The sample complexity of agnostic learning with deterministic labels

Shai Ben-David
Cheriton School of Computer Science
University of Waterloo
Waterloo, ON, N2L 3G1
CANADA
shai@uwaterloo.ca

Ruth Urner
College of Computing
School of Computer Science
Georgia Institute of Technology
Atlanta, Georgia 30303
USA
rurner@cc.gatech.edu

Abstract

We investigate agnostic learning when there is no noise in the labeling function, that is, the labels are *deterministic*. We show that in this setting, in contrast to the fully agnostic learning setting (with possibly noisy labeling functions), the sample complexity of learning a binary hypothesis class is not fully determined by the VC-dimension of the class. For any d, we present classes of VC-dimension d that are learnable from $O(d/\epsilon)$ many samples and classes that require samples of sizes $\Omega(d/\epsilon^2)$. Furthermore, we show that in this setting, there exist classes where ERM algorithms are not optimal: While the class can be learned with sample complexity $O(d/\epsilon)$, the convergence rate of any ERM algorithm is only $\Omega(d/\epsilon^2)$. We introduce a new combinatorial parameter of a class of binary valued functions and show that it provides a full combinatorial characterization of the sample complexity of deterministic - label agnostic learning of a class.

Introduction

We investigate the sample complexity of binary classification learning with respect to a concept class, when the labeling function is deterministic but does not necessarily belong to the class (agnostic learning of deterministic labels). As far as we are aware, this case has not been fully analyzed before. It is well known that in the PAC framework of machine learning theory, a class H of classifiers is learnable if and only if it has finite VC-dimension. The analysis of the sample complexity is then divided into two main cases, according to whether the class H contains a zero-error classifier or not.

In the first case, usually referred to as the *realizable case*, the sample complexity of learning H is known to be (roughly) $\Theta\left(\frac{\mathrm{VCdim}(H)}{\epsilon}\right)$. In the second case, referred to as the *agnostic case*, no such assumption is made. In particular, the labeling function of the underlying data-generating distribution is not necessarily deterministic. In this case the sample complexity is known to be (roughly) $\Theta\left(\frac{\mathrm{VCdim}(H)}{\epsilon^2}\right)$. Proving the lower bound for this case usually involves taking advantage of the stochasticity in the labeling function.

The work of Mammen and Tsybakov initiated one type of finer grained analysis (Mammen and Tsybakov 1999). Allowing stochastic labels, but assuming that the Bayes optimal classifier is a member of the class, they analyze the convergence of ERM classifiers and prove a bound on their error rates that interpolate between the realizable and the agnostic case. These bounds involve a parameter, usually denoted by α, that restricts the amount of noise in the labeling (Tsybakov noise condition). We discuss generalizations of this initial analysis in the related work section.

Table 1: Sample complexity of (PAC)-learning H

	realizable: $b \in H$	agnostic
deterministic labels	$\frac{\mathrm{VCdim}(H)}{\epsilon}$	**?**
probabilistic labels	under Tsybakov noise: $\frac{\mathrm{VCdim}(H)}{\epsilon^\alpha}$	$\frac{\mathrm{VCdim}(H)}{\epsilon^2}$

Table 1 illustrates the relationship between these three cases (realizable, fully agnostic, and realizable under Tsybakov noise condition). In this work, we turn the focus on the upper right corner: agnostic learning under deterministic labellings. We point out a curious gap in the label complexity of learning VC-classes with respect to the class of distributions with deterministic labeling functions. We show that, for any d, there exist classes of VC-dimension d that are learnable with sample complexity $O(d/\epsilon)$, regardless of the approximation error, and classes for which learning requires sample sizes of $\Omega(d/\epsilon^2)$. This is in contrast to the fully agnostic setting (the lower right corner) and the realizable case (upper left). In both these scenarios the sample complexity is fully determined by the VC-dimension of the hypothesis class. We introduce a simple combinatorial parameter of a class, the *class diameter* and prove that it fully characterize the distinction between classes with fast agnostic deterministic learning rates (where the sample complexity groans linearly with $1/\epsilon$) and those with slow rates ($m = \Omega(1/\epsilon^2)$). The class diameter is an upper bound on the class VC dimension but can be arbitrarily higher. It follows that the sample complexity of agnostically learning a class w.r.t. deterministic labelings behaves differently than the sample complexity w.r.t. arbitrary labelings.

The second issue that we investigate is the optimality

(from the point of view of sample complexity) of ERM learners in the setting of deterministic labelings. We show that, for every d, there exist classes of VC-dimension d, for which an ERM algorithm is not optimal. We show that some classes are learnable with sample complexity $O(d/\epsilon)$ while any ERM algorithm requires $1/\epsilon^2$ many samples (for agnostically learning these classes w.r.t deterministic labelings). This is again in contrast with the fully agnostic and the realizable setting, where the optimal rates are achieved by any ERM procedure for every hypothesis class. We provide full characterization of the classes that demonstrate such a sample complexity gap.

Our results open the way to several directions of more detailed analysis. In particular, it will be interesting to understand the sample complexity of learning classes as a function of some parameter of the data-generating distribution. This direction is analogous to the above mentioned analysis of learning under the Tsybakov noise condition.

Related work

The PAC framework for binary classification learning was first introduced by (Valiant 1984). (Blumer et al. 1989) characterize learnability of a binary hypothesis class in terms of its VC dimension. Essentially, this characterization goes back to (Vapnik and Chervonenkis 1971). The agnostic PAC model was introduced by (Haussler 1992). Those papers also provide the tight characterizations (up to logarithmic factors) mentioned in the introduction of the sample complexity in terms of the VC-dimension, for both the realizable and agnostic settings. In both cases, the sample complexity of any empirical risk minimization (ERM) learner is equal to that of the best possible learning algorithm.

The gap between the error rates in the two settings have attracted quite a lot of attention. Most notably, (Mammen and Tsybakov 1999) provide convergence bounds that interpolate between the realizable case and the agnostic case. They introduce the *Tsybakov noise condition*, a abound on the stochasticity (or noisiness) of the labels. They prove convergence bounds under this condition and the additional assumption that the Bayes optimal classifier is a member of the hypothesis class. Note that these bounds depend on parameters of the data-generating distribution, in contrast with the sample complexity bounds in the PAC and agnostic PAC frameworks that are *distribution-free*. (Tsybakov 2004) generalizes these results to the case where the Bayes classifier is only assumed to be a member of some collection of known hypothesis classes. (Boucheron, Bousquet, and Lugosi 2005) provide an analysis (under the Tsybakov noise condition) that does not impose restrictions on the Bayes classifier. However the obtained convergence rates depend on the approximation error of the class (they become weaker as the approximation error grows).

The setup where the labeling rule is deterministic, but yet does not belong to the learnt class has been addressed o a lesser degree. (Kääriäinen 2006) presents a lower bound of order $1/\epsilon^2$ for agnostic learning of any class that contains the two constant functions when the labeling is deterministic. However, these lower bounds do not grow with the complexity (e.g., VC-dimension) of the learnt class and their dependence on the domain size is not discussed there.

We are not aware of any previous work that shows lower bonds of order $\mathrm{VCdim}(H)/\epsilon^2$ on the sample complexity of learning deterministic labellings, or the upper bounds of order $\mathrm{VCdim}(H)/\epsilon$ that hold for arbitrary deterministic labellings, regardless of the approximation error of H, as the ones we present in this paper. Another aspect of this paper that does not seem to have been brought up by earlier work is the existence of classes for which there are learners with similarly low sample complexity (of order $\mathrm{VCdim}(H)/\epsilon$ for arbitrary deterministic-label distributions) in cases where any ERM learner requires order of $1/\epsilon^2$ sample sizes.

Definitions

We let $\mathcal{X}$ denote a *domain* set and let $\{0,1\}$ be the *label* set. A *hypothesis* (or *label predictor* or *classifier*), is a binary function $h : \mathcal{X} \to \{0,1\}$, and a *hypothesis class* H is a set of hypotheses. We model a *learning task* as some distribution P over $\mathcal{X} \times \{0,1\}$ that generates data. We denote the marginal distribution of P over $\mathcal{X}$ by $P_{\mathcal{X}}$ and let $l : \mathcal{X} \to [0,1]$ denote the induced conditional label probability function, $l(x) = P(y = 1|x)$. We call l the *labeling function* or *labeling rule* of the distribution P. We say that the labeling function is *deterministic*, if $l(x) \in \{0,1\}$ for all $x \in \mathcal{X}$. Otherwise, we call the labeling function *probabilistic*.

For some function $h : \mathcal{X} \to \{0,1\}$ we define the *error* of h with respect to P as

$$\mathrm{Err}_P(h) = \Pr_{(x,y)\sim P}[y \neq h(x)].$$

For a class H of hypotheses on $\mathcal{X}$, we let the smallest error of a hypothesis $h \in H$ with respect to P be denoted by

$$\mathrm{opt}_P(H) := \inf_{h\in H} \mathrm{Err}_P(h).$$

We call $\mathrm{opt}_P(H)$ the *approximation error* of the hypothesis class H with respect to P. We let $\mathrm{opt}_P = \inf_{h\in\{0,1\}^{\mathcal{X}}} \mathrm{Err}_P(h)$ denote the smallest possible error of any classifier over $\mathcal{X}$ with respect to P. We refer to opt_P as the *Bayes optimal error* of P. The *Bayes optimal classifier* b is defined as $b(x) = 1$ if $l(x) \geq 1/2$ and $b(x) = 0$ if $l(x) < 1/2$. The Bayes optimal classifier attains the minimum error opt_P. Note that, if the labeling function is deterministic, then $\mathrm{opt}_P = 0$.

Let $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \{0,1\})^n$ be a finite sequence of labeled domain points. We define the *empirical error* of a hypothesis with respect to S as

$$\mathrm{Err}_S(h) = \frac{1}{|S|} \sum_{(x,y)\in S} |y - h(x)|.$$

A *standard learner* $\mathcal{A}$ is an algorithm that takes a sequence $S = ((x_1, y_1), \ldots, (x_n, y_n))$ and outputs a hypothesis $h : X \to \{0,1\}$. Formally,

$$\mathcal{A} : \bigcup_{m=1}^{\infty} (\mathcal{X} \times \{0,1\})^m \to \{0,1\}^{\mathcal{X}}.$$

Definition 1 (Learnability)**.** Let $\mathcal{X}$ denote some domain. We say that an algorithm $\mathcal{A}$ *learns* some class of binary classifiers $H \subseteq \{0,1\}^{\mathcal{X}}$ with respect to a set of distributions $\mathcal{Q}$ over $\mathcal{X} \times \{0,1\}$, if there exists a function $m : (0,1) \times (0,1) \to \mathbb{N}$ such that, for all distributions $P \in \mathcal{Q}$, and for all $\epsilon > 0$ and $\delta > 0$, when given an *i.i.d.* sample of size at least $m(\epsilon, \delta)$ from P, then, with probability at least $1-\delta$ over the sample, $\mathcal{A}$ outputs a classifier $h : X \to \{0,1\}$ with error at most $\mathrm{opt}_P(H) + \epsilon$. In this case, for given ϵ and δ, we also say that the algorithm (ϵ, δ)-*learns* H with respect to $\mathcal{Q}$ from $m(\epsilon, \delta)$ examples.

For most of this paper we consider classes of distributions that have a deterministic labeling function. For a domain $\mathcal{X}$, we let $\mathcal{Q}^{\mathrm{det}}_{\mathcal{X}}$ denote the set of all such distributions. In the following definition we use the use the term "smallest function" to denote the pointwise smallest function.

Definition 2 (Sample Complexity)**.** We call the smallest function $m : (0,1) \times (0,1) \to \mathbb{N}$ that satisfies the condition of Definition 1 the *sample complexity of the algorithm $\mathcal{A}$ for learning H with respect to $\mathcal{Q}$*. We denote this function by $m[\mathcal{A}, \mathcal{Q}, H]$. We call the smallest function $m : (0,1) \times (0,1) \to \mathbb{N}$ such that there exists a learner $\mathcal{A}$ with $m[\mathcal{A}, \mathcal{Q}, H] \leq m$ the *sample complexity of learning H with respect to $\mathcal{Q}$* and denote this function by $m[\mathcal{Q}, H]$. We omit $\mathcal{Q}$ in this notation, when $\mathcal{Q}$ is the set of all distributions over $\mathcal{X} \times \{0,1\}$, and call $m[H]$ the *sample complexity of learning H*. For the set $\mathcal{Q}^{\mathrm{det}}_{\mathcal{X}}$ of distributions with deterministic labeling functions, we use the notation $m^{\mathrm{det}}[H]$.

Sample complexity gap between classes of the same VC dimension

We start by proving that, for every VC-dimension d, there exists classes that require samples of sizes $\Omega(d/\epsilon^2)$. This shows that restricting our attention to distributions with deterministic labeling functions does not necessarily render learning easier (as might be expected from the results on agnostic learning under the Tsybakov noise condition).

For any domain $\mathcal{X}$, we let $H_{1,0}$ be the hypothesis class that contains only the constant function 1 and the constant function 0. The following lemma establishes a lower bound on the sample complexity of learning this class and also provides an upper bound on the required domain size for this result. The lemma was shown in (Urner 2013), but has not been published before.

Lemma 3. *Let $0 < \epsilon < 1/4$ and $0 < \delta < 1/32$, let $\mathcal{X}$ be a finite domain of size at least $1/\epsilon^3$ and let $\mathcal{Q}$ be the set of distributions over $\mathcal{X} \times \{0,1\}$ whose marginal distribution $P_{\mathcal{X}}$ is uniform over $\mathcal{X}$ and whose labeling function deterministically labels a $(1/2-\epsilon)$-fraction of the points 0 and $(1/2+\epsilon)$-fraction of the points 1, or the other way around. Let $H_{1,0}$ be the hypothesis class that contains only the constant function 1 and the constant function 0. Then, $(\epsilon/2, \delta)$-learning H with respect to $\mathcal{Q}$ requires a sample size of $\Omega(1/\epsilon^2)$.*

Proof. For every distribution P in $\mathcal{Q}$ we have $\mathrm{opt}_P(H) = 1/2 - \epsilon$. Consider the majority algorithm $\mathcal{M}$ that, given a sample $S = ((x_1, y_1) \ldots (x_m, y_m))$, predicts with a function that agrees with the labels of the sample points on S and outside the sample predicts with the majority label in S. We will now first argue that, for every distribution $P \in \mathcal{Q}$, this algorithm needs to see $\Omega(1/\epsilon^2)$ many points to succeed at the task. Then we show that for any other learning algorithm $\mathcal{A}$, there exits a distribution in $\mathcal{Q}$ where $\mathcal{A}$ performs worse than $\mathcal{M}$. These two steps together imply the claim.

Step 1: Assume that the sample size is $|S| \leq \frac{1}{2\epsilon^2}$. Note that this corresponds to at most an $\epsilon/2$-fraction of the sample points. Thus, if $\mathcal{M}$ predicts (outside of S) with a label that is not the overall (true) majority label, then the error of $\mathcal{M}(S)$ is at least $1/2+\epsilon-|S|/|\mathcal{X}| \geq 1/2+\epsilon/2 > \mathrm{opt}_P(H)+\epsilon/2$. This implies that, for $\mathcal{M}$, $(\epsilon/2, \delta)$-learning H with respect to $\mathcal{Q}$ reduces to correctly learning what the majority label is, that is, it reduces to correctly predicting the bias of a coin. The lower bound in Lemma 5.1 by (Anthony and Bartlett 1999) now implies that $\mathcal{M}$ requires a sample larger than $\frac{1}{2\epsilon^2}$ for $\epsilon < 1/4$ and $\delta < 1/32$.

Step 2: Consider some algorithm $\mathcal{A}$ and assume that this algorithm $(\epsilon/2, \delta)$-learns H with respect to $\mathcal{Q}$ with samples of size m. Fix a sequence of m domain points $(x_1, \ldots, x_m)$. We now consider the expected performance of the learner $\mathcal{A}$ averaged over all distributions in $\mathcal{Q}$, given that the domain points in the sample are $S_{\mathcal{X}} = (x_1, \ldots, x_m)$. Recall that every distribution in $\mathcal{Q}$ has uniform marginal over $\mathcal{X}$, thus the different distributions are distinguished solely by their labeling functions. Slightly abusing the notation, we denote this set of labeling functions also by $\mathcal{Q}$.

Consider a test point x that is not one of the $(x_1, \ldots, x_m)$. Note that, for a fixed labeling of the points in $S_{\mathcal{X}}$, among the labeling functions of distributions in $\mathcal{Q}$ agreeing with that labeling on $S_{\mathcal{X}}$, there are more functions that label x with the majority label on $S_{\mathcal{X}}$ than functions that label x with the minority label on $S_{\mathcal{X}}$. For a labeling function $l \in \mathcal{Q}$, we let S_l denote the points in $S_{\mathcal{X}}$ labeled with l. This implies that

$$\begin{aligned} &\mathbb{E}_{x\sim P_{\mathcal{X}}}\, \mathbb{E}_{l\sim\mathcal{Q}}[\mathbb{1}_{\mathcal{A}(S_l)(x)\neq l(x)} \mid x \notin S_{\mathcal{X}}] \\ \geq\ &\mathbb{E}_{x\sim P_{\mathcal{X}}}\, \mathbb{E}_{l\sim\mathcal{Q}}[\mathbb{1}_{\mathcal{M}(S_l)(x)\neq l(x)} \mid x \notin S_{\mathcal{X}}], \end{aligned}$$

where l is chosen uniformly at random from the set $\mathcal{Q}$. As the expectation is commutative, we get

$$\begin{aligned} &\mathbb{E}_{l\sim\mathcal{Q}}\mathbb{E}_{x\sim P_{\mathcal{X}}}[\mathbb{1}_{\mathcal{A}(S_l)(x)\neq l(x)} \mid x \notin S_{\mathcal{X}}] \\ \geq\ &\mathbb{E}_{l\sim\mathcal{Q}}\mathbb{E}_{x\sim P_{\mathcal{X}}}[\mathbb{1}_{\mathcal{M}(S_l)(x)\neq l(x)} \mid x \notin S_{\mathcal{X}}]. \end{aligned}$$

As this is independent of the choice of $S_{\mathcal{X}}$, we further obtain

$$\begin{aligned} &\mathbb{E}_{S_{\mathcal{X}}\sim\mathcal{P}^m}\mathbb{E}_{l\sim\mathcal{Q}}\mathbb{E}_{x\sim P_{\mathcal{X}}}[\mathbb{1}_{\mathcal{A}(S_l)(x)\neq l(x)} \mid x \notin S_{\mathcal{X}}] \\ \geq\ &\mathbb{E}_{S_{\mathcal{X}}\sim\mathcal{P}^m}\mathbb{E}_{l\sim\mathcal{Q}}\mathbb{E}_{x\sim P_{\mathcal{X}}}[\mathbb{1}_{\mathcal{M}(S_l)(x)\neq l(x)} \mid x \notin S_{\mathcal{X}}]. \end{aligned}$$

This yields

$$\begin{aligned} &\mathbb{E}_{l\sim\mathcal{Q}}\mathbb{E}_{S_{\mathcal{X}}\sim\mathcal{P}^m}\mathbb{E}_{x\sim P_{\mathcal{X}}}[\mathbb{1}_{\mathcal{A}(S_l)(x)\neq l(x)} \mid x \notin S_{\mathcal{X}}] \\ \geq\ &\mathbb{E}_{l\sim\mathcal{Q}}\mathbb{E}_{S_{\mathcal{X}}\sim\mathcal{P}^m}\mathbb{E}_{x\sim P_{\mathcal{X}}}[\mathbb{1}_{\mathcal{M}(S_l)(x)\neq l(x)} \mid x \notin S_{\mathcal{X}}]. \end{aligned}$$

This implies that there exists a function $l \in \mathcal{Q}$ such that

$$\begin{aligned} &\mathbb{E}_{S_{\mathcal{X}}\sim\mathcal{P}^m}\mathbb{E}_{x\sim P_{\mathcal{X}}}[\mathbb{1}_{\mathcal{A}(S_l)(x)\neq l(x)} \mid x \notin S_{\mathcal{X}}] \\ \geq\ &\mathbb{E}_{S_{\mathcal{X}}\sim\mathcal{P}^m}\mathbb{E}_{x\sim P_{\mathcal{X}}}[\mathbb{1}_{\mathcal{M}(S_l)(x)\neq l(x)} \mid x \notin S_{\mathcal{X}}]. \end{aligned}$$

That is, for this distribution with labeling function l, the expected error of $\mathcal{A}$ is larger than the expected error of $\mathcal{M}$ (outside the sample). This completes the proof of the lemma. □

For learning over an infinite domain Lemma 3 thus immediately yields:

Theorem 4. *Let $\mathcal{X}$ be an infinite domain. Then for every class H that contains the two constant functions, learning the class H with respect to the class of all distributions with deterministic labeling functions has sample complexity*

$$m^{\det}[H](\epsilon,\delta) \geq \frac{1}{\epsilon^2}$$

for every $\delta < 1/32$.

In fact, it is easy to see that the sample complexity lower bound applies to any class that contains two functions h, h' such that $\{x \in \mathcal{X} : h(x) \neq h'(x)\}$ is infinite.

We now show how to extend the lower bound of the above result to classes of arbitrary VC-dimension. we show that, for every d, there is a class with VC-dimension d with sample complexity $\Omega\left(d/\epsilon^2\right)$. We use the following notation for the next result: For a hypothesis class H over some domain set $\mathcal{X}$ and a sample sizes m, let $\epsilon_H(m)$ denote the "inverse of the sample complexity", that is,

$$\epsilon_H^{\det}(m) = \mathrm{Inf}_{\mathcal{A}}\mathrm{Sup}_{P\in\mathcal{D}}\mathbb{E}_{S\sim P^m}[\mathrm{Err}_P(\mathcal{A}(S))] - \mathrm{opt}_P(H)$$

(where $\mathcal{D}$ is the family of all probability distributions over $\mathcal{X}\times\{0,1\}$ whose marginal is a deterministic function). Note that for the class $H_{1,0}$ in Lemma 3 we have $\epsilon_H^{\det}(m) = 1/\sqrt{m}$, which is a convex function (that is, the restriction to $\mathbb{N}$ of a convex function over $\mathbb{R}$).

Theorem 5. *Let $\mathcal{X}$ be an infinite domain and $d \in \mathbb{N}$. Then there exists a class H^d over $\mathcal{X}$ with VC-dimension $\mathrm{VCdim}(H) = d$ such that learning this class H^d with respect to the class of all distributions with deterministic labeling functions has sample complexity*

$$m^{\det}[H^d](\epsilon,\delta) \geq \frac{d}{\epsilon^2}$$

for all $\delta < 1/32$.

Proof. Since $\mathcal{X}$ is infinite we can embed d disjoint copies $\mathcal{X}_1,\ldots\mathcal{X}_d$ of $\mathcal{X}$ into $\mathcal{X}$. We construct a class H^d over $\mathcal{X}$ such that, for each $\mathcal{X}_i$, the restriction of H^d to $\mathcal{X}_i$ is (isomorphic to) $H_{0,1}$. That is, H^d is the class of all functions that are constant on each of the sub-domains, $\mathcal{X}_i$. Note that the VC-dimension of H^d is d. For a distribution P over $\mathcal{X}$ we let P_i denote its restriction to $\mathcal{X}_i$.

For a sample S we let $S_i = S \cap \mathcal{X}_i$ and $n_i = |S \cap \mathcal{X}_i|$. Further, we let N_S denote the vector $(n_1,\ldots,n_d)$. For any given m, we let $\mathcal{N}_m$ denote the set of all possible vectors N_S for samples of size m.

Let $\mathcal{M}^d$ be the learning algorithm that, given a sample $S = ((x_1,y_1)\ldots(x_m,y_m))$, predicts with a function that agrees with the labels of the sample points on S and outside the sample chooses the function that agrees, for each sub-domain $\mathcal{X}_i$, with the constant function on $\mathcal{X}_i$ that has the majority label in S_i. The argument now is similar to that in the proof of Lemma 3. Given any $\epsilon > 0$, pick, for each $i \leq d$ a subset $A_i \subseteq \mathcal{X}_i$ of size $1/\epsilon^3$, and let $\mathcal{Q}_\epsilon$ be the family of all probability distributions over $\mathcal{X}\times\{0,1\}$ whose marginal distribution $P_{\mathcal{X}}$ is uniform over $\cup_{i=1}^d A_i$ and whose labeling function deterministically labels, for each $i \leq d$ a $(1/2-\epsilon)$-fraction of the points in A_i with 0 and $(1/2+\epsilon)$-fraction of the points in A_i with 1, or the other way around.

First we claim that, for every distribution $P \in \mathcal{Q}_\epsilon$, the algorithm $\mathcal{M}_d$ needs to see $\Omega(1/\epsilon^2)$ many points to succeed at the task. Then we show that for any other learning algorithm $\mathcal{A}$, there exits a distribution in $\mathcal{Q}_\epsilon$ where $\mathcal{A}$ performs worse than $\mathcal{M}_d$. These two steps together imply the claim. Repeating the argument of Step 2 in the proof of Lemma 3, there exists some distribution $P \in \mathcal{Q}_\epsilon$ such that

$$\begin{aligned}
&\mathbb{E}_{S\sim P^m}[\mathrm{Err}_P(\mathcal{M}_d(S))]\\
=\ &\sum_{N\in\mathcal{N}_m}\left(\Pr_{S\sim P^m}[N_S = N]\sum_{i=1}^d \frac{1}{d}\mathrm{Err}_{P_i}(\mathcal{M}_d(S))\right)\\
\geq\ &\sum_{N\in\mathcal{N}_m}\left(\Pr_{S\sim P^m}[N_S = N]\sum_{i=1}^d \frac{1}{d}\epsilon(n_i)\right)\\
\geq\ &\sum_{i=1}^d \frac{1}{d}\epsilon(m/d)\\
=\ &\epsilon(m/d)
\end{aligned}$$

The last inequality follows from Jensen's inequality, since $\epsilon(m)$ is convex by assumption and the expected size of each of the n_i is m/d. This implies the claim. □

We now show, that there also exist classes of arbitrary VC-dimension that are easy to learn with respect to distributions with deterministic labeling functions. For this, we consider the classes H_d of all functions that label at most d domain points 1 and every other point 0. Note that the class H_d has VC-dimension d.

Theorem 6. *Let $\mathcal{X}$ be an infinite domain and $d \in \mathbb{N}$. There exists a class of VC-dimension d, namely H_d, with sample complexity satisfies $m^{\det}[H_d](\epsilon,\delta) \leq \frac{d}{\epsilon}\log\left(\frac{d}{\epsilon}\right)$.*

Proof. We consider the algorithm $\mathcal{A}$ that on any input sample $S = ((x_1,y_1),\ldots,(x_m,y_m))$ outputs the classifier that labels every $x \in \{x_1,\ldots x_m\}$ by the corresponding label y_i from the sample S (such a label is uniquely defined since we assume that the labeling function is deterministic), and labels any point that is not in the domain of S by 0.

To analyze the sample complexity of this algorithm, we first show that (with high probability) S hits every *heavy* domain point, that is points whose weight (with respect to the marginal distribution) is at least $\frac{\epsilon}{d}$. Since there are at most d/ϵ such domain points, a sample of size $\geq \frac{d}{\epsilon}\log\left(\frac{d}{\epsilon}\right)$ guarantees that with probability greater than $(1-\delta)$ the sample S will hit every such domain point. Now, since the best hypothesis on the class labels at most d points with 1, the error of $\mathcal{A}(S)$ is at most $\mathrm{opt}_P(H) + d\cdot\frac{\epsilon}{d} = \mathrm{opt}_P(H) + \epsilon$, whenever S hits all the heavy points. This implies the claim. □

Note that the algorithm $\mathcal{A}$, described in the proof, is not an ERM algorithm – whenever the training sample S contains more than d points labeled 1, $\mathcal{A}(S)$ is not a member of H_d.

Characterization of learning rates

We now provide a characterization of the family of all classes that have fast learning rates. Namely the classes for which there exists a learning algorithm that learn the class from $\tilde{O}(1/\epsilon)$ sample sizes. The characterization turns out to depend on the following simple combinatorial parameter.

Definition 7. The *diameter of a class* H is $D(H) = \sup_{h,h' \in H} |\{x : h(x) \neq h'(x)\}|$.

Relationship with VCdim:

Claim 1. *1. For every class H, $VCdim(H) \leq D(H)$.*

2. There exist classes with $VCdim(H) = 1$ and $D(H) = \infty$.

Proof. 1. If H shatters a set A, then there exist functions, $h_0, h_1 \in H$ such that for all $x \in A$, $h_0(x) = 0 \neq h_1(x)$. It follows that $D(H) \geq |A|$.

2. Consider the class of all initial segments over the unit interval $[0, 1]$. Its VC-dimension is 1 and its diameter is infinite.

□

Theorem 8 (Characterizing the deterministic sample complexity)**.** *The deterministic sample complexity of a class is determined by its diameter. Namely, for any class H of binary valued functions,*

1. If $D(H)$ is finite then the deterministic sample complexity of a class H is $\tilde{O}(1/\epsilon)$. Furthermore, if $D(H) = k < \infty$ then for some constant, C, for all (ϵ, δ),

$$m_H^{det}(\epsilon, \delta) \leq C \frac{k \log(1/\epsilon) + \log(1/\delta)}{\epsilon}.$$

2. If $D(H)$ is infinite then the deterministic sample complexity of H is $\Omega\left(1/\epsilon^2\right)$.

Proof outline: In the first case, we can repeat the argument we had for the class of at-most-d-ones. For the second case, if the diameter is infinite, then for every n, H contains a pair of functions that disagree on at least n many points. Learning H is therefore at least as hard as learning the class $H_{1,0}$ of the two constant functions over an n-size domain. We have shown in Lemma 3 that for every ϵ there exists some n such that such learning requires $\Omega\left(1/\epsilon^2\right)$ for deterministic labelings.

□

The sample complexity of ERM algorithms.

As mentioned above, one of the fundamental features of both the PAC model and the agnostic-pac model is that the sample complexity of learning by any ERM learner is, up to constant factors, as good as that of any possible learner. Somewhat surprisingly, we notice that this feature is no longer true when one restricts the data-generating distributions to those with deterministic labelings. As shown in Theorem 6, the algorithm $\mathcal{A}$ there requires only $\frac{d}{\epsilon} \log\left(\frac{d}{\epsilon}\right)$ examples to reach accuracy ϵ over any label-deterministic distribution. Our next result shows that any ERM algorithm for the same class H_d requires at least d/ϵ^2 examples to achieve ϵ accuracy with probability greater than $1 - 1/32$ with respect to the same family of all label-deterministic distributions. Namely, in the case of deterministic distributions, there exists a class for which any ERM learner is sub-optimal in terms of its sample complexity. We first present an example of such a class, and then we provide a general characterization of the classes for with ERM enjoys fast convergence rates. As a corollary, we also get a characterization of the family of classes for which ERM algorithms are not optimal (from the sample complexity perspective).

We denote an ERM algorithm for some class H by $\mathrm{ERM}(H)$.

Theorem 9. *Let $\mathcal{X}$ be some infinite domain. Then the sample complexity of the algorithm $\mathrm{ERM}(H_d)$ for the class of all distributions with deterministic labeling functions is lower bounded by*

$$m^{\mathrm{det}}[\mathrm{ERM}(H_d), H_d](\epsilon, \delta) \geq \frac{d}{\epsilon^2}.$$

for any $\delta < 1/32$.

In fact, this lower bound holes for any proper *learning algorithm (an algorithm for learning a class H is called proper if its outputs are always members of H).*

Proof. For $d = 1$ consider a domain of two points x_1 and x_2 and the two distributions that label both of these points with 1 and give weight $1/2 - \epsilon$ to one of the points and weight $1/2 + \epsilon$ to the other. Then, learning H_1 with respect to this set of distributions corresponds to estimating the bias of a coin. Thus Lemma 5.1 of (Anthony and Bartlett 1999) implies that the sample complexity of such an estimation task is larger than $1/\epsilon^2$.

For general d, we consider a domain $D \subseteq \mathcal{X}$ of $2d$ points. We divide them into pairs $\{(x_i, x_i') \mid i \leq d\}$. Let $\mathcal{Q}_D$ be the family of all distributions that label all of these points 1 and for every pair its marginal gives weight $\frac{1}{d}(1/2 + \epsilon)$ to one of these points and weight $\frac{1}{d}(1/2 - \epsilon)$ to the other. Note that any member of H_d can label at most d points with 1.

Without loss of generality, any optimal ERM algorithm, picks one member of each pair according to the number of times it occurs in the training sample. More precisely, let us say that an algorithm $\mathcal{A}$ is *balanced* if for every input sample, S, $\mathcal{A}(S)$ is a hypothesis in H_d that assigned the value 1 to one point in each pair (x_i, x_i'). We claim that for every proper learning algorithm $\mathcal{B}$ for H_d there exists a balanced algorithm $\mathcal{A}$ such that for every distribution $P \in \mathcal{Q}_D$, and every sample size m, $\mathbb{E}_{S \sim P^m}[\mathrm{Err}_P(\mathcal{B}(S))] \geq \mathbb{E}_{S \sim P^m}[\mathrm{Err}_P(\mathcal{A}_d(S))]$. This follows from noting that whenever $\mathcal{B}(S)$ is not a balanced hypothesis, there is some i such that $\mathcal{B}(S)$ assigned the label 1 to both x_i and x_i' and some j such that $\mathcal{B}(S)$ assigned 0 to both x_j and x_j'. However, it is easy to realize that an algorithm that on such S outputs an $h \in H_d$ that agrees with $\mathcal{B}(S)$ on all the points except that it picks one of $\{x_i, x_i'\}$ uniformly at random and assigned to it 1 and assigned 0 to the other member of that pair, and does similarly for

$\{x_j, x'_j\}$, has, in expectation over these coin tosses, an error that is not worse than that of $\mathcal{B}$. Now, applying a similar argument as in the proof of Lemma **??**, we conclude that, as long as the sample size is below $\frac{d}{\epsilon^2}$ the expected error of such an ERM algorithm is at larger than $\text{opt}_P(H) + \epsilon/2$. □

Characterizing the sample complexity of ERM algorithms

We now provide a complete characterization of fast vs. slow rates of convergence of their ERM algorithms over deterministic label distributions.

The sample complexity of any ERM algorithm is lower bounded by that of learning H in the *known label*, KLCL, model. For that model, Ben-David et al, (for review 2011), establish a lower bound of the KLCL sample complexity of learning. To state that result, we need the following definition:

Definition 10. Given a class H of binary-valued functions over some domain set X, let $A_H = \{x \in X : \exists h, h' \in H$ such that $h(x) \neq h'(x)\}$. We classify the possible classes of functions H over a domain set X into three mutually exclusive types, based on their behavior on A_H.

1. We say that H is *simple* if H shatters A_H.
2. We say that H is *pseudo-simple* if A_H is infinite and H does not shatter A_H, but shatters every finite subset of A_H.
3. We say that H is *non-simple* if H there exists some finite subset of A_H that is not shattered by H.

It is straightforward to check that each class of functions H is of exactly one of the above three types. In addition, if H has finite VC dimension, then H cannot be pseudo-simple.

Ben-David et al (Ben-David and Ben-David 2011) provides the following characterization of the KLCL sample complexity of a given class as a function of the accuracy parameter ϵ.

Theorem 11 (The KLCL Theorem)**.** *For any hypothesis class H,*

1. *If H is simple then the KLCL sample complexity of H is zero.*
2. *If H is pseudo-simple and X is countable, then the KLCL sample complexity of H is $\Theta\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$.*
3. *If H is non-simple, then the KLCL sample complexity of H is $\Omega\left(\frac{1}{k}\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$, assuming $\epsilon < \frac{1}{k+1}$.*

Corollary 12 (Characterization of the deterministic sample complexity of ERM)**.** *For every class H,*

1. *If A_H is finite and H shatters it, then ERM learns H with respect to deterministic labelings from $\tilde{O}(1/\epsilon)$ samples. More precisely, for some constant, C, for all (ϵ, δ),*

$$m^{det}_{ERM(H)}(\epsilon, \delta) \leq C\frac{d\log(d/\epsilon) + \log(1/\delta)}{\epsilon},$$

where d is the size of A_H.

2. *Otherwise, (if A_H is infinite, or it is not shattered by H), then ERM requires $\Omega(1/\epsilon^2)$ samples. More precisely, for some constant, C' (that may depend on H,*

$$m^{det}_{ERM(H)}(\epsilon, \delta) \geq C'\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right).$$

Here again, as in Theorem 9, the lower bonds applies to any proper learning algorithm.

Proof. In the first case, ERM will have error $\leq \epsilon$ as soon as the training sample hits every member of A_H that has weight at least ϵ/d. The sample size bound in part 1 of the corollary guarantees that that will be the case with probability $\geq 1-\delta$.

For the second case of the corollary, note that either there is a finite subset of A_H that H does not shatter, in which case H is non-simple and part 3 of Theorem 11 holds, or H has infinite VC dimension, in which case its diameter $D(H)$ is also infinity, and part 2 of Theorem 8 applies. □

We are now in a position to characterize the classes for which ERM algorithms are not sample complexity optimal. Namely, classes for which there exists a gap between their learning sample complexity and the sample complexity of their ERM algorithms (all with respect to data distributions with deterministic labelings).

Corollary 13. *A class H can be learnt fast (i.e. from $\tilde{O}(1/\epsilon)$ examples), while any ERM algorithm for that class requires $\Omega\left(1/\epsilon^2\right)$ size samples, if and only if either A_H is finite but not shattered by H or A_H is infinite while $D(H)$ is finite.*

Discussion

Our investigation reveals that the task of agnostic learning classes in the deterministic label setting is quite different from the similar task when no deterministic restriction is imposed on the labeling function. In particular, the sample complexity of classes in the setup investigated here is not fully determined by the VC dimension of the class, and ERM algorithms do not always achieve optimal error rates. We have provided characterization of learning rates, both for general learning algorithms and for proper learning algorithms (including ERM algorithms) in terms of simple combinatorial parameters of the learnt classes.

In this work, we have considered notions of distribution independent learnability and sample complexity rates. In contrast, the known results on learning under the Tsybakov noise condition provide convergence rates that are distribution dependent. The Tsybakov noise condition can be viewed as a "niceness" or "benignness" parameter on the data-generating distribution. It is then shown that ERM classifiers converge faster, the nicer the underlying distribution is.

Our results imply that learning a hypothesis class of finite VC-dimension does not necessarily get easier in the deterministic labeling setting (as one might expect from the results on faster convergence rates under the Tsybakov noise

condition). Thus, it may be interesting to also investigate learnability in the distribution dependent framework here. That is, we would like to identify parameters of distributions with deterministic labeling function that model some sort of "benignness" of the distribution in this setting.

Acknowledgement

We would like to thank Peter Bartlett for suggesting the classes H_d for Theorem 6.

References

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

Ben-David, S., and Ben-David, S. 2011. Learning a classifier when the labeling is known. In Kivinen, J.; Szepesvári, C.; Ukkonen, E.; and Zeugmann, T., eds., *ALT*, volume 6925 of *Lecture Notes in Computer Science*, 440–451. Springer.

Blumer, A.; Ehrenfeucht, A.; Haussler, D.; and Warmuth, M. K. 1989. Learnability and the vapnik-chervonenkis dimension. *J. ACM* 36(4):929–965.

Boucheron, S.; Bousquet, O.; and Lugosi, G. 2005. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics* 9(11):323–375.

for review, A. 2011. Private communication.

Haussler, D. 1992. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Inf. Comput.* 100(1):78–150.

Kääriäinen, M. 2006. Active learning in the non-realizable case. In *Proceedings of the Conference on Algorithmic Learning Theory (ALT)*, 63–77.

Mammen, E., and Tsybakov, A. B. 1999. Smooth discrimination analysis. *Annals of Statistics* 27(6):1808–1829.

Tsybakov, A. B. 2004. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics* 32(1):135–166.

Urner, R. 2013. Learning with non-standard supervision. `http://uwspace.uwaterloo.ca/handle/10012/7925`.

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

Vapnik, V. N., and Chervonenkis, A. J. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications* 16(2):264–280.