A Theory of PAC Learnability of Partial Concept Classes

Noga Alon NALON@MATH.PRINCETON.EDU
Princeton University

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Toyota Technological Institute at Chicago

Ron Holzman HOLZMAN@TECHNION.AC.IL
Technion

Shay Moran SMORAN@TECHNION.AC.IL
Technion and Google Research

Abstract

We extend the classical theory of PAC learning in a way which allows to model a rich variety of practical learning tasks where the data satisfy special properties that ease the learning process. For example, tasks where the distance of the data from the decision boundary is bounded away from zero, or tasks where the data lie on a lower dimensional surface. The basic and simple idea is to consider *partial concepts*: these are functions that can be undefined on certain parts of the space. When learning a partial concept, we assume that the source distribution is supported only on points where the partial concept is defined.

This way, one can naturally express assumptions on the data such as lying on a lower dimensional surface, or that it satisfies margin conditions. In contrast, it is not at all clear that such assumptions can be expressed by the traditional PAC theory using learnable total concept classes, and in fact we exhibit easy-to-learn partial concept classes which provably cannot be captured by the traditional PAC theory. This also resolves, in a strong negative sense, a question posed by Attias, Kontorovich, and Mansour (2019).

We characterize PAC learnability of partial concept classes and reveal an algorithmic landscape which is fundamentally different than the classical one. For example, in the classical PAC model, learning boils down to *Empirical Risk Minimization* (ERM). This basic principle follows from *Uniform Convergence* and the *Fundamental Theorem of PAC Learning* (Vapnik and Chervonenkis, 1971, 1974a; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989).

In stark contrast, we show that the ERM principle fails spectacularly in explaining learnability of partial concept classes. In fact, we demonstrate classes that are incredibly easy to learn, but such that any algorithm that learns them must use an hypothesis space with unbounded VC dimension. We also find that the sample compression conjecture of Littlestone and Warmuth fails in this setting. Our impossibility results hinge on the recent breakthroughs in communication complexity and graph theory by Göös (2015); Ben-David, Hatami, and Tal (2017); Balodis, Ben-David, Göös, Jain, and Kothari (2021).

Thus, this theory features problems that cannot be represented in the traditional way and cannot be solved in the traditional way. We view this as evidence that it might provide insights on the nature of learnability in realistic scenarios which the classical theory fails to explain. We include in the paper suggestions for future research and open problems in several contexts, including combinatorics, geometry, and learning theory.

Keywords: PAC Learning, Learnability, VC Dimension, Margin, Online Learning, Empirical Risk Minimization

Contents

1. Introduction

In many practical learning problems the data satisfy special properties that ease the learning process. For example, imagine a learning task where the distance of the data from the decision boundary is bounded below by some margin > 0, or learning tasks where the data lie on a low-dimensional surface. (E.g. consider the task of classifying photographs of animals by whether the animal is a cat; arguably, the representations of such images lie on a low-dimensional subset of the space of all possible representations, most of which do not even represent a possible photograph.)

Common approaches for modelling such tasks often use data-dependent assumptions which are not captured by the traditional theory of PAC learning: namely, they are not expressed by a PAC learnable concept class. A classical example is the task of learning a high dimensional linear classifier with margin. Standard learning algorithms for this task, such as the classical Perceptron algorithm (Rosenblatt, 1958), use hypothesis classes which are not PAC learnable. Indeed, the Perceptron uses the hypothesis class of all linear classifiers, whose VC dimension scales linearly with the Euclidean dimension, and is therefore not PAC learnable when the dimension is unbounded. To the best of our knowledge, the same applies to all learning algorithms in this context.[1] Thus, learnability of large-margin linear classifiers is not expressed as the PAC learnability of a natural concept class.

Consequently, the general framework for data dependent analysis deviated from the traditional PAC setting while relying on additional modeling assumptions (Shawe-Taylor, Bartlett, Williamson, and Anthony, 1998; Herbrich and Williamson, 2002). Technically, this is done by introducing a data-dependent "luckiness" function which induces a (data-dependent) hierarchy of hypotheses (luckier hypotheses precede less lucky ones, as we discuss in more detail in Section 4.1). For example, in the case of large margin linear classifiers, the luckiness of each linear separator is its margin with respect to the input sample. While this framework has been successfully applied in various contexts, it does not yield a crisp notion of learnability in the spirit of PAC learning. Moreover, the general results in this framework assume rather arcane technical conditions and, while these conditions suffice for proving bounds on a case-by-case basis in various situations, it is not clear whether they are necessary in general.

To address the above shortcomings, we aim to develop a mathematical theory that is able to capture some of the above features of practical learning systems, yet admits a complete characterization of learnability in the spirit of the PAC theory. Towards this end, we take a complementary approach for modeling data-dependent assumptions: instead of modeling the algorithm's bias using a luckiness function, we extend the type of learning tasks and the notion of learnability. As will be discussed below, this provides a natural generalization of the traditional learning theory, which allows a unified treatment of data-dependent bounds and model-dependent bounds.

Partial Concepts. The basic idea is simple: rather than learning a class of concepts $\mathbb{H} \subseteq \{0,1\}^{\mathcal{X}}$, where each concept $c \in \mathbb{H}$ is a total function $c\colon \mathcal{X} \to \{0,1\}$, we consider partial concept classes $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$, where each concept c is a *partial* function; specifically, if x is such that $c(x) = \star$ then c is *undefined* at x. The *support* of a partial concept $c\colon \mathcal{X} \to \{0,1,\star\}$ is the set $\mathsf{supp}(c) := c^{-1}(\{0,1\}) = \{x \in \mathcal{X} : c(x) \neq \star\}$.

We then note that all the classical parameters such as VC dimension, Littlestone dimension, etc., naturally extend to partial concept classes without modification. In particular, a key quantity of

1. In fact, in Section 3.1 we conjecture that any algorithm that learns this task satisfies that its image, i.e. the set of hypotheses it can output, has an unbounded VC dimension.

interest in this work is the *VC dimension* of a partial concept class $\mathbb{H}$, denoted by $\mathsf{VC}(\mathbb{H})$, which is defined as the maximum size of a shattered set $U \subseteq \mathcal{X}$, where U is *shattered* if every binary pattern $u \in \{0,1\}^U$ is realized by some $h \in \mathbb{H}$ (i.e. $h|_U = u$). This allows us to express formal connections between these parameters and learnability in a unified way that also applies to partial concept classes and covers various data-dependent assumptions. For instance, the learnability of linear separators with margin then reduces to merely noting that the VC dimension of the corresponding partial concept class is bounded by a function of the margin. We note, however, that the algorithmic approach to establishing this connection is necessarily quite different from the algorithms typically used in the analysis of total concept classes, as we discuss at length below.

2. Results

In the next sections we give an overview of the main contributions in this work. Some of the formal statements rely on standard terminology (such as VC dimension, PAC learnability, etc) which is formally defined in the later technical sections.

2.1. Expressivity

Allowing for partial concepts enables modelling data-dependent assumptions in a natural way: indeed, given any such assumption, consider all legal samples S which satisfy the assumption and define the corresponding partial class of all (partial) concepts such that every sample realizable by them is legal.

For example, consider again the task of learning a γ-margin linear separator in $\mathbb{R}^N$. A sample $S \in \left(\mathbb{R}^N \times \{0,1\}\right)^n$ here is legal if the zero- and one-labelled examples are linearly separable with margin γ, and the corresponding partial class $\mathbb{H}$ consists of all partial concepts $h\colon \mathbb{R}^N \to \{0,1,\star\}$ such that $h^{-1}(0)$ and $h^{-1}(1)$ are linearly separable with margin at least γ. (See Section 3.1 for a more elaborate discussion of this partial class.) Similarly, one can easily model tasks in which the data lie on a low-dimensional subspace/manifold; such assumptions are naturally captured by partial concepts which are undefined outside some such low-dimensional subset.

In contrast, it is not at all clear that these learning tasks can be expressed in the traditional PAC model using a class of total concepts. In fact, our first result demonstrates an incredibly easy-to-learn class of partial concepts that cannot be represented by *any* learnable class of total concepts. To state this result we need to formally define when a total concept class $\bar{\mathbb{H}}$ represents a partial concept class $\mathbb{H}$: intuitively, we want that every learning task definable by $\mathbb{H}$ is also definable by $\bar{\mathbb{H}}$. Formally, let us say that a total class $\bar{\mathbb{H}} \subseteq \{0,1\}^{\mathcal{X}}$ *strongly*[2] *disambiguates* a partial class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ if every partial concept $h \in \mathbb{H}$ is extended by some total concept $\bar{h} \in \bar{\mathbb{H}}$. Namely:

$$(\forall h \in \mathbb{H})(\exists \bar{h} \in \bar{\mathbb{H}}): \quad \bar{h}(x) = h(x) \quad \text{for all } x \in \mathrm{supp}(h).$$

Theorem 1 (Partial Concepts Are More Expressive Than Total Concepts) *There exists a partial concept class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathbb{N}}$ whose VC dimension is 1 such that every total class $\bar{\mathbb{H}} \subseteq \{0,1\}^{\mathbb{N}}$ which strongly disambiguates $\mathbb{H}$ must have an infinite VC dimension, i.e. $\mathsf{VC}(\bar{\mathbb{H}}) = \infty$.*

2. We will later define a weaker notion.

As we will discuss below, the above partial class $\mathbb{H}$ is easy to learn: since its VC dimension is one, we show there is an algorithm that PAC learns whenever the examples have distribution with support contained in the support of a partial concept from $\mathbb{H}$, and the sample complexity is $O(\log(1/\delta)/\varepsilon)$, where ε, δ are the standard accuracy and confidence parameters. However, the above theorem implies that if one tries to extend each partial concept in $\mathbb{H}$ by a total concept by disambiguating the $\star$'s in it, then one must end up with a class whose VC dimension is unbounded, and hence is not PAC learnable.

Theorem 1 answers an open question posed by Attias, Kontorovich, and Mansour (2019). It might be interesting to note that our proof of it exploits a surprising connection with the theory of communication complexity. In particular, it hinges on the recent breakthroughs concerning the clique vs. independent set problem by Göös (2015); Ben-David, Hatami, and Tal (2017); Balodis, Ben-David, Göös, Jain, and Kothari (2021). We discuss this further in Section 2.4.3 below.

2.2. PAC Learnability

We next present a characterization of the PAC learnable partial concept classes. But first, we should clarify the definition of PAC learning in this context. Let us begin with the noiseless and realizable setting: intuitively, we want realizability to express the premise that the data drawn from the source distribution satisfy the data-dependent assumptions captured by the partial concept class $\mathbb{H}$. This gives rise to the following definition: a distribution P on $\mathcal{X} \times \{0,1\}$ is *realizable by* $\mathbb{H}$ if almost surely (i.e., with probability 1), a sample $S = ((x_i, y_i))_{i=1}^n \sim P^n$ (for any n) is realizable by some partial concept $h \in \mathbb{H}$: that is, $\{x_i\}_{i=1}^n \subseteq \mathrm{supp}(h)$, and $h(x_i) = y_i$ for all $i \leq n$. The connection between this definition of a realizable distribution and the one used in the classical PAC model is clarified in Lemma 33. For a partial concept h and a distribution P on $\mathcal{X} \times \{0,1\}$, we define the *prediction error*: $\mathrm{er}_P(h) := P(\{(x,y) : h(x) \neq y\})$. To be clear, this means we *always* count the case $h(x) = \star$ as a prediction mistake.

Definition 2 (PAC Learnability) *A partial concept class $\mathbb{H}$ is* PAC learnable *if, for every $\varepsilon, \delta \in (0,1)$, there exists a finite $\mathcal{M}(\varepsilon,\delta) \in \mathbb{N}$ and a learning algorithm $\mathbb{A}$ such that, for every distribution P on $\mathcal{X} \times \{0,1\}$ realizable w.r.t. $\mathbb{H}$, for $S \sim P^{\mathcal{M}(\varepsilon,\delta)}$, with probability at least $1-\delta$,*

$$\mathrm{er}_P(\mathbb{A}(S)) \leq \varepsilon.$$

The value $\mathcal{M}(\varepsilon,\delta)$ is called the sample complexity *of $\mathbb{A}$, and the* optimal *sample complexity is the minimum achievable value of $\mathcal{M}(\varepsilon,\delta)$ for every given ε, δ.*

In Section C.2 we define learnability in the agnostic case in a similar manner, using the convention that $\star$'s are always treated as errors.

We begin by addressing the following fundamental question:

Which Partial Concept Classes Are PAC Learnable and How?

The *Fundamental Theorem of PAC Learning* asserts that a total concept class $\mathbb{H}$ is PAC learnable if and only if its VC dimension is finite (Vapnik and Chervonenkis, 1974a; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989; Shalev-Shwartz and Ben-David, 2014). This theorem also yields the celebrated *Empirical Risk Minimization* principle: any algorithm which outputs an hypothesis $h \in \mathbb{H}$ which minimizes the empirical error learns $\mathbb{H}$. Such algorithms are called Empirical Risk Minimizers (ERMs). In the following theorem we show that the characterization of PAC learnability in terms of the VC dimension extends to partial concept classes:

Theorem 3 (A Characterization of PAC Learnability.) *The following statements are equivalent for any partial concept class* $\mathbb{H} \subseteq \{0, 1, \star\}^{\mathcal{X}}$.

- $\mathsf{VC}(\mathbb{H}) < \infty$.
- $\mathbb{H}$ *is PAC learnable.*
- $\mathbb{H}$ *is agnostically PAC learnable.*

It is important to note that our proof of Theorem 3 is fundamentally different from the classical uniform-convergence-based argument, and it does not yield any version of the ERM principle. (We discuss this in more detail below.) Instead, our proof hinges on a combination of sample compression and a variant of the *1-Inclusion-Graph Algorithm* due to Haussler, Littlestone, and Warmuth (1994). The obtained algorithm is transductive, in the sense that its output hypothesis is not computed explicitly: rather, given any test point, it uses the entire training set to compute its label (as is the case, e.g., for the k-Nearest Neighbor Algorithm). An interesting property of our algorithm (as well as other transductive algorithms) is that the complexity of the model (hypothesis) it outputs can increase with the size of the input sample. Below we show that in general, this property is inevitable: there exist partial classes $\mathbb{H}$ with $\mathsf{VC}(\mathbb{H}) = 1$ such that any algorithm which PAC learns them must satisfy that its range (i.e. the set of hypotheses it can output) has an unbounded VC dimension.

2.3. Failure of Traditional Learning Principles

One of the conceptual contributions of the traditional PAC learning theory is the ERM principle: any learnable class $\mathcal{H}$ is learned by any algorithm which outputs a concept $h \in \mathcal{H}$ that minimizes the empirical error on the training set. Moreover, any ERM algorithm achieves the optimal sample complexity, up to lower order factors (Vapnik and Chervonenkis, 1974a; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989). This simple principle is attractive from an algorithmic perspective as it reduces a learning problem (in which the goal is to minimize an unknown function er_P), to an optimization problem (in which the goal is to minimize the known empirical loss).

However, recent machine learning breakthroughs demonstrate important phenomena that lack explanations, and sometimes even contradict conventional wisdom (see e.g. (Zhang et al., 2017; Nagarajan and Kolter, 2019; Maennel et al., 2020; Unterthiner et al., 2020; Feldman, 2020; Brown et al., 2020)). For example, consider the modern approach of training very rich models to (and often beyond) the point of complete interpolation of the training-set. In the lens of traditional learning theory, this would constitute a clear example of overfitting; however, this approach achieves excellent results in practice when implemented in deep neural networks, as well as in other hypothesis spaces such as ensembles of decision trees, kernel machines, and minimum norm linear regressors (Belkin et al., 2019; Nakkiran et al., 2020).

One reason for the incapacity of traditional generalization theory to model modern machine learning is because the traditional theory reduces learning to an empirical risk minimization task over not-too-large hypothesis spaces. In contrast, modern algorithms typically train hypotheses with a huge number of parameters.

Thus, it is interesting to seek extensions of the classical PAC theory which necessitate alternative principles beyond ERM. Theorem 3 implies that the equivalence between finite VC dimension and PAC learnability extends to partial concept classes. However, we next demonstrate that the ERM principle has no useful analogue here. In order to address this, we first need to specify what empirical risk minimization even means in this context.

Naive ERM Fails. One natural option is to define an empirical risk minimizer to be any algorithm which outputs a partial concept $h \in \mathbb{H}$ that minimizes the empirical loss (i.e., that interpolates the input data in the realizable case). However, it is easy to see that such algorithms fail to learn even very simple classes:

Proposition 4 *There exists a partial concept class $\mathbb{H}$ with $\mathsf{VC}(\mathbb{H}) = 0$ such that any proper algorithm (i.e., which outputs a partial concept from $\mathbb{H}$) fails to PAC learn $\mathbb{H}$.*

Proof Sketch Let $n \in \mathbb{N}$ be even and consider the class $\mathbb{H} \subseteq \{0, 1, \star\}^{[n]}$ defined by

$$\mathbb{H} = \{h_A : A \subseteq [n], |A| = n/2\}, \text{ where } h_A(x) = \begin{cases} 0 & x \in A, \\ \star & x \in [n] \setminus A. \end{cases}$$

Note that $\mathbb{H}$ has VC dimension 0 and that it is trivially PAC learnable by the algorithm which always outputs the all-zero function $h_0 \equiv 0$ (which is not in $\mathbb{H}$). However, any algorithm which is restricted to outputting partial concepts from $\mathbb{H}$ (and in particular any such ERM) will fail to learn this class unless it gets at least $\Omega(n)$ examples; indeed, this follows by a similar argument as in the standard no-free-lunch argument for VC classes: let the target concept $c \in \mathbb{H}$ be drawn uniformly at random and let the marginal distribution be uniform over $\mathsf{supp}(c)$; if the learner observes fewer than $n/4$ examples, and must output a hypothesis $\hat{h}_n \in \mathbb{H}$, it must *guess* the locations of at least $n/4$ elements of $\mathsf{supp}(c)$ not observed in the data, and very likely will guess incorrectly for a constant fraction of them. An infinite variant of this construction yields a 0-dimensional class that cannot be PAC learned by any ERM: namely, on $\mathcal{X} = \mathbb{N}$, let $\mathbb{H}$ be all $\{0, \star\}$-valued functions h with, $\forall t \geq 2$, exactly 2^{t-2} points $x \in [2^t] \setminus [2^{t-1}]$ with $h(x) = 0$; then the above argument can be applied in any region $[2^t] \setminus [2^{t-1}]$ to show 2^{t-3} examples do not suffice for proper learners, for any t. ■

General ERM fails. A stronger (and natural) family of empirical risk minimization algorithms in this context are algorithms which learn $\mathbb{H}$ by performing empirical risk minimization over an appropriate class $\mathbb{H}' \subseteq \{0, 1\}^{\mathcal{X}}$. For example, for the class $\mathbb{H}$ discussed above, we can pick $\mathbb{H}' = \{h_0\}$ to be the class consisting only of the all-zero function. Observe that indeed any ERM for $\mathbb{H}'$ successfully learns $\mathbb{H}$. The existence of such an $\mathbb{H}'$ yields a reduction from PAC learning $\mathbb{H}$ to PAC learning $\mathbb{H}'$. Does the ERM principle apply in this sense? That is:

Given a partial concept class $\mathbb{H}$, does there always exist a class $\mathbb{H}'$ such that any ERM w.r.t $\mathbb{H}'$ learns $\mathbb{H}$?

Can the task of learning a given partial class $\mathbb{H}$ be reduced to the task of empirical risk minimization over some total class $\mathbb{H}'$? The following theorem provides a negative answer (Proof in Section D):

Theorem 5 (Failure of Empirical Risk Minimization) *There exists a partial concept class $\mathbb{H}$ with $\mathsf{VC}(\mathbb{H}) = 1$ such that, for any total concept class $\bar{\mathbb{H}}$, there exists an ERM algorithm for $\bar{\mathbb{H}}$ that is not a PAC learning algorithm for $\mathbb{H}$.*

The next theorem (also proved in Section D) shows that regardless of ERM, a partial concept class may require that any learning algorithm that outputs total concepts must have a large image (in the sense of VC dimension).

Theorem 6 *There exists a partial concept class $\mathbb{H}$ with $\mathsf{VC}(\mathbb{H}) = 1$ such that any learning algorithm $\mathbb{A}$ that only outputs total concepts must have image with infinite VC dimension.*

Algorithmic Principles That Complement ERM? Let us conclude this section with a suggestion for future work: *Explore for general algorithmic principles that apply in this more general setting and complement the traditional ERM Principle.* As noted above, while Theorem 3 asserts that indeed every partial VC class is PAC learnable, our proof of it does not seem to give rise to a general principle in the spirit of ERM.

2.4. The Landscape of Partial VC Classes

In this section we investigate basic properties of partial VC classes and their relationship with total classes. We begin by exhibiting two learning-theoretical differences between partial and total classes: in the contexts of sample compression (Section 2.4.1) and differentially private learning (Section 2.4.2). Then, in Section 2.4.3 we investigate the following question which is central to this work: given a partial class $\mathbb{H}$ with VC dimension d, can one find a "small" class $\bar{\mathbb{H}}$ which *disambiguates* $\mathbb{H}$? We provide negative as well as positive results in this context.

2.4.1. Sample Compression Schemes

Sample compression is a fundamental technique for proving generalization bounds. Littlestone and Warmuth (1986b) proposed it as an intuitive, algorithm-dependent, technique for establishing PAC learnability of concept classes of interest. Later works have demonstrated its usefulness in various statistical learning settings, including semi-supervised and even unsupervised learning (Graepel, Herbrich, and Shawe-Taylor, 2005; Wiener, Hanneke, and El-Yaniv, 2015; David, Moran, and Yehudayoff, 2016; Kontorovich, Sabato, and Weiss, 2017; Gottlieb, Kontorovich, and Nisnevitch, 2018; Hanneke, Kontorovich, and Sadigurschi, 2019; Ashtiani, Ben-David, Harvey, Liaw, Mehrabian, and Plan, 2020). In fact, David, Moran, and Yehudayoff (2016) established that this technique is in a sense universal by proving that learnability is equivalent to compressibility in a general and abstract learning setting.

A sample compression scheme can be seen as a protocol between a *compressor* κ and a *reconstructor* ρ (see Figure 1): the compressor gets the input sample S, from which she picks a small subsample S'. The compressor sends to the reconstructor the subsample S', along with a short binary string B of additional information: i.e., $(S', B) = \kappa(S)$. The reconstructor then, based on S' and B, outputs a concept $h = \rho(S', B)$. For a given partial concept class $\mathbb{H}$, we say (κ, ρ) is a compression scheme *for* $\mathbb{H}$ if, for all finite data sequences S realizable w.r.t. $\mathbb{H}$, the above $h = \rho(\kappa(S))$ returned by the reconstructor is correct on the entire sample S (including the examples in S that were not sent to the reconstructor). The size of the compression scheme on S is defined to be $|B| + |S'|$; the size of the compression scheme for a given sample size m is the maximum size $|B| + |S'|$ over all $S \in (\mathcal{X} \times \{0, 1\})^m$. The formal definition is given in Section A in Definition 29.

A classical example of an algorithm that can be presented as a compression scheme is the *Support Vector Machine* algorithm in $\mathbb{R}^d$. Here, the compressor sends to the reconstructor the $d + 1$ support vectors which determine the maximum margin separating hyperplane (see Figure 2).

Warmuth's $600 Sample Compression Question. Sample compression is the topic of one of the longest-standing and most well-studied open problems in learning theory:

Does every concept class $\mathbb{H}$ have a compression scheme of size $O(\mathsf{VC}(\mathbb{H}))$?

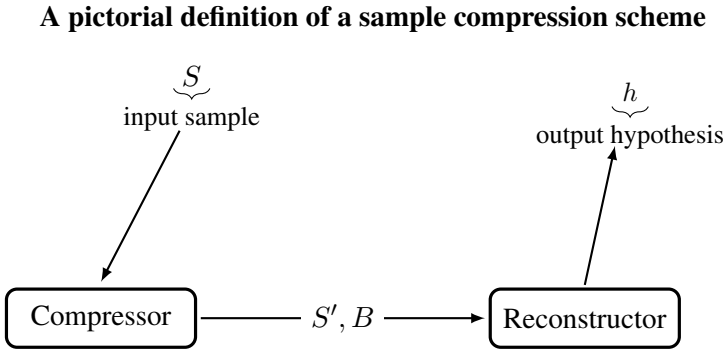

Figure 1: S' is a subsample of S and B is a binary string of additional information.

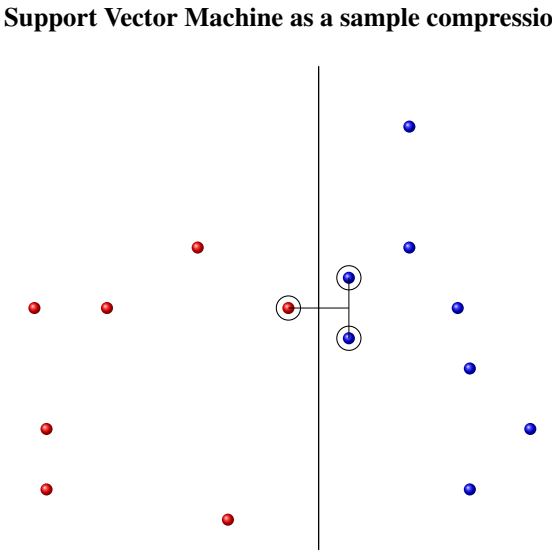

Figure 2: The input sample S consists of all red and blue points; the separating hyperplane with maximum margin is determined by the subsample of the $d+1$ *support points*. Thus, the compression scheme uses only these points.

This question has been studied since the pioneering work by Littlestone and Warmuth (1986b), and later Warmuth (2003) even announced a $600 reward for solving it! For a discussion of this question in a broader context, we refer the reader to the book by Wigderson (2019).

It is therefore interesting to explore sample compression schemes in the setting of partial concept classes. Perhaps surprisingly, it turns out that in this context the answer to the sample compression question is negative (in a strong sense). On the positive side, we show that every partial VC class has a compression scheme whose size scales logarithmically with the input sample size:

Theorem 7 (Sample Compression for Partial Concept Classes)

1. *Let $\mathbb{H}$ be a partial concept class. Then, there exists a sample compression scheme for $\mathbb{H}$ of size $\tilde{O}(\mathsf{VC}(\mathbb{H})\log(m))$, where m is the size of the input sample.*

2. *There exists a partial concept class $\mathbb{H}$ with $\mathsf{VC}(\mathbb{H}) = 1$ such that any sample compression scheme for $\mathbb{H}$ must have size $\Omega((\log(m))^{1-o(1)})$, where m is the size of the input sample, and the $o(1)$ term vanishes as $m \to \infty$. In particular, the bounded-size sample compression conjecture is **false** for partial concept classes.*

The proof of this result is in Section D. Theorem 7 demonstrates a stark difference between total and partial VC classes: Moran and Yehudayoff (2016) proved that every total VC class has a sample compression scheme whose size is bounded by a function of the VC dimension. By Item 2 above, this result does not extend to partial VC classes, even with VC dimension one.

2.4.2. Littlestone Dimension vs Private Learning

Differentially private PAC learning is an additional setting which demonstrates a curious difference between partial classes and total classes.

Differential privacy (DP) (Dwork, McSherry, Nissim, and Smith, 2006) is a sound theoretical approach to reason about privacy in a precise and quantifiable fashion. It has become the gold standard of statistical data privacy (Dwork and Roth, 2014) and been implemented in practice, notably by Google (Erlingsson, Pihur, and Korolova, 2014), Apple (app, 2016a,b), and in the 2020 US census (Dajani, Lauger, Singer, Kifer, Reiter, Machanava-jjhala, Garfinkel, Dahl, Graham, Karwa, Kim, Lelerc, Schmutte, Sexton, Vilhuber, and Abowd).

A recent line of work revealed a qualitative characterization of DP-learnability in the PAC model: A total concept class $\mathbb{H}$ can be PAC learned by a DP-algorithm if and only if its *Littlestone dimension* $\mathsf{LD}(\mathbb{H})$ is finite (Alon, Livni, Malliaris, and Moran, 2019; Gonen, Hazan, and Moran, 2019; Bun, Livni, and Moran, 2020; Ghazi, Golowich, Kumar, and Manurangsi, 2020). (The Littlestone dimension is a combinatorial parameter which arises in the context of online learning, see Section A for a formal definition.) It is therefore natural to ask whether this characterization extends to partial concept classes:

Open Question 1 (Private PAC Learnability) *Does the characterization of differentially private PAC learning extend to partial classes?: Let $\mathbb{H}$ be a partial class. Is it the case that $\mathbb{H}$ is PAC learnable by a differentially private algorithm if and only if it has a finite Littlestone dimension?*

Despite the fact that natural partial classes with finite Littlestone dimension are known to be DP learnable (e.g. halfspaces with margin (Nguyen, Ullman, and Zakynthinou, 2020)), it is not clear

how to generally prove either of the implications "$\mathsf{LD}(\mathbb{H}) < \infty \implies \mathbb{H}$ is DP-learnable" or "$\mathbb{H}$ is DP-learnable $\implies \mathsf{LD}(\mathbb{H}) < \infty$".

The known proofs of the direction "$\mathsf{LD}(\mathbb{H}) < \infty \implies \mathbb{H}$ is DP-learnable" for total concept classes (Bun, Livni, and Moran, 2020; Ghazi, Golowich, Kumar, and Manurangsi, 2020) utilize (among other things) the ERM principle and uniform convergence which, as discussed earlier, is not satisfied by partial concept classes.

As for the direction "$\mathsf{LD}(\mathbb{H}) = \infty \implies \mathbb{H}$ is not DP-learnable ", the very first step of the proof by Alon, Livni, Malliaris, and Moran (2019) fails for partial classes: the proof proceeds by first reducing an arbitrary class with an unbounded Littlestone dimension to the class of one-dimensional thresholds, and then proving that one-dimensional thresholds are not privately PAC learnable.

The reduction to one-dimensional thresholds boils down to a combinatorial parameter called the *threshold dimension*: the threshold dimension of a class $\mathbb{H}$, denoted by $\mathrm{TD}(\mathbb{H})$, is the maximum integer d for which there exist $x_1, \ldots, x_d \in \mathcal{X}$ and $h_1, \ldots h_d \in \mathbb{H}$ such that $h_i(x_j) = \mathbb{1}[i \leq j]$. For total concept classes $\mathbb{H}$ it is known[3] $\mathrm{TD}(\mathbb{H}) \geq \lfloor \log(\mathsf{LD}(\mathbb{H})) \rfloor$; this essentially implies that any class with a large Littlestone dimension contains a large subclass of thresholds. Interestingly, this relation fails to extend to partial classes, as shown in the next theorem (proved in Section D):

Theorem 8 *There exists a partial concept class $\mathbb{H}$ with $\mathrm{TD}(\mathbb{H}) \leq 2$ but $\mathsf{LD}(\mathbb{H}) = \infty$.*

2.4.3. Disambiguations

We next present one of the main focuses of this work which concerns the following questions: Can partial VC classes be represented by total VC classes? Relatedly, can one reduce the task of learning a given partial VC class to the task of learning a total VC class? We begin with the following central definition of disambiguation:

Definition 9 (Disambiguation) *A total concept class $\bar{\mathbb{H}}$ is a special type of partial concept class such that every $h \in \bar{\mathbb{H}}$ has range $\{0,1\}$: i.e., is a total concept. A total concept class $\bar{\mathbb{H}} \subseteq \{0,1\}^{\mathcal{X}}$ is said to* disambiguate *a partial concept class $\mathbb{H}$ if every finite data sequence $S \in (\mathcal{X} \times \{0,1\})^*$ realizable w.r.t. $\mathbb{H}$ is also realizable w.r.t. $\bar{\mathbb{H}}$. In this case, $\bar{\mathbb{H}}$ is called a* disambiguation *of $\mathbb{H}$.*

Note the difference between Definition 9 and the definition used in Theorem 1: the latter poses a stricter requirement, namely that each partial concept in $\mathbb{H}$ is extended by some total concept in $\bar{\mathbb{H}}$. We note that the two definitions are equivalent when $\mathcal{X}$ is finite (more generally, when $\mathrm{supp}(h)$ is finite for every $h \in \mathbb{H}$), and are essentially equivalent when $\mathcal{X}$ is countable.[4]

Definition 9 is more suitable in the context of learning because it suffices to guarantee that every PAC learner for $\bar{\mathbb{H}}$ is a PAC learner for $\mathbb{H}$, and hence reduces the task of PAC learning the partial class $\mathbb{H}$ to PAC learning the total class $\bar{\mathbb{H}}$. One could further relax Definition 9 by allowing errors and by only requiring to disambiguate short samples, in a way that implies that a learner for $\bar{\mathbb{H}}$ is a weak learner for $\mathbb{H}$. However, the next proposition implies that such relaxations are essentially equivalent to Definition 9.

3. It is also known that $\mathsf{LD}(\mathbb{H}) \geq \lfloor \log(\mathrm{TD}(\mathbb{H})) \rfloor$, but this inequality extends also to partial classes with the same proof (see Alon, Livni, Malliaris, and Moran (2019)).
4. In the sense that whenever $\mathbb{H}$ can be disambiguated according to the weaker definition by $\bar{\mathbb{H}}$ then it can also be disambiguated according to the stronger definition by $\bar{\mathbb{H}}'$ such that $\mathsf{VC}(\bar{\mathbb{H}}) = \mathsf{VC}(\bar{\mathbb{H}}')$.

Proposition 10 (Approximate Disambiguation $\implies$ Disambiguation) *Let $\mathbb{H}$ be a partial class and let $\gamma > 0$. Assume that there exists a total class $\bar{\mathbb{H}}$ with $\mathsf{VC}(\bar{\mathbb{H}}) = d$ that "weakly disambiguates" $\mathbb{H}$ in the following sense: for every sample $S = \{(x_i, y_i)\}_{i=1}^n$ realizable by $\mathbb{H}$ of size $|S| = n = O(\frac{d}{\gamma^2})$ there exists $\bar{h} \in \bar{\mathbb{H}}$ such that*

$$\hat{\mathrm{er}}_S(\bar{h}) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[\bar{h}(x_i) \neq y_i] \leq \frac{1-\gamma}{2}.$$

Then, $\mathbb{H}$ can be disambiguated (in the sense of Definition 9) by a total class whose VC-dimension is at most $\tilde{O}(\frac{d \cdot d^\star}{\gamma^2})$, where $d^\star \leq 2^{d+1}$ is the dual VC dimension of $\bar{\mathbb{H}}$.

Proposition 10 might be viewed as a kind of compactness theorem; for example, it implies that in order to disambiguate $\mathbb{H}$ it suffices to represent only the samples realizable by $\mathbb{H}$ of size at most $100d$ by a total class $\bar{\mathbb{H}}$ with $\mathsf{VC}(\bar{\mathbb{H}}) = d$. The proofs of all the results in this subsection appear in Section B.

The following result demonstrates a one-dimensional class which cannot be disambiguated while retaining a bounded VC dimension.

Theorem 11 (A VC Class Which Cannot be Disambiguated) *For any $n \in \mathbb{N}$ there exists a partial concept class $\mathbb{H}_n \subseteq \{0,1,\star\}^{[n]}$ with $\mathsf{VC}(\mathbb{H}_n) = 1$ and $\mathrm{TD}(\mathbb{H}_n) \leq 2$ such that any disambiguation $\bar{\mathbb{H}}$ of $\mathbb{H}_n$ has size at least $n^{(\log(n))^{1-o(1)}}$, where the $o(1)$ term tends to 0 as $n \to \infty$. In particular, this implies $\mathsf{LD}(\bar{\mathbb{H}}) \geq \mathsf{VC}(\bar{\mathbb{H}}) \geq (\log(n))^{1-o(1)}$, and shows that for infinite $\mathcal{X}$ there exists $\mathbb{H}_\infty \subseteq \{0,1,\star\}^{\mathcal{X}}$ with $\mathsf{VC}(\mathbb{H}_\infty) = 1$ and $\mathrm{TD}(\mathbb{H}_\infty) \leq 2$, while $\mathsf{LD}(\bar{\mathbb{H}}) = \mathsf{VC}(\bar{\mathbb{H}}) = \infty$ for every disambiguation $\bar{\mathbb{H}}$ of $\mathbb{H}_\infty$.*

Below, in Theorem 12 we show that the bound in Theorem 11 is nearly tight. Theorem 11 resolves, in a strong negative sense, an open problem presented by Attias, Kontorovich, and Mansour (2019), which sought a disambiguation whose VC dimension is bounded by a (linear) function of $\mathsf{VC}(\mathbb{H})$. Further, Theorem 11 is our workhorse for proving the impossibility results discussed in the previous sections regarding expressivity (Theorem 1) the failure of the ERM principle (Theorem 5), the image of any learning algorithm (Theorem 6), sample compression schemes (Theorem 7), and private PAC learning (Theorem 8).

Interestingly, its proof hinges on a recent breakthrough in communication complexity and its implications in graph theory: Göös (2015); Ben-David, Hatami, and Tal (2017); Balodis, Ben-David, Göös, Jain, and Kothari (2021). Despite the advantage that our proof of Theorem 11 is short and simple, it unfortunately provides only little insight on the structure of the concluded class $\mathbb{H}$. In part, this is due to the complexity of the relevant result in graph theory, which is obtained by a series of reductions, some of which are unintuitive. It will be interesting to exhibit a natural partial VC class which demonstrates this separation. Towards this end, we propose a geometric candidate in Section 3.1.

A Sauer-Shelah-Perles Lemma for Partial VC Classes? So far we discussed several differences between partial and total VC classes. All of these differences boil down to Theorem 11. We next investigate which properties of total VC classes are *retained* by partial classes.

Arguably the most basic property of VC classes is manifested by the Sauer-Shelah-Perles Lemma (SSP) (Sauer, 1972). This lemma bounds the cardinality of a class $\mathbb{H} \subseteq \{0,1\}^n$ with $\mathsf{VC}(\mathbb{H}) = d$

by

$$|\mathbb{H}| \leq \binom{n}{\leq d}.$$

Is there an analogue of the SSP Lemma for partial classes? An immediate and direct generalization of it to partial classes would be that $|\mathbb{H}| \leq \binom{n}{\leq \mathsf{VC}(\mathbb{H})}$ for every partial class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$. However it is easy to see that this is false, as witnessed e.g. by the class $\mathbb{H} = \{0,\star\}^n$ which satisfies $\mathsf{VC}(\mathbb{H}) = 0$ and $|\mathbb{H}| = 2^n$. A more mature candidate for extending the SSP Lemma to partial classes is via disambiguations:

Can every $\mathbb{H} \subseteq \{0,1,\star\}^n$ be disambiguated by a total class $\bar{\mathbb{H}} \subseteq \{0,1\}^n$ such that $|\bar{\mathbb{H}}| \leq \binom{n}{\leq \mathsf{VC}(\mathbb{H})}$?

Indeed, the above class $\mathbb{H} = \{0,\star\}^n$ is disambiguated by $\bar{\mathbb{H}} = \{0^n\}$ which satisfies this inequality. Unfortunately, Theorem 11 also refutes this version: it demonstrates a one-dimensional class such that every disambiguating class has size which is at least a quasipolynomial in n. On the positive side, it turns out that Theorem 11 is the only obstacle for this version in the sense that relaxing the polynomial bound to a quasipolynomial one works:

Theorem 12 (Quasipolynomial Sauer-Shelah-Perles Lemma) *Let $\mathbb{H}$ be a partial concept class on a finite $\mathcal{X}$ with $\mathsf{VC}(\mathbb{H}) = d$. Then there exists a disambiguation $\bar{\mathbb{H}}$ of $\mathbb{H}$ of size*

$$|\bar{\mathbb{H}}| = |\mathcal{X}|^{O(d \log(|\mathcal{X}|))}.$$

Like the SSP Lemma, also Theorem 12 yields a dichotomy for partial classes: for every partial class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$, either there are arbitrarily large finite $\mathcal{X}' \subseteq \mathcal{X}$ of size n such that any disambiguation $\bar{\mathbb{H}}$ of $\mathbb{H}|_{\mathcal{X}'}$ has size

$$|\bar{\mathbb{H}}| = 2^n,$$

or there exists a polynomial poly such that for every finite $\mathcal{X}' \subseteq \mathcal{X}$ of size n there exists a disambiguation $\bar{\mathbb{H}}$ of $\mathbb{H}|_{\mathcal{X}'}$ whose size

$$|\bar{\mathbb{H}}| \leq \mathsf{poly}(n^{\log(n)}).$$

Note that in the latter case, the disambiguation $\bar{\mathbb{H}}$ depends on $\mathcal{X}'$. Further, Theorem 11 implies that such dependence is, in general, necessary; that is, there cannot be a single universal disambiguation $\bar{\mathbb{H}}$ of $\mathbb{H}$ satisfying $|\bar{\mathbb{H}}|_{\mathcal{X}'}| = o(2^n)$ for all finite $\mathcal{X}' \subseteq \mathcal{X}$, where $n = |\mathcal{X}'|$. Indeed, such an $\bar{\mathbb{H}}$ would have a finite VC dimension, which would contradict Theorem 11. Nevertheless, the next result (of which Theorem 12 is a corollary) shows that it is possible to (strongly) disambiguate any partial VC class $\mathbb{H}$ while maintaining a quasipolynomial bound on the growth function for initial finite subsets $\mathcal{X}' \subseteq \mathcal{X}$:

Theorem 13 *Let $\mathcal{X} = \mathbb{N} = \{1,2,\ldots\}$ and let $\mathbb{H}$ be a partial concept class on $\mathcal{X}$ with $\mathsf{VC}(\mathbb{H}) = d < \infty$. Then there exists a strong disambiguation $\bar{\mathbb{H}}$ of $\mathbb{H}$, so that for every finite m, the projection of $\bar{\mathbb{H}}$ on $[m]$ has size at most $(m+1)^{(d+1)\log_2(m)+2} = m^{O(d\log(m))}$.*

This result is new; however, after expressing it to others in personal communications, it has recently been applied in the work of Attias, Kontorovich, and Mansour (2021) in order to prove a bound on the fat-shattering dimension of k-fold maxima of real valued function classes.

Let us conclude this discussion about disambiguations and the SSP Lemma with a question:

Open Question 2 (Polynomial Growth $\implies$ Disambiguation?) *Let $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ and assume there exists a polynomial* poly *such that for every finite $\mathcal{X}' \subseteq \mathcal{X}$ there exists a disambiguation $\bar{\mathbb{H}}$ of $\mathbb{H}|_{\mathcal{X}'}$ of size $|\bar{\mathbb{H}}| \leq \mathsf{poly}(n)$, where $n = |\mathcal{X}'|$. Does there exist a disambiguation $\bar{\mathbb{H}}$ of $\mathbb{H}$ such that $\mathsf{VC}(\bar{\mathbb{H}}) < \infty$?*

We next discuss two techniques for disambiguating which will be useful in our proofs.

Disambiguating by Sample-Compression. Sample compression schemes naturally imply disambiguations: indeed, consider a partial concept class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ over a finite domain of size $|\mathcal{X}| = n$, and assume we are given a sample compression scheme for $\mathbb{H}$ of size k. Therefore, for any partial concept $h \in \mathbb{H}$ there exist $x_1, \ldots, x_k \in \mathrm{supp}(h)$ such that h is extended by the total concept

$$\bar{h} = \rho\Big((x_i, h(x_i))_{i=1}^{k}, B\Big),$$

where ρ is the reconstruction function of the compression scheme, and B is a bit-string of side information of length at most k. In particular, by applying ρ on all such sequences of length at most k and all such bit-strings B, we obtain a disambiguation of $\mathbb{H}$ of size $n^{O(k)}$. This is summarized in the following proposition:

Proposition 14 *For any finite $\mathcal{X}$ and any partial concept class $\mathbb{H}$, if $\mathbb{H}$ has a compression scheme of size k, there exists a disambiguation $\bar{\mathbb{H}}$ of $\mathbb{H}$ of size at most $(c|\mathcal{X}|/k)^k$ for a numerical constant c.*

Disambiguating by Majority-Votes. We conclude this section with highlighting one idea which is used in the proofs of Theorems 12 and 13. Let $\mathbb{H} \subseteq \{0,1,\star\}^n$ be a partial class. Consider an online learning setting in which an adversary picks a target partial concept $h \in \mathbb{H}$, and then in each round $i = 1, \ldots, n$, the learner first guesses a label $\hat{y}_i$. Then, if $i \in \mathrm{supp}(h)$ and $\hat{y}_i \neq h(i)$ then the learner is given the correct value $h(i)$. (Otherwise, if $i \notin \mathrm{supp}(h)$ or $\hat{y}_i = h(i)$ then the learner gets no feedback.) Notice that a learner which makes at most k mistakes, in the worst case over all $h \in \mathbb{H}$, defines a disambiguation $\bar{\mathbb{H}}$ of $\mathbb{H}$ whose size $|\bar{\mathbb{H}}|$ is at most $\binom{n}{\leq k}$.

Our proofs follow by exhibiting a learner which makes at most $O(\mathsf{VC}(\mathbb{H})\log(n))$ mistakes. This is done by considering a kind of majority-vote using the family of sets which are shattered by $\mathbb{H}$. We refer the reader to Section B for more details.

2.5. Online Learning

We conclude Section 2 with a characterization of online learnability. The following theorem shows that the Littlestone dimension retains its role of characterzing online learnability. See Section E for a precise definition of the online learning setting, in both the realizable (mistake-bound) case and the agnostic (regret-bound) case, along with the formal proof, and more-detailed quantitative results.

Theorem 15 *The following statements are equivalent for a partial concept class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$.*

- $\mathsf{LD}(\mathbb{H}) < \infty$.
- *$\mathbb{H}$ is online learnable in the realizable (mistake-bound) setting.*
- *$\mathbb{H}$ is online learnable in the agnostic (regret-bound) setting.*

Like partial VC classes, also partial classes with finite Littlestone dimension (= Littlestone classes) exhibit different behaviour from their total counterparts. One such example was discussed in Section 2.4.2. However, our understanding of partial Littlestone classes is more limited. In particular, we conclude with the following basic question:

Open Question 3 *Let $\mathbb{H}$ be a partial class with $\mathsf{LD}(\mathbb{H}) < \infty$. Does there exist a disambiguation of $\mathbb{H}$ by a total class $\bar{\mathbb{H}}$ such that $\mathsf{LD}(\bar{\mathbb{H}}) < \infty$? Is there one with $\mathsf{VC}(\bar{\mathbb{H}}) < \infty$?*

We remark that if the answer to Open Question 2 is affirmative, then so is the answer to the second part (about VC dimension) of Open Question 3. This follows because Littlestone classes can be disambiguated using the SOA algorithm (see Appendix E).

3. Three Examples and Two Open Questions

We next present three examples of partial concept classes which capture the well-studied learning tasks corresponding to linear classification with margin guarantees, boosting, and general classifiers with margin. We also pose two open problems regarding disambiguating these classes.

3.1. Geometric Margin

We next demonstrate the expressivity of partial concepts by presenting the classical results regarding learnability of linear classifiers with margin as the PAC learnability of a partial concept class. Since this basic result cannot be expressed as the PAC learnability of a natural (total) concept class, its presentation in introductory classes to machine learning usually deviates from the classical PAC learning theory. Thus, this demonstrates a possible didactic value of the theory of partial concept classes.

Let V be a (possibly infinite dimensional) real Hilbert space, and let $R, \gamma > 0$ be the margin parameters.

Definition 16 (Separability with Margin) *A sample $(x_1, y_1), \ldots, (x_n, y_n) \in V \times \{0, 1\}$ is (R, γ)-separable if:*

1. *there exists a ball $B \subseteq V$ of radius R such that $x_1, \ldots, x_n \in B$, and*
2. *the distance between the convex hull of $\{x_i : y_i = 1\}$ and the convex hull of $\{x_i : y_i = 0\}$ is at least 2γ.*

In other words, a sample is (R, γ)-separable, if the 0-labelled examples and 1-labelled examples can be separated by a linear classifier with margin γ and all examples lie in a ball of radius R.

Let $\mathbb{H}_{R,\gamma}$ denote the class

$$\mathbb{H}_{R,\gamma} := \Big\{ h \in \{0, 1, \star\}^V : \big(\forall x_1, \ldots, x_n \in \mathrm{supp}(h)\big) : \\ (x_1, h(x_1)), \ldots, (x_n, h(x_n)) \text{ is } (R, \gamma)\text{-separable} \Big\}.$$

The following proposition provides tight bounds on the VC dimension and the Littlestone dimension of $\mathbb{H}_{R,\gamma}$ (in order to focus on the parameters R, γ and not on the dimension of V, we assume that the latter is large, specifically $\dim(V) \geq R^2/\gamma^2$). It is based on classical results concerning linear classifiers with margin, dating back to Rosenblatt (1958).

Proposition 17 *For all* $\gamma, R > 0$*:* $\mathsf{VC}(\mathbb{H}_{R,\gamma}) = \Theta\left(\frac{R^2}{\gamma^2}\right)$ *and* $\mathsf{LD}(\mathbb{H}_{R,\gamma}) = \Theta\left(\frac{R^2}{\gamma^2}\right)$.

Proof Since $\mathsf{VC} \leq \mathsf{LD}$, it suffices to show that

$$\mathsf{VC}(\mathbb{H}_{R,\gamma}) = \Omega\left(\frac{R^2}{\gamma^2}\right) \quad \text{and} \quad \mathsf{LD}(\mathbb{H}_{R,\gamma}) = O\left(\frac{R^2}{\gamma^2}\right).$$

The upper bound on $\mathsf{LD}(\mathbb{H}_{R,\gamma})$ follows by the classical mistake-bound analysis of the Perceptron algorithm (Rosenblatt, 1958), which implies that $\mathbb{H}_{R,\gamma}$ is online learnable in the realizable setting with at most $O\left(\frac{R^2}{\gamma^2}\right)$ mistakes, and therefore $\mathsf{LD}(\mathbb{H}_{R,\gamma}) = O\left(\frac{R^2}{\gamma^2}\right)$.

To obtain a lower bound on $\mathsf{VC}(\mathbb{H}_{R,\gamma})$, let $e_1, e_2, \ldots$ be an orthonormal basis for V, and consider the set

$$C = \left\{ Re_i : i \leq \frac{R^2}{\gamma^2} \right\}.$$

Note that C is shattered: indeed, C is contained in the ball of radius R centered at the origin, and for every partition of $\{i : i \leq R^2/\gamma^2\}$ into two sets A, B, let w denote the vector

$$w = \frac{\gamma}{R}\left(\sum_{i \in A} e_i - \sum_{i \in B} e_i\right).$$

Note that $\|w\|^2 = \frac{\gamma^2}{R^2}(|A| + |B|) \leq 1$ and that $w \cdot Re_i = \gamma$ for $i \in A$ and $w \cdot Re_i = -\gamma$ for $i \in B$. Thus, w witnesses that the distance between the convex-hull of $\{Re_i : i \in A\}$ and the convex-hull of $\{Re_i : i \in B\}$ is $\geq 2\gamma$. ■

We conclude this example with an open question: Can learnability of linear classifiers under margin assumptions be modeled by the PAC learnability of a total concept class?

Open Question 4 *Does there exist a disambiguation of* $\mathbb{H}_{R,\gamma}$ *by a total class* $\bar{\mathbb{H}} \subseteq \{0,1\}^V$ *whose VC/Littlestone dimensions are bounded by a function of* R, γ*?*

It seems plausible that the answer to this question is no: in particular, our attempts to find "natural" (geometrically defined) disambiguations resulted with classes whose VC dimension depends on the dimension of the underlying Hilbert space. Note that if the answer here is indeed negative, then so is the answer to Open Questions 2 and 3.

3.2. Boosting

Boosting is a celebrated machine learning approach which is based on the idea of combining weak and moderately inaccurate hypotheses to a strong and accurate one. The following example concerns boosting under the assumption that the weak hypotheses belong to a class of bounded capacity. This setting was explored in detail by Alon, Gonen, Hazan, and Moran (2020), and is inspired by the common understanding that weak hypotheses are "rules-of-thumbs" from an "easy-to-learn class". (Schapire and Freund '12, Shalev-Shwartz and Ben-David '14.) Formally, it is assumed the class of weak hypotheses has a bounded VC dimension.

One of the main goals addressed by Alon, Gonen, Hazan, and Moran (2020) is to characterize which target concepts can be learned by boosting weak hypotheses from a given base-class $\mathcal{B}$. As

we will now demonstrate, *the setting introduced by Alon, Gonen, Hazan, and Moran (2020) can be naturally expressed by partial concept classes.*

The starting point of Alon, Gonen, Hazan, and Moran (2020) is a reformulation of the weak learnability assumption: Recall that the γ-weak learnability assumption asserts that if $c : \mathcal{X} \to \{0,1\}$ is the target concept then, if the weak learner is given enough c-labeled examples drawn from any input distribution over $\mathcal{X}$, it will return an hypothesis which is γ-correlated with c. One can rephrase the weak learnability assumption only in terms of $\mathcal{B}$ using the following notion:[5]

Definition 18 (γ-realizable samples (Alon, Gonen, Hazan, and Moran (2020))) *Let $\mathcal{B} \subseteq \{0,1\}^{\mathcal{X}}$ be the base-class and let $\gamma \in (0,1]$. A sample $S = ((x_1,y_1),\ldots,(x_n,y_n))$ is γ-realizable with respect to $\mathcal{B}$ if for any probability distribution D over S there exists $b \in \mathcal{B}$ such that*

$$\Pr_{(x,y)\sim D}[b(x) \neq y] \leq \frac{1-\gamma}{2}.$$

Note that for $\gamma = 1$ the notion of γ-realizability specializes to the classical notion of realizability (i.e., consistency with the class). Also note that as $\gamma \to 0$, the set of γ-realizable samples becomes larger.

Using this notion one can describe the (partial) class of concepts which can be learned by boosting γ-accurate hypotheses from $\mathcal{B}$. We denote this class by $\mathbb{H}_\gamma$ and it is defined as follows:

$$\mathbb{H}_\gamma = \Big\{ h \in \{0,1,\star\}^{\mathcal{X}} : \big(\forall x_1,\ldots,x_n \in \mathrm{supp}(h)\big) : \\ (x_1,h(x_1)),\ldots,(x_n,h(x_n)) \text{ is } \gamma\text{-realizable by } \mathcal{B} \Big\}.$$

Although Alon, Gonen, Hazan, and Moran (2020) lacked the terminology of partial concept classes, they explicitly studied bounds on the VC dimension of $\mathbb{H}_\gamma$ (which they denoted by γ-VC dimension). They provided the following upper bounds:

Theorem 19 (Alon, Gonen, Hazan, and Moran (2020)) *Let $\mathcal{B}$ be a class with VC dimension d, and let $\gamma > 0$. Then, the following upper bounds on $\mathsf{VC}(\mathbb{H}_\gamma)$ hold:*

$$\mathsf{VC}(\mathbb{H}_\gamma) = O\left(\frac{d}{\gamma^2}\log(d/\gamma)\right) = \tilde{O}\Big(\frac{d}{\gamma^2}\Big),$$

and

$$\mathsf{VC}(\mathbb{H}_\gamma) = O_d\left(\frac{1}{\gamma^{\frac{2d}{d+1}}}\right),$$

where $O_d(\cdot)$ conceals a multiplicative constant that depends only on d.

Alon, Gonen, Hazan, and Moran (2020) further demonstrated base-classes $\mathcal{B}$ which imply tightness in some ranges of the parameters γ, d. However, in general it remains open to establish tight bounds on $\mathsf{VC}(\mathbb{H}_\gamma)$ in both γ, d.

5. In fact, γ-realizability corresponds to the *empirical weak learning assumption* by Schapire and Freund (2012)[Chapter 2.3.3]. The latter is a weakening of the standard weak PAC learning assumption which suffices to guarantee generalization.

A Disambiguation. In contrast with the previous example of linear classifiers, here we can prove that there exists a disambiguation with a bounded VC dimension:

Theorem 20 (A Disambiguation) *Let $\mathcal{B}$ be a class with VC dimension d, and let $\gamma > 0$. Then, there exists a disambiguation of $\mathbb{H}_\gamma$ by a total concept class $\bar{\mathbb{H}}_\gamma$, such that*

$$\mathsf{VC}\left(\bar{\mathbb{H}}_\gamma\right) = \tilde{O}\left(\frac{d \cdot d^\star}{\gamma^2}\right),$$

where $d^\star \leq 2^{d+1}$ is the dual VC dimension of $\mathcal{B}$.

Proof Let $k = O(\frac{d^\star}{\gamma^2})$. Define $\bar{\mathbb{H}}_\gamma$ to be the class of all majority-votes of k hypotheses from $\mathcal{B}$. By standard bounds on the VC dimension of composed classes we have $\mathsf{VC}(\bar{\mathbb{H}}_\gamma) = \tilde{O}\left(\frac{d \cdot d^\star}{\gamma^2}\right)$.

It remains to show that $\bar{\mathbb{H}}_\gamma$ disambiguates $\mathbb{H}_\gamma$. This follows by a standard combination of a Minimax argument and uniform convergence (a.k.a ε-approximation): let $(x_1, y_1), \ldots, (x_n, y_n)$ be a sample realizable by $\mathbb{H}_\gamma$. Thus, for every distribution D over $\{(x_i, y_i)\}_{i=1}^n$ there exists a weak-hypothesis $b \in \mathcal{B}$ such that

$$\Pr_{(x,y)\sim D}[b(x) \neq y] \leq \frac{1-\gamma}{2}.$$

By the Minimax Theorem (von Neumann and Morgenstern, 1944), there exists a distribution P over $\mathcal{B}$ such that

$$\big(\forall (x_i, y_i)\big) : \Pr_{b\sim P}[b(x_i) \neq y_i] \leq \frac{1-\gamma}{2}.$$

Thus, by the uniform convergence theorem (Vapnik and Chervonenkis, 1968) applied to the dual class $\mathcal{B}^\star$, it follows that with positive probability, the majority vote of a sequence of independently drawn $b_1, \ldots, b_k \sim P$ (where $k = O(\frac{d^\star}{\gamma^2})$) satisfies:

$$\big(\forall (x_i, y_i)\big) : \quad \big(\mathsf{Majority}(b_1, \ldots, b_k)\big)(x_i) = y_i,$$

as required. ■

Note that the dual VC dimension $d^\star$ can be exponential in the VC dimension. Thus, the following question remains:

Open Question 5 *Let $\mathcal{B}$ be a class with VC dimension d, and let $\gamma > 0$. Does there exist a disambiguation of $\mathbb{H}_\gamma$ by a total class $\bar{\mathbb{H}} \subseteq \{0,1\}^{\mathcal{X}}$ whose VC dimension is bounded by a polynomial in d, γ^{-1}?*

3.3. General Separators with Margin

As a final example, we present an example of a partial concept class that *can* be disambiguated without significantly increasing the VC dimension. Specifically, consider the case $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ for $d \in \mathbb{N}$, and for $\gamma > 0$ let $\mathbb{G}_{d,\gamma}$ be the set of *all* partial functions $h : \mathcal{X} \to \{0, 1, \star\}$ with $\min(\{\|x_0 - x_1\| : x_0, x_1 \in \mathcal{X}, h(x_0) = 0, h(x_1) = 1\} \cup \{\infty\}) \geq \gamma$: that is, $\mathbb{G}_{d,\gamma}$ is the set of all partial functions having a margin γ separation between all points labeled 0 and all points labeled 1. This class effectively arises in many works (e.g., von Luxburg and Bousquet, 2004; Gottlieb, Kontorovich, and Nisnevitch, 2018), where it is typically expressed as a margin condition on a data set

(or, equivalently, a Lipschitz constant for the smoothest real-valued function that fits the data). For distributions P producing data sets satisfying this separation, there are immediate implications for prediction error bounds for various simple neighborhood-based prediction algorithms such as the *nearest neighbor* algorithm (e.g., Cover and Hart, 1967; Chaudhuri and Dasgupta, 2014; Urner and Ben-David, 2013).

For this class $\mathbb{G}_{d,\gamma}$, we first observe that its VC dimension is roughly γ^{-d}. Specifically, let $M_d(\gamma)$ be the γ-packing number of $\mathcal{X}$: that is, the largest number m s.t there exist $x_1, \ldots, x_m \in \mathcal{X}$ with $\min_{i \neq j} \|x_i - x_j\| \geq \gamma$. Then we have the following proposition.

Proposition 21 $\mathsf{VC}(\mathbb{G}_{d,\gamma}) = M_d(\gamma)$. *In particular, there exist numerical constants* $c, C \in (0, \infty)$ *such that* $\left(\frac{c}{\gamma}\right)^d \leq \mathsf{VC}(\mathbb{G}_{d,\gamma}) \leq \left(\frac{C}{\gamma}\right)^d$.

Proof Let $m = M_d(\gamma)$. To show a lower bound on $\mathsf{VC}(\mathbb{G}_{d,\gamma})$, let $x_1, \ldots, x_m$ be any γ-packing of $\mathcal{X}$. Since any classification of these points has its closest 1 and 0 -labeled points at distance $\geq \gamma$, it follows that $x_1, \ldots, x_m$ is shattered by $\mathbb{G}_{d,\gamma}$; indeed, $\mathbb{G}_{d,\gamma}$ contains 2^m functions h with $\mathrm{supp}(h) = \{x_1, \ldots, x_m\}$ which witness the shattering. Thus, $\mathsf{VC}(\mathbb{G}_{d,\gamma}) \geq m$. To show an upper bound, for $n \in \mathbb{N}$ and a sequence $x_1, \ldots, x_n$, for $(i^*, j^*) = \mathrm{argmin}_{(i,j):i \neq j} \|x_i - x_j\|$, if $\|x_{i^*} - x_{j^*}\| < \gamma$, then for any $y_1, \ldots, y_n \in \{0, 1\}$ such that $(x_1, y_1), \ldots, (x_n, y_n)$ is realizable w.r.t. $\mathbb{G}_{d,\gamma}$, it must be that $y_{i^*} = y_{j^*}$, and hence $x_1, \ldots, x_n$ is not shattered by $\mathbb{G}_{d,\gamma}$. Thus, *every* shattered set is γ-separated. Since m is the maximum size of a γ-separated set, it follows that $\mathsf{VC}(\mathbb{G}_{d,\gamma}) \leq m$. The claimed inequalities in terms of c, C then follow from the well-known bounds on $M_d(\gamma)$ (e.g., Szarek, 1998). ■

Next, we argue that, unlike the other two examples above, which seem unlikely to have disambiguations of similar VC dimension, the class $\mathbb{G}_{d,\gamma}$ *does* have a (strong) disambiguation with VC dimension of comparable size.

Proposition 22 $\mathbb{G}_{d,\gamma}$ *has a strong disambiguation* $\bar{\mathbb{G}}_{d,\gamma}$ *with* $\mathsf{VC}(\bar{\mathbb{G}}_{d,\gamma}) = M_d(\gamma/2) \leq \left(\frac{2C}{\gamma}\right)^d$.

Proof Fix a maximum-size $(\gamma/2)$-packing $S = \{x_1, \ldots, x_m\}$ of $\mathcal{X}$, where $m = M_d(\gamma/2)$. Let $\mathcal{V} = \{V_1, \ldots, V_m\}$ be the Voronoi partition induced by S: that is, $V_i = \{x \in \mathcal{X} : i = \mathrm{argmin}_{i'} \|x - x_{i'}\|\}$ (breaking ties in the argmin to favor smaller i', so that $\mathcal{V}$ is indeed a partition of $\mathcal{X}$). Since S is of maximum size, any $x \in \mathcal{X}$ has $\min_i \|x - x_i\| < \gamma/2$. Thus, each V_i has diameter strictly less than γ by the triangle inequality. For each $\mathbf{y} = (y_1, \ldots, y_m) \in \{0, 1\}^m$, let $\bar{h}_{\mathbf{y}}(x) = \sum_{i=1}^m y_i \mathbb{1}[x \in V_i]$. Finally, define $\bar{\mathbb{G}}_{d,\gamma} = \{\bar{h}_{\mathbf{y}} : \mathbf{y} \in \{0, 1\}^m\}$.

To see that $\mathsf{VC}(\bar{\mathbb{G}}_{d,\gamma}) = M_d(\gamma/2)$, note that the sequence $x_1, \ldots, x_m$ is shattered by $\bar{\mathbb{G}}_{d,\gamma}$, since each x_i is the unique closest point to itself (as the other $x_{i'}$ points are all $\gamma/2$-far); thus, $\mathsf{VC}(\bar{\mathbb{G}}_{d,\gamma}) \geq m$. Moreover, since $|\bar{\mathbb{G}}_{d,\gamma}| = 2^m$, it necessarily has $\mathsf{VC}(\bar{\mathbb{G}}_{d,\gamma}) \leq m$. The inequality, upper bounding $M_d(\gamma/2)$, follows from the well-known bounds on packing numbers in bounded subsets of a Euclidean space (e.g., Szarek, 1998).

To complete the proof, we argue that $\bar{\mathbb{G}}_{d,\gamma}$ strongly disambiguates $\mathbb{G}_{d,\gamma}$. Let $h \in \mathbb{G}_{d,\gamma}$. Since each V_i has diameter strictly less than γ, h can assume only one value on $V_i \cap \mathrm{supp}(h)$. Choosing that value as y_i (or an arbitrary value if $V_i \cap \mathrm{supp}(h) = \emptyset$), we get a $\mathbf{y} = (y_1, \ldots, y_m)$ such that $\bar{h}_{\mathbf{y}} \in \bar{\mathbb{G}}_{d,\gamma}$ agrees with h on its support. ■

4. Connections to Other Notions in the Literature

4.1. Data-Dependent Generalization Guarantees

In this section we describe how one can view data-dependent generalization bounds as *Structural Risk Minimization* over partial concept classes.

Already in 1974, Vapnik and Chervonenkis (Vapnik and Chervonenkis, 1974a) showed that standard VC-dimension-based bounds can be significantly improved in the case of linear classifiers that correctly classify the data with a large margin. More generally, data-dependent guarantees provide bounds on the *generalization error* of a classifier that can be computed using *the same data* that was used to train the classifier (Shawe-Taylor, Bartlett, Williamson, and Anthony, 1998; Herbrich and Williamson, 2002). This makes such bounds particularly appealing in the context of model selection.[6] Recently, data-dependent bounds have been used to study generalization in deep neural networks; see e.g. (Bartlett, Foster, and Telgarsky, 2017; Dziugaite and Roy, 2017; Neyshabur, Bhojanapalli, McAllester, and Srebro, 2017; Dziugaite, Drouin, Neal, Rajkumar, Caballero, Wang, Mitliagkas, and Roy, 2020a; Dziugaite, Hsu, Gharbieh, and Roy, 2020b).

Data-dependent analysis is often based on assumptions which cannot be modeled in the traditional PAC learning setting: namely, it cannot be expressed as the PAC learnability of a given concept class. Consider for example the task of learning a high-dimensional linear classifier with γ-margin on the unit ball; the distribution-free sample complexity of this task is proportional to $1/\gamma^2$, as witnessed e.g. by the classical Perceptron algorithm (Rosenblatt, 1958). However, note that the hypotheses outputted by the Perceptron — namely the class of linear classifiers — has PAC sample complexity (or VC dimension) that scales linearly with the Euclidean dimension, and can therefore be arbitrarily larger than $1/\gamma^2$ and even infinite. To the best of our knowledge, the same applies to all learning algorithms in this context. Thus, it seems that *learnability of large-margin linear classifiers cannot be expressed as the PAC learnability of a concept class.* In any case, there is certainly no simple and natural VC class of total concepts which disambiguates large-margin linear classifiers.

Consequently, the general framework for data-dependent analysis deviated from the traditional PAC setting (Shawe-Taylor, Bartlett, Williamson, and Anthony, 1998; Herbrich and Williamson, 2002). Technically, this is done by introducing a data-dependent "luckiness" function which induces a (data-dependent) hierarchy of hypotheses (luckier hypotheses precede less lucky ones, as we discuss in more detail below). For example, in the case of large margin linear classifiers, the luckiness of each linear separator is its margin with respect to the input sample.

While the luckiness framework has been successfully applied in various contexts, *it does not yield a crisp notion of learnability in the spirit of PAC learning.* Moreover, the general results in this context require the luckiness function to satisfy rather arcane technical conditions and, while these conditions suffice for proving bounds on a case-by-case basis in various situations, it is not clear whether they are necessary in general.

Data-Dependent Generalization Guarantees via Partial Concept Classes. An attractive feature of partial concept classes is that they allow to express a variety of learning guarantees for specific types of data as "standard" learning guarantees with respect to a partial concept class: for example, the study of learning guarantees for linear classifiers with margin reduces to the PAC learnability of the partial concept class $\mathbb{H}_{R,\gamma}$ defined in Section 3.1. Furthermore, this framework leads

6. Namely, given two competing classifiers, prioritize the one for which the data-dependent bound is better.

to a natural approach for proving *data-dependent* learning guarantees: i.e., bounds on $\mathrm{er}_P(\hat{h}_n)$ that do not require assumptions on P, but rather are expressed in terms of properties of the data set. This can be achieved via a standard application of the principle of *Structural Risk Minimization* (SRM): that is, rather than constructing a *data-dependent* hierarchy of total concept classes as considered by (Shawe-Taylor, Bartlett, Williamson, and Anthony, 1998), we can establish data-dependent error bounds based on a *fixed and data-independent* sequence of *partial* concept classes, so that we can apply *standard* SRM arguments as in (Vapnik and Chervonenkis, 1974a,b; Vapnik, 1998).

Specifically, consider any sequence $\mathbb{H}_1, \mathbb{H}_2, \ldots$ of partial concept classes, and for each i let $\mathbb{A}_i$ be a learning algorithm designed for the class $\mathbb{H}_i$. For any data sequence $S = \{(x_i, y_i)\}_{i=1}^n$ in $\mathcal{X} \times \{0,1\}$, define $\hat{\mathrm{er}}_S(\mathbb{H}_i) := \min_{h \in \mathbb{H}_i} \frac{1}{|S|} \sum_i \mathbb{1}[h(x_i) \neq y_i]$. First we describe a realizable version of SRM. For each i, suppose there is a bound $B_i(n, \delta)$ such that, for any P, for $S \sim P^n$, with probability at least $1 - \delta$, if $\hat{\mathrm{er}}_S(\mathbb{H}_i) = 0$, then $\mathrm{er}_P(\mathbb{A}_i(S)) \leq B_i(n, \delta)$. Then we can easily produce a method with a corresponding data-dependent error bound: choose $\hat{\imath}$ of minimal $B_{\hat{\imath}}(n, \delta/\hat{\imath}(\hat{\imath}+1))$ subject to $\hat{\mathrm{er}}_S(\mathbb{H}_{\hat{\imath}}) = 0$ (if it exists), and output $\hat{h} = \mathbb{A}_{\hat{\imath}}(S)$. The corresponding guarantee is that, with probability at least $1 - \delta$, if $\hat{\imath}$ exists, then

$$\mathrm{er}_P(\hat{h}) \leq B_{\hat{\imath}}(n, \delta/\hat{\imath}(\hat{\imath}+1)).$$

This holds by a simple union bound, so that the $B_i(n, \delta/i(i+1))$ guarantees hold simultaneously for all i with probability at least $1 - \sum_i \delta/i(i+1) = 1 - \delta$. In Section C.2 (Lemma 43), we give a general algorithm that can always achieve the type of guarantee for $\mathbb{A}_i$ required above, specifically with

$$B_i(n, \delta) = O\left(\frac{\mathsf{VC}(\mathbb{H}_i)}{n} \log^2(n) + \frac{1}{n} \log\left(\frac{1}{\delta}\right)\right).$$

For instance, for the margin example in Section 3.1, since $\mathsf{VC}(\mathbb{H}_{R,\gamma}) = \Theta\left(\frac{R^2}{\gamma^2}\right)$ (from Proposition 17), we can recover the data-dependent margin bounds of (Shawe-Taylor, Bartlett, Williamson, and Anthony, 1998) by taking the classes $\mathbb{H}_i$ in the hierarchy as the partial concept classes $\mathbb{H}_{R_i,\gamma_i}$ for an appropriate sequence of (R_i, γ_i): for instance, it suffices to define $R_i = j_i$ and $\gamma_i = 1/k_i$, where (j_i, k_i) is an enumeration of $\mathbb{N}^2$ satisfying $i \leq (j_i+1)^2(k_i+1)^2$. Thus, with probability at least $1-\delta$, if the data set S has x's contained in a ball of radius $\hat{R}$ and S is linearly separable with margin $\hat{\gamma}$, then we may choose the class $\mathbb{H}_{\lceil \hat{R} \rceil, 1/\lceil 1/\hat{\gamma} \rceil}$ to recover the bound $\mathrm{er}_P(\hat{h}) = O\left(\frac{\hat{R}^2}{\hat{\gamma}^2} \frac{1}{n} \log^2(n) + \frac{1}{n} \log\left(\frac{1}{\delta}\right)\right)$ from (Shawe-Taylor, Bartlett, Williamson, and Anthony, 1998).

We can similarly derive a bound that does not require $\hat{\mathrm{er}}_S(\mathbb{H}_{\hat{\imath}}) = 0$, recovering the full spirit of the SRM principle. Specifically, suppose that for each $\mathbb{H}_i$, the learning algorithm $\mathbb{A}_i$ guarantees that, for any P, for $S \sim P^n$, with probability at least $1 - \delta$, $\mathrm{er}_P(\mathbb{A}_i(S)) \leq \hat{\mathrm{er}}_S(\mathbb{H}_i) + B_i(n, \delta)$. Then let us choose $\hat{\imath}$ to minimize $\hat{\mathrm{er}}_S(\mathbb{H}_{\hat{\imath}}) + B_{\hat{\imath}}(n, \delta/\hat{\imath}(\hat{\imath}+1))$, and output $\hat{h} = \mathbb{A}_{\hat{\imath}}(S)$. As above, by the union bound, we have that with probability at least $1 - \delta$,

$$\mathrm{er}_P(\hat{h}) \leq \hat{\mathrm{er}}_S(\mathbb{H}_{\hat{\imath}}) + B_{\hat{\imath}}(n, \delta/\hat{\imath}(\hat{\imath}+1)).$$

Again, in Section C.2 (Lemma 43), we propose a general algorithm $\mathbb{A}_i$ that can provide guarantees as required above with

$$B_i(n, \delta) = O\left(\sqrt{\frac{\mathsf{VC}(\mathbb{H}_i)}{n} \log^2(n) + \frac{1}{n} \log\left(\frac{1}{\delta}\right)}\right).$$

We can also extend this to capture both cases, by supposing $\mathbb{A}_i$ has the guarantee that, for any P, for $S \sim P^n$, with probability at least $1-\delta$, $\mathrm{er}_P(\mathbb{A}_i(S)) \leq \hat{\mathrm{er}}_S(\mathbb{H}_i) + B_i(\hat{\mathrm{er}}_S(\mathbb{H}_i), n, \delta)$. Choosing $\hat{i}$ to minimize $\hat{\mathrm{er}}_S(\mathbb{H}_{\hat{i}}) + B_{\hat{i}}(\hat{\mathrm{er}}_S(\mathbb{H}_{\hat{i}}), n, \delta/\hat{i}(\hat{i}+1))$ and outputting $\hat{h} = \mathbb{A}_{\hat{i}}(S)$, we get that with probability at least $1-\delta$, $\mathrm{er}_P(\hat{h}) \leq \hat{\mathrm{er}}_S(\mathbb{H}_{\hat{i}}) + B_{\hat{i}}(\hat{\mathrm{er}}_S(\mathbb{H}_{\hat{i}}), n, \delta/\hat{i}(\hat{i}+1))$. In particular, in Section C.2 (Lemma 43), we propose a general algorithm $\mathbb{A}_i$ that provides a guarantee B_i as required above, with

$$B_i(\hat{\varepsilon}, n, \delta) = O\left(\sqrt{\hat{\varepsilon}\left(\frac{\mathsf{VC}(\mathbb{H}_i)}{n}\log^2(n) + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right)} + \frac{\mathsf{VC}(\mathbb{H}_i)}{n}\log^2(n) + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right).$$

Comparison to SRM with Data-dependent Hierarchies. The SRM framework by Shawe-Taylor, Bartlett, Williamson, and Anthony (1998) is based on a data-dependent regularization function which they call *luckiness*: let $\mathbb{H}$ be a total concept class, and let m denote any input-sample size. A luckiness function is a mapping $L\colon \mathcal{X}^m \times \mathbb{H} \to \mathbb{R}^+$ which, given an input sample $S = \{(x_i, y_i)\}_{i \leq m}$, assigns to each hypothesis $h \in \mathbb{H}$ a real number $L(x_1, \ldots, x_m; h)$ which measures its "luckiness". The intuition is that when choosing between two competing concepts with equal empirical error rate on the data, we should prefer the one with a larger value of $L(x_1, \ldots, x_m; \cdot)$. For example, in the context of linear classification with margin (on a bounded space), the luckiness function $L(x_1, \ldots, x_m; h)$ assigns to each linear classifier its margin with respect to $x_1, \ldots, x_m$.

While a complete formal comparison of the two frameworks is beyond the scope of this work, we note that nearly all of the essential features of the data-dependent SRM framework of (Shawe-Taylor, Bartlett, Williamson, and Anthony, 1998) can be captured and generalized by the present framework of SRM with data-*independent* hierarchies of *partial* concept classes.

Let us call a luckiness function L *projective* if, for any $x_1, \ldots, x_n$ and $h \in \mathbb{H}$, every $m \in \mathbb{N}$ and $i_1, \ldots, i_m \in [n]$ satisfy $L(x_{i_1}, \ldots, x_{i_m}, h) \geq L(x_1, \ldots, x_n, h)$. All of the examples of luckiness functions given by (Shawe-Taylor, Bartlett, Williamson, and Anthony, 1998) are projective (including the margin example), and it is not hard to see that one can convert any luckiness function into a projective one by defining $L'(x_1, \ldots, x_n, h) = \min_m \min_{i_1,\ldots,i_m} L(x_{i_1}, \ldots, x_{i_m}, h)$. Given any projective luckiness function L, we can construct a hierarchy of partial concept classes $\mathbb{H}_1 \subseteq \mathbb{H}_2 \subseteq \cdots$ as follows. For $r > 0$, we say that a partial concept $h\colon \mathcal{X} \to \{0, 1, \star\}$ is *r-lucky* if there exists a total concept $h' \in \mathbb{H}$ such that $h(x) = h'(x)$ for all $x \in \mathsf{supp}(h)$ (i.e., h' extends h), and $L(\boldsymbol{x}; h') \geq r$ for every $\boldsymbol{x} \in (\mathsf{supp}(h))^*$. Let $\tilde{\mathbb{H}}_r$ denote the class of all r-lucky partial concepts.[7] Note that $\tilde{\mathbb{H}}_r : r \in \mathbb{R}^+$ is a hierarchy of partial concept classes (i.e., $\tilde{\mathbb{H}}_r \supseteq \tilde{\mathbb{H}}_s$ for $r \leq s$). Moreover, for any given r and data sequence $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \{0,1\})^m$, S is realizable w.r.t. the data-dependent total concept class $\{h \in \mathbb{H} : L(x_1, \ldots, x_m; h) \geq r\}$ *if and only if* S is realizable w.r.t. the partial concept class $\tilde{\mathbb{H}}_r$. Thus, the data-independent hierarchy $\tilde{\mathbb{H}}_r$ captures the essential information given by the luckiness function. Moreover, the rather-complex technical requirements on L imposed by (Shawe-Taylor, Bartlett, Williamson, and Anthony, 1998) imply, in particular, a bound on the VC dimension of the partial concept classes $\tilde{\mathbb{H}}_r$.[8] Thus, we can recover the types of data-dependent error bounds provided by (Shawe-Taylor, Bartlett, Williamson,

7. For measurability, it may be desirable to restrict to only those h with finite support. This does not affect the validity of the claims.

8. Technically, the assumption in Shawe-Taylor, Bartlett, Williamson, and Anthony (1998) bounds the *effective* VC dimension of $\tilde{\mathbb{H}}_r$: i.e., the VC dimension w.r.t. typical samples; but also all other results discussed here apply under this assumption.

and Anthony, 1998) using the above SRM technique with partial concept classes $\mathbb{H}_i = \tilde{\mathbb{H}}_{r_i}$, for a suitable discretization $r_1 \geq r_2 \geq \cdots$ (e.g., chosen so that $\mathsf{VC}(\mathbb{H}_i) = i$). On the other hand, our framework allows us to use SRM with *any* sequence of partial concept classes, including those not induced by a luckiness function on a class of total concepts.

4.2. Multiclass Classification

One basic question that immediately arises when considering partial concepts is how this setting differs from the 3-label *multiclass* classification problem (Ben-David, Cesa-Bianchi, Haussler, and Long, 1995). In both cases, there is a class $\mathbb{H}$ of functions $\mathcal{X} \to \{0, 1, \star\}$. The only distinction is in the definition of PAC learning, where the multiclass setting would allow distributions P on $\mathcal{X} \times \{0, 1, \star\}$, whereas the setting of partial concepts restricts to distributions supported on $\mathcal{X} \times \{0, 1\}$. That is, a distribution P on $\mathcal{X} \times \{0, 1, \star\}$ is realizable w.r.t. $\mathbb{H}$ in the 3-label multiclass setting if $\inf_{h \in \mathbb{H}} P(\{(x, y) : h(x) \neq y\}) = 0$, and otherwise the definition of PAC learnability remains the same as in Definition 2.

As it turns out, any partial concept class $\mathbb{H}$ that is PAC learnable in the 3-label multiclass setting is also PAC learnable in the partial concepts setting. However, there are simple examples where the reverse implication fails. A simple example of this is the class $\mathbb{H}$ of all functions $\mathbb{N} \to \{0, 1, \star\}$ whose image is $\{0, \star\}$. Generally, we can relate learnability in these two settings by considering the VC dimension of the *supports* of the partial concepts, as shown in the following simple result.

Proposition 23 *Let $\mathbb{H}$ be a class of functions $\mathcal{X} \to \{0, 1, \star\}$. The following are equivalent.*

1. *$\mathbb{H}$ is PAC learnable in the 3-label multiclass settting.*
2. *$\mathbb{H}$ is PAC learnable in the partial concepts setting* ***and*** *$\mathsf{VC}(\{\mathrm{supp}(h) : h \in \mathbb{H}\}) < \infty$.*

Proof PAC learnability in the 3-label multiclass setting is known to be completely characterized by a family of combinatorial complexity measures (Ben-David, Cesa-Bianchi, Haussler, and Long, 1995). In particular, the *Graph dimension* is defined as $d_G := \sup_{h_0} \mathsf{VC}(\{x \mapsto \mathbb{1}[h(x) = h_0(x)] : h \in \mathbb{H}\})$, where h_0 ranges over all functions $\mathcal{X} \to \{0, 1, \star\}$. Another complexity measure, known as the *Natarajan dimension* (Natarajan, 1989), denoted d_N, is defined as the largest d such that there exist $(x_1, y_1^{(0)}, y_1^{(1)}), \ldots, (x_d, y_d^{(0)}, y_d^{(1)}) \in \mathcal{X} \times \{0, 1, \star\}^2$ with $y_i^{(0)} \neq y_i^{(1)}$ for all i, and with the property that $\forall b_1, \ldots, b_d \in \{0, 1\}$, $\exists h \in \mathbb{H}$ with $\forall i \leq d, h(x_i) = y_i^{(b_i)}$. In the case of 3-label multiclass classification, Ben-David, Cesa-Bianchi, Haussler, and Long (1995) show that $d_N \leq d_G \leq c d_N$ for a finite numerical constant c. Moreover, Natarajan (1989); Ben-David, Cesa-Bianchi, Haussler, and Long (1995) have shown that $\mathbb{H}$ is PAC learnable in the 3-label multiclass setting if and only if $d_N < \infty$.

In particular, note that $\mathsf{VC}(\mathbb{H})$ is merely the quantity resulting from restricting each $y_i^{(0)} = 0$ and $y_i^{(1)} = 1$ in the definition of Natarajan dimension, so that we always have $d_N \geq \mathsf{VC}(\mathbb{H})$. Thus, any $\mathbb{H}$ that is PAC learnable in the 3-label multiclass setting has $\mathsf{VC}(\mathbb{H}) < \infty$, so that our Theorem 3 implies $\mathbb{H}$ is also PAC learnable in the partial concepts setting. Moreover, note that for any sequence $x_1, \ldots, x_d$ shattered by $\{\mathrm{supp}(h) : h \in \mathbb{H}\}$, if we take $h_0 : \mathcal{X} \to \{0, 1, \star\}$ as any function equal $\star$ on all of $x_1, \ldots, x_d$, then this sequence is also shattered by $\{x \mapsto \mathbb{1}[h(x) = h_0(x)] : h \in \mathbb{H}\}$. Therefore, $d_G \geq \mathsf{VC}(\{\mathrm{supp}(h) : h \in \mathbb{H}\})$, so that if $\mathbb{H}$ is PAC learnable in the 3-label multiclass setting, then $\mathsf{VC}(\{\mathrm{supp}(h) : h \in \mathbb{H}\}) < \infty$.

In the other direction, let $(x_1, y_1^{(0)}, y_1^{(1)}), \ldots, (x_d, y_d^{(0)}, y_d^{(1)})$ be as in the definition of d_N. Then $\{x_i : \star \notin \{y_i^{(0)}, y_i^{(1)}\}\}$ is shattered by the partial concept class $\mathbb{H}$, while $\{x_i : \star \in \{y_i^{(0)}, y_i^{(1)}\}\}$ is shattered by $\{\mathrm{supp}(h) : h \in \mathbb{H}\}$. Therefore, we have $d_N \leq \mathsf{VC}(\mathbb{H}) + \mathsf{VC}(\{\mathrm{supp}(h) : h \in \mathbb{H}\})$. Thus, since any $\mathbb{H}$ that is PAC learnable in the partial concepts setting must have $\mathsf{VC}(\mathbb{H}) < \infty$ (by our Theorem 3), we find that if $\mathsf{VC}(\{\mathrm{supp}(h) : h \in \mathbb{H}\}) < \infty$ as well, then $d_N < \infty$, and hence $\mathbb{H}$ is also PAC learnable in the 3-label multiclass setting. ■

The comparison to 3-label multiclass classification yields some further interesting observations. For instance, one can show that Proposition 23 also implies that, when $\mathsf{VC}(\{\mathrm{supp}(h) : h \in \mathbb{H}\}) < \infty$, the ERM principle *does* hold for learning in the partial concepts setting.[9] This contrasts with the discussion above where we found that ERM learners can fail spectacularly for some partial concept classes with $\mathsf{VC}(\mathbb{H}) < \infty$ (cf Proposition 4).

Moreover, this connection to multiclass classification has a further implication for disambiguation. Specifically, we have the following result.

Proposition 24 *Any partial concept class $\mathbb{H}$ can be strongly disambiguated to a total concept class $\bar{\mathbb{H}}$ with $\mathsf{VC}(\bar{\mathbb{H}}) = O(\mathsf{VC}(\mathbb{H}) + \mathsf{VC}(\{\mathrm{supp}(h) : h \in \mathbb{H}\}))$.*

Proof Define $\bar{\mathbb{H}} = \{x \mapsto \mathbb{1}[h(x) = 1] : h \in \mathbb{H}\}$. Continuing the notation introduced in the proof of Proposition 23, we have $\mathsf{VC}(\bar{\mathbb{H}}) \leq d_G = O(d_N)$ (where the last equality is from Ben-David, Cesa-Bianchi, Haussler, and Long, 1995, in this case of 3-class multiclass classification). Then, as established in the proof of Proposition 23, we have $d_N \leq \mathsf{VC}(\mathbb{H}) + \mathsf{VC}(\{\mathrm{supp}(h) : h \in \mathbb{H}\})$, which completes the proof. ■

In particular, this means that if $\mathbb{H}$ is a PAC learnable partial concept class, and $\mathsf{VC}(\{\mathrm{supp}(h) : h \in \mathbb{H}\}) < \infty$, then it can be disambiguated to a learnable total concept class. This contrasts with the general case discussed in the sections above, where we found that there exist learnable partial concept classes $\mathbb{H}$ that cannot be disambiguated to learnable total concept classes (see Theorems 1 and 11).

Another property enjoyed by classes $\mathbb{H}$ that are PAC learnable in the 3-label multiclass setting is that they satisfy *closure properties*. Specifically, for any finite k and classes $\mathbb{H}_1, \ldots, \mathbb{H}_k$ that are PAC learnable in the 3-label multiclass setting, and any function $U : \{0, 1, \star\}^k \to \{0, 1, \star\}$, the class $\{x \mapsto U(h_1(x), \ldots, h_k(x)) : \forall i, h_i \in \mathbb{H}_i\}$ is also learnable in the 3-label multiclass setting. Together with Proposition 23, we may conclude that, for any partial concept classes $\mathbb{H}_1, \ldots, \mathbb{H}_k$ with $\mathsf{VC}(\mathbb{H}_i) < \infty$ and $\mathsf{VC}(\{\mathrm{supp}(h) : h \in \mathbb{H}_i\}) < \infty$, and for any function $U : \{0, 1, \star\}^k \to \{0, 1, \star\}$, the partial concept class $\{x \mapsto U(h_1(x), \ldots, h_k(x)) : \forall i, h_i \in \mathbb{H}_i\}$ is PAC learnable in the partial concepts setting. Interestingly, this property is *not* true for general partial concept classes, with unrestricted supports. Specifically, we have the following result.

Proposition 25 *Define the* 2*-argument* majority *function U: if $1 \in \{y, y'\} \subseteq \{1, \star\}$, let $U(y, y') = 1$; if $0 \in \{y, y'\} \subseteq \{0, \star\}$, let $U(y, y') = 0$; otherwise let $U(y, y') = \star$. If $|\mathcal{X}| = \infty$, there exist*

9. This follows from the fact that, with $\mathsf{VC}(\mathbb{H}) < \infty$ and $\mathsf{VC}(\{\mathrm{supp}(h) : h \in \mathbb{H}\}) < \infty$, we have a Sauer-Shelah-Perles type bound on the total number of $\{0, 1, \star\}$ patterns possible on any data set, from which uniform convergence guarantees follow for the losses.

partial concept classes $\mathbb{H}_1, \mathbb{H}_2$ *with* $\mathsf{VC}(\mathbb{H}_1) = \mathsf{VC}(\mathbb{H}_2) = 0$ *such that* $\mathsf{VC}(\{x \mapsto U(h_1(x), h_2(x)) : h_1 \in \mathbb{H}_1, h_2 \in \mathbb{H}_2\}) = \infty$.

Proof Take $\mathbb{H}_1$ as the set of all functions with image contained in $\{0, \star\}$, and $\mathbb{H}_2$ the set of all functions with image contained in $\{1, \star\}$. For any $d \in \mathbb{N}$ and any distinct $x_1, \ldots, x_d \in \mathcal{X}$, for any $y_1, \ldots, y_d \in \{0, 1\}$, take $h_1 \in \mathbb{H}_1$ with $\mathbb{1}[h_1(x_i) = 0] = \mathbb{1}[y_i = 0]$ for all i, and $h_2 \in \mathbb{H}_2$ with $\mathbb{1}[h_2(x_i) = 1] = \mathbb{1}[y_i = 1]$ for all i. In particular, note that $\mathrm{supp}(h_1) \cap \{x_1, \ldots, x_d\}$ and $\mathrm{supp}(h_2) \cap \{x_1, \ldots, x_d\}$ are disjoint, and their union is $\{x_1, \ldots, x_d\}$. Moreover, $U(h_1(x_i), h_2(x_i)) = y_i$ for all i. Thus, the sequence $x_1, \ldots, x_d$ is shattered by $\{x \mapsto U(h_1(x), h_2(x)) : h_1 \in \mathbb{H}_1, h_2 \in \mathbb{H}_2\}$. Since d can be chosen arbitrarily large, this completes the proof. ■

In fact, one can even show such a negative result for functions U having a *single* argument: that is, $U : \{0, 1, \star\} \to \{0, 1, \star\}$. For instance, taking $U(y) = \mathbb{1}[y = 1]$, the class $\{x \mapsto U(h(x)) : h \in \mathbb{H}\}$ represents a strong disambiguation of $\mathbb{H}$, so that Theorem 1 indicates that, even if $\mathbb{H}$ is learnable, it can happen that the class $\{x \mapsto U(h(x)) : h \in \mathbb{H}\}$ is not learnable.

Appendix A. Formal Definitions of Complexity Measures

Before getting into the detailed results and proofs, we first elaborate on the definitions of the combinatorial complexity measures appearing in our results. As mentioned in Section 2, when suitably expressed, the complexity measures all inherit precisely the same definitions as for the traditional setting of total concept classes. Nevertheless, for those readers not familiar with the original definitions for total concept classes, we state the definitions here in full detail.

Definition 26 (Vapnik-Chervonenkis Dimension) *For a partial concept class* $\mathbb{H}$*, the* VC dimension *of* $\mathbb{H}$*, denoted* $\mathsf{VC}(\mathbb{H})$*, is the maximum number* $d \in \mathbb{N} \cup \{0\}$ *such that* $\exists x_1, \ldots, x_d$ *with* $\{(h(x_1), \ldots, h(x_d)) : h \in \mathbb{H}\} \supseteq \{0, 1\}^d$*. Such a sequence* $\{x_1, \ldots, x_d\}$ *is said to be* shattered *by* $\mathbb{H}$*. If there is no largest such* d*, then define* $\mathsf{VC}(\mathbb{H}) = \infty$.

Next we state the definition of the Littlestone dimension. Recall that a sequence of examples $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{0, 1\}$ is said to be *realizable* w.r.t. $\mathbb{H}$ if $\exists h \in \mathbb{H}$ with $\forall i \leq n$, $h(x_i) = y_i$.

Definition 27 (Littlestone dimension) *For any partial concept class* $\mathbb{H}$*, the* Littlestone dimension *of* $\mathbb{H}$*, denoted by* $\mathsf{LD}(\mathbb{H})$*, is the largest integer* d *such that there exists a set* $\{x_{\mathbf{y}} : \mathbf{y} \in \bigcup_{0 \leq i \leq d-1} \{0, 1\}^i\} \subseteq \mathcal{X}$ *with the property that, for every* $y_1, \ldots, y_d \in \{0, 1\}$*, the sequence*

$$(x_{()}, y_1), (x_{(y_1)}, y_2), (x_{(y_1, y_2)}, y_3), \ldots, (x_{(y_1, \ldots, y_{d-1})}, y_d)$$

is realizable w.r.t. $\mathbb{H}$*. In particular, if no* x *has more than one realizable label in* $\{0, 1\}$*, then* $\mathsf{LD}(\mathbb{H}) = 0$*. On the other hand, if no such largest* d *exists, we define* $\mathsf{LD}(\mathbb{H}) = \infty$.

To interpret the definition, it is conventional to think of the $x_{\mathbf{y}}$ points as being organized into a binary *tree* based on the prefixes of $\mathbf{y}$, where edges corresponding to a left branch are labeled 0 and edges corresponding to a right branch are labeled 1: that is, the bits of $\mathbf{y}$ determine whether to branch left or right at each level along the path leading to the node $x_{\mathbf{y}}$. Then $\mathsf{LD}(\mathbb{H})$ is the maximum

depth d of a complete tree of this type, such that all descending paths from the root correspond to a sequence realizable w.r.t. $\mathbb{H}$ when each non-terminal node $x_{\mathbf{y}}$ on the path is given the label of the edge followed next in the path.

Next we restate the definition of the Threshold dimension:

Definition 28 (Threshold dimension) *The* Threshold dimension *of a partial concept class $\mathbb{H}$, denoted by $\mathrm{TD}(\mathbb{H})$, is the maximum integer d for which there exist $x_1, \ldots, x_d \in \mathcal{X}$ and $h_1, \ldots h_d \in \mathbb{H}$ such that $h_i(x_j) = \mathbb{1}[i \leq j]$. If no such largest d exists, define $\mathrm{TD}(\mathbb{H}) = \infty$.*

We conclude by restating the definition of *compression scheme* in formal detail. Specifically, the following definition is originally due to Littlestone and Warmuth (1986a).

Definition 29 (Compression Scheme) *A* compression scheme *is a pair (κ, ρ), consisting of a compression function $\kappa : (\mathcal{X} \times \{0,1\})^* \to (\mathcal{X} \times \{0,1\})^* \times \{0,1\}^*$ and a reconstruction function $\rho : (\mathcal{X} \times \{0,1\})^* \times \{0,1\}^* \to \{0,1\}^{\mathcal{X}}$, satisfying the following property. For any sequence $S \in (\mathcal{X} \times \{0,1\})^*$, $\kappa(S)$ evaluates to some $(S', B) \in (\mathcal{X} \times \{0,1\})^* \times \{0,1\}^*$ where S' is a sequence of elements of S (possibly re-ordered, and possibly including copies, having length at most $|S|$).*[10]

The size *of the compression scheme for a given sample size m is $\max_{S \in (\mathcal{X} \times \{0,1\})^m} |\kappa(S)|$ (i.e., the length of the sequence S' plus the number of bits in B), and the (unqualified)* size *of the compression scheme is the maximum size over all m, or infinite if the size can be unbounded.*

For any partial concept class $\mathbb{H}$, a sample compression scheme for *$\mathbb{H}$ is a compression scheme (κ, ρ) with the additional property that, for every finite data sequence S realizable w.r.t. $\mathbb{H}$, $\rho(\kappa(S))$ is correct on S: i.e., $\hat{\mathrm{er}}_S(\rho(\kappa(S))) = 0$.*

The intention in the above definition is that $\rho(\kappa(\cdot))$ is interpreted as a learning algorithm. For brevity, we will sometimes leave off the bit sequence B, simply specifying $\kappa : (\mathcal{X} \times \{0,1\})^* \to (\mathcal{X} \times \{0,1\})^*$ and $\rho : (\mathcal{X} \times \{0,1\})^* \to \{0,1\}^{\mathcal{X}}$, reflecting the special case where no extra bits are ever output by the compression function.

Appendix B. Proofs of Disambiguation

Proof of Proposition 10 The proof is similar to that of Theorem 20. Let $k = O(d^\star/\gamma^2)$. Consider the class $\bar{\mathbb{H}}_k$ of all k-wise majority votes of concepts from $\bar{\mathbb{H}}$. We will show that $\bar{\mathbb{H}}_k$ disambiguates $\mathbb{H}$ (note that by standard bounds on the variability of the VC dimension under composition/aggregation, we have that $\mathsf{VC}(\bar{\mathbb{H}}_k) = \tilde{O}(\mathsf{VC}(\bar{\mathbb{H}}) \cdot k) = \tilde{O}(\frac{d \cdot d^*}{\gamma^2})$, as required).

Let S be a sample realizable by $\mathbb{H}$. We claim that for every distribution P over S there exists $\bar{h} \in \bar{\mathbb{H}}$ such that $\mathrm{er}_P(\bar{h}) \leq \frac{1-\gamma/2}{2}$. This will suffice, because then an application of the Minimax Theorem yields a distribution Q over $\bar{\mathbb{H}}$ such that $\Pr_{\bar{h} \sim Q}[\bar{h}(x) \neq y] \leq \frac{1-\gamma/2}{2}$ for every $(x, y) \in S$. Then, an application of the VC theorem (ε-approximation/uniform-convergence) to the dual class $\bar{\mathbb{H}}^\star$ implies that with positive probability, a random i.i.d sample $\bar{h}_1, \ldots, \bar{h}_k \sim Q$, where $k = O(d^\star/\gamma^2)$, will satisfy that its majority vote realizes S entirely.

10. We also constrain (κ, ρ) to be such that, for any $n \in \mathbb{N}$, $((x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), x_n) \mapsto \rho(\kappa((x_1, y_1), \ldots, (x_{n-1}, y_{n-1})))(x_n)$ is a measurable function.

Thus, it remains to show that for every distribution P over S there exists $\bar{h} \in \bar{\mathbb{H}}$ such that $\mathrm{er}_P(\bar{h}) \leq \frac{1-\gamma/2}{2}$. Indeed, consider such a distribution P, and draw a random sample S' of size $O(d/\gamma^2)$ from P. By assumption, there must exist $\bar{h} = \bar{h}(S') \in \bar{\mathbb{H}}$ whose empirical loss on the sample satisfies $\hat{\mathrm{er}}_{S'}(\bar{h}) \leq \frac{1-\gamma}{2}$. By the VC theorem (again, ε-approximation/uniform-convergence), it follows that with positive probability (over the generation of S'), we have that the population loss satisfies

$$\mathrm{er}_P(\bar{h}) \leq \hat{\mathrm{er}}_{S'}(\bar{h}) + \gamma/4 \leq \frac{1-\gamma/2}{2},$$

as required.

■

Proof of Theorem 11 Interestingly, our proof exploits a recent line of breakthroughs in complexity-theory and combinatorics. For our purpose, it will be convenient to use the following combinatorial formulation of these results, which provides a nearly tight bound to a question posed by Alon, Saks, and Seymour (for background on this question, see the survey by Bousquet, Lagoutte, and Thomassé (2014)). Let $G = (V, E)$ be a simple graph. Recall that the chromatic number of G, denoted by $\chi(G)$, is the minimum k for which there exists a coloring $c : V \to [k]$ such that every edge $\{u, v\} \in E$ satisfies $c(u) \neq c(v)$. The *biclique partition number* of G, denoted $\mathsf{bp}(G)$ is the minimum number of complete bipartite graphs needed to partition the edge set of G. The next result follows from a recent line of works by Göös (2015); Ben-David, Hatami, and Tal (2017); Balodis, Ben-David, Göös, Jain, and Kothari (2021):

Theorem 30 *For every n, there exists a simple graph $G = (V, E)$ with $\mathsf{bp}(G) = n$ such that*

$$\chi(G) \geq n^{(\log(n))^{1-\varepsilon(n)}},$$

where $\varepsilon(n)$ is a sequence satisfying $\varepsilon(n) \to_{n\to\infty} 0$.

Let $G = (V, E)$ be a graph as promised by Theorem 30, and let $B_i = (L_i, R_i, E_i)$ be n complete bipartite graphs which witness that $\mathsf{bp}(G) = n$. Define a partial concept class $\mathbb{H}_n \subseteq \{0, 1, \star\}^n$ as follows: for each vertex $v \in V$ there is a partial concept $c_v \in \mathbb{H}$ such that for every $i \in [n]$:

$$c_v(i) = \begin{cases} 0 & v \in L_i, \\ 1 & v \in R_i, \\ \star & \text{otherwise.} \end{cases}$$

We finish the proof with the following two lemmas:

Lemma 31 $\mathsf{VC}(\mathbb{H}_n) = 1$ *and* $\mathrm{TD}(\mathbb{H}_n) \leq 2$. *In fact, on every pair of coordinates $\{i, j\}$ the class $\mathbb{H}$ realizes at most 2 patterns, that is:*

$$(\forall i, j \in [n]) : \quad \Big|\{0,1\}^2 \cap \{\big(h(i), h(j)\big) : h \in \mathbb{H}\}\Big| \leq 2$$

Proof $\mathsf{VC}(\mathbb{H}_n) > 0$ because every edge $\{u, v\} \in E$ satisfies $\{u, v\} \in E_i$ for some $i \in [n]$ and thus $\{c_v(i), c_u(i)\} = \{0, 1\}$ which implies that $\{i\}$ is shattered by $\mathbb{H}_n$ and hence $\mathsf{VC}(\mathbb{H}_n) \geq 1$.

Note that $\mathsf{VC}(\mathbb{H}_n) < 2$ and $\mathrm{TD}(\mathbb{H}_n) \leq 2$ follow from

$$(\forall i, j \in [n]): \quad \Big|\{0,1\}^2 \cap \{\big(h(i), h(j)\big) : h \in \mathbb{H}\}\Big| \leq 2,$$

and thus it suffices to prove the latter. Let $\{i, j\} \subseteq [n]$ be a pair of distinct coordinates. Assume towards contradiction that

$$\Big|\{0,1\}^2 \cap \{\big(h(i), h(j)\big) : h \in \mathbb{H}\}\Big| \geq 3.$$

Thus, either both patterns 00 and 11 are realized, or both patterns 01 and 10 are realized. We first rule out the former: assume towards contradiction that both 00 and 11 are realized. Thus, there exist two partial concepts $c_u, c_v \in \mathbb{H}$ for $u, v \in V$ such that $c_u(i) = c_u(j) = 0$ and $c_v(i) = c_v(j) = 1$. Thus, by the definitions of c_u, c_v it follows that $u \in L_i \cap L_j$ and $v \in R_i \cap R_j$ and therefore the edge $\{u, v\}$ is covered by both B_i and B_j, which contradicts the assumption that the edges of $B_1, \ldots, B_n$ partition the edges of G. Similarly, the realization of both patterns 01 and 10 also implies an edge which is covered twice and hence a contradiction. ■

Lemma 32 *Let $\bar{\mathbb{H}} \subseteq \{0,1\}^n$ be a disambiguation of $\mathbb{H}_n$. Then $\bar{\mathbb{H}}$ defines a coloring of G using $|\bar{\mathbb{H}}|$ colors. Therefore,*

$$|\bar{\mathbb{H}}| \geq n^{(\log(n))^{1-\varepsilon(n)}},$$

as required.

Proof Assign to each vertex $v \in V$ a color $\bar{c}_v \in \bar{\mathbb{H}}$ such that $\bar{c}_v$ disambiguates the partial concept $c_v \in \mathbb{H}_n$. (I.e., $\bar{c}_v$ extends c_v to a total concept in $\{0,1\}^n$.) Indeed, this is a proper coloring since for every edge $\{u, v\} \in E$ there exists a complete bipartite $B_i = (L_i, R_i, E_i)$ such that $u \in L_i$ and $v \in R_i$ or vice versa. Thus, $c_u(i) \neq c_v(i)$ and both are in $\{0,1\}$, therefore also $\bar{c}_u(i) \neq \bar{c}_v(i)$. Hence, u, v are assigned different colors, as required. ■

Thus, we have shown the first part in Theorem 11 by demonstrating the classes $\mathbb{H}_n$, for $n \in \mathbb{N}$. For the second part, we need to show that over an infinite $\mathcal{X}$ (say $\mathcal{X} = \mathbb{N}$), there exists $\mathbb{H}_\infty \subseteq \{0, 1, \star\}^{\mathcal{X}}$ such that $\mathsf{VC}(\mathbb{H}_\infty) = 1$, $\mathrm{TD}(\mathbb{H}_\infty) \leq 2$ and every total concept class $\bar{\mathbb{H}}$ that disambiguates $\mathbb{H}_\infty$ satisfies $\mathsf{VC}(\bar{\mathbb{H}}) = \infty$. One way to construct $\mathbb{H}_\infty$ is by taking disjoint copies of the classes $\mathbb{H}_n$ (i.e., each $\mathbb{H}_n$ has its domain $\mathcal{X}_n$, the domains $\mathcal{X}_n$ are mutually disjoint, and $\mathbb{H}_\infty$ is defined by taking the union $\cup_n \widetilde{\mathbb{H}}_n$, where $\widetilde{\mathbb{H}}_n$ is obtained from $\mathbb{H}_n$ by adding $\star$'s outside its domain). It is easy to see that $\mathsf{VC}(\mathbb{H}_\infty) = 1$ and $\mathrm{TD}(\mathbb{H}_\infty) \leq 2$. To see that every disambiguating class $\bar{\mathbb{H}}$ must have an unbounded VC dimension, notice that such a class $\bar{\mathbb{H}}$ simultaneously disambiguates all of the $\mathbb{H}_n$'s. Thus, by the (contra-positive of) the Sauer-Shelah-Perles Lemma (Sauer, 1972), $\mathsf{VC}(\bar{\mathbb{H}})$ must be unbounded.

■

Proof of Theorem 12 Assume without loss of generality that $\mathcal{X} = [n] = \{1, \ldots, n\}$.

For any $\mathbb{H}' \subseteq \mathbb{H}$, we define its shattering strength:

$$s\big(\mathbb{H}'\big) = \Big|\big\{S \subseteq [n] \ : \ S \text{ is shattered by } \mathbb{H}'\big\}\Big|.$$

Note in particular that $s(\mathbb{H}') \leq \binom{n}{\leq \mathsf{VC}(\mathbb{H}')} \leq \binom{n}{\leq d}$. Also, for $(x,y) \in [n] \times \{0,1\}$ we denote

$$\mathbb{H}'|(x,y) = \{h \in \mathbb{H}' : h(x) = y\}.$$

Define the VC-majority function associated with a subclass $\mathbb{H}'$ by letting $M_{\mathbb{H}'}(x)$ for $x \in [n]$ be the value $y \in \{0,1\}$ which maximizes $s(\mathbb{H}'|(x,y))$, with an arbitrary tie-breaking rule. Observe that

$$s(\mathbb{H}') \geq s(\mathbb{H}'|(x,0)) + s(\mathbb{H}'|(x,1)).$$

Indeed, every S that is shattered by one of the subclasses $\mathbb{H}'|(x,0)$ or $\mathbb{H}'|(x,1)$ is also shattered by $\mathbb{H}'$ and if S is shattered by both $\mathbb{H}'|(x,0)$ and $\mathbb{H}'|(x,1)$ then both S and $S \cup \{x\}$ are shattered by $\mathbb{H}'$.

Consider the following strong disambiguation algorithm. Given a partial concept $h \in \mathbb{H}$, write the entries of its disambiguation in the natural order. Start with $\mathbb{H}'$ equal the entire class $\mathbb{H}$ and compute the value of its VC-majority function at $x = 1$. Write this value as the entry at x and leave the class intact, except if $x \in \mathrm{supp}(h)$ and $h(x) \neq M_{\mathbb{H}}(x)$. In the latter case, write the opposite value, and update the class by adding $(x, h(x))$ as a constraint: i.e., update $\mathbb{H}'$ to $\mathbb{H}'|(x,h(x))$. Proceed to the next value of x and carry out this procedure with the updated subclass, and so on, until reaching $x = n$.

We claim that given any partial concept $h \in \mathbb{H}$, the number $u(h)$ of updates is at most

$$\log_2(s(\mathbb{H})) \leq \log_2\left(\binom{n}{\leq d}\right) \leq 1 + d\log_2(n).$$

Indeed, by the above, after each update the strength of the updated subclass is at most half of the strength of the subclass before the update, and at all times the maintained subclass contains the partial concept h. When running this disambiguation algorithm, the output is determined by the location of the updates. By the above bound on $u(h)$, the number of ways to place the updates is at most

$$\binom{n}{\leq 1 + d\log_2(n)} = n^{O(d\log(n))}.$$

■

Proof of Theorem 13 The proof follows a similar idea like the proof of Theorem 12, but we use a carefully tailored weighted VC-majority rather than an unweighted one.

For a finite sequence $(x_1,y_1),\ldots,(x_k,y_k)$ with $x_i \in \mathbb{N}$, $y_i \in \{0,1\}$ and $x_1 < \ldots < x_k$, denote by $\mathbb{H}|(x_1,y_1),\ldots,(x_k,y_k)$ the subclass of those partial concepts $h \in \mathbb{H}$ such that $h(x_i) = y_i$ for all i. For such a constrained subclass, we define its weight:

$$w(\mathbb{H}|(x_1,y_1),\ldots,(x_k,y_k)) = \sum_S \frac{1}{n(S)^{d+1}},$$

where the summation is over all nonempty subsets S of $\mathbb{N} \setminus \{1,\ldots,x_k\}$ that are shattered by this subclass, and $n(S)$ denotes the largest element of S. In the special case when $k = 0$, i.e., the class is the entire $\mathbb{H}$, the definition is the same, taking $\{1,\ldots,x_k\} = \emptyset$. Observe that in this case (and hence in every case) the sum is finite:

$$w(\mathbb{H}) \leq \sum_n \frac{n^{d-1}}{n^{d+1}} = \sum_n \frac{1}{n^2} = \frac{\pi^2}{6},$$

where the inequality holds because for a fixed n, the number of terms $\frac{1}{n^{d+1}}$ is at most the number of subsets of $[n]$ of size d or less that include the element n, and this number is $\sum_{i=0}^{d-1}\binom{n-1}{i} \leq n^{d-1}$.

Define the VC-weighted majority function associated with such a constrained subclass by letting $M_{\mathbb{H}|(x_1,y_1),\ldots,(x_k,y_k)}(x)$ for $x \in \mathbb{N} \setminus \{1,\ldots,x_k\}$ be the value $y \in \{0,1\}$ which maximizes $w(\mathbb{H}|(x_1,y_1),\ldots,(x_k,y_k),(x,y))$, with an arbitrary tie-breaking rule. Observe that

$$w(\mathbb{H}|(x_1,y_1),\ldots,(x_k,y_k)) \geq w(\mathbb{H}|(x_1,y_1),\ldots,(x_k,y_k),(x,0)) + w(\mathbb{H}|(x_1,y_1),\ldots,(x_k,y_k),(x,1)).$$

Indeed, every term $\frac{1}{n(S)^{d+1}}$ that appears in one of the two sums on the right-hand side, also appears in the sum on the left-hand side (for the same S). If a term appears in both sums on the right-hand side, then it appears twice in the sum on the left-hand side: once for S and once for $\{x\} \cup S$.

Consider the following strong disambiguation algorithm. Given a partial concept $h \in \mathbb{H}$, write the entries of its disambiguation in the natural order. Start with the entire class $\mathbb{H}$ and compute the value of its VC-weighted majority function at $x = 1$. Write this value as the entry at x and leave the class intact, except if $x \in \mathrm{supp}(h)$ and $h(x) \neq M_{\mathbb{H}}(x)$. In the latter case, write the opposite value, and update the class by adding $(x, h(x))$ as a constraint. Proceed to the next value of x and carry out this procedure with the updated subclass, and so on.

We claim that given any partial concept $h \in \mathbb{H}$, the number $u(m)$ of updates up to $x = m$ is at most $(d+1)\log_2(m)+2$. Indeed, by the above, after each update the weight of the updated subclass is at most half of the weight of the subclass before the update. Consider the subclass before the last update up to $x = m$ (assuming there is at least one – otherwise there is nothing to prove). For the last update to occur, that subclass must shatter at least the singleton $S = \{x\}$, so its weight then is at least $\frac{1}{m^{d+1}}$. Hence

$$\frac{1}{m^{d+1}} 2^{u(m)-1} \leq w(\mathbb{H}) \leq \frac{\pi^2}{6},$$

which implies that $u(m) \leq \log_2(\frac{\pi^2}{6} m^{d+1}) + 1 \leq (d+1)\log_2(m) + 2$, as claimed.

When running this disambiguation algorithm, the first m entries of the output are determined by the location of the updates up to $x = m$. By the above bound on $u(m)$, the number of ways to place the updates is at most $\sum_{i=0}^{\lfloor (d+1)\log_2(m)+2 \rfloor}\binom{m}{i} \leq (m+1)^{(d+1)\log_2(m)+2} = m^{O(d\log(m))}$. ■

Appendix C. PAC Learnability: Proofs and Sample Complexity Bounds

This section presents the formal proofs associated with PAC learnability, both in the realizable case and agnostic case. In particular, these results will imply Theorem 3 from Section 2.2, but we will also establish quantitative versions that provide upper and lower bounds on the optimal sample complexity in each case.

Here, and in later sections, we use the notation $\log(x) := \max\{\ln(x), 1\}$.

C.1. Realizable PAC Learning

The following lemma describes basic properties of the notion of realizability with respect to partial concepts (as used in Definition 2): it demonstrates the relationship with the classical definition for classes $\mathbb{H} \subseteq \{0,1\}^{\mathcal{X}}$ which contain only total concepts, and implies that the definition we use is more general.

Lemma 33 (Connection with the classical notion of realizability) *Let $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ and let P be a distribution over $\mathcal{X} \times \{0,1\}$ such that*

$$\inf_{h\in\mathbb{H}} \mathrm{er}_P(h) = 0.$$

Then, for every n, $S \sim P^n$ is realizable by $\mathbb{H}$ with probability 1. Conversely, if P is a distribution with finite or countable support such that for every n, $S \sim P^n$ is realizable by $\mathbb{H}$ with probability 1, then $\inf_{h\in\mathbb{H}} \mathrm{er}_P(h) = 0$. Moreover, if $\mathbb{H}$ contains only total concepts and $\mathsf{VC}(\mathbb{H}) < \infty$ then the converse holds regardless of the support of P.

Thus, our definition generalizes the classical one in the sense that any distribution P which is realizable in the classical sense ($\inf_{h\in\mathbb{H}} \mathrm{er}_P(h) = 0$) is also realizable according to our definition (every sample drawn from it is realizable with probability 1). Hence, any algorithm which learns $\mathbb{H}$ in the realizable case according to our definition, also learns it according to the traditional sense. On the other hand, our definition of realizable PAC learning in the partial concept class setting may admit realizable distributions for which $\inf_{h\in\mathbb{H}} \mathrm{er}_P(h) \neq 0$. For example, if $\mathcal{X}$ is the interval $[0,1]$ and $\mathbb{H} \subseteq \{0,\star\}^{\mathcal{X}}$ consists of the functions satisfying that $h^{-1}(0)$ has Lebesgue measure $1/2$, then the uniform P over $\mathcal{X} \times \{0\}$ is realizable according to our definition, yet $\inf_{h\in\mathbb{H}} \mathrm{er}_P(h) = 1/2$.

Proof Notice that for every $h \in \mathbb{H}$:

$$\mathrm{er}_P(h) = \mathop{\mathbb{E}}_{S\sim P^n} \hat{\mathrm{er}}_S(h) \geq \mathop{\mathbb{E}}_{S\sim P^n} \big[\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_S(h)\big],$$

and therefore also

$$\inf_{h\in\mathbb{H}} \mathrm{er}_P(h) \geq \mathop{\mathbb{E}}_{S\sim P^n} \big[\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_S(h)\big] \geq 0.$$

Thus, if $\inf_{h\in\mathbb{H}} \mathrm{er}_P(h) = 0$ then $\mathbb{E}_{S\sim P^n}\big[\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_S(h)\big] = 0$ and hence $S \sim P^n$ is realizable with probability 1.

For the converse, assume first that P has finite support of size n. Taking S to be the support of P, which occurs with positive probability for $S \sim P^n$, shows that the entire support is realizable by $\mathbb{H}$, that is, $\min_{h\in\mathbb{H}} \mathrm{er}_P(h) = 0$. If the support of P is countably infinite, we can enumerate it and repeat the above argument for its first n elements, showing the existence of $h \in \mathbb{H}$ realizing those n elements. Taking $n \to \infty$ implies that $\inf_{h\in\mathbb{H}} \mathrm{er}_P(h) = 0$. Finally, assume that $\mathbb{H} \subseteq \{0,1\}^{\mathcal{X}}$ contains only total concepts and satisfies $\mathsf{VC}(\mathbb{H}) < \infty$, and P satisfies that $S \sim P^n$ is realizable with probability 1. Then:

$$\begin{aligned}
0 &= \mathop{\mathbb{E}}_{S\sim P^n} \big[\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_S(h)\big] \\
&\geq \inf_{h\in\mathbb{H}} \Big[\mathrm{er}_P(h) - O\Big(\sqrt{\frac{\mathsf{VC}(\mathbb{H})}{n}}\Big)\Big] && \text{(by uniform convergence)} \\
&= \inf_{h\in\mathbb{H}} \mathrm{er}_P(h) - o(1).
\end{aligned}$$

Thus, by letting $n \to \infty$ we see that $\inf_{h\in\mathbb{H}} \mathrm{er}_P(h) = 0$ as claimed. ■

The following theorem establishes upper and lower bounds on the optimal sample complexity of PAC learning for any given partial concept class $\mathbb{H}$. In particular, this supplies part of the proof of Theorem 3 (i.e., the part concerning the realizable case).[11]

Theorem 34 (PAC Sample Complexity) *For any partial concept class $\mathbb{H}$ with $\mathsf{VC}(\mathbb{H}) < \infty$, the optimal sample complexity of PAC learning $\mathbb{H}$, $\mathcal{M}(\varepsilon, \delta)$, satisfies the following bounds:*

- $\mathcal{M}(\varepsilon, \delta) = O\Big(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon} \log\big(\frac{1}{\delta}\big)\Big)$.
- $\mathcal{M}(\varepsilon, \delta) = O\Big(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon} \log^2\Big(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon}\Big) + \frac{1}{\varepsilon} \log\big(\frac{1}{\delta}\big)\Big)$.
- $\mathcal{M}(\varepsilon, \delta) = \Omega\Big(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon} + \frac{1}{\varepsilon} \log\big(\frac{1}{\delta}\big)\Big)$.

Moreover, if $\mathsf{VC}(\mathbb{H}) = \infty$, then $\mathbb{H}$ is not PAC learnable.

The following lemma was proven by Haussler, Littlestone, and Warmuth (1994) for total concept classes. Here we merely note that the result trivially also holds for partial concept classes.

Lemma 35 (One-inclusion Graph Predictor) *For any partial concept class $\mathbb{H}$ with $\mathsf{VC}(\mathbb{H}) < \infty$, there is a function $\mathbb{A} : (\mathcal{X} \times \{0,1\})^* \times \mathcal{X} \to \{0,1\}$ such that, for any $n \in \mathbb{N}$ and any sequence $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \{0,1\})^n$ that is realizable w.r.t. $\mathbb{H}$,*

$$\frac{1}{n!} \sum_{\sigma \in \mathrm{Sym}(n)} \mathbb{1}\big[\mathbb{A}(x_{\sigma(1)}, y_{\sigma(1)}, \ldots, x_{\sigma(n-1)}, y_{\sigma(n-1)}, x_{\sigma(n)}) \neq y_{\sigma(n)}\big] \leq \frac{\mathsf{VC}(\mathbb{H})}{n}, \tag{1}$$

where $\mathrm{Sym}(n)$ denotes the symmetric group (of permutations of $\{1, \ldots, n\}$).

Proof As Haussler, Littlestone, and Warmuth (1994) proved this result for all total concept classes $\mathbb{H}$, we need only note that it easily extends to partial concept classes, as follows. For any $n \in \mathbb{N}$ and $S = \{x_1, \ldots, x_n\}$, let $\mathcal{X}_S$ denote the set of distinct elements of the sequence S, and define $\mathbb{H}_{\mathcal{X}_S}$ as the class of all total functions $h : \mathcal{X}_S \to \{0,1\}$ such that the sequence $\{(x, h(x)) : x \in \mathcal{X}_S\}$ is realizable w.r.t. $\mathbb{H}$. In the case that $\mathbb{H}_{\mathcal{X}_S} \neq \emptyset$, let $\mathbb{A}_{\mathcal{X}_S}$ be the function guaranteed by the lemma for the case of instance space equal $\mathcal{X}_S$ and for the total concept class $\mathbb{H}_{\mathcal{X}_S}$ defined on this space. Then for any $y_1, \ldots, y_n \in \{0,1\}$ such that $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ is realizable w.r.t. $\mathbb{H}$ (and therefore also realizable w.r.t. $\mathbb{H}_{\mathcal{X}_S}$), define $\mathbb{A}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) = \mathbb{A}_{\mathcal{X}_S}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$. Since any permutation of the sequence $x_1, \ldots, x_n$ leaves the spaces $\mathcal{X}_S$ and $\mathbb{H}_{\mathcal{X}_S}$ unchanged, and since it is clear from the definition of VC dimension that $\mathsf{VC}(\mathbb{H}_{\mathcal{X}_S}) \leq \mathsf{VC}(\mathbb{H})$, it follows that (1) holds for the sequence $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.

Thus, for any $n \in \mathbb{N}$ and any sequence $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ realizable w.r.t. $\mathbb{H}$, we have defined the value $\mathbb{A}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n)$ in a way that altogether satisfies (1). To complete the definition, we may (arbitrarily) define $\mathbb{A}(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) = 0$ for all sequences $(x_1, \ldots, x_n) \in \mathcal{X}^n$ and $(y_1, \ldots, y_{n-1}) \in \{0,1\}^{n-1}$ such that $\{h \in \mathbb{H} : \forall i \leq n-1, h(x_i) =$

11. In all of the results on PAC learning, we implicitly suppose $\mathbb{H}$ satisfies appropriate mild conditions necessary to guarantee measurability of the learning algorithms involved in the proofs of upper bounds. We refer the interested reader to van der Vaart and Wellner (1996); van Handel (2013) for thorough discussions of such issues, which we will not discuss further in this article. We also note that such restrictions are only required if $\mathcal{X}$ is uncountably infinite.

y_i, and $h(x_n) \in \{0,1\}\} = \emptyset$. ■

Parts of the proof also rely on a well-known generalization bound for compression schemes, together with a construction of a particular compression scheme based on Boosting. The following lemma is a classic result due to Littlestone and Warmuth (1986a).

Lemma 36 (Consistent Compression Generalization Bound) *There exists a finite numerical constant $c \geq 1$ such that, for any compression scheme (κ, ρ), for any $n \in \mathbb{N}$ and $\delta \in (0,1)$, for any distribution P on $\mathcal{X} \times \{0,1\}$, for $S \sim P^n$, with probability at least $1-\delta$, if $\hat{\mathrm{er}}_S(\rho(\kappa(S))) = 0$, then*

$$\mathrm{er}_P(\rho(\kappa(S))) \leq \frac{c}{n - |\kappa(S)|}\left(|\kappa(S)|\log(n) + \log\left(\frac{1}{\delta}\right)\right).$$

The next component is based on a well-known Boosting algorithm, known as α-Boost, which yields a compression scheme of a quantifiable size that is sample-consistent, given access to a "weak" learning algorithm; see the book of Schapire and Freund (2012) for a proof.

Lemma 37 (Boosting) *For any $k, n \in \mathbb{N}$ and sequence $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{0,1\}$, suppose $\mathbb{A}_w : (\mathcal{X} \times \{0,1\})^k \to \{0,1\}^{\mathcal{X}}$ is an algorithm that, for any distribution P on $\mathcal{X} \times \{0,1\}$ with $P(\{(x_1,y_1),\ldots,(x_n,y_n)\}) = 1$, there exists $S_P \in \{(x_1,y_1),\ldots,(x_n,y_n)\}^k$ with $\mathrm{er}_P(\mathbb{A}_w(S_P)) \leq 1/3$. Then there is a numerical constant $c \geq 1$ such that, for $T = \lceil c\log(n)\rceil$, there exist sequences $S_1, \ldots, S_T \in \{(x_1,y_1),\ldots,(x_n,y_n)\}^k$ such that, for $\hat{h}(\cdot) := \mathrm{Majority}(\mathbb{A}_w(S_1)(\cdot), \ldots, \mathbb{A}_w(S_T)(\cdot))$, it holds that $\hat{h}(x_i) = y_i$ for all $i \in \{1,\ldots,n\}$.*

We are now ready for the proof of Theorem 34.

Proof of Theorem 34 The proof that classes with $\mathsf{VC}(\mathbb{H}) = \infty$ are not PAC learnable, and indeed also the lower bound $\mathcal{M}(\varepsilon,\delta) = \Omega\left(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon} + \frac{1}{\varepsilon}\log(\frac{1}{\delta})\right)$, follow by standard arguments from Vapnik and Chervonenkis (1974a); Blumer, Ehrenfeucht, Haussler, and Warmuth (1989); Ehrenfeucht, Haussler, Kearns, and Valiant (1989). Specifically, for any finite $k \leq \mathsf{VC}(\mathbb{H})$, letting $\mathcal{X}_k = \{x_1,\ldots,x_k\}$ be a set shattered by $\mathbb{H}$, and letting $\mathbb{H}_k$ be the class of all total functions $\mathcal{X}_k \to \{0,1\}$, any distribution P on $\mathcal{X}_k \times \{0,1\}$ realizable w.r.t. $\mathbb{H}_k$ can be extended to a distribution on $\mathcal{X} \times \{0,1\}$ realizable w.r.t. $\mathbb{H}$ with $P((\mathcal{X} \setminus \mathcal{X}_k) \times \{0,1\}) = 0$. Thus, any lower bound on the sample complexity of PAC learning the total concept class $\mathbb{H}_k$ is also a lower bound on the sample complexity of learning $\mathbb{H}$. In particular, Vapnik and Chervonenkis (1974a); Blumer, Ehrenfeucht, Haussler, and Warmuth (1989); Ehrenfeucht, Haussler, Kearns, and Valiant (1989) show a lower bound proportional to $\frac{k}{\varepsilon} + \frac{1}{\varepsilon}\log(\frac{1}{\delta})$ for the sample complexity of PAC learning $\mathbb{H}_k$, which is therefore also a lower bound on the sample complexity of PAC learning $\mathbb{H}$. Since this holds for all finite $k \leq \mathsf{VC}(\mathbb{H})$, it follows that partial concept classes with $\mathsf{VC}(\mathbb{H}) = \infty$ are not PAC learnable, and partial concept classes $\mathbb{H}$ with $\mathsf{VC}(\mathbb{H}) < \infty$ have optimal sample complexity $\mathcal{M}(\varepsilon,\delta) = \Omega\left(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon} + \frac{1}{\varepsilon}\log(\frac{1}{\delta})\right)$.

To prove the first upper bound on the optimal sample complexity when $\mathsf{VC}(\mathbb{H}) < \infty$ (which also implies the claim of PAC learnability for such classes), we begin by studying the function $\mathbb{A}$ from Lemma 35, following the analogous proof for total concept classes given by Haussler, Littlestone, and Warmuth (1994). In particular, for any distribution P realizable w.r.t. $\mathbb{H}$, and for

$(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ i.i.d. P, letting $\tilde{h}_n(\cdot) := \mathbb{A}(X_1, Y_1, \ldots, X_n, Y_n, \cdot)$, then by exchangeability of these $n+1$ samples and linearity of the expectation, we have that

$$\mathbb{E}\Big[\mathrm{er}_P(\tilde{h}_n)\Big] = \mathbb{E}[\mathbb{1}[\mathbb{A}(X_1, Y_1, \ldots, X_n, Y_n, X_{n+1}) \neq Y_{n+1}]]$$
$$= \frac{1}{(n+1)!} \sum_{\sigma \in \mathrm{Sym}(n+1)} \mathbb{E}\big[\mathbb{1}\big[\mathbb{A}(X_{\sigma(1)}, Y_{\sigma(1)}, \ldots, X_{\sigma(n)}, Y_{\sigma(n)}, X_{\sigma(n+1)}) \neq Y_{\sigma(n+1)}\big]\big]$$
$$= \mathbb{E}\left[\frac{1}{(n+1)!} \sum_{\sigma \in \mathrm{Sym}(n+1)} \mathbb{1}\big[\mathbb{A}(X_{\sigma(1)}, Y_{\sigma(1)}, \ldots, X_{\sigma(n)}, Y_{\sigma(n)}, X_{\sigma(n+1)}) \neq Y_{\sigma(n+1)}\big]\right] \leq \frac{\mathsf{VC}(\mathbb{H})}{n+1},$$

where this last inequality follows from the property (1) for $\mathbb{A}$ in Lemma 35, which holds with probability one for the sequence $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ since P is realizable w.r.t. $\mathbb{H}$.

To complete the proof, we again follow an argument of Haussler, Littlestone, and Warmuth (1994) to convert this algorithm, guaranteeing $\mathbb{E}[\mathrm{er}_P(\tilde{h}_n)] \leq \frac{\mathsf{VC}(\mathbb{H})}{n+1}$, into an algorithm with a bound on $\mathrm{er}_P(\hat{h})$ holding with high probability $1-\delta$. Specifically, let $m = \left\lfloor \frac{4\mathsf{VC}(\mathbb{H})}{\varepsilon} \right\rfloor \left\lceil \log_2\left(\frac{2}{\delta}\right) \right\rceil + \left\lceil \frac{32}{\varepsilon} \ln\left(\frac{2\lceil \log_2(2/\delta) \rceil}{\delta}\right) \right\rceil = O\left(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon} \log\left(\frac{1}{\delta}\right)\right)$, and let $(X_1, Y_1), \ldots, (X_m, Y_m)$ be i.i.d. P. Let $n = \left\lfloor \frac{4\mathsf{VC}(\mathbb{H})}{\varepsilon} \right\rfloor$, and let S_1 be the first n samples, S_2 the next n samples, and so on up to S_k, for $k = \left\lceil \log_2\left(\frac{2}{\delta}\right) \right\rceil$. Let T be the remaining $t := m - nk = \left\lceil \frac{32}{\varepsilon} \ln\left(\frac{2\lceil \log_2(2/\delta) \rceil}{\delta}\right) \right\rceil$ samples. For each $i \in \{1, \ldots, k\}$, let $h_i(\cdot) = \mathbb{A}(S_i, \cdot)$. Let $\hat{h}_m = \mathrm{argmin}_{h \in \{h_i : i \leq k\}} \sum_{(x,y) \in T} \mathbb{1}[h(x) \neq y]$. Define the learning algorithm for $\mathbb{H}$ as returning this $\hat{h}_m$, given $(X_1, Y_1), \ldots, (X_m, Y_m)$.

To show that this meets the PAC learning requirement, note that each $i \leq k$ has $\mathbb{E}[\mathrm{er}_P(h_i)] \leq \frac{\varepsilon}{4}$. Thus, by Markov's inequality, with probability at least $\frac{1}{2}$, $\mathrm{er}_P(h_i) \leq \frac{\varepsilon}{2}$. Since these h_i are independent, we have that with probability at least $1 - 2^{-k} \geq 1 - \frac{\delta}{2}$, at least one h_{i^*} has $\mathrm{er}_P(h_{i^*}) \leq \frac{\varepsilon}{2}$. Also, by a Chernoff bound, for each $i \leq k$, on the event $\mathrm{er}_P(h_i) \leq \frac{\varepsilon}{2}$,

$$\Pr\left(\frac{1}{t} \sum_{(x,y) \in T} \mathbb{1}[h_i(x) \neq y] > \frac{3}{4}\varepsilon \,\middle|\, h_i\right) \leq e^{-t\varepsilon/24},$$

while on the event $\mathrm{er}_P(h_i) > \varepsilon$,

$$\Pr\left(\frac{1}{t} \sum_{(x,y) \in T} \mathbb{1}[h_i(x) \neq y] \leq \frac{3}{4}\varepsilon \,\middle|\, h_i\right) \leq e^{-t\varepsilon/32}.$$

Thus, by the law of total probability (over these two events), and a union bound (over all $i \leq k$), with probability at least $1 - ke^{-t\varepsilon/32} \geq 1 - \frac{\delta}{2}$, if any i has $\mathrm{er}_P(h_i) \leq \frac{\varepsilon}{2}$, then the returned classifier $\hat{h}_m$ has $\mathrm{er}_P(\hat{h}_m) \leq \varepsilon$. By a union bound over the above two events, each of probability at least $1 - \frac{\delta}{2}$, we have that with probability at least $1 - \delta$, $\mathrm{er}_P(\hat{h}_m) \leq \varepsilon$. This completes the proof of the first upper bound.

To prove the second upper bound, again suppose $\mathsf{VC}(\mathbb{H}) < \infty$, and again let $\mathbb{A}$ be the function from Lemma 35. Let P be realizable w.r.t. $\mathbb{H}$, let $n \in \mathbb{N}$, and $S = \{(X_i, Y_i)\}_{i \in [n]} \sim P^n$. Since P is realizable w.r.t. $\mathbb{H}$, with probability one, any distribution P_0 supported on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$

is also realizable w.r.t. $\mathbb{H}$. Therefore, as established above, for any such distribution P_0, for $k = 3\mathsf{VC}(\mathbb{H})$ and $S_{P_0} \sim P_0^k$, $\mathbb{E}[\mathrm{er}_{P_0}(\mathbb{A}(S_{P_0}, \cdot))] \leq 1/3$. In particular, this implies that, given S and P_0, there exists a deterministic choice of $S_{P_0} \in \{(X_1, Y_1), \ldots, (X_n, Y_n)\}^k$ with $\mathrm{er}_{P_0}(\mathbb{A}(S_{P_0})) \leq 1/3$. Thus, $\mathbb{A}$ satisfies the requirement for $\mathbb{A}_w$ in Lemma 37, so that Lemma 37 implies that for a value $T = \lceil c_1 \log(n) \rceil$ (for numerical constant $c_1 \geq 1$), there exist $S_1, \ldots, S_T \in \{(X_1, Y_1), \ldots, (X_n, Y_n)\}^k$ such that, for $\hat{h}_n(\cdot) := \mathrm{Majority}(\mathbb{A}(S_1, \cdot), \ldots, \mathbb{A}(S_T, \cdot))$, it holds that $\hat{\mathrm{er}}_S(\hat{h}_n) = 0$. Moreover, note that $\hat{h}_n$ can be expressed as a compression scheme, with compression function κ such that $\kappa(S) = (S_1, \ldots, S_T)$ and reconstruction function ρ such that $\rho(S_1, \ldots, S_T) = \hat{h}_n$. Therefore, Lemma 36 implies that, with probability at least $1 - \delta$,

$$\mathrm{er}_P(\hat{h}_n) \leq \frac{c_2}{n - kT}\left(kT\log(n) + \log\left(\frac{1}{\delta}\right)\right)$$

for a numerical constant $c_2 \geq 1$. For any given $\varepsilon \in (0, 1)$, the right hand side above can be made less than ε for an appropriate choice of

$$n = O\left(\frac{1}{\varepsilon}\left(\mathsf{VC}(\mathbb{H})\log^2\left(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right),$$

so that we have that with probability at least $1 - \delta$, $\mathrm{er}_P(\hat{h}_n) \leq \varepsilon$. This completes the proof. ■

We conclude this section by noting the gap between the upper and lower bounds in Theorem 34. In the case of total concept classes $\mathbb{H}$, Hanneke (2016) showed that the optimal sample complexity of PAC learning is exactly $\Theta\left(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon} + \frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right)$. Whether this remains true for the more-general setting of partial concept classes is an interesting open question:

Open Question 6 *Does the optimal sample complexity $\mathcal{M}(\varepsilon, \delta)$ of PAC learning any partial concept class $\mathbb{H}$ always satisfy $\mathcal{M}(\varepsilon, \delta) = \Theta\left(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon} + \frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right)$?*

C.2. Agnostic PAC Learning

This section extends the learnability results to the agnostic setting, thus, together with the result above, fulfilling the complete claim of Theorem 3. We first provide a precise definition of agnostic learning, as formulating a definition appropriate for partial concept classes requires some care.

Recall that we think of partial concept classes as a general way for expressing assumptions on the data. Thus, we would like to define agnostic learning of a partial concept class $\mathbb{H}$ in a way that reflects the "distance from realizability" of a typical sample drawn from the source distribution P. This gives rise to the following quantities: for any $n \in \mathbb{N}$ and data sequence $S \in (\mathcal{X} \times \{0, 1\})^n$, define the *empirical error rate* of any partial concept h as $\hat{\mathrm{er}}_S(h) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}[h(x_i) \neq y_i]$. For a distribution P on $\mathcal{X} \times \{0, 1\}$, define the approximation error of $\mathbb{H}$ with respect to samples of size n as

$$\varepsilon^*(n) = \mathop{\mathbb{E}}_{S \sim P^n}\left[\min_{h \in \mathbb{H}} \hat{\mathrm{er}}_S(h)\right].$$

We will later see that $\varepsilon^\star(n)$ is non-decreasing in n and hence $\lim_{n\to\infty} \varepsilon^\star(n)$ exists and equals $\sup_n \varepsilon^\star(n)$ (see Lemma 39). We define the *approximation error* of $\mathbb{H}$ as

$$\mathrm{er}_P(\mathbb{H}) := \lim_{n\to\infty} \mathop{\mathbb{E}}_{S \sim P^n}\left[\min_{h \in \mathbb{H}} \hat{\mathrm{er}}_S(h)\right].$$

This measures how well a given partial concept class can *fit* data sets that can be sampled from P. In particular, note that $\mathrm{er}_P(\mathbb{H}) = 0$ if and only if P is *realizable* w.r.t. $\mathbb{H}$. In the agnostic PAC setting, we will be interested in achieving prediction error not-much-worse than $\mathrm{er}_P(\mathbb{H})$, as stated in the following definition.

Definition 38 (Agnostic PAC Learnability) *We say a partial concept class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ is* agnostically PAC learnable *if $\forall \varepsilon, \delta \in (0,1)$, $\exists \mathcal{M}(\varepsilon,\delta) \in \mathbb{N}$ and a learning algorithm $\mathbb{A}$ such that, for all distributions P on $\mathcal{X} \times \{0,1\}$, for $S \sim P^{\mathcal{M}(\varepsilon,\delta)}$, with probability at least $1-\delta$, $\mathrm{er}_P(\mathbb{A}(S)) \leq \mathrm{er}_P(\mathbb{H}) + \varepsilon$. The quantity $\mathcal{M}(\varepsilon,\delta)$ is known as the* sample complexity *of $\mathbb{A}$ for agnostic PAC learning, and the* optimal *sample complexity of agnostic PAC learning for $\mathbb{H}$ is defined as the minimum achievable value of $\mathcal{M}(\varepsilon,\delta)$ for each given ε, δ.*

The following lemma shows that $\mathrm{er}_P(\mathbb{H})$ is indeed well-defined for every $\mathbb{H}$. Also, it shows that for total classes, our definition of learnability is not easier to satisfy than the classical one.[12] (I.e., any learning algorithm which agnostically learns a total class $\mathbb{H}$ according to Definition 38 also agnostically learns $\mathbb{H}$ in the classical PAC sense).

Lemma 39 *Let $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$ and let P be a distribution over $\mathcal{X} \times \{0,1\}$. Define*

$$\varepsilon^\star(n) = \mathop{\mathbb{E}}_{S\sim P^n}\left[\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_S(h)\right].$$

Then:

1. *For every $n \geq 2$, $\varepsilon^\star(n) \geq \varepsilon^\star(n-1)$, and in particular $\lim_{n\to\infty} \varepsilon^\star(n)$ exists.*
2. *Also,*
$$\mathrm{er}_P(\mathbb{H}) = \lim_{n\to\infty} \varepsilon^\star(n) \leq \inf_{h\in\mathbb{H}} \mathrm{er}_P(h),$$
and if in addition $\mathbb{H}$ contains only total concepts and $\mathsf{VC}(\mathbb{H}) < \infty$ then the above inequality is satisfied with an equality.

Proof

We begin with the first item. Let $n \geq 2$, and let S be a sequence $(X_1,Y_1),\ldots,(X_n,Y_n)$ drawn i.i.d. from P. For $i \in [n]$, denote by S_{-i} the subsequence obtained by removing (X_i,Y_i). For any $h \in \mathbb{H}$ we get from the definition of the empirical error rate and elementary double counting that

$$\hat{\mathrm{er}}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathrm{er}}_{S_{-i}}(h).$$

12. Note that one could naturally adapt the classical definition of agnostic PAC learning to partial concept classes: namely, aiming for $\mathrm{er}_P(\mathbb{A}(S)) \leq \inf_{h\in\mathbb{H}} \mathrm{er}_P(h) + \varepsilon$, with probability at least $1-\delta$. However, unlike the criterion in Definition 38, this alternative criterion would *not* recover the implication that, for any given distribution P realizable w.r.t. $\mathbb{H}$, agnostic PAC learning under P implies PAC learning under P. For instance, for $\mathcal{X} = [0,1]$ and $\mathbb{H}$ all partial functions with image $\{0,\star\}$ and having finite support, the distribution P uniform on $\mathcal{X} \times \{0\}$ is realizable w.r.t. $\mathbb{H}$, but $\inf_{h\in\mathbb{H}} \mathrm{er}_P(h) = 1$, so that even the algorithm that returns $x \mapsto 1$ satisfies the alternative agnostic PAC criterion. This is another reason to use the definition in terms of $\mathrm{er}_P(\mathbb{H})$ as stated in Definition 38. Note however that all of our results on agnostic learning apply for the alternative definition.

This implies that

$$\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_S(h) \geq \frac{1}{n}\sum_{i=1}^{n} \min_{h\in\mathbb{H}} \hat{\mathrm{er}}_{S_{-i}}(h).$$

Taking $\mathbb{E}_{S\sim P^n}$ on both sides of this inequality, and noting that for each i, $S_{-i} \sim P^{n-1}$, we get

$$\mathop{\mathbb{E}}_{S\sim P^n}\left[\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_S(h)\right] \geq \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{S_{-i}\sim P^{n-1}}\left[\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_{S_{-i}}(h)\right],$$

so

$$\varepsilon^\star(n) \geq \frac{1}{n}\sum_{i=1}^{n} \varepsilon^\star(n-1) = \varepsilon^\star(n-1),$$

as required.

Let us now prove the second item. Notice that for every $h \in \mathbb{H}$ and $n \in \mathbb{N}$:

$$\mathrm{er}_P(h) = \mathop{\mathbb{E}}_{S\sim P^n} \hat{\mathrm{er}}_S(h) \geq \varepsilon^\star(n),$$

and therefore also

$$\inf_{h\in\mathbb{H}} \mathrm{er}_P(h) \geq \varepsilon^\star(n).$$

Letting $n \to \infty$ yields the stated inequality. In addition if $\mathbb{H} \subseteq \{0,1\}^{\mathcal{X}}$ contains only total concepts and satisfies $\mathsf{VC}(\mathbb{H}) < \infty$ then, by uniform-convergence, for $S \sim P^n$, with probability at least $1-\delta$ we have:

$$\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_S(h) \geq \inf_{h\in\mathbb{H}} \mathrm{er}_P(h) - O\Big(\sqrt{\frac{\mathsf{VC}(\mathbb{H}) + \log(\frac{1}{\delta})}{n}}\Big).$$

Taking $\delta = \frac{1}{n}$ and letting $n \to \infty$ we have the converse inequality $\lim_{n\to\infty} \varepsilon^\star(n) \geq \inf_{h\in\mathbb{H}} \mathrm{er}_P(h)$, and hence an equality. ■

In particular, in the case of total concept classes $\mathbb{H}$, this implies that any learning algorithm that is an agnostic PAC learner by our definition is also an agnostic PAC learner in the traditional definition, and vice versa.

The portion of Theorem 3 concerning agnostic PAC learnability is summarized in the following specialized statement.

Theorem 40 (Agnostic PAC Learnability) *The following statements are equivalent for any partial concept class $\mathbb{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$.*

- $\mathsf{VC}(\mathbb{H}) < \infty$.
- $\mathbb{H}$ *is agnostically PAC learnable.*

Thus, the conditions for agnostic PAC learnability of $\mathbb{H}$ are the same as for (realizable) PAC learnability, and recover the known conditions for agnostic learnability of total concept classes $\mathbb{H}$ (both realizable and agnostic).

As we did for the realizable case, we will prove a more-detailed result on agnostic PAC learning, which also establishes upper and lower bounds on the optimal sample complexity. In particular, Theorem 40 follows as an immediate implication.

Theorem 41 (Agnostic PAC Sample Complexity) *For any partial concept class $\mathbb{H}$ with $\mathsf{VC}(\mathbb{H}) < \infty$, the optimal sample complexity of agnostically PAC learning $\mathbb{H}$, $\mathcal{M}(\varepsilon, \delta)$, satisfies the following bounds:*

- $\mathcal{M}(\varepsilon, \delta) = O\left(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon^2} \log^2\left(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon}\right) + \frac{1}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right)$.
- $\mathcal{M}(\varepsilon, \delta) = \Omega\left(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon^2} + \frac{1}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right)$.

Moreover, if $\mathsf{VC}(\mathbb{H}) = \infty$, then $\mathbb{H}$ is not agnostically PAC learnable.

While the lower bound will follow from standard approaches, to prove the upper bound we will use a technique introduced by David, Moran, and Yehudayoff (2016), which reduces agnostic learning to realizable learning. This technique makes use of two main components: generalization bounds for sample compression schemes, and a construction of a compression scheme based on Boosting (to which the generalization bounds are then applied). Both of these components are well known. For the Boosting component, we rely on Lemma 37. We restate the relevant result for compression schemes here for completeness. Specifically, the following lemma is a variation on a classic result due to Graepel, Herbrich, and Shawe-Taylor (2005).[13]

Lemma 42 (Agnostic Compression Generalization Bound) *There exists a finite numerical constant $c > 0$ such that, for any compression scheme (κ, ρ), for any $n \in \mathbb{N}$ and $\delta \in (0, 1)$, for any distribution P on $\mathcal{X} \times \{0, 1\}$, for $S \sim P^n$, letting $B(S, \delta) := \frac{1}{n}\left(|\kappa(S)| \log(n) + \log\left(\frac{1}{\delta}\right)\right)$, with probability at least $1 - \delta$,*

$$|\mathrm{er}_P(\rho(\kappa(S))) - \hat{\mathrm{er}}_S(\rho(\kappa(S)))| \leq c\sqrt{\hat{\mathrm{er}}_S(\rho(\kappa(S)))B(S, \delta)} + cB(S, \delta).$$

We apply Lemma 42 to a boosting-based compression scheme to obtain the following intermediate result, representing the key component in the upper bound claimed in Theorem 41. This also supplies the algorithm supporting the claims regarding structural risk minimization in Section 4.1.

Lemma 43 *For any partial concept class $\mathbb{H}$ with $\mathsf{VC}(\mathbb{H}) < \infty$, there is a learning algorithm $\mathbb{A}$ such that, for any distribution P on $\mathcal{X} \times \{0, 1\}$, any $m \in \mathbb{N}$, and $\delta \in (0, 1)$, for $S \sim P^m$, letting $\hat{\mathrm{er}}_S(\mathbb{H}) := \min_{h \in \mathbb{H}} \hat{\mathrm{er}}_S(h)$, with probability at least $1 - \delta$, the output $\hat{h} := \mathbb{A}(S)$ satisfies*

$$\mathrm{er}_P(\hat{h}) \leq \hat{\mathrm{er}}_S(\mathbb{H}) + c\sqrt{\hat{\mathrm{er}}_S(\mathbb{H})\frac{1}{m}\left(\mathsf{VC}(\mathbb{H})\log^2(m) + \log\left(\frac{1}{\delta}\right)\right)} + \frac{c}{m}\left(\mathsf{VC}(\mathbb{H})\log^2(m) + \log\left(\frac{1}{\delta}\right)\right)$$

for a finite numerical constant c.

13. The original result of Graepel, Herbrich, and Shawe-Taylor (2005) does not include the $\hat{\mathrm{er}}_S(\rho(\kappa(S)))$ factor inside the square root. However, this variant follows by the same argument, simply substituting the empirical Bernstein bound rather than Hoeffding's inequality. See (Maurer and Pontil, 2009) for a similar result.

Proof We follow a reduction-to-realizable technique of David, Moran, and Yehudayoff (2016). Specifically, let $\mathbb{A}_w$ be a PAC learning algorithm for $\mathbb{H}$ (for the realizable case) achieving the optimal sample complexity $\mathcal{M}_{\mathrm{RE}}(\varepsilon', \delta')$ for PAC learning $\mathbb{H}$ (in the realizable case), where $\varepsilon' = \delta' = 1/3$. Note that, without loss of generality, we may suppose $\mathbb{A}_w$ outputs total functions (e.g., replacing all $\star$ with 0 cannot increase $\mathrm{er}_P(h)$ for any h). We will first explain that this may serve as a weak learning algorithm $\mathbb{A}_w$ in Lemma 37. Specifically, let $k = \mathcal{M}_{\mathrm{RE}}(1/3, 1/3)$. Given any $n \in \mathbb{N}$ and any sequence $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{0,1\}$ realizable w.r.t. $\mathbb{H}$, any distribution P on $\mathcal{X} \times \{0,1\}$ with $P(\{(x_1, y_1), \ldots, (x_n, y_n)\}) = 1$ is realizable w.r.t. $\mathbb{H}$, and therefore guarantees that, for $S \sim P^k$, with probability at least $2/3$, $\mathrm{er}_P(\mathbb{A}_w(S)) \leq 1/3$. In particular, this implies there exists at least one $S_P \in \{(x_1, y_1), \ldots, (x_n, y_n)\}^k$ with $\mathrm{er}_P(\mathbb{A}_w(S)) \leq 1/3$.

Now, let P be any distribution on $\mathcal{X} \times \{0,1\}$, let $m \in \mathbb{N}$, and let $S = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$ have distribution P^m. Let R denote the longest subsequence of $(X_1, Y_1), \ldots, (X_m, Y_m)$ that is realizable w.r.t. $\mathbb{H}$ (breaking ties based on a fixed measurable total ordering of such sequences). If $|R| = 0$, then (arbitrarily) define $\hat{h}$ as the all-0 function $\hat{h}(x) = 0$. Otherwise, if $|R| > 0$, by the above property of $\mathbb{A}_w$, together with Lemma 37, for $T = \lceil c' \log(|R|) \rceil$ (for a numerical constant c'), there exist $S_1, \ldots, S_T \in R^k$ such that, letting $\hat{h}(\cdot) := \mathrm{Majority}(\mathbb{A}_w(S_1)(\cdot), \ldots, \mathbb{A}_w(S_T)(\cdot))$, we have $\hat{\mathrm{er}}_R(\hat{h}) = 0$.

In particular, this implies $\hat{\mathrm{er}}_S(\hat{h}) \leq \frac{m-|R|}{m} = \hat{\mathrm{er}}_S(\mathbb{H})$. Moreover, $\hat{h}$ is the output of the compression scheme that selects $\kappa(S) = (S_1, \ldots, S_T)$ and $\rho(\kappa(S)) = \hat{h}$ (or in the case $|R| = 0$, $\kappa(S) = \{\}$ and $\rho(\kappa(S)) = \hat{h}$). Therefore, Lemma 42 implies that, with probability at least $1 - \delta$,

$$\mathrm{er}_P(\hat{h}) \leq \hat{\mathrm{er}}_S(\hat{h}) + c'' \sqrt{\hat{\mathrm{er}}_S(\hat{h}) \frac{1}{m} \left(kT \log(m) + \log\left(\frac{1}{\delta}\right)\right)} + c'' \frac{1}{m} \left(kT \log(m) + \log\left(\frac{1}{\delta}\right)\right)$$

$$\leq \hat{\mathrm{er}}_S(\mathbb{H}) + c''' \sqrt{\hat{\mathrm{er}}_S(\mathbb{H}) \frac{1}{m} \left(\mathsf{VC}(\mathbb{H}) \log^2(m) + \log\left(\frac{1}{\delta}\right)\right)} + c''' \frac{1}{m} \left(\mathsf{VC}(\mathbb{H}) \log^2(m) + \log\left(\frac{1}{\delta}\right)\right)$$

for appropriate numerical constants c'', c''', where the last inequality is due to Theorem 34. ■

We are now ready for the proof of Theorem 41.

Proof of Theorem 41 As was true in the proof of Theorem 34, the proof that classes with $\mathsf{VC}(\mathbb{H}) = \infty$ are not agnostically PAC learnable, and the lower bound $\mathcal{M}(\varepsilon, \delta) = \Omega\left(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon^2} + \frac{1}{\varepsilon^2} \log(\frac{1}{\delta})\right)$, follow by standard arguments from Vapnik and Chervonenkis (1974a); Anthony and Bartlett (1999). Specifically, for any finite $k \leq \mathsf{VC}(\mathbb{H})$, letting $\mathcal{X}_k = \{x_1, \ldots, x_k\}$ be a set shattered by $\mathbb{H}$, and letting $\mathbb{H}_k$ be the class of total functions $h : \mathcal{X}_k \to \{0,1\}$, all distributions $P^{(k)}$ on $\mathcal{X}_k \times \{0,1\}$ can be extended to a distribution P on $\mathcal{X} \times \{0,1\}$ with $P((\mathcal{X} \setminus \mathcal{X}_k) \times \{0,1\}) = 0$. Moreover, letting $(X_1, Y_1), (X_2, Y_2), \ldots$ be i.i.d. P, we always have

$$\begin{aligned}
\mathrm{er}_P(\mathbb{H}) &= \sup_n \mathop{\mathbb{E}}_{S \sim P^n} \left[\min_{h \in \mathbb{H}} \hat{\mathrm{er}}_S(h) \right] \\
&\leq \sup_n \inf_{h \in \mathbb{H}} \mathop{\mathbb{E}}_{S \sim P^n} [\hat{\mathrm{er}}_S(h)] = \inf_{h \in \mathbb{H}} \mathrm{er}_P(h) = \min_{h \in \mathbb{H}_k} \mathrm{er}_{P^{(k)}}(h).
\end{aligned}$$

Additionally, note that

$$\begin{aligned}\mathrm{er}_P(\mathbb{H}) &\geq \lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{k} \min_{y\in\{0,1\}} \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}[X_t = x_i]\mathbb{1}[Y_t \neq y]\right] \\ &= \sum_{i=1}^{k} \mathbb{E}\left[\min_{y\in\{0,1\}} \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}[X_t = x_i]\mathbb{1}[Y_t \neq y]\right] \\ &= \sum_{i=1}^{k} \min_{y\in\{0,1\}} P(\{(x_i, 1-y)\}) = \min_{h\in\mathbb{H}_k} \mathrm{er}_{P^{(k)}}(h),\end{aligned}$$

where we have used the Dominated Convergence Theorem, continuity of the $\min$, and the Strong Law of Large Numbers. Thus, $\mathrm{er}_P(\mathbb{H}) = \min_{h\in\mathbb{H}_k} \mathrm{er}_{P^{(k)}}(h)$. This means that $\mathcal{M}(\varepsilon,\delta)$ can be no smaller than the sample complexity $M_k(\varepsilon,\delta)$ of agnostically learning the total concept class $\mathbb{H}_k$ on $\mathcal{X}_k$, in the traditional (total concepts) sense: that is, there exists an algorithm that, for all distributions $P^{(k)}$ on $\mathcal{X}_k \times \{0,1\}$, from $M_k(\varepsilon,\delta)$ i.i.d. samples from $P^{(k)}$, outputs an $\hat{h}$ with $\mathrm{er}_{P^{(k)}}(\hat{h}) - \min_{h\in\mathbb{H}_k} \mathrm{er}_{P^{(k)}}(h) \leq \varepsilon$, with probability at least $1-\delta$. Since $\mathcal{X}_k$ is shattered, the standard lower bounds based on VC dimension imply a lower bound $M_k(\varepsilon,\delta) = \Omega\big(\frac{k}{\varepsilon^2} + \frac{1}{\varepsilon^2}\log\big(\frac{1}{\delta}\big)\big)$ (Vapnik and Chervonenkis, 1974a; Anthony and Bartlett, 1999; Kontorovich and Pinelis, 2019). In particular, the sample complexity lower bound $\mathcal{M}(\varepsilon,\delta) = \Omega\Big(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon^2} + \frac{1}{\varepsilon^2}\log\big(\frac{1}{\delta}\big)\Big)$ for the case of $\mathsf{VC}(\mathbb{H}) < \infty$ follows immediately, as does the fact that having $\mathsf{VC}(\mathbb{H}) = \infty$ implies that $\mathbb{H}$ is not agnostically PAC learnable.

To prove the upper bound on the optimal sample complexity when $\mathsf{VC}(\mathbb{H}) < \infty$ (which also implies the claim of agnostic PAC learnability for such classes), consider the algorithm $\mathbb{A}$ from Lemma 43. From that lemma, we have that for any distribution P, sample size $m \in \mathbb{N}$, and $\delta \in (0,1)$, for $S = \{(X_1,Y_1),\ldots,(X_m,Y_m)\} \sim P^m$, with probability at least $1-\delta/2$, the returned classifier $\hat{h} = \mathbb{A}(S)$ satisfies

$$\begin{aligned}\mathrm{er}_P(\hat{h}) &\leq \hat{\mathrm{er}}_S(\mathbb{H}) + c\sqrt{\hat{\mathrm{er}}_S(\mathbb{H})\frac{1}{m}\left(\mathsf{VC}(\mathbb{H})\log^2(m) + \log\left(\frac{2}{\delta}\right)\right)} + \frac{c}{m}\left(\mathsf{VC}(\mathbb{H})\log^2(m) + \log\left(\frac{2}{\delta}\right)\right) \\ &\leq \hat{\mathrm{er}}_S(\mathbb{H}) + c'\sqrt{\frac{1}{m}\left(\mathsf{VC}(\mathbb{H})\log^2(m) + \log\left(\frac{1}{\delta}\right)\right)}\end{aligned}$$

for appropriate numerical constants c, c'.

Taking $m = \left\lceil c''\frac{1}{\varepsilon^2}\left(\mathsf{VC}(\mathbb{H})\log^2\left(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon}\right) + \log\big(\frac{1}{\delta}\big)\right)\right\rceil$ for a suitable numerical constant c'' we get from the above that with probability at least $1-\delta/2$,

$$\mathrm{er}_P(\hat{h}) \leq \hat{\mathrm{er}}_S(\mathbb{H}) + \frac{\varepsilon}{2}.$$

It remains to replace the empirical value $\hat{\mathrm{er}}_S(\mathbb{H}) := \min_{h\in\mathbb{H}} \hat{\mathrm{er}}_S(h)$ with our benchmark $\mathrm{er}_P(\mathbb{H})$. We do this by observing that the random variable $\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_S(h)$ is concentrated around its mean. Indeed, this random variable is a function of the i.i.d. sequence $(X_1,Y_1),\ldots,(X_m,Y_m)$, and changing the value of any (X_i,Y_i) can change $\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_S(h)$ by at most $\frac{1}{m}$. By McDiarmid's inequality

$$\Pr\left(\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_S(h) > \mathop{\mathbb{E}}_{S'\sim P^m}\left[\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_{S'}(h)\right] + \frac{\varepsilon}{2}\right) \leq e^{-\frac{\varepsilon^2 m}{2}},$$

and for our value of m this probability is less than $\delta/2$. By the union bound, with probability at least $1-\delta$,

$$\mathrm{er}_P(\hat{h}) \leq \min_{h\in\mathbb{H}} \hat{\mathrm{er}}_S(h) + \frac{\varepsilon}{2} \leq \mathop{\mathbb{E}}_{S'\sim P^m}\left[\min_{h\in\mathbb{H}} \hat{\mathrm{er}}_{S'}(h)\right] + \varepsilon \leq \mathrm{er}_P(\mathbb{H}) + \varepsilon.$$

So taking the algorithm $\mathbb{A}$ from Lemma 43 meets the requirement. ■

We note that, in the special case of total concept classes, the optimal sample complexity of agnostic PAC learning is known to be exactly $\mathcal{M}(\varepsilon,\delta) = \Theta\left(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon^2} + \frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ (Talagrand, 1994; see van der Vaart and Wellner, 1996, Theorems 2.14.1 and 2.6.7). However, the proof of this refined upper bound relies on a technique known as *chaining*, and moreover relies on uniform convergence, which can fail for partial concept classes. Thus, the following question remains open:

Open Question 7 *Does the optimal sample complexity, $\mathcal{M}(\varepsilon,\delta)$, of agnostically PAC learning any partial concept class $\mathbb{H}$ satisfy $\mathcal{M}(\varepsilon,\delta) = \Theta\left(\frac{\mathsf{VC}(\mathbb{H})}{\varepsilon^2} + \frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$?*

Appendix D. Proofs of Results on Traditional Learning Principles

Proof of Theorem 5 Let $\mathcal{X} = \mathbb{N}$ and let $\mathbb{H}_\infty$ be the partial concept class from the proof of Theorem 11, having $\mathsf{VC}(\mathbb{H}_\infty) = 1$ while any disambiguation $\bar{\mathbb{H}}_\infty$ of $\mathbb{H}_\infty$ must have $\mathsf{VC}(\bar{\mathbb{H}}_\infty) = \infty$. Also let h_0 be a concept with $h_0(x) = 0$ everywhere, and define $\mathbb{H} = \mathbb{H}_\infty \cup \{h_0\}$. In particular, it follows from Lemma 31 that any pair of points $(x_1, x_2) \in \mathcal{X}^2$ have at most two patterns of labels in $\{0,1\}^2$ realizable w.r.t. $\mathbb{H}_\infty$. Therefore, adding h_0 to the class does not increase its VC dimension, since it can at most increase the number of patterns to three, but not to four. Hence, $\mathsf{VC}(\mathbb{H}) = 1$.

Now consider any total concept class $\bar{\mathbb{H}}$. First, if $\bar{\mathbb{H}}$ is not a disambiguation of $\mathbb{H}$, then there exists a finite sequence S of points $(x,y) \in \mathcal{X}\times\{0,1\}$ that is realizable w.r.t. $\mathbb{H}$ but not realizable w.r.t. $\bar{\mathbb{H}}$. Without loss of generality, suppose all elements of S are distinct. Then letting P be the uniform distribution on S, P is realizable w.r.t. $\mathbb{H}$. However, for any learning algorithm producing hypotheses $\hat{h}$ in $\bar{\mathbb{H}}$, regardless of how many i.i.d. samples it is provided with, it will always have $\mathrm{er}_P(\hat{h}) \geq 1/|S|$, so that its sample complexity for any $\varepsilon < 1/|S|$ is infinite: that is, it does not satisfy the PAC learning requirement.

On the other hand, consider the case where $\bar{\mathbb{H}}$ is a disambiguation of $\mathbb{H}$. Then $\bar{\mathbb{H}}$ is also a disambiguation of $\mathbb{H}_\infty$, and therefore, by the property of $\mathbb{H}_\infty$ from Theorem 11, it must be that $\mathsf{VC}(\bar{\mathbb{H}}) = \infty$. Now let $U_1, U_2, \ldots$ be a sequence of disjoint subsets of $\mathcal{X}$ with $|U_i| = i$, such that each U_i is shattered by $\bar{\mathbb{H}}$; for instance, such a sequence can be constructed by first considering shattered sets of sizes 4^i for each i, each of which has at least $(1/2)4^i$ elements not appearing in any of the smaller sets. Then consider an ERM algorithm $\mathbb{A}$ for $\bar{\mathbb{H}}$ that, given any data sequence S whose elements are all contained in $U_i \times \{0\}$ for one of the $i \in \mathbb{N}$, $\mathbb{A}(S)$ returns a function h with $h(x) = 1$ for every $x \in U_i$ that does not appear in S. $\mathbb{A}(S)$ may be defined as any ERM in the case S is not contained in any of the $U_i \times \{0\}$ sets.

Now, given any sample size m, take P uniform on $U_{cm} \times \{0\}$ for an integer $c \geq 2$. Note that this P is realizable w.r.t. $\mathbb{H}$ since $h_0 \in \mathbb{H}$. However, for $S \sim P^m$, we will have S contained in $U_{cm} \times \{0\}$, but at least $(c-1)m$ of the cm elements of this set will not be present in S, and therefore $\mathbb{A}(S)(x)$ will be 1 for at least $(c-1)m$ elements of U_{cm}. Thus, $\mathrm{er}_P(\mathbb{A}(S)) \geq 1 - \frac{1}{c}$. By choosing c large, this can be made arbitrarily close to 1.

In particular, this implies there is no finite sample size m at which $\Pr_{S\sim P^m}(\mathrm{er}_P(\mathbb{A}(S)) < 1-\frac{1}{c}) > 0$ holds for all P realizable w.r.t. $\mathbb{H}$, so that $\mathbb{A}$ is not a PAC learning algorithm for $\mathbb{H}$. ■

Proof of Theorem 6 Let $\mathbb{H}_\infty$ be the class from Theorem 11. Consider a finite realizable data sequence S, and assume without loss of generality that its elements are all distinct. Letting P be the uniform distribution on S, and aiming at a prediction error at most $\varepsilon < 1/|S|$, the algorithm running on a large enough sample must output with positive probability a function that realizes S. Since we can choose S as any finite data sequence realizable w.r.t. $\mathbb{H}$, the image of the learning algorithm disambiguates $\mathbb{H}_\infty$, and hence by Theorem 11 it must have infinite VC dimension. ■

Proof of Theorem 7 For the first claim, note that Theorem 34 implies that, for any realizable data set, and any distribution supported on those points, there exists a sequence of $O(\mathsf{VC}(\mathbb{H}))$ data points from the m points that can be fed into the algorithm from Theorem 34 to get a hypothesis $\hat{h}$ with prediction error at most $1/3$ under that distribution. In conjunction with a standard boosting algorithm (e.g., α-boost; see Lemma 37 of Section C.2), we arrive at a sequence $\hat{h}_1, \ldots, \hat{h}_T$, for $T = O(\log(m))$, where each $\hat{h}_i$ is based on applying the algorithm from Theorem 34 to some subset of $O(\mathsf{VC}(\mathbb{H}))$ data points, and where $\mathrm{Majority}(\hat{h}_1, \ldots, \hat{h}_T)$ is correct on the entire set of m data points. Thus, we can specify this classifier using $k = O(\mathsf{VC}(\mathbb{H})\log(m))$ data points. Together with $O(k\log(k))$ bits to encode an order of the points so as to recover precisely which points correspond to which $\hat{h}_i$, this yields the claimed compression scheme.

For the second claim, consider the partial concept class $\mathbb{H}_\infty$ constructed in the proof of Theorem 11. Recall that $\mathsf{VC}(\mathbb{H}_\infty) = 1$, and $\mathbb{H}_\infty$ is formed by taking disjoint copies of $\mathbb{H}_n$ so that $|\bar{\mathbb{H}}_n| \geq n^{(\log(n))^{1-o(1)}}$ for any disambiguation $\bar{\mathbb{H}}_n$ of $\mathbb{H}_n$.

Now suppose there is a compression scheme of some size s_m depending on the sample size m. Then note that this also supplies a compression scheme for $\mathbb{H}_m$ for data sets of size m. But then Proposition 14 implies there exists a disambiguation $\bar{\mathbb{H}}_m$ of $\mathbb{H}_m$ of size at most $(cm/s_m)^{s_m}$ for a numerical constant c. On the other hand, Theorem 11 provides that $\bar{\mathbb{H}}_m$ must have size at least $m^{(\log(m))^{1-o(1)}}$. Therefore,

$$\left(\frac{cm}{s_m}\right)^{s_m} \geq m^{(\log(m))^{1-o(1)}},$$

which implies $s_m \geq c'(\log(m))^{1-o(1)}$ for a numerical constant c'. ■

To prove Theorem 8, we will rely on the following lemma.

Lemma 44 *Any partial concept class $\mathbb{H}$ with $\mathsf{LD}(\mathbb{H}) < \infty$ admits a compression scheme of size $\mathsf{LD}(\mathbb{H})$.*

Proof Consider the optimal online learning algorithm $\mathbb{A}_{\mathrm{SOA}}$ of Littlestone (1988), defined formally (and for partial concept classes) in the proof of Theorem 47 in Section E.2. For any data sequence $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{0, 1\}$ realizable w.r.t. $\mathbb{H}$, we initialize $K = \{\}$. We then find a point (x_i, y_i) in the sequence with $\mathbb{A}_{\mathrm{SOA}}(K)(x_i) \neq y_i$, if one exists, and we add (x_i, y_i) to the set K. We repeat this until there are no remaining points (x_i, y_i) with $\mathbb{A}_{\mathrm{SOA}}(K)(x_i) \neq y_i$. This specifies a compression function $\kappa((x_1, y_1), \ldots, (x_n, y_n)) = K$, and the reconstruction function is simply $\mathbb{A}_{\mathrm{SOA}}(K)$ (noting that this is invariant to the order of K). By Theorem 47, we have $|K| \leq \mathsf{LD}(\mathbb{H})$,

so that this specifies a compression scheme of size $\mathsf{LD}(\mathbb{H})$. ■

Proof of Theorem 8 Consider again the partial concept class $\mathbb{H}_\infty$ defined in the proof of Theorem 11, and recall that $\mathrm{TD}(\mathbb{H}_\infty) \leq 2$. On the other hand, as established in Theorem 7, $\mathbb{H}_\infty$ does not admit a bounded-size compression scheme. Together with Lemma 44, it must be that $\mathsf{LD}(\mathbb{H}_\infty) = \infty$. ■

Appendix E. Online Learning: Detailed Results and Specific Bounds

The *online learning* model for total concept classes is a classic framework, introduced by Littlestone (1988), in which there is a data sequence $(X_1, Y_1), (X_2, Y_2), \ldots$, which is considered *arbitrary* (possibly adversarially-chosen), and we are interested in the number of *mistakes* made by a learning algorithm $\mathbb{A}$: that is, the number of times t with $\mathbb{A}((X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1}))(X_t) \neq Y_t$.

E.1. Online Learning in the Realizable Case

As with the PAC model, the online model has two main variants, depending on whether we suppose the sequence (X_i, Y_i) is realizable w.r.t. $\mathbb{H}$ (known as the *realizable* case, or the *mistake bound* model), or is arbitrary (known as *agnostic* online learning). We start by presenting the realizable case. Specifically, we have the following definition.

Definition 45 (Realizable Online Learnability) *A partial concept class $\mathbb{H}$ is* online learnable *if there exists a bound $\mathsf{MB}(\mathbb{H}) < \infty$ such that, for every $T \in \mathbb{N}$, there exists a learning algorithm $\mathbb{A}$ that, for every sequence $(X_1, Y_1), \ldots, (X_T, Y_T) \in \mathcal{X} \times \{0, 1\}$ realizable w.r.t. $\mathbb{H}$, $\sum_{t=1}^{T} \mathbb{1}[\mathbb{A}((X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1}))(X_t) \neq Y_t] \leq \mathsf{MB}(\mathbb{H})$.*

Littlestone (1988) proved that, in the realizable case w.r.t. any *total* concept class $\mathbb{H}$, there exists a learning algorithm making a *bounded* number of mistakes if and only if the combinatorial parameter $\mathsf{LD}(\mathbb{H})$ is finite; this parameter has since become known as the *Littlestone dimension*, and is a key quantity of interest in many learning models. Here we use the extended definition of Littlestone dimension stated in Definition 27 for *partial* concept classes.

We have the following result, extending Littlestone's classic theorem to hold for partial concept classes. In particular, this supplies part of the claim in Theorem 15 from Section 2.5.

Theorem 46 (Realizable Online Learnability) *The following statements are equivalent for a partial concept class $\mathbb{H} \subseteq \{0, 1, \star\}^{\mathcal{X}}$.*

- $\mathsf{LD}(\mathbb{H}) < \infty$.
- *$\mathbb{H}$ is online learnable.*

E.2. Proof of Theorem 46 and Quantitative Mistake Bounds

Theorem 46 will follow as an immediate implication of the following more detailed result, which also supplies a concrete expression for the optimal mistake bound $\mathsf{MB}(\mathbb{H})$.

Theorem 47 (Optimal Mistake Bound) *For any partial concept class $\mathbb{H}$, there exists an online learning algorithm making at most $\mathsf{LD}(\mathbb{H})$ mistakes on any realizable data sequence. Moreover, for any finite $d \leq \mathsf{LD}(\mathbb{H})$ and any (possibly randomized) learning algorithm $\mathbb{A}$, there exists a realizable data sequence of length d on which $\mathbb{A}$ makes an expected number of mistakes at least $d/2$.*

Proof The proof of this result is essentially identical to the well-known proof for the special case of total concept classes. We include the full details nonetheless, for completeness.

We begin with the upper bound. For any sequence $S \in (\mathcal{X} \times \{0,1\})^*$, define $\mathbb{H}_S = \{h \in \mathbb{H} : \hat{\mathrm{er}}_S(h) = 0\}$. Then, following Littlestone (1988), we first note that the Littlestone dimension can be interpreted inductively, and in particular, for any non-empty $\mathbb{H}' \subseteq \mathbb{H}$, for every $x \in \mathcal{X}$, there exists $y \in \{0,1\}$ such that $\mathsf{LD}(\mathbb{H}'_{(x,y)}) < \mathsf{LD}(\mathbb{H}')$, where we interpret $\mathsf{LD}(\{\}) = -1$. Based on this, if $\mathsf{LD}(\mathbb{H}) < \infty$, we define the *Standard Optimal Algorithm* (SOA) $\mathbb{A}_{\mathrm{SOA}}$ as follows. For any $S \in (\mathcal{X} \times \{0,1\})^*$ and $x \in \mathcal{X}$, let $\mathbb{A}_{\mathrm{SOA}}(S)(x) = \mathrm{argmax}_{y \in \{0,1\}} \mathsf{LD}(\mathbb{H}_{S \cup \{(x,y)\}})$, breaking ties to favor $y = 0$ (or by any other rule). In particular, note that if S is realizable w.r.t. $\mathbb{H}$, then at most one $y \in \{0,1\}$ can have $\mathsf{LD}(\mathbb{H}_{S\cup\{(x,y)\}}) = \mathsf{LD}(\mathbb{H}_S)$ (it is also possible that neither value y has this property). Thus, for every sequence $(x_1, y_1), \ldots, (x_T, y_T)$ realizable w.r.t. $\mathbb{H}$, at each t for which $\mathbb{A}_{\mathrm{SOA}}((x_1,y_1),\ldots,(x_{t-1},y_{t-1}))(x_t) \neq y_t$, it must be that $\mathsf{LD}(\mathbb{H}_{\{(x_1,y_1),\ldots,(x_t,y_t)\}}) \leq \mathsf{LD}(\mathbb{H}_{\{(x_1,y_1),\ldots,(x_{t-1},y_{t-1})\}}) - 1$. Thus, there can be at most $\mathsf{LD}(\mathbb{H})$ times t at which this occurs: that is, $\mathbb{A}_{\mathrm{SOA}}$ makes at most $\mathsf{LD}(\mathbb{H})$ mistakes on any realizable sequence.

To prove the lower bound, we consider the set $\{x_{\mathbf{y}} : \mathbf{y} \in \bigcup_{0 \leq i \leq d-1} \{0,1\}^i\} \subseteq \mathcal{X}$ from the definition of $\mathsf{LD}(\mathbb{H})$. We construct a data sequence via the probabilistic method. Choose $\mathbf{y} = (y_1, \ldots, y_d) \in \{0,1\}^d$ uniformly at random, and define the data sequence as $X_1 = x_{()}$, $X_2 = x_{(y_1)}$, $X_3 = x_{(y_1,y_2)}$, …, $X_d = x_{(y_1,\ldots,y_{d-1})}$, and for each $i \in \{1,\ldots,d\}$ let $Y_i = y_i$. In particular, note that each Y_t is independent of $(X_1,Y_1),\ldots,(X_{t-1},Y_{t-1}),X_t$. Thus, for any learning algorithm $\mathbb{A}$, we have $\Pr(\mathbb{A}((X_1,Y_1),\ldots,(X_{t-1},Y_{t-1}))(X_t) \neq Y_t) \geq \frac{1}{2}$ (it would be equal $1/2$ if the algorithm were restricted to outputting 0 or 1, not $\star$). Therefore

$$\mathbb{E}\left[\sum_{t=1}^{d} \mathbb{1}[\mathbb{A}((X_1,Y_1),\ldots,(X_{t-1},Y_{t-1}))(X_t) \neq Y_t]\right] \geq \frac{d}{2}$$

by linearity of the expectation. In particular, by the law of total expectation, this implies there exists a deterministic sequence $(X_1,Y_1),\ldots,(X_d,Y_d)$ with this property. ■

E.3. Online Learning in the Agnostic Case

Similarly to the PAC model, the online learning model also has an *agnostic* variant (Ben-David, Pál, and Shalev-Shwartz, 2009), which makes *no* assumptions about the data sequence, and rather than bounding the number of mistakes, it bounds the *excess* of the number of mistakes made by the algorithm compared to the number made by the best hypothesis in the class $\mathbb{H}$: known as the *regret*. Specifically, we have the following definition.

Definition 48 (Agnostic Online Learnability) *A partial concept class $\mathbb{H}$ is* agnostically online learnable *if there exists a sequence $\mathsf{Reg}(\mathbb{H},T) = o(T)$ such that, for every $T \in \mathbb{N}$, there exists*

a (possibly randomized) learning algorithm $\mathbb{A}$ *that, for every sequence* $(X_1, Y_1), \ldots, (X_T, Y_T) \in \mathcal{X} \times \{0, 1\}$,

$$\mathbb{E} \sum_{t=1}^{T} \mathbb{1}[\mathbb{A}((X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1}))(X_t) \neq Y_t] \leq \mathsf{Reg}(\mathbb{H}, T) + \min_{h \in \mathbb{H}} \sum_{t=1}^{T} \mathbb{1}[h(X_t) \neq Y_t].$$

For the special case of $\mathbb{H}$ a *total* concept class, Ben-David, Pál, and Shalev-Shwartz (2009) proved that $\mathbb{H}$ is agnostically online learnable if and only if $\mathsf{LD}(\mathbb{H}) < \infty$. Here we extend this result to partial concept classes. In particular, this supplies the second part of the claim in Theorem 15 from Section 2.5, so that proving this result will complete the proof of Theorem 15 as well.

Theorem 49 (Agnostic Online Learnability) *The following statements are equivalent for a partial concept class* $\mathbb{H} \subseteq \{0, 1, \star\}^{\mathcal{X}}$.

- $\mathsf{LD}(\mathbb{H}) < \infty$.
- $\mathbb{H}$ *is agnostically online learnable.*

E.4. Proof of Theorem 49 and Quantitative Regret Bounds for the Agnostic Case

As above, Theorem 49 will follow as an immediate implication of the following more detailed result, which also supplies a concrete bound on the form of the optimal regret $\mathsf{Reg}(\mathbb{H}, T)$.

Theorem 50 (Optimal Regret Bound) *For any partial concept class* $\mathbb{H}$ *with* $\mathsf{LD}(\mathbb{H}) > 0$, *there exists an online learning algorithm achieving an expected regret guarantee*

$$\mathsf{Reg}(\mathbb{H}, T) = O\left(\sqrt{\mathsf{LD}(\mathbb{H}) T \ln(T/\mathsf{LD}(\mathbb{H}))}\right).$$

Moreover, for any finite $d \leq \mathsf{LD}(\mathbb{H})$, *any* $T \in \mathbb{N}$ *such that* $T \geq d$, *and any learning algorithm* $\mathbb{A}$, *there exists a data sequence of length* T *on which* $\mathbb{A}$ *has expected regret* $\Omega\left(\sqrt{dT}\right)$.

The proof makes use of a classic result for learning from expert advice (Vovk, 1990, 1992; Littlestone and Warmuth, 1994; Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, and Warmuth, 1997; Kivinen and Warmuth, 1999; Singer and Feder, 1999); see Theorem 2.2 of Cesa-Bianchi and Lugosi (2006). To simplify notation, we write $x_{1:t} := (x_1, \ldots, x_t)$.

Lemma 51 (Experts; Cesa-Bianchi and Lugosi, 2006, Theorem 2.2) *For any* $N, T \in \mathbb{N}$ *and* $f_1, \ldots, f_N$ *functions* $\mathcal{X}^* \to [0, 1]$, *letting* $\eta = \sqrt{(8/T)\ln(N)}$, *for any* $(x_1, y_1), \ldots, (x_T, y_T) \in \mathcal{X} \times [0, 1]$, *letting* $w_{0,i} = 1$ *and* $w_{t,i} = e^{-\eta \sum_{s \leq t} |f_i(x_{1:s}) - y_s|}$ *for each* $t \leq T$, $i \leq N$, *letting* $\bar{f}_t(x_{1:t}, y_{1:(t-1)}) = \sum_i w_{t-1,i} f_i(x_{1:t}) / \sum_{i'} w_{t-1,i'}$, *it holds that*

$$\sum_{t=1}^{T} \left| \bar{f}_t(x_{1:t}, y_{1:(t-1)}) - y_t \right| - \min_{1 \leq i \leq N} \sum_{t=1}^{T} |f_i(x_{1:t}) - y_t| \leq \sqrt{(T/2)\ln(N)}.$$

We are now ready for the proof of Theorem 50.

Proof of Theorem 50 As was the case of Theorem 47, the proof of this result is essentially based on existing proofs for the special case of total concept classes, though in this case a few important changes are required. We include the full details for completeness.

The upper bound is based on the work of Ben-David, Pál, and Shalev-Shwartz (2009). Consider any data sequence $(X_1, Y_1), \ldots, (X_T, Y_T)$, and let $h^* = \operatorname{argmin}_{h \in \mathbb{H}} \sum_{t=1}^{T} \mathbb{1}[h(X_t) \neq Y_t]$ (breaking ties arbitrarily). Let $t_1, \ldots, t_q$ denote the subsequence of $1, \ldots, T$ such that $h^*(X_{t_i}) \neq \star$. In particular, $(X_{t_1}, h^*(X_{t_1})), \ldots, (X_{t_q}, h^*(X_{t_q}))$ is realizable w.r.t. $\mathbb{H}$. Let $\mathbb{A}_{\mathrm{SOA}}$ be as in the proof of Theorem 47, and recall from that proof that $\mathbb{A}_{\mathrm{SOA}}$ makes at most $\mathsf{LD}(\mathbb{H})$ mistakes on any realizable data sequence. Now construct a subsequence J^* of $\{t_1, \ldots, t_q\}$ as follows. Initialize $J^* = ()$. For each $i \in \{1, \ldots, q\}$ in increasing order, if $\mathbb{A}_{\mathrm{SOA}}(\{(X_t, h^*(X_t))\}_{t \in J^*})(X_{t_i}) \neq h^*(X_{t_i})$ then append t_i to the sequence J^* and continue. In particular, note that the sequence $\{(X_t, h^*(X_t))\}_{t \in J^*}$ is realizable w.r.t. $\mathbb{H}$, so that $\mathbb{A}_{\mathrm{SOA}}$ makes at most $\mathsf{LD}(\mathbb{H})$ mistakes. On the other hand, enumerating $J^* = \{j_1, \ldots, j_{|J^*|}\}$, for each $k \leq |J^*|$ we have $\mathbb{A}_{\mathrm{SOA}}(\{(X_{j_i}, h^*(X_{j_i}))\}_{i<k})(X_{j_k}) \neq h^*(X_{j_k})$. Thus, it must be that $|J^*| \leq \mathsf{LD}(\mathbb{H})$. Moreover, for each $t \in \{t_1, \ldots, t_q\}$ with $t \notin J^*$, we have $\mathbb{A}_{\mathrm{SOA}}(\{(X_{j_i}, h^*(X_{j_i}))\}_{j_i<t})(X_t) = h^*(X_t)$.

Now, for each subsequence J of $\{1, \ldots, T\}$ with $|J| \leq \mathsf{LD}(\mathbb{H})$, for each $j \in J$, inductively define $\hat{Y}_j^J = 1 - \mathbb{A}_{\mathrm{SOA}}(\{(X_{j'}, \hat{Y}_{j'}^J)\}_{j' \in J : j' < j})(X_j)$. Then define an algorithm $\mathbb{A}^J$ that, for each $t \in \{1, \ldots, T\}$ with $t \notin J$, has $\mathbb{A}^J(X_1, \ldots, X_t) = \mathbb{A}_{\mathrm{SOA}}(\{(X_j, \hat{Y}_j^J)\}_{j \in J : j < t})(X_t)$, and for each $t \in J$, $\mathbb{A}^J(X_1, \ldots, X_t) = 1 - \mathbb{A}_{\mathrm{SOA}}(\{(X_j, \hat{Y}_j^J)\}_{j \in J : j < t})(X_t)$. In particular, note that for all $i \in \{1, \ldots, q\}$, $\mathbb{A}^{J^*}(X_1, \ldots, X_{t_i}) = h^*(X_{t_i})$. Therefore, since $h^*(X_t) \neq Y_t$ at every $t \notin \{t_1, \ldots, t_q\}$, we have $\sum_{t=1}^{T} \mathbb{1}\left[\mathbb{A}^{J^*}(X_1, \ldots, X_t) \neq Y_t\right] \leq \sum_{t=1}^{T} \mathbb{1}[h^*(X_t) \neq Y_t]$.

Now to describe the online learning algorithm achieving the regret guarantee from the theorem, we will apply the classic exponential weights algorithm with these $\mathbb{A}^J$ predictors as the "experts". Specifically, applying Lemma 51 with the above value of T, with $N = \sum_{i=0}^{\mathsf{LD}(\mathbb{H})} \binom{T}{i}$, and with functions $f_1, \ldots, f_N$ given by an enumeration of the algorithms $\mathbb{A}^J$, $J \subseteq \{1, \ldots, T\}$, $|J| \leq \mathsf{LD}(\mathbb{H})$, we conclude that, for $\bar{f}_t$ as defined in Lemma 51,

$$\sum_{t=1}^{T} \left| \bar{f}_t(X_{1:t}, Y_{1:(t-1)}) - Y_t \right| - \min_{1 \leq i \leq N} \sum_{t=1}^{T} |f_i(X_{1:t}) - Y_t| \leq \sqrt{(T/2)\ln(N)}.$$

Let us then define a randomized predictor $\mathbb{A}$ such that, for any $t \in \{1, \ldots, T\}$, $\mathbb{A}((X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1}))(X_t)$ evaluates to 1 with probability $\bar{f}_t(X_{1:t}, Y_{1:(t-1)})$, and other-

wise evaluates to 0 (where this random evaluation occurs independently for each t). We then have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[\mathbb{A}((X_1,Y_1),\dots,(X_{t-1},Y_{t-1}))(X_t) \neq Y_t]\right] = \sum_{t=1}^{T} \left|\bar{f}_t(X_{1:t}, Y_{1:(t-1)}) - Y_t\right| \\
&\leq \min_{1\leq i\leq N} \sum_{t=1}^{T} |f_i(X_{1:t}) - Y_t| + \sqrt{(T/2)\ln(N)} \\
&\leq \sum_{t=1}^{T} \mathbb{1}\left[\mathbb{A}^{J^*}(X_1,\dots,X_t) \neq Y_t\right] + \sqrt{(T/2)\ln\left(\sum_{i\leq \mathsf{LD}(\mathbb{H})} \binom{T}{i}\right)} \\
&\leq \sum_{t=1}^{T} \mathbb{1}[h^*(X_t) \neq Y_t] + O\left(\sqrt{\mathsf{LD}(\mathbb{H}) T \ln\left(\frac{T}{\mathsf{LD}(\mathbb{H})}\right)}\right).
\end{aligned}$$

This completes the proof of the upper bound.

The lower bound proof is essentially identical to the existing proof for total concepts from Ben-David, Pál, and Shalev-Shwartz (2009), but we include the details for completeness. Given $0 < d \leq \mathsf{LD}(\mathbb{H})$ and $T \geq d$, let $k = \lfloor T/d \rfloor$. Define the sequence $Y_1,\dots,Y_T$ as independent $\mathrm{Bernoulli}(1/2)$ random variables. Consider the set $\{x_{\mathbf{y}} : \mathbf{y} \in \bigcup_{0\leq i\leq d-1}\{0,1\}^i\}$ from the definition of $\mathsf{LD}(\mathbb{H})$ (Definition 27). Let $T_0 = 0$, and for each $i \in \{1,\dots,d-1\}$, let $T_i = ki$, and let $T_d = T$. Then, for each $i \in \{1,\dots,d\}$, let $y_i = \mathrm{Majority}(Y_{T_{i-1}+1},\dots,Y_{T_i})$ and for each $t \in \{T_{i-1}+1,\dots,T_i\}$, define $X_t = x_{\{y_{i'}\}_{i'<i}}$.

Since the Y_t values are independent, for any learning algorithm we certainly have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[\mathbb{A}((X_1,Y_1),\dots,(X_{t-1},Y_{t-1}))(X_t) \neq Y_t]\right] \geq \frac{T}{2}$$

(with equality if $\mathbb{A}$ outputs 0 or 1). It remains only to upper bound the value $\mathbb{E}\left[\min_{h\in\mathbb{H}} \sum_{t=1}^{T} \mathbb{1}[h(X_t)\neq Y_t]\right]$. In particular, consider a partial concept $\bar{h}\in\mathbb{H}$ with $\bar{h}(x_{\{y_{i'}\}_{i'<i}}) = y_i$ for each $i \in \{1,\dots,d\}$, which exists by definition of $x_{\mathbf{y}}$ from Definition 27. Then, for each $i \in \{1,\dots,d\}$,

$$\sum_{t=T_{i-1}+1}^{T_i} \left(2\mathbb{1}\left[\bar{h}(X_t) = Y_t\right] - 1\right) = \sum_{t=T_{i-1}+1}^{T_i} (2\mathbb{1}[Y_t = y_i] - 1) = \left|\sum_{t=T_{i-1}+1}^{T_i} (2Y_t - 1)\right|,$$

and Khinchine's inequality (see Lemma A.9 of Cesa-Bianchi and Lugosi, 2006) implies

$$\mathbb{E}\left[\left|\sum_{t=T_{i-1}+1}^{T_i} (2Y_t - 1)\right|\right] \geq \sqrt{(T_i - T_{i-1})/2}.$$

Thus, since

$$\mathbb{E}\left[\sum_{t=T_{i-1}+1}^{T_i} \left(2\mathbb{1}\left[\bar{h}(X_t) = Y_t\right] - 1\right)\right] = (T_i - T_{i-1}) - 2\,\mathbb{E}\left[\sum_{t=T_{i-1}+1}^{T_i} \mathbb{1}\left[\bar{h}(X_t) \neq Y_t\right]\right],$$

we conclude that

$$\mathbb{E}\left[\sum_{t=T_{i-1}+1}^{T_i} \mathbb{1}\left[\bar{h}(X_t) \neq Y_t\right]\right] \leq \frac{T_i - T_{i-1}}{2} - \sqrt{\frac{T_i - T_{i-1}}{8}}.$$

Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\bar{h}(X_t) \neq Y_t\right]\right] \leq \sum_{i=1}^{d} \frac{T_i - T_{i-1}}{2} - \sqrt{\frac{T_i - T_{i-1}}{8}} \leq \frac{T}{2} - \sqrt{\frac{d^2 k}{8}} \leq \frac{T}{2} - \frac{1}{4}\sqrt{dT}.$$

Altogether,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[\mathbb{A}((X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1}))(X_t) \neq Y_t] - \min_{h \in \mathbb{H}} \sum_{t=1}^{T} \mathbb{1}[h(X_t) \neq Y_t]\right] \geq (1/4)\sqrt{dT}.$$

In particular, by the law of total expectation, this also implies there exists a ($\mathbb{A}$-dependent) deterministic choice of the sequence $(X_1, Y_1), \ldots, (X_T, Y_T)$ satisfying this. ■

We note that the upper bound for total concept classes has been refined by Alon, Ben-Eliezer, Dagan, Moran, Naor, and Yogev (2021) to match the lower bound up to numerical constants: that is $\mathsf{Reg}(\mathbb{H}, T) = \Theta\left(\sqrt{\mathsf{LD}(\mathbb{H})T}\right)$. However, that proof uses techniques for which it is unclear whether they can be extended to partial concept classes. Thus, there remains an open question:

Open Question 8 *Is the optimal regret for partial concept classes always* $\Theta\left(\sqrt{\mathsf{LD}(\mathbb{H})T}\right)$*?*

References

Apple tries to peek at user habits without violating privacy. *The Wall Street Journal*, 2016a.

Apple promises to deliver AI smarts without sacrificing your privacy. *The Verge*, 2016b.

Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite Littlestone dimension. In *Proceedings of the 51st Annual ACM Symposium on the Theory of Computing*, STOC '19, New York, NY, USA, 2019. ACM.

Noga Alon, Alon Gonen, Elad Hazan, and Shay Moran. Boosting simple learners. *CoRR*, abs/2001.11704, 2020. URL https://arxiv.org/abs/2001.11704.

Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. Adversarial laws of large numbers and optimal regret in online classification. In *Proceedings of the 53rd Annual ACM Symposium on Theory of Computing*, 2021.

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

Hassan Ashtiani, Shai Ben-David, Nicholas J. A. Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *J. ACM*, 67(6):32:1–32:42, 2020. doi: 10.1145/3417994. URL https://doi.org/10.1145/3417994.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In Aurélien Garivier and Satyen Kale, editors, *Algorithmic Learning Theory, ALT 2019, 22-24 March 2019, Chicago, Illinois, USA*, volume 98 of *Proceedings of Machine Learning Research*, pages 162–183. PMLR, 2019. URL http://proceedings.mlr.press/v98/attias19a.html.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. *arXiv:1810.02180v4*, 2021.

Kaspars Balodis, Shalev Ben-David, Mika Göös, Siddhartha Jain, and Robin Kothari. Unambiguous DNFs and Alon-Saks-Seymour. *CoRR*, abs/2102.08348, 2021.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6240–6249, 2017. URL http://papers.nips.cc/paper/7204-spectrally-normalized-margin-bounds-for-neural-networks.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1903070116. URL https://www.pnas.org/content/116/32/15849.

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.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

Shalev Ben-David, Pooya Hatami, and Avishay Tal. Low-sensitivity functions from unambiguous certificates. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, volume 67 of *LIPIcs*, pages 28:1–28:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017. doi: 10.4230/LIPIcs.ITCS.2017.28. URL https://doi.org/10.4230/LIPIcs.ITCS.2017.28.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.

Nicolas Bousquet, Aurélie Lagoutte, and Stéphan Thomassé. Clique versus independent set. *Eur. J. Comb.*, 40:73–92, 2014. doi: 10.1016/j.ejc.2014.02.003. URL https://doi.org/10.1016/j.ejc.2014.02.003.

Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? *CoRR*, abs/2012.06421, 2020. URL https://arxiv.org/abs/2012.06421.

Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *61th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020.*, 2020.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the Association for Computing Machinery*, 44(3):427–485, 1997.

K. Chaudhuri and S. Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems 27*, 2014.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Aref N. Dajani, Amy D. Lauger, Phyllis E. Singer, Daniel Kifer, Jerome P. Reiter, Ashwin Machanava-jjhala, Simson L. Garfinkel, Scot A. Dahl, Matthew Graham, Vishesh Karwa, Hang Kim, Philip Lelerc, Ian M. Schmutte, William N. Sexton, Lars Vilhuber, and John M. Abowd. The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau. Presented at the September 2017 meeting of the Census Scientific Advisory Committee.

O. David, S. Moran, and A. Yehudayoff. Supervised learning through the lens of compression. In *Advances in Neural Information Processing Systems 29*, 2016.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In Gal Elidan, Kristian Kersting, and Alexander T. Ihler, editors, *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017. URL http://auai.org/uai2017/proceedings/papers/173.pdf.

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. In search of robust measures of generalization. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/86d7c8a08b4aaa1bc7c599473f5dddda-Abstract.html.

Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, and Daniel M. Roy. On the role of data in PAC-Bayes bounds. *CoRR*, abs/2006.10929, 2020b. URL https://arxiv.org/abs/2006.10929.

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 1054–1067, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329576. doi: 10.1145/2660267.2660348. URL https://doi.org/10.1145/2660267.2660348.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 954–959. ACM, 2020. doi: 10.1145/3357713.3384290. URL https://doi.org/10.1145/3357713.3384290.

Badih Ghazi, Noah Golowich, Ravi Kumar, and Pasin Manurangsi. Sample-efficient proper PAC learning with approximate differential privacy. *CoRR*, abs/2012.03893, 2020. URL https://arxiv.org/abs/2012.03893.

Alon Gonen, Elad Hazan, and Shay Moran. Private learning implies online learning: An efficient reduction. *NeurIPS*, 2019.

Mika Göös. Lower bounds for clique vs. independent set. In Venkatesan Guruswami, editor, *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 1066–1076. IEEE Computer Society, 2015. doi: 10.1109/FOCS.2015.69. URL https://doi.org/10.1109/FOCS.2015.69.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. *IEEE Trans. Inf. Theory*, 64(6):4120–4128, 2018. doi: 10.1109/TIT.2018.2822267. URL https://doi.org/10.1109/TIT.2018.2822267.

T. Graepel, R. Herbrich, and J. Shawe-Taylor. PAC-Bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

S. Hanneke. The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016.

S. Hanneke, A. Kontorovich, and M. Sadigurschi. Sample compression for real-valued learners. In *Proceedings of the* 30^{th} *International Conference on Algorithmic Learning Theory*, 2019.

D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

Ralf Herbrich and Robert C. Williamson. Algorithmic luckiness. *J. Mach. Learn. Res.*, 3:175–212, 2002. URL http://jmlr.org/papers/v3/herbrich02a.html.

J. Kivinen and M. K. Warmuth. Averaging expert predictions. In *Proceedings of the* 4^{th} *European Conference on Computational Learning Theory*, 1999.

A. Kontorovich and I. Pinelis. Exact lower bounds for the agnostic probably-approximately-correct (PAC) machine learning model. *The Annals of Statistics*, 47(5):2822–2854, 2019.

Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Nearest-neighbor sample compression: Efficiency, consistency, infinite dimensions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1573–1583, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/934815ad542a4a7c5e8a2dfa04fea9f5-Abstract.html.

N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

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

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. 1986b.

Hartmut Maennel, Ibrahim M. Alabdulmohsin, Ilya O. Tolstikhin, Robert J. N. Baldock, Olivier Bousquet, Sylvain Gelly, and Daniel Keysers. What do neural networks learn when trained with random labels? In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e4191d610537305de1d294adb121b513-Abstract.html.

A. Maurer and M. Pontil. Empirical Bernstein bounds and sample-variance penalization. In *Proceedings of the* 22^{nd} *Conference on Learning Theory*, 2009.

Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *J. ACM*, 63(3): 21:1–21:10, 2016. doi: 10.1145/2890490. URL https://doi.org/10.1145/2890490.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11611–11622, 2019. URL http://papers.nips.cc/paper/9336-uniform-convergence-may-be-unable-to-explain-generalization-in-deep-lear

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *8th International Conference*

on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1g5sA4twr.

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

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5947–5956, 2017. URL http://papers.nips.cc/paper/7176-exploring-generalization-in-deep-learning.

Huy Le Nguyen, Jonathan R. Ullman, and Lydia Zakynthinou. Efficient private algorithms for learning large-margin halfspaces. In Aryeh Kontorovich and Gergely Neu, editors, *Algorithmic Learning Theory, ALT 2020, 8-11 February 2020, San Diego, CA, USA*, volume 117 of *Proceedings of Machine Learning Research*, pages 704–724. PMLR, 2020. URL http://proceedings.mlr.press/v117/nguy-en20a.html.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. ISSN 0033-295X. doi: 10.1037/h0042519. URL http://dx.doi.org/10.1037/h0042519.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (A)*, 13(1):145–147, 1972.

R. E. Schapire and Y. Freund. *Boosting*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2012.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inf. Theory*, 44(5):1926–1940, 1998. doi: 10.1109/18.705570. URL https://doi.org/10.1109/18.705570.

A. Singer and M. Feder. Universal linear prediction by model order weighting. *IEEE Transactions on Signal Processing*, 47(10):2685–2699, 1999.

S. Szarek. Metric entropy of homogeneous spaces. *Quantum Probability, Banach Center Publications*, 43:395–410, 1998.

M. Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22: 28–76, 1994.

Thomas Unterthiner, Daniel Keysers, Sylvain Gelly, Olivier Bousquet, and Ilya O. Tolstikhin. Predicting neural network accuracy from weights. *CoRR*, abs/2002.11448, 2020. URL https://arxiv.org/abs/2002.11448.

R. Urner and S. Ben-David. Probabilistic Lipschitzness a niceness assumption for deterministic labels. In *Learning Faster from Easy Data Workshop*, 2013.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

R. van Handel. The universal Glivenko-Cantelli property. *Probability and Related Fields*, 155: 911–934, 2013.

V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Proc. USSR Acad. Sci.*, 181(4):781–783, 1968.

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

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974a.

V. Vapnik and A. Chervonenkis. On the method of ordered risk minimization. i. *Avtomatika i Telemekhanika*, (8):21—-30, 1974b.

Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

V. Vovk. Aggregating strategies. In *Proceedings of the* 3^{rd} *Annual Workshop on Computational Learning Theory*, 1990.

V. Vovk. Universal forecasting algorithms. *Information and Computation*, 96(2):245–277, 1992.

Manfred K. Warmuth. Compressing to VC dimension many points. In *COLT*, volume 2777 of *Lecture Notes in Computer Science*, pages 743–744. Springer, 2003.

Y. Wiener, S. Hanneke, and R. El-Yaniv. A compression technique for analyzing disagreement-based active learning. *Journal of Machine Learning Research*, 16(4):713–745, 2015.

Avi Wigderson. *Mathematics and Computation*. Princeton University Press, 2019. doi: doi:10.1515/9780691192543. URL https://doi.org/10.1515/9780691192543.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.