\epsilon}\Big) + \ln\Big(\frac{1}{\delta}\Big)\Big)\Big).$$

□

5 Discussion

The problem has four parameters, d, k, ϵ and δ, so there are many ways to compare the sample complexity of the algorithms. In the non-realizable setting there is one more parameter α, but this is set to be a constant in the beginning of the algorithms. Our sample complexity upper bounds are summarized in the following table.

Table 1: Sample complexity upper bounds

	Algorithm 1	Algorithm 2
Realizable	$O\Big(\frac{\ln(k)}{\epsilon}\Big(d\ln\Big(\frac{1}{\epsilon}\Big) + k\ln\Big(\frac{k}{\delta}\Big)\Big)\Big)$	$O\Big(\frac{\ln(k/\delta)}{\epsilon}\Big(d\ln\Big(\frac{1}{\epsilon}\Big) + k + \ln\Big(\frac{1}{\delta}\Big)\Big)\Big)$
Non-realizable ($2+\alpha$ approx.)	$O\Big(\frac{\ln(k)}{\alpha^4\epsilon}\Big(d\ln\Big(\frac{1}{\epsilon}\Big) + k\ln\Big(\frac{k}{\delta}\Big)\Big)\Big)$	$O\Big(\frac{\ln(k/\delta)}{\alpha^4\epsilon}\Big(d\ln\Big(\frac{1}{\epsilon}\Big) + k\ln\Big(\frac{1}{\alpha}\Big) + \ln\Big(\frac{1}{\delta}\Big)\Big)\Big)$
Non-realizable (randomized)	$O\Big(\frac{\ln(k)}{\alpha^3\epsilon^2}\Big(d\ln\Big(\frac{1}{\epsilon}\Big) + k\ln\Big(\frac{k}{\delta}\Big)\Big)\Big)$	$O\Big(\frac{\ln(k/\delta)}{\alpha^3\epsilon^2}\Big((d+k)\ln\Big(\frac{1}{\epsilon}\Big) + \ln\Big(\frac{1}{\delta}\Big)\Big)\Big)$

Usually δ can be considered constant, since it represents the required error probability, or, in the high success probability regime, $\delta = \frac{1}{poly(k)}$. For both of these natural settings, we can see that Algorithm 2 is

better than Algorithm 1, except for the case of the expected error guarantee. If we assume $k = \Theta(d)$, then Algorithm 2 is always better than Algorithm 1.

In the realizable setting, Algorithm R1 is always better than the algorithm of [5] for the centralized variant of the problem and matches their number of samples in the personalized variant. In addition, Theorem 4.1 of [5] states that the sample complexity of any algorithm in the collaborative model is $\Omega\left(\frac{k}{\epsilon}\ln\left(\frac{k}{\delta}\right)\right)$, given that $d = \Theta(k)$ and $\epsilon, \delta \in (0, 0.1)$, and this holds even for the personalized variant. For $d = \Theta(k)$, the sample complexity of Algorithm R2 is exactly $\ln\left(\frac{k}{\delta}\right)$ times the sample complexity for learning one task. Furthermore, when $|\mathcal{F}| = 2^d$ (e.g. the hard instance for the lower bound of [5]), only $m_{\epsilon,\delta} = O\left(\frac{1}{\epsilon}\left(d + \ln\left(\frac{1}{\delta}\right)\right)\right)$ samples are required in the non-collaborative setting instead of the general bound of the VC theorem, so the sample complexity bound for Algorithm R2 is $O\left(\ln\left(\frac{k}{\delta}\right)\frac{1}{\epsilon}\left(d + k + \ln\left(\frac{1}{\delta}\right)\right)\right)$ and matches exactly the lower bound of [5] up to lower order terms.

In the non-realizable setting, our generalization of algorithms R1 and R2, NR1 and NR2 respectively, have the same sample complexity as in the realizable setting and match the error guarantee for $\texttt{OPT} = 0$. If $\texttt{OPT} \neq 0$, they guarantee an error of a factor 2 multiplicatively on $\texttt{OPT}$. The randomized classifiers returned by Algorithms NR1-AVG and NR2-AVG avoid this factor of 2 in their expected error guarantee. However, to learn such classifiers, there are required $O\left(\frac{1}{\epsilon}\right)$ times more samples.

References

[1] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, New York, NY, USA, 1st edition, 2009. ISBN 052111862X, 9780521118620.

[2] Maria-Florina Balcan, Avrim Blum, Shai Fine, and Yishay Mansour. Distributed learning, communication complexity and privacy. In *Proceedings of the 25th Conference on Computational Learning Theory (COLT)*, pages 26.1–26.22, 2012.

[3] Jonathan Baxter. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28(1):7–39, July 1997. ISSN 0885-6125. doi: 10.1023/A:1007327622663.

[4] Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12(1): 149–198, March 2000. ISSN 1076-9757.

[5] Avrim Blum, Nika Haghtalab, Ariel D. Procaccia, and Mingda Qiao. Collaborative PAC learning. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2389–2398, 2017.

[6] Jiecao Chen, Qin Zhang, and Yuan Zhou. Tight bounds for collaborative PAC learning via multiplicative weights, 2018.

[7] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 713–720, 2011.

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1126–1135, 2017.

[9] Christos Koufogiannakis and Neal E. Young. A nearly linear-time PTAS for explicit fractional packing and covering linear programs. *Algorithmica*, 70(4):648–674, December 2014. ISSN 0178-4617. doi: 10.1007/s00453-013-9771-6.

[10] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of the 22nd Conference on Computational Learning Theory (COLT)*, pages 19–30, 2009.

[11] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1041–1048, 2009.

[12] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, 2nd edition, 2017. ISBN 110715488X, 9781107154889.

[13] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, November 1984. ISSN 0001-0782. doi: 10.1145/1968.1972.

[14] Jialei Wang, Mladen Kolar, and Nathan Srebro. Distributed multi-task learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 751–760, 2016.