The Optimal Distribution-Free Sample Complexity of Distribution-Dependent Learning

Steve Hanneke STEVE.HANNEKE@GMAIL.COM

Abstract

This work establishes a new upper bound on the worst-case number of labeled samples sufficient for PAC learning in the realizable case, if the learning algorithm is allowed dependence on the data distribution, or an additional pool of unlabeled samples. The bound matches known lower bounds up to constant factors. This resolves a long-standing open problem on the sample complexity of distribution-dependent PAC learning. We also discuss potential avenues for extension to distribution-independent learning. The technique and analysis build on a recent breakthrough by Hans Simon.

1. Introduction

Probably approximately correct learning (or *PAC* learning; Valiant, 1984) is a classic criterion for supervised learning, which has been the focus of much research in the past three decades. The objective in PAC learning is to produce a classifier that, with probability at least $1-\delta$, has error rate is at most ε. To qualify as a PAC learning algorithm, it must satisfy this guarantee for all possible target concepts in a given family, under all possible data distributions. To achieve this objective, the learning algorithm is supplied with a number m of i.i.d. training examples (data points), along with the corresponding correct classifications. One of the central questions in the study of PAC learning is determining the minimum number $\mathcal{M}(\varepsilon,\delta)$ of training examples necessary and sufficient such that there exists a PAC learning algorithm requiring at most $\mathcal{M}(\varepsilon,\delta)$ samples (for any given ε and δ). This quantity $\mathcal{M}(\varepsilon,\delta)$ is known as the *sample complexity*.

Determining the sample complexity of PAC learning is a long-standing open problem. There have been upper and lower bounds established for decades, but they differ by a logarithmic factor. It is widely believed that this logarithmic factor gap can be removed for certain well-designed learning algorithms, and attempting to prove this has been the subject of much effort. Simon (2015) has very recently made an enormous leap forward toward resolving this issue. Specifically, he proves that a simple ensemble-based learning algorithm is able to reduce this factor to a very slowly-growing function. However, that work does not quite completely resolve the gap, so that determining the optimal sample complexity remains an open problem.

The present work makes progress on this problem by completely eliminating the logarithmic factor, but under a variant of the setting in which the learning algorithm may have *direct dependence* on the data distribution, or alternatively, has access to a somewhat larger pool of *unlabeled* data (i.e., semi-supervised learning). In particular, this resolves the gap between the best previously-known upper and lower bounds for that setting, thus establishing a precise expression for the optimal sample complexity of distribution-dependent PAC learning (up to numerical constant factors).

While establishing sharp worst-case sample complexity bounds for distribution-dependent (or semi-supervised) PAC learning is certainly interesting in its own right, this work may also be useful toward establishing optimal sample complexity bounds for distribution-independent learning as well. For instance, the techniques employed here may potentially have distribution-independent (and strictly supervised) variants for which the stated sample complexity guarantees remain valid. Additionally, we note that a weakened version of a conjecture of Ben-David, Lu, and Pál (2008) (which we discuss in Section 6) states that allowing dependence on the data distribution can improve the sample complexity by at most a numerical constant factor. If this conjecture is correct, the present work would establish optimal sample complexity bounds for (distribution-independent) PAC learning as well (up to numerical constant factors). Indeed, it follows from the present work that this (weakened) conjecture is *equivalent* to the classic conjecture that the lower bounds of Blumer, Ehrenfeucht, Haussler, and Warmuth (1989) and Ehrenfeucht, Haussler, Kearns, and Valiant (1989) on the sample complexity of distribution-independent PAC learning are generally sharp up to numerical constant factors (see e.g., Warmuth, 2004, for discussion of this conjecture).

2. Notation

We begin by introducing some basic notation essential to the discussion. Fix a nonempty set $\mathcal{X}$, called the *instance space*; we suppose $\mathcal{X}$ is equipped with a σ-algebra, defining the measurable subsets of $\mathcal{X}$. Also denote by $\mathcal{Y} = \{-1, +1\}$, called the *label space*. A *classifier* is any measurable function $h : \mathcal{X} \to \mathcal{Y}$. Fix a nonempty set $\mathbb{C}$ of classifiers, called the *concept space*. To focus the discussion on nontrivial cases,[1] we suppose $|\mathbb{C}| \geq 3$; other than this, the results in this article will be valid for *any* choice of $\mathbb{C}$.

In the learning problem, there is a probability measure $\mathcal{P}$ over $\mathcal{X}$, called the *data distribution*, and a sequence $X_1(\mathcal{P}), X_2(\mathcal{P}), \ldots$ of independent $\mathcal{P}$-distributed random variables, called the *unlabeled examples* (or unlabeled data); for $m \in \mathbb{N}$, also define $\mathbb{X}_{1:m}(\mathcal{P}) = \{X_1(\mathcal{P}), \ldots, X_m(\mathcal{P})\}$, and for completeness denote $\mathbb{X}_{1:0}(\mathcal{P}) = \{\}$. There is also a special element of $\mathbb{C}$, denoted $f^\star$, called the *target function*. For any sequence $S = \{x_1, \ldots, x_k\}$ in $\mathcal{X}$, denote by $(S, f^\star(S)) = \{(x_1, f^\star(x_1)), \ldots, (x_k, f^\star(x_k))\}$. For any probability measure P over $\mathcal{X}$, and any classifier h, denote by $\mathrm{er}_P(h; f^\star) = P(x : h(x) \neq f^\star(x))$. A *learning algorithm* $\mathbb{A}$ is a map,[2] mapping any sequence $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ in $\mathcal{X} \times \mathcal{Y}$ (of arbitrary length $m \in \mathbb{N} \cup \{0\}$) to a classifier $h : \mathcal{X} \to \mathcal{Y}$ (not necessarily in $\mathbb{C}$).

Definition 1 *For any $\varepsilon, \delta \in (0, 1)$, the* distribution-free sample complexity of distribution-dependent (ε, δ)-PAC learning*, denoted $\tilde{\mathcal{M}}(\varepsilon, \delta)$, is defined as the smallest $m \in \mathbb{N} \cup \{0\}$ such that, for every possible data distribution $\mathcal{P}$, there exists a learning algorithm $\mathbb{A}_{\mathcal{P}}$ such that, $\forall f^\star \in \mathbb{C}$, denoting $\hat{h} = \mathbb{A}_{\mathcal{P}}((\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P}))))$,*

$$\mathbb{P}\left(\mathrm{er}_{\mathcal{P}}\left(\hat{h}; f^\star\right) \leq \varepsilon\right) \geq 1 - \delta. \tag{1}$$

1. The sample complexities for $|\mathbb{C}| = 1$ and $|\mathbb{C}| = 2$ are already quite well understood in the literature, the former having sample complexity 0, and the latter having sample complexity either 1 or $\Theta(\frac{1}{\varepsilon} \ln \frac{1}{\delta})$ (depending on whether the two classifiers are exact complements or not).
2. We also admit randomized algorithms, where $\forall S \in (\mathcal{X} \times \mathcal{Y})^m$, the (random) value of $\mathbb{A}(S)$ is independent of all $X_i(P)$.

If no such m exists, we define $\tilde{\mathcal{M}}(\varepsilon, \delta) = \infty$.

Definition 1 is our primary object of study in this work. Note that the order of quantifiers in this definition allows that, for each possible choice of $\mathcal{P}$, we can find a different algorithm $\mathbb{A}_{\mathcal{P}}$ to satisfy the "probably approximately correct" guarantee (1) (though it must still satisfy this guarantee for all choices of $f^\star \in \mathbb{C}$). This may be interpreted as allowing distribution-dependence in the learning algorithm: that is, we may intuitively think of the collection $\{\mathbb{A}_{\mathcal{P}} : \mathcal{P} \text{ is a prob. measure}\}$ together as a *distribution-dependent* learning algorithm. For any such distribution-dependent learning algorithm $\mathbb{A}_{\mathcal{P}}$, we say $\mathbb{A}_{\mathcal{P}}$ *achieves* a sample complexity $\tilde{M}(\varepsilon, \delta)$ if, for every $\mathcal{P}$, every $f^\star \in \mathbb{C}$, and every $m \geq \tilde{M}(\varepsilon, \delta)$, the classifier $\hat{h} = \mathbb{A}_{\mathcal{P}}((\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P}))))$ satisfies (1). Thus, $\tilde{\mathcal{M}}(\varepsilon, \delta)$ can equivalently be defined as the smallest (integer-valued) sample complexity $\tilde{M}(\varepsilon, \delta)$ achievable by distribution-dependent learning algorithms.

We require a few additional definitions before proceeding. For any sequence $S = \{(x_1, y_1), \ldots, (x_k, y_k)\}$ in $\mathcal{X} \times \mathcal{Y}$, denote by $\mathbb{C}[S] = \{h \in \mathbb{C} : \forall (x, y) \in S, h(x) = y\}$, referred to as the set of classifiers *consistent* with S. Following Vapnik and Chervonenkis (1971), we say a sequence $\{x_1, \ldots, x_k\}$ in $\mathcal{X}$ is *shattered* by $\mathbb{C}$ if $\forall y_1, \ldots, y_k \in \mathcal{Y}$, $\exists h \in \mathbb{C}$ such that $\forall i \in \{1, \ldots, k\}$, $h(x_i) = y_i$: that is, there are 2^k distinct classifications of $\{x_1, \ldots, x_k\}$ realized by classifiers in $\mathbb{C}$. The Vapnik-Chervonenkis dimension (or *VC dimension*) of $\mathbb{C}$ is then defined as the largest integer k for which there exists a sequence $\{x_1, \ldots, x_k\}$ in $\mathcal{X}$ shattered by $\mathbb{C}$; if no such largest k exists, the VC dimension is said to be infinite. We denote by d the VC dimension of $\mathbb{C}$. This quantity is of fundamental importance in characterizing the sample complexity of PAC learning. In particular, it is well known that the distribution-free sample complexity (either with or without distribution-dependence in the learning algorithm) is finite for any $\varepsilon, \delta \in (0, 1)$ if and only if $d < \infty$ (Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989; Benedek and Itai, 1991; Kulkarni, 1989; Devroye, Györfi, and Lugosi, 1996); for simplicity of notation, for the remainder of this article we suppose $d < \infty$; furthermore, note that our assumption of $|\mathbb{C}| \geq 3$ implies $d \geq 1$.

We adopt a common variation on big-O asymptotic notation, used in much of the learning theory literature. Specifically, for functions $f, g : (0, 1)^2 \to [0, \infty)$, we let $f(\varepsilon, \delta) = O(g(\varepsilon, \delta))$ denote the assertion that $\exists \varepsilon_0, \delta_0 \in (0, 1)$ and $c_0 \in (0, \infty)$ such that, $\forall \varepsilon \in (0, \varepsilon_0)$, $\forall \delta \in (0, \delta_0)$, $f(\varepsilon, \delta) \leq c_0 g(\varepsilon, \delta)$; however, we also require that the values $\varepsilon_0, \delta_0, c_0$ in this definition be *numerical constants*, meaning that they are *independent of* $\mathbb{C}$ *and* $\mathcal{X}$. For instance, this means c_0 cannot depend on d. We equivalently write $f(\varepsilon, \delta) = \Omega(g(\varepsilon, \delta))$ to assert that $g(\varepsilon, \delta) = O(f(\varepsilon, \delta))$. Finally, we write $f(\varepsilon, \delta) = \Theta(g(\varepsilon, \delta))$ to assert that both $f(\varepsilon, \delta) = O(g(\varepsilon, \delta))$ and $f(\varepsilon, \delta) = \Omega(g(\varepsilon, \delta))$ hold. We also sometimes write $O(g(\varepsilon, \delta))$ in an expression, as a place-holder for some function $f(\varepsilon, \delta)$ satisfying $f(\varepsilon, \delta) = O(g(\varepsilon, \delta))$: for instance, the statement $N(\varepsilon, \delta) \leq d + O(\mathrm{Log}(1/\delta))$ expresses that $\exists f(\varepsilon, \delta) = O(\mathrm{Log}(1/\delta))$ for which $N(\varepsilon, \delta) \leq d + f(\varepsilon, \delta)$. Also, for any value $z \geq 0$, define $\mathrm{Log}(z) = \ln(\max\{z, e\})$.

As is commonly required in the learning theory literature, we adopt the assumption that any quantity appearing in a probability or expectation expression is indeed measurable. For our purposes, this comes into effect only in the application of classic generalization bounds for sample-consistent classifiers (Lemma 4 below). See Blumer, Ehrenfeucht, Haussler, and Warmuth (1989) and van der Vaart and Wellner (2011) for discussion of conditions on $\mathbb{C}$ sufficient for this measurability assumption to hold.

3. Background

Our objective in this work is to establish *sharp* sample complexity bounds. As such, we should first review the known *lower bounds* on $\tilde{\mathcal{M}}(\varepsilon,\delta)$. A basic lower bound of $\frac{1-\varepsilon}{\varepsilon}\ln\left(\frac{1}{\delta}\right)$ was established by Blumer, Ehrenfeucht, Haussler, and Warmuth (1989) for $0<\varepsilon<1/2$ and $0<\delta<1$. Although this lower bound was stated in that work as a lower bound on the sample complexity of distribution-independent learning algorithms, its proof fixes a particular distribution $\mathcal{P}$ and reasons only about worst-case performance over the choice of $f^\star\in\mathbb{C}$. The lower bound therefore also applies to $\tilde{\mathcal{M}}(\varepsilon,\delta)$ as well. A second lower bound of $\frac{d-1}{32\varepsilon}$ was supplied by Ehrenfeucht, Haussler, Kearns, and Valiant (1989), for $0<\varepsilon\leq 1/8$ and $0<\delta\leq 1/100$. Again, although that work studied distribution-independent learning algorithms, the proof of this lower bound fixes a particular distribution $\mathcal{P}$ and discusses only the worst-case choice of $f^\star\in\mathbb{C}$ while holding $\mathcal{P}$ fixed. Thus, this too supplies a lower bound on $\tilde{\mathcal{M}}(\varepsilon,\delta)$. Taken together, these results imply that, for any $\varepsilon\in(0,1/8]$ and $\delta\in(0,1/100]$,

$$\tilde{\mathcal{M}}(\varepsilon,\delta)\geq\max\left\{\frac{d-1}{32\varepsilon},\frac{1-\varepsilon}{\varepsilon}\ln\left(\frac{1}{\delta}\right)\right\}=\Omega\left(\frac{1}{\varepsilon}\left(d+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\right). \tag{2}$$

This lower bound is also complemented by classic *upper bounds* on the sample complexity. In particular, Vapnik (1982) and Blumer, Ehrenfeucht, Haussler, and Warmuth (1989) established an upper bound of

$$\tilde{\mathcal{M}}(\varepsilon,\delta)=O\left(\frac{1}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right)+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\right). \tag{3}$$

They prove that this bound in fact holds for a *distribution-independent* learning algorithm: namely, any algorithm that returns a classifier $h\in\mathbb{C}[(\mathbb{X}_{1:m}(\mathcal{P}),f^\star(\mathbb{X}_{1:m}(\mathcal{P})))]$, also known as a *sample-consistent learning algorithm* (or *empirical risk minimization* algorithm). A sometimes-better upper bound was established by Haussler, Littlestone, and Warmuth (1994):

$$\tilde{\mathcal{M}}(\varepsilon,\delta)=O\left(\frac{d}{\varepsilon}\mathrm{Log}\left(\frac{1}{\delta}\right)\right). \tag{4}$$

This bound is achieved by a modified variant of the *one-inclusion graph prediction algorithm*, a learning algorithm also proposed by Haussler, Littlestone, and Warmuth (1994).

In very recent work, Simon (2015) produced a breakthrough insight. Specifically, by analyzing a classifier based on a simple majority vote among classifiers consistent with distinct subsets of the training sample $(\mathbb{X}_{1:m}(\mathcal{P}),f^\star(\mathbb{X}_{1:m}(\mathcal{P})))$, Simon (2015) established that, for any choice of $K\in\mathbb{N}$,

$$\tilde{\mathcal{M}}(\varepsilon,\delta)=O\left(\frac{2^{2K}\sqrt{K}}{\varepsilon}\left(d\log^{(K)}\left(\frac{1}{\varepsilon}\right)+K+\log\left(\frac{1}{\delta}\right)\right)\right),$$

where $\log^{(K)}(x)$ is the K-times iterated logarithm: $\log^{(0)}(x)=\max\{x,1\}$ and $\log^{(K)}(x)=\max\{\log_2(\log^{(K-1)}(x)),1\}$. In particular, a natural choice would be $K\approx\log^*\left(\frac{1}{\varepsilon}\right)$,[3] which

3. The function $\log^*(x)$ is the iterated logarithm: the smallest $K\in\mathbb{N}\cup\{0\}$ for which $\log^{(K)}(x)\leq 1$. It is an extremely slowly growing function of x.

(one can show) optimizes the asymptotic dependence on ε in the above bound,[4] yielding

$$\tilde{\mathcal{M}}(\varepsilon,\delta) = O\left(\frac{1}{\varepsilon}2^{O(\log^*(1/\varepsilon))}\left(d+\mathrm{Log}\left(\frac{1}{\delta}\right)\right)\right). \tag{5}$$

Note that there is a range of ε, δ, and d values for which this bound is strictly better than both (3) and (4) (i.e., where $\mathrm{Log}(1/\delta) \ll d\mathrm{Log}(1/\varepsilon)/(2^{2\log^*(1/\varepsilon)}\sqrt{\log^*(1/\varepsilon)})$ and $2^{2\log^*(1/\varepsilon)}\sqrt{\log^*(1/\varepsilon)} \ll \min\{\mathrm{Log}(1/\delta), d\}$). However, it still does not quite match the lower bound (2).

There have also been many special-case analyses, studying restricted types of concept spaces $\mathbb{C}$ for which the above gaps can be closed (e.g., Auer and Ortner, 2007; Darnstädt, 2015; Hanneke, 2015). However, these special conditions do not include many of the most-commonly studied concept spaces, such as linear separators and multilayer neural networks. There have also been a variety of studies that, in addition to restricting to specific concept spaces $\mathbb{C}$, also introduce strong restrictions on the data distribution $\mathcal{P}$, and establish an upper bound of the same form as the lower bound (2) under these restrictions (e.g., Long, 2003; Giné and Koltchinskii, 2006; Bshouty, Li, and Long, 2009; Hanneke, 2009, 2015; Balcan and Long, 2013). However, there are many interesting classes $\mathbb{C}$ and distributions $\mathcal{P}$ for which these results do not imply any improvements over (3). Thus, in the present literature, there persists a gap between the lower bound (2) and the minimum of all of the known upper bounds (3), (4), and (5) applicable to the *general case* of an arbitrary concept space of a given VC dimension d (under arbitrary data distributions).

We note that all of the above upper bounds, (3), (4), and (5), were established for *distribution-independent* learning algorithms, and are therefore also bounds on the sample complexity of distribution-independent learning. However, the best known general upper bounds on the distribution-free sample complexity of *distribution-dependent* learning offer improvements over the above (in the general case of arbitrary $\mathbb{C}$ of VC dimension d) *only in constant factors*: for instance, a construction of Bshouty, Li, and Long (2009) reveals that for any given d, there exist distributions $\mathcal{P}$ over $\mathbb{R}^d$ for which, with $\mathbb{C}$ the class of homogeneous linear separators (which has VC dimension d; Cover, 1965), the distribution-dependent sample complexity bounds established by Benedek and Itai (1991) and Bshouty, Li, and Long (2009) for distribution-dependent learning algorithms are no smaller than the upper bound (3), aside from numerical constant factors.

In the present work, we establish a new upper bound for a distribution-dependent learning algorithm, which holds for *any* concept space $\mathbb{C}$, and which improves over all of the above upper bounds in its joint dependence on ε,δ, and d. In particular, it is *optimal*, in the sense that it matches the lower bound (2) up to numerical constant factors. This work thus resolves a long-standing open problem, by establishing the precise form of the optimal sample complexity, up to numerical constant factors.

4. In general, the entire form of the bound is optimized (up to numerical constant factors) by choosing $K = \max\left\{\log^*\left(\frac{1}{\varepsilon}\right) - \log^*\left(\frac{1}{d}\log\left(\frac{1}{\delta}\right)\right) + 1, 1\right\}$.

4. Main Result

For simplicity of the analysis, we begin by supposing the algorithm has direct access to the data distribution $\mathcal{P}$. In Section 5 below, we discuss a simple technique for relaxing this to suppose only that it has access to a pool of $\mathcal{P}$-distributed unlabeled data.

4.1 Sketch of the Approach

Before getting into the formal details, we first outline the high-level motivation for this algorithm.

The general approach used here builds on an argument of Simon (2015), which itself has roots in the analysis of sample-consistent learning algorithms by Hanneke (2009). The essential idea is that, if two classifiers are each consistent with respective distinct labeled training samples, we can analyze the probability that they *both* make a mistake on a random point by bounding the error rate of the first classifier under the distribution $\mathcal{P}$, and bounding the error rate of the second classifier under the *conditional* distribution given that the first classifier makes a mistake. Then, if the first classifier's error rate is larger than our desired bound, we can argue that at least a certain number of points in the second training sample are contained in that classifier's error region, and we can then bound the conditional error rate of the second classifier in terms of the number of such points. Multiplying these two bounds then yields a bound on the probability they both make a mistake. We then apply this reasoning to three classifiers, bounding the probability that any two of them make a mistake; combined with the union bound, the sum of these three resulting bounds provides an upper bound on the error rate of the majority vote of the three classifiers.

The original analysis of Simon (2015) applied this reasoning, together with classic VC bounds, with only the constraint that the three classifiers be consistent with their respective samples, and otherwise allowed them to be chosen arbitrarily. In the present work, we instead suppose that the first classifier in each pair in the above argument is produced by a recursive call to our proposed algorithm. Thus, we are able to use our refined bound, taken as an inductive hypothesis, for the bound on the error rate of the first classifier in each pair, which thereby refines the final bound as desired.

However, in order to obtain a bound on the conditional error rate of the second classifier given the error region of the first, based on classic VC bounds, we need to ensure that the second classifier is indeed a sample-consistent element of $\mathbb{C}$. We thus have a constraint: among all pairings of these three classifiers, at least one of the two classifiers should be a recursive call to the proposed algorithm, and at least one of the two classifiers should be a sample-consistent element of $\mathbb{C}$. One can easily see that, to satisfy this requirement, at least one of the three classifiers must be *both* a sample-consistent element of $\mathbb{C}$ *and* the result of a recursive call to the proposed algorithm. However, majority vote classifiers are not themselves guaranteed to be representable as elements of $\mathbb{C}$. To resolve this, we *project* the majority vote classifier onto the set of sample-consistent elements of $\mathbb{C}$, which at most doubles the error rate. We emphasize that this projection step is the *only* use of distribution-dependence in the entire algorithm.

4.2 Formal Details

For any three values $y_1, y_2, y_3 \in \mathcal{Y}$, define the majority function: $\text{Majority}(y_1, y_2, y_3) = \text{sign}(y_1 + y_2 + y_3) = 2\mathbb{1}[y_1 + y_2 + y_3 \geq 0] - 1$. For any set A, function $f : A \to [0, \infty)$, and value $\gamma \geq 0$, define the set of γ-near minimizers of f:

$$\underset{a \in A}{\text{argmin}^{\gamma}} f(a) = \left\{ a \in A : f(a) \leq \inf_{a' \in A} f(a') + \gamma \right\}.$$

Additionally, for any $m \in \mathbb{N} \cup \{0\}$ and $\delta \in (0, 1)$, let $\gamma(m, \delta)$ be any value in $(0, \infty)$ satisfying $\gamma(m, \delta) \leq \frac{1}{3(m+1)} \left(d + \ln \left(\frac{12}{\delta} \right) \right)$. Finally, let c_0 be any real value satisfying $e^7 \leq c_0 \leq e^{176}$. Now consider the following recursive algorithm.[5]

Algorithm: $\mathbb{A}_{\mathcal{P}}^{(\delta)}(S)$
0. If $|S| \leq c_0(d + \ln(12/\delta))$
1. Return any $\hat{h} \in \mathbb{C}[S]$
2. Let S_1 denote the first $\lfloor |S|/2 \rfloor$ elements of S, S_2 the remaining $\lceil |S|/2 \rceil$ elements
3. Let $\hat{h}_1 = \mathbb{A}_{\mathcal{P}}^{(\delta/6)}(S_1)$, $\hat{h}_2 = \mathbb{A}_{\mathcal{P}}^{(\delta/6)}(S_2)$, and $\hat{h}_3$ is an arbitrary element of $\mathbb{C}[S]$
4. Return any $\hat{h} \in \underset{h \in \mathbb{C}[S]}{\text{argmin}^{\gamma(|S|, \delta)}} \mathcal{P}(x : h(x) \neq \text{Majority}(\hat{h}_1(x), \hat{h}_2(x), \hat{h}_3(x)))$

Theorem 2 *For every $\varepsilon, \delta \in (0, 1)$,*

$$\tilde{\mathcal{M}}(\varepsilon, \delta) = O\left(\frac{1}{\varepsilon} \left(d + \text{Log}\left(\frac{1}{\delta} \right) \right) \right).$$

In particular, the above distribution-dependent algorithm $\mathbb{A}_{\mathcal{P}}^{(\delta)}$ achieves a sample complexity of the form expressed on the right hand side.

Combined with (2), this immediately implies the following corollary.

Corollary 3

$$\tilde{\mathcal{M}}(\varepsilon, \delta) = \Theta\left(\frac{1}{\varepsilon} \left(d + \text{Log}\left(\frac{1}{\delta} \right) \right) \right).$$

4.3 Proof of Theorem 2

The following classic result will be needed in the proof. It is implied by a result of Vapnik (1982), and was obtained via a direct proof by Blumer, Ehrenfeucht, Haussler, and Warmuth (1989). The version stated here features slightly smaller constant factors, due to Anthony and Bartlett (1999).

5. We will suppose that the steps left partially ambiguous in this definition (i.e., the choices of $\hat{h}$ in Step 1, $\hat{h}_3$ in Step 3, and $\hat{h}$ in Step 4) are performed in a way that depends only on the input S (and δ, $\mathcal{P}$), and possibly some "internal" randomness (independent from all other $X_i(P)$); thus, $\mathbb{A}_{\mathcal{P}}^{(\delta)}$ satisfies the requirements of a learning algorithm discussed in Section 2. In particular, when $S = (\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P})))$, in Step 3, the classifier $\hat{h}_1$ is independent from the sample S_2, and $\hat{h}_2$ is independent from S_1.

Lemma 4 *For any $\delta \in (0,1)$, $m \in \mathbb{N}$, $f^\star \in \mathbb{C}$, and any probability measure P over $\mathcal{X}$, letting $Z_1, \ldots, Z_m$ be independent P-distributed random variables, with probability at least $1-\delta$, every $h \in \mathbb{C}[\{(Z_i, f^\star(Z_i))\}_{i=1}^m]$ satisfies*

$$\mathrm{er}_P(h; f^\star) \leq \frac{2}{m}\left(d\mathrm{Log}\left(\frac{2em}{d}\right) + \mathrm{Log}\left(\frac{2}{\delta}\right)\right).$$

We are now ready for the proof of Theorem 2.

Proof of Theorem 2 Fix any $f^\star \in \mathbb{C}$ and probability measure $\mathcal{P}$ over $\mathcal{X}$. We will prove by induction that, for any $m' \in \mathbb{N}$, and any $\delta' \in (0,1)$, with probability at least $1-\delta'$, the classifier $\hat{h}_{m',\delta'} = \mathbb{A}_{\mathcal{P}}^{(\delta')}((\mathbb{X}_{1:m'}(\mathcal{P}), f^\star(\mathbb{X}_{1:m'}(\mathcal{P}))))$ satisfies

$$\mathrm{er}_{\mathcal{P}}(\hat{h}_{m',\delta'}; f^\star) \leq \frac{1}{m'+1}\left(358d + 97\ln\left(\frac{12}{\delta'}\right)\right). \tag{6}$$

First, as a base case, note that for any $m' \in \mathbb{N}$ and $\delta' \in (0,1)$ with $m' \leq c_0(d + \ln(12/\delta'))$, $\mathbb{A}_{\mathcal{P}}^{(\delta')}((\mathbb{X}_{1:m'}(\mathcal{P}), f^\star(\mathbb{X}_{1:m'}(\mathcal{P}))))$ terminates in Step 1, in which case Lemma 4 implies that with probability at least $1-\delta'$, the returned classifier $\hat{h}_{m',\delta'}$ satisfies

$$\begin{aligned}
\mathrm{er}_{\mathcal{P}}(\hat{h}_{m',\delta'}; f^\star) &\leq \frac{2}{m'}\left(d\mathrm{Log}\left(\frac{2em'}{d}\right) + \mathrm{Log}\left(\frac{2}{\delta'}\right)\right) \\
&\leq \frac{2}{m'}\left(d\mathrm{Log}\left(\frac{2ec_0\left(d+\ln(12/\delta')\right)}{d}\right) + \mathrm{Log}\left(\frac{2}{\delta'}\right)\right) \\
&= \frac{2}{m'}\left(d\mathrm{Log}\left(e^{-3}c_0\left(2e^4 + \frac{2e^4}{d}\ln\left(\frac{12}{\delta'}\right)\right)\right) + \mathrm{Log}\left(\frac{2}{\delta'}\right)\right) \\
&\leq \frac{2}{m'}\left(d\ln(e^{-3}c_0(2e^4+e)) + \left(1+2e^3\right)\ln\left(\frac{12}{\delta'}\right)\right) \\
&\leq \frac{1}{m'}\left(357d + 96\ln\left(\frac{12}{\delta'}\right)\right),
\end{aligned}$$

where the inequality on the second-to-last line is due to Lemma 8 in Appendix A. If $m' \geq 357$, then the expression on this last line is at most

$$\frac{1}{m'+1}\left(358d + 97\ln\left(\frac{12}{\delta'}\right)\right),$$

and otherwise, if $m' < 357$, then we trivially have $\mathrm{er}_{\mathcal{P}}(\hat{h}_{m',\delta'}; f^\star) \leq 1 < \frac{1}{m'+1}\left(358d + 97\ln\left(\frac{12}{\delta'}\right)\right)$. Also note that, for any $m \in \mathbb{N}$ and $\delta \in (0,1)$ with $m \leq c_0(d+\ln(12/\delta))$, every $m' \in \mathbb{N}$ and $\delta' \in (0,1)$ with $m' \leq m$ and $\delta' \leq \delta$ also satisfy $m' \leq c_0(d+\ln(12/\delta'))$. Together we have that, for any $m \in \mathbb{N}$ and $\delta \in (0,1)$ with $m \leq c_0(d+\ln(12/\delta))$, for every $m' \in \mathbb{N}$ and $\delta' \in (0,1)$ with $m' \leq m$ and $\delta' \leq \delta$, with probability at least $1-\delta'$, (6) holds.

Now take as an inductive hypothesis that, for some $m \in \mathbb{N}$ and $\delta \in (0,1)$ with $m > c_0(d+\ln(12/\delta))$, for every $m' \in \mathbb{N}$ and $\delta' \in (0,1)$ with $m' < m$ and $\delta' < \delta$, with probability at least $1-\delta'$, (6) is satisfied. Let S_1, S_2, $\hat{h}_1$, $\hat{h}_2$, and $\hat{h}_3$ be as in the definition of the algorithm $\mathbb{A}_{\mathcal{P}}^{(\delta)}(S)$, with $S = (\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P})))$. In particular, note that $m \geq 2$, so that S_1 and S_2 are both nonempty, and $\max\{|S_1|, |S_2|\} < m$. Furthermore, by definition

of $\mathbb{A}_{\mathcal{P}}^{(\delta/6)}$, we have $\hat{h}_1 \in \mathbb{C}[S_1]$, $\hat{h}_2 \in \mathbb{C}[S_2]$, and $\hat{h}_3 \in \mathbb{C}[S]$. For each $i \in \{1,2,3\}$, define $R_i = \{x \in \mathcal{X} : \hat{h}_i(x) \neq f^\star(x)\}$, the error region of $\hat{h}_i$.

Note that $S_1 = (\mathbb{X}_{1:\lfloor m/2 \rfloor}(\mathcal{P}), f^\star(\mathbb{X}_{1:\lfloor m/2 \rfloor}(\mathcal{P})))$. Also note that (by the i.i.d. assumption) S_2 is distributionally equivalent to $(\mathbb{X}_{1:\lceil m/2 \rceil}(\mathcal{P}), f^\star(\mathbb{X}_{1:\lceil m/2 \rceil}(\mathcal{P})))$. Thus, by the inductive hypothesis, for each $i \in \{1,2\}$, there is an event E_i of probability at least $1-\delta/6$, on which

$$\mathcal{P}(R_i) \leq \frac{1}{|S_i|+1}\left(358d + 97\ln\left(\frac{72}{\delta}\right)\right) \leq \frac{1}{\lfloor m/2 \rfloor}\left(358d + 97\ln\left(\frac{72}{\delta}\right)\right). \tag{7}$$

Next, fix any $i \in \{1,2\}$ and denote by $\bar{i}$ the element of $\{1,2\}$ not equal to i. In particular, note that every $j \in \{2,3\}$ with $j > i$ has $\hat{h}_j \in \mathbb{C}[S_{\bar{i}}]$. Denote by $N_{\bar{i}} = |\{(x,y) \in S_{\bar{i}} : \hat{h}_i(x) \neq y\}|$, and let $(Z_{\bar{i},1}, f^\star(Z_{\bar{i},1})), \ldots, (Z_{\bar{i},N_{\bar{i}}}, f^\star(Z_{\bar{i},N_{\bar{i}}}))$ denote the subsequence of $S_{\bar{i}}$ for which $Z_{\bar{i},t} \in R_i$, $t \in \{1,\ldots,N_{\bar{i}}\}$. Note that $Z_{\bar{i},1}, \ldots, Z_{\bar{i},N_{\bar{i}}}$ are conditionally independent given $\hat{h}_i$ and $N_{\bar{i}}$, each with conditional distribution $\mathcal{P}(\cdot|R_i)$ (if $N_{\bar{i}} > 0$). Thus, applying Lemma 4 under the conditional distribution given $\hat{h}_i$ and $N_{\bar{i}}$, combined with the law of total probability, we have that on an event E_i' of probability at least $1-\delta/6$, if $N_{\bar{i}} > 0$, then every $h \in \mathbb{C}[\{(Z_{\bar{i},t}, f^\star(Z_{\bar{i},t}))\}_{t=1}^{N_{\bar{i}}}]$ satisfies

$$\mathrm{er}_{\mathcal{P}(\cdot|R_i)}(h; f^\star) \leq \frac{2}{N_{\bar{i}}}\left(d\mathrm{Log}\left(\frac{2eN_{\bar{i}}}{d}\right) + \mathrm{Log}\left(\frac{12}{\delta}\right)\right).$$

Furthermore, we have $\mathbb{C}[\{(Z_{\bar{i},t}, f^\star(Z_{\bar{i},t}))\}_{t=1}^{N_{\bar{i}}}] \supseteq \mathbb{C}[S_{\bar{i}}]$, so that every $j \in \{2,3\}$ with $j > i$ has $\hat{h}_j \in \mathbb{C}[\{(Z_{\bar{i},t}, f^\star(Z_{\bar{i},t}))\}_{t=1}^{N_{\bar{i}}}]$. Thus, on the event E_i', if $N_{\bar{i}} > 0$, every such j has

$$\begin{aligned}\mathcal{P}(R_i \cap R_j) &= \mathcal{P}(R_i)\mathcal{P}(R_j|R_i) = \mathcal{P}(R_i)\mathrm{er}_{\mathcal{P}(\cdot|R_i)}(\hat{h}_j; f^\star) \\ &\leq \mathcal{P}(R_i)\frac{2}{N_{\bar{i}}}\left(d\mathrm{Log}\left(\frac{2eN_{\bar{i}}}{d}\right) + \mathrm{Log}\left(\frac{12}{\delta}\right)\right).\end{aligned} \tag{8}$$

Additionally, by a Chernoff bound (applied under the conditional distribution given $\hat{h}_i$) and the law of total probability, there is an event E_i'' of probability at least $1-\delta/6$, on which, if $\mathcal{P}(R_i) \geq \frac{8}{\lfloor m/2 \rfloor}\ln\left(\frac{6}{\delta}\right)$, then

$$N_{\bar{i}} \geq \mathcal{P}(R_i)|S_{\bar{i}}|/2 \geq \mathcal{P}(R_i)\lfloor m/2 \rfloor/2.$$

This also means that, on E_i'', if $\mathcal{P}(R_i) \geq \frac{8}{\lfloor m/2 \rfloor}\ln\left(\frac{6}{\delta}\right)$, then $\mathcal{P}(R_i)\lfloor m/2 \rfloor/2 > 0$, so that $N_{\bar{i}} > 0$.

Combining this with (7) and (8), and noting that $x \mapsto \frac{1}{x}\mathrm{Log}(cx)$ is nonincreasing on $(0,\infty)$ (for any fixed $c > 0$), we have that on $E_i \cap E_i' \cap E_i''$, if $\mathcal{P}(R_i) \geq \frac{8}{\lfloor m/2 \rfloor}\ln\left(\frac{6}{\delta}\right)$, then

$$\begin{aligned}\mathcal{P}(R_i \cap R_j) &\leq \frac{4}{\lfloor m/2 \rfloor}\left(d\mathrm{Log}\left(\frac{e\mathcal{P}(R_i)\lfloor m/2 \rfloor}{d}\right) + \mathrm{Log}\left(\frac{12}{\delta}\right)\right) \\ &\leq \frac{4}{\lfloor m/2 \rfloor}\left(d\mathrm{Log}\left(\frac{e\left(358d + 97\ln\left(\frac{72}{\delta}\right)\right)}{d}\right) + \mathrm{Log}\left(\frac{12}{\delta}\right)\right) \\ &\leq \frac{4}{\lfloor m/2 \rfloor}\left(d\mathrm{Log}\left(97\left(\frac{358e}{97} + e\ln(6) + \frac{e}{d}\ln\left(\frac{12}{\delta}\right)\right)\right) + \mathrm{Log}\left(\frac{12}{\delta}\right)\right) \\ &\leq \frac{4}{\lfloor m/2 \rfloor}\left(d\ln\left(358e + 97e\ln(6) + 97e\right) + 2\ln\left(\frac{12}{\delta}\right)\right),\end{aligned}$$

where this last inequality is due to Lemma 8 in Appendix A. Additionally, if $\mathcal{P}(R_i) < \frac{8}{\lfloor m/2 \rfloor} \ln\left(\frac{6}{\delta}\right)$, then monotonicity of probability measures implies

$$\mathcal{P}(R_i \cap R_j) \le \mathcal{P}(R_i) < \frac{8}{\lfloor m/2 \rfloor} \ln\left(\frac{6}{\delta}\right) \le \frac{4}{\lfloor m/2 \rfloor}\left(d \ln\left(358e + 97e\ln(6) + 97e\right) + 2\ln\left(\frac{12}{\delta}\right)\right).$$

Thus, regardless of the value of $\mathcal{P}(R_i)$, on the event $E_i \cap E_i' \cap E_i''$, we have

$$\mathcal{P}(R_i \cap R_j) \le \frac{4}{\lfloor m/2 \rfloor}\left(d \ln\left(358e + 97e\ln(6) + 97e\right) + 2\ln\left(\frac{12}{\delta}\right)\right).$$

Next, define the majority vote classifier: $\hat{h}_{\text{Maj}}(x) = \text{Majority}(\hat{h}_1(x), \hat{h}_2(x), \hat{h}_3(x))$ (for all $x \in \mathcal{X}$). Note that for any $x \in \mathcal{X}$ with $\hat{h}_{\text{Maj}}(x) \ne f^\star(x)$, at least two of the values $\hat{h}_1(x)$, $\hat{h}_2(x)$, $\hat{h}_3(x)$ must differ from $f^\star(x)$: that is, $x \in (R_1 \cap R_2) \cup (R_1 \cap R_3) \cup (R_2 \cap R_3)$. Therefore, by the union bound,

$$\text{er}_{\mathcal{P}}(\hat{h}_{\text{Maj}}; f^\star) \le \mathcal{P}(R_1 \cap R_2) + \mathcal{P}(R_1 \cap R_3) + \mathcal{P}(R_2 \cap R_3).$$

We therefore have that, on the event $\bigcap_{i\in\{1,2\}} E_i \cap E_i' \cap E_i''$,

$$\text{er}_{\mathcal{P}}(\hat{h}_{\text{Maj}}; f^\star) \le \frac{12}{\lfloor m/2 \rfloor}\left(d \ln\left(358e + 97e\ln(6) + 97e\right) + 2\ln\left(\frac{12}{\delta}\right)\right). \tag{9}$$

Finally, since $\{x : \hat{h}_{m,\delta}(x) \ne f^\star(x)\} \subseteq \{x : \hat{h}_{m,\delta}(x) \ne \hat{h}_{\text{Maj}}(x)\} \cup \{x : \hat{h}_{\text{Maj}}(x) \ne f^\star(x)\}$, the union bound implies

$$\text{er}_{\mathcal{P}}(\hat{h}_{m,\delta}; f^\star) = \mathcal{P}(x : \hat{h}_{m,\delta}(x) \ne f^\star(x)) \le \mathcal{P}(x : \hat{h}_{m,\delta}(x) \ne \hat{h}_{\text{Maj}}(x)) + \mathcal{P}(x : \hat{h}_{\text{Maj}}(x) \ne f^\star(x)).$$

By definition (in Step 4),

$$\mathcal{P}(x : \hat{h}_{m,\delta}(x) \ne \hat{h}_{\text{Maj}}(x)) \le \inf_{h \in \mathbb{C}[S]} \mathcal{P}(x : h(x) \ne \hat{h}_{\text{Maj}}(x)) + \gamma(m, \delta),$$

and since $f^\star \in \mathbb{C}[S]$, this is at most $\mathcal{P}(x : f^\star(x) \ne \hat{h}_{\text{Maj}}(x)) + \gamma(m, \delta)$. Therefore, $\text{er}_{\mathcal{P}}(\hat{h}_{m,\delta}; f^\star) \le 2\mathcal{P}(x : \hat{h}_{\text{Maj}}(x) \ne f^\star(x)) + \gamma(m, \delta) = 2\text{er}_{\mathcal{P}}(\hat{h}_{\text{Maj}}; f^\star) + \gamma(m, \delta)$. Combining this with (9) we have that, on the event $\bigcap_{i\in\{1,2\}} E_i \cap E_i' \cap E_i''$,

$$\text{er}_{\mathcal{P}}(\hat{h}_{m,\delta}; f^\star) \le \frac{24}{\lfloor m/2 \rfloor}\left(d \ln\left(358e + 97e\ln(6) + 97e\right) + 2\ln\left(\frac{12}{\delta}\right)\right) + \gamma(m, \delta). \tag{10}$$

Since we are considering a value $m > c_0(d + \ln(12/\delta)) > c_0 \ln(12e)$, we have that $\lfloor m/2 \rfloor > \frac{m-2}{2} > \left(1 - \frac{2}{c_0 \ln(12e)}\right)\frac{m}{2} > \left(1 - \frac{2}{c_0 \ln(12e)}\right)\frac{c_0 \ln(12e)}{c_0 \ln(12e)+1}\frac{m+1}{2} = \frac{c_0 \ln(12e)-2}{c_0 \ln(12e)+1}\frac{m+1}{2}$. Therefore, the right hand side of the inequality (10) is at most

$$\frac{c_0 \ln(12e) + 1}{c_0 \ln(12e) - 2}\frac{48}{m+1}\left(d \ln\left(358e + 97e\ln(6) + 97e\right) + 2\ln\left(\frac{12}{\delta}\right)\right) + \gamma(m, \delta).$$

By a direct calculation of the numerical value of the logarithmic factor, together with the facts that $c_0 \geq e^7$ and $\gamma(m,\delta) \leq \frac{1}{3(m+1)}\left(d + \ln\left(\frac{12}{\delta}\right)\right)$, the above expression is at most

$$\frac{1}{m+1}\left(358d + 97\ln\left(\frac{12}{\delta}\right)\right).$$

Furthermore, by the union bound, the event $\bigcap_{i\in\{1,2\}} E_i \cap E_i' \cap E_i''$ has probability at least $1-\delta$. Thus, we have succeeded in extending the inductive hypothesis to include $m' = m$ and $\delta' = \delta$.

By the principle of induction, we have established the claim that, for every $m \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\mathrm{er}_{\mathcal{P}}(\hat{h}_{m,\delta}; f^\star) \leq \frac{1}{m+1}\left(358d + 97\ln\left(\frac{12}{\delta}\right)\right). \tag{11}$$

To complete the proof, we simply note that, for any $\varepsilon, \delta \in (0,1)$, for any value of $m \in \mathbb{N}$ of size at least

$$\left\lfloor \frac{1}{\varepsilon}\left(358d + 97\ln\left(\frac{12}{\delta}\right)\right)\right\rfloor, \tag{12}$$

the right hand side of (11) is less than ε, so that $\mathbb{A}_{\mathcal{P}}^{(\delta)}$ achieves a sample complexity equal the expression in (12). In particular, this implies

$$\tilde{\mathcal{M}}(\varepsilon,\delta) \leq \frac{1}{\varepsilon}\left(358d + 97\ln\left(\frac{12}{\delta}\right)\right) = O\left(\frac{1}{\varepsilon}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\right).$$

■

5. Replacing Distribution-Dependence with Unlabeled Data

The algorithm above is stated as a distribution-dependent algorithm. Since arbitrary access to the probability measure $\mathcal{P}$ is typically considered a strong requirement, it is worth investigating the nature of the dependence on $\mathcal{P}$ actually required by this algorithm. In particular, in this section we find that it suffices to have an additional number of independent $\mathcal{P}$-distributed *unlabeled* samples, a setting known as *semi-supervised* learning. Specifically, we note that the dependence on $\mathcal{P}$ is isolated to a projection step (Step 4), in which we find a sample-consistent classifier relatively close to the majority vote classifier. Fortunately, as discussed above, this does not need to be the *closest* classifier in $\mathbb{C}[S]$, but merely one whose distance to the majority classifier is within $\gamma(m,\delta)$ of the minimal distance. This slack enables us to provide the required guarantee (with high probability) by projecting onto $\mathbb{C}[S]$ using L_1 distances calculated via the *empirical* probability measure $\hat{\mathcal{P}}$ based on the additional unlabeled samples. Thus, we have a *distribution-independent* semi-supervised learning algorithm, which still achieves the sample complexity guarantee of Theorem 2 above (aside from a small increase in a constant). We also discuss the number of unlabeled samples sufficient for this method.

Formally, let $X'_1(\mathcal{P}), X'_2(\mathcal{P}), \ldots$ be independent $\mathcal{P}$-distributed random variables (also independent from all $X_i(P)$). For any $u \in \mathbb{N}$, define $\mathbb{X}'_{1:u}(\mathcal{P}) = \{X'_1(\mathcal{P}), \ldots, X'_u(\mathcal{P})\}$. For any $u \in \mathbb{N}$, any $\mathcal{U} \in \mathcal{X}^u$, and any measurable set $A \subseteq \mathcal{X}$, define

$$\hat{\mathcal{P}}_{\mathcal{U}}(A) = \frac{1}{u}\sum_{x \in \mathcal{U}} \mathbb{1}[x \in A],$$

the *empirical probability* of A. Now consider the following modification of $\mathbb{A}^{(\delta)}_{\mathcal{P}}$, where $\mathcal{U}$ is a finite sequence of points in $\mathcal{X}$, and S is a finite sequence of points in $\mathcal{X} \times \mathcal{Y}$ (with $\mathbb{C}[S] \neq \emptyset$).[6]

Algorithm: $\hat{\mathbb{A}}^{(\delta)}_{\mathcal{U}}(S)$
0. If $|S| \leq c_0(d + \ln(12/\delta))$
1. Return any $\hat{h} \in \mathbb{C}[S]$
2. Let S_1 denote the first $\lfloor |S|/2 \rfloor$ elements of S, S_2 the remaining $\lceil |S|/2 \rceil$ elements
3. Let $\hat{h}_1 = \hat{\mathbb{A}}^{(\delta/6)}_{\mathcal{U}}(S_1)$, $\hat{h}_2 = \hat{\mathbb{A}}^{(\delta/6)}_{\mathcal{U}}(S_2)$, and $\hat{h}_3$ is an arbitrary element of $\mathbb{C}[S]$
4. Return $\hat{h} = \operatorname{argmin}_{h \in \mathbb{C}[S]} \hat{\mathcal{P}}_{\mathcal{U}}(\{x : h(x) \neq \text{Majority}(\hat{h}_1(x), \hat{h}_2(x), \hat{h}_3(x))\})$

Note that $\hat{\mathbb{A}}^{(\delta)}_{\mathcal{U}}$ is defined precisely as $\mathbb{A}^{(\delta)}_{\mathcal{P}}$ above, except that in Step 4, $\mathcal{P}$ is replaced by $\hat{\mathcal{P}}_{\mathcal{U}}$ and $\gamma(m, \delta)$ is repalced by 0. The idea is that, if $\hat{\mathcal{P}}_{\mathcal{U}}$ is within $\gamma(m, \delta)$ of $\mathcal{P}$ for all evaluations of the latter in $\mathbb{A}^{(\delta)}_{\mathcal{P}}$, then $\hat{\mathbb{A}}^{(\delta)}_{\mathcal{U}}$ will be a valid *instantiation* of $\mathbb{A}^{(\delta)}_{\mathcal{P}}$. In order to argue that this is indeed the case (for $\mathcal{U} = \mathbb{X}'_{1:u}(\mathcal{P})$, for a sufficiently large u), we first state the following lemma. This result is largely based on a well-known generalization of Lemma 4, due to Vapnik (1982) (along with a complementary result of Bousquet, Boucheron, and Lugosi, 2004), plus some additional reasoning about the VC dimension of loss regions for majority vote classifiers based on a result of Vidyasagar (2003). For completeness, we include a brief derivation of the result in its stated form, starting from the theorem of Vapnik (1982).

Lemma 5 *There is a finite numerical constant $\tilde{c}_1 \geq 1$ such that, for any nonempty set $\mathcal{H}$ of classifiers with VC dimension at most d, for any probability measure $\mathcal{P}$ over $\mathcal{X}$, any $\gamma, \delta \in (0,1)$ and $\nu \in [0,1]$, and any $u \in \mathbb{N}$ satisfying*

$$u \geq \frac{\tilde{c}_1(\nu + \gamma)}{\gamma^2}\left(d\text{Log}\left(\frac{1}{\gamma}\right) + \text{Log}\left(\frac{1}{\delta}\right)\right), \tag{13}$$

with probability at least $1 - \delta$, for every $h_1, h_2, h_3 \in \mathbb{C}$ such that $\exists h' \in \mathcal{H}$ with $\mathcal{P}(x : h'(x) \neq \text{Majority}(h_1(x), h_2(x), h_3(x))) \leq \nu$, the classifier

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}(x : h(x) \neq \text{Majority}(h_1(x), h_2(x), h_3(x)))$$

satisfies

$$\mathcal{P}\left(x : \hat{h}(x) \neq \text{Majority}(h_1(x), h_2(x), h_3(x))\right) \leq \inf_{h \in \mathcal{H}} \mathcal{P}\left(x : h(x) \neq \text{Majority}(h_1(x), h_2(x), h_3(x))\right) + \gamma.$$

6. As above, we suppose any partially ambiguous steps are instantiated in a way that depends only on the input S (and $\delta, \mathcal{U}$), and any internal randomness, so that for any $\mathcal{U}$, this is a valid learning algorithm.

Proof Define a collection of sets

$$\mathcal{D} = \{\{x : \text{Majority}(h_1(x), h_2(x), h_3(x)) \neq h_0(x)\} : h_1, h_2, h_3 \in \mathbb{C}, h_0 \in \mathcal{H}\}.$$

Since each set $\{x : \text{Majority}(h_1(x), h_2(x), h_3(x)) \neq h_0(x)\} \in \mathcal{D}$ is determined by the classifiers $h_1, h_2, h_3 \in \mathbb{C}$ and $h_0 \in \mathcal{H}$, a theorem of Vidyasagar (2003, Theorem 4.5) implies that the VC dimension of $\mathcal{D}$ is less than $8\log_2(4e)d$.[7] Denote by $d_{\mathcal{D}} = 8\log_2(4e)d$.

Next, a theorem of Vapnik (1982, Theorem 6.8), along with the fact that $\sqrt{1+a} \leq 1+\sqrt{a}$ for any $a \geq 0$, implies that, with probability at least $1-\delta/2$, every $D \in \mathcal{D}$ satisfies

$$\mathcal{P}(D) \leq \hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}(D) + \frac{4}{u}\left(d_{\mathcal{D}}\text{Log}\left(\frac{2eu}{d_{\mathcal{D}}}\right) + \text{Log}\left(\frac{24}{\delta}\right)\right) + \sqrt{\hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}(D)\frac{4}{u}\left(d_{\mathcal{D}}\text{Log}\left(\frac{2eu}{d_{\mathcal{D}}}\right) + \text{Log}\left(\frac{24}{\delta}\right)\right)}.$$

In particular, for a sufficiently large choice of the constant $\tilde{c}_1$, combining the above inequality with the constraint (13) on u implies (see e.g., Vidyasagar, 2003, Corollary 4.1)

$$\mathcal{P}(D) \leq \hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}(D) + \frac{\gamma^2}{16(\nu+\gamma)} + \sqrt{\hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}(D)\frac{\gamma^2}{16(\nu+\gamma)}}. \tag{14}$$

Additionally, a result of Bousquet, Boucheron, and Lugosi (2004, Theorem 7), together with the VC-Sauer lemma (Vapnik and Chervonenkis, 1971; Sauer, 1972), implies that, with probability at least $1-\delta/2$, every $D \in \mathcal{D}$ satisfies

$$\hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}(D) \leq \mathcal{P}(D) + 2\sqrt{\hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}(D)\frac{1}{u}\left(d_{\mathcal{D}}\text{Log}\left(\frac{2em}{d_{\mathcal{D}}}\right) + \text{Log}\left(\frac{8}{\delta}\right)\right)}.$$

Furthermore, noting that the above is a quadratic inequality in $\sqrt{\hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}(D)}$, solving this inequality (and again using the fact that $\sqrt{1+a} \leq 1+\sqrt{a}$ for $a \geq 0$) yields that, on the above event,

$$\hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}(D) \leq \mathcal{P}(D) + \frac{4}{u}\left(d_{\mathcal{D}}\text{Log}\left(\frac{2em}{d_{\mathcal{D}}}\right) + \text{Log}\left(\frac{8}{\delta}\right)\right) + \sqrt{\mathcal{P}(D)\frac{4}{u}\left(d_{\mathcal{D}}\text{Log}\left(\frac{2em}{d_{\mathcal{D}}}\right) + \text{Log}\left(\frac{8}{\delta}\right)\right)}.$$

As above, for a sufficiently large choice of the constant $\tilde{c}_1$, combining the above inequality with the constraint (13) on u implies

$$\hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}(D) \leq \mathcal{P}(D) + \frac{\gamma^2}{16(\nu+\gamma)} + \sqrt{\mathcal{P}(D)\frac{\gamma^2}{16(\nu+\gamma)}}. \tag{15}$$

7. In the terminology introduced above, the VC dimension of the collection $\mathcal{D}$ of sets is defined as the VC dimension of the corresponding set of classifiers $\{x \mapsto 2\mathbb{1}_D(x) - 1 : D \in \mathcal{D}\}$.

By the union bound, with probability at least $1-\delta$, every $D \in \mathcal{D}$ satisfies both (14) and (15). In particular, fix any $h_1, h_2, h_3 \in \mathbb{C}$ for which $\exists h' \in \mathcal{H}$ with $\mathcal{P}(x : h'(x) \neq \text{Majority}(h_1(x), h_2(x), h_3(x))) \leq \nu$, and fix any such $h' \in \mathcal{H}$. Denote by $h_{\text{Maj}}(x) = \text{Majority}(h_1(x), h_2(x), h_3(x))$ (for all $x \in \mathcal{X}$), and let $\hat{h} = \text{argmin}_{h \in \mathcal{H}} \hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}(x : h(x) \neq h_{\text{Maj}}(x))$. In particular, we have

$$\hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}\left(x : \hat{h}(x) \neq h_{\text{Maj}}(x)\right) \leq \hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}\left(x : h'(x) \neq h_{\text{Maj}}(x)\right).$$

Thus, on the above event, (14) implies

$$\begin{aligned}
&\mathcal{P}\left(x : \hat{h}(x) \neq h_{\text{Maj}}(x)\right) \\
&\leq \hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}\left(x : \hat{h}(x) \neq h_{\text{Maj}}(x)\right) + \frac{\gamma^2}{16(\nu+\gamma)} + \sqrt{\hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}\left(x : \hat{h}(x) \neq h_{\text{Maj}}(x)\right)\frac{\gamma^2}{16(\nu+\gamma)}} \\
&\leq \hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}\left(x : h'(x) \neq h_{\text{Maj}}(x)\right) + \frac{\gamma^2}{16(\nu+\gamma)} + \sqrt{\hat{\mathcal{P}}_{\mathbb{X}'_{1:u}(\mathcal{P})}(x : h'(x) \neq h_{\text{Maj}}(x))\frac{\gamma^2}{16(\nu+\gamma)}},
\end{aligned}$$

while (15), and the fact that $\mathcal{P}(x : h'(x) \neq h_{\text{Maj}}(x)) \leq \nu$, imply this last line is at most

$$\begin{aligned}
&\mathcal{P}\left(x : h'(x) \neq h_{\text{Maj}}(x)\right) + \frac{\gamma^2}{16(\nu+\gamma)} + \sqrt{\nu\frac{\gamma^2}{16(\nu+\gamma)}} + \frac{\gamma^2}{16(\nu+\gamma)} \\
&\qquad + \sqrt{\left(\nu + \frac{\gamma^2}{16(\nu+\gamma)} + \sqrt{\nu\frac{\gamma^2}{16(\nu+\gamma)}}\right)\frac{\gamma^2}{16(\nu+\gamma)}} \\
&\leq \mathcal{P}\left(x : h'(x) \neq h_{\text{Maj}}(x)\right) + \gamma.
\end{aligned}$$

Since this holds for every $h' \in \mathcal{H}$ with $\mathcal{P}(x : h'(x) \neq h_{\text{Maj}}(x)) \leq \nu$, we have that as long as some such $h' \in \mathcal{H}$ exists, on the above event,

$$\begin{aligned}
\mathcal{P}\left(x : \hat{h}(x) \neq h_{\text{Maj}}(x)\right) &\leq \inf_{\substack{h' \in \mathcal{H}: \\ \mathcal{P}(x:h'(x) \neq h_{\text{Maj}}(x)) \leq \nu}} \mathcal{P}\left(x : h'(x) \neq h_{\text{Maj}}(x)\right) + \gamma \\
&= \inf_{h \in \mathcal{H}} \mathcal{P}\left(x : h(x) \neq h_{\text{Maj}}(x)\right) + \gamma.
\end{aligned}$$

■

With this lemma in hand, we have the following result, which indicates that the semi-supervised learning algorithm $\hat{\mathbb{A}}_{\mathcal{U}}^{(\delta/2)}$ can indeed be used in place of $\mathbb{A}_{\mathcal{P}}^{(\delta)}$ to achieve the optimal distribution-free sample complexity (for the number of *labeled* data) without any direct dependence on the distribution $\mathcal{P}$, when given access to a pool of unlabeled data $\mathbb{X}'_{1:u}(\mathcal{P})$.

Theorem 6 *There is a finite numerical constant $\tilde{c}_2 \geq 1$ and a $\mathbb{N}$-valued function $M_{\text{SS}}(\varepsilon, \delta)$ with*

$$M_{\text{SS}}(\varepsilon, \delta) = O\left(\frac{1}{\varepsilon}\left(d + \text{Log}\left(\frac{1}{\delta}\right)\right)\right) \tag{16}$$

such that, for any probability measure $\mathcal{P}$ over $\mathcal{X}$, any $f^\star \in \mathbb{C}$, any $\varepsilon, \delta \in (0,1)$, any integer $m \geq M_{\mathrm{SS}}(\varepsilon, \delta)$, and any integer

$$u \geq \tilde{c}_2 m \frac{d\mathrm{Log}\left(\frac{m}{d}\right) + \mathrm{Log}\left(\frac{24}{\delta}\right)}{d + \mathrm{Log}\left(\frac{24}{\delta}\right)}, \tag{17}$$

with probability at least $1-\delta$, the classifier $\hat{h} = \hat{\mathbb{A}}^{(\delta/2)}_{\mathbb{X}'_{1:u}(\mathcal{P})}((\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P}))))$ satisfies $\mathrm{er}_{\mathcal{P}}(\hat{h}; f^\star) \leq \varepsilon$.

In particular, Theorem 6 immediately implies the following corollary.

Corollary 7 *There are $\mathbb{N}$-valued functions $M_{\mathrm{SS}}(\varepsilon, \delta)$ and $U_{\mathrm{SS}}(\varepsilon, \delta)$ with*

$$M_{\mathrm{SS}}(\varepsilon, \delta) = O\left(\frac{1}{\varepsilon}\left(d + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\right)$$

and

$$U_{\mathrm{SS}}(\varepsilon, \delta) = O\left(\frac{1}{\varepsilon}\left(d\mathrm{Log}\left(\frac{1}{\varepsilon}\right) + \mathrm{Log}\left(\frac{1}{\delta}\right)\right)\right) \tag{18}$$

such that, for any probability measure $\mathcal{P}$ over $\mathcal{X}$, any $f^\star \in \mathbb{C}$, and any $\varepsilon, \delta \in (0,1)$, taking $m = M_{\mathrm{SS}}(\varepsilon, \delta)$ and $u = U_{\mathrm{SS}}(\varepsilon, \delta)$, with probability at least $1-\delta$, the classifier $\hat{h} = \hat{\mathbb{A}}^{(\delta/2)}_{\mathbb{X}'_{1:u}(\mathcal{P})}((\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P}))))$ satisfies $\mathrm{er}_{\mathcal{P}}(\hat{h}; f^\star) \leq \varepsilon$.

Proof Let $M_{\mathrm{SS}}(\varepsilon, \delta)$ be as in Theorem 6, and let $U_{\mathrm{SS}}(\varepsilon, \delta)$ equal the right hand side of (17) (rounded up to an integer) for $m = M_{\mathrm{SS}}(\varepsilon, \delta)$. Then note that Lemma 8 and some simple algebra reveal that $U_{\mathrm{SS}}(\varepsilon, \delta)$ satisfies (18). Finally, since choosing m and u with $m = M_{\mathrm{SS}}(\varepsilon, \delta)$ and $u = U_{\mathrm{SS}}(\varepsilon, \delta)$ satisfies the constraints of Theorem 6, the conclusion that $\mathbb{P}\left(\mathrm{er}_{\mathcal{P}}(\hat{h}; f^\star) \leq \varepsilon\right) \geq 1-\delta$ follows from the theorem. ■

We now present the proof of Theorem 6.

Proof of Theorem 6 We first note that, for any $\mathcal{P}$, the algorithm $\mathbb{A}^{(\delta)}_{\mathcal{P}}$ is technically an entire family of algorithms, distinguished by the choices of c_0 and $\gamma(\cdot,\cdot)$ (subject to the stated constraints), how they select the classifier $\hat{h}$ in Step 1, the classifier $\hat{h}_3$ in Step 3, and the classifier $\hat{h}$ in Step 4. Each of these may have several options to choose from. However, regardless of how these choices are made (based on the input set S and parameter δ for the given call, the distribution $\mathcal{P}$, and any internal independent randomness), the analysis in the proof of Theorem 2, and the conclusion of Theorem 2 itself, remain valid.

Now consider a special subfamily of these algorithms $\mathbb{A}^{(\delta)}_{\mathcal{P}}$. Specifically, let $\tilde{\gamma}(m, \delta) = \frac{1}{3(m+1)}\left(d + \ln\left(\frac{12}{\delta}\right)\right)$, for $m \in \mathbb{N}$ and $\delta \in (0,1)$, and for any sequence $\mathcal{U} \in \mathcal{X}^u$ and $\delta \in (0,1)$, let $\tilde{\mathbb{A}}^{(\delta)}_{\mathcal{P},\mathcal{U}}$ denote an instantiation of $\mathbb{A}^{(\delta)}_{\mathcal{P}}$ defined as follows.

Algorithm: $\tilde{\mathbb{A}}^{(\delta)}_{\mathcal{P},\mathcal{U}}(S)$
0. If $|S| \leq c_0(d + \ln(12/\delta))$
1. Return any $\hat{h} \in \mathbb{C}[S]$
2. Let S_1 denote the first $\lfloor |S|/2 \rfloor$ elements of S, S_2 the remaining $\lceil |S|/2 \rceil$ elements
3. Let $\hat{h}_1 = \tilde{\mathbb{A}}^{(\delta/6)}_{\mathcal{P},\mathcal{U}}(S_1)$, $\hat{h}_2 = \tilde{\mathbb{A}}^{(\delta/6)}_{\mathcal{P},\mathcal{U}}(S_2)$, and $\hat{h}_3$ is an arbitrary element of $\mathbb{C}[S]$
4. Let $\hat{h}' = \operatorname{argmin}_{h \in \mathbb{C}[S]} \hat{\mathcal{P}}_{\mathcal{U}}(x : h(x) \neq \text{Majority}(\hat{h}_1(x), \hat{h}_2(x), \hat{h}_3(x)))$
and return $\hat{h} = \hat{h}'$ if $\hat{h}' \in \operatorname{argmin}^{\tilde{\gamma}(|S|,\delta)}_{h \in \mathbb{C}[S]} \mathcal{P}(x : h(x) \neq \text{Majority}(\hat{h}_1(x), \hat{h}_2(x), \hat{h}_3(x)))$
and otherwise return any
$\hat{h} \in \operatorname{argmin}^{\tilde{\gamma}(|S|,\delta)}_{h \in \mathbb{C}[S]} \mathcal{P}(x : h(x) \neq \text{Majority}(\hat{h}_1(x), \hat{h}_2(x), \hat{h}_3(x)))$

In particular, note that $\tilde{\mathbb{A}}^{(\delta)}_{\mathcal{P},\mathcal{U}}(S)$ uses valid choices of $\hat{h}$ in Steps 1 and 4, and $\hat{h}_3$ in Step 3, according to the definition of $\mathbb{A}^{(\delta)}_{\mathcal{P}}(S)$, so that $\tilde{\mathbb{A}}^{(\delta)}_{\mathcal{P},\mathcal{U}}$ is indeed a member of the family of $\mathbb{A}^{(\delta)}_{\mathcal{P}}$ algorithms (and hence, the analysis in the proof of Theorem 2 applies to it).

For any finite sequence $\mathcal{U}$ in $\mathcal{X}$, the semi-supervised algorithm $\hat{\mathbb{A}}^{(\delta)}_{\mathcal{U}}$ also describes an entire family of algorithms. In this case, the choices of c_0, $\hat{h}$ in Step 1, and $\hat{h}_3$ in Step 3 are subject to the same criteria as in $\mathbb{A}^{(\delta)}_{\mathcal{P}}$, but the choice of $\hat{h}$ in Step 4 is subject to a different constraint. Let us fix any particular instantiation of these choices, so that for any given $\mathcal{U}$, $\hat{\mathbb{A}}^{(\delta)}_{\mathcal{U}}$ is a particular algorithm in this family. For any $\mathcal{P}$, let us further consider the instantiation of $\tilde{\mathbb{A}}^{(\delta)}_{\mathcal{P},\mathcal{U}}(S)$ which makes these choices in the same way: that is, the choice of c_0, the choice of $\hat{h}$ in Step 1, the choice of $\hat{h}_3$ in Step 3, and the choice of $\hat{h}'$ in Step 4, are each identical to the respective choices of c_0, $\hat{h}$ in Step 1, $\hat{h}_3$ in Step 3, and $\hat{h}$ in Step 4, by $\hat{\mathbb{A}}^{(\delta)}_{\mathcal{U}}(S)$, at the root level and also in every recursive call, whenever these respective steps are reached under otherwise identical conditions (so, for instance, $\hat{h}'$ in $\tilde{\mathbb{A}}^{(\delta)}_{\mathcal{P},\mathcal{U}}(S)$ would be the same as $\hat{h}$ in $\hat{\mathbb{A}}^{(\delta)}_{\mathcal{U}}(S)$ in Step 4 as long as $\hat{h}_1$,$\hat{h}_2$,$\hat{h}_3$ from the previous step are the same in $\tilde{\mathbb{A}}^{(\delta)}_{\mathcal{P},\mathcal{U}}(S)$ as in $\hat{\mathbb{A}}^{(\delta)}_{\mathcal{U}}(S)$). In particular, note that if

$$\hat{h}' \in \operatorname{argmin}^{\tilde{\gamma}(|S'|,\delta')}_{h \in \mathbb{C}[S']} \mathcal{P}(x : h(x) \neq \text{Majority}(\hat{h}_1(x), \hat{h}_2(x), \hat{h}_3(x))) \tag{19}$$

for every $S' \subseteq S$ and $\delta' \leq \delta$ for which $\tilde{\mathbb{A}}^{(\delta')}_{\mathcal{P},\mathcal{U}}(S')$ is called at some point in the execution of $\tilde{\mathbb{A}}^{(\delta)}_{\mathcal{P},\mathcal{U}}(S)$ (including the root call), then we have $\tilde{\mathbb{A}}^{(\delta)}_{\mathcal{P},\mathcal{U}}(S) = \hat{\mathbb{A}}^{(\delta)}_{\mathcal{U}}(S)$.

Next, we note that the proof of Theorem 2 in fact establishes more than stated in the theorem. Specifically, let us fix a probability measure $\mathcal{P}$ over $\mathcal{X}$ and an $f^\star \in \mathbb{C}$ (both arbitrary). Then, for any $m \in \mathbb{N}$ and $\delta \in (0,1)$, from (9) and (10) and the principle of induction (with base case as described in the proof), the proof establishes that, on an event $\tilde{H}(m,\delta)$ of probability at least $1-\delta$, for every call to $\mathbb{A}^{(\delta')}_{\mathcal{P}}(S')$ (with $S' \subseteq (\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P})))$ and $\delta' \leq \delta$) appearing in the execution of $\mathbb{A}^{(\delta)}_{\mathcal{P}}((\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P}))))$ (including the root call and recursive calls), we have that the classifier $h_{S',\delta'} = \mathbb{A}^{(\delta')}_{\mathcal{P}}(S')$ satisfies

$$\text{er}_{\mathcal{P}}(h_{S',\delta'}; f^\star) \leq \frac{1}{|S'|+1}\left(358d + 97\ln\left(\frac{12}{\delta'}\right)\right), \tag{20}$$

and if the call is also *non-terminal* (i.e., $|S'| > c_0(d + \ln(12/\delta'))$), then the classifier $\hat{h}_{\mathrm{Maj}}(x) = \mathrm{Majority}(\hat{h}_1(x), \hat{h}_2(x), \hat{h}_3(x))$ (for $\hat{h}_1$,$\hat{h}_2$,$\hat{h}_3$ in Step 3 of $\mathbb{A}^{(\delta')}_{\mathcal{P}}(S')$) satisfies

$$\begin{aligned}\mathrm{er}_{\mathcal{P}}(\hat{h}_{\mathrm{Maj}}; f^\star) &\leq \frac{12}{\lfloor |S'|/2 \rfloor}\left(d\ln\left(358e + 97e\ln(6) + 97e\right) + 2\ln\left(\frac{12}{\delta'}\right)\right) \\ &\leq \frac{25}{|S'|+1}\left(8d + 2\ln\left(\frac{12}{\delta'}\right)\right). \end{aligned} \tag{21}$$

In particular, this is the case for $\tilde{\mathbb{A}}^{(\delta)}_{\mathcal{P},\mathcal{U}}((\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P}))))$, where the event $\tilde{H}(m,\delta)$ now depends on the particular sequence $\mathcal{U}$. By the law of total probability (to integrate out this dependence), for any $u \in \mathbb{N}$, there is an event $\tilde{H}'(m,u,\delta)$ of probability at least $1-\delta$, on which (20) holds for the classifier $h_{S',\delta'} = \tilde{\mathbb{A}}^{(\delta')}_{\mathcal{P},\mathbb{X}'_{1:u}(\mathcal{P})}(S')$ for every call to $\tilde{\mathbb{A}}^{(\delta')}_{\mathcal{P},\mathbb{X}'_{1:u}(\mathcal{P})}(S')$ (with $S' \subseteq (\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P})))$ and $\delta' \leq \delta$) appearing in the execution of $\tilde{\mathbb{A}}^{(\delta)}_{\mathcal{P},\mathbb{X}'_{1:u}(\mathcal{P})}((\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P}))))$, and (21) holds for every non-terminal such call (for $\hat{h}_1$,$\hat{h}_2$,$\hat{h}_3$ as in Step 3 of $\tilde{\mathbb{A}}^{(\delta')}_{\mathcal{P},\mathbb{X}'_{1:u}(\mathcal{P})}(S')$).

In particular, this implies that, for any $m, u \in \mathbb{N}$ with

$$m \geq \left\lfloor \frac{1}{\varepsilon}\left(358d + 97\ln\left(\frac{24}{\delta}\right)\right)\right\rfloor, \tag{22}$$

on the event $\tilde{H}'(m,u,\delta/2)$, the classifier $\tilde{h}_{u,m} = \tilde{\mathbb{A}}^{(\delta/2)}_{\mathcal{P},\mathbb{X}'_{1:u}(\mathcal{P})}((\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P}))))$ satisfies $\mathrm{er}_{\mathcal{P}}(\tilde{h}_{u,m}; f^\star) \leq \varepsilon$. Take $M_{\mathrm{SS}}(\varepsilon,\delta)$ equal to the right hand side of (22), which one can easily verify satisfies the requirement (16).

Now let $\hat{h}_{u,m} = \hat{\mathbb{A}}^{(\delta/2)}_{\mathbb{X}'_{1:u}(\mathcal{P})}((\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P}))))$. By the union bound, we have that

$$\begin{aligned}&\mathbb{P}\left(\mathrm{er}_{\mathcal{P}}(\hat{h}_{u,m}; f^\star) > \varepsilon\right) \leq \mathbb{P}\left(\mathrm{er}_{\mathcal{P}}(\hat{h}_{u,m}; f^\star) > \varepsilon \text{ and } \tilde{H}'(m,u,\delta/2)\right) + 1 - \mathbb{P}\left(\tilde{H}'(m,u,\delta/2)\right) \\ &\leq \mathbb{P}\left(\mathrm{er}_{\mathcal{P}}(\tilde{h}_{u,m}; f^\star) > \varepsilon \text{ and } \tilde{h}_{u,m} = \hat{h}_{u,m} \text{ and } \tilde{H}'(m,u,\delta/2)\right) \\ &\quad + \mathbb{P}\left(\mathrm{er}_{\mathcal{P}}(\hat{h}_{u,m}; f^\star) > \varepsilon \text{ and } \tilde{h}_{u,m} \neq \hat{h}_{u,m} \text{ and } \tilde{H}'(m,u,\delta/2)\right) + \delta/2 \\ &\leq \mathbb{P}\left(\mathrm{er}_{\mathcal{P}}(\tilde{h}_{u,m}; f^\star) > \varepsilon \text{ and } \tilde{H}'(m,u,\delta/2)\right) + \mathbb{P}\left(\tilde{h}_{u,m} \neq \hat{h}_{u,m} \text{ and } \tilde{H}'(m,u,\delta/2)\right) + \delta/2 \\ &= \mathbb{P}\left(\tilde{h}_{u,m} \neq \hat{h}_{u,m} \text{ and } \tilde{H}'(m,u,\delta/2)\right) + \delta/2. \end{aligned} \tag{23}$$

Thus, it suffices to prove that, for an appropriate choice of $\tilde{c}_2$, for a value of u as in (17), there is an event $\tilde{H}''(m,u,\delta/2)$ of probability at least $1-\delta/2$, such that on $\tilde{H}''(m,u,\delta/2) \cap \tilde{H}'(m,u,\delta/2)$, $\tilde{h}_{u,m} = \hat{h}_{u,m}$.

We will prove this by induction. Define $\tilde{c}_2 = 4808\tilde{c}_1$, and fix a value of $u \in \mathbb{N}$ satisfying (17). As a base case, any call of $\tilde{\mathbb{A}}^{(\delta')}_{\mathcal{P},\mathbb{X}'_{1:u}(\mathcal{P})}(S')$ with $|S'| \leq c_0(d + \ln(12/\delta'))$ trivially has $\tilde{\mathbb{A}}^{(\delta')}_{\mathcal{P},\mathbb{X}'_{1:u}(\mathcal{P})}(S') = \hat{\mathbb{A}}^{(\delta')}_{\mathbb{X}'_{1:u}(\mathcal{P})}(S')$, since both algorithms return immediately in Step 1 (which is assumed to be identically executed in both algorithms), without any executions of Step 4. In particular, if $m \leq c_0(d + \ln(24/\delta))$, then this trivially completes the proof. Now for the nontrivial case, suppose $m > c_0(d + \ln(24/\delta))$.

Now take as an inductive hypothesis that for some $S' \subseteq (\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P})))$ and $\delta' \leq \delta/2$ with $|S'| > c_0(d+\ln(12/\delta'))$, $\tilde{\mathbb{A}}^{(\delta')}_{\mathcal{P},\mathbb{X}'_{1:u}(\mathcal{P})}(S')$ is called at some point in the execution of $\tilde{\mathbb{A}}^{(\delta/2)}_{\mathcal{P},\mathbb{X}'_{1:u}(\mathcal{P})}((\mathbb{X}_{1:m}(\mathcal{P}), f^\star(\mathbb{X}_{1:m}(\mathcal{P}))))$ (either the root call itself, or as a recursive call), and for S_i and $\hat{h}_i = \tilde{\mathbb{A}}^{(\delta'/2)}_{\mathcal{P},\mathbb{X}'_{1:u}(\mathcal{P})}(S_i)$ (for each $i \in \{1,2\}$) as in the definition of $\tilde{\mathbb{A}}^{(\delta')}_{\mathcal{P},\mathbb{X}'_{1:u}(\mathcal{P})}(S')$, there are events $\tilde{H}'''_i$ of probability at least $1-(\delta'/2)^2$ (for each $i \in \{1,2\}$), such that for each $i \in \{1,2\}$, on the event $\tilde{H}'''_i \cap \tilde{H}'(m,u,\delta/2)$, we have $\hat{\mathbb{A}}^{(\delta'/2)}_{\mathbb{X}'_{1:u}(\mathcal{P})}(S_i) = \hat{h}_i$.

Our objective in the inductive step is to prove that there is an event $\tilde{H}'''_0$ of probability at least $1-(\delta')^2$ such that, on $\tilde{H}'''_0 \cap \tilde{H}'(m,u,\delta/2)$, (19) is satisfied for this S' and δ'. Denote $\tilde{\nu}(|S'|,\delta') = \frac{25}{|S'|+1}\left(8d + 2\ln\left(\frac{12}{\delta'}\right)\right)$. Based on this definition, and the definition of $\tilde{\gamma}(|S'|,\delta')$, along with the facts that $e^7(d+\ln(12/\delta')) < |S'| \leq m$ and $0 < \delta' \leq \delta/2$, some simple algebra reveals that

$$\begin{aligned}
&\frac{\tilde{c}_1(\tilde{\nu}(|S'|,\delta') + \tilde{\gamma}(|S'|,\delta'))}{\tilde{\gamma}(|S'|,\delta')^2}\left(d\mathrm{Log}\left(\frac{1}{\tilde{\gamma}(|S'|,\delta')}\right) + \mathrm{Log}\left(\frac{2}{(\delta')^2}\right)\right) \\
&\leq (3\cdot 25\cdot 8+1)\tilde{c}_1\frac{1}{\tilde{\gamma}(|S'|,\delta')}\left(d\mathrm{Log}\left(\frac{1}{\tilde{\gamma}(|S'|,\delta')}\right) + \mathrm{Log}\left(\frac{2}{(\delta')^2}\right)\right) \\
&\leq 3(3\cdot 25\cdot 8+1)\tilde{c}_1(|S'|+1)\frac{d\mathrm{Log}\left(\frac{3(|S'|+1)}{d}\right) + \mathrm{Log}\left(\frac{2}{(\delta')^2}\right)}{d+\mathrm{Log}\left(\frac{12}{\delta'}\right)} \\
&\leq 8(3\cdot 25\cdot 8+1)\tilde{c}_1|S'|\frac{d\mathrm{Log}\left(\frac{|S'|}{d}\right) + \mathrm{Log}\left(\frac{12}{\delta'}\right)}{d+\mathrm{Log}\left(\frac{12}{\delta'}\right)} \\
&\leq \tilde{c}_2 m\frac{d\mathrm{Log}\left(\frac{m}{d}\right) + \mathrm{Log}\left(\frac{24}{\delta}\right)}{d+\mathrm{Log}\left(\frac{24}{\delta}\right)}.
\end{aligned}$$

Therefore, for u satisfying (17), an application of Lemma 5 implies that there is an event $\tilde{G}$ of probability at least $1-(\delta')^2/2$, on which, if $\exists h' \in \mathbb{C}[S']$ with $\mathcal{P}(x : h'(x) \neq \hat{h}_{\mathrm{Maj}}(x)) \leq \tilde{\nu}(|S'|,\delta')$, then

$$\mathcal{P}(x : \hat{h}'(x) \neq \hat{h}_{\mathrm{Maj}}(x)) \leq \inf_{h\in\mathbb{C}[S']} \mathcal{P}(x : h(x) \neq \hat{h}_{\mathrm{Maj}}(x)) + \tilde{\gamma}(|S'|,\delta').$$

That is, (19) is satisfied. In particular, note that since $f^\star \in \mathbb{C}[S']$, on the event $\tilde{H}'(m,u,\delta/2)$, (21) implies that $\mathcal{P}(x : f^\star(x) \neq \hat{h}_{\mathrm{Maj}}(x)) \leq \tilde{\nu}(|S'|,\delta')$, so that taking $h' = f^\star$ suffices. This means that on the event $\tilde{G} \cap \tilde{H}'''_1 \cap \tilde{H}'''_2 \cap \tilde{H}'(m,u,\delta/2)$, (19) is satisfied, so that $\hat{\mathbb{A}}^{(\delta')}_{\mathbb{X}'_{1:u}(\mathcal{P})}(S') = \tilde{\mathbb{A}}^{(\delta')}_{\mathcal{P},\mathbb{X}'_{1:u}(\mathcal{P})}(S')$. Defining $\tilde{H}'''_0 = \tilde{G} \cap \tilde{H}'''_1 \cap \tilde{H}'''_2$, and noting that, by the union bound, this event has probability at least $1-(\delta')^2$, this extends the inductive hypothesis.

By the principle of induction, we have established that there is an event $\tilde{H}''(m,u,\delta/2)$ of probability at least $1-(\delta/2)^2 \geq 1-\delta/2$ such that, on $\tilde{H}'(m,u,\delta/2) \cap \tilde{H}''(m,u,\delta/2)$, $\tilde{h}_{u,m} = \hat{h}_{u,m}$. Together with (23), this implies that $\mathbb{P}\left(\mathrm{er}_{\mathcal{P}}(\hat{h}_{u,m}; f^\star) > \varepsilon\right) \leq \delta$, which completes the proof. ∎

6. Closing Thoughts: Toward an Optimal Distribution-Independent Learning Algorithm

At this time, proving that the sample complexity of *distribution-independent* (and strictly supervised) PAC learning is $\Theta\left(\frac{1}{\varepsilon}\left(d + \operatorname{Log}\left(\frac{1}{\delta}\right)\right)\right)$ remains an open problem. In light of the present work, one natural route toward a possible solution would be to attempt to reduce the number of samples required to perform the projection step (Step 4) in the semi-supervised algorithm $\hat{\mathbb{A}}_{\mathcal{U}}^{(\delta/2)}(S)$ to a total of $O\left(\frac{1}{\varepsilon}\left(d + \operatorname{Log}\left(\frac{1}{\delta}\right)\right)\right)$, when we also have $|S| = O\left(\frac{1}{\varepsilon}\left(d + \operatorname{Log}\left(\frac{1}{\delta}\right)\right)\right)$, while preserving the fact that it instantiates Step 4 of $\mathbb{A}_{\mathcal{P}}^{(\delta/2)}(S)$ (with probability at least $1 - \delta/2$). In this case, we could simply use *labeled* samples to perform this step, thus obtaining a distribution-independent supervised learning algorithm achieving sample complexity $O\left(\frac{1}{\varepsilon}\left(d + \operatorname{Log}\left(\frac{1}{\delta}\right)\right)\right)$. At this time, it is unclear whether this number of samples suffices to supply the required guarantee in this projection step. This is closely related to a problem studied by Ben-David and Ben-David (2011), concerning the sample complexity of finding a classifier h in a given set $\mathcal{H}$, with $\mathcal{P}(x : h(x) \neq f(x)) \leq \inf_{h' \in \mathcal{H}} \mathcal{P}(x : h'(x) \neq f(x)) + \gamma$, for a *known* classifer f not in $\mathcal{H}$. As revealed by the proof of Theorem 6, our projection step can be viewed as a special case of a restriction of this problem, in which f is constrained to be within distance $O(\gamma)$ of $\mathcal{H} = \mathbb{C}[S]$. That said, it is conceivable that the particular details of our problem (e.g., f being a majority vote of three classifiers from $\mathbb{C}$) offer considerable advantages over the general case, which may further reduce the number of samples sufficient for this projection.

It is not clear (to the author) at this time whether the sample complexity guarantee from Theorem 2 would remain valid if Step 4 of $\mathbb{A}_{\mathcal{P}}^{(\delta)}$ were to instead simply return the majority classifier $x \mapsto \operatorname{Majority}(\hat{h}_1(x), \hat{h}_2(x), \hat{h}_3(x))$ (i.e., without projecting to $\mathbb{C}[S]$). This modification would remove all direct dependence on $\mathcal{P}$ from $\mathbb{A}_{\mathcal{P}}^{(\delta)}$. That algorithm appears to have interesting behavior on several example scenarios, but if a general analysis is possible, it seems to require additional insights beyond the conditional error analysis used here.

More broadly, even aside from considering modifications of the algorithm proposed above, Theorem 2 would establish optimal sample complexity bounds for distribution-independent PAC learning, given the validity of another open conjecture in the learning theory literature. Specifically, Ben-David, Lu, and Pál (2008) have conjectured that the sample complexity achieved by any distribution-dependent learning algorithm is always at most a *constant factor* smaller than the sample complexity achieved by some corresponding distribution-independent learning algorithm; that conjecture even claims this is true when the sample complexity itself is allowed to depend on the data distribution. If correct, this would be a surprisingly strong (and beautiful) result. If we interpret the conjecture to mean that the constant factor is invariant to the choice of concept space, then this general conjecture also implies a much weaker one:[8]

8. This conjecture has been investigated to a small extent in the literature. For instance, in the case of threshold classifiers, an analysis of Ben-David, Lu, and Pál (2008) indicates that the constant factor difference between optimal sample complexities for distribution-independent vs distribution-dependent learning is asymptotically at most 2, and have conjectured that this is indeed the correct value. For this same concept space, Helmbold and Long (2012) have argued that (in the closely-related *prediction model* of learning) this constant factor difference is at least $\frac{6}{5}$, implying that there actually is a slight advantage to allowing distribution-dependence in the learning algorithm.

The worst-case (over both the target and distribution) sample complexity achievable by any distribution-dependent PAC learning algorithm is at most a numerical constant factor smaller than the best worst-case sample complexity achievable by distribution-independent PAC learning algorithms.

If this weaker conjecture is valid, then Theorem 2 (together with existing lower bounds) would also establish a precise expression for the sample complexity of distribution-independent PAC learning (up to a constant factor). Indeed, it follows from Corollary 3 that this conjecture is in fact *equivalent* to the well-known conjecture that the optimal sample complexity of distribution-independent PAC learning is $\Theta\left(\frac{1}{\varepsilon}\left(d+\operatorname{Log}\left(\frac{1}{\delta}\right)\right)\right)$.

Appendix A. A Technical Lemma

The following basic lemma is useful in the proof of Theorem 2.[9]

Lemma 8 *For any $a, b, c_1 \in [1, \infty)$ and $c_2 \in [0, \infty)$,*

$$a \ln\left(c_1\left(c_2 + \frac{b}{a}\right)\right) \leq a \ln\left(c_1(c_2 + e)\right) + \frac{1}{e}b.$$

Proof If $\frac{b}{a} \leq e$, then monotonicity of $\ln(\cdot)$ implies

$$a \ln\left(c_1\left(c_2 + \frac{b}{a}\right)\right) \leq a \ln(c_1(c_2 + e)),$$

which is clearly no greater than $a \ln(c_1(c_2 + e)) + \frac{1}{e}b$.

On the other hand, if $\frac{b}{a} > e$, then

$$a \ln\left(c_1\left(c_2 + \frac{b}{a}\right)\right) \leq a \ln\left(c_1 \max\{c_2, 2\}\frac{b}{a}\right) = a \ln\left(c_1 \max\{c_2, 2\}\right) + a \ln\left(\frac{b}{a}\right).$$

The first term in the rightmost expression is at most $a \ln(c_1(c_2+2)) \leq a \ln(c_1(c_2+e))$. The second term in the rightmost expression can be rewritten as $b\frac{\ln(b/a)}{b/a}$. Since $x \mapsto \ln(x)/x$ is nonincreasing on (e, ∞), in the case $\frac{b}{a} > e$, this is at most $\frac{1}{e}b$. Together, we have that

$$a \ln\left(c_1\left(c_2 + \frac{b}{a}\right)\right) \leq a \ln(c_1(c_2 + e)) + \frac{1}{e}b$$

in this case as well. ■

References

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. 4.3

9. This lemma and proof also appear in a sibling paper (Hanneke, 2015).

P. Auer and R. Ortner. A new PAC bound for intersection-closed concept classes. *Machine Learning*, 66(2-3):151–163, 2007. 3

M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of the* 26$^{\text{th}}$ *Conference on Learning Theory*, 2013. 3

S. Ben-David and S. Ben-David. Learning a classifier when the labeling is known. In *Proceedings of the* 22$^{\text{nd}}$ *International Conference on Algorithmic Learning Theory*, 2011. 6

S. Ben-David, T. Lu, and D. Pál. Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning. In *Proceedings of the* 21$^{\text{st}}$ *Conference on Learning Theory*, 2008. 1, 6, 8

G. M. Benedek and A. Itai. Learnability by fixed distributions. *Theoretical Computer Science*, 86:377–389, 1991. 2, 3

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4): 929–965, 1989. 1, 2, 3, 3, 4.3

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. *Lecture Notes in Artificial Intelligence*, 3176:169–207, 2004. 5, 5

N. H. Bshouty, Y. Li, and P. M. Long. Using the doubling dimension to analyze the generalization of learning algorithms. *Journal of Computer and System Sciences*, 75(6): 323–335, 2009. 3

T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14 (3):326–334, 1965. 3

M. Darnstädt. The optimal PAC bound for intersection-closed concept classes. *Information Processing Letters*, 115(4):458–461, 2015. 3

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, Inc., 1996. 2

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989. 1, 3

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006. 3

S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2009. 3, 4.1

S. Hanneke. Refined error bounds for several learning algorithms. *Manuscript*, 2015. 3, 9

D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115:248–292, 1994. 3, 3

D. P. Helmbold and P. M. Long. New bounds for learning intervals with implications for semi-supervised learning. In *Proceedings of the* 25^{th} *Conference on Learning Theory*, 2012. 8

S. R. Kulkarni. On metric entropy, Vapnik-Chervonenkis dimension, and learnability for a class of distributions. Technical Report CICS-P-160, Center for Intelligent Control Systems, 1989. 2

P. M. Long. An upper bound on the sample complexity of PAC learning halfspaces with respect to the uniform distribution. *Information Processing Letters*, 87(5):229–234, 2003. 3

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (A)*, 13: 145–147, 1972. 5

H. Simon. An almost optimal PAC algorithm. In *Proceedings of the* 28^{th} *Conference on Learning Theory*, 2015. 1, 3, 4.1

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984. 1

A. van der Vaart and J. A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011. 2

V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982. 3, 4.3, 5, 5

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971. 2, 5

M. Vidyasagar. *Learning and Generalization with Applications to Neural Networks*. Springer-Verlag, 2^{nd} edition, 2003. 5, 5

M. Warmuth. The optimal PAC algorithm. In *Proceedings of the* 17^{th} *Conference on Learning Theory*, 2004. 1