The Information-Theoretic Value of Unlabeled Data in Semi-Supervised Learning

Alexander Golovnev [1] **Dávid Pál** [2] **Balázs Szörényi** [2]

Abstract

We quantify the separation between the numbers of labeled examples required to learn in two settings: Settings *with* and *without* the knowledge of the distribution of the unlabeled data. More specifically, we prove a separation by $\Theta(\log n)$ multiplicative factor for the class of projections over the Boolean hypercube of dimension n. We prove that there is no separation for the class of all functions on domain of any size. Learning with the knowledge of the distribution (a.k.a. *fixed-distribution learning*) can be viewed as an idealized scenario of semi-supervised learning where the number of unlabeled data points is so great that the unlabeled distribution is known exactly. For this reason, we call the separation the *value of unlabeled data*.

1. Introduction

Hanneke (2016) showed that for any class C of Vapnik-Chervonenkis dimension d there exists an algorithm that ϵ-learns any target function from C under any distribution from $O\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$ labeled examples with probability at least $1-\delta$. For this paper, it is important to stress that Hanneke's algorithm does *not* receive the distribution of unlabeled data as input. On the other hand, Benedek & Itai (1991) showed that for any class C and any distribution there exists an algorithm that ϵ-learns any target from C from $O\left(\frac{\log N_{\epsilon/2}+\log(1/\delta)}{\epsilon}\right)$ labeled examples with probability at least $1-\delta$ where $N_{\epsilon/2}$ is the size of an $\frac{\epsilon}{2}$-cover of C with respect to the disagreement metric $d(f,g) = \Pr[f(x) \neq g(x)]$. Here, it is important to note that Benedek and Itai construct for each distribution a separate algorithm. In other words, they construct a family of algorithms indexed by the (uncountably many) distributions over the domain. Alternatively, we can think of Benedek-Itai's family of algorithms as a single algorithm that receives the distribution as an input. It is known that $N_\epsilon = O(1/\epsilon)^{O(d)}$; see Dudley (1978). Thus, ignoring $\log(1/\epsilon)$ factor, Benedek-Itai bound is never worse than Hanneke's bound.

As we already mentioned, Benedek-Itai's algorithm receives as input the distribution of unlabeled data. The algorithm uses it to construct an $\frac{\epsilon}{2}$-cover. Unsurprisingly, there exist distributions which have a small $\frac{\epsilon}{2}$-cover and thus sample complexity of Benedek-Itai's algorithm on such distributions is significantly lower then the Hanneke's bound. For instance, a distribution concentrated on a single point has an $\frac{\epsilon}{2}$-cover of size 2 for any positive ϵ.

However, an algorithm does not need to receive the unlabeled distribution in order to enjoy low sample complexity. For example, empirical risk minimization (ERM) algorithm needs significantly less labeled examples to learn any target under some unlabeled distributions. For instance, if the distribution is concentrated on a single point, ERM needs only one labeled example to learn any target. One could be lead to believe that there exists an algorithm that does *not* receive the unlabeled distribution as input and achieves Benedek-Itai bound (or a slightly worse bound) for *every* distribution. In fact, one could think that ERM or Hanneke's algorithm could be such algorithms. If ERM, Hanneke's algorithm, or some other distribution-independent algorithm had sample complexity that matches (or nearly matches) the optimal distribution-specific sample complexity for *every* distribution, we could conclude that the knowledge of unlabeled data distribution is completely useless.

As Darnstädt et al. (2013) showed this is not the case. They showed that *any* algorithm for learning projections over $\{0,1\}^n$ that does not receive the unlabeled distribution as input, requires, for some data unlabeled distributions, more labeled examples than the Benedek-Itai bound. However, they did not quantify this gap beside stating that it grows without bound as n goes to infinity.

In this paper, we quantify the gap by showing that *any* distribution-independent algorithm for learning the class

[1]Harvard University, Cambridge, MA, USA [2]Yahoo Research, New York, NY, USA. Correspondence to: Dávid Pál <davidko.pal@gmail.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

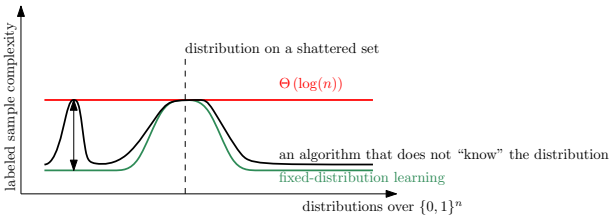

Figure 1. The graph shows sample complexity bounds of learning a class of projections over the domain $\{0,1\}^n$ under various unlabeled distributions. We assume that ϵ and δ are constant, say, $\epsilon = \delta = \frac{1}{100}$. The graph shows three lines. The red horizontal line is Hanneke's bound for the class of projections, which is $\Theta(\mathrm{VC}(C_n)) = \Theta(\log n)$. The green line is the Benedek-Itai bound. The green line touches the red line for certain distributions, but is lower for other distributions. In particular, for certain distributions the green line is $O(1)$. The dashed line corresponds to a particular distribution on a shattered set. This is where the green line and red line touch. Furthermore, here the upper bound coincides with the lower bound for that particular distribution. The black line is the sample complexity of an arbitrary *distribution-independent* algorithm. For example, the reader can think of the ERM or Hanneke's algorithm. We prove that there exist a distribution where the black line is $\Omega(\log n)$ times higher than the green line. This separation is indicated by the double arrow.

of projections over $\{0,1\}^n$ requires, for some unlabeled distributions, $\Omega(\log n)$ times as many labeled examples as Benedek-Itai bound. Darnstädt et al. (2013) showed the gap for any class with Vapnik-Chervonenkis dimension d is at most $O(d)$. It is well known that Vapnik-Chervonenkis dimensions of projections over $\{0,1\}^n$ is $\Theta(\log n)$. Thus our lower bound matches the upper bound $O(d)$. To better understand the relationship of the upper and lower bounds, we illustrate the situation for the class of projections over $\{0,1\}^n$ in Figure 1.

In contrast, we show that for the class of *all* functions (on any domain) there is no gap between the two settings. In other words, for learning a target from the class of all functions, unlabeled data are in fact useless. This illustrates the point that the gap depends in a non-trivial way on the combinatorial structure of the function class rather than just on the Vapnik-Chervonenkis dimension.

The paper is organized as follows. In Section 2 we review prior work. Section 3 gives the necessary definitions and basic probabilistic tools. In Section 4 we give the proof of the separation result for projections. In Section 5 we prove that there is no gap for the class of all functions. For completeness, in the supplementary material, we give a proof of a simple upper bound $O(1/\epsilon)^{O(d)}$ on the size of the minimum ϵ-cover and a proof of Benedek-Itai's $O\left(\frac{\log N_{\epsilon/2}+\log(1/\delta)}{\epsilon}\right)$ sample complexity upper bound.

2. Related Work

The question of whether knowledge of unlabeled data distribution helps was proposed and initially studied by Ben-David et al. (2008); see also Lu (2009). However, they considered only classes with Vapnik-Chervonenkis dimension at most 1, or classes with Vapnik-Chervonenkis dimension d but only distributions for which the size of the ϵ-cover is $\Theta(1/\epsilon)^{\Theta(d)}$, i.e. the ϵ-cover is as large as it can be.[1] In these settings, for constant ϵ and δ, the separation of labeled sample complexities is at most a constant factor, which is exactly what Ben-David et al. (2008) proved. In these settings, unlabeled data are indeed useless. However, these results say nothing about distributions with ϵ-cover of small size and it ignores the dependency on the Vapnik-Chervonenkis dimension.

The question was studied in earnest by Darnstädt et al. (2013) who showed two major results. First, they show that for any non-trivial concept class C and for every distribution, the ratio of the labeled sample complexities between distribution-independent and distribution-dependent algorithms is bounded by the Vapnik-Chervonenkis dimension. Second, they show that for the class of projections over $\{0,1\}^n$, there are distributions where the ratio grows to infinity as a function of n.

In learning theory, the disagreement metric and ϵ-cover were introduced by Benedek & Itai (1991) but the ideas are much older; see e.g. Dudley (1978; 1984). The $O(1/\epsilon)^{O(d)}$ upper bound on size of the smallest ϵ-cover is by Dudley (1978, Lemma 7.13); see also Devroye & Lugosi (2000, Chapter 4) and Haussler (1995). We present a proof of $O(1/\epsilon)^{O(d)}$ upper bound in Appendix A in the supplementary material.

For any distribution-independent algorithm and any class C of Vapnik-Chervonenkis dimension $d \geq 2$ and any $\epsilon \in (0,1)$ and $\delta \in (0,1)$, there exists a distribution over the domain and a concept which requires at least $\Omega\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$ labeled examples to ϵ-learn with probability at least $1-\delta$; see Anthony & Bartlett (1999, Theorem 5.3) and Blumer et al. (1989); Ehrenfeucht et al. (1989). The proof of the lower bound constructs a distribution that does *not* depend on the algorithm. The distribution is a particular distribution over a fixed set shattered by C. So even an algorithm that knows the distribution requires $\Omega\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$ labeled examples.

[1] For any concept class with Vapnik-Chervonenkis dimension d and any distribution, the size of the smallest ϵ-cover is at most $O(1/\epsilon)^{O(d)}$.

3. Preliminaries

Let $\mathcal{X}$ be a non-empty set. We denote by $\{0,1\}^{\mathcal{X}}$ the class of all functions from $\mathcal{X}$ to $\{0,1\}$. A *concept class over a domain* $\mathcal{X}$ is a subset $C \subseteq \{0,1\}^{\mathcal{X}}$. A *labeled example* is a pair $(x,y) \in \mathcal{X} \times \{0,1\}$.

A *distribution-independent learning algorithm* is a function $A : \bigcup_{m=0}^{\infty} (\mathcal{X} \times \{0,1\})^m \to \{0,1\}^{\mathcal{X}}$. In other words, the algorithm gets as input a sequence of labeled examples $(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)$ and outputs a function from $\mathcal{X}$ to $\{0,1\}$. We allow the algorithm to output a function that does not belong to C, i.e., the algorithm can be improper. A *distribution-dependent algorithm* is a function that maps any probability distribution over $\mathcal{X}$ to a distribution-independent algorithm.

Let P be a probability distribution over a domain $\mathcal{X}$. For any two functions $f : \mathcal{X} \to \{0,1\}$, $g : \mathcal{X} \to \{0,1\}$ we define the disagreement pseudo-metric

$$d_P(f,g) = \Pr_{X \sim P}[f(X) \neq g(X)] \; .$$

Let C be a concept class over $\mathcal{X}$, let $c \in C$, let $\epsilon, \delta \in (0,1)$. Let $X_1, X_2, \ldots, X_m$ be an i.i.d. sample from P. We define the corresponding labeled sample $T = ((X_1, c(X_1)), (X_2, c(X_2)), \ldots, (X_m, c(X_m)))$. We say that an algorithm A, *ϵ-learns* target c from m samples with probability at least $1-\delta$ if

$$\Pr\left[d_P(c, A(T)) \leq \epsilon\right] \geq 1 - \delta \; .$$

The smallest non-negative integer m such that for any target $c \in C$, the algorithm A, ϵ-learns the target c from m samples with probability at least $1-\delta$ is denoted by $m(A, C, P, \epsilon, \delta)$.

We recall the standard definitions from learning theory. For any concept $c : \mathcal{X} \to \{0,1\}$ and any $S \subseteq \mathcal{X}$ we define $\pi(c,S) = \{x \in S \; : \; c(x) = 1\}$. In other words, $\pi(c,S)$ is the set of examples in S which c labels 1. A set $S \subseteq \mathcal{X}$ is *shattered* by a concept class C if for any subset $S' \subseteq S$ there exists a classifier $c \in C$ such that $\pi(c,S) = S'$. *Vapnik-Chervonenkis dimension* of a concept class C is the size of the largest set $S \subseteq \mathcal{X}$ shattered by C. A subset C' of a concept class C is an *ϵ-cover* of C for a probability distribution P if for any $c \in C$ there exists $c' \in C'$ such that $d_P(c,c') \leq \epsilon$.

To prove our lower bounds we need three general probabilistic results. The first one is the standard Hoeffding bound. The other two are simple and intuitive propositions. The first proposition says that if average error $d_P(c, A(T))$ is high, the algorithm fails to ϵ-learn with high probability. The second proposition says that the best algorithm for predicting a bit based on some side information, is to compute conditional expectation of the bit and thresholds it at $1/2$.

Theorem 1 (Hoeffding bound)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables that lie in interval $[a,b]$ with probability one and let $p = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[X_i]$. Then, for any $t \geq 0$,*

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq p + t\right] \leq e^{-2nt^2/(a-b)^2} \; ,$$

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \leq p - t\right] \leq e^{-2nt^2/(a-b)^2} \; .$$

Proposition 2 (Error probability vs. Expected error)**.** *Let Z be a random variable such that $Z \leq 1$ with probability one. Then,*

$$\Pr[Z > t] \geq \frac{\mathbf{E}[Z] - t}{1 - t} \qquad \textit{for any } t \in [0,1).$$

Proof. We have

$$\begin{aligned} \mathbf{E}[Z] &\leq t \cdot \Pr[Z \leq t] + 1 \cdot \Pr[Z > t] \\ &= t \cdot (1 - \Pr[Z > t]) + \Pr[Z > t] \; . \end{aligned}$$

Solving for $\Pr[Z > t]$ finishes the proof. □

Proposition 3 (Predicting Single Bit)**.** *Let $\mathcal{U}$ be a finite non-empty set. Let U, V be random variables (possibly correlated) such that $U \in \mathcal{U}$ and $V \in \{0,1\}$ with probability one. Let $f : \mathcal{U} \to \{0,1\}$ be a predictor. Then,*

$$\Pr\left[f(U) \neq V\right] \geq \sum_{u \in \mathcal{U}} \left(\frac{1}{2} - \left|\frac{1}{2} - \mathbf{E}\left[V \,|\, U = u\right]\right|\right) \cdot \Pr[U = u] \; .$$

Proof. We have

$$\Pr\left[f(U) \neq V\right] = \sum_{u \in \mathcal{U}} \Pr\left[f(U) \neq V \,|\, U = u\right] \cdot \Pr[U = u] \; .$$

It remains to show that

$$\Pr\left[f(U) \neq V \,|\, U = u\right] \geq \frac{1}{2} - \left|\frac{1}{2} - \mathbf{E}\left[V \,|\, U = u\right]\right| \; .$$

Since if $U = u$, the value $f(U) = f(u)$ is fixed, and hence

$$\begin{aligned} &\Pr\left[f(U) \neq V \,|\, U = u\right] \\ &\quad \geq \min\left\{\Pr\left[V = 1 \,|\, U = u\right], \; \Pr\left[V = 0 \,|\, U = u\right]\right\} \\ &\quad = \min\left\{\mathbf{E}\left[V \,|\, U = u\right], \; 1 - \mathbf{E}\left[V \,|\, U = u\right]\right\} \\ &\quad = \frac{1}{2} - \left|\frac{1}{2} - \mathbf{E}\left[V \,|\, U = u\right]\right| \end{aligned}$$

We used the fact that $\min\{x, 1-x\} = \frac{1}{2} - \left|\frac{1}{2} - x\right|$ for all $x \in \mathbb{R}$ which can be easily verified by considering two cases: $x \geq \frac{1}{2}$ and $x < \frac{1}{2}$. □

4. Projections

In this section, we denote by C_n the class of *projections* over the domain $\mathcal{X} = \{0,1\}^n$. The class C_n consists of n functions $c_1, c_2, \ldots, c_n$ from $\{0,1\}^n$ to $\{0,1\}$. For any $i \in \{1,2,\ldots,n\}$, for any $x \in \{0,1\}^n$, the function c_i is defined as $c_i((x[1], x[2], \ldots, x[n])) = x[i]$.

For any $\epsilon \in (0, \frac{1}{2})$ and $n \geq 2$, we consider a family $\mathcal{P}_{n,\epsilon}$ consisting of n probability distributions $P_1, P_2, \ldots, P_n$ over the Boolean hypercube $\{0,1\}^n$. In order to describe the distribution P_i, for some i, consider a random vector $X = (X[1], X[2], \ldots, X[n])$ drawn from P_i. The distribution P_i is a product distribution, i.e., $\Pr[X = x] = \prod_{j=1}^{n} \Pr[X[j] = x[j]]$ for any $x \in \{0,1\}^n$. The marginal distributions of the coordinates are

$$\Pr[X[j] = 1] = \begin{cases} \frac{1}{2} & \text{if } j = i, \\ \epsilon & \text{if } j \neq i, \end{cases} \qquad \text{for } j = 1, 2, \ldots, n.$$

The reader should think of ϵ as a constant that does not depend on n, say, $\epsilon = \frac{1}{100}$.

The following result is folklore. We include its proof for completeness.

Proposition 4. *Vapnik-Chervonenkis dimension of C_n is $\lfloor \log_2 n \rfloor$.*

Proof. Let us denote the Vapnik-Chervonenkis dimension by d. Recall that d is the size of the largest shattered set. Let S be any shattered set of size d. Then, there must be at least 2^d distinct functions in C_n. Hence, $d \leq \log_2 |C_n| = \log_2 n$. Since d is an integer, we conclude that $d \leq \lfloor \log_2 n \rfloor$.

On the other hand, we construct a shattered set of size $\lfloor \log_2 n \rfloor$. The set will consists of points $x_1, x_2, \ldots, x_{\lfloor \log_2 n \rfloor} \in \{0,1\}^n$. For any $i \in \{1,2,\ldots,\lfloor \log_2 n \rfloor\}$ and any $j \in \{0,1,2,\ldots,n-1\}$, we define $x_i[j]$ to be the i-th bit in the binary representation of the number j. (The bit at position $i = 1$ is the least significant bit.) It is not hard to see that for any $v \in \{0,1\}^{\lfloor \log_2 n \rfloor}$, there exists $c \in C_n$ such that $v = (c(x_1), c(x_2), \ldots, c(x_{\lfloor \log_2 n \rfloor}))$. Indeed, given v, let $k \in \{0, 1, \ldots, 2^{\lfloor \log_2 n \rfloor} - 1\}$ be the number with binary representation v, then we can take $c = c_{k+1}$. □

Lemma 5 (Small cover)**.** *Let $n \geq 2$ and $\epsilon \in (0, \frac{1}{2})$. Any distribution in $\mathcal{P}_{n,\epsilon}$ has 2ϵ-cover of size 2.*

Proof. Consider a distribution $P_i \in \mathcal{P}_{n,\epsilon}$ for some $i \in \{1,2,\ldots,n\}$. Let j be an arbitrary index in $\{1,2,\ldots,n\} \setminus \{i\}$. Consider the projections $c_i, c_j \in C_n$. We claim that $C' = \{c_i, c_j\}$ is a 2ϵ-cover of C_n.

To see that C' is a 2ϵ-cover of C_n, consider any $c_k \in C_n$. We need to show that $d_{P_i}(c_i, c_k) \leq 2\epsilon$ or $d_{P_i}(c_j, c_k) \leq 2\epsilon$. If $k = i$ or $k = j$, the condition is trivially satisfied. Consider $k \in \{1,2,\ldots,n\} \setminus \{i,j\}$. Let $X \sim P_i$. Then,

$$\begin{aligned} d_{P_i}(c_j, c_k) &= \Pr[c_j(X) \neq c_k(X)] \\ &= \Pr[c_j(X) = 1 \wedge c_k(X) = 0] \\ &\quad + \Pr[c_j(X) = 0 \wedge c_k(X) = 1] \\ &= \Pr[X[j] = 1 \wedge X[k] = 0] \\ &\quad + \Pr[X[j] = 0 \wedge X[k] = 1] \\ &= \Pr[X[j] = 1] \Pr[X[k] = 0] \\ &\quad + \Pr[X[j] = 0] \Pr[X[k] = 1] \\ &= 2\epsilon\,(1 - \epsilon) \\ &< 2\epsilon\,. \end{aligned}$$

□

Using Benedek-Itai bound (see Theorem 12 in Appendix B in the supplementary material) we obtain the corollary below. The corollary states that the distribution-dependent sample complexity of learning target in C_n under any distribution from $P_{n,\epsilon}$ does *not* depend on n.

Corollary 6 (Learning with knowledge of the distribution)**.** *Let $n \geq 2$ and $\epsilon \in (0, \frac{1}{2})$. There exists a distribution-dependent algorithm such that for any distribution from $\mathcal{P}_{n,\epsilon}$, any $\delta \in (0,1)$, any target function $c \in C_n$, if the algorithm gets*

$$m \geq \frac{12 \ln(2/\delta)}{\epsilon}$$

labeled examples, with probability at least $1 - \delta$, it 4ϵ-learns the target.

The next theorem states that without knowing the distribution, learning a target under a distribution from $\mathcal{P}_{n,\epsilon}$ requires at least $\Omega(\log n)$ labeled examples. It is important to note that ϵ in this bound is the parameter of the distribution, and not the accuracy of the PAC learning model.

Theorem 7 (Learning without knowledge of the distribution)**.** *For any distribution-independent algorithm, any $\epsilon \in (0, \frac{1}{4})$ and any $n \geq 600/\epsilon^3$ there exists a distribution $P \in \mathcal{P}_{n,\epsilon}$ and a target concept $c \in C_n$ such that if the algorithm gets*

$$m \leq \frac{\ln n}{3 \ln(1/\epsilon)}$$

labeled examples, it fails to $\frac{1}{16}$-learn the target concept with probability more than $\frac{1}{16}$.

The main idea of the proof is the following. Assume that the learner is restricted to output some function that belongs to C_n (i.e., the learner is *proper*). Then with high probability, the number of coordinates that coincide with the target on a random sample is $\Omega(\epsilon n)$, and, thus, the number of projections that output the same value on each of the m

random samples is $\Omega(\epsilon^m n)$. Therefore, with high probability, at least one other projection produces the exact same output as the target. In this case, the learner has to choose randomly, and the probability of choosing a wrong answer is at least $1/2$. This implies that the learner must see at least $m \geq \Omega(\frac{\ln n}{\ln(1/\epsilon)})$ samples. In the proof below we make this intuition formal, and generalize it to the case of improper learners, too.

Proof of Theorem 7. Let A be any learning algorithm. For ease of notation, we formalize it is a function

$$A : \bigcup_{m=0}^{\infty} \left(\{0,1\}^{m\times n} \times \{0,1\}^m\right) \to \{0,1\}^{\{0,1\}^n} .$$

The algorithm receives an $m \times n$ matrix and a binary vector of length m. The rows of the matrix corresponds to unlabeled examples and the vector encodes the labels. The output of A is any function from $\{0,1\}^n \to \{0,1\}$.

We demonstrate the existence of a pair $(P, c) \in \mathcal{P}_{n,\epsilon} \times C_n$ which cannot be learned with m samples by the probabilistic method. Let I be chosen uniformly at random from $\{1, 2, \ldots, n\}$. We consider the distribution $P_I \in \mathcal{P}_{n,\epsilon}$ and target $c_I \in C_n$. Let $X_1, X_2, \ldots, X_m$ be an i.i.d. sample from P_I and let $Y_1 = c_I(X_1), Y_2 = c_I(X_2), \ldots, Y_m = c_I(X_m)$ be the target labels. Let X be the $m \times n$ matrix with entries $X_i[j]$ and let $Y = (Y_1, Y_2, \ldots, Y_m)$ be the vector of labels. The output of the algorithm is $A(X, Y)$. We will show that

$$\mathbf{E}\left[d_{P_I}(c_I, A(X,Y))\right] \geq \frac{1}{8} . \tag{1}$$

This means that there exists $i \in \{1, 2, \ldots, n\}$ such that

$$\mathbf{E}\left[d_{P_i}(c_i, A(X,Y)) \mid I = i\right] \geq \frac{1}{8} .$$

By Proposition 2,

$$\Pr\left[d_{P_i}(c_i, A(X,Y)) > \frac{1}{16} \,\middle|\, I = i\right] \geq \frac{\frac{1}{8} - \frac{1}{16}}{1 - \frac{1}{16}} > \frac{1}{16} .$$

It remains to prove (1). Let Z be a test sample drawn from P_I. That is, conditioned on I, the sequence $X_1, X_2, \ldots, X_m, Z$ is i.i.d. drawn from P_I. Then, by Proposition 3,

$$\mathbf{E}\left[d_{P_I}(c_I, A(X,Y))\right] = \Pr\left[A(X,Y)(Z) \neq c_I(Z)\right] \geq \sum_{\substack{x\in\{0,1\}^{m\times n}\\ y\in\{0,1\}^m\\ z\in\{0,1\}^n}} \left(\frac{1}{2} - \left|\frac{1}{2} - \mathbf{E}\left[c_I(Z) \mid X = x, Y = y, Z = z\right]\right|\right) \cdot \Pr\left[X = x, Y = y, Z = z\right] . \tag{2}$$

We need to compute $\mathbf{E}\left[c_I(Z) \mid X = x, Y = y, Z = z\right]$. For that we need some additional notation. For any matrix $x \in \{0,1\}^{m\times n}$, let $x[1], x[2], \ldots, x[n]$ be its columns. For any matrix $x \in \{0,1\}^{m\times n}$ and vector $y \in \{0,1\}^m$ let

$$k(x, y) = \{i \in \{1, 2, \ldots, n\} \ : \ x[i] = y\}$$

be the set of indices of columns of x equal to the vector y. Also, we define $\|\cdot\|$ to be the sum of absolute values of entries of a vector or a matrix. (Since we use $\|\cdot\|$ only for binary matrices and binary vectors, it will be just the number of ones.)

For any $i \in \{1, 2, \ldots, n\}$,

$$\Pr\left[I = i, X = x, Y = y\right] = \begin{cases} \frac{1}{n}\left(\frac{1}{2}\right)^m \epsilon^{\|x\|-\|y\|}(1-\epsilon)^{mn-\|x\|+\|y\|} & \text{if } i \in k(x,y), \\ 0 & \text{if } i \notin k(x,y). \end{cases}$$

Therefore, for any $i \in \{1, 2, \ldots, n\}$,

$$\begin{aligned} &\Pr\left[I = i \mid X = x, Y = y\right] \\ &\quad = \frac{\Pr\left[I = i, X = x, Y = y\right]}{\Pr\left[X = x, Y = y\right]} \\ &\quad = \frac{\Pr\left[I = i, X = x, Y = y\right]}{\sum_{j\in k(x,y)} \Pr\left[I = j, X = x, Y = y\right]} \\ &\quad = \begin{cases} \frac{1}{|k(x,y)|} & \text{if } i \in k(x,y), \\ 0 & \text{if } i \notin k(x,y). \end{cases} \end{aligned}$$

Conditioned on I, the variables Z and (X, Y) are independent. Thus, for any $x \in \{0,1\}^n$, and $i = 1, 2, \ldots, n$,

$$\begin{aligned} &\Pr\left[Z = z \mid I = i, X = x, Y = y\right] \\ &\quad = \Pr\left[Z = z \mid I = i\right] \\ &\quad = \begin{cases} \frac{1}{2}\epsilon^{\|z\|-1}(1-\epsilon)^{n-\|z\|} & \text{if } z[i] = 1, \\ \frac{1}{2}\epsilon^{\|z\|}(1-\epsilon)^{n-1-\|z\|} & \text{if } z[i] = 0. \end{cases} \end{aligned}$$

This allows us to compute the conditional probability

$$\begin{aligned} &\Pr\left[I = i, Z = z \mid X = x, Y = y\right] \\ &\quad = \Pr\left[Z = z \mid I = i, X = x, Y = y\right] \\ &\qquad\qquad \cdot \Pr\left[I = i \mid X = x, Y = y\right] \\ &\quad = \begin{cases} \frac{\epsilon^{\|z\|-1}(1-\epsilon)^{n-\|z\|}}{2|k(x,y)|} & \text{if } i \in k(x,y) \text{ and } z[i] = 1, \\ \frac{\epsilon^{\|z\|}(1-\epsilon)^{n-1-\|z\|}}{2|k(x,y)|} & \text{if } i \in k(x,y) \text{ and } z[i] = 0, \\ 0 & \text{if } i \notin k(x,y). \end{cases} \end{aligned}$$

For any $z \in \{0,1\}^n$, let

$$s(x, y, z) = \{i \in k(x, y) \ : \ z[i] = 1\} ,$$

and note that $s(x, y, z) \subseteq k(x, y)$. Then,

$$\Pr\left[Z = z \mid X = x, Y = y\right]$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \Pr\left[Z = z, I = i \,|\, X = x, Y = y\right] \\
&= \sum_{i \in k(x,y)} \Pr\left[Z = z, I = i \,|\, X = x, Y = y\right] \\
&= \sum_{i \in s(x,y,x)} \Pr\left[Z = z, I = i \,|\, X = x, Y = y\right] \\
&\quad + \sum_{i \in k(x,y) \setminus s(x,y,z)} \Pr\left[Z = z, I = i \,|\, X = x, Y = y\right] \\
&= \frac{1}{2|k(x,y)|} \cdot |s(x,y,z)| \cdot \epsilon^{\|z\|-1}(1-\epsilon)^{n-\|z\|} \\
&\quad + \frac{(|k(x,y)| - |s(x,y,z)|) \cdot \epsilon^{\|z\|}(1-\epsilon)^{n-1-\|z\|}}{2|k(x,y)|} \\
&= \frac{\epsilon^{\|z\|-1}(1-\epsilon)^{n-1-\|z\|}}{2|k(x,y)|} \\
&\qquad \cdot \left(|s(x,y,z)| \cdot (1-2\epsilon) + |k(x,y)| \cdot \epsilon\right) .
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\mathbf{E}\left[c_I(Z) \,|\, X = x, Y = y, Z = z\right] \\
&\quad = \Pr\left[Z[I] = 1 \,|\, X = x, Y = y, Z = z\right] \\
&\quad = \frac{\Pr\left[Z[I] = 1, Z = z \,|\, X = x, Y = y\right]}{\Pr\left[Z = z \,|\, X = x, Y = y\right]} \\
&\quad = \frac{\sum_{i=1}^{n} \Pr\left[I = i, Z[i] = 1, Z = z \,|\, X = x, Y = y\right]}{\Pr\left[Z = z \,|\, X = x, Y = y\right]} \\
&\quad = \frac{\frac{|s(x,y,z)|}{2|k(x,y)|} \cdot \epsilon^{\|z\|-1}(1-\epsilon)^{n-\|z\|}}{\frac{\epsilon^{\|z\|-1}(1-\epsilon)^{n-1-\|z\|}}{2|k(x,y)|}} \\
&\qquad \cdot \frac{1}{(|s(x,y,z)| \cdot (1-2\epsilon) + |k(x,y,z)| \cdot \epsilon)} \\
&\quad = \frac{|s(x,y,z)| \cdot (1-\epsilon)}{|s(x,y,z)| \cdot (1-2\epsilon) + |k(x,y)| \cdot \epsilon} \\
&\quad = \frac{1-\epsilon}{1-2\epsilon + \frac{|k(x,y)| \cdot \epsilon}{|s(x,y,z)|}}
\end{aligned}
$$

We now show that the last expression is close to $1/2$. It is easy to check that

$$
\frac{|k(x,y)| \cdot \epsilon}{|s(x,y,z)|} \in \left[\frac{5}{6}, 2\right] \;\Rightarrow\; \frac{1-\epsilon}{1-2\epsilon + \frac{|k(x,y)| \cdot \epsilon}{|s(x,y,z)|}} \in \left[\frac{1}{4}, \frac{3}{4}\right] .
$$

Indeed, since $\epsilon \in (0, \frac{1}{4})$,

$$
\frac{1-\epsilon}{1-2\epsilon + \frac{|k(x,y)| \cdot \epsilon}{|s(x,y,z)|}} \geq \frac{1-\epsilon}{1-2\epsilon+2} \geq \frac{1-1/4}{1+2} = \frac{1}{4}
$$

and

$$
\begin{aligned}
\frac{1-\epsilon}{1-2\epsilon + \frac{|k(x,y)| \cdot \epsilon}{|s(x,y,z)|}} &\leq \frac{1-\epsilon}{1-2\epsilon+5/6} \\
&\leq \frac{1}{1-1/2+5/6} = \frac{3}{4} .
\end{aligned}
$$

We now substitute this into the (2). We have

$$
\begin{aligned}
&\sum_{\substack{x \in \{0,1\}^{m \times n} \\ y \in \{0,1\}^m \\ z \in \{0,1\}^n}} \left(\frac{1}{2} - \left|\frac{1}{2} - \mathbf{E}\left[c_I(Z) \,|\, X = x, Y = y, Z = z\right]\right|\right) \\
&\qquad \cdot \Pr\left[X = x, Y = y, Z = z\right] \\
&= \sum_{\substack{x \in \{0,1\}^{m \times n} \\ y \in \{0,1\}^m \\ z \in \{0,1\}^n}} \left(\frac{1}{2} - \left|\frac{1}{2} - \frac{1-\epsilon}{1-2\epsilon + \frac{|k(x,y)| \cdot \epsilon}{|s(x,y,z)|}}\right|\right) \\
&\qquad \cdot \Pr\left[X = x, Y = y, Z = z\right] \\
&\geq \sum_{\substack{x \in \{0,1\}^{m \times n} \\ y \in \{0,1\}^m \\ z \in \{0,1\}^n \\ \frac{|k(x,y,z)|\epsilon}{|s(x,y,z)|} \in [\frac{5}{6}, 2]}} \left(\frac{1}{2} - \left|\frac{1}{2} - \frac{1-\epsilon}{1-2\epsilon + \frac{|k(x,y)| \cdot \epsilon}{|s(x,y,z)|}}\right|\right) \\
&\qquad \cdot \Pr\left[X = x, Y = y, Z = z\right] \\
&\geq \sum_{\substack{x \in \{0,1\}^{m \times n} \\ y \in \{0,1\}^m \\ z \in \{0,1\}^n \\ \frac{|k(x,y,z)|\epsilon}{|s(x,y,z)|} \in [\frac{5}{6}, 2]}} \left(\frac{1}{2} - \frac{1}{4}\right) \cdot \Pr\left[X = x, Y = y, Z = z\right] \\
&= \frac{1}{4} \Pr\left[\frac{|k(X,Y)| \cdot \epsilon}{|s(X,Y,Z)|} \in \left[\frac{5}{6}, 2\right]\right] .
\end{aligned}
$$

In order to prove (1), we need to show that $\frac{|k(X,Y)| \cdot \epsilon}{|s(X,Y,Z)|} \in \left[\frac{5}{6}, 2\right]$ with probability at least $1/2$. To that end, we define two additional random variables

$$
K = |k(X,Y)| \qquad \text{and} \qquad S = |s(X,Y,Z)| .
$$

The condition $\frac{|k(X,Y)| \cdot \epsilon}{|s(X,Y,Z)|} \in \left[\frac{5}{6}, 2\right]$ is equivalent to

$$
\frac{1}{2}\epsilon \leq \frac{S}{K} \leq \frac{6}{5}\epsilon . \tag{3}
$$

First, we lower bound K. For any $y \in \{0,1\}^m$ and any $i, j \in \{1, 2, \ldots, n\}$,

$$
\begin{aligned}
&\Pr\left[j \in k(X,Y) \,|\, Y = y, I = i\right] \\
&\qquad = \begin{cases} 1 & \text{if } j = i, \\ \epsilon^{\|y\|}(1-\epsilon)^{m-\|y\|} & \text{if } j \neq j. \end{cases}
\end{aligned}
$$

Conditioned on $Y = y$ and $I = i$, the random variable $K - 1 = |k(X,Y) \setminus \{I\}|$ is a sum of $n-1$ Bernoulli variables with parameter $\epsilon^{\|y\|}(1-\epsilon)^{m-\|y\|}$, one for each column except for column i. Hoeffding bound with $t = \epsilon^m/2$ and the loose lower bound $\epsilon^{\|y\|}(1-\epsilon)^{m-\|y\|} \geq \epsilon^m$ gives

$$\begin{aligned}
&\Pr\left[\frac{K-1}{n-1} > \frac{\epsilon^m}{2} \,\middle|\, Y = y, I = i\right] \\
&= \Pr\left[\frac{K-1}{n-1} > \epsilon^m - t \,\middle|\, Y = y, I = i\right] \\
&\geq \Pr\left[\frac{K-1}{n-1} > \epsilon^{\|y\|}(1-\epsilon)^{m-\|y\|} - t \,\middle|\, Y = y, I = i\right] \\
&\geq 1 - e^{-2(n-1)t^2} \ .
\end{aligned}$$

Since $m \leq \frac{\ln n}{3\ln(1/\epsilon)}$, we lower bound $t = \frac{\epsilon^m}{2}$ as

$$t = \epsilon^m/2 > \frac{1}{2}\epsilon^{\frac{\ln n}{3\ln(1/\epsilon)}} = \frac{1}{2\sqrt[3]{n}} \ .$$

Since the lower bound is uniform for all choices of y and i, we can remove the conditioning and conclude that

$$\Pr\left[K > 1 + \frac{(n-1)}{2\sqrt[3]{n}}\right] \geq 1 - \exp\left(-\frac{(n-1)}{2n^{2/3}}\right) \ .$$

For $n \geq 25$, we can simplify it further to

$$\Pr\left[K \geq \frac{n^{2/3}}{2}\right] \geq \frac{3}{4} \ .$$

Second, conditioned on $K = r$, the random variable S is a sum of $r-1$ Bernoulli random variables with parameter ϵ and one Bernoulli random variable with parameter $1/2$. Hoeffding bound for any $t \geq 0$ gives that

$$\Pr\left[\left|\frac{S}{K} - \frac{\epsilon(K-1)+1/2}{K}\right| < t \,\middle|\, K = r\right] \geq 1 - 2e^{-2rt^2} \ .$$

Thus,

$$\begin{aligned}
&\Pr\left[\left|\frac{S}{K} - \frac{\epsilon(K-1)+1/2}{K}\right| < t \text{ and } K \geq \frac{n^{2/3}}{2}\right] \\
&\geq \sum_{r=\lceil n^{2/3}/2\rceil}^{n} \Pr\left[\left|\frac{S}{K} - \frac{\epsilon(K-1)+1/2}{K}\right| < t \,\middle|\, K = r\right] \\
&\qquad \cdot \Pr[K = r] \\
&\geq \sum_{r=\lceil n^{2/3}/2\rceil}^{n} \left(1 - 2e^{-2rt^2}\right) \cdot \Pr[K = r] \\
&\geq \left(1 - 2e^{-n^{2/3}t^2/2}\right) \cdot \Pr\left[K \geq \frac{n^{2/3}}{2}\right] \ .
\end{aligned}$$

We choose $t = \epsilon/4$. Since $n \geq 600/\epsilon^3$, we have $e^{-n^{2/3}t^2/2} < \frac{1}{8}$ and thus

$$\begin{aligned}
&\Pr\left[\left|\frac{S}{K} - \frac{\epsilon(K-1)+1/2}{K}\right| < t \text{ and } K \geq \frac{n^{2/3}}{2}\right] \\
&\geq \left(1 - 2e^{-n^{2/3}t^2/2}\right) \cdot \Pr\left[K \geq \frac{n^{2/3}}{2}\right] \\
&\geq \frac{3}{4}\left(1 - 2e^{-n^{2/3}t^2/2}\right) \\
&> \frac{3}{4}\left(1 - \frac{1}{4}\right) = \frac{9}{16} > \frac{1}{2} \ .
\end{aligned}$$

We claim that $t = \epsilon/4$, $\left|\frac{S}{K} - \frac{\epsilon(K-1)+1/2}{K}\right| < t$ and $K \geq \frac{n^{2/3}}{2}$ imply (3). To see that, note that $\left|\frac{S}{K} - \frac{\epsilon(K-1)+1/2}{K}\right| < t$ is equivalent to

$$\frac{\epsilon(K-1)+1/2}{K} - t < \frac{S}{K} < \frac{\epsilon(K-1)+1/2}{K} + t$$

which implies that

$$p\left(1 - \frac{1}{K}\right) - t < \frac{S}{K} < \epsilon\left(1 - \frac{1}{K}\right) + \frac{1}{2K} + t \ .$$

Since $K \geq \frac{n^{2/3}}{2}$ and $n \geq 25$ we have $K > 4$, which implies that

$$\frac{3}{4}\epsilon - t < \frac{S}{K} < \frac{3}{4}\epsilon + \frac{1}{2K} + t \ .$$

Since $K \geq \frac{n^{2/3}}{2}$ and $n \geq \frac{12}{\epsilon^{3/2}}$ we have $K > \frac{5}{2\epsilon}$, which implies that

$$\frac{3}{4}\epsilon - t < \frac{S}{K} < \frac{3}{4}\epsilon + \frac{\epsilon}{5} + t \ .$$

Since $t = \epsilon/4$, the condition (3) follows. □

5. All Functions

Let $\mathcal{X}$ be some finite domain. We say a sample $T = ((x_1,y_1),\ldots,(x_m,y_m)) \in (\mathcal{X} \times \{0,1\})^m$ of size m is *self-consistent* if for any $i,j \in \{1,2,\ldots,m\}$, $x_i = x_j$ implies that $y_i = y_j$. A distribution independent algorithm A is said to be *consistent* if for any self-consistent sample $T = ((x_1,y_1),\ldots,(x_m,y_m)) \in (\mathcal{X} \times \{0,1\})^m$, $A(T)(x_i) = y_i$ holds for any $i = 1,2,\ldots,m$.

In this section we show that for $C_{\text{all}} = \{0,1\}^{\mathcal{X}}$, any consistent distribution independent learner is almost as powerful as any distribution independent learner. Note that, in particular, the ERM algorithm for C_{all} is consistent. In other words, for the class C_{all} unlabeled data do *not* have any information theoretic value.

Theorem 8 (No Gap)**.** *Let $\mathcal{X}$ be some finite domain, $C_{all} = \{0,1\}^{\mathcal{X}}$ and A be any consistent learning algorithm. Then, for any distribution P over $\mathcal{X}$, any (possibly distribution dependent) learning algorithm B and any $\epsilon,\delta \in (0,1)$,*

$$m(A, C_{all}, P, 2\epsilon, 2\delta) \leq m(B, C_{all}, P, \epsilon, \delta) \ .$$

Proof. Fix any integer $m \geq 0$ and any distribution P over $\mathcal{X}$. Let $X, X_1, X_2, \ldots, X_m$ be an i.i.d. sample from P. Define the random variable

$$Z = \Pr[X \notin \{X_1, X_2, \ldots, X_m\} \mid X_1, X_2, \ldots, X_m] \,.$$

In other words, Z is the probability mass not covered by $X_1, X_2, \ldots, X_m$. For any $c \in C_{\text{all}}$, let $T_c = ((X_1, c(X_1)), (X_2, c(X_2)), \ldots, (X_m, c(X_m)))$ be the sample labeled according to c. Since A is consistent, with probability one, for any $c \in C_{\text{all}}$,

$$d_P(A(T_c), c) \leq Z \,. \tag{4}$$

Let $\widetilde{c}$ be chosen uniformly at random from C_{all}, independently of $X, X_1, X_2, \ldots, X_m$. Additionally, define $\widehat{c} \in C_{\text{all}}$ as

$$\widehat{c}(x) = \begin{cases} \widetilde{c}(x) & \text{if } x \in \{X_1, X_2, \ldots, X_m\}, \\ 1 - \widetilde{c}(x) & \text{otherwise.} \end{cases}$$

and note that $\widehat{c}$ and $\widetilde{c}$ are distributed identically and $T_{\widetilde{c}} = T_{\widehat{c}}$, and thus

$$\begin{aligned} &\mathbf{E}\left[\mathbf{1}\left[d_P\left(B\left(T_{\widetilde{c}}\right), \widetilde{c}\right) \geq \epsilon\right] \mid T_{\widetilde{c}}\right] \\ &= \mathbf{E}\left[\mathbf{1}\left[d_P\left(B\left(T_{\widehat{c}}\right), \widehat{c}\right) \geq \epsilon\right] \mid T_{\widetilde{c}}\right] \end{aligned} \tag{5}$$

We have

$$\begin{aligned} &\sup_{c \in C_{\text{all}}} \Pr[d_P(B(T_c), c) \geq \epsilon] \\ &= \sup_{c \in C_{\text{all}}} \mathbf{E}\left[\mathbf{1}\left[d_P\left(B\left(T_c\right), c\right) \geq \epsilon\right]\right] \\ &\geq \mathbf{E}\left[\mathbf{1}\left[d_P\left(B\left(T_{\widetilde{c}}\right), \widetilde{c}\right) \geq \epsilon\right]\right] \\ &= \mathbf{E}\left[\mathbf{E}\left[\mathbf{1}\left[d_P\left(B\left(T_{\widetilde{c}}\right), \widetilde{c}\right) \geq \epsilon\right] \mid T_{\widetilde{c}}\right]\right] \\ &= \mathbf{E}\left[\mathbf{E}\left[\frac{1}{2}\mathbf{1}\left[d_P\left(B\left(T_{\widetilde{c}}\right), \widetilde{c}\right) \geq \epsilon\right] \right.\right. \\ &\qquad \left.\left. + \frac{1}{2}\mathbf{1}\left[d_P\left(B\left(T_{\widehat{c}}\right), \widehat{c}\right) \geq \epsilon\right] \,\middle|\, T_{\widetilde{c}}\right]\right] \qquad (6) \\ &\geq \mathbf{E}\left[\mathbf{E}\left[\frac{1}{2}\mathbf{1}\left[Z \geq 2\epsilon\right] \,\middle|\, T_{\widetilde{c}}\right]\right] \qquad (7) \\ &= \frac{1}{2}\,\mathbf{E}\left[\mathbf{1}\left[Z \geq 2\epsilon\right]\right] \\ &= \frac{1}{2}\Pr\left[Z \geq 2\epsilon\right] \\ &= \frac{1}{2}\sup_{c \in C}\Pr\left[Z \geq 2\epsilon\right] \\ &\geq \frac{1}{2}\sup_{c \in C}\Pr\left[d_P(A(T_c), c) \geq 2\epsilon\right] \,. \qquad (8) \end{aligned}$$

Equation (6) follows from (5). To justify inequality (7), note that since the classifiers $\widetilde{c}$ and $\widehat{c}$ disagree on the missing mass, if $Z \geq 2\epsilon$ then $d_P(B(T_{\widetilde{c}}), \widetilde{c}) \geq \epsilon$ or $d_P(B(T_{\widehat{c}}), \widehat{c}) \geq \epsilon$ or both. By symmetry between $\widetilde{c}$ and $\widehat{c}$, if $Z \geq 2\epsilon$ then with probability at least $1/2$, $d_P(B(T_{\widetilde{c}}), \widetilde{c}) \geq \epsilon$. Inequality (8) follows from (4).

Since the inequality

$$\sup_{c \in C_{\text{all}}} \Pr[d_P(B(T_c), c) \geq \epsilon] \geq \frac{1}{2}\sup_{c \in C}\Pr\left[d_P(A(T_c), c) \geq 2\epsilon\right]$$

holds for arbitrary m, it implies $m(A, C_{\text{all}}, P, 2\epsilon, 2\delta) \leq m(B, C_{\text{all}}, P, \epsilon, \delta)$ for any $\epsilon, \delta \in (0, 1)$. □

6. Conclusion and Open Problems

Darnstädt et al. (2013) showed that the gap between the number of samples needed to learn a class of functions of Vapnik-Chervonenkis dimension d *with* and *without* knowledge of the distribution is upper-bounded by $O(d)$. We show that this bound is tight for the class of Boolean projections. On the other hand, for the class of all functions, this gap is only constant. These observations lead to the following research directions.

First, it will be interesting to understand the value of the gap for larger classes of functions. For example, one might consider the classes of (monotone) disjunctions over $\{0,1\}^n$, (monotone) conjuctions over $\{0,1\}^n$, parities over $\{0,1\}^n$, and halfspaces over $\mathbb{R}^n$. The Vapnik-Chervonenkis dimension of these classes is $\Theta(n)$ thus the gap for these classes is at least $\Omega(1)$ and at most $O(n)$. Other than these crude bounds, the question of what is the gap for these classes is wide open.

Second, as the example with class of all functions shows, the gap is *not* characterized by the Vapnik-Chervonenkis dimension. It will be interesting to study other parameters which determine this gap. In particular, it will be interesting to obtain upper bounds on the gap in terms of other quantities.

Finally, we believe that studying this question in the agnostic extension of the PAC model (Anthony & Bartlett, 1999, Chapter 2) will be of great interest, too.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. The work of the first author is supported by a Rabin postdoctoral fellowship.

References

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

Ben-David, S., Lu, T., and Pál, D. Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning. In *Proceedings*

of the 21st Annual Conference on Learning Theory, Helsinki, Finland, 9–12, July 2008, pp. 33–44. Omnipress, 2008.

Benedek, G. M. and Itai, A. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86 (2):377–389, 1991.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Darnstädt, M., Simon, H. U., and Szörényi, B. Unlabeled data does provably help. In *30th International Symposium on Theoretical Aspects of Computer Science, STACS 2013, February 27 - March 2, 2013, Kiel, Germany*, pp. 185–196, 2013.

Devroye, L. and Lugosi, G. *Combinatorial Methods in Density Estimation*. Springer, 2000.

Dudley, R. M. Central limit theorems for empirical measures. *The Annals of Probability*, pp. 899–929, 1978.

Dudley, R. M. *A course on empirical processes*, pp. 1–142. Springer, 1984.

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

Hanneke, S. The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016.

Haussler, D. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69 (2):217–232, 1995.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Lu, T. Fundamental limitations of semi-supervised learning. Master's thesis, David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1, 2009. Available at https://uwspace.uwaterloo.ca/bitstream/handle/10012/4387/lumastersthesis_electronic.pdf.

A. Size of ϵ-cover

In this section, we present $(4e/\epsilon)^{d/(1-1/e)}$ upper bound on the size of the ϵ-cover of any concept class of Vapnik-Chervonenkis dimension d. To prove our result, we need Sauer's lemma. Its proof can be found, for example, in Anthony & Bartlett (1999, Chapter 3).

Lemma 9 (Sauer's lemma)**.** *Let $\mathcal{X}$ be a non-empty domain and let $C \subseteq \{0,1\}^{\mathcal{X}}$ be a concept class with Vapnik-Chervonenkis dimension d. Then, for any $S \subseteq \mathcal{X}$,*

$$|\{\pi(c,S) \ : \ c \in C\}| \leq \sum_{i=0}^{d} \binom{|S|}{i} .$$

We remark that if $n \geq d \geq 1$ then

$$\sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{ne}{d}\right)^{d} \tag{9}$$

where $e = 2.71828\ldots$ is the base of the natural logarithm. This follows from the following calculation

$$\begin{aligned}\left(\frac{d}{n}\right)^{d} \cdot \sum_{i=0}^{d} \binom{n}{i} &\leq \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^{i} \\ &\leq \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^{i} \\ &= \left(1+\frac{d}{n}\right)^{n} \leq e^{d}\end{aligned}$$

where we used in the last step that $1 + x \leq e^{x}$ for any $x \in \mathbb{R}$.

Theorem 10 (Size of ϵ-cover)**.** *Let $\mathcal{X}$ be a non-empty domain and let $C \subseteq \{0,1\}^{\mathcal{X}}$ be a concept class with Vapnik-Chervonenkis dimension d. Let P be any distribution over $\mathcal{X}$. For any $\epsilon \in (0,1]$, there exists a set $C' \subseteq C$ such that*

$$|C'| \leq \left(\frac{4e}{\epsilon}\right)^{d/(1-1/e)} \tag{10}$$

and for any $c \in C$ there exists $c' \in C'$ such that $d_P(c,c') \leq \epsilon$.

Proof. We say that a set $B \subseteq C$ is an *ϵ-packing* if

$$\forall c, c' \in B \qquad c \neq c' \quad \Longrightarrow \quad d_P(c,c') > \epsilon$$

We claim that there exists a maximal ϵ-packing. In order to show that a maximal set exists we to appeal to Zorn's lemma. Consider the collection of all ϵ-packings. We impose partial order on them by set inclusion. Notice that any totally ordered collection $\{B_i \ : \ i \in I\}$ of ϵ-packings has an upper bound $\bigcup_{i\in I} B_i$ that is an ϵ-packing. Indeed, if $c, c' \in \bigcup_{i\in I} B_i$ such that $c \neq c'$ then there exists $i \in I$ such that $c, c' \in B_i$ since $\{B_i \ : \ i \in I\}$ is totally ordered. Since B_i is an ϵ-packing, $d_P(c,c') > \epsilon$. We conclude that $\bigcup_{i\in I} B_i$ is an ϵ-packing. By Zorn's lemma, there exists a maximal ϵ-packing.

Let C' be a maximal ϵ-packing. We claim that C' is also an ϵ-cover of C. Indeed, for any $c \in C$ there exists $c' \in C'$ such that $d_P(c,c') \leq \epsilon$ since otherwise $C' \cup \{c\}$ would be an ϵ-packing, which would contradict maximality of C'.

It remains to upper bound $|C'|$. Consider any finite subset $C'' \subseteq C'$. It suffices to show an upper bound on $|C''|$ and since C'' is arbitrary, the same upper bound holds for $|C'|$. Let $M = |C''|$ and let $c_1, c_2, \ldots, c_M$ be concepts in C''. For any $i, j \in \{1, 2, \ldots, M\}$, $i < j$, let

$$A_{i,j} = \{x \in \mathcal{X} \ : \ c_i(x) \neq c_j(x)\} .$$

Let $X_1, X_2, \ldots, X_K$ be an i.i.d. sample from P. We will choose K later. Since $d_P(c_i, c_j) > \epsilon$,

$$\Pr[X_k \in A_{i,j}] > \epsilon \qquad \text{for } k = 1, 2, \ldots, K.$$

Since there are $\binom{M}{2}$ subsets $A_{i,j}$, we have

$$\begin{aligned}&\Pr\left[\forall i,j, i<j, \ \exists k, \ X_k \in A_{i,j}\right] \\ &\quad = 1 - \Pr\left[\exists i,j, i<j, \ \forall k, \ X_k \notin A_{i,j}\right] \\ &\quad \geq 1 - \sum_{1\leq i<j\leq M} \Pr\left[\forall k, \ X_k \notin A_{i,j}\right] \\ &\quad \geq 1 - \sum_{1\leq i<j\leq M} (1-\epsilon)^{K} \\ &\quad = 1 - \binom{M}{2}(1-\epsilon)^{K} .\end{aligned}$$

For $K = \left\lceil \frac{\ln\binom{M}{2}}{\epsilon} \right\rceil + 1$, the above probability is strictly positive. This means there exists a set $S = \{x_1, x_2, \ldots, x_K\} \subseteq X$ such that $A_{i,j} \cap S$ is non-empty for every $i < j$. This means that for every for every $i \neq j$, $c_i(S) \neq c_j(S)$ and hence $M = |C''| = |\{\pi(c,S) \ : \ c \in C\}|$. Thus by Sauer's lemma

$$M \leq \sum_{i=0}^{d} \binom{K}{i} .$$

We now show that this inequality implies that $M \leq (4e/\epsilon)^{d/(1-1/e)}$. We consider several cases.

Case 1: $d = -\infty$. That is, no set is shattered, and $C = \emptyset$. Then, $M = 0$ and inequality trivially follows.

Case 2: $d = 0$. Then, $M \leq 1$ and the inequality trivially follows.

Case 3a: $d \geq 1$ and $M \leq e^{d}$. Clearly, $M \leq e^{d} \leq (4e/\epsilon)^{d/(1-1/e)}$.

Case 3b: $d \geq 1$ and $M > e^d$. Then, $K \geq \ln M \geq d$ and hence by (9),

$$M \leq \sum_{i=0}^{d} \binom{K}{i} \leq \left(\frac{Ke}{d}\right)^d .$$

Thus,

$$\begin{aligned}
\ln M &\leq d \ln\left(\frac{Ke}{d}\right) \\
&\leq d \ln\left(\frac{e\left(\left\lceil \frac{\ln\binom{M}{2}}{\epsilon}\right\rceil + 1\right)}{d}\right) \\
&\leq d \ln\left(\frac{e\left(\frac{\ln\binom{M}{2}}{\epsilon} + 2\right)}{d}\right) \\
&\leq d \ln\left(\frac{e\left(\frac{\ln\binom{M}{2}+2}{\epsilon}\right)}{d}\right) \\
&\leq d \ln\left(\frac{e\left(\frac{2\ln M+2}{\epsilon}\right)}{d}\right) \\
&\leq d \ln\left(\frac{e\left(\frac{4\ln M}{\epsilon}\right)}{d}\right) \\
&= d\left[\ln\left(\frac{4e}{\epsilon}\right) + \ln\left(\frac{\ln M}{d}\right)\right] \\
&\leq d \ln\left(\frac{4e}{\epsilon}\right) + \frac{1}{e}\ln M .
\end{aligned}$$

where in the last step we used that $\ln x \leq x/e$ for any $x > 0$. Hence,

$$(1 - 1/e)\ln M \leq d \ln\left(\frac{4e}{\epsilon}\right)$$

which implies the lemma. □

B. Fixed Distribution Learning

Theorem 11 (Chernoff–Hoeffding bound, (Hoeffding, 1963)). *Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli random variables with $\mathbf{E}[X_i] = p$. Then, for any $\epsilon \in [0, \min\{p, 1-p\}$),*

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq p + \epsilon\right] \leq e^{-nD(p+\epsilon\|p)} ,$$
$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \leq p - \epsilon\right] \leq e^{-nD(p-\epsilon\|p)} .$$

where

$$D\left(x\|y\right) = x\ln\left(\frac{x}{y}\right) + (1-x)\ln\left(\frac{1-x}{1-y}\right)$$

is the Kullback-Leibler divergence between Bernoulli distributions with parameters $x, y \in [0, 1]$.

We further use the following inequality

$$D\left(x\|y\right) \geq \frac{(x-y)^2}{2\max\{x, y\}}$$

Theorem 12 (Benedek-Itai). *Let $C \subseteq \{0,1\}^{\mathcal{X}}$ be a concept class over a non-empty domain $\mathcal{X}$. Let P be a distribution over $\mathcal{X}$. Let $\epsilon \in (0, 1]$ and assume that C has an $\frac{\epsilon}{2}$-cover of size at most N. Then, there exists an algorithm, such that for any $\delta \in (0, 1)$, any target $c \in C$, if it gets*

$$m \geq 48\left(\frac{\ln N + \ln(1/\delta)}{\epsilon}\right)$$

labeled samples then with probability at least $1 - \delta$, it ϵ-learns the target.

Proof. Given a labeled sample $T = ((x_1, y_1), \ldots, (x_m, y_m))$, for any $c \in C$, we define

$$\mathrm{err}_T(c) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left[c(x_i) \neq y_i\right] .$$

Let $C' \subseteq C$ be an $(\epsilon/2)$-cover of size at most N. Consider the algorithm A that given a labeled sample T outputs

$$\widehat{c} = \operatorname*{argmin}_{c' \in C'} \mathrm{err}_T(c')$$

breaking ties arbitrarily. We prove that A, with probability at least $1 - \delta$, ϵ-learns any target $c \in C$ under the distribution P.

Consider any target $c \in C$. Then there exists $\widetilde{c} \in C'$ such that $d_P(c, \widetilde{c}) \leq \epsilon/2$. Let $C'' = \{c' \,:\, d_P(c, c') > \epsilon\}$. We claim that with probability at least $1 - \delta$, for all $c' \in C''$, $\mathrm{err}_T(c') > \frac{2}{3}\epsilon$ and $\mathrm{err}_T(\widetilde{c}) < \frac{2}{3}\epsilon$ and hence A outputs $\widehat{c} \in C' \setminus C''$.

Consider any $c' \in C''$ and note that $\mathrm{err}_T(c')$ is an average of Bernoulli random variables with mean $d_P(c, c') > \epsilon$. Thus, by Chernoff bound,

$$\begin{aligned}
\Pr\left[\mathrm{err}_T(c') > \frac{2}{3}\epsilon\right] &> 1 - \exp\left(-mD\left(\frac{2}{3}\epsilon\middle\|d_P(c,c')\right)\right) \\
&> 1 - \exp\left(-m\frac{(\frac{2}{3}\epsilon - d_P(c,c'))^2}{2d_P(c,c')}\right) \\
&> 1 - \exp\left(-m\epsilon/18\right)
\end{aligned}$$

where the last inequality follows from the inequality

$$\left(\frac{2}{3}\epsilon - x\right)^2 \geq \frac{1}{9}\epsilon x$$

valid for any $x \geq \epsilon > 0$. Similarly, $\mathrm{err}_T(\widetilde{c})$ is an average of Bernoulli random variables with mean $d_P(c, \widetilde{c}) < \epsilon/2$. Thus, by Chernoff bound,

$$\begin{aligned}
\Pr\left[\mathrm{err}_T(\widetilde{c}) < \frac{2}{3}\epsilon\right] &> 1 - \exp\left(-mD\left(\frac{2}{3}\epsilon \middle\| d_P(c,\widetilde{c})\right)\right) \\
&> 1 - \exp\left(-m\frac{(\frac{2}{3}\epsilon - d_P(c,\widetilde{c}))^2}{\frac{4}{3}\epsilon}\right) \\
&> 1 - \exp\left(-m\epsilon/48\right) \ .
\end{aligned}$$

Since $|C''| \leq N - 1$, by union bound, with probability at least $1 - (N-1)\exp(-m\epsilon/48)$, for all $c' \in C''$, $\mathrm{err}_T(c') > \frac{2}{3}\epsilon$. Finally, with probability at least $1 - N\exp(-m\epsilon/48) \geq 1 - \delta$, $\mathrm{err}_T(\widetilde{c}) < \frac{2}{3}\epsilon$ and for all $c' \in C''$, $\mathrm{err}_T(c') > \frac{2}{3}\epsilon$. □