Unlabeled Data Does Provably Help *

Malte Darnstädt[1], Hans Ulrich Simon[1], and Balázs Szörényi[2,3]

1 **Department of Mathematics, Ruhr-University Bochum, Germany**
{malte.darnstaedt,hans.simon}@rub.de
2 **INRIA Lille, SequeL project, France**
3 **MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary**
szorenyi@inf.u-szeged.hu

Abstract

A fully supervised learner needs access to correctly labeled examples whereas a semi-supervised learner has access to examples part of which are labeled and part of which are not. The hope is that a large collection of unlabeled examples significantly reduces the need for labeled-ones. It is widely believed that this reduction of "label complexity" is marginal unless the hidden target concept and the domain distribution satisfy some "compatibility assumptions". There are some recent papers in support of this belief. In this paper, we revitalize the discussion by presenting a result that goes in the other direction. To this end, we consider the PAC-learning model in two settings: the (classical) fully supervised setting and the semi-supervised setting. We show that the "label-complexity gap" between the semi-supervised and the fully supervised setting can become arbitrarily large for concept classes of infinite VC-dimension (or sequences of classes whose VC-dimensions are finite but become arbitrarily large). On the other hand, this gap is bounded by $O(\ln |C|)$ for each finite concept class C that contains the constant zero- and the constant one-function. A similar statement holds for all classes C of finite VC-dimension.

1998 ACM Subject Classification I.2.6 Concept Learning

Keywords and phrases algorithmic learning, sample complexity, semi-supervised learning

Digital Object Identifier 10.4230/LIPIcs.xxx.yyy.p

1 Introduction

In the PAC[1]-learning model [11], a learner's input are samples, labeled correctly according to an unknown target concept, and two parameters $\varepsilon, \delta > 0$. He has to infer, with high probability of success, an approximately correct binary classification rule, which is called "hypothesis" in this context. In the non-agnostic setting (that we focus on in this paper), the following assumptions are made:

- There is a concept class C (known to the learner) so that the "correct" labels are assigned to the instances x from the underlying domain X by a function $c : X \to \{0, 1\}$ from C (the unknown target function).
- There is a probability distribution P on X (unknown to the learner) so that the samples (labeled according to c) are independently chosen at random according to P.

The learner is considered successful if his hypothesis h satisfies $P[h(x) \neq c(x)] < \varepsilon$ (approximate correctness).[2] The probability for success should be larger than $1 - \delta$ (so the learner's

* This work was supported by the bilateral Research Support Programme between Germany (DAAD 50751924) and Hungary (MÖB 14440).

[1] PAC = Probably Approximately Correct

[2] Note, that we don't require the learner to observe that his hypothesis is accurate to be successful.

Conference title on which this volume is based on.
Editors: Billy Editor, Bill Editors; pp. 1–12

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

hypothesis is probably approximately correct). The learner is called *proper* if he commits himself to picking his hypothesis from C. We refer to ε as the *accuracy parameter*, or simply as the *accuracy*, and we refer to δ as the *confidence parameter*, or simply as the *confidence*.

Providing a learner with a large collection of labeled samples is expensive because reliable classification labels are typically generated by a human expert. On the other hand, unlabeled samples are easy to get (e.g., can be collected automatically from the web). This raised the question whether the "label complexity" of a learning problem can be significantly reduced when learning is "semi-supervised", i.e., if the learner is not only provided with labeled samples but also with unlabeled-ones.[3] The existing analysis of the semi-supervised setting can be summarized roughly as follows:

- The benefit of unlabeled samples can be enormous if the target concept and the domain distribution satisfy some suitable "compatibility assumptions" (see [1]).
- On the other hand, the benefit seems to be marginal if we do not impose any extra-assumptions (see [2, 7]).

These findings perfectly match with the common belief that some kind of compatibility between the target concept and the domain distribution is needed for adding horsepower to semi-supervised algorithms. However, the results of the second type are not yet fully convincing:

- The paper [2] provides some upper bounds on the label complexity in the fully supervised setting and some lower bounds, that match up to a small constant factor, in the semi-supervised setting (or even in the setting with a distribution P that is known to the learner). These bounds however are established only for some special concept classes over the real line. It is unclear whether they generalize to a broader variety of concept classes.
- The paper [7] analyzes arbitrary finite concept classes and shows the existence of a purely supervised "smart" PAC-learning algorithm whose label consumption exceeds the label consumption of the best learner with full prior knowledge of the domain distribution at most by a constant factor for the "vast majority" of pairs (c, P). This however does not exclude the possibility that there still exist "bad pairs" (c, P) leading to a poor performance of the smart learner.

In this paper, we reconsider the question whether unlabeled samples can be of significant help to a learner even when we do not impose any extra-assumptions on the PAC-learning model. A comparably old paper, [8], indicates that an affirmative answer to this question is thinkable (despite of the fact that it was written a long time before semi-supervised learning became an issue). In [8] it is shown that there exists a concept class C_∞ and a family $\mathcal{P}_\infty$ of domain distributions such that the following holds:

1. For each $P \in \mathcal{P}_\infty$, C_∞ is properly PAC-learnable under the fixed distribution P (where "fixed" means that the learner has full prior knowledge of P).
2. C_∞ is *not properly* PAC-learnable under unknown distributions taken from $\mathcal{P}_\infty$.

These results point into the right direction for our purpose, but they are not precisely what we want:

- Although "getting a large unlabeled sample" comes close to "knowing the domain distribution", it is not quite the same. (In fact, one can show that C_∞, with domain distributions taken from $\mathcal{P}_\infty$, is *not* PAC-learnable in the semi-supervised setting.)

[3] In contrast to the setting of "active learning", we do however not assume that the learner can actively decide for which samples the labels are uncovered.

- The authors of [8] do *not* show that C_∞ is *not PAC-learnable* under unknown distributions P taken from $\mathcal{P}_\infty$. In fact, their proof uses a target concept that almost surely (w.r.t. P) assigns 1 to every instance in the domain. But the (proper!) learner must not return the constant 1-function of error 0 because of his commitment to hypotheses from C_∞.

Main Results:

The precise statement of our main results requires some more notation. For any concept class C over domain X and any domain distribution P, let $m_{C,P}(\varepsilon,\delta)$ denote the smallest number of labeled samples (in dependence of the accuracy ε and the confidence δ) needed to PAC-learn C under fixed distribution P. For any concept class C and any (semi-supervised or fully supervised) PAC-learning algorithm A, let $m^A_{C,P}(\varepsilon,\delta)$ denote the smallest number of labeled samples such that the resulting hypothesis of A is ε-accurate with confidence δ provided that P, unknown to A, is the underlying domain distribution. We first investigate the conjecture (up to minor differences identical to Conjecture 4 in [2])[4] that there is a purely supervised learner whose label consumption exceeds the label consumption of the best learner with full prior knowledge of the domain distribution at most by a factor $k(C)$ that depends on C only, as opposed to a dependence on ε or δ. The following result, whose proof is found in Section 3.1, confirms this conjecture to a large extent for finite classes, and to a somewhat smaller extent for classes of finite VC-dimension:

▶ **Theorem 1.** *Let C be a concept class over domain X that contains the constant zero- and the constant one-function. Then:*

1. *If C is finite, there exists a fully supervised PAC-learning algorithm A such that, for every domain distribution P, $m^A_{C,P}(2\varepsilon,\delta) = O(\ln|C|)\cdot m_{C,P}(\varepsilon,\delta)$.*
2. *If the VC-dimension of C is finite, there exists a fully supervised PAC-learning algorithm A such that, for every domain distribution P, $m^A_{C,P}(2\varepsilon,\delta) = O(\mathrm{VCdim}(C)\cdot\log(1/\varepsilon))\cdot m_{C,P}(\varepsilon,\delta) = \tilde{O}(\mathrm{VCdim}(C))\cdot m_{C,P}(\varepsilon,\delta)$.*

Can we generalize Theorem 1 to concept classes C of infinite VC-dimension provided that the domain distribution is taken from a family $\mathcal{P}$ such that $m_{C,P}(\varepsilon,\delta)<\infty$ for all $P\in\mathcal{P}$? This question will be answered to the negative by the following result (proved in Section 3.2):

▶ **Theorem 2.** *There exists a concept class C_* over domain $\{0,1\}^*$ and a family $\mathcal{P}_*$ of domain distributions such that the following holds:*

1. *There exists a semi-supervised algorithm A such that, for all $P\in\mathcal{P}_*$, $m^A_{C_*,P} = O(1/\varepsilon^2 + \log(1/\delta)/\varepsilon)$. (This implies the same upper bound on $m_{C_*,P}$ for all $P\in\mathcal{P}_*$.)*
2. *For every fully supervised algorithm A and for all $\varepsilon<1/2, \delta<1$: $\sup_{P\in\mathcal{P}_*} m^A_{C_*,P}(\varepsilon,\delta)=\infty$.*

Does there exist a universal constant k (not depending on C) such that we get a result similar to Theorem 1 but with $k(C)$ replaced by k? The following result (proved in Section 3.2) shows that, even for classes of finite VC-dimension, such a universal constant does not exist.

▶ **Theorem 3.** *There exists a sequence $(C_n)_{n\geq1}$ of concept classes over domains $(\{0,1\}^n)_{n\geq1}$ such that $\lim_{n\to\infty}\mathrm{VCdim}(C_n)=\infty$ and a sequence $(\mathcal{P}_n)_{n\geq1}$ of domain distribution families such that the following holds:*

1. *There exists a semi-supervised algorithm A that PAC-learns $(C_n)_{n\geq1}$ under any unknown distribution and, for all $P\in\mathcal{P}_n$, $m^A_{C_n,P}(\varepsilon,\delta) = O(1/\varepsilon^2+\log(1/\delta)/\varepsilon)$. (This implies the same upper bound on $m_{C_n,P}$ for all $P\in\mathcal{P}_n$.)*

[4] In contrast to [2], we allow the supervised learner to be twice as inaccurate as the semi-supervised learner because, otherwise, it can be shown that results in the manner of Theorem 1 are impossible even for simple classes.

2. *For every fully supervised algorithm A and all $\varepsilon < 1/2, \delta < 1$:* $\sup_{n\geq 1, P\in\mathcal{P}_n} m^A_{C_n,P}(\varepsilon,\delta) = \infty$.

Some comments are in place here:

- Since the class C_* from Theorem 2 has a countable domain, namely $\{0,1\}^*$, C_* occurs (via projection) as a subclass in every concept class that shatters a set of infinite cardinality. A similar remark applies to the sequence $(C_n)_{n\geq 1}$ and concept classes that shatter finite sets of arbitrary size. Thus every concept class of infinite VC-dimension contains subclasses that are significantly easier to learn in the semi-supervised setting of the PAC-model (in comparison to the full supervised setting).
- An error bound $\varepsilon = 1/2$ is trivially achieved by random guesses for the unknown label. Let α and β be two arbitrary small, but strictly positive, constants. Theorems 2 and 3 imply that even the modest task of returning, with a success probability of at least α, a hypothesis of error at most $1/2 - \beta$ cannot be achieved in the fully supervised setting unless the number of labeled examples becomes arbitrarily large.
- Theorem 3 implies that the results from [2] do *not* generalize (from the simple classes discussed there) to arbitrary finite classes. It implies furthermore that the "bad pairs" (c, P) occurring in the main result from [7] are unavoidable and not an artifact of the analysis in that paper.
- C_n is not an artificially constructed or exotic class: it is in fact the class of non-negated literals over n boolean variables, which occurs as a subset of many popular concept classes (e.g. monomials, decision lists, half spaces). The class C_* is a natural generalization of C_n to the set of boolean strings of arbitrary length.
- The classes $C_*, \mathcal{P}_*$ from Theorem 2, defined in Section 3.2, are close relatives of the classes $C_\infty, \mathcal{P}_\infty$ from [8], but the adversary argument that we have to employ is much more involved than the corresponding argument in [8] (where the learner was assumed to be proper and had been fooled mainly because of his commitment to hypotheses from C_∞).

2 Definitions, Notations and Facts

For any $n \in \mathbb{N}$, we define $[n] = \{1, \ldots, n\}$. The symmetric difference between two sets A and B is denoted $A \oplus B$, i.e., $A \oplus B = (A \setminus B) \cup (B \setminus A)$. The indicator function $\mathbb{I}(\text{cond})$ yields 1 if "cond" is a true condition, and 0 otherwise.

2.1 Prerequisites from Probability Theory

Let X be an integer-valued random variable. As usual, a most likely value a for X is called a *mode* of X. In this paper, the largest integer that is a mode of X is denoted $\text{mode}(X)$. As usual, X is said to be *unimodal* if $\Pr[X = x]$ is increasing with x for all $x \leq \text{mode}(X)$, and decreasing with x for all $x \geq \text{mode}(X)$.

Let Ω be a space equipped with a σ-algebra of events and with a probability measure P. For any sequence $(A_n)_{n\geq 1}$ of events, $\limsup_{n\to\infty} A_n$ is defined as the set of all $\omega \in \Omega$ that occur in infinitely many of the sets A_n, i.e., $\limsup_{n\to\infty} A_n = \cap_{n=1}^\infty \cup_{m=n}^\infty A_m$. We briefly remind the reader of the Borel-Cantelli Lemma:

▶ **Lemma 4** ([9]). *Let $(A_n)_{n\geq 1}$ be a sequence of independent events, and let $A = \limsup_{n\to\infty} A_n$. Then $P(A) = 1$ if $\sum_{n=1}^\infty P(A_n) = \infty$, and $P(A) = 0$ otherwise.*

▶ **Corollary 5.** *Let $(A_n)_{n\geq 1}$ be a sequence of independent events such that $\sum_{n=1}^{\infty} P(A_n) = \infty$. Let $B_{k,n}$ be the set of all $\omega \in \Omega$ that occur in at least k of the events $A_1, \ldots, A_n$. Then, for any $k \in \mathbb{N}$, $\lim_{n\to\infty} P(B_{k,n}) = 1$.*

Proof. Note that $B_{k,n} \subseteq B_{k,n+1}$ for every n. Since probability measures are continuous from below, it follows that $\lim_{n\to\infty} P(B_{k,n}) = P(\cup_{n=1}^{\infty} B_{k,n})$. Since, obviously, $\limsup_{n\to\infty} A_n \subseteq \cup_{n=1}^{\infty} B_{k,n}$, an application of the Borel-Cantelli Lemma yields the result. ◀

The following result, which is a variant of the Central Limit Theorem for triangular arrays, is known in the literature as the Lindeberg-Feller Theorem:

▶ **Theorem 6** ([6])**.** *Let $(X_{n,i})_{n\in\mathbb{N},i\in[n]}$ be a (triangular) array of random variables such that*
1. *$\mathbb{E}[X_{n,i}] = 0$ for all $n \in \mathbb{N}$, $i = 1, \ldots, n$.*
2. *$X_{n,1}, \ldots, X_{n,n}$ are independent for every $n \in \mathbb{N}$.*
3. *$\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}[X_{n,i}^2] = \sigma^2 > 0$.*
4. *For each $\varepsilon > 0$, $\lim_{n\to\infty} s_n(\varepsilon) = 0$ where $s_n(\varepsilon) = \sum_{i=1}^{n} \mathbb{E}[X_{n,i}^2 \mathbb{I}(|X_{n,i}| \geq \varepsilon)]$.*

Then $\lim_{n\to\infty} P\left[a < \frac{1}{\sigma} \cdot \sum_{i=1}^{n} X_{n,i} < b\right] = \varphi(b) - \varphi(a)$ where φ denotes the density function of the standard normal distribution.

An easy padding argument shows that this theorem holds "mutatis mutandis" for triangular arrays of the form $(X_{n_k,i})$ where $i = 1, \ldots, n_k$ and $(n_k)_{k\geq 1}$ is an increasing and unbounded sequence of positive integers. (The limes is then taken for $k \to \infty$.) We furthermore note that, for the special case of independent Bernoulli variables $X_{n,i}$ with probability p_i of success, Theorem 6 applies to the triangular array $(X_{n,i} - p_i)/\sigma_n$ where $\sigma_n^2 = \sum_{i=1}^{n} p_i(1-p_i)$. (A similar remark applies to the more general case of bounded random variables.)

The following result is an immediate consequence of Theorem 6 (plus the remarks thereafter):

▶ **Lemma 7.** *Let $l(k) = o(\sqrt{k})$. Let $(n_k)_{k\geq 1}$ be an increasing and unbounded sequence of positive integers. Let $(p_{k,i})_{k\in\mathbb{N},i\in[n_k]}$ range over all triangular arrays of parameters in $[0,1]$ such that*

$$\forall k \in \mathbb{N} : \sum_{i=1}^{n_k} p_{k,i}(1-p_{k,i}) \geq k \ . \tag{1}$$

Let $(X_{k,i})_{k\in\mathbb{N},i\in[n_k]}$ be the corresponding triangular array of row-wise independent Bernoulli variables. Then the function h given by

$$h(k) = \sup_{(p_{k,i})} \sup_{s\in\{0,\ldots,n_k\}} P\left[\left|\sum_{i=1}^{n_k} X_{k,i} - s\right| < l(k)\right]$$

approaches 0 as k approaches infinity.

Proof. Assume for sake of contradiction that $\limsup_{k\to\infty} h(k) > 0$. Then there exist $(p_{k,i})$ satisfying (1) and $s_k \in \{0, \ldots, n_k\}$ such that

$$\limsup_{k\to\infty} P\left[\left|\sum_{i=1}^{n_k} X_{k,i} - s_k\right| < l(k)\right] > 0 \ . \tag{2}$$

The random variable $S_k = \sum_{i=1}^{n_k} X_{k,i}$ has mean $\mu_k = \sum_{i=1}^{n_k} p_{k,i}$ and variance $\sigma_k^2 = \sum_{i=1}^{n_k} p_{k,i} \cdot (1-p_{k,i}) \geq k$. The Lindeberg-Feller Theorem applied to the triangular array $\left(\frac{X_{k,i}-p_i}{\sigma_k}\right)$ yields

$$\lim_{k\to\infty} P\left[a < \frac{S_k - \mu_k}{\sigma_k} < b\right] = \varphi(b) - \varphi(a) \ . \tag{3}$$

For S_k to hit a given interval of length $2l(k)$ (like the interval $[s_k - l(k), s_k + l(k)]$ in (2)) it is necessary for $(S_k - \mu_k)/\sigma_k$ to hit a given interval of length $2l(k)/\sigma_k$. Note that $\lim_{k\to\infty} l(k)/\sigma_k = 0$ because $\sigma_k \geq \sqrt{k}$ and $l(k) = o(\sqrt{k})$. Thus the hitting probability approaches 0 as k approaches infinity. This contradicts to (2). ◀

For ease of later reference, we let $k(\beta)$ for $\beta > 0$ be a function such that $h(k) \leq \beta$ for all $k \geq k(\beta)$. (Such a function must exist according to Lemma 7.)

▶ **Corollary 8.** *With the notation and assumptions from Lemma 7, the following holds: the probability mass of the mode of $\sum_{i=1}^{n_k} X_{k,i}$ is at most β for all $k \geq k(\beta)$.*

The following result implies the unimodality of binomially distributed random variables:

▶ **Lemma 9** ([10]). *Every sum of independent Bernoulli variables (with possibly different probabilities of success) is unimodal.*

2.2 Prerequisites from Learning Theory

A *concept class C over domain X* is a family of functions from X to $\{0,1\}$. C is said to be *PAC-learnable with sample size $m(\varepsilon,\delta)$* if there exists a (possibly randomized) algorithm A with the following property. For every concept $c \in C$, for every distribution P on X, and for all $\varepsilon, \delta > 0$ and $m = m(\varepsilon,\delta)$, if $\vec{x} = (x_1, \ldots, x_m)$ is drawn at random according to P^m, $\vec{b} = (c(x_1), \ldots, c(x_m))$, and A is given access to $\varepsilon, \delta, \vec{x}, \vec{b}$, then, with probability greater than $1-\delta$, A outputs a hypothesis $h : X \to \{0,1\}$ such that $P[h(x) = c(x)] > 1 - \varepsilon$. We say that h is *ε-accurate* (resp. ε-inaccurate) if $P[h(x) = c(x)] > 1 - \varepsilon$ (resp. $P[h(x) \neq c(x)] \geq \varepsilon$). We say the learner *fails* when he returns an ε-inaccurate hypothesis. As mentioned in the introduction already, we refer to ε as the *accuracy* and to δ as the *confidence*. In this paper, we consider the following variations of the basic model:

Proper PAC-learnability: The hypothesis $h : X \to \{0,1\}$ must be a member of C.

PAC-learnability under a fixed distribution: P is fixed and known to the learner.

The semi-supervised setting: The input of the learning algorithm is augmented by a finite number (depending on the various parameters of the learning task) of unlabeled samples. All samples, labeled- and unlabeled-ones, are drawn independently from X according to the domain distribution P.

Note that PAC-learnability with sample size $m(\varepsilon,\delta)$ under a fixed distribution follows from PAC-learnability with sample size $m(\varepsilon,\delta)$ in the semi-supervised setting because, if A knows the domain distribution P, it can first generate sufficiently many unlabeled samples and then run a simulation of the semi-supervised learning algorithm.

Throughout the paper, a mapping from X to $\{0,1\}$ is identified with the set of instances from X that are mapped to 1. Thus, concepts are considered as mappings from X to $\{0,1\}$ or, alternatively, as subsets of X. (E.g., we may write $P(h \oplus c)$ instead of $P[h(x) \neq c(x)]$.) $X' \subseteq X$ is said to be *shattered by C* if $\{X' \cap c |\ c \in C\}$ coincides with the powerset of X'. The VC-dimension of C, denoted $\mathrm{VCdim}(C)$, is infinite if there exist arbitrarily large sets that are shattered by C, and it is the size of the largest set shattered by C otherwise. We remind the reader to the following well-known results:

▶ **Lemma 10** ([4]). *A finite class C is properly PAC-learnable by any consistent hypothesis finder from $\lceil \ln(|C|/\delta)/\varepsilon \rceil$ labeled samples.*

▶ **Lemma 11** ([5]). *A class C of finite VC-dimension is properly PAC-learnable by any consistent hypothesis finder from $O((\mathrm{VCdim}(C) \cdot \log(1/\varepsilon) + \log(1/\delta))/\varepsilon)$ labeled samples.*

$C' \subseteq C$ is called an *ε-covering of C with respect to P* if for any $c \in C$ there exists $c' \in C'$ such that $P(c \oplus c') < \varepsilon$. The *covering number* $N_{C,P}(\varepsilon)$ is the size of the smallest ε-covering of C with respect to P. With this notation, the following holds:

▶ **Lemma 12** ([3]). *A concept class C is properly PAC-learnable under a fixed distribution P from $O(\log(N_{C,P}(\varepsilon/2)/\delta)/\varepsilon)$ labeled samples.*

A result by Balcan and Blum[5] implies the same upper bound on the label complexity for semi-supervised algorithms and concept classes of finite VC-dimension:

▶ **Lemma 13** ([1]). *Let C be a concept class of finite VC-dimension. Then C is PAC-learnable in the semi-supervised setting from $O(\mathrm{VCdim}(C)\log(1/\varepsilon)/\varepsilon^2 + \log(1/\delta)/\varepsilon^2)$ unlabeled and $O(\log(N_{C,P}(\varepsilon/6)/\delta)/\varepsilon)$ labeled samples.*

The following game between the learner and his adversary is useful for proving lower bounds on the sample size m:

Step 1: An "adversary" fixes a probability distribution D on pairs of the form (c, P) where $c \in C$ and P is a probability distribution on the domain X.

Step 2: The target concept c and the domain distribution P (representing the learning task) are chosen at random according to D.

Step 3: $(x_1, \ldots, x_m)$ is drawn at random according to P^m, and $\varepsilon, \delta, (x_1, \ldots, x_m), (b_1, \ldots, b_m)$ such that $b_i = c(x_i)$ is given as input to the learner.

Step 4: The adversary might give additional pieces of information to the learner.[6]

Step 5: The learner returns a hypothesis h. He "fails" if $P[h(x) \neq c(x)] \geq \varepsilon$.

This game differs from the PAC-learning model mainly in two respects. First, the learner is not evaluated against the pair (c, P) on which he performs worst but on a pair (c, P) chosen at random according to D (albeit D is chosen by an adversary). Second, the learner possibly obtains additional pieces of information in Step 4. Since both maneuvers can be to the advantage of the learner only, they do not compromise the lower bound argument. Thus, if we can show that, with probability at least δ, the learner fails in the above game, we may conclude that the sample size m does not suffice to meet the (ε, δ)-criterion of PAC-learning. Moreover, according to Yao's principle [12], lower bounds obtained by this technique even apply to randomized learning algorithms.

3 The Semi-supervised Versus the Purely Supervised Setting

This section is devoted to the proofs of our main results. The proof for Theorem 1 is presented in Section 3.1. The proofs for Theorems 2 and 3 are presented in Section 3.2.

3.1 Proof of Theorem 1

We start with the following lower bound on $m_{C,P}(\varepsilon, \delta)$:

▶ **Lemma 14.** *Let C be a concept class and let P be a distribution on domain X. For any $\varepsilon > 0$, let*

$$\lceil \varepsilon \rceil_{C,P} = \min\{\varepsilon' |\ (\varepsilon' \geq \varepsilon) \wedge (\exists c, c' \in C : P(c \oplus c') = \varepsilon'\}$$

[5] Apply Theorem 13 from [1] with a constant compatibility of 1 for all concepts and distributions.

[6] This step has purely proof-technical reasons: sometimes the analysis becomes simpler when the power of the learner is artificially increased.

where, by convention, the minimum of an empty set equals ∞. With this notation, the following holds:

1. *If $\lceil 2\varepsilon \rceil_{C,P} \leq 1$, then $m_{C,P}(\varepsilon, \delta) \geq 1$.*
2. *Let $\gamma = 1 - \lceil 2\varepsilon \rceil_{C,P}$. If $\lceil 2\varepsilon \rceil_{C,P} < 1$, then*

$$m_{C,P}(\varepsilon, \delta) \geq \log_{1/\gamma} \frac{1}{2\delta} = \Omega\left(\log_{1/\gamma} \frac{1}{\delta}\right) \quad . \tag{4}$$

3. *If $\lceil 2\varepsilon \rceil_{C,P} \leq 1/4$, then*

$$m_{C,P}(\varepsilon, \delta) \geq \left\lfloor \frac{\ln(1/(2\delta))}{2\lceil 2\varepsilon \rceil_{C,P}} \right\rfloor = \Omega\left(\frac{\ln(1/\delta)}{\lceil 2\varepsilon \rceil_{C,P}}\right) \quad . \tag{5}$$

Proof. It is easy to see that at least one labeled sample is needed if $\lceil 2\varepsilon \rceil_{C,P} \leq 1$. Let us now assume that $\lceil 2\varepsilon \rceil_{C,P} < 1$. Let $c, c' \in C$ be chosen such that $P(c \oplus c') = \lceil 2\varepsilon \rceil_{C,P}$. The adversary picks c and c' as target concept with probability $1/2$, respectively. With a probability of $(1 - \lceil 2\varepsilon \rceil_{C,P})^m$, none of the labeled samples hits $c \oplus c'$. Since $P(c \oplus c') \geq 2\varepsilon$, the learner has no hypothesis at his disposal that is ε-accurate for c and c'. Thus, if none the samples distinguishes between c and c', the learner will fail with a probability of $1/2$. We can conclude that the learner fails with an overall probability of at least $\frac{1}{2}(1 - \lceil 2\varepsilon \rceil_{C,P})^m = \frac{1}{2}\gamma^m$.

Setting this probability less than or equal to δ and solving for m leads to the lower bound (4). If $\lceil 2\varepsilon \rceil_{C,P} \leq 1/4$, a straightforward computation shows that $\frac{1}{2}\gamma^m$ is bounded from below by $\frac{1}{2}\exp(-2\lceil 2\varepsilon \rceil_{C,P} m)$. Setting this expression less than or equal to δ and solving for m leads to the lower bound (5). ◀

We are ready now for the **Proof of Theorem 1:**
We use the notation from Lemma 14. We first present the main argument under the (wrong!) assumption that $\lceil \varepsilon \rceil_{C,P}$ is known to the learner. At the end of the proof, we explain how a fully supervised learning algorithm can compensate for not knowing P. The first important observation, following directly from the definition of $\lceil \varepsilon \rceil_{C,P}$, is that, in order to achieve an accuracy of ε, it suffices to achieve an accuracy $\lceil \varepsilon \rceil_{C,P}$ with a hypothesis from C. Thus, for the purpose of Theorem 1, it suffices to have a supervised proper learner that achieves accuracy $\lceil 2\varepsilon \rceil_{C,P}$ with confidence δ. We proceed with the following case analysis:

Case 1: $\lceil 2\varepsilon \rceil_{C,P} \leq 1/4$.
There is a gap of $O(\ln |C|)$ only between the upper bound from Lemma 10 (with $\lceil 2\varepsilon \rceil_{C,P}$ in the role of ε) and the lower bound (5). Returning a consistent hypothesis, so that Lemma 10 applies, is appropriate in this case.

Case 2: $1/4 < \lceil 2\varepsilon \rceil_{C,P} < 15/16$.
We may argue similarly as in Case 1 except that the upper bound from Lemma 10 is compared to the lower bound (4). (Note that $\gamma = \theta(1)$ in this case.) As in Case 1, returning a consistent hypothesis is appropriate.

Case 3: $15/16 < \lceil 2\varepsilon \rceil_{C,P} < 1$.
In this case $0 < \gamma = 1 - \lceil 2\varepsilon \rceil_{C,P} < 1/16$. The learner will exploit the fact that one of the hypotheses $\emptyset$ and X is a good choice. He returns hypothesis X if label "1" has the majority within the labeled samples, and hypothesis $\emptyset$ otherwise. Let c, as usual, denote the target concept. If $\gamma < P(c) < 1 - \gamma$, then both of $\emptyset$ and X are $\lceil 2\varepsilon \rceil_{C,P}$-accurate. Let us assume that $P(c) \leq \gamma$. (The case $P(c) \geq 1 - \gamma$ is symmetric.) The learner will fail only if, despite of the small probability γ for label "1", these labels have the majority. It is easy to see that the probability for this to happen is bounded by $(m/2)\binom{m}{m/2}\gamma^{m/2}$ and therefore also bounded by $2^{3m/2}\gamma^{m/2} = (8\gamma)^{m/2}$. Setting the last expression less than

or equal to δ and solving for m reveals that $O(\log_{1/\gamma}(1/\delta))$ many labeled samples are enough. This matches the lower bound (4) modulo a constant factor.

Case 4: $\lceil 2\varepsilon \rceil_{C,P} = 1$.
This is a trivial case where each labeled sample almost surely makes inconsistent any hypothesis $h \in C$ of error at least ε. The learner may return any hypothesis that is supported by at least one labeled sample.

Case 5: $\lceil 2\varepsilon \rceil_{C,P} = \infty$.
This is another trivial case where any concept from C is 2ε-accurate with respect to any other concept from C. The learner needs no labeled example and may return any $h \in C$.

In any case, the "label-complexity" gap is bounded by $O(\ln |C|)$. We finally have to explain how this can be exploited by a supervised learner A who does not have any prior knowledge of P. The main observation is that, according to the bound in Lemma 10, the condition $m > \lceil \ln(|C|/\delta)/(15/16) \rceil$ indicates that the sample size is large enough to achieve an accuracy below 15/16 so that returning a consistent hypothesis is the appropriate action (as in Cases 1 and 2 above). If, on the other hand, the above condition on m is violated, then A will set either $h = \emptyset$ or $h = X$ depending on which label holds the majority (which would also be an appropriate choice in Cases 3 and 4 above). It is not hard to show that this procedure leads to the desired performance, which concludes the proof for the first part of Theorem 1.
As for the second part, one can use a similar argument that employs Lemma 11 instead of Lemma 10.

3.2 Proof of Theorems 2 and 3

Throughout this section, we set $X_n = \{0,1\}^n$ and $X_* = \{0,1\}^*$. We will identify a finite string $x \in X_*$ with the infinite string that starts with x and ends with an infinite sequence of zeros. C_* denotes the family of functions $c_i : X_* \to \{0,1\}$, $i \in \mathbb{N} \cup \{0\}$, given by $c_0(x) = 0$ and $c_i(x) = x_i$ for all $i \geq 1$. Note that $c_i(x) = 0$ for all $i > |x|$. C_n denotes the class of functions obtained by restricting a function from C_* to the subdomain X_n. For every $i \geq 1$, let $p_i = 1/\log(3+i)$. For every permutation σ of $1, \ldots, n$, let P_σ be the probability measure on X_n obtained by setting $x_{\sigma(i)} = 1$ with probability p_i (resp. $x_{\sigma(i)} = 0$ with probability $1 - p_i$) independently for $i = 1, \ldots, n$. $\mathcal{P}_n = \{P_\sigma\}$ denotes the family of all such probability measures on X_n. Note that P_σ can also be considered as a probability measure on X_* (that is centered on X_n). $\mathcal{P}_*$, a family of probability measures on X_*, is defined as $\cup_{n \geq 1} \mathcal{P}_n$.

▶ **Lemma 15.** *1. C_* is properly PAC-learnable under any fixed distribution $P_\sigma \in \mathcal{P}_*$ from $O(1/\varepsilon^2 + \log(1/\delta)/\varepsilon)$ labeled samples.*
2. For any (unknown) $P_\sigma \in \mathcal{P}_*$, C_* is properly PAC-learnable in the semi-supervised setting from $O(\log(n/\delta)/\varepsilon)$ unlabeled and $O(1/\varepsilon^2 + \log(1/\delta)/\varepsilon)$ labeled samples. Here, n denotes the smallest index such that $P_\sigma \in \mathcal{P}_n$.
3. There exists a semi-supervised algorithm A that PAC-learns C_n under any unknown domain distribution. Moreover, for all $P \in \mathcal{P}_n$, $m^A_{C_n,P}(\varepsilon, \delta) = O(1/\varepsilon^2 + \log(1/\delta)/\varepsilon)$.

Proof. 1. Let σ be a permutation of $1, \ldots, n$. For all $i > n$: $c_i = \emptyset$ almost surely w.r.t. P_σ. For all $2^{2/\varepsilon} - 3 \leq i \leq n$: $P_\sigma[c_{\sigma(i)} \oplus \emptyset] = P_\sigma[c_{\sigma(i)}] = p_i \leq \varepsilon/2$. Thus, setting $N = \lceil 2^{2/\varepsilon} \rceil - 4$, $\{\emptyset, c_{\sigma(1)}, \ldots, c_{\sigma(N)}\}$ forms an $\varepsilon/2$-covering of C_* with respect to P_σ. An application of Lemma 12 now yields the result.

2. The very first unlabeled sample reveals the parameter n such that the unknown measure P_σ is centered on X_n. Note that, for every $i \in [n]$, $x_i = 1$ with probability $p_{\sigma^{-1}(i)}$. It is an easy application of the multiplicative Chernov-bound (combined with the Union-bound) to see that $O(\log(n/\delta)/\varepsilon)$ unlabeled samples suffice to retrieve (with probability $1 - \delta/2$

of success) an index set $I \subset [n]$ with the following properties. On one hand, I includes all $i \in [n]$ such that $p_{\sigma^{-1}(i)} \geq \varepsilon/2$. On the other hand, I excludes all $i \in [n]$ such that $p_{\sigma^{-1}(i)} \leq \varepsilon/8$. Consequently $\{\emptyset\} \cup \{c_i |\ i \in I\}$ is an $\varepsilon/2$-covering of C_n with respect to P_σ and its size is bounded by $1 + |I| \leq 2^{8/\varepsilon}$. Another application of Lemma 12 now yields the result.

3. The third statement in Lemma 15 is an immediate consequence of Lemma 13 and the fact that, as proved above, $N_{C_n}(\varepsilon/6) = 2^{O(1/\varepsilon)}$ (regardless of the value of n). ◀

▶ **Lemma 16.** *Let A be a fully supervised algorithm designed to PAC-learn C_* under any unknown distribution taken from $\mathcal{P}_*$. For every finite sample size m and for all $\alpha, \beta > 0$, an adversary can achieve the following: with a probability of at least $1 - \alpha$ the hypothesis returned by A has an error of at least $1/2 - \beta$.*[7]

Proof. The proof will run through the following stages:

1. We first fix some technical notations and conditions (holding in probability) which the proof builds on.
2. Then we specify the strategy of the learner's adversary.
3. We argue that, given the strategy of the adversary, the learner has probably almost no advantage over random guesses.
4. We finally verify the technical conditions.

Let us start with Stage 1. (Though somewhat technical it will help us to provide a precise description of the subsequent stages.) Let $M \in \{0,1\}^{(m+1)\times(\mathbb{N}\setminus\{1\})}$ be a random matrix (with columns indexed by integers not smaller than 2) such that the entries are independent Bernoulli variables where the variable $M_{i,j}$ has probability $p_j = 1/\log(3+j) < 1/2$ of success. Let $M(n)$ denote the finite matrix composed of the first $n-1$ columns of M. Let $k = \max\{\lceil 1/\alpha \rceil, k(2\beta)\}$ where $k(\beta)$ is the function from the remark right after Lemma 7. In Stage 4 of the proof, we will show that there exists $n = n_k \in \mathbb{N}$ such that, with probability at least $1 - 1/k$, the following conditions are valid for each bit pattern $b \in \{0,1\}^{m+1}$:

(A) $b \in \{0,1\}^{m+1}$ coincides with at least $4k^2$ columns of $M(n)$.

(B) Let $b' \in \{0,1\}^m$ be the bit pattern obtained from b by omission of the final bit. Call column $j \geq 2$ of $M(n)$ "marked" if its first m bits yield pattern b'. Let $I \subseteq \{2,\ldots,n\}$ denote the set of indices for marked columns. Then, $\sum_{i\in I} p_i \geq 2k$ so that $\sum_{i\in I} p_i(1-p_i) \geq k$ (because $p_i < 1/2$).

The strategy of the adversary (Stage 2 of the proof) is as follows: she sets $n = n_k$, picks a permutation σ of $1,\ldots,n$ uniformly at random, chooses domain distribution P_σ, and selects the target concept c_t such that $t = \sigma(1)$. In the sequel, probabilities are simply denoted $P[\cdot]$. Note that the component x_t of a sample x can be viewed as a fair coin since $P[x_t = 1] = p_1 = 1/\log(4) = 1/2$. The learning task resulting from this setting is related to the technical definitions and conditions from Stage 1 as follows:

- The first m rows of the matrix $M(n)$ are the components $\sigma(2),\ldots,\sigma(n)$ of the m labeled samples.
- The bits of $b' \in \{0,1\}^m$ are the t-th components of the m labeled samples. These bits are perfectly random, and they are identical to the classification labels.
- The set $I \subseteq \{2,\ldots,n\}$ points to all marked columns of $M(n)$, i.e., it points to all columns of $M(n)$ which are duplicates of b'.
- Row $m+1$ of M represents an unlabeled test sample that has to be classified by the learner.

[7] Loosely speaking, the learner has "probably almost no advantage over random guesses".

The adversary passes also the set $J = \{\sigma(i)|\ i \in I \cup \{1\}\}$, with the understanding that index t of the target concept is an element of J, and the set $I \subseteq \{2, \ldots, n\}$ as additional information to the learner. This maneuver marks the end of Stage 2 in our proof.

We now move on to Stage 3 of the proof and explain why the strategy of the adversary leads to a poor learning performance (thereby assuming that conditions (A) and (B) hold). Note that, by symmetry, every index in J has the same a-posteriori probability to coincide with t. Because the learner has no way to break the symmetry between the indices in J before he sees the test sample x, the best prediction for the label of x does not depend on the individual bits in x but only on the number of ones in the bit positions from J, i.e., it only depends on the value of

$$Y' = \sum_{j \in J} x_j = x_{\sigma(1)} + \sum_{i \in I} x_{\sigma(i)} = x_{\sigma(1)} + Y \quad \text{where} \quad Y = \sum_{i \in I} x_{\sigma(i)} = \sum_{i \in I} M_{m+1,i} \ .$$

Note that the learner knows the distribution of Y (given by the parameters $(p_i)_{i \in I}$) since the set I had been passed on to him by the adversary. For sake of brevity, let $\ell = x_{\sigma(1)}$ denote the classification label of the test sample x. Given a value s of Y' (and the fact that the a-priori probabilities for $\ell = 0$ and $\ell = 1$ are equal), the Bayes decision is in favor of the label $\ell \in \{0, 1\}$ which maximizes $P[Y' = s|\ell]$. Clearly, $P[Y' = s|\ell = 1] = P[Y = s - 1]$ and $P[Y' = s|\ell = 0] = P[Y = s]$. Thus, the Bayes decision is in favor of $\ell = 0$ if and only if $P[Y = s] \geq P[Y = s - 1]$. Since Y is a sum of independent Bernoulli variables, we may apply Lemma 9 and conclude that Y has a unimodal distribution. It follows that the Bayes decision is the following threshold function: be in favor of $\ell = 0$ iff $Y' \leq \text{mode}(Y)$. The punchline of this discussion is as follows: the Bayes decision is independent of the true label ℓ unless Y hits its mode (so that $Y' = Y + \ell$ is either $\text{mode}(Y)$ or $\text{mode}(Y) + 1$). It follows that the Bayes error is at least $(1 - P[Y = \text{mode}(Y)])/2$. Because of Condition (B) and the fact that $k \geq k(2\beta)$, we may apply Corollary 8 and obtain $P[Y = \text{mode}(Y)] \leq 2\beta$ so that the Bayes error is at least $1/2 - \beta$.

We finally enter Stage 4 of the proof and show that conditions (A) and (B) hold with a probability of at least $1 - \alpha$ provided that $n = n_k$ is large enough. Let b range over all bit patterns from $\{0, 1\}^{m+1}$. Consider the events

- $A_r(b)$: $b \in \{0, 1\}^{m+1}$ coincides with the r-th column of M.
- $B_{k,n}(b)$: $b \in \{0, 1\}^{m+1}$ coincides with at least k columns of $M(n)$.

It is easy to see that $\sum_{r=1}^{\infty} P(A_r(b)) = \infty$. Applying the Borel-Cantelli Lemma to the events $(A_r(b))_{r \geq 1}$ and Corollary 5 to the events $(B_{4k^2,n}(b))_{k,n \geq 1}$, we arrive at the following conclusion. There exists $n_k(b) \in \mathbb{N}$ such that, for all $n \geq n_k(b)$, the probability of $B_{4k^2,n}(b)$ is at least $1 - 1/(2^{m+3}k)$. We set $n = n_k = \max_b n_k(b)$. Then, the probability of $B_{4k^2,n} = \bigcap_{b \in \{0,1\}^{m+1}} B_{4k^2,n}(b)$ is at least $1 - 1/(4k)$. In other words: with a probability of at least $1 - 1/(4k)$, each $b \in \{0, 1\}^{m+1}$ coincides with at least $4k^2$ columns of $M(n)$. Thus condition (A) is violated with a probability of at most $1/(4k)$.

We move on to condition (B). With $p = \sum_{i \in I} p_i$, we can decompose $P[B_{4k^2,n}]$ as follows:

$$P[B_{4k^2,n}] = P\left[B_{4k^2,n}\,|p < 2k\right] \cdot P\left[p < 2k\right] + P\left[B_{4k^2,n}\,|p \geq 2k\right] \cdot P\left[p \geq 2k\right]$$

Note that, according to the definitions of $B_{4k^2,n}(b)$ and $B_{4k^2,n}$, event $B_{4k^2,n}$ implies that $Y \geq 4k^2$ because there must be at least $4k^2$ occurrences of 1 in row $m + 1$ and in the marked columns of $M(n)$. On the other hand, $\mathbb{E}[Y] = p$. According to Markov's inequality, $P\left[Y \geq 4k^2|p < 2k\right] \leq (2k)/(4k^2) = 1/(2k)$. Thus, $P[B_{4k^2,n}] \leq 1/(2k) + P\left[p \geq 2k\right]$. Recall that, according to condition (A), $1 - 1/(4k) \leq P[B_{4k^2,n}]$. Thus, $P\left[p \geq 2k\right] \geq 1 - 1/(4k) - 1/(2k) = 1 - 3/(4k)$. Since $k \geq 1/\alpha$, we conclude that the probability to violate one of the conditions (A) and (B) is bounded by $1/k \leq \alpha$. ◀

We are now ready to complete the proofs of our main results. Theorem 2 is a direct consequence of the second statement in Lemma 15 and of Lemma 16. The first part of Theorem 3 is a direct consequence of the third statement in Lemma 15. As for the second part, an inspection of the proof of Lemma 16 reveals that the adversary argument uses a "finite part" C_n of C_* only (with n chosen sufficiently large).

4 Final Remarks:

As we have seen in this paper, it is impossible to show in full generality that unlabeled samples have a marginal effect only in the absence of any compatibility assumptions. It would be interesting to explore which concept classes are similar in this respect to the artificial classes C_* and $(C_n)_{n\geq 1}$ that were discussed in this paper. We would also like to know if the bounds of Theorem 1 are tight (either for special classes or for the general case). It would be furthermore interesting to extend our results to the agnostic setting.

References

1 Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the Association on Computing Machinery*, 57(3):19:1–19:46, 2010.

2 Shai Ben-David, Tyler Lu, and Dávid Pál. 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*, pages 33–44, 2008.

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

4 Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Information Processing Letters*, 24:377–380, 1987.

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

6 Kai Lai Chung. *A Course in Probability Theory*. Academic Press, 1974.

7 Malte Darnstädt and Hans U. Simon. Smart PAC-learners. *Theoretical Computer Science*, 412(19):1756–1766, 2011.

8 Richard M. Dudley, Sanjeev R. Kulkarni, Thomas J. Richardson, and Ofer Zeitouni. A metric entropy bound is not sufficient for learnability. *IEEE Transactions on Information Theory*, 40(3):883–885, 1994.

9 William Feller. *An Introduction to Probability Theory and its Applications*, volume 1. John Wiley & Sons, 1968.

10 Julian Keilson and Hans Gerber. Some results for discrete unimodality. *Journal of the American Statistical Association*, 66(334):386–389, 1971.

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

12 Andrew Yao. Probabilistic computations: Toward a unified measure of complexity. In *Proceedings of the 18th Symposium on Foundations of Computer Science*, pages 222–227, 1977.