A Theory of PAC Learnability of Partial Concept Classes

Noga Alon
Princeton University, NJ, USA
nalon@math.princeton.edu

Steve Hanneke
Purdue University
West Lafayette, IN, USA
steve.hanneke@gmail.com

Ron Holzman
Technion
Haifa, Israel
holzman@technion.ac.il

Shay Moran
Technion, Haifa, Israel
Google Research, Tel-Aviv
smoran@technion.ac.il

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, 1974b; Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989; Hodges, 1993).

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 Goos (2015), Ben-David, Hatami, and Tal (2017); Balodis, Ben-David, Góös, 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.

I. 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 $\delta > 0$, 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 IV-B). 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

1In fact, in Section III-A, 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.

Index Terms—PAC Learning; Learnability; VC Dimension; Margin; Online Learning; Empirical Risk Minimization;
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.

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

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 interest in this work is the VC dimension of a partial concept class $\tilde{H}$, denoted by $\text{VC}(\tilde{H})$, which is defined as the maximum size of a shattered set $U \subseteq X$, where $U$ is shattered if every binary pattern $u \in \{0, 1\}^U$ is realized by some $h \in \tilde{H}$ (i.e., $\tilde{h}[U] = u$). 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 hinges on the recent breakthroughs concerning the clique vs. $\theta$-hypothesis. However, the above theorem implies that if one tries to extend each partial concept in $\tilde{H}$ by a total concept by disambiguating the $\ast$’s in it, then one must end up with a class whose VC dimension is unbounded, and hence is not PAC learnable.

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

As we will discuss below, the above partial class $\tilde{H}$ 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 $\tilde{H}$, and the sample complexity is $O(\log(1/\delta)/\epsilon)$, where $\epsilon, \delta$ are the standard accuracy and confidence parameters. However, the above theorem implies that if one tries to extend each partial concept in $\tilde{H}$ by a total concept by disambiguating the $\ast$’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 II-D3 below.

B. 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 \( H \). This gives rise to the following definition: a distribution \( P \) on \( \mathcal{X} \times \{0,1\} \) is realizable by \( H \) 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 \( h \in \mathcal{H} \): that is, \( \{x_i\}_{i=1}^n \subseteq \text{supp}(h) \), 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: 

\[
\varepsilon_P(h) = P(\{x,y \mid h(x) \neq y\}).
\]

To be clear, this means we always count the case \( h(x) = \ast \) as a prediction mistake.

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

\[
\varepsilon_P(\hat{h}(S)) \leq \varepsilon.
\]

The value \( \mathcal{M}(\varepsilon, \delta) \) is called the sample complexity of \( \mathcal{H} \), and the optimal sample complexity is the minimum achievable value of \( \mathcal{M}(\varepsilon, \delta) \) for every given \( \varepsilon, \delta \).

In Section C.2 we define learnability in the agnostic case in a similar manner, using the convention that \( \ast \)'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 \( \mathcal{H} \) is PAC learnable if and only if its VC dimension is finite (Vapnik and Chervonenkis 1974b; 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 \( h \in \mathcal{H} \) which minimizes the empirical error learns \( \mathcal{H} \). 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 \( \mathcal{H} \subseteq \{0,1,\ast\}^X \):

- \( \mathcal{H} \) is PAC learnable.
- \( \mathcal{H} \) 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 \( \mathcal{H} \) with \( \text{VC}(\mathcal{H}) = 1 \) 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.

C. 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 (i.e., \( \mathcal{H} \) is learnable by any ERM algorithm). Moreover, any ERM algorithm achieves the optimal sample complexity, up to lower order factors (Vapnik and Chervonenkis 1974b; 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 \( \varepsilon_P \)), 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; Nakiran 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.

a) 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 $\text{VC}(\mathbb{H}) = 0$ 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 $\mathbb{H} \subseteq \{0, 1, \ast\}^n$ defined by

$$\mathbb{H} = \{ h_A : A \subseteq [n], |A| = n/2 \},$$

where, $h_A(x) = \begin{cases} 0 & x \in A, \\ \ast & x \in [n] \setminus A. \end{cases}$ 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 $\text{supp}(c)$; 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 $\text{supp}(c)$ 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 $\{0, \ast\}$-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 $t$.

b) 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 $\mathbb{H}' \subseteq \{0, 1\}^X$. 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:

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 $\text{VC}(\mathbb{H}) = 1$ such that, for any total concept class $\mathbb{H}'$, there exists an ERM algorithm for $\mathbb{H}$ 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 $\text{VC}(\mathbb{H}) = 1$ such that any learning algorithm $\mathbb{H}$ 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.

## D. 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 II-D1) and differentially private learning (Section II-D2). Then, in Section II-D3 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 $\mathbb{H}'$ which disambiguates $\mathbb{H}$? We provide negative as well as positive results in this context.

1) **Sample Compression Schemes:** Sample compression is a fundamental technique for proving generalization bounds. Littlestone and Warmuth (1986) 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 Nisnevich, 2018; Hanneke, Kontorovich, and Sadgurschi, 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 $\kappa$ and a reconstructor $\rho$ (see Figure 1): the compressor gets the input sample $S$, from which she picks
This question has been studied since the pioneering work by Littlestone and Warmuth [1986], 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 Concepts).**

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

2) There exists a partial concept class $\mathbb{H}$ with $\text{VC}(\mathbb{H}) = 1$ 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 4. 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) **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 even been implemented in practice, notably by Google [Erling, Pihur, and Korolova 2014], Apple [McMillan 2016; Vincent 2016], and in the 2020 US census [Dajani, Lauger, Singer, Kifer, Reiter, Machanavajjhala, Garfinkel, Dahl, Graham, Karwa, Kim, Lelerc, Schmutte, Sexton, Vilhuber, and Abowd 2017].

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 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:

**Question 8 (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 Zakythiou 2020)), it is not clear how to generally prove either of the implications “LD(H) < \infty \implies H is DP-learnable” or “H is DP-learnable \implies \text{LD}(H) < \infty”.

The known proofs of the direction “LD(H) < \infty \implies H is DP-learnable” for total concept classes (Bun, Livni, and Moran 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 “LD(H) = \infty \implies 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 H, denoted by TD(H), is the maximum integer d for which there exist x_1, \ldots, x_d \in X' and h_1, \ldots, h_d \in H such that h_i(x_j) = 1[j \leq j]. For total concept classes H it is known that TD(H) \geq \log(\text{LD}(H)); 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 9.** There exists a partial concept class H with TD(H) \leq 2 but LD(H) = \infty.

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? Relatively, 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 10 (Disambiguation).** A total concept class \(\overline{\text{H}}\) is a special type of partial concept class such that every \(h \in H\) has range \(\{0, 1\}\); i.e., is a total concept. A total concept class \(\overline{\text{H}} \subseteq \{0, 1\}^X\) is said to disambiguate a partial concept class \(H\) if every finite data sequence \(S \in (X \times \{0, 1\})^\ast\) realizable w.r.t. \(\overline{\text{H}}\) is also realizable w.r.t. \(H\). In this case, \(\overline{\text{H}}\) is called a disambiguation of \(H\).

Note the difference between Definition 10 and the definition used in Theorem 1; the latter poses a stricter requirement, namely, that each partial concept in \(\overline{\text{H}}\) is extended by some total concept in \(H\). We note that the two definitions are equivalent when \(X\) is finite (more generally, when \(\text{supp}(h)\) is finite for every \(h \in \overline{\text{H}}\)), and are essentially equivalent when \(X\) is countable.

Definition 10 is more suitable in the context of learning because it suffices to guarantee that every PAC learner for \(\overline{\text{H}}\) is a PAC learner for \(H\), and hence reduces the task of PAC learning the partial class \(H\) to PAC learning the total class \(\overline{\text{H}}\). One could further relax Definition 10 by allowing errors and by only requiring to disambiguate short samples, in a way that implies that a learner for \(\overline{\text{H}}\) is a weak learner for \(H\). However, the next proposition implies that such relaxations are essentially equivalent to Definition 10.

**Proposition 11 (Approximate Disambiguation \implies Disambiguation).** Let \(\overline{\text{H}}\) be a partial class and let \(\gamma > 0\). Assume that there exists a total class \(\overline{\text{H}}\) with \(VC(\overline{\text{H}}) = d\) that “weakly disambiguates” \(H\) in the following sense: for every sample \(S = \{(x_i, y_i)\}_{i=1}^n\) realizable by \(H\) of size \(|S| = n\leq O(D/\delta)\) there exists \(h \in \overline{\text{H}}\) such that

\[
\epsilon_S(h) := \frac{1}{n} \sum_{i=1}^n |h(x_i) \neq y_i| \leq 1 - \frac{\gamma}{2}.
\]

Then, \(\overline{\text{H}}\) can be disambiguated (in the sense of Definition 10) by a total class whose VC-dimension is at most \(O(D/\delta^2)\), where \(d^* \leq 2d^+ + 1\) is the dual VC dimension of \(\overline{\text{H}}\).

Proposition 11 might be viewed as a kind of compactness theorem; for example, it implies that in order to disambiguate \(\overline{\text{H}}\) it suffices to represent only the samples realizable by \(\overline{\text{H}}\) of size at most \(100d\) by a total class \(\overline{\text{H}}\) with \(VC(\overline{\text{H}}) = d\). The proofs of all the results in this subsection appear in Section D.

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

**Theorem 12 (A VC Class Which Cannot Be Disambiguated).** For any \(n \in \mathbb{N}\) there exists a partial concept class \(H_n \subseteq \{0, 1, \ast\}^n\) with \(VC(H_n) = 1\) and \(TD(H_n) \leq 2\) such that any disambiguation \(\overline{\text{H}}_n\) of \(H_n\) 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 \(LD(\overline{\text{H}}) \geq VC(\overline{\text{H}}) \geq (\log(n))^{1-o(1)}\), and shows that for infinite \(X\) there exists \(\overline{\text{H}}_\infty \subseteq \{0, 1, \ast\}^X\) with \(VC(\overline{\text{H}}_\infty) = 1\) and \(TD(\overline{\text{H}}_\infty) \leq 2\), while \(LD(\overline{\text{H}}) = VC(\overline{\text{H}}) = \infty\) for every disambiguation \(\overline{\text{H}}_\infty\).

Below, in Theorem 13 we show that the bound in Theorem 12 is nearly tight. Theorem 12 resolves, in a strong negative sense, an open problem presented by Attias, Kondorovich, and Mansour (2019), which sought a disambiguation whose VC dimension is bounded by a (linear) function of \(VC(H)\). Further, Theorem 12 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

\footnote{It is also known that \(LD(\overline{\text{H}}) \geq |\log(TD(\overline{\text{H}}))|\), but this inequality extends also to partial classes with the same proof (see Alon, Livni, Malliaris, and Moran 2019).}
(Theorem 6), sample compression schemes (Theorem 7), and private PAC learning (Theorem 9). In particular, note that Theorem 12 immediately implies Theorem 11.

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 12 is short and simple, it unfortunately provides only little insight on the structure of the concluded class $H$. 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 III-A.

### a) 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 12. 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 $H \subseteq \{0,1\}^n$ with $\text{VC}(H) = d$ by $|H| \leq \binom{n}{d}$. Is there an analogue of the SSP Lemma for partial classes? An immediate and direct generalization of it to partial classes would be that $|H| \leq \binom{|\text{VC}(H)|}{d} \leq 2^n$. However it is easy to see that this is false, as witnessed e.g. by the class $H = \{0,1\}^n$ which satisfies $\text{VC}(H) = 0$ and $|H| = 2^n$. A more natural candidate for extending the SSP Lemma to partial classes is via disambiguations:

**Question 15** (Polynomial Growth $\implies$ Disambiguation?). Let $H \subseteq \{0,1,*\}^X$ and assume there exists a polynomial poly such that for every finite $X' \subseteq X$ there exists a disambiguation $H'$ of $H|_{X'}$ of size $|H'| \leq \text{poly}(n)$, where $n = |X'|$. Does there exist a disambiguation $H'$ of $H$ such that $\text{VC}(H') < \infty$?

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

#### b) Disambiguating by Sample-Compression.

Sample compression schemes naturally imply disambiguations: indeed, consider a partial concept class $H \subseteq \{0,1,*\}^X$ over a finite domain of size $|X| = n$, and assume we are given a sample compression scheme for $H$ of size $k$. Therefore, for any partial concept $h \in H$ there exist $x_1, \ldots, x_k \in \text{supp}(h)$ such that $h$ is extended by the total concept $\tilde{h} = \rho\left(\left\{x_i, h(x_i)\right\}_{i=1}^k, B\right)$. where $\rho$ 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 $\rho$ on all such sequences of length at most $k$ and all such bit-strings $B$, we obtain a disambiguation of $H$ of size $n^{O(k)}$. This is summarized in the following proposition:

**Proposition 16.** For any finite $X$ and any partial concept class $H$, if $H$ has a compression scheme of size $k$, there exists a disambiguation $H'$ of $H$ of size at most $(c|X|/k)^k$ for a numerical constant $c$.

**c) Disambiguating by Majority-Votes.** We conclude this section with highlighting one idea which is used in the proofs of Theorems 13 and 14. Let $H \subseteq \{0,1,*\}^n$ be a partial class. Consider an online learning setting in which an adversary picks a target partial concept $h \in H$, and then in each round...
at most one 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 $O(VC(\mathbb{H}) \log(n))$ 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 III for more details.

E. Online Learning

We conclude Section III with a characterization of online learnability. The following theorem shows that the Littlestone dimension retains its role of characterizing online learnability. See Section II-A 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 17.** The following statements are equivalent for a partial concept class $\mathbb{H} \subseteq \{0, 1, \ast\}^Y$.

- $LD(\mathbb{H}) < \infty$.
- $\mathbb{H}$ 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 II-D2. However, our understanding of partial Littlestone classes is more limited. In particular, we conclude with the following basic question:

**Question 18.** Let $\mathbb{H}$ be a partial class with $LD(\mathbb{H}) < \infty$. Does there exist a disambiguation of $\mathbb{H}$ by a total class $\bar{\mathbb{H}}$ such that $LD(\bar{\mathbb{H}}) < \infty$? Is there one with $VC(\bar{\mathbb{H}}) < \infty$?

We remark that if the answer to Open Question 15 is affirmative, then so is the answer to the second part (about VC dimension) of Open Question 18. This follows because Littlestone classes can be disambiguated using the SOA algorithm (see Appendix E).

III. 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.

A. 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 19** (Separability with Margin). A sample $(x_1, y_1), \ldots, (x_n, y_n) \in V \times \{0, 1\}$ is $(R, \gamma)$-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 $2 \gamma$.

In other words, a sample is $(R, \gamma)$-separable, if the 0-labelled examples and 1-labelled examples can be separated by a linear classifier with margin $\gamma$ and all examples lie in a ball of radius $R$. Let $\mathbb{H}_{R, \gamma}$ denote the class $\mathbb{H}_{R, \gamma} := \{h \in \{0, 1, \ast\}^V : (\forall x_1, \ldots, x_n \in supp(h)) : (x_1, h(x_1)), \ldots, (x_n, h(x_n)) \text{ is } (R, \gamma)\text{-separable}\}$.

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 $R, \gamma$ 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 20.** For all $\gamma, R > 0$: $VC(\mathbb{H}_{R, \gamma}) = \Theta\left(\frac{R^2}{\gamma^2}\right)$ and $LD(\mathbb{H}_{R, \gamma}) = \Theta\left(\frac{R^2}{\gamma^2}\right)$.

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?

**Question 21.** Does there exist a disambiguation of $\mathbb{H}_{R, \gamma}$ by a total class $\bar{\mathbb{H}} \subseteq \{0, 1\}^Y$ whose VC/Littlestone dimensions are bounded by a function of $R, \gamma$?

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 15 and 18.

B. 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 \(B\). 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 \(\gamma\)-weak learnability assumption asserts that if \(c : \mathcal{X} \rightarrow \{0, 1\}\) 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 \(\gamma\)-correlated with \(c\). One can rephrase the weak learnability assumption only in terms of \(B\) as follows:

**Definition 22** (\(\gamma\)-realizable samples [Alon, Gonen, Hazan, and Moran (2020)]). Let \(B \subseteq \{0, 1\}^\mathcal{X}\) be the base-class and let \(\gamma \in (0, 1]\). A sample \(S = \{(x_1, y_1), \ldots, (x_n, y_n)\}\) is \(\gamma\)-realizable with respect to \(B\) if for any probability distribution \(\Pr\) over \(S\) there exists \(b \in B\) such that \(\Pr_{(x,y) \sim D} b(x) \neq y \leq \frac{1-\gamma^2}{\gamma^2}\).

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

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

\[
\mathbb{H}_\gamma = \{ h \in \{0,1, \ast\}^\mathcal{X} : \forall x_1, \ldots, x_n \in \operatorname{supp}(h), (x_1, h(x_1)), \ldots, (x_n, h(x_n)) \text{ is } \gamma\text{-realizable by } B \}.
\]

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 \(\gamma\)-VC dimension). They provided the following upper bounds:

**Theorem 23** [Alon, Gonen, Hazan, and Moran (2020)]. Let \(B\) be a class with VC dimension \(d\), and let \(\gamma > 0\). Then, the following upper bounds on \(\text{VC}(\mathbb{H}_\gamma)\) hold:

\[
\text{VC}(\mathbb{H}_\gamma) = O\left(\frac{d}{\gamma^2} \log(d/\gamma)\right) = \tilde{O}\left(\frac{d}{\gamma^2}\right),
\]

and

\[
\text{VC}(\mathbb{H}_\gamma) = O_d\left(\frac{1}{\gamma}\right),
\]

where \(O_d(\cdot)\) conceals a multiplicative constant that depends only on \(d\).
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. 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 III-A. Furthermore, this framework leads to a natural approach for proving data-dependent learning guarantees: i.e., bounds on $\operatorname{er}_P(\hat{h}_n)$ 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 [1974b]; Vapnik [1998]).

Specifically, consider any sequence $\mathbb{H}_1, \mathbb{H}_2, \ldots$ of partial concept classes, and for each $i$ let $\mathbb{H}_i$ be a learning algorithm designed for the class $\mathbb{H}_i$. For any data sequence $S = \{(x_i, y_i)\}_{i=1}^n$ in $X \times \{0, 1\}$, define $\operatorname{er}_S(\mathbb{H}_i) := \min_{\hat{h} \in \mathbb{H}_i} \frac{1}{n} \sum_{i}[\hat{h}(x_i) \neq y_i]$. First we describe a realizable version of SRM. For each $i$, suppose there is a bound $B_i(n, \delta, \gamma)$ such that, for any $P$, for $S \sim P^n$, with probability at least $1 - \delta$, if $\operatorname{er}_S(\mathbb{H}_i) = 0$, then $\operatorname{er}_P(\mathbb{H}_i(S)) \leq B_i(n, \delta, \gamma)$. Then we can easily produce a method with a corresponding data-dependent error bound: choose $i$ of minimal $B_i(n, \delta/i(i+1))$ subject to $\operatorname{er}_S(\mathbb{H}_i) = 0$ (if it exists), and output $\hat{h}_0 = \mathbb{H}_i(S)$. The corresponding guarantee is that, with probability at least $1 - \delta$, if $i$ exists,

$$\operatorname{er}_P(\hat{h}_0) \leq B_i(n, \delta/i(i+1)).$$

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 $\mathbb{H}_i$ required above, specifically with

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

For instance, for the margin example in Section III-A since $\text{VC}(\mathbb{H}_{R,\gamma}) = \Theta\left(\frac{\gamma^2}{\pi^2}\right)$ (from Proposition 20), 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 $\mathbb{H}_{R,\gamma}$ for an appropriate sequence of $(R_i, \gamma_i)$: 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 $\bar{R}$ and $S$ is linearly separable with margin $\gamma$, then we may choose the class $\mathbb{H}_{[\bar{R},1/\sqrt{\gamma}]}$ to recover the bound

$$\operatorname{er}_P(\hat{h}) = O\left(\frac{\bar{R}^2}{n} \log^2(n) + \frac{1}{n} \log\left(\frac{1}{\gamma}\right)\right)$$

from (Shawe-Taylor et al. [1998]).

We can similarly derive a bound that does not require $\operatorname{er}_S(\mathbb{H}_i) = 0$, 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

7 Namely, given two competing classifiers, prioritize the one for which the data-dependent bound is better.
at least $1 - \delta$, $\text{er}_P(A_i(S)) \leq \text{er}_S(H_i) + B_i(n, \delta)$. Then let us choose $i$ to minimize $\text{er}_S(H_i) + B_i(n, \delta/i(i + 1))$, and output $\hat{h} = A_i(S)$. As above, by the union bound, we have that with probability at least $1 - \delta$,

$$
\text{er}_P(\hat{h}) \leq \text{er}_S(H_i) + B_i(n, \delta/i(i + 1)).
$$

Again, in Section 2.2 (Lemma 43), we propose a general algorithm $A_i$ that can provide such guarantees with

$$
B_i(n, \delta) = O\left(\frac{\text{VC}(H_i)}{n} \log^2(n) + \frac{1}{n} \log \left(\frac{1}{\delta}\right)\right).
$$

b) 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 $H$ be a total concept class, and let $m$ denote any input-sample size. A luckiness function is a mapping $L: \mathcal{X}^m \times H \to \mathbb{R}^+$ which, given an input sample $S = \{(x_i, y_i)\}_{i \leq m}$, assigns to each hypothesis $h \in H$ 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 $h \in H$, 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 $H_1 \subseteq H_2 \subseteq \cdots$ as follows. For $r > 0$, we say that a partial concept $h: \mathcal{X} \to \{0, 1, \ast\}$ is $r$-lucky if there exists a total concept $h' \in H$ such that $h(x) = h'(x)$ for all $x \in \text{supp}(h)$ (i.e., $h'$ extends $h$), and $L(x; h') \geq r$ for every $x \in \text{supp}(h)$. Let $\mathbb{H}_r$ denote the class of all $r$-lucky partial concepts. Note that $\mathbb{H}_r \subseteq \mathbb{R}^+$ is a hierarchy of partial concept classes (i.e., $\mathbb{H}_r \supseteq \mathbb{H}_{r'}$ 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 $\{h \in H : L(x_1, \ldots, x_m; h) \geq r\}$ if and only if $S$ is realizable w.r.t. the partial concept class $\mathbb{H}_r$. Thus, the data-independent hierarchy $\mathbb{H}_r$ 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 $\mathbb{H}_r$.

Thus, we can recover the types of data-dependent error bounds provided by Shawe-Taylor, Bartlett, Williamson, and Anthony [1998] using the above SRM technique with partial concept classes $\mathbb{H}_i = \mathbb{H}_r$, for a suitable discretization $r_1 \geq r_2 \geq \cdots$ (e.g., chosen so that $\text{VC}(\mathbb{H}_i) = i$). 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.

C. 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 of functions $\mathcal{X} \to \{0, 1, \ast\}$. The only distinction is in the definition of PAC learning, where the multiclass setting would allow distributions $P$ on $\mathcal{X} \times \{0, 1, \ast\}$, whereas the setting of partial concepts restricts to distributions supported on $\mathcal{X} \times \{0, 1\}$. That is, a distribution $P$ on $\mathcal{X} \times \{0, 1, \ast\}$ is realizable w.r.t. $H$ in the 3-label multiclass setting if $\inf_{h \in H} P(\{(x, y) : h(x) \neq y\}) = 0$; the definition of PAC learnability is otherwise the same as Definition 2.

As it turns out, any partial concept class $H$ 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. For example of this is the class $H$ of all functions $N \to \{0, 1, \ast\}$ whose image is $\{0, \ast\}$. 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 26.** Let $H \subseteq \{0, 1, \ast\}^X$. The following are equivalent: (1) $H$ is PAC learnable in the 3-label multiclass setting. (2) $H$ is PAC learnable in the partial concepts setting and $\text{VC}(\{\text{supp}(h) : h \in H\}) < \infty$.

The comparison to 3-label multiclass classification yields some further interesting observations. For instance, one can show that Proposition 26 also implies that, when $\text{VC}(\{\text{supp}(h) : h \in H\}) < \infty$, the ERM principle does hold for learning in the partial concepts setting\cite{Shawe-Taylor, Bartlett, Williamson, and Anthony 1998}. This contrasts with the discussion above where we found that ERM learners can fail spectacularly for some partial concept classes with $\text{VC}(H) < \infty$ (cf Proposition 4). Moreover, this connection to

\footnote{Technically, the assumption in Shawe-Taylor, Bartlett, Williamson, and Anthony [1998] bounds the effective VC dimension of $H_r$: i.e., the VC dimension w.r.t. typical samples; but also all other results discussed here apply under this assumption.}

\footnote{This follows from the fact that, with $\text{VC}(H) < \infty$ and $\text{VC}(\{\text{supp}(h) : h \in H\}) < \infty$, we have a Sauer-Shelah-Perles type bound on the total number of $\{0, 1, \ast\}$ patterns possible on any data set, from which uniform convergence guarantees follow for the losses.}
multiclass classification has a further implication for disambiguation. Specifically, we have the following result.

**Proposition 27.** Any partial concept class $\mathcal{H}$ can be strongly disambiguated to a total concept class $\mathcal{H}$ with $\text{VC}(\mathcal{H}) = O(\text{VC}(\mathcal{H}) + \text{VC}(\{\text{supp}(h) : h \in \mathcal{H}\}))$.

In particular, this means that if $\mathcal{H}$ is a PAC learnable partial concept class, and $\text{VC}(\{\text{supp}(h) : h \in \mathcal{H}\}) < \infty$, 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 $\mathcal{H}$ that cannot be disambiguated to learnable total concept classes (see Theorems 1 and 2).

**ACKNOWLEDGMENT**

Noga Alon’s research is supported in part by NSF grant DMS-1855464, BSF grant 2018267 and the Simons Foundation. This work was done while Ron Holzman visited Princeton University. Shay Moran is a Robert J. Shillman Fellow and his research is supported by the ISF, grant no. 1225/20, by an Azrieli Faculty Fellowship, and by BSF grant 2018385.

**REFERENCES**

Alon, N., Livni, R., Malliaris, M., and Moran, S. (2019). 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. ACM.

Alon, N., Gonen, A., Hazan, E., and Moran, S. (2020). Boosting simple learners. *CoRR, abs/2001.11704*.

Ashtiani, H., Ben-David, S., Harvey, N. J. A., Liaw, C., Mehrabian, A., and Plan, Y. (2020). Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *J. ACM, 67*(6), 32:1–32:42.

Attias, I., Kontorovich, A., and Mansour, Y. (2019). Improved generalization bounds for robust learning. In A. Garivier and S. 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.

Attias, I., Kontorovich, A., and Mansour, Y. (2021). Improved generalization bounds for robust learning. *arXiv:1810.02180v4*.

Balodis, K., Ben-David, S., Göös, M., Jain, S., and Kothari, R. (2021). Unambiguous DNFs and Alon-Saks-Seymour. *CoRR, abs/2102.08348*.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. (2017). Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. 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.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences, 116*(32), 15849–15854.

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

Ben-David, S., Hatami, P., and Tal, A. (2017). Low-sensitivity functions from unambiguous certificates. In C. 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.

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

Brown, G., Bun, M., Feldman, V., Smith, A., and Talwar, K. (2020). When is memorization of irrelevant training data necessary for high-accuracy learning? *CoRR, abs/2012.06421*.

Bun, M., Livni, R., and Moran, S. (2020). An equivalence between private classification and online prediction. In *61th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020*.

Daji, A. N., Lauger, A. D., Singer, P. E., Kifer, D., Reiter, J. P., Machanava-jhala, A., Garfinkel, S. L., Dahl, S. A., Graham, M., Karwa, V., Kim, H., Lelerc, P., Schmutte, I. M., Sexton, W. N., Vilhuber, L., and Abowd, J. M. (2017). The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau. Presented at the September 2017 meeting of the Census Scientific Advisory Committee.

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

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

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer.

Dziugaite, G. K. and Roy, D. M. (2017). Computing non-vacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In G. Elidan, K. Kersting, and A. T. Ihler, editors, *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press.

Dziugaite, G. K., Drouin, A., Neal, B., Rajkumar, N., Caballero, E., Wang, L., Mitliagkas, I., and Roy, D. M. (2020). In search of robust measures of generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and A. Blum, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020*, virtual.
Dziugaite, G. K., Hsu, K., Gharbieh, W., and Roy, D. M. (2020b). On the role of data in PAC-Bayes bounds. CoRR, abs/2006.10929.

Erlingsson, U., Pihur, V., and Korolova, A. (2014). 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. Association for Computing Machinery.

Feldman, V. (2020). Does learning require memorization? a short talk about a long tail. In K. Makarychev, Y. Makarychev, M. Tulsiani, G. Kamath, and J. Chuzhoy, editors, Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020, pages 954–959. ACM.

Ghazi, B., Golowich, N., Kumar, R., and Manurangsi, P. (2020). Near-optimal sample compression for nearest neighbors. In V. 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.

Gottlieb, L., Kontorovich, A., and Nisnevitch, P. (2018). Lower bounds for online learning: An efficient reduction. NeurIPS.

Göös, M. (2015). Lower bounds for clique vs. independent set. In In V. 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.

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

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

Herbrich, R., and Williamson, R. C. (2002). Algorithmic luckiness. J. Mach. Learn. Res., 3, 175–212.

Hodges, W. (1993). Model theory, volume 42 of Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge.

Kontorovich, A., Sabato, S., and Weiss, R. (2017). Nearest-neighbor sample compression: Efficiency, consistency, infinite dimensions. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. 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.

Littlestone, N. and Warmuth, M. (1986). Relating data compression and learnability.

Long, P. (2001). On agnostic learning with \{0, *, 1\}-valued and real-valued hypotheses. In Proceedings of the 14th Annual Conference on Learning Theory and 5th European Conference on Computational Learning Theory.

Maennel, H., Alabdulmohsin, I. M., Tolstikhin, I. O., Baldock, R. J. N., Bousquet, O., Gelly, S., and Keyser, D. (2020). What do neural networks learn when trained with random labels? In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

McMillan, R. (2016). Apple tries to peek at user habits without violating privacy. The Wall Street Journal.

Moran, S. and Yehudayoff, A. (2016). Sample compression schemes for VC classes. J. ACM, 63(3), 21:1–21:10.

Nagarajan, V. and Kolter, J. Z. (2019). Uniform convergence may be unable to explain generalization in deep learning. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox, and R. 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.

Nakkiran, P., Kaplin, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2020). Deep double descent: Where bigger models and more data hurt. In 9th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. (2017). Exploring generalization in deep learning. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. 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.

Nguyen, H. L., Ullman, J. R., and Zakythinou, L. (2020). Efficient private algorithms for learning large-margin halfspaces. In A. Kontorovich and G. 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.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408.

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

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

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

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., and Anthony, M. (1998). Structural risk minimization over data-dependent hierarchies. IEEE Trans. Inf. Theory, 44(5), 1926–1940.
Unterthiner, T., Keysers, D., Gelly, S., Bousquet, O., and Tolstikhin, I. O. (2020). Predicting neural network accuracy from weights. *CoRR, abs/2002.11448*.

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

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

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

Vapnik, V. and Chervonenkis, A. (1974b). *Theory of Pattern Recognition*. Nauka, Moscow.

Vincent, J. (2016). Apple promises to deliver AI smarts without sacrificing your privacy. *The Verge*.

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

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

Wigderson, A. (2019). *Mathematics and Computation*. Princeton University Press.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). 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.