Teaching and compressing for low VC-dimension

Shay Moran* Amir Shpilka† Avi Wigderson‡ Amir Yehudayoff§

Abstract

In this work we study the quantitative relation between VC-dimension and two other basic parameters related to learning and teaching. We present relatively efficient constructions of *sample compression schemes* and *teaching sets* for classes of low VC-dimension. Let C be a finite boolean concept class of VC-dimension d. Set $k = O(d2^d \log\log|C|)$.

We construct sample compression schemes of size k for C, with additional information of $k\log(k)$ bits. Roughly speaking, given any list of C-labelled examples of arbitrary length, we can retain only k labeled examples in a way that allows to recover the labels of all others examples in the list.

We also prove that there always exists a concept c in C with a teaching set (i.e. a list of c-labelled examples uniquely identifying c) of size k. Equivalently, we prove that the recursive teaching dimension of C is at most k.

The question of constructing sample compression schemes for classes of small VC-dimension was suggested by Littlestone and Warmuth (1986), and the problem of constructing teaching sets for classes of small VC-dimension was suggested by Kuhlmann (1999). Previous constructions for general concept classes yielded size $O(\log|C|)$ for both questions, even when the VC-dimension is constant.

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

†Department of Computer Science, Tel Aviv University, Israel. shpilka@post.tau.ac.il. 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).

‡School of Mathematics, Institute for Advanced Study, Princeton NJ. avi@ias.edu.

§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.

1 Introduction

The study of mathematical foundations of learning and teaching has been very fruitful, revealing fundamental connections to various other areas of mathematics, such as geometry, topology, and combinatorics. Many key ideas and notions emerged from this study: Vapnik and Chervonenkis's VC-dimension [41], Valiant's seminal definition of PAC learning [40], Littlestone and Warmuth's sample compression schemes [30], Goldman and Kearns's teaching dimension [18], recursive teaching dimension (RT-dimension, for short)[43, 11, 36] and more.

While it is known that some of these measures 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.

In this work, we make progress on these two questions. First, we prove that the RT-dimension of a boolean concept class C having VC-dimension d is upper bounded by[1] $O(d2^d \log\log|C|)$. Secondly, we give a sample compression scheme of size $O(d2^d \log\log|C|)$ that uses additional information.

1.1 VC-dimension

VC-dimension and size. A concept class over the universe X is a set $C \subseteq \{0,1\}^X$. When X is finite, we denote $|X|$ by $n(C)$. 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 $x \in Y - Z$.

The most basic result concerning VC-dimension is the Sauer-Shelah-Perles Lemma, that upper bounds $|C|$ in terms of $n(C)$ and $\mathrm{VC}(C)$. It has been independently proved several times, e.g. in [38].

Theorem 1.1 (Sauer-Shelah-Perles)**.** *Let C be a boolean concept class with VC-dimension d. Then,*

$$|C| \leq \sum_{k=0}^{d} \binom{n(C)}{k}.$$

In particular, if $d \geq 2$ then $|C| \leq n(C)^d$

VC-dimension and PAC learning. The VC-dimension is one of the most basic complexity measures for concept classes. It is perhaps mostly known in the context of the PAC learning model. PAC learning was introduced in Valiant's seminal work [40] as a theoretical model for learning from random examples drawn from an unknown distribution

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

(see the book [26] for more details). A foundational connection between VC-dimension and PAC learning was discovered by Blumer, Ehrenfeucht, Haussler and Warmuth [7], who showed that PAC learning sample complexity is equivalent to VC-dimension. The proof of this theorem uses Theorem 1.1 and an argument commonly known as double sampling (see Section A in the appendix for a short and self contained description of this well known argument).

Theorem 1.2 (Blumer et al.)**.** *Let X be a set and $C \subseteq \{0,1\}^X$ be a concept class of VC-dimension d. Let μ be a distribution over X. Let $\epsilon, \delta > 0$ and m an integer satisfying $2(2m+1)^d(1-\epsilon/4)^m < \delta$. Let $c \in C$ and $Y = (x_1, \ldots, x_m)$ be a multiset of m independent samples from μ. Then, the probability that there is $c' \in C$ so that $c|_Y = c'|_Y$ but $\mu(\{x : c(x) \neq c'(x)\}) > \epsilon$ is at most δ.*

VC-dimension and the metric structure. Another fundamental result in this area is Haussler's [22] description of the metric structure of concept classes with low VC-dimension (see also the work of Dudley [13]). Roughly, it says that a concept class C of VC-dimension d, when thought of as an L_1 metric space, behaves like a d dimensional space in the sense that the size of an ϵ-separated set in C is at most $(1/\epsilon)^d$. More formally, every probability distribution μ on X induces the (pseudo) metric

$$\mathsf{dist}_\mu(c, c') = \mu(\{x : c(x) \neq c'(x)\})$$

on C. A set $S \subseteq C$ is called ϵ-separated with respect to μ if for every two concepts $c \neq c'$ in S we have $\mathsf{dist}_\mu(c, c') > \epsilon$. A set $A = A_\mu(C, \epsilon) \subseteq C$ is called an *ϵ-approximating set*[2] for C with respect to μ if it is a maximal ϵ-separated set with respect to μ. The maximality of A implies that for every $c \in C$ there is some rounding $r = r(c, \mu, C, \epsilon)$ in A so that r is a good approximation to c, that is, $\mathsf{dist}_\mu(c, r) \leq \epsilon$. We call r a rounding of c in A.

An approximating set can be thought of as a metric approximation of the possibly complicated concept class C, and for many practical purposes it is a good enough substitute for C. Haussler proved that there are always small approximating sets.

Theorem 1.3 (Haussler)**.** *Let $C \subseteq \{0,1\}^X$ be a concept class with VC-dimension d. Let μ be a distribution on X. Let $\epsilon \in (0,1]$. If S is ϵ-separated with respect to μ then*

$$|S| \leq e(d+1)\left(\frac{2e}{\epsilon}\right)^d \leq \left(\frac{4e^2}{\epsilon}\right)^d.$$

[2] In metric spaces such a set is called an ϵ-net, however in learning theory and combinatorial geometry the term ϵ-net has a different meaning, so we use ϵ-approximating instead.

A proof of a weaker statement. For $m \geq 2\log(|S|)/\epsilon$ let $x_1, \ldots, x_m$ be independent samples from μ. For every $c \neq c'$ in S,

$$\Pr_{\mu^m}(\forall i \in [m] \;\; c(x_i) = c'(x_i)) < (1-\epsilon)^m \leq e^{-m\epsilon} \leq 1/|S|^2.$$

The union bound implies that there is a choice of $Y \subseteq X$ of size $|Y| \leq m$ so that $|S|_Y| = |S|$. Theorem 1.1 implies $|S| \leq (|Y|+1)^d$. Thus, $|S| < (30d\log(2d/\epsilon)/\epsilon)^d$. □

1.2 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 [39], Jackson and Tomkins [25], Goldman, Rivest and Schapire [20], Goldman and Kearns [18], Goldman and Mathias [19] Angluin and Krikis [2], Balbach [5], and Kobayashi and Shinohara [27]. We now discuss some of these models in more detail.

Teaching sets. The first mathematical models for teaching [18, 39, 3] 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, which was independently introduced by Goldman and Kearns [18], Shinohara and Miyano [39] and Anthony et al. [3]. 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 complexity in these models is captured by the hardest concept to teach, i.e., $\max_{c \in C} \min\{|Y| : Y \text{ is a teaching set for } c \text{ in } C\}$.

Teaching sets also appear in other areas of learning theory: Hanneke [21] used it in his study of the label complexity in active learning, and the authors of [42] used variants of it to design efficient algorithms for learning distributions using imperfect data.

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 [19] and Angluin and Krikis [2] therefore suggested less restrictive teaching models, and more efficient teaching schemes were indeed discovered in these models. One approach, studied by Zilles et al. [43], Doliwa et al. [11], and Samei et al. [36], 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 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, 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.

A similar notion to RT-dimension was independently suggested in [42] 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 [14].

Previous results. Doliwa et al. [11] and Zilles et al. [43] asked whether small VC-dimension implies small recursive teaching dimension. An equivalent question was asked 10 years earlier by Kuhlmann [28]. 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 [28] 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[3] class then its recursive teaching dimension is equal to its VC-dimension [11, 35]. Other families of concept classes for which the recursive teaching dimension is at most the VC-dimension are discussed in [11]. In the other direction, [28] provided examples of concept classes with VC-dimension d and recursive teaching dimension at least $\frac{3}{2}d$.

The only bound on the recursive teaching dimension for general classes was observed by both [11, 42]. 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. Our first main result is the following general bound, which exponentially improves over the $\log|C|$ bound when the VC-dimension is small (the proof is given in Section 3).

Theorem 1.4 (RT-dimension)**.** *Let C be a concept class of VC-dimension d. Then there exists $c \in C$ with a teaching set of size at most*

$$d2^{d+3}(\log(4e^2) + \log\log|C|).$$

It follows that the recursive teaching dimension of concept classes of VC-dimension d is at most $d2^{d+3}(\log(4e^2) + \log\log|C|)$ as well.

1.3 Sample compression schemes

A fundamental and well known statement in learning theory says that if the VC-dimension of a concept class C is small, then any consistent[4] algorithm successfully PAC learns concepts from C after seeing just a few labelled examples [41, 8]. In practice, however, a major challenge one has to face when designing a learning algorithm is the construction of an hypothesis that is consistent with the examples seen. Many learning algorithms share the property that the output hypothesis is constructed using a small subset of the examples. For example, in support vector machines, only the set of support vectors is needed to construct the separating hyperplane [10]. Sample compression schemes provide a formal meaning for this algorithmic property.

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[5] if x is in the interval $I \subset \mathbb{R}$. Can we remember just a few of

[3] That is, C satisfies Sauer-Shelah-Perles Lemma with equality.
[4] An algorithm that outputs an hypothesis in C that is consistent with the input examples.
[5] That is $c_I(x) = 1$ iff $x \in I$.

the samples in a way that allows to recover all the 100 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 formal definition. Littlestone and Warmuth [30] formally defined sample compression schemes as follows. Let $C \subseteq \{0,1\}^X$ with $|X| = n$. Let

$$L_C(k_1, k_2) = \{(Y, y) : Y \subseteq X,\ k_1 \leq |Y| \leq k_2,\ y \in C|_Y\},$$

the set of labelled samples from C, of sizes between k_1 and k_2. A k-sample compression scheme for C with information Q, for a finite set Q, consists of two maps κ, ρ for which the following hold:

(κ) The *compression map*

$$\kappa : L_C(1, n) \to L_C(0, k) \times Q$$

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

(ρ) The *reconstruction map*

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

is so that for all (Y, y) in $L_C(1, n)$,

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

Intuitively, the compression map takes a long list of samples (Y, y) and encodes it as a short sub-list of samples (Z, z) together with some small amount of side information $q \in Q$, which helps in the reconstruction phase. The reconstruction takes a short list of samples (Z, z) and decodes it using the side information q, without any knowledge of (Y, y), to an hypothesis in a way that essentially inverts the compression. Specifically, the following property must always hold: if the compression of $(Y, c|_Y)$ is the same as that of $(Y', c'|_{Y'})$ then $c|_{Y \cap Y'} = c'|_{Y \cap Y'}$.

A different perspective of the side information is as a list decoding in which the small set of labelled examples (Z, z) is mapped to the set of hypothesis $\{\rho((Z, z), q) : q \in Q\}$, one of which is correct.

We note that 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 $(Y, y) \in L_C(1, n)$ such that $h = \rho(\kappa(Y, y))$ we have that $h|_Y = y$. Thus, h has to be consistent only on the sampled coordinates that were compressed and not elsewhere.

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.

Connections to learning. Sample compression schemes are known to yield practical learning algorithms (see e.g. [32]), and allow learning for multi labelled concept classes [37].

They can also be interpreted as a formal manifestation of Occam's razor. Occam's razor is a philosophical principle attributed to William of Ockham from the late middle ages. 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 [30, 8, 15, 33, 24, 16, 12]. In the context of sample compression schemes, simplicity is captured by the size of the compression scheme. Interestingly, this manifestation of Occam's razor is provably useful [30]: Sample compression schemes imply PAC learnability.

Theorem 1.5 (Littlestone-Warmuth)**.** *Let $C \subseteq \{0,1\}^X$, 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 Q. Let $h = \rho(\kappa(Y, y))$. Then,*

$$\Pr_{\mu^m}(\mathsf{dist}_\mu(h, c) > \epsilon) < |Q| \sum_{j=0}^{k} \binom{m}{j} (1-\epsilon)^{m-j}.$$

Proof sketch. There are $\sum_{j=0}^{k} \binom{m}{j}$ subsets T of $[m]$ of size at most k. There are $|Q|$ choices for $q \in Q$. Each choice of T, q yields a function $h_{T,q} = \rho((T, y_T), q)$ that is measurable with respect to $x_T = (x_t : t \in T)$. The function h is one of the functions in $\{h_{T,q} : |T| \leq k, q \in Q\}$. For each $h_{T,q}$, the coordinates in $[m] - T$ are independent, and so if $\mathsf{dist}_\mu(h_{T,q}, c) > \epsilon$ 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. □

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, [16] proved that there are concept classes of VC-dimension d for which any sample compression scheme has size at least d.

This is part of the motivation for the following basic question that was asked by Littlestone and Warmuth [30] nearly 30 years ago and is still open: Does a concept class of VC-dimension d have a sample compression scheme of size depending only on d (and not on the universe size)?

In fact, unlike the VC-dimension, the definition of sample compression schemes as well as the fact that they imply PAC learnability naturally generalizes to multi-labelled concept classes [37]. Thus, Littlestone and Warmuth's question above can be seen as the boolean 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 of this class?

Previous constructions. Floyd [15] and Floyd and Warmuth [16] constructed sample compression schemes of size $\log |C|$. The construction in [16] uses a transformation that converts certain online learning algorithms to compression schemes. Helmbold and Warmuth [24] and Freund [17] showed how to compress a sample of size m to a sample of size $O(\log(m))$ using some side information for classes of constant VC-dimension (the implicit constant in the $O(\cdot)$ depends on the VC-dimension).

In a long line of works, several interesting compression schemes for special cases were constructed. A partial list includes Helmbold et al. [23], Floyd and Warmuth [16], Ben-David and Litman [6], Chernikov and Simon [9], Kuzmin and Warmuth [29], Rubinstein et al. [34], Rubinstein and Rubinstein [35], Livni and Simon [31] and more. These works provided connections between compression schemes and geometry, topology and model theory.

Our contribution. Here we make the first quantitive progress on this question, since the work of Floyd [15]. The following theorem shows that low VC-dimension implies the existence of relatively efficient compression schemes. The constructive proof is provided in Section 4.

Theorem 1.6 (Sample compression scheme)**.** *If C has VC-dimension d then it has a k-sample compression scheme with additional information Q where $k = O(d2^d \log\log |C|)$ and $\log |Q| \leq O(k \log(k))$.*

1.4 Discussion and open problems

This work provides relatively efficient constructions of teaching sets and sample compression schemes.

The main questions, however, remain open. Is there always a concept with a teaching set of size depending only on the VC-dimension? (The interesting case is finite con-

cept classes, as mentioned above.) Are there always sample compression schemes of size depending only on the VC-dimension?

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 3.

Lemma 1.7. *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 [28, 1]). Therefore, assume that the VC-dimension of C is 2. Every shattered pair $\{x, x'\} \subseteq X$ partitions C to 4 nonempty sets:

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

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

2 The dual class

We shall repeatedly use the dual concept class to C and its properties. The dual concept class $C^* \subseteq \{0,1\}^C$ of C is defined by $C^* = \{c_x : x \in X\}$, where $c_x : C \to \{0,1\}$ is the map so that $c_x(c) = 1$ iff $c(x) = 1$. If we think of C as a binary matrix whose rows are the concepts in C, then C^* corresponds to the distinct rows of the transposed matrix (so it may be that $|C^*| < |n(C)|$).

We use the following well known property (see [4]).

Claim 2.1 (Assouad)**.** *If the VC-dimension of C is d then the VC-dimension of C^* is at most 2^{d+1}.*

Proof sketch. If the VC-dimension of C^* is 2^{d+1} then in the matrix representing C there are 2^{d+1} rows that are shattered, and in these rows there are $d+1$ columns that are shattered. □

We also define the dual approximating set (recall the definition of $A_\mu(C,\epsilon)$ from Section 1.1). Denote by $A^*(C,\epsilon)$ the set $A_U(C^*,\epsilon)$, where U is the uniform distribution on C^*.

3 Teaching sets

In this section we prove Theorem 1.4. The high level idea is to use Theorem 1.3 and Claim 2.1 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.

Proof of Theorem 1.4. For classes with VC-dimension 1 there is $c \in C$ with a teaching set of size 1, see e.g. [11]. 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 [11, 42]. 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 2.1, the VC-dimension of C^* is at most 2^{d+1}. Theorem 1.3 thus implies that

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

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:

$$\begin{aligned} |C| &\leq (n(C))^d && \text{(by Theorem 1.1, since } VC(C) \geq 2) \\ &< \left(\frac{1}{\epsilon}\right)^{d \cdot 2^{d+2}} && \text{(by Equation 2 above)} \\ &= |C|. && \text{(by definition of } \epsilon) \end{aligned}$$

□

4 Sample compression schemes

In this section we prove Theorem 1.6. The theorem statement and the definition of sample compression schemes appear in Section 1.3.

While the details are somewhat involved, due to the complexity of the definitions, the high level idea may be (somewhat simplistically) summarized as follows.

For an appropriate choice of ϵ, we pick an ϵ-approximating set A^* of the dual class C^*. It is helpful to think of A^* as a subset of the domain X. Now, either A^* faithfully represents the sample (Y, y) or it does not (we do not formally define "faithfully represents" here). We identify the following win-win situation: In both cases, we can reduce the compression task to that in a much smaller set of concepts of size at most $\epsilon|C| \approx |C|^{1-2^{-d}}$, similarly to as for teaching sets in Section 3. This yields the same double-logarithmic behavior.

In the case that A^* faithfully represents (Y, y), Case 2 below, we recursively compress in the small class $C|_{A^*}$. In the unfaithful case, Case 1 below, we recursively compress in a (small) set of concepts for which disagreement occurs on some point of Y, just as in Section 3. In both cases, we have to extend the recursive solution, and the cost is adding one sample point to the compressed sample (and some small amount of additional information by which we encode whether Case 1 or 2 occurred).

The compression we describe is inductively defined, and has the following additional structure. Let $((Z, z), q)$ be in the image of κ. The information q is of the form $q = (f, T)$, where $T \geq 0$ is an integer so that $|Z| \leq T + O(d \cdot 2^d)$, and $f : \{0, 1, \ldots, T\} \to Z$ is a partial one-to-one function[6]. This implies that $|Q| \leq 2^{O(T \log(|Z|))}$.

The rest of this section is organized as follows. In Section 4.1 we define the compression map κ. In Section 4.2 we give the reconstruction map ρ. The proof of correctness is given in Section 4.3 and the upper bound on the size of the compression is calculcated in Section 4.4.

[6]That is, it is defined over a subset of $\{0, 1, \ldots, T\}$ and it is injective on its domain.

4.1 Compression map: defining κ

Let C be a concept class. The compression map is defined by induction on $n = n(C)$. For simplicity of notation, let $d = VC(C) + 2$.

In what follows we shall routinely use $A^*(C, \epsilon)$. There are several ϵ-approximating sets and so we would like to fix one of them, say, the one obtained by greedily adding columns to $A^*(C, \epsilon)$ starting from the first[7] column (recall that we can think of C as a matrix whose rows correspond to concepts in C and whose columns are concepts in the dual class C^*). To keep notation simple, we shall use $A^*(C, \epsilon)$ to denote both the approximating set in C^* and the subset of X composed of columns that give rise to $A^*(C, \epsilon)$. This is a slight abuse of notation but the relevant meaning will always be clear from the context.

Induction base. The base of the induction applies to all concept classes C so that $|C| \leq (4e^2)^{d \cdot 2^d + 1}$. In this case, we use the compression scheme of Floyd and Warmuth [15, 16] which has size $\log(|C|) = O(d \cdot 2^d)$, using Theorem 1.1. This compression scheme has no additional information. Therefore, to maintain the structure of our compression scheme we append to it redundant additional information by setting $T = 0$ and f to be empty.

Induction step. Let C be so that $|C| > (4e^2)^{d \cdot 2^d + 1}$. Let $0 < \epsilon < 1$ be so that

$$\epsilon |C| = \left(\frac{1}{\epsilon}\right)^{d \cdot 2^d}. \tag{3}$$

This choice balances the recursive size. By Claim 2.1, the VC-dimension of C^* is at most 2^{d-1} (recall that $d = VC(C) + 2$). Theorem 1.3 thus implies that

$$|A^*(C, \epsilon)| \leq \left(\frac{4e^2}{\epsilon}\right)^{2^{d-1}} < \left(\frac{1}{\epsilon}\right)^{2^d} < n(C). \tag{4}$$

(Where the second last inequality follows from the definition of ϵ and the assumption on the size of C and the last inequality follows from the definition of ϵ and Theorem 1.1).

Let $(Y, y) \in L_C(1, n)$. Every $x \in X$ has a rounding[8] $r(x)$ in $A^*(C, \epsilon)$. We distinguish between two cases:

Case 1: There exist $x \in Y$ and $c \in C$ such that $c|_Y = y$ and $c(r(x)) \neq c(x)$.

[7] We shall assume w.l.o.g. that there is some well known order on X.

[8] The choice of $r(x)$ also depends on C, ϵ, but to simplify the notation we do not explicitly mention it. In the introduction we denoted $r(x)$ by r_x. The reason for the change in notation is, again, its simplicity.

This is the unfaithful case in which we compress as in Section 3. Let

$$\begin{aligned}
&C' = \{c'|_{X-\{x,r(x)\}} : c' \in C, c'(x) = c(x), c'(r(x)) = c(r(x))\},\\
&Y' = Y - \{x, r(x)\},\\
&y' = y|_{Y'}.
\end{aligned}$$

Apply recursively κ on C' and the sample $(Y', y') \in L_{C'}(1, n(C'))$. Let $((Z', z'), (f', T'))$ be the result of this compression. Output $((Z, z), (f, T))$ defined as[9]

$$\begin{aligned}
&Z = Z' \cup \{x\},\\
&z|_{Z'} = z',\ \ z(x) = y(x),\\
&T = T' + 1,\\
&f|_{\{0,\ldots,T-1\}} = f'|_{\{0,\ldots,T-1\}},\\
&f(T) = x \qquad\qquad (f \text{ is defined on } T, \text{ marking that Case 1 occurred})
\end{aligned}$$

Case 2: For all $x \in Y$ and $c \in C$ such that $c|_Y = y$, we have $c(x) = c(r(x))$.

This is the faithful case, in which we compress by restricting C to A^*. Consider $r(Y) = \{r(y) : y \in Y\} \subseteq A^*(C, \epsilon)$. For each $x' \in r(Y)$, pick[10] $s(x') \in Y$ to be an element such that $r(s(x')) = x'$. Let

$$\begin{aligned}
&C' = C|_{A^*(C,\epsilon)},\\
&Y' = r(Y),\\
&y'(x') = y(s(x'))\ \forall x' \in Y'.
\end{aligned}$$

By (4), we know $|A^*(C, \epsilon)| < n(C)$. Therefore, we can recursively apply κ on C' and $(Y', y') \in L_{C'}(1, n(C'))$ and get $((Z', z'), (f', T'))$. Output $((Z, z), (f, T))$ defined as

$$\begin{aligned}
&Z = \{s(x') : x' \in Z'\},\\
&z(x) = z'(r(x))\ \forall x \in Z, \qquad\qquad (r(x) \in Z')\\
&T = T' + 1,\\
&f = f'. \qquad\qquad (f \text{ is not defined on } T, \text{ marking that Case 2 occurred})
\end{aligned}$$

The following lemma summarizes two key properties of the compression scheme. The

[9] Remember that f is a partial function.

[10] The function s can be thought of as the inverse of r. Since r is not necessarily invertible we use a different notation than r^{-1}.

correctness of this lemma follows directly from the definitions of Cases 1 and 2 above.

Lemma 4.1. *Let $(Y, y) \in L_C(1, n(C))$ and $((Z, z), (T, f))$ be the compression of (Y, y) described above, where $T \geq 1$. The following properties hold:*

1. *f is defined on T and $f(T) = x$ iff $x \in Y$ and there exists $c \in C$ such that $c|_Y = y$ and $c(r(x)) \neq c(x)$.*
2. *f is not defined on T iff for all $x \in Y$ and $c \in C$ such that $c|_Y = y$, it holds that $c(x) = c(r(x))$.*

4.2 Reconstruction map: defining ρ

The reconstruction map is similarly defined by induction on $n(C)$. Let C be a concept class and let $((Z, z), (f, T))$ be in the image[11] of κ with respect to C. Let $\epsilon = \epsilon(C)$ be as in (3).

Induction base. The induction base here applies to the same classes like the induction base of the compression map. This is the only case where $T = 0$, and we apply the reconstruction map of Floyd and Warmuth [15, 16]

Induction step. Distinguish between two cases:

Case 1: f is defined on T.

Let $x = f(T)$. Denote

$$\begin{aligned}
X' &= X - \{x, r(x)\}, \\
C' &= \{c'|_{X'} : c' \in C, c'(x) = z(x), c'(r(x)) = 1 - z(x)\}, \\
Z' &= Z - \{x, r(x)\}, \\
z' &= z|_{Z'}, \\
T' &= T - 1, \\
f' &= f|_{\{0,\ldots,T'\}}.
\end{aligned}$$

Apply recursively ρ on $C', ((Z', z'), (f', T'))$. Let $h' \in \{0,1\}^{X'}$ be the result. Output h where

$$\begin{aligned}
h|_{X'} &= h', \\
h(x) &= z(x), \\
h(r(x)) &= 1 - z(x).
\end{aligned}$$

[11] For $((Z, z), (f, T))$ not in the image of κ we set $\rho((Z, z), (f, T))$ to be some arbitrary concept.

Case 2: f is not defined on T.

Consider $r(Z) = \{r(x) : x \in Z\} \subseteq A^*(C, \epsilon)$. For each $x' \in r(Z)$, pick $s(x') \in Z$ to be an element such that $r(s(x')) = x'$. Let

$$\begin{aligned} X' &= A^*(C, \epsilon), \\ C' &= C|_{X'}, \\ Z' &= r(Z), \\ z'(x') &= z(s(x')) \ \forall x' \in Z', \\ T' &= T - 1, \\ f' &= f|_{\{0,\ldots,T'\}}. \end{aligned}$$

Apply recursively ρ on $C', ((Z', z'), (f', T'))$ and let $h' \in \{0,1\}^{X'}$ be the result. Output h satisfying

$$h(x) = h'(r(x)) \ \forall x \in X.$$

4.3 Correctness

The following lemma yields the correctness of the compression scheme.

Lemma 4.2. *Let C be a concept class, $(Y, y) \in L_C(1, n)$, $\kappa(Y, y) = ((Z, z), (f, T))$ and $h = \rho(\kappa(Y, y))$. Then,*

1. *$Z \subseteq Y$ and $z|_Z = y|_Z$, and*
2. *$h|_Y = y|_Y$.*

Proof. We proceed by induction on $n(C)$. In the base case, $|C| \leq (4e^2)^{d \cdot 2^d + 1}$ and the lemma follows from the correctness of Floyd and Warmuth's compression scheme (this is the only case in which $T = 0$). In the induction step, assume $|C| > (4e^2)^{d \cdot 2^d + 1}$. We distinguish between two cases:

Case 1: f is defined on T.

Let $x = f(T)$. This case corresponds to Case 1 in the definitions of κ and Case 1 in the definition of ρ. By Item 1 of Lemma 4.1, $x \in Y$ and there exists $c \in C$ and $x \in Y$ such that $c|_Y = y$ and $c(r(x)) \neq c(x)$. Let $C', (Y', y')$ be the class defined in Case 1 in the definition of κ. Since $n(C') < n(C)$, we know that κ, ρ on C' satisfy the induction hypothesis. Let

$$\begin{aligned} ((Z', z'), (f', T')) &= \kappa(C', (Y', y')), \\ h' &= \rho(C', ((Z', z'), (f', T'))), \end{aligned}$$

be the resulting compression and reconstruction. Since we are in Case 1 in the definitions of κ and Case 1 in the definition of ρ, $((Z, z), (f, T))$ and h have the following form:

$$\begin{aligned}
&Z = Z' \cup \{x\},\\
&z|_{Z'} = z',\ z(x) = y(x),\\
&T = T' + 1,\\
&f|_{\{0,\ldots,T-1\}} = f'|_{\{0,\ldots,T-1\}},\\
&f(T) = x,
\end{aligned}$$

and

$$\begin{aligned}
&h|_{X-\{x,r(x)\}} = h',\\
&h(x) = z(x) = y(x) = c(x),\\
&h(r(x)) = 1 - z(x) = 1 - y(x) = 1 - c(x) = c(r(x)).
\end{aligned}$$

Consider item 1 in the conclusion of the lemma. By the definition of Y' and x,

$$\begin{aligned}
Y' \cup \{x\} &\subseteq Y, && \text{(by the definition of } Y')\\
Z' &\subseteq Y'. && \text{(by the induction hypothesis)}
\end{aligned}$$

Therefore, $Z = Z' \cup \{x\} \subseteq Y$.

Consider item 2 in the conclusion of the lemma. By construction and induction,

$$h|_{Y\cap\{x,r(x)\}} = c|_{Y\cap\{x,r(x)\}} = y|_{Y\cap\{x,r(x)\}} \quad \text{and} \quad h|_{Y'} = h'|_{Y'} = y'.$$

Thus, $h|_Y = y$.

Case 2: f is not defined on T.

This corresponds to Case 2 in the definitions of κ and Case 2 in the definition of ρ. Let $C', (Y', y')$ be the result of Case 2 in the definition of κ. Since $n(C') < n(C)$, we know that κ, ρ on C' satisfy the induction hypothesis. Let

$$\begin{aligned}
&((Z', z'), (f', T')) = \kappa(C', (Y', y')),\\
&h' = \rho(C', ((Z', z'), (f', T'))),\\
&s : Y' \to Y,
\end{aligned}$$

as defined in Case 2 in the definitions of κ and Case 2 in the definition of ρ. By

construction, $((Z,z),(f,T))$ and h have the following form:

$$\begin{aligned}
Z &= \{s(x') : x' \in Z'\},\\
z(x) &= z'(r(x))\ \forall x \in Z,\\
T &= T' + 1,\\
f &= f',
\end{aligned}$$

and

$$h(x) = h'(r(x))\ \forall x \in X.$$

Consider item 1 in the conclusion of the lemma. Let $x \in Z$. By the induction hypothesis, $Z' \subseteq Y'$. Thus, $x = s(x')$ for some $x' \in Z' \subseteq Y'$. Since the range of s is Y, it follows that $x \in Y$. This shows that $Z \subseteq Y$.

Consider item 2 in the conclusion of the lemma. For $x \in Y$,

$$\begin{aligned}
h(x) &= h'(r(x)) && \text{(by the definition of } h\text{)}\\
&= y'(r(x)) && \text{(by the induction hypothesis)}\\
&= y(s(r(x))) && \text{(by the definition of } y' \text{ in Case 2 of } \kappa\text{)}\\
&= y(x), &&
\end{aligned}$$

where the last equality holds due to item 2 of Lemma 4.1: Indeed, let $c \in C$ be so that $c|_Y = y$. Since f is not defined on T, for all $x \in Y$ we have $c(x) = c(r(x))$. In addition, for all $x \in Y$ it holds that $r(s(r(x))) = r(x)$ and $s(r(x)) \in Y$. Hence, if $y(s(r(x))) \neq y(x)$ then one of them is different than $c(r(x))$, contradicting the assumption that we are in Case 2 of κ. □

4.4 The compression size

Consider a concept class C which is not part of the induction base (i.e. $|C| > (4e^2)^{d\cdot 2^d+1}$). Let $\epsilon = \epsilon(C)$ be as in (3). We show the effect of each case in the definition of κ on either $|C|$ or $n(C)$:

1. Case 1 in the definition of κ: Here the size of C' becomes smaller

$$|C'| \leq \epsilon|C|.$$

Indeed, this holds as in the dual set system C^*, the normalized hamming distance between c_x and $c_{r(x)}$ is at most ϵ and therefore the number of $c \in C$ such that $c(x) \neq c(r(x))$ is at most $\epsilon|C|$.

2. Case 2 in the definition of κ: here $n(C')$ becomes smaller as

$$n(C') = |A^*(C, \epsilon)| \leq \left(\frac{1}{\epsilon}\right)^{2^d}.$$

We now show that in either cases, $|C'| \leq |C|^{1-\frac{1}{d\cdot 2^d+1}}$, which implies that after

$$O((d \cdot 2^d + 1) \log\log |C|)$$

iterations, we reach the induction base.
In Case 1:

$$|C'| \leq \epsilon|C| = |C|^{1-\frac{1}{d\cdot 2^d+1}}. \qquad \text{(by the definition of } \epsilon)$$

In Case 2:

$$\begin{aligned} |C'| &\leq (n(C'))^d && \text{(by Theorem 1.1, since } VC(C') \leq d-2) \\ &\leq \left(\frac{1}{\epsilon}\right)^{d\cdot 2^d} && \text{(by Theorem 1.3, since } n(C') = |A^*(C,\epsilon)|) \\ &= |C|^{1-\frac{1}{d\cdot 2^d+1}}. && \text{(by definition of } \epsilon) \end{aligned}$$

Remark. *Note the similarity between the analysis of the cases above, and the analysis of the size of a teaching set in Section 3. Case 1 corresponds to the rate of the progress performed in each iteration of the construction of a teaching set. Case 2 corresponds to the calculation showing that in each iteration significant progress can be made.*

Thus, the compression map κ performs at most

$$O((d \cdot 2^d + 1) \log\log |C|)$$

iterations. In every step of the recursion the sizes of Z and T increase by at most 1. In the base of the recursion, T is 0 and the size of Z is at most $O(d \cdot 2^d)$. Hence, the total size of the compression satisfies

$$\begin{aligned} |Z| \leq k &= O(2^d d \log\log |C|), \\ \log(|Q|) &\leq O(k \log(k)). \end{aligned}$$

This completes the proof of Theorem 1.6.

References

[1] Noga Alon, Shay Moran, and Amir Yehudayoff. Sign rank, VC dimension and spectral gaps. *Electronic Colloquium on Computational Complexity (ECCC)*, 21:135, 2014. 10

[2] D. Angluin and M. Krikis. Learning from different teachers. *Machine Learning*, 51(2):137–163, 2003. 4, 5

[3] M. Anthony, G. Brightwell, D. A. Cohen, and J. Shawe-Taylor. On exact specification by examples. In *COLT*, pages 311–318, 1992. 4

[4] P. Assouad. Densite et dimension. *Ann. Institut Fourter*, 3:232–282, 1983. 11

[5] F. Balbach. *Models for algorithmic teaching*. PhD thesis, University of Lübeck, 2007. 4

[6] 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. 9

[7] 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. 3, 23

[8] A. Blumer, A.j Ehrenfeucht, D. Haussler, and M. K. Warmuth. Occam's razor. *Inf. Process. Lett.*, 24(6):377–380, 1987. 6, 8

[9] A. Chernikov and P. Simon. Externally definable sets and dependent pairs. *Israel Journal of Mathematics*, 194(1):409–425, 2013. 9

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

[11] T. Doliwa, H.-U. Simon, and S. Zilles. Recursive teaching dimension, learning complexity, and maximum classes. In *ALT*, pages 209–223, 2010. 2, 5, 6, 11

[12] P. Domingos. The role of occam's razor in knowledge discovery. *Data Min. Knowl. Discov.*, 3(4):409–425, 1999. 8

[13] R.M. Dudley. Central limit theorems for empirical measures. *Ann. Probab.*, 6:899–929, 1978. 3

[14] Zeev Dvir, Anup Rao, Avi Wigderson, and Amir Yehudayoff. Restriction access. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 19–33, 2012. 5

[15] S. Floyd. Space-bounded learning and the vapnik-chervonenkis dimension. In *COLT*, pages 349–364, 1989. 8, 9, 13, 15

[16] S. Floyd and M. K. Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995. 8, 9, 13, 15

[17] Yoav Freund. Boosting a weak learning algorithm by majority. *Inf. Comput.*, 121(2):256–285, 1995. 9

[18] S. A. Goldman and M. J. Kearns. On the complexity of teaching. *J. Comput. Syst. Sci.*, 50(1):20–31, 1995. 2, 4

[19] S. A. Goldman and H. D. Mathias. Teaching a smarter learner. *J. Comput. Syst. Sci.*, 52(2):255–267, 1996. 4, 5

[20] S. A. Goldman, R. L. Rivest, and R. E. Schapire. Learning binary relations and total orders. *SIAM J. Comput.*, 22(5):1006–1034, 1993. 4

[21] S. Hanneke. Teaching dimension and the complexity of active learning. In *COLT*, pages 66–81, 2007. 4

[22] D. Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Comb. Theory, Ser. A*, 69(2):217–232, 1995. 3

[23] D. P. Helmbold, R. H. Sloan, and M. K. Warmuth. Learning integer lattices. *SIAM J. Comput.*, 21(2):240–266, 1992. 9

[24] D. P. Helmbold and M. K. Warmuth. On weak learning. *J. Comput. Syst. Sci.*, 50(3):551–573, 1995. 8, 9

[25] J. C. Jackson and A. Tomkins. A computational model of teaching. In *COLT*, pages 319–326, 1992. 4

[26] M. J. Kearns and U. V. Vazirani. *An introduction to computational learning theory.* MIT Press, Cambridge, MA, USA, 1994. 3

[27] H. Kobayashi and A. Shinohara. Complexity of teaching by a restricted number of examples. In *COLT*, 2009. 4

[28] C. Kuhlmann. On teaching and learning intersection-closed concept classes. In *EuroCOLT*, pages 168–182, 1999. 5, 6, 10

[29] D. Kuzmin and M. K. Warmuth. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 2007. 9

[30] N. Littlewood and M. Warmuth. Relating data compression and learnability. *Unpublished*, 1986. 2, 7, 8, 9

[31] R. Livni and P. Simon. Honest compressions and their application to compression schemes. In *COLT*, pages 77–92, 2013. 9

[32] M. Marchand and J. Shawe-Taylor. The set covering machine. *Journal of Machine Learning Research*, 3:723–746, 2002. 8

[33] J. R. Quinlan and R. L. Rivest. Inferring decision trees using the minimum description length principle. *Inf. Comput.*, 80(3):227–248, 1989. 8

[34] 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. 9

[35] B. I. P. Rubinstein and J. H. Rubinstein. A geometric approach to sample compression. *Journal of Machine Learning Research*, 13:1221–1261, 2012. 6, 9

[36] R. Samei, P. Semukhin, B. Yang, and S. Zilles. Algebraic methods proving sauer's bound for teaching complexity. *Theor. Comput. Sci.*, 558:35–50, 2014. 2, 5

[37] R. Samei, P. Semukhin, B. Yang, and S. Zilles. Sample compression for multi-label concept classes. In *COLT*, volume 35, pages 371–393, 2014. 8, 9

[38] N. Sauer. On the density of families of sets. *J. Comb. Theory, Ser. A*, 13:145–147, 1972. 2

[39] A. Shinohara and S. Miyano. Teachability in computational learning. In *ALT*, pages 247–255, 1990. 4

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

[41] 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

[42] A. Wigderson and A. Yehudayoff. Population recovery and partial identification. In *FOCS*, pages 390–399, 2012. 4, 5, 6, 11

[43] S. Zilles, S. Lange, R. Holte, and M. Zinkevich. Models of cooperative teaching and learning. *J. Mach. Learn. Res.*, 12:349–384, 2011. 2, 5

A Double sampling

Here we provide our version of the double sampling argument from [7] that upper bounds the sample complexity of PAC learning for classes of constant VC-dimension. We use the following simple general lemma.

Lemma A.1. *Let $(\Omega, \mathcal{F}, \mu)$ and $(\Omega', \mathcal{F}', \mu')$ be countable[12] probability spaces. Let*

$$F_1, F_2, F_3, \ldots \in \mathcal{F}, \;\; F'_1, F'_2, F'_3, \ldots \in \mathcal{F}'$$

be so that $\mu'(F'_i) \geq 1/2$ for all i. Then

$$\mu \times \mu' \left(\bigcup_i F_i \times F'_i \right) \geq \frac{1}{2} \mu \left(\bigcup_i F_i \right),$$

where $\mu \times \mu'$ is the product measure.

Proof. Let $F = \bigcup_i F_i$. For every $\omega \in F$, let $F'(\omega) = \bigcup_{i:\omega \in F_i} F'_i$. As there exists i such that $\omega \in F_i$ it holds that $F'_i \subseteq F'(\omega)$ and hence $\mu'(F'(\omega)) \geq 1/2$. Thus,

$$\mu \times \mu' \left(\bigcup_i F_i \times F'_i \right) = \sum_{\omega \in F} \mu(\{\omega\}) \cdot \mu'(F'(\omega)) \geq \sum_{\omega \in F} \mu(\{\omega\})/2 = \mu(F)/2.$$

□

We now give a proof of Theorem 1.2. To ease the reading we repeat the statement of the theorem.

Theorem. *Let X be a set and $C \subseteq \{0,1\}^X$ be a concept class of VC-dimension d. Let μ be a distribution over X. Let $\epsilon, \delta > 0$ and m an integer satisfying $2(2m+1)^d(1-\epsilon/4)^m < \delta$. Let $c \in C$ and $Y = (x_1, \ldots, x_m)$ be a multiset of m independent samples from μ. Then, the probability that there is $c' \in C$ so that $c|_Y = c'|_Y$ but $\mu(\{x : c(x) \neq c'(x)\}) > \epsilon$ is at most δ.*

Proof of Theorem 1.2. Let $Y' = (x'_1, \ldots, x'_m)$ be another m independent samples from μ, chosen independently of Y. Let

$$H = \{h \in C : \mathsf{dist}_\mu(h, c) > \epsilon\}.$$

For $h \in C$, define the event

$$F_h = \{Y : c|_Y = h|_Y\},$$

[12] A similar statement holds in general.

and let $F = \bigcup_{h \in H} F_h$. Our goal is thus to upper bound $\Pr(F)$. For that, we also define the independent event

$$F'_h = \{Y' : \mathsf{dist}_{Y'}(h, c) > \epsilon/2\}.$$

We first claim that $\Pr(F'_h) \geq 1/2$ for all $h \in H$. This follows from Chernoff's bound, but even Chebyshev's inequality suffices: For every $i \in [m]$, let V_i be the indicator variables of the event $h(x'_i) \neq c(x'_i)$ (i.e., $V_i = 1$ if and only if $h(x'_i) \neq c(x'_i)$). The event F'_h is equivalent to $V = \sum_i V_i/m > \epsilon/2$. Since $h \in H$, we have $p := \mathbb{E}[V] > \epsilon$. Since elements of Y' are chosen independently, it follows that $\mathrm{Var}(V) = p(1-p)/m$. Thus, the probability of the complement of F'_h satisfies

$$\Pr((F'_h)^c) \leq \Pr(|V - p| \geq p - \epsilon/2) \leq \frac{p(1-p)}{(p-\epsilon/2)^2 m} < \frac{4}{\epsilon m} \leq 1/2.$$

We now give an upper bound on $\Pr(F)$. We note that

$$\Pr(F) \leq 2 \Pr\left(\bigcup_{h \in H} F_h \times F'_h\right). \qquad \text{(Lemma A.1)}$$

Let $S = Y \cup Y'$, where the union is as multisets. Conditioned on the value of S, the multiset Y is a uniform subset of half of the elements of S. Thus,

$$\begin{aligned}
2 \Pr\left(\bigcup_{h \in H} F_h \times F'_h\right) &= 2 \mathop{\mathbb{E}}_S \left[\mathbb{E}\left[\mathbb{1}_{\{\exists h \in H : h|_Y = c|_Y,\ \mathsf{dist}_{Y'}(h,c) > \epsilon/2\}} \middle| S\right]\right] \\
&= 2 \mathop{\mathbb{E}}_S \left[\mathbb{E}\left[\mathbb{1}_{\{\exists h' \in H|_S : h'|_Y = c|_Y,\ \mathsf{dist}_{Y'}(h',c) > \epsilon/2\}} \middle| S\right]\right] \\
&\leq 2 \mathop{\mathbb{E}}_S \left[\sum_{h' \in H|_S} \mathbb{E}\left[\mathbb{1}_{\{h'|_Y = c|_Y,\ \mathsf{dist}_{Y'}(h',c) > \epsilon/2\}} \middle| S\right]\right]. \\
&\qquad \text{(by the union bound)}
\end{aligned}$$

Notice that if $\mathsf{dist}_{Y'}(h', c) > \epsilon/2$ then $\mathsf{dist}_S(h', c) > \epsilon/4$, hence the probability that we choose Y such that $h'|_Y = c|_Y$ is at most $(1 - \epsilon/4)^m$. Using Theorem 1.1 we get

$$\Pr(F) \leq 2 \mathop{\mathbb{E}}_S \left[\sum_{h' \in H|_S} (1 - \epsilon/4)^m\right] \leq 2(2m+1)^d (1 - \epsilon/4)^m.$$

□