Labeled Compression Schemes for Extremal Classes

Shay Moran[1,2,3](✉) and Manfred K. Warmuth[4]

[1] Technion, Israel Institute of Technology, 32000 Haifa, Israel
shaymoran1@gmail.com
[2] Microsoft Research, Herzliya, Israel
[3] Max Planck Institute for Informatics, Saarbrücken, Germany
[4] Computer Science Department, University of California,
Santa Cruz, USA
manfred@ucsc.edu

Abstract. It is a long-standing open problem whether there exists a compression scheme whose size is of the order of the Vapnik-Chervonienkis (VC) dimension d. Recently compression schemes of size exponential in d have been found for any concept class of VC dimension d. Previously, compression schemes of size d have been given for maximum classes, which are special concept classes whose size equals an upper bound due to Sauer-Shelah. We consider a generalization of maximum classes called extremal classes. Their definition is based on a powerful generalization of the Sauer-Shelah bound called the Sandwich Theorem, which has been studied in several areas of combinatorics and computer science. The key result of the paper is a construction of a sample compression scheme for extremal classes of size equal to their VC dimension. We also give a number of open problems concerning the combinatorial structure of extremal classes and the existence of unlabeled compression schemes for them.

1 Introduction

Generalization and compression/simplification are two basic facets of "learning". Generalization concerns the expansion of existing knowledge and compression concerns simplifying our explanations of it. In machine learning, compression and generalization are deeply related: learning algorithms perform compression and the ability to compress guarantees good generalization.

A simple form of this connection is how Occam's Razor [5] is manifested in Machine Learning: if the input sample can be compressed to a small number of bits which encodes a hypothesis consistent with the input sample, then good generalization is guaranteed. A more sophisticated notion of compression is given by "sample compression schemes" [20]. In these schemes the input sample is compressed to a carefully chosen small subsample that encodes a hypothesis consistent with the input sample. For example support vector machine can be

Supported by NSF grant IIS-1118028. See [25] for a slightly more detailed version.

R. Ortner et al. (Eds.): ALT 2016, LNAI 9925, pp. 34–49, 2016.
DOI: 10.1007/978-3-319-46379-7_3

seen as compressing the original sample to the subset of support vectors which represent a maximum margin hyperplane that is consistent with the entire original sample.

What is the connection to generalization? In the Occam's razor setting, the generalization error decreases with the number of bits that are used to encode the output hypothesis. Similarly for compression schemes, the generalization error decreases with the sample size.

A core question is what parameter of the concept class characterizes the sample size required for good generalization? The Vapnik-Chervonenkis (VC) dimension serves as such a parameter [4], where the exact definition of generalization underlying our discussion is specified by the Probably Approximately Correct (PAC) model of learning [32]. The size of the best compression scheme is an alternate parameter and has several additional advantages: (i) Compression schemes frame many natural algorithms (e.g. support vector machines). (ii) Unlike the VC dimension, the definition of sample compression schemes as well as the fact that they yield low generalization error extends naturally to multi label concept classes [29]. This is particularly interesting when the number of labels is very large (or possibly infinite), because for that case there is no known combinatorial parameter that characterizes the sample complexity in the PAC model (See [9]).

Previous Work. In 1986, [20] defined *sample compression schemes* and showed that the sample size required for learning grows linearly with the size of the subsamples the scheme compresses to. They have also posed the other direction as an open question: Does every concept class have a compression scheme of size depending only on its VC dimension? Later [12,33], refined this question: Does every class of VC dimension d have a sample compression scheme of size $O(d)$.

[3] proved a compactness theorem for sample compression schemes. It essentially says that existence of compression schemes for infinite classes follows[1] from the existence of such schemes for finite classes. Thus, it suffices to consider only finite concept classes. [12] constructed sample compression schemes of size $\log |C|$ for every concept class C. More recently [24] have constructed sample compression schemes of size $\exp(d) \log \log |C|$ where $d = VCdim(C)$. Finally, [26] have constructed sample compression scheme of size $\exp(d)$, resolving Littlestone and Warmuth's question. Their compression scheme is based on an earlier compression scheme which was defined in the context of boosting (see [13]). This sample compression scheme is of variable size: It compresses samples of size m to subsamples of size $O(d \log m)$.

For many natural and important families of concept classes, sample compression schemes of size equal the VC dimension were constructed (e.g. [3,8,21,28]). However, the question whether there exists a compression scheme whose size is equal or linear in the VC dimension remains open.

[12] observed that in order to prove the conjecture it suffices to consider only maximal classes (A class C is maximal if no concept can be added without

[1] The proof of that theorem is however non-constructive.

increasing the VC dimension). Furthermore, they constructed sample compression schemes of size d for every *maximum class* of VC dimension d. These classes are maximum in the sense that their size equals an upper bound (due to Sauer-Shelah) on the size of any concept class of VC dimension d. Later, [17,28] provided even more efficient sample compression schemes for maximum classes that are called unlabeled compression schemes because the labels of the subsample are not needed to encode the output hypothesis.

One possibility of making a progress on Floyd and Warmuth's question is by extending the optimal compression schemes for maximum classes to a more general family. In this paper we consider a natural and rich generalization of maximum classes which are known as extremal classes (or shattering extremal classes). Similar to maximum classes, these classes are defined when a certain inequality, which generalizes the Sauer-Shelah bound, and known as The Sandwich Theorem is tight. The Sandwich Theorem as well as extremal classes were discovered several times and independently by several groups of researchers and in several contexts such as Functional analysis [27], Discrete-geometry [18], Phylogenetic Combinatorics [2,11] and Extremal Combinatorics [6,7]. Even though a lot of knowledge regarding the structure of extremal classes has been accumulated, the understanding of these classes is still considered incomplete by several authors [6,15].

Our Results. Our main result is a construction of sample compression scheme of size d for every extremal class of VC dimension d. When the concept class is maximum, then our scheme specializes to the compression scheme for maximum classes given in [12]. Our generalized sample compression scheme for extremal classes is still easy to describe. However its analysis requires more combinatorics and heavily exploits the rich structure of extremal classes. Despite being more general, the construction is simple. We also give explicit examples of maximal classes that are extremal but not maximum (see Example 5).

We also discuss a certain greedy peeling method for producing an unlabeled compressions scheme. Such schemes were first conjectured in [17] and later proven to exist for maximum classes [28]. However the existence of such schemes for extremal classes remains open. We relate the existence of such schemes to basic open questions concerning the combinatorial structure of extremal classes.

Organization. In Sect. 2 we give some preliminary definitions and define extremal classes. We also discuss some basic properties and give some examples of extremal classes which demonstrate their generality over maximum classes. In Sect. 3 we give a labeled compression scheme for any extremal class of VC dimension d. Finally, in Sect. 4 we relate unlabeled compression schemes for extremal classes with basic open questions concerning extremal classes.

2 Extremal Classes

2.1 Preliminaries

Concepts, Concept Classes, and the One-Inclusion Graph. A concept c is a mapping from some domain to $\{0, 1\}$. We assume that the domain of c (denoted by $dom(c)$) is finite and allow the case that $dom(c) = \emptyset$. A concept c can also be viewed as a characteristic function of a subset of $dom(c)$, i.e. for any domain point $x \in dom(c)$, $c(x) = 1$ iff $x \in c$. A concept class C is a set of concepts with the same domain (denoted by $dom(C)$). A concept class can be represented by a binary table (see Fig. 1), where the rows correspond to concepts and the columns to the elements of $dom(C)$. Whenever the elements in $dom(C)$ are clear from the context, then we represent concepts as bit strings of length $|dom(C)|$ (See Fig. 1).

The concept class C can also be represented as a subgraph of the Boolean hypercube with $|dom(C)|$ dimensions. Each dimension corresponds to a particular domain element, the vertices are the concepts in C and two concepts are connected with an edge if they disagree on the label of a single element (Hamming distance 1). This graph is called the *one-inclusion graph* of C. Note that each edge is naturally labeled by the single dimension/element on which the incident concepts disagree (See Fig. 1).

	x_1	x_2	x_3	x_4	x_5	x_6
c_1	0	0	0	0	0	0
c_2	0	0	1	0	0	0
c_3	0	1	0	0	0	0
c_4	1	0	0	0	0	0
c_5	0	0	1	0	1	0
c_6	0	0	1	1	0	0
c_7	1	0	1	0	0	0
c_8	1	1	0	0	0	0
c_9	0	0	1	0	1	1
c_{10}	0	0	1	1	1	0
c_{11}	0	0	1	1	0	1
c_{12}	1	0	1	1	0	0
c_{13}	1	1	1	0	0	0
c_{14}	1	1	0	1	0	0
c_{15}	0	0	1	1	1	1
c_{16}	1	0	1	1	0	1
c_{17}	1	1	1	1	0	0
c_{18}	1	0	1	1	1	1

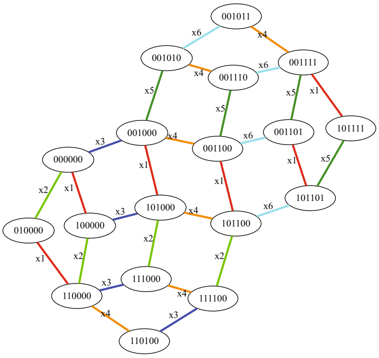

Fig. 1. The table and the one-inclusion graph of an extremal class C of VC dimension 2. The reduction $C^{x_2} = \{00000, 10000, 11000, 11100\}$ has the domain $\{x_1, x_3, x_4, x_5, x_6\}$. Notice that each concept in C^{x_2} corresponds to an edge labelled with x_2. Similarly $C^{\{x_3,x_4\}}$ consists of the single concept $\{1100\}$ over the reduced domain $\{x_1, x_2, x_5, x_6\}$. Notice that this concept corresponds to the single cube of C with dimension set $\{x_3, x_4\}$.

Restrictions and Samples. We denote the *restriction/sample* of a concept c onto $S \subseteq dom(c)$ as $c|S$. This concept has the restricted domain S and labels this domain consistently with c. Essentially concept $c|S$ is obtained by removing from row c in the table all columns not in S. The *restriction/set of samples* of an entire class C onto $S \subseteq dom(C)$ is denoted as $C|S$. A table for $C|S$ is produced by simply removing all columns not in S from the table for C and collapsing identical rows.[2] Also the one-inclusion graph for the restriction $C|S$ is now a subgraph of the boolean hypercube with $|S|$ dimensions instead of the full dimension $|dom(C)|$. We also use $C - S$ as shorthand for $C|(dom(C) \setminus S)$ (since the columns labeled with S are removed from the table). Note that the sub domain $S \subseteq dom(C)$ induces an equivalence class on C: Two concepts $c, c' \in C$ are equivalent iff $c|S = c'|S$. Thus there is one equivalence class per concept of $C|S$.

Cubes. A concept class B is called a cube if for some subset S of the domain $dom(B)$, the restriction $B|S$ is the set of all $2^{|S|}$ concepts over the domain S and the class $B - S$ contains a single concept. We denote this single concept by $\text{tag}(B)$. In this case, we say that S is the dimension set of B (denoted as $\dim(B)$). For example, if B contains two concepts that are incident to an edge labeled x then B is a cube with $\dim(B) = \{x\}$. We say that B *is a cube of* concept class C if B is a cube that is a subset of C. We say that B is a maximal cube of C if there exists no other cube of C which strictly contains B. When the dimensions are clear from the context, then a concept is described as a bit string of length $dom(C)$. Similarly a cube, B, is described as an expression in $\{0, 1, *\}^{|dom(C)}$, where the dimensions of $\dim(B)$ are the *'s and the remaining bits is the concept $\text{t}ag(B)$.

Reductions. In addition to the restriction it is common to define a second operation on concept classes. We will describe this operation using cubes. The *reduction C^S* is a concept class on the domain $dom(C) \setminus S$ which has one concept per cube with dimensions set S

$$C^S := \{\text{tag}(B) : B \text{ is a cube of } C \text{ such that } \dim(B) = S\}.$$

The reduction with respect to a single dimension x is denoted as C^x. See Fig. 1 for some examples.

Shattering and Strong Shattering. We say that $S \subseteq dom(C)$ is *shattered* by C, if $C|S$ is the set of all $2^{|S|}$ concepts over the domain S. Furthermore, S is *strongly shattered* by C, if C has a cube with dimensions set S. We use $s(C)$ to denote all shattered sets of C and $\text{st}(C)$ to denote all strongly shattered sets, respectively. Clearly, both $s(C)$ and $st(C)$ are closed under the subset relation, and $st(C) \subseteq s(C)$.

The following theorem is the result of accumulated work by different authors, and parts of it were rediscovered independently several times [1,6,11,27].

[2] We define $c|\emptyset = \emptyset$. Note that $C|\emptyset = \{\emptyset\}$ if $C \neq \emptyset$ and $\emptyset$ otherwise.

Theorem 1 (Sandwich Theorem). *For any concept class C, $|\mathrm{st}(C)| \leq |C| \leq |s(C)|$.*

This theorem has been discovered independently several times and has several proofs (see [23] for more details).

The inequalities in this theorem can be strict: Let $C \subseteq \{0,1\}^n$ be such that C contains all boolean vectors with an even number of $1's$. Then $st(C)$ contains only the empty set and $s(C)$ contains all subsets of $\{1,\ldots,n\}$ of size at most $n-1$. Thus in this example, $|st(C)| = 1$, $|C| = 2^{n-1}$, and $|s(C)| = 2^n - 1$.

The *VC dimension* [4] is defined as: $VCdim(C) = \max\{|S| : S \in s(C)\}$. Clearly, $s(C) \subseteq \{S \subseteq dom(C) : |S| \leq VCdim(C)\}$ and hence the cardinality $|C| \leq \sum_{i=0}^{VCdim(C)} \binom{|dom(C)|}{i}$. Thus the Sandwich theorem implies the well-known Sauer-Shelah Lemma [30,31].

2.2 Definition of Extremal Classes and Examples

Maximum classes are defined as concept classes which satisfy the Sauer-Shelah inequality with equality. Analogously, *extremal classes* are defined as concept classes which satisfy the inequalities[3] in the Sandwich Theorem with equality: A concept class C is *extremal* if for every shattered set S of C there is a cube of C with dimension set S, i.e. $s(C) = \mathrm{st}(C)$.

Every maximum class is an extremal class. Moreover, maximum classes of VC dimension d are precisely the extremal classes for which the shattered sets consist of all subsets of the domain of size up to d. The other direction does not hold - there are extremal classes that are not maximum. All the following examples are extremal but not maximum.

Example 1. Consider the concept class C over the domain $\{x_1,\ldots,x_6\}$ given in Fig. 1. In this example $st(C) = s(C) = \big\{\emptyset, \{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}, \{x_1,x_2\}, \{x_1,x_3\}, \{x_1,x_4\}, \{x_1,x_5\}, \{x_1,x_6\},\{x_2,x_3\}, \{x_2,x_4\}, \{x_3,x_4\}, \{x_4,x_5\}, \{x_4,x_6\}, \{x_5,x_6\}\big\}$. This example also demonstrates the cubical structure of extremal classes.

Example 2 **(Downward-Closed Classes).** A standard example of a maximum class of VC dimension d is

$$C = \{c \in \{0,1\}^n : \text{ the number of 1's in } c \text{ is at most } d\}.$$

This is simply the hamming ball of radius d around the all 0's concept. A natural generalization of such classes are downward closed classes. We say that C is downward closed if for all $c \in C$ and for all $c' \leq c$, also $c' \in C$. Here $c' \leq c$ means that for every $x \in dom(C)$, $c'(x) \leq c(x)$. It is not hard to verify that every downward closed class is extremal.

[3] There are two inequalities in the Sandwich Theorem, but every class which satisfies one of them with equality also satisfies the other with equality (See Theorem 2).

Example 3 **(Hyper-Planes Arrangements in a Convex Domain).** Another standard set of examples for maximum classes comes from geometry (see e.g. [14]). Let H be an arrangement of hyperplanes in $\mathbb{R}^d$. For each hyperplane $p_i \in H$, pick one of half-planes determined by p_i to be its *positive side* and the other its *negative side*. The hyperplanes of H cut $\mathbb{R}^d$ into open regions (*cells*). Each cell defines a binary mapping with domain H:

$$c(p_i) = \begin{cases} 1 & \text{if } c \text{ is in the positive side of } p_i \\ 0 & \text{if } c \text{ is in the negative side of } p_i. \end{cases}$$

It is known that if the hyperplanes are in general position, then the set C of all cells is a maximum class of VC dimension d.

Consider the following generalization of these classes: Let $K \subseteq \mathbb{R}^d$ be a convex set. Instead of taking the vectors corresponding to all of the cells, take only those that correspond to cells that intersect K:

$$C_K = \{c : c \text{ corresponds to a cell that intersects } K\}.$$

C_K is extremal. In fact, for C_K to be extremal it is not even required that the hyperplanes are in general position. It suffices to require that no $d+1$ hyperplanes have a non-empty intersection (e.g. parallel hyperplanes are allowed). Figure 2 illustrates such a class C_K in the plane. These classes were studied in [23].

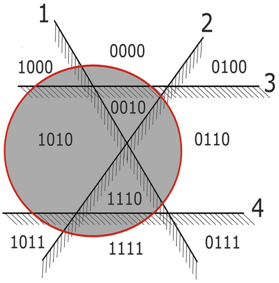

Fig. 2. An extremal class that correspond to the cells of a hyperplane arrangement of a convex set. An arrangement of 4 lines is given which partitions the plane to 10 cells. Each cell corresponds to a binary vector which specifies its location relative to the lines. For example the cell corresponding to 1010 is on the positive sides of lines 1 and 3 and on the negative side of lines 3 and 4. Here the convex set K is an ellipse and the extremal concept class consisting of the cells the ellipse intersects is $C_K = \{1000, 1010, 1011, 1111, 1110, 0010, 0000, 0110\}$ (the cells 0100, 0111 are not intersected by the ellipse). The class C_K here has VC dimension 2. Note that it's shattered sets of size 2 are exactly the pairs of lines whose intersection point lies in the ellipse K.

Interestingly, extremal classes also arise in the context of graph theory:

Example 4 **(Edge-Orientations Which Preserve Connectivity [16]).** Let $G = (V, E)$ be an undirected simple graph and let $\overrightarrow{E}$ be a fixed reference orientation. Now an arbitrary orientation of E is a function $d : E \to \{0, 1\}$: If $d(e) = 0$ then e is oriented as in $\overrightarrow{E}$ and if $d(e) = 1$ then e is oriented opposite to $\overrightarrow{E}$. Now let $s, t \in V$ be two fixed vertices, and consider all orientations of E for which there exists a directed path from s to t. The corresponding class of orientations $E \to \{0, 1\}$ is an extremal concept class over the domain E.

Moreover, the extremality of this class yields the following result in graph theory: The number of orientations for which there exists a directed path from s to t equals the number of subgraphs for which there exists an undirected path from s to t. For a more thorough discussion and other examples of extremal classes related to graph orientations see [16].

Example 5 **(General Construction of a Maximal Class that is Extremal but not Maximum).** Take a k-dimensional cube and glue to each of its vertices an edge of a new distinct dimension. The resulting class has 2^{k+1} concepts and $n = 2^k + k$ dimensions. Let C be the complement of that class.

Claim. C is an extremal maximal class of VC dimension $n - 2$ which is not maximum.

A proof of this claim is given in the full version of this paper [25]. Note that $|C| = 2^n - 2^{k+1} = 2^{2^k+k} - 2^{k+1}$ and maximum classes of VCdim $d = n - 2$ over n dimensions have size $2^n - n - 1 = 2^{2^k+k} - 2^k - k - 1$. So the maximum classes of VCdim $n - 2$ are by $2^k - k - 1$ larger than the constructed extremal maximal class of VCdim $n - 2$.

2.3 Basic Properties of Extremal Classes

Extremal classes have a rich combinatorial structure (See [23] and references within for more details). We discuss some of parts which are relevant to compression schemes.

The following theorem provides alternative characterizations of extremal classes:

Theorem 2 **([2,6]).** *The following statements are equivalent:*

1. *C is extremal, i.e. $s(C) = \mathrm{st}(C)$.*
2. *$|s(C)| = |\,\mathrm{st}(C)|$.*
3. *$|st(C)| = |C|$.*
4. *$|C| = |s(C)|$.*
5. *$\{0, 1\}^n \setminus C$ is extremal.*

The following theorem shows that the property of "being an extremal class" is preserved under standard operations. It was also proven independently by several authors (e.g. [2,6]).

Theorem 3. *Let C be any extremal class, $S \subseteq dom(C)$, and B be any cube such that $dom(B) = dom(C)$. Then $C - S$ and C^S are extremal concept classes over the domain $dom(C) - S$ and $B \cap C$ is an extremal concept class over the domain $dom(C)$.*

Note that if C is maximum then $C - S$ and C^S are also maximum, but $B \cap C$ is not necessarily maximum. This is an example of the advantage extremal classes have over the more restricted notion of maximum classes.

Interestingly, the fact that extremal classes are preserved under intersecting with cubes yields a rather simple proof (communicated to us by Ami Litman) of the fact that every extremal class is "distance preserving". This property also holds for maximum classes [14], however the proof for extremal classes is much simpler than the previous proof for maximum classes (given in [14]):

Theorem 4 [15]. *Let C be any extremal class. Then for every $c_0, c_1 \in C$, the distance between c_0 and c_1 in the one-inclusion graph of C equals the hamming distance between c_1 and c_2.*

The proof is given in the full version of this paper [25].

The following lemma brings out the special cubical structure of extremal classes. We will use it to prove the correctness of the compression scheme given in the following section. It shows that if B_1 and B_2 are two maximal cubes of an extremal class C then their dimensions sets $\dim(B_1)$ and $\dim(B_2)$ are incomparable.

Lemma 1. *Given B_1 and B_2 are two cubes of an extremal class C. If B_1 is maximal, then*

$$\dim(B_1) \subseteq \dim(B_2) \implies B_1 = B_2.$$

A proof is given in the full version of this paper [25].

3 A Labeled Compression Scheme for Extremal Classes

Let C be a concept class. On a high level, a sample compression scheme for C compresses every sample of C to a subsample of size at most k and this subsample represents a hypothesis on the entire domain of C that must be consistent with the original sample. More formally, a labeled compression scheme of size k for C consists of a compression map κ and a reconstruction map ρ. The domain of the compression map consists of all samples from concepts in C: For each sample s, κ compresses it to a subsample s' of size at most k. The domain of the reconstruction function ρ is the set of all samples of C of size at most k. Each such sample is used by ρ to reconstruct a concept h with $dom(h) = dom(C)$. The sample compression scheme must satisfy that for all samples s of C, $\rho(\kappa(s))\ |dom(s) = s$. The sample compression scheme is said to be *proper* if the reconstructed hypothesis h always belongs to the original concept class C.

A proper labeled compression scheme for extremal classes of size at most the VC dimension is given in Algorithm 1. Let C be an extremal concept class and s be a sample of C. In the compression phase the algorithm finds any *maximal cube* B of $C|dom(s)$ that contains the sample s and compresses s to the subsample determined by the dimensions set of that maximal cube. Note that the size of the dimension set (and the compression scheme) is bounded by the VC dimension.

How should we reconstruct? Consider all concepts of C that are consistent with the sample s:

$$H_s = \{h \in C : h|dom(s) = s\}.$$

Correctness means that we need to reconstruct to one of those concepts. Let s' be the input for the reconstruction function and let $D := dom(s')$. During the reconstruction, the domain $dom(s)$ of the original sample s is not known. All that is known at this point is that D is the dimensions set of a maximal cube B of $C|dom(s)$ that contained the sample s. The reconstruction map of the algorithm outputs a concept in the following set H_B (Fig. 3):

$$\{h \in C : h \text{ in cube } B' \text{ of } C \text{ s.t. } \dim(B') = \dim(B) \text{ and } h|\dim(B) = s|\dim(B)\}.$$

For the correctness of the compression scheme it suffices to show that for all choices of the maximal cube B of $C|dom(s)$, H_B is non-empty and a subset of H_s. The following Lemma guarantees the non-emptiness.

Lemma 2. *Let C be an extremal class and let $D \subseteq dom(C)$ be the dimensions set of some cube of $C|dom(s)$. Then D is also the dimensions set of some cube of C.*

Proof. Clearly the dimension set D is shattered by $C|dom(s)$ and therefore it is also shattered by C. By the extremality of C, D is also strongly shattered by it, and thus there exists a cube B of C with dimensions set D.

The second lemma show that for each choice of the maximal cube B, $H_B \subseteq H_s$.

Algorithm 1. (Labeled compression scheme for any extremal classes C)

The compression map.

- Input: A sample s of C.
- Output: A subsample $s' = s|\dim(B)$, where B is any maximal cube of $C|dom(s)$ that contains the sample s.

The reconstruction map.

- Input: A sample s' of size at most $VCdim(C)$.
- Output: Any concept h which is consistent with s' on $dom(s')$ and belongs to a cube B of C with dimensions set $dom(s')$.

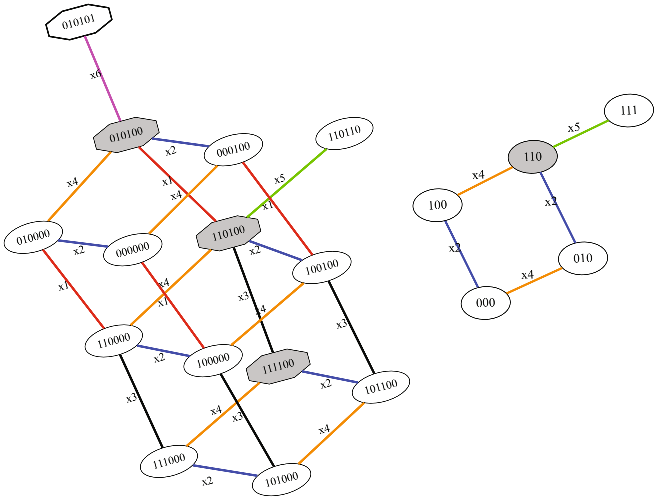

Fig. 3. The one-inclusion graph of an extremal concept class C is given on the left. Consider the sample $s = \overset{x_2 x_4 x_5}{\mathbf{1\,1\,0}}$. There are 4 concepts $c \in C$ consistent with this sample (the octagonal vertices), i.e. $H_s = \{1\mathbf{1}1\mathbf{10}0, 1\mathbf{1}0\mathbf{10}0, 0\mathbf{1}0\mathbf{10}0, 0\mathbf{1}0\mathbf{10}1\}$. There are 2 maximal cubes of $C|dom(s)$ (graph on right) that contain the sample s (in grey) with dimension sets $\{x_5\}$ and $\{x_2, x_4\}$, respectively. Let B be the maximal cube with dimension set $D = \{x_2, x_4\}$. There are 3 cubes of C (on left) with the same dimension set D. Each contains a concept h (shaded grey) that is consistent with the original sample on D, i.e. $h|D = s|D = \overset{x_2 x_4}{\mathbf{1\,1}}$ and therefore $H_B = \{1\mathbf{1}1\mathbf{1}00, 1\mathbf{1}0\mathbf{1}00, 0\mathbf{1}0\mathbf{1}00\}$. For the correctness we need that H_B (grey nodes on left) is non-empty and a subset of H_s (octagon nodes on left). Note that in this case H_B is a strict subset.

Lemma 3. *Let s be a sample of an extremal class C, let B be any maximal cube of $C|dom(s)$ that contains s, and let D denote the dimensions set of B. Then for any cube B' of C with $\dim(B') = D$, the concept $h \in B'$ that is consistent with s on D is also consistent with s on $dom(s) \setminus D$.*

Proof. Since B is a cube with dimensions set D, $B|(dom(s) \setminus D)$ contains the single concept $tag(B)$.

Let B' be any cube of C with $\dim(B') = D$, and let h be the concept in B' which is consistent with s on D. Now consider the cube $B'|dom(s)$. We will show that $B'|dom(s) = B$. This will finish the proof as it shows that both $h|dom(s)$ and s belong to $B'|dom(s) = B$ which means that $tag(B) = h|(dom(s) \setminus D) = s|(dom(s) \setminus D)$. Moreover, by the definition of h, $h|D = s|D$, and therefore $h|dom(s) = s$ as required.

We now show that $B'|dom(s) = B$. Indeed, since B' is a cube of C with dimension set $D \subseteq dom(s)$, the cube $B'|dom(s)$ is a cube of $C|dom(s)$ with the same dimension set D. Thus the dimension set of $B'|dom(s)$ contains the dimension set of the maximal cube B of $C|dom(s)$. Therefore, since $C|dom(s)$ is extremal (Theorem 3) it follows by Lemma 1 that $B'|dom(s) = B$.

4 Unlabeled Sample Compression Schemes and Combinatorial Conjectures

Alternate “unlabeled” compression schemes have also been found for maximum classes and a natural question is whether these schemes again generalize to extremal classes. As we shall see there is an excellent match between the combinatorics of unlabeled compression schemes and extremal classes. The existence of such schemes remains open at this point. We can however relate their existence to some natural conjectures about extremal classes.

An unlabeled compression schemes compresses a sample s of the concept class C to an (unlabeled) subset of the domain of the sample s. In other words, in an unlabeled compression scheme the labels of the original sample are not used by the reconstruction map. The size of the compression scheme is now the maximum size of the subset that the sample is compressed to. Consider an unlabeled compression scheme for C of size $VCdim(C)$. For a moment restrict your attention to samples of C over some fixed domain $S \subseteq dom(C)$. Each such sample is a concept in the restriction $C|S$. Note that two different concepts in $C|S$ must be compressed to different subsets of S, otherwise if they were compressed to the same subset, the reconstruction of it would not be consistent with one of them. For maximum classes, the number of concepts in $C|S$ is exactly the number of subsets of S of size up to the VC dimension. Intuitively, this “tightness” makes unlabeled compression schemes combinatorially rich and interesting.

Previous unlabeled compression schemes for maximum classes were based on “representation maps”; these are one-to-one mappings between C and subsets of $dom(C)$ of size at most $VCdim(C)$. Representation maps were used in the following way: each sample s is compressed to a subset of $dom(s)$ which represents a consistent hypothesis with s, and each subset of size at most $VCdim(C)$ of $dom(C)$ is reconstructed to the hypothesis it represents. The key combinatorial property that enabled finding representation maps for maximum classes was a “non clashing” condition [17]. This property was used to show that for any sample s of C there is exactly one concept c that is consistent with s and $r(c) \subseteq dom(s)$. This immediately implies an unlabeled compression scheme based on non clashing representation maps: Compress to the unique subset of the domain of the sample that represents a concept consistent with the given sample.

The first representation maps for maximum classes were derived via a recursive construction [17]. Alternate representation maps were also proposed in [17] based on a certain greedy “peeling” algorithm that iteratively assigns a representation to a concept and removes this concept from the class. The correctness of the representation maps based on peeling was finally established in [28].

Representation Maps. For any concept class C a *representation map* is any one-to-one mapping from concepts to subsets of the domain, i.e. $r : C \rightarrow \mathcal{P}(dom(C))$. We say that $c \in C$ *is represented* by the *representation set* $r(c)$. Furthermore we say that two different concepts c, c' *clash* with respect to r if they are consistent with each other on the union of their representation sets, i.e. $c|\,(r(c) \cup r(c')) = c'|\,(r(c) \cup r(c'))$. If no two concepts clash then we say that r is *non clashing.*

Example 6 **(Non Clashing Representation Map for Distance Preserving Classes).** Let C be a distance preserving class, that is for every $u, v \in C$, the distance between u, v in the one-inclusion graph of C equals to their hamming distance. For every $c \in C$, define $deg_C(c) = \{x \in dom(C) : c \text{ is incident to an } x\text{-edge of } C\}$. The representation map $r(c) := deg_C(c)$ has the property that for every $c \neq c' \in C$, c and c' disagree on $r(c)$. To see this, note that any shortest path from c to c' in C traverses exactly the dimensions on which c and c' disagrees. In particular, the first edge leaving c in this path traverses a dimension x for which $c(x) \neq c'(x)$. By the definition of $deg_C(c)$ we have that $x \in deg_C(c)$ and indeed c and c' disagree on $deg_C(c)$.

In fact, this gives a stronger property for distance preserving classes, which is summarized in the following lemma. This lemma will be useful in our analysis.

Lemma 4. *Let C be a distance preserving class and let $c \in C$. Then $deg_C(c)$ is a teaching set for c with respect to C. That is, for all $c' \in C$, such that $c' \neq c$ there is $x \in deg_C(c)$ such that $c(x) \neq c'(x)$.*

Clearly the representation map $r(c) = deg_C(c)$ is non clashing. The following lemma establishes that certain non clashing representation maps immediately give unlabeled compression schemes:

Algorithm 2. (Unlabeled compression scheme from a representation map)

The compression map.

- Input: A sample s of C.
  Let $c \in C$ be the unique concept which satisfies $c|dom(s) = s$, and $r(c) \subseteq dom(s)$.
- Output $r(c)$.

The reconstruction map.

- Input: a set $S' \in st(C)$.
  Since r is a bijection between C and $st(C)$, there is a unique c such that $r(c) = S'$.
- Output c.

Lemma 5. *Let r be any representation map that is a bijection between an extremal class C and $st(C)$. Then the following two statements are equivalent: (i) r is non clashing. (ii) For every sample s of C, there is exactly one concept $c \in C$ that is consistent with s and $r(c) \subseteq dom(s)$.*

A proof is given in the full version of this paper [25].

Based on this lemma it is easy to see that a representation mapping r for an extremal concept class C defines a compression scheme as follows (See Algorithm 2). For any sample s of C we *compress* s to the unique representative $r(c)$ such that c is consistent with s and $r(c) \subseteq dom(s)$. Reconstruction is even simpler, since r is bijective: If s is compressed to the set $r(c)$, then we reconstruct $r(c)$ to the concept c.

Corner Peeling Yields Good Representation Maps. We now present a natural conjecture concerning extremal classes and relate it to the construction of non clashing representation maps. A concept c of an extremal class C is a *corner* of C if $C \setminus \{c\}$ is extremal. By Lemma 1 we have that for each $S \subseteq dom(C)$ there is at most one maximal cube with dimension set S and if S. Therefore $\mathrm{st}(C \setminus \{c\}) = \mathrm{st}(C) \setminus \{\dim(B) : B \text{ is maximal cube of } C \text{ containing } c\}$.

For $C \setminus \{c\}$ to be extremal, $|\,\mathrm{st}(C \setminus \{c\})|$ must be $|C| - 1$ (by Theorem 2) and therefore c is a corner of an extremal class C iff c lies in exactly one maximal cube of C.

Conjecture 1. Every non empty extremal class C has at least one corner.

A related conjecture for maximum classes was presented in [17]. For these latter classes, the conjecture was finally proved in [28]. This conjecture also has been proven for other special cases such as extremal classes of VC dimension at most 2 [19,22].

In fact [19] proved a stronger statement: For every two extremal classes $C_1 \subseteq C_2$ such that $VCdim(C_2) \leq 2$ and $|C_2 \setminus C_1| \geq 2$, there exists an extremal class C such that $C_1 \subset C \subset C_2$ (i.e. C is a strict subset of C_2 and a strict superset of C_1). Indeed, this statement is stronger as by repeatedly picking a larger extremal class $C_1 \subseteq C_2$ eventually a $c \in C_2$ is obtained such that $C_2 - \{c\}$ is extremal. For general extremal classes this stronger statement also remains open.

Conjecture 2. For every two extremal classes $C_1 \subseteq C_2$ with $|C_2 \setminus C_1| \geq 2$ there exists an extremal class C such that $C_1 \subset C \subset C_2$.

How does Conjecture 1 yield a representation map? Define an order[4] $c_1 \ldots c_{|C|}$ on the concept clase C such that for every i, c_i is a corner of $C_i = \{c_j : j \geq i\}$, and define a map $r : C \rightarrow st(C)$ such that $r(c_i) = \dim(B_i)$ where B_i is the unique maximal cube of C_i that c_i belongs to. We claim that r is a representation map. Indeed, r is a one-to-one mapping from C to $st(C)$ (and since C is extremal r is a bijection). To see that r is non clashing, note that $r(c_i) = \dim(B_i) = \deg_{C_i}(c_i)$. C_i is extremal and therefore distance preserving (Theorem 4). Thus, Lemma 4 implies that $r(c_i)$ is a teaching set of c_i with respect to C_i. This implies that r is indeed non clashing.

[4] Such orderings are related to the recursive teaching dimension which was studied by [10].

5 Discussion

We studied the conjecture of [12] which asserts that every concept classes has a sample compression scheme of size linear in its VC dimension. We extended the family of concept classes for which the conjecture is known to hold by showing that every extremal class has a sample compression scheme of size equal to its VC dimension. We demonstrated that extremal classes form a natural and rich generalization of maximum classes for which the conjecture had been proved before [12], and further related basic conjectures concerning the combinatorial structure of extremal classes with the existence of optimal unlabeled compression schemes. These connections may also be used in the future to provide a better understanding on the combinatorial structure of extremal classes, which is considered to be incomplete by several authors [6,15].

Our compression schemes for extremal classes yield another direction of attacking the general conjecture of Floyd and Warmuth: it is enough to show that an arbitrary maximal concept class of VC dimension d can be covered by $\exp(d)$ extremal classes of VC dimension $O(d)$. Note it takes additional $O(d)$ bits to specify which of the $\exp(d)$ extremal classes is used in the compression.

Acknowledgements. We thank Michał Dereziński for a good feedback on the writing of the paper and Ami Litman for helpful combinatorial insights.

References

1. Anstee, R., Rónyai, L., Sali, A.: Shattering news. Graphs Comb. **18**(1), 59–73 (2002)
2. Bandelt, H., Chepoi, V., Dress, A., Koolen, J.: Combinatorics of lopsided sets. Eur. J. Comb. **27**(5), 669–689 (2006)
3. Ben-David, S., Litman, A.: Combinatorial variability of Vapnik-Chervonenkis classes with applications to sample compression schemes. Discret. Appl. Math. **86**(1), 3–25 (1998)
4. Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K.: Learnability and the Vapnik-Chervonenkis dimension. J. Assoc. Comput. Mach. **36**(4), 929–965 (1989)
5. Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K.: Occam's razor. Inf. Process. Lett. **24**(6), 377–380 (1987)
6. Bollobás, B., Radcliffe, A.J.: Defect Sauer results. J. Comb. Theory Ser. A **72**(2), 189–208 (1995)
7. Bollobás, B., Radcliffe, A.J., Leader, I.: Reverse Kleitman inequalities. Proc. Lond. Math. Soc. Ser. A (3) **58**, 153–168 (1989)
8. Chernikov, A., Simon, P.: Externally definable sets and dependent pairs. Isr. J. Math. **194**(1), 409–425 (2013)
9. Daniely, A., Shalev-Shwartz, S.: Optimal learners for multiclass problems. In: COLT, pp. 287–316 (2014)
10. Doliwa, T., Simon, H.U., Zilles, S.: Recursive teaching dimension, learning complexity, and maximum classes. In: Hutter, M., Stephan, F., Vovk, V., Zeugmann, T. (eds.) ALT 2015. Lecture Notes in Artificial Intelligence (LNAI), vol. 6331, pp. 209–223. Springer, Heidelberg (2010). doi:10.1007/978-3-642-16108-7_19

11. Dress, A.: Towards a theory of holistic clustering. DIMACS Ser. Discret. Math. Theoret. Comput. Sci. **37**, 271–289 (1997). (Amer. Math. Soc.)
12. Floyd, S., Warmuth, M.K.: Sample compression, learnability, and the Vapnik-Chervonenkis dimension. Mach. Learn. **21**(3), 269–304 (1995)
13. Freund, Y., Schapire, R.E.: Boosting: Foundations and Algorithms. Adaptive Computation and Machine Learning. MIT Press, Cambridge (2012)
14. Gartner, B., Welzl, E.: Vapnik-Chervonenkis dimension and (pseudo-)hyperplane arrangements. Discret. Comput. Geom. (DCG) **12**, 399–432 (1994)
15. Greco, G.: Embeddings and the trace of finite sets. Inf. Process. Lett. **67**(4), 199–203 (1998)
16. Kozma, L., Moran, S.: Shattering, graph orientations, and connectivity. Electron. J. Comb. **20**(3), P44 (2013)
17. Kuzmin, D., Warmuth, M.K.: Unlabeled compression schemes for maximum classes. J. Mach. Learn. Res. **8**, 2047–2081 (2007)
18. Lawrence, J.: Lopsided sets and orthant-intersection by convex sets. Pac. J. Math. **104**(1), 155–173 (1983)
19. Litman, A., Moran, S.: Unpublished results (2012)
20. Littlestone, N., Warmuth, M.: Relating data compression and learnability (1986, Unpublished)
21. Livni, R., Simon, P.: Honest compressions and their application to compression schemes. In: COLT, pp. 77–92 (2013)
22. Mészáros, T., Rónyai, L.: Shattering-extremal set systems of VC dimension at most 2. Electron. J. Comb. **21**(4), P4.30 (2014)
23. Moran, S.: Shattering-extremal systems (2012). CoRR abs/1211.2980
24. Moran, S., Shpilka, A., Wigderson, A., Yehudayoff, A.: Teaching and compressing for low VC-dimension. In: ECCC TR15-025 (2015)
25. Moran, S., Warmuth, M.K.: Labeled compression schemes for extremal classes (2015). CoRR abs/1506.00165. http://arXiv.org/abs/1506.00165
26. Moran, S., Yehudayoff, A.: Sample compression schemes for VC classes. J. ACM **63**(3), 21:1–21:10 (2016)
27. Pajor, A.: Sous-espaces l_1^n des espaces de banach. Travaux en Cours. Hermann, Paris (1985)
28. Rubinstein, B.I.P., Rubinstein, J.H.: A geometric approach to sample compression. J. Mach. Learn. Res. **13**, 1221–1261 (2012)
29. Samei, R., Yang, B., Zilles, S.: Generalizing labeled and unlabeled sample compression to multi-label concept classes. In: Auer, P., Clark, A., Zeugmann, T., Zilles, S. (eds.) ALT 2015. Lecture Notes in Artificial Intelligence (LNAI), vol. 8776, pp. 275–290. Springer, Heidelberg (2014). doi:10.1007/978-3-319-11662-4_20
30. Sauer, N.: On the density of families of sets. J. Comb. Theory Ser. A **13**, 145–147 (1972)
31. Shelah, S.: A combinatorial problem; stability and order for models and theories in infinitary languages. Pac. J. Math. **41**, 247–261 (1972)
32. Valiant, L.: A theory of the learnable. Commun. ACM **27**, 1134–1142 (1984)
33. Warmuth, M.K.: Compressing to VC dimension many points. In: Schölkopf, B., Warmuth, M.K. (eds.) COLT/Kernel 2003. LNCS (LNAI), vol. 2777, pp. 743–744. Springer, Heidelberg (2003)