Journal of Machine Learning Research 16 (2015) 2377-2404 Submitted 8/13; Revised 1/15; Published 12/15

Multiclass Learnability and the ERM Principle

Amit Daniely AMIT.DANIELY@MAIL.HUJI.AC.IL
Dept. of Mathematics, The Hebrew University, Givat-Ram Campus, Jerusalem 91904, Israel

Sivan Sabato SIVAN.SABATO@MICROSOFT.COM
Dept. of Computer Science, Ben-Gurion University of the Negev, Beer Sheva 8410501, Israel

Shai Ben-David SHAI@CS.UWATERLOO.CA
David R. Cheriton School of Computer Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1

Shai Shalev-Shwartz SHAIS@CS.HUJI.AC.IL
School of Computer Science and Engineering, The Hebrew University, Givat-Ram Campus, Jerusalem 91904, Israel

Editor: Peter Auer

Abstract

We study the sample complexity of multiclass prediction in several learning settings. For the PAC setting our analysis reveals a surprising phenomenon: In sharp contrast to binary classification, we show that there exist multiclass hypothesis classes for which some Empirical Risk Minimizers (ERM learners) have lower sample complexity than others. Furthermore, there are classes that are learnable by some ERM learners, while other ERM learners will fail to learn them. We propose a principle for designing good ERM learners, and use this principle to prove tight bounds on the sample complexity of learning *symmetric* multiclass hypothesis classes—classes that are invariant under permutations of label names. We further provide a characterization of mistake and regret bounds for multiclass learning in the online setting and the bandit setting, using new generalizations of Littlestone's dimension.

Keywords: multiclass, sample complexity, ERM

1. Introduction

Multiclass prediction is the problem of classifying an object into one of several possible target classes. This task surfaces in many domains. Common practical examples include document categorization, object recognition in computer vision, and web advertisement.

The centrality of the multiclass learning problem has spurred the development of various approaches for tackling this task. Most of these approaches fall under the following general description: There is an instance domain $\mathcal{X}$ and a set of possible class labels $\mathcal{Y}$. The goal of the learner is to learn a mapping from instances to labels. The learner receives training examples, and outputs a predictor which belongs to some hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, where $\mathcal{Y}^{\mathcal{X}}$ is the set of all functions from $\mathcal{X}$ to $\mathcal{Y}$. We study the sample complexity of the task of learning $\mathcal{H}$, namely, how many random training examples are needed for learning an accurate predictor from $\mathcal{H}$. This question has been extensively studied and is quite well understood for the binary case (i.e, where $|\mathcal{Y}| = 2$). In contrast, as we shall see, existing theory of the multiclass case is less complete.

In the first part of the paper we consider multiclass learning in the classical PAC setting of Valiant (1984). Since the 1970's, following Vapnik and Chervonenkis's seminal work on binary classification (Vapnik and Chervonenkis, 1971), it was widely believed that excluding trivialities, if a problem is at all learnable, then uniform convergence holds, and the problem is also learnable by every Empirical Risk Minimizer (ERM learner). The equivalence between learnability and uniform convergence has been proved for binary classification and for regression problems (Kearns et al., 1994; Bartlett et al., 1996; Alon et al., 1997). Recently, Shalev-Shwartz et al. (2010) have shown that in the general setting of learning of Vapnik (1995), learnability is not equivalent to uniform convergence. Moreover, some learning problems are learnable, but not with every ERM. In particular, this was shown for an unsupervised learning problem in the class of stochastic convex learning problems. The conclusion in Shalev-Shwartz et al. (2010) is that the conditions for learnability in the general setting are significantly more complex than in supervised learning. In this work we show that even in multiclass learning, uniform convergence is not equivalent to learnability. We find this result surprising, since multiclass prediction is very similar to binary classification.

This result raises once more the question of determining the true sample complexity of multiclass learning, and the optimal learning algorithm in this setting. We provide conditions under which tight characterization of the sample complexity of a multiclass hypothesis class can be provided. Specifically, we consider the important case of hypothesis classes which are invariant to renaming of class labels. We term such classes *symmetric* hypothesis classes. We show that the sample complexity for symmetric classes is tightly characterized by a known combinatorial measure called the Natarajan dimension. We conjecture that this result holds for non-symmetric classes as well.

We further study multiclass sample complexity in other learning models. Overall, we consider the following categorization of learning models:

- Interaction with the data source (batch vs. online protocols): In the batch protocol, we assume that the training data is generated i.i.d. by some distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. The goal is to find, with a high probability over the training samples, a predictor h such that $\Pr_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$ is as small as possible. In the online protocol we receive examples one by one, and are asked to predict the label of each given example on the fly. Our goal is to make as few prediction mistakes as possible in the worst case (see Littlestone 1987).

- The type of feedback (full information vs. bandits): In the full information setting, we receive the correct label of every example. In the bandit setting, the learner first sees an unlabeled example, and then outputs its prediction for the label. Then, a binary feedback is received, indicating only whether the prediction was correct or not, but not revealing the correct label in the case of a wrong guess (see for example Auer et al. 2003, 2002; Kakade et al. 2008).

The batch/full-information model is the standard PAC setting, while the online/full-information model is the usual online setting. The online/bandits model is the usual multiclass-bandits setting. We are not aware of a treatment of the batch/bandit model in previous works.

1.1 Paper Overview

After presenting formal definitions and notations in Section 2, we begin our investigation of multiclass sample complexity in the classical PAC learning setting. Previous results have provided upper and lower bounds on the sample complexity of multiclass learning in this setting when using any ERM algorithm. The lower bounds are controlled by the *Natarajan dimension*, a combinatorial measure which generalizes the VC dimension for the multiclass case, while the upper bounds are controlled by the *graph dimension*, which is another generalization of the VC dimension. The ratio between these two measures can be as large as $\Theta(\ln(k))$, where $k = |\mathcal{Y}|$ is the number of class labels. In Section 3 we survey known results, and also present a new improvement for the upper bound in the realizable case. All the bounds here are uniform, that is, they hold for all ERM learners.

These uniform bounds are the departure point of our research. Our goal is to find a combinatorial measure, similar to the VC-Dimension, that characterizes the sample complexity of a given class, up to logarithmic factors, *independent of the number of classes*. We delve into this challenge in Section 4. First, we show that no uniform bound on arbitrary ERM learners can tightly characterize the sample complexity: We describe a family of concept classes for which there exist 'good' ERM learners and 'bad' ERM learners, with a ratio of $\Theta(\ln(k))$ between their sample complexities. We further show that if k is infinite, then there are hypothesis classes that are learnable by some ERM learners but not by other ERM learners. Moreover, we show that for any hypothesis class, the sample complexity of the *worst* ERM learner in the realizable case is characterized by the graph dimension.

These results indicate that classical concepts which are commonly used to provide upper bounds for all ERM learners of some hypothesis class, such as the growth function, cannot lead to tight sample complexity characterization for the multiclass case. We thus propose algorithmic-dependent versions of these quantities, that allow bounding the sample complexity of specific ERM learners.

We consider three cases in which we show that the true sample complexity of multiclass learning in the PAC setting is fully characterized by the Natarajan dimension. The first case includes any ERM algorithm that does not use too many class labels, in a precise sense that we define via the new notion of *essential range* of an algorithm. In particular, the requirement is satisfied by any ERM learner which only predicts labels that appeared in the sample. The second case includes any ERM learner for symmetric hypothesis classes. The third case is the scenario where we have no prior knowledge on the different class labels, which we defined precisely in Section 4.3.

We conjecture that the upper bound obtained for symmetric classes holds for non-symmetric classes as well. Such a result cannot be implied by uniform convergence alone, since, by the results mentioned above, there always exist ERM learners with a sample complexity that is higher than this conjectured upper bound. It therefore follows that a proof of our conjecture will require the derivation of new learning rules. We hope that this would lead to new insights in other statistical learning problems as well.

In Section 5 we study multiclass learnability in the online model and in the bandit model. We introduce two generalizations of the Littlestone dimension, which characterize multiclass learnability in each of these models respectively. Our bounds are tight for the realizable case.

2. Problem Setting and Notation

Let $\mathcal{X}$ be a space, $\mathcal{Y}$ a discrete space[1] and $\mathcal{H}$ a class of functions from $\mathcal{X}$ to $\mathcal{Y}$. Denote $k = |\mathcal{Y}|$ (note that k can be infinite). For a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the error of a function $f : \mathcal{X} \to \mathcal{Y}$ with respect to $\mathcal{D}$ is defined as $\mathrm{Err}(f) = \mathrm{Err}_{\mathcal{D}}(f) = \Pr_{(x,y)\sim\mathcal{D}}(f(x) \neq y)$. The best error achievable by $\mathcal{H}$ on $\mathcal{D}$, namely, $\mathrm{Err}_{\mathcal{D}}(\mathcal{H}) := \inf_{f\in\mathcal{H}} \mathrm{Err}_{\mathcal{D}}(f)$, is called the *approximation error* of $\mathcal{H}$ on $\mathcal{D}$.

In the PAC setting, a *learning algorithm* for a class $\mathcal{H}$ is a function, $\mathcal{A} : \cup_{n=0}^{\infty}(\mathcal{X}\times\mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}$. We denote a training sequence by $S_m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. An *ERM learner* for class $\mathcal{H}$ is a learning algorithm that for any sample S_m returns a function that minimizes the empirical error relative to any other function in $\mathcal{H}$. Formally, the empirical error of a function f on a sample S_m is

$$\underset{S_m}{\mathrm{Err}}(f) = \frac{1}{m}|\{i \in [m] : f(x_i) \neq y_i\}|.$$

A learning algorithm $\mathcal{A}$ of class $\mathcal{H}$ is an ERM learner if $\mathrm{Err}_{S_m}(\mathcal{A}(S_m)) = \min_{f\in\mathcal{H}} \mathrm{Err}_{S_m}(f)$.

The *agnostic sample complexity* of a learning algorithm $\mathcal{A}$ is the function $m^a_{\mathcal{A},\mathcal{H}}$ defined as follows: For every $\epsilon, \delta > 0$, $m^a_{\mathcal{A},\mathcal{H}}(\epsilon, \delta)$ is the minimal integer such that for every $m \geq m^a_{\mathcal{A},\mathcal{H}}(\epsilon, \delta)$ and every distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$,

$$\Pr_{S_m\sim\mathcal{D}^m}\left(\underset{\mathcal{D}}{\mathrm{Err}}(\mathcal{A}(S_m)) > \underset{\mathcal{D}}{\mathrm{Err}}(\mathcal{H}) + \epsilon\right) \leq \delta. \tag{1}$$

Here and in subsequent definitions, we omit the subscript $\mathcal{H}$ when it is clear from context. If there is no integer satisfying the inequality above, define $m^a_{\mathcal{A}}(\epsilon, \delta) = \infty$. $\mathcal{H}$ is learnable with $\mathcal{A}$ if for all ϵ and δ the agnostic sample complexity is finite. The agnostic sample complexity of a class $\mathcal{H}$ is

$$m^a_{\mathrm{PAC},\mathcal{H}}(\epsilon, \delta) = \inf_{\mathcal{A}} m^a_{\mathcal{A},\mathcal{H}}(\epsilon, \delta) \ ,$$

where the infimum is taken over all learning algorithms for $\mathcal{H}$. The *agnostic ERM sample complexity* of $\mathcal{H}$ is the sample complexity that can be guaranteed for any ERM learner. It is defined by

$$m^a_{\mathrm{ERM},\mathcal{H}}(\epsilon, \delta) = \sup_{\mathcal{A}\in\mathrm{ERM}} m^a_{\mathcal{A},\mathcal{H}}(\epsilon, \delta) \ ,$$

where the supremum is taken over all ERM learners for $\mathcal{H}$. Note that always $m_{\mathrm{PAC}} \leq m_{\mathrm{ERM}}$.

We say that a distribution $\mathcal{D}$ is *realizable* by a hypothesis class $\mathcal{H}$ if there exists some $f \in \mathcal{H}$ such that $\mathrm{Err}_{\mathcal{D}}(f) = 0$. The *realizable sample complexity* of an algorithm $\mathcal{A}$ for a class $\mathcal{H}$, denoted $m^r_{\mathcal{A}}$, is the minimal integer such that for every $m \geq m^r_{\mathcal{A}}(\epsilon, \delta)$ and every distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ which is realizable by $\mathcal{H}$, Equation (1) holds. The realizable sample complexity of a class $\mathcal{H}$ is $m^r_{\mathrm{PAC},\mathcal{H}}(\epsilon, \delta) = \inf_{\mathcal{A}} m^r_{\mathcal{A}}(\epsilon, \delta)$, where the infimum is taken over all learning algorithms for $\mathcal{H}$. The realizable ERM sample complexity of a class $\mathcal{H}$ is $m^r_{\mathrm{ERM},\mathcal{H}}(\epsilon, \delta) = \sup_{\mathcal{A}\in\mathrm{ERM}} m^r_{\mathcal{A}}(\epsilon, \delta)$, where the supremum is taken over all ERM learners for $\mathcal{H}$.

Given a subset $S \subseteq \mathcal{X}$, we denote $\mathcal{H}|_S = \{f|_S : f \in \mathcal{H}\}$, where $f|_S$ is the restriction of f to S, namely, $f|_S : S \to \mathcal{Y}$ is such that for all $x \in S$, $f|_S(x) = f(x)$.

1. To avoid measurability issues, we assume that $\mathcal{X}$ and $\mathcal{Y}$ are countable.

3. Uniform Sample Complexity Bounds for ERM Learners

We first recall some known results regarding the sample complexity of multiclass learning. Recall the definition of the Vapnik-Chervonenkis dimension (Vapnik, 1995):

Definition 1 (VC dimension) *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class. A subset $S \subseteq \mathcal{X}$ is* shattered *by $\mathcal{H}$ if $\mathcal{H}|_S = \{0,1\}^S$. The* VC-dimension *of $\mathcal{H}$, denoted $\mathrm{VC}(\mathcal{H})$, is the maximal cardinality of a subset $S \subseteq \mathcal{X}$ that is shattered by $\mathcal{H}$.*

The VC-dimension, a cornerstone in statistical learning theory, characterizes the sample complexity of learning *binary* hypothesis classes, as the following bounds suggest.

Theorem 2 (Vapnik, 1995 and Bartlett and Mendelson, 2002) *There are absolute constants $C_1, C_2 > 0$ such that for every $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$,*

$$C_1 \left(\frac{\mathrm{VC}(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon} \right) \leq m^r_{\mathrm{PAC}}(\epsilon, \delta) \leq m^r_{\mathrm{ERM}}(\epsilon, \delta) \leq C_2 \left(\frac{\mathrm{VC}(\mathcal{H}) \ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon} \right),$$

and

$$C_1 \left(\frac{\mathrm{VC}(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon^2} \right) \leq m^a_{\mathrm{PAC}}(\epsilon, \delta) \leq m^a_{\mathrm{ERM}}(\epsilon, \delta) \leq C_2 \left(\frac{\mathrm{VC}(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon^2} \right).$$

One of the important implications of this result is that in binary classification, *all* ERM learners are as good, up to a multiplicative factor of $\ln(1/\epsilon)$.

It is natural to seek a generalization of the VC-dimension to hypothesis classes of non-binary functions. We recall two generalizations, both introduced by Natarajan (1989). In both generalizations, shattering of a set S is redefined by requiring that for any partition of S into T and $S \setminus T$, there exists a $g \in \mathcal{H}$ whose behavior on T differs from its behavior on $S \setminus T$. The two definitions are distinguished by their definition of "different behavior".

Definition 3 (Graph dimension) *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and let $S \subseteq \mathcal{X}$. We say that $\mathcal{H}$* G-shatters *S if there exists an $f : S \to \mathcal{Y}$ such that for every $T \subseteq S$ there is a $g \in \mathcal{H}$ such that*

$$\forall x \in T,\ g(x) = f(x),\ \text{and}\ \forall x \in S \setminus T,\ g(x) \neq f(x).$$

The graph dimension *of $\mathcal{H}$, denoted $d_G(\mathcal{H})$, is the maximal cardinality of a set that is G-shattered by $\mathcal{H}$.*

Definition 4 (Natarajan dimension) *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and let $S \subseteq \mathcal{X}$. We say that $\mathcal{H}$* N-shatters *S if there exist $f_1, f_2 : S \to \mathcal{Y}$ such that $\forall y \in S,\ f_1(y) \neq f_2(y)$, and for every $T \subseteq S$ there is a $g \in \mathcal{H}$ such that*

$$\forall x \in T,\ g(x) = f_1(x),\ \text{and}\ \forall x \in S \setminus T,\ g(x) = f_2(x).$$

The Natarajan dimension *of $\mathcal{H}$, denoted $d_N(\mathcal{H})$, is the maximal cardinality of a set that is N-shattered by $\mathcal{H}$.*

Both of these dimensions coincide with the VC-dimension for $k = 2$. Note also that we always have $d_N \leq d_G$. By reductions to and from the binary case, similarly to Natarajan (1989) and Ben-David et al. (1995) one can show the following result:

Theorem 5 *For the constants C_1, C_2 from Theorem 2, for every $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ we have*

$$C_1 \left(\frac{d_N(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon} \right) \leq m^r_{\text{PAC}}(\epsilon, \delta) \leq m^r_{\text{ERM}}(\epsilon, \delta) \leq C_2 \left(\frac{d_G(\mathcal{H}) \ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon} \right),$$

and

$$C_1 \left(\frac{d_N(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon^2} \right) \leq m^a_{\text{PAC}}(\epsilon, \delta) \leq m^a_{\text{ERM}}(\epsilon, \delta) \leq C_2 \left(\frac{d_G(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon^2} \right).$$

Proof (sketch) For the lower bound, let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class of Natarajan dimension d and Let $\mathcal{H}_d := \{0,1\}^{[d]}$. We claim that $m^r_{\text{PAC},\mathcal{H}_d} \leq m^r_{\text{PAC},\mathcal{H}}$, and similarly for the agnostic sample complexity, so the lower bounds are obtained by Theorem 2. Let $\mathcal{A}$ be a learning algorithm for $\mathcal{H}$. Consider the learning algorithm, $\bar{\mathcal{A}}$, for $\mathcal{H}_d$ defined as follows. Let $S = \{s_1, \ldots, s_d\} \subseteq X$ be a set and let f_0, f_1 be functions that witness the N-shattering of $\mathcal{H}$. Given a sample $((x_i, y_i))_{i=1}^m \subseteq [d] \times \{0,1\}$, let $g = \mathcal{A}((s_{x_i}, f_{y_i}(s_{x_i}))_{i=1}^m)$. $\bar{\mathcal{A}}$ returns $f : [d] \to \{0,1\}$ such that $f(i) = 1$ if and only if $g(s_i) = f_1(s_i)$. It is not hard to see that $m^r_{\bar{\mathcal{A}},\mathcal{H}_d} \leq m^r_{\mathcal{A},\mathcal{H}}$, thus $m^r_{\text{PAC},\mathcal{H}_d} \leq m^r_{\text{PAC},\mathcal{H}}$ and similarly for the agnostic case.

For the upper bound, let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class of graph dimension d. For every $f \in \mathcal{H}$ define $\bar{f} : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ by setting $\bar{f}(x, y) = 1$ if and only if $f(x) = y$ and let $\bar{\mathcal{H}} = \{\bar{f} : f \in \mathcal{H}\}$. It is not hard to see that $\text{VC}(\bar{\mathcal{H}}) = d_G(\mathcal{H})$. Let $\mathcal{A}$ be an ERM algorithm for $\mathcal{H}$. Let $\bar{\mathcal{A}}$ be an ERM algorithm for $\bar{\mathcal{H}}$ such that for a sample $(((x_i, z_i), y_i))_{i=1}^m \subseteq \mathcal{X} \times \mathcal{Y} \times \{0,1\}$, if for all i, $y_i = 1$, $\bar{\mathcal{A}}$ returns $\bar{f}$, where $f = \mathcal{A}((x_i, z_i)_{i=1}^m)$. It is easy to check that $\bar{\mathcal{A}}$ is consistent and therefore can be extended to an ERM learner for $\bar{\mathcal{H}}$, and that $m^r_{\mathcal{A},\mathcal{H}} \leq m^r_{\bar{\mathcal{A}},\bar{\mathcal{H}}}$. Thus $m^r_{\text{ERM},\mathcal{H}} \leq m^r_{\text{ERM},\bar{\mathcal{H}}}$. The analogous inequalities hold for the agnostic sample complexity as well. Thus the desired upper bounds follow from Theorem 2. ■

This theorem shows that the finiteness of the Natarajan dimension is a necessary condition for learnability, and the finiteness of the graph dimension is a sufficient condition for learnability. In Ben-David et al. (1995) it was proved that for every hypotheses class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$,

$$d_N(\mathcal{H}) \leq d_G(\mathcal{H}) \leq 4.67 \log_2(k) d_N(\mathcal{H}) \ . \tag{2}$$

It follows that if $k < \infty$ then the finiteness of the Natarajan dimension is both a necessary and a sufficient condition for learnability.[2] Incorporating Equation (2) into Theorem 5, it can be seen that the Natarajan dimension, as well as the graph dimension, characterize the sample complexity of $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ up to a multiplicative factor of $O(\ln(k) \ln(\frac{1}{\epsilon}))$. Precisely, the following result can be derived:

2. The result of Ben-David et al. (1995) in fact holds also for a rich family of generalizations of the VC dimension, of which the Graph dimension is one example.

Theorem 6 *There are constants C_1, C_2 such that, for every $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$,*

$$C_1\left(\frac{d_N(\mathcal{H})+\ln(\frac{1}{\delta})}{\epsilon}\right) \leq m^r_{\text{PAC}}(\epsilon,\delta) \leq m^r_{\text{ERM}}(\epsilon,\delta) \leq C_2\left(\frac{d_N(\mathcal{H})\ln(k)\cdot\ln(\frac{1}{\epsilon})+\ln(\frac{1}{\delta})}{\epsilon}\right),$$

and

$$C_1\left(\frac{d_N(\mathcal{H})+\ln(\frac{1}{\delta})}{\epsilon^2}\right) \leq m^a_{\text{PAC}}(\epsilon,\delta) \leq m^a_{\text{ERM}}(\epsilon,\delta) \leq C_2\left(\frac{d_N(\mathcal{H})\ln(k)+\ln(\frac{1}{\delta})}{\epsilon^2}\right).$$

3.1 An Improved Upper Bound for the Realizable Case

The following theorem provides a sample complexity upper bound which provides a tighter dependence on ϵ.

Theorem 7 *For every concept class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$,*

$$m^r_{\text{ERM}}(\epsilon,\delta) = O\left(\frac{d_N(\mathcal{H})\left(\ln(\frac{1}{\epsilon})+\ln(k)+\ln(d_N(\mathcal{H}))\right)+\ln(\frac{1}{\delta})}{\epsilon}\right).$$

The proof of this theorem is immediate given Theorem 13, which is provided in Section 4. We give the short proof of this theorem thereafter. While a proof for the Theorem can be established by a simple adaptation of previous techniques, we find it valuable to present this result here, as we could not find it in the literature.

4. PAC Sample Complexity with ERM Learners

In this section we study the sample complexity of multiclass ERM learners in the PAC setting. First, we show that unlike the binary case, in the multiclass setting different ERM learners can have very different sample complexities.

Example 1 (A Large Gap Between ERM Learners) *Let $\mathcal{X}$ be any finite or countable domain set. Let $\mathcal{P}_f(\mathcal{X})$ denote the collection of finite and co-finite subsets $A \subseteq \mathcal{X}$. We will take the label space to be $\mathcal{P}_f(\mathcal{X})$ together with a special label, denoted by $*$ (I.e. $\mathcal{Y} = \mathcal{P}_f(\mathcal{X}) \cup \{*\}$). For every $A \in \mathcal{P}_f(\mathcal{X})$, define $f_A : \mathcal{X} \to \mathcal{Y}$ by*

$$f_A(x) = \begin{cases} A & \text{if } x \in A \\ * & \text{otherwise,} \end{cases}$$

and consider the hypothesis class $\mathcal{H}_{\mathcal{X}} = \{f_A : A \in \mathcal{P}_f(\mathcal{X})\}$. It is not hard to see that $d_N(\mathcal{H}_{\mathcal{X}}) = 1$. On the other hand, if $\mathcal{X}$ is finite then $\mathcal{X}$ is G-shattered using the function $f_\emptyset$, therefore $d_G(\mathcal{H}_{\mathcal{X}}) = |\mathcal{X}|$. If $\mathcal{X}$ is infinite, then every finite subset of $\mathcal{X}$ is G-shattered, thus $d_G(\mathcal{H}_{\mathcal{X}}) = \infty$.

Consider two ERM algorithms for $\mathcal{H}_{\mathcal{X}}$, $\mathcal{A}_{\text{bad}}$ and $\mathcal{A}_{\text{good}}$, which satisfy the following properties. For $\mathcal{A}_{\text{bad}}$, whenever a sample of the form $S_m = \{(x_1,*),\ldots,(x_m,*)\}$ is observed, $\mathcal{A}_{\text{bad}}$ returns $f_{\{x_1,\ldots,x_m\}^c}$. Intuitively, while $\mathcal{A}_{\text{bad}}$ selects a hypothesis that minimizes the

empirical error, its choice for S_m seems to be sub-optimal. We will show later, based on Theorem 9, that the sample complexity of $\mathcal{A}_{\text{bad}}$ is $\Omega\left(\frac{|\mathcal{X}|+\ln(\frac{1}{\delta})}{\epsilon}\right)$.

For $\mathcal{A}_{\text{good}}$, we require that the algorithm only ever returns either $f_\emptyset$, or a hypothesis A such that the label A appeared in the sample—One can easily verify that there exists an ERM algorithm that satisfies this condition. Specifically, this means that for the sample $S_m = \{(x_1, *), \ldots, (x_m, *)\}$, $\mathcal{A}_{\text{good}}$ necessarily returns $f_\emptyset$. We have the following guarantee for $\mathcal{A}_{\text{good}}$:

Claim 1 *$m^r_{\mathcal{A}_{\text{good}},\mathcal{H}_\mathcal{X}}(\epsilon,\delta) \leq \frac{1}{\epsilon}\ln\frac{1}{\delta}$, and $m^a_{\mathcal{A}_{\text{good}},\mathcal{H}_\mathcal{X}}(\epsilon,\delta) \leq \frac{1}{\epsilon^2}\ln(\frac{1}{\epsilon})\ln\frac{1}{\delta}$.*

Proof *We prove the bound for the realizable case. The bound for the agnostic case will be immediate using Cor. 15, which we prove later.*

Let $\mathcal{D}$ be a distribution over $\mathcal{X}\times\mathcal{Y}$ and suppose that the correct labeling for $\mathcal{D}$ is f_A. Let m be the size of the sample. For any sample, $\mathcal{A}_{\text{good}}$ returns either $f_\emptyset$ or f_A. If it returns f_A then its error on $\mathcal{D}$ is zero. On the other hand, $\text{Err}_\mathcal{D}(f_\emptyset) = \Pr_{(X,Y)\sim\mathcal{D}}(X \in A)$. Thus, $\mathcal{A}_{\text{good}}$ returns a hypothesis with error ϵ or more only if $\Pr_{(X,Y)\sim\mathcal{D}}(X \in A) \geq \epsilon$ and all the m examples in the sample are from A^c. Assume $m \geq \frac{1}{\epsilon}\ln(\frac{1}{\delta})$, then the probability of the latter event is $(P(A^c))^m \leq (1-\epsilon)^m \leq e^{-\epsilon m} \leq \delta$. ■

This example shows that the gap between two different ERM learners can be as large as the gap between the Natarajan dimension and the graph dimension. By considering $\mathcal{H}_\mathcal{X}$ with an infinite $\mathcal{X}$, we conclude the following corollary.

Corollary 8 *There exist sets $\mathcal{X}$, $\mathcal{Y}$ and a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^\mathcal{X}$, such that $\mathcal{H}$ is learnable by some ERM learner but is not learnable by some other ERM learner.*

In Example 1, the bad ERM indeed requires as many examples as the graph dimension, while the good ERM requires only as many as the Natarajan dimension. Do such a 'bad' ERM and a 'good' ERM always exist? Our next result answers the question for the 'bad' ERM in the affirmative. Indeed, the graph dimension determines the learnability of $\mathcal{H}$ using the *worst* ERM learner.

Theorem 9 *There are constants $C_1, C_2 > 0$ such that the following holds. For every hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^\mathcal{X}$ of Graph dimension ≥ 2, there exists an ERM learner $\mathcal{A}_{\text{bad}}$ such that for every $\epsilon < \frac{1}{12}$ and $\delta < \frac{1}{100}$,*

$$C_1\left(\frac{d_G(\mathcal{H})+\ln(\frac{1}{\delta})}{\epsilon}\right) \leq m^r_{\mathcal{A}_{\text{bad}}}(\epsilon,\delta) \leq m^r_{\text{ERM}}(\epsilon,\delta) \leq C_2\left(\frac{d_G(\mathcal{H})\ln(\frac{1}{\epsilon})+\ln(\frac{1}{\delta})}{\epsilon}\right).$$

Proof The upper bound is simply a restatement of Theorem 5. It remains to prove that there exists an ERM learner, $\mathcal{A}_{\text{bad}}$, with $m^r_{\mathcal{A}_{bad}}(\epsilon,\delta) \geq C_1\left(\frac{d_G(\mathcal{H})+\ln(\frac{1}{\delta})}{\epsilon}\right)$.

First, assume that $d = d_G(\mathcal{H}) < \infty$. Let $S = \{x_0, \ldots, x_{d-1}\} \subseteq \mathcal{X}$ be a set which is G-Shattered by $\mathcal{H}$ using the function f_0. Let $\mathcal{A}_{\text{bad}}$ be an ERM learner with the following property. Upon seeing a sample $T \subseteq S$ which is consistent with f_0, $\mathcal{A}_{\text{bad}}$ returns a function

that coincides with f_0 on T and disagrees with f_0 on $S \setminus T$. Such a function exists since S is G-shattered using f_0.

Fix $\delta < \frac{1}{100}$ and $\epsilon < \frac{1}{12}$. Note that $1 - 2\epsilon \geq e^{-4\epsilon}$. Define a distribution on $\mathcal{X}$ by setting $\Pr(x_0) = 1 - 2\epsilon$ and for all $1 \leq i \leq d-1$, $\Pr(x_i) = \frac{2\epsilon}{d-1}$. Suppose that the correct hypothesis is f_0 and let $\{(X_i, f_0(X_i))\}_{i=1}^m$ be a sample. Clearly, the hypothesis returned by $\mathcal{A}_{\text{bad}}$ will err on all the examples from S which are not in the sample. By Chernoff's bound, if $m \leq \frac{d-1}{6\epsilon}$, then with probability at least $\frac{1}{100} \geq \delta$, the sample will include no more than $\frac{d-1}{2}$ examples from $S \setminus \{x_0\}$, so that the returned hypothesis will have error at least ϵ. To see that, define r.v. Y_i, $1 \leq i \leq m$ by setting $Y_i = 1$ if $X_i \neq x_0$ and 0 otherwise. By Chernoff's bound, if $r = \lfloor \frac{d-1}{6\epsilon} \rfloor$ then

$$\Pr\left(\sum_{i=1}^{m} Y_i \geq \frac{d-1}{2}\right) \leq \Pr\left(\sum_{i=1}^{r} Y_i \geq 3\epsilon k\right) \leq \exp\left(-\frac{\frac{1}{2}^2}{3} 2\epsilon r\right) < 0.99 \ .$$

Moreover, the probability that the sample includes only x_0 (and thus $\mathcal{A}_{\text{bad}}$ will return a hypothesis with error 2ϵ) is $(1-2\epsilon)^m \geq e^{-4\epsilon m}$, which is more than δ if $m \leq \frac{1}{4\epsilon}\ln(\frac{1}{\delta})$. We therefore obtain that

$$m^r_{\mathcal{A}_{\text{bad}}}(\epsilon, \delta) \geq \max\left\{\frac{d-1}{6\epsilon}, \frac{1}{2\epsilon}\ln(1/\delta)\right\} \geq \frac{d-1}{12\epsilon} + \frac{1}{4\epsilon}\ln(1/\delta) \ ,$$

as required.

If $d_G(\mathcal{H}) = \infty$, let S_n, $n = 2, 3, \ldots$ be a sequence of pairwise disjoint shattered sets such that $|S_n| = n$. For every n, suppose that f_0^n indicated that S_n is G-shattered. Let $\mathcal{A}_{\text{bad}}$ be an ERM learner with the following property. Upon seeing a sample $T \subseteq S_n$ labeled by f_0^n, $\mathcal{A}_{\text{bad}}$ returns a function that coincides with f_0^n on T and disagrees with f_0 on $S_n \setminus T$. Repeating the argument of the finite case for S_n instead of S shows that for every $\epsilon < \frac{1}{12}$ and $\delta < \frac{1}{100}$ it holds that $m_{\mathcal{A}_{\text{bad}}}(\epsilon, \delta) \geq C_1\left(\frac{n+\ln(\frac{1}{\delta})}{\epsilon}\right)$. Since it holds for every n, we conclude that $m^r_{\mathcal{A}_{\text{bad}}}(\epsilon, \delta) = \infty$. ■

To get the sample complexity lower bound for the ERM learner $\mathcal{A}_{\text{bad}}$ in Example 1, observe that this algorithm satisfies the specifications of a bad ERM algorithm from the proof above.

We conclude that for any multiclass learning problem there exists a 'bad' ERM learner. The existence of 'good' ERM learners turns out to be a more involved question. We conjecture that for every class there exists a 'good' ERM learner – that is, a learning algorithm whose realizable sample complexity is $\tilde{O}\left(\frac{d_N}{\epsilon}\right)$ (where the $\tilde{O}$ notation may hide poly-logarithmic factors of $\frac{1}{\epsilon}, d_N$ and $1/\delta$ but *not* of $|Y|$). As we describe in the rest of this section, in this work we prove this conjecture for several families of hypothesis classes.

What is the crucial feature that makes $\mathcal{A}_{\text{good}}$ better than $\mathcal{A}_{\text{bad}}$ in Example 1? For the realizable case, if the correct labeling is $f_A \in \mathcal{H}_{\mathcal{X}}$, then for *any* sample, $\mathcal{A}_{\text{good}}$ would return only one of at most two functions: either f_A or $f_\emptyset$. On the other hand, if the correct labeling is $f_\emptyset$, then $\mathcal{A}_{\text{bad}}$ might return *every* function in $\mathcal{H}_{\mathcal{X}}$. Thus, to return a hypothesis with error at most ϵ, $\mathcal{A}_{\text{good}}$ needs to reject at most one hypothesis, while $\mathcal{A}_{\text{bad}}$ might need to reject many more. Following this intuition, we propose the following rough principle: *A good ERM learner is one that, for every target hypothesis, considers a small number of hypotheses.*

We would like to use this intuition to design ERMs with a better sample complexity than the one that can be guaranteed for a general ERM as in Theorem 7. Classical sample complexity upper bounds that hold for all ERM learners hinge on the notion of a *growth function*, which counts the number of different hypotheses induced by the hypothesis class on a sample of a certain size. To bound the sample complexity of a specific ERM learner, we define algorithm-dependent variants of the concept of a growth function.

Definition 10 (Algorithm-dependent growth function) *Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. Let $\mathcal{A}$ be a learning algorithm for $\mathcal{H}$. For $m > 0$ and a sample $S = ((x_i, y_i))_{i=1}^{2m}$ of size $2m$, let $\mathcal{X}_S = \{x_1, \ldots, x_{2m}\}$, and define*

$$F_{\mathcal{A}}(S) = \{\mathcal{A}(S')|_{\mathcal{X}_S} \mid S' \subseteq S,\ |S'| = m\}.$$

Let $R(\mathcal{H})$ be the set of samples which are consistent with $\mathcal{H}$, that is $S = ((x_i, f(x_i)))_{i=1}^{2m}$ for some $f \in \mathcal{H}$. Define the realizable algorithm-dependent growth function *of $\mathcal{A}$ by*

$$\Pi^r_{\mathcal{A}}(m) = \sup_{S \in R(\mathcal{H}), |S|=2m} |F_{\mathcal{A}}(S)|.$$

Define the agnostic algorithm-dependent growth function *of $\mathcal{A}$ for sample S by*

$$\Pi^a_{\mathcal{A}}(m) = \sup_{S \in (\mathcal{X}\times\mathcal{Y})^{2m}} |F_{\mathcal{A}}(S)|.$$

These definitions enable the use of a 'double sampling' argument, similarly to the one used with the classical growth function (see Anthony and Bartlett, 1999, chapter 4). This argument is captured by the following lemma.

Lemma 11 (The Double Sampling Lemma) *Let $\mathcal{A}$ be an ERM learner, and let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. Denote $\epsilon = \mathrm{Err}_{\mathcal{D}}(\mathcal{A}(S_m)) - \mathrm{Err}_{\mathcal{D}}(\mathcal{H})$, and let $\delta \in (0, 1)$.*

1. *If $\mathcal{D}$ is realizable by $\mathcal{H}$ then with probability at least $1 - \delta$,*

$$\epsilon \leq 12 \ln(2\Pi^r_{\mathcal{A}}(m)/\delta)/m.$$

2. *For any $\mathcal{D}$, with probability at least $1 - \delta$,*

$$\epsilon \leq \sqrt{\frac{32 \ln((4\Pi^a_{\mathcal{A}}(m) + 4)/\delta)}{m}}.$$

Proof The proof idea of the this lemma is similar to the one of the 'double sampling' results of Anthony and Bartlett (1999) (see their Theorems 4.3 and 4.8).

For the first part of the claim, let $\mathcal{D}$ be a realizable distribution for $\mathcal{H}$. For $m \leq 8$, the claim trivially holds, therefore assume $m \geq 8$. Let $\nu = 12 \ln(2\Pi^r_{\mathcal{A}}(m)/\delta)/m$ and assume w.l.o.g. that $\nu \leq 1$.

Suppose that for some $S \in (\mathcal{X} \times \mathcal{Y})^m$, $\mathrm{Err}_{\mathcal{D}}(\mathcal{A}(S)) \geq \nu$. Let $T \in (\mathcal{X} \times \mathcal{Y})^m$ be another sample drawn from D^m, independently from S. We show that $Err_T(\mathcal{A}(S)) \geq \nu/2$ with probability at least $\frac{1}{2}$. For $\nu \leq \frac{1}{2}$, by Chernoff's bound, this holds with probability at least $1 - \exp(-m\nu/16)$, which is larger than $\frac{1}{2}$ by the definition of ν. For $\nu \geq \frac{1}{2}$, by Hoeffding's

inequality, this holds with probability at least $1-\exp(-m\nu^2/2) \geq 1-\exp(-m/8)$, which is larger than $\frac{1}{2}$, since $m \geq 8$. It follows that

$$\frac{1}{2} \Pr_{S\sim\mathcal{D}^m}(\operatorname*{Err}_{\mathcal{D}}(\mathcal{A}(S)) \geq \nu) \leq \Pr_{(S,T)\sim\mathcal{D}^{2m}}(\operatorname*{Err}_{T}(\mathcal{A}(S)) \geq \nu/2). \tag{3}$$

Let $Z = (z_1, \ldots, z_{2m}) \in R(\mathcal{H})$, and let $\sigma : [2m] \to [2m]$ be a permutation. We write Z^1_σ to mean $(z_{\sigma(1)}, \ldots, z_{\sigma(m)})$ and Z^2_σ to mean $(z_{\sigma(m+1)}, \ldots, z_{\sigma(2m)})$.

Similarly to Lemma 4.5 in Anthony and Bartlett (1999), for σ drawn uniformly from the set of permutations,

$$\begin{aligned}\Pr_{(S,T)\in\mathcal{D}^{2m}}(\operatorname*{Err}_{T}(\mathcal{A}(S)) \geq \nu/2) &= \mathop{\mathbb{E}}_{Z\sim\mathcal{D}^{2m}}(\Pr_{\sigma}(\operatorname*{Err}_{Z^2_\sigma}(\mathcal{A}(Z^1_\sigma)) \geq \nu/2)) \\ &\leq \sup_{Z\in R(\mathcal{H}),|Z|=2m} \Pr_{\sigma}(\operatorname*{Err}_{Z^2_\sigma}(\mathcal{A}(Z^1_\sigma)) \geq \nu/2).\end{aligned} \tag{4}$$

To bound the right hand side, note that since $\mathcal{A}$ is an ERM algorithm, for any fixed $Z \in R(\mathcal{H})$ and any σ, $\operatorname{Err}_{Z^1_\sigma}(\mathcal{A}(Z^1_\sigma)) = 0$. Thus

$$\Pr_{\sigma}(\operatorname*{Err}_{Z^2_\sigma}(\mathcal{A}(Z^1_\sigma)) \geq \nu/2) \leq \Pr_{\sigma}(\exists h \in F_{\mathcal{A}}(Z),\ \operatorname*{Err}_{Z^1_\sigma}(h) = 0 \text{ and } \operatorname*{Err}_{Z^2_\sigma}(h) \geq \nu/2).$$

For any fixed h, if the right hand side is not zero, then there exist at least $\nu m/2$ elements (x, y) in Z such that $h(x) \neq y$. In the latter case, the probability (over σ) that all such elements are in Z^2_σ is at most $2^{-\nu m/2}$. With a union bound over $h \in F_{\mathcal{A}}(Z)$, we conclude that for any Z,

$$\Pr_{\sigma}(\operatorname*{Err}_{Z^2_\sigma}(\mathcal{A}(Z^1_\sigma)) \geq \nu/2) \leq |F_{\mathcal{A}}(Z)|2^{-\nu m/2}.$$

Combining with Equation (4) gives

$$\Pr_{(S,T)\in\mathcal{D}^{2m}}(\operatorname*{Err}_{T}(\mathcal{A}(S)) \geq \nu/2) \leq \sup_{Z\in R(\mathcal{H})} |F_{\mathcal{A}}(Z)|2^{-\nu m/2} = \Pi^r_{\mathcal{A}}(m)2^{-\nu m/2}.$$

By Equation (3) and the definition of ν,

$$\Pr_{S\sim\mathcal{D}^m}(\operatorname*{Err}_{\mathcal{D}}(\mathcal{A}(S)) \geq \nu) \leq 2\Pi^r_{\mathcal{A}}(m)2^{-\nu m/2} \leq \delta.$$

This proves the first part of the claim.

For the second part of the claim, let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. Denote $\epsilon^* = \operatorname{Err}_{\mathcal{D}}(\mathcal{H})$, and let $h^* \in \mathcal{H}$ such that $\operatorname{Err}_{\mathcal{D}}(h^*) = \epsilon^*$.

Let $\nu = \sqrt{\frac{32\ln((4\Pi^a_{\mathcal{A}}(m)+4)/\delta)}{m}}$. Suppose that for some $S \in (\mathcal{X} \times \mathcal{Y})^m$, $\operatorname{Err}_{\mathcal{D}}(\mathcal{A}(S)) \geq \epsilon^* + \nu$. Let $T \in (\mathcal{X} \times \mathcal{Y})^m$ be a random sample drawn from D^m independently from S. By Hoeffding's inequality, with probability at least $1-\exp(-m\nu^2/2)$, which is at least $\frac{1}{2}$ by the definition of ν^2, $\operatorname{Err}_T(\mathcal{A}(S)) \geq \epsilon^* + \nu/2$. It follows that

$$\frac{1}{2} \Pr_{S\sim\mathcal{D}^m}(\operatorname*{Err}_{\mathcal{D}}(\mathcal{A}(S)) \geq \epsilon^* + \nu) \leq \Pr_{(S,T)\sim\mathcal{D}^{2m}}(\operatorname*{Err}_{T}(\mathcal{A}(S)) \geq \epsilon^* + \nu/2). \tag{5}$$

Let $Z = (z_1, \ldots, z_{2m}) \in (\mathcal{X} \times \mathcal{Y})^{2m}$, and let $\sigma : [2m] \to [2m]$ be a permutation. Denote Z^1_σ and Z^2_σ as above.

Denote $\mathcal{Z} = \{Z \in (\mathcal{X} \times \mathcal{Y})^{2m} \mid \mathrm{Err}_Z(\mathcal{A}(Z^1_\sigma)) \leq \epsilon^* + \nu/8\}$. By lemma 4.5 in Anthony and Bartlett (1999) again, for σ drawn uniformly from the set of permutations,

$$\begin{aligned}\Pr_{(S,T)\in\mathcal{D}^{2m}}(\underset{T}{\mathrm{Err}}(\mathcal{A}(S)) \geq \epsilon^* + \nu/2) &= \mathop{\mathbb{E}}_{Z\sim\mathcal{D}^{2m}}(\Pr_\sigma(\underset{Z^2_\sigma}{\mathrm{Err}}(\mathcal{A}(Z^1_\sigma)) \geq \epsilon^* + \nu/2)) \qquad (6)\\ &\leq \mathop{\mathbb{E}}_{Z\sim\mathcal{D}^{2m}}\left(\Pr_\sigma(\underset{Z^2_\sigma}{\mathrm{Err}}(\mathcal{A}(Z^1_\sigma)) \geq \epsilon^* + \nu/2)\Big| Z \in \mathcal{Z}\right) + \Pr(Z \notin \mathcal{Z}).\end{aligned}$$

To bound the right hand side, first note that by Hoeffding's inequality, the second term is bounded by

$$\Pr(Z \notin \mathcal{Z}) \leq \exp(-\nu^2 m/16). \qquad (7)$$

For the first term, $\mathrm{Err}_{Z^2_\sigma}(\mathcal{A}(Z^1_\sigma)) \geq \epsilon^* + \nu/2$ implies that unless $\mathrm{Err}_{Z^1_\sigma}(\mathcal{A}(Z^1_\sigma)) > \epsilon^* + \nu/4$, necessarily $\mathrm{Err}_{Z^2_\sigma}(\mathcal{A}(Z^1_\sigma)) - \mathrm{Err}_{Z^1_\sigma}(\mathcal{A}(Z^1_\sigma)) \geq \nu/4$. Since $\mathcal{A}$ is an ERM algorithm, $\mathrm{Err}_{Z^1_\sigma}(\mathcal{A}(Z^1_\sigma)) > \epsilon^* + \nu/4$ only if also $\mathrm{Err}_{Z^1_\sigma}(h^*) > \epsilon^* + \nu/4$. Therefore, for any Z,

$$\begin{aligned}&\Pr_\sigma(\underset{Z^2_\sigma}{\mathrm{Err}}(\mathcal{A}(Z^1_\sigma)) \geq \epsilon^* + \nu/2) \leq \\ &\quad \Pr_\sigma(\underset{Z^1_\sigma}{\mathrm{Err}}(h^*) > \epsilon^* + \nu/4) + \Pr_\sigma(\underset{Z^2_\sigma}{\mathrm{Err}}(\mathcal{A}(Z^1_\sigma)) - \underset{Z^1_\sigma}{\mathrm{Err}}(\mathcal{A}(Z^1_\sigma)) > \nu/4). \qquad (8)\end{aligned}$$

$\mathrm{Err}_{Z^1_\sigma}(h^*)$ is an average of m random variables of the form $\mathbb{I}[h^*(x_i) \neq y_i]$, that are sampled without replacement from the finite population Z, with population average $\mathrm{Err}_Z(h^*)$. For $Z \in \mathcal{Z}$, $\mathrm{Err}_Z(h^*) \leq \epsilon^* + \nu/8$. Therefore, by Hoeffding's inequality for sampling without replacements from a finite population (Hoeffding, 1963), for $Z \in \mathcal{Z}$,

$$\Pr_\sigma(\underset{Z^1_\sigma}{\mathrm{Err}}(h^*) > \epsilon^* + \nu/4) \leq \Pr_\sigma(\underset{Z^1_\sigma}{\mathrm{Err}}(h^*) - \underset{Z}{\mathrm{Err}}(h^*) > \nu/8) \leq \exp(-\nu^2 m/32). \qquad (9)$$

In addition, by the same inequality, and applying the union bound over $h \in F_\mathcal{A}(Z)$, for any Z

$$\begin{aligned}\Pr_\sigma(\underset{Z^2_\sigma}{\mathrm{Err}}(\mathcal{A}(Z^1_\sigma)) - \underset{Z^1_\sigma}{\mathrm{Err}}(\mathcal{A}(Z^1_\sigma)) > \nu/4) &\leq \Pr_\sigma(\exists h \in F_\mathcal{A}(Z), \underset{Z^2_\sigma}{\mathrm{Err}}(h) - \underset{Z^1_\sigma}{\mathrm{Err}}(h) > \nu/4)\\ &\leq \Pr_\sigma(\exists h \in F_\mathcal{A}(Z), \underset{Z^2_\sigma}{\mathrm{Err}}(h) - \underset{Z}{\mathrm{Err}}(h) > \nu/8) + \Pr_\sigma(\exists h \in F_\mathcal{A}(Z), \underset{Z^1_\sigma}{\mathrm{Err}}(h) - \underset{Z}{\mathrm{Err}}(h) > \nu/8)\\ &\leq 2\Pi^a_\mathcal{A}(m)\exp(-\nu^2 m/32). \qquad (10)\end{aligned}$$

Combined with Equation (8) and Equation (9), it follows that for $Z \in \mathcal{Z}$,

$$\Pr_\sigma(\underset{Z^2_\sigma}{\mathrm{Err}}(\mathcal{A}(Z^1_\sigma)) \geq \epsilon^* + \nu/2) \leq (2\Pi^a_\mathcal{A}(m) + 1)\exp(-\nu^2 m/32).$$

With Equation (5), Equation (6), and Equation (7), we conclude that

$$\Pr_{S\sim\mathcal{D}^m}(\underset{\mathcal{D}}{\mathrm{Err}}(\mathcal{A}(S)) \geq \epsilon^* + \nu) \leq (4\Pi^a_\mathcal{A}(m) + 4)\exp(-\nu^2 m/32) \equiv \delta.$$

The claim follows since $\epsilon = \mathrm{Err}_\mathcal{D}(\mathcal{A}(S)) - \epsilon^*$. ■

As we shall presently see, Lemma 11 can be used to provide better sample complexity bounds for some 'good' ERM learners.

4.1 Learning with a Small Essential Range

A key tool that we will use for providing better bounds is the notion of *essential range*, defined below. The essential range of an algorithm quantifies the number of different labels that can be emitted by the functions the algorithm might return for samples of a given size. In this definition we use the notion of the range of a function. Formally, for a function $f : \mathcal{X} \to \mathcal{Y}$, its range is the set of labels to which it maps $\mathcal{X}$, denoted by $\text{range}(f) = \{f(x) \mid x \in \mathcal{X}\}$.

Definition 12 (Essential range) *Let $\mathcal{A}$ be a learning algorithm for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. The* realizable essential range *of $\mathcal{A}$ is the function $r^r_{\mathcal{A}} : \mathbb{N} \to \mathbb{N}$, defined as follows.*

$$r^r_{\mathcal{A}}(m) = \sup_{S \in R(\mathcal{H}), |S|=2m} \left| \cup_{S' \subset S,\, |S'|=m} \text{range}(\mathcal{A}(S')) \right| .$$

The agnostic essential range *of $\mathcal{A}$ is the function $r^a_{\mathcal{A}} : \mathbb{N} \to \mathbb{N}$, defined as follows.*

$$r^a_{\mathcal{A}}(m) = \sup_{S \subseteq \mathcal{X} \times \mathcal{Y}, |S|=2m} \left| \cup_{S' \subset S,\, |S'|=m} \text{range}(\mathcal{A}(S')) \right| .$$

Intuitively, an algorithm with a small essential range uses a smaller set of labels for any particular distribution, thus it enjoys better convergence guarantees. This is formally quantified in the following result.

Theorem 13 *Let $\mathcal{A}$ be an ERM learning algorithm for $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with essential ranges $r^r_{\mathcal{A}}(m)$ and $r^a_{\mathcal{A}}(m)$. Denote $\epsilon = \text{Err}_{\mathcal{D}}(\mathcal{A}(S_m)) - \text{Err}_{\mathcal{D}}(\mathcal{H})$. Then,*

- *If $\mathcal{D}$ is realizable by $\mathcal{H}$ and $\delta < 0.1$ then with probability at least $1 - \delta$,*

$$\epsilon \leq O\left(\frac{d_N(\mathcal{H})(\ln(m) + \ln(r^r_{\mathcal{A}}(m))) + \ln(1/\delta)}{m} \right) .$$

- *For any probability distribution D, with probability at least $1 - \delta$,*

$$\epsilon \leq O\left(\sqrt{\frac{d_N(\mathcal{H})(\ln(m) + \ln(r^a_{\mathcal{A}}(m)) + \ln(1/\delta)}{m}} \right) .$$

To prove the realizable part of this theorem, we use the following combinatorial lemma by Natarajan:

Lemma 14 *(Natarajan, 1989) For every hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, $|\mathcal{H}| \leq |\mathcal{X}|^{d_N(\mathcal{H})} |\mathcal{Y}|^{2d_N(\mathcal{H})}$.*

Proof [of Theorem 13] For the realizable sample complexity, the growth function can be bounded as follows. Let $S \in R(\mathcal{H})$ such that $|S| = 2m$, and consider the function class $F_{\mathcal{A}}(S)$ (see Definition 10). By definition, the domain of $F_{\mathcal{A}}(S)$ is $\mathcal{X}_S$ of size $2m$, and the range of $F_{\mathcal{A}}(S)$ is of size at most $r^r_{\mathcal{A}}(m)$. Lastly, the Natarajan dimension of $F_{\mathcal{A}}(S)$ is at most $d_N(\mathcal{H})$, since $F_{\mathcal{A}}(S) \subseteq \mathcal{H}|_S$.

Therefore, by Lemma 14, $|F_{\mathcal{A}}(S)| \leq (2m)^{d_N(\mathcal{H})} r^r_{\mathcal{A}}(m)^{2d_N(\mathcal{H})}$. Taking the supremum over all such S, we get

$$\Pi^r_{\mathcal{A}}(m) \leq (2m)^{d_N(\mathcal{H})} r^r_{\mathcal{A}}(m)^{2d_N(\mathcal{H})}.$$

The bound on ϵ follows from the first part of Lemma 11.

For the agnostic sample complexity, a similar argument shows that

$$\Pi^a_{\mathcal{A}}(m) \leq (2m)^{d_N(\mathcal{H})} r^a_{\mathcal{A}}(m)^{2d_N(\mathcal{H})},$$

and the bound on ϵ follows from the second part of Lemma 11. ■

Theorem 7, which provides an improved bound for the realizable case, now follows from the fact that the essential range is never more than k. But the essential range can also be much smaller than k. For example, the essential range of the algorithm from Example 1 is bounded by $2m+1$ (the $2m$ labels appearing in the sample together with the $*$ label). In fact, we can state a more general bound, for any algorithm which never 'invents' labels it did not observe in the sample.

Corollary 15 *Let $\mathcal{A}$ be an ERM learner for a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. Suppose that for every sample S, the function $\mathcal{A}(S)$ never outputs labels which have not appeared in S. Then*

$$m^r_{\mathcal{A}}(\epsilon,\delta) = O\left(\frac{d_N(\mathcal{H})(\ln(\frac{1}{\epsilon}) + \ln(d_N(\mathcal{H}))) + \ln(\frac{1}{\delta})}{\epsilon}\right),$$

and

$$m^a_{\mathcal{A}}(\epsilon,\delta) = O\left(\frac{d_N(\mathcal{H})(\ln(\frac{1}{\epsilon}) + \ln(d_N(\mathcal{H}))) + \ln(\frac{1}{\delta})}{\epsilon^2}\right).$$

This corollary is immediate from Theorem 13 by setting $r^r_{\mathcal{A}}(m) = r^a_{\mathcal{A}}(m) = 2m$.

From this corollary, we immediately get that every hypothesis class which admits such algorithms, and has a large gap between the Natarajan dimension and the graph dimension realizes a gap between the sample complexities of different ERM learners. Indeed, the graph dimension can even be unbounded, while the Natarajan dimension is finite and the problem is learnable. This is demonstrated by the following example.

Example 2 *Denote the ball in $\mathbb{R}^n$ with center z and radius r by $B_n(z,r) = \{x \mid \|x-z\| \leq r\}$. For a given ball $B = B_n(z,r)$ with $z \in \mathbb{R}^n$ and $r > 0$, let $h_B : \mathbb{R}^n \to \mathbb{R}^n \cup \{*\}$ be the function defined by $h_B(x) = z$ if $x \in B$ and $h_B(x) = *$ otherwise. Let h_* be a hypothesis that always returns $*$. Define the hypothesis class $\mathcal{H}_n$ of hypotheses from $\mathbb{R}^n$ to $\mathbb{R}^n \cup \{*\}$ by*

$$\mathcal{H}_n = \{h_B \mid \exists z \in \mathbb{R}^n, \infty \geq r > 0, \text{ such that } B = B_n(z,r)\} \cup \{h_*\}.$$

Relying on the fact that the VC dimension of balls in $\mathbb{R}^n$ is $n+1$, it is not hard to see that $d_G(\mathcal{H}_n) = n+1$. Also, it is easy to see that $d_N(\mathcal{H}_n) = 1$. It is not hard to see that there exists an ERM, $\mathcal{A}_{\text{good}}$, satisfying the requirements of Corollary 15. Thus,

$$m^r_{\mathcal{A}_{\text{good}}}(\epsilon,\delta) \leq O\left(\frac{\ln(1/\epsilon) + \ln(1/\delta)}{\epsilon}\right), \quad m^a_{\mathcal{A}_{\text{good}}}(\epsilon,\delta) \leq O\left(\frac{\ln(1/\delta)}{\epsilon^2}\right).$$

On the other hand, Theorem 9 implies that there exists a bad ERM learner, $\mathcal{A}_{\text{bad}}$ with

$$m^a_{\mathcal{A}_{\text{bad}}}(\epsilon,\delta) \geq m^r_{\mathcal{A}_{\text{bad}}}(\epsilon,\delta) \geq C_1\left(\frac{n + \ln(1/\delta)}{\epsilon}\right).$$

Our results so far show that whenever an ERM learner with a small essential range exists, the sample complexity of learning the multiclass problem can be improved over the worst ERM learner. In the next section we show that this is indeed the case for hypothesis classes which satisfy a natural condition of *symmetry*.

4.2 Learning with Symmetric Classes

We say that a hypothesis class $\mathcal{H}$ is symmetric if for any $f \in \mathcal{H}$ and any permutation $\phi : \mathcal{Y} \to \mathcal{Y}$ on labels we have that $\phi \circ f \in \mathcal{H}$ as well. Symmetric classes are a natural choice if there is no prior knowledge on properties of specific labels in $\mathcal{Y}$ (See also the discussion in Section 4.3.1 below). We now show that for symmetric classes, the Natarajan dimension characterizes the optimal sample complexity up to logarithmic factors. It follows that a finite Natarajan dimension is a necessary and sufficient condition for learnability of a symmetric class. We will make use of the following lemma, which provides a key observation on symmetric classes.

Lemma 16 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a symmetric hypothesis class of Natarajan dimension d. Then any $h \in \mathcal{H}$ has a range of size at most $2d+1$.*

Proof If $k \leq 2d+1$ we are done. Thus assume that there are $2d+2$ distinct elements $y_1, \ldots, y_{2d+2} \in \mathcal{Y}$. Assume to the contrary that there is a hypothesis $h \in \mathcal{H}$ with a range of more than $2d+1$ values. Thus there is a set $S = \{x_1, \ldots, x_{d+1}\} \subseteq \mathcal{X}$ such that $h|_S$ has $d+1$ values in its range. Since $\mathcal{H}$ is symmetric, we can show that $\mathcal{H}$ N-shatters S as follows: Since $\mathcal{H}$ is symmetric, we can rename all the labels in the range of $h|_S$ as we please and get another function in $\mathcal{H}$. Thus there are two functions $f_1, f_2 \in \mathcal{H}$ such that for all $i \leq d+1$, $f_1(x_i) = y_i$ and $f_2(x_i) = y_{d+1+i}$. Now, let $S \subseteq T$. Since $\mathcal{H}$ is symmetric we can again rename the labels in the range of $h|_S$ to get a function $g \in \mathcal{H}$ such that $g(x) = f_1(x)$ for every $x \in T$ and $g(x) = f_2(x)$ for every $x \in S \setminus T$. Therefore the set S is shattered, thus the Natarajan dimension of $\mathcal{H}$ is at least $d+1$, contradicting the assumption. ■

First, we provide an upper bound on the sample complexity of ERM in the realizable case.

Theorem 17 *There are absolute constants C_1, C_2 such that for every symmetric hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$*

$$C_1 \left(\frac{d_N(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon} \right) \leq m^r_{\text{ERM}}(\epsilon, \delta) \leq C_2 \left(\frac{d_N(\mathcal{H}) \left(\ln(\frac{1}{\epsilon}) + \ln(d_N(\mathcal{H})) \right) + \ln(\frac{1}{\delta})}{\epsilon} \right)$$

Proof The lower bound is a restatement of Theorem 5. For the upper bound, first note that if $k \leq 4d_N(\mathcal{H}) + 2$ the upper bound trivially follows from Theorem 7. Thus assume $k > 4d_N(\mathcal{H}) + 2$. We define an ERM learner $\mathcal{A}$ with a small essential range, as required in Theorem 13: Fix a set $Z \subseteq \mathcal{Y}$ of size $|Z| = 2d_N(\mathcal{H}) + 1$. Assume an input sample $(x_1, f(x_1)), \ldots, (x_m, f(x_m))$, and denote the set of labels that appear in the sample by $L = \{f(x_i) \mid i \in [m]\}$. We require that $\mathcal{A}$ return a hypothesis which is consistent with the sample and has range in $L \cup Z$.

To see that such an ERM learner exists, observe that by Lemma 16, the range of f has at most $2d_N(\mathcal{H})+1$ distinct labels. Therefore, there is a set $R \subseteq \mathcal{Y}$ such that $|R| \leq 2d_N(\mathcal{H})+1$

and the range of f is $L \cup R$. Due to the symmetry of $\mathcal{H}$, we can rename the labels in R to labels in Z, and get another function $g \in \mathcal{H}$, that is consistent with the sample and has range in $L \cup Z$. This function can be returned by $\mathcal{A}$.

The range of $\mathcal{A}$ over all samples that are labeled by a fixed function $f \in \mathcal{H}$ is thus in the union of Z and the range of f. $|Z| \leq 2d_N(\mathcal{H}) + 1$ and by Lemma 16, the range of f is also at most $2d_N(\mathcal{H}) + 1$. Therefore the realizable essential range of $\mathcal{A}$ is at most $4d_N(\mathcal{H}) + 2$. The desired bound for the sample complexity of $\mathcal{A}$ thus follows from Theorem 13.

We now show that the same bound in fact holds for all ERM learners for $\mathcal{H}$. Suppose that $\mathcal{A}'$ is an ERM learner for which the bound does not hold. Then there is a function f and a distribution D over $\mathcal{X} \times \mathcal{Y}$ which is consistent with f, and there are m, ϵ and δ for which $m \geq m^r_{\mathcal{A}}(\epsilon, \delta)$, such that with probability greater than δ over samples S_m, $\mathrm{Err}_{\mathcal{D}}(\mathcal{A}'(S_m)) - \mathrm{Err}_{\mathcal{D}}(\mathcal{H}) > \epsilon$. Consider $\mathcal{A}$ as defined above, with a set Z that does not overlap with the range of f. For every sample S_m consistent with f, denote $\hat{f} = \mathcal{A}'(S_m)$, and let $\mathcal{A}$ return g which results from renaming the labels in $\hat{f}$ as follows: For any label that appeared in S_m, the same label is used in g. For any label that did not appear in S_m, a label from Z is used instead. Clearly, $\mathrm{Err}_{\mathcal{D}}(\mathcal{A}(S_m)) \geq \mathrm{Err}_{\mathcal{D}}(\mathcal{A}'(S_m))$. But this contradicts the upper bounds on $m^r_{\mathcal{A}}(\epsilon, \delta)$. We conclude that the upper bound holds for all ERM learners. ■

Second, we have the following upper bound for the agnostic case.

Theorem 18 *There are absolute constants C_1, C_2 such that for every symmetric hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$*

$$C_1 \left(\frac{d_N(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon^2} \right) \leq m^a_{\mathrm{ERM}}(\epsilon, \delta) \leq C_2 \left(\frac{d_N(\mathcal{H}) \ln(\min\{d_N(\mathcal{H}), k\}) + \ln(\frac{1}{\delta})}{\epsilon^2} \right),$$

Proof [3] The lower bound is a restatements of Theorem 6. For the upper bound, first note that if $k \leq 6d_N(\mathcal{H})$ then the upper bound follows from Theorem 6. Thus assume $k \geq 6d_N(\mathcal{H}) \geq 4d_N(\mathcal{H}) + 2$. Fix a set $Z \subseteq \mathcal{Y}$ of size $|Z| = 4d_N(\mathcal{H}) + 2$. Denote $\mathcal{H}' = \{f \in \mathcal{H} : f(\mathcal{X}) \subseteq Z\}$. By Lemma 16, the range of every function in $\mathcal{H}$ contains at most $\frac{|Z|}{2}$ elements. Thus, by symmetry, it is easy to see that $d_G(\mathcal{H}) = d_G(\mathcal{H}')$ and $d_N(\mathcal{H}) = d_N(\mathcal{H}')$. By equation (2) and the fact that the range of functions in $\mathcal{H}'$ is Z, we conclude that

$$\begin{aligned} d_G(\mathcal{H}) = d_G(\mathcal{H}') &= O(d_N(\mathcal{H}') \ln(|Z|)) \\ &= O(d_N(\mathcal{H}') \ln(\min\{d_N(\mathcal{H}'), k\}) = O(d_N(\mathcal{H}) \ln(d_N(\mathcal{H}))). \end{aligned}$$

Using Theorem 5 we obtain the desired upper bounds. ■

These results indicate that for symmetric classes, the sample complexity is determined by the Natarajan dimension up to logarithmic factors. Moreover, the ratio between the sample complexities of worst ERM and the best ERM in this case is also at most logarithmic in ϵ and the Natarajan dimension. We present the following open question:

Open question 19 *Are there symmetric classes such that there are two different ERM learners with a sample complexity ratio of $\Omega(\ln(d_N))$ between them?*

3. We note that this proof show that for symmetric classes $d_G = O(d_N \log(d_N))$. Hence, it can be adopted to give a simpler proof of Theorem 17, but with a multiplicative (rather than additive) factor of $\log\left(\frac{1}{\epsilon}\right)$.

4.3 Learning with No Prior Knowledge on Labels

Suppose we wish to learn some multiclass problem and have some hypothesis class that we wish to use for learning. The hypothesis class is defined using arbitrary label names, say $\mathcal{Y} = \{1, \ldots, k\} = [k]$. In many learning problems, we do not have any prior knowledge on a preferred mapping between these arbitrary label names and the actual real-world labels (e.g., names of topics of documents). Thus, any mapping between the real-world class labels and the arbitrary labels in $[k]$ is as reasonable as any other. We formalize the last assertion by *assuming that this mapping is chosen uniformly at random* [4]. In this section we show that in this scenario, when $k = \Omega(d_N(\mathcal{H}))$, it is likely that we will achieve poor classification accuracy.

Formally, let $\mathcal{H} \subset [k]^{\mathcal{X}}$ be a hypothesis class. Let $\mathcal{L}$ be the set of real-world labels, $|\mathcal{L}| = k$. A mapping of the label names $[k]$ to the true labels $\mathcal{L}$ is a bijection $\phi : [k] \to \mathcal{L}$. For such ϕ we let $\mathcal{H}_\phi = \{\phi \circ f : f \in \mathcal{H}\}$. [5]

The following theorem lower-bounds the approximation error when ϕ is chosen at random. The result holds for any distribution with fairly balanced label frequencies. Formally, we say that $\mathcal{D}$ over $\mathcal{X} \times \mathcal{L}$ is *balanced* if for any $l \in \mathcal{L}$, the probability that a random pair drawn from $\mathcal{D}$ has label l is at most $10/k$.

Theorem 20 *Fix $\alpha > 0$. There exist a constant $C_\alpha > 0$ such that for any $k > 0$, any hypothesis class $\mathcal{H} \subseteq [k]^{\mathcal{X}}$ such that $d_N(\mathcal{H}) \leq C_\alpha k$, and any balanced distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{L}$, with probability at least $1 - o(2^{-k})$ over the choice of ϕ, $\mathrm{Err}_{\mathcal{D}}(\mathcal{H}_\phi) \geq 1 - \alpha$.*

Remark 21 *Theorem 20 is tight, in the sense that a similar proposition cannot be obtained for all $d_N \leq f(k)$ for some $f(k) \in \omega(k)$. To see this, consider the class $\mathcal{H} = [k]^{[k]}$, for which $d_N(\mathcal{H}) = k$. For any ϕ, $\mathcal{H}_\phi = \mathcal{H}$. Thus, for any distribution such that $\mathrm{Err}_{\mathcal{D}}(\mathcal{H}) = 0$, we have $\mathrm{Err}_{\mathcal{D}}(\mathcal{H}_\phi) = 0$.*

To prove Theorem 20, we prove the following lemma, which provides a lower bound on the error of any hypothesis with a random bijection.

Lemma 22 *Let $h : \mathcal{X} \to [k]$ and let $\phi : [k] \to \mathcal{L}$ be a bijection chosen uniformly at random. Let $S = \{(x_1, l_1), \ldots, (x_m, l_m)\} \subseteq \mathcal{X} \times \mathcal{L}$. Denote, for $l \in \mathcal{L}$, $\hat{p}_l = \frac{|\{j : l_j = l\}|}{m}$. Fix $\alpha > 0$, and let $\gamma = \frac{\alpha^2}{\sum_{l \in \mathcal{L}} \hat{p}_l^2}$. Then*

$$\Pr[\operatorname*{Err}_S(\phi \circ h) < 1 - \alpha] \leq \left(\frac{8ke}{\gamma^2}\right)^{\frac{\gamma}{2}}.$$

Proof Denote $P = \sqrt{\sum_{l \in \mathcal{L}} \hat{p}_l^2}$. For a sample $S \subset \mathcal{X} \times \mathcal{L}$ and a function $f : \mathcal{X} \to \mathcal{L}$ denote $\mathrm{Gain}_S(f) = 1 - \mathrm{Err}_S(f)$. For $l \in \mathcal{L}$ denote $S_l = ((x_i, l_i))_{i : l_i = l}$. By Cauchy-Schwartz, we have

$$\operatorname*{Gain}_S(\phi \circ h) = \sum_{l \in \mathcal{L}} \hat{p}_l \cdot \operatorname*{Gain}_{S_l}(\phi \circ h) \leq P \cdot \sqrt{\sum_{l \in \mathcal{L}} \left(\operatorname*{Gain}_{S_l}(\phi \circ h)\right)^2}.$$

4. We note also that choosing this mapping at random is sometimes advocated for multiclass learning, e.g., for a filter tree Beygelzimer et al. (2007) and for an Error Correcting Output Code (Dietterich and Bakiri, 1995; Allwein et al., 2000).

5. Several notions, originally defined w.r.t. functions from $\mathcal{X}$ to $\mathcal{Y}$ (e.g. $\mathrm{Err}_{\mathcal{D}}(h)$), can be naturally extended to functions from $\mathcal{X}$ to $\mathcal{L}$. We will freely use these extensions.

Assume that $\mathrm{Err}_S(\phi \circ h) \leq 1 - \alpha$. Then

$$\sum_{l \in \mathcal{L}} \operatorname*{Gain}_{S_l}(\phi \circ h) \geq \sum_{l \in \mathcal{L}} \left(\operatorname*{Gain}_{S_l}(\phi \circ h) \right)^2 \geq \frac{\left(\mathrm{Gain}_S(\phi \circ h)\right)^2}{P^2} \geq \frac{\alpha^2}{P^2} = \gamma.$$

Note first that the left hand side is at most k, thus $\gamma \leq k$. Since for every $l \in \mathcal{L}$ it holds that $0 \leq \mathrm{Gain}_{S_l}(\phi \circ h) \leq 1$, we conclude that there are at least $n = \lceil \frac{\gamma}{2} \rceil$ labels $l \in \mathcal{L}$ such that

$$\operatorname*{Gain}_{S_l}(\phi \circ h) \geq \frac{\gamma}{2k} \ .$$

For a fixed set of n labels $l_1, \ldots, l_n \in \mathcal{L}$, the probability that $\forall i,\ \mathrm{Gain}_{S_{l_i}}(\phi \circ h) \geq \frac{\gamma}{2k}$ is at most

$$\prod_{i=1}^{n} \frac{2k}{(k+1-i)\gamma} \leq \left(\frac{2k}{(k+1-n)\gamma} \right)^n \ .$$

To see that, suppose that ϕ is sampled by first choosing the value of $\phi^{-1}(l_1)$ then $\phi^{-1}(l_2)$ and so on. For every l_i, there are at most $\frac{2k}{\gamma}$ values for $\phi^{-1}(l_i)$ for which $\mathrm{Gain}_{S_{l_i}}(\phi \circ h) \geq \frac{\gamma}{2k}$. Thus, after the values of $\phi^{-1}(l_1), \ldots, \phi^{-1}(l_{i-1})$ have been determined, the probability that $\phi^{-1}(l_i)$ is one of these values is at most $\frac{2k}{(k+1-i)\cdot\gamma}$.

It follows that the probability that $\mathrm{Gain}_{S_l}(\phi \circ h) \geq \frac{\gamma}{2k}$ for n different labels l is at most

$$\begin{aligned}
\binom{k}{n} \cdot \left(\frac{2k}{(k+1-n)\gamma} \right)^n &\leq \left(\frac{ek}{n} \right)^n \cdot \left(\frac{2k}{(k+1-n)\gamma} \right)^n \\
&\leq \left(\frac{2ke}{\gamma} \right)^n \cdot \left(\frac{2k}{(k-\gamma/2)\gamma} \right)^n \\
&\leq \left(\frac{8ke}{\gamma^2} \right)^n .
\end{aligned}$$

If $\frac{8ke}{\gamma^2} \geq 1$ then the bound in the statement of the lemma holds trivially. Otherwise, the bound follows since $n \geq \gamma/2$. ■

Proof [Proof of Theorem 20] Denote $p_l = \Pr_{(X,L)\sim\mathcal{D}}[L = l]$. Let $S = \{(x_1, l_1), \ldots, (x_m, l_m)\} \subseteq \mathcal{X} \times \mathcal{L}$ be an i.i.d. sample drawn according to $\mathcal{D}$. Denote $\hat{p}_l = \frac{|\{j : l_j = l\}|}{m}$.

For any fixed bijection ϕ, by Theorem 6, with probability $1 - \delta$ over the choice of S,

$$\operatorname*{Err}_{\mathcal{D}}(\mathcal{H}_\phi) \geq \inf_{h \in \mathcal{H}} \mathrm{Err}_S(\phi \circ h) - O\left(\sqrt{\frac{\ln(k) d_N(\mathcal{H}) + \ln(1/\delta)}{m}} \right).$$

Since there are less than k^k such bijections, we can apply the union bound to get that with probability $1 - \delta$ over the choice of S,

$$\forall \phi, \quad \operatorname*{Err}_{\mathcal{D}}(\mathcal{H}_\phi) \geq \inf_{h \in \mathcal{H}} \mathrm{Err}_S(\phi \circ h) - O\left(\sqrt{\frac{\ln(k) d_N(\mathcal{H}) + k\ln(k) + \ln(1/\delta)}{m}} \right).$$

Assume $k \geq C \cdot d_N(\mathcal{H})$ for some constant $C > 0$, and let $m = \Theta\left(\frac{k \cdot \ln(k)}{\alpha^2}\right)$ such that with probability at least 3/4,

$$\forall \phi, \quad \operatorname*{Err}_{\mathcal{D}}(\mathcal{H}_\phi) \geq \inf_{h \in \mathcal{H}} \operatorname{Err}_S(\phi \circ h) - \alpha/2. \tag{11}$$

We have

$$E[\sum_{l \in \mathcal{L}} \hat{p}_l^2] = 2\frac{1}{m^2} \sum_{l \in \mathcal{L}} \left(\binom{m}{2} p_l^2 + m p_l \right) \leq 2k \cdot \left(\frac{m(m-1)}{2m^2} \frac{100}{k^2} + \frac{10}{mk} \right) \leq \frac{120}{k}.$$

Thus, by Markov's inequality, with probability at least $\frac{1}{2}$ over the samples we have

$$\sum_{l \in \mathcal{L}} \hat{p}_l^2 < \frac{240}{k}. \tag{12}$$

Thus, with probability at least 1/4, both (12) and (11) hold. In particular, there exists a single sample S for which both (12) and (11) hold. Let us fix such an $S = \{(x_1, l_1), \ldots, (x_m, l_m)\}$.

Assume now that $\phi : \mathcal{Y} \to \mathcal{L}$ is sampled uniformly. For a fixed $h \in \mathcal{H}$ and for $\gamma = (\alpha/2)^2 / \sum_{l \in \mathcal{L}} \hat{p}_l^2 \geq k\alpha^2/960$, we have, by Lemma 22 that

$$\Pr_{\phi}\left[\operatorname{Err}_S(\phi \circ h) < 1 - \frac{\alpha}{2}\right] \leq \left(\frac{8ke}{\gamma^2}\right)^{\frac{\gamma}{2}} \leq (C_1 k \alpha^4)^{-C_2 k \alpha^2} := \eta,$$

for constants $C_1, C_2 > 0$. By Lemma 14, $|\mathcal{H}|_{\{x_1,\ldots,x_m\}}| \leq (m \cdot k)^{2d_N(\mathcal{H})}$. Thus, with probability $\geq 1 - (m \cdot k)^{2d} \cdot \eta$ over the choice of ϕ, $\inf_{h \in \mathcal{H}} \operatorname{Err}_S(\phi \circ h) \geq 1 - \frac{\alpha}{2}$ and by (11) also

$$\operatorname{Err}_{\mathcal{D}}(\mathcal{H}_\phi) \geq 1 - \alpha. \tag{13}$$

By our choice of m, and since $k \geq d_N(\mathcal{H})$, for some universal constant $C_1 \geq 1$, $m \leq C_1 \cdot \frac{k^2}{\alpha^2}$. Considering α a constant, we have, for some constants $C_i > 0$,

$$(m \cdot k)^{2d_N(\mathcal{H})} \cdot \eta \leq (C_3 k)^{6d_N(\mathcal{H})} \cdot (C_4 k)^{-C_5 k}.$$

By requiring that $k \geq 12 d_N(\mathcal{H})/C_5$, we get that the right hand side is at most $o(2^{-k})$.

■

4.3.1 Symmetrization

From Theorem 20 it follows that if there is no prior knowledge about the labels, and the label frequencies are balanced, we must use a class of Natarajan dimension $\Omega(k)$ to obtain reasonable approximation error. As we show next, in this case, there is almost no loss in the sample complexity if one instead uses the *symmetrization* of the class, obtained by considering all the possible label mappings $\phi : [k] \to \mathcal{L}$. Formally, let $\mathcal{H} \subset [k]^{\mathcal{X}}$ be some hypothesis class and let $\mathcal{L}$ be a set with $|\mathcal{L}| = k$. The symmetrization of $\mathcal{H}$ is the symmetric class

$$\mathcal{H}_{\text{sym}} = \{\phi \circ h \mid h \in \mathcal{H},\ \phi : [k] \to \mathcal{L} \text{ is a bijection}\}.$$

Lemma 23 *Let $\mathcal{H} \subseteq [k]^{\mathcal{X}}$ be a hypothesis class with Natarajan dimension d. Then*

$$d_N(\mathcal{H}_{\text{sym}}) = O(\max\{d\log(d), k\log(k)\}).$$

Proof Let $d_s = d_N(\mathcal{H}_{\text{sym}})$. Let $X \subset \mathcal{X}$ be a set of cardinality d_s that is N-shattered by $\mathcal{H}_{\text{sym}}$. By Lemma 14, $|\mathcal{H}|_X| \leq (d_s k^2)^d$. It follows that $|\mathcal{H}_{\text{sym}}|_X| \leq k!(d_s k^2)^d$. On the other hand, since $\mathcal{H}_{\text{sym}}$ N-shatters X, $|\mathcal{H}_{\text{sym}}|_X| \geq 2^{|X|} = 2^{d_s}$. It follows that $2^{d_s} \leq k!(d_s k^2)^d$. Taking logarithms we obtain that $d_s \leq k\log(k) + d(\ln(d_s) + 2\ln(k))$. The Lemma follows. ■

5. Other Learning Settings

In this section we consider the characterization of learnability in other learning settings: The online setting and the bandit setting.

5.1 The Online Model

Learning in the online model is conducted in a sequence of consecutive rounds. On each round $t = 1, 2, \ldots, T$, the environment presents a sample $x_t \in \mathcal{X}$, then the algorithm should predict a value $\hat{y}_t \in \mathcal{Y}$, and finally the environment reveals the correct value $y_t \in \mathcal{Y}$. The prediction at time t can be based only on the examples $x_1, \ldots, x_t$ and the previous outcomes $y_1, \ldots, y_{t-1}$. Our goal is to minimize the number of prediction mistakes in the worst case, where the number of mistakes on the first T rounds is $L_T = |\{t \in [T] : \hat{y}_t \neq y_t\}|$. Assume a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. In the realizable setting, we assume that for some function $f \in \mathcal{H}$ all the outcomes are evaluations of f, namely, $y_t = f(x_t)$.

Learning in the realizable online model has been studied by Littlestone (1987), who showed that a combinatorial measure, called the Littlestone dimension, characterizes the min-max optimal number of mistakes for *binary* hypotheses classes in the realizable case. We propose a generalization of the Littlestone dimension to multiclass hypotheses classes.

Consider a rooted tree T whose internal nodes are labeled by elements from $\mathcal{X}$ and whose edges are labeled by elements from $\mathcal{Y}$, such that the edges from a single parent to its child-nodes are each labeled with a different label. The tree T is *shattered* by $\mathcal{H}$ if, for every path from root to leaf which traverses the nodes $x_1, \ldots, x_k$, there is a function $f \in \mathcal{H}$ such that $f(x_i)$ is the label of the edge (x_i, x_{i+1}). We define the *Littlestone dimension* of a multiclass hypothesis class $\mathcal{H}$, denoted L-Dim($\mathcal{H}$), to be the maximal depth of a complete binary tree that is shattered by $\mathcal{H}$ (or ∞ if there are a shattered trees for arbitrarily large depth).

As we presently show, the number L-Dim($\mathcal{H}$) fully characterizes the worst-case mistake bound for the online model in the realizable setting. The upper bound is achieved using the following algorithm.

Algorithm: Standard Optimal Algorithm (SOA)
Initialization: $V_0 = \mathcal{H}$.
For $t = 1, 2 \ldots$,
receive x_t

for $y \in \mathcal{Y}$, let $V_t^{(y)} = \{f \in V_{t-1} : f(x_t) = y\}$
predict $\hat{y}_t \in \arg\max_y \text{L-Dim}(V_t^{(y)})$
receive true answer y_t
update $V_t = V_t^{(y_t)}$

Theorem 24 *The SOA algorithm makes at most* L-Dim($\mathcal{H}$) *mistakes on any realizable sequence. Furthermore, the worst-case number of mistakes of* any *deterministic online algorithm is at least* L-Dim($\mathcal{H}$)*. For any randomized online algorithm, the expected number of mistakes on the worst sequence is at least* $\frac{1}{2}$ L-Dim($\mathcal{H}$)*.*

Proof (sketch) First, we show that the *SOA* algorithm makes at most L-Dim($\mathcal{H}$) mistakes. The proof is a simple adaptation of the proof of the binary case (see Littlestone, 1987; Shalev-Shwartz, 2012). We note that for each t there is at most one $y \in \mathcal{Y}$ with L-Dim($V_t^{(y)}$) = L-Dim(V_t), and for the rest of the labels we have L-Dim($V_t^{(y)}$) < L-Dim(V_t) (otherwise, it is not hard to construct a tree of depth L-Dim(V_t) + 1, whose root is x_t, that is shattered by V_t). Thus, whenever the algorithm errs, the Littlestone dimension of V_t decreases by at least 1, so after L-Dim($\mathcal{H}$) mistakes, V_t is composed of a single function.

For the second part of the theorem, it is not hard to see that, given a shattered tree of depth L-Dim($\mathcal{H}$), the environment can force any deterministic online learning algorithm to make L-Dim($\mathcal{H}$) mistakes. Note also that allowing the algorithm to make randomized predictions cannot be too helpful. It is easy to see that given a shattered tree of depth L-Dim($\mathcal{H}$), the environment can enforce any randomized online learning algorithm to make at least L-Dim($\mathcal{H}$)/2 mistakes on average, by traversing the shattered tree, and providing at every round the label that the randomized algorithm is less likely to predict. ■

In the agnostic case, the sequence of outcomes, $y_1, \ldots, y_m$, is not necessarily consistent with some function $f \in \mathcal{H}$. Thus, one wishes to bound the *regret* of the algorithm, instead of its absolute number of mistakes. The regret is the difference between the number of mistakes made by the algorithm and the number of mistakes made by the best-matching function $f \in \mathcal{H}$. The agnostic case for classes of binary-output functions has been studied in Ben-David et al. (2009). It was shown that, as in the realizable case, the Littlestone dimension characterizes the optimal regret bound.

We show that the generalized Littlestone dimension characterizes the optimal regret bound for the multiclass case as well. The proof follows the paradigm of 'learning with expert advice' (see e.g. Cesa-Bianchi and Lugosi, 2006; Shalev-Shwartz, 2012), which we now briefly describe. Suppose that at each step, t, before the algorithm chooses its prediction, it observes N *advices* $(f_1^t, \ldots, f_N^t) \in \mathcal{Y}^N$, which can be used to determine its prediction. We think of f_i^t as the prediction made by the *expert* i at time t and denote the *loss* of the expert i at time T by $L_{i,T} = |\{t \in [T] : f_{i,t} \neq y_t\}|$. The goal here it to devise an algorithm that achieves a loss which is comparable with the loss of the best expert. Given T, the following algorithm (Cesa-Bianchi and Lugosi, 2006, chapter 2) achieves expected loss at most $\min_{i \in [N]} L_{i,T} + \sqrt{\frac{1}{2}\ln(N)T}$.

Algorithm: Learning with Expert Advice (LEA)
Parameters: Time horizon – T

Set $\eta = \sqrt{8\ln(N)/T}$
For $t = 1, 2 \ldots, T$
receive expert advices $(f_1^t, \ldots, f_N^t) \in \mathcal{Y}^N$
predict $\hat{y}_t = f_{i,t}$ with probability proportional to $\exp(-\eta L_{i,t-1})$
receive true answer y_t

We use this algorithm and its guarantee to prove the following theorem.

Theorem 25 *In the agnostic online multiclass setting, the expected loss of the optimal algorithm on the worst-case sequence is at most* $\min_{f\in\mathcal{H}} L_{f,T} + \sqrt{\frac{1}{2}\,\text{L-Dim}(\mathcal{H})T\log(Tk)}$.

Proof First, we construct an expert for every $f \in \mathcal{H}$, whose advice at time t is $f(x_t)$. Denote the loss of the expert corresponding to f at time t by $L_{f,t}$. Running the algorithm LEA with this set of experts yields an algorithm whose expected error is at most $\min_{f\in\mathcal{H}} L_{f,T} + \sqrt{\frac{1}{2}\ln(|\mathcal{H}|)T}$. Our goal now is to construct a more compact set of experts, which will allow us to bound the loss in terms of $\text{L-Dim}(\mathcal{H})$ instead of $\ln(|\mathcal{H}|)$.

Given time horizon T, let $A_T = \{A \subset [T] \mid |A| \leq \text{L-Dim}(\mathcal{H})\}$. For every $A \in A_T$ and $\phi : A \to \mathcal{Y}$, we define an expert $E_{A,\phi}$. The expert $E_{A,\phi}$ imitates the SOA algorithm when it errs exactly on the examples $\{x_t \mid t \in A\}$ and the true labels of these examples are determined by ϕ. Formally, the expert $E_{A,\phi}$ proceeds as follows:

Set $V_1 = \mathcal{H}$.
For $t = 1, 2 \ldots, T$
Receive x_t.
Set $l_t = \text{argmax}_{y\in\mathcal{Y}}\,\text{L-Dim}(\{f \in V_t : f(x_t) = y\})$.
If $t \in A$, Predict $\phi(t)$ and update $V_{t+1} = \{f \in V_t : f(x_t) = \phi(t)\}$.
If $t \notin A$, Predict l_t and update $V_{t+1} = \{f \in V_t : f(x_t) = l_t\}$.

The number of experts we constructed is $\sum_{j=0}^{\text{L-Dim}(\mathcal{H})} \binom{T}{j}(k-1)^j \leq (Tk)^{\text{L-Dim}(\mathcal{H})}$. Denote the number of mistakes made by the expert $E_{A,\phi}$ after T rounds by $L_{A,\phi,T}$. If we apply the LEA algorithm with the set of experts we have constructed, the resulting algorithm makes at most

$$\min_{A,\phi} L_{A,\phi,T} + \sqrt{\frac{1}{2}T\,\text{L-Dim}(\mathcal{H})\ln(Tk)}$$

mistakes. We claim that $\min_{A,\phi} L_{A,\phi,T} \leq \min_{f\in\mathcal{H}} L_{f,T}$: Let $f \in \mathcal{H}$. Denote by $A \subset [T]$ the set of rounds in which the SOA algorithm errs when running on the sequence $(x_1, f(x_1)), \ldots, (x_T, f(x_T))$ and define $\phi : A \to \mathcal{Y}$ by $\phi(t) = f(x_t)$. Since the SOA algorithm makes at most $\text{L-Dim}(\mathcal{H})$ mistakes, $|A| \leq \text{L-Dim}(\mathcal{H})$. It is not hard to see that the predictions of the expert $E_{A,\phi}$ coincide with the predictions of the expert E_f. Thus, $L_{A,\phi,T} = L_{f,T}$. ■

Adapting the proof of Lemma 14 from Ben-David et al. (2009), we conclude a corresponding lower bound:

Theorem 26 *In the agnostic online multiclass setting, the expected loss of every algorithm on the worst-case sequence is at least* $\min_{f\in\mathcal{H}} L_{f,T} + \sqrt{\frac{1}{8}\,\text{L-Dim}(\mathcal{H})T}$.

We leave as an open question to close the gap between the bounds in the above Theorems. Note that this gap is analogous to the sample complexity gap for ERM learners in the PAC setting, seen in Theorem 6.

5.2 The Bandit Setting

So far we have assumed that the label of each training example is fully revealed. In this section we deal with the bandit setting. In this setting, the learner does not get to see the correct label of a training example. Instead, the learner first receives an instance $x \in \mathcal{X}$, and should guess a label, $\hat{y}$. The learner then receives a binary response, which indicates only whether the guess was correct or not. If the guess is correct then the learner knows the identity of the correct label. If the guess is wrong, the learner only knows that $\hat{y}$ is not the correct label, and not the identity of the correct label.

5.2.1 Bandit vs. Full Information in the Batch Model

In this section we consider the bandit setting in the batch model. In this setting the sample is drawn i.i.d. as before, but the learner first observes only the instances $x_1, \ldots, x_m$. The learner then guesses a label for each of the instances, and receives a binary response indicating for each label whether it was the correct one.

Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and let $k = |\mathcal{Y}|$. Our goal is to analyze the *realizable bandit sample complexity* of $\mathcal{H}$, which we denote by $m_b^r(\epsilon, \delta)$, and the *agnostic bandit sample complexity* of $\mathcal{H}$, which we denote by $m_b^a(\epsilon, \delta)$. The following theorem provides upper bounds on the sample complexities.

Theorem 27 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class. Then,*

$$m_b^r(\epsilon,\delta) = O\left(k \cdot \frac{d_G(\mathcal{H}) \cdot \ln\left(\frac{1}{\epsilon}\right) + \ln(\frac{1}{\delta})}{\epsilon}\right) \text{ and } m_b^a(\epsilon,\delta) = O\left(k \cdot \frac{d_G(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon^2}\right) .$$

Proof Let $\mathcal{A}_f$ be a (full information) ERM learner for $\mathcal{H}$. Consider the following algorithm, denoted $\mathcal{A}_b$, for the bandit setting: Given a sample $(x_i, y_i)_{i=1}^m$, for each i the algorithm guesses a label $\hat{y}_i \in \mathcal{Y}$ drawn uniformly at random. Then the algorithm calls $\mathcal{A}_f$ with an input sample which consists only of the sample pairs for which the binary response indicated that the guess $\hat{y}_i$ was correct. Thus, the input sample is $\{(x_i, \hat{y}_i) \mid \hat{y}_i = y_i\}$. $\mathcal{A}_b$ then returns whatever hypothesis $\mathcal{A}_f$ returned.

We show that $m^r_{\mathcal{A}_b}(\epsilon,\delta) \leq 3k \cdot m^r_{\mathcal{A}_f}(\epsilon, \frac{\delta}{2}) + \frac{3}{2}\log\left(\frac{2}{\delta}\right) =: m'$ and similarly for the agnostic case, so that the theorem is implied by the bounds in the full information setting (Theorem 5). Indeed, suppose that m examples suffice for $\mathcal{A}_f$ to return a hypothesis with excess error at most ϵ, with probability at least $1 - \frac{\delta}{2}$. Let $(x_i, y_i)_{i=1}^{m'}$ be a sample for the bandit algorithm. By Chernoff's bound, with probability at least $1 - \frac{\delta}{2}$, $\mathcal{A}_b$ guesses correctly the label of at least m examples. Therefore $\mathcal{A}_f$ runs on a sample of at least this size. The sample that $\mathcal{A}_f$ receives is a conditionally i.i.d. sample, given the size of the sample, with

the same conditional distribution as the one the original sample was sampled from. Thus, with probability at least $1 - \frac{\delta}{2}$, $\mathcal{A}_f$ (and, consequently, $\mathcal{A}_b$) returns a hypothesis with excess error at most ϵ. ■

An interesting quantity to consider is the price of bandit information in the batch model: Let $\mathcal{H}$ be a hypotheses class, and define $\mathrm{PBI}_{\mathcal{H}}(\epsilon, \delta) = m^r_{b,\mathcal{H}}(\epsilon, \delta)/m^r_{\mathrm{PAC},\mathcal{H}}(\epsilon, \delta)$. By Theorems 27 and 6 and Equation 2 we see that, $\mathrm{PBI}(\epsilon, \delta) = O(\ln(\frac{1}{\epsilon})k\ln(k))$. This is essentially tight since it is not hard to see that if both $\mathcal{X}, \mathcal{Y}$ are finite and we let $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$, then $\mathrm{PBI}_{\mathcal{H}} = \Omega(k)$.

Using Theorems 27 and 5 and Equation 2 we can further conclude that, as in the full information case, the finiteness of the Natarajan dimension is necessary and sufficient for learnability in the bandit setting as well. However, the ratio between the upper bound due to Theorem 27 and the lower bound, due to Theorem 5, is $\Omega(\ln(k) \cdot k)$. It would be interesting to find a more tight characterization of the sample complexity in the bandit setting. This characterization cannot depend solely on the Natarajan dimension, or other quantities which are strongly related to it (such as the graph dimension or other notion of dimension defined in Ben-David et al. (1995)): For example, the classes $[k]^{[d]}$ and $[2]^{[d]}$ have the same Natarajan dimension, but their bandit sample complexity differs by a factor of $\Omega(k)$.

5.2.2 Bandit vs. Full Information in the Online Model

We now consider Bandits in the online learning model. We focus on the realizable case, in which the feedback provided to the learner is consistent with some function $f_0 \in \mathcal{H}$. We define a new notion of dimension of a class, that determines the sample complexity in this setting.

As in Section 5.1, consider a rooted tree T whose internal nodes are labeled by elements from $\mathcal{X}$ and whose edges are labeled by elements from $\mathcal{Y}$, such that the edges from a single parent to its child-nodes are each labeled with a different label. The tree T is *BL-shattered* by $\mathcal{H}$ if, for every path from root to leaf $x_1, \ldots, x_k$, there is a function $f \in \mathcal{H}$ such that for every i, $f(x_i)$ is *different* from the label of (x_i, x_{i+1}). The **Bandit-Littlestone dimension** of $\mathcal{H}$, denoted BL-dim($\mathcal{H}$), is the maximal depth of a complete k-ary tree that is BL-shattered by $\mathcal{H}$.

Theorem 28 *Let $\mathcal{H}$ be a hypothesis class with $L = \mathrm{BL\text{-}Dim}(\mathcal{H})$. Then every deterministic online bandit learning algorithm for $\mathcal{H}$ will make at least L mistakes in the worst case. Moreover, there is an online learning algorithm that makes at most L mistakes on every realizable sequence.*

Proof First, let T be a BL-shattered tree of depth L. We show that for every deterministic learning algorithm there is a sequence $x_1, \ldots, x_L$ and a labeling function $f_0 \in \mathcal{H}$ such that the algorithm makes L mistakes on this sequence. The sequence consists of the instances attached to nodes of T, when traversing the tree from the root to one of its leaves, such that the label of each edge (x_i, x_{i+1}) is equal to the algorithm's prediction $\hat{y}_i$. The labeling function $f_0 \in \mathcal{H}$ is one such that for all i, $f_0(x_i)$ is different from the label of edge (x_i, x_{i+1}). Such a function exists since T is BL-shattered, and the algorithm will clearly make L mistakes on this sequence.

Second, the following online learning algorithm makes at most L mistakes on any realizable input sequence.

Algorithm: Bandit Standard Optimal Algorithm (BSOA)
Initialization: $V_0 = \mathcal{H}$.
For $t = 1, 2 \ldots$,
Receive x_t
For $y \in \mathcal{Y}$, let $V_t^{(y)} = \{f \in V_{t-1} : f(x_t) \neq y\}$
Predict $\hat{y}_t \in \arg\min_y \text{BL-Dim}(V_t^{(y)})$
Receive an indication whether $\hat{y}_t = f(x_t)$
If the prediction is wrong, update $V_t = V_t^{(\hat{y}_t)}$.

To see that BSOA makes at most L mistakes, note that at each time t, there is at least one $V_t^{(y)}$ with $\text{BL-Dim}(V_t^{(y)}) < \text{BL-Dim}(V_{t-1})$. This can be seen by assuming to the contrary that this is not so, and concluding that if $\text{BL-Dim}(V_t^{(y)}) = \text{BL-Dim}(V_{t-1})$ for all $y \in [k]$, then one can construct a shattered tree of size $\text{BL-Dim}(V_{t-1}) + 1$ for V_{t-1}, thus reaching a contradiction.

Thus, whenever the algorithm errs, the dimension of V_t decreases by one. Thus, after L mistakes, the dimension is 0, which means that there is a single function that is consistent with the sample, so no more mistakes can occur. ■

The price of bandit information: Let $\text{PBI}(\mathcal{H}) = \text{BL-Dim}(\mathcal{H})/\,\text{L-Dim}(\mathcal{H})$ and fix $k \geq 2$. How large can $\text{PBI}(\mathcal{H})$ be when $\mathcal{H}$ is a class of functions from a domain $\mathcal{X}$ to a range $\mathcal{Y}$ of cardinality k? We refer the reader to Daniely and Helbertal (2013), where it is shown that $\text{PBI}(\mathcal{H}) \leq 4k\log(k)$. This bound is tight up to the logarithmic factor.

6. Discussion

We have shown in this work that even in the simple case of multiclass learning, different ERM learners for the same problem can have large gaps in their sample complexities. To put our results in a more general perspective, consider the *General Setting of Learning* introduced by Vapnik (1998). In this setting, a *learning problem* is a triplet $(\mathcal{H}, \mathcal{Z}, l)$, where $\mathcal{H}$ is a hypothesis class, $\mathcal{Z}$ is a data domain, and $l : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ is a loss function. We emphasize that $\mathcal{H}$ is not necessarily a class of functions but rather an abstract set of models. The goal of the learner is, given a sample $S \in \mathcal{Z}^m$, sampled from some (unknown) distribution $\mathcal{D}$ over $\mathcal{Z}$, to find a hypothesis $h \in \mathcal{H}$ that minimizes the *expected loss*, $l(h) = \mathbb{E}_{z\sim\mathcal{D}}[l(h, z)]$.

The general setting of learning encompasses multiclass learning as follows: given a hypotheses class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$, take $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and define $l : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ by $l(h, (x, y)) = 1[h(x) \neq y]$. However, the general learning setting encompasses many other problems as well, for instance:

- **Regression with the squared loss:** Here, $\mathcal{Z} = \mathbb{R}^n \times \mathbb{R}$, $\mathcal{H}$ is a set of real-valued functions over $\mathbb{R}^n$ and $l(h, (x, y)) = (h(x) - y)^2$.

- **k-means:** Here, $\mathcal{Z} = \mathbb{R}^n$, $\mathcal{H} = (\mathbb{R}^n)^k$ and, for $h = (c_1, \ldots, c_k) \in \mathcal{H}$ and $x \in \mathcal{Z}$, the loss is $l((c_1, \ldots, c_k), x) = \min_{j \in [k]} ||c_j - x||^2$.
- **Density estimation:** Here, $\mathcal{Z}$ is an arbitrary finite set, $\mathcal{H}$ is some set of probability density functions over $\mathcal{Z}$, and the loss function is the log loss, $l(p, x) = -\ln(p(x))$.

A learning problem is *learnable* in the general setting of learning if there exists a function $\mathcal{A} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$ such that for every $\epsilon > 0$ and $\delta > 0$ there exists an m such that for every distribution $\mathcal{D}$ over $\mathcal{Z}$,

$$\Pr_{S \sim \mathcal{Z}^m} \left(l(\mathcal{A}(S)) \geq \inf_{h \in \mathcal{H}} l(h) + \epsilon \right) < \delta$$

A learning problem *converges uniformly* if, for every $\epsilon > 0$,

$$\lim_{m \to \infty} \Pr_{S \sim \mathcal{Z}^m} \left(\sup_{h \in \mathcal{H}} |l(h) - l_S(h)| > \epsilon \right) = 0$$

where for $S = (z_1, \ldots, z_m) \in \mathcal{Z}^m$, $l_S(h) = \frac{1}{m} \sum_{i=1}^{m} l(h, z_i)$ is the empirical loss of h on the sample S. An easy observation is that uniform convergence implies learnability, and a classical result is that for binary classification and for regression (with absolute or squared loss), the inverse implication also holds. Thus, it was believed that excluding some trivialities, learnability is equivalent to uniform convergence. In Shalev-Shwartz et al. (2010) it is shown that for stochastic convex optimization, learnability does not imply uniform convergence, giving an evidence that the above belief might be misleading. Our results in this work can be seen as another step in this direction, as we have shown that even in multiclass classification – a simple, natural and popular generalization of binary classification, the above mentioned equivalence no longer holds.

We conclude with an open question. In view of our results in Section 4, the following conjecture suggests itself.

Conjecture 29 *There exists a constant C such that, for* every *hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$,*

$$m^r_{\text{PAC}}(\epsilon, \delta) \leq C \left(\frac{d_N(\mathcal{H}) \ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon} \right)$$

In light of Theorem 9 and the fact that there are cases where $d_G \geq \log_2(k-1) d_N$, the conjecture can only be proved if this learning rate can be achieved by a learning algorithm that is not just an *arbitrary* ERM learner. So far, all the general upper bounds that we are aware of are valid for *any* ERM learner. Understanding how to select among ERM learners is fundamental as it teaches us what is the optimal way to learn. We hope that our examples from section 4 and our result for symmetric classes will lead to a better understanding of the optimal learning method.

Remark 30 *A subsequent paper (Daniely and Shalev-Shwartz, 2014) established several results that are highly related to the subject of this paper. First, they have shown that the ERM rule is suboptimal even for multiclass classification with linear classes. Second, they have shown that for some classes, an optimal learner must be improper – that is, it must have the ability to return a hypothesis that does not belong to the learnt class. Finally, they have show that the one-inclusion algorithm (Rubinstein et al., 2006) is optimal for multiclass classification. We note that Conjecture 29 is still open.*

Acknowledgments

We wish to thank Ohad Shamir for valuable comments. Amit Daniely is a recipient of the Google Europe Fellowship in Learning Theory, and this research is supported in part by this Google Fellowship. Sivan Sabato is partly supported by the ISRAEL SCIENCE FOUNDATION (grant No. 555/15).

References

E. L. Allwein, R.E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

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

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SICOMP: SIAM Journal on Computing*, 32, 2003.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

PL Bartlett, PM Long, and RC Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996.

S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. Long. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. *Journal of Computer and System Sciences*, 50: 74–86, 1995.

S. Ben-David, D. Pal, , and S. Shalev-Shwartz. Agnostic online learning. In *COLT*, 2009.

A. Beygelzimer, J. Langford, and P. Ravikumar. Multiclass classification with filter trees. *Preprint, June*, 2007.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

A. Daniely and S. Shalev-Shwartz. Optimal learners to multiclass problems. In *COLT*, 2014.

Amit Daniely and Tom Helbertal. The price of bandit information in multiclass online classification. In *Conference on Learning Theory*, pages 93–104, 2013.

T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, January 1995.

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

S.M. Kakade, S. Shalev-Shwartz, and A. Tewari. Efficient bandit algorithms for online multiclass prediction. In *International Conference on Machine Learning*, 2008.

Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

N. Littlestone. Learning when irrelevant attributes abound. In *FOCS*, pages 68–77, October 1987.

B. K. Natarajan. On learning sets and functions. *Mach. Learn.*, 4:67–97, 1989.

Benjamin I Rubinstein, Peter L Bartlett, and J Hyam Rubinstein. Shifting, one-inclusion mistake bounds and tight multiclass expected risk bounds. In *Advances in Neural Information Processing Systems*, pages 1193–1200, 2006.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 9999:2635–2670, 2010.

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

V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

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

V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.