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

Proper PAC learning is compressing

Shay Moran* Amir Yehudayoff†

Abstract

We prove that proper PAC learnability implies compression. Namely, if a concept $C \subseteq \Sigma^X$ is properly PAC learnable with d samples, then C has a sample compression scheme of size $2^{O(d)}$. In particular, every boolean concept class with constant VC dimension has a sample compression scheme of constant size. This answers a question of Littlestone and Warmuth (1986). The proof uses an approximate minimax phenomenon for boolean matrices of low VC dimension.

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]), and compression can be used to boost the accuracy of learning procedures (see [16, 11] and Chapter 4 in [6]).

About thirty years ago, Littlestone and Warmuth [16] provided a mathematical framework for studying compression in the context of learning theory. In a nutshell, they showed that compression implies learnability and asked whether learnability implies compression.

1.1 Definitions

Concepts and samples. Let Σ, X be finite sets (we focus on this case to eliminate measurability and similar issues but the arguments presented here are more general). A concept is a function $c : X \to \Sigma$. A concept class $C \subseteq \Sigma^X$ is a collection of concepts.

A subset Y of X is thought of as a collection of sample points. For $Y \subseteq X$ and $c \in C$, let $c|_Y$ be the restriction of the function c to the set Y. We think of $c|_Y$ as the labeling

*Departments of Computer Science, Technion-IIT, Israel and Max Planck Institute for Informatics, Saarbrücken, Germany. shaymrn@cs.technion.ac.il.

†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

of Y according to c. A C-labelled sample is a pair (Y, y), where $Y \subseteq X$ and $y = c|_Y$ for some $c \in C$. The size of a C-labelled sample (Y, y) is $|Y|$. For an integer k, denote by $L_C(k)$ the set of C-labelled samples of size at most k. Denote by $L_C(\infty)$ the set of all C-labelled samples of finite size.

PAC learning. Probably approximately correct (PAC) learning was defined in Valiant's seminal work [25]. We use the following definition. The concept class C is PAC learnable with d samples if there is a map that generates hypotheses $H : L_C(d) \to \Sigma^X$ so that for every $c \in C$ and for every probability distribution μ on X,

$$\Pr_{\mu^d}\left[\left\{Y \in X^d : \mu(\{x \in X : h_Y(x) \neq c(x)\}) \leq 1/3\right\}\right] \geq 2/3,$$

where $h_Y = H(Y, c|_Y)$. Roughly speaking, an hypothesis generated by H using d independent samples is a μ-approximation of c with reasonable probability. If the image of H is contained in C, we say that C is properly PAC learnable.

VC dimension. A boolean concept class is $C \subseteq \{0,1\}^X$. A set $Y \subseteq X$ is shattered in C 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 Vapnik-Chervonenkis (VC) dimension of C, denoted $\mathrm{VC}(C)$, is the maximum size of a shattered set in C [26].

A fundamental and well-known result of Blumer, Eherenfeucht, Haussler, and Warmuth [4], which is based on an earlier work of Vapnik and Chervonenkis [26], states that every boolean concept class C can be properly PAC learned with[1] $O(\mathrm{VC}(C))$ examples (in fact the sample complexity of PAC learning for boolean classes is captured by the VC dimension).

Sample compression schemes. Sample compression schemes were defined by Littlestone and Warmuth [16]. 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. Formally, 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*

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

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

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

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

The size of the scheme is[2] $k + \log(|I| + 1)$, 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 indicates which element in the list is the correct one.

See [9, 10, 18] for more discussions of this definition, and some insightful examples.

1.2 Background

Littlestone and Warmuth [16] proved that compression implies learnability (see Theorem 1.1 below), and asked whether learnability implies compression for boolean concept classes: *"Are there concept classes with finite dimension for which there is no scheme with bounded kernel size and bounded additional information?"*

This 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 [9, 10] constructed sample compression schemes of size $\log |C|$ for every concept class C. Freund [11] showed how to compress a sample of size m to a sample of size $O(d \log(m))$ with some side information for boolean classes of VC dimension d.

As the study of sample compression schemes deepened, many insightful and optimal schemes for special cases have been constructed: Floyd [9], Helmbold et al. [12], Floyd and Warmuth [10], Ben-David and Litman [3], Chernikov and Simon [5], Kuzmin and Warmuth [13], Rubinstein et al. [22], Rubinstein and Rubinstein [23], Livni and Simon [17] and more.

Finally, in our recent work with Shpilka and Wigderson [18], we constructed sample compression schemes of size $O(d \cdot 2^d \cdot \log\log |C|)$ using some side information for every boolean concept class C of VC dimension d.

Compression implies learnability. Littlestone and Warmuth proved that the sample complexity of PAC learning is at most (roughly) the size of a compression scheme [16].

Theorem 1.1 (Compression implies learnability [16])**.** *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 $Y = (x_1, \ldots, x_m)$*

[2] Logarithms in this text are of base two.

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* $\epsilon > 0$,

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

In particular, C *can be PAC learned with* $O(k \log(k) + \log(|I|+1))$ *samples.*

Proof sketch. There are $\sum_{j=0}^{k} \binom{m}{j}$ subsets T of $[m]$ of size at most k. There are $|I|$ choices for $i \in I$. Each choice of T, i yields a function $h_{T,i} = \rho((T, y_T), i)$ that is measurable with respect to $x_T = (x_t : t \in T)$. The function h is one of the functions in $\{h_{T,i} : |T| \leq k, i \in I\}$. For each $h_{T,i}$, the coordinates in $[m] - T$ are independent, and so if $\mu(\{x \in X : h_{T,i}(x) \neq c(x)\}) > \epsilon$ then the probability that all these $m - |T|$ samples agree with c is less than $(1-\epsilon)^{m-|T|}$. The union bound completes the proof. □

1.3 Learning is compressing

Our main theorem says that proper PAC learnability implies sample compression schemes of constant size.

Theorem 1.2 (Proper learnability implies compression)**.** *If* $C \subseteq \Sigma^X$ *is properly PAC learnable with* d *samples, then* C *has a sample compression scheme of size* $2^{O(d)}$.

The theorem specifically answers Littlestone and Warmuth's question [16]; every boolean concept class of finite VC dimension has a sample compression scheme of finite size. The theorem, however, only provides an exponential dependence on d, whereas many of the known compression schemes for special cases (e.g. [10, 3, 13, 23, 17]) have size $O(d)$. Warmuth's question [27] whether $O(d)$-sample compression schemes always exist remains open.

Our construction (see Section 3) of sample compression schemes is overall quite short and simple, but uses a different perspective of the problem than in previous work (mentioned above). 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 (these appear in Section 2).

2 Preliminaries

Sample complexity. There are many generalization of VC dimension to non-boolean concept classes (see [2] and references within). Here we use the following one. Let $C \subseteq \Sigma^X$.

For every $c \in C$, define a boolean 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 the distinguishing dimension of C as

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

This definition of dimension is similar to notions used in [19, 7, 2]. If C is boolean then $\mathrm{VC}(C) = \mathrm{DD}(C)$.

Vapnik and Chervonenkis [26] and Blumer et al. [4] proved that VC dimension is equivalent to the sample complexity of PAC learning. The distinguishing dimension is a lower bound on the sample complexity of PAC learning (see [4, 8, 2]).

Theorem 2.1 (Lower bound for sample complexity [4, 8, 2])**.** *The number of samples needed to PAC learn C is at least $\Omega(DD(C))$.*

Dual classes. Let $C \subseteq \{0,1\}^X$ be a boolean concept class. The dual concept class $C^* \subseteq \{0,1\}^C$ of C is defined as the set of all functions $f_x : C \to \{0,1\}$ so that $f_x(c) = 1$ if and only if $c(x) = 1$. 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. Assouad [1] bounded $\mathrm{VC}(C^*)$ in terms of $\mathrm{VC}(C)$.

Claim 2.2 (VC dimension of dual [1])**.** *If $VC(C) = d$ then $VC(C^*) \leq 2^{d+1}$.*

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 [26], and was later quantitively improved in [14, 24].

Theorem 2.3 (Approximations for bounded VC dimension [26, 14, 24])**.** *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 [20] is a seminal result in game theory (see the textbook [21]). 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 boolean 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 (see [21]).

Theorem 2.4 (Minimax [20]). *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 1.2 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, then for every $\epsilon > 0$, there is a multi-set of $O(2^d/\epsilon^2)$ 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, Markakis and Mehta [15] 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, and used this to find an approximate Nash equilibrium in quasi-polynomial time. The ideas presented here show that if 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.

3 A compression scheme

In the proof of Theorem 1.2, we use 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 3.1 says that if C has constant 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 $O(2^d/\epsilon^2)$ points from C.

Lemma 3.1 (Sampling for bounded 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 $|F| \leq O(2^d/\epsilon^2)$ 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. By Claim 2.2, the VC dimension of the dual class C^* is at most 2^{d+1}. 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 2.3 applied to C^*. □

3.1 The construction

Proof of Theorem 1.2. Since C is properly PAC learnable with d samples, let

$$H : L_C(d) \to C$$

be so that for every $c \in C$ and for every probability distribution q on X, there is $Z \subseteq \text{supp}(q)$ of size $|Z| \leq d$ 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

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

The compression is based on the following claim.

Claim 3.2. *There are T sets $Z_1, Z_2, \ldots, Z_T \subseteq Y$, each of size at most d, with $T \leq K := 2^{O(d)}$ so that the following holds. For $t \in [T]$, let*

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

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 $k := K \cdot d + 1 \leq 2^{O(d)}$.

Proof of Claim 3.2. 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 2.4, there is a distribution p on $\mathcal{H}$ such that for every $x \in Y$,

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

Let $B \subseteq \{0,1\}^Y$ be the set of concepts b_h, for $h \in \mathcal{H}$, defined by $b_h(x) = 1$ if and only if $h(x) = y(x)$. The distribution p induces a distribution p_B on B so that for every $x \in Y$,

$$p_B(\{b \in B : b(x) = 1\}) \geq 2/3.$$

Since C is PAC learnable with d samples, Theorem 2.1 implies $\mathrm{DD}(C) \leq O(d)$. Hence, $\mathrm{VC}(B) \leq O(d)$. By Lemma 3.1 applied to B and p_B with $\epsilon = 1/8$, there is a multi-set $E \subseteq B$ of size $|E| \leq K := 2^{O(d)}$ so that for every $x \in Y$,

$$\frac{|\{e \in E : e(x) = 1\}|}{|E|} \geq p_B(\{b \in B : b(x) = 1\}) - 1/8 > 1/2.$$

The multi-set $E \subseteq B$ corresponds to a multi-set $F = \{f_1, f_2, \ldots, f_T\} \subseteq \mathcal{H}$ of size $T = |E|$ so that for every $x \in Y$,

$$|\{t \in [T] : f_t(x) = y(x)\}| > T/2. \tag{3}$$

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 $\lambda_x((Z, z), i)$; this follows from Equation (1), from that i allows to correctly recover $Z_1, \ldots, Z_T$, and from that $y|_{Z_t} = z|_{Z_t}$ for all $t \in [T]$. By (3), for every $x \in Y$, the symbol $y(x)$ appears in more than half of the list $\lambda_x((Z, z), i)$ so indeed $h(x) = y(x)$. □

Acknowledgements

We thank Amir Shpilka and Avi Wigderson for helpful discussions. We also thank Ben Lee Volk 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. 4, 5

[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. 3, 4

[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

[5] A. Chernikov and P. Simon. Externally definable sets and dependent pairs. *Israel Journal of Mathematics*, 194(1):409–425, 2013. 3

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

[7] R. M. Dudley. Universal Donsker classes and metric entropy. *Ann. Probab.*, 15(4):1306–1326, 10 1987. 5

[8] A. Ehrenfeucht, D. Haussler, M. J. Kearns, and L. G. Valiant. A general lower bound on the number of examples needed for learning. *Inf. Comput.*, 82(3):247–261, 1989. 5

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

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

[11] Y. Freund. Boosting a weak learning algorithm by majority. *Inf. Comput.*, 121(2):256–285, 1995. 1, 3, 4

[12] D. P. Helmbold, R. H. Sloan, and M. K. Warmuth. Learning integer lattices. *SIAM J. Comput.*, 21(2):240–266, 1992. 3

[13] D. Kuzmin and M. K. Warmuth. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 2007. 3, 4

[14] Y. Li, P. M. Long, and A. Srinivasan. Improved bounds on the sample complexity of learning. In *SODA*, pages 309–318, 2000. 5

[15] 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. 6

[16] N. Littlewood and M. Warmuth. Relating data compression and learnability. *Unpublished*, 1986. 1, 2, 3, 4

[17] R. Livni and P. Simon. Honest compressions and their application to compression schemes. In *COLT*, pages 77–92, 2013. 3, 4

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

[19] B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989. 5

[20] J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928. 5, 6

[21] G. Owen. *Game Theory*. Academic Press, 1995. 5, 6

[22] 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. 3

[23] B. I. P. Rubinstein and J. H. Rubinstein. A geometric approach to sample compression. *Journal of Machine Learning Research*, 13:1221–1261, 2012. 3, 4

[24] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1):28–76, 1994. 5

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

[26] 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, 5

[27] M. K. Warmuth. Compressing to VC dimension many points. In *COLT/Kernel*, pages 743–744, 2003. 4

ECCC ISSN 1433-8092

http://eccc.hpi-web.de