2015 IEEE 56th Annual Symposium on Foundations of Computer Science

Compressing and teaching for low VC-dimension

Shay Moran*, Amir Shpilka†, Avi Wigderson‡, and Amir Yehudayoff§
*Department of Computer Science, Technion-IIT, Israel and
Max Planck Institute for Informatics, Saarbrücken, Germany
Email: shaymrn@cs.technion.ac.il
†Department of Computer Science, Tel Aviv University, Israel
Email: shpilka@post.tau.ac.il
‡School of Mathematics, Institute for Advanced Study, Princeton NJ
Email: avi@ias.edu
§Department of Mathematics, Technion-IIT, Israel
Email: amir.yehudayoff@gmail.com

Abstract

In this work we study the quantitative relation between VC-dimension and two other basic parameters related to learning and teaching. Namely, the quality of sample compression schemes and of teaching sets for classes of low VC-dimension. Let C be a binary concept class of size m and VC-dimension d. Prior to this work, the best known upper bounds for both parameters were log(m), while the best lower bounds are linear in d. We present significantly better upper bounds on both as follows.

We construct sample compression schemes of size exp(d) for C. This resolves a question of Littlestone and Warmuth (1986). Roughly speaking, we show that given an arbitrary set of labeled examples from an unknown concept in C, one can retain only a subset of exp(d) of them, in a way that allows to recover the labels of all other examples in the set, using additional exp(d) information bits.

We further show that there always exists a concept c in C with a teaching set (i.e. a list of c-labeled examples uniquely identifying c in C) of size exp(d) loglog(m). This problem was studied by Kuhlmann (1999). Our construction also implies that the recursive teaching (RT) dimension of C is at most exp(d) loglog(m) as well. The RT-dimension was suggested by Zilles et al. and Doliwa et al. (2010). The same notion (under the name partial-ID width) was independently studied by Wigderson and Yehudayoff (2013). An upper bound on this parameter that depends only on d is known just for the very simple case d=1, and is open even for d=2. We also make small progress towards this seemingly modest goal.

Keywords

PAC learning; VC dimension; sample compression schemes; recursive teaching dimension

I. Introduction

The study of mathematical foundations of learning and teaching has been very fruitful, revealing fundamental connections to various other areas of computer science and mathematics, such as computational complexity, geometry, topology, combinatorics, and probability theory. Many key ideas and notions emerged from this study: Sample complexity, Vapnik and Chervonenkis's VC-dimension [1], Valiant's definition of PAC learning [2], Littlestone and Warmuth's sample compression schemes [3], Goldman and Kearns's teaching dimension [4], recursive teaching (RT) dimension [5]–[7] and more.

VC-dimension. The VC-dimension is one of the most basic complexity measures for binary-labeled concept classes (i.e. families of boolean functions). A binary-labeled concept class over the domain[1] X is a set $C \subseteq \{0,1\}^X$. The VC-dimension of C, denoted $\mathrm{VC}(C)$, is the maximum size of a shattered subset of X, where a set $Y \subseteq X$ is shattered if for every $Z \subseteq Y$ there is $c \in C$ so that $c(x) = 1$ for all $x \in Z$ and $c(x) = 0$ for all $x \in Y - Z$. In other words, viewing a concept class as a $C \times X$ binary matrix, its VC dimension is the largest number d of columns in which each of the 2^d possible binary patterns occurs. Many natural families of mathematical objects, as well as real world data sets, have low VC dimension.

It is perhaps mostly known in the context of the PAC learning model. PAC learning was introduced in Valiant's seminal work [2] as a theoretical model for learning from random examples drawn from an unknown distribution (see the book [8] for more details). This model is deeply related to Vapnik and Chervonenkis's foundational work on uniform convergence in probability theory [1]. A well-known result of Blumer, Eherenfeucht, Haussler, and Warmuth [9], which is based on [1], states that PAC learning sample complexity is equivalent to VC-dimension.

This text combines two separate papers. One with the same authors titled "Teaching and compressing for low VC-dimension" (ECCC TR15-025) and one with the first and last authors titled "Sample compression schemes for VC classes" (ECCC TR15-040).

[1]In order to eliminate measurability and similar issues, we focus on the case where the set X is a countable set. However, the arguments presented here are more general.

0272-5428/15 $31.00 © 2015 IEEE
DOI 10.1109/FOCS.2015.12

This work. While it is known that some of the complexity measures mentioned above are tightly linked, the exact relationship between them is still not well understood. In particular, it is a long standing question whether the VC-dimension can be used to give a universal bound on the size of sample compression schemes or on the RT-dimension. That is, whether these two parameters can be bounded in terms of the VC-dimension alone, regardless of the concept class. In this work, we make significant progress on both questions.

We start with a brief summary of our contributions, and by stating the two main results of this work. A reader that is not familiar with the definitions is directed to Sections II and III below. Section II contains formal definitions, background, motivation and theorems concerning sample compression schemes. Section III contains a similar discussion concerning teaching sets.

Sample compression schemes were defined by Littlestone and Warmuth [3] as a natural abstraction that captures a common property of many learning procedures, like procedures for learning geometric shapes or algebraic structures (see also [10], [11]). Roughly speaking, a sample compression scheme takes a long list of samples and compresses it to a short sub-list of samples in a way that allows to invert the compression. They showed that sample compression schemes of small size yield PAC learning algorithms with small sample complexity. They additionally asked whether the other direction holds. Our first result answers their question affirmatively. Thus, for binary-labeled concept classes, compression characterizes PAC learning sample complexity; the existence of compression schemes of finite size is equivalent to finite PAC learning sample complexity.

Theorem 1.1 (Compression): If $C \subseteq \{0,1\}^X$ has VC dimension d, then C has a sample compression scheme of size[2] $2^{O(d)}$.

The key property of this compression is that its size does not depend on the size of the given sample. Our construction of sample compression schemes is overall quite short and simple. It is inspired by Freund's work [12] where majority is used to boost the accuracy of learning procedures. It also uses several known properties of PAC learnability and VC-dimension, together with von Neumann's minimax theorem, and it reveals approximate but efficient equilibrium strategies for zero-sum games of low VC-dimension. The running time of the construction is polynomial in $|C|^{2^{O(d)}}, |X|^d$.

Teaching sets were defined in several works [4], [13], [14] as a tool for studying teaching theory (they were also studied in other contexts, e.g. [15], [16]). A teaching set for a concept c in $C \subseteq \{0,1\}^X$ is a subset of the domain X that uniquely identifies c in C. A natural complexity measure for teaching that was defined in [5]–[7] is the recursive teaching (RT) dimension of C. Roughly speaking, this measure is t if there is a linear order on C according to which every concept c in C can be distinguished from later ones by specifying its values in some t coordinates. Doliwa et al. [6] and Zilles et al. [5] asked whether small VC-dimension implies small RT-dimension. An equivalent question was asked ten years earlier by Kuhlmann [17]. Our second contribution is a construction of small teaching sets. While our bound depends on $|C|$, it provides an exponential improvement over the previous $\log|C|$ bound.

Theorem 1.2 (RT-dimension): If $C \subseteq \{0,1\}^X$ has VC dimension d, then there is $c \in C$ with a teaching set of size at most $O(d2^d \log\log|C|)$. Hence, the RT-dimension of C is at most $O(d2^d \log\log|C|)$.

The time it takes to construct a teaching set of this size is polynomial in $|C|, |X|$.

Comment. While we state the bound above in terms of the number of concepts $|C|$, it can be stated in terms of the domain size $|X|$, which may be considered a more natural input size parameter. The well known Sauer-Perles-Shelah lemma (see e.g. [18]) asserts that $|C| = O(|X|^d)$, and so the bound above can be replaces by $O(d2^d \log(d) + d2^d \log\log|X|)$.

II. Sample compression schemes

Learning and compression are known to be deeply related to each other. Learning procedures perform compression, and compression is an evidence of and is useful in learning. There are many examples of this tight link. For example, support vector machines, a fundamental algorithm commonly applied to solve classification problems, performs compression by saving only the "support vectors" seen so far and discarding the rest (see Chapter 6 in [19]). Another example is the use of compression to boost the accuracy of learning procedures (see [3], [12] and Chapter 4 in [20]).

The reasons for this tight connection are both practical and philosophical. When presented with a huge amount of data about some phenomena, say, in the form of labeled examples, a learner would like to store as little of the data as possible in manner that allows the rest to be inferred (possibly using a small amount of side information). A practical motivation is that storage is expensive. A different motivation relates to that the ability to compress indicates the possibility of generalization, which we associate with learning, and often unveils the structure underlying the concepts learned. Occam's razor, which is

[2]In this text $O(f)$ means at most $\alpha f + \beta$ for $\alpha, \beta > 0$ constants.

a philosophical principle attributed to William of Ockham from the late middle ages, also fits this connection. It says that in the quest for an explanation or an hypothesis, one should prefer the simplest one which is consistent with the data. There are many works on the role of Occam's razor in learning theory (a partial list includes [3], [10], [11], [21]–[24]).

Before giving the formal definition of compression schemes, let us consider a simple illustrative example. Assume we are interested in learning the concept class of intervals on the real line. We get a collection of 100 samples of the form $(x, c_I(x))$ where $x \in \mathbb{R}$ and $c_I(x) \in \{0, 1\}$ indicates[3] if x is in the interval $I \subset \mathbb{R}$. Can we remember just a few of the samples in a way that allows to recover the labels of all other samples? In this case, the answer is affirmative and in fact it is easy to do so. Just remember two locations, those of the left most 1 and of the right most 1 (if there are no 1s, just remember one of the 0s). From this data, we can reconstruct the value of c_I on all the other 100 samples.

The above example brings several questions to mind. What is special about the family of intervals that allows this type of compression? Is it related to the geometry or topology of the underlying universe? How can we abstractly define the above procedure?

Definition (sample compression schemes). We start with a high level description. A sample compression scheme comprises of two maps: a compression and a reconstruction. The compression maps a labeled sample of arbitrary size to a sub-sample of bounded size, plus a bounded amount of additional side information. The reconstruction recovers from a labeled sample of bounded size a full hypothesis h that must agree with the labels of *every* labeled sample that is compressed to it.

More formally, a C-labeled sample is a pair (Y, y), where $Y \subseteq X$ and $y = c|_Y$ for some $c \in C$. The size of a labeled sample (Y, y) is $|Y|$. For an integer k, denote by $L_C(k)$ the set of C-labeled samples of size at most k. Denote by $L_C(\infty)$ the set of all C-labeled samples of finite size. A k-sample compression scheme for C with information I, where I is a finite set, consists of two maps κ, ρ for which the following hold:

(κ) The *compression map*

$$\kappa : L_C(\infty) \to L_C(k) \times I$$

takes (Y, y) to $((Z, z), i)$ with $Z \subseteq Y$ and $y|_Z = z$.

(ρ) The *reconstruction map*

$$\rho : L_C(k) \times I \to \Sigma^X$$

is so that for all (Y, y) in $L_C(\infty)$,

$$\rho(\kappa(Y, y))|_Y = y.$$

The size of the scheme is $k + \log(|I|)$, and its kernel size is k. In the language of coding theory, the side information I can be thought of as list decoding; the map ρ has a short list of possible reconstructions of a given (Z, z), and the information $i \in I$ indicates which element in the list is the correct one. The following property must always hold: if the compression of $(Y, c|_Y)$ is the same as that of $(Y', c'|_{Y'})$ then $c|_{Y \cap Y'} = c'|_{Y \cap Y'}$. It is not necessarily the case that the reconstructed hypothesis belongs to the original class C. All it has to satisfy is that for any $(Y, y) \in L_C(\infty)$ such that $h = \rho(\kappa(Y, y))$ we have that $h|_Y = y$. In other words, h has to be consistent only on the sampled coordinates that were compressed and not elsewhere. See [10], [11] for more discussions of this definition, and some insightful examples.

Let us consider a simple example of a sample compression scheme, to help digest the definition. Let C be a concept class and let r be the rank over, say, $\mathbb{R}$ of the matrix whose rows correspond to the concepts in C. We claim that there is an r-sample compression scheme for C with no side information. Indeed, for any $Y \subseteq X$, let Z_Y be a set of at most r columns that span the columns of the matrix $C|_Y$. Given a sample (Y, y) compress it to $\kappa(Y, y) = (Z_Y, z)$ for $z = y|_{Z_Y}$. The reconstruction maps ρ takes (Z, z) to any concept $h \in C$ so that $h|_Z = z$. This sample compression scheme works since if $(Z, z) = \kappa(Y, y)$ then every two different rows in $C|_Y$ must disagree on Z.

Background and motivation. Sample compression schemes are known to yield practical learning algorithms (see e.g. [25]), and allow learning for multi-labeled concept classes [26]. Interestingly, every compression scheme also yields a natural and general learning procedure: Given a labeled sample (Y, y), the learner compresses it to $\kappa(Y, y)$ and outputs the hypothesis $h = \rho(\kappa(Y, y))$. Littlestone and Warmuth's [3] provided a straightforward proof that this procedure is indeed a PAC learner.

Theorem 2.1 (Compression implies learnability [3]): Let $C \subseteq \Sigma^X$ and $c \in C$. Let μ be a distribution on X, and $x_1, \ldots, x_m$ be m independent samples from μ. Let $Y = (x_1, \ldots, x_m)$ and $y = c|_Y$. Let κ, ρ be a k-sample compression scheme for C with additional information I. Let $h = \rho(\kappa(Y, y))$. Then, for every $\epsilon > 0$,

$$\Pr_{\mu^m}\left[\mu(\{x \in X : h(x) \neq c(x)\}) > \epsilon\right] < |I| \sum_{j=0}^{k} \binom{m}{j} (1 - \epsilon)^{m-j}.$$

[3]That is $c_I(x) = 1$ iff $x \in I$.

In particular, C can be PAC learned with $O\Big(\big(k\log(1/\epsilon)+\log(1/\delta)+\log(|I|)\big)/\epsilon\Big)$ samples, generalization error ϵ and success probability $1-\delta$.

Proof sketch: Consider the sample space X^m with the probability function μ^m. We define several events in this space. First, the event $E \subseteq X^m$ contains all $\bar{x} \in X^m$ for which the function $h_{\bar{x}} = \kappa(\rho(\bar{x}, c|_{\bar{x}}))$ is ϵ-far from c according to μ. Second, for every $T \subseteq [m]$ of size $|T| \leq k$ and for every information $i \in I$, define the event $E_{T,i}$ as follows. Let $x_T = (x_t : t \in T)$. The event $E_{T,i}$ is the set of all $\bar{x}$ for which the function $h_{T,i,x_T} = \rho((T, c|_{x_T}), i)$ is ϵ-far from c according to μ but h_{T,i,x_T} agrees with c on $\bar{x}$.

We now claim that $E \subseteq \bigcup_{T,i} E_{T,i}$. Indeed, let $\bar{x} \in E$. Let $((Z,z),i) = \kappa(\bar{x}, c|_{\bar{x}})$, and let $S = \{s \in [m] : x_s \in Z\}$. Thus, $\bar{x} \in E_{S,i}$.

The theorem follows using the union bound, since $\Pr[E_{T,i}] \leq (1-\epsilon)^{m-|T|}$ for every T, i. Indeed, for fixed T, i, the function h_{T,i,x_T} depends only on x_T and is independent of $x_{[m]-T}$. So, if h_{T,i,x_T} is ϵ-far from c, then the probability that it agrees with c on $x_{[m]-T}$ is less than $(1-\epsilon)^{m-|T|}$. ∎

The definition of sample compression schemes naturally generalizes to multi-labeled concept classes. Moreover, the above proof remains valid and therefore compression implies learnability for general concept classes [26]. Littlestone and Warmuth's question can thus be seen as the binary instance of a much broader question: Is it true that the size of an optimal sample compression scheme for a given concept class is the sample complexity of PAC learning it?

Since the sample complexity of PAC learning is essentially the VC-dimension, a lower bound on the size of sample compression schemes in terms of VC-dimension should hold. Indeed, [11] proved that there are concept classes of VC-dimension d for which any sample compression scheme has size at least d.

Further motivation for considering compression schemes comes from the problem of boosting a weak learner to a strong learner. Boosting is a central theme in learning theory that was initiated by Kearns and Valiant [27], [28]. The boosting question, roughly speaking, is: given a learning algorithm with generalization error 0.49, can we use it to get an algorithm with generalization error ϵ of our choice? Theorem 2.1 implies that if the learning algorithm yields a sample compression scheme, then boosting follows with roughly a multiplicative overhead of $1/\epsilon$ in the sample size. Therefore, constructions of efficient compression schemes immediately imply boosting.

Schapire [29] and later on Freund [12] solved the boosting problem, and showed how to efficiently boost the generalization error of PAC learners. They showed that if C is PAC learnable with d samples and generalization error 0.49, then C is PAC learnable with $O(d\log^2(d/\epsilon)/\epsilon)$ samples and generalization error ϵ (see e.g. Corollary 3.3 in [12]). Interestingly, their boosting is based on a weak type of compression. They showed how to compress a sample of size m to a sample of size $O(d\log m)$, and that such compression already implies boosting.

Previous constructions. Littlestone and Warmuth's question and variants of it lead to a rich body of work that revealed profound properties of VC-dimension and learning. These works also discovered and utilized connections between sample compression schemes, and model theory, topology, combinatorics, and geometry. Floyd and Warmuth [10], [11] constructed sample compression schemes of size $\log|C|$ for every concept class C. They also constructed optimal compression schemes of size d for maximum classes[4] of VC-dimension d, as a first step towards solving the general question. As the study of sample compression schemes deepened, many insightful and optimal schemes for special cases have been constructed: Floyd [10], Helmbold et al. [30], Floyd and Warmuth [11], Ben-David and Litman [31], Chernikov and Simon [32], Kuzmin and Warmuth [33], Rubinstein et al. [34], Rubinstein and Rubinstein [35], Livni and Simon [36] and more. Finally, in our work [37], where the teaching sets we present here are constructed, we also constructed sample compression schemes of size roughly $\exp(d)\log\log|C|$ for classes C of VC-dimension d.

Our contribution. Theorem 1.1 says that every concept class with VC-dimension d has a sample compression scheme of size $\exp(d)$. The compression is even more efficient when the dual class is also under control. The dual concept class $C^* \subseteq \{0,1\}^C$ of C is defined as the set of all functions $f_x : C \to \{0,1\}$ defined by $f_x(c) = c(x)$. If we think of C as a binary matrix whose rows are concepts in C and columns are elements of X, then C^* corresponds to the distinct rows of the transposed matrix.

Theorem 2.2 (Compression using data on dual): If $C \subseteq \{0,1\}^X$ has VC dimension d and C^* has VC dimension d^*, then C has a sample compression scheme of size $O(d^* \cdot d \cdot \log(d^* \cdot d))$.

Theorem 1.1 follows from Theorem 2.2 via the following bound, which was proved by Assouad [38].

Claim 2.3 (Dual VC dimension [38]): If $\mathrm{VC}(C) \leq d$, then $\mathrm{VC}(C^*) < 2^{d+1}$.

[4]That is, C satisfies Sauer-Shelah-Perles lemma with equality; $|C| = \sum_{k=0}^{d} \binom{|X|}{k}$.

A specific and natural example for which the dual class is well behaved is geometrically defined classes. Assume, for example, that C represents the incidence relation among halfspaces and points in r-dimensional real space (a.k.a. sign rank or Dudely dimension r). That is, for every $c \in C$ there is a vector $a_c \in \mathbb{R}^r$ and for every $x \in X$ there is a vector $b_x \in \mathbb{R}^r$ so that $c(x) = 1$ if and only if the inner product $\langle a_c, b_x \rangle = \sum_{j=1}^{r} a_c(j) b_x(j)$ is positive. It follows that $\mathrm{VC}(C) \leq r$, but the symmetric structure also implies that $\mathrm{VC}(C^*) \leq r$. So overall the compression scheme constructed here for this C actually has size $O(r^2 \log r)$ and not $2^{O(r)}$.

Proof background and overview. Freund [12] and later on Freund and Schapire [39] showed that, roughly speaking, for every class C that is PAC learnable with d samples via a learning map H, there exists a compression scheme that compresses a sample (Y, y) of size m to a sub-sample of size $O(d \log m)$. Their constructive proof is iterative. In each iteration t, a distribution μ_t on Y is carefully and adaptively chosen. Then, d points from Y are drawn according to μ_t, and fed into H to produce an hypothesis h_t. They show that after $T = O(\log(1/\epsilon))$ iterations, the majority vote over $h_1, \ldots, h_T$ provides an ϵ-approximation of the desired concept (for a more detailed discussion, see also Sections 1.2 and 13.1.5 in [20]). Choosing $\epsilon < 1/m$ yields a sample compression scheme from m to $O(d \log m)$.

A first observation towards removing the $\log m$ factor is that this constructive argument from [12], [39] can be replaced by a combination of von Neumann's minimax theorem and a Chernoff bound. The $\log m$ factor eventually comes from a union bound over the m samples.

The compression scheme presented in this text replaces the union bound with a more accurate analysis for classes of low VC dimension which gives the $O(d^*)$ factor instead of the $\log m$ factor (the connection between PAC learnability and VC dimension comes from [9]).

Finally, here is a high level description of the compression process. Given a sample of the form (Y, y), the compression identifies $T \leq O(d^*)$ subsets $Z_1, \ldots, Z_T$ of Y, each of size at most d. It then compresses (Y, y) to (Z, z) with $Z = \bigcup_{t \in [T]} Z_t$ and $z = y|_Z$. The additional information $i \in I$ allows to recover $Z_1, \ldots, Z_T$ from Z. The reconstruction process gets $((Z, z), i)$ as input. It uses the information i to generate T subsets $Z_1, \ldots, Z_T$ of Z. It then uses the PAC learner for C to generate T hypotheses $h_1, \ldots, h_T$; the hypothesis h_t is the output of the learner with input $(Z_t, z|_{Z_t})$. The final reconstruction hypothesis h is the majority vote over $h_1, \ldots, h_T$. The full details of the construction and its analysis appear in Section IV.

III. Teaching

Imagine a teacher that helps a student to learn a concept c by picking insightful examples. The concept c is known only to the teacher, but c belongs to a class of concepts C known to both the teacher and the student. The teacher carefully chooses a set of examples that is tailored for c, and then provides these examples to the student. Now, the student should be able to recover c from these examples.

A central issue that is addressed in the design of mathematical teaching models is "collusions." Roughly speaking, a collusion occurs when the teacher and the student agree in advance on some unnatural encoding of information about c using the bit description of the chosen examples, instead of using attributes that separate c from other concepts. Many mathematical models for teaching were suggested: Shinohara and Miyano [13], Jackson and Tomkins [40], Goldman, Rivest and Schapire [41], Goldman and Kearns [4], Goldman and Mathias [42] Angluin and Krikis [43], Balbach [44], and Kobayashi and Shinohara [45]. We now discuss some of these models in more detail.

Teaching sets. The first mathematical models for teaching [4], [13], [14] handle collusions in a fairly restrictive way, by requiring that the teacher provides a set of examples Y that uniquely identifies c. Formally, this is captured by the notion of a teaching set. A set $Y \subseteq X$ is a teaching set for c in C if for all $c' \neq c$ in C, we have $c'|_Y \neq c|_Y$. The teaching dimension [4] captures the hardest concept to teach, i.e., it is defined as $\max_{c \in C} \min\{|Y| : Y \text{ is a teaching set for } c \text{ in } C\}$.

Teaching sets were studied in several works and under different names: Shinohara and Miyano [13] studied them implicitly in the context of teaching; Natarajan [46] called them discriminants, and studied them in the context of machine learning; Anthony et al. [14] called them specifying sets, and studied their combinatorial properties and related algorithmic problems; Goldman and Kearns [4] called them teaching sequences, and defined the teaching dimension; Kushilevitz et al. [47] called them witness sets, and studied their average size; and Hanneke [15] used them in his study of the complexity of active learning.

Defining the teaching complexity using the hardest concept is often too restrictive. Consider for example the concept class consisting of all singletons and the empty set over a domain X of size n. Its teaching complexity in these models is n, since the only teaching set for the empty set is X. This is a fairly simple concept class that has the maximum possible complexity.

Recursive teaching dimension. Goldman and Mathias [42] and Angluin and Krikis [43] therefore suggested less restrictive teaching models, and more efficient teaching schemes were indeed discovered in these models. One approach, studied by Zilles et al. [5], Doliwa et al. [6], and Samei et al. [7], uses a natural hierarchy on the concept class C which is defined as follows. The first layer in the hierarchy consists of all concepts whose teaching set has minimal size. Then, these concepts are removed and the second layer consists of all concepts whose teaching set with respect to the remaining concepts has minimal size. Then, these concepts are removed and so on, until all concepts are removed. The maximum size of a teaching set that is chosen in this process is called the *recursive teaching (RT)* dimension.

For example, the concept class consisting of singletons and the empty set, which was considered earlier and has full teaching dimension, has recursive teaching dimension 1: the first layer in the hierarchy consists of all singletons, which have teaching sets of size 1; once all singletons are removed, we are left with a concept class of size 1, the concept class $\{\emptyset\}$, and in it the empty set has a teaching set of size 0. This recursive structure can be interpreted as the following teaching plan. If the teacher wishes to teach one of the singletons, say $\{x\}$, then he provides the sample $(x, 1)$ to the student, which immediately reveals the identity of the hidden concept. If the teacher, however, wishes to teach the empty set, which does not have a small teaching set, then he provides the empty sample $\emptyset$ to the student. The student observes this info, and thinks "if the hidden concept was one of the singletons then the teacher would not provide this sample, so the hidden concept must be the empty set" and indeed the student is correct.

A similar notion to RT-dimension was independently suggested in [16] under the terminology partial IDs. There the focus was on getting a simultaneous upper bound on the size of the sets, as well as the number of layers in the recursion, and it was shown that for any concept class C both can be made at most $\log |C|$. Motivation for this study comes from the population recovery learning problem defined in [48], and indeed [16] used partial IDs to design efficient algorithms for learning distributions using imperfect data.

Previous results. Doliwa et al. [6] and Zilles et al. [5] asked whether small VC-dimension implies small recursive teaching dimension. An equivalent question was asked 10 years earlier by Kuhlmann [17]. Since the VC-dimension does not increase when concepts are removed from the class, this question is equivalent to asking whether every class with small VC-dimension has some concept in it with a small teaching set. Given the semantics of the recursive teaching dimension and the VC-dimension, an interpretation of this question is whether exact teaching is not much harder than approximate learning (i.e., PAC learning).

For infinite classes the answer to this question is negative. There is an infinite concept class with VC-dimension 1 so that every concept in it does not have a finite teaching set. An example for such a class is $C \subseteq \{0,1\}^{\mathbb{Q}}$ defined as $C = \{c_q : q \in \mathbb{Q}\}$ where c_q is the indicator function of all rational numbers that are smaller than q. The VC-dimension of C is 1, but every teaching set for some $c_q \in C$ must contain a sequence of rationals that converges to q.

For finite classes this question is open. However, in some special cases it is known that the answer is affirmative. In [17] it is shown that if C has VC-dimension 1, then its recursive teaching dimension is also 1. It is known that if C is a maximum class then its recursive teaching dimension is equal to its VC-dimension [6], [35]. Other families of concept classes for which the recursive teaching dimension is at most the VC-dimension are discussed in [6]. In the other direction, [17] provided examples of concept classes with VC-dimension d and recursive teaching dimension at least $3d/2$.

The only bound on the recursive teaching dimension for general classes was observed by both [6], [16]. It states that the recursive teaching dimension of C is at most $\log |C|$. This bound follows from a simple halving argument which shows that for all C there exists some $c \in C$ with a teaching set of size $\log |C|$.

Our contribution. Theorem 1.2 exponentially improves over the $\log |C|$ bound when the VC-dimension is small. It says that if C has VC-dimension d then there is a concept c in C with a teaching set of size $\exp(d) \log\log |C|$. In particular, the RT-dimension of C is at most $\exp(d) \log\log |C|$.

The proof. The high level idea is to identify two distinct x, x' in X so that the set of $c \in C$ so that $c(x) \neq c(x')$ is much smaller than $|C|$, add x, x' to the teaching set, and continue inductively. To achieve this, we think of C as a metric space, with the normalized hamming distance as the metric, and we use the following optimal and beautiful result of Haussler [49].

Theorem 3.1 (Haussler): Let $C \subseteq \{0,1\}^X$ be a concept class with VC-dimension d. Let $\epsilon \in (0,1]$. Let $S \subseteq C$ be so that for every two distinct concepts $c, c' \in S$, the normalized hamming distance between c, c' is at least ϵ. Then, $|S| \leq e(d+1)\left(2e/\epsilon\right)^d \leq \left(4e^2/\epsilon\right)^d$.

Roughly speaking, Haussler's theorem allows us to think of C as an L_1 metric space of dimension d. This perspective is the starting point for our construction. Think of C as a binary matrix with rows indexed by concepts in C and columns by elements of X. Haussler's theorem says that if C is large then there are two distinct rows $c \neq c'$ that are close to each other.

We actually apply Haussler's theorem on the transposed matrix C^*, whose VC-dimension is at most 2^{d+1}, by Claim 2.3. We thus conclude that there are two distinct columns $x \neq x'$ that are close to each other. This means that the number of concepts $c \in C$ so that $c(x) \neq c(x')$ is much smaller than $|C|$.

Proof of Theorem 1.2: For classes with VC-dimension 1 there is $c \in C$ with a teaching set of size 1, see e.g. [6]. We may therefore assume that $d \geq 2$.

We show that if $|C| > (4e^2)^{d \cdot 2^{d+2}}$, then there exist $x \neq x'$ in X such that

$$0 < |\{c \in C : c(x) = 0 \text{ and } c(x') = 1\}| \leq |C|^{1 - \frac{1}{d2^{d+2}}}. \tag{1}$$

From this the theorem follows, since if we iteratively add such x, x' to the teaching set, then after at most $d2^{d+2} \log\log |C|$ iterations the size of the concept class is reduced to less than $(4e^2)^{d \cdot 2^{d+2}}$. At this point we can identify a unique concept by adding at most $\log((4e^2)^{d \cdot 2^{d+2}})$ additional indices to the teaching set, using the halving argument of [6], [16]. This gives a teaching set of size at most $2d2^{d+2} \log\log |C| + d2^{d+2} \log(4e^2)$ for some $c \in C$, as required.

In order to prove (1), it is enough to show that there exist $c_x \neq c_{x'}$ in C^* such that the normalized hamming distance between $c_x, c_{x'}$ is at most $\epsilon := |C|^{-\frac{1}{d2^{d+2}}}$. Assume towards contradiction that the distance between every two concepts in C^* is more than ϵ, and assume without loss of generality that $n(C) = |C^*|$ (that is, all the columns in C are distinct). By Claim 2.3, the VC-dimension of C^* is at most 2^{d+1}. Theorem 3.1 thus implies that

$$n(C) = |C^*| \leq \left(4e^2/\epsilon\right)^{2^{d+1}} < (1/\epsilon)^{2^{d+2}}. \tag{2}$$

Where the last inequality follows from the definition of ϵ and the assumption on the size of C. Therefore, we arrive at the following contradiction:

$$\begin{aligned} |C| &\leq (n(C))^d && \text{(by Sauer-Perles-Shelah, since } VC(C) \geq 2) \\ &< (1/\epsilon)^{d \cdot 2^{d+2}} && \text{(by Equation 2 above)} \\ &= |C|. && \text{(by definition of } \epsilon) \end{aligned}$$

■

IV. Sample compression schemes

A. Preliminaries

Sample complexity. A fundamental and well-known result of Blumer, Eherenfeucht, Haussler, and Warmuth [9], which is based on an earlier work of Vapnik and Chervonenkis [1], states that PAC learning sample complexity is equivalent to VC-dimension.

Theorem 4.1 (Sample complexity of PAC learning [1], [9]): Let $C \subseteq \{0,1\}^X$ has VC dimension d then C is properly PAC learnable with $O((d\log(1/\epsilon) + \log(1/\delta))/\epsilon)$ examples, generalization error ϵ and success probability $1 - \delta$.

Approximations. The following theorem shows that every distribution can be approximated by a distribution of small support, when the statistical tests belong to a class of small VC dimension. This phenomenon was first proved by Vapnik and Chervonenkis [1], and was later quantitively improved in [50], [51].

Theorem 4.2 (Approximations for bounded VC dimension [1], [50], [51]): Let $C \subseteq \{0,1\}^X$ of VC dimension d. Let μ be a distribution on X. For all $\epsilon > 0$, there exists a multi-set $Y \subseteq X$ of size $|Y| \leq O(d/\epsilon^2)$ such that for all $c \in C$,

$$\left| \mu(\{x \in X : c(x) = 1\}) - \frac{|\{x \in Y : c(x) = 1\}|}{|Y|} \right| \leq \epsilon.$$

Minimax. Von Neumann's minimax theorem [52] is a seminal result in game theory (see the textbook [53]). Assume that there are 2 players, a row player and a column player. A pure strategy of the row player is $r \in [m]$ and a pure strategy of the column player is $j \in [n]$. Let M be a binary matrix so that $M(r,j) = 1$ if and only if the row player wins the game when the pure strategies r, j are played.

The minimax theorem says that if for every mixed strategy (a distribution on pure strategies) q of the column player, there is a mixed strategy p of the row player that guarantees the row player wins with probability at least V, then there is a mixed strategy p of the row player so that for all mixed strategies q of the column player, the row player wins with probability at least V. A similar statement holds for the column player. This implies that there is a pair of mixed strategies that form a Nash equilibrium for the zero-sum game M defines (see [53]).

Theorem 4.3 (Minimax [52]): Let $M \in \mathbb{R}^{m \times n}$ be a real matrix. Then,

$$\min_{p \in \Delta^m} \max_{q \in \Delta^n} \; p^t M q = \max_{q \in \Delta^n} \min_{p \in \Delta^m} \; p^t M q,$$

where Δ^ℓ is the set of distributions on $[\ell]$.

The arguments in the proof of Theorem 2.2 below imply the following variant of the minimax theorem, which may be of interest in the context of game theory. The minimax theorem holds for a general matrix M. In other words, there is no assumption on the set of winning/losing states in the game.

We observe that a combinatorial restriction on the winning/losing states in the game implies that there is an approximate efficient equilibrium state. Namely, if the rows of M have VC dimension d and the columns of M have VC dimension d^*, then for every $\epsilon > 0$, there is a multi-set of $O(d^*/\epsilon^2)$ pure strategies $R \subseteq [m]$ for the row player, and a multi-set of $O(d/\epsilon^2)$ pure strategies $J \subseteq [n]$ for the column player, so that a uniformly random choice from R, J guarantees the players a gain that is ϵ-close to the gain in the equilibrium strategy.

Lipton and Young [54] and Lipton, Markakis and Mehta [55] call such a pair of mixed strategies an ϵ-Nash equilibrium. They showed that in every game there are ϵ-Nash equilibriums with logarithmic support ([54] for zero-sum games and [55] for general games). They used this to find an approximate Nash equilibrium in quasi-polynomial time. The ideas presented here show that if say the rows of the matrix of the game has constant VC dimension then there are ϵ-Nash equilibriums with constant support, and that consequently an approximate Nash equilibrium can be found in polynomial time.

Carathéodory's theorem. The compression scheme uses the following simple lemma. The lemma can be seen as an approximate, combinatorial version of Carathéodory's theorem from convex geometry. Let $C \subseteq \{0,1\}^n \subset \mathbb{R}^n$ and denote by K the convex hull of C in $\mathbb{R}^n$. Carathéodory's theorem says that every point $p \in K$ is a convex combination of at most $n+1$ points from C. Lemma 4.4 imply that if C^* has small VC dimension then every $p \in K$ can be approximated by a convex combination of small support. Namely, if $\mathrm{VC}(C^*) = d^*$ then p can be ϵ-approximated in ℓ_∞ by a convex combination of at most $O(d^*/\epsilon^2)$ points from C.

Lemma 4.4 (Sampling for dual VC dimension): Let $C \subseteq \{0,1\}^X$ of VC dimension d. Let p be a distribution on concepts in C, and let $\epsilon > 0$. Then, there is a multi-set $F \subseteq C$ of size $|F| \leq O(\mathrm{VC}(C^*)/\epsilon^2)$ so that for every $x \in X$,

$$\left| p(\{c \in C : c(x) = 1\}) - \frac{|\{f \in F : f(x) = 1\}|}{|F|} \right| \leq \epsilon.$$

Proof: Every $x \in X$ corresponds to a concept in C^*. The distribution p is a distribution on the domain of the functions in C^*. The lemma follows by Theorem 4.2 applied to C^*. ■

B. The construction

Proof of Theorem 2.2: Since the VC dimension of C is d, by Theorem 4.1, there is some $s = O(d)$ and a proper learning map $H : L_C(s) \to C$ so that for every $c \in C$ and for every probability distribution q on X, there is $Z \subseteq \mathrm{supp}(q)$ of size $|Z| \leq s$ so that $q(\{x \in X : h_Z(x) \neq c(x)\}) \leq 1/3$ where $h_Z = H(Z, c|_Z)$.

Compression. Let $(Y, y) \in L_C(\infty)$. Let

$$\mathcal{H} = \mathcal{H}_{Y,y} = \{H(Z,z) : Z \subseteq Y, |Z| \leq s, z = y|_Z\} \subseteq C.$$

The compression is based on the following claim.

Claim 4.5: There are T sets $Z_1, Z_2, \ldots, Z_T \subseteq Y$, each of size at most s, with $T \leq K := O(d^*)$ so that the following holds. For $t \in [T]$, let

$$f_t = H(Z_t, y|_{Z_t}). \tag{3}$$

Then, for every $x \in Y$,

$$|\{t \in [T] : f_t(x) = y(x)\}| > T/2. \tag{4}$$

Given the claim, the compression $\kappa(Y, y)$ is defined as

$$Z = \bigcup_{t \in [T]} Z_t \quad \text{and} \quad z = y|_Z.$$

The additional information $i \in I$ allows to recover the sets $Z_1, \ldots, Z_T$ from the set Z. There are many possible ways to encode this information, but the size of I can be chosen to be at most k^k with $k := K \cdot s + 1 \leq O(d^* \cdot d)$.

Proof of Claim 4.5: By choice of H, for every distribution q on Y, there is $h \in \mathcal{H}$ so that

$$q\left(\{x \in Y : h(x) = y(x)\}\right) \geq 2/3.$$

By Theorem 4.3, there is a distribution p on $\mathcal{H}$ such that for every $x \in Y$,

$$p(\{h \in \mathcal{H} : h(x) = y(x)\}) \geq 2/3.$$

By Lemma 4.4 applied to $\mathcal{H}$ and p with $\epsilon = 1/8$, there is a multi-set $F = \{f_1, f_2, \ldots, f_T\} \subseteq \mathcal{H}$ of size $T \leq K = O(d^*)$ so that for every $x \in Y$,

$$\frac{|\{t \in [T] : f_t(x) = y(x)\}|}{T} \geq p(\{h \in \mathcal{H} : h(x) = y(x)\}) - 1/8 > 1/2.$$

For every $t \in [T]$, let Z_t be a subset of Y of size $|Z_t| \leq d$ so that

$$H(Z_t, y|_{Z_t}) = f_t.$$

■

Reconstruction. Given $((Z, z), i)$, the information i is interpreted as a list of T subsets $Z_1, \ldots, Z_T$ of Z, each of size at most d. For $t \in [T]$, let

$$h_t = H(Z_t, z|_{Z_t}).$$

Define $h = \rho((Z, z), i)$ as follows: For every $x \in X$, let $h(x)$ be a symbol that appears most in the list

$$\lambda_x((Z, z), i) = (h_1(x), h_2(x), \ldots, h_T(x)),$$

where ties are arbitrarily broken.

Correctness. Fix $(Y, y) \in L_C(\infty)$. Let $((Z, z), i) = \kappa(Y, y)$ and $h = \rho((Z, z), i)$. For $x \in Y$, consider the list

$$\phi_x(Y, y) = (f_1(x), f_2(x), \ldots, f_T(x))$$

defined in the compression process of (Y, y). The list $\phi_x(Y, y)$ is identical to the list $\lambda_x((Z, z), i)$, due to the following three reasons: Equation (3); the information i allows to correctly recover $Z_1, \ldots, Z_T$; and $y|_{Z_t} = z|_{Z_t}$ for all $t \in [T]$. Finally, by (4), for every $x \in Y$, the symbol $y(x)$ appears in more than half of the list $\lambda_x((Z, z), i)$ so indeed $h(x) = y(x)$. ■

V. Discussion and open problems

We proved that every concept class C with VC-dimension d has a sample compression scheme of size $\exp(d)$. Many of the known compression schemes for special concept classes (e.g. [11], [31], [33], [35], [36]), however, have size $O(d)$ which is optimal up to a constant factor. Warmuth's question [56] whether optimal $O(d)$-sample compression schemes always exist remains open. It is worth recalling, nevertheless, that in some of these special cases (e.g. low sign rank) our construction is more efficient and has size polynomial in d.

We have also proved that every C of VC-dimension d has a concept with a teaching set of size $\exp(d) \log\log |C|$. The main question about teaching sets, therefore, remains open: Is there always a concept with a teaching set of size depending only on the VC-dimension? Recall that this question is only interesting for finite concept classes.

The simplest case that is still open is VC-dimension 2. One can refine this case even further. VC-dimension 2 means that on any three coordinates $x, y, z \in X$, the projection $C|_{\{x,y,z\}}$ has at most 7 patterns. A more restricted family of classes is $(3, 6)$ concept classes, for which on any three coordinates there are at most 6 patterns. We can show that the recursive teaching dimension of $(3, 6)$ classes is at most three (the proof can be found in Section **??**).

Lemma 5.1: Let C be a finite $(3, 6)$ concept class. Then there exists some $c \in C$ with a teaching set of size at most 3.

Proof: Assume that $C \subseteq \{0, 1\}^X$ with $X = [n]$. If C has VC-dimension 1 then there exists $c \in C$ with a teaching set of size 1 (see [17], [57]). Therefore, assume that the VC-dimension of C is 2. Every shattered pair $\{x, x'\} \subseteq X$ partitions C to 4 nonempty sets:

$$C_{b,b'}^{x,x'} = \{c \in C : c(x) = b, c(x') = b'\},$$

for $b, b' \in \{0, 1\}$. Pick a shattered pair $\{x_*, x'_*\}$ and b_*, b'_* for which the size of $C_{b_*,b'_*}^{x_*,x'_*}$ is minimal. Without loss of generality assume that $\{x_*, x'_*\} = \{1, 2\}$ and that $b_* = b'_* = 0$. To simplify notation, we denote $C_{b,b'}^{1,2}$ simply by $C_{b,b'}$.

We prove below that $C_{0,0}$ has VC-dimension 1. This completes the proof since then there is some $c \in C_{0,0}$ and some $x \in [n] \setminus \{1,2\}$ such that $\{x\}$ is a teaching set for c in $C_{0,0}$. Therefore, $\{1,2,x\}$ is a teaching set for c in C.

First, a crucial observation is that since C is a $(3,6)$ class, no pair $\{x,x'\} \subseteq [n] \setminus \{1,2\}$ is shattered by both $C_{0,0}$ and $C \setminus C_{0,0}$. Indeed, if $C \setminus C_{0,0}$ shatters $\{x,x'\}$ then either $C_{1,0} \cup C_{1,1}$ or $C_{0,1} \cup C_{1,1}$ has at least 3 patterns on $\{x,x'\}$. If in addition $C_{0,0}$ shatters $\{x,x'\}$ then C has at least 7 patterns on $\{1,x,x'\}$ or $\{2,x,x'\}$, contradicting the assumption that C is a $(3,6)$ class.

Now, assume towards contradiction that $C_{0,0}$ shatters $\{x,x'\}$. Thus, $\{x,x'\}$ is not shattered by $C \setminus C_{0,0}$ which means that there is some pattern $p \in \{0,1\}^{\{x,x'\}}$ so that $p \notin (C \setminus C_{0,0})|_{\{x,x'\}}$. This implies that $C^{x,x'}_{p(x),p(x')}$ is a proper subset of $C_{0,0}$, contradicting the minimality of $C_{0,0}$. ∎

A more general and high level direction of research is the investigation of the combinatorial structure of and other related phenomena in concept classes of low VC-dimension. This goal is directly connected to many beautiful and useful ideas, like Vapnik and Chervonenkis's analysis of uniform convergence in statistics and probability theory [1], Valiant's influential work [2], and other practical algorithms as well as abstract mathematical theories.

ACKNOWLEDGMENT

We thank Noga Alon, Gillat Kol, Ben Lee Volk and Manfred Warmuth for helpful discussions in various stages of this work.

For the second author, the research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 257575, and from the Israel Science Foundation (grant number 339/10). For the third author, this research was partially supported by NSF grant CCF-1412958. The fourth author is a Horev fellow – supported by the Taub foundation, and his research is also supported by ISF and BSF.

REFERENCES

[1] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities." *Theory Probab. Appl.*, vol. 16, pp. 264–280, 1971.

[2] L. Valiant, "A theory of the learnable." *Commun. ACM*, vol. 27, pp. 1134–1142, 1984.

[3] N. Littlestone and M. Warmuth, "Relating data compression and learnability," *Unpublished*, 1986.

[4] S. A. Goldman and M. J. Kearns, "On the complexity of teaching," *J. Comput. Syst. Sci.*, vol. 50, no. 1, pp. 20–31, 1995. [Online]. Available: http://dx.doi.org/10.1006/jcss.1995.1003

[5] S. Zilles, S. Lange, R. Holte, and M. Zinkevich, "Models of cooperative teaching and learning." *J. Mach. Learn. Res.*, vol. 12, pp. 349–384, 2011.

[6] T. Doliwa, H. Simon, and S. Zilles, "Recursive teaching dimension, learning complexity, and maximum classes," in *ALT*, 2010, pp. 209–223.

[7] R. Samei, P. Semukhin, B. Yang, and S. Zilles, "Algebraic methods proving sauer's bound for teaching complexity," *Theor. Comput. Sci.*, vol. 558, pp. 35–50, 2014.

[8] M. J. Kearns and U. V. Vazirani, *An introduction to computational learning theory*. Cambridge, MA, USA: MIT Press, 1994.

[9] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension." *J. Assoc. Comput. Mach.*, vol. 36, no. 4, pp. 929–965, 1989.

[10] S. Floyd, "Space-bounded learning and the vapnik-chervonenkis dimension," in *COLT*, 1989, pp. 349–364.

[11] S. Floyd and M. K. Warmuth, "Sample compression, learnability, and the vapnik-chervonenkis dimension," *Machine Learning*, vol. 21, no. 3, pp. 269–304, 1995.

[12] Y. Freund, "Boosting a weak learning algorithm by majority," *Inf. Comput.*, vol. 121, no. 2, pp. 256–285, 1995.

[13] A. Shinohara and S. Miyano, "Teachability in computational learning," in *ALT*, 1990, pp. 247–255.

[14] M. Anthony, G. Brightwell, D. A. Cohen, and J. Shawe-Taylor, "On exact specification by examples," in *COLT*, 1992, pp. 311–318. [Online]. Available: http://doi.acm.org/10.1145/130385.130420

[15] S. Hanneke, "Teaching dimension and the complexity of active learning," in *COLT*, 2007, pp. 66–81.

[16] A. Wigderson and A. Yehudayoff, "Population recovery and partial identification," in *FOCS*, 2012, pp. 390–399.

[17] C. Kuhlmann, "On teaching and learning intersection-closed concept classes," in *EuroCOLT*, 1999, pp. 168–182.

[18] N. Sauer, "On the density of families of sets." *J. Comb. Theory, Ser. A*, vol. 13, pp. 145–147, 1972.

[19] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[20] Y. Freund and R. E. Schapire, *Boosting: Foundations and Algorithms*, ser. Adaptive computation and machine learning. MIT Press, 2012.

[21] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Occam's razor," *Inf. Process. Lett.*, vol. 24, no. 6, pp. 377–380, 1987.

[22] J. R. Quinlan and R. L. Rivest, "Inferring decision trees using the minimum description length principle," *Inf. Comput.*, vol. 80, no. 3, pp. 227–248, 1989.

[23] D. P. Helmbold and M. K. Warmuth, "On weak learning," *J. Comput. Syst. Sci.*, vol. 50, no. 3, pp. 551–573, 1995.

[24] P. Domingos, "The role of occam's razor in knowledge discovery," *Data Min. Knowl. Discov.*, vol. 3, no. 4, pp. 409–425, 1999.

[25] M. Marchand and J. Shawe-Taylor, "The set covering machine," *Journal of Machine Learning Research*, vol. 3, pp. 723–746, 2002.

[26] R. Samei, P. Semukhin, B. Yang, and S. Zilles, "Sample compression for multi-label concept classes," in *COLT*, 2014, pp. 371–393.

[27] M. Kearns, "Thoughts on hypothesis boosting," *Unpublished manuscript*, 1988.

[28] M. Kearns and L. G. Valiant, "Cryptographic limitations on learning boolean formulae and finite automata," in *STOC*, 1989, pp. 433–444.

[29] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, pp. 197–227, 1990.

[30] D. P. Helmbold, R. H. Sloan, and M. K. Warmuth, "Learning integer lattices," *SIAM J. Comput.*, vol. 21, no. 2, pp. 240–266, 1992.

[31] S. Ben-David and A. Litman, "Combinatorial variability of vapnik-chervonenkis classes with applications to sample compression schemes," *Discrete Applied Mathematics*, vol. 86, no. 1, pp. 3–25, 1998.

[32] A. Chernikov and P. Simon, "Externally definable sets and dependent pairs," *Israel Journal of Mathematics*, vol. 194, no. 1, pp. 409–425, 2013.

[33] D. Kuzmin and M. K. Warmuth, "Unlabeled compression schemes for maximum classes," *Journal of Machine Learning Research*, vol. 8, pp. 2047–2081, 2007.

[34] B. I. P. Rubinstein, P. L. Bartlett, and J. H. Rubinstein, "Shifting: One-inclusion mistake bounds and sample compression," *J. Comput. Syst. Sci.*, vol. 75, no. 1, pp. 37–59, 2009.

[35] B. I. P. Rubinstein and J. H. Rubinstein, "A geometric approach to sample compression," *Journal of Machine Learning Research*, vol. 13, pp. 1221–1261, 2012.

[36] R. Livni and P. Simon, "Honest compressions and their application to compression schemes," in *COLT*, 2013, pp. 77–92.

[37] S. Moran, A. Shpilka, A. Wigderson, and A. Yehudayoff, "Teaching and compressing for low VC-dimension," *ECCC*, vol. TR15-025, 2015.

[38] P. Assouad, "Densite et dimension," *Ann. Institut Fourter*, vol. 3, pp. 232–282, 1983.

[39] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[40] J. C. Jackson and A. Tomkins, "A computational model of teaching," in *COLT*, 1992, pp. 319–326.

[41] S. A. Goldman, R. L. Rivest, and R. E. Schapire, "Learning binary relations and total orders," *SIAM J. Comput.*, vol. 22, no. 5, pp. 1006–1034, 1993.

[42] S. A. Goldman and H. D. Mathias, "Teaching a smarter learner," *J. Comput. Syst. Sci.*, vol. 52, no. 2, pp. 255–267, 1996.

[43] D. Angluin and M. Krikis, "Learning from different teachers," *Machine Learning*, vol. 51, no. 2, pp. 137–163, 2003.

[44] F. Balbach, "Models for algorithmic teaching," Ph.D. dissertation, University of Lübeck, 2007.

[45] H. Kobayashi and A. Shinohara, "Complexity of teaching by a restricted number of examples," in *COLT*, 2009.

[46] B. K. Natarajan, *Machine Learning: A Theoretical Approach*. Morgan Kaufmann, 1991.

[47] E. Kushilevitz, N. Linial, Y. Rabinovich, and M. E. Saks, "Witness sets for families of binary vectors," *J. Comb. Theory, Ser. A*, vol. 73, no. 2, pp. 376–380, 1996.

[48] Z. Dvir, A. Rao, A. Wigderson, and A. Yehudayoff, "Restriction access," in *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, 2012, pp. 19–33.

[49] D. Haussler, "Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension." *J. Comb. Theory, Ser. A*, vol. 69, no. 2, pp. 217–232, 1995.

[50] Y. Li, P. M. Long, and A. Srinivasan, "Improved bounds on the sample complexity of learning," in *SODA*, 2000, pp. 309–318.

[51] M. Talagrand, "Sharper bounds for Gaussian and empirical processes." *Ann. Probab.*, vol. 22, no. 1, pp. 28–76, 1994.

[52] J. v. Neumann, "Zur theorie der gesellschaftsspiele," *Mathematische Annalen*, vol. 100, pp. 295–320, 1928.

[53] G. Owen, *Game Theory*. Academic Press, 1995.

[54] R. J. Lipton and N. E. Young, "Simple strategies for large zero-sum games with applications to complexity theory," *CoRR*, vol. cs.CC/0205035, 2002.

[55] R. J. Lipton, E. Markakis, and A. Mehta, "Playing large games using simple strategies," in *ACM Conference on Electronic Commerce*. New York, NY, USA: ACM, 2003, pp. 36–41.

[56] M. K. Warmuth, "Compressing to VC dimension many points," in *COLT/Kernel*, 2003, pp. 743–744.

[57] N. Alon, S. Moran, and A. Yehudayoff, "Sign rank, VC dimension and spectral gaps," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 21, p. 135, 2014.