Multiclass Learnability and the ERM principle

Amit Daniely AMIT.DANIELY@MAIL.HUJI.AC.IL
Dept. of Mathematics, The Hebrew University, Jerusalem, Israel
Sivan Sabato SIVAN_SABATO@CS.HUJI.AC.IL
School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel
Shai Ben-David SHAI@CS.UWATERLOO.CA
David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada
Shai Shalev-Shwartz SIVANS@CS.HUJI.AC.IL
School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel

Editor: Sham Kakade, Ulrike von Luxburg

Abstract

Multiclass learning is an area of growing practical relevance, for which the currently available theory is still far from providing satisfactory understanding. We study the learnability of multiclass prediction, and derive upper and lower bounds on the sample complexity of multiclass hypothesis classes in different learning models: batch/online, realizable/unrealizable, full information/bandit feedback. Our analysis reveals a surprising phenomenon: In the multiclass setting, in sharp contrast to binary classification, not all Empirical Risk Minimization (ERM) algorithms are equally successful. We show that there exist hypotheses classes for which some ERM learners have lower sample complexity than others. Furthermore, there are classes that are learnable by some ERM learners, while other ERM learner 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 (that is, classes that are invariant under any permutation of label names). We demonstrate the relevance of the theory by analyzing the sample complexity of two widely used hypothesis classes: generalized linear multiclass models and reduction trees. We also obtain some practically relevant conclusions.

Keywords: List of keywords

1. Introduction

The task of multiclass learning, that is learning to classify an object into one of many candidate classes, surfaces in many domains including 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 the task. Many of the methods define a set of possible multiclass predictors, $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ (where $\mathcal{X}$ is the data domain and $\mathcal{Y}$ is the set of labels), called the hypothesis class, and then use the training examples to choose a predictor from $\mathcal{H}$ (for instance Crammer and Singer, 2003). In this paper we study the sample complexity of such hypothesis classes, namely, how many training examples are needed for learning an accurate predictor. This question has been extensively studied and is quite well understood for the binary case, where $|\mathcal{Y}| = 2$. In contrast, the existing theory of the multiclass case, where $|\mathcal{Y}| > 2$, is much less complete.

We study multiclass sample complexity in several learning models. These models vary in three aspects:

- 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 a predictor h with a small probability to err, $\Pr_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$, with a high probabilty over training samples. In the online protocol we receive examples one by one and are asked to predict the labels on the fly. Our goal is to make as few prediction mistakes as possible in the worst case (see Littlestone (1987)).
- The underlying labeling mechanism (realizable vs. agnostic): In the realizable case, we assume that the labels of the instances are determined by some $h^\star \in \mathcal{H}$. In the agnostic case no restrictions on the labeling rule are imposed, and our goal is to make predictions which are not much worse than the best predictor in $\mathcal{H}$.
- The type of feedback (full information vs. bandits): In the full information setting, each example is revealed to the learner along with its correct label. In the bandit setting, the learner first sees an unlabeled example, and then outputs its guess for the label. Then a binary feedback is received, indicating only whether the guess 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)).

In Section 2 we consider multiclass sample complexity in the PAC model (namely, the batch protocol with full information). Natarajan (1989) provides a characterization of multiclass PAC learnability in terms of a parameter of $\mathcal{H}$ known as the Natarajan dimension and denoted $d_N(\mathcal{H})$ (see section 2.2 for the relevant definitions). For the realizable case we show in Section 2.3 that there are constants C_1, C_2 such that the sample complexity of learning $\mathcal{H}$ with error ϵ and confidence $1 - \delta$ satisfies

$$C_1 \left(\frac{d + \ln(\frac{1}{\delta})}{\epsilon} \right) \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \left(\frac{d \left(\ln(\frac{1}{\epsilon}) + \ln(|\mathcal{Y}|) + \ln(d) \right) + \ln(\frac{1}{\delta})}{\epsilon} \right), \tag{1}$$

where $d = d_N(\mathcal{H})$. This improves the best previously known upper bound (theorem 5), in which there is a dependence on $\ln(|\mathcal{Y}|) \cdot \ln(\frac{1}{\epsilon})$.

The Natarajan dimension is equal to the VC dimension when $|\mathcal{Y}| = 2$. However, for larger label sets $\mathcal{Y}$, the bound on the sample complexity is not as tight as the known bound for the binary case, where the gap between the lower and upper bounds is only logarithmic in $1/\epsilon$. This invokes the challenge of tightening these sample complexity bounds for the multiclass case. A common approach to proving sample complexity bounds for PAC learning is to carefully analyze the sample complexity of ERM learners. In the case of PAC learning, all ERM learners have the same sample complexity (up to a logarithmic factor, see (Vapnik, 1995)). However, rather surprisingly, this is not the case for multiclass learning[1].

[1] Note that Shalev-Shwartz et al. (2010) established gaps between ERM learners in the general learning setting. However, here we consider multiclass learning, which seems very similar to binary classification.

In Section 2.4 we describe a family of concept classes for which there exist “good” ERM learner and “bad” ERM learner with a large gap between their sample complexities. Analyzing these examples, we deduce a rough principle on how to choose a good ERM learner. We also determine the sample complexity of the worst ERM learner for a given concept class, $\mathcal{H}$, up to a multiplicative factor of $O(\ln(\frac{1}{\epsilon}))$. We further show that if $|\mathcal{Y}|$ is infinite, then there are hypotheses classes that are learnable by *some* ERM learners but not by *all* ERM learners. In Section 2.5 we employ the suggested principle to derive an improved sample complexity upper bound for *symmetric* classes ($\mathcal{H}$ is symmetric if $\phi \circ f \in \mathcal{H}$ whenever $f \in \mathcal{H}$ and ϕ is a permutation of $\mathcal{Y}$). Symmetric classes are useful, since they are a natural choice when there is no prior knowledge about the relations between the possible labels. Moreover, many popular hypothesis classes that are used in practice are symmetric.

We conjecture that the upper bound obtained for symmetric classes holds for the sample complexity of non-symmetric classes as well. Such a result cannot be implied by uniform convergence alone, since, by the results mentioned above, there always exist bad ERM learners whose sample complexity is higher than this conjectured upper bound. It therefore seems that a proof for 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 3 we study multiclass learnability in the online model. We describe a simple generalization of the Littlestone dimension, and derive tight lower and upper bounds on the number, in terms of that dimension, of mistakes the optimal online algorithm will make in the worst case. Section 4 is devoted to a discussion of sample complexity of multiclass learning in the Bandit settings. Finally, in Section 5 we calculate the sample complexity of some popular families of hypothesis classes, which include linear multiclass hypotheses and filter trees, and discuss some practical implications of our bounds.

2. Multiclass Learning in the PAC Model

2.1. Problem Setting and Notation

For a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the error of a function $f \in \mathcal{H}$ with respect to $\mathcal{D}$ is $\mathrm{Err}(f) = \mathrm{Err}_{\mathcal{D}}(f) = \Pr_{(x,y)\sim\mathcal{D}}(f(x) \neq y)$. A *learning algorithm* for a class $\mathcal{H}$ is a function, $A : \cup_{n=0}^{\infty}(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$. 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 $f \in \mathcal{H}$ that minimizes the number of sample errors $|\{i \in [m] : f(x_i) \neq y_i\}|$. This work focuses on statistical properties of the learning algorithms and ignores computatational complexity aspects.

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

$$\Pr_{S_m \sim \mathcal{D}^m}\left(\mathop{\mathrm{Err}}_{\mathcal{D}}(A(S_m)) > \inf_{f \in \mathcal{H}} \mathop{\mathrm{Err}}_{\mathcal{D}}(f) + \epsilon\right) \leq \delta. \qquad (2)$$

If there is no integer satisfying these requirements, define $m^a_A(\epsilon, \delta) = \infty$. The *(agnostic) sample complexity* of a class $\mathcal{H}$ is

$$m^a_{\mathcal{H}}(\epsilon, \delta) = \inf_A m^a_A(\epsilon, \delta) \ ,$$

where the infimum is taken over all learning algorithms.

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* A for a class $\mathcal{H}$, denoted m^r_A, is the minimal integer such that for every $m \geq m^r_A(\epsilon, \delta)$ and every distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ which is realizable by $\mathcal{H}$, Equation. (2) holds. The realizable sample complexity of a class $\mathcal{H}$ is $m^r_{\mathcal{H}}(\epsilon, \delta) = \inf_A m^r_A(\epsilon, \delta)$ where the infimum is taken over all learning algorithms.

2.2. Known Sample Complexity Results

We first survey some known results regarding the sample complexity of multiclass learning. We start with the realizable case and then discuss the agnostic case. Given a subset $S \subseteq \mathcal{X}$, we denote $\mathcal{H}|_S = \{f|_S : f \in \mathcal{H}\}$. 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* VC($\mathcal{H}$)*, is the maximal cardinality of a subset $S \subseteq \mathcal{X}$ that is shattered by $\mathcal{H}$.*

The VC-dimension is cornerstone in statistical learning theory as it characterizes the sample complexity of a *binary* hypothesis class. Namely

Theorem 2 (Vapnik, 1995) *There are absolute constants $C_1, C_2 > 0$ such that the realizable sample complexity of every hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ satisfies*

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

Moreover, the upper bound is attained by any ERM learner.

It is natural to seek a generalization of the VC-Dimension to hypothesis classes of non-binary functions. A straightforward attempt is to redefine shattering of $S \subset \mathcal{X}$ by the property $\mathcal{H}|_S = \mathcal{Y}^S$. However, this requirement is too strong and does not lead to tight bounds on the sample complexity. Instead, we recall two alternative generalizations, introduced by Natarajan (1989). In both definitions, shattering is redefined to require 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 differ in how "different behavior" is defined.

Definition 3 (Graph dimension and Natarajan dimension) *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and let $S \subseteq \mathcal{X}$. We say that $\mathcal{H}$* 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*

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

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*

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

The graph dimension *of $\mathcal{H}$, denoted $d_G(\mathcal{H})$, is the maximal cardinality of a set that is G-shattered by $\mathcal{H}$. 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 $|\mathcal{Y}| = 2$. Note also that we always have $d_N \leq d_G$.

By reductions from and to the binary case, it is not hard to show, similarly to Natarajan (1989) and Ben-David et al. (1995) (see Appendix A for a full proof), that

Theorem 4 *For the constants C_1, C_2 from theorem 2, for every $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ we have*

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

Moreover, the upper bound is attained by any ERM learner.

From this theorem it follows 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 concept class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$,

$$d_N(\mathcal{H}) \leq d_G(\mathcal{H}) \leq 4.67 \log_2(|\mathcal{Y}|) d_N(\mathcal{H}) \ . \tag{3}$$

It follows that if $|\mathcal{Y}| < \infty$ then the finiteness of the Natarajan dimension is a necessary and sufficient condition for learnability. Incorporating Equation. (3) into theorem 4, 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 $O(\log(|\mathcal{Y}|) \log(\frac{1}{\epsilon}))$. Precisely,

Theorem 5 *(Ben-David et al., 1995) For the constants C_1, C_2 from theorem 2,*

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

Moreover, the upper bound is attained by any ERM learner.

A similar analysis can be performed for the agnostic case. For binary classification we have that for every hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$,

$$m^a_{\mathcal{H}}(\epsilon, \delta) = \Theta\left(\frac{1}{\epsilon^2} \left(VC(\mathcal{H}) + \ln(\frac{1}{\delta}) \right) \right) , \tag{4}$$

and this is attained by any ERM learner. Here too it is possible to obtain by reduction from and to the binary case that for every hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$,

$$\Omega\left(\frac{1}{\epsilon^2} \left(d_N(\mathcal{H}) + \ln(\frac{1}{\delta}) \right) \right) \leq m^a_{\mathcal{H}}(\epsilon, \delta) \leq O\left(\frac{1}{\epsilon^2} \left(d_G(\mathcal{H}) + \ln(\frac{1}{\delta}) \right) \right) . \tag{5}$$

By Equation. (3) we have

$$m^a_{\mathcal{H}}(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2} \left(\log(|\mathcal{Y}|) \cdot d_N(\mathcal{H}) + \ln(\frac{1}{\delta}) \right) \right) . \tag{6}$$

Thus in the agnostic case as well, the Natarajan dimension characterizes the agnostic sample complexity up to a multiplicative factor of $O(\log(|\mathcal{Y}|))$. Here too, all of these bounds are attained by any ERM learner.

2.3. An Improved Result for the Realizable Case

The following theorem provides a sample complexity upper bound which can be better than Theorem 5 when $\ln(d_N(\mathcal{H})) \ll \ln(|\mathcal{Y}|) \cdot \ln(\frac{1}{\epsilon})$. The proof of the theorem is given in Appendix A. While the proof is a simple adaptation of previous results, we find it valuable to present this result here, as we could not find it in the research literature.

Theorem 6 *For every concept class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$,*

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

Moreover, the bound is attained by any ERM learner.

Theorem 6 is the departure point of our research. As indicated above, one of our objectives is to prove sample complexity bounds for the multiclass case with a ratio of $O(\ln(\frac{1}{\epsilon}))$ between the upper bound and the lower bound, as in the binary case. In the next section we show that such an improvement cannot be attained by uniform convergence analysis, since the ratio between the sample complexity of the worst ERM learner and the best ERM learner of a given hypothesis class might be as large as $\ln(|\mathcal{Y}|)$.

2.4. The Gap between "Good ERM" and "Bad ERM"

The tight bounds in the binary case given in Theorem 2 are attained by *any* ERM learner. In contrast to the binary case, we now show that in the multiclass case there can be a significant sample complexity gap between different ERM learners. Moreover, in the case of classification with an infinite number of classes, there are learnable hypothesis classes that some ERM learners fail to learn. We begin with showing that the graph dimension determines the sample complexity of the worst ERM learner up to a multiplicative factor of $O(\ln(\frac{1}{\epsilon}))$.

Theorem 7 *There are absolute constants $C_1, C_2 > 0$ such that for every hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and every ERM learner A,*

$$m^r_A(\epsilon, \delta) \le C_2\left(\frac{d_G(\mathcal{H})\ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon}\right).$$

Moreover, there is an ERM learner A_{bad} such that

$$m^r_{A_{\text{bad}}}(\epsilon, \delta) \ge C_1\left(\frac{d_G(\mathcal{H}) + \ln(\frac{1}{\delta})}{\epsilon}\right). \tag{7}$$

Proof The upper bound on m^r_A is just a restatement of theorem 4. It remains to prove that there exists an ERM learner, A_{bad}, satisfying (7). We shall first consider the case where $d = d_G(\mathcal{H}) < \infty$.

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 A_{bad} be an ERM learner with the property that upon seeing a sample whose instances are in $T \subseteq S$, and whose labels are determined by f_0, it returns $f \in \mathcal{H}$ such that f equals to f_0 on T and f is different from f_0 on $S \setminus T$. The existence of such an f follows form the assumption that S is G-shattered using f_0.

Fix $\delta < e^{-1/6}$ and let ϵ small enough such 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 the sample size be m. Clearly, the hypothesis returned by A_{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 $\geq e^{-\frac{1}{6}} \geq \delta$, the sample will include no more than $\frac{d-1}{2}$ examples from S. Thus the returned hypothesis will have error $\geq \epsilon$. Moreover, the probability that the sample includes only x_0 (and thus A_{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

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

as required. If $d_G(\mathcal{H}) = \infty$ then the argument above can be repeated for a sequence of pairwise disjoint G-shattered sets S_n, $n = 1, 2, \ldots$ with $|S_n| = n$. ■

The following example shows that in some cases there are learning algorithms that are much better than the worst ERM:

Example 1 *(A Large Gap Between ERM Learners) Let $\mathcal{X}_0$ be any finite or countable domain set and let $\mathcal{X}$ be some subset of $\mathcal{X}_0$. Let $\mathcal{P}_f(\mathcal{X})$ denote the collection of finite and co-finite subsets $A \subseteq \mathcal{X}$. For every $A \in \mathcal{P}_f(\mathcal{X})$, define $f_A : \mathcal{X}_0 \to \mathcal{P}_f(\mathcal{X}) \cup \{*\}$ by*

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

and consider the concept family $\mathcal{H}_{\mathcal{X}} = \{f_A : A \in \mathcal{P}_f(\mathcal{X})\}$. We first note that any ERM learner that sees an example of the form (x, A) for some $A \subseteq \mathcal{X}$ must return the hypothesis f_A, thus to define an ERM learner we only have to specify the hypothesis it returns upon seeing a sample of the form $S_m = \{(x_1, *), \ldots, (x_m, *)\}$. Note also that $\mathcal{X}$ is G-shattered using the function $f_\emptyset$, and therefore $d_G(\mathcal{H}_{\mathcal{X}}) \geq |\mathcal{X}|$ (it is easy to see that, in fact $d_G(\mathcal{H}_{\mathcal{X}}) = |\mathcal{X}|$).

We consider two ERM learners – A_{good}, which on a sample of the form S_m returns the hypothesis $f_\emptyset$, and A_{bad}, which, upon seeing S_m, returns $f_{\{x_1,\ldots,x_m\}^c}$, thus satisfying the specification of a bad ERM algorithm from the proof of Theorem 7. It follows that the sample complexity of A_{bad} is $\Omega\left(\frac{|\mathcal{X}|}{\epsilon} + \frac{1}{\epsilon}\ln(\frac{1}{\delta})\right)$. On the other hand,

Claim 1 *The sample complexity of A_{good} is at most $\frac{1}{\epsilon}\ln\frac{1}{\delta}$.*

Proof *Let $\mathcal{D}$ be a distribution over $\mathcal{X}_0$ and suppose that the correct labeling is f_A. Let m be the size of the sample. For any sample, A_{good} returns either $f_\emptyset$ or f_A. If it returns f_A then its generalization error is zero. Thus, it returns a hypothesis with error $\geq \epsilon$ only if $\Pr_{\mathcal{D}}(A) \geq \epsilon$ and all the m examples in the sample are from A^c. Assume $m \geq \frac{1}{\epsilon}\ln(\frac{1}{\delta})$, then probability of the latter event is no more than $(1-\epsilon)^m \leq e^{-\epsilon m} \leq \delta$.* ■

Since $\mathcal{X}$ can be infinite in the above example we conclude that

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.*

What is the crucial feature that makes A_{good} better than A_{bad}? If the correct labeling is $f_A \in \mathcal{H}_{\mathcal{X}}$, then for *any* sample, A_{good} might return at most one of two functions – namely f_A or $f_\emptyset$. On the other hand, if the sample is labeled by the function $f_\emptyset$, A_{bad} might return *every* function in $\mathcal{H}_{\mathcal{X}}$. Thus, to return a hypothesis with error $\le \epsilon$, A_{good} needs to reject only one hypothesis while A_{bad} needs to reject many more. We conclude the following (rough) principle: *A good ERM is an ERM that, for every target hypothesis, consider a small number of hypotheses.*

Next, we formalize the above intuition by proving a general theorem that enables us to derive sample complexity bounds for ERM learners that are designed using the above principle. Fix a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. We view an ERM learner as an operator that for any $f \in \mathcal{H}, S \subseteq \mathcal{X}$ takes the partial function $f|_S$ as input and extends it to a function $g = A(f|_S) \in \mathcal{H}$ such that $g|_S = f|_S$. For every $f \in \mathcal{H}$, denote by $F_A(f)$ the set of all the functions that the algorithm A might return upon seeing a sample of the form $\{(x_i, f(x_i))\}_{i=1}^m$ for some $m \ge 0$. Namely,

$$F_A(f) = \{A(f|_S) : S \subseteq \mathcal{X},\ |S| < \infty\}$$

To provide an upper bound on $m^r_A(\epsilon,\delta)$, it suffices to show that for every $f \in \mathcal{H}$, with probability at least $1-\delta$, all the functions with error at least ϵ in $F_A(f)$ will be rejected after seeing m examples. This is formalized in the following theorem.

Theorem 9 *Let A be an ERM learner for a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. Define the* growth function *of A by $\Pi_A(m) = \sup_{f\in\mathcal{H}} \Pi_{F_A(f)}(m)$, where for $F \subseteq \mathcal{Y}^{\mathcal{X}}$, $\Pi_F(m) = \sup\{|F|_S| : S \subseteq \mathcal{X},\ |S| \le m\}$. Then*

$$m^r_A(\epsilon,\delta) \le \min\{m \,:\, \Pi_A(2m)\, 2^{1-\frac{\epsilon m}{2}} < \delta\}\ .$$

The theorem immediately follows from the following lemma.

Lemma 10 (The Double Sampling Lemma) *Let A be an ERM learner. Fix a distribution $\mathcal{D}$ over $\mathcal{X}$ and a function $f_0 \in \mathcal{H}$. Denote by A_m the event that, after seeing m i.i.d. examples drawn from $\mathcal{D}$ and labeled by f_0, A returns a hypothesis with error at least ϵ. Then $\Pr(A_m) \le 2 \cdot \Pi_A(2m) 2^{-\frac{\epsilon m}{2}}$.*

Proof Let S_1 and S_2 be two samples of m i.i.d. examples labeled by f_0. Let B_m be the event that there exists a function $f \in \mathcal{H}$ with error at least ϵ, such that (1) f is not rejected by S_1 (i.e. $f_0(x) = f(x)$ for all examples x in S_1), and (2) there exist at least $\frac{\epsilon m}{2}$ examples $(x, f_0(x))$ in S_2 for which $f(x) \ne f_0(x)$. By Chernoff's bound, for $m = \Omega(\frac{1}{\epsilon})$, $\Pr(B_m) = \Pr(B_m|A_m)\Pr(A_m) \ge \frac{1}{2}\Pr(A_m)$. W.l.o.g., we can assume that S_1, S_2 are generated as follows: First, $2m$ examples are drawn to create a sample U. Then S_1 and S_2 are generated by selecting a random partition of U into two samples of size m. Now, $\Pr(B_m)$ is bounded by the probability that there is an $f \in \mathcal{H}|_U$ such that (1) there are at least $\frac{\epsilon m}{2}$ examples in U such that f disagrees with f_0 on these examples and (2) all of these examples are located in S_2. For a single $f \in \mathcal{H}|_U$ that disagrees with f_0 on $l \geq \frac{\epsilon m}{2}$ samples, the probability that all these examples are located in S_2 is $\binom{m}{l}/\binom{2m}{l} \leq 2^{-l} \leq 2^{-\frac{\epsilon m}{2}}$. Thus, using the union bound we obtain that $\Pr(B_m) \leq |\mathcal{H}|_U|\, 2^{-\frac{\epsilon m}{2}} \leq \Pi(2m) 2^{-\frac{\epsilon m}{2}}$. ■

The bound in theorem 6 is based on the (trivial) inequality $\Pi_A \leq \Pi_{\mathcal{H}}$. However, as Example 1 shows, Π_A can be much smaller than $\Pi_{\mathcal{H}}$. As we shall see in the sequel, we can apply the double sampling lemma to get better sample complexity bounds for "good" ERM learners. The key tool for these sample complexity bounds is Lemma 12, that is, in turn, based on the following combinatorial result:

Lemma 11 *(Natarajan, 1989) For every hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, $|\mathcal{H}| \leq |\mathcal{X}|^{d_N(\mathcal{H})} |\mathcal{Y}|^{2d_N(\mathcal{H})}$.*

Lemma 12 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a class of functions. Assume that for some number r, for every $h \in \mathcal{H}$, the size of the range of h is at most r. Let A be an algorithm such that, for some set of values $Y' \subseteq \mathcal{Y}$, for every $f \in \mathcal{H}$, and every sample $S_m = ((x_1, f(x_1)), \ldots (x_m, f(x_m)))$, the function returned by A on input S_m is consistent with S_m and has its values in the set $\{f(x_1), \ldots, f(x_m)\} \cup Y'$. Then,*

$$m_A^r(\epsilon, \delta) = O\left(\frac{d_N(\mathcal{H})(\ln(\frac{1}{\epsilon}) + \ln(\max\{r, |Y'|\})) + \ln(\frac{1}{\delta})}{\epsilon}\right) .$$

Proof The assumptions of the lemma imply that, for every $f \in \mathcal{H}$, the range of the functions in $F_A(f)$ is contained in the union of Y' and the range of f. Therefore, using Lemma 11 we obtain that $\Pi_A(2m) \leq (2m)^{d_N(\mathcal{H})}(|Y'|+r)^{2d_N(\mathcal{H})}$, and the bound follows from Theorem 9. ■

Note that classes in which each function $h \in H$ uses at most r values, for some $r < d_N(H)\log(|\mathcal{Y}|)$, can have a large range $\mathcal{Y}$ and a graph dimension that is significantly larger than their Natarajan dimension. In such cases, we may be able to show a gap between the sample complexity of bad and good ERM learners, by applying the lower bound from Theorem 7. In particular, we get such a result for the following family of hypotheses classes, which generalizes Example 1.

Corollary 13 *Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to some range set $\mathcal{Y}$, such that, for some value $y_0 \in \mathcal{Y}$, for every $h \in H$, the range of h contains at most one value besides y_0. Assume also that $\mathcal{H}$ contains the constant y_0 function. Let d denote the Natarajan dimension of $\mathcal{H}$. Then there exists an ERM learning algorithm A for H such that the (ϵ, δ) sample complexity of A is*

$$O\left(\frac{d \cdot \ln(1/\epsilon) + \ln(1/\delta)}{\epsilon}\right).$$

Every class in that family that has a large graph dimension will therefore realize a gap between the sample complexities of different ERM learners.

Example 2 *Consider the set of all balls in $\mathbb{R}^n$ and, for each such ball, $B = B(z, r)$ with center z and radius r, let h_B be the function defined by $h_B(x) = z$ if $x \in B$ and $h_B(x) = \star$*

otherwise. Let $\mathcal{H}_{\mathcal{B}^n} = \{h_B : B = B(z,r) \text{ for some } z \in \mathbb{R}^n,\ r \in \mathbb{R}\} \cup \{h_\star\}$ (where $h_\star$ is the constant $\star$ function). It is not hard to see that $d_N(\mathcal{H}_{\mathcal{B}^n}) = 1$ and $d_G(\mathcal{H}_{\mathcal{B}^n}) = n+1$. Furthermore, let A_{good} be the ERM learner that for every sample $S = (x_1, f(x_1)), \dots (x_m, f(x_m))$, returns h_{B_S}, where B_S is the minimal ball that is consistent with the sample. Note that this algorithm uses, for every $f \in \mathcal{H}_{\mathcal{B}^n}$ and every sample S labeled by such f, at most one value (the value $\star$) on top of the values $\{f(x_1), \dots, f(x_m)\}$.

In this case, Theorem 7 implies that for some constant C_1, there exists a bad ERM learner, A_{bad} such that

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

On the other hand, Lemma 12 implies that there is a good ERM learner, A_{good} and a constant C_2 for which

$$m^r_{A_{\text{good}}}(\epsilon,\delta) \leq C_2 \left(\frac{\ln(1/\epsilon) + \ln(1/\delta)}{\epsilon}\right).$$

Note that, if one restricts the hypothesis class to allow only balls that have their centers in some finite set of grid points, the class uses only a finite range of labels. However, if such a grid is sufficiently dense, the sample complexities of both algorithms, A_{bad} and A_{good}, would not change.

2.5. Symmetric Classes

The principle for choosing a good ERM leads to tight bounds on the sample complexity of *symmetric classes*. Recall that a class $\mathcal{H}$ is called symmetric if for any $f \in \mathcal{H}$ and any permutation ϕ on labels, we have that $\phi \circ f \in \mathcal{H}$ as well.

Theorem 14 *There are absolute constants C_1, C_2 such that for every symmetric hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$*

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

A key observation that enables us to employ our principle in this case is:

Lemma 15 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a symmetric hypothesis class of Natarajan dimension d. Then, the range of any $f \in \mathcal{H}$ is of size at most $2d+1$.*

Proof If $|\mathcal{Y}| \leq 2d+1$ we are done. Thus assume that there are $2d+2$ distinct elements $y_1, \dots, y_{2d+2} \in \mathcal{Y}$. Assume to the contrary that there is a hypothesis $f \in \mathcal{H}$ with a range of more than d values. Thus there is a set $S = \{x_1, \dots, x_{d+1}\} \subseteq \mathcal{X}$ such that $f|_S$ has $d+1$ values in its range. It follows that $\mathcal{H}$ N-shatters S, thus reaching a contradiction. Indeed, since $\mathcal{H}$ is symmetric, there are functions $f_0, f_1 \in \mathcal{H}$ such that $f_j(x_i) = y_{j(d+1)+i}$. Similarly, for every $T \subseteq S$, there is a $g \in \mathcal{H}$ such that $g(x) = f_0(x)$ for every $x \in T$ and $g(x) = f_1(x)$ for every $x \in S \setminus T$. ■

We are now ready to prove Theorem 14.

Proof (of Theorem 14) The lower bound is a restatement of Theorem 4. For the upper bound, we define an algorithm A that conforms to the conditions in Lemma 12: Fix a set $\mathcal{Y}' \subseteq \mathcal{Y}$ of size $|\mathcal{Y}'| = \min\{|\mathcal{Y}|, 2d_N(\mathcal{H}) + 1\}$. Given a sample $(x_1, f(x_1)), \ldots, (x_m, f(x_m))$, A returns a hypothesis that is consistent with the sample and that attains only values in $\{f(x_1), \ldots, f(x_m)\} \cup \mathcal{Y}'$. It is possible due to symmetry and Lemma 15. ■

A similar analysis can be performed for the agnostic case. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a symmetric hypothesis class. Let $\mathcal{Y}' \subseteq \mathcal{Y}$ be an arbitrary set of size $\min\{|\mathcal{Y}|, 4d_N(G)+2\}$. Denote $\mathcal{H}' = \{f \in \mathcal{H} : f(\mathcal{X}) \subseteq \mathcal{Y}'\}$. Using lemma 15 and 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 3, we conclude that $d_G(\mathcal{H}) = O(\log(d_N(\mathcal{H})) \cdot d_N(\mathcal{H}))$. Using equation 5 we obtain a sample complexity bound of

$$m^a_{\mathcal{H}}(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2}\left(\log(\min\{d_N(\mathcal{H}), |\mathcal{Y}|\}) \cdot d_N(\mathcal{H}) + \ln(\frac{1}{\delta})\right)\right),$$

which is better than Equation. (6). Moreover, the ratio between this bound and the lower bound (Equation. (5)) is $O(\log(d_N(\mathcal{H})))$ regardless of $|\mathcal{Y}|$. Note that this bound is attained by any ERM. We present the following open question:

Open question 16 *Examples 1 and 2 show that there are (non-symmetric) hypothesis classes with a ratio of* $\Omega(\ln(|\mathcal{Y}|))$ *between the sample complexities of the worst ERM learner and the best ERM learner. How large can this gap be for symmetric hypothesis classes?*

3. Multiclass Learning in the Online Model

Learning in the online model is conducted in a sequence of consecutive rounds. On each round $t = 1, 2, \ldots$, the environment presents a sample $x_t \in \mathcal{X}$, the algorithm should predict a value $\hat{y}_t \in \mathcal{Y}$, and then 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 $y_1, \ldots, y_{t-1}$. We start with the realizable case, in which we assume that for some function $f \in \mathcal{H}$, all the outcomes are evaluations of f, namely, $y_t = f(x_t)$. Given an online learning algorithm, A, define its *(realizable) sample complexity*, $\mathcal{M}(A)$, to be the maximal number of wrong predictions that it might make on a legal sequence of any length.

The sample complexity of online learning has been studied by Littlestone (1987), who showed that a combinatorial measure, called the Littlestone dimension, characterizes the sample complexity of online learning. We now propose a generalization of the Littlestone dimension to classes of non-binary functions.

Consider a rooted tree T whose internal nodes are labeled by $\mathcal{X}$ and whose edges are labeled by $\mathcal{Y}$, such that the labels on edges from a parent to its child nodes are all different from each other. The tree T is *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 $f(x_i)$ equals the label of (x_i, x_{i+1}). The *Littlestone dimension*, L-dim($\mathcal{H}$), of $\mathcal{H}$ is the maximal depth of a complete binary tree that is shattered by $\mathcal{H}$.

It is not hard to see that, given a shattered tree of depth l, the environment can force any online learning algorithm to make l mistakes. Thus, for any algorithm A, $\mathcal{M}(A) \geq$ L-Dim($\mathcal{H}$). We shall now present an algorithm whose sample complexity is upper bounded by L-Dim($\mathcal{H}$).

```
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 $V_t = V_t^{(t_t)}$
```

Theorem 17 *$\mathcal{M}(SOA) = \text{L-Dim}(\mathcal{H})$.*

The proof is a simple adaptation of the proof of the binary case (see Littlestone, 1987). The idea is to note that for each t there is at most one $y \in \mathcal{Y}$ with $\text{L-Dim}(V_t^{(y)}) = \text{L-Dim}(V_t)$, and for the rest of the labels we have $\text{L-Dim}(V_t^{(y)}) < \text{L-Dim}(V_t)$. Thus, whenever the algorithm errs, the Littlestone dimension of V_t decreases by at least 1, so after $\text{L-Dim}(\mathcal{H})$ mistakes, V_t is composed of a single function.

Note that we only considered deterministic algorithms. However, allowing the algorithm to make randomized predictions does not substantially improve its sample complexity. It is easy to see that given a shattered tree of depth l, the environment can enforce any randomized online learning algorithm to make at least $l/2$ mistakes on average.

In the agnostic case, the sequence of outcomes, $y_1, \ldots, y_m$, is not necessarily realizable by some target function $f \in \mathcal{H}$. In that case, our goal is to have a *regret* of at most ϵ, where the regret is defined as

$$\frac{1}{m}|\{t \in [m] : \hat{y}_t \neq y_t\}| - \min_{f \in \mathcal{H}} \frac{1}{m}|\{t \in [m] : f(x_t) \neq y_t\}| \ .$$

We denote by $m_A^a(\epsilon)$ the number of examples required so that the regret of an algorithm A will be at most ϵ and by $m^a(\epsilon)$ the infimum, over all algorithms A, of $m_A^a(\epsilon)$.

Online learnability in the agnostic case, for classes of binary-output functions, has been studied in Ben-David et al. (2009), who showed that the Littlestone dimension characterizes the sample complexity in the agnostic case as well. The basic idea is to construct a set of experts by running the SOA algorithm on all sub-sequences of the examples whose length is at most $\text{L-Dim}(\mathcal{H})$, and then to run an online algorithm for learning with experts. This idea can be generalized to the multiclass case, but we leave this generalization to a longer version of this manuscript.

4. The Bandit Setting

So far we have assumed that each learning example is comprised of an instance and its corresponding label. In this section we deal with the so-called bandit setting. In the bandit model, 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 feedback, indicating whether its guess is correct or not.

4.1. Bandit vs Full Information in the Batch Model

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

Theorem 18 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class. Then,*

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

Proof Since the claim is trivial if $|\mathcal{Y}| = \infty$, we can assume that $k := |\mathcal{Y}| < \infty$. Let A_{full} be a (full information) ERM learner for $\mathcal{H}$. Consider the following algorithm 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 returns the hypothesis returned by A_{full} with the input sample which consists of the pairs (x_i, y_i) for which $\hat{y}_i = y_i$. We claim that $m_{A_{\text{bandit}}}(\epsilon, \delta) \le 3k \cdot m_{A_{\text{full}}}(\epsilon, \frac{\delta}{2})$ (for both the agnostic and the realizable case), so the theorem is implied by the bounds in the full information setting (theorem 7 and equation 5). Indeed, suppose that m examples suffice for A_{full} to return, with probability at least $1 - \frac{\delta}{2}$ a hypothesis with regret at most ϵ. Let $(x_i, y_i)_{i=1}^{3km}$ be a sample for the bandit algorithm. By Chernoff bound, with probability at least $1 - \frac{\delta}{2}$, the sample A_{bandit} transfers to A_{full} consist of at least m examples. Note that the sample that A_{full} receives is an i.i.d. sample according to the same distribution from which the original sample was sampled. Thus, with probability at least $1 - \frac{\delta}{2}$, A_{full} (and, consequently, A_{bandit}) returns a hypothesis with regret at most ϵ. ■

The price of bandit information in the batch model: Let $\mathcal{H}$ be a hypotheses class. Define $PBI_{\mathcal{H}}(\epsilon, \delta) = \frac{m^{r,b}_{\mathcal{H}}(\epsilon,\delta)}{m^{r}_{\mathcal{H}}(\epsilon,\delta)}$. By Theorems 18,4 and Equation 3 we see that, $PBI_{\mathcal{H}}(\epsilon, \delta) = O(\ln(|\mathcal{Y}|) \cdot \ln(\frac{1}{\epsilon}) \cdot |\mathcal{Y}|)$. 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 $PBI_{\mathcal{H}} = \Omega(|\mathcal{Y}|)$.

Using Theorems 18,4 and Equations 5,3 we see 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 and the lower bounds is $\Omega(\ln(|\mathcal{Y}|) \cdot |\mathcal{Y}|)$. It would be interesting to find a more tight characterization of the sample complexity in the bandit setting. The Natarajan dimension (as well as the graph dimension and other known notions of dimension defined in (Ben-David et al., 1995), as they are all closely related to the Natarajan dimension) is deemed to fail for the following reason: For every k, d, there are classes $\mathcal{H} \subseteq [k]^{[d]}$ of Natarajan dimension d where the realizable bandit sample complexity is $O(\frac{d}{\epsilon} + \frac{\ln(\frac{1}{\delta})}{\epsilon})$ (e.g. every class $\mathcal{H}$ such that $d_N(\mathcal{H}) = d$ and for every $x \in [d]$, $\#\{f(x) : f \in \mathcal{H}\} = 2$). On the other hand, the realizable bandit sample complexity of $[k]^{[d]}$ is $\Omega\left(k \cdot \left(\frac{d}{\epsilon} + \frac{\ln(\frac{1}{\delta})}{\epsilon}\right)\right)$.

4.2. 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. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and denote $k = |\mathcal{Y}|$. Consider a rooted tree T whose internal nodes are labeled by $\mathcal{X}$ and such that the labels on edges from a parent to its child nodes are all different from each other. 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 19 *Let $\mathcal{H}$ be a hypothesis class with $L = \text{BL-Dim}(\mathcal{H})$. The sample complexity of every deterministic online learning algorithm for $\mathcal{H}$ is at least L. Moreover, there is an online learning algorithm whose sample complexity is exactly L.*

Proof First, let T be a BL-shattered tree of depth L. We first 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.

Second, the following online learning algorithm makes at most L mistakes.

```
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 $\mathcal{M}(BSOA) \leq L$, note that at each time t, there is at least one $V_t^{(y)}$ with $\text{BL-Dim}(V_t^{(y)}) < \text{BL-Dim}(V_{t-1})$. 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. ■

We conclude with an open question on the price of bandit information in the online model:

Open question 20 *Let $PBI(\mathcal{H}) = \frac{\text{BL-Dim}(\mathcal{H})}{\text{L-Dim}(\mathcal{H})}$ and fix $k \geq 2$. How large can $PBI(\mathcal{H})$ be when $\mathcal{H}$ is a class of functions from a domain $\mathcal{X}$ to a range $\mathcal{Y}$ of cardinality k?*

5. The Sample Complexity of Known Multiclass Hypothesis Classes

In this section we analyze the sample complexity of two families of hypothesis classes for multiclass classification: the generalized linear construction (Duda and Hart, 1973; Vapnik, 1998; Hastie and Tibshirani, 1995; Freund and Schapire, 1997; Schapire and Singer, 1999; Collins, 2002; Taskar et al., 2003), and multiclass reduction trees (Beygelzimer et al., 2007, 2009; Fox, 1997). In particular, a special case of the generalized linear construction is the multi-vector construction (e.g. Crammer and Singer, 2003; Fink et al., 2006). We show that the sample complexity of the multi-vector construction and the reduction trees construction is similar and depends approximately linearly on the number of class labels. Due to the lack of space, proofs are omitted and can be found in the appendix.

5.1. The Generalized Linear Multiclass Construction

A generalized linear multiclass hypothesis class is defined with respect to a class specific feature mapping $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^t$, for some integer t. For any such ϕ define the hypothesis class $\mathcal{M}_\phi^t = \{h[w] \mid w \in \mathbb{R}^t\}$, where

$$h[w](x) = \operatorname*{argmax}_{y \in \mathcal{Y}} \langle w, \phi(x, y) \rangle,$$

where we ignore tie-breaking issues w.l.o.g. . A popular special case is the linear construction used in multiclass SVM (Crammer and Singer, 2003) where $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = [k]$, $t = dk$, and $\phi = \psi_{d,k}$, defined by

$$\psi_{d,k}(x, i) \triangleq (0, \ldots, 0, x[1], \ldots, x[d], 0, \ldots, 0),$$

where $x[1]$ is in coordinate $d(i-1)+1$. We abbreviate $\mathcal{L}_d^k \triangleq \mathcal{M}_{\psi_{d,k}}^{dk}$. We first consider a general ϕ and show that the sample complexity for any ϕ is upper-bounded by a function of t.

Theorem 21 *Let d_N be the Natarajan-dimension of $\mathcal{M}_\phi^t$. Then $d_N \le O(t \log(t))$.*

For the linear construction a matching lower bound on the Natarajan dimension is shown in the following theorem. Thus, as one might expect, the sample complexity of learning with $\mathcal{L}_k^d$ is of the order of dk.

Theorem 22 *For $d \ge 0$ and $k \ge 2$, let d_N be the Natarajan-dimension of $\mathcal{L}_k^d$. Then*

$$\Omega(dk) \le d_N \le O(dk \log(dk)).$$

5.2. Reduction trees

Reduction trees provide a way of constructing multiclass hypotheses from binary classifiers. A reduction tree consists of a tree structure, where each internal node is mapped to a binary classifier and each leaf is mapped to one of the multiclass labels. Classification of an example is done by traversing the tree, starting from the root and ending in one of the leaves, where in each node the result of the binary classifier determines whether to go left or right.

It has been shown that by using appropriate learning algorithms, one can guarantee a multiclass classification error of no more than $\log_2(k)\epsilon$, where k is the number of classes, and ϵ is the average error of the binary classifiers (Fox, 1997; Beygelzimer et al., 2009). However, this result does not directly provide sample complexity guarantees for these algorithms, since the value of ϵ itself depends on the sample and on the learning algorithm.

In the following we analyze the sample complexity of any fixed reduction tree, under the assumption that the binary classifiers all belong to some fixed hypothesis class with a finite VC-dimension d. We provide bounds on the Natarajan dimension of the resulting multiclass hypothesis class, and show that it can be as large as $\Omega(dk)$ for some hypothesis classes. We further analyze the special case where the binary hypothesis class is the class of linear separators in $\mathbb{R}^d$, and show that a similar result, though slightly weaker, holds for this class as well.

We now formally define a reduction tree and the hypothesis class related to it (see Figure 1 in the appendix for illustration). Let $\mathcal{X}$ be the domain of examples and let $[k]$ be the set of possible labels. A reduction tree is a full binary tree T. Denote the head node of T by $H(T)$. The sub-tree which is the left child of $H(T)$ is denoted by $L(T)$ and the sub-tree which is the right child of $H(T)$ is denoted by $R(T)$. The set of internal nodes of T is denoted by $N(T)$, and the set of leaf nodes of T is denoted by $\mathrm{leaf}(T)$. A multiclass classifier is a triplet $[T, \lambda, C]$ where T is a reduction tree, λ is a one-to-one mapping $\lambda[\cdot] : \mathrm{leaf}(T) \to [k]$, and $C[\cdot] : N(T) \to \{0,1\}^{\mathcal{X}}$ is a mapping from the internal nodes of T to binary classifiers on the domain $\mathcal{X}$. $[T, \lambda, C] : \mathcal{X} \to [k]$ is defined recursively as follows:

$$[T,\lambda,C](x) = \begin{cases} [L(T),\lambda,C](x) & H(T) \notin \mathrm{leaf}(T) \text{ and } C[H(T)](x) = 0, \\ [R(T),\lambda,C](x) & H(T) \notin \mathrm{leaf}(T) \text{ and } C[H(T)](x) = 1, \\ \lambda[H(T)](x) & H(T) \in \mathrm{leaf}(T). \end{cases}$$

Unless otherwise mentioned, we assume a fixed λ, and identify T with the pair (T, λ). Accordingly, $[T, \lambda, C]$ is abbreviated to $[T, C]$. Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class of binary classifiers on $\mathcal{X}$. The hypothesis class induced by $\mathcal{H}$ on the tree T with label mapping λ, denoted by $\mathcal{H}_{(T,\lambda)}$, is the set of multiclass classifiers which can be generated on T using binary classifiers from $\mathcal{H}$. Formally,

$$\mathcal{H}_{(T,\lambda)} = \{[T,\lambda,C] \mid \forall n \in N(T), C[n] \in \mathcal{H}\}.$$

We abbreviate $\mathcal{H}_{(T,\lambda)}$ to $\mathcal{H}_T$ when the labeling λ is fixed.

Suppose that the VC-dimension of $\mathcal{H}$ is d. What can be said about the sample complexity of $\mathcal{H}_T$ for a given tree T? First, a simple counting argument provides an upper bound on the graph-dimension and the Natarajan-dimension of $\mathcal{H}_T$: Any hypothesis in $\mathcal{H}_T$ is a function of the values of $|N(T)| = k-1$ binary hypotheses from $\mathcal{H}$. Therefore, the number of possible labelings of A by $\mathcal{H}_T$ for any $A \subseteq \mathcal{X}$ is bounded by $|\mathcal{H}|_A|^{k-1}$. By Sauer's lemma, $|\mathcal{H}|_A| \le |A|^d$. Thus $|\mathcal{H}_T|_A| \le |A|^{d(k-1)}$. If A is G-shattered or N-shattered by $\mathcal{H}_T$, then $|\mathcal{H}_T|_A| \ge 2^{|A|}$. Thus $2^{|A|} \le |A|^{d(k-1)}$. It follows that $|A| \le O(dk\log(dk))$, thus the same upper bound holds for the graph-dimension and the Natarajan-dimension. A closely matching lower bound is provided in the following theorem.

Theorem 23 *Let $k \ge 2$ and $d \ge 2$ be integers. For any reduction tree T with $k \ge 2$ leafs, there exists a binary hypothesis class $\mathcal{H}$ with VC-dimension d such that $\mathcal{H}_T$ has Natarajan dimension $d(k-1)$.*

Theorem 23 shows that for every tree there exists a binary hypothesis class which induces a high sample complexity on the resulting multiclass hypothesis class. The following theorem shows that moreover, the popular hypothesis class of linear separators in $\mathbb{R}^d$ induces reduction trees with a sample complexity which is almost as large, up to a logarithmic factor.

Let $\mathcal{W}^d$ be the class of non-homogeneous linear separators in $\mathbb{R}^d$, that is $\mathcal{W}^d = \{x \to \text{sign}(\langle x, w\rangle + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$. For a full binary tree T with k leaves, denote by $n_1(T)$ the number of internal nodes with one leaf child and one non-leaf child, and by $n_2(T)$ the number of internal nodes with two leaf children.

Theorem 24 *For any multiclass-to-binary tree T with k leaves, the graph dimension of $\mathcal{W}^d_T$ is at least $(d+1)\cdot n_2(T) + d \cdot n_1(T) \geq dk/2$. Consequently the Natarajan dimension is $\Omega(dk/\log(k))$.*

We conclude that the sample complexity of different reduction trees is similar, and that this sample complexity is also similar to that of the multi-vector construction. This implies that when choosing between the different hypothesis classes, considerations other than the sample complexity should determine the choice. One such important consideration is the approximation error. Since sample complexity analysis bounds only the estimation error, one wishes to have the approximation error as low as possible. Thus if there is some prior knowledge on the match between the hypothesis class and the source distribution, this might guide the choice of the hypothesis class. The following theorem shows, however, that for fairly balanced reduction trees this match is highly dependent on the assignment of labels to leaf nodes. For any reduction tree T denote by Λ the set of one-to-one mappings from the leaf(T) to $[k]$, and let U be the uniform distribution over Λ.

Theorem 25 *Let T be a full binary tree with k leaves, and let n be the number of leaves on the left sub-tree. For any hypothesis class $\mathcal{H}$ with VC-dimension d, and for any distribution D over $\mathcal{X} \times [k]$ which assigns non-zero probability to each label in $[k]$,*

$$\Pr_{\lambda \sim U}[\mathcal{H}_{(T,\lambda)} \text{ separates } D] \leq \left(\frac{ek}{d}\right)^d \binom{k}{n}^{-1}.$$

Thus if $k \gg d$ and n is a constant fraction of k, this probability decreases exponentially with k.

6. Conclusions and Open Problems

In this paper we have studied several new aspects of multiclass sample complexity. Many interesting questions arise and some are listed below.

Consider the two example classes from section 2.4. It is interesting to note that, in both cases, $d_N(\mathcal{H}) = 1$, and $m^r_{\mathcal{H}}(\epsilon, \delta) = \Theta(\frac{1}{\epsilon}\ln(\frac{1}{\delta}))$. It seems like the Natarajan dimension is the parameter that controls the sample complexity for those examples. That is also the case for symmetric classes as well as some other classes that we have examined but did not include in this paper. We therefore raise:

Conjecture 26 *There exists a constant C such that, for every hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$,*

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

In light of theorem 7 and the fact that there are cases where $d_G \geq \log_2(|\mathcal{Y}| - 1) d_N$, in order to prove the conjecture we will have to find 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 correct way to learn. We suspect that such an understanding might lead to improved bounds in the binary case as well. We hope that our examples from section 2.4 and our result for symmetric classes will prove to be the first steps in the search for the best ERM.

Another direction is the study of learnability conditions for additional hypotheses classes. Section 5 shows that some well known multiclass constructions have surprisingly similar sample complexity properties. It is of practical significance and theoretical interest to study learnability conditions for other constructions, and especially to develop a fuller understanding of the relationship between different constructions, in a manner that could guide an informed choice of a hypothesis class.

Acknowledgments

Sivan Sabato is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities.

References

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.

S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. Long. Characterizations of learnability for classes of $\{0, \dots, n\}$-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.

Alina Beygelzimer, John Langford, and Pradeep Ravikumar. Error-correcting tournaments. *CoRR*, 2009.

M. Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Conference on Empirical Methods in Natural Language Processing*, 2002.

K. Crammer and Y. Singer. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991, 2003.

R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley, 1973.

Michael Fink, Shai Shalev-Shwartz, Yoram Singer, and Shimon Ullman. Online multiclass learning by interclass hypothesis sharing. In *International Conference on Machine Learning*, 2006.

J. Fox. *Applied Regression Analysis, Linear Models, and Related Methods.* SAGE Publications, 1997.

Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

T.J. Hastie and R.J. Tibshirani. *Generalized additive models.* Chapman & Hall, 1995.

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

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.

R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):1–40, 1999.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *NIPS*, 2003.

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

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

Appendix A. Proofs Omitted from the Text

Proof (of theorem 4)
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 $m_{\mathcal{H}_d} \le m_{\mathcal{H}}$, so the lower bound is obtained by theorem 2. Let A be a learning algorithm for $\mathcal{H}$. Consider the learning algorithm, $\bar{A}$, for $\mathcal{H}_d$ defined as follows. Let $S = \{s_1, \ldots, s_d\} \subseteq X,\ f_0, f_1$ be a set and functions that indicate that $d_N(\mathcal{H}) = d$. Given a sample $(x_i, y_i) \in [d] \times \{0,1\},\ i = 1, \ldots, m$, let $g = A((s_{x_i}, f_{y_i}(s_{x_i}))_{i=1}^m)$. Define $f = \bar{A}((x_i, y_i)_{i=1}^m)$ by setting $f(i) = 1$ if and only if $g(s_i) = f_1(s_i)$. It is not hard to see that $m_{\bar{A}} \le m_A^r$, thus, $m_{\mathcal{H}_d} \le m_{\mathcal{H}}$.
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 $VC(\bar{\mathcal{H}}) = d_G(\mathcal{H})$.

Suppose that $f \in \mathcal{H}$ is consistent with a sample $(x_i, f_0(x_i))_{i=1}^m$ of $m = \Omega(\frac{d}{\epsilon}\ln(\frac{1}{\epsilon}) + \frac{1}{\epsilon}\ln(\frac{1}{\delta}))$ examples, drawn i.i.d. according to some distribution $\mathcal{D}$ on $\mathcal{X}$. We must show that, with probability $\geq 1-\delta$, $\mathrm{Err}_{\mathcal{D},f_0}(f) \leq \epsilon$. However, by theorem 2,

$$\mathrm{Err}_{\mathcal{D},f_0}(f) = \Pr_{x\sim\mathcal{D}}(\bar{f}(x, f_0(x)) \neq 1) \leq \epsilon$$

With probability $\geq 1-\delta$. ■

Proof (of Theorem 6) Let A be an ERM learner. Since $F_A(f) \subseteq \mathcal{H}$ for every f, it follows that $\Pi_A \leq \Pi_{\mathcal{H}}$. By lemma 11, $\Pi_{\mathcal{H}}(m) \leq m^{d_N(\mathcal{H})}|\mathcal{Y}|^{2d_N(\mathcal{H})}$. Incorporating it into Theorem 9 we get the desired bound. ■

Proof (of Theorem 21) Let $S = \{x_1, \ldots, x_{d_N}\} \subseteq \mathbb{R}^d$ be a set which is N-shattered by $\mathcal{M}_\phi^t$, and let $f_1, f_2 : S \to \mathcal{Y}$ be the functions that witness the shattering. For every $i \in [d_N]$ let $z_i = \phi(x_i, f_1(x_i)) - \phi(x_i, f_2(x_i)) \in \mathbb{R}^t$. Denote $Z = \{z_i\}_{i\in[d_N]}$. Consider the hypothesis class of homogeneous linear separators in $\mathbb{R}^t$, defined by $\{z \to \mathrm{sign}(\langle w, z\rangle) \mid w \in \mathbb{R}^t\}$. Since the VC-dimension of this class is t, by Sauer's lemma the number of possible labelings of Z with this class is upper-bounded by $(d_N)^t$. We now show that there is a one-to-one mapping from subsets $T \subseteq S$ to labelings of Z: For any $T \subseteq S$, let $w \in \mathbb{R}^t$ such that

$$\{x \in S \mid h[w](x) = f_1(x)\} = T, \text{ and } \{x \in S \mid h[w](x) = f_2(x)\} = S \setminus T.$$

Then $T = \{x \in S \mid \langle w, \phi(x, f_1(x))\rangle \geq \langle w, \phi(x, f_2(x))\rangle\} = \{x_i \mid \langle w, z_i\rangle \geq 0\}$. Thus every T induces a different labeling of Z. It follows that the number of subsets of S is bounded by the number of labelings of Z, thus $2^{d_N} \leq (d_N)^t$. It follows that $d_N \leq O(t\log(t))$. ■

Proof (of Theorem 22) The upper bound is a direct consequence of Theorem 21. For the lower bound, we show that there exists an N-shattered set of size $\lfloor d/2\rfloor \cdot \lfloor k/2\rfloor$. Let $b = \lfloor k/2\rfloor$. Let $x_1, \ldots, x_b \in \mathbb{R}^2$ be b different vectors such that $\forall i \in [b], ||x_i|| = 1$. Let $S = \{y_{i,j}\}_{i\in[b], j\in[\lfloor d/2\rfloor]} \subseteq \mathbb{R}^d$, where for $s \in [d]$:

$$y_{i,j}[s] = \begin{cases} x_i[1] & s = 2j-1 \\ x_i[2] & s = 2j \\ 0 & \text{otherwise.} \end{cases}$$

We show that S is N-shattered, thus $d_N \geq |S| = \lfloor k/2\rfloor \cdot \lfloor d/2\rfloor$. Define functions $f_1, f_2 : S \to [k]$ such that for $y_{i,j} \in S$, $f_1(y_{i,j}) = i$ and $f_2(y_{i,j}) = b + i$. For a subset $T \subseteq S$, let $w \in \mathbb{R}^{dk}$ such that for $i \in [b], s \in [d]$

$$w[d(i-1)+s] = \begin{cases} x_i[1] & y_{i,j} \in Z \text{ and } s = 2j-1, \\ x_i[2] & y_{i,j} \in Z \text{ and } s = 2j, \\ 0 & \text{otherwise.} \end{cases}$$

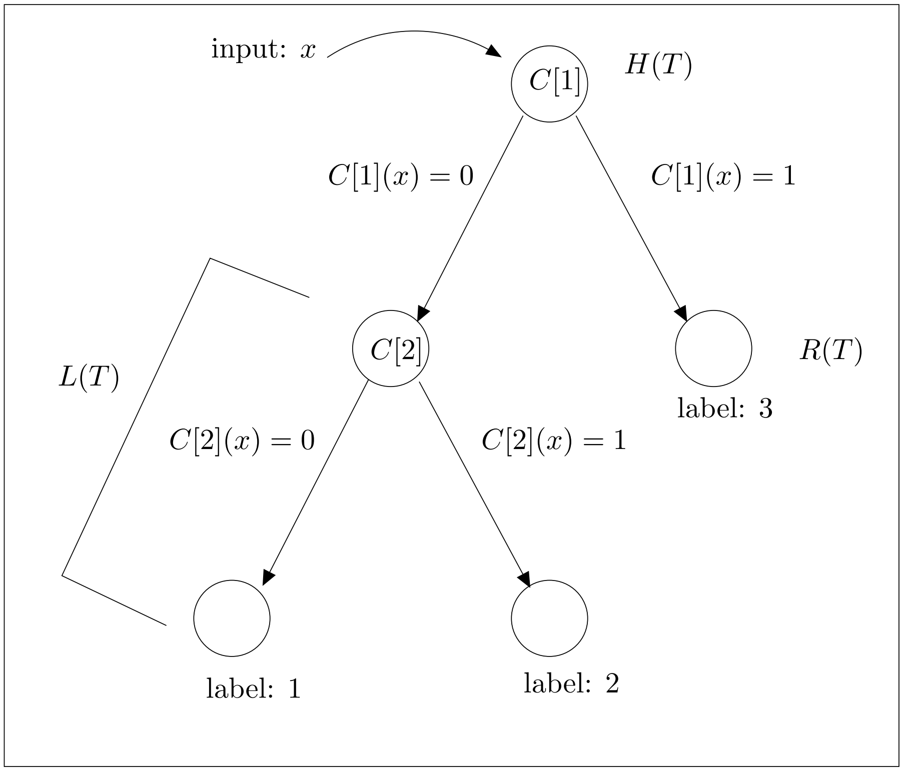

Figure 1: Illustration of a reduction tree

and for $i \in \{b+1, \ldots, 2b\}, s \in [d]$,

$$w[d(i-1)+s] = \begin{cases} x_i[1] & y_{i-b,j} \notin Z \text{ and } s = 2j-1, \\ x_i[2] & y_{i-b,j} \notin Z \text{ and } s = 2j, \\ 0 & \text{otherwise.} \end{cases}$$

Then $h[w] = f_1(y)$ for $y \in T$ and $h[w] = f_2(y)$ for $y \in S \setminus T$. Thus S is N-shattered. ■

Proof (of Theorem 23) Let $\mathcal{H}(T)$ be a binary hypothesis class for tree T. We construct $\mathcal{H}(T)$ inductively on the structure of the tree. For every tree T, the domain of the binary hypotheses in $\mathcal{H}(T)$ will be $[d] \times N(T)$.

Induction basis: Assume that both $L(T)$ and $R(T)$ are leafs, thus $k = 2$ and $|N(T)| = 1$. Define $\mathcal{H}(T) = \{h \mid h : [d] \times \{H(T)\} \to \{0, 1\}\}$.

Inductive step: Assume T has two children $L(T)$ and $R(T)$, and at least one of them is not a leaf. By the induction hypothesis, if $L(T)$ is a non-leaf then $\mathcal{H}(L(T))$ is a set of binary hypotheses with domain $[d] \times N(L(T))$. $\mathcal{H}(L(T))$ has VC-dimension d, and the Natarajan dimension of $\mathcal{H}(L(T))_{L(T)}$ is $d \cdot |N(L(T))|$. The same holds for $R(T)$. Define $\mathcal{H}(T) = \{h_0, h_1\} \cup \mathcal{H}_L \cup \mathcal{H}_R \cup \mathcal{H}_H$, where:

- $h_0(x) = 0$ and $h_1(x) = 1$ for all $x \in [d] \times N(T)$,

- If $L(T)$ is a leaf, $\mathcal{H}_L = \emptyset$. Otherwise,

$$\mathcal{H}_L = \Big\{h : [d] \times N(T) \to \{0,1\} \mid \exists h_L \in \mathcal{H}(L(T)), \forall x \in [d] \times N(T), \\ h(x) = \begin{cases} h_L(x) & x \in [d] \times N(L(T)), \\ 0 & \text{otherwise.} \end{cases} \Big\}$$

- $\mathcal{H}_R$ is defined similarly, for $R(T)$ instead of $L(T)$.

- $\mathcal{H}_H$ is defined as follows:

$$\mathcal{H}_H = \{h : [d] \times N(T) \to \{0,1\} \mid \forall x \in [d] \times N(L(T)), h(x) = 0, \\ \forall x \in [d] \times N(R(T)), h(x) = 1\}.$$

We now prove by induction that for every tree T the following claims hold:

- $\mathcal{H}(T)$ has VC-dimension d,

- $\mathcal{H}(T)_T$ has Natarajan dimension $d \cdot |N(T)|$.

- An auxiliary claim: $\mathcal{H}(T)$ includes the hypotheses h_0 and h_1.

Induction Basis: If both $L(T)$ and $R(T)$ are leafs, then the VC-dimension of $\mathcal{H}(T)$ is clearly d. The induced multiclass hypothesis class $\mathcal{H}(T)_T$ is in fact a set of binary hypotheses which is isomorphic to $\mathcal{H}(T)$, thus its Natarajan dimension is also $d = d(k-1)$. The zero hypothesis is clearly in $\mathcal{H}(T)$ by construction.

Induction Step: Assume T has two children $L(T)$ and $R(T)$, and at least one of them is not a leaf. By the construction of $\mathcal{H}(T)$, the auxiliary claim clearly holds. The following lemmas, whose proofs follows, prove the two other claims:

Lemma 27 *$\mathcal{H}(T)$ has VC-dimension d.*

Lemma 28 *$\mathcal{H}(T)_T$ has Natarajan dimension $d|N(T)|$.*

Thus the induction hypothesis holds. ∎

Proof (of Lemma 27) The VC-dimension of $\mathcal{H}(T)$ is at least d, since the VC-dimension of at least one of $\mathcal{H}(L(T))$ and $\mathcal{H}(R(T))$ is d. Assume to the contrary that it is larger than d, then there exists a set $A = \{x_1, \ldots, x_{d+1}\} \subseteq [d] \times N(T)$ which is shattered by $\mathcal{H}(T)$. Denote for brevity $S_L = [d] \times N(L(T))$, $S_R = [d] \times N(R(T))$ and $S_H = [d] \times H(T)$. By the construction of $\mathcal{H}(T)$ and the auxiliary claim, $\mathcal{H}(T)|_{S_L} = \mathcal{H}(L(T))$ and $\mathcal{H}(T)|_{S_R} = \mathcal{H}(R(T))$ whenever $L(T)$ and $R(T)$ are not leaves respectively. In addition, since $|S_H| = d$, $A \nsubseteq S_H$. Since $|A| \geq 3$, there exist three different elements in $x, y, z \in A$ such that at least two of them are in different sets out of S_L, S_H, S_R. We consider the different cases (where names of elements are w.l.o.g.) and show for each case a labeling l_x, l_y, l_z for x, y, z that cannot be achieved with a hypothesis in $\mathcal{H}(T)$:

- If $x \in S_H$, $y \in S_R$ then $l_x = 1, l_y = 0$ cannot be achieved.

- If $x, y \in S_L,\ z \in S_H \cup S_R$ then $l_x = 1, l_y = 0, l_z = 1$ cannot be achieved.
- If $x \in S_L,\ y, z \in S_R$ then $l_x = 1, l_y = 0, l_z = 1$ cannot be achieved.
- If $x \in S_L,\ y, z \in S_H$ then $l_x = 1, l_y = 0, l_z = 1$ cannot be achieved.

We have reached a contradiction, therefore no such A exists. ■

Proof (of Lemma 28) The Natarajan dimension is upper bounded by the size of the domain, which is $d|N(T)|$. By the induction hypothesis, $\mathcal{H}(L(T))_{L(T)}$ and $\mathcal{H}(R(T))_{R(T)}$ have Natarajan dimension $d_L = d|L(T)|$ and $d_R = d|R(T)|$ respectively. Thus $[d] \times N(L(T))$ and $[d] \times N(R(T))$ are N-shattered by $\mathcal{H}(L(T))_{L(T)}$ and $\mathcal{H}(R(T))_{R(T)}$ respectively. Let f_1^L, f_2^L, and f_1^R, f_2^R be the pairs of functions that witness the N-shattering of $\mathcal{H}(L(T))_{L(T)}$ and $\mathcal{H}(R(T))_{R(T)}$ respectively. Let c_L be the class of the left-most child in $L(T)$, and let c_R be the class of the left-most child in $R(T)$. define g_1 and g_2 as follows:

$$g_1(x) = \begin{cases} f_1^L(x) & x \in [d] \times N(L(T)) \\ f_1^R(x) & x \in [d] \times N(R(T)) \\ c_L & x \in [d] \times \{H(T)\} \end{cases}$$

$$g_2(x) = \begin{cases} f_2^L(x) & x \in [d] \times N(L(T)) \\ f_2^R(x) & x \in [d] \times N(R(T)) \\ c_R & x \in [d] \times \{H(T)\} \end{cases}$$

It is easy to verify that $[d] \times N(T)$ is N-shattered using g_1 and g_2. ■

Proof (of Theorem 24) The proof is by induction on the structure of the tree.

Induction basis: Assume that T is a tree with one internal node and two leaf children. Then $\mathcal{W}_T^d$ is isomorphic up to label names to $\mathcal{W}^d$. Thus the graph dimension of $\mathcal{W}_T^d$ is equal to the VC-dimension of $\mathcal{W}^d$, that is $d + 1 = (d + 1) \cdot n_1(T)$.

Inductive step: We consider two cases: Either both $R(T)$ and $L(T)$ are non-leaves or ons is a leaf and one is not.

Case 1: Let T be a tree where both $L(T)$ and $R(T)$ are non-leaves. By the induction hypothesis, the graph dimension of $\mathcal{W}_{L(T)}^d$ is at least $d_L = (d+1) \cdot n_2(L(T)) + d \cdot n_1(L(T))$ and the graph dimension of $\mathcal{W}_{R(T)}^d$ is at least $d_R = (d+1) \cdot n_2(R(T)) + d \cdot n_1(R(T))$. Thus there exist sets $A_L = \{a_1, \ldots, a_{d_L}\}$ and $B_R = \{b_1, \ldots, b_{d_R}\}$ which are G-shattered by $L(T)$ and $R(T)$ respectively, using functions f_L and f_R respectively. Let

$$a_L = (\min_{i \in [d_L]} \{a_i[1]\} + 1, 0, \ldots, 0) \in \mathbb{R}^d$$

$$b_R = (-\max_{i \in [d_R]} \{b_i[1]\} - 1, 0, \ldots, 0) \in \mathbb{R}^d$$

Let $\tilde{A}_L = \{a_1 + a_L, \ldots, a_{d_L} + a_L\}$ and let $\tilde{B}_R = \{b_1 + b_R, \ldots, b_{d_L} + b_R\}$. Then $\forall x \in \tilde{A}_L$, $x[1] > 0$, and $\forall x \in \tilde{B}_R$, $x[1] < 0$.

We show that the set $\tilde{A}_L \cup \tilde{B}_R$ is G-shattered by $\mathcal{W}^d_T$: Define

$$f(x) = \begin{cases} f_R(x) & x[1] > 0 \\ f_L(x) & \text{otherwise.} \end{cases}$$

Let $Z \subseteq \tilde{A}_L \cup \tilde{B}_R$. We construct a mapping $C : N(T) \to \mathcal{H}$ such that

$$\{x \in \tilde{A}_L \cup \tilde{B}_R \mid [T, C](x) = f(x)\} = Z.$$

Let $Y \subseteq A_L \cap B_R = \{a_i \mid a_i + a_L \in Z\} \cap \{b_i \mid b_i + b_R \in Z\}$. Since A_L and B_R are G-shattered with f_L and f_R, there exist mappings $C_L : N(L(T)) \to \mathcal{W}^d$ and $C_R : N(R(T)) \to \mathcal{W}^d$ such that

$$\{x \in A_L \mid [L(T), C_L](x) = f_L(x)\} = Y \cap A_L,$$
$$\{x \in B_R \mid [R(T), C_R](x) = f_R(x)\} = Y \cap B_R.$$

Define the mapping C as a translation of the mappings C_L and C_R, defined by:

$$\forall n \in L(T), C_L[n] = (w, b) \Rightarrow C[n] = (w, b - \langle w, a_L \rangle),$$
$$\forall n \in R(T), C_R[n] = (w, b) \Rightarrow C[n] = (w, b - \langle w, b_R \rangle).$$

Then

$$\{x \in \tilde{A}_L \mid [L(T), C](x) = f_L(x)\} = Z \cap \tilde{A}_L,$$
$$\{x \in \tilde{B}_R \mid [R(T), C](x) = f_R(x)\} = Z \cap \tilde{B}_R.$$

Now, set $C[H(T)](x) = \text{sign}(\langle x, w \rangle + b)$ where $w = (1, 0, \dots, 0)$ and $b = 0$. Then

$$\forall x \in \tilde{A}_L, [T, C](x) = [L(T), C](x) = f_L(x) = f(x),$$
$$\forall x \in \tilde{B}_R, [T, C](x) = [R(T), C](x) = f_R(x) = f(x).$$

Thus $\tilde{A}_L \cup \tilde{B}_R$ is G-shattered by $\mathcal{W}^d_T$. It follows that the graph dimension of $\mathcal{W}^d_T$ is at least $|\tilde{A}_L \cup \tilde{B}_R| = d_L + d_R = (d+1) \cdot n_2(T) + d \cdot n_1(T)$.

Case 2: Assume w.l.o.g. that T is a tree where $L(T)$ is not a leaf node and $R(T)$ is a leaf node with $\lambda[R(T)] = t$. By the induction hypothesis, the graph dimension of $\mathcal{W}^d_{L(T)}$ is at least $d_L = (d+1) \cdot n_2(L(T)) + d \cdot n_1(L(T))$. Thus there exists a set $A = \{a_1, \dots, a_{d_L}\}$ which is G-shattered by $L(T)$ using the function f_L.

Denote by e_i the i'th unit vector in $\mathbb{R}^d$, and let $q > 0$ be large enough such that $\{(0, \dots, 0), qe_1, \dots, qe_d\}$ is shattered with a margin of $2M$, where $M = \max_{x \in A} ||x||_2$. Let $B = A \cup \{qe_1, \dots, qe_d\}$. Then we show B is G-shattered using the following function f:

$$f(x) = \begin{cases} f_L & ||x|| \leq q \\ t & \text{otherwise.} \end{cases}$$

Let $Z \subseteq B$. We construct a mapping $C : N(T) \to \mathcal{H}$ such that

$$\{x \in B \mid [T, C](x) = f(x)\} = Z. \tag{8}$$

Since A is G-shattered using f_L, there exists a mapping $C_L : N(L(T)) \to \mathcal{W}^d$ such that $\{x \in A \mid [L(T), C_L](x) = f_L(x)\} = Z \cap A$. Define C such that $\forall n \in N(L(T)), C[n] = C_L[n]$. In addition, Let $C[H(T)] \in \mathcal{W}^d$ be a hypothesis such that $\forall i, e_i \in Z \iff h(e_i) = 1$, and $\forall x, ||x||_2 \leq M \to h(0) = 0$. Then Equation. (8) holds. Thus the graph dimension of $\mathcal{W}^d_T$ is at least $|B| = d_L + d \geq (d+1) \cdot n_2(T) + d \cdot n_1(T)$. ■

Proof (of Theorem 25) If suffices to consider distributions with deterministic labeling, such that the correct label is a function $f : \mathcal{X} \to [k]$. Let $A = \{x_1, \ldots, x_k\} \in \mathcal{X}$ such that for all $i \in [k]$, $f(x_i) = i$. For any labeling $\lambda \in \Lambda$, let $f_\lambda : A \to \{0, 1\}$ be the indicator function of the set of labels assigned to leaves in $L(T)$, that is $f_\lambda(x_i) = \mathbf{1}[\exists n \in \text{leaf}(L(T)), \lambda[n] = i]$. If D is separable with $\mathcal{H}_{(T,\lambda)}$ then $f_\lambda = C[H((T,\lambda))]|_A \in \mathcal{H}|_A$. By Sauer's lemma, $|\mathcal{H}|_A| \leq \left(\frac{ek}{d}\right)^d$. There are $\binom{k}{n}$ possible indicator functions f_λ for a labeling λ, and they all have equal probability for $\lambda \sim U$. Thus $\mathbb{P}_{\lambda \sim U}[f_\lambda \in \mathcal{H}|_A] \leq \left(\frac{ek}{d}\right)^d / \binom{k}{n}$. ■