JMLR: Workshop and Conference Proceedings vol 35:1–16, 2014

The sample complexity of agnostic learning under deterministic labels

Shai Ben-David SHAI@UWATERLOO.CA
Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, CANADA

Ruth Urner RURNER@CC.GATECH.EDU
School of Computer Science, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

Abstract

With the emergence of Machine Learning tools that allow handling data with a huge number of features, it becomes reasonable to assume that, over the full set of features, the true labeling is (almost) fully determined. That is, the labeling function is deterministic, but not necessarily a member of some known hypothesis class. However, agnostic learning of deterministic labels has so far received little research attention. We investigate this setting and show that it displays a behavior that is quite different from that of the fundamental results of the common (PAC) learning setups. First, we show that the sample complexity of learning a binary hypothesis class (with respect to deterministic labeling functions) 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 $\tilde{O}(d/\epsilon)$-many samples and classes that require samples of size $\Omega(d/\epsilon^2)$. Furthermore, we show that in this setup, there are classes for which any proper learner has suboptimal sample complexity. While the class can be learned with sample complexity $\tilde{O}(d/\epsilon)$, any *proper* (and therefore, any ERM) algorithm requires $\Omega(d/\epsilon^2)$ samples. We provide combinatorial characterizations of both phenomena, and further analyze the utility of unlabeled samples in this setting. Lastly, we discuss the error rates of nearest neighbor algorithms under deterministic labels and additional niceness-of-data assumptions.

1. Introduction

We investigate the sample complexity of binary classification learning with respect to a concept class, when the true 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. Situations in which the labeling rule may be deterministic include setups where data is represented by a very large number of features. For example, document categorization using a feature space containing all of the language's dictionary. Such tasks are often handled with sparsity-inducing regularization (over, say, linear classifiers). The motivation behind inducing sparsity can either be due to some prior knowledge that the data model is sparse, but often it is due to the computational and interpretability merits of sparse models. In such cases, the set of potential output classifiers is not assumed to include a zero error classifier (Bach et al. (2012)).

The analysis of the sample complexity in the PAC model is 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* setup, the sample complexity of learning H is (roughly) $\Theta\left(\frac{\mathrm{VCdim}(H)}{\epsilon}\right)$. In the second case, the *agnostic* setup, 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 (roughly) $\Theta\left(\frac{\mathrm{VCdim}(H)}{\epsilon^2}\right)$. Proving the lower bound in the agnostic case usually involves taking advantage of the stochasticity in the labeling function. Non-realizability, the lack of a zero-error classifier in the learner's search space, can be caused by stochasticity, or "noise", of the labeling rule (with respect to the features considered) or by the limitations of the learner's search space.

The first case has attracted a lot of attention in the past decade. Mammen and Tsybakov (1999) initiated some finer grained analysis. Assuming that the Bayes optimal classifier is a member of the class, they analyze the sample complexity 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, Tsybakov's noise exponent, usually denoted by α, that restricts the amount of noise in the labeling. Table 1 illustrates the relationship between these three cases (realizable, fully agnostic, and realizable with controlled noise). In this work, we turn the focus on the upper right corner: agnostic learning under deterministic labellings. One can get the impression that the stochasticity of the labeling procedure is the only reason for slow learning rates. One contribution of our work is to show that slow rates (sample sizes $\Omega(\mathrm{VCdim}/\epsilon^2)$) are common for fully deterministic labellings as well, once the assumption that the Bayes classifier is (almost) in the class is dropped. It turns

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

	Bayes in H	Agnostic
Deterministic labels	$\frac{\mathrm{VCdim}(H)}{\epsilon}$	**?**
Probabilistic labels	under Tsybakov noise: $\frac{\mathrm{VCdim}(H)}{\epsilon^{2-\alpha}}$	$\frac{\mathrm{VCdim}(H)}{\epsilon^2}$

out that learning bounds in the agnostic-deterministic setup behave differently from what we know in the other cases. First, the sample complexity of learning a binary hypothesis class is not fully determined by the VC-dimension of the class. We show that, for any d, there exist classes of VC-dimension d that are learnable with sample complexity $\tilde{O}(d/\epsilon)$, regardless of the approximation error, and classes for which learning requires sample sizes of $\Omega(d/\epsilon^2)$. Note that in the case we focus on, the noise condition is fixed (there is no noise whatsoever) and yet both fast-rate and slow-rate convergence occurs. Furthermore, every class falls into one of these two categories. There are no intermediate rates. We introduce a simple combinatorial parameter of a class, the *class diameter* and prove that a class has fast convergence rates if and only if its class diameter is finite. We also show that most "interesting classes" have infinite diameter, and thus slow convergence rates even in our noise free setting.

We then analyze the sample complexity of proper learners. It is well known that proper learning, where the learner is required to output a member of the hypothesis class, is often *computationally* harder than unrestricted learning (see for example Section 1.4 of Kearns and Vazirani (1995)). However, in the common PAC setups, the restriction to proper outputs does not bear any *sample complexity* consequences – optimal learning rates are achieved by any ERM procedure for every hypothesis class. In contrast, for agnostic deterministic learning, we prove an inherent weakness of proper learners *from the sample complexity perspective*. We show that, for every d, there exist classes of VC-dimension d that are learnable with sample complexity $\tilde{O}(d/\epsilon)$ while any proper learner for these classes (and, in particular, any ERM learner) requires $\Omega\left(1/\epsilon^2\right)$ size samples. We provide a combinatorial characterization of the classes that demonstrate such a sample complexity

gap. Furthermore, we show that in all cases where proper (and ERM) learners are sample complexity suboptimal, access to unlabeled samples can fully overcome this gap. We propose a semi-supervised version of ERM that becomes optimal once it has access to sufficiently many unlabeled examples. Lastly, we briefly discuss non-parametric (Nearest Neighbor) learning of deterministic labels under additional data assumptions.

We point out that, while all the lower bounds presented in this work involve an approximation error that is close to $1/2$, using standard techniques, these lower bounds can be converted to hold for any arbitrary, but fixed, smaller approximation error. For this, one considers distributions, where most of the weight is on one heavy point, and concentrates the lower bound construction on the rest.

2. 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). 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 realizable and the agnostic case has attracted quite a lot of attention. These can be viewed along two directions. An assumption of a small approximation error by a class allows leveraging Bennett and Bernstein inequalities and yields bounds that imply fast rates in the range of generalization error that is higher than the approximation error of the class. This analysis goes back to Vapnik and Chervonenkis (1971). More recently, Mammen and Tsybakov (1999) considered the setting in which the Bayes classifier belongs to the class H, and therefore the stochasticity of the labeling (or its "noise") is the only source of the approximation error. They introduce the *Tsybakov noise condition*, a bound on the stochasticity (or noisiness) of the labels and prove convergence rates under that condition (and the additional assumption that the Bayes optimal classifier is a member of the hypothesis class). 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 et al. (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). Our results indicate that the relationship between noise and fast rates, as captures by these conditions, is indeed restricted to the case where the Bayes classifier is in the class (or is very well approximated by a member of the class).

The setup where the labeling rule is deterministic, but yet does not belong to the learned class has been addressed to 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 learned 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 $\mathrm{VCdim}(H)/\epsilon^2$ for 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. Sample complexity suboptimality of *some* ERM learners has recently been detected in context of multi-label classification by Daniely et al. (2011). The existence of classes for which there are learners with low sample complexity (of order $\mathrm{VCdim}(H)/\epsilon$ for arbitrary deterministic-label distributions) while *any ERM*

learner still requires sample sizes of order $1/\epsilon^2$ does not seem to have been brought up by earlier work.

3. 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* 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.

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, $\mathcal{A} : \bigcup_{m=1}^{\infty}(\mathcal{X} \times \{0,1\})^m \to \{0,1\}^{\mathcal{X}}$. A learner is an *empirical risk minimizer (ERM)* for a class H if, for a sample S, it outputs a member of H of minimal empirical error. We denote ERM algorithms for a class H by $\mathrm{ERM}(H)$. We call a learner a *proper learner* for a class H, if it always outputs a function from H. Note that any $\mathrm{ERM}(H)$ algorithm is a proper learner for H.

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 $h : \mathcal{X} \to \{0,1\}$ 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.*

In 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* $\mathcal{Q}^{\text{det}}_{\mathcal{X}}$ *of distributions with deterministic labeling functions, we use the notation* $m^{\text{det}}[H]$.

4. Sample complexity gap between classes of the same VC dimension

4.1. Classes that are hard to learn in the agnostic-deterministic model

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 quantifies a sufficient 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* $H_{1,0}$ *with respect to* $\mathcal{Q}$ *requires a sample size of* $\Omega(1/\epsilon^2)$.

Proof For every distribution P in $\mathcal{Q}$ we have $\text{opt}_P(H) = 1/2 - \epsilon$. 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 $\Omega(d/\epsilon^2)$ 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 $1/2 + \epsilon - |S|/|\mathcal{X}| \geq 1/2 + \epsilon/2 > \text{opt}_P(H) + \epsilon/2$. 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 of 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

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

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

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

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

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

This yields

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

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

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

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. ■

Note that enlarging the class $H_{1,0}$, can only yield a smaller approximation error in the setup of the above proof. Thus, the lower bound in Lemma 3 also holds for any class H with $H_{1,0} \subseteq H$. For learning over an infinite domain Lemma 3 now immediately yields:

Theorem 4 *Let $\mathcal{X}$ be an infinite domain. Then for every class H that contains the two constant functions, for every $\delta < 1/32$, learning the class H with respect to the class of all distributions with deterministic labeling functions has sample complexity $m^{\mathrm{det}}[H](\epsilon, \delta) \geq \frac{1}{\epsilon^2}$.*

Remark 5 *In fact, it is easy to see that the sample complexity lower bound of Theorem 4 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 strengthen the lower bound above to take into account the VC-dimension of the class. We will show that, for every d, there are classes with VC-dimension d with sample complexity $\Omega\left(d/\epsilon^2\right)$. In fact, the family of such classes is rather rich and includes some of the most popular classes used in learning, like linear classifiers, neural networks, decision trees and more. However, as we show below, there are also classes of arbitrarily large VC-dimension that have faster learning rates. To state our lower bound we need the following variation on the notion of shattering.

Definition 6 *Let $\mathcal{X}$ be any domain set, H a class of $\{0,1\}$ valued functions over $\mathcal{X}$ and let $A_1, \ldots A_d$ be subsets of $\mathcal{X}$. We say that H* set-shatters *$A_1, \ldots A_d$ if for every binary vector $\sigma = (\sigma_1, \ldots \sigma_d) \in \{0,1\}^d$ there exists some $h_\sigma \in H$ such that for all $i \leq d$ and $x \in \mathcal{X}$, if $x \in A_i$ then $h_\sigma(x) = \sigma_i$.*

Theorem 7 *Let $\mathcal{X}$ be any domain set and H a class of binary-valued functions over $\mathcal{X}$. If H* set-shatters *$A_1, \ldots A_d$ for some infinite subsets $A_1, \ldots A_d$ of $\mathcal{X}$, then, for all $\delta < 1/32$, the deterministic sample complexity of H satisfies $m^{\det}[H](\epsilon, \delta) \geq \frac{d}{\epsilon^2}$.*

For the proof, we use the following notation: For a hypothesis class H over some domain set $\mathcal{X}$ and a sample size m, let $\epsilon_H^{\det}(m)$ denote the "inverse of the sample complexity", that is,

$$\epsilon_H^{\det}(m) = \inf_{\mathcal{A}} \sup_{P \in \mathcal{Q}_{\mathcal{X}}^{\det}} \mathbb{E}_{S \sim P^m}[\mathrm{Err}_P(\mathcal{A}(S)) \; - \; \mathrm{opt}_P(H)]$$

Recall that $\mathcal{Q}_{\mathcal{X}}^{\det}$ denotes the family of all probability distributions over $\mathcal{X} \times \{0,1\}$ with deterministic labeling function. Note that for the class $H_{1,0}$ in Lemma 3 we have $\epsilon_{H_{0,1}}^{\det}(m) = 1/\sqrt{m}$, which is a convex function (that is, the restriction to $\mathbb{N}$ of a convex function over $\mathbb{R}$).

Proof [of Theorem 7] For a distribution P over $\mathcal{X}$ we let P_i denote its restriction to A_i. For a sample S we let $S_i = S \cap A_i$ and $n_i = |S_i|$. 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.

Given any $\epsilon > 0$, pick, for each $i \leq d$ a subset $B_i \subseteq A_i$ 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.

The argument now is similar to that in the proof of Lemma 3. 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 subdomain A_i, with the constant function on A_i that has the majority label in S_i. First we show 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 argue 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.

For the first step, let P be any distribution in $\mathcal{Q}_\epsilon$. Then

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

$$\geq \; \sum_{N \in \mathcal{N}_m} \left(\Pr_{S \sim P^m}[N_S = N] \sum_{i=1}^d \frac{1}{d} \epsilon_{H_{0,1}}^{\det}(n_i) \right) \; \geq \; \sum_{i=1}^d \frac{1}{d} \epsilon_{H_{0,1}}^{\det}(m/d) \; = \; \epsilon_{H_{0,1}}^{\det}(m/d)$$

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 that the sample complexity of M_d on this task is lower bounded by d/ϵ^2. Repeating the argument of Step 2 in the proof of Lemma 3, one can show that for any other learner $\mathcal{A}$, there exists a distribution where the expected error (over all samples) of $\mathcal{A}$ is larger than the expected error of $\mathcal{M}_d$. This completes the proof. ■

Next, we wish to argue that for most of the "natural" learning classes, H, over Euclidean spaces, like linear classifiers, neural network classifiers and decision trees over decision stumps, the assumptions of Theorem 7 hold for $d = \mathrm{VCdim}(H)$. It follows that, for all of these classes, learning with respect to deterministic labels is (with respect to the sample complexity) as hard as learning them in the fully agnostic setup, where labellings can be arbitrarily noisy.

Recall that for a subset A of a topological space, its interior, $\mathrm{Int}(A)$ is the maximal open set contained in A, and its closure, $\overline{A}$, is the minimal closed set that contains A. Let us call a set, A, *substantial* if it is contained in the closure of its interior. Namely, $A \subseteq \overline{\mathrm{Int}(A)}$. Intuitively speaking, substantial sets have the full space dimension in every ball. Note that every open set is substantial, since in such cases $A = \mathrm{Int}(A)$. It follows that for every continuous $f : \mathbb{R}^n \to \mathbb{R}$, the set $\mathrm{pos}(f) = \{x \in \mathbb{R}^n : f(x) > 0\}$ is a substantial set, and if f is monotone in all variables, the complement of $\mathrm{pos}(f)$ is also substantial. Furthermore, the union of any finite number of substantial sets is substantial. Given these facts, it is not hard to see that every classifier in any of the common classes mentioned above, partitions the input space into two substantial sets.

Lemma 8 *Let H be a class of $\{0,1\}$- valued functions over $\mathbb{R}^n$ for some n. Assume that $h^{-1}(1)$ is a substantial set for every $h \in H$. Then if H shatters some finite set $D \subseteq \mathbb{R}^n$ then H set-shatters some $|D|$-size collection of infinite sets.*

Proof [Outline] First, a standard set-topological argument shows that that for every substantial set A, every $x \in A$ and every open ball B that contains x, there exists a non-empty open ball $B' \subseteq B$ (possibly not containing x) such that $B' \subseteq A$.

Given any finite shattered set $D = \{x_1, \ldots x_d\}$, let $H' = \{h_1, \ldots h_{2^d}\}$ be a subset of H that shatters D. Now, find, for every $i \leq d$ an open ball B_i that contains x_i and, for every $j \leq 2^d$ such that $h_j(x_i) = 0$, B_i is disjoint from $h_j^{-1}(1)$. Finally, use the above property of substantial sets to find non-empty open balls, $B'_i : i \leq d$, such that for very $i \leq d$ and every $j \leq 2^d$, if $h_j(x_i) = 1$ then $B'_i \subseteq h_j^{-1}(1)$ and if $h_j(x_i) = 0$ then $B'_i \cap h_j^{-1}(1) = \emptyset$. Clearly, H set-shatters $\{B'_i : i \leq d\}$. ■

4.2. Classes with fast learning rates

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 9 *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 satisfying $m^{\mathrm{det}}[H_d](\epsilon, \delta) \leq \frac{d}{\epsilon}\left(\log\left(\frac{d}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)$.*

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 $x_i \in \{x_1, \ldots x_m\}$ 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 larger than

$\frac{d}{\epsilon}\left(\log\left(\frac{d}{\epsilon}\right)+\log\left(\frac{1}{\delta}\right)\right)$ 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.

4.3. 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 learns the class from $\tilde{O}\left(1/\epsilon\right)$ sample sizes. The characterization turns out to depend on the following simple combinatorial parameter.

Definition 10 *We define the* diameter of a class H *as* $D(H)=\sup_{h,h'\in H}|\{x:h(x)\neq h'(x)\}|$.

Claim 11 (Relationship of diameter and VC-dimension) *For every class H,* $\mathrm{VCdim}(H)\leq D(H)$. *Further, there exist classes with* $\mathrm{VCdim}(H)=1$ *and* $D(H)=\infty$.

Proof 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 $D(H)\geq|A|$. For the second claim, consider the class of all initial segments over the unit interval $[0,1]$. Its VC-dimension is 1 and its diameter is infinite. ■

Theorem 12 (Characterizing the deterministic learning rates) *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 all (ϵ,δ), we have $m_H^{det}(\epsilon,\delta)\leq\frac{k}{\epsilon}\left(\log\left(\frac{k}{\epsilon}\right)+\log\left(\frac{1}{\delta}\right)\right)$.
2.) If $D(H)$ is infinite then the deterministic sample complexity of H is $\Omega\left(1/\epsilon^2\right)$.

Proof 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 labellings. ■

5. 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. Surprisingly, this feature is no longer true when one restricts the data-generating distributions to those with deterministic labellings. As shown in Theorem 9, 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 accuracy ϵ 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 which ERM enjoys fast convergence rates. As a corollary, we also get a characterization of the family of classes for which ERM is not optimal (from the sample complexity perspective). f minimal empirical error from H_d) has sample complexity strictly worse than the the algorithm $\mathcal{A}$ from the proof of Theorem 9. We denote an ERM algorithm of some class H by $\mathrm{ERM}(H)$.

Theorem 13 *Let $\mathcal{X}$ be some infinite domain. Then, for any $\delta < 1/32$, the sample complexity of any proper learner $\mathcal{A}$ for H_d, in particular any $\mathrm{ERM}(H_d)$ algorithm, with respect to the class of all distributions with deterministic labeling functions is lower bounded by $m^{\mathrm{det}}[\mathcal{A}, H_d](\epsilon, \delta) \geq d/\epsilon^2$.*

Proof [Idea] 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 $\frac{1}{d}(1/2 + 2\epsilon)$ to one of these points and weight $\frac{1}{d}(1/2 - 2\epsilon)$ to the other. This yields a lower bound of d/ϵ^2. The full proof has been moved to the Appendix A. ■

5.1. Characterizing the sample complexity of ERM algorithms

The sample complexity of any ERM algorithm is lower bounded by that of learning H in the *known label*, KLCL, model. In this model, the true labeling function is given to the learner, together with a sample, and the task is to find the best hypothesis in a class H. Ben-David and Ben-David (2011), establish a lower bound for the KLCL sample complexity. To employ that, we need the following:

Definition 14 *Given a class H of binary-valued functions over some domain set $\mathcal{X}$, let $A_H = \{x \in \mathcal{X} : \exists h, h' \in H$ such that $h(x) \neq h'(x)\}$. We call A_H the* effective domain *of the class H.*

We classify classes of binary functions H over a domain set $\mathcal{X}$ into three mutually exclusive types, based on their behavior on A_H:

Definition 15 *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 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 can not be pseudo-simple. We employ the following characterization of the KLCL sample complexity of a given class:

Theorem 16 (The KLCL Theorem, Ben-David and Ben-David (2011)) *For any 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 $\epsilon < \frac{1}{k+1}$, where k is the largest integer such that all subsets of A_H of size k are shattered by H.

Corollary 17 (Characterization of the deterministic sample complexity of ERM) *For every H:*
1.) If A_H is finite and H shatters it, then ERM learns H with respect to deterministic labellings from $\tilde{O}(1/\epsilon)$ samples. More precisely, there exists a constant, C, such that for all (ϵ, δ):

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

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

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

Proof In the first case, ERM will have error at most ϵ 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 at least $1 - \delta$.

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 16 holds, or H has infinite VC dimension, in which case its diameter $D(H)$ is also infinity, and part 2 of Theorem 12 applies. ■

Corollary 18 (Characterization of the optimality of ERM under deterministic labels) *A class H can be learned from $\tilde{O}\left(1/\epsilon\right)$ 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.*

6. Using unlabeled data

In this section, we show that in all the cases, where learning with ERM is not optimal, unlabeled data can make up for the gap between the sample complexity of learning and the sample complexity of learning with ERM. We propose a weighted version of ERM, where sample points receive weights according to their frequency in an unlabeled sample. We show that this weighted ERM achieves *labeled sample complexity* as low as the sample complexity of general learning in these cases. Thus proper learning with optimal labeled sample complexity is possible provided that the learner also has access to a (slightly larger) unlabeled sample.

Definition 19 *We say that a function $\mathcal{A} : (\bigcup_{m=0}^{\infty}(\mathcal{X} \times \{0,1\})^m) \bigcup (\bigcup_{m=0}^{\infty} \mathcal{X}^m) \to H$ is an SSL-ERM learner for H if, for every labeled sample S and unlabeled sample U,*

$$\mathcal{A}(S, U) \in \operatorname{argmin}_{h \in H} \sum_{(x,y) \in S} \mathrm{SU}(x) |h(x) - y|,$$

where we let $\mathrm{SU}(x)$ denote the weight of x in the the union $S \cup U$, that is, the number of occurrences of x in $S \cup U$ divided by $|S| + |U|$.

Recall that A_H denotes the effective domain of the class H (Definition 14) and that $D(H)$ denotes its diameter (Definition 10). Note that, whenever the effective domain A_H of some class H is finite, the diameter $D(H)$ is finite as well. Thus, the theorem below covers all cases, where learning with ERM is not optimal from a sample complexity point of view (compare Corollary 18). The proof has been moved to the Appendix B for space reasons.

Theorem 20 *Let H be such that $D(H)$ is finite, and let $\mathcal{A}$ be any SSL-ERM learner for H. Then, for some constant C, for any deterministic data generating distribution P over $X \times \{0,1\}$, if $n \geq C\left(\frac{1+\log(1/\delta)}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ and $m \geq C\frac{1}{\epsilon^2}\left(\ln\left(\frac{1}{\epsilon^2}\right)+\ln\left(\frac{1}{\delta}\right)\right)$ then with probability at least $1-\delta$ over a labeled sample $S \sim P^n$ and an unlabeled sample $U \sim P_{\mathcal{X}}^m$, we have $\mathrm{Err}_P(\mathcal{A}(S,U) \leq \mathrm{opt}_P(H)+\epsilon$.*

7. Discussion of sample complexity gaps under data niceness conditions

In this work, the sample complexity rates we have been considering were all *distribution independent* - they refer to learning with respect to any data distribution (as long as its labeling rule is deterministic). An interesting follow-up direction is to limit the set of possible data distributions, or parameterize it according to some "data tameness" parameter, and analyze how such restrictions affect the learning rates.

Obviously, noise controlling conditions, like that of Mammen and Tsybakov (1999), are vacuous in the case of deterministic labels that we focus on. However, one may consider different reasonable data "niceness" parameters. Probabilistic Lipschitzness (PL) is a measure of the coherence between a data distribution's marginal and its labeling rule. It quantifies the extent to which similar instances tend to have similar labels. Such a coherence is implicit in many machine learning algorithmic paradigms. Indeed, assuming PL the convergence rates of a Nearest Neighbor algorithms provably improve (see Urner and Ben-David (2013) for an overview). It is shown there, that, under deterministic labeling rules, $\frac{2}{\epsilon\,\delta}\left(\frac{2^{1/n}\sqrt{d}}{\epsilon^{1/n}}\right)^d = O\left(\left(\frac{1}{\epsilon}\right)^{\frac{d+n}{n}}\right)$ samples suffice for 1-Nearest Neighbor to have error at most ϵ, where n is a PL-parameter that quantifies the niceness of the distribution, and d the euclidean dimension. These bounds also imply a distribution depended learning rates for unrestricted (not necessarily proper) learning of hypothesis classes. For sufficiently large n (that is, for sufficiently nice data), unrestricted learning of *any class* has sample complexity $o(1/\epsilon^2)$.

However, it is easy to see that these optimistic rates, do not hold for proper learning of a hypothesis class (a Nearest Neighbor predictor will not likely be a member of the class). The proofs of $\Omega(1/\epsilon^2)$ lower bounds for proper learning, even under deterministic labels, consider distributions supported on very few discrete points. Any labeling on those points satisfies a Lipschitz condition (and thereby the above PL condition for arbitrarily large n). Thus, this setting of distribution depended learning, also exhibits a gap between the sample complexity (as a function of $1/\epsilon$) of unrestricted learning and that of proper learning (and in particular ERM). Interestingly, this gap between the sample complexity of learning and the sample complexity of ERM learning under deterministic labels and Probabilistic Lipschitzness, can also be overcome by unlabeled data. Urner et al. (2011) present an SSL algorithm for proper learning in this exact setting that achieves optimal rates. On the downside, while having favorable dependence on $(1/\epsilon)$ both the Nearest Neighbor non-proper learner, and the SSL proper learning paradigm, exhibit a bad dependence on the dimension of the space d. It would be interesting to investigate whether learning rates that are efficient both in terms of the $1/\epsilon$ and d are possible.

Acknowledgments

We thank Peter Bartlett for suggesting the classes H_d for Theorem 9.
This work was supported in part by AFOSR grant FA9550-09-1-0538.

References

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

Francis R. Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

Shalev Ben-David and Shai Ben-David. Learning a classifier when the labeling is known. In *Proceedings of the Conference on Algorithmic Learning Theory (ALT)*, pages 440–451, 2011.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9(11):323–375, 2005.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. In *COLT*, pages 207–232, 2011.

D Haussler and E Welzl. Epsilon-nets and simplex range queries. In *Proceedings of the second annual symposium on Computational geometry*, SCG '86, pages 61–71, 1986.

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

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

Michael J. Kearns and Umesh V. Vazirani. Computational learning theory. *SIGACT News*, 26(1): 43–45, 1995.

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

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

Ruth Urner. *Learning with non-Standard Supervision*. PhD thesis, University of Waterloo, 2013. URL http://uwspace.uwaterloo.ca/handle/10012/7925.

Ruth Urner and Shai Ben-David. Probabilistic lipschitzness: A niceness assumption for deterministic labels. Learning Faster from Easy Data Workshop@NIPS, 2013. URL http://www.cc.gatech.edu/~rurner/PLEasyDataNIPS2013.pdf.

Ruth Urner, Shai Ben-David, and Shai Shalev-Shwartz. Unlabeled data can speed up prediction time. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 641–648, 2011.

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

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

Appendix A. Proof of Theorem 13

Proof For general d, we consider a domain $D \subseteq \mathcal{X}$ of $2d$ points. We divide them into pairs $\{p_i = (x_i, x_i') \mid i \leq d\}$. 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 $\frac{1}{d}(1/2 + 2\epsilon)$ to one of these points and weight $\frac{1}{d}(1/2 - 2\epsilon)$ to the other.

We now first define and analyze a slightly different task, that we call the *learning d pairs problem*. We then first present a $\Omega(d/\epsilon^2)$ lower bound for this problem and then proceed to to reducing this problem to properly learning H_d. This reduction then implies the desired lower bound of $\Omega(d/\epsilon^2)$ for properly learning H_d, and in particular for any $\mathrm{ERM}(H_d)$ algorithm.

Learning d pairs We define the problem of learning d pairs over the domain described above as learning the smaller class H_d^p, where H_d^p contains all the functions that label exactly one point of every pair with 1. That is, we consider learning H_d with the additional requirement that the output hypothesis contains exactly one point from each pair $p_i = (x_i, x_i')$ (here we identify a binary function with the subset that the binary function labels with 1).

Note that $\mathrm{opt}_P(H_d^p) = 1/2 - 2\epsilon$. Thus, in order to learn with accuracy ϵ, an algorithm has to output a hypothesis of error smaller than $1/2 - \epsilon$. This implies that the output hypothesis needs to choose the heavier point from at least $3d/4$ of the pairs.

Note that detecting the heavier point from any pair p_i corresponds to determining the bias of a coin with bias 2ϵ. Thus, the lower bound of Lemma 5.1 in Anthony and Bartlett (1999) implies that a learning algorithm needs to see at least $3d/4(2\epsilon)^2 = 3d/16\epsilon^2$ random examples, in order to estimate the bias (determining the heavier point) of at least $3d/4$ of the pairs correctly.

Reduction to properly learning H_d Note that $\mathrm{opt}_P(H_d) = 1/2 - 2\epsilon$. A hypothesis in H_d can choose both or neither or one point from a pair p_i. For a distribution P and a hypothesis $h \in H_d$, we let $\mathrm{good}(P, h)$ the number of *good pairs* where h chooses the lighter point, $\mathrm{bad}(P, h)$ the number of *bad pairs*, where h chooses the heavier point, and $\mathrm{even}(P, h)$ the number of *even pairs*, where h chooses either both or none of the points in the pair.

Claim 21 *For a distribution $P \in \mathcal{Q}_D$ and an $h \in H_d$, we have $\mathrm{Err}(h) \leq 1/2 - 3/2\epsilon = \mathrm{opt}_P(H_d) + \epsilon/2$, if and only if*

$$\mathrm{good}(P, h) - \mathrm{bad}(P, h) \geq 3d/4.$$

It is easy to see that the reduction from properly ϵ-learning H_d^p to properly $\epsilon/2$-learning H_d follows from this claim. Given a successful proper learner for H_d, its output h_d can be turned into a hypothesis h_d^p from H_d^p by choosing a random point in any even pair of h_d and otherwise agreeing with h_d. The claim implies that if $\mathrm{Err}(h) \leq 1/2 - 3/2\epsilon$, then h_d^p will contain the lighter point of at least $3d/4$ of the pairs. Thus, independently of what h_d^p chooses on the other pairs, we have $\mathrm{Err}_P(h_d^p) \leq 1/2 - \epsilon$.

Proof [Proof of Claim 21] For a hypothesis $h \in H_d$, we let $\mathrm{Err}_{p_i}(h)$ denote the error of h on the pair p_i. This is, $\mathrm{Err}_{p_i}(h) = 1/2 - \epsilon$ if p_i is a good pair for h, $\mathrm{Err}_{p_i}(h) = 1/2 + \epsilon$ if p_i is a bad pair for h, and $\mathrm{Err}_{p_i}(h) = 0$ or $\mathrm{Err}_{p_i}(h) = 1$ if p_i is an even pair for h.

Note that for every $h \in H_d$, there are equally many even pairs where h chooses both points as there are even pairs where h chooses none of the points. Thus, we have $\mathrm{Err}_{p_i}(h) = 1/2$ on average over all even pairs. The total error of h can be decomposed as follows:

$$\begin{aligned}\mathrm{Err}_P(h) &= \frac{1}{d}\sum_{i=1}^{d}\mathrm{Err}_{p_i}(h)\\ &= \frac{1}{d}[1/2\mathrm{even}(P,h) + (1/2+\epsilon)\mathrm{bad}(P,h) + (1/2-\epsilon)\mathrm{good}(P,h)]\\ &= \frac{1}{2} + \frac{2\epsilon}{d}\mathrm{bad}(P,h) - \frac{2\epsilon}{d}\mathrm{good}(P,h)\end{aligned}$$

It is easy to verify that this is larger than $1/2 - 3/2\epsilon$ if and only if $\mathrm{good}(P,h) - \mathrm{bad}(P,h) \geq 3d/4$. This completes the proof of the claim. ■

■

Appendix B. Proof of Theorem 20

Proof Let $d = D(H)$ denote the diameter of the class H. We let $S_{\mathcal{X}}$ denote the projection of S on $\mathcal{X}$. With the sample sizes indicated above, the sample $S_{\mathcal{X}}$ is an $\epsilon/3d$-net and the union of samples $S \cup U$ an $\epsilon/3d$-approximation for the set of singletons $\{\{x\} : x \in X\}$ with respect to P with probability at least $1 - \delta$ (see Haussler and Welzl (1986)). From here on, we assume that this is the case. That is, for every $x \in \mathcal{X}$ with $P_{\mathcal{X}}(x) \geq \epsilon/3d$, we have $(x, l(x)) \in S$ and for all $x \in \mathcal{X}$ we have $|\mathrm{SU}(x) - P_{\mathcal{X}}(x)| \leq \epsilon/3d$.

Let h^* be a hypothesis of minimal error in H and let $h_{SU} = \mathcal{A}(S,U)$. For two function h and h' in H let $h\Delta h' = \{x \in X : h(x) \neq h'(x)\}$. Note that $|h^*\Delta h_{SU}| \leq d$ since $h_{SU} = \mathcal{A}(S,U) \in H$. Now we have

$$\begin{aligned}&\mathrm{Err}_P(h_{SU}) - \mathrm{Err}_P(h^*)\\ =\ &\Pr_{x\sim P_{\mathcal{X}}}[x \in (h^*\Delta h_{SU}) \wedge h_{SU} \neq l(x)] - \Pr_{x\sim P_{\mathcal{X}}}[x \in (h^*\Delta h_{SU}) \wedge h^* \neq l(x)]\\ =\ &P_{\mathcal{X}}((h^*\Delta h_{SU}) \cap (h_{SU}\Delta l)) - P_{\mathcal{X}}((h^*\Delta h_{SU}) \cap (h^*\Delta l))\\ =\ &P_{\mathcal{X}}(S_{\mathcal{X}} \cap (h^*\Delta h_{SU}) \cap (h_{SU}\Delta l)) - P_{\mathcal{X}}(S_{\mathcal{X}} \cap (h^*\Delta h_{SU}) \cap (h^*\Delta l))\\ &\quad + P_{\mathcal{X}}((\mathcal{X}\setminus S_{\mathcal{X}}) \cap (h^*\Delta h_{SU}) \cap (h_{SU}\Delta l)) - P_{\mathcal{X}}((\mathcal{X}\setminus S_{\mathcal{X}}) \cap (h^*\Delta h_{SU}) \cap (h^*\Delta l))\end{aligned}$$

The contribution of the second part of the above sum can be bounded by $\epsilon/3$ since the set $h^*\Delta h_{SU}$ contains at most d points and each point in $\mathcal{X}\setminus S_{\mathcal{X}}$ has weight at most $\epsilon/3d$. We now bound the fist part of the sum. By definition of $\mathcal{A}$, we have

$$\mathrm{SU}((h_{SU}\Delta l) \cap S_{\mathcal{X}}) \leq \mathrm{SU}((h^*\Delta l) \cap S_{\mathcal{X}})$$

and this implies

$$\mathrm{SU}((h_{SU}\Delta l) \cap S_{\mathcal{X}} \cap (h^*\Delta h_{SU})) \leq \mathrm{SU}((h^*\Delta l) \cap S_{\mathcal{X}} \cap (h^*\Delta h_{SU}))$$

From this we get

$$\mathrm{SU}((h_{SU}\Delta l) \cap S_{\mathcal{X}} \cap (h^*\Delta h_{SU})) - \mathrm{SU}((h^*\Delta l) \cap S_{\mathcal{X}} \cap (h^*\Delta h_{SU})) \leq 0 \qquad (1)$$

Now, since $|(h^*\Delta h_{SU}))| \leq d$ and since we have $\mathrm{SU}(x) - P_{\mathcal{X}}(x) \leq \epsilon/3d$ for all $x \in \mathcal{X}$, we have

$$|\mathrm{SU}((h_{SU}\Delta l) \cap S_{\mathcal{X}} \cap (h^*\Delta h_{SU})) - P_{\mathcal{X}}((h_{SU}\Delta l) \cap S_{\mathcal{X}} \cap (h^*\Delta h_{SU}))| \leq \epsilon/3$$

and

$$|\mathrm{SU}((h^*\Delta l) \cap S_{\mathcal{X}} \cap (h^*\Delta h_{SU})) - P_{\mathcal{X}}((h^*\Delta l) \cap S_{\mathcal{X}} \cap (h^*\Delta h_{SU}))| \leq \epsilon/3.$$

This implies for the fist part of the sum that

$$\begin{aligned} & P_{\mathcal{X}}(S_{\mathcal{X}} \cap (h^*\Delta h_{SU}) \cap (h_{SU}\Delta l)) - P_{\mathcal{X}}(S_{\mathcal{X}} \cap (h^*\Delta h_{SU}) \cap (h^*\Delta l)) \\ \leq\ & \mathrm{SU}(S_{\mathcal{X}} \cap (h^*\Delta h_{SU}) \cap (h_{SU}\Delta l)) + \epsilon/3 - \mathrm{SU}(S_{\mathcal{X}} \cap (h^*\Delta h_{SU}) \cap (h^*\Delta l)) + \epsilon/3 \\ \leq\ & 2\epsilon/3, \end{aligned}$$

where the last inequality follows from (1). This completes the proof. ■