Majority-of-Three: The Simplest Optimal Learner?

Ishaq Aden-Ali ADENALI@BERKELEY.EDU
Department of EECS, UC Berkeley.

Mikael Møller Høgsgaard HOGSGAARD@CS.AU.DK
Computer Science Department, Aarhus University.

Kasper Green Larsen LARSEN@CS.AU.DK
Computer Science Department, Aarhus University.

Nikita Zhivotovskiy ZHIVOTOVSKIY@BERKELEY.EDU
Department of Statistics, UC Berkeley.

Editors: Shipra Agrawal and Aaron Roth

Abstract

Developing an optimal PAC learning algorithm in the realizable setting, where empirical risk minimization (ERM) is suboptimal, was a major open problem in learning theory for decades. The problem was finally resolved by Hanneke a few years ago. Unfortunately, Hanneke's algorithm is quite complex as it returns the majority vote of many ERM classifiers that are trained on carefully selected subsets of the data. It is thus a natural goal to determine the simplest algorithm that is optimal. In this work we study the arguably simplest algorithm that could be optimal: returning the majority vote of three ERM classifiers. We show that this algorithm achieves the optimal in-expectation bound on its error which is provably unattainable by a single ERM classifier. Furthermore, we prove a near-optimal high-probability bound on this algorithm's error. We conjecture that a better analysis will prove that this algorithm is in fact optimal in the high-probability regime.
Keywords: PAC learning, Risk Bounds, Generalization Bounds, Sample Complexity

1. Introduction

In the setting of realizable Probably Approximately Correct (PAC) learning Valiant (1984); Vapnik and Chervonenkis (1964, 1974), the goal is to learn or approximate an unknown target function $f^\star \in \{0,1\}^{\mathcal{X}}$ from a labelled training sample $(S, f^\star(S)) = ((X_1, f^\star(X_1)), \dots, (X_n, f^\star(X_n)))$, where the X_i's are i.i.d. samples from an unknown distribution P over an instance space $\mathcal{X}$. In the realizable setting, we are furthermore promised that $f^\star$ belongs to a known function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ of Vapnik-Chervonenkis (VC) dimension d.

Given a labelled training sample $(S, f^\star(S))$, a learning algorithm produces a function $\widehat{f}_S \in \{0,1\}^{\mathcal{X}}$ with the goal of minimizing the probability of mispredicting the label of a new sample from P, where we denote this error by $\mathrm{err}_P\left(\widehat{f}_S\right) := \mathbf{Pr}_{X\sim P}[\widehat{f}_S(X) \neq f^\star(X)]$. The simplest reasonable learning algorithm, known as *empirical risk minimization* (ERM), simply reports an arbitrary function $\widehat{f}_S \in \mathcal{F}$ that is consistent with $f^\star$ on the training data, i.e. $\widehat{f}_S(X_i) = f^\star(X_i)$ for all $i = 1, \dots, n$. Classic work by Blumer et al. Blumer et al. (1989) (the same bound also essentially follows from the earlier works Vapnik and Chervonenkis (1968, 1971)) shows that for any $\delta > 0$, it holds with probability $1-\delta$ over S that any $\widehat{f}_S \in \mathcal{F}$ consistent with $f^\star$ on S has

$$\mathrm{err}_P\left(\widehat{f}_S\right) = O\left(\frac{d}{n}\log\left(\frac{n}{d}\right) + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right). \tag{1}$$

On the lower bound side, there exists an instance space $\mathcal{X}$ and function class $\mathcal{F}$ such that for a certain ERM algorithm, there is a target function $f^\star \in \mathcal{F}$ and hard distribution P for which Eq. (1) is tight Haussler et al. (1994); Auer and Ortner (2007); Simon (2015); Hanneke (2016b). Learning algorithms that always output a function in $\mathcal{F}$ are referred to as *proper* learning algorithms. Generally, it is known that not only ERM, but all proper learners fail to achieve the optimal error bound in the PAC learning framework. See the corresponding lower bounds in Bousquet et al. (2020).

For *improper* learning algorithms — algorithms that are allowed to output an arbitrary function $\widehat{f}_S \in \{0,1\}^{\mathcal{X}}$ — known lower bounds on the error only imply that we must have

$$\operatorname{err}_P\left(\widehat{f}_S\right) = \Omega\left(\frac{d}{n} + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right). \tag{2}$$

Developing an algorithm with a matching error upper bound, or strengthening the lower bound, was a major open problem for decades. This was finally resolved in 2015 when Hanneke Hanneke (2016a), building on the work of Simon Simon (2015), proposed the first optimal algorithm with an error upper bound matching Eq. (2), leading to the optimal error bound

$$\Theta\left(\frac{d}{n} + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right). \tag{3}$$

Hanneke's algorithm is based on constructing a large number ($\approx n^{0.79}$) of sub-samples $S_1, S_2, \cdots \subseteq S$ of the training data. This algorithm then runs ERM on each $(S_i, f^\star(S_i))$ to obtain functions $\widehat{f}_{S_1}, \widehat{f}_{S_2}, \ldots$ and finally combines them via a majority vote. The sub-samples S_i are constructed to have a carefully designed overlapping structure, and an intricate inductive argument exploiting this structure is then used to argue optimality. Recent work by Larsen Larsen (2023) shows that the carefully designed overlap structure may instead be replaced by the significantly simpler strategy of sampling each S_i as $\Theta(n)$ samples with replacement from S. This algorithm is precisely the classic heuristic known as Bagging, or bootstrap aggregation, due to Breiman Breiman (1996). Furthermore, the proof shows that a mere $O(\log(n/\delta))$ sub-samples suffice for an optimal sample complexity. The proof is however even more involved than Hanneke's and uses his analysis at its core.

Another line of work studied an alternative learning algorithm, the one-inclusion graph algorithm of Haussler, Littlestone, and Warmuth Haussler et al. (1994) that returns a function $\widehat{f}_{\mathrm{OIG}}$. This work also introduces the *prediction model* of learning, which focuses on achieving bounds on the *expected error* rather than *high probability* bounds on the error. The one-inclusion graph algorithm was initially shown to have an expected error of

$$\mathop{\mathbb{E}}_{S\sim P^n}\left[\operatorname{err}_P\left(\widehat{f}_{\mathrm{OIG}}\right)\right] \le \frac{d}{n+1}, \tag{4}$$

which was later proven to be optimal within this prediction model Li et al. (2001). Because of the tightness of the in-expectation bound Eq. (4), Warmuth conjectured Warmuth (2004) that the one-inclusion graph algorithm achieves an error upper bound matching the general lower bound Eq. (2) in the high probability regime.

Recent work by Aden-Ali, Cherapanamjeri, Shetty, and Zhivotovskiy Aden-Ali et al. (2023a) unfortunately refutes this conjecture. Concretely, they show that for any $d \in \mathbb{N}$, sample size $n \ge d$

and confidence parameter $\delta \geq cd/n$, there exists a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d and a hard distribution P such that a certain implementation of the one-inclusion graph algorithm has, with probability at least δ,

$$\mathrm{err}_P\left(\widehat{f}_{\mathrm{OIG}}\right) = \Omega\left(\frac{d}{\delta n}\right).$$

This result essentially says that, in general, the one-inclusion graph algorithm's high-probability guarantee cannot be better than applying Markov's inequality to the in-expectation guarantee in Eq. (4). In recent work also by Aden-Ali et al. Aden-Ali et al. (2023b), it was shown that if one combines the output of $\Omega(n)$ predictions made by one-inclusion algorithms on prefixes of the training data $((X_1, f^\star(X_1)), \ldots, (X_m, f^\star(X_m)))$ for $m = n/2, \ldots, n$ via a majority vote, then the resulting function is optimal in the high probability regime and, therefore, matches the error bound Eq. (3). Unfortunately, the one-inclusion graph algorithm (and this extension) is much less intuitive than the aforementioned algorithms based on taking majority votes of ERMs.

1.1. The simplest possible optimal algorithm?

In light of prior work, we have several provably optimal algorithms for PAC learning in the realizable setting. The algorithms and their analyses vary in complexity and a natural question remains: What is the simplest possible optimal algorithm? We know from lower bounds that the algorithm has to be improper and as such must be more complicated than ERM. Bagging is arguably the simplest algorithm among previous proposed algorithms, but has the most difficult analysis. The voting among one-inclusion algorithms has a somewhat simple proof, but the algorithm is not the simplest. In this work, we consider what is perhaps the simplest imaginable improper algorithm, *Majority-of-Three* (*ERMs*): Partition S into three equal-sized disjoint pieces S_1, S_2, S_3, run the same ERM algorithm on each $(S_i, f^\star(S_i))$ to obtain $\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3}$, and combine them via a majority vote to produce the function $\mathrm{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})$. Since a majority vote of two functions is undefined when the functions disagree, this is arguably the simplest possible improper algorithm. Our first main result shows that this concrete majority vote of three ERMs, which we will refer to as Majority-of-Three throughout, is optimal in expectation.

Theorem 1 *Fix a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$ and target function $f^\star \in \mathcal{F}$. For any ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$ it follows that*

$$\mathop{\mathbb{E}}_{S_1,S_2,S_3 \sim P^n}\left[\mathrm{err}_P\left(\mathrm{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})\right)\right] = O\left(\frac{d}{n}\right).$$

This result shows that Majority-of-Three matches the optimal expectation bound Eq. (4) achieved by the one-inclusion graph algorithm, up to a universal constant. Furthermore, our proof of Theorem 1 is in fact quite simple, especially compared to the previous proof that Bagging is optimal.

We note here that a single ERM alone is sub-optimal by a multiplicative $\ln(n/d)$ factor in-expectation (see the well-known lower bound in (Haussler et al., 1994, Theorem 4.2)). We emphasize that in Theorem 1, the ERMs corresponding to S_1, S_2 and S_3 can be chosen by *any* algorithm $\widehat{f}$ that outputs functions consistent with the sample. The only restriction is that it is the same algorithm $\widehat{f}$ that is run on each S_i (and that the subsets S_i are disjoint and thus i.i.d.).

We now turn our attention to the high-probability regime, where we prove the following result.

Theorem 2 *Fix a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$ and target function $f^\star \in \mathcal{F}$. Fix any ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$. For any parameter $\delta \in (0, 1/2]$ it holds with probability at least $1-\delta$ over the randomness of $S_1, S_2, S_3 \sim P^n$ that*

$$\operatorname{err}_P\left(\operatorname{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})\right) = O\left(\frac{d}{n}\log\left(\log\left(\min\left\{\frac{n}{d}, \frac{1}{\delta}\right\}\right)\right) + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right).$$

This bound is sub-optimal due to the $\log(\log(\min\{n/d, 1/\delta\}))$ term, however the additive $\log(1/\delta)$ term dominates for $\delta \le d^{-d}$. Thus, Majority-of-Three is optimal both in the constant (Theorem 1) and high-probability regimes (Theorem 2). Because of this, we conjecture that Majority-of-Three is in fact optimal for all δ and leave this as an open question for future research.

Conjecture 3 *Fix a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$ and target function $f^\star \in \mathcal{F}$. Fix any ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$. For any parameter $\delta \in (0, 1)$ it holds with probability at least $1-\delta$ over the randomness of $S_1, S_2, S_3 \sim P^n$ that*

$$\operatorname{err}_P\left(\operatorname{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})\right) = O\left(\frac{d}{n} + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right).$$

1.2. An alternative by Simon

In his breakthrough work, Simon Simon (2015) proposed taking majority votes of three ERMs trained on certain sub-samples of the training sample.[1] However, his algorithm is slightly different than ours. Concretely, he proposed the following algorithm: given an ERM algorithm and labelled training sample, partition S into three equal-sized disjoint pieces S_1, S_2, S_3 and for $i = 1, 2, 3$, run any ERM algorithm on $((S_1, \dots, S_i), f^\star((S_1, \dots, S_i)))$ to obtain $\widehat{f}_{S_1}, \widehat{f}_{(S_1,S_2)}, \widehat{f}_{(S_1,S_2,S_3)}$, and combine them via a majority vote to produce the function $\operatorname{Maj}(\widehat{f}_{S_1}, \widehat{f}_{(S_1,S_2)}, \widehat{f}_{(S_1,S_2,S_3)})$. Intuitively, more training data for the ERM should be better and Simon also proved the following high-probability upper bound on his algorithm's error:

$$\operatorname{err}_P\left(\operatorname{Maj}\left(\widehat{f}_{S_1}, \widehat{f}_{(S_1,S_2)}, \widehat{f}_{(S_1,S_2,S_3)}\right)\right) = O\left(\frac{d}{n}\log\left(\log\left(\frac{n}{d}\right)\right) + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right). \tag{5}$$

This bound is asymptotically smaller than the tight bound Eq. (1) that holds for a single ERM.

We note that Simon also discusses the applicability of his analysis to more general majorities of ERMs including the Majority-of-Three function $\operatorname{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})$ analyzed in this work.[2] However, adopting the approach in Simon (2015), the error of Majority-of-Three is controlled by the same upper bound as expressed in (5), which is suboptimal as demonstrated by Theorem 1. Furthermore, we additionally remark that a similar in spirit construction based on the majority of three functions has been extensively studied in Schapire's PhD thesis Schapire (1992). However, his approach (inspired by what we now know as boosting) works with essentially any learning algorithm and is not necessarily limited to ERM.

In the same work Simon (2015), Simon further showed that for a specific function class $\mathcal{F}$ for which there is a choice of target function $f^\star \in \mathcal{F}$ and hard distribution P that certify the tightness

1. Simon studied majority votes over any odd number L of ERMs trained on specific sub-samples of the data. He also proved bounds on the error of these majority votes that shrunk as L increased.
2. Simon's analysis applies to any majority where each of the participating ERMs is trained on an independent constant fraction of the training sample.

of Eq. (1) for a certain choice ERM, his algorithm can actually achieve an optimal upper bound matching Eq. (2) for $\mathcal{F}$ regardless of the choice of $f^\star \in \mathcal{F}$ and P. Unfortunately, we prove the following lower bound that shows that the upper bound Eq. (5) cannot be improved in general, answering a question posed by Simon.

Theorem 4 *For any sample size n that is divisible by 6 and positive integer $d \le n$, there is a function class $\mathcal{F} \subseteq \{0,1\}^{[0,1]}$ with VC dimension $4d$, distribution P over $[0,1]$, target function $f^\star \in \mathcal{F}$, and an ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$ such that the following holds: given i.i.d. training samples $S_1, S_2, S_3 \sim P^n$,*

$$\mathrm{err}_P\left(\mathrm{Maj}\left(\widehat{f}_{S_1}, \widehat{f}_{(S_1,S_2)}, \widehat{f}_{(S_1,S_2,S_3)}\right)\right) = \Omega\left(\frac{d}{n}\log\left(\log\left(\frac{n}{d}\right)\right)\right),$$

with probability at least $2/3$ over the randomness of $S = (S_1, S_2, S_3)$.

This result shows that Simon's algorithm unfortunately cannot achieve the optimal bound Eq. (3) in general. This indicates that it is important that the ERM algorithm used in Majority-of-Three is instantiated on disjoint subsets of the training sample.

1.3. Notation

We use $\mathcal{X}$ to denote the *instance space*, $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ to denote a function class, and let $\mathcal{Z} = \mathcal{X} \times \{0,1\}$. Throughout, P is a distribution over $\mathcal{X}$ and $f^\star \in \mathcal{F}$ is the unknown *target function* in the class. For $n \in \mathbb{N}$ and a distribution P, we denote by P^n the product distribution of P. We say that a sequence $S = (X_1 \ldots, X_n)$ is a *training sample* of size n where X_i are i.i.d. samples from a distribution P. For a training sample $S = (X_1, \ldots, X_n)$, we find it useful to write $(S, f^\star(S)) = ((X_1, f^\star(X_1)), \ldots, (X_n, f^\star(X_n)))$, and we call this the *labelled training sample*. For training samples $S_1 = (X_1, \ldots, X_n)$ and $S_2 = (X_{n+1}, \ldots, X_{n+m})$ we let $(S_1, S_2) = (X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m})$, and for S_1, S_2 and S_3 we take $(S_1, S_2, S_3) = ((S_1, S_2), S_3)$. We define the error of a binary function f under distribution P and target function $f^\star$ to be $\mathrm{err}_P(f) = \mathbf{Pr}_{X\sim P}[f(X) \neq f^\star(X)]$. For any measurable set $R \subseteq \mathcal{X}$, we define P_R to be the conditional distribution of P restricted to R, i.e. for $X \sim P_R$ we have that for any measurable function g that $\mathbb{E}_{X\sim P_R}[g(X)] = \mathbb{E}_{X\sim P}[g(X)\mathbf{1}\{X \in R\}] / \mathbf{Pr}_{X\sim P}[X \in R]$.

For a function class $\mathcal{F}$ and subset $U = \{x_1, \ldots, x_d\} \subseteq \mathcal{X}$ of d points we let $\mathcal{F}|_U$ denote the set $\{y \in \{0,1\}^d \mid \exists f \in \mathcal{F} : \forall i \in [d],\ f(x_i) = y_i\}$. The *Vapnik-Chervonenkis (VC) dimension* is then defined as the largest number d such that there exists a point set $U \subseteq \mathcal{X}$ of size d such that the cardinality of $\mathcal{F}|_U$ is 2^d. We use $\log(x)$ and $\ln(x)$ to denote $\log_2(x)$ and $\log_e(x)$ respectively and we also use $\mathrm{Log}(x) := \max\{2, \log_2(x)\}$ to denote a truncated logarithm.

Let $\mathcal{Z}^* = \cup_{i=1}^{\infty}\mathcal{Z}^i$ be the set of all possible labelled training samples. We define a *learning algorithm* $\widehat{f}$ to be a mapping $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$. That is, given a labelled training sample $(S, f^\star(S))$ as input, $\widehat{f}(\cdot; (S, f^\star(S))) : \mathcal{X} \to \{0,1\}$ is the function that is learned from $(S, f^\star(S))$. For ease of reading, we often denote the learned function by $\widehat{f}_S := \widehat{f}(\cdot; (S, f^\star(S)))$. A learning algorithm $\widehat{f}$ is an *Empirical Risk Minimizer (ERM)* for the class $\mathcal{F}$ if, given a labelled training sample $(S, f^\star(S))$ as input, it output a function $\widehat{f}_S$ in $\mathcal{F}$ that satisfies $\widehat{f}_S(X_i) = f^\star(X_i)$ for every X_i that appears in S. We define the majority vote of k binary functions $f_1, \ldots, f_k : \mathcal{X} \to \{0,1\}$ to be the function

$$\mathrm{Maj}(f_1, \ldots, f_k)(x) := \mathbf{1}\{f_1(x) + \cdots + f_k(x) > k/2\}.$$

2. Majority-of-Three is optimal in-expectation

In this section, we prove the main in-expectation result for the Majority-of-Three algorithm. Before we prove our result, we will find it helpful to introduce some auxiliary notation. Throughout this section, we set $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$ to be a fixed (but arbitrary) ERM algorithm. Fix a distribution P over $\mathcal{X}$ and let $f^\star \in \mathcal{F}$ be the target function. For any $x \in \mathcal{X}$ we let

$$p_x(n, f^\star, P) = \Pr_{S \sim P^n}[\widehat{f}_S(x) \neq f^\star(x)].$$

In words, $p_x(n, f^\star, P)$ is the chance that $\widehat{f}_S$ errs on the point x for an *average* sample $S \sim P^n$. We now define a partition of $\mathcal{X}$ based on $p_x(n, f^\star, P)$. Consider the following sets for any $i \in \mathbb{N}$:

$$R_i(n, f^\star, P) = \{x \in \mathcal{X} : p_x(n, f^\star, P) \in (2^{-i}, 2^{-i+1}]\}.$$

We often write $R_i = R_i(n, f^\star, P)$ and $p_x = p_x(n, f^\star, P)$ since n, P, and $f^\star$ will always be clear from the context. With this notation in place, we are now ready to prove that Majority-of-Three has an optimal in-expectation upper bound on its error.

Theorem 1 *Fix a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$ and target function $f^\star \in \mathcal{F}$. For any ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$ it follows that*

$$\mathop{\mathbb{E}}_{S_1,S_2,S_3 \sim P^n}\left[\mathrm{err}_P\left(\mathrm{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})\right)\right] = O\left(\frac{d}{n}\right).$$

To prove Theorem 1, we require the following lemma which says that two ERMs trained on 2 i.i.d. training samples of the same size rarely makes a mistake on the same point.

Lemma 5 *Fix a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$ and target function $f^\star \in \mathcal{F}$. For any ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$ it follows that*

$$\mathop{\mathbb{E}}_{S_1,S_2 \sim P^n}\left[\Pr_{X \sim P}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right]\right] \le c\frac{d}{n},$$

where c is a universal constant.

We postpone the proof of Lemma 5 for now and show how it implies Theorem 1.

Proof [Proof of Theorem 1] For any fixed $x \in \mathcal{X}$ and fixed samples S_1, S_2, S_3, if it is the case that $\mathrm{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})(x) \neq f^\star(x)$, then there must be at least two distinct indices $i, j \in [3]$ such that $\widehat{f}_{S_i}(x) \neq f^\star(x)$ *and* $\widehat{f}_{S_j}(x) \neq f^\star(x)$. So,

$$\begin{aligned}\mathrm{err}_P\left(\mathrm{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})\right) &= \Pr_{X \sim P}\left[\mathrm{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})(X) \neq f^\star(X)\right] \\ &\le \sum_{\substack{i,j \in [3] \\ i<j}} \Pr_{X \sim P}\left[\widehat{f}_{S_i}(X) \neq f^\star(X) \wedge \widehat{f}_{S_j}(X) \neq f^\star(X)\right].\end{aligned}$$

Combining the above and Lemma 5 gives us

$$\mathop{\mathbb{E}}_{S_1,S_2,S_3 \sim P^n}\left[\mathrm{err}_P\left(\mathrm{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})\right)\right] \le 3c\frac{d}{n}.$$

This concludes the proof. ■

We now move on to proving Lemma 5, where we will use the following lemma.

Lemma 6 *Fix a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$, target function $f^\star \in \mathcal{F}$, and $R \subseteq \mathcal{X}$ such that $\mathbf{Pr}_{X\sim P}[X \in R] \neq 0$. For any ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$ it follows that*

$$\mathop{\mathbb{E}}_{S\sim P^n}\left[\mathrm{err}_{P_R}\left(\widehat{f}_S\right)\right] \leq 20\frac{d\,\mathrm{Log}(e\,\mathbf{Pr}_{X\sim P}[X \in R]n/d)}{\mathbf{Pr}_{X\sim P}[X \in R]n}.$$

Lemma 6 is an immediate consequence of the celebrated uniform convergence principle and a simple proof can be found in Appendix A.1. We now prove Lemma 5.

Proof [Proof Lemma 5] Let S_1 and S_2 be independent samples from P^n. By the independence of S_1 and S_2 and the definition of p_x we have, for any $x \in \mathcal{X}$, that

$$\mathop{\mathbf{Pr}}_{S_1,S_2\sim P^n}\left[\widehat{f}_{S_1}(x) \neq f^\star(x) \wedge \widehat{f}_{S_2}(x) \neq f^\star(x)\right] = \prod_{i=1}^{2}\mathop{\mathbf{Pr}}_{S_i\sim P^n}\left[\widehat{f}_{S_i}(x) \neq f^\star(x)\right]$$
$$= \mathop{\mathbf{Pr}}_{S_1\sim P^n}\left[\widehat{f}_{S_1}(x) \neq f^\star(x)\right]^2 = p_x^2.$$

Using the above, the law of total expectation with partitioning $(R_i)_{i\in\mathbb{N}}$, and swapping the order of expectations (X and (S_1, S_2) being independent), we get that

$$\mathop{\mathbb{E}}_{S_1,S_2\sim P^n}\left[\mathop{\mathbf{Pr}}_{X\sim P}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right]\right]$$
$$= \sum_{i=1}^{\infty}\mathop{\mathbf{Pr}}_{X\sim P}[X \in R_i]\mathop{\mathbb{E}}_{X\sim P}\left[p_X^2|X \in R_i\right] \leq \sum_{i=1}^{\infty}\mathop{\mathbf{Pr}}_{X\sim P}[X \in R_i]\,2^{-2i+2},$$

where the inequality follows from the fact that $p_x \leq 2^{-i+1}$ for every $x \in R_i$. We will now show that $\mathbf{Pr}_{X\sim P}[X \in R_i] \leq cdi2^i/n$ for every $i \in \mathbb{N}$ (for a universal constant $c \geq 1$ chosen below), which combined with the above gives us

$$\mathop{\mathbb{E}}_{S_1,S_2\sim P^n}\left[\mathop{\mathbf{Pr}}_{X\sim P}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right]\right] \leq \frac{4cd}{n}\sum_{i=1}^{\infty} i2^{-i} \leq 8c\frac{d}{n}.$$

This yields the claim with the constant $8c$. Towards a contradiction, assume there is an $i \in \mathbb{N}$ such that $\mathbf{Pr}_{X\sim P}[X \in R_i] > cdi2^i/n$, which is equivalent to $\mathbf{Pr}_{X\sim P}[X \in R_i]\,n/d > ci2^i \geq 1$. Using this assumption, the fact that $x \to \mathrm{Log}(ex)/x$ is decreasing for $x > 0$, and Lemma 6, we have

$$\mathop{\mathbb{E}}_{S_1\sim P^n}\left[\mathrm{err}_{P_{R_i}}\left(\widehat{f}_{S_1}\right)\right] \leq 20\frac{\mathrm{Log}(e\,\mathbf{Pr}_{X\sim P}[X \in R_i]\,n/d)}{(\mathbf{Pr}_{X\sim P}[X \in R_i]\,n/d)} \leq 20\frac{\mathrm{Log}\left(eci2^i\right)}{ci2^i}. \tag{6}$$

By changing the order of expectations of the left hand side of the above and using the fact that $p_x > 2^{-i}$ for every $x \in R_i$, we also have

$$\mathop{\mathbb{E}}_{S_1\sim P^n}\left[\mathrm{err}_{P_{R_i}}\left(\widehat{f}_{S_1}\right)\right] = \mathop{\mathbb{E}}_{X\sim P_{R_i}}[p_X] > 2^{-i}. \tag{7}$$

Combining the upper bound (6), the lower bound (7), and the fact that the function $x \to \mathrm{Log}(ex)/x$ is decreasing for $x > 0$, we get

$$1 < 20\frac{\mathrm{Log}\left(eci2^i\right)}{ci} \leq 20\left(\frac{\mathrm{Log}\left(eci\right)}{ci} + \frac{2}{c}\right) \leq 20\frac{\mathrm{Log}(ec) + 2}{c}.$$

However, for c large enough, the right hand side of the above is strictly less than 1. This gives us the desired contradiction and concludes the proof. ■

3. A lower bound for certain majorities

In this section, we prove that not all majorities of 3 ERMs trained on subsets of the data are optimal. In particular, we show that Simon's Simon (2015) original partitioning scheme of the training sample into 3 sub-samples can produce a majority of 3 ERMs with sub-optimal error. Recall Simon's algorithm: partition the training sample $S = (S_1, S_2, S_3)$ into 3 equal pieces S_1, S_2, S_3, train 3 ERMs $\widehat{f}_{S_1}, \widehat{f}_{(S_1,S_2)}, \widehat{f}_{(S_1,S_2,S_3)}$, and return the majority vote $\mathrm{Maj}(\widehat{f}_{S_1}, \widehat{f}_{(S_1,S_2)}, \widehat{f}_{(S_1,S_2,S_3)})$. Simon proved that this algorithm enjoys the PAC upper bound

$$\mathrm{err}_P\left(\mathrm{Maj}\left(\widehat{f}_{S_1}, \widehat{f}_{(S_1,S_2)}, \widehat{f}_{(S_1,S_2,S_3)}\right)\right) = O\left(\frac{d}{n}\log\left(\log\left(\frac{n}{d}\right)\right) + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right).$$

The next theorem shows that this algorithm unfortunately has a matching lower bound on its error.

Theorem 4 *For any sample size n that is divisible by 6 and positive integer $d \le n$, there is a function class $\mathcal{F} \subseteq \{0,1\}^{[0,1]}$ with VC dimension $4d$, distribution P over $[0,1]$, target function $f^\star \in \mathcal{F}$, and an ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$ such that the following holds: given i.i.d. training samples $S_1, S_2, S_3 \sim P^n$,*

$$\mathrm{err}_P\left(\mathrm{Maj}\left(\widehat{f}_{S_1}, \widehat{f}_{(S_1,S_2)}, \widehat{f}_{(S_1,S_2,S_3)}\right)\right) = \Omega\left(\frac{d}{n}\log\left(\log\left(\frac{n}{d}\right)\right)\right),$$

with probability at least $2/3$ over the randomness of $S = (S_1, S_2, S_3)$.

Comparing the above bound with the upper bound in Theorem 1, we see that if the ERMs did not overlap in their sub-samples the log factor would not be present. The construction we use in our lower bound is a modification of the usual construction used to prove a lower bound on the error of a single ERM (see Auer and Ortner (2007); Simon (2015)). In these constructions, one takes the domain $\mathcal{X}$ to be a finite set of size roughly $n/\log(n/d)$ where $n \ge d$ is the sample size[3] and the function class $\mathcal{F}$ is taken to be all functions that assign the value 1 to at most d points on $\mathcal{X}$. Furthermore, the target function is set to be the 0 function, and the sampling distribution is the uniform distribution over $\mathcal{X}$. Finally, the "bad" ERM algorithm returns any function that assigns as many 1's to the domain as possible, while being consistent on the observed samples. The error of this ERM is tightly connected to the number of unique elements we sample from the domain. One can then use a coupon collector argument to show that the error is $\Omega(d\log(n/d)/n)$ with constant probability.

Simon noticed that we cannot directly use this "hard instance" to prove a lower bound on his algorithm due to the structure of the class $\mathcal{F}$ (Simon, 2015, Theorem 7). We get around this by considering a version of this construction that uses a continuous domain (instead of finite) and a function class consisting of functions that are unions of intervals (instead of points).

Before we prove Theorem 4, it will be convenient to introduce the following notation. For a positive integer d and non-empty set A, we define $A_{\lfloor d \rfloor}$ to be the set consisting of the smallest d

3. These results are often stated as lower bounds on the *sample complexity* for some target error ϵ.

elements of A with respect to an ordering of the elements of A. The ordering we use will be clarified when needed. We now prove Theorem 4.

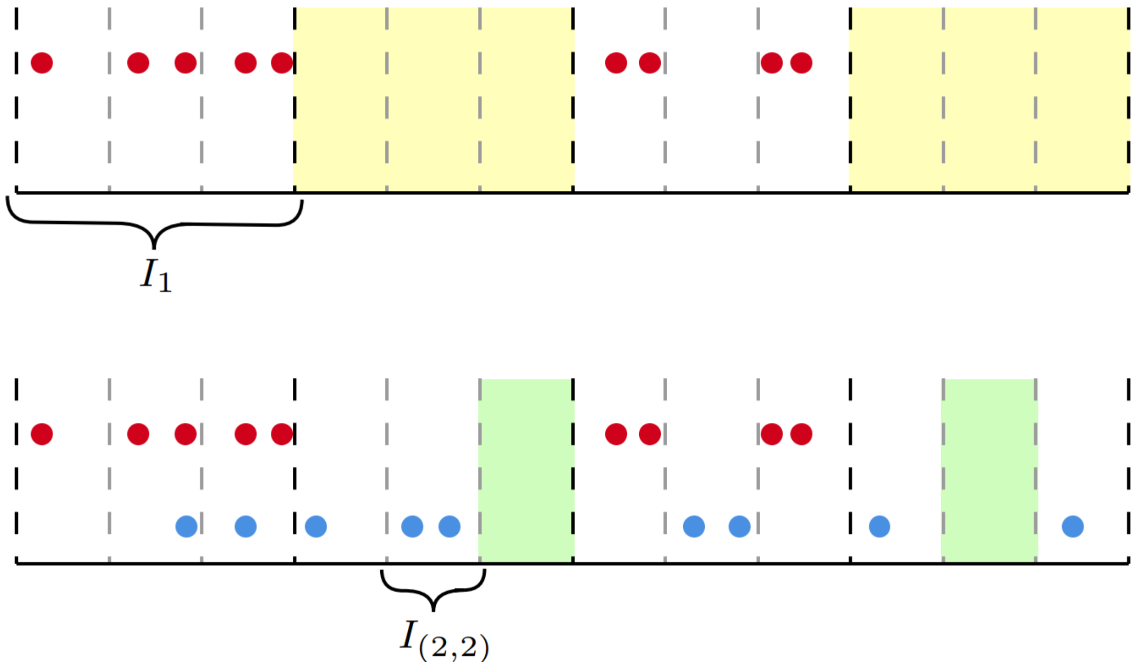

Figure 1: An illustration of the partitioning of the interval $(0, 1]$ for a training sample consisting of $m = 18$ points with $d = 2$. The interval $(0, 1]$ is partitioned into 4 intervals $I_1, \ldots, I_4$. Each interval I_i is further partitioned into the 4 subintervals $I_{(i,1)}, \ldots, I_{(i,4)}$. The red points correspond to the first half of the sample $(X_1, \ldots, X_9)$ and the blue points correspond to the second half of the sample $(X_{10}, \ldots, X_{18})$. The yellow highlighted regions are the first d intervals I_2 and I_4 that contain no points from $(X_1, \ldots, X_9)$. The green highlighted regions are the first d subintervals of I_2 and I_4 that contain no points from $(X_{10}, \ldots, X_{18})$. The green intervals are added to the union of intervals used by $\widehat{f}_S$ as their indices correspond to the set $L_1(S)$.

Proof [Proof of Theorem 4] Fix a sample size n divisible by 6 and positive integer $d \leq n$. Throughout, we will assume any interval considered is left-open and right-closed. A collection of intervals $I_1, \ldots, I_t \subseteq [0, 1]$ can be viewed as the binary function $f_{I_1 \cup \cdots \cup I_t}$ that satisfies $f_{I_1 \cup \cdots \cup I_t}(x) = 1$ if and only if there is an index j such that $x \in I_j$. We will consider the function class $\mathcal{F}$ that is the collection of all functions corresponding to the union of at most $2d$ interval. It is not hard to show that this class has VC dimension $4d$. We take P to be the uniform distribution on the domain $[0, 1]$ and choose the target function $f^\star$ to be the 0 function on the domain $[0, 1]$.

We now describe the "bad" ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0, 1\}$. For the remainder of the proof, $C > 0$ is a large universal constant that we will determine below. For a training sample size m, we define three collections of sets:

1. $\{I_i(m) : i \in [m_1]\}$ is the unique partition of $(0,1]$ into $m_1 := \lceil Cm/\ln(Cm/d)\rceil$ intervals of the same length.

2. $\{I_{i,j}(m) : i \in [m_1], j \in [m_2]\}$ where, for a fixed i, $\{I_{i,j}(m) : j \in [m_2]\}$ is the unique partition of $I_i(m)$ into $m_2 := \lceil 4Cm/(m_1 \ln(\ln(Cm/d)))\rceil$ intervals of the same length.

3. $\{J_i(m) : i \in m_3\}$ is the unique partition of $(0,1]$ into $m_3 := \lceil 2Cm/\ln(2Cm/d)\rceil$ intervals of the same length.

Given a labelled training sample $(S, f^\star(S)) = ((X_1, 0), \ldots, (X_m, 0))$ as input, the ERM algorithm $\widehat{f}$ constructs the function $\widehat{f}_S = \widehat{f}(\cdot; (S, f^\star(S)))$ in the following way:[4]

1. For $i \in [m_1]$, $j \in [m_2]$, and $k \in [m_3]$ define the sets

$$\tilde{I}_i(S) = \{x_1, \ldots, x_{\lfloor m/2\rfloor}\} \cap I_i(m),$$
$$\tilde{I}_{(i,j)}(S) = \{x_{\lfloor m/2\rfloor+1}, \ldots, x_m\} \cap I_{(i,j)}(m),$$
$$\tilde{J}_k(S) = \{x_1, \ldots, x_m\} \cap J_k(m).$$

2. Define the index sets

$$L_1(S) = \{(i,j) : i \in \{i' : \tilde{I}_{i'}(S) = \emptyset\}_{\lfloor d\rfloor}, \tilde{I}_{(i,j)}(S) = \emptyset\}_{\lfloor d\rfloor},$$ [5]
$$L_2(S) = \{k : \tilde{J}_k(S) = \emptyset\}_{\lfloor d\rfloor}.$$

3. Define the union of intervals

$$I_S = \left(\bigcup_{(i,j)\in L_1(S)} I_{i,j}(m)\right) \bigcup \left(\bigcup_{i\in L_2(S)} J_i(m)\right).$$

4. Finally, define the function $\widehat{f}_S = f_{I_S}$.

Observe that I_S is the union of at most $2d$ disjoint intervals, so $\widehat{f}_S$ will always be in the class $\mathcal{F}$. Furthermore, $\widehat{f}_S$ is always consistent with the sample S by construction. See Fig. 1 for an example of the resulting intervals considered by the set $L_1(S)$. Let $m = n/3$. From now on we use m_1 and m_2 to denote the number of intervals of the form $I_i(2m)$ and $I_{(i,j)}(2m)$ considered by $\widehat{f}_{(S_1,S_2)}$ respectively. Consider the unions of intervals I_{S_1} and $I_{(S_1,S_2)}$ corresponding to the ERM functions $\widehat{f}_{S_1}$ and $\widehat{f}_{(S_1,S_2)}$. The number m is divisible by 2 from our choice of n, so it follows that $J_i(m) = I_i(2m)$ and $\tilde{J}_i(S_1) = \tilde{I}_i(S_1, S_2)$, which implies $L_2(S_1) = \{k : \tilde{I}_k((S_1, S_2)) = \emptyset\}_{\lfloor d\rfloor}$. Thus, $\widehat{f}_{S_1}$ and $\widehat{f}_{(S_1,S_2)}$ agree, and simultaneously err, on every subinterval $I_{(i,j)}(2m)$ with $(i,j) \in L_1(S_1, S_2)$. Because P is the uniform distribution and every interval $I_{(i,j)}(2m)$ has length $1/(m_1 m_2) = \Theta(\ln(\ln(n/d))/n)$, it follows that the error of the majority vote satisfies

$$\mathrm{err}_P\left(\mathrm{Maj}\left(\widehat{f}_{S_1}, \widehat{f}_{(S_1,S_2)}, \widehat{f}_{(S_1,S_2,S_3)}\right)\right) \geq \frac{|L_1(S_1,S_2)|}{m_1 m_2} = \Omega\left(\frac{|L_1(S_1,S_2)|}{n}\ln\left(\ln\left(\frac{n}{d}\right)\right)\right).$$

4. This defines $\widehat{f}_S$ when $(S, f^\star(S))$ contains only 0 labels. On any $(S, f^\star(S))$ that contains a 1 label we return an arbitrary consistent function.

5. The ordering used for pairs (i,j) and (i',j') is the natural one: $(i,j) \leq (i',j')$ if $i < i'$ or $i = i'$ and $j \leq j'$.

Thus, if $|L_1(S_1, S_2)| = d$, we have

$$\mathrm{err}_P\left(\mathrm{Maj}\left(\widehat{f}_{S_1}, \widehat{f}_{(S_1,S_2)}, \widehat{f}_{(S_1,S_2,S_3)}\right)\right) = \Omega\left(\frac{d}{n}\log\left(\log\left(\frac{n}{d}\right)\right)\right),$$

so the claim of the theorem follows once we prove that

$$\mathop{\mathbf{Pr}}_{(S_1,S_2)\sim P^{2m}}\left[|L_1(S_1, S_2)| = d\right] \geq 2/3.$$

To this end, let $E_1 = E_1(S_1)$ be the event that S_1 satisfies $|L_2(S_1)| = d$ and let $E_2 = E_2((S_1, S_2))$ be the event that (S_1, S_2) satisfies $|L_1(S_1, S_2)| = d$. Using the law of total probability we get,

$$\mathop{\mathbf{Pr}}_{(S_1,S_2)\sim P^{2n/3}}\left[|L_1(S_1, S_2)| = d\right] = \mathop{\mathbf{Pr}}_{(S_1,S_2)\sim P^{2m}}\left[E_2\right] \geq \mathop{\mathbf{Pr}}_{(S_1,S_2)\sim P^{2m}}\left[E_2 \mid E_1\right] \mathop{\mathbf{Pr}}_{S_1\sim P^m}\left[E_1\right],$$

so it suffices to prove that $\mathbf{Pr}_{(S_1,S_2)\sim P^{2m}}\left[E_2 \mid E_1\right] \geq \sqrt{2/3}$ and $\mathbf{Pr}_{S_1\sim P^m}\left[E_1\right] \geq \sqrt{2/3}$. We omit the proof of the later inequality since it is very similar to the proof of the former inequality.

When E_1 occurs, we have $|L_2(S_1)| = |\{k : \tilde{J}_k(S_1) = \emptyset\}_{\lfloor d\rfloor}| = |\{k : \tilde{I}_k((S_1, S_2)) = \emptyset\}_{\lfloor d\rfloor}| = d$. So, showing that the event E_2 occurs conditioned on E_1 is equivalent to showing that *at least* d subintervals in the collection $\{I_{(i,j)} : i \in L_2(S_1), j \in [m_2]\}$ do not contain any points from the sample S_2. Let $Y \sim Q$ be the random variable that counts the number of points required to sample from P until $m_2d - d$ subintervals in $\{I_{(i,j)} : i \in L_2(S_1), j \in [m_2]\}$ contain one of the sampled points. Furthermore, let $Y_t \sim Q_t$ denote the random variable that counts the number of trials required to cover $(t + 1)$ subintervals given that we have covered t. Notice that Y_t is a geometric random variable with parameter $p_t = \frac{m_2d-t}{m_1m_2} = \frac{d}{m_1} - \frac{t}{m_1m_2}$ and $Y = \sum_{t=0}^{m_2d-d-1} Y_t$. It follows that

$$\mathop{\mathbf{Pr}}_{(S_1,S_2)\sim P^{2m}}\left[E_2 \mid E_1\right] \geq \mathop{\mathbf{Pr}}_{Y\sim Q}\left[Y \geq m\right] = \mathop{\mathbf{Pr}}_{Y_t\sim Q_t}\left[\sum_{t=0}^{m_2d-d-1} Y_t \geq m\right].$$

We can use a concentration inequality for sums of geometric random variables together with some simple calculations to show that

$$\mathop{\mathbf{Pr}}_{Y_t\sim Q_t}\left[\sum_{t=0}^{m_2d-d-1} Y_t \geq m\right] \geq \sqrt{2/3}, \tag{8}$$

when C is large enough. We defer these calculations to Appendix B. This concludes the proof. ■

4. High probability upper bound

In this section we prove our high-probability upper bound for the Majority-of-Three algorithm which we now restate for convenience.

Theorem 2 *Fix a function class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$ and target function $f^\star \in \mathcal{F}$. Fix any ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0, 1\}$. For any parameter $\delta \in (0, 1/2]$ it holds with probability at least $1 - \delta$ over the randomness of $S_1, S_2, S_3 \sim P^n$ that*

$$\mathrm{err}_P\left(\mathrm{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})\right) = O\left(\frac{d}{n}\log\left(\log\left(\min\left\{\frac{n}{d}, \frac{1}{\delta}\right\}\right)\right) + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right).$$

In this section it will be convenient to use the following notation: for a probability distribution P over $\mathcal{X}$ and set $R \subseteq \mathcal{X}$, we define $P(R) = \mathbf{Pr}_{X \sim P}[X \in R]$. Theorem 2 is a consequence of the following technical lemma.

Lemma 7 *Fix a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$ and target function $f^\star \in \mathcal{F}$. Fix an ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$. For any parameter $\delta \in (0, 1/2]$ it holds with probability at least $1-\delta$ over the randomness of $S_1, S_2 \sim P^n$ that*

$$\Pr_{X \sim P}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right]$$
$$\leq c\left(\frac{d}{n}\operatorname{Log}\left(\operatorname{Log}\left(\min\left\{\frac{n}{d}, \frac{1}{\delta}\right\}\right)\right) + \frac{1}{n}\operatorname{Log}\left(\frac{1}{\delta}\right)\right),$$

where c is a universal constant.

We now prove Theorem 2 using Lemma 7 and postpone the proof of Lemma 7.

Proof [Proof of Theorem 2] Since $\operatorname{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})(x) \neq f^\star(x)$ happens if and only if there exists two distinct indices $i, j \in [3]$ such that $\widehat{f}_{S_i}(X) \neq f^\star(X)$ *and* $\widehat{f}_{S_j}(X) \neq f^\star(X)$, we get that

$$\operatorname{err}_P\left(\operatorname{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})\right) \leq \sum_{\substack{i,j \in [3] \\ i<j}} \Pr_{X \sim P}\left[\widehat{f}_{S_i}(X) \neq f^\star(X) \wedge \widehat{f}_{S_j}(X) \neq f^\star(X)\right].$$

Using Lemma 7 with confidence parameter $\delta/3$ for every distinct pair $i, j \in [3]$ together with a union bound gives us, with probability at least $1-\delta$ over the randomness of (S_1, S_2, S_3), that

$$\operatorname{err}_P\left(\operatorname{Maj}(\widehat{f}_{S_1}, \widehat{f}_{S_2}, \widehat{f}_{S_3})\right) = O\left(\frac{d}{n}\operatorname{Log}\left(\operatorname{Log}\left(\min\left\{\frac{n}{d}, \frac{1}{\delta}\right\}\right)\right) + \frac{1}{n}\operatorname{Log}\left(\frac{1}{\delta}\right)\right).$$

This concludes the proof. ■

Before we prove Lemma 7, we provide a short overview of the proof. Our first step is to reuse the idea from Section 2 to partition the instance space $\mathcal{X}$ into sets $\{R_i\}_{i \in \mathbb{N}}$ based on the chance that an average ERM errs on a point in $x \in \mathcal{X}$. However, we use a different way to quantify the errors defining R_i by incorperating the failure parameter δ. For $i \geq 2$, we can actually reuse our in-expectation analysis from Section 2 together with a simple application of Markov's inequality and a sequence of union bounds. This gives us an upper bound on the joint error of two ERMs on the conditional distributions for all $\{R_i\}_{i \geq 2}$, with high probability. The major technical work of the proof lies in controlling the joint error of two ERMs on the conditional distribution of R_1. To do this, we borrow an idea from Simon Simon (2015) that views the probability of $\widehat{f}_{S_1}$ and $\widehat{f}_{S_2}$ jointly erring as the probability that $\widehat{f}_{S_1}$ errs times the probability that $\widehat{f}_{S_2}$ errs conditioned on $\widehat{f}_{S_1}$ erring.[6] A crucial technicality that differentiates our setting from Simon's is that the probability that $\widehat{f}_{S_1}$ and $\widehat{f}_{S_2}$ jointly err is taken over a *conditional distribution* P_R rather than the distribution P from which the samples S_1 and S_2 are drawn.

The following lemma formalizes how we can control the joint error of two ERMs under P_R.

6. This idea used by Simon in fact builds upon even earlier work of Hanneke Hanneke (2009) in the context of active learning. It was also applied in the context of PAC learning by Darnstädt Darnstädt (2015).

Lemma 8 *Fix a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$, target function $f^\star \in \mathcal{F}$ and $R \subseteq \mathcal{X}$ such that $P(R) \neq 0$. Fix an ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$. For any parameter $\delta \in (0, 1/2]$ it holds with probability at least $1-\delta$ over the randomness of $S_1, S_2 \sim P^n$ that*

$$\Pr_{X \sim P_R}[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)] \leq \frac{8 \max\{d \operatorname{Log}(8e \operatorname{Log}(eP(R)\, n/d)), \operatorname{Log}(8/\delta)\}}{P(R)\, n}.$$

We now prove Lemma 7 using Lemma 8, the proof of Lemma 8 can be found in Appendix A.4.
Proof [Proof of Lemma 7] We use the same definition for p_x (see Section 2) but redefine the sets $\{R_i\}_{i \in \mathbb{N}}$ to be

$$R_1 = \left\{x \in \mathcal{X} : p_x \in (2^{-1}\delta/\operatorname{Log}(1/\delta), 1]\right\},$$

and for any integer $i \geq 2$,

$$R_i = \left\{x \in \mathcal{X} : p_x \in \left(2^{-i}\delta/\operatorname{Log}(1/\delta), 2^{-i+1}\delta/\operatorname{Log}(1/\delta)\right]\right\}.$$

Using the law of total probability we have

$$\begin{aligned}
&\mathbf{Pr}_{X \sim P}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right] \\
&= P(R_1) \Pr_{X \sim P_{R_1}}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right] \\
&+ \sum_{i=2}^{\infty} P(R_i) \Pr_{X \sim P_{R_i}}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right].
\end{aligned}$$

We will prove that there is a universal constant $c > 0$ such that the events

$$\begin{aligned}
E_1 = E_1((S_1, S_2)) \coloneqq \Bigg\{ P(R_1) \Pr_{X \sim P_{R_1}}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right] \leq \\
c \max\left\{\frac{d \operatorname{Log}(\operatorname{Log}(\min\{n/d, 1/\delta\}))}{n}, \frac{\operatorname{Log}(1/\delta)}{n}\right\} \Bigg\},
\end{aligned}$$

and

$$E_2 = E_2((S_1, S_2)) \coloneqq \left\{ \sum_{i=2}^{\infty} P(R_i) \Pr_{X \sim P_{R_i}}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right] \leq c\frac{d}{n} \right\}$$

each happen with probability at least $1 - \delta/2$ over the randomness of (S_1, S_2). The claim of Lemma 7 then follows from a union bound. Define the set $I = \{i \geq 2 : P(R_i) \neq 0\}$. To prove that E_1 and E_2 each occur with high probability, we will use the following proposition.

Proposition 9 *In the setting of Lemma 7 we have the following:*

1. *There is a universal constant c' such that for any $i \in \mathbb{N}$*

$$P(R_i) \leq \frac{c' 2^i d \operatorname{Log}(2^i \operatorname{Log}(1/\delta)/\delta) \operatorname{Log}(1/\delta)}{\delta n}.$$

2. *With probability at least* $1-\delta/2$ *over the randomness of* (S_1,S_2) *we have, simultaneously for all* $i \in I = \{i \geq 2 : P(R_i) \neq 0\}$*, that*

$$\Pr_{X\sim P_{R_i}}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right] \leq \frac{5\cdot 2^{-1.1i}\delta}{\mathrm{Log}^2(1/\delta)}.$$

We defer the proof of Proposition 9 to Appendix A.3 as its proof is similar to that of Lemma 5.

We first prove that the event E_1 occurs with high probability. If $P(R_1) = 0$, then we immediately have that $\mathbf{Pr}_{(S_1,S_2)}[E_1] = 1$. We now consider the case that $P(R_1) \neq 0$. From Item 1 of Proposition 9 we can conclude there is a universal constant $\tilde{c}$ such that $P(R_1) \leq \min\{1, \tilde{c}\frac{d\,\mathrm{Log}^2(1/\delta)}{\delta n}\}$. Using this combined with Lemma 8 we have, with probability at least $1-\delta/2$ over the randomness of (S_1,S_2), that

$$\begin{aligned}
&P(R_1)\Pr_{X\sim P_{R_1}}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right]\\
&\leq 8\max\left\{\frac{d\,\mathrm{Log}(8e\,\mathrm{Log}(\min\{en/d, e\tilde{c}\,\mathrm{Log}^2(1/\delta)/\delta\}))}{n}, \frac{\mathrm{Log}(16/\delta)}{n}\right\}\\
&\leq c\max\left\{\frac{d\,\mathrm{Log}(\mathrm{Log}(\min\{n/d, 1/\delta\}))}{n}, \frac{\mathrm{Log}(1/\delta)}{n}\right\},
\end{aligned}$$

where the last inequality holds for c large enough.

We now prove that the event E_2 occurs with high probability. Combining Items 1 and 2 of Proposition 9 we have, with probability at least $1-\delta/2$ over the randomness of (S_1,S_2), that

$$\begin{aligned}
&\sum_{i=2}^{\infty}P(R_i)\Pr_{X\sim P_{R_i}}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right]\\
&\quad\leq \sum_{i\notin I}0 + \sum_{i\in I}\frac{2^i c' d\,\mathrm{Log}(2^i\,\mathrm{Log}(1/\delta)/\delta)\,\mathrm{Log}(1/\delta)}{\delta n}\cdot\frac{5\cdot 2^{-1.1i}\delta}{\mathrm{Log}^2(1/\delta)}\\
&\quad\leq \frac{5c'd}{n}\sum_{i=2}^{\infty}\frac{2^{-0.1i}\,\mathrm{Log}(2^i\,\mathrm{Log}(1/\delta)/\delta)}{\mathrm{Log}(1/\delta)}\\
&\quad\leq \frac{5c'd}{n}\sum_{i=2}^{\infty}\frac{2^{-0.1i}\cdot(i\,\mathrm{Log}(2)+\mathrm{Log}(\mathrm{Log}(1/\delta))+\mathrm{Log}(1/\delta))}{\mathrm{Log}(1/\delta)}\\
&\quad\leq c\frac{d}{n},
\end{aligned}$$

where the last inequality holds for c large enough. This concludes the proof. ■

Acknowledgments

M. M. Høgsgaard and K. G. Larsen are supported by a DFF Sapere Aude Research Leader Grant No. 9064-00068B.

References

Ishaq Aden-Ali, Yeshwanth Cherapanamjeri, Abhishek Shetty, and Nikita Zhivotovskiy. The one-inclusion graph algorithm is not always optimal. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*, volume 195 of *Proceedings of Machine Learning Research*, pages 72–88. PMLR, 2023a. URL https://proceedings.mlr.press/v195/aden-ali23a.html.

Ishaq Aden-Ali, Yeshwanth Cherapanamjeri, Abhishek Shetty, and Nikita 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, 2023b.

Peter Auer and Ronald Ortner. A new PAC bound for intersection-closed concept classes. *Machine Learning*, 66(2):151–163, 2007.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. In *Conference on Learning Theory*, pages 582–609. PMLR, 2020.

Leo Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, aug 1996. ISSN 0885-6125. doi: 10.1023/A:1018054314350. URL https://doi.org/10.1023/A:1018054314350.

Malte Darnstädt. The optimal PAC bound for intersection-closed concept classes. *Information Processing Letters*, 115(4):458–461, 2015.

Steve Hanneke. *Theoretical Foundations of Active Learning*. Doctoral thesis, Carnegie-Mellon University, Machine Learning Department, 2009.

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

Steve Hanneke. Refined error bounds for several learning algorithms. *The Journal of Machine Learning Research*, 17(1):4667–4721, 2016b.

David Haussler, Nick Littlestone, and Manfred K Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

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

Kasper Green 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. URL https://proceedings.mlr.press/v195/larsen23a.html.

Yi Li, Philip M Long, and Aravind Srinivasan. The one-inclusion graph algorithm is near-optimal for the prediction model of learning. *IEEE Transactions on Information Theory*, 47(3):1257–1261, 2001.

Robert E. Schapire. *The Design and Analysis of Efficient Learning Algorithms*. ACM Doctoral Dissertation Awards. The MIT Press, 1992. ISBN 9780262193252.

Hans U Simon. An almost optimal PAC algorithm. In *Conference on Learning Theory*, pages 1552–1563. PMLR, 2015.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Vladimir Vapnik and Alexey Chervonenkis. A class of algorithms for pattern recognition learning. *Avtomatika i Telemekhanika*, 25(6):937–945, 1964.

Vladimir Vapnik and Alexey Chervonenkis. Algorithms with complete memory and recurrent algorithms in the problem of learning pattern recognition. *Avtomatika i Telemekhanika*, pages 95–106, 1968.

Vladimir Vapnik and Alexey Chervonenkis. On uniform convergence of the frequencies of events to their probabilities. *Teoriya Veroyatnostei i ee Primeneniya*, 16(2):264–279, 1971.

Vladimir Vapnik and Alexey Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

Manfred K Warmuth. The optimal PAC algorithm. In *International Conference on Computational Learning Theory*, pages 641–642. Springer, 2004.

Appendix A. Ommited proofs from Sections 2 and 4

In this appendix we prove Lemma 6, Lemma 8, Proposition 9, and Lemma 11. These results are by-products of the classic *uniform convergence* result which uniformly bounds the error of any function in $\mathcal{F}$ that is consistent with the training sample. To state the result, we first introduce some notation. For a training sample $S = ((X_1, f^\star(X_1)), \ldots, (X_n, f^\star(X_n)))$, let $\mathcal{F}_S$ denote the functions in $\mathcal{F}$ that are consistent with S, i.e., $f \in \mathcal{F}_S$ if and only if $f(X_i) = f^\star(X_i)$ for every $i \in [n]$.

Lemma 10 *[Uniform convergence Blumer et al. (1989)] Fix a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$ and target function $f^\star \in \mathcal{F}$. For any parameter $\delta \in (0, 1/2]$ it holds with probability at least $1-\delta$ over $S \sim P^n$ that*

$$\sup_{f \in \mathcal{F}_S} \mathrm{err}_P(f) \leq 2\left(\frac{d\log(2en/d) + \log(2/\delta)}{n}\right).$$

In what follows, we will use the slightly weaker bound

$$\sup_{f \in \mathcal{F}_S} \mathrm{err}_P(f) \leq 4\max\left\{\frac{d\,\mathrm{Log}(2en/d)}{n}, \frac{\mathrm{Log}(2/\delta)}{n}\right\}. \tag{9}$$

A.1. Proof of Lemma 6

We now prove Lemma 6 which we restate here for convenience.

Lemma 6 *Fix a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$, target function $f^\star \in \mathcal{F}$, and $R \subseteq \mathcal{X}$ such that $\mathbf{Pr}_{X\sim P}[X \in R] \neq 0$. For any ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$ it follows that*

$$\mathop{\mathbb{E}}_{S\sim P^n}\left[\mathrm{err}_{P_R}\left(\widehat{f}_S\right)\right] \le 20\frac{d\,\mathrm{Log}(e\,\mathbf{Pr}_{X\sim P}[X\in R]n/d)}{\mathbf{Pr}_{X\sim P}[X\in R]n}.$$

Proof Consider the case that $\mathbf{Pr}_{X\sim P}[X \in R]n \le 4d$. In this case the claim follows easily since $\mathrm{err}_{P_R}\left(\widehat{f}_S\right) \le 1$ and $x \to \mathrm{Log}(ex)/x$ is decreasing in x for $x > 0$, so $20\frac{d\,\mathrm{Log}(e\,\mathbf{Pr}_{X\sim P}[X\in R]n/d)}{\mathbf{Pr}_{X\sim P}[X\in R]n} > 1$. We now consider the case that $\mathbf{Pr}_{X\sim P}[X \in R]n > 4d$. For any $m \in \mathbb{N}$ we define the event $E_m = E_m(S) = \{|\{i \in [n] : X_i \in R\}| = m\}$. Similarly, we define the event

$$E = E(S) = \bigcup_{m \ge \mathbf{Pr}_{X\sim P}[X\in R]n/2} E_m.$$

It follows from a Chernoff bound and our assumption that $\mathbf{Pr}_{X\sim P}[X \in R]n > 4d$ that

$$\mathop{\mathbf{Pr}}_{S\sim P^n}[E] \ge 1 - \exp\left(-\frac{\mathbf{Pr}_{X\sim P}[X\in R]n}{8}\right) \ge 1 - \frac{8}{\mathbf{Pr}_{X\sim P}[X\in R]n}.$$

Using the law of total probability we have

$$\mathop{\mathbb{E}}_{S\sim P^n}\left[\mathrm{err}_{P_R}\left(\widehat{f}_S\right)\right] \le \mathop{\mathbb{E}}_{S\sim P^n}\left[\mathrm{err}_{P_R}\left(\widehat{f}_S\right) \;\middle|\; E\right] + \frac{8}{\mathbf{Pr}_{X\sim P}[X\in R]n}. \tag{10}$$

So, if we show that for any $m \ge \mathbf{Pr}_{X\sim P}[X \in R]n/2$ that

$$\mathop{\mathbb{E}}_{S\sim P^n}\left[\mathrm{err}_{P_R}\left(\widehat{f}_S\right) \;\middle|\; E_m\right] \le \frac{12d\,\mathrm{Log}\left(e\,\mathbf{Pr}_{X\sim P}[X\in R]n/d\right)}{\mathbf{Pr}_{X\sim P}[X\in R]n}, \tag{11}$$

the claim follows from one more application of the law of total probability applied to the first term on the right hand side of Eq. (10).

We now prove Eq. (11). Using the non-negativity of $\mathrm{err}_{P_R}\left(\widehat{f}_S\right)$ we have

$$\begin{aligned}\mathop{\mathbb{E}}_{S\sim P^n}\left[\mathrm{err}_{P_R}\left(\widehat{f}_S\right) \;\middle|\; E_m\right] &= \int_0^\infty \mathop{\mathbf{Pr}}_{S\sim P^n}\left[\mathrm{err}_{P_R}\left(\widehat{f}_S\right) > x \;\middle|\; E_m\right] dx \\ &\le \frac{4d\,\mathrm{Log}\left(2em/d\right)}{m} + \int_{\frac{4d\,\mathrm{Log}(2em/d)}{m}}^1 \mathop{\mathbf{Pr}}_{S\sim P^n}\left[\mathrm{err}_{P_R}\left(\widehat{f}_S\right) > x \;\middle|\; E_m\right] dx.\end{aligned} \tag{12}$$

Notice that conditioned on E_m, the m samples that land in R form an i.i.d. sample from the conditional distribution P_R. Thus, any ERM trained on S is also consistent with m i.i.d. samples from P_R, so we can apply uniform convergence (Lemma 10) to control the error of any ERM when measured with respect to the conditional distribution P_R. Setting $\delta = 2^{1-\frac{mx}{4}}$ we have

$$\int_{\frac{4d\,\mathrm{Log}(2em/d)}{m}}^1 \mathop{\mathbf{Pr}}_{S\sim P^n}\left[\mathrm{err}_{P_R}\left(\widehat{f}_S\right) > x \;\middle|\; E_m\right] dx$$

$$= \int_{\frac{4d\,\mathrm{Log}(2em/d)}{m}}^{1} \mathop{\mathbf{Pr}}_{S\sim P^n}\left[\mathrm{err}_{P_R}\left(\widehat{f}_S\right) > \frac{4\,\mathrm{Log}(2/\delta)}{m} \,\middle|\, E_m\right] dx$$
$$\leq 2\int_{\frac{4d\,\mathrm{Log}(2em/d)}{m}}^{1} 2^{-\frac{mx}{4}}\,dx$$
$$\leq 2\left(\frac{4\cdot 2^{-d\,\mathrm{Log}(2em/d)}}{m\ln(2)}\right) \leq \frac{2d\,\mathrm{Log}\,(2em/d)}{m}.$$

Here, the first equality follows from the fact that $m \geq \mathbf{Pr}_{X\sim P}[X \in R]n/2 \geq 2d$ and our choice of δ. The second inequality follows from Eq. (9) and the final inequality follows from the fact that $d\,\mathrm{Log}(2em/d) \geq 2$. Now, using the fact that $x \to \mathrm{Log}(2ex)/x$ is decreasing for $x > 0$ together with the fact that $m \geq \mathbf{Pr}_{X\sim P}[X \in R]n/2 \geq 2d$, we conclude

$$\mathop{\mathbb{E}}_{S\sim P^n}\left[\mathrm{err}_{P_R}\left(\widehat{f}_S\right) \,\middle|\, E_m\right] \leq \frac{6d\,\mathrm{Log}\,(2em/d)}{m} < \frac{12d\,\mathrm{Log}\,(e\,\mathbf{Pr}_{X\sim P}[X\in R]n/d)}{\mathbf{Pr}_{X\sim P}[X\in R]n},$$

as claimed. ■

A.2. Proof of Lemma 8.

In this section we prove Lemma 8. To do so, we will need the following lemma which is a simple consequence of uniform convergence. We defer the proof of this lemma to Appendix A.4.

Lemma 11 *Fix a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$, target function $f^\star \in \mathcal{F}$, and $R \subseteq \mathcal{X}$ such that $P(R) \neq 0$. Fix an ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$. For any parameter $\delta \in (0,1/2]$ it holds with probability at least $1-\delta$ over the randomness of $S \sim P^n$ that*

$$\mathrm{err}_{P_R}\left(\widehat{f}_S\right) \leq 8\max\left\{\frac{d\,\mathrm{Log}(eP\,(R)\,n/d)}{P\,(R)\,n}, \frac{\mathrm{Log}(4/\delta)}{P\,(R)\,n}\right\}.$$

We are now ready to prove Lemma 8, which we restate before the proof for convenience

Lemma 8 *Fix a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$, target function $f^\star \in \mathcal{F}$ and $R \subseteq \mathcal{X}$ such that $P(R) \neq 0$. Fix an ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$. For any parameter $\delta \in (0,1/2]$ it holds with probability at least $1-\delta$ over the randomness of $S_1, S_2 \sim P^n$ that*

$$\Pr_{X\sim P_R}[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)] \leq \frac{8\max\{d\,\mathrm{Log}(8e\,\mathrm{Log}(eP\,(R)\,n/d)), \mathrm{Log}(8/\delta)\}}{P\,(R)\,n}.$$

Proof [Proof of Lemma 8] We will prove that the event

$$E = E((S_1,S_2)) := \left\{\Pr_{X\sim P_R}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right]\right.$$
$$\left.\leq 8\max\left\{\frac{d\,\mathrm{Log}(8e\,\mathrm{Log}(eP\,(R)\,n/d))}{P\,(R)\,n}, \frac{\mathrm{Log}(8/\delta)}{P\,(R)\,n}\right\}\right\}$$

occurs with high probability. Let B_1 denote the (random) set $\{x \in \mathcal{X} : \widehat{f}_{S_1}(x) \neq f^\star(x)\}$ and define the event $E_1 = E_1(S_1) := \{P(R \cap B_1) \neq 0\}$. By the law of total probability, we have

$$\underset{(S_1,S_2)\sim P^{2n}}{\mathbf{Pr}}[E] = \underset{(S_1,S_2)\sim P^{2n}}{\mathbf{Pr}}[E_1 \cap E] + \underset{(S_1,S_2)\sim P^{2n}}{\mathbf{Pr}}\left[\bar{E}_1 \cap E\right]. \tag{13}$$

Furthermore, we can write the second term on the right hand side of Eq. (13) as

$$\underset{(S_1,S_2)\sim P^{2n}}{\mathbf{Pr}}\left[\bar{E}_1 \cap E\right] = \underset{(S_1,S_2)\sim P^{2n}}{\mathbf{Pr}}\left[E \mid \bar{E}_1\right] \underset{(S_1,S_2)\sim P^{2n}}{\mathbf{Pr}}\left[\bar{E}_1\right] = \underset{(S_1,S_2)\sim P^{2n}}{\mathbf{Pr}}\left[\bar{E}_1\right].$$

Combining the identities above, it suffices to show that

$$\underset{(S_1,S_2)\sim P^{2n}}{\mathbf{Pr}}[E \cap E_1] \geq \underset{(S_1,S_2)\sim P^{2n}}{\mathbf{Pr}}[E_1] - \delta.$$

Notice that when E_1 occurs, then for any measurable set $C \subseteq \mathcal{X}$, the distribution $(P_R)_{B_1}$ (which is the conditional distribution of P_R restricted to B_1) satisfies

$$(P_R)_{B_1}(C) = \frac{P_R\left(C \cap B_1\right)}{P_R\left(B_1\right)} = \frac{P\left(C \cap B_1 \cap R\right)}{P\left(B_1 \cap R\right)} = P_{R\cap B_1}\left(C\right),$$

i.e., $(P_R)_{B_1} = P_{R\cap B_1}$. Thus on E_1, the probability that both $\widehat{f}_{S_1}$ and $\widehat{f}_{S_2}$ simultaneously err on a new data point drawn from P_R can be written as

$$\Pr_{X\sim P_R}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right] = \mathrm{err}_{P_R}\left(\widehat{f}_{S_1}\right)\mathrm{err}_{P_{R\cap B_1}}\left(\widehat{f}_{S_2}\right). \tag{14}$$

We now bound the right side of Eq. (14). To do this, we define the following events over (S_1, S_2):

$$E_2 = E_2(S_1) := \left\{\mathrm{err}_{P_R}\left(\widehat{f}_{S_1}\right) \leq 8\max\left\{\frac{d\,\mathrm{Log}(eP\left(R\right)n/d)}{P\left(R\right)n}, \frac{\mathrm{Log}(8/\delta)}{P\left(R\right)n}\right\}\right\}$$

and for outcomes of S_1 in E_1

$$\begin{aligned}&E_3 = E_3((S_1, S_2))\\ &:= \left\{\mathrm{err}_{P_{R\cap B_1}}\left(\widehat{f}_{S_2}\right) \leq 8\max\left\{\frac{d\,\mathrm{Log}(eP\left(R\right)n\,\mathrm{err}_{P_R}\left(\widehat{f}_{S_1}\right)/d)}{P\left(R\right)n\,\mathrm{err}_{P_R}\left(\widehat{f}_{S_1}\right)}, \frac{\mathrm{Log}(8/\delta)}{P\left(R\right)n\,\mathrm{err}_{P_R}\left(\widehat{f}_{S_1}\right)}\right\}\right\}.\end{aligned}$$

We now show that the event $E_1 \cap E_2 \cap E_3$ happens with probability at least $P[E_1] - \delta$ and that it implies the event $E_1 \cap E$. Assume that $E_1 \cap E_2 \cap E_3$ occurs. If

$$8\max\left\{\frac{d\,\mathrm{Log}(eP\left(R\right)n/d)}{P\left(R\right)n}, \frac{\mathrm{Log}(8/\delta)}{P\left(R\right)n}\right\} = 8\frac{d\,\mathrm{Log}(eP\left(R\right)n/d)}{P\left(R\right)n},$$

we have

$$\mathrm{err}_{P_R}\left(\widehat{f}_{S_1}\right)\mathrm{err}_{P_{R\cap B_1}}\left(\widehat{f}_{S_2}\right) \leq 8\max\left\{\frac{d\,\mathrm{Log}(eP\left(R\right)n\,\mathrm{err}_{P_R}\left(\widehat{f}_{S_1}\right)/d)}{P\left(R\right)n}, \frac{\mathrm{Log}(8/\delta)}{P\left(R\right)n}\right\}$$

$$\leq 8 \max \left\{ \frac{d \operatorname{Log}(8e \operatorname{Log}(eP(R)n/d))}{P(R)n}, \frac{\operatorname{Log}(8/\delta)}{P(R)n} \right\}.$$

Otherwise, if

$$8 \max \left\{ \frac{d \operatorname{Log}(eP(R)n/d)}{P(R)n}, \frac{\operatorname{Log}(8/\delta)}{P(R)n} \right\} = 8 \frac{\operatorname{Log}(8/\delta)}{P(R)n},$$

we have

$$\operatorname{err}_{P_R}\left(\widehat{f}_{S_1}\right) \operatorname{err}_{P_{R \cap B_1}}\left(\widehat{f}_{S_2}\right) \leq 8 \frac{\operatorname{Log}(8/\delta)}{P(R)n} \cdot 1.$$

We can thus conclude that $E_1 \cap E_2 \cap E_3$ implies $E_1 \cap E$. Towards showing the bound lower bound of $\mathbf{Pr}_{(S_1,S_2)\sim P^{2n}}[E_1 \cap E_2 \cap E_3] \geq P[E_1] - \delta$, notice that the bound $\mathbf{Pr}_{(S_1,S_2)\sim P^{2n}}\left[\bar{E}_2\right] \leq \delta/2$ can be established from Lemma 11 directly. Furthermore, for *any fixed realization* of S_1 such that $P(R \cap B_1) \neq 0$, Lemma 11 implies that

$$\operatorname{err}_{P_{R \cap B_1}}\left(\widehat{f}_{S_2}\right) \leq 8 \max \left\{ \frac{d \operatorname{Log}\left(eP(R) \operatorname{err}_{P_R}\left(\widehat{f}_{S_1}\right) n/d\right)}{P(R) \operatorname{err}_{P_R}\left(\widehat{f}_{S_1}\right) n}, \frac{\operatorname{Log}(8/\delta)}{P(R) \operatorname{err}_{P_R}\left(\widehat{f}_{S_1}\right) n} \right\},$$

with probability at least $1 - \delta/2$ over the randomness of S_2. Using the independence of S_1 and S_2 we have

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{(S_1,S_2)\sim P^{2n}}[E_1 \cap E_2 \cap E_3] &= \mathop{\mathbb{E}}_{S_1\sim P^n}\left[\mathbf{1}_{E_1}\mathbf{1}_{E_2} \mathop{\mathbf{Pr}}_{S_2\sim P^n}[E_3]\right] \\
&\geq \mathop{\mathbb{E}}_{S_1\sim P^n}[\mathbf{1}_{E_1}\mathbf{1}_{E_2}](1-\delta/2) \\
&\geq (1 - \mathop{\mathbf{Pr}}_{S_1\sim P^n}[\bar{E}_1] - \mathop{\mathbf{Pr}}_{S_1\sim P^n}[\bar{E}_2])(1-\delta/2) \\
&\geq (\mathop{\mathbf{Pr}}_{S_1\sim P^n}[E_1] - \delta/2)(1-\delta/2) \\
&\geq \mathop{\mathbf{Pr}}_{S_1\sim P^n}[E_1] - \delta.
\end{aligned}$$

This completes the proof. ■

A.3. Proof of Proposition 9

We now prove Proposition 9 which we restate here for convenience.

Proposition 9 *In the setting of Lemma 7 we have the following:*

1. *There is a universal constant c' such that for any $i \in \mathbb{N}$*

$$P(R_i) \leq \frac{c' 2^i d \operatorname{Log}(2^i \operatorname{Log}(1/\delta)/\delta) \operatorname{Log}(1/\delta)}{\delta n}.$$

2. *With probability at least* $1-\delta/2$ *over the randomness of* (S_1,S_2) *we have, simultaneously for all* $i \in I = \{i \geq 2 : P(R_i) \neq 0\}$*, that*

$$\Pr_{X\sim P_{R_i}}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right] \leq \frac{5\cdot 2^{-1.1i}\delta}{\operatorname{Log}^2(1/\delta)}.$$

Proof We first prove Item 1. Towards a contradiction, assume that there is an $i \in \mathbb{N}$ such that

$$\frac{P\left(R_i\right)n}{d} \geq \frac{c' 2^i \operatorname{Log}(2^i \operatorname{Log}(1/\delta)/\delta)\operatorname{Log}(1/\delta)}{\delta}$$

for a constant c' that we will choose below. By changing the order of expectations we have

$$\mathop{\mathbb{E}}_{S_1\sim P^n}\left[\operatorname{err}_{P_{R_i}}\left(\widehat{f}_{S_1}\right)\right] = \mathop{\mathbb{E}}_{X\sim P_{R_i}}\left[\mathop{\mathbf{Pr}}_{S_1\sim P^n}\left[\widehat{f}_{S_1}(X)\neq f^\star(X)\right]\right] = \mathop{\mathbb{E}}_{X\sim P_{R_i}}\left[p_X\right].$$

Using the above together with the fact that $p_X \geq 2^{-i}\delta/\operatorname{Log}(1/\delta)$ for any $X \in R_i$, we can conclude that

$$\mathop{\mathbb{E}}_{S_1\sim P^n}\left[\operatorname{err}_{P_{R_i}}\left(\widehat{f}_{S_1}\right)\right] > \frac{\delta}{2^i\operatorname{Log}(1/\delta)}. \tag{15}$$

On the other hand, using Lemma 6 we conclude that there is a universal constant $\hat{c}$ such that

$$\mathop{\mathbb{E}}_{S_1\sim P^n}\left[\operatorname{err}_{P_{R_i}}\left(\widehat{f}_{S_1}\right)\right] \leq \hat{c}\frac{d\operatorname{Log}\left(P\left(R_i\right)n/d\right)}{P\left(R_i\right)n}.$$

Combining this inequality with the fact that $\operatorname{Log}(x)/x$ is a decreasing function for $x>0$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{S_1\sim P^n}\left[\operatorname{err}_{P_{R_i}}\left(\widehat{f}_{S_1}\right)\right] &\leq \hat{c}\frac{\delta\operatorname{Log}\left(\frac{c'2^i\operatorname{Log}(2^i\operatorname{Log}(1/\delta)/\delta)\operatorname{Log}(1/\delta)}{\delta}\right)}{c'2^i\operatorname{Log}(2^i\operatorname{Log}(1/\delta)/\delta)\operatorname{Log}(1/\delta)}\\
&= \hat{c}\frac{2^{-i}\delta}{\operatorname{Log}(1/\delta)}\cdot\frac{\operatorname{Log}\left(\frac{c'2^i\operatorname{Log}(2^i\operatorname{Log}(1/\delta)/\delta)\operatorname{Log}(1/\delta)}{\delta}\right)}{c'\operatorname{Log}\left(\frac{2^i\operatorname{Log}(1/\delta)}{\delta}\right)}\\
&\leq \hat{c}\frac{2^{-i}\delta}{\operatorname{Log}(1/\delta)}\cdot\frac{\operatorname{Log}\left(c'\right)+\operatorname{Log}\left(\operatorname{Log}\left(\frac{2^i\operatorname{Log}(1/\delta)}{\delta}\right)\right)+\operatorname{Log}\left(2^i\frac{\operatorname{Log}(1/\delta)}{\delta}\right)}{c'\operatorname{Log}\left(\frac{2^i\operatorname{Log}(1/\delta)}{\delta}\right)}.
\end{aligned}$$

However, for c' large enough, the above is less than $2^{-i}\delta/\operatorname{Log}(1/\delta)$ which contradicts the lower bound Eq. (15). Thus, we have shown that there is a constant c' such that

$$\frac{P\left(R_i\right)n}{d} \leq \frac{c'2^i\operatorname{Log}(2^i\operatorname{Log}(1/\delta)/\delta)\operatorname{Log}(1/\delta)}{\delta},$$

which proves Item 1.

We now prove Item 2. We will show that for each $i \in I$ with probability at least $1-2^{-0.9i+2}\delta/5$ we have

$$\Pr_{X\sim P_{R_i}}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right] \leq \frac{5\cdot 2^{-1.1i}\delta}{\operatorname{Log}^2(1/\delta)}. \tag{16}$$

Applying a union bound implies that the above holds simultaneously for every $i \in I$ with probability at least $1 - \sum_{i \geq 2} 2^{-0.9i+2}\delta/5 \geq 1 - \delta/2$. To see that Eq. (16) holds for each $i \in I$, notice that we can use the fact that S_1 and S_2 are i.i.d. samples together with the fact that $p_X \leq 2^{-i+1}\delta/\operatorname{Log}(1/\delta)$ for $X \in R_i$ to conclude that

$$\mathop{\mathbb{E}}_{X \sim P_{R_i}}\left[\Pr_{S_1,S_2 \sim P^n}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right]\right] = \mathop{\mathbb{E}}_{X \sim P_{R_i}} p_X^2 \leq \frac{2^{-2i+2}\delta^2}{\operatorname{Log}^2(1/\delta)}.$$

Combining this with Markov's inequality we have

$$\begin{aligned}&\Pr_{S_1,S_2 \sim P^n}\left[\Pr_{X \sim P_{R_i}}\left[\widehat{f}_{S_1}(X) \neq f^\star(X) \wedge \widehat{f}_{S_2}(X) \neq f^\star(X)\right] > \frac{5 \cdot 2^{-1.1i}\delta}{\operatorname{Log}^2(1/\delta)}\right] \\ &\leq \frac{2^{-2i+2}\delta^2}{\operatorname{Log}^2(1/\delta)} \frac{\operatorname{Log}^2(1/\delta)}{5 \cdot 2^{-1.1i}\delta} = 2^{-0.9i+2}\delta/5,\end{aligned}$$

which proves the claim. ■

A.4. Proof of Lemma 11

We now prove Lemma 11 which we restate here for convenience.

Lemma 11 *Fix a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Fix a distribution P over $\mathcal{X}$, target function $f^\star \in \mathcal{F}$, and $R \subseteq \mathcal{X}$ such that $P(R) \neq 0$. Fix an ERM algorithm $\widehat{f} : \mathcal{X} \times \mathcal{Z}^* \to \{0,1\}$. For any parameter $\delta \in (0, 1/2]$ it holds with probability at least $1 - \delta$ over the randomness of $S \sim P^n$ that*

$$\operatorname{err}_{P_R}\left(\widehat{f}_S\right) \leq 8 \max\left\{\frac{d \operatorname{Log}(eP(R)\,n/d)}{P(R)\,n}, \frac{\operatorname{Log}(4/\delta)}{P(R)\,n}\right\}.$$

Proof If $8\operatorname{Log}(4/\delta)/(P(R)n) \geq 1$ we are done as $\operatorname{err}_{P_R}(f) \leq 1$. Thus, for the remainder of the proof we will assume that $8\operatorname{Log}(4/\delta)/(P(R)n) < 1$, which is equivalent to $P(R)n \geq 8\operatorname{Log}(4/\delta)$. Define the event

$$E = E(S) = \{|\{i \in [n] : X_i \in R\}| \geq P(R)n/2\}.$$

Using the Chernoff bound and our assumption that $P(R)n \geq 8\operatorname{Log}(4/\delta)$, we have

$$\Pr_{S \sim P^n}[E] \geq 1 - \exp\left(-\frac{P(R)n}{8}\right) \geq 1 - \delta/2.$$

Notice that conditioned on E, the $M \geq P(R)n/2 \geq 1$ samples[7] that land in R form an i.i.d. sample from the conditional distribution P_R. Thus when E occurs, any ERM trained on S is also consistent with $M \geq P(R)n/2 \geq 1$ i.i.d. samples from P_R, so Lemma 10 yields, with probability at least $1 - \delta/2$ over the randomness of S, that

$$\operatorname{err}_{P_R}\left(\widehat{f}_S\right) \leq 4 \max\left\{\frac{d \operatorname{Log}(2eM/d)}{M}, \frac{\operatorname{Log}(4/\delta)}{M}\right\}$$

7. The number of samples M is random.

$$\leq 8 \max \left\{ \frac{d\,\mathrm{Log}(eP(R)n/d)}{P(R)n}, \frac{\mathrm{Log}(4/\delta)}{P(R)n} \right\}.$$

Here, the second inequality follows from the fact that $x \to \mathrm{Log}(2ex)/x$ is decreasing for $x > 0$. Using the law of total probability we get that

$$\begin{aligned}
&\mathop{\mathbf{Pr}}_{S\sim P^n}\left[\mathrm{err}_{P_R}\left(\widehat{f}_S\right) > 8\max\left\{\frac{d\,\mathrm{Log}(eP(R)n/d)}{P(R)n}, \frac{\mathrm{Log}(4/\delta)}{P(R)n}\right\}\right] \\
&\leq \mathop{\mathbf{Pr}}_{S\sim P^n}\left[\bar{E}\right] + \mathop{\mathbf{Pr}}_{S\sim P^n}\left[\mathrm{err}_{P_R}\left(\widehat{f}_S\right) > 8\max\left\{\frac{d\,\mathrm{Log}(eP(R)n/d)}{P(R)n}, \frac{\mathrm{Log}(4/\delta)}{P(R)n}\right\} \,\middle|\, E\right] \\
&\leq \delta/2 + \delta/2 = \delta.
\end{aligned}$$

This concludes the proof. ■

Appendix B. Ommited proofs from Section 3

In this appendix we prove Eq. (8). We will show that

$$\mathop{\mathbf{Pr}}_{Y\sim Q}[Y \geq m] = \mathop{\mathbf{Pr}}_{Y_t\sim Q_t}\left[\sum_{t=0}^{m_2 d - d - 1} Y_t \geq m\right] \geq \sqrt{\frac{2}{3}}.$$

Let $p^\star = p_{m_2 d - d - 1} = \frac{d+1}{m_1 m_2}$ be the smallest parameter p_t of the geometric random variables $\{Y_t\}_{t=0}^{m_2 d - d - 1}$ that we consider. We make use of the following well known concentration inequality for sums of geometric random variables:

$$\mathop{\mathbf{Pr}}_{Y\sim Q}\left[Y \leq \lambda \mathop{\mathbb{E}}_{Y\sim Q}[Y]\right] \leq \exp\left(-p^\star \mathop{\mathbb{E}}_{Y\sim Q}[Y](\lambda - 1 - \ln(\lambda))\right), \tag{17}$$

which holds for any $0 < \lambda \leq 1$ (see (Janson, 2018, Theorem 3.1)). Let $\lambda = 1/4$. We will show $\mathbb{E}_{Y\sim Q}[Y] \geq 4m$ and $p^\star\, \mathbb{E}_{Y\sim Q}[Y] \geq 4$ which combined with $1/4 - 1 - \ln(1/4) \geq 1/2$ and Eq. (17) gives us

$$\mathop{\mathbf{Pr}}_{Y\sim Q}[Y \leq m] \leq \mathop{\mathbf{Pr}}_{Y\sim Q}\left[Y \leq \mathop{\mathbb{E}}_{Y\sim Q}[Y]/4\right] \leq \exp\left(-p^\star \mathop{\mathbb{E}}_{Y\sim Q}[Y]/2\right) \leq \exp(-2) \leq 1 - \sqrt{2/3}.$$

This implies $\mathbf{Pr}_{Y\sim Q}[Y \geq m] \geq \sqrt{2/3}$ as required. We first show that $\mathbb{E}_{Y\sim Q}[Y] \geq 4m$. We have

$$\mathop{\mathbb{E}}_{Y\sim Q}[Y] = \sum_{t=0}^{m_2 d - d - 1} \frac{m_1 m_2}{m_2 d - t} = \sum_{i=d+1}^{m_2 d} \frac{m_1 m_2}{i} \geq m_1 m_2 \ln\left(\frac{m_2}{2}\right). \tag{18}$$

Plugging in the definition of m_1 and m_2 into Eq. (18) and using the fact that $\lceil x \rceil \leq 2x$ for $x \geq 0.5$ gives us

$$\mathop{\mathbb{E}}_{Y\sim Q}[Y] \geq \frac{8Cm}{\ln(\ln(2Cm/d))} \ln\left(\frac{\ln(2Cm/d)}{\ln(\ln(2Cm/d))}\right).$$

For C large enough we have $\ln(\ln(2Cm/d)) > 0$, $\frac{\ln(2Cm/d)}{\ln(\ln(2Cm/d))} \geq \sqrt{\ln(2Cm/d)}$ and $C > 1$, so

$$\mathop{\mathbb{E}}_{Y \sim Q}[Y] \geq \frac{8Cm}{\ln\left(\ln\left(2Cm/d\right)\right)} \ln\left(\sqrt{\ln(2Cm/d)}\right) \geq 4m.$$

We now show that $p^\star \mathbb{E}_{Y \sim Q}[Y] \geq 4$. Using the fact that $p^\star \geq 2/(m_1 m_2)$ together with Eq. (18) gives us

$$p^\star \mathop{\mathbb{E}}_{Y \sim Q}[Y] \geq 2\ln\left(\frac{m_2}{2}\right) \geq 2\ln\left(\frac{\ln(2Cm/d)}{\ln\left(\ln\left(2Cm/d\right)\right)}\right) \geq 4,$$

where the last inequality holds for C large enough.