Optimal Learners for Multiclass Problems

Amit Daniely *Dept. of Mathematics, The Hebrew University, Jerusalem, Israel*

Shai Shalev-Shwartz *School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel*

Abstract

The fundamental theorem of statistical learning states that for *binary* classification problems, any Empirical Risk Minimization (ERM) learning rule has close to optimal sample complexity. In this paper we seek for a generic optimal learner for *multiclass* prediction. We start by proving a surprising result: a generic optimal multiclass learner must be *improper*, namely, it must have the ability to output hypotheses which do not belong to the hypothesis class, even though it knows that all the labels are generated by some hypothesis from the class. In particular, no ERM learner is optimal. This brings back the fundmamental question of "how to learn"? We give a complete answer to this question by giving a new analysis of the one-inclusion multiclass learner of Rubinstein et al. (2006) showing that its sample complexity is essentially optimal. Then, we turn to study the popular hypothesis class of generalized linear classifiers. We derive optimal learners that, unlike the one-inclusion algorithm, are computationally efficient. Furthermore, we show that the sample complexity of these learners is better than the sample complexity of the ERM rule, thus settling in negative an open question due to Collins (2005).

1. Introduction

Multiclass classification is the problem of learning a classifier h from a domain $\mathcal{X}$ to a label space $\mathcal{Y}$, where $|\mathcal{Y}| > 2$ and the error of a prediction is measured by the probability that $h(x)$ is not the correct label. It is a basic problem in machine learning, surfacing a variety of domains, including object recognition, speech recognition, document categorization and many more. Over the years, multiclass classification has been subject to intense study, both theoretical (Natarajan, 1989; Ben-David et al., 1995; Rubinstein et al., 2006; Daniely et al., 2011, 2012) and practical (e.g. (Shalev-Shwartz et al., 2004; Collins, 2005; Keshet et al., 2005; Torralba et al., 2007)). Many methods have been developed to tackle this problem, starting from the the naive one-vs-all method, to more complex methods, such as structured output prediction (Collins, 2000, 2002; Lafferty et al., 2001; Taskar et al., 2003; Tsochantaridis et al., 2004), error correcting output codes (Dietterich and Bakiri, 1995) and others. These developments made it possible to handle a variety of multiclass classification problems, including even problems that have a very complex label space, that is structured and exponentially large (e.g. speech recognition, OCR, and multiple object categorization).

Despite being very basic and natural, and despite these developments and efforts, our theoretical understanding of multiclass classification is still far from being satisfactory, in particular relatively to our understanding of binary classification (i.e., when $|\mathcal{Y}| = 2$). In this work, we focus on the sample complexity of (distribution free) learning of hypothesis classes $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. The two most fundamental questions are:

1. What is learnable? More quantitatively, what is the sample complexity of a given class $\mathcal{H}$?

2. How to learn? In particular, is there a generic algorithm with optimal sample complexity?

For binary classification problems, these two questions are essentially solved (up to log-factors of the error and confidence parameters ϵ and δ): The fundamental result of Vapnik and Chervonenkis (1971) asserts that the VC dimension characterizes the sample complexity, and that any Empirical Risk Minimization (ERM) algorithm enjoys close-to-optimal sample complexity.

In a recent surprising result, Daniely et al. (2011) have shown that in multiclass classification there might be substantial gaps between the sample complexity of different ERMs. We start by showing an even stronger "peculiarity", discriminating binary from multiclass classification. Recall that an algorithm is called *improper* if it might return a hypothesis that does not belong to the learnt class. Traditionally, improper learning has been applied to enable efficient computations. It seems counter intuitive that computationally unbounded learner would benefit from returning a hypothesis outside of the learnt class. Surprisingly, we show that an optimal learning algorithm *must* be improper! Namely, we show that there are classes that are learnable *only* by an improper algorithm. Pointing out that we actually do not understand how to learn optimally, these results "reopen" the above two basic questions for multiclass classification.

In this paper we essentially resolve these two questions. We give a new analysis of the multiclass one inclusion algorithm (Rubinstein et al. (2006) based on Haussler et al. (1988), see also Simon and Szörényi (2010)), showing that it is optimal up to a constant factor of 2 in

a transductive setting. This improves on the original analysis, that yielded optimality only up to a factor of $\log(|\mathcal{Y}|)$ (which, as explained, might be quite large in several situations). By showing reductions from transductive to inductive learning, we consequently obtain an optimal learner in the PAC model, up to a logarithmic factor of $\frac{1}{\delta}$ and $\frac{1}{\epsilon}$. The analysis of the one inclusion algorithm results with a characterization of the sample complexity of a class $\mathcal{H}$ by a sequence of numbers $\mu_{\mathcal{H}}(m)$. Concretely, it follows that the best possible guarantee on the error, after seeing m examples, is $\Theta\left(\frac{\mu_{\mathcal{H}}(m)}{m}\right)$.

Comparing to binary classification, we should still strive for a better characterization: We would like to have a characterization of the sample complexity by a *single number* (i.e. some notion of dimension) rather than a sequence. Our analysis of the one inclusion algorithm naturally leads to a new notion of dimension, of somewhat different character than previously studied notions. We show that this notion have certain advantages comparing to other previously studied notions, and formulate a concrete combinatorial conjecture that, if true, would lead to a crisper characterization of the sample complexity.

Departing general theory, we turn our focus to investigate hypothesis classes that are used in practice, in light of the above results and the result of Daniely et al. (2011). We consider classes of multiclass linear classifiers that are learnt by several popular learning paradigms, including multiclass SVM with kernels (Crammer and Singer, 2001), structured output prediction (Collins, 2000, 2002; Lafferty et al., 2001; Taskar et al., 2003; Tsochantaridis et al., 2004), and others. Arguably, the two most natural questions in this context are: (i) is the ERM rule still sub-optimal even for such classes? and (ii) If yes, are there *efficient* optimal learnears for these classes?

Regarding the first question, we show that even though the sample complexity of these classes is upper bounded in terms of the dimension or the margin, there are sub-optimal ERMs whose sample complexity has additional multiplicative factor that depends on the number of labels. This settles in negative an open question due to Collins (2005). Regarding the second question above, as opposed to the one-inclusion algorithm, which is in general inefficient, for linear classes we derive computationally efficient learners (provided that the hypotheses can be evaluated efficiently), that enjoy optimal sample complexity.

Basic definitions: Let $\mathcal{X}$ be an instance space and $\mathcal{Y}$ a label space. To account for margin-based classifiers as well, it would be convenient to allow classifiers to return the label $\ominus$ that will stand for "don't know". A classifier (or hypothesis) is a mapping $h : \mathcal{X} \to (\mathcal{Y} \cup \{\ominus\})$. A hypothesis class is a set of classifiers, $\mathcal{H} \subset (\mathcal{Y} \cup \{\ominus\})^{\mathcal{X}}$. The error of a classifier with respect to a joint distribution over $\mathcal{X} \times \mathcal{Y}$ is the probability that $h(x) \neq y$. Throughout this paper, we mainly consider learning in the realizable case, which means that there is $h^* \in \mathcal{H}$ which has zero error (extensions to agnostic learning are discussed in section A). Therefore, we can focus on the marginal distribution $\mathcal{D}$ over $\mathcal{X}$ and denote the error of a classifier h with respect to the realizing classifier h^* as $\mathrm{Err}_{\mathcal{D},h^*}(h) := \Pr_{x\sim\mathcal{D}}\left(h(x) \neq h^*(x)\right)$.

A *learning algorithm* is a function $\mathcal{A}$ that receives a training set of m instances, $S \in \mathcal{X}^m$, together with their labels according to h^*. We denote the restriction of h^* to the instances in S by $h^*|_S$. The output of the algorithm $\mathcal{A}$, denoted $\mathcal{A}(S, h^*|_S)$ is a classifier. A learning algorithm is *proper* if it always outputs a hypothesis from $\mathcal{H}$. A learning algorithm is an *ERM learner* for the class $\mathcal{H}$ if, for any sample, it returns a function in $\mathcal{H}$ that minimizes the empirical error relative to any other function in $\mathcal{H}$. The *(PAC) sample complexity*

of a learning algorithm $\mathcal{A}$ is the function $m_{\mathcal{A},\mathcal{H}}$ defined as follows: For every $\epsilon, \delta > 0$, $m_{\mathcal{A},\mathcal{H}}(\epsilon, \delta)$ is the minimal integer such that for every $m \geq m_{\mathcal{A},\mathcal{H}}(\epsilon, \delta)$, every distribution $\mathcal{D}$ on $\mathcal{X}$, and every target hypothesis $h^* \in \mathcal{H}$, $\Pr_{S\sim\mathcal{D}^m}\left(\mathrm{Err}_{\mathcal{D},h^*}(\mathcal{A}(S, h^*|_S)) > \epsilon\right) \leq \delta$. Here and in subsequent definitions, we omit the subscript $\mathcal{H}$ when it is clear from context. If no integer satisfying the inequality above, define $m_{\mathcal{A}}(\epsilon, \delta) = \infty$. $\mathcal{H}$ is learnable with $\mathcal{A}$ if for all ϵ and δ the sample complexity is finite. The *(PAC) sample complexity* of a class $\mathcal{H}$ is $m_{\mathrm{PAC},\mathcal{H}}(\epsilon, \delta) = \inf_{\mathcal{A}} m_{\mathcal{A},\mathcal{H}}(\epsilon, \delta)$, where the infimum is taken over all learning algorithms. The *ERM sample complexity* (a.k.a. the *uniform convergence sample complexity*) of $\mathcal{H}$ is the sample complexity that can be guaranteed for any ERM learner. It is defined by $m_{\mathrm{ERM},\mathcal{H}}(\epsilon, \delta) = \sup_{\mathcal{A}\in\mathrm{ERM}} m^a_{\mathcal{A},\mathcal{H}}(\epsilon, \delta)$ where the supremum is taken over all ERM learners for $\mathcal{H}$. Clearly, we always have $m_{\mathrm{PAC}} \leq m_{\mathrm{ERM}}$.

We use $[m]$ to denote the set $\{1, \ldots, m\}$. We treat vectors as column vectors. We denote by $e_i \in \mathbb{R}^d$ the i'th vector in the standard basis of $\mathbb{R}^d$. We denote by B^d the closed unit ball in $\mathbb{R}^d$. We denote by $M_{d\times k}$ the space of real matrices with d rows and k columns. For a matrix $X \in M_{d\times k}$ and $i \in [k]$, we denote by $X^i \in \mathbb{R}^d$ the i'th column of X. Given a subset $A \subseteq \mathcal{X}$, we define $\mathcal{H}|_A = \{h|_A : h \in \mathcal{H}\}$.

2. No optimal learner can be proper

Our first result shows that, surprisingly, any learning algorithm with a close to optimal sample complexity must be improper.

Theorem 1 *For every $1 \leq d \leq \infty$ there exists a hypothesis class $\mathcal{H}_d$, with $2^d + 1$ labels such that:*

- *The PAC sample complexity of $\mathcal{H}_d$ is $O\left(\frac{\log(1/\delta)}{\epsilon}\right)$.*
- *The PAC sample complexity of any proper learning algorithm for $\mathcal{H}_d$ is $\Omega\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$.*
- *In particular, $\mathcal{H}_\infty$ is a learnable class that is not learnable by a proper algorithm.*

A detailed proof is given in the appendix, and here we sketch the main idea of the proof. Let $\mathcal{X}$ be some finite set and let $\mathcal{Y} = 2^{\mathcal{X}} \cup \{*\}$. For every $A \subseteq \mathcal{X}$ define $h_A : \mathcal{X} \to \mathcal{Y}$ by $h_A(x) = \begin{cases} A & x \in A \\ * & \text{otherwise} \end{cases}$. Consider the hypothesis class $\mathcal{H}_{\mathcal{X},\mathrm{Cantor}} = \{h_A \mid A \subset \mathcal{X}\}$. This class is due to Daniely et al. (2011) and we call it *the first Cantor class* due to the resemblance to the construction used for proving the famous theorem of Cantor from set theory (e.g., http://en.wikipedia.org/wiki/Cantor's_theorem). Daniely et al. (2011) employed this class to establish gaps between the sample complexity of different ERM learners. In particular, they have shown that there is an ERM learner with sample complexity $\leq \frac{\ln(1/\delta)}{\epsilon}$, while there are other ERMs whose sample complexity is $\Omega\left(\frac{|\mathcal{X}|+\ln(1/\delta)}{\epsilon}\right)$.

To show that no proper learner can be optimal, let $\mathcal{X}_d$ be a set consisting of d elements and define the following subclass of $\mathcal{H}_{\mathcal{X}_d,\mathrm{Cantor}}$: $\mathcal{H}_d = \left\{h_A \mid |A| = \left\lfloor \frac{d}{2} \right\rfloor\right\}$. Since $\mathcal{H}_d \subset \mathcal{H}_{\mathcal{X}_d,\mathrm{Cantor}}$, we can apply the "good" ERM learner described in Daniely et al. (2011) with respect to the class $\mathcal{H}_{\mathcal{X}_d,\mathrm{Cantor}}$ and obtain an algorithm for $\mathcal{H}_d$ whose sample complexity is $\leq \frac{\ln(1/\delta)}{\epsilon}$. Note that this algorithm is improper — it might output a hypothesis from

$\mathcal{H}_{\mathcal{X}_d,\text{Cantor}}$ which is not in $\mathcal{H}_d$. As we show, no proper algorithm is able to learn $\mathcal{H}_d$ using $o\left(\frac{d}{\epsilon}\right)$ examples. To understand the main point in the proof, suppose that an adversary chooses $h_A \in \mathcal{H}_d$ uniformly at random, and let the algorithm learn it, where the distribution on $\mathcal{X}_d$ is uniform on the complement of A, denoted A^c. Now, the error of every hypothesis $h_B \in \mathcal{H}_d$ is $\frac{|B \setminus A|}{d}$. Therefore, to return a hypothesis with small error, the algorithm must recover a set that is almost disjoint from A, and therefore should recover A. However, if it sees only $o(d)$ examples, all it knows is that some $o(d)$ elements in $\mathcal{X}$ do not belong to A. It is not hard to be convinced that with this little information, the probability that the algorithm will succeed is negligible.

3. An optimal learner for general classes

In this section we describe and analyze a generic optimal learning algorithm. We start with an algorithm for a transductive learning setting, in which the algorithm observes $m-1$ labeled examples and an additional unlabeled example, and it should output the label of the unlabeled example. Later, in Section 3.3 we show a generic reduction from the transductive setting to the usual inductive learning model (that is, the vanilla PAC model).

Formally, in the transductive model, the algorithm observes a set of m unlabeled examples, $S \in \mathcal{X}^m$, and then one of them is picked uniformly at random, $x \sim U(S)$. The algorithm observes the labels of all the examples but the chosen one, and should predict the label of the chosen example. That is, the input of the algorithm, $\mathcal{A}$, is the set $S \in \mathcal{X}^m$, and the restriction of some $h^* \in \mathcal{H}$ to $S \setminus x$, denoted $h^*|_{S \setminus x}$. The algorithm should output $y \in \mathcal{Y}$. The error rate of a transductive algorithm $\mathcal{A}$ is the function $\epsilon_{\mathcal{A},\mathcal{H}} : \mathbb{N} \to [0,1]$ defined as $\epsilon_{\mathcal{A},\mathcal{H}}(m) = \sup_{S \in \mathcal{X}^m, h^* \in \mathcal{H}} \left[\Pr_{x \sim U(S)} \left(\mathcal{A}(S, h^*|_{S \setminus x}) \neq h^*(x)\right)\right]$. The error rate of a class $\mathcal{H}$ in the transductive model is defined as $\epsilon_{\mathcal{H}}(m) = \inf_{\mathcal{A}} \epsilon_{\mathcal{A},\mathcal{H}}(m)$, where the infimum is over all transductive learning algorithms.

3.1. The one-inclusion algorithm

We next describe the one-inclusion transductive learning algorithm of Rubinstein et al. (2006). Let $S = \{x_1, \ldots, x_m\}$ be an unlabelled sample. For every $i \in [m]$ and $h \in \mathcal{H}|_S$, let $e_{i,h} \subset \mathcal{H}|_S$ be all the hypotheses in $\mathcal{H}|_S$ whose restriction to $S \setminus \{x_i\}$ equals to $h|_{S \setminus \{x_i\}}$. That is, $h' \in e_{i,h}$ iff for all $j \neq i$ we have $h'(x_j) = h(x_j)$. Note that if $h' \in e_{i,h}$ then $e_{i,h'} = e_{i,h}$.

Given $(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_m, y_m)$ let $h \in \mathcal{H}|_S$ be some hypothesis for which $h(x_j) = y_j$ for all $j \neq i$. We know that the target hypothesis can be any hypothesis in $e_{i,h}$. Therefore, we can think on the transductive algorithm as an algorithm that obtains some $e_{i,h}$ and should output one hypothesis from $e_{i,h}$. Clearly, if $|e_{i,h}| = 1$ we know that the target hypothesis is h. But, what should the algorithm do when $|e_{i,h}| > 1$?

The idea of the one-inclusion algorithm is to think on the collection $E = \{e_{i,h}\}_{i \in [m], \mathcal{H} \in \mathcal{H}|_S}$ as a collection of hyperedges of a hypergraph $G = (V, E)$. Recall that in a hypergraph, V is some set of vertices and each hyperedge $e \in E$ is some subset of V. In our case, the vertex set is $V = \mathcal{H}|_S$. This hypergraph is called the one-inclusion hypergraph. Note that if $|e| = 2$ for every $e \in E$ we obtain the usual definition of a graph. In such a case, an *orientation* of an undirected edge $e = \{v_1, v_2\}$ is picking one of the vertices (e.g. v_1) to be the "head" of the edge. Similarly, an orientation of a hyperedge is choosing one $v \in e$ to be the "head" of

the hyperedge. And, an orientation of the entire hypergraph is a function $f : E \to V$ such that for all $e \in E$ we have that $f(e) \in e$.

Getting back to our transductive learning task, it is easy to see that any (deterministic) transductive learning algorithm is equivalent to an orientation function $f : E \to V$ of the one-inclusion hypergraph. The error rate of such an algorithm, assuming the target function is $h^* \in \mathcal{H}|_S$, is

$$\Pr_{i \sim U([m])}[f(e_{i,h^*}) \neq h^*] = \frac{1}{m}\sum_{i=1}^{m} 1[f(e_{i,h^*}) \neq h^*] = \frac{|\{e \in E : h^* \in e \wedge f(e) \neq h^*\}|}{m} \ . \quad (1)$$

The quantity $|\{e \in E : h^* \in e \wedge f(e) \neq h^*\}|$ is called the *out-degree* of the vertex h^* and denoted $d^+(h^*)$. It follows that the error rate of an orientation f is $\max_{h^* \in \mathcal{H}|_S} \frac{d^+(h^*)}{m}$. It follows that the best deterministic transductive algorithm should find an orientation of the hypergraph that minimizes the maximal out degree. This leads to the one-inclusion algorithm.

Algorithm 1 Multiclass one inclusion algorithm for $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$

1: **Input:** unlabeled examples $S = (x_1, \ldots, x_m)$, labels $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m)$
2: Define the one-inclusion graph $G = (V, E)$ where $V = \mathcal{H}|_S$ and $E = \{e_{j,h}\}_{j \in [m], h \in V}$
3: Find orientation $f : E \to V$ that minimizes the maximal out-degree of G
4: Let $h \in V$ be s.t. $h(x_j) = y_j$ for all $j \neq i$, and let $\hat{h} = f(e_{i,h})$
5: **Output:** predict $\hat{h}(x_i)$

3.2. Analysis

The main result of this section is a new analysis of the one inclusion algorithm, showing its optimality in the transductive model, up to a constant factor of $1/2$. In the next subsection we deal with the PAC model.

To state our results, we need a few definitions. Let $G = (V, E)$ be a hypergraph. Throughout, we only consider hypergraphs for which E is an antichain (i.e., there are no $e_1, e_2 \in E$ such that e_1 is strictly contained in e_2). Given $U \subseteq V$, define the *induced* hypergraph, $G[U]$, as the hypergraph whose vertex set is U and whose edge set is all sets $e \subseteq U$ such that $e = U \cap e'$ for some $e' \in E$, $|e| \geq 2$, and e is maximal w.r.t. these conditions.

The *degree* of a vertex v in a hypergraph $G = (V, E)$ is the number of hyperedges, $e \in E$, such that $|e| \geq 2$ and $v \in e$. The *average degree* of G is $d(G) = \frac{1}{|V|}\sum_{v \in V} d(v)$. The *maximal average degree* of G is $\mathrm{md}(G) = \max_{U \subseteq V : |U| < \infty} d(G[U])$. For a hypothesis class $\mathcal{H}$ define

$$\mu_{\mathcal{H}}(m) = \max\{\mathrm{md}(G(\mathcal{H}|_S)) \mid S \in \mathcal{X}^m\} \ ,$$

where $G(\mathcal{H}|_S)$ is the one-inclusion hypergraph defined in Algorithm 1.

Theorem 2 *For every class $\mathcal{H}$, $\frac{1}{2}\frac{\mu_{\mathcal{H}}(m)}{m} \leq \epsilon_{\mathcal{H}}(m) \leq \frac{\mu_{\mathcal{H}}(m)}{m}$.*

Proof To prove the upper bound, recall that the one inclusion algorithm uses an orientation of the one-inclusion hypergraph that minimizes the maximal out-degree, and recall that in

(1) we have shown that the error rate of an orientation function is upper bounded by the maximal out-degree over m. Therefore, the proof of the upper bound of the theorem follows directly from the following lemma:

Lemma 3 *Let $G = (V, E)$ be a hypergraph with maximal average degree d. Then, there exists an orientation of G with maximal out-degree of at most d.*

The proof of the lemma is given in the appendix.

While the above proof of the upper bound is close in spirit to the arguments used by Haussler et al. (1988) and Rubinstein et al. (2006), the proof of the lower bound relies on a new argument. As opposed to Rubinstein et al. (2006) who lower bounded $\epsilon_{\mathcal{H}}(m)$ using the Natarajan dimension, we give a direct analysis.

Let $S \in \mathcal{X}^m$ be a set such that $\mathrm{md}(G(\mathcal{H}|_S)) = \mu_{\mathcal{H}}(m)$. For simplicity we assume that $|S| = m$ (i.e., S does not contain multiple elements). Since $\mathrm{md}(G(\mathcal{H}|_S)) = \mu_{\mathcal{H}}(m)$, there is finite $\mathcal{F} \subset \mathcal{G}$ with $d(G(\mathcal{F}|_S)) = \mu_{\mathcal{H}}(m)$. Consider the following scenario. Suppose that $h^* \in \mathcal{F}|_S$ is chosen uniformly at random, and in addition, a point $x \in S$ is also chosen uniformly at random. Now, suppose that a learner $\mathcal{A}$ is given the sample S with all points labelled by h^* except x that is unlabelled. It is enough to show that the probability that $\mathcal{A}$ errs is $\geq \frac{\mu_{\mathcal{H}}(m)}{2m}$.

Denote by U the event that x correspond to an edge in $G(\mathcal{F}|_S)$ coming out of h^*. Given U, the value of $h^*(x)$, given what the algorithm sees, is distributed uniformly in the set $\{h(x) \mid h \in \mathcal{F} \text{ and } h|_{S\setminus\{x\}} = h^*|_{S\setminus\{x\}}\}$. Since this set consists of at least two elements, given U, the algorithm errs with probability $\geq \frac{1}{2}$.

It is therefore enough to prove that $\Pr(U) \geq \frac{\mu_{\mathcal{H}}(m)}{m}$. Indeed, given h^*, the probability that x corresponds to an edge coming out of h^* is exactly the degree of h^* over m. Therefore, the probability that x corresponds to an edge coming out of a randomly chosen h^* is the average degree of $G(\mathcal{F}|_S)$ over m, i.e., $\frac{\mu_{\mathcal{H}}(m)}{m}$. ■

3.3. PAC optimality: from transductive to inductive learning

In the previous section we have analyzed the optimal error rate of learning in the transductive learning. We now turn to the inductive PAC model. By a simple reduction from inductive to transductive learning, we will show that a variant of the one-inclusion algorithm is essentially optimal in the PAC model.

First, any transductive algorithm $\mathcal{A}$ can be naturally interpreted as an inductive algorithm, which we denote by $\mathcal{A}^i$. Specifically, $\mathcal{A}^i$ returns, after seeing the sample $S = \{(x_i, y_i)\}_{i=1}^{m-1}$, the hypothesis $h : \mathcal{X} \to \mathcal{Y}$ such that $h(x)$ is the label $\mathcal{A}$ would have predicted for x after seeing the labelled sample S.

It holds that (see the appendix) the (worst case) expectation of the error of the hypothesis returned by $\mathcal{A}^i$ operating on m points sample, is the same, up to a factor of e to $\epsilon_{\mathcal{A}}(m)$. Using this fact and a simple amplification argument, it is not hard to show that a variant of the one-inclusion algorithm is essentially optimal in the PAC model.

Namely, we consider the algorithm $\overline{\mathcal{I}}$ that splits the sample into $2\log(1/\delta)$ parts, run the one inclusion algorithm on $\log(1/\delta)$ different parts to obtain $\log(1/\delta)$ candidate hypotheses, and finally chooses the best one, by validation on the remaining points. As the following

theorem (whose proof is given in the appendix) shows, $\overline{\mathcal{I}}$ is optimal up to a factor of $O\left(\log\left(\frac{1}{\delta}\right)\log\left(\frac{1}{\epsilon}\right)\right)$ in the PAC model, in the following sense:

Theorem 4 *For some $c > 0$, and every class $\mathcal{H}$, $m_{\overline{\mathcal{I}},\mathcal{H}}(\epsilon,\delta) \leq m_{\mathrm{PAC},\mathcal{H}}(c\epsilon,\delta)\cdot\frac{1}{c}\log(1/\delta)\log(1/\epsilon)$.*

4. Efficient optimal learning and gaps for linear classes

In this section we study the family of linear hypothesis classes. This family is widely used in practice and received a lot of attention in the literature—see for example Crammer and Singer (2001); Collins (2000, 2002); Lafferty et al. (2001); Taskar et al. (2003); Tsochantaridis et al. (2004). We show that, rather surprisingly, even for such simple classes, there can be gaps between the ERM sample complexity and the PAC sample complexity. This settles in negative an open question raised by Collins (2005). We also derive computationally efficient optimal learners for linear classes, based on the concept of compression schemes. This is in contrast to the one-inclusion algorithm from the previous section, which in general is inefficient. Due to the lack of space, most proofs are deferred to the appendix.

4.1. Linear hypothesis classes

We first define the various hypothesis classes of multiclass linear classifiers that we study. All of these classes depend on a class-specific feature mapping, $\Psi : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^d$. We will provide several examples of feature mappings that are widely used in practice.

4.1.1. Dimension based linear classifiers (denoted $\mathcal{H}_\Psi$)

For $w \in \mathbb{R}^d$ and $x \in \mathcal{X}$, define the multiclass predictor $h_w(x) = \mathrm{argmax}_{y\in\mathcal{Y}}\langle w, \Psi(x,y)\rangle$. In case of a tie, $h_w(x)$ is assumed to be the "don't know label", $\ominus$. The corresponding hypothesis class is defined as $\mathcal{H}_\Psi = \{h_w \mid w \in \mathbb{R}^d\}$.

Example 1 (multivector construction) *If the labels are unstructured, a canonical choice of Ψ is the so called* multivector construction. *Here, $\mathcal{Y} = [k]$, $\mathcal{X} = \mathbb{R}^d$ and $\Psi : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^{dk}$ is defined as follows: $\Psi(x,y)$ is the $d \times k$ matrix whose y'th column is x, while the rest are 0. In this case, every classifier corresponds to a matrix W, and the prediction on an instance $x \in \mathbb{R}^d$ is the index of the column that maximizes the inner product with x.*

4.1.2. Large margin linear classifiers (denoted $\mathcal{H}_{\Psi,R}$)

The second kind of hypothesis class induced by Ψ is margin based. Here, we assume that the range of Ψ is contained in the unit ball of $\mathbb{R}^d$. Every vector $w \in \mathbb{R}^d$ defines a function $h_w : \mathcal{X} \to (\mathcal{Y}\cup\{\ominus\})$ by

$$\forall x \in \mathcal{X},\quad h_w(x) = \begin{cases} y & \text{if} \langle w, \Psi(x,y) - \Psi(x,y')\rangle \geq 1 \text{ for every } y' \neq y \\ \ominus & \text{if no such } y \text{ exists} \end{cases}$$

The class of *linear classifiers of complexity $R > 0$ induced by Ψ* is $\mathcal{H}_{\Psi,R} = \left\{h_w \mid \|w\|^2 \leq R\right\}$.

Example 2 (multivector construction with margin) *The margin based analogue to example 1 is defined similarly. This class is the class that is learnt by multiclass SVM.*

4.1.3. The classes $\mathcal{H}_{d,t,q}$ and $\mathcal{H}_{d,t,q,R}$ for structured output prediction

Next we consider an embedding Ψ that is specialized and used in classification tasks where the number of possible labels is exponentially large, but the labels are structured (e.g. Taskar et al. (2003)). For example, in speech recognition, the label space might me the collection of all sequences of ≤ 20 English words.

To motivate the definition, consider the case that we are to recognize a t-letter word appearing in an image. Let q be the size of the alphabet. The set of possible labels is naturally associated with $[q]^t$. A popular method to tackle this task (see for example Taskar et al. (2003)) is the following: The image is broken into t parts, each of which contains a single letter. Each letter is represented as a vector in $\mathbb{R}^d$. Thus, each image is represented as a matrix in $M_{d\times t}$. To devise a linear hypothesis class to this problem, we should specify a mapping $\Psi : M_{d\times t} \times [q]^t \to \mathbb{R}^n$ for some n. Given $X \in M_{d\times t}$ and $y \in [q]^t$, $\Psi(X, y)$ will be a pair $(\Psi_1(X, y), \Psi_2(X, y))$. The mapping Ψ_1 allows the classifiers to take into account the shape of the letters appearing in the different t parts the word was broken into. The mapping Ψ_2 allows the classifiers to take into account the structure of the language (e.g. the fact that the letter "u" usually appears after the letter "q"). $\Psi_1(X, y) \in M_{d\times q}$ is the matrix whose j'th column is the sum of the columns X^i with $y_i = j$ (in other words, the j'th column is the sum of the letters in the image that are predicted to be j by y). $\Psi_2(X, y) \in M_{q,q}$ will be the matrix with 1 in the (i, j) entry if the letter j appears after the letter i somewhere in the word y, and 0 in all other entries. Even though the number of labels is exponential in t, this class (in the realizable case) can be learnt in time polynomial in d, t and q (see Collins (2005)).

We will show gaps in the performance of different ERMs for the class $\mathcal{H}_\Psi$. If fact, we will prove a slightly stronger result. We will consider the class $\mathcal{H}_{\Psi_1}$, that we will denote by $\mathcal{H}_{d,t,q}$. It is easy to see that $\mathcal{H}_{\Psi_1}$ can be realized by $\mathcal{H}_\Psi$. Therefore, any lower bound for $\mathcal{H}_{\Psi_1}$ automatically lower bounds also $\mathcal{H}_\Psi$. As for upper bounds, as long as $q = O(d)$, the upper bounds we show are the same for $\mathcal{H}_\Psi$ and $\mathcal{H}_{\Psi_1}$. To summarize, the gaps we show for $\mathcal{H}_{\Psi_1}$ automatically (as long as $q = O(d)$) hold for $\mathcal{H}_\Psi$ as well.

Finally, we define a margin-based analogue to $\mathcal{H}_{d,t,q}$. The instance space is $(B^d)^t$, and we treat each $X \in (B^d)^t$ as a matrix with t columns, each of which is a vector in B^d. The labels are $[q]^t$. Define $\Psi : (B^d)^t \times [q]^k \to M_{d\times q}$ as follows: for $X \in (B^d)^t$ and $y \in [q]^t$, $\Psi(X, y)$ is the matrix whose j'th column is $\frac{1}{q}$ of the average of all columns X^i such that $y_i = j$. Note that the range of Ψ is contained in the unit ball. For $R > 0$, define $\mathcal{H}_{d,t,q,R} := \mathcal{H}_{\Psi,R}$.

4.2. Results

We begin with linear predictors without margin. The first part of the following theorem asserts that for every $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ there is some algorithm that learns $\mathcal{H}_\Psi$ with sample complexity $O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon}\right)$. The second part of the theorem shows that in several cases (i.e., for some Ψ's), this algorithm outperforms other ERMs, by a factor of $\log(|\mathcal{Y}|)$.

Theorem 5

- *For every* $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$*, the PAC sample complexity of* $\mathcal{H}_\Psi$ *is* $O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon}\right)$*, and is achievable by a new efficient*[1] *compression scheme.*
- *For every* $\mathcal{Y}$ *and* $d > 0$*, there is some* $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ *for which the ERM sample complexity of* $\mathcal{H}_\Psi$ *is* $\Omega\left(\frac{d\log(|\mathcal{Y}|)+\log(1/\delta)}{\epsilon}\right)$.

To put the result in the relevant context, it was known (e.g. Daniely et al. (2011)) that the sample complexity of every ERM for this class is $O\left(\frac{d\log(|\mathcal{Y}|)\log(1/\epsilon)+\log(1/\delta)}{\epsilon}\right)$. In particular, the second part of the theorem is tight, up to the logarithmic dependence over $\frac{1}{\epsilon}$. However, it was not known whether the factor of $\log(|\mathcal{Y}|)$ for general ERM is necessary. The second part of the theorem shows that this factor is indeed necessary.

As to the tightness of the first part, for certain embeddings, including the multivector construction (example 1), a lower bound of $\Omega\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$ is known for every algorithm. Hence, the first part of the theorem is also tight up to the logarithmic dependence over $\frac{1}{\epsilon}$.

Our second theorem for linear classes is analogous to theorem 5 for margin based classes. The first part shows that for every $\Psi : \mathcal{X} \times \mathcal{Y} \to B^d$ there is some algorithm that learns $\mathcal{H}_{\Psi,R}$ with sample complexity $O\left(\frac{R\log(1/\epsilon)+\log(1/\delta)}{\epsilon}\right)$. The second part of the theorem shows that in several cases, the above algorithm outperforms other ERMs, by a factor of $\log(|\mathcal{Y}|)$.

Theorem 6

- *For every* $\Psi : \mathcal{X} \times \mathcal{Y} \to B^d$ *and* $R > 0$*, the PAC sample complexity of* $\mathcal{H}_{\Psi,R}$ *is* $O\left(\frac{R\log(1/\epsilon)+\log(1/\delta)}{\epsilon}\right)$.
- *For every* $\mathcal{Y}$ *and* $R > 0$*, there is some* $\Psi : \mathcal{X} \times \mathcal{Y} \to B^d$ *for*[2] *which the ERM sample complexity of* $\mathcal{H}_{\Psi,R}$ *is* $\Omega\left(\frac{R\log(|\mathcal{Y}|)+\log(1/\delta)}{\epsilon}\right)$.

The first part of the theorem is not new. An algorithm that achieves this bound is the perceptron. It was known (e.g. Collins (2005)) that the sample complexity of every ERM for this class is $O\left(\frac{R\log(|\mathcal{Y}|/\epsilon)+\log(1/\delta)}{\epsilon}\right)$. In particular, the second part of the theorem is tight, up to the logarithmic dependence over $\frac{1}{\epsilon}$. However, it was not known whether the gap is real: In (Collins, 2005), it was left as an open question to show whether the perceptron's bound holds for every ERM. The second part of the theorem answers this open question in negative. Regarding lower bounds, as in the case of $\mathcal{H}_\Psi$, for certain embeddings, including the multivector construction with margin (example 1), a lower bound of $\Omega\left(\frac{R+\log(1/\delta)}{\epsilon}\right)$ is known and valid for every learning algorithm. In particular, the first part of the theorem is also tight up to the logarithmic dependence over $\frac{1}{\epsilon}$.

An additional result that we report on shows that, for every $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, the Natarajan dimension of $\mathcal{H}_\Psi$ is at most d (the definition of the Natarajan dimension is recalled in the appendix). This strengthens the result of (Daniely et al., 2011) who showed that it is bounded by $O(d\log(d))$. It is known (e.g. Daniely et al. (2012)) that for the multivector construction (example 1), in which the dimension of the range of Ψ is dk, the

1. Assuming we have an appropriate separation oracle.
2. Here, d can be taken to be polynomial in R and $\log(|\mathcal{Y}|)$.

Natarajan dimension is lower bounded by $(d-1)(k-1)$. Therefore, the theorem is tight up to a factor of $1+o(1)$.

Theorem 7 *For every $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, $\mathrm{Ndim}(\mathcal{H}_\Psi) \leq d$.*

Next, we give analogs to theorems 5 and 6 for the structured output classes $\mathcal{H}_{d,k}$ and $\mathcal{H}_{d,k,R}$. These theorems show that the phenomenon of gaps between different ERMs, as reported in (Daniely et al., 2011), happens also in hypothesis classes that are used in practice.

Theorem 8

- *For every $d,t,q>0$, the PAC sample complexity of $\mathcal{H}_{d,t,q}$ is $O\left(\frac{dq\log(1/\epsilon)+\log(1/\delta)}{\epsilon}\right)$.*
- *For every $d,t,q>0$ the ERM sample complexity of $\mathcal{H}_{d,t,q}$ is $\Omega\left(\frac{dq\log(t)+\log(1/\delta)}{\epsilon}\right)$.*

Theorem 9

- *For every $d,t,q,R>0$, the PAC sample complexity of $\mathcal{H}_{d,t,q,R}$ is $O\left(\frac{R\log(1/\epsilon)+\log(1/\delta)}{\epsilon}\right)$.*
- *For every $t,q,R>0$ and $d \geq (t+1)R$, the ERM sample complexity of $\mathcal{H}_{d,t,q,R}$ is $\Omega\left(\frac{R\log(t)+\log(1/\delta)}{\epsilon}\right)$.*

The first parts of theorems 8 and 9 are direct consequences of theorems 5 and 6. These results are also tight up to the logarithmic dependence over $\frac{1}{\epsilon}$. The second parts of the theorems do not follow from theorems 5 and 6. Regarding the tightness of the second part, the best known upper bounds for the ERM sample complexity of $\mathcal{H}_{d,t,q}$ and $\mathcal{H}_{d,t,q,R}$ are $O\left(\frac{dqt\log(\frac{1}{\epsilon})+\log(1/\delta)}{\epsilon}\right)$ and $O\left(\frac{Rt\log(\frac{1}{\epsilon})+\log(1/\delta)}{\epsilon}\right)$ respectively. Closing the gap between these upper bounds and the lower bounds of theorems 8 and 9 is left as an open question.

4.3. The compression-based optimal learners

Each of the theorems 5, 6, 8 and 9 are composed of two statements. The first claims that some algorithm have a certain sample complexity, while the second claims that there exists an ERM whose sample complexity is worse than the sample complexity of the algorithm from the first part. As explained in this subsection, the first parts of these theorems are established by devising (efficient) compression schemes. In the next subsection we will elaborate on the proof of the second parts (the lower bounds on specific ERMs). Unfortunately, due to lack of space, we must be very brief.

We now show that for linear classes, it is possible to derive optimal learners which are also computationally efficient. For the case of margin-based classes, this result is not new — an efficient algorithm based on the multiclass perceptron has been proposed in Collins (2002). For completeness, we briefly survey this approach in the appendix. For dimension based linear classes, we give a new efficient algorithm.

The algorithm relies on compression based generalization bounds (see Theorem 18 in the appendix). Based on this theorem, it is enough to show that for every $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, $\mathcal{H}_\Psi$ has a compression scheme of size d. We consider the following compression scheme. Given a realizable sample $(x_1,y_1),\ldots,(x_m,y_m)$, let $Z \subseteq \mathbb{R}^d$ be the set of all vectors of the form $\Psi(x_i,y_i)-\Psi(x_i,y)$ for $y \neq y_i$. Let w be the vector of minimal norm in the convex hull of Z, $\mathrm{conv}(Z)$. Note that by the convexity of $\mathrm{conv}(Z)$, w is unique and can be found efficiently using a convex optimization procedure. Represent w as a convex combination of

d vectors from Z. This is possible since, by claim 1 below, $0 \notin \text{conv}(Z)$. Therefore, w is in the boundary of the polytope $\text{conv}(Z)$. Thus, w lies in a convex polytope whose dimension is $\leq d-1$, and is the convex hull of points from Z. Therefore, by Caratheodory's theorem (and using its efficient constructive proof), w is a convex combination of $\leq d$ points from Z. Output the examples in the sample that correspond to the vectors in the above convex combination. If there are less than d such examples, arbitrarily output more examples.

The *De-Compression* procedure is as follows. Given $(x_1, y_1), \ldots, (x_d, y_d)$, let $Z' \subseteq \mathbb{R}^d$ be the set of all vectors of the form $\Psi(x_i, y_i) - \Psi(x_i, y)$ for $y \neq y_i$. Then, output the minimal norm vector $w \in \text{conv}(Z')$.

In the appendix (Section D.5) we show that this is indeed a valid compression scheme, that is, if we start with a realizable sample $(x_1, y_1), \ldots, (x_m, y_m)$, compress it, and then de-compress it, we are left with a hypothesis that makes no errors on the original sample.

4.4. Lower bounds for specific ERMs

Next, we explain how we prove the second parts of theorems 5, 6, 8 and 9. For theorems 5 and 6, the idea is to start with the first Cantor class (introduced in section 2) and by a geometric construction, realize it by a linear class. This realization enables us to extend the "bad ERM" for the first Cantor class, to a "bad ERM" for that linear class. The idea behind the lower bounds of theorems 8 and 9 is similar, but technically more involved. Instead of the first Cantor class, we introduce a new discrete class, *the second Cantor class*, which may be of independent interest. This class, which can be viewed as a dual to the first Cantor class, is defined as follows. Let $\tilde{\mathcal{Y}}$ be some non-empty finite set. Let $\mathcal{X} = 2^{\tilde{\mathcal{Y}}}$ and let $\mathcal{Y} = \tilde{\mathcal{Y}} \cup \{*\}$. For every $y \in \tilde{\mathcal{Y}}$ define a function $h_y : \mathcal{X} \to \mathcal{Y}$ by $h_y(A) = \begin{cases} y & y \in A \\ * & \text{otherwise} \end{cases}$. Also, let $h_* : \mathcal{X} \to \mathcal{Y}$ be the constant function $*$. Finally, let $\mathcal{H}_{\mathcal{Y},\text{Cantor}} = \{h_y \mid y \in \mathcal{Y}\}$. In section C we show that the graph dimension (see a definition in the appendix) of $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$ is $\Theta(\log(|\mathcal{Y}|))$. The analysis of the graph dimension of this class is more involved than the first Cantor class: by a probabilistic argument, we show that a random choice of $\Omega\left(\log(|\mathcal{Y}|)\right)$ points from $\mathcal{X}$ is shattered with positive probability. We show also (see section C) that the PAC sample complexity of $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$ is $\leq \frac{\log(1/\delta)}{\epsilon}$. Since the graph dimension characterizes the ERM sample complexity (see the appendix), this class provides another example of a hypothesis class with gaps between ERM and PAC learnability.

5. A new dimension

Consider again the question of characterizing the sample complexity of learning a class $\mathcal{H}$. Theorem 2 shows that the sample complexity of a class $\mathcal{H}$ is characterized by the sequence of densities $\mu_{\mathcal{H}}(m)$. A better characterization would be a notion of dimension that assigns a single number, $\dim(\mathcal{H})$, that controls the growth of $\mu_{\mathcal{H}}(m)$, and consequently, the sample complexity of learning $\mathcal{H}$. To reach a plausible generalization, let us return for a moment to binary classification, and examine the relationships between the VC dimension and the sequence $\mu_{\mathcal{H}}(m)$. It is not hard to see that

- The VC dimension of $\mathcal{H}$ is the maximal number d such that $\mu_{\mathcal{H}}(d) = d$.

Moreover, a beautiful result of Haussler et al. (1988) shows that

- If $|\mathcal{Y}| = 2$, then $\text{VCdim}(\mathcal{H}) \leq \mu_{\mathcal{H}}(m) \leq 2\,\text{VCdim}(\mathcal{H})$ for every $m \geq \text{VCdim}(\mathcal{H})$.

These definition and theorem naturally suggest a generalization to multiclass classification:

Definition 10 *The* dimension, $\dim(\mathcal{H})$, *of the class* $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ *is the maximal number* d *such that* $\mu_{\mathcal{H}}(d) = d$.

Conjecture 11 *There exists a constant* $C > 0$ *such that for every* $\mathcal{H}$ *and* $m \geq \dim(\mathcal{H})$, $\dim(\mathcal{H}) \leq \mu_{\mathcal{H}}(m) \leq C \cdot \dim(\mathcal{H})$. *Consequently, by Theorem 2,*

$$\epsilon_{\mathcal{H}}(m) = \Theta\left(\frac{\dim(\mathcal{H})}{m}\right) \quad \textit{and} \quad \Omega\left(\frac{\dim(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)}{\epsilon}\right) \leq m_{\mathcal{H}}(\epsilon, \delta) \leq O\left(\frac{\dim(\mathcal{H})\log\left(\frac{1}{\delta}\right)}{\epsilon}\right)$$

For concreteness, we give an equivalent definition of $\dim(\mathcal{H})$ and a formulation of conjecture 11 that are somewhat simpler, and do not involve the sequence $\mu_{\mathcal{H}}(m)$

Definition 12 *Let* $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$. *We say that* $A \subset \mathcal{X}$ *is* shattered *by* $\mathcal{H}$ *is there exists a finite* $\mathcal{F} \subset \mathcal{H}$ *such that for every* $x \in A$ *and* $f \in \mathcal{F}$ *there is* $g \in \mathcal{F}$ *such that* $g(x) \neq f(x)$ *and* $g|_{A\setminus\{x\}} = f|_{A\setminus\{x\}}$. *The* dimension *of* $\mathcal{H}$ *is the maximal cardinality of a shattered set.*

Recall that the *degree* (w.r.t. $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$) of $f \in \mathcal{H}$ is the number of points $x \in \mathcal{X}$ for which there exists $g \in \mathcal{H}$ that disagree with f only on x. We denote the *average degree* of $\mathcal{H}$ by $d(\mathcal{H})$.

Conjecture 13 *There exists* $C > 0$ *such that for every finite* $\mathcal{H}$, $d(\mathcal{H}) \leq C \cdot \dim(\mathcal{H})$.

By combination of theorems 2 and Rubinstein et al. (2006), a weaker version of conjecture 11 is true. Namely, that for some absolute constant $C > 0$

$$\dim(\mathcal{H}) \leq \mu_{\mathcal{H}}(m) \leq C \cdot \log(|\mathcal{Y}|) \cdot \dim(\mathcal{H}) \ . \tag{2}$$

In addition, it is not hard to see that the new dimension is bounded between the Natarajan and Graph dimensions, $\text{Ndim}(\mathcal{H}) \leq \dim(\mathcal{H}) \leq \text{Gdim}(\mathcal{H})$. For the purpose of characterizing the sample complexity, this inequality is appealing for two reasons. First, it is known (Daniely et al., 2011) that the graph dimension does not characterize the sample complexity, since it can be substantially larger than the sample complexity in several cases. Therefore, any notion of dimension that do characterize the sample complexity must be upper bounded by the graph dimension. As for the Natarajan dimension, it is known to lower bound the sample complexity. By Theorem 2 and equation (2), the new dimension also lower bounds the sample complexity. Therefore, the left inequality shows that the new dimension always provides a lower bound that is at least as good as the Natarajan dimension's lower bound.

References

N. Alon and J. H. Spencer. *The Probabilistic Method*. Wiley-Interscience, second edition, 2000.

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.

M. Collins. Discriminative reranking for natural language parsing. In *Machine Learning*, 2000.

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.

Michael Collins. Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods. In *New developments in parsing technology*, pages 19–55. Springer, 2005.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz. Multiclass learnability and the erm principle. In *COLT*, 2011.

A. Daniely, S. Sabato, and S. Shalev-Shwartz. multiclass learning approaches: A theoretical comparision with implications. In *NIPS*, 2012.

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.

David Haussler, Nick Littlestone, and Manfred K. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. In *FOCS*, pages 100–109, October 1988.

J. Keshet, S. Shalev-Shwartz, Y. Singer, and D. Chazan. Phoneme alignment based on discriminative learning. In *Interspeech*, 2005.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, pages 282–289, 2001.

N. Littlestone and M. Warmuth. Relating data compression and learnability. Unpublished manuscript, November 1986.

J. Matousek. *Lectures on discrete geometry*, volume 212. Springer, 2002.

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, J. Keshet, and Y. Singer. Learning to align polyphonic music. In *Proceedings of the 5th International Conference on Music Information Retrieval*, 2004.

Hans Ulrich Simon and Balázs Szörényi. One-inclusion hypergraph density revisited. *Information Processing Letters*, 110(8):341–344, 2010.

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

A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(5):854–869, 2007.

I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.

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.

Appendix A. Agnostic learning and further directions

In this work we focused on learning in the realizable setting. For general hypothesis classes, it is left as an open question to find an optimal algorithm for the agnostic setting. However, for linear classes, our upper bounds are attained by compression schemes. Therefore, as indicated by Theorem 18, our results can be extended to the agnostic setting, yielding algorithms for $\mathcal{H}_\Psi$ and $\mathcal{H}_{\Psi,R}$ whose sample complexity is $O\left(\frac{d\log(d/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$ and $O\left(\frac{R\log(R/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$ respectively. We note that these upper bounds are optimal, up to the factors of $\log(d/\epsilon)$ and $\log(R/\epsilon)$. Our lower bounds clearly hold for agnostic learning (this is true for any lower bound on the realizable case). Yet, we would be excited to see better lower bounds for the agnostic setting. Specifically, are there classes $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ of Natarajan dimension d with ERMs whose agnostic sample complexity is $\Omega\left(\frac{d\log(|\mathcal{Y}|)}{\epsilon^2}\right)$?

Except extensions to the agnostic settings, the current work suggests several more directions for further research. First, it would be very interesting to go beyond multiclass classification, and to devise generic optimal algorithms for other families of learning problems. Second, as noted before, naive implementation of the one-inclusion algorithm is prohibitively inefficient. Yet, we still believe that the ideas behind the one-inclusion algorithm might lead to better efficient algorithms. In particular, it might be possible to derive efficient algorithms based on the principles behind the one-inclusion algorithm, and maybe even give an efficient implementation of the one-inclusion algorithm for concrete hypothesis classes.

Appendix B. Background

B.1. The Natarajan and Graph Dimensions

We recall two of the main generalizations of the VC dimension to multiclass hypothesis classes.

Definition 14 (Graph dimension) *Let $\mathcal{H} \subseteq (\mathcal{Y} \cup \{\ominus\})^{\mathcal{X}}$ be a hypothesis class. We say that $A \subseteq \mathcal{X}$ is* G-shattered *if there exists $h : A \to \mathcal{Y}$ such that for every $B \subseteq A$ there is $h' \in \mathcal{H}$ with $h(A) \subset \mathcal{Y}$ for which*

$$\forall x \in B,\ h'(x) = h(x) \text{ while } \forall x \in A \setminus B,\ h'(x) \neq h(x) \ .$$

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

As the following theorem shows, the graph dimension essentially characterizes the ERM sample complexity.

Theorem 15 (Daniely et al. (2011)) *For every hypothesis class $\mathcal{H}$ with graph dimension d,*

$$\Omega\left(\frac{d+\log(1/\delta)}{\epsilon}\right) \leq m_{\mathrm{ERM}}(\epsilon,\delta) \leq O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon}\right) \ .$$

Definition 16 (Natarajan dimension) *Let $\mathcal{H} \subseteq (\mathcal{Y} \cup \{\ominus\})^{\mathcal{X}}$ be a hypothesis class. We say that $A \subseteq \mathcal{X}$ is* N-shattered *if there exist $h_1, h_2 : A \to \mathcal{Y}$ such that $\forall x \in A,\ h_1(x) \neq h_2(x)$ and for every $B \subseteq A$ there is $h \in \mathcal{H}$ for which*

$$\forall x \in B,\ h(x) = h_1(x) \text{ while } \forall x \in A \setminus B,\ h(x) = h_2(x) \ .$$

The Natarajan dimension *of $\mathcal{H}$, denoted* $\mathrm{Ndim}(\mathcal{H})$*, is the maximal cardinality of an N-shattered set.*

Theorem 17 (essentially Natarajan (1989)) *For every hypothesis class $\mathcal{H} \subset (\mathcal{Y} \cup \{\ominus\})^{\mathcal{X}}$ with Natarajan dimension d,*

$$\Omega\left(\frac{d+\log(1/\delta)}{\epsilon}\right) \leq m_{\mathrm{PAC}}(\epsilon,\delta) \leq O\left(\frac{d\log(|\mathcal{Y}|)\log(1/\epsilon)+\log(1/\delta)}{\epsilon}\right) \ .$$

We note that the upper bound in the last theorem follows from theorem 15 and the fact that (see Ben-David et al. (1995)) for every hypothesis class $\mathcal{H}$,

$$\mathrm{Gdim}(\mathcal{H}) \leq 5\log(|\mathcal{Y}|)\,\mathrm{Ndim}(\mathcal{H}) \ . \tag{3}$$

We also note that (Daniely et al., 2011) conjectured that the logarithmic factor of $|\mathcal{Y}|$ in Theorem 17 can be eliminated (maybe with the expense of poly-logarithmic factors of $\frac{1}{\epsilon}, \frac{1}{\delta}$ and $\mathrm{Ndim}(\mathcal{H})$).

B.2. Compression Schemes

A *compression scheme* of size d for a class $\mathcal{H}$ is a pair of functions:

$$\text{Com} : \cup_{m=d}^{\infty}(\mathcal{X}\times\mathcal{Y})^m \to (\mathcal{X}\times\mathcal{Y})^d \text{ and } \text{DeCom} : (\mathcal{X}\times\mathcal{Y})^d \to \mathcal{Y}^{\mathcal{X}} \ ,$$

with the property that for every realizable sample

$$S = (x_1, y_1), \ldots, (x_m, y_m)$$

it holds that, if $h = \text{DeCom} \circ \text{Com}(S)$ then

$$\forall 1 \leq i \leq m, \ \ y_i = h(x_i) \ .$$

Each compression scheme yields a learning algorithm, namely, $\text{DeCom} \circ \text{Com}$. It is known that the sample complexity of this algorithm is upper bounded by the size of the compression scheme. Precisely, we have:

Theorem 18 (Littlestone and Warmuth (1986)) *Suppose that there exists a compression scheme of size d for a class $\mathcal{H}$. Then:*

- *The PAC sample complexity of $\mathcal{H}$ is upper bounded by $O\left(\frac{d\log(1/\epsilon)+\frac{1}{\delta}}{\epsilon}\right)$*
- *The agnostic PAC sample complexity of $\mathcal{H}$ is upper bounded by $O\left(\frac{d\log(d/\epsilon)+\frac{1}{\delta}}{\epsilon^2}\right)$*

Appendix C. The Cantor classes

C.1. The first Cantor class

Let $\mathcal{X}$ be some finite set and let $\mathcal{Y} = 2^{\mathcal{X}} \cup \{*\}$. For every $A \subseteq \mathcal{X}$ define $h_A : \mathcal{X} \to \mathcal{Y}$ by

$$h_A(x) = \begin{cases} A & x \in A \\ * & \text{otherwise} \end{cases} \ .$$

Finally, let

$$\mathcal{H}_{\mathcal{X},\text{Cantor}} = \{h_A \mid A \subset \mathcal{X}\} \ .$$

Lemma 19 (Daniely et al. (2011))

- *The graph dimension of $\mathcal{H}_{\mathcal{X},\text{Cantor}}$ is $|\mathcal{X}|$. Therefore, the ERM sample complexity of $\mathcal{H}_{\mathcal{X},\text{Cantor}}$ is $\Omega\left(\frac{|\mathcal{X}|+\log(1/\delta)}{\epsilon}\right)$.*
- *The Natarajan dimension of $\mathcal{H}_{\mathcal{X},\text{Cantor}}$ is 1. Furthermore, the PAC sample complexity of $\mathcal{H}_{\mathcal{X},\text{Cantor}}$ is $O\left(\frac{\log(1/\delta)}{\epsilon}\right)$.*

Proof For the first part, it is not hard to see that the function $f_\emptyset$ witnesses the G-shattering of $\mathcal{X}$. The second part follows directly from Lemma 20, given below. ■

Lemma 20 (essentially Daniely et al. (2011)) *Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class with the following property: There is a label $* \in \mathcal{Y}$ such that, for every $f \in \mathcal{H}$ and $x \in \mathcal{X}$, either $f(x) = *$ or f is the only function in $\mathcal{H}$ whose value at x is $f(x)$. Then,*

- *The PAC sample complexity of $\mathcal{H}$ is $\leq \frac{\log(1/\delta)}{\epsilon}$.*
- $\mathrm{Ndim}(\mathcal{H}) \leq 1$.

Proof We first prove the second part. Assume on the way of contradiction that $\mathrm{Ndim}(\mathcal{H}) > 1$. Let $\{x_1, x_2\} \subseteq \mathcal{X}$ be an N-shattered set of cardinality 2 and let f_1, f_2 be two functions that witness the shattering. Since $f_1(x_1) \neq f_2(x_1)$, at least one of $f_1(x_1), f_2(x_1)$ is different from $*$. W.l.o.g, assume that $f_1(x_1) \neq *$. Now, by the definition of N-shattering, there is a function $f \in \mathcal{H}_{\mathcal{Y},\mathrm{Cantor}}$ such that $f(x_1) = f_1(x_1)$ and $f(x_2) = f_2(x_2) \neq f_1(x_2)$. However, the only function in $\mathcal{H}$ satisfying $f(x_1) = f_1(x_1)$ is f_1. A contradiction.

We proceed to the first part. Assume w.l.o.g. the the function $f_* \equiv *$ is in $\mathcal{H}$. Consider the following algorithm. Given a (realizable) sample

$$(x_1, y_1), \ldots, (x_m, y_m),$$

if $y_i = *$ for every i then return the function f_*. Otherwise, return the hypothesis $h \in \mathcal{H}$, that is consistent with the sample. Note the the existence of a consistent hypothesis is guaranteed, as the sample is realizable. This consistent hypothesis is also unique: if $y_i \neq *$ then, by the assumption on $\mathcal{H}$, there is at most one function $f \in \mathcal{H}$ for which $h(x_i) = y_i$.

This algorithm is an ERM with the following property: For every learnt hypothesis and underlying distribution, the algorithm might return only one out of two functions – either f_* or the learnt hypothesis. We claim that the sample complexity of such an ERM must be $\leq \frac{\log(1/\delta)}{\epsilon}$. Indeed such an algorithm returns a hypothesis with error $\geq \epsilon$ only if:

- $\mathrm{Err}(f_*) \geq \epsilon$.
- For every $i \in [m]$, $y_i = *$.

However, if $\mathrm{Err}(f_*) \geq \epsilon$, the probability that $y_i = *$ is $\leq 1 - \epsilon$. Therefore, the probability of the the second condition is $\leq (1-\epsilon)^m \leq e^{-m\epsilon}$, which is $\leq \delta$ if $m \geq \frac{\log(1/\delta)}{\epsilon}$. ■

C.2. The second Cantor class

Let $\tilde{\mathcal{Y}}$ be some non-empty finite set. Let $\mathcal{X} = 2^{\tilde{\mathcal{Y}}}$ and let $\mathcal{Y} = \tilde{\mathcal{Y}} \cup \{*\}$. For every $y \in \tilde{\mathcal{Y}}$ define a function $h_y : \mathcal{X} \to \mathcal{Y}$ by

$$h_y(A) = \begin{cases} y & y \in A \\ * & \text{otherwise} \end{cases}.$$

Also, let $h_* : \mathcal{X} \to \mathcal{Y}$ be the constant function $*$. Finally, let $\mathcal{H}_{\mathcal{Y},\mathrm{Cantor}} = \{h_y \mid y \in \mathcal{Y}\}$.

Lemma 21

- *The graph dimension of* $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$ *is* $\Theta\left(\log\left(|\mathcal{Y}|\right)\right)$. *Therefore, the ERM sample complexity of* $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$ *is* $\Omega\left(\frac{\log(|\mathcal{Y}|)+\log(1/\delta)}{\epsilon}\right)$.
- *The Natarajan dimension of* $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$ *is* 1. *Furthermore, the PAC sample complexity of* $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$ *is* $O\left(\frac{\log(1/\delta)}{\epsilon}\right)$.

Proof The second part of the lemma follows from Lemma 20. We proceed to the first part. First, by equation (3) and the second part, $\text{Gdim}(\mathcal{H}_{\mathcal{Y},\text{Cantor}}) \leq 5\log(|\mathcal{Y}|)$. It remains to show that $\text{Gdim}(\mathcal{H}_{\mathcal{Y},\text{Cantor}}) \geq \Omega\left(\log(|\mathcal{Y}|)\right)$. To do so, we must show that there are $r = \Omega\left(\log(|\mathcal{Y}|)\right)$ sets $\mathcal{A} = \{A_1, \ldots, A_r\} \subseteq \mathcal{X}$ such that $\mathcal{A}$ is G-shattered by $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$. To do so, we will use the probabilistic method (see e.g. Alon and Spencer (2000)). We will choose $A_1, \ldots, A_r \subseteq \tilde{\mathcal{Y}}$ at random, such that each A_i is chosen uniformly at random from all subsets of $\tilde{\mathcal{Y}}$ (i.e., each $y \in \tilde{\mathcal{Y}}$ is independently chosen to be in A_i with probability $\frac{1}{2}$) and the different A_i's are independent. We will show that if $r = \lfloor\frac{\log(|\mathcal{Y}|-1)}{2}\rfloor - 2$, then with positive probability $\mathcal{A} = \{A_1, \ldots, A_r\}$ is G-shattered and $|\mathcal{A}| = r$ (i.e., the A_i's are different).

Denote $d = |\tilde{\mathcal{Y}}|$. Let $\psi : [r] \to \mathcal{X}$ be the (random) function $\psi(i) = A_i$ and let $\phi : \mathcal{Y} \to \{0, 1\}$ be the function that maps each $y \in \tilde{\mathcal{Y}}$ to 1 and $*$ to 0. Consider the (random) binary hypothesis class $\mathcal{H} = \{\phi \circ h_y \circ \psi \mid y \in \tilde{\mathcal{Y}}\}$. As we will show, for $r = \lfloor\frac{\log(d)}{2}\rfloor - 2$, $E[|\mathcal{H}|] > 2^r - 1$. In particular, there exists some choice of $\mathcal{A} = \{A_1, \ldots, A_r\}$ for which $|\mathcal{H}| > 2^r - 1$. Fix those sets for a moment. Since always $|\mathcal{H}| \leq 2^r$, it must be the case that $|\mathcal{H}| = 2^r$, i.e., $\mathcal{H} = 2^{[r]}$. By the definition of $\mathcal{H}$, it follows that for every $B \subseteq [r]$, there is $h_y \in \mathcal{H}_{\mathcal{Y},\text{Cantor}}$ such that for every $i \in B$, $h_y(A_i) = *$, while for every $i \notin B$, $h_y(A_i) \neq *$. It follows that $|\mathcal{A}| = r$ and $\mathcal{A}$ is G-shattered.

It remains to show that indeed, for $r = \lfloor\frac{\log(d)}{2}\rfloor - 2$, $E[|\mathcal{H}|] > 2^r - 1$. For every $S \subseteq [r]$, Let χ_S be the indicator random variable that is 1 if and only if $1_S \in \mathcal{H}$. We have

$$E[|\mathcal{H}|] = E[\sum_{S\subseteq[r]} \chi_S] = \sum_{S\subseteq[r]} E[\chi_S] \ . \tag{4}$$

Fix some $S \subseteq [r]$. For every $y \in \tilde{\mathcal{Y}}$ let $\chi_{S,y}$ be the indicator function that is 1 if and only if $1_S = \phi \circ h_y \circ \psi$. Note that $\sum_{y\in\tilde{\mathcal{Y}}} \chi_{S,y} > 0$ if and only if $\chi_S = 1$. Therefore, $E[\chi_S] = \Pr\left(\chi_S = 1\right) = \Pr\left(\sum_{y\in\tilde{\mathcal{Y}}} \chi_{S,y} > 0\right)$. Observe that

$$E[\sum_{y\in\tilde{\mathcal{Y}}} \chi_{S,y}] = \sum_{y\in\tilde{\mathcal{Y}}} \Pr\left(y \in A_i \text{ iff } i \in S\right) = d \cdot 2^{-r} \ .$$

We would like to use Chebyshev's inequality for the sum $\sum_{y\in\tilde{\mathcal{Y}}} \chi_{S,y}$. For this to be effective, we show next that for different $y_1, y_2 \in \tilde{\mathcal{Y}}$, χ_{S,y_1} and χ_{S,y_2} are uncorrelated. Note that $E[\chi_{S,y_1}\chi_{S,y_2}]$ is the probability that for every $i \in S$, $y_1, y_2 \in A_i$ while for every $i \notin S$, $y_1, y_2 \notin A_i$. It follows that

$$E[\chi_{S,y_1}\chi_{S,y_2}] = 2^{-2r} \ .$$

Therefore, $\mathrm{cov}(\chi_{S,y_1}\chi_{S,y_2}) = E[\chi_{S,y_1}\chi_{S,y_2}] - E[\chi_{S,y_1}]E[\chi_{S,y_2}] = 2^{-2r} - 2^{-r}2^{-r} = 0$. We conclude that χ_{S,y_1} and χ_{S,y_2} are uncorrelated. Thus, by Chebyshev's inequality,

$$\begin{aligned}
\Pr\left(\chi_S = 0\right) &= \Pr\left(\sum_{y\in\tilde{\mathcal{Y}}} \chi_{S,y} = 0\right) \\
&\leq \Pr\left(\left|\sum_{y\in\tilde{\mathcal{Y}}} \chi_{S,y} - d\cdot 2^{-r}\right| \geq d\cdot 2^{-r-1}\right) \\
&\leq \frac{2^{2r+2}}{d^2}\mathrm{var}\left(\sum_{y\in\tilde{\mathcal{Y}}} \chi_{S,y}\right) \\
&= \frac{2^{2r+2}}{d^2}\sum_{y\in\tilde{\mathcal{Y}}} \mathrm{var}\left(\chi_{S,y}\right) \\
&\leq \frac{2^{2r+2}}{d^2}\sum_{y\in\tilde{\mathcal{Y}}} E[\chi_{S,y}] \\
&= \frac{2^{2r+2}}{d^2} d 2^{-r} = \frac{2^{r+2}}{d} \ .
\end{aligned}$$

Remember that $r = \lfloor \frac{\log(d)}{2} \rfloor - 2$, so that $d > 2^{2r+2}$. Hence, $E[\chi_S] = 1 - \Pr(\chi_S = 0) \geq 1 - 2^{-r}$. Using equation (4), we conclude that

$$E[|\mathcal{H}|] > (1-2^{-r})2^r = 2^r - 1.$$

■

Appendix D. Proofs

D.1. Some lemmas and additional notations

Let $\mathcal{X}', \mathcal{Y}'$ be another instance and label spaces. Let $\Gamma : \mathcal{X}' \to \mathcal{X}$ and $\Lambda : \mathcal{Y}\cup\{\ominus\} \to \mathcal{Y}'\cup\{\ominus\}$. We denote

$$\Lambda \circ \mathcal{H} \circ \Gamma = \{\Lambda \circ h \circ \Gamma \mid h \in \mathcal{H}\} \ .$$

If Γ (respectively Λ) is the identity function we simplify the above notation to $\Lambda \circ \mathcal{H}$ (respectively $\mathcal{H} \circ \Gamma$). We say that a hypothesis class $\mathcal{H}' \subseteq (\mathcal{Y}' \cup \{\ominus\})^{\mathcal{X}'}$ is *realizable* by $\mathcal{H} \subseteq (\mathcal{Y} \cup \{\ominus\})^{\mathcal{X}}$ if $\mathcal{H}' \subseteq \Lambda \circ \mathcal{H} \circ \Gamma$ for some functions Γ and Λ. Note that in this case, the different notions of sample complexity with respect to $\mathcal{H}'$ are never larger than the corresponding notions with respect to $\mathcal{H}$.

Let $\mathcal{H} \subset (\mathcal{Y} \cup \{\ominus\})^{\mathcal{X}}$ be a hypothesis class. The *disjoint union* of m copies of $\mathcal{H}$ is the hypothesis class $\mathcal{H}_m$ whose instance space is $\mathcal{X}_m := \mathcal{X} \times [m]$, whose label space is $\mathcal{Y} \cup \{\ominus\}$, and that is composed of all functions $f : \mathcal{X}_m \to \mathcal{Y} \cup \{\ominus\}$ whose restriction to each copy of $\mathcal{X}$ is a function in $\mathcal{H}$ (namely, for every $i \in [m]$, the function $x \mapsto f(x,i)$ belongs to $\mathcal{H}$).

Lemma 22 *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class. Let $\mathcal{H}_m$ be a disjoint union of m copies of $\mathcal{H}$.*

1. *If $\mathcal{H}$ is realized by $\mathcal{H}_\Psi$ for some $\Psi : \mathcal{X}' \times \mathcal{Y}' \to \mathbb{R}^d$, then $\mathcal{H}_m$ is realized by $\mathcal{H}_{\Psi_m}$ for some $\Psi_m : \mathcal{X}'_m \times \mathcal{Y}' \to \mathbb{R}^{dm}$. Here, $\mathcal{X}'_m$ is a disjoint union of m copies of $\mathcal{X}'$.*
2. *If $\mathcal{H}$ is realized by $\mathcal{H}_{\Psi,R}$ for some $\Psi : \mathcal{X}' \times \mathcal{Y}' \to B^d$, then $\mathcal{H}_m$ is realized by $\mathcal{H}_{\Psi_m,mR}$ for some $\Psi_m : \mathcal{X}'_m \times \mathcal{Y}' \to B^{dm}$. Here, $\mathcal{X}'_m$ is a disjoint union of m copies of $\mathcal{X}'$.*
3. *If $\mathcal{H}$ is realized by $\mathcal{H}_{d,k}$, then $\mathcal{H}_m$ is realized by $\mathcal{H}_{dm,k}$.*
4. *If $\mathcal{H}$ is realized by $\mathcal{H}_{d,k,R}$, then $\mathcal{H}_m$ is realized by $\mathcal{H}_{dm,k,mR}$.*

Proof We prove only part 1. The remaining three are very similar. Let $\Gamma : \mathcal{X} \to \mathcal{X}', \Lambda : \mathcal{Y}' \to \mathcal{Y}$ be two mappings for which

$$\mathcal{H} \subseteq \Lambda \circ \mathcal{H}_\Psi \circ \Gamma \ .$$

Let $\mathcal{X}_m = \mathcal{X} \times [m]$ be a disjoint union of m copies of $\mathcal{X}$. Let $T_i : \mathbb{R}^d \to \mathbb{R}^{dm}$ be the linear mapping that maps e_j to $e_{(i-1)d+j}$. Define $\Psi_m : \mathcal{X}_m \times \mathcal{Y} \to \mathbb{R}^{dm}$ by $\Psi_m((x,i),y) = T_i(\Psi(x,y))$. Define $\Gamma_m : \mathcal{X}_m \to \mathcal{X}'_m$ by $\Gamma_m(x,i) = (\Gamma(x),i)$. It is not hard to check that

$$\mathcal{H}_m \subseteq \Lambda \circ \mathcal{H}_{\Psi_m} \circ \Gamma_m \ .$$

■

Lemma 23 *Let $\mathcal{H} \subseteq (\mathcal{Y} \cup \{\ominus)^{\mathcal{X}}$ be a hypothesis class and let $\mathcal{H}_m$ be a disjoint union of m copies of $\mathcal{H}$. Then* $\mathrm{Gdim}(\mathcal{H}_m) = m \cdot \mathrm{Gdim}(\mathcal{H})$.

Proof A routine verification. ■

D.2. Proof of Theorem 1

For simplicity, we prove the theorem for d even and $d = \infty$. For finite d, fix some d-elements set $\mathcal{X}_d$. Let $\mathcal{Y}_d = 2^{\mathcal{X}_d} \cup \{*\}$. For $A \subseteq \mathcal{X}_d$ define $h_A : \mathcal{X}_d \to \mathcal{Y}_d$ by

$$h_A(x) = \begin{cases} A & x \in A \\ * & \text{otherwise} \end{cases} \ .$$

Finally, let

$$\mathcal{H}_d = \left\{ h_A \mid |A| = \frac{d}{2} \right\} \ .$$

We next define a "limit" of the classes $\mathcal{H}_d$. Suppose that the sets $\{\mathcal{X}_d\}_{d \text{ is even integer}}$ are pairwise disjoint. Let $\mathcal{X}_\infty = \cup_{d \text{ is even}} \mathcal{X}_d$ and $\mathcal{Y}_\infty = \left(\cup_{d \text{ is even}} 2^{\mathcal{X}_d}\right) \cup \{*\}$. For $A \subseteq \mathcal{X}_d$, extend $h_A : \mathcal{X}_d \to \mathcal{Y}_d$ to a function $h_A : \mathcal{X}_\infty \to \mathcal{Y}_\infty$ by defining it to be $*$ outside of $\mathcal{X}_d$. Finally, let

$$\mathcal{H}_\infty = \left\{ h_A \mid \text{for some } d,\ A \subseteq \mathcal{X}_d \text{ and } |A| = \frac{d}{2} \right\} \ .$$

We will use the following version of Chernoff's bound:

Theorem 24 *Let $X_1, \ldots, X_n \in \{0,1\}$ be independent random variables, $X = X_1 + \ldots + X_n$ and $\mu = \mathbb{E}[X]$. Then $\Pr(X \geq 2\mu) \leq \exp\left(-\frac{\mu}{3}\right)$.*

We are now ready to prove Theorem 1. The first part follows from Lemma 20. The last part is a direct consequence of the first and second part. We proceed to the second part. For $d < \infty$, the task of properly learning $\mathcal{H}_d$ can be easily reduced to the task of properly learning $\mathcal{H}_\infty$. Therefore, the sample complexity of learning $\mathcal{H}_\infty$ by a proper learning algorithm is lower bounded by the sample complexity of properly learning $\mathcal{H}_d$. Therefore, it is enough to prove the second part for finite d.

Fix some $x_0 \in \mathcal{X}$. Let $\epsilon > 0$. Let $A \subset \mathcal{X}_d \setminus \{x_0\}$ be a set with $\frac{d}{2}$ elements. Let $\mathcal{D}_A$ be a distribution on $\mathcal{X}_d \times \mathcal{Y}_d$ that assigns a probability of $1 - 16\epsilon$ to some point $(x_0, h_A(x_0)) \in \mathcal{X}_d \times \mathcal{Y}_d$ and is uniform on the remaining points of the form $\{(x, h_A(x)) \mid x \notin A\}$.

We claim that there is some A such that whenever $\mathcal{A}$ runs on $\mathcal{D}_A$ with $m \leq \frac{1}{128}\frac{d}{\epsilon}$ examples, it outputs with probability $\geq \frac{1}{2}$ a hypothesis with error $\geq \epsilon$. This shows that for every $\delta < \frac{1}{2}$, $m_{\mathcal{A}}(\epsilon, \delta) \geq \frac{1}{128}\frac{d}{\epsilon}$. Also, since $\mathcal{H}_d$ contains two different function that agree on some point, by a standard argument, we have $m_{\mathcal{A}}(\epsilon, \delta) = \Omega\left(\frac{\log(1/\delta)}{\epsilon}\right)$. Combining these two estimates, the proof is established.

It remains to show the existence of such A. Suppose that A is chosen uniformly at random among all subsets of $\mathcal{X}_d \setminus \{x_0\}$ of size $\frac{d}{2}$. Let X be the random variable counting the number of samples, out of $\frac{1}{128}\frac{d}{\epsilon}$ i.i.d. examples drawn from $\mathcal{D}_A$, which are not $(x_0, h_A(x_0))$. We have $\mathbb{E}[X] = \frac{1}{8}d$. Therefore, by Chernoff's bound 24, with probability $> 1 - \exp\left(-\frac{d}{24}\right) > \frac{1}{2}$, the algorithm will see less than $\frac{d}{4}$ examples whose instance is from $\mathcal{X} \setminus \{x_0\} \setminus A$. Conditioning on this event, A is a uniformly chosen random set of size $\frac{d}{2}$ that is chosen uniformly from all subsets of a set $\mathcal{X}' \subset \mathcal{X}$ with $|\mathcal{X}'| \geq \frac{3}{4}d$ ($\mathcal{X}'$ is the set of all points that are not present in the sample), and the hypothesis returned by the algorithm is h_B, where $B \subset \mathcal{X}$ is a subset of size $\frac{d}{2}$ that is independent from A. It is not hard to see that in this case $\mathbb{E}|B \setminus A| \geq \frac{1}{6}d$. Hence, there exists some A for which, with probability $> \frac{1}{2}$ over the choice of the sample, $|B \setminus A| \geq \frac{1}{6}d$. For such A we have, since h_B errs on all elements in $B \setminus A$ and the probability of each such element is $\geq \frac{16\epsilon}{\frac{d}{2}} = \frac{32}{d}\epsilon$,

$$\mathrm{Err}_{\mathcal{D}_A}(h_B) \geq |B \setminus A|\frac{32\epsilon}{d} \geq \frac{d}{6}\frac{32\epsilon}{d} > \epsilon$$

with probability $> \frac{1}{2}$ over the choice of the sample.

D.3. Proof of Lemma 3

We first prove it to finite hypergraphs. We use induction on the number of vertices. By assumption, $d(G) \leq d$. Therefore, there is $v_0 \in V$ with $d(v_0) \leq d$. Let $G' = (V', E') = G[V \setminus \{v_0\}]$. By the induction hypothesis, there exists an orientation $h' : E' \to V'$ with maximal out-degree d. We define an orientation $h : E \to V$ by

$$h(e) = \begin{cases} v & e = \{v_0, v\} \\ h'(e \setminus \{v_0\}) & \text{otherwise} \end{cases}$$

The lemma extend to the case where $\mathcal{Y}$ is infinite by a standard application of the compactness theorem for propositional calculus.

D.4. Proof of theorem 4

Let $\mathcal{A}$ be some learning algorithm, and denote by $\mathcal{I}$ the one inclusion algorithm. Suppose that we run $\mathcal{A}$ on $m_{\mathcal{A},\mathcal{H}}\left(\frac{\epsilon}{2},\frac{\epsilon}{2}\right)$ examples, obtain a hypothesis h and predict $h(x)$ on some new example. The probability of error if $\leq \left(1-\frac{\epsilon}{2}\right)\frac{\epsilon}{2}+\frac{\epsilon}{2} \leq \epsilon$. By theorem 2, it follows that

$$m_{\mathcal{A},\mathcal{H}}\left(\frac{\epsilon}{2},\frac{\epsilon}{2}\right) \geq \min\left\{m \mid \frac{1}{2e}\frac{\mu_{\mathcal{H}}(m)}{m} \leq \epsilon\right\} =: \bar{m} \ .$$

Now, if we run the one inclusion algorithm on $\bar{m}$ examples then, again by theorem 2, the probability that the hypothesis it return will err a new example is $\leq 2e\epsilon$. Therefore, the probability that the error of the returned hypothesis is $\geq 4e\epsilon$ is $\leq \frac{1}{2}$. In follows that

$$\bar{m} \geq m_{\mathcal{I},\mathcal{H}}\left(4e\epsilon,\frac{1}{2}\right) \ .$$

Combining the two inequalities, we obtain that

$$m_{\mathcal{I},\mathcal{H}}\left(4e\epsilon,\frac{1}{2}\right) \leq m_{\mathcal{A},\mathcal{H}}\left(\frac{\epsilon}{2},\frac{\epsilon}{2}\right)$$

Since this is true for every algorithm $\mathcal{A}$, we have

$$m_{\mathcal{I},\mathcal{H}}\left(4e\epsilon,\frac{1}{2}\right) \leq m_{\text{PAC},\mathcal{H}}\left(\frac{\epsilon}{2},\frac{\epsilon}{2}\right) \leq m_{\text{PAC},\mathcal{H}}\left(\frac{\epsilon}{4},\frac{1}{2}\right)\cdot O\left(\log(1/\epsilon)\right)$$

Here, the last inequality follows by a standard repetition argument. Equivalently,

$$m_{\mathcal{I},\mathcal{H}}\left(\epsilon,\frac{1}{2}\right) \leq m_{\text{PAC},\mathcal{H}}\left(\frac{\epsilon}{16e},\frac{1}{2}\right)\cdot O\left(\log(1/\epsilon)\right)$$

Again, using a repetition argument we conclude that

$$m_{\overline{\mathcal{I}},\mathcal{H}}(\epsilon,\delta) \leq m_{\mathcal{I},\mathcal{H}}\left(\frac{\epsilon}{2},\frac{1}{2}\right)\cdot O\left(\log(1/\delta)\right) \leq m_{\text{PAC},\mathcal{H}}\left(\frac{\epsilon}{32e},\frac{1}{2}\right)\cdot O\left(\log(1/\delta)\log(1/\epsilon)\right)$$

D.5. Validity of the compression scheme given in Section 4.3

It is not hard to see that the hypothesis we output is the minimal-norm vector $w \in \text{conv}(Z)$ (where Z is the set defined in the compression step). It is left to show that w makes no errors on the original sample. Indeed, otherwise there exists $z \in Z$ for which $\langle w,z\rangle \leq 0$. By claim 1, $z \neq 0$. For $\alpha = \frac{\|w\|^2}{\|z\|^2+\|w\|^2} \in (0,1)$, let $w' = (1-\alpha)w+\alpha z$. We have that $w' \in \text{conv}(Z)$. Moreover,

$$\begin{aligned}\|w'\|^2 &= (1-\alpha)^2\|w\|^2+\alpha^2\|z\|^2+2\alpha(1-\alpha)\langle w,z\rangle \leq (1-\alpha)^2\|w\|^2+\alpha^2\|z\|^2 \\ &= \frac{\|z\|^4\|w\|^2+\|w\|^4\|z\|^2}{\left(\|z\|^2+\|w\|^2\right)^2} = \frac{\|z\|^2\|w\|^2}{\|z\|^2+\|w\|^2} < \|w\|^2 \ .\end{aligned}$$

This contradicts the minimality of w. It only remains to prove the following claim, which was used in the analysis.

Claim 1 *Let $(x_1, y_1), \ldots, (x_m, y_m)$ be a realizable sample and let Z be the set of all vectors of the form $\Psi(x_i, y_i) - \Psi(x_i, y)$ for $y \neq y_i$. Then $0 \notin \text{conv}(Z)$.*

Proof Since the sample is realizable, there exists a vector w in $\mathbb{R}^d$ for which, $\forall z \in Z$, $\langle w, z \rangle > 0$. Clearly, this holds also for every $z \in \text{conv}(Z)$, hence $0 \notin \text{conv}(Z)$. ■

D.6. Proof of the second part of Theorem 5

Without loss of generality, we assume that $\mathcal{Y}$ consists of $2^n + 1$ elements for some natural number n (otherwise, use only $2^n + 1$ labels, where n is the largest number satisfying $2^n + 1 \leq |\mathcal{Y}|$). Let $\mathcal{X}$ be a set consisting of n elements. By renaming the names of the labels, we can assume that $\mathcal{Y} = 2^{\mathcal{X}} \cup \{*\}$. By Lemma 21, the ERM sample complexity of $\mathcal{H}_{\mathcal{X},\text{Cantor}}$ is $\Omega\left(\frac{\log(|\mathcal{Y}|)+\log(1/\delta)}{\epsilon}\right)$. We will show that there exists a function $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^3$, such that $\mathcal{H}_{\mathcal{X},\text{Cantor}}$ is realized by $\mathcal{H}_\Psi$. It follows that the ERM sample complexity of $\mathcal{H}_\Psi$ is also $\Omega\left(\frac{\log(|\mathcal{Y}|)+\log(1/\delta)}{\epsilon}\right)$. Therefore, the second part of Theorem 5 is proved for $d = 3$. The extension of the result to general d follows from Lemma 22.

Definition of Ψ: Denote $k = 2^{|\mathcal{X}|}$ and let $f : 2^{\mathcal{X}} \to \{0, 1, \ldots, k-1\}$ be some one-to-one mapping. For $A \subseteq \mathcal{X}$ define

$$\phi(A) = \left(\cos\left(\frac{2\pi f(A)}{k}\right), \sin\left(\frac{2\pi f(A)}{k}\right), 0\right) .$$

Also, define

$$\phi(*) = (0, 0, 1) .$$

Note that for different subsets $A, B \subseteq \mathcal{X}$ we have that

$$\langle \phi(A), \phi(B) \rangle = \cos\left(\frac{2\pi(f(A) - f(B))}{k}\right) \leq \cos\left(\frac{2\pi}{k}\right) < \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi}{k}\right) < 1 \qquad (5)$$

Define $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^3$ by

$$\forall A \subset \mathcal{X}, \quad \Psi(x, A) = \begin{cases} \phi(A) & x \in A \\ 0 & x \notin A \end{cases}$$

$$\Psi(x, *) = \left(\frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi}{k}\right)\right) \cdot \phi(*)$$

Claim 2 *$\mathcal{H}_{\mathcal{X},\text{Cantor}}$ is realized by $\mathcal{H}_\Psi$.*

Proof We will show that $\mathcal{H}_{\mathcal{X},\text{Cantor}} \subseteq \mathcal{H}_\Psi$. Let $B \subseteq \mathcal{X}$. We must show that $h_B \in \mathcal{H}_\Psi$. Let $w \in \mathbb{R}^3$ be the vector

$$w = \phi(B) + \phi(*) .$$

We claim that for the function $h_w \in \mathcal{H}_\Psi$, defined by w we have $h_w = h_B$. Indeed, let $x \in \mathcal{X}$ we split into the cases $x \in B$ and $x \notin B$.

Case 1 ($x \in B$): We must show that $h_w(x) = B$. That is, for every $y \in \mathcal{Y} \setminus \{B\}$,

$$\langle w, \Psi(x, B)\rangle > \langle w, \Psi(x, y)\rangle \ .$$

Note that

$$\langle w, \Psi(x, B)\rangle = \langle \phi(B) + \phi(*), \phi(B)\rangle = 1 \ .$$

Therefore, for every $y \in \mathcal{Y} \setminus \{B\}$, we must show that $1 > \langle w, \Psi(x, y)\rangle$. We split into three cases. If $y = A$ for some $A \subseteq \mathcal{X}$ and $x \in A$ then, using equation (5),

$$\langle w, \Psi(x, y)\rangle = \langle \phi(B) + \phi(*), \phi(A)\rangle = \langle \phi(B), \phi(A)\rangle < 1 \ .$$

If $y = A$ for some $A \subseteq \mathcal{X}$ and $x \notin A$ then,

$$\langle w, \Psi(x, y)\rangle = \langle \phi(B) + \phi(*), 0\rangle = 0 < 1 \ .$$

If $y = *$ then,

$$\langle w, \Psi(x, y)\rangle = \left\langle \phi(B) + \phi(*), \left(\frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi}{k}\right)\right) \cdot \phi(*)\right\rangle = \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi}{k}\right) < 1 \ .$$

Case 2 ($x \notin B$): We must show that $h_w(x) = *$. That is, for every $A \in \mathcal{Y} \setminus \{*\}$,

$$\langle w, \Psi(x, *)\rangle > \langle w, \Psi(x, A)\rangle \ .$$

Note that

$$\langle w, \Psi(x, *)\rangle = \left\langle \phi(B) + \phi(*), \left(\frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi}{k}\right)\right) \phi(*)\right\rangle = \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi}{k}\right) \ .$$

Therefore, for every $A \in \mathcal{Y} \setminus \{B\}$, we must show that $\frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi}{k}\right) > \langle w, \Psi(x, A)\rangle$. Indeed, if $x \in A$ then $A \neq B$ (since $x \notin B$). Therefore, using equation (5),

$$\langle w, \Psi(x, A)\rangle = \langle \phi(B) + \phi(*), \phi(A)\rangle = \langle \phi(B), \phi(A)\rangle < \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi}{k}\right) \ .$$

If $x \notin A$ then

$$\langle w, \Psi(x, A)\rangle = \langle \phi(B) + \phi(*), 0\rangle = 0 < \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi}{k}\right) \ .$$

D.7. Proof of Theorem 6

To prove the first part of Theorem 6, we will rely again on Theorem 18. We will show a compression scheme of size $O(R)$, which is based on the multiclass perceptron. This compression scheme is not new. However, for completeness, we briefly survey it next. Recall that the multiclass perceptron is an online classification algorithm. At each step it receives an instance and tries to predict its label, based on the observed past. The two crucial properties of the preceptron that we will rely on are the following:

- If the perceptron runs on a sequence of examples that is realizable by $\mathcal{H}_{\Psi,R}$, then it makes at most $O(R)$ mistakes.

- The predictions made by the perceptron algorithm, are affected only by previous *erroneous* predictions.

Based on these two properties, the compression scheme proceeds as follows: Given a realizable sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, it runs the preceptron algorithm $\Omega(R)$ times on the sequence $(x_1, y_1), \ldots, (x_m, y_m)$ (without a reset between consecutive runs). By the first property, in at least one of these runs, the preceprton will make no mistakes on the sequence $(x_1, y_1), \ldots, (x_m, y_m)$ (otherwise, there would be $\Omega(R)$ mistakes in total). The output of the compression step would be the erroneous examples previous to this sequence. By the first property, the number of such examples is $O(R)$. The decompression will run the preceptron on these examples, and output the hypothesis $h : \mathcal{X} \to \mathcal{Y}$, such that $h(x)$ is the prediction of the perceptron on x, after operating on these examples. By the second property, h is correct on every x_i.

We proceed to the second part. By Lemma 23 and Lemma 19, it is enough to show that a disjoint union of $\Omega(R)$ copies of $\mathcal{H}_{\mathcal{X},\text{Cantor}}$, with $|\mathcal{X}| = \Omega(\log(|\mathcal{Y}|))$, can be realized by $\mathcal{H}_{\Psi,R}$ for an appropriate mapping $\Psi : \mathcal{X} \times \mathcal{Y} \to B^d$ for some $d > 0$. By Lemma 22, it is enough to show that, for some universal constant $C > 0$, $\mathcal{H}_{\mathcal{X},\text{Cantor}}$, with $|\mathcal{X}| = \Omega(\log(|\mathcal{Y}|))$, can be realized by $\mathcal{H}_{\Psi,C}$ for an appropriate mapping $\Psi : \mathcal{X} \times \mathcal{Y} \to B^d$ for some $d > 0$.

Without loss of generality, we assume that $|\mathcal{Y}| - 1$ is a power of 2 (otherwise, use only k labels, where k is the largest integer such that $k - 1$ is a poser of 2 and $k \leq |\mathcal{Y}|$). Denote $k = |\mathcal{Y}| - 1$. Fix some finite set $\mathcal{X}$ of cardinality $\log(|\mathcal{Y}| - 1)$. By renaming the labels, we can assume that $\mathcal{Y} = 2^{\mathcal{X}} \cup \{*\}$.

Let $\{e_y\}_{y \in \mathcal{Y}}$ be a collection of unit vectors in $\mathbb{R}^d$ with the property that for $y_1 \neq y_2$,

$$|\langle e_{y_1}, e_{y_2} \rangle| < \frac{1}{100} \ . \tag{6}$$

Remark 25 *Clearly, it is possible to find such a collection when $d = k + 1$ (simply take $\{e_y\}_{y \in \mathcal{Y}}$ to be an orthogonal basis of $\mathbb{R}^{k+1}$). However, equation (6) requires the collection to be just "almost orthogonal". Such a collection can be found in $\mathbb{R}^d$ for $d = O(\log(k))$ (see, e.g. Matousek (2002), chapter 13).*

Define $\Psi : \mathcal{X} \times \mathcal{Y} \to B^d$ by

$$\forall A \subset \mathcal{X}, \quad \Psi(x, A) = \begin{cases} e_A & x \in A \\ 0 & x \notin A \end{cases}$$

$$\Psi(x, *) = e_*$$

The following claim establishes the proof of Theorem 6.

Claim 3 *$\mathcal{H}_{\mathcal{X},\text{Cantor}}$ is realized by $\mathcal{H}_{\Psi,8}$.*

Proof We will show that $\mathcal{H}_{\mathcal{X},\text{Cantor}} \subseteq \mathcal{H}_{\Psi,8}$. Let $B \subseteq \mathcal{X}$. We must show that $h_B \in \mathcal{H}_{\Psi,8}$. Let $w = W \cdot (e_B + \frac{1}{2} e_*)$ for $W = \frac{100}{45}$. We claim that the hypothesis in $\mathcal{H}_{\Psi,8}$ that corresponds to w is h_B. Indeed, let $x \in \mathcal{X}$. We split into the cases $x \in B$ and $x \notin B$.

Case 1 ($x \in B$)**:** We must show that for every $y \in \mathcal{Y} \setminus \{B\}$,

$$\langle w, \Psi(x, B)\rangle \geq 1 + \langle w, \Psi(x, y)\rangle \ .$$

Note that

$$\langle w, \Psi(x, B)\rangle = \left\langle W \cdot \left(e_B + \frac{1}{2}e_*\right), e_B \right\rangle = W\left(1 + \frac{1}{2}\langle e_*, e_B\rangle\right) \geq W\left(1 - \frac{1}{100}\right) \ .$$

Now, if $y \in \mathcal{Y} \setminus \{B\}$ then either $y \subseteq \mathcal{X}$ and $x \notin y$. In this case, $\langle w, \Psi(x, y)\rangle = \langle w, 0\rangle = 0$. In the remaining cases,

$$\langle w, \Psi(x, y)\rangle = \left\langle W \cdot \left(e_B + \frac{1}{2}e_*\right), e_y \right\rangle = W\left(\langle e_y, e_B\rangle + \frac{1}{2}\langle e_*, e_B\rangle\right) \leq W\frac{1}{50} \ .$$

It follows that

$$\langle w, \Psi(x, B)\rangle - \langle w, \Psi(x, y)\rangle \geq \frac{24}{25}W \geq 1 \ .$$

Case 2 ($x \notin B$)**:** We must show that for every $y \in \mathcal{Y} \setminus \{*\}$,

$$\langle w, \Psi(x, *)\rangle \geq 1 + \langle w, \Psi(x, y)\rangle \ .$$

Note that

$$\langle w, \Psi(x, *)\rangle = \left\langle W \cdot \left(e_B + \frac{1}{2}e_*\right), e_* \right\rangle = W\left(\langle e_B, e_*\rangle + \frac{1}{2}\right) \geq W\left(\frac{1}{2} - \frac{1}{100}\right) \ .$$

Now, suppose that $A = y \in \mathcal{Y} \setminus \{*\}$. If $x \notin A$ then,

$$\langle w, \Psi(x, y)\rangle = \left\langle W \cdot \left(e_B + \frac{1}{2}e_*\right), 0 \right\rangle = 0 \leq \frac{1}{25}W \ .$$

If $x \in A$ then $A \neq B$. Therefore,

$$\langle w, \Psi(x, y)\rangle = \left\langle W \cdot \left(e_B + \frac{1}{2}e_*\right), e_A \right\rangle = W\left(\langle e_B, e_A\rangle + \frac{1}{2}\langle e_*, e_A\rangle\right) \leq \frac{1}{25}W \ .$$

It follows that

$$\langle w, \Psi(x, *)\rangle - \langle w, \Psi(x, y)\rangle \geq \frac{45}{100}W \geq 1 \ .$$

D.8. Proof of Theorem 9

The first part of the theorem follows directly from the first part of Theorem 6. We proceed to the second part. First, we note that $\mathcal{H}_{d,t,2,R}$ can be realized by $\mathcal{H}_{d,t,q,R}$. Therefore, it is enough to restrict to the case $q = 2$. To simplify notations, we denote $\mathcal{H}_{d,t,2,R}$ by $\mathcal{H}_{d,t,R}$. Also, the label space of $\mathcal{H}_{d,t,R}$ will be $\{0,1\}^t$ instead of $[2]^q$.

By Lemma 23 and Lemma 21, it is enough to show that a disjoint union of $\Omega(R)$ copies of $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$, with $|\mathcal{Y}| = \Omega(t)$, can be realized by $\mathcal{H}_{d,t,R}$ for $d \geq (t+1)R$. By Lemma 22, it is enough to show that, for some universal constant $C > 0$, $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$, with $|\mathcal{Y}| = t+1$, can be realized by $\mathcal{H}_{t+1,t,C}$. Indeed:

Claim 4 *Let $\tilde{\mathcal{Y}} = [t]$. The class $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$ is realized by $\mathcal{H}_{t+1,t,128}$.*

Proof Recall that the instance space of $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$ is $\mathcal{X} = 2^{[t]}$. Also, let $e_* := e_{t+1} \in B^{t+1}$. Consider the mapping $\Gamma : \mathcal{X} \to (B^{t+1})^t$ defined as follows. For every $A \in \mathcal{X}$, $\Gamma(A)$ is the matrix whose i'th column is $\frac{1}{2}e_i + \frac{1}{4}e_*$ if $i \in A$ and $\frac{1}{4}e_*$ otherwise. Let $\Lambda : \{0,1\}^t \cup \{\ominus\} \to [t] \cup \{*\}$ be any mapping that maps $e_i \in \{0,1\}^t$ to i and $0 \in \{0,1\}^t$ to $*$. To establish the claim we will show that

$$\mathcal{H}_{\mathcal{Y},\text{Cantor}} \subseteq \Lambda \circ \mathcal{H}_{t+1,t,128} \circ \Gamma \ .$$

We must show that for every $i \in [t]$, $h_i \in \Lambda \circ \mathcal{H}_{t+1,t,128} \circ \Gamma$ and that $h_* \in \Lambda \circ \mathcal{H}_{t+1,t,128} \circ \Gamma$. We start with h_i. Let $W \in M_{(t+1)\times 2}$ be the matrix whose left column is 0 and whose right column is $8e_i - 8e_*$. Let $h_W \in \mathcal{H}_{t+1,t,128}$ be the hypothesis corresponding to W. We claim that $h_i = \Lambda \circ h_W \circ \Gamma$. Indeed, let $A \in \mathcal{X}$. We must show that $\Lambda(h_W(\Gamma(A))) = h_i(A)$. By the definition of Λ and h_i, it is enough to show that $h_W(\Gamma(A)) = e_i$ if $i \in A$, while $h_W(\Gamma(A)) = 0$ otherwise. Let $\Psi : (B^{t+1})^t \times \{0,1\}^t \to M_{t+1,2}$ be the mapping for which $\mathcal{H}_{t+1,t,128} = \mathcal{H}_{\Psi,128}$. Since the left column of W is zero, we have that $\langle W, \Psi(\Gamma(A), 0)\rangle = 0$, and for $0 \neq y \in \{0,1\}^t$,

$$\begin{aligned}
\langle W, \Psi(\Gamma(A), y)\rangle &= \frac{1}{2 \cdot |\{j \mid y_j = 1\}|} \sum_{j|y_j=1} \langle 4e_i - 4e_*, (\Gamma(A))^j \rangle \\
&= \frac{1}{|\{j \mid y_j = 1\}|} \sum_{j|y_j=1} (2 \cdot 1[i = j \text{ and } i \in A] - 1) \\
&= \frac{2 \cdot 1[i \in A \text{ and } y_i = 1]}{|\{j \mid y_j = 1\}|} - 1 \ .
\end{aligned}$$

It follows that if $i \in A$ then $\langle W, \Psi(\Gamma(A), e_i)\rangle = 1$ while $\langle W, \Psi(\Gamma(A), y)\rangle \leq 0$ for every $y \neq e_i$. Therefore, $h_W(\Gamma(A)) = e_i$. If $i \notin A$ then $\langle W, \Psi(\Gamma(A), 0)\rangle = 0$ while $\langle W, \Psi(\Gamma(A), y)\rangle \leq -1$ for every $y \neq 0$. Therefore $h_W(\Gamma(A)) = 0$.

The fact that $h_* \in \Lambda \circ \mathcal{H}_{t+1,t,128} \circ \Gamma$ follows from a similar argument, where $W \in M_{(t+1)\times 2}$ is the matrix whose left column is 0 and whose right column is $-8e_*$. It is not hard to see that if $h_W \in \mathcal{H}_{t+1,t,128}$ is the hypothesis corresponding to W, we have $h_* = \Lambda \circ h_W \circ \Gamma$. ■

D.9. Proof of Theorem 8

The first part of the theorem follows directly from the first part of Theorem 5. We proceed to the second part. First, by the following lemma, it is enough to restrict ourselves to the case $q = 2$. Given two hypothesis classes $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and $\mathcal{H}' \subseteq \mathcal{Y}'^{\mathcal{X}'}$, we say that $\mathcal{H}'$ *finitely realizes* $\mathcal{H}$ if, for every finite $\mathcal{U} \subset \mathcal{X}$, $\mathcal{H}'$ realizes $\mathcal{H}|_{\mathcal{U}}$. It is clear that in this case $\text{Gdim}(\mathcal{H}') \geq \text{Gdim}(\mathcal{H})$.

Lemma 26 *For every d, t and $q \geq 2$, a disjoint union of $\lfloor \frac{q}{2} \rfloor$ copies of $\mathcal{H}_{d,t,2}$ is finitely realized by $\mathcal{H}_{d+2,t,q}$*

Proof For simplicity, assume that q is even and let $r = \frac{q}{2}$. Let $X_1, \ldots, X_r$ be finite subsets of $M_{d,t}$. We should show that there is a mapping $\Gamma : X_1 \dot{\cup} \ldots \dot{\cup} X_r \to M_{d+2,t}$ and a mapping

$\Lambda : [q]^t \to [2]^t$ such that

$$\left(\mathcal{H}_{d,t,2}\right)_m |_{X_1 \dot\cup \ldots \dot\cup X_r} \subset \left(\Lambda \circ \mathcal{H}_{d+2,t,q} \circ \Gamma\right) |_{X_1 \dot\cup \ldots \dot\cup X_r} \tag{7}$$

For $x \in X_j$ we define

$$\Gamma(x) = \left(x^T, \cos\left(j\frac{2\pi}{r}\right), \sin\left(j\frac{2\pi}{r}\right)\right)^T$$

Also, let $\lambda : [q] \to [2]$ be the function that maps odd numbers to 1 and even numbers to 2. Finally, define $\Lambda : [q]^t \to [2]^t$ by $\Lambda(y_1, \ldots, y_t) = (\lambda(y_1), \ldots, \lambda(y_t))$. We claim that (7) holds with these Λ and Γ.

Indeed, let $W_1, \ldots, W_r \in M_{d\times 2}$. We should show that the function $g \in \left(\mathcal{H}_{d,t,2}\right)_m |_{X_1 \dot\cup \ldots \dot\cup X_r}$ defined by these function is of the form $\left(\Lambda \circ h \circ \Gamma\right) |_{X_1 \dot\cup \ldots \dot\cup X_r}$ for some $h \in \mathcal{H}_{d+2,t,q}$. Fix $M > 0$ and let h be the hypothesis defined by the matrix $W \in M_{d+2,q}$ defined as follows

$$W = \begin{bmatrix} W_1^1 & W_1^2 & W_2^1 & W_2^2 & \cdots & W_r^1 & W_r^2 \\ M\cos\left(\frac{2\pi}{r}\right) & M\cos\left(\frac{2\pi}{r}\right) & M\cos\left(2\frac{2\pi}{r}\right) & M\cos\left(2\frac{2\pi}{r}\right) & \cdots & M\cos\left(r\frac{2\pi}{r}\right) & M\cos\left(r\frac{2\pi}{r}\right) \\ M\sin\left(\frac{2\pi}{r}\right) & M\sin\left(\frac{2\pi}{r}\right) & M\sin\left(2\frac{2\pi}{r}\right) & M\sin\left(2\frac{2\pi}{r}\right) & \cdots & M\sin\left(r\frac{2\pi}{r}\right) & M\sin\left(r\frac{2\pi}{r}\right) \end{bmatrix}$$

It is not hard to check that for large enough M, $g = \left(\Lambda \circ h \circ \Gamma\right) |_{X_1 \dot\cup \ldots \dot\cup X_r}$ ■

Next we prove Theorem 8 for $q = 2$. To simplify notation, we let $\mathcal{H}_{d,t} := \mathcal{H}_{d,t,2}$. We make one further reduction, showing that it is enough to prove the theorem for the case $d = 3$. Indeed, by Lemma 23 and Lemma 21, it is enough to show that a disjoint union of $\Omega(d)$ copies of $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$, with $|\mathcal{Y}| = \Omega(k)$, can be realized by $\mathcal{H}_{d,t}$. By Lemma 22, it is enough to show that, for some universal constant $C > 0$ (we will take $C = 3$), $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$, with $|\mathcal{Y}| = t + 1$, can be realized by $\mathcal{H}_{C,t}$. Indeed:

Claim 5 *Let $\tilde{\mathcal{Y}} = [t]$. The class $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$ is realized by $\mathcal{H}_{3,t}$.*

Proof [(sketch)] The proof is similar to the proof of the second part of Theorem 9. Recall that the instance space of $\mathcal{H}_{\mathcal{Y},\text{Cantor}}$ is $\mathcal{X} = 2^{[t]}$. For $i \in [t]$ define $\phi(i) = \left(\cos\left(\frac{i2\pi}{t}\right), \sin\left(\frac{i2\pi}{t}\right), 0\right)$. Also, let

$$\phi(*) = \left(0, 0, \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi}{t}\right)\right) .$$

Consider the mapping $\Gamma : \mathcal{X} \to (B^3)^t$ defined as follows. For every $A \in \mathcal{X}$, $\Gamma(A)$ is the matrix whose i'th column is $\frac{1}{2}\phi(i) + \frac{1}{2}\phi(*)$ if $i \in A$ and $\frac{1}{2}\phi(*)$ otherwise. Let $\Lambda : \{0,1\}^t \cup \{\ominus\} \to [k] \cup \{*\}$ be any mapping that maps $e_i \in \{0,1\}^t$ to i and $0 \in \{0,1\}^t$ to $*$. To establish the claim we will show that

$$\mathcal{H}_{\mathcal{Y},\text{Cantor}} \subseteq \Lambda \circ \mathcal{H}_{3,t} \circ \Gamma .$$

We must show that for every $i \in [t]$, $h_i \in \Lambda \circ \mathcal{H}_{3,t} \circ \Gamma$ and that $h_* \in \Lambda \circ \mathcal{H}_{3,t} \circ \Gamma$. We start with h_i. Let $W \in M_{3\times 2}$ be the matrix whose left column is 0 and whose right column is $\phi(i) - e_3$. It is not hard to see that if $h_W \in \mathcal{H}_{3,t}$ is the hypothesis corresponding to W, we have $h_i = \Lambda \circ h_W \circ \Gamma$.

For h_*, let $W \in M_{3\times 2}$ be the matrix whose left column is 0 and whose right column is $-e_3$. It is not hard to see that for $h_W \in \mathcal{H}_{3,t}$, we have $h_* = \Lambda \circ h_W \circ \Gamma$. ■

D.10. Proof of Theorem 7

Theorem 27 *For every* $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, $\mathrm{Ndim}(\mathcal{H}_\Psi) \leq d$.

Proof Let $C \subseteq \mathcal{X}$ be an N-shattered set, and let $f_0, f_1 : C \to \mathcal{Y}$ be two functions that witness the shattering. We must show that $|C| \leq d$. For every $x \in C$ let $\rho(x) = \Psi(x, f_0(x)) - \Psi(x, f_1(x))$. We claim that $\rho(C) = \{\rho(x) \mid x \in C\}$ consists of $|C|$ elements (i.e. ρ is one to one) and is shattered by the binary hypothesis class of homogeneous linear separators on $\mathbb{R}^d$,

$$\mathcal{H} = \{x \mapsto \mathrm{sign}(\langle w, x \rangle) \mid w \in \mathbb{R}^d\} \ .$$

Since $\mathrm{VCdim}(\mathcal{H}) = d$, it will follow that $|C| = |\rho(C)| \leq d$, as required.

To establish our claim it is enough to show that $|\mathcal{H}|_{\rho(C)}| = 2^{|C|}$. Indeed, given a subset $B \subseteq C$, by the definition of N-shattering, there exists $h_B \in \mathcal{H}_\Psi$ for which

$$\forall x \in B, h_B(x) = f_0(x) \text{ and } \forall x \in C \setminus B, h_B(x) = f_1(x) \ .$$

It follows that there exists a vector $w_B \in \mathbb{R}^d$ such that, for every $x \in B$,

$$\langle w, \Psi(x, f_0(x)) \rangle > \langle w, \Psi(x, f_1(x)) \rangle \Rightarrow \langle w, \rho(x) \rangle > 0 \ .$$

Similarly, for every $x \in C \setminus B$,

$$\langle w, \rho(x) \rangle < 0 \ .$$

It follows that the hypothesis $g_B \in \mathcal{H}$ defined by $w \in \mathbb{R}^d$ label the points in $\rho(B)$ by 1 and the points in $\rho(C \setminus B)$ by 0. It follows that if $B_1, B_2 \subseteq C$ are two different sets then $(h_{B_1})|_{\rho(C)} \neq (h_{B_2})|_{\rho(C)}$. Therefore $|\mathcal{H}|_C| = 2^{|C|}$ as required. ■

Remark 28 (Tightness of Theorem 7) *Theorem 7 is tight for some functions* $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$. *For example, consider the case that* $\mathcal{X} = [d]$, $\mathcal{Y} = \{\pm 1\}$ *and* $\Psi(x, y) = y \cdot e_x$. *It is not hard to see that* $\mathcal{H}_\Psi = \mathcal{Y}^{\mathcal{X}}$. *Therefore,* $\mathrm{Ndim}(\mathcal{H}_\Psi) = \mathrm{VCdim}(\mathcal{H}_\Psi) = d$. *On the other hand, the theorem is not tight for every* Ψ. *For example, if* $|\mathcal{X}| < d$, *then for every* Ψ, $\mathrm{Ndim}(\mathcal{H}_\Psi) \leq |\mathcal{X}| < d$.