Does Unlabeled Data Provably Help?
Worst-case Analysis of the Sample Complexity of Semi-Supervised Learning

(This paper is eligible for the Mark Fulk Award (Tyler Lu and Dávid Pál).)

Shai Ben-David and **Tyler Lu** and **Dávid Pál**
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, ON, Canada
{shai,ttlu,dpal}@cs.uwaterloo.ca

Abstract

We study the potential benefits of unlabeled data to classification prediction to the learner. We compare learning in the semi-supervised model to the standard, supervised PAC (distribution free) model, considering both the realizable and the unrealizable (agnostic) settings.

Roughly speaking, our conclusion is that access to unlabeled samples cannot provide sample size guarantees that are better than those obtainable without access to unlabeled data, unless one postulates very strong assumptions about the distribution of the labels.

In particular, we prove that for basic hypothesis classes over the real line, if the distribution of unlabeled data is 'smooth', knowledge of that distribution cannot improve the labeled sample complexity by more than a constant factor (e.g., 2). We conjecture that a similar phenomena holds for any hypothesis class and any unlabeled data distribution. We also discuss the utility of semi-supervised learning under the common *cluster assumption* concerning the distribution of labels, and show that even in the most accommodating cases, where data is generated by two uni-modal label-homogeneous distributions, common SSL paradigms may be misleading and inflict poor prediction performance.

1 Introduction

While supervised classification has received a lot of research attention and is reasonably well understood, in many practical learning scenarios, labeled data is hard to come by and unlabeled data is readily available. Consequently, users try to utilize available unlabeled data to assist with the classification learning process. Learning from both labeled and unlabeled data is commonly called semi-supervised learning (SSL). Due to its wide potential applications, this approach is gaining attention in both the application oriented and the theoretical machine learning communities.

However, theoretical analysis of semi-supervised learning has, so far, been scarce and it falls short of providing unequivocal explanation of merits of using unlabeled examples in learning. We take steps toward rectifying this theory-practice gap by providing formal analysis of some semi-supervised learning settings. The question we focus on is whether unlabeled data can be utilized to provably improve the sample complexity of classification learning.

We investigate what type of assumptions about the data generating distribution (or which circumstances) are sufficient to make the SSL approach yield better predictions than fully supervised learning. The bulk of this paper focuses on showing that without prior knowledge about the distribution of *labels*, SSL cannot guarantee any significant advantages in sample complexity (no more than a constant factor for learning tasks over the real line).

The basis for our theory is a simplified, utopian, model of semi-supervised learning, in which the learning algorithm has perfect knowledge of the probability distribution of the unlabeled data. We focus on estimating the *labeled sample* complexity of learning. Since our model provides the learner with more information than just a sample of the unlabeled data distribution, lower bounds on the labeled sample complexity of learning in our model imply similar lower bounds for common notions of semi-supervised learning. Upper bounds, or sample size sufficiency results (for the labeled samples) in our model, apply to the common SSL setting only once sufficiently large unlabeled samples are available to the learner. In this paper we mainly discuss lower bounds, and when we address upper bounds we settle for stating that they apply eventually as the unlabeled sample sizes grow.

Our model of semi-supervised learning can be viewed as learning with respect to a fixed distribution, (see Benedek and Itai [5]). However, our emphasis is different. Our goal is to compare how the *knowledge* of the unlabeled distribution helps, as opposed to learning when the only access to the underlying unlabeled data distribution is via the training labeled sample. We call the former setting *semi-supervised* and the latter *supervised* or *fully supervised* learning.

We present explicit formalization of different ways in which the merits of the semi-supervised paradigm can be measured. We then investigate the extent by which SSL can provide provable advantages over fully supervised learning with respect to these measures.

Roughly speaking, we conclude that no special unlabeled data distribution (like, say, one that breaks into clear data clusters) suffices to render SSL an advantage over fully supervised learning. Unlabeled data can make a difference only under strong assumptions (or prior knowledge) about the conditional *labeled* distribution.

One should note however, that in many cases such knowledge can also be utilized by a fully supervised algorithm. The search for justification to the SSL paradigm therefore leaves us with one setting - the cases where there exists prior knowledge about the *relationship* between the labels and the unlabeled data structure (and not just about the labels per se). However, we show in Section 3 that common applications of SSL paradigms for utilizing such relationship (like the popular *cluster assumption* or the related algorithmic bias towards class boundaries that pass through low-density data regions) may lead to poor prediction accuracy, even when the data does comply with the underlying data model (say, the data is generated by a mixture of two Gaussian distributions, one for each label, each generating a homogeneously labeled set of examples).

The potential merits of SSL, in both settings - either with or without making assumptions about the labeled distribution, have been investigated before. Vapnik's model of transductive learning [21], as well as Kääriäinen's paper [17] address the setting without restrictions on the way labels are generated while Balcan-Blum's augmented PAC model for semi-supervised learning [3, 4] offers a framework for formalizing prior knowledge about the relationship between labels and the structure of the unlabeled distribution. We elaborate more about these in the next section on related work. One basic difference between these works and ours is that they try to provide explanations of the success of the SSL paradigm while we focus on investigating its inherent limitations.

We do not resolve the issue of the utility of unlabeled data in full generality. Rather, we demonstrate the answers for relatively simple classes of concepts over the real line (thresholds and unions of d intervals). We believe that the answers generalize to other classes in an obvious way. Along the way we also pose some conjectures and open questions.

The paper is organized as follows. We start by discussing previous related work in section 2. Then, we take a detour in section 3 and show that a commonly held assumption can result in performance degradation of SSL. We continue on our main path in section 4 where we formally define our model of semi-supervised learning and introduce notation. Section 5 casts the previous paradigms in our model and formally poses the question of in what sense unlabeled data can help. This question will guide the rest of the paper as we tackle it. Then section 6 analyzes this question for basic learning tasks on the real line. The section finishes off by asking a slightly different question how one might compare SSL and supervised learning. We conclude our paper in section 7 where we also discuss open questions and directions for further research.

2 Related Work

As we mentioned above, analysis of performance guarantees for semi-supervised learning can be carried out in two main setups. The first focuses on the unlabeled marginal data distribution and does not make any prior assumptions about the conditional label distribution. The second approach focuses on assumptions about the conditional labeled distribution, under which the SSL approach has potentially better label prediction performance than learning based on just labeled samples. The investigation of the first setup was pioneered by Vapnik in the late 70s in his model of transductive learning, e.g. [21]. There has been growing interest in this model in the recent years due to the popularity of using unlabeled data in practical label prediction tasks. This model assumes that unlabeled examples are drawn IID from an unknown distribution, and then the labels of some randomly picked subset of these examples are revealed to the learner. The goal of the learner is to label the remaining examples minimizing the error. The main difference with SSL is that the error of learner's hypothesis is judged only with respect to the known initial sample.

However, there are no known bounds in the transductive setting that are strictly better than supervised learning bounds (Vapnik's bounds [21] are almost identical). El-Yaniv and Pechyony [14] prove bounds that are similar to the usual margin bounds using Rademacher complexity, except that the learner is allowed to decide *a posteriori* the concept class given the unlabeled examples. But they do not show whether it can be advantageous to choose the class in this way. Their earlier paper [13] give bounds in terms of a notion of *uniform stability* of the learning algorithm, and in the broader setting where examples are not assumed to come IID from an unknown distribution. But again, it's not clear whether and when it beats the supervised learning bounds.

Methods for semi-supervised learning without prior assumption of conditional label distributions have been developed by Benedek and Itai [5], and Kääriäinen [17]. The idea of Benedek and Itai's algorithm is to construct a minimum ϵ-cover and apply empirical risk minimization (ERM) on the functions in the cover. Of course this ϵ-cover algorithm makes sense when we have knowledge of the unlabeled distribution. The algorithm of Kääriäinen is inspired by the clever observation that one can output the function that minimizes the distance to all other functions of the version space. This algorithm *can* be twice as good as in supervised ERM. For more details on these algorithms, see section 5.

The second, certainly more popular, set of semi-supervised approaches focuses on assumptions about the conditional labeled distributions. A recent PAC model of SSL proposed by Balcan and Blum [3, 4] attempts to formally capture such assumptions. They propose a notion of a compatibility function that assigns a higher score to classifiers which "fit nicely" with respect to the unlabeled distribution. The rational is that by narrowing down the set of classifiers to only compatible ones, the capacity of the set of potential classifiers goes down and the generalization bounds of empirical risk minimization improve. However, since the set of potential classifiers is trimmed down by a compatibility threshold, if the presumed label-structure relationship fails to hold, the learner may be left with only poorly performing classifiers. One serious concern about this approach is that it provides no way of verifying these crucial modeling assumptions. In section 3 we demonstrate that this approach may damage learning even when the underlying assumptions seem to hold. In Claim 2 we show that without prior knowledge of such relationship that the Balcan and Blum approach has poor worst-case generalization performance.

Common assumptions include the *smoothness assumption* and the related *low density assumption* [10] which sug-

gests that the decision boundary should lie in a low density region. In section 3, we give examples of mixtures of two Gaussians showing that the low density assumption may be misleading even under favourable data generation models, resulting in low density boundary SSL classifiers with larger error than the outcome of straightforward supervised learning that ignores the unlabeled data.

Many other assumptions about the labels/unlabeled data structure relationship have been investigated, most notably co-training [6] and explicit generative data models [11].

However, in all these approaches, the assumptions limiting the family of distributions P belongs to are quite strong and hard to verify.

3 Issues with Approaches Based on the Cluster Assumption

This paper has several results of the form "as long as one does not make any assumptions about the behavior of the *labels*, SSL cannot help much over algorithms that ignore the unlabeled data."

However, two arguments can be raised against such claims. First, SSL is not really intended to be used without any prior assumption about the distribution of labels. In fact, SSL can be viewed as applying some prior knowledge (or just belief) that the labels are somehow correlated with the unlabeled structure of the data. Can we say anything (anything negative, naturally ...) under such an assumption?

Second, maybe using unlabeled data can't *always* help you, but if it can help *sometimes* why not use it (always)? Well, can we show that in some cases the use of unlabeled data can indeed hurt the learner? Of course, nothing of that kind can apply for all potential learners, since a learner can choose to ignore the unlabeled data and then of course not get hurt by "using" it. We are therefore left with asking, "can the use of unlabeled data hurt the performance of *concrete* common SSL paradigms?"

We briefly address these two questions below by demonstrating that for certain *common* SSL strategies ("low density cut" and Balcan-Blum style use of "compatibility threshold") SSL can sometimes hurt you even when the (vaguely stated) "cluster assumption" does hold (when the data breaks into clear clusters).

In Figures 1, 2, and 3 we depict three examples of simple data distributions in which the data is generated by a mixture of two uni-modal distributions, and if each of these modes generated examples labeled homogeneously, each by a different label, then the minimum density of the unlabeled mixture data is significantly off the optimal label prediction decision boundary. Figure 1 shows a mixture of two equal-variance symmetric Gaussians, Figure 2 is a mixture of different Gaussians and Figure 3 shows an extreme case of unimodal density functions for which the error of the minimum density partition has classification error that is twice that of the optimal decision boundary.

Note that in all such examples, not only does the minimum-density bias mislead the learning process, but also, if one follows the paradigm suggested by Balcan and Blum [4], a wrong choice of the compatibility threshold level will doom the learning process to failure (whereas a simple empirical risk minimization that ignores unlabeled data will succeed based on a small number of labeled samples).

4 A No Prior Knowledge Model of Semi-Supervised Learning

We adopt the common (agnostic) PAC in which a learning problem is modeled by a probability distribution P over $X \times \{0,1\}$ for some domain set, X. Any function from X to $\{0,1\}$ is called a *hypothesis*. *Examples* are pairs, $(x,y) \in X \times \{0,1\}$, and a *sample* is a finite sequence $S = \{(x_i, y_i)\}_{i=1}^m$ of examples. The fundamental definition of our paper is:

Definition 1 (SL and SSL).

- *A* supervised *learning (SL) algorithm is a function,* $L : \bigcup_{m\in\mathbb{N}}(X \times \{0,1\})^m \to \{0,1\}^X$*, that mapping samples to a hypotheses.*
- *A* semi-supervised *learning (SSL) algorithm is a function* $L : \bigcup_{m\in\mathbb{N}}(X \times \{0,1\})^m \times \mathcal{P} \to \{0,1\}^X$*, where* $\mathcal{P}$ *is a set of probability distributions over* X*. Namely, an SSL algorithm takes as input not only a finite labeled sample but also a probability distribution over the domain set (and outputs a hypothesis, as before).*

For such a distribution P, let $\mathcal{D}(P)$ denote the marginal distribution over X. That is, formally, for $X' \subseteq X$ we define $\mathcal{D}(P)(X') = P(X' \times \{0,1\})$ provided that $X' \times \{0,1\}$ is P-measurable. For a learning problem P, we call $\mathcal{D}(P)$ the *unlabeled distribution* of P.

The *error* of a hypothesis h, with respect to P, is $\mathrm{Err}^P(h) = \Pr_{(x,y)\sim P}[h(x) \neq y]$. For a class H of hypotheses, the *sample complexity* of a semi-supervised learning algorithm A with respect to P, *confidence* $\delta > 0$ and *accuracy* $\epsilon > 0$, is

$$m(A, H, P, \epsilon, \delta) = \min\left\{m \in \mathbb{N} \ : \ \Pr_{S\sim P^m}[\mathrm{Err}^P(A(S, \mathcal{D}(P))) - \inf_{h'\in H} \mathrm{Err}^P(h') > \epsilon] < \delta\right\}.$$

The sample complexity of a supervised learning algorithm A is defined similarly, except that the second input parameter $\mathcal{D}(P)$ is omitted.

We consider two settings, realizable and agnostic. In the *agnostic* setting, P can be arbitrary. The *realizable* setting is defined by assuming that there exists hypothesis $h \in H$ such that $\mathrm{Err}^P(h) = 0$; consequently $\inf_{h'\in H} \mathrm{Err}^P(h') = 0$. In particular, this implies that for any $x \in X$, the conditional probabilities, $P(y = 0|\ x)$ and $P(y = 1|\ x)$ are always either 0 or 1. In the agnostic setting we do not make any such requirement.

Following the common PAC terminology and notation, the *empirical error*, $\mathrm{Err}^S(h)$, of a hypothesis h on a sample S is defined as $\mathrm{Err}^S(h) = \frac{1}{m}|\{i \ : \ i \in \{1, 2, \ldots, m\},\ h(x_i) \neq y_i\}|$.

Without reference to any learning problem, an *unlabeled distribution* D is simply any distribution over X. We use $\mathrm{Ext}(D)$ to denote all possible *extensions* of D, that is, $\mathrm{Ext}(D)$ is the family of all possible distributions P such that $\mathcal{D}(P) = D$. For an unlabeled distribution D and hypothesis h, D_h denotes the probability distribution in $\mathrm{Ext}(D)$ over $X \times \{0,1\}$ such that $D_h(y = h(x) \mid x) = 1$.

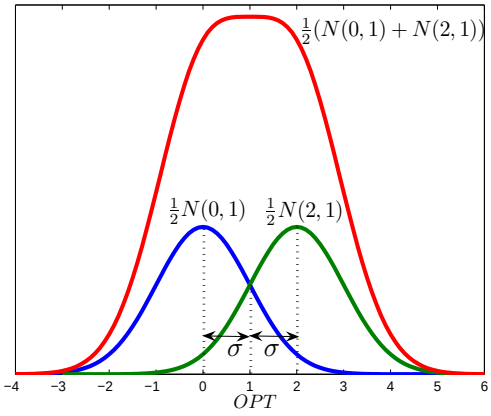

Figure 1: Mixture of two Gaussians $\mathcal{N}(0,1)$ (-) and $\mathcal{N}(2,1)$ (+) shows that the optimum threshold is at 1, the densest point. The sum of these two Gaussians is unimodal.

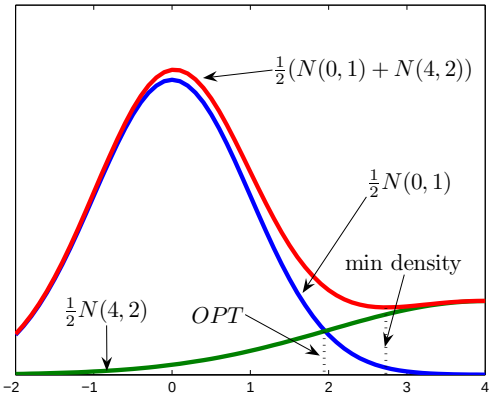

Figure 2: Mixture of two Gaussians $\mathcal{N}(0,1)$ (-) and $\mathcal{N}(4,2)$ (+) with difference variances. The minimum density point does not coincide with the optimum threshold where the two Gaussians intersect. The error of optimum is ≈ 0.17 and that of the minimum density point is ≈ 0.21.

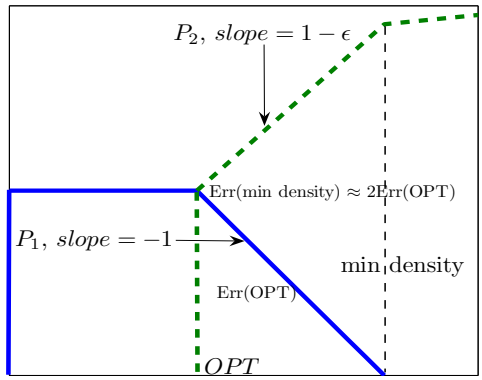

Figure 3: The solid line indicates the distribution P_1 (-) and the dotted line is P_2 (+). Their intersection is the optimum. The slope of the solid line is slightly steeper than that of the dotted line ($|-1| > 1-\epsilon$). The minimum density point occurs where P_1 falls to 0. So error of the minimum density threshold is twice that of the optimum.

For a subset T of some universal set, we use $\mathbf{1}T$ to denote its characteristic function. In particular, if $T \subseteq X$ then $\mathbf{1}T$ is a hypothesis over X. For two hypothesis g, h we use $g\Delta h$ to denote their "symmetric difference", that is, $g\Delta h$ is a hypothesis defined as $\mathbf{1}\{x \in X \ : \ g(x) \neq h(x)\}$. Let us also define $\mathrm{VC}(H)$ to be the VC-dimension [20] of hypothesis class H.

5 Previous No Prior Knowledge Paradigms

Previous approaches to SSL algorithms for the no prior knowledge paradigm have used the unlabeled sample to figure out the "geometry" of the hypothesis space with respect to the unlabeled (marginal) distribution. A common approach is to use that knowledge to reduce the hypothesis search space. In doing so, one may improve the generalization upper bounds.

Recall that given an unlabeled distribution D and a hypothesis class H, an *ϵ-cover* is a subset $H' \subseteq H$ such that for any $h \in H$ there exists $g \in H'$ such that $D(g\Delta h) \leq \epsilon$. Note that if H' is an ϵ-cover for H with respect to D, then for every extension $P \in \mathrm{Ext}(D)$ the $\inf_{g \in H'} \mathrm{Err}^P(g) \leq \inf_{h \in H} \mathrm{Err}^P(h) + \epsilon$. The smaller an ϵ-cover is the better its generalization bound one for the ERM algorithm over this cover.

In some cases the construction of a small ϵ-cover is a major use of unlabeled data. Benedek and Itai [5] analyze the approach, in the case when the unlabeled distribution is fixed and therefore can thought of as known to the learner.

The Balcan-Blum [4] suggest a different way of using the unlabeled data to reduce the hypothesis space. However, we claim that without making any prior assumptions about the relationship between the labeled and unlabeled distributions, their approach boils down to the ϵ-cover construction described above.

Claim 2. *Let H be any hypotheses class, $\epsilon, \delta > 0$, and D be any unlabeled distribution. Let $H' \subseteq H$ be the set of "compatible hypotheses." Suppose A is an SSL algorithm that outputs any hypothesis in H'. If H' does not contain an ϵ-cover of H with respect to D, the error of the hypothesis that A outputs is at least ϵ regardless of the size of the labeled sample.*

Proof. Since H' does not ϵ-cover of H, there exist a hypothesis $h \in H$ such that for all $g \in H'$, $D(g\Delta h) > \epsilon$. Thus, for any $g \in H'$, $\mathrm{Err}^{D_h}(g) > \epsilon$. Algorithm A outputs some $g \in H'$ and the proof follows. □

Kääriäinen [17] utilizes the unlabeled data in a different way. Given the labeled data his algorithm constructs the version space $F \subseteq H$ of all sample-consistent hypotheses, and then applies the knowledge of the unlabeled distribution D to find the "center" of that version space. Namely, a hypothesis $g \in F$ that minimizes $\max_{h \in F} D(g\Delta h)$.

Clearly, all the above paradigms depend on the knowledge of the unlabeled distribution D. In return, better upper bounds on the sample complexity of the respective algorithms (or equivalently on the errors of the hypotheses produced by such algorithms) can be shown. For example, Benedek and Itai give (for the realizable case) an upper bound on the sample complexity that depends on the size

of the ϵ-cover—the smaller ϵ-cover, the smaller the upper bound.

In the next section we analyze the gains that such knowledge of unlabeled data distribution can make in the no prior knowledge setting. We prove that over the real line for any "smooth" unlabeled distribution D, ERM over the full hypothesis class H has worst case sample complexity that is at most by constant factor bigger than the worst case sample complexity of *any* SSL algorithm. We conjecture that this a more general phenomenon.

Conjecture 3. For any hypothesis class H, there exists a constant $c \geq 1$ and a supervised algorithm A, such that for any distribution D over the domain and any semi-supervised learning algorithm B,

$$\sup_{h \in H} m(A, H, D_h, \epsilon, \delta) \leq c \cdot \sup_{h \in H} m(B, H, D_h, \epsilon, \delta)$$

for any ϵ and δ small enough, say smaller than $1/c$.

Conjecture 4. For any hypothesis class H, there exists a constant $c \geq 1$ and a supervised algorithm A, such that for any distribution D over the domain and any semi-supervised learning algorithm B,

$$\sup_{P \in \text{Ext } D} m(A, H, P, \epsilon, \delta) \leq c \cdot \sup_{P \in \text{Ext } D} m(B, H, P, \epsilon, \delta)$$

for any ϵ and δ small enough, say smaller than $1/c$.

6 Inherent Limitations of Semi-Supervised Learning

This section is devoted to proving the inherent limitations of SSL paradigm in the no prior knowledge model over the real line. In section 6.2 we prove Conjecture 3 for thresholds on the real line in the realizable setting, under the condition that the unlabeled distribution is absolutely continuous. In section 6.3 we prove Conjecture 4 for thresholds and union of d intervals over the real line in the agnostic setting (under the same unlabeled distribution condition).

The former follows from Theorems 7 and 9. The latter follows from Corollary 12 (for thresholds) and from Corollary 15 (for union of d intervals).

Let us start by defining the hypothesis classes. The class of thresholds is defined as $H = \{\mathbf{1}(-\infty, t] \ : \ t \in \mathbb{R}\}$ and the class of union of d intervals

$$UI_d = \{[a_1, a_2) \cup [a_3, a_4) \cup \cdots \cup [a_{2\ell-1}, a_{2\ell}) \ : \ \ell \leq d, \ a_1 \leq a_2 \leq \cdots \leq a_{2\ell}\} \ .$$

To prove the results we rely on a simple "rescaling trick" that we explain section 6.1.

In section 6.4 we discuss other potential formulations of the comparison between SL and SSL algorithms.

6.1 Rescaling Trick

In this section we show that learning any "natural" hypothesis class on the real has the same sample complexity for any absolutely continuous unlabeled distribution independent of its shape. Intuitively, if we imagine the real axis made of rubber, then a natural hypothesis class is one that is closed under rescaling (stretching) of the axis. Classes of thresholds and union of d intervals are examples of such natural classes, since under any rescaling an interval remains an interval. The rescaling will apply also on the unlabeled distribution over the real line and it will allow us to go from any absolutely continuous distribution to the uniform distribution over $(0, 1)$.

More formally, a *rescaling* is a continuous increasing function f from an open interval I onto an open interval J. We denote by $H|_A$ the restriction of a class H to a subset A, that is, $H|_A = \{h|_A \ : \ h \in H\}$. We use $\circ$ to denote function composition. We say that a hypothesis class H over $\mathbb{R}$ is *closed under rescaling* whenever for any rescaling $f : I \to J$, if $h|_J \in H|_J$, then $h|_J \circ f \in H|_I$. If H is any class closed under rescaling, then any rescaling f induces a bijection between $h|_J \mapsto h|_J \circ f$ bijection between $H|_I$ and $H|_J$. (This follows since f^{-1} is also rescaling.) Clearly, the class of thresholds and the class of unions of d intervals are closed under rescaling.

We show that the sample complexity of is unaffected by rescalings provided the the hypothesis class is closed under rescalings. We split the results into two lemmas—Lemma 5 and Lemma 6. The first lemma shows that if we have a supervised algorithm with certain sample complexity for the case when the unlabeled distribution is the uniform distribution over $(0, 1)$, then the algorithm can be translated into an SSL algorithm with the same sample complexity for the case when the unlabeled distribution is any absolutely continuous distribution. The second lemma shows the translation in the other direction. Namely, that a SSL algorithm with certain sample complexity on some absolutely continuous unlabeled distribution can be translated to a supervised algorithm for the case when unlabeled distribution is uniform over $(0, 1)$.

Lemma 5 (Rescaling trick I). *Let H be a hypothesis class over $\mathbb{R}$ closed under rescaling. Let U be the uniform distribution over $(0, 1)$. Let $\epsilon, \delta > 0$.*

(a) (Realizable case)*: If A is any supervised or semi-supervised algorithm, then there exists an semi-supervised learning algorithm B such that for any distribution D over an open interval I which is absolutely continuous with respect to Lebesgue measure on I*

$$\sup_{h \in H} m(B, H, D_h, \epsilon, \delta) \leq \sup_{g \in H} m(A, H, U_g, \epsilon, \delta) \ . \quad (1)$$

(b) (Agnostic case)*: If A is any supervised or semi-supervised algorithm, then there exists an semi-supervised learning algorithm B such that for any distribution D over an open interval I which is absolutely continuous with respect to Lebesgue measure on I*

$$\sup_{P \in \text{Ext}(D)} m(B, H, P, \epsilon, \delta) \leq \sup_{Q \in \text{Ext}(U)} m(A, H, Q, \epsilon, \delta) \ . \quad (2)$$

Proof. Fix H and A. We construct algorithm B as follows. The algorithm B has two inputs, a sample $S = \{(x_i, y_i)\}_{i=1}^m$ and a distribution D. Based on D the algorithm computes the cumulative distribution function $F : I \to (0, 1)$, $F(t) = D(I \cap (-\infty, t])$. Then, B computes from S transformed sample $S' = \{(x'_i, y_i)\}_{i=1}^m$ where $x'_i = F(x_i)$. On a sample S' the algorithm B simulates algorithm A and computes $h = A(S')$. (If A is semi-supervised we fix its second input to be U). Finally, B outputs $g = h \circ F$.

It remains to show that for any D with continuous cumulative distribution function (1) and (2) holds for any $\epsilon, \delta > 0$. We prove (2), the other equality is proved similarly.

Let $P \in \mathrm{Ext}(D)$. Slightly abusing notation, we define the "image" distribution $F(P)$ over $(0,1) \times \{0,1\}$ to be

$$F(P)(M) = P(\{(x,y) \,:\, (F(x),y) \in M\})$$

for any (measurable) $M \subseteq (0,1) \times \{0,1\}$. It is not hard to see that if S is distributed according to P^m, then S' is distributed according to $(F(P))^m$. Clearly, $\mathcal{D}(F(P)) = U$ i.e. $F(P) \in \mathrm{Ext}(U)$. Further note that since D is absolutely continuous, F is a rescaling. Hence $\mathrm{Err}^{F(P)}(h) = \mathrm{Err}^{P}(h \circ F)$ and $\inf_{h \in H} \mathrm{Err}^{P}(h) = \inf_{h \in H} \mathrm{Err}^{F(P)}(h)$. Henceforth, for any ϵ and any $m \in N$

$$\begin{aligned}
&\Pr_{S \sim P^m}[\mathrm{Err}^{P}(B(S,D)) - \inf_{h \in H} \mathrm{Err}^{P}(h) > \epsilon] \\
&= \Pr_{S' \sim F(P)^m}[\mathrm{Err}^{P}(A(S') \circ F) - \inf_{h \in H} \mathrm{Err}^{F(P)}(h) > \epsilon] \\
&= \Pr_{S' \sim F(P)^m}[\mathrm{Err}^{F(P)}(A(S')) - \inf_{h \in H} \mathrm{Err}^{F(P)}(h) > \epsilon] \,.
\end{aligned}$$

Therefore, for any $\epsilon, \delta > 0$,

$$\begin{aligned}
m(B,H,P,\epsilon,\delta) &= m(A,H,F(P),\epsilon,\delta) \\
&\leq \sup_{Q \in \mathrm{Ext}(P)} m(A,H,Q,\epsilon,\delta) \,.
\end{aligned}$$

Taking supremum over $P \in \mathrm{Ext}(D)$ finishes the proof. □

Lemma 6 (Rescaling trick II). *Let H be a hypothesis class over $\mathbb{R}$ closed under rescaling. Let U be the uniform distribution over $(0,1)$. Let $\epsilon, \delta > 0$.*

(a) (Realizable case)*: If B is any supervised or semi-supervised algorithm and D is any distribution over an open interval I, which is absolutely continuous with respect to the Lebesgue measure on I, then there exists a supervised learning algorithm A such that*

$$\sup_{g \in H} m(A,H,U_g,\epsilon,\delta) \leq \sup_{h \in H} m(B,H,D_h,\epsilon,\delta) \,. \qquad (3)$$

(b) (Agnostic case)*: If B is any supervised or semi-supervised algorithm and D is any distribution over an open interval I, which is absolutely continuous with respect to the Lebesgue measure on I, then there exists a supervised learning algorithm A such that*

$$\sup_{Q \in \mathrm{Ext}(U)} m(A,H,Q,\epsilon,\delta) \leq \sup_{P \in \mathrm{Ext}(D)} m(B,H,P,\epsilon,\delta) \,. \quad (4)$$

Proof. Fix H, B and D. Let $F : I \to (0,1)$ be the be cumulative distribution function of D, that is, $F(t) = D(I \cap (-\infty, t))$. Since D is absolutely continuous, F is a rescaling and inverse F^{-1} exists.

Now, we construct algorithm A. Algorithm A maps input sample $S' = \{(x'_i, y_i)\}_{i=1}^m$ to sample $S = \{(x_i, y_i)\}_{i=1}^m$ where $x_i = F^{-1}(x'_i)$. On a sample S the algorithm A simulates algorithm B and computes $g = B(S,D)$. (If B is supervised, then the second input is omitted.) Finally, A outputs $h = g \circ F^{-1}$.

It remains to show that for any D with continuous cumulative distribution function (3) and (4) holds for any $\epsilon, \delta > 0$. We prove (4), the other equality is proved similarly.

Let $Q \in \mathrm{Ext}(U)$. Slightly abusing notation, we define the "pre-image" distribution $F^{-1}(Q)$ over $I \times \{0,1\}$ to be

$$F^{-1}(Q)(M) = Q\left(\{(F(x),y) \,:\, (x,y) \in M\}\right)$$

for any (measurable) $M \subseteq I \times \{0,1\}$. It is not hard to see that if S' is distributed according to Q, then S is distributed according to $(F^{-1}(Q))^m$. Clearly, $\mathcal{D}(F^{-1}(U)) = D$ i.e. $F^{-1}(Q) \in \mathrm{Ext}(D)$. Since F^{-1} is a rescaling, $\mathrm{Err}^{F^{-1}(Q)}(h) = \mathrm{Err}^{Q}(h \circ F^{-1})$ and $\inf_{h \in H} \mathrm{Err}^{Q}(h) = \inf_{h \in H} \mathrm{Err}^{F^{-1}(Q)}(h)$. Henceforth, for any $\epsilon > 0$ and any $m \in \mathbb{N}$

$$\begin{aligned}
&\Pr_{S' \sim Q^m}[\mathrm{Err}^{Q}(A(S')) - \inf_{h \in H} \mathrm{Err}^{Q}(h)] \\
&= \Pr_{S \sim F^{-1}(Q)^m}[\mathrm{Err}^{Q}(B(S,D) \circ F^{-1}) - \inf_{h \in H} \mathrm{Err}^{F^{-1}(Q)}(h)] \\
&= \Pr_{S \sim F^{-1}(Q)^m}[\mathrm{Err}^{F^{-1}(Q)}(B(S,D)) - \inf_{h \in H} \mathrm{Err}^{F^{-1}(Q)}(h)] \,.
\end{aligned}$$

Therefore, for any $\epsilon, \delta > 0$,

$$\begin{aligned}
m(A,H,Q,\epsilon,\delta) &= m(B,H,F^{-1}(Q),\epsilon,\delta) \\
&\leq \sup_{P \in \mathrm{Ext}(D)} m(B,H,P,\epsilon,\delta)
\end{aligned}$$

Taking supremum over $Q \in \mathrm{Ext}(U)$ finishes the proof. □

6.2 Sample Complexity of Learning Thresholds in the Realizable Case

In this section we consider learning the class of thresholds, $H = \{\mathbf{1}(-\infty, t] \,:\, t \in \mathbb{R}\}$, on the real line in the realizable setting and show that for absolutely continuous unlabeled distributions SSL has at most factor 2 advantage over SL in the sample complexity.

First, in Theorem 7, we show $\frac{\ln(1/\delta)}{\epsilon}$ upper bound on the sample complexity of supervised learning. This seems to be a folklore result. Second, we consider sample complexity of semi-supervised learning in the case when $\mathcal{D}(P)$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. In Theorems 8 and 9 we show that the sample complexity is between $\frac{\ln(1/\delta)}{2\epsilon} + O(\frac{1}{\epsilon})$ and $\frac{\ln(1/\delta)}{2.01\,\epsilon} - O(\frac{1}{\epsilon})$.[1] Ignoring the lower order terms, we see that the sample complexity of supervised learning is (asymptotically) at most 2-times larger than that of semi-supervised learning.

We will make use the following of two algorithms: supervised algorithm L and semi-supervised algorithm B proposed by Kääriäinen [17]. Both algorithms on a sample $S = ((x_1,y_2),(x_2,y_2),\ldots,(x_m,y_m))$ first compute

$$\begin{aligned}
\ell &= \max\{x_i \,:\, i \in \{1,2,\ldots,m\},\ y_i = 1\} \,, \\
r &= \min\{x_i \,:\, i \in \{1,2,\ldots,m\},\ y_i = 0\} \,.
\end{aligned}$$

Algorithm L simply outputs the hypothesis $\mathbf{1}(-\infty, \ell]$. Algorithm B makes use of its second input, distribution D. Provided that $\ell < r$, B computes $t'' = \sup\{t' \,:\, D((\ell,t']) \leq D((\ell,r])/2\}$ and outputs hypothesis $\mathbf{1}(-\infty, t]$.

[1] The 2.01 in the lower bound can be replaced by arbitrary number strictly greater than 2. This slight imperfection is a consequence of that the true dependence of the sample complexity on ϵ, in this case, is of the form $1/\ln(1-2\epsilon)$ and not $1/(2\epsilon)$.

Theorem 7 (SL upper bound). *Let H be the class of thresholds and L be the supervised learning algorithm defined above. For any D, for any $\epsilon, \delta > 0$, and any "target" $h \in H$,*

$$m(A, H, D_h, \epsilon, \delta) \leq \frac{\ln(1/\delta)}{\epsilon} .$$

Proof. Let $h = \mathbf{1}(-\infty, t)$ and let $s = \sup\{s : D((s,t]) \geq \epsilon\}$. The event $\mathrm{Err}^{D_h}(L(S)) \geq \epsilon$ occurs precisely when the interval $(s,t]$ does not contain any sample points. This happens with probability $(1 - D((s,t])))^m \leq (1-\epsilon)^m$. If $m \geq \frac{\ln(1/\delta)}{\epsilon}$, then $(1-\epsilon)^m \leq \exp(-\epsilon m) \leq \delta$. □

Theorem 8 (SSL upper bound). *Let H be the class of thresholds and B be the semi-supervised learning algorithm defined above. For any absolutely continuous distribution D over an open interval, any $\epsilon \in (0, \frac{1}{4})$, $\delta \in (0, \frac{1}{2})$, and any "target" $h \in H$,*

$$m(B, H, D_h, \epsilon, \delta) \leq \frac{\ln(1/\delta)}{2\epsilon} + \frac{\ln 2}{2\epsilon} .$$

Proof. By rescaling trick (Lemma 5 part (a)) we can assume that D is uniform over $(0,1)$. Fix $\epsilon \in (0, \frac{1}{4})$, $\delta \in (0, \frac{1}{2})$ and $h \in H$. We show that, for any $m \geq 2$,

$$\Pr_{S \sim D_h^m}[\mathrm{Err}^{D_h}(B(S, D_h)) \geq \epsilon] \leq 2(1-2\epsilon)^m , \qquad (5)$$

from which the theorem easily follows, since if $m \geq \frac{\ln(1/\delta)}{2\epsilon} + \frac{\ln 2}{2\epsilon}$, then $m \geq 2$ and $2(1-2\epsilon)^m \leq 2\exp(-2m\epsilon) \leq \delta$.

In order to prove (5), let $h = \mathbf{1}(-\infty, t]$ be the "target". Without loss of generality $t \in [0, \frac{1}{2}]$. With a little abuse, we assume that $\ell \in [0,t]$ and $r \in [t,1]$. For convenience, we define $a : [0,t] \to [t,1]$, $b : [0,t] \to [t,1]$ as $a(\ell) = \max(2t - \ell - 2\epsilon, t)$ and $b(\ell) = \min(2t - \ell + 2\epsilon, 1)$ respectively. It is easily verified that $\mathrm{Err}^{D_h}(B(S, D_h)) \leq \epsilon$ if and only if $r \in [a(\ell), b(\ell)]$.

We lower bound the probability of success

$$p = \Pr_{S \sim D_h^m}[\mathrm{Err}^{D_h}(B(S, D_h)) \leq \epsilon] .$$

There are two cases:

Case 1: If $t > 2\epsilon$, then we integrate over all possible choices of the rightmost positive example in S (which determines ℓ) and leftmost negative example in S (which determines r). There are $m(m-1)$ choices for the rightmost positive example and leftmost negative example. We have

$$p \geq p_1 = m(m-1) \int_0^t \int_{a(\ell)}^{b(\ell)} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell .$$

Case 2: If $t \leq 2\epsilon$, then we integrate over all possible choices of the rightmost positive example in S and leftmost negative example in S. Additionally we also consider samples without positive examples, and integrate over all possible choices of the leftmost (negative) example. We have

$$p \geq p_2 = m(m-1) \int_0^t \int_{a(\ell)}^{b(\ell)} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell + m \int_t^{2\epsilon} (1-r)^{m-1} \, \mathrm{d}r$$

Both cases split into further subcases.

Subcase 1a: If $t > 2\epsilon$ and $t + 4\epsilon \leq 1$ and $t + \epsilon \geq 1/2$, then $0 \leq 2t + 2\epsilon - 1 \leq t - 2\epsilon \leq t$ and

$$\begin{aligned}
p_1 &= m(m-1)\Big[\int_0^{2t+2\epsilon-1} \int_{a(\ell)}^{b(\ell)} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{2t+2\epsilon-1}^{t-2\epsilon} \int_{a(\ell)}^{b(\ell)} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{t-2\epsilon}^{t} \int_{a(\ell)}^{b(\ell)} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \Big] \\
&= m(m-1)\Big[\int_0^{2t+2\epsilon-1} \int_{2t-\ell-2\epsilon}^{1} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{2t+2\epsilon-1}^{t-2\epsilon} \int_{2t-\ell-2\epsilon}^{2t-\ell+2\epsilon} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{t-2\epsilon}^{t} \int_{t}^{2t-\ell+2\epsilon} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \Big] \\
&= 1 - \frac{1}{2}(1 - 2t - 2\epsilon)^m - \frac{1}{2}(-1 + 2t + 6\epsilon)^m - (1-2\epsilon)^m \\
&\geq 1 - 2(1-2\epsilon)^m .
\end{aligned}$$

Subcase 1b: If $t > 2\epsilon$ and $t + \epsilon \leq 1/2$, then $2t + 2\epsilon - 1 \leq 0 \leq t - 2\epsilon \leq t$ and

$$\begin{aligned}
p_1 &= m(m-1)\Big[\int_0^{t-2\epsilon} \int_{a(\ell)}^{b(\ell)} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{t-2\epsilon}^{t} \int_{a(\ell)}^{b(\ell)} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \Big] \\
&= m(m-1)\Big[\int_0^{t-2\epsilon} \int_{2t-\ell-2\epsilon}^{2t-\ell+2\epsilon} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{t-2\epsilon}^{t} \int_{t}^{2t-\ell+2\epsilon} (1 - r + \ell)^{m-2} \, \mathrm{d}r\mathrm{d}\ell \Big] \\
&= 1 - (1-2\epsilon)^m + \frac{1}{2}(1 - 2t - 2\epsilon)^m - \frac{1}{2}(1 - 2t + 2\epsilon)^m \\
&\geq 1 - \frac{3}{2}(1-2\epsilon)^m .
\end{aligned}$$

Subcase 1c: If $t > 2\epsilon$ and $t + 4\epsilon \geq 1$, then $0 \leq t - 2\epsilon \leq$

$2t + 2\epsilon - 1 \le t$, and

$$\begin{aligned}
p_1 &= m(m-1)\Bigg[\int_0^{t-2\epsilon}\int_{a(\ell)}^{b(\ell)}(1-r+\ell)^{m-2}\,\mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{t-2\epsilon}^{2t+2\epsilon-1}\int_{a(\ell)}^{b(\ell)}(1-r+\ell)^{m-2}\,\mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{2t+2\epsilon-1}^{t}\int_{a(\ell)}^{b(\ell)}(1-r+\ell)^{m-2}\,\mathrm{d}r\mathrm{d}\ell\Bigg] \\
&= m(m-1)\Bigg[\int_0^{t-2\epsilon}\int_{2t-\ell-2\epsilon}^{1}(1-r+\ell)^{m-2}\,\mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{t-2\epsilon}^{2t+2\epsilon-1}\int_{t}^{1}(1-r+\ell)^{m-2}\,\mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{2t+2\epsilon-1}^{t}\int_{t}^{2t-\ell+2\epsilon}(1-r+\ell)^{m-2}\,\mathrm{d}r\mathrm{d}\ell\Bigg] \\
&= 1-(1-2\epsilon)^m - \frac{1}{2}(1-2t+2\epsilon)^m - \frac{1}{2}(2t+2\epsilon-1)^m \\
&\ge 1-2(1-2\epsilon)^m\ .
\end{aligned}$$

Subcase 2a: If $t \le 2\epsilon$ and $t+\epsilon \ge 1/2$, then $t-2\epsilon \le 0 \le 2t+2\epsilon-1 \le t$ and

$$\begin{aligned}
p_2 &= m(m-1)\Bigg[\int_0^{2t+2\epsilon-1}\int_{a(\ell)}^{b(\ell)}(1-r+\ell)^{m-2}\,\mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{2t+2\epsilon-1}^{t}\int_{a(\ell)}^{b(\ell)}(1-r+\ell)^{m-2}\,\mathrm{d}r\mathrm{d}\ell\Bigg] \\
&\quad + m\int_t^{2\epsilon}(1-r)^{m-1}\,\mathrm{d}r \\
&= m(m-1)\Bigg[\int_0^{2t+2\epsilon-1}\int_{t}^{1}(1-r+\ell)^{m-2}\,\mathrm{d}r\mathrm{d}\ell \\
&\quad + \int_{2t+2\epsilon-1}^{t}\int_{t}^{2t-\ell+2\epsilon}(1-r+\ell)^{m-2}\,\mathrm{d}r\mathrm{d}\ell\Bigg] \\
&\quad + (1-t)^m - (1-2\epsilon)^m \\
&= 1-\frac{3}{2}(1-2\epsilon)^m - \frac{1}{2}(2t+2\epsilon-1)^m \\
&\ge 1-2(1-2\epsilon)^m\ .
\end{aligned}$$

Subcase 2b: If $t \le 2\epsilon$ and $t+\epsilon \le 1/2$, then $t-2\epsilon \le 0$, $2t+2\epsilon-1 \le 0$ and

$$\begin{aligned}
p_2 &= m(m-1)\int_0^{t}\int_{a(\ell)}^{b(\ell)}(1-r+\ell)^{m-2}\,\mathrm{d}r\mathrm{d}\ell \\
&\quad + m\int_t^{2\epsilon}(1-r)^{m-1}\,\mathrm{d}r \\
&= m(m-1)\int_0^{t}\int_{t}^{2t-\ell+2\epsilon}(1-r+\ell)^{m-2}\,\mathrm{d}r\mathrm{d}\ell \\
&\quad + (1-t)^m - (1-2\epsilon)^m \\
&= 1-\frac{3}{2}(1-2\epsilon)^m - \frac{1}{2}(1-2t-2\epsilon)^m \\
&\ge 1-2(1-2\epsilon)^m\ .
\end{aligned}$$

□

Theorem 9 (SSL lower bound). *For any (randomized) semi-supervised algorithm A, any $\epsilon \in (0, 0.001)$, any $\delta > 0$, any absolutely continuous probability distribution D over an open interval, there exists $h \in H$, such that*

$$m(A, H, D_h, \epsilon, \delta) \ge \frac{\ln(1/\delta)}{2.01\epsilon} - \frac{\ln 2}{2.01\epsilon}\ .$$

Proof. By rescaling trick (Lemma 6 part (a)) we can assume that D is uniform over $(0,1)$. Fix A, ϵ, δ. We show the existence of required h by a probabilistic argument. We consider picking t uniformly at random from $(0,1)$ and let $h = \mathbf{1}(-\infty, t]$. We prove that for any $m \ge 0$,

$$\mathop{\mathbb{E}}_{t}\Pr_{S\sim D_h^m}[\mathrm{Err}^{D_h}(A(S, D_h)) \ge \epsilon] \ge \frac{1}{2}(1-2\epsilon)^m$$

or equivalently

$$\mathop{\mathbb{E}}_{S\sim D_h^m}\Pr_{t}[\mathrm{Err}^{D_h}(A(S, D_h)) \ge \epsilon] \ge \frac{1}{2}(1-2\epsilon)^m\ . \tag{6}$$

To lower bound the left side, fix unlabeled points $0 \le x_1 \le x_2 \le \cdots \le x_m \le 1$. For convenience, let $x_0 = 0$ and $x_{m+1} = 1$. We claim that

$$\Pr_{t}[\mathrm{Err}^{D_h}(A(S, D_h)) \ge \epsilon] \ge \sum_{i=0}^{m}\max(x_{i+1} - x_i - 2\epsilon, 0)\ . \tag{7}$$

To prove that we also fix $i \in \{0, 1, 2, \ldots, m\}$ and restrict t to lie in the interval $(x_i, x_{i+1}]$. The labels in S are hence fixed. If we also fix the random bits used by A for random its internal randomization, the hypothesis $g = A(S, D_h)$ is fixed. It is not hard to see that regardless of g

$$\int_{x_i}^{x_{i+1}} \mathbf{1}\left\{t\ :\ \mathrm{Err}^{D_h}(g) \ge \epsilon\right\}\mathrm{d}t \ge \max(x_{i+1} - x_i - 2\epsilon, 0)\ ,$$

it follows from that the set $\{t\ :\ \mathrm{Err}^{D_h}(g) < \epsilon\}$ is contained in an interval of length at most 2ϵ. We obtain (7) by taking expectation over the random bits used by A and summing over all i.

In order to prove (6) we will compute expectation over $S \sim D_h^m$ of both sides of (7). Expectation of the left side of (7) equals to the left side of (6). The expectation of the right side of (7) is equal to

$$I_m = m!\underbrace{\int_0^{x_{m+1}}\int_0^{x_m}\int_0^{x_{m-1}}\cdots\int_0^{x_2}}_{m\text{ times}} \sum_{i=0}^{m}\max(x_{i+1} - x_i - 2\epsilon, 0)\ \mathrm{d}x_1\cdots\mathrm{d}x_{m-2}\mathrm{d}x_{m-1}\mathrm{d}x_m\ ,$$

since there are $m!$ equiprobable choices for the order of the points $x_1, x_2, \ldots, x_m$ among which we choose, without loss of generality, the one with $x_1 \le x_2 \le \cdots \le x_m$. We look at I_m as a function of x_{m+1} and we prove that

$$I_m(x_{m+1}) = (\max(x_{m+1} - 2\epsilon, 0))^{m+1}\ , \tag{8}$$

for any $m \ge 0$ and any $x_{m+1} \in [0, 1]$. The bound (6) follows from (8), since $I_m = I_m(1) = (1-2\epsilon)^{m+1} \ge \frac{1}{2}(1-2\epsilon)^m$

for $\epsilon \leq 1/4$. In turn, (8) follows, by induction on m, from the recurrence

$$I_m(x_{m+1}) = m \int_0^{x_{m+1}} I_{m-1}(x_m) + \max(x_{m+1} - x_m - 2\epsilon, 0) \cdot x_m^{m-1} \, \mathrm{d}x_m \ ,$$

which is valid for all $m \geq 1$. In the base case, $m = 0$, $I_0(x_1) = \max(x_1 - 2\epsilon, 0)$ trivially follows by definition. In the inductive case, $m \geq 1$, we consider two cases. First case, $x_{m+1} < 2\epsilon$, holds since $\max(x_{i+1} - x_i - 2\epsilon, 0) = 0$ and hence by definition $I_m(x_{m+1}) = 0$. In the second case, $x_{m+1} \geq 2\epsilon$, from the recurrence and the induction hypothesis we have

$$\begin{aligned} I_m(x_{m+1}) &= m \int_0^{x_{m+1}} (\max(x_m - 2\epsilon, 0))^m \\ &\quad + \max(x_{m+1} - x_m - 2\epsilon, 0) \cdot x_m^{m-1} \, \mathrm{d}x_m \\ &= m \int_{2\epsilon}^{x_{m+1}} (x_m - 2\epsilon)^m \, \mathrm{d}x_m \\ &\quad + m \int_0^{x_{m+1}-2\epsilon} (x_{m+1} - x_m - 2\epsilon) x_m^{m-1} \, \mathrm{d}x_m \\ &= \frac{m}{m+1}(x_{m+1} - 2\epsilon)^{m+1} \\ &\quad + \frac{1}{m+1}(x_{m+1} - 2\epsilon)^{m+1} \\ &= (x_{m+1} - 2\epsilon)^{m+1} \ . \end{aligned}$$

To finish the proof of the theorem, suppose $m < \frac{\ln(1/\delta)}{2.01\epsilon} - \frac{\ln 2}{2.01\epsilon}$. Then $\frac{1}{2}(1-2\epsilon)^m > \delta$, since

$$\ln\left(\frac{1}{2}(1-2\epsilon)^m\right) = -\ln 2 + m\ln(1-2\epsilon) > -\ln 2 - m(2.01\epsilon) > \ln\delta \ ,$$

where we have used that $\ln(1-2\epsilon) > -2.01\epsilon$ for any $\epsilon \in (0, 0.001)$. Therefore from (6), for at least one target $h = \mathbf{1}(-\infty, t]$, with probability greater than δ, algorithm A fails to output a hypothesis with error less than ϵ. □

Remark. The $\frac{\ln(1/\delta)}{2.01\,\epsilon} - O(\frac{1}{\epsilon})$ lower bound applies to supervised learning as well. However, we do not know of any supervised algorithm (deterministic or randomized) that has asymptotic sample complexity $c\frac{\ln(1/\delta)}{\epsilon}$ for any constant $c < 1$. For example, the randomized algorithm that outputs with probability $1/2$ the hypothesis $\mathbf{1}(-\infty, \ell]$ and with probability $1/2$ the hypothesis $\mathbf{1}(-\infty, r)$ still cannot achieve the SSL sample complexity. We conjecture that all supervised algorithms for learning thresholds on real line in the realizable setting have asymptotic sample complexity $\frac{\ln(1/\delta)}{\epsilon}$.

6.3 Sample Complexity in Agnostic Case

In this section, we show that even in the agnostic setting SSL does not have more than constant factor improvement over SL. We prove some lower bounds for some classes over the real line. We introduce the notion of a b-shatterable distribution, which intuitively, are distributions where there are b "clusters" that can be shattered by the concept class. The main lower bound of this section are for such distributions (see Theorem 14). We show how this lower bound results in tight sample complexity bounds for two concrete problems. The first is learning thresholds on the real line where we show a bound of $\Theta(\ln(1/\delta)/\epsilon^2)$. Then we show sample complexity of $\Theta\left(\frac{2d+\ln(1/\delta)}{\epsilon^2}\right)$ for the union of d intervals on the real line.

The sample complexity of the union of d intervals for a fixed distribution in a noisy setting has also been investigated by Gentile and Helmbold [15]. They show a lower bound of $\Omega\left(2d\log\frac{1}{\Delta}/(\Delta(1-2\eta)^2)\right)$ where Δ is the distance to the target that the learning algorithm should guarantee with high probability, and η is the probability of a wrong label appearing (see classification noise model of [1]). This notation implies that the difference in true error of target and the algorithm's output is $\epsilon = (1-2\eta)\Delta$. Setting $\eta = 1/2 - \epsilon/4$ gives $\Omega(2d/\epsilon^2)$. We note that we do not make the assumption of a constant level of noise for each unlabeled example. It turns out, however, that in our proofs we do construct worst case distributions that have a constant noise rate that is slightly below $1/2$.

We point out two main differences between our results and that of Gentile and Helmbold. The first being that we explicitly construct noisy distributions to obtain ϵ^2 in the denominator. The second difference is that our technique appears to be quite different from theirs, which uses an information theory approach, whereas we make use of known techniques based on lower bounding how well one can distinguish similar noisy distributions, and then applying an averaging argument. The main tools used in this section come from Anthony and Bartlett [2, Chapter 5].

We first cite a result on how many examples are needed to distinguish two similar, Bernoulli distributions in Lemma 10. Then in Lemma 11 we prove an analogue of this for arbitrary unlabeled distributions. The latter result is used to give us a lower bound in Theorem 14 for b-shatterable distributions (see Definition 13). Corollary 12 and 15 gives us tight sample complexity bounds for thresholds and union of intervals on $\mathbb{R}$.

Lemma 10 (Anthony and Bartlett [2]). *Suppose that P is a random variable uniformly distributed on $\{P_1, P_2\}$ where P_1, P_2 are Bernoulli distributions over $\{0,1\}$ with $P_1(1) = 1/2 - \gamma$ and $P_2(1) = 1/2 + \gamma$ for $0 < \gamma < 1/2$. Suppose that $\xi_1, \ldots, \xi_m$ are IID $\{0,1\}$ valued random variables with $\Pr(\xi_i = 1) = P(1)$ for each i. Let f be a function from $\{0,1\}^m \to \{P_1, P_2\}$. Then*

$$\mathop{\mathbb{E}}_{P} \Pr_{\xi \sim P^m}[f(\xi) \neq P] > \frac{1}{4}\left(1 - \sqrt{1 - \exp\left(\frac{-4m\gamma^2}{1-4\gamma^2}\right)}\right) =: F(m,\gamma).$$

One can view the lemma this way: if one randomly picks two weighted coins with similar biases, then there's a lower bound on the confidence with which one can accurately predict the coin that was picked.

The next result is similar except an unlabeled distribution D is fixed, and the distributions we want to distinguish will be extensions of D.

Lemma 11. *Fix any X, H, D over X, and $m > 0$. Suppose there exists $h, g \in H$ with $D(h\Delta g) > 0$. Let P_h and P_g be the extension of D such that $P_h((x, h(x))|x) = P_g((x, g(x))|x) = 1/2+\gamma$. Let $A_D : (h\Delta g \times \{0,1\})^m \to H$ be any function. Then for any $x_1, \ldots, x_m \in h\Delta g$, there exists $P \in \{P_h, P_g\}$ such that if $y_i \sim P_{x_i}$ for all i,*

$$\Pr_{y_i}[\mathrm{Err}^P(A_D((x_1,y_1),\ldots,(x_m,y_m))) - OPT_P > \gamma D(h\Delta g)] > F(m,\gamma) .$$

Where P_x is the conditional distribution of P given x, and $OPT_P = 1/2 - \gamma$. Thus if the probability of failure is at most δ, we require

$$m \geq \left(\frac{1}{4\gamma^2} - 1\right) \ln \frac{1}{8\delta}. \tag{9}$$

Proof. Suppose for a contradiction this is not true. Let $\mathcal{P} = \{P_h, P_g\}$. Then there exists an A_D and $x_1, \ldots, x_m$ such that

$$\forall P \in \mathcal{P},\ \Pr_{y_i}[\mathrm{Err}^P(A_D((x_1,y_1),\ldots,(x_m,y_m))) - OPT_P > \gamma D(h\Delta g)] \leq F(m,\gamma). \tag{10}$$

Then we will show that the lower bound in Lemma 10 can be violated. Now $h\Delta g$ can be partitioned into $\Delta_0 = \{x : h(x) = 0\}$ and $\Delta_1 = \{x : h(x) = 1\}$. Without loss of generality assume $\{x_1, \ldots, x_l\} \subseteq \Delta_0$ and $\{x_{l+1}, \ldots, x_m\} \subseteq \Delta_1$. Let $A = A_D((x_1,y_1),\ldots,(x_m,y_m))$.

From the triangle inequality $D(A\Delta h) + D(A\Delta g) \geq D(h\Delta g)$. Thus if A is closer to h then $D(A\Delta g) \geq D(h\Delta g)/2$ and vice versa. Let P be a random variable uniformly distributed on $\mathcal{P}$. We have $\Pr(y_1 = 1) = \cdots = \Pr(y_l = 1) = P_{\Delta_0}(1) = \Pr(y_{l+1} = 0) = \cdots = \Pr(y_m = 0) = P_{\Delta_1}(0)$.

Let $\xi_1, \ldots, \xi_m \sim P_{\Delta_0}$ so that $\Pr(\xi_i = 1) = 1/2 - \gamma$ when $P = P_h$ and equal to $1/2 + \gamma$ when $P = P_g$. Let us define the function $f : \{0,1\}^m \to \mathcal{P}$ as follows. It will take as input $\xi_1, \ldots, \xi_m$ then transform this to an input of A_D as $I = (x_1, \xi_1), \ldots, (x_l, \xi_l), (x_{l+1}, 1-\xi_{l+1}), \ldots, (x_m, 1-\xi_m)$ so that ξ_i and $1 - \xi_j$ is from the same distribution as y_i and y_j, respectively, for $i \leq l, j > l$. Now define

$$f(\xi_1,\ldots,\xi_l) = \begin{cases} P_h & \text{if } D(A_D(I)\Delta h) < D(A_D(I)\Delta g) \\ P_g & \text{otherwise} \end{cases}.$$

We have

$$\begin{aligned} &\mathop{\mathbb{E}}_{P} \Pr_{\xi \sim P_{\Delta_0}^m}[f(\xi) \neq P] \\ &\quad \leq \mathop{\mathbb{E}}_{P} \Pr_{\xi}[D(A_D(I)\Delta OPT_P) > D(h\Delta g)/2] \\ &\quad \leq \mathop{\mathbb{E}}_{P} \Pr_{\xi}\left[\mathrm{Err}^P(A_D(I)) - OPT_P > \gamma D(h\Delta g)\right] \\ &\quad \leq F(m,\gamma) \end{aligned}$$

where the last inequality follows from (10). This is a contradiction, so the lower bound from Lemma 10 must apply. If the probability of failure $F(m,\gamma)$ is at most δ, solving the inequality for m gives (9). □

Corollary 12. *The SSL sample complexity of learning thresholds over the uniform distribution over $(0,1)$ is $\Theta(\ln(1/\delta)/\epsilon^2)$.*

Proof. Upper bound comes from any ERM algorithm. Let $h = \mathbf{1}(-\infty, 0]$ and $g = \mathbf{1}(-\infty, 1]$ so $D(h\Delta g) = 1$. Set $\gamma = \epsilon$ as in Lemma 11. □

Definition 13. *The triple (X, H, D) is b-shatterable if there exists disjoint sets $C_1, C_2, \ldots, C_b$ with $D(C_i) = 1/b$ for each i, and for each $S \subseteq \{1, 2, \ldots, b\}$, there exists $h \in H$ such that*

$$h \cap \left(\bigcup_{i=1}^{b} C_i\right) = \bigcup_{i \in S} C_i.$$

Theorem 14. *If (X, H, D) is b-shatterable and H contains h, g with $D(h\Delta g) = 1$ then a lower bound on the SSL sample complexity for $0 < \epsilon, \delta < 1/64$ is*

$$\Omega\left(\frac{b + \ln\frac{1}{\delta}}{\epsilon^2}\right).$$

Proof. The proof is similar to Theorem 5.2 in Anthony and Bartlett [2]. Let $G = \{h_1, h_2, \ldots, h_{2^b}\}$ be the class of functions that b-shatters D with respect to $C = \{C_1, \ldots, C_b\}$. We construct noisy extensions of D, $\mathcal{P} = \{P_1, P_2, \ldots, P_{2^b}\}$ so that for each i, $P_i((x, h_i(x))) = (1 + 2\gamma)/(2b)$. For any $h \in H$ let $\mathrm{snap}(h) = \mathrm{argmin}_{h' \in G} D(h\Delta h')$. Suppose $P \in \mathcal{P}$, let h^* denote the optimal classifier which is some $g \in G$ depending on the choice of P. If $i \neq j$ and $N(h_i, h_j)$ is the number of sets in C where h_i and h_j disagree, then $D(h_i\Delta h_j) \geq N(h_i, h_j)/b$, and since G is a $1/b$-packing,

$$\begin{aligned} \mathrm{Err}^P(h) &\geq \mathrm{Err}^P(h^*) + \frac{\gamma}{b} N(\mathrm{snap}(h), h^*) \\ &= \frac{1}{2}\left(\mathrm{Err}^P(\mathrm{snap}(h)) + \mathrm{Err}^P(h^*)\right). \end{aligned} \tag{11}$$

Modifying the proof of Anthony and Bartlett with the use of Lemma 11 rather than Lemma 10 we get that there exists a $P \in \mathcal{P}$ such that whenever $m \leq b/(320\epsilon^2)$,

$$\Pr_{S \sim P^m}\left[\mathrm{Err}^P(\mathrm{snap}(A(D,S))) - \mathrm{Err}^P(h^*) > 2\epsilon\right] > \delta.$$

Whenever A fails, we get from (11)

$$\begin{aligned} &\mathrm{Err}^P(A(D,S)) - \mathrm{Err}^P(h^*) \\ &\qquad \geq \frac{1}{2}\left(\mathrm{Err}^P(\mathrm{snap}(h)) + \mathrm{Err}^P(h^*)\right) \geq \epsilon. \end{aligned}$$

To get $\Omega(\ln(1/\delta)/\epsilon^2)$, apply Lemma 11 with h and g. □

We will now apply the above theorem to give the sample complexity for learning union of intervals on the real line. Recall that by the rescaling trick, we only need to consider the sample complexity with respect to the uniform distribution on $(0,1)$.

Corollary 15. *The SSL sample complexity for learning the class of union of at most d intervals $UI_d = \{[a_1, a_2) \cup \cdots \cup [a_{2l-1}, a_{2l}) : l \leq d, 0 \leq a_1 \leq a_2 \leq \cdots \leq a_{2l} \leq 1\}$ over uniform distribution on $(0,1)$ is*

$$\Theta\left(\frac{2d + \ln\frac{1}{\delta}}{\epsilon^2}\right).$$

Proof. We have $\mathrm{VC}(UI_d) = 2d$, thus the upper bound follows immediately. Construct $2d$-shatterable sets by letting $C_i = [(i-1)/2d, i/2d)$ for $i = 1, \ldots, 2d$. For any $S \subseteq \{1, \ldots, 2d\}$ define $h_S = \bigcup_{i \in S} C_i$. Now if $|S| \leq d$ then clearly $h_S \in UI_d$, if $|S| > d$ then $h_{\overline{S}} \in UI_d$ since $|\overline{S}| < d$. But then $[0,1) \backslash h_{\overline{S}}$ can be covered by at most d intervals, so $h_S \in UI_d$. Thus the set $\{h_S : S \subseteq \{1, \ldots, 2d\}\}$ $2d$-shatters D on $[0,1]$. Also let $h = [0,0) = \emptyset$ and $g = [0,1)$. Now apply Theorem 14 for the bound. □

6.4 No Optimal Semi-Supervised Algorithm

One could imagine a different formulation of the comparison between SL and SSL paradigms. For example, one might ask naively whether, for given class H, there is a semi-supervised algorithm A, such that for any supervised algorithm B, and any ϵ, δ, on any probability distribution P the sample complexity of A is no higher than the sample complexity of B. The answer to the question is easily seen to be negative, because for any P there exists a supervised learning algorithm B_P that ignores the labeled examples and simply outputs hypothesis $h \in H$ with minimum error $\mathrm{Err}^P(h)$ (or even Bayes optimal classifier for P). On P the sample complexity of B_P is zero, unfortunately, on P', sufficiently different from P, the sample complexity of B_P is infinite.

One might disregard algorithms such as B_P and ask the same question as above, except that one quantifies over only the subset of algorithms that on *any* distribution over $X \times \{0,1\}$ have sample complexity that is polynomial in $1/\epsilon$ and $\ln(1/\delta)$. Such algorithms are often called PAC (Probably Approximately Correct). The following theorem demonstrates that such restriction does not help and the answer to the question is still negative.

Theorem 16. *Let $H = \{\mathbf{1}(-\infty, t] : t \in \mathbb{R}\}$ be the class of thresholds over the real line. For any absolutely continuous distribution D (with respect to Lebesgue measure on $\mathbb{R}$), any semi-supervised algorithm A, any $\epsilon > 0$ and $\delta \in (0, \frac{1}{2})$, there exists a distribution $P \in \mathrm{Ext}(D)$ and a supervised PAC learning algorithm B such that*

$$m(A, H, P, \epsilon, \delta) > m(B, H, P, \epsilon, \delta) \ .$$

Proof. Fix any A, D and m. Let L be the algorithm that chooses the left most empirical error minimizer, that is, on a sample S, L outputs $\mathbf{1}(-\infty, \ell]$, where

$$\ell = \inf \left\{ t \in \mathbb{R} \ : \ \mathrm{Err}^S(\mathbf{1}(-\infty, t]) = \min_{h' \in H} \mathrm{Err}^S(h') \right\} \ .$$

For any $h \in H$ we also define algorithm L_h, which outputs h if $\mathrm{Err}^S(h) = 0$, and otherwise L_h outputs $L(S)$. First, note that $L \equiv L_{\mathbf{1}\emptyset}$. Second, for any h, L_h outputs a hypothesis that minimizes empirical error, and since $\mathrm{VC}(H) = 1$, it is a PAC algorithm. Third, clearly the sample complexity of L_h on D_h is zero (regardless of ϵ and δ).

Theorem 9 shows that there exists $h \in H$ such that the sample complexity of A on D_h is positive, in fact, it is increasing as ϵ and δ approach zero. Thus there exists supervised algorithm $B = L_h$ with lower sample complexity than A. □

7 Conclusion

We provide a formal analysis of the sample complexity of semi-supervised learning compared to that of learning from labeled data only. We focus on bounds that do not depend on assumptions concerning the relationship between the labels and unlabeled data distribution.

Our main conclusion is that in such a setting semi-supervised learning has limited advantage. Formally, we show that for basic concept classes over the real line this advantage is never more than a constant factor of the sample size. We believe that this phenomena applies much more widely.

We also briefly address the error bounds under common assumptions on the relationship between unlabeled data and the labels. We demonstrate that even when such assumptions apply common SSL paradigms may be inferior to standard empirical risk minimization. We conclude that prior beliefs like the cluster assumption should be formulated more precisely to reflect the known practical merits of SSL.

The paper calls attention to and formalizes some natural fundamental questions about the theory-practice gap concerning semi-supervised learning. The major open question we raise is whether any semi-supervised learning algorithm can achieve sample size guarantees that are unattainable without access to unlabeled data. This is formalized in Conjectures 4 and 3.

Acknowledgements. We like to thank Nati (Nathan) Srebro and Vitaly Feldman for useful discussions.

References

[1] Dana Angluin and Philip D. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1987.

[2] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, January 1999.

[3] Maria-Florina Balcan and Avrim Blum. A PAC-style model for learning from labeled and unlabeled data. In *Proceedings of 18th Annual Conference on Learning Theory 2005*, pages 111–126. Springer, 2005.

[4] Maria-Florina Balcan and Avrim Blum. An augmented PAC model for semi-supervised learning. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, chapter 21, pages 61–89. MIT Press, September 2006.

[5] Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, 1991.

[6] Avrim Blum and Tom M. Mitchell. Combining labeled and unlabeled sata with co-training. In *COLT*, pages 92–100, 1998.

[7] R. Bruce. Semi-supervised learning using prior probabilities and em. In *IJCAI Workshop on Text Learning: Beyond Supervision*, August 2001.

[8] Vittorio Castelli. *The relative value of labeled and unlabeled samples in pattern recognition*. PhD thesis, Stanford University, Stanford, CA, December 1994.

[9] Vittorio Castelli and Thomas M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing param-

eter. *IEEE Transactions on Information Theory*, 42(6):2102–2117, 1996.

[10] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. MIT Press, September 2006.

[11] Fabio Cozman and Ira Cohen. Risks of semi-supervised learning: How unlabeled data can degrade performance of generative classifiers. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, chapter 4, pages 57–72. MIT Press, September 2006.

[12] Sanjoy Dasgupta, Michael L. Littman, and David A. McAllester. Pac generalization bounds for co-training. In *NIPS*, pages 375–382, 2001.

[13] Ran El-Yaniv and Dmitry Pechyony. Stable transductive learning. In *COLT*, pages 35–49, 2006.

[14] Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. In *COLT*, pages 157–171, 2007.

[15] Claudio Gentile and David P. Helmbold. Improved lower bounds for learning from noisy examples: and information-theoretic approach. In *Proceedings of COLT 1998*, pages 104–115. ACM, 1998.

[16] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209, 1999.

[17] Matti Kääriäinen. Generalization error bounds using unlabeled data. In *Proceedings of COLT 2005*, pages 127–142. Springer, 2005.

[18] Joel Ratsaby and Santosh S. Venkatesh. Learning from a mixture of labeled and unlabeled examples with parametric side information. In *COLT*, pages 412–417, 1995.

[19] F. Oles T. Zhang. A probability analysis on the value of unlabeled data for classification problems. In *ICML*, pages 1191–1198, 2000.

[20] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, September 1998.

[21] Vladimir N. Vapnik. Transductive inference and semi-supervised learning. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, chapter 24, pages 453–472. MIT Press, September 2006.

[22] Xiaojin Zhu. *Semi-supervised learning with graphs*. PhD thesis, Carnegie Mellon University, 2005.

[23] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Science, University of Wisconsin Madison, 2007.