

Revisiting Agnostic PAC Learning

Steve Hanneke*
Purdue University

Kasper Green Larsen†
Aarhus University

Nikita Zhivotovskiy‡
UC Berkeley

Abstract

PAC learning, dating back to Valiant’84 and Vapnik and Chervonenkis’64,’74, is a classic model for studying supervised learning. In the *agnostic* setting, we have access to a hypothesis set \mathcal{H} and a training set of labeled samples $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$ drawn i.i.d. from an unknown distribution \mathcal{D} . The goal is to produce a classifier $h : \mathcal{X} \rightarrow \{-1, 1\}$ that is competitive with the hypothesis $h_{\mathcal{D}}^* \in \mathcal{H}$ having the least probability of mispredicting the label y of a new sample $(x, y) \sim \mathcal{D}$.

Empirical Risk Minimization (ERM) is a natural learning algorithm, where one simply outputs the hypothesis from \mathcal{H} making the fewest mistakes on the training data. This simple algorithm is known to have an optimal error in terms of the VC-dimension of \mathcal{H} and the number of samples n .

In this work, we revisit agnostic PAC learning and first show that ERM is in fact sub-optimal if we treat the performance of the best hypothesis, denoted $\tau := \Pr_{\mathcal{D}}[h_{\mathcal{D}}^*(x) \neq y]$, as a parameter. Concretely we show that ERM, and any other proper learning algorithm, is sub-optimal by a $\sqrt{\ln(1/\tau)}$ factor. We then complement this lower bound with the first learning algorithm achieving an optimal error for nearly the full range of τ . Our algorithm introduces several new ideas that we hope may find further applications in learning theory.

1 Introduction

One of the most basic theoretical models for studying binary classification in a supervised learning setup, is the Probably Approximately Correct (PAC) learning framework of Valiant [25], and Vapnik and Chervonenkis [27, 28]. In this framework, a training data set consists of n i.i.d. samples $\mathbf{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ from an unknown data distribution \mathcal{D} over $\mathcal{X} \times \{-1, 1\}$. Here \mathcal{X} is an input domain and $\{-1, 1\}$ are the two possible labels. The goal is to design a learning algorithm \mathcal{A} , that on a training set \mathbf{S} , produces a classifier/hypothesis $h_{\mathbf{S}} : \mathcal{X} \rightarrow \{-1, 1\}$ minimizing the probability of mispredicting the label of a fresh sample from \mathcal{D} , denoted by $\text{er}_{\mathcal{D}}(h) := \Pr_{\mathcal{D}}[h(\mathbf{x}) \neq \mathbf{y}]$.

In the PAC learning framework, the algorithm \mathcal{A} is further given a hypothesis set $\mathcal{H} \subseteq \mathcal{X} \rightarrow \{-1, 1\}$, and the performance of the hypothesis $h_{\mathbf{S}}$ produced by \mathcal{A} must be competitive with the best hypothesis $h_{\mathcal{D}}^*$ in \mathcal{H} , where $h_{\mathcal{D}}^* := \text{argmin}_{h \in \mathcal{H}} \text{er}_{\mathcal{D}}(h)$ (breaking ties arbitrarily). Classic work on PAC learning distinguishes two important cases, namely *realizable* and *agnostic* learning. In the realizable setting, it is assumed that $\text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*) = 0$, i.e. that there is a hypothesis in \mathcal{H} perfectly classifying all data. Here the goal is to achieve $\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) \leq \varepsilon$ for ε going to 0 as fast as possible with n . In the often more realistic setup of agnostic learning, the goal is instead to guarantee $\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) \leq \text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*) + \varepsilon$, thus being competitive with $h_{\mathcal{D}}^*$.

Realizable setting. The realizable setting is by now very well understood, in particular following a surge of results over the past few years. The most natural learning algorithm in this setting is Empirical Risk Minimization (ERM), that simply outputs an arbitrary hypothesis $h_{\mathbf{S}} \in \mathcal{H}$ that achieves $\text{er}_{\mathbf{S}}(h_{\mathbf{S}}) = 0$. Here, and throughout the paper, we let $\text{er}_S(h)$ for a set of samples S denote $\Pr_S[h(\mathbf{x}) \neq \mathbf{y}]$, and when we subscript a probability by S , we let (\mathbf{x}, \mathbf{y}) be a uniform random sample from S . Note that a hypothesis $h_{\mathbf{S}}$ with

*steve.hanneke@gmail.com

†larsen@cs.au.dk

‡zhivotovskiy@berkeley.edu

$\text{er}_{\mathbf{S}}(h_{\mathbf{S}}) = 0$ is guaranteed to exist since $h_{\mathcal{D}}^*$ is one such hypothesis. Classic work [28, 26, 6] shows that ERM guarantees, with probability $1 - \delta$ over $\mathbf{S} \sim \mathcal{D}^n$, that $\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) = O((d \ln(n/d) + \ln(1/\delta))/n)$. Here d denotes the VC-dimension [29] of \mathcal{H} and is defined as the largest number of points $x_1, \dots, x_d \in \mathcal{X}$ for which \mathcal{H} can generate all 2^d possible labelings of x_1, \dots, x_d . This *sample complexity* is known [17, 5, 24, 8] to be optimal for any *proper* learning algorithm, i.e. there exists an input domain \mathcal{X} , hypothesis set \mathcal{H} of VC-dimension d and a data distribution \mathcal{D} , such that any \mathcal{A} that outputs a hypothesis $h_{\mathbf{S}}$ from \mathcal{H} must have $\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) = \Omega((d \ln(n/d) + \ln(1/\delta))/n)$ with probability at least δ . Determining the optimal sample complexity for *improper* learning algorithms, i.e., algorithms that are allowed to output an arbitrary hypothesis $h_{\mathbf{S}} : \mathcal{X} \rightarrow \{-1, 1\}$, and not only hypotheses from \mathcal{H} , was a major open problem for decades. Finally, in work by Hanneke [14], building on ideas of Simon [24], an optimal learning algorithm guaranteeing $\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) = O((d + \ln(1/\delta))/n)$ with probability $1 - \delta$ was finally developed. This matches previous lower bounds [6, 12] and thus settled the complexity of realizable PAC learning. Over the past few years, there have been several works proving optimality of other and arguably simpler learning algorithms, including for the practical heuristic bagging [19, 10], a variant of the one-inclusion graph learning algorithm [1, 17], and most recently for a simple majority vote among three ERM classifiers [2].

Agnostic setting. ERM is also a very natural learning algorithm in the agnostic setting. Instead of outputting a hypothesis $h_{\mathbf{S}}$ with $\text{er}_{\mathbf{S}}(h_{\mathbf{S}}) = 0$ (which might not exist), ERM instead outputs the hypothesis $h_{\mathbf{S}}^* = \text{argmin}_{h \in \mathcal{H}} \text{er}_{\mathbf{S}}(h)$ achieving the best performance on the training data (breaking ties arbitrarily). This strategy is well understood and is known [16] to guarantee $\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) = \text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*) + O(\sqrt{(d + \ln(1/\delta))/n})$. Note that always use $h_{\mathcal{D}}^*$ to denote the hypothesis $\text{argmin}_h \text{er}_{\mathcal{D}}(h)$. Unlike the realizable setting, there is a matching lower bound [3] (Theorem 5.2) even for improper learning algorithms. Thus in contrast to the realizable setting, simple ERM is provably optimal. While this might seem the end of the story, the picture is however more complicated. In particular, one would expect there to be some form of transition between the agnostic and realizable setting, i.e. for sufficiently small $\tau = \text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*)$, ERM must become sub-optimal. The bounds with the explicit dependence on τ are quite standard in the literature and are sometimes called the *first-order bounds*, especially in the contexts of online learning and optimization. For the state-of-the-art upper and lower bounds in the agnostic PAC learning setup, we refer to [7, Corollary 5.3] and the corresponding lower bounds in [11, Chapter 14] and in [4]. Since we revisit ERM and state its sample complexity also as a function of τ , we start with the following upper bounds (with $0 \ln(1/0) = 0$):

Theorem 1 (ERM Theorem, derived from [20]). *For any input domain \mathcal{X} , hypothesis set \mathcal{H} of VC-dimension d , number of samples n , distribution \mathcal{D} over $\mathcal{X} \times \{-1, 1\}$ and any $0 < \delta < 1$, it holds with probability at least $1 - \delta$ over a sample $\mathbf{S} \sim \mathcal{D}^n$ that every hypothesis $h \in \mathcal{H}$ satisfies*

$$|\text{er}_{\mathcal{D}}(h) - \text{er}_{\mathbf{S}}(h)| = O\left(\sqrt{\frac{\text{er}_{\mathcal{D}}(h)(d \ln(1/\text{er}_{\mathcal{D}}(h)) + \ln(1/\delta))}{n}} + \frac{d \ln(n/d) + \ln(1/\delta)}{n}\right).$$

In particular, this implies that running ERM returns a hypothesis $h_{\mathbf{S}} \in \mathcal{H}$ satisfying

$$\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) = \tau + O\left(\sqrt{\frac{\tau(d \ln(1/\tau) + \ln(1/\delta))}{n}} + \frac{d \ln(n/d) + \ln(1/\delta)}{n}\right),$$

where $\tau = \text{er}_{\mathcal{D}}(h_{\mathcal{D}}^)$.*

Observe the $\sqrt{\tau \ln(1/\tau)}$ dependency in $\text{er}_{\mathcal{D}}(h_{\mathbf{S}})$ that smoothly interpolates between the agnostic and realizable setting. By the lower bounds in [11, Chapter 14], we have that any learning algorithm \mathcal{A} produces with probability at least δ a hypothesis $h_{\mathbf{S}}$ with

$$\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) = \tau + \Omega\left(\sqrt{\frac{\tau(d + \ln(1/\delta))}{n}} + \frac{d + \ln(1/\delta)}{n}\right).$$

Thus there is a $\sqrt{\ln(1/\tau)}$ gap between ERM and the lower bound. Furthermore, and unlike the realizable setting, there are no known algorithms that bridge this gap and we have no proof that optimal algorithms need to be improper (except when $\tau = 0$).

Our Contributions. In this work, we close this gap for almost the full range of τ . First, we prove that any proper learning algorithm must incur this $\sqrt{\ln(1/\tau)}$ factor in its sample complexity:

Theorem 2. *There is a constant $C > 0$ such that for any VC-dimension d , number of samples n and τ satisfying $Cd \ln(n/d)/n \leq \tau \leq 1/C$, there is an input domain \mathcal{X} and hypothesis set \mathcal{H} of VC-dimension d , satisfying that for every **proper** learning algorithm \mathcal{A} , there is a distribution \mathcal{D} over $\mathcal{X} \times \{-1, 1\}$ such that:*

1. *There is a hypothesis $h \in \mathcal{H}$ with $\text{er}_{\mathcal{D}}(h) = \tau$.*
2. *With probability at least $1/16$ over a sample $\mathbf{S} \sim \mathcal{D}^n$, it holds that the hypothesis $h_{\mathbf{S}} \in \mathcal{H}$ produced by \mathcal{A} on \mathbf{S} has $\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) = \tau + \Omega(\sqrt{\tau d \ln(1/\tau)/n})$.*

Motivated by this lower bound, we design a new improper learning algorithm that avoids this $\sqrt{\ln(1/\tau)}$ penalty and achieves an optimal sample complexity except for very small values of τ :

Theorem 3. *For any input domain \mathcal{X} , hypothesis set \mathcal{H} of VC-dimension d , number of samples n , distribution \mathcal{D} over $\mathcal{X} \times \{-1, 1\}$ and any $0 < \delta < 1$, there is an algorithm, DISAGREEINGEXPERTS, that when given samples $\mathbf{S} \sim \mathcal{D}^n$ and failure probability δ , returns with probability at least $1 - \delta$ a hypothesis $h_{\mathbf{S}} : \mathcal{X} \rightarrow \{-1, 1\}$ satisfying*

$$\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) = \tau + O\left(\sqrt{\frac{\tau(d + \ln(1/\delta))}{n}} + \frac{\ln^5(n/d)(d + \ln(1/\delta))}{n}\right),$$

where $\tau = \text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*)$.

This is the first known learning algorithm to provably outperform ERM in the agnostic setting. Furthermore, we stress that despite the recent progress on realizable PAC learning, none of the ideas in those works seem to generalize easily to the agnostic setting. Instead, our algorithm is based on a new paradigm of recursively training pairs of nearly optimal classifiers that disagree in many of their predictions. We elaborate on this new approach in Section 1.1 and hope it may find further applications in learning theory.

1.1 Proof Overview

In this section, we present the high level ideas of both our new agnostic PAC learning algorithm, DISAGREEINGEXPERTS, as well as our lower bound for proper learners. We begin with the upper bound.

New algorithm. Our improved algorithm relies on several new insights regarding Empirical Risk Minimization. To set the stage for describing these ideas, consider a data distribution \mathcal{D} over $\mathcal{X} \times \{-1, 1\}$ and let $\tau = \text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*)$. If we run ERM on a data set $\mathbf{S} \sim \mathcal{D}^n$ of size n , then by the ERM Theorem (Theorem 1), this ensures that for sufficiently large constant $c > 0$, ERM will not return a hypothesis h with $\text{er}_{\mathcal{D}}(h) > \tau + c\sqrt{\tau d \ln(1/\tau)/n}$ (let us ignore δ and the additive $d \ln(n/d)/n$ term in the ERM Theorem for simplicity). This is, of course, a $\sqrt{\ln(1/\tau)}$ factor worse than what we are aiming for. To improve this bound, we show that there is always a *win-win* situation we may exploit to shave the $\sqrt{\ln(1/\tau)}$ factor.

To understand this win-win scenario, consider the set $\bar{\mathcal{H}} \subseteq \mathcal{H}$ of near-optimal hypotheses h with $\text{er}_{\mathcal{D}}(h) \leq \tau + c\sqrt{\tau d \ln(1/\tau)/n}$, i.e., the hypotheses that might be returned by a typical execution of ERM. These are what we think of as *experts* in our algorithm name DISAGREEINGEXPERTS. In the proof of the ERM Theorem, the basic idea is to union bound over all $h \in \bar{\mathcal{H}}$ (with a chaining argument), to show that $|\text{er}_{\mathcal{D}}(h) - \text{er}_{\mathbf{S}}(h)| = O(\sqrt{\tau d \ln(1/\tau)/n})$ for all $h \in \bar{\mathcal{H}}$ simultaneously and thus returning the hypothesis $h_{\mathbf{S}}^*$ with smallest error on \mathbf{S} is a good strategy. Our first new insight is, that if the hypotheses in $\bar{\mathcal{H}}$ are sufficiently similar, then this union bound improves for $\bar{\mathcal{H}}$. Concretely, assume that all pairs of hypotheses

$h_1, h_2 \in \bar{\mathcal{H}}$ have $\Pr_{\mathcal{D}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] = O(\tau/\ln(1/\tau))$. We argue that this implies that all hypotheses in $\bar{\mathcal{H}}$ satisfy the stronger guarantee that $|\text{er}_{\mathbf{S}}(h) - \text{er}_{\mathcal{D}}(h)| = O(\sqrt{\tau d/n})$ rather than just $O(\sqrt{\tau d \ln(1/\tau)/n})$, hence improving the accuracy obtained from ERM. The intuitive reason for this improvement is, that when all the hypotheses in $\bar{\mathcal{H}}$ are very similar, it suffices to bound $|\text{er}_{\mathbf{S}}(h) - \text{er}_{\mathcal{D}}(h)|$ for one $h \in \bar{\mathcal{H}}$ and to bound $|(\text{er}_{\mathbf{S}}(h') - \text{er}_{\mathbf{S}}(h)) - (\text{er}_{\mathcal{D}}(h') - \text{er}_{\mathcal{D}}(h))|$ for all other $h' \in \bar{\mathcal{H}}$. Since $|\text{er}_{\mathcal{D}}(h') - \text{er}_{\mathcal{D}}(h)| = O(\Pr_{\mathcal{D}}[h(\mathbf{x}) \neq h'(\mathbf{x})]) = O(\tau/\ln(1/\tau))$, we get stronger concentration on $|(\text{er}_{\mathbf{S}}(h') - \text{er}_{\mathbf{S}}(h)) - (\text{er}_{\mathcal{D}}(h') - \text{er}_{\mathcal{D}}(h))|$ than each individual $|\text{er}_{\mathbf{S}}(h') - \text{er}_{\mathcal{D}}(h')|$.

Unfortunately, we have no guarantee that all pairs of hypotheses $h_1, h_2 \in \bar{\mathcal{H}}$ have $\Pr_{\mathcal{D}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] = O(\tau/\ln(1/\tau))$. Our next contribution is thus to find a way of exploiting the existence of two near-optimal hypotheses h_1, h_2 with $\Pr_{\mathcal{D}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] = \Omega(\tau/\ln(1/\tau))$ (i.e., a pair of disagreeing experts). Here we show that the conditional distribution of a sample (\mathbf{x}, \mathbf{y}) from \mathcal{D} with $h_1(\mathbf{x}) = h_2(\mathbf{x})$ is "easier"¹ than the distribution \mathcal{D} . In more detail, we know that $\text{er}_{\mathcal{D}}(h_j) = \tau + O(\sqrt{\tau d \ln(1/\tau)/n})$ for $j = 1, 2$. Since precisely one of them errs whenever $h_1(x) \neq h_2(x)$, and both or none err when $h_1(x) = h_2(x)$, we have that $\text{er}_{\mathcal{D}}(h_1) + \text{er}_{\mathcal{D}}(h_2) = \Pr_{\mathcal{D}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] + 2\Pr_{\mathcal{D}}[h_1(\mathbf{x}) \neq \mathbf{y} \wedge h_1(\mathbf{x}) = h_2(\mathbf{x})]$. Since they are both near-optimal, this implies $\Pr_{\mathcal{D}}[h_1(\mathbf{x}) \neq \mathbf{y} \wedge h_1(\mathbf{x}) = h_2(\mathbf{x})] = \tau + O(\sqrt{\tau d \ln(1/\tau)/n}) - \Omega(\tau/\ln(1/\tau))$. This is $\tau - \Omega(\tau/\ln(1/\tau))$ for τ sufficiently large (this assumption is one of the causes of the additive $\ln^5(n/d)(d + \ln(1/\delta))/n$ term in our upper bounds). Rewriting this also gives $\Pr_{\mathcal{D}}[h_1(\mathbf{x}) \neq \mathbf{y} \mid h_1(\mathbf{x}) = h_2(\mathbf{x})] = \Pr_{\mathcal{D}}[h_1(\mathbf{x}) = h_2(\mathbf{x})]^{-1}(\tau - \Omega(\tau/\ln(1/\tau)))$. Recalling that precisely one of h_1 and h_2 errs when they disagree, and that they are both near-optimal implies $\Pr_{\mathcal{D}}[h_1(\mathbf{x}) = h_2(\mathbf{x})] \geq 1 - O(\tau)$ and thus $\Pr_{\mathcal{D}}[h_1(\mathbf{x}) \neq \mathbf{y} \mid h_1(\mathbf{x}) = h_2(\mathbf{x})] = (1 + O(\tau))(\tau - \Omega(\tau/\ln(1/\tau))) = \tau - \Omega(\tau/\ln(1/\tau))$. What we have just argued is, that under the conditional distribution $\mathcal{D}_=$ of a sample $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$ with $h_1(\mathbf{x}) = h_2(\mathbf{x})$, there is a hypothesis $h_{\mathcal{D}_=}^* \in \mathcal{H}$ with $\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) = \tau - \Omega(\tau/\ln(1/\tau))$ (in particular, both h_1 and h_2 have this property). The distribution $\mathcal{D}_=$ is thus somewhat easier than \mathcal{D} since the optimal error under \mathcal{D} is τ .

Our next idea is to repeat the above argument recursively in order to drive $\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*)$ further down. More formally, if we can again find a pair of disagreeing experts h_1, h_2 for the distribution $\mathcal{D}_=$ and repeat this t times, then we end up with a list of pairs $(h_1^1, h_2^1), \dots, (h_1^t, h_2^t)$ such that under the distribution $\mathcal{D}_=$ of a sample $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$ conditioned on $\forall i : h_1^i(\mathbf{x}) = h_2^i(\mathbf{x})$, we have $\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) \leq \tau(1 - 1/\ln(1/\tau))^t$. After $t = O(\ln(1/\tau) \ln \ln(1/\tau))$ iterations, we have ensured $\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) \leq \tau/\ln(1/\tau)$. Empirical Risk Minimization on samples $\mathbf{S}_=$ with $\forall i : h_1^i(\mathbf{x}) = h_2^i(\mathbf{x})$ then gives a hypothesis with $\text{er}_{\mathcal{D}_=}(h_{\mathbf{S}_=}^*) = \text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) + O(\sqrt{(\tau/\ln(1/\tau))d \ln(\ln(1/\tau)/\tau)/n}) = \text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) + O(\sqrt{\tau d/n})$.

What remains is to handle samples with $h_1^i(\mathbf{x}) \neq h_2^i(\mathbf{x})$ for some i . We let \mathcal{D}_{\neq} denote the distribution \mathcal{D} conditioned on such a sample. Our key observation is that we can control the probability of receiving such a sample. Concretely, we show that $\Pr_{\mathcal{D}}[\exists i : h_1^i(\mathbf{x}) \neq h_2^i(\mathbf{x})] = \Theta(\tau)$. We thus expect to see $\Theta(\tau n)$ samples, denoted \mathbf{S}_{\neq} , from \mathcal{D}_{\neq} in a training set $\mathbf{S} \sim \mathcal{D}^n$. A completely naive invocation of the ERM Theorem, only assuming $\tau = O(1)$, shows that we find a hypothesis $h_{\mathbf{S}_{\neq}}^*$ with $\text{er}_{\mathcal{D}_{\neq}}(h_{\mathbf{S}_{\neq}}^*) = \text{er}_{\mathcal{D}_{\neq}}(h_{\mathcal{D}_{\neq}}^*) + O(\sqrt{d/|\mathbf{S}_{\neq}|}) = \text{er}_{\mathcal{D}_{\neq}}(h_{\mathcal{D}_{\neq}}^*) + O(\sqrt{d/(\tau n)})$. Note that the dependency on τ is very bad for this hypothesis, i.e., a $\sqrt{1/\tau}$ rather than $\sqrt{\tau}$. However, since samples with $h_1^i(\mathbf{x}) \neq h_2^i(\mathbf{x})$ are so rare, this turns out to be sufficient.

We now have all the ingredients for our algorithm. If we have obtained the pairs of disagreeing experts $(h_1^1, h_2^1), \dots, (h_1^t, h_2^t)$ and the two hypotheses $h_{\mathbf{S}_=}^*$ and $h_{\mathbf{S}_{\neq}}^*$, our final classifier does as follows on a new point $x \in \mathcal{X}$ without a label: First, it checks whether there is a pair with $h_1^i(x) \neq h_2^i(x)$. If so, it returns $h_{\mathbf{S}_{\neq}}^*(x)$. Otherwise, it returns $h_{\mathbf{S}_=}^*(x)$. If p denotes $\Pr_{\mathcal{D}}[\exists i : h_1^i(\mathbf{x}) \neq h_2^i(\mathbf{x})]$, then $p = O(\tau)$ and our final classifier $h_{\mathbf{S}}$

¹A similar argument was used in [9, 22] in the context of classification with an abstention option. The authors also use the disagreement sets of what we call the *experts* — candidates for being an output of a typical ERM. However, the authors of [9, 22] focus on either abstaining or learning the labels of the set of disagreements of pairs of experts, while in this work we use the fact that the conditional distribution of the set where two experts agree is "easier".

satisfies

$$\begin{aligned}
\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) &= p \text{er}_{\mathcal{D}_{\neq}}(h_{\mathbf{S}_{\neq}}^*) + (1-p) \text{er}_{\mathcal{D}_{=}}(h_{\mathbf{S}_{=}}^*) \\
&= p \text{er}_{\mathcal{D}_{\neq}}(h_{\mathcal{D}_{\neq}}^*) + O(p\sqrt{d/(\tau n)}) + (1-p) \text{er}_{\mathcal{D}_{=}}(h_{\mathcal{D}_{=}}^*) + O(\sqrt{\tau d/n}) \\
&= p \text{er}_{\mathcal{D}_{\neq}}(h_{\mathcal{D}}^*) + (1-p) \text{er}_{\mathcal{D}_{=}}(h_{\mathcal{D}}^*) + O(\sqrt{\tau d/n}) \\
&= \text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*) + O(\sqrt{\tau d/n}).
\end{aligned}$$

This completes the high level description of the key ideas in our new algorithm. Let us finally remark that we clearly do not have t training sets of size n each for training h_1^i, h_2^i for $i = 1, \dots, t$. Instead, we allocate around n/t samples for each iteration. This of course reduces the performance of any estimates based on ERM. However, we can show that this only matters for very small values of τ and thus is a second source of the additive $\ln^5(n/d)(d + \ln(1/\delta))/n$ term.

Lower bound for proper learners. Our lower bound proof is quite simple. Assume we wish to prove a lower bound on the error of a proper learner when the hypothesis set has VC-dimension d , we have n samples and $\text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*) = \tau$ for some τ . We construct an instance where the input domain \mathcal{X} is the discrete set x_1, \dots, x_u with $u \approx d/\tau$. We let the hypothesis set \mathcal{H} consist of all hypotheses returning -1 on precisely d of the u points. Finally, the unknown concept we are trying to learn is the all-1 concept. This hypothesis class is routinely used in the existing lower bounds, and in fact corresponds to the hardest case under Massart's noise condition [21, 23, 15, 30]. Note that \mathcal{H} does not contain the all-1 concept; thus, we cannot simply choose the proper learner that always outputs this concept.

Assume we have some proper learning algorithm \mathcal{A} for this hypothesis set and input domain \mathcal{X} . We now consider a number of different data distributions $\mathcal{D}_1, \mathcal{D}_2, \dots$, and argue that there is at least one of the distributions \mathcal{D}_i under which \mathcal{A} often (with constant probability) produces a hypothesis $h_{\mathbf{S}}$ with $\text{er}_{\mathcal{D}_i}(h_{\mathbf{S}}) = \tau + \Omega(\sqrt{\tau \ln(1/\tau)d/n})$ when $\mathbf{S} \sim \mathcal{D}_i^n$.

The distributions we consider each corresponds to a hypothesis $h \in \mathcal{H}$. The distribution \mathcal{D}_h returns each point $x \in \mathcal{X}$ such that $h(x) = -1$ with probability $1/u - \alpha$. For the remaining points, the distribution returns them with probability $1/u + f(\alpha, d, u)$ such that we get a probability distribution (thus $f(\alpha, d, u) < \alpha$ for $u \geq 2d$). Since the unknown concept/true labeling function is the all-1 function, we have that any hypothesis $h \in \mathcal{H}$ errs precisely when it returns -1 . Thus in particular, the best hypothesis under \mathcal{D}_h is h and that hypothesis has $\text{er}_{\mathcal{D}_h}(h) = d(1/u - \alpha)$. This is the value we set to τ by choosing u and α appropriately.

Now consider choosing one of the distributions \mathcal{D}_h uniformly at random and running \mathcal{A} on $\mathbf{S} \sim \mathcal{D}_h^n$. Since \mathcal{A} is proper, it has to return a hypothesis in \mathcal{H} . This means that it has to choose d points x_i on which to return -1 . Now crucially, if a constant fraction of those are chosen such that $h(x_i) = 1$, then $\text{er}_{\mathcal{D}_h}(h_{\mathbf{S}}) = \tau + \Omega(d\alpha)$. Intuitively, since the points with $h(x_i) = -1$ receive the least probability mass under \mathcal{D}_h , and \mathcal{A} does not know the distribution \mathcal{D}_h , the best strategy for \mathcal{A} is to output the hypothesis $h_{\mathbf{S}}$ returning -1 on the d points x_i from which there are fewest copies in the sample \mathbf{S} . We expect to see $n(1/u + f(\alpha)) \leq n/u + \alpha n$ copies of each x_i with $h(x_i) = 1$ and we expect to see $n(1/u - \alpha) = n/u - \alpha n$ copies of each point with $h(x_i) = -1$. A simple application of Chebyshev's inequality implies that with constant probability, it holds for at least half the points x_i with $h(x_i) = -1$ that we see $n_i \geq n/u - \alpha n - O(\sqrt{n/u})$ copies of it in \mathbf{S} . Now for the points x_i with $h(x_i) = 1$, by anti-concentration, we see no more than $n_i = n/u + \alpha n - \Omega(\sqrt{n \ln(u/d)/u})$ copies with probability roughly d/u . We thus expect to see $\Omega(d)$ such points with $n_i = n/u + \alpha n - \Omega(\sqrt{n \ln(u/d)/u})$. If $\sqrt{n \ln(u/d)/u} > c\alpha n$ for a large enough $c > 0$, this implies we have fewer copies of these points and \mathcal{A} will return -1 on at least $d/2$ of them. We can thus choose $\alpha = \Theta(\sqrt{\ln(u/d)/(un)}) = \Theta(\sqrt{\tau \ln(1/\tau)/(dn)})$ and conclude $\text{er}_{\mathcal{D}_h}(h_{\mathbf{S}}) = \tau + \Omega(d\alpha) = \tau + \Omega(\sqrt{\tau d \ln(1/\tau)/n})$ as claimed.

2 Near-Optimal Agnostic PAC Learner

In this section, we present our new agnostic PAC learner, DISAGREEINGEXPERTS (Algorithm 1), with an optimal error bound except for very small values of $\tau = \inf_{h \in \mathcal{H}} \text{er}_{\mathcal{D}}(h)$. The guarantees of Algorithm 1 are

stated in our main upper bound result, Theorem 3. To simplify the analysis, DISAGREEINGEXPERTS ensures that we may focus on analysing a subroutine COREDISAGREEINGEXPERTS (Algorithm 2) under the following simplifying assumptions:

1. $n \geq c_n \ln^{3.5}(n/d)(d + \ln(1/\delta))$ for large enough constant $c_n > 0$.
2. $c_\tau \ln^9(n/d)(d + \ln(1/\delta))/n \leq \tau \leq 1/c_\tau$ for large enough constant $c_\tau > 0$.
3. We have an estimate $\tilde{\tau} \in [\tau/2, 2\tau]$ available.

Under these assumptions, we show that the subroutine COREDISAGREEINGEXPERTS (Algorithm 2) with probability at least $1 - \delta$ over a training set $\mathbf{S} \sim \mathcal{D}^n$, returns a hypothesis $h_{\mathbf{S}}$ with $\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) \leq \tau + O(\sqrt{\tau(d + \ln(1/\delta))/n})$.

We start by justifying these assumptions before delving into the details of the analysis. *Crucially*, our full algorithm DISAGREEINGEXPERTS needs *none* of these assumption, they are merely to ease the presentation and analysis of the main part of our algorithm, COREDISAGREEINGEXPERTS.

Simplifying assumptions. For assumption 1., notice that our claimed upper bound in Theorem 3 on the error of $h_{\mathbf{S}}$ exceeds 1 for smaller n and thus is trivially true. The algorithm DISAGREEINGEXPERTS, shown as Algorithm 1, takes care of assumptions 2. and 3.

Algorithm 1: DISAGREEINGEXPERTS($S, \mathcal{H}, d, \delta$)

Input: Training set S of n samples $\{(x_i, y_i)\}_{i=1}^n$ with $(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$, hypothesis set \mathcal{H} of VC-dimension d , failure parameter $\delta > 0$.

Result: Classifier $h_S : \mathcal{X} \rightarrow \{-1, 1\}$.

- 1 Partition S into three sets S_1, S_2, S_3 of $n/3$ samples each.
 - 2 Let $\tilde{\tau} \leftarrow \text{er}_{S_1}(h_{S_1}^*)$.
 - 3 Run COREDISAGREEINGEXPERTS($S_2, \mathcal{H}, d, \delta, \tilde{\tau}$) to obtain hypothesis h_1 .
 - 4 Run ERM on S_2 to obtain hypothesis h_2 .
 - 5 **return** $h_S \in \{h_1, h_2\}$ with smallest $\text{er}_{S_3}(h_S)$.
-

Given a training set $\mathbf{S} \sim \mathcal{D}^n$, DISAGREEINGEXPERTS first splits the training set into 3 sets $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3$ of $n/3$ samples each. It then computes the error $\tilde{\tau}$ of the best hypothesis $h_{\mathbf{S}_1}^*$ in \mathcal{H} on \mathbf{S}_1 . By the ERM Theorem (Theorem 1) and assumption 1. (that we already justified), the estimate $\tilde{\tau}$ satisfies $\tilde{\tau} \in [\tau/2, 2\tau]$ with probability $1 - \delta$.

It then invokes COREDISAGREEINGEXPERTS on \mathbf{S}_2 using this estimate $\tilde{\tau}$ to obtain a hypothesis \mathbf{h}_1 . This justifies assumption 3. (by rescaling δ by a constant factor).

It also runs ERM on \mathbf{S}_2 to obtain a hypothesis \mathbf{h}_2 . Finally, it uses \mathbf{S}_3 as a validation set to estimate $\text{er}_{\mathcal{D}}(\mathbf{h}_1)$ and $\text{er}_{\mathcal{D}}(\mathbf{h}_2)$ to within additive (by Chernoff):

$$|\text{er}_{\mathbf{S}_3}(\mathbf{h}_i) - \text{er}_{\mathcal{D}}(\mathbf{h}_i)| = O(\sqrt{\text{er}_{\mathcal{D}}(\mathbf{h}_i) \ln(1/\delta)/n} + \ln(1/\delta)/n).$$

Returning the hypothesis among $\mathbf{h}_1, \mathbf{h}_2$ with the least $\text{er}_{\mathbf{S}_3}(\mathbf{h}_i)$ ensures that the final hypothesis $h_{\mathbf{S}}$ has error at most

$$\min_i \text{er}_{\mathcal{D}}(\mathbf{h}_i) + O(\sqrt{\text{er}_{\mathcal{D}}(\mathbf{h}_i) \ln(1/\delta)/n} + \ln(1/\delta)/n).$$

By the guarantee claimed above for Algorithm 2, this is at most $\tau + O(\sqrt{\tau(d + \ln(1/\delta))/n})$ when $c_\tau \ln^9(n/d)(d + \ln(1/\delta))/n \leq \tau \leq 1/c_\tau$ (i.e. under assumption 2.). For smaller τ , the ERM Theorem (Theorem 1) guarantees that \mathbf{h}_2 has an error of at most

$$\tau + O\left(\sqrt{\frac{\tau(d \ln(n/d) + \ln(1/\delta))}{n}} + \frac{d \ln(n/d) + \ln(1/\delta)}{n}\right) = \tau + O\left(\frac{\ln^5(n/d)(d + \ln(1/\delta))}{n}\right).$$

Note that we have upper bounded $\tau \ln(1/\tau)$ in the first term by $\tau \ln(n/d)$ since the second term dominates for $\tau \ll d/n$. Similarly, for $\tau > 1/c_\tau$, the ERM Theorem guarantees that \mathbf{h}_2 has an error of at most

$$\tau + O\left(\sqrt{(d + \ln(1/\delta))/n}\right) = \tau + O\left(\sqrt{\tau(d + \ln(1/\delta))/n}\right).$$

This completes the justifications for assumptions 1., 2., and 3. We now proceed to analyzing the main part of our new algorithm, denoted COREDISAGREEINGEXPERTS, under these assumptions.

2.1 Core algorithm

Our algorithm COREDISAGREEINGEXPERTS is shown as Algorithm 2, where we define

$$\alpha(n, d, \delta, \beta) := c_\alpha \left(\sqrt{\frac{\beta(d \ln(1/\beta) + \ln(1/\delta))}{n}} + \frac{d \ln(n/d) + \ln(1/\delta)}{n} \right)$$

with c_α a sufficiently large constant. In Algorithm 2, the two parameters c_t, c_Z are also sufficiently large constants (in particular, c_Z is sufficiently larger than $c_\alpha + c_t$ and c_α is sufficiently larger than the constant hiding in the $O(\cdot)$ -notation of the ERM Theorem (Theorem 1)).

Algorithm 2: COREDISAGREEINGEXPERTS($S, \mathcal{H}, d, \delta, \tilde{\tau}$)

Input: Training set S of $2n$ samples $\{(x_i, y_i)\}_{i=1}^{2n}$ with $(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$, hypothesis set \mathcal{H} of VC-dimension d , failure parameter $\delta > 0$, estimate $\tilde{\tau} \in [\tau/2, 2\tau]$.

Result: Classifier $h_S : \mathcal{X} \rightarrow \{-1, 1\}$.

- 1 Partition S into two sets B, C of n samples each.
 - 2 Let $t \leftarrow c_t \ln(1/\tilde{\tau}) \ln \ln(1/\tilde{\tau})$.
 - 3 Let $Z_t \leftarrow c_Z \cdot t \ln^2(n/d) (d \ln(n/d) + \ln(1/\delta)) / n$.
 - 4 Partition B into t sets B^1, \dots, B^t of n/t samples each.
 - 5 $r \leftarrow 0$
 - 6 **for** $i = 1, \dots, t$ **do**
 - 7 Let $T^i \subseteq B^i$ be the samples in B^i with $h_1^j(x) = h_2^j(x)$ for all $j < i$.
 - 8 Run ERM on T^i to obtain a hypothesis $h_{T^i}^*$.
 - 9 Let $\gamma_i \leftarrow \text{er}_{T^i}(h_{T^i}^*)$.
 - 10 **if** $\gamma_i \leq Z_t$ **then**
 - 11 | **break**
 - 12 Let $\mathcal{H}^i \subseteq \mathcal{H}$ be the hypotheses $h \in \mathcal{H}$ with $\text{er}_{T^i}(h) \leq \gamma_i + \alpha(n/t, d, \delta, \gamma_i)$.
 - 13 **if** there is no pair $h_1, h_2 \in \mathcal{H}^i$ with $\Pr_{T^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq \gamma_i / \ln(1/\gamma_i)$ **then**
 - 14 | **break**
 - 15 **else**
 - 16 | Let $h_1^i \leftarrow h_1$ and $h_2^i \leftarrow h_2$ for a pair $h_1, h_2 \in \mathcal{H}^i$ with $\Pr_{T^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq \gamma_i / \ln(1/\gamma_i)$.
 - 17 | $r \leftarrow i$
 - 18 Partition C into two sets $C_=$ and C_\neq where $C_=$ contains all x with $h_1^i(x) = h_2^i(x)$ for all $i = 1, \dots, r$ and C_\neq contains the remaining.
 - 19 Run ERM on $C_=$ to obtain a hypothesis $h_{C_=}^*$.
 - 20 Run ERM on C_\neq to obtain a hypothesis $h_{C_\neq}^*$.
 - 21 Let h_S be the classifier that on an input x checks whether $h_1^i(x) = h_2^i(x)$ for all $i = 1, \dots, r$. If so, h_S returns $h_{C_=}^*(x)$ and otherwise it returns $h_{C_\neq}^*(x)$.
 - 22 **return** h_S .
-

Recall that our goal is to show that under assumptions 1., 2. and 3., it holds with probability at least $1 - \delta$ over a training set $\mathbf{S} \sim \mathcal{D}^{2n}$, that Algorithm 2 returns a hypothesis $h_{\mathbf{S}}$ with $\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) \leq$

$\tau + O(\sqrt{\tau(d + \ln(1/\delta))/n})$. We note that the algorithm is presented as if given a training set of size $2n$, not n . This is merely to make the constants simpler and only affects the generalization error by a constant factor after rescaling n with $n/2$.

Brief overview. Before giving the details of the analysis, let us briefly discuss the steps of Algorithm 2, introduce some notation and give the high level ideas in the analysis. Assume we run Algorithm 2 on a data set $\mathbf{S} \sim \mathcal{D}^{2n}$. The data set is first split into two pieces \mathbf{B}, \mathbf{C} of size n each.

We start by partitioning \mathbf{B} into t pieces $\mathbf{B}^1, \dots, \mathbf{B}^t$ of n/t samples each and execute the for-loop in steps 6-17. The goal of these steps is to obtain hypotheses \mathbf{h}_1^i and \mathbf{h}_2^i that are both close to optimal and yet disagree a lot in their predictions, i.e., disagreeing experts. In each step of the loop, we gather the set \mathbf{T}^i of samples $(\mathbf{x}, \mathbf{y}) \in \mathbf{B}^i$ for which $\mathbf{h}_1^j(\mathbf{x}) = \mathbf{h}_2^j(\mathbf{x})$ for all $j < i$, i.e., none of the previous pairs disagree on \mathbf{x} (for $i = 1$, we have $\mathbf{T}^1 = \mathbf{B}^1$). Now consider any fixed outcome B^1, \dots, B^{i-1} of $\mathbf{B}^1, \dots, \mathbf{B}^{i-1}$ and $h_1^1, h_2^1, \dots, h_1^{i-1}, h_2^{i-1}$ of $\mathbf{h}_1^1, \mathbf{h}_2^1, \dots, \mathbf{h}_1^{i-1}, \mathbf{h}_2^{i-1}$. The samples in \mathbf{T}^i are i.i.d. from \mathcal{D} conditioned on $h_1^j(\mathbf{x}) = h_2^j(\mathbf{x})$ for all $j < i$. Denote this conditional distribution by \mathcal{D}^i . Steps 8-9 estimate the best possible error $\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)$ achievable under \mathcal{D}^i . If this error is sufficiently small, we exit the for-loop in step 11.

If not, we gather the subset of hypotheses \mathbf{H}^i that are near-optimal on the data set \mathbf{T}^i (step 12). Among these, we look for a pair h_1, h_2 that disagree on many predictions in \mathbf{T}^i . If there is no such pair, we exit the for-loop in step 14. Finally, if there is, we let \mathbf{h}_1^i and \mathbf{h}_2^i be an arbitrary such pair.

Once the for-loop has completed, we use the obtained pairs $\mathbf{h}_1^i, \mathbf{h}_2^i$ to partition the samples in \mathbf{C} into two sets $\mathbf{C}_=$ and \mathbf{C}_\neq , where \mathbf{C}_\neq contains the samples $(x, y) \in \mathbf{C}$ where at least one i has $\mathbf{h}_1^i(x) \neq \mathbf{h}_2^i(x)$ and $\mathbf{C}_=$ contains the remaining. We finally run ERM on each of the two sets to obtain hypotheses $h_{\mathbf{C}_=}^*$ and $h_{\mathbf{C}_\neq}^*$.

The intuition for why the above works was also discussed in Section 1.1. We repeat the main ideas here in context of the full algorithm description. First, if we exit the for-loop before having completed all t steps, then either it was possible to obtain a very small error on \mathbf{T}^i (step 10-11) or there was no pair h_1^i, h_2^i that disagree on many predictions (step 13-14). In the former case, ERM on $\mathbf{C}_=$ ensures that $h_{\mathbf{C}_=}^*$ makes few mistakes on samples from \mathcal{D} where $\mathbf{h}_1^i(\mathbf{x}) = \mathbf{h}_2^i(\mathbf{x})$ for all i . Denote the distribution of such a sample by $\mathbf{D}_=$. In the latter case, since all hypotheses that are near-optimal on \mathbf{T}^i make almost the same predictions, the ERM bounds improve for $\mathbf{C}_=$. If we complete all t iterations of the for-loop, then we will show that each step decreases $\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*)$ enough that $\text{er}_{\mathbf{D}_=}(h_{\mathbf{D}_=}^*) \leq \tau/\ln(1/\tau)$. With this reduced error, the additive mistakes resulting from ERM is down-scaled sufficiently to cancel out the $\sqrt{\ln(1/\tau)}$ factor of sub-optimality.

Finally, for the set \mathbf{C}_\neq , we will show that $\Pr_{\mathcal{D}}[\exists i : \mathbf{h}_1^i(\mathbf{x}) \neq \mathbf{h}_2^i(\mathbf{x})] \leq O(\tau)$. Thus, when we run ERM on \mathbf{C}_\neq , we can afford to merely upper bound the error of $h_{\mathbf{C}_\neq}^*$ by $\text{er}_{\mathbf{D}_\neq}(h_{\mathbf{D}_\neq}^*) + O(\sqrt{(d + \ln(1/\delta))/(\tau n)})$. This is because we only see such a sample with probability $O(\tau)$ and thus the additive error contributes only $O(\tau\sqrt{(d + \ln(1/\delta))/(\tau n)}) = O(\sqrt{\tau(d + \ln(1/\delta))/n})$. Here \mathbf{D}_\neq denotes the conditional distribution of a sample (\mathbf{x}, \mathbf{y}) from \mathcal{D} conditioned on there being at least one i for which $\mathbf{h}_1^i(\mathbf{x}) \neq \mathbf{h}_2^i(\mathbf{x})$.

Analysis. We are now ready for the formal correctness proof. Let $\mathbf{S} \sim \mathcal{D}^{2n}$ denote a random training set of $2n$ samples and let \mathbf{B} and \mathbf{C} be the respective sets of size n constructed by Algorithm 2.

We first argue that once the for-loop in steps 6-17 of Algorithm 2 terminates, the two distributions $\mathbf{D}_=$ and \mathbf{D}_\neq have the following desirable properties:

Lemma 1. *It holds with probability at least $1 - \delta/2$ over $\mathbf{B} \sim \mathcal{D}^n$, that upon termination of the for-loop, we have $\Pr_{\mathcal{D}}[\exists i : \mathbf{h}_1^i(\mathbf{x}) \neq \mathbf{h}_2^i(\mathbf{x})] \leq 8\tau$ and:*

- *For any $m \geq n/2$, it holds with probability at least $1 - \delta/8$ over a set $\mathbf{C}_= \sim \mathbf{D}_=^m$ that $\text{er}_{\mathbf{D}_=}(h_{\mathbf{C}_=}^*) = \text{er}_{\mathbf{D}_=}(h_{\mathbf{D}_=}^*) + O(\sqrt{\tau(d + \ln(1/\delta))/n})$.*

Before proving Lemma 1, we show that it suffices to establish our claim on $\text{er}_{\mathcal{D}}(h_{\mathbf{S}})$ and thus completes the proof of Theorem 3.

Fix an arbitrary outcome B of \mathbf{B} for which the properties in Lemma 1 are satisfied upon termination. This also fixed $\mathbf{h}_1^i, \mathbf{h}_2^i$ to some h_1^i, h_2^i and $\mathbf{D}_=$ and \mathbf{D}_\neq to some $\mathcal{D}_=$ and \mathcal{D}_\neq . The set \mathbf{C} still consists of i.i.d. samples from \mathcal{D} as \mathbf{C} is not used in the for-loop.

Define $p := \Pr_{\mathcal{D}}[\exists i : h_1^i(\mathbf{x}) \neq h_2^i(\mathbf{x})]$. We start by showing properties of $h_{\mathbf{C}_{\neq}}^*$ when $p \geq c_p \ln(1/\delta)/n$ for a large enough constant c_p . Under this assumption on p , by Chernoff, we have $|\mathbf{C}_{\neq}| \geq (p/2)n$ except with probability $1 - \delta/8$. In this case, it follows from the ERM Theorem (Theorem 1) that with probability $1 - \delta/8$, $h_{\mathbf{C}_{\neq}}^*$ has

$$\text{er}_{\mathcal{D}_{\neq}}(h_{\mathbf{C}_{\neq}}^*) = \text{er}_{\mathcal{D}_{\neq}}(h_{\mathcal{D}_{\neq}}^*) + O(\sqrt{(d + \ln(1/\delta))/(pn)}).$$

Next, we show properties of $h_{\mathbf{C}_{=}}^*$. Since we assume $\tau \leq 1/c_\tau$ for a big enough constant c_τ , we must have $(1 - p) \geq 1 - 8\tau \geq 3/4$. Since n is assumed sufficiently large, this implies that with probability at least $1 - \delta/8$, we have $|\mathbf{C}_{=}| \geq n/2$. Conditioned on this, by the properties in Lemma 1, we have with probability at least $1 - \delta/8$ that

$$\text{er}_{\mathcal{D}_{=}}(h_{\mathbf{C}_{=}}^*) = \text{er}_{\mathcal{D}_{=}}(h_{\mathcal{D}_{=}}^*) + O\left(\sqrt{\frac{\tau(d + \ln(1/\delta))}{n}}\right).$$

The returned hypothesis $h_{\mathbf{S}}$ thus satisfies

$$\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) = p \text{er}_{\mathcal{D}_{\neq}}(h_{\mathbf{C}_{\neq}}^*) + (1 - p) \text{er}_{\mathcal{D}_{=}}(h_{\mathbf{C}_{=}}^*).$$

If $p < c_p \ln(1/\delta)/n$ (and by assumption that $\tau \geq c_\tau \ln^9(n/d)(d + \ln(1/\delta))/n$), this gives

$$\begin{aligned} \text{er}_{\mathcal{D}}(h_{\mathbf{S}}) &\leq p + (1 - p) \text{er}_{\mathcal{D}_{=}}(h_{\mathbf{C}_{=}}^*) \\ &= p + (1 - p) \text{er}_{\mathcal{D}_{=}}(h_{\mathcal{D}_{=}}^*) + O\left(\sqrt{\frac{\tau(d + \ln(1/\delta))}{n}}\right) \\ &= (1 - p) \text{er}_{\mathcal{D}_{=}}(h_{\mathcal{D}_{=}}^*) + O\left(\sqrt{\frac{\tau(d + \ln(1/\delta))}{n}}\right) \\ &= \text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*) + O\left(\sqrt{\frac{\tau(d + \ln(1/\delta))}{n}}\right). \end{aligned}$$

If $p \geq c_p \ln(1/\delta)/n$, we have from Lemma 1 that $p \leq 8\tau$ and thus

$$\begin{aligned} \text{er}_{\mathcal{D}}(h_{\mathbf{S}}) &= p \text{er}_{\mathcal{D}_{\neq}}(h_{\mathbf{C}_{\neq}}^*) + (1 - p) \text{er}_{\mathcal{D}_{=}}(h_{\mathbf{C}_{=}}^*) + p \cdot O\left(\sqrt{\frac{d + \ln(1/\delta)}{pn}}\right) + O\left(\sqrt{\frac{\tau(d + \ln(1/\delta))}{n}}\right) \\ &= p \text{er}_{\mathcal{D}_{\neq}}(h_{\mathcal{D}_{\neq}}^*) + (1 - p) \text{er}_{\mathcal{D}_{=}}(h_{\mathcal{D}_{=}}^*) + O\left(\sqrt{\frac{p(d + \ln(1/\delta))}{n}}\right) + O\left(\sqrt{\frac{\tau(d + \ln(1/\delta))}{n}}\right) \\ &= \text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*) + O\left(\sqrt{\frac{\tau(d + \ln(1/\delta))}{n}}\right). \end{aligned}$$

This completes the proof of Theorem 3. What remains is thus to establish Lemma 1, which is the focus of the next subsection.

2.2 Progress on termination (proof of Lemma 1)

In this section, we prove Lemma 1. Intuitively, termination of the for-loop results in the properties claimed in Lemma 1 provided that the estimates based on performance on the \mathbf{T}^i 's are sufficiently accurate. To formalize this, we define a number of natural *failure events* relating to the accuracy of these.

Failure events. We define a number of bad events that we argue rarely occur. Let \mathbf{D}^i be the random variable giving the distribution of a sample $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$ conditioned on $\forall j < i : \mathbf{h}_1^j(\mathbf{x}) = \mathbf{h}_2^j(\mathbf{x})$. Note that this random variable is determined from $\mathbf{B}^1, \dots, \mathbf{B}^{i-1}$. The samples \mathbf{T}^i are then i.i.d. from \mathbf{D}^i . We define two failure events relating to how well \mathbf{T}^i represents \mathbf{D}^i :

1. Let $E_{i,0}$ be the event that Algorithm 2 reaches iteration i , none of the events $E_{j,0}, E_{j,1}$ occurred for $i < j$ and there is a hypothesis $h \in \mathcal{H}$ with $|\text{er}_{\mathbf{D}^i}(h) - \text{er}_{\mathbf{T}^i}(h)| > (1/32)\alpha(n/t, d, \delta, \min\{\text{er}_{\mathbf{D}^i}(h), \text{er}_{\mathbf{T}^i}(h)\})$.
2. Let $E_{i,1}$ be the event that Algorithm 2 reaches iteration i , none of the events $E_{j,0}, E_{j,1}$ occurred for $i < j$ and there is a pair of hypotheses h_1, h_2 with $|\text{Pr}_{\mathbf{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] - \text{Pr}_{\mathbf{T}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]| > (1/32)\alpha(n/t, d, \delta, \min\{\text{Pr}_{\mathbf{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})], \text{Pr}_{\mathbf{T}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]\})$.

We will show that these events are unlikely:

Lemma 2. *For all i , we have $\Pr[E_{i,0}] \leq \delta/(4t)$ and $\Pr[E_{i,1}] \leq \delta/(4t)$.*

We also show that when none of the events occur, the hypotheses and execution of Algorithm 2 satisfies the following:

Observation 1. *Assume none of the events $E_{j,0}$ and $E_{j,1}$ occur for $j \leq i$ and that Algorithm 2 does not terminate before iteration i . Then if $\gamma_i = \text{er}_{\mathbf{T}^i}(h_{\mathbf{T}^i}^*) \leq Z_t$, it holds that $\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*) \leq 2Z_t$. If $\gamma_i > Z_t$, then each of the following hold:*

- $\gamma_i \leq \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*) + \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*)/\ln(1/\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*)) \leq 2\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*)$.
- Every hypothesis h in \mathbf{H}^i satisfies $\text{er}_{\mathbf{D}^i}(h) \leq \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*) + (1/8)\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*)/\ln(1/\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*))$.
- Every hypothesis $h \in \mathcal{H}$ with $\text{er}_{\mathbf{D}^i}(h) \leq \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*) + (1/8)\alpha(n/t, d, \delta, \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*))$ is in \mathbf{H}^i .
- Every pair of hypotheses h_1, h_2 with $\text{Pr}_{\mathbf{T}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq \gamma_i/\ln(1/\gamma_i)$ satisfy $\text{Pr}_{\mathbf{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq (1/2)\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*)/\ln(1/\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*))$.
- Every pair of hypotheses h_1, h_2 with $\text{Pr}_{\mathbf{T}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] < \gamma_i/\ln(1/\gamma_i)$ satisfies $\text{Pr}_{\mathbf{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \leq 4\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*)/\ln(1/\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*))$.

The proof of Lemma 2 mostly uses standard concentration results for classes with bounded VC-dimension and has thus been deferred to Appendix A.1. Similarly, the proof of Observation 1 merely uses the definition of α and Z_t and has thus been deferred to Appendix A.2.

More interestingly, we show that if none of the events $E_{i,j}$ occur, then the for-loop makes progress towards reducing $\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*)$ in each iteration:

Lemma 3. *For any integer $1 \leq i \leq t+1$, assume none of the events $E_{j,0}$ and $E_{j,1}$ occurred for $j < i$ and that Algorithm 2 did not terminate with $r < i-1$. Then $\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*) \leq \text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*)(1 - 1/(32\ln(1/\text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*))))^{i-1}$ and $\Pr_{\mathcal{D}}[\exists j \leq i : \mathbf{h}_1^j(\mathbf{x}) \neq \mathbf{h}_2^j(\mathbf{x})] \leq 8(\text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*) - \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*))$.*

We prove Lemma 3 in the next subsection and for now focus on completing the proof of Lemma 1 from Lemma 2, Observation 1 and Lemma 3. We have restated it here for convenience:

Restatement of Lemma 1. *It holds with probability at least $1 - \delta/2$ over $\mathbf{B} \sim \mathcal{D}^n$, that upon termination of the for-loop, we have $\Pr_{\mathcal{D}}[\exists i : \mathbf{h}_1^i(\mathbf{x}) \neq \mathbf{h}_2^i(\mathbf{x})] \leq 8\tau$ and:*

- For any $m \geq n/2$, it holds with probability at least $1 - \delta/8$ over a set $\mathbf{C}_= \sim \mathbf{D}_=^m$ that $\text{er}_{\mathbf{D}_=}(h_{\mathbf{C}_=}^*) = \text{er}_{\mathbf{D}_=}(h_{\mathbf{D}_=}^*) + O(\sqrt{\tau(d + \ln(1/\delta))/n})$.

Proof of Lemma 1. From Lemma 2 and a union bound, we conclude that with probability at least $1 - \delta/2$, none of the events $E_{i,j}$ occur. We show that conditioned on this, the properties claimed in Lemma 1 hold. So fix an outcome B of \mathbf{B} where the events did not occur. This also fixes $\mathbf{h}_1^i, \mathbf{h}_2^i, \mathbf{T}^i, \mathbf{H}^i, \mathbf{D}_i, \mathbf{D}_=$ and

\mathcal{D}_{\neq} to some $h_1^i, h_2^i, T^i, \mathcal{H}^i, \mathcal{D}^i, \mathcal{D}_=$ and \mathcal{D}_{\neq} . Let r denote the value of the variable r in Algorithm 2 upon termination.

First, recall that upon termination, we have $\mathcal{D}_= \simeq \mathcal{D}^{r+1}$. Since none of the events $E_{i,j}$ occurred for any $i \leq t$, we have from Lemma 3 (with $i = r + 1$) that $\Pr_{\mathcal{D}}[\exists j \leq r : h_1^j(\mathbf{x}) \neq h_2^j(\mathbf{x})] \leq 8(\text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*) - \text{er}_{\mathcal{D}^{r+1}}(h_{\mathcal{D}^{r+1}}^*)) \leq 8\tau$. This establishes the first claim in Lemma 1.

For the second claim, we split the proof in several cases depending on how the for-loop in Algorithm 2 terminates on B . The main observations, which we will expand upon below, are: 1. if we terminate in step 11, then $\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) = O(Z_t) = O(\sqrt{\tau(d + \ln(1/\delta))/n})$, 2. if we terminate by completing all iterations of the for-loop, then $\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) \leq \tau/\ln(1/\tau)$. Both of these are sufficient to show that the ERM Theorem (Theorem 1) on \mathcal{C}_{\neq} is good enough. Finally, if we terminate in step 14, we carefully exploit that all near-optimal hypotheses agree on most samples. This allows for a better guarantee on ERM than invoking the ERM Theorem.

The easiest cases are termination in step 11 and completion of the for-loop, so we argue for those first.

Termination in Step 11. Since we terminate in step 11, we must have $\gamma_{r+1} \leq Z_t$. By Observation 1, this implies $\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) = \text{er}_{\mathcal{D}^{r+1}}(h_{\mathcal{D}^{r+1}}^*) \leq 2Z_t$.

Since $\tilde{\tau} \in [\tau/2, 2\tau]$, we have $t = O(\ln(1/\tau) \ln \ln(1/\tau))$ and since we assume $\tau > d/n$, we have $Z_t = O(\ln(n/d) \ln \ln(n/d) \ln^3(n/d)(d + \ln(1/\delta))/n)$. By the ERM Theorem (Theorem 1) and since we assume n sufficiently large, we have that for any $m \geq n/2$, it holds with probability at least $1 - \delta/8$ over a set $\mathcal{C}_= \sim \mathcal{D}_=^m$ that $\text{er}_{\mathcal{D}_=}(h_{\mathcal{C}_=}^*) \leq 4Z_t$. Since we assume $\tau \geq c_{\tau} \ln^9(n/d)(d + \ln(1/\delta))/n$ for large enough c_{τ} , we have $\sqrt{\tau(d + \ln(1/\delta))/n} \geq 4Z_t$ and thus $\text{er}_{\mathcal{D}_=}(h_{\mathcal{C}_=}^*) \leq \sqrt{\tau(d + \ln(1/\delta))/n}$ with probability at least $1 - \delta/8$ over $\mathcal{C}_=$.

Termination by completion. Since none of the events $E_{i,j}$ occurred and we terminate upon completing the for-loop, we have $r \leftarrow t$ and we get from Lemma 3 (with $i = t + 1 = r + 1$) that $\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) = \text{er}_{\mathcal{D}^{r+1}}(h_{\mathcal{D}^{r+1}}^*) \leq \tau(1 - 1/(32 \ln(1/\tau)))^t$. This is at most $\tau \exp(-t/(32 \ln(1/\tau)))$. Since $\tilde{\tau} \geq \tau/2$ we have $t \geq 32 \ln(1/\tau) \ln \ln(1/\tau)$ for c_t large enough. Thus $\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) \leq \tau/\ln(1/\tau)$. The ERM Theorem (Theorem 1) now implies that with probability at least $1 - \delta/8$ over a set $\mathcal{C}_= \sim \mathcal{D}_=^m$ with $m \geq n/2$, we have

$$\text{er}_{\mathcal{D}_=}(h_{\mathcal{C}_=}^*) = \text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) + O\left(\sqrt{\frac{\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*)(d \ln(\frac{1}{\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*)}) + \ln(1/\delta))}{n}} + \frac{d \ln(n/d) + \ln(1/\delta)}{n}\right).$$

For $\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) \leq \tau/\ln(1/\tau)$, we have $\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) \ln(1/\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*)) = O(\tau)$ and thus since we assume $\tau \geq c_{\tau} \ln^9(n/d)(d + \ln(1/\delta))$, we conclude

$$\text{er}_{\mathcal{D}_=}(h_{\mathcal{C}_=}^*) = \text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) + O\left(\sqrt{\frac{\tau(d + \ln(1/\delta))}{n}}\right).$$

Termination in Step 14. Assume we terminate in step 14 of some iteration i and let $r \leftarrow i - 1$. Then by definition of Algorithm 2, there is no pair $h_1, h_2 \in \mathcal{H}^i$ with $\Pr_{T^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq \gamma_i/\ln(1/\gamma_i)$ with $\gamma_i = \text{er}_{T^i}(h_{T^i}^*)$.

Now define $\bar{\mathcal{H}}^i \subseteq \mathcal{H}$ as the set of all hypotheses $h \in \mathcal{H}$ with $\text{er}_{\mathcal{D}^i}(h) \leq \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) + (1/8)\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$. By Observation 1, all pairs $h_1, h_2 \in \bar{\mathcal{H}}^i$ are in \mathcal{H}^i and thus have $\Pr_{T^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] < \gamma_i/\ln(1/\gamma_i)$. From Observation 1, this further implies that $\Pr_{\mathcal{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \leq 4 \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)/\ln(1/\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$.

Consider now a set $\mathcal{C}_= \sim \mathcal{D}_=^m$ for an $m \geq n/2$ and recall $\mathcal{D}_= \simeq \mathcal{D}_i$ when we terminate in iteration i of the for-loop. By the ERM Theorem (Theorem 1) and for the constant c_{α} in the definition of α large enough, we have that with probability at least $1 - \delta/24$, all hypotheses $h \in \mathcal{H} \setminus \bar{\mathcal{H}}^i$ have $\text{er}_{\mathcal{C}_=}(h) \geq \text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) + (1/16)\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$.

Finally, for the hypotheses in $\bar{\mathcal{H}}^i$, we have by definition that any $h \in \bar{\mathcal{H}}^i$ has $\Pr_{\mathcal{D}_=} [h(\mathbf{x}) \neq h_{\mathcal{D}_=}^*(\mathbf{x})] \leq 4 \text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) / \ln(1/\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*))$. We now invoke the following improved version of the ERM Theorem for hypothesis sets with such properties:

Lemma 4. *Let \mathcal{D} be a distribution over \mathcal{X} and $\mathcal{H} \subset \mathcal{X} \rightarrow \{-1, 1\}$ a hypothesis set of VC-dimension d . Assume there is a hypothesis $h_0 \in \mathcal{H}$ such that for all $h \in \mathcal{H}$, we have $\Pr_{\mathcal{D}}[h(\mathbf{x}) \neq h_0(\mathbf{x})] \leq p$. Then for any $0 < \delta < 1$, it holds with probability $1 - \delta$ over a set \mathbf{S} of n i.i.d. samples from \mathcal{D} that*

$$\sup_{h \in \mathcal{H}} |\text{er}_{\mathbf{S}}(h) - \text{er}_{\mathcal{D}}(h)| = |\text{er}_{\mathbf{S}}(h_0) - \text{er}_{\mathcal{D}}(h_0)| + O\left(\sqrt{\frac{p(\ln(1/p)d + \ln(1/\delta))}{n}} + \frac{d \ln(n/d) + \ln(1/\delta)}{n}\right).$$

Applying Lemma 4 on $\bar{\mathcal{H}}^i$ with $p = 4 \text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) / \ln(1/\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*))$ and $h_0 = h_{\mathcal{D}_=}^*$ gives with probability at least $1 - \delta/24$ over $\mathbf{C}_=$ that

$$\sup_{h \in \bar{\mathcal{H}}^i} |\text{er}_{\mathbf{C}_=}(h) - \text{er}_{\mathcal{D}_=}(h)| = |\text{er}_{\mathbf{C}_=}(h_{\mathcal{D}_=}^*) - \text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*)| + O\left(\sqrt{\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*)(d + \ln(1/\delta))/n} + (d \ln(n/d) + \ln(1/\delta))/n\right).$$

At the same time, for the fixed hypothesis $h_{\mathcal{D}_=}^*$, we have with probability at least $1 - \delta/24$ (by Chernoff) that

$$|\text{er}_{\mathbf{C}_=}(h_{\mathcal{D}_=}^*) - \text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*)| = O\left(\sqrt{\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) \ln(1/\delta)/n} + \ln(1/\delta)/n\right).$$

It follows that ERM on $\mathbf{C}_=$ will return a hypothesis $h_{\mathbf{C}_=}^*$ from $\bar{\mathcal{H}}^i$ and that hypothesis has

$$\text{er}_{\mathbf{C}_=}(h_{\mathbf{C}_=}^*) = \text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) + O\left(\sqrt{\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*)(d + \ln(1/\delta))/n} + (d \ln(n/d) + \ln(1/\delta))/n\right). \quad (1)$$

Finally, from Lemma 3, we have $\Pr_{\mathcal{D}}[\exists j \leq r : h_1^j(\mathbf{x}) \neq h_2^j(\mathbf{x})] \leq 8\tau$. Since $\tau \leq 1/c_\tau$ for large enough c_τ , this implies $\Pr_{\mathcal{D}}[\forall j \leq r : h_1^j(\mathbf{x}) = h_2^j(\mathbf{x})] \geq 1/2$ and thus it must be the case that $\text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) \leq 2\tau$. Inserting this in (1) and using the assumption $\tau \geq c_\tau \ln^9(n/d)(d + \ln(1/\delta))/n$ yields

$$\text{er}_{\mathbf{C}_=}(h_{\mathbf{C}_=}^*) = \text{er}_{\mathcal{D}_=}(h_{\mathcal{D}_=}^*) + O\left(\sqrt{\tau(d + \ln(1/\delta))/n}\right).$$

This completes the proof of Lemma 1 subject to proving Lemma 2, Observation 1, Lemma 3 and Lemma 4. As mentioned earlier, we prove Lemma 2 and Observation 1 in Appendix A.1 and Appendix A.2. We prove Lemma 3 in Section 2.3 and Lemma 4 in Section 2.4. \square

2.3 Progress in For-Loop (proof of Lemma 3)

In this section, we prove Lemma 3 stating that each iteration of the for-loop reduces $\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*)$ while only increasing $\Pr_{\mathcal{D}}[\exists j \leq i : \mathbf{h}_1^j(\mathbf{x}) \neq \mathbf{h}_2^j(\mathbf{x})]$ slightly. We have restated Lemma 3 here for convenience:

Restatement of Lemma 3. *For any integer $1 \leq i \leq t + 1$, assume none of the events $E_{j,0}$ and $E_{j,1}$ occurred for $j < i$ and that Algorithm 2 did not terminate with $r < i - 1$. Then $\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*) \leq \text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*)(1 - 1/(32 \ln(1/\text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*))))^{i-1}$ and $\Pr_{\mathcal{D}}[\exists j \leq i : \mathbf{h}_1^j(\mathbf{x}) \neq \mathbf{h}_2^j(\mathbf{x})] \leq 8(\text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*) - \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*))$.*

The main idea behind the proof is to consider the pair of hypotheses \mathbf{h}_1^{i-1} and \mathbf{h}_2^{i-1} . These two hypotheses have a near-optimal error under \mathbf{D}^{i-1} and yet disagree on the classification of many points. Since one of them is incorrect when they disagree, this intuitively implies that they have to err significantly less when they agree. Observing that \mathbf{D}^i is the distribution \mathcal{D} conditioned on $\mathbf{h}_1^j(\mathbf{x}) = \mathbf{h}_2^j(\mathbf{x})$ for all $j < i$, this implies that $\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*)$ is smaller than $\text{er}_{\mathbf{D}^{i-1}}(h_{\mathbf{D}^{i-1}}^*)$ and thus we have made progress. We formalize this intuition in the following proof.

Proof of Lemma 3. We prove the lemma by induction in i . In the base case $i = 1$, we have $\mathcal{D}^1 = \mathcal{D}$ and $\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) = \text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*) = \tau$ and the claim clearly holds.

For the inductive step, consider any fixed outcome B^1, \dots, B^{i-1} of $\mathbf{B}^1, \dots, \mathbf{B}^{i-1}$ for which the algorithm did not terminate with $r < i - 1$ and where none of the events $E_{j,0}, E_{j,1}$ occurred for $j < i$. This also fixes an outcome $h_1^j, h_2^j, T^j, \mathcal{H}^j$ of $\mathbf{h}_1^j, \mathbf{h}_2^j, \mathbf{T}^j, \mathbf{H}^j$ for $j < i$ and an outcome \mathcal{D}^j of \mathbf{D}^j for $j \leq i$.

Since the algorithm did not terminate with $r < i - 1$, in iteration $i - 1$, there was a pair $h_1, h_2 \in \mathcal{H}^{i-1}$ with $\Pr_{T^{i-1}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq \gamma_{i-1}/\ln(1/\gamma_{i-1})$ and we have $h_1^{i-1} = h_1$ and $h_2^{i-1} = h_2$ for some such pair. Furthermore, the events $E_{j,0}, E_{j,1}$ did not occur for any $j < i$. Thus by Observation 1, both h_1 and h_2 satisfy $\text{er}_{\mathcal{D}^{i-1}}(h_j) \leq \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) + (1/8) \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*)/\ln(1/\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*)) \leq 2 \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*)$. Also, from Observation 1, we have

$$\Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq (1/2) \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*)/\ln(1/\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*)).$$

We now have

$$\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) \leq \text{er}_{\mathcal{D}^i}(h_1) = (1/2)(\text{er}_{\mathcal{D}^i}(h_1) + \text{er}_{\mathcal{D}^i}(h_2)) \quad (2)$$

as h_1 and h_2 agree under \mathcal{D}^i . We see that

$$2 \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) + (1/4) \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*)/\ln(1/\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*)) \geq \text{er}_{\mathcal{D}^{i-1}}(h_1) + \text{er}_{\mathcal{D}^{i-1}}(h_2).$$

Using that precisely one of h_1 and h_2 errs when they disagree, and that the distribution \mathcal{D}^{i-1} conditioned on $h_1(\mathbf{x}) = h_2(\mathbf{x})$ is the distribution \mathcal{D}^i we get

$$\begin{aligned} & \text{er}_{\mathcal{D}^{i-1}}(h_1) + \text{er}_{\mathcal{D}^{i-1}}(h_2) \\ &= \Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) = h_2(\mathbf{x})](\text{er}_{\mathcal{D}^i}(h_1) + \text{er}_{\mathcal{D}^i}(h_2)) + \Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]. \end{aligned}$$

Hence

$$\begin{aligned} & \text{er}_{\mathcal{D}^i}(h_1) + \text{er}_{\mathcal{D}^i}(h_2) \\ & \leq \Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) = h_2(\mathbf{x})]^{-1} \left(2 \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) + \frac{1}{4} \cdot \frac{\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*)}{\ln(1/\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*))} - \Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \right) \\ & \leq \Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) = h_2(\mathbf{x})]^{-1} \left(2 \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) - \Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]/2 \right). \end{aligned}$$

At the same time, we have, for $\Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \leq 1/2$, that $\Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) = h_2(\mathbf{x})]^{-1} \leq 1 + 2 \Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]$.

To see that $\Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \leq 1/2$, we first get from the induction hypothesis that $\Pr_{\mathcal{D}}[\exists j < i - 1 : h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \leq 8\tau$. Hence for $\tau \leq 1/c_\tau$ for large enough c_τ , we have $\Pr_{\mathcal{D}}[\forall j < i - 1 : h_1(\mathbf{x}) = h_2(\mathbf{x})] \geq 1/2$. This further implies $\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) \leq 2\tau$. But then $\Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \leq \text{er}_{\mathcal{D}^{i-1}}(h_1) + \text{er}_{\mathcal{D}^{i-1}}(h_2) \leq 4 \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) \leq 8\tau \leq 1/2$ as claimed.

Again, for $\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) \leq 2\tau < 2/c_\tau$ for large enough $c_\tau > 0$, we finally conclude

$$\begin{aligned} & \text{er}_{\mathcal{D}^i}(h_1) + \text{er}_{\mathcal{D}^i}(h_2) \\ & \leq (1 + 2 \Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]) \left(2 \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) - \Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]/2 \right) \\ & \leq 2 \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) - \Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]/4. \end{aligned}$$

It follows from the above and (2) that

$$\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) \leq (1/2)(\text{er}_{\mathcal{D}^i}(h_1) + \text{er}_{\mathcal{D}^i}(h_2)) \leq \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) - \Pr_{\mathcal{D}^{i-1}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]/8. \quad (3)$$

This is at most

$$\begin{aligned}
& \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) - \frac{\Pr[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]}{8} \\
& \leq \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) - (1/16) \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) / \ln(1/\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*)) \\
& \leq (1 - 1/(16 \ln(1/\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*)))) \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*).
\end{aligned}$$

We now have two cases. If we already have $\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) \leq \tau^2$, then for $\tau \leq 1/c_\tau$ (and using $t = O(\ln(1/\tau) \ln \ln(1/\tau))$), we conclude $\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) \leq \tau(1 - 1/(32 \ln(1/\tau)))^{i-1}$ as claimed. If instead $\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) > \tau^2$, we have

$$(1 - 1/(16 \ln(1/\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*)))) \leq (1 - 1/(32 \ln(1/\tau)))$$

It finally follows from the induction hypothesis that

$$\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) \leq \tau(1 - 1/(32 \ln(1/\tau)))^{i-1}$$

From (3), it also follows that:

$$\begin{aligned}
& \Pr_{\mathcal{D}}[\exists j \leq i : h_1^j(\mathbf{x}) \neq h_2^j(\mathbf{x})] - \Pr_{\mathcal{D}}[\exists j \leq i-1 : h_1^j(\mathbf{x}) \neq h_2^j(\mathbf{x})] \\
& = \Pr_{\mathcal{D}}[h_1^i(\mathbf{x}) \neq h_2^i(\mathbf{x}) \mid \forall j < i : h_1^j(\mathbf{x}) = h_2^j(\mathbf{x})] \Pr_{\mathcal{D}}[\forall j < i : h_1^j(\mathbf{x}) = h_2^j(\mathbf{x})] \\
& \leq \Pr_{\mathcal{D}}[h_1^i(\mathbf{x}) \neq h_2^i(\mathbf{x}) \mid \forall j < i : h_1^j(\mathbf{x}) = h_2^j(\mathbf{x})] \\
& = \Pr_{\mathcal{D}^{i-1}}[h_1^i(\mathbf{x}) \neq h_2^i(\mathbf{x})] \\
& \leq 8(\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) - \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)).
\end{aligned}$$

From the induction hypothesis, we conclude

$$\begin{aligned}
& \Pr_{\mathcal{D}}[\exists j \leq i : h_1^j(\mathbf{x}) \neq h_2^j(\mathbf{x})] \\
& = \Pr_{\mathcal{D}}[\exists j \leq i-1 : h_1^j(\mathbf{x}) \neq h_2^j(\mathbf{x})] + (\Pr_{\mathcal{D}}[\exists j \leq i : h_1^j(\mathbf{x}) \neq h_2^j(\mathbf{x})] - \Pr_{\mathcal{D}}[\exists j \leq i-1 : h_1^j(\mathbf{x}) \neq h_2^j(\mathbf{x})]) \\
& \leq 8(\text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*) - \text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*)) + 8(\text{er}_{\mathcal{D}^{i-1}}(h_{\mathcal{D}^{i-1}}^*) - \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)) \\
& = 8(\text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*) - \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)).
\end{aligned}$$

The claim follows. □

2.4 Tighter ERM for near-identical hypotheses (proof of Lemma 4)

In this section, we prove that Empirical Risk Minimization performs better than the general ERM Theorem (Theorem 1) when the input distribution satisfies that all hypotheses in a set \mathcal{H} rarely disagree. We have restated Lemma 4 here for convenience:

Restatement of Lemma 4. *Let \mathcal{D} be a distribution over \mathcal{X} and $\mathcal{H} \subset \mathcal{X} \rightarrow \{-1, 1\}$ a hypothesis set of VC-dimension d . Assume there is a hypothesis $h_0 \in \mathcal{H}$ such that for all $h \in \mathcal{H}$, we have $\Pr_{\mathcal{D}}[h(\mathbf{x}) \neq h_0(\mathbf{x})] \leq p$. Then for any $0 < \delta < 1$, it holds with probability $1 - \delta$ over a set \mathbf{S} of n i.i.d. samples from \mathcal{D} that*

$$\begin{aligned}
& \sup_{h \in \mathcal{H}} |\text{er}_{\mathbf{S}}(h) - \text{er}_{\mathcal{D}}(h)| \\
& = |\text{er}_{\mathbf{S}}(h_0) - \text{er}_{\mathcal{D}}(h_0)| + O\left(\sqrt{p(\ln(1/p)d + \ln(1/\delta))/n} + (d \ln(n/d) + \ln(1/\delta))/n\right).
\end{aligned}$$

Proof of Lemma 4. We assume that \mathcal{D} is a distribution over $\mathcal{X} \times \{-1, 1\}$ for which the label y is uniquely determined from x , i.e. one of $\Pr_{\mathcal{D}}[y = 1 \mid \mathbf{x} = x]$ and $\Pr_{\mathcal{D}}[y = -1 \mid \mathbf{x} = x]$ is 0 for all $x \in \mathcal{X}$. This can be assumed wlog. by replacing each $x \in \mathcal{X}$ with two point x_{-1} and x_1 and letting the probability

density function p of \mathcal{D} satisfy $p((x_{-1}, -1)) = p(x) \Pr_{\mathcal{D}}[\mathbf{y} = -1 \mid \mathbf{x} = x]$, $p((x_{-1}, 1)) = 0$ and $p((x_1, 1)) = p(x) \Pr_{\mathcal{D}}[\mathbf{y} = 1 \mid \mathbf{x} = x]$, $p((x_1, -1)) = 0$. Finally, for each $h \in \mathcal{H}$ we let $h(x_{-1}) = h(x_1) = h(x)$. This leaves the VC-dimension of \mathcal{H} and $\text{er}_{\mathcal{D}}(h)$ for any h unchanged. Let $c : \mathcal{X} \rightarrow \{-1, 1\}$ denote the concept giving the label of each $x \in \mathcal{X}$.

Now consider the hypothesis set $\mathcal{H}_{=} \subseteq \mathcal{X} \rightarrow \{-1, 1\}$ containing for each $h \in \mathcal{H}$ the hypothesis $h_{=}$ taking the value 1 on $x \in \mathcal{X}$ with $h(x) \neq h_0(x) \wedge h(x) = c(x)$ and the value -1 otherwise. Also consider the set \mathcal{H}_{\neq} containing for each $h \in \mathcal{H}$ the hypothesis h_{\neq} taking the value 1 on x with $h(x) \neq h_0(x) \wedge h(x) \neq c(x)$, and -1 otherwise. The VC-dimension of $\mathcal{H}_{=}$ and \mathcal{H}_{\neq} are both at most $O(d)$. Furthermore, every h in $\mathcal{H}_{=}$ and \mathcal{H}_{\neq} satisfy $\Pr_{\mathcal{D}}[h(\mathbf{x}) = 1] \leq p$. Consider now the distribution \mathcal{D}' obtained by sampling $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$ and replacing \mathbf{y} by -1 . Then $\text{er}_{\mathcal{D}'}(h) = \Pr_{\mathcal{D}}[h(\mathbf{x}) = 1]$ for any $h \in \mathcal{H}_{=}$ and $h \in \mathcal{H}_{\neq}$. The ERM Theorem (Theorem 1) on $\mathcal{H}_{=}$ and \mathcal{H}_{\neq} with distribution \mathcal{D}' implies that with probability $1 - \delta$ over $\mathbf{S} \sim \mathcal{D}^n$, it holds that

$$\sup_{h \in \mathcal{H}_{=} \cup \mathcal{H}_{\neq}} \left| \Pr_{\mathbf{S}}[h(\mathbf{x}) = 1] - \Pr_{\mathcal{D}}[h(\mathbf{x}) = 1] \right| \leq c_0 \cdot \left(\sqrt{p(d \ln(1/p) + \ln(1/\delta))/n} + (d \ln(n/d) + \ln(1/\delta))/n \right)$$

for a constant $c_0 > 0$. Letting $\mathbf{S}(\cdot \mid h_0 \neq h)$ denote the uniform distribution over samples $(x, y) \in \mathbf{S}$ with $h_0(x) \neq h(x)$, and $\mathcal{D}(\cdot \mid h_0 \neq h)$, $\mathbf{S}(\cdot \mid h_0 = h)$, $\mathcal{D}(\cdot \mid h_0 = h)$ defined symmetrically, we thus have for any $h \in \mathcal{H}$,

$$\begin{aligned} & |\text{er}_{\mathbf{S}}(h) - \text{er}_{\mathcal{D}}(h)| \\ &= \left| \Pr_{\mathbf{S}}[h_0 \neq h] \text{er}_{\mathbf{S}(\cdot \mid h_0 \neq h)}(h) + \Pr_{\mathbf{S}}[h_0 = h] \text{er}_{\mathbf{S}(\cdot \mid h_0 = h)}(h) - \Pr_{\mathcal{D}}[h_0 \neq h] \text{er}_{\mathcal{D}(\cdot \mid h_0 \neq h)}(h) - \Pr_{\mathcal{D}}[h_0 = h] \text{er}_{\mathcal{D}(\cdot \mid h_0 = h)}(h) \right| \\ &= \left| \Pr_{\mathbf{S}}[h_{\neq} = 1] + \Pr_{\mathbf{S}}[h_0 = h] \text{er}_{\mathbf{S}(\cdot \mid h_0 = h)}(h) - \Pr_{\mathcal{D}}[h_{\neq} = 1] - \Pr_{\mathcal{D}}[h_0 = h] \text{er}_{\mathcal{D}(\cdot \mid h_0 = h)}(h) \right| \\ &= \left| \Pr_{\mathbf{S}}[h_{\neq} = 1] + (\text{er}_{\mathbf{S}}(h_0) - \Pr_{\mathbf{S}}[h_0 \neq h] \text{er}_{\mathbf{S}(\cdot \mid h_0 \neq h)}(h_0)) - \Pr_{\mathcal{D}}[h_{\neq} = 1] - (\text{er}_{\mathcal{D}}(h_0) - \Pr_{\mathcal{D}}[h_0 \neq h] \text{er}_{\mathcal{D}(\cdot \mid h_0 \neq h)}(h_0)) \right| \\ &= \left| \Pr_{\mathbf{S}}[h_{\neq} = 1] + (\text{er}_{\mathbf{S}}(h_0) - \Pr_{\mathbf{S}}[h_{=} = 1]) - \Pr_{\mathcal{D}}[h_{\neq} = 1] - (\text{er}_{\mathcal{D}}(h_0) - \Pr_{\mathcal{D}}[h_{=} = 1]) \right| \\ &\leq |\text{er}_{\mathbf{S}}(h_0) - \text{er}_{\mathcal{D}}(h_0)| + \left| \Pr_{\mathbf{S}}[h_{\neq} = 1] - \Pr_{\mathcal{D}}[h_{\neq} = 1] \right| + \left| \Pr_{\mathbf{S}}[h_{=} = 1] - \Pr_{\mathcal{D}}[h_{=} = 1] \right| \\ &\leq |\text{er}_{\mathbf{S}}(h_0) - \text{er}_{\mathcal{D}}(h_0)| + 2c_0 \cdot \left(\sqrt{p(\ln(1/p)d + \ln(1/\delta))/n} + (d \ln(n/d) + \ln(1/\delta))/n \right). \end{aligned}$$

The claim follows. \square

3 Lower Bound for Proper Agnostic Learning

In this section, we prove a lower bound for proper agnostic PAC learning, stated formally in Theorem 2. So let $C > 0$ be a sufficiently large constant, let d be a target VC-dimension, n a number of samples and let τ satisfy $Cd \ln(n/d)/n \leq \tau \leq 1/C$. Note that the assumption on τ also implies $n \geq C^2 d \ln(n/d)$.

We define the input domain \mathcal{X} to be the discrete domain $\mathcal{X} = \{x_1, \dots, x_u\}$ for a $u \geq d$ to be determined. The hypothesis set \mathcal{H} contains all hypotheses h that predict -1 on precisely d of the points in \mathcal{X} . The target concept $c : \mathcal{X} \rightarrow \{-1, 1\}$ to learn has $c(x) = 1$ for all $x \in \mathcal{X}$.

Consider now an arbitrary *proper* learning algorithm \mathcal{A} for \mathcal{H} and c . Recall that a proper learning algorithm always returns a hypothesis $h \in \mathcal{H}$. Our goal is to show that there is a distribution \mathcal{D} over \mathcal{X} such that the hypothesis $h_{\mathbf{S}}$ returned by \mathcal{A} on a sample $\mathbf{S} \sim \mathcal{D}^n$ often satisfies

$$\text{er}_{\mathcal{D}}(h_{\mathbf{S}}) = \tau + \Omega(\sqrt{\tau \ln(1/\tau)d/n})$$

while there is a hypothesis $h \in \mathcal{H}$ with $\text{er}_{\mathcal{D}}(h) = \tau$.

For proving this, we consider multiple distributions over \mathcal{X} , one for each $h \in \mathcal{H}$. For a hypothesis $h \in \mathcal{H}$, the distribution \mathcal{D}_h returns a uniform point among those x_i with $h(x_i) = -1$ with probability $(1 - \alpha)d/u$

and it returns a uniform point among the remaining with probability $1 - (1 - \alpha)d/u$. Here $0 < \alpha < 1$ is a parameter to be determined. Observe that since $\alpha > 0$, we have that h achieves the smallest error under \mathcal{D}_h among all $h \in \mathcal{H}$. Furthermore

$$\text{er}_{\mathcal{D}_h}(h) = (1 - \alpha)d/u. \quad (4)$$

We will later fix α and u such that $\tau = (1 - \alpha)d/u$, i.e. $\text{er}_{\mathcal{D}_h}(h) = \tau$.

To prove a lower bound on the error of \mathcal{A} , we now draw a random $\mathbf{h} \in \mathcal{H}$ (unknown to \mathcal{A}) and measure the performance of \mathcal{A} under the distribution $\mathcal{D}_{\mathbf{h}}$. For this, we consider the following *failure* event of \mathcal{A} . We say that \mathcal{A} fails on a sample S from \mathcal{D}_h if it returns a hypothesis $h_S \in \mathcal{H}$ such that $|\{i \in [u] : h_S(x_i) = h(x_i) = -1\}| \leq d/2$. We first observe

Observation 2. *If \mathcal{A} fails on a sample S from \mathcal{D}_h^n for an $h \in \mathcal{H}$, then*

$$\text{er}_{\mathcal{D}_h}(h_S) \geq \tau + \alpha d/(2u).$$

Proof. Since the target concept c is the all-1 concept, we have that $\text{er}_{\mathcal{D}_h}(h_S) = \Pr_{x \sim \mathcal{D}}[h_S(x) = -1]$. Every i for which $h(x_i) = -1$ has $\Pr_{x \sim \mathcal{D}}[x = x_i] = (1 - \alpha)/u$ and every i with $h(x_i) = 1$ has $\Pr_{x \sim \mathcal{D}}[x = x_i] = (1 - (1 - \alpha)d/u)/(u - d/u) > 1/u$. It follows that $\text{er}_{\mathcal{D}_h}(h_S) \geq (d/2)/u + (d/2)(1 - \alpha)/u = d/u - (d/2)\alpha/u = \tau + \alpha d/(2u)$. \square

The second part of the proof shows that \mathcal{A} fails with constant probability over a random choice of \mathbf{h} and sample $\mathbf{S} \sim \mathcal{D}_{\mathbf{h}}^n$.

Lemma 5. *There is a universal constant $C' > 2$, such that for any proper learning algorithm \mathcal{A} , if $\alpha \leq \min\{\sqrt{u \ln(u/d)/(nC')}, 1/C'\}$ and u satisfies $dC' \leq u \leq n/C'$, then it holds with probability at least $1/16$ over a random \mathbf{h} from \mathcal{H} and a random sample $\mathbf{S} \sim \mathcal{D}_{\mathbf{h}}^n$ that \mathcal{A} fails on \mathbf{S} .*

Before giving the proof of Lemma 5, let us derive our lower bound. Recall that τ satisfies $Cd \ln(n/d)/n \leq \tau \leq 1/C$ for a sufficiently large constant $C > 0$. Using (4) and $\tau = \text{er}_{\mathcal{D}_h}(\mathbf{h})$, we let

$$u = (1 - \alpha)d/\tau.$$

Also, fix $\alpha = \min\{\sqrt{u \ln(u/d)/(nC')}, 1/C'\}$ where C' is the constant from Lemma 5. For C large enough, we have $u = (1 - \alpha)d/\tau \geq C(1 - \alpha)d \geq C(1 - 1/C')d \geq dC'$. Similarly, we have $u = (1 - \alpha)d/\tau \leq (1 - \alpha)n/C \leq n/C \leq n/C'$. Thus u and α satisfy the constraints in Lemma 5. It follows that for any proper learning algorithm \mathcal{A} and a random \mathbf{h} from \mathcal{H} , \mathcal{A} fails with probability at least $1/16$ on a sample $\mathbf{S} \sim \mathcal{D}_{\mathbf{h}}^n$. From Observation 2, we get that in this case, $\text{er}_{\mathcal{D}_h}(h_{\mathbf{S}}) \geq \tau + \alpha d/(2u)$. By our choice of parameters, we have

$$\alpha d/(2u) = \min\{\sqrt{\ln(u/d)d^2/(4C'nu)}, d/(2C'u)\}$$

Since $u = (1 - \alpha)d/\tau \in [(1 - 1/C')d/\tau, d/\tau] \subseteq [d/(2\tau), d/\tau]$, this is at least

$$\min\{\sqrt{\tau \ln(1/(2\tau))d/(4C'n)}, \tau/(2C')\}.$$

For $\tau \geq Cd \ln(n/d)/n$ and C large enough, the former term is the minimum and we conclude:

$$\text{er}_{\mathcal{D}_h}(h_{\mathbf{S}}) = \tau + \Omega\left(\sqrt{\frac{\tau d \ln(1/\tau)}{n}}\right).$$

This concludes the proof of Theorem 2. What remains is thus to establish Lemma 5:

Proof of Lemma 5. It is not hard to see (naive Bayes) that the optimal strategy for any \mathcal{A} minimizing the probability of failure (over \mathbf{h} and \mathbf{S}) when given a sample \mathbf{S} , is to output the hypothesis $h_{\mathbf{S}}$ returning -1 on the d points among x_1, \dots, x_u from which fewest samples were seen. We thus upper bound the probability that this set contains more than $d/2$ samples x_i with $\mathbf{h}(x_i) = -1$.

For this, fix an arbitrary outcome h of \mathbf{h} , let $U_h \subset \{x_1, \dots, x_n\}$ denote the set of x_i with $h(x_i) = -1$. We have $|U_h| = d$. For any $x_i \in U_h$, let \mathbf{n}_i give the number of copies of x_i in the sample $\mathbf{S} \sim \mathcal{D}_h^n$. We have $\mathbb{E}[\mathbf{n}_i] = (1 - \alpha)n/u$. Since \mathbf{n}_i is a sum of independent indicator random variables, its variance is at most its expectation. It follows from Chebyshev's inequality that $\Pr[|\mathbf{n}_i - \mathbb{E}[\mathbf{n}_i]| > 8\sqrt{\mathbb{E}[\mathbf{n}_i]}] \leq 1/64$. Markov's inequality implies that with probability at least $15/16$, there are no more than $d/4$ indices $i \in U_h$ for which $\mathbf{n}_i < (1 - \alpha)n/u - 8\sqrt{(1 - \alpha)n/u}$.

We next show that there is a good chance that at least $d/2$ of the points x_i with $x_i \notin U_h$ have $\mathbf{n}_i < (1 - \alpha)n/u - 8\sqrt{(1 - \alpha)n/u}$. For this, we require $u \geq 2d$. Then for any $x_i \notin U_h$, we have $\Pr_{x \sim \mathcal{D}_h}[x = x_i] = q/u$ for $q \in [1, 1 + \alpha]$. We now invoke the following anti-concentration result:

Lemma 6. *Consider throwing n balls independently into u bins, such that the i 'th bin is hit with probability p_i . Let S be a subset of m bins such that each bin in S has $p_i = p$ for some $12/n \leq p \leq 1/2$. Then for every integer $k \leq m/C$ for a sufficiently large constant $C > 0$, it holds with probability at least $1/8$ that there are at least k bins in S containing less than $\max\{pn - \sqrt{pn \ln(m/k)}/6, pn/2\}$ balls each.*

We prove the lemma in Appendix A.3.

By Lemma 6 (setting $p = q/u$, $m = u - d \geq u/2$, $k = d/2$), with probability at least $1/8$ provided $Cd \leq u \leq n/C$ for a large enough constant $C > 0$, there are at least $d/2$ points x_i with $h(x_i) = 1$ such that we see no more than

$$qn/u - \min\{\sqrt{(qn/u) \ln(u/d)}/6, qn/(2u)\} \leq (1 + \alpha)n/u - \min\{\sqrt{(n/u) \ln(u/d)}/6, n/(2u)\}$$

copies of x_i .

We aim to choose α such that $\min\{\sqrt{(n/u) \ln(u/d)}/6, n/(2u)\} \geq 2\alpha n/u + 8\sqrt{n/u}$, since then the number of copies we see of these x_i is no more than $(1 - \alpha)n/u - 8\sqrt{n/u}$.

For this, we first constrain u to satisfy $Cd \leq u \leq n/C$ for a big enough constant $C > 0$ so that

$$8\sqrt{n/u} \leq \max\{\sqrt{(n/u) \ln(u/d)}/12, n/(4u)\}.$$

The constraint on α is now satisfied when

$$2\alpha n/u \leq \min\{\sqrt{(n/u) \ln(u/d)}/12, n/(4u)\}$$

which is

$$\alpha \leq \min\{\sqrt{u \ln(u/d)/(24^2 n)}, 1/8\}.$$

We conclude that for such α and u , with probability at least $1/8 - 1/16 = 1/16$ over \mathbf{S} and \mathbf{h} , we have that \mathcal{A} fails on \mathbf{S} . \square

4 Conclusion and Open Problems

In this work, we established that ERM, and all other proper learning algorithms, are sub-optimal for agnostic PAC learning when treating $\tau = \text{er}_{\mathcal{D}}(h_{\mathcal{D}}^*)$ as a parameter. We then complemented the lower bound with a new improper learning algorithm that achieves an optimal sample complexity except for very small values of τ . However, a number of intriguing questions remain. First, can we develop an algorithm that is optimal for the full range of τ ? In particular, our new algorithm implies that it suffices to consider the near-realizable case of $\tau = O(\ln^9(n/d)d/n)$. Secondly, we know that variants of majority voting (bagging, etc.) are optimal for realizable PAC learning. The analysis tools used when proving their optimality breaks down for the agnostic setting. Can we somehow analyse them in a different way and prove that e.g. bagging is optimal both in the agnostic and realizable setting? If not, can we prove a lower bound for concrete algorithms, such as bagging or Hanneke's majority voter, proving that they are sub-optimal in the agnostic case? Thirdly, it could be the case that there is a higher lower bound for all learning algorithms when $\tau \approx d/n$. Can we prove this? Another interesting question is whether we can design an optimal agnostic learning algorithm that automatically adapts to δ ? In more detail, our new algorithm requires knowledge of the failure probability

δ and thus works only for a fixed user defined δ . ERM on the other hand automatically works for all values of δ simultaneously. Next, our algorithm is not necessarily efficient in terms of running time. In particular, even if ERM over \mathcal{H} is efficient, it is unclear how to determine if there are two hypotheses h_1, h_2 that are both near-optimal and yet disagree in the classification of many samples. Can we design an efficient learning algorithm with sample complexity similar to our new algorithm? Finally, the authors find the idea of recursively training near-optimal, but highly disagreeing classifiers, to be promising. Are there other applications of this idea in learning theory?

Acknowledgment

Kasper Green Larsen is co-funded by the European Union (ERC, TUCOA, 101125203) and Independent Research Fund Denmark (DFR) Sapere Aude Research Leader Grant No. 9064-00068B. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

- [1] I. Aden-Ali, Y. Cherapanamjeri, A. Shetty, and N. Zhivotovskiy. Optimal PAC bounds without uniform convergence. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1203–1223. IEEE Computer Society, 2023.
- [2] I. Aden-Ali, M. M. Høgsgaard, K. G. Larsen, and N. Zhivotovskiy. Majority-of-three: The simplest optimal learner? *CoRR*, abs/2403.08831, 2024.
- [3] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, USA, 1st edition, 2009.
- [4] J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009.
- [5] P. Auer and R. Ortner. A new PAC bound for intersection-closed concept classes. *Machine Learning*, 66(2):151–163, 2007.
- [6] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.
- [7] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.
- [8] O. Bousquet, S. Hanneke, S. Moran, and N. Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. In *Conference on Learning Theory*, pages 582–609. PMLR, 2020.
- [9] O. Bousquet and N. Zhivotovskiy. Fast classification rates without standard margin assumptions. *Information and Inference: A Journal of the IMA*, 10(4):1389–1421, 2021.
- [10] L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, aug 1996.
- [11] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.
- [12] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

- [13] A. Grønlund, L. Kamma, and K. G. Larsen. Margins are insufficient for explaining gradient boosting. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.
- [14] S. Hanneke. The optimal sample complexity of pac learning. *The Journal of Machine Learning Research*, 17(1):1319–1333, 2016.
- [15] S. Hanneke. Refined error bounds for several learning algorithms. *The Journal of Machine Learning Research*, 17(1):4667–4721, 2016.
- [16] D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.
- [17] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0, 1\}$ -functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.
- [18] P. N. Klein and N. E. Young. On the number of iterations for dantzig-wolfe optimization and packing-covering approximation algorithms. *SIAM J. Comput.*, 44(4):1154–1172, 2015.
- [19] K. G. Larsen. Bagging is an optimal PAC learner. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*, volume 195 of *Proceedings of Machine Learning Research*, pages 450–468. PMLR, 2023.
- [20] Y. Li, P. Long, and A. Srinivasan. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62:516 – 527, 2001.
- [21] P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34, 10 2006.
- [22] N. Puchkin and N. Zhivotovskiy. Exponential savings in agnostic active learning through abstention. In *Conference on Learning Theory*, pages 3806–3832. PMLR, 2021.
- [23] M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. *Advances in Neural Information Processing Systems*, 24, 2011.
- [24] H. U. Simon. An almost optimal PAC algorithm. In *Conference on Learning Theory*, pages 1552–1563. PMLR, 2015.
- [25] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
- [26] V. Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 1982.
- [27] V. Vapnik and A. Chervonenkis. A class of algorithms for pattern recognition learning. *Avtomatika i Telemekhanika*, 25(6):937–945, 1964.
- [28] V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.
- [29] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.
- [30] N. Zhivotovskiy and S. Hanneke. Localization of vc classes: Beyond local rademacher complexities. *Theoretical Computer Science*, 742:27–49, 2018.

A Appendix

A.1 Failures are rare (proof of Lemma 2)

In this section, we show that the failure events $E_{i,0}$ and $E_{i,1}$ are unlikely. Formally, we prove:

Restatement of Lemma 2. *For all i , we have $\Pr[E_{i,0}] \leq \delta/(4t)$ and $\Pr[E_{i,1}] \leq \delta/(4t)$.*

Proof of Lemma 2. We consider each type of event in turn. Since both of the events only occur when none of the events $E_{j,0}, E_{j,1}$ occurred for any $j < i$, we bound the probability of $E_{i,0}$ and $E_{i,1}$ under this assumption. So fix an outcome B^1, \dots, B^{i-1} of $\mathbf{B}^1, \dots, \mathbf{B}^{i-1}$ such that none of the events occurred for $j < i$. This also fixes an outcome h_1^j, h_2^j of $\mathbf{h}_1^j, \mathbf{h}_2^j$ for $j < i$ and \mathcal{D}^i of \mathbf{D}^i . Note that \mathbf{B}^i is independent of the events $E_{j,0}, E_{j,1}$ for $j < i$ and thus \mathbf{B}^i still consists of n/t i.i.d. samples from \mathcal{D} .

Observe that by Lemma 3, $\Pr_{\mathcal{D}}[\exists j < i : h_1^j(\mathbf{x}) \neq h_2^j(\mathbf{x})] \leq 8\tau$. Since $\tau \leq 1/c_\tau$ for large enough $c_\tau > 0$, this further implies that $\Pr_{\mathcal{D}}[\forall j < i : h_1^j(\mathbf{x}) = h_2^j(\mathbf{x})] \geq 1 - 8\tau \geq 1/2$. We further have $|\mathbf{B}^i| = n/t > c_n \ln(t/\delta)$ by assumptions $n \geq c_n \ln^{3.5}(n/d)(d + \ln(1/\delta))$, $t = O(\ln(1/\bar{\tau}) \ln \ln(1/\bar{\tau})) = O(\ln(n/d) \ln \ln(n/d))$, $\bar{\tau} \in [\tau/2, 2\tau]$ and $\tau > d/n$. It follows that $|\mathbf{T}^i| \geq n/(2t)$ except with probability $\delta/(8t)$. We thus bound the probabilities under the assumption that \mathbf{T}^i consists of $m \geq n/(2t)$ i.i.d. samples from \mathcal{D}^i .

Event $E_{i,0}$. By the ERM Theorem (Theorem 1) on \mathbf{T}^i , we have with probability $1 - \delta/(8t)$ that for all $h \in \mathcal{H}$

$$|\text{er}_{\mathbf{T}^i}(h) - \text{er}_{\mathcal{D}^i}(h)| = O\left(\sqrt{\frac{t \text{er}_{\mathcal{D}^i}(h)(d \ln(n/d) + \ln(t/\delta))}{n}} + \frac{t(d \ln(n/d) + \ln(t/\delta))}{n}\right). \quad (5)$$

The $\ln(t/\delta) = \ln(t) + \ln(1/\delta)$ term may be replaced by $\ln(1/\delta)$ as $\ln(t)$ is dominated by the $d \ln(n/d)$ term. Thus for the constant c_α in the definition of α large enough, this implies

$$|\text{er}_{\mathbf{T}^i}(h) - \text{er}_{\mathcal{D}^i}(h)| \leq (1/32)\alpha(n/t, d, \delta, \min\{\text{er}_{\mathcal{D}^i}(h), \text{er}_{\mathbf{T}^i}(h)\}).$$

To see that we may insert $\min\{\text{er}_{\mathcal{D}^i}(h), \text{er}_{\mathbf{T}^i}(h)\}$ instead of $\text{er}_{\mathcal{D}^i}(h)$, we consider two cases. First, if $\text{er}_{\mathcal{D}^i}(h) > \sqrt{c_\alpha t}(d \ln(n/d) + \ln(t/\delta))$ for a sufficiently large constant c_α in the definition of α , we have from (5) that $\text{er}_{\mathbf{T}^i}(h) \geq (1/2)\text{er}_{\mathcal{D}^i}(h)$ and thus $|\text{er}_{\mathbf{T}^i}(h) - \text{er}_{\mathcal{D}^i}(h)| \leq (1/32)\alpha(n/t, d, \delta, \min\{\text{er}_{\mathcal{D}^i}(h), \text{er}_{\mathbf{T}^i}(h)\})$. If on the other hand $\text{er}_{\mathcal{D}^i}(h) \leq \sqrt{c_\alpha t}(d \ln(n/d) + \ln(t/\delta))$ then we still have $(1/32)\alpha(n/t, d, \delta, 0) \geq |\text{er}_{\mathbf{T}^i}(h) - \text{er}_{\mathcal{D}^i}(h)|$ by (5) and large enough constant c_α .

Event $E_{i,1}$. Consider the set of hypotheses $\mathcal{H}' = \mathcal{H} \oplus \mathcal{H}$ consisting of all hypotheses that may be written as $g_{h_1, h_2}(x) = h_1(x) \cdot h_2(x)$ where \cdot denotes multiplication and $h_1, h_2 \in \mathcal{H}$. Then $\Pr_{\mathcal{D}}[g_{h_1, h_2}(\mathbf{x}) = 1] = \Pr_{\mathcal{D}}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]$ for any distribution \mathcal{D} . Furthermore, the VC-dimension of \mathcal{H}' is $O(d)$.

It follows by the ERM Theorem (Theorem 1) and the constant c_α in the definition of α large enough, that with probability at least $1 - \delta/(8t)$, any pair h_1, h_2 satisfy $|\Pr_{\mathcal{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] - \Pr_{\mathbf{T}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]| \leq (1/32)\alpha(n/t, d, \delta, \min\{\Pr_{\mathcal{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})], \Pr_{\mathbf{T}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]\})$. Here we may insert the min by the same arguments as above. \square

A.2 Properties when no failures (proof of Observation 1)

In this section, we show that when the events $E_{j,0}$ and $E_{j,1}$ do not occur, the hypotheses in \mathbf{H}^i and \mathcal{H} behave nicely. Concretely, we prove Observation 1, which we have restated here:

Restatement of Observation 1. *Assume none of the events $E_{j,0}$ and $E_{j,1}$ occur for $j \leq i$ and that Algorithm 2 does not terminate before iteration i . Then if $\gamma_i = \text{er}_{\mathbf{T}^i}(h_{\mathbf{T}^i}^*) \leq Z_t$ it holds that $\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*) \leq 2Z_t$. If $\gamma_i > Z_t$, then each of the following hold:*

- $\gamma_i \leq \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*) + \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*)/\ln(1/\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*)) \leq 2\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*)$.

- Every hypothesis h in \mathbf{H}^i satisfies $\text{er}_{\mathbf{D}^i}(h) \leq \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*) + (1/8) \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*) / \ln(1/\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*))$.
- Every hypothesis $h \in \mathcal{H}$ with $\text{er}_{\mathbf{D}^i}(h) \leq \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*) + (1/8)\alpha(n/t, d, \delta, \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*))$ is in \mathbf{H}^i .
- Every pair of hypotheses h_1, h_2 with $\Pr_{\mathbf{T}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq \gamma_i / \ln(1/\gamma_i)$ satisfy $\Pr_{\mathbf{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq (1/2) \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*) / \ln(1/\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*))$.
- Every pair of hypotheses h_1, h_2 with $\Pr_{\mathbf{T}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] < \gamma_i / \ln(1/\gamma_i)$ satisfy $\Pr_{\mathbf{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \leq 4 \text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*) / \ln(1/\text{er}_{\mathbf{D}^i}(h_{\mathbf{D}^i}^*))$.

Before proving Observation 1, we state and prove an auxiliary result regarding α and Z_t :

Observation 3. For $x \geq Z_t/2$, we have

$$\alpha(n/t, d, \delta, x) \leq \frac{2c_\alpha}{\sqrt{c_Z}} \cdot \frac{x}{\ln(1/x)}.$$

Proof of Observation 3. Notice that for any $x \geq Z_t/2$, we have $\ln(1/x) \leq \ln(n/d)$ and thus for $x \geq Z_t/2$:

$$\begin{aligned} \alpha(n/t, d, \delta, x) &\leq c_\alpha \cdot \left(\sqrt{\frac{tx(d \ln(1/x) + \ln(1/\delta))}{n}} + \frac{t(d \ln(n/d) + \ln(1/\delta))}{n} \right) \\ &= c_\alpha \cdot \left(\sqrt{\frac{tx \ln^2(1/x)(d \ln(1/x) + \ln(1/\delta))}{n \ln^2(1/x)}} + \frac{t \ln(1/x)(d \ln(n/d) + \ln(1/\delta))}{n \ln(1/x)} \right) \\ &\leq c_\alpha \cdot \left(\sqrt{\frac{tx \ln^2(n/d)(d \ln(n/d) + \ln(1/\delta))}{n \ln^2(1/x)}} + \frac{t \ln(n/d)(d \ln(n/d) + \ln(1/\delta))}{n \ln(1/x)} \right) \\ &\leq c_\alpha \cdot \left(\sqrt{\frac{x^2}{c_Z \ln^2(1/x)}} + \frac{x}{c_Z \ln(1/x)} \right) \\ &\leq \frac{2c_\alpha}{\sqrt{c_Z}} \cdot \frac{x}{\ln(1/x)}. \end{aligned}$$

□

Let us also restate the failure events $E_{i,0}$ and $E_{i,1}$ here for convenience:

1. Let $E_{i,0}$ be the event that Algorithm 2 reaches iteration i , none of the events $E_{j,0}, E_{j,1}$ occurred for $i < j$ and there is a hypothesis $h \in \mathcal{H}$ with $|\text{er}_{\mathbf{D}^i}(h) - \text{er}_{\mathbf{T}^i}(h)| > (1/32)\alpha(n/t, d, \delta, \min\{\text{er}_{\mathbf{D}^i}(h), \text{er}_{\mathbf{T}^i}(h)\})$.
2. Let $E_{i,1}$ be the event that Algorithm 2 reaches iteration i , none of the events $E_{j,0}, E_{j,1}$ occurred for $i < j$ and there is a pair of hypotheses h_1, h_2 with $|\Pr_{\mathbf{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] - \Pr_{\mathbf{T}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]| > (1/32)\alpha(n/t, d, \delta, \min\{\Pr_{\mathbf{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})], \Pr_{\mathbf{T}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]\})$.

We are ready to prove Observation 1.

Proof of Observation 1. Since we only claim something when the events $E_{j,0}$ and $E_{j,1}$ did not occur for $j \leq i$ and that Algorithm 2 did not terminate before iteration i , we assume this. So fix such an outcome B^1, \dots, B^i of $\mathbf{B}^1, \dots, \mathbf{B}^i$. This also fixes an outcome $h_1^j, h_2^j, T^j, \mathcal{D}^j, \mathcal{H}^j$ of $\mathbf{h}_1^j, \mathbf{h}_2^j, \mathbf{T}^j, \mathbf{D}^j, \mathbf{H}^j$ for $j \leq i$.

Assume first that $\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) > 2Z_t$. We wish to show $\gamma_i > Z_t$. To see this, note that by definition of $E_{i,0}$, we have $\text{er}_{\mathbf{T}^i}(h_{\mathbf{T}^i}^*) \geq 2Z_t - (1/32)\alpha(n/t, d, \delta, 2Z_t)$. By Observation 3, for large enough c_Z , this is at least $2Z_t - Z_t/2 > Z_t$. This proves the part of Observation 1 stating that if $\gamma_i = \text{er}_{\mathbf{T}^i}(h_{\mathbf{T}^i}^*) \leq Z_t$ then $\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) \leq 2Z_t$.

For the remainder of the proof, assume $\gamma_i > Z_t$. We start by proving bounds on $\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)$.

Since the event $E_{i,0}$ did not occur, we have

- $\text{er}_{T^i}(h_{\mathcal{D}^i}^*) \geq \gamma_i \geq Z_t$ and thus $\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) \geq Z_t - \alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$. We claim this inequality implies $\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) \geq Z_t/2$. To see this, assume for contradiction that $\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) < Z_t/2$, then since α is increasing in its last argument, we have by Observation 3 that $\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) + \alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)) \leq Z_t/2 + \alpha(n/t, d, \delta, Z_t/2)$. For c_Z large enough, the right hand side is less than Z_t , which contradicts the inequality $\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) \geq Z_t - \alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$.

In summary, we have

$$\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) \geq Z_t/2. \quad (6)$$

Next, we upper bound γ_i .

- Since $E_{i,0}$ did not occur, we have

$$\gamma_i = \text{er}_{T^i}(h_{T^i}^*) \leq \text{er}_{T^i}(h_{\mathcal{D}^i}^*) \leq \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) + (1/32)\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)).$$

By Observation 3 for large enough c_Z and using (6), this implies

$$\gamma_i \leq \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) + \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)/\ln(1/\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)) \leq 2 \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*).$$

This establishes the first bullet in Observation 1.

- To establish the second bullet of Observation 1, note that by definition of \mathcal{H}^i , every hypothesis h in \mathcal{H}^i satisfies $\text{er}_{T^i}(h) \leq \gamma_i + \alpha(n/t, d, \delta, \gamma_i)$ and since $\gamma_i \leq 2 \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)$ (by the previous bullet) and $\alpha(n/t, d, \delta, 2x) \leq 2\alpha(n/t, d, \delta, x)$ we have $\text{er}_{T^i}(h) \leq \gamma_i + 2\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$. Also using that $\gamma_i \leq \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) + (1/32)\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$ gives $\text{er}_{T^i}(h) \leq \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) + 3\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)) \leq 6 \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)$. Finally, since $E_{i,0}$ did not occur, we have $\text{er}_{\mathcal{D}^i}(h) \leq \text{er}_{T^i}(h) + (1/32)\alpha(n/t, d, \delta, \text{er}_{T^i}(h)) \leq \text{er}_{T^i}(h) + (1/32)\alpha(n/t, d, \delta, 6 \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)) \leq \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) + (1/5)\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$. By (6), it follows by Observation 3 that $\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)) \leq \frac{2c_\alpha}{\sqrt{c_Z}} \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)/\ln(1/\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$. For c_Z large enough, we thus have $\text{er}_{\mathcal{D}^i}(h) \leq \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) + (1/8) \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)/\ln(1/\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$. This establishes the second bullet of Observation 1.
- For the third bullet, let h be a hypothesis with $\text{er}_{\mathcal{D}^i}(h) \leq \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) + (1/8)\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$. Since $E_{i,0}$ did not occur, we have $\text{er}_{T^i}(h) \leq \text{er}_{\mathcal{D}^i}(h) + (1/32)\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h))$. By Observation 3 and (6), we have $(1/8)\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)) \leq \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)$ for c_Z large enough. Thus we have $\text{er}_{\mathcal{D}^i}(h) \leq 2 \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)$. This gives $\text{er}_{T^i}(h) \leq \text{er}_{\mathcal{D}^i}(h) + (1/16)\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)) \leq \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) + (3/8)\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$. Finally, since $E_{i,0}$ did not occur, we have $\gamma_i \geq \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) - (1/32)\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$, implying $\text{er}_{T^i}(h) \leq \gamma_i + (1/2)\alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$. We also have $\gamma_i \geq (1/2) \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)$ by the first bullet. Hence we conclude $\text{er}_{T^i}(h) \leq \gamma_i + \alpha(n/t, d, \delta, \gamma_i)$, which puts h in \mathcal{H}^i .
- For the fourth bullet, consider a pair of hypotheses h_1, h_2 with $\Pr_{T^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq \gamma_i/\ln(1/\gamma_i)$. Since $E_{i,1}$ did not occur, we have

$$\Pr_{\mathcal{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq \Pr_{T^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] - (1/32)\alpha(n/t, d, \delta, \Pr_{T^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]).$$

Since $\gamma_i/\ln(1/\gamma_i) \geq Z_t/\ln(1/Z_t)$ and the expression $x - (1/32)\alpha(n/t, d, \delta, x)$ is increasing in x for $x > Z_t/\ln(1/Z_t)$ and c_Z large enough, we conclude

$$\Pr_{\mathcal{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] \geq \gamma_i/\ln(1/\gamma_i) - \alpha(n/t, d, \delta, \gamma_i/\ln(1/\gamma_i)) \geq \gamma_i/\ln(1/\gamma_i) - \alpha(n/t, d, \delta, \gamma_i).$$

Since $\gamma_i \geq Z_t$, it follows from Observation 3 that this is at least $(3/4)\gamma_i/\ln(1/\gamma_i)$ for c_Z large enough. Finally, since $\gamma_i \geq \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) - \alpha(n/t, d, \delta, \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$. We have $\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) \geq Z_t/2$ so by Observation 3, this is at least $\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)(1 - \ln(1/\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)))$. We conclude $\Pr_{\mathcal{D}^i}[h_1(x) \neq h_2(x)] \geq (3/4)\gamma_i/\ln(1/\gamma_i) \geq (1/2) \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*)/\ln(1/\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$.

- For the fifth bullet, consider a pair h_1, h_2 with $\Pr_{\mathcal{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] > 4 \text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*) / \ln(1/\text{er}_{\mathcal{D}^i}(h_{\mathcal{D}^i}^*))$. From the first bullet, this implies $\Pr_{\mathcal{D}^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] > 2\gamma_i / \ln(2/\gamma_i)$. Since $E_{i,1}$ did not occur, this implies

$$\begin{aligned} \Pr_{T^i}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})] &\geq 2\gamma_i / \ln(2/\gamma_i) - (1/32)\alpha(n/t, d, \delta, 2\gamma_i / \ln(2/\gamma_i)) \\ &\geq 2\gamma_i / \ln(2/\gamma_i) - (1/16)\alpha(n/t, d, \delta, \gamma_i). \end{aligned}$$

By definition of α and Z_t (using $\gamma_i \geq Z_t$), this is more than $\gamma_i / \ln(1/\gamma_i)$ and hence not both of h_1 and h_2 are in \mathcal{H}^i . □

A.3 Points with few copies (proof of Lemma 6)

In this section, we give the proof of the following lemma

Restatement of Lemma 6. *Consider throwing n balls independently into u bins, such that the i 'th bin is hit with probability p_i . Let S be a subset of m bins such that each bin in S has $p_i = p$ for some $12/n \leq p \leq 1/2$. Then for every integer $k \leq m/C$ for a sufficiently large constant $C > 0$, it holds with probability at least $1/8$ that there are at least k bins in S containing less than $\max\{np - \sqrt{np \ln(m/k)}/6, np/2\}$ balls each.*

Our proof of Lemma 6 follows previous work by Grønlund, Kamma and Larsen [13]. Let S be a subset of m bins out of u bins such that each bin in S is hit with probability $12/n \leq p \leq 1/2$. Now fix a bin $b \in S$ and let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be indicator random variables, where \mathbf{X}_j is 1 if the j 'th ball is in b and 0 otherwise. We now invoke the following lemma

Lemma 7 (Klein and Young [18]). *Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be independent indicator random variables with success probability $p \leq 1/2$. For every $\sqrt{3/(np)} < \delta < 1/2$,*

$$\Pr \left[\sum_i \mathbf{X}_i \leq (1 - \delta)np \right] \geq \exp(-9np\delta^2).$$

Using the requirement $m \geq Ck$ for a sufficiently large constant $C > 0$, we have by Lemma 7 with $\delta = \min\{\sqrt{\ln(m/k)/(np)}/6, 1/2\}$, that $\Pr[\sum_i \mathbf{X}_i \leq np - \min\{\sqrt{np \ln(m/k)}/6, np/2\}] \geq \exp(-\ln(m/k)/4) = (k/m)^{1/4} = (m/k)^{3/4} \cdot (k/m) \geq C^{3/4}(k/m) \geq 2(k/m)$. Note that $\ln(m/k) \geq \ln(C)$ and thus $\delta > \sqrt{3/(np)}$ when C is large enough. We also remark that $1/2 \geq \sqrt{3/(np)}$ since we assume $p \geq 12/n$.

Now define indicator random variables \mathbf{Y}_i for each bin in S , taking the value 1 if the number of balls in the bin is no more than $np - \min\{\sqrt{np \ln(m/k)}/6, np/2\} = \max\{np - \sqrt{np \ln(m/k)}/6, np/2\}$. By the argument above (and symmetry of the \mathbf{Y}_i 's), we have $\mathbb{E}[\mathbf{Y}_i] = q$ for some $q \geq 2(k/m)$. At the same time, we have $\mathbb{E}[(\sum_i \mathbf{Y}_i)^2] = \sum_i \sum_j \mathbb{E}[\mathbf{Y}_i \mathbf{Y}_j]$. Since the random variables \mathbf{Y}_i are negatively correlated, we have $\mathbb{E}[\mathbf{Y}_i \mathbf{Y}_j] \leq \mathbb{E}[\mathbf{Y}_i] \mathbb{E}[\mathbf{Y}_j] = q^2$ for $i \neq j$. For $i = j$, we have $\mathbb{E}[\mathbf{Y}_i^2] = \mathbb{E}[\mathbf{Y}_i] = q$. Hence $\mathbb{E}[\sum_i \mathbf{Y}_i] = mq$ and $\mathbb{E}[(\sum_i \mathbf{Y}_i)^2] \leq mq + m(m-1)q^2 \leq mq + m^2q^2$. By Paley-Zygmund, we conclude $\Pr[\sum_i \mathbf{Y}_i > mq/2] \geq (1/4)(mq)^2 / (mq + m^2q^2)$. Since $mq \geq 2k > 1$, this is at least $1/8$. Since $mq/2 \geq k$, the conclusion follows.