Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 40 (2015)

Sample compression schemes for VC classes

Shay Moran[*] Amir Yehudayoff[†]

Abstract

Sample compression schemes were defined by Littlestone and Warmuth (1986) as an abstraction of the structure underlying many learning algorithms. Roughly speaking, a sample compression scheme of size k means that given an arbitrary list of labeled examples, one can retain only k of them in a way that allows to recover the labels of all other examples in the list. They showed that compression implies PAC learnability for binary-labeled classes, and asked whether the other direction holds. We answer their question and show that every concept class C with VC dimension d has a sample compression scheme of size exponential in d. The proof uses an approximate minimax phenomenon for binary matrices of low VC dimension, which may be of interest in the context of game theory.

1 Introduction

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. For example, support vector machines, which are commonly applied to solve classification problems, perform compression (see Chapter 6 in [6]). Another example is the use of compression to boost the accuracy of learning procedures (see [23, 11] and Chapter 4 in [13]).

About thirty years ago, Littlestone and Warmuth [23] provided a mathematical framework for studying compression in the context of learning theory. In a nutshell, they showed that compression indeed implies learnability and asked whether learnability implies compression.

[*] Departments of Computer Science, Technion-IIT, Israel and Max Planck Institute for Informatics, Saarbrücken, Germany. shaymrn@cs.technion.ac.il. Research is supported by ISF and BSF.

[†] Department of Mathematics, Technion-IIT, Israel. amir.yehudayoff@gmail.com. Horev fellow – supported by the Taub foundation. Research is also supported by ISF and BSF.

ISSN 1433-8092

1.1 Learning

Here we provide a brief description of standard learning terminology. For more information, see the books [18, 13, 6].

Imagine a student who wishes to learn a concept $c : X \to \{0, 1\}$ by observing some training examples. In order to eliminate measurability issues, we focus on the case that X is a finite or countable set (although the arguments we use are more general). The high level goal of the student is to come up with an hypothesis $h : X \to \{0, 1\}$ that is close to the unknown concept c using the least number of training examples. There are many possible ways to formally define the student's objective. An important one is Valiant's probably approximately correct (PAC) learning model [34], which is closely related to an earlier work of Vapnik and Chervonenkis [35]. This model is defined as follows.

The training examples are modeled as a pair (Y, y) where $Y \subseteq X$ is the multiset of points the student observes and $y = c|_Y$ is their labels according to c. The collection of all possible training examples is defined as follows. Let $C \subseteq \{0, 1\}^X$ be a concept class. A C-labeled sample is a pair (Y, y), where $Y \subseteq X$ is a multiset and $y = c|_Y$ for some $c \in C$. The size of a labeled sample (Y, y) is the size of Y as a multiset. 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.

The concept class C is PAC learnable with d samples, generalization error ϵ, and probability of success $1 - \delta$ if there is a learning map $H : L_C(d) \to \{0, 1\}^X$ so that the hypothesis H generates is accurate with high probability. Formally, for every $c \in C$ and for every probability distribution μ on X,

$$\Pr_{\mu^d}\Big[\big\{Y \in X^d : \mu(\{x \in X : h_Y(x) \neq c(x)\}) \leq \epsilon\big\}\Big] \geq 1 - \delta,$$

where $h_Y = H(Y, c|_Y)$. In this text, when the parameters ϵ, δ are not explicitly stated we mean that their value is $1/3$. If the image of H is contained in C, we say that C is properly PAC learnable.

A fundamental question that emerges is characterizing the sample complexity of PAC learning. The work of Blumer, Eherenfeucht, Haussler, and Warmuth [4], which is based on [35], provides such a characterization. The characterization is based on the Vapnik-Chervonenkis (VC) dimension of C, which is defined as follows. A set $Y \subseteq X$ is C-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 $x \in Y - Z$. The VC dimension of C, denoted $\mathrm{VC}(C)$, is the maximum size of a C-shattered set (it may be infinite). They proved that the sample complexity of PAC learning C is $\mathrm{VC}(C)$, up to constant factors[1].

Theorem 1.1 (Sample complexity of PAC learning [35, 4])**.** *If $C \subseteq \{0, 1\}^X$ has VC*

[1] Big O and Ω notation means up to absolute constants.

dimension d, then C is properly PAC learnable with $O((d\log(2/\epsilon)+\log(2/\delta))/\epsilon)$ samples, generalization error ϵ and success probability $1-\delta$.

1.2 Compression

Littlestone and Warmuth [23] defined sample compression schemes as a natural abstraction that captures a common property of many learning procedures, like procedures for learning geometric shapes or algebraic structures (see also [9, 10]).

Definition. 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. Formally, a sample compression scheme for C with kernel size k and side 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 $z = y|_Z$.

(ρ) The *reconstruction map*

$$\rho : L_C(k) \times I \to \{0,1\}^X$$

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

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

The size of the scheme is[2] $k + \log(|I|)$. 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. See [9, 10, 25] for more discussions of this definition, and some insightful examples.

Motivation and background. Littlestone and Warmuth showed that every compression scheme yields a natural learning procedure: Given a labeled sample (Y, y), the learner compresses it to $\kappa(Y, y)$ and outputs the hypothesis $h = \rho(\kappa(Y, y))$. They proved that this is indeed a PAC learner.

Theorem 1.2 (Compression implies learnability [23])**.** *Let $C \subseteq \{0,1\}^X$, and let κ, ρ be a sample compression scheme for C of size k. Let $d \geq 8\big(k\log(2/\epsilon) + \log(1/\delta)\big)/\epsilon$. Then, the learning map $H : L_C(d) \to \{0,1\}^X$ defined by $H(Y, y) = \rho(\kappa(Y, y))$ is PAC learning C with d samples, generalization error ϵ and success probability $1-\delta$.*

[2] Logarithms in this text are base 2.

Proof sketch. Let μ be a distribution on X, and $x_1, \ldots, x_d$ be d independent samples from μ. There are $\sum_{j=0}^{k} \binom{d}{j}$ subsets T of $[d]$ of size at most k. There are $|I|$ choices for information $i \in I$. Every fixing of T, i yields a random function $h_{T,i} = \rho((T, c|_T), i)$ that is measurable with respect to $x_T = (x_t : t \in T)$. The random function $h_{T,i}$ is independent of $x_{[d]-T}$. For every fixed T, i, x_T, therefore, if $\mu(\{x \in X : h_{T,i}(x) \neq c(x)\}) > \epsilon$ then the probability that $h_{T,i}$ agrees with c on all samples in $[d] - T$ is less than $(1-\epsilon)^{d-|T|}$. The function h is one of the functions in the random set $\{h_{T,i} : |T| \leq k, i \in I\}$, and it satisfies $h|_Y = c|_Y$. The union bound completes the proof. □

Littlestone and Warmuth also asked whether the other direction holds: *"Are there concept classes with finite dimension for which there is no scheme with bounded kernel size and bounded additional information?"*

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 [16, 17]. 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 1.2 implies that if the learning algorithm yields a sample compression scheme, then boosting follows with a multiplicative overhead of roughly $1/\epsilon$ in the sample size. In other words, efficient compression schemes immediately yield boosting.

Schapire [32] and later on Freund [11] 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 [11]). 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 roughly $d \log m$, and that such compression already implies boosting (see Section 1.3 below for more details).

Additional motivation for studying sample compression schemes relates to feature selection, which is about identifying meaningful features of the underlying domain that are sufficient for learning purposes (see e.g. [14]). The existence of efficient compression schemes, loosely speaking, shows that in any arbitrarily big data there is a small set of features that already contains all the relevant information. More concretely, a construction of an efficient compression scheme provides tools that may be helpful for feature selection.

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. Floyd and Warmuth [9, 10] constructed sample compression schemes of size $\log |C|$ for every finite concept class C. They also constructed optimal compression schemes of size d for maximum classes[3] of VC dimension d, as a first step towards solving the general

[3] That is, $C \subseteq \{0,1\}^X$ of size $|C| = \sum_{j=0}^{d} \binom{|X|}{j}$ with $d = \mathrm{VC}(C)$.

question. As the study of sample compression schemes deepened, many insightful and optimal schemes for special cases have been constructed: Floyd [9], Helmbold et al. [15], Floyd and Warmuth [10], Ben-David and Litman [3], Chernikov and Simon [5], Kuzmin and Warmuth [19], Rubinstein et al. [29], Rubinstein and Rubinstein [30], Livni and Simon [24] and more. These works discovered and utilized connections between sample compression schemes, and model theory, topology, combinatorics, and geometry. Finally, in our recent work with Shpilka and Wigderson [25], we constructed sample compression schemes of size roughly $2^{O(d)} \cdot \log\log |C|$ for every finite concept class C of VC dimension d.

1.3 Our contribution

Our main theorem states that VC classes have sample compression schemes of finite size. The key property of this compression is that its size does not depend on the size of the given sample (Y, y).

Theorem 1.3 (Compression)**.** *If $C \subseteq \{0,1\}^X$ has VC dimension d, then C has a sample compression scheme of size $2^{O(d)}$.*

Our construction (see Section 3) of sample compression schemes is overall quite short and simple. It is inspired by Freund's work [11] 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 (see Section 2 below).

The construction 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 1.4 (Compression using dual VC dimension)**.** *If $C \subseteq \{0,1\}^X$ has VC dimension $d > 0$ and C^* has VC dimension $d^* > 0$, then C has a sample compression scheme of size $k \log k$ with $k = O(d^* \cdot d)$.*

Theorem 1.3 follows from Theorem 1.4 via the following bound, which was observed by Assouad [1].

Claim 1.5 (Dual VC dimension [1])**.** *If $VC(C) \leq d$, then $VC(C^*) < 2^{d+1}$.*

A 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, 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 [11] and later on Freund and Schapire [12] showed that for every class C that is PAC learnable with d samples, there exists a compression scheme that compresses a C-labeled sample (Y, y) of size m to a sub-sample of size $k = O(d \log m)$ with additional information of $k \log k$ bits (for a more detailed discussion, see Sections 1.2 and 13.1.5 in [13]). Their constructive proof is iterative: In each iteration t, a distribution μ_t on Y is carefully and adaptively chosen. Then, d independent points from Y are drawn according to μ_t, and fed into the learning map to produce an hypothesis h_t. They showed that after $T = O(\log(1/\epsilon))$ iterations, the majority vote h over $h_1, \ldots, h_T$ is an ϵ-approximation of y with respect to the uniform measure on Y. In particular, if we choose $\epsilon < 1/m$, then h completely agrees with y on Y. This makes $T = O(\log m)$ and gives a sample compression scheme from a sample of size m to a sub-sample of size $d \cdot T = O(d \log m)$.

The size of Freund and Schapire's compression scheme is not uniformly bounded, it depends on $|Y|$. A first step towards removing this dependence is observing that their proof can be replaced by a combination of von Neumann's minimax theorem and a Chernoff bound. In this argument, 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 that utilizes the VC dimension of the dual class. This analysis ultimately replaces the $\log m$ factor by a d^* factor.

2 Preliminaries

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 [35], and was later quantitively improved in [20, 33].

Theorem 2.1 (Approximations for bounded VC dimension [35, 20, 33]). *Let $C \subseteq \{0,1\}^X$ of VC dimension d. Let μ be a distribution on X. For all $\epsilon > 0$, there exists a multiset $Y \subseteq X$ of size $|Y| \leq O(d/\epsilon^2)$ such that for all $c \in C$,*

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

Carathéodory's theorem. The following simple lemma can be thought of as an approximate and 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. The lemma says that if $\mathrm{VC}(C^*)$ is small then every $p \in K$ can be approximated by a convex combination with a small support.

Lemma 2.2 (Sampling for dual VC dimension)**.** *Let $C \subseteq \{0,1\}^X$ and let $d^* = VC(C^*)$. Let p be a distribution on C and let $\epsilon > 0$. Then, p can be ϵ-approximated in L^∞ by an average of at most $O(d^*/\epsilon^2)$ points from C. That is, there is a multiset $F \subseteq C$ of size $|F| \leq O(d^*/\epsilon^2)$ 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 2.1 applied to C^*. □

Minimax. Von Neumann's minimax theorem [27] is a seminal result in game theory (see e.g. the textbook [28]). Assume that there are 2 players[4], 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 $j \in [n]$. A mixed strategy is a distribution on pure strategies. 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 q of the column player, there is a mixed strategy p of the row player that guarantees that 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 p^*, q^* that form a Nash equilibrium for the zero-sum game M defines (see [28]).

Theorem 2.3 (Minimax [27])**.** *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 1.4 below imply the following variant of the minimax theorem, which may be of interest in the context of game theory. The minimax

[4] We focus on the case of zero-sum games.

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 multiset of $O(d^*/\epsilon^2)$ pure strategies $R \subseteq [m]$ for the row player, and a multiset 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. Such a pair of mixed strategies is called an ϵ-Nash equilibrium. Lipton and Young [22] showed that in every zero-sum game there are ϵ-Nash equilibriums with logarithmic support[5]. The ideas presented here show that if, say, the rows of the matrix of the game have constant VC dimension, then there are ϵ-Nash equilibriums with constant support.

3 A sample compression scheme

We start with a high level description of the compression process (Theorem 1.4). 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 uses the information $i \in I$ to recover $Z_1, \ldots, Z_T$ from Z, and then uses the PAC learning map H to generate T hypotheses $h_1, \ldots, h_T$ defined as $h_t = H(Z_t, z|_{Z_t})$. The final reconstruction hypothesis $h = \rho((Z, z), i)$ is the majority vote over $h_1, \ldots, h_T$.

Proof of Theorem 1.4. Since the VC dimension of C is d, by Theorem 1.1, there is $s = O(d)$ 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 $Z \subseteq \text{supp}(q)$ 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 3.1. *There are $T \leq O(d^*)$ sets $Z_1, Z_2, \ldots, Z_T \subseteq Y$, each of size at most s, so that the following holds. For $t \in [T]$, let*

$$f_t = H(Z_t, y|_{Z_t}). \tag{1}$$

[5] Lipton, Markakis and Mehta [21] proved a similar statement for general games.

Then, for every $x \in Y$,

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

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 $k = 1 + O(d^*) \cdot s \leq O(d^* \cdot d)$.

Proof of Claim 3.1. 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 2.3, 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 2.2 applied to $\mathcal{H}$ and p with $\epsilon = 1/8$, there is a multiset $F = \{f_1, f_2, \ldots, f_T\} \subseteq \mathcal{H}$ of size $T \leq O(d^*)$ 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 in $\{0, 1\}$ 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 (1); 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 (2), for every $x \in Y$, the symbol $y(x)$ appears in more than half of the list $\lambda_x((Z, z), i)$ so indeed $h(x) = y(x)$. □

4 Concluding remarks and questions

We have shown that every VC class admits a sample compression scheme with size exponential in its VC dimension. This is the first bound that depends only on the VC dimension, and holds for all binary-labeled classes. It is worth noting that many of the known compression schemes for special cases, like [10, 3, 19, 30, 24], have size d or $O(d)$ which is essentially optimal. In many of these cases, our construction is in fact of size polynomial in d, since the VC dimension of the dual class is small as well. Nevertheless, Floyd and Warmuth's question [10, 36] whether sample compression schemes of size $O(d)$ always exist remains open.

Multi-labeled classes. Unlike VC dimension, sample compression schemes as well as the fact that they imply PAC learnability naturally generalizes to multi-labeled concept classes (see e.g. [31].) Littlestone and Warmuth's question is therefore an instance of a more general question: Does the size of an optimal sample compression scheme for a given class capture the sample complexity of PAC learning of this class? A positive answer to this question will yield a universal and natural parameter that captures the sample complexity of PAC learning.

There are many generalization of VC dimension to multi-labeled concept classes $C \subseteq \Sigma^X$, see [2] and references within. An example that naturally comes up in our analysis is the distinguishing dimension $\mathrm{DD}(C)$: For every $c \in C$, define a binary concept class $B_c \subseteq \{0, 1\}^X$ as the set of all b_h, for $h \in C$, defined by $b_h(x) = 1$ if and only if $h(x) = c(x)$. Define

$$\mathrm{DD}(C) = \sup\{\mathrm{VC}(B_c) : c \in C\}.$$

If C is binary then $\mathrm{VC}(C) = \mathrm{DD}(C)$. This definition of dimension is similar to notions used in [26, 8, 2]. It can be verifies that if C is multi-labeled then our compression scheme for C has size exponential in $\mathrm{DD}(C)$. However, although $\Omega(\mathrm{VC}(C))$ is a lower bound on the sample complexity of PAC learning for a binary-labeled C, the distinguishing dimension $\mathrm{DD}(C)$ is not a lower bound on the sample complexity of PAC learning for a multi-labeled C. Indeed, an example constructed by Danieli and Shalev-Schwartz [7] implies that there is a concept class $C \subseteq \Sigma^X$ that is properly PAC learnable with $O(1)$ samples but $\mathrm{DD}(C) \geq \Omega(\log |\Sigma|)$.

Learners' complexity. The efficiency of our construction relies on the fact that every binary-labeled concept class C has a proper learner with optimal sample complexity. A closer look at the proof reveals that it is valid even if the learner is not proper; it suffices that the set of hypotheses produced by the learner have low VC dimension.

This motivates the following natural question: Is it true that for every learning map H for $C \subseteq \{0,1\}^X$ with $\mathrm{VC}(C) = d$ and for every $c \in C$, the set of hypotheses that H outputs when learning c has VC dimension $O(d)$ as well?

The answer is negative; some students learn although they make things more complicated than necessary. Here is an example. Let n be a power of 2, and consider the concept class $C = \{(00\ldots0)\} \subset \{0,1\}^X$ with $X = [n + 3\log n]$ consisting only of the all zero concept. The learning map H gets as input a labeled sample $(Y, y) \in L_C(3)$ of size 3, and outputs the following hypothesis h. If $Y \not\subseteq [n]$ then h is defined to be 0 everywhere. Otherwise, h is defined as 0 on $[n]$ and on the last $3\log n$ coordinates h is defined as $\psi(Y)$, where ψ is a bijection from $[n]^3$ to $\{0,1\}^{[3\log n]}$. First, the image of H has VC dimension $3\log n$ since the last $3\log n$ coordinates are shattered by it. Second, the map H is a PAC learner for C. Indeed, let μ be a distribution on X. If $\mu([n]) \geq 2/3$ then the error of h is always smaller than $1/3$. If $\mu([n]) < 2/3$ then the only case that h has positive error is that $Y \subseteq [n]$, which happens with probability $(2/3)^3 < 1/3$.

A variation of the question above is: Does every multi-labeled class C have a learner H that makes a nearly optimal number of samples with an image that is not much more complicated than C?

The answer for binary-labeled classes is affirmative; C has a nearly optimal proper learner. Danieli and Shalev-Schwartz [7] showed that there are multi-labeled concept classes that are PAC learnable with $O(1)$ samples but are not properly PAC learnable with $O(1)$ samples. In their example, however, the image of H has just one more concept than C. This question therefore remains open.

Acknowledgements

We thank Amir Shpilka and Avi Wigderson for helpful discussions. We also thank Ben Lee Volk and Manfred Warmuth for comments on an earlier version of this text.

References

[1] P. Assouad. Densite et dimension. *Ann. Institut Fourter*, 3:232–282, 1983. 5

[2] S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. M. Long. Characterizations of learnability for classes of {0,...,n}-valued functions. *J. Comput. Syst. Sci.*, 50(1):74–86, 1995. 10

[3] S. Ben-David and A. Litman. Combinatorial variability of Vapnik-Chervonenkis classes with applications to sample compression schemes. *Discrete Applied Mathematics*, 86(1):3–25, 1998. 5, 10

[4] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989. 2

[5] A. Chernikov and P. Simon. Externally definable sets and dependent pairs. *Israel Journal of Mathematics*, 194(1):409–425, 2013. 5

[6] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000. 1, 2

[7] A. Daniely and S. Shalev-Shwartz. Optimal learners for multiclass problems. In *COLT*, pages 287–316, 2014. 11

[8] R. M. Dudley. Universal Donsker classes and metric entropy. *Ann. Probab.*, 15(4):1306–1326, 10 1987. 10

[9] S. Floyd. Space-bounded learning and the vapnik-chervonenkis dimension. In *COLT*, pages 349–364, 1989. 3, 4, 5

[10] S. Floyd and M. K. Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995. 3, 4, 5, 10

[11] Y. Freund. Boosting a weak learning algorithm by majority. *Inf. Comput.*, 121(2):256–285, 1995. 1, 4, 5, 6

[12] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997. 6

[13] Y. Freund and R. E. Schapire. *Boosting: Foundations and Algorithms*. Adaptive computation and machine learning. MIT Press, 2012. 1, 2, 6

[14] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003. 4

[15] D. P. Helmbold, R. H. Sloan, and M. K. Warmuth. Learning integer lattices. *SIAM J. Comput.*, 21(2):240–266, 1992. 5

[16] M. Kearns. Thoughts on hypothesis boosting. *Unpublished manuscript*, 1988. 4

[17] M. Kearns and L. G. Valiant. Cryptographic limitations on learning boolean formulae and finite automata. In David S. Johnson, editor, *STOC*, pages 433–444. ACM, 1989. 4

[18] M. Kearns and U. V. Vazirani. *An introduction to computational learning theory.* MIT Press, Cambridge, MA, USA, 1994. 2

[19] D. Kuzmin and M. K. Warmuth. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 2007. 5, 10

[20] Y. Li, P. M. Long, and A. Srinivasan. Improved bounds on the sample complexity of learning. In *SODA*, pages 309–318, 2000. 6

[21] R. J. Lipton, E. Markakis, and A. Mehta. Playing large games using simple strategies. In *ACM Conference on Electronic Commerce*, pages 36–41, New York, NY, USA, 2003. ACM. 8

[22] R. J. Lipton and N. E. Young. Simple strategies for large zero-sum games with applications to complexity theory. *CoRR*, cs.CC/0205035, 2002. 8

[23] N. Littlestone and M. Warmuth. Relating data compression and learnability. *Unpublished*, 1986. 1, 3

[24] R. Livni and P. Simon. Honest compressions and their application to compression schemes. In *COLT*, pages 77–92, 2013. 5, 10

[25] S. Moran, A. Shpilka, A. Wigderson, and A. Yehudayoff. Teaching and compressing for low VC-dimension. *ECCC*, TR15-025, 2015. 3, 5

[26] B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989. 10

[27] J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928. 7

[28] G. Owen. *Game Theory.* Academic Press, 1995. 7

[29] B. I. P. Rubinstein, P. L. Bartlett, and J. H. Rubinstein. Shifting: One-inclusion mistake bounds and sample compression. *J. Comput. Syst. Sci.*, 75(1):37–59, 2009. 5

[30] B. I. P. Rubinstein and J. H. Rubinstein. A geometric approach to sample compression. *Journal of Machine Learning Research*, 13:1221–1261, 2012. 5, 10

[31] R. Samei, B. Yang, and S. Zilles. Generalizing labeled and unlabeled sample compression to multi-label concept classes. In *ALT*, pages 275–290, 2014. 10

[32] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990. 4

[33] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1):28–76, 1994. 6

[34] L.G. Valiant. A theory of the learnable. *Commun. ACM*, 27:1134–1142, 1984. 2

[35] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971. 2, 6

[36] M. K. Warmuth. Compressing to VC dimension many points. In *COLT/Kernel*, pages 743–744, 2003. 10

ECCC ISSN 1433-8092

http://eccc.hpi-web.de