Order Compression Schemes*

Malte Darnstädt[1], Thorsten Doliwa[1], Hans Ulrich Simon[1], and Sandra Zilles[2]

[1] Fakultät für Mathematik, Ruhr-Universität Bochum
D-44780 Bochum, Germany
{malte.darnstaedt,thorsten.doliwa,hans.simon}@rub.de
[2] Department of Computer Science, University of Regina
Regina, SK, Canada S4S 0A2
zilles@cs.uregina.ca

Abstract. Sample compression schemes are schemes for "encoding" a set of examples in a small subset of examples. The long-standing open sample compression conjecture states that, for any concept class $\mathcal{C}$ of VC-dimension d, there is a sample compression scheme in which samples for concepts in $\mathcal{C}$ are compressed to samples of size at most d.
We show that every order over $\mathcal{C}$ induces a special type of sample compression scheme for $\mathcal{C}$, which we call order compression scheme. It turns out that order compression schemes can compress to samples of size at most d if $\mathcal{C}$ is maximum, intersection-closed, a Dudley class, or of VC-dimension 1–and thus in most cases for which the sample compression conjecture is known to be true.
Since order compression schemes are much simpler than sample compression schemes in general, their study seems to be a promising step towards resolving the sample compression conjecture. We reveal a number of fundamental properties of order compression schemes, which are helpful in such a study. In particular, order compression schemes exhibit interesting graph-theoretic properties as well as connections to the theory of learning from teachers.

1 Introduction

In the context of concept learning, sample compression schemes are schemes for "encoding" a set of examples in a small subset of examples. For instance, from the set of examples they process, learning algorithms often extract a subset of particularly "significant" examples in order to represent their hypotheses. This way sample bounds for PAC-learning of a concept class $\mathcal{C}$ can be obtained from the size of a smallest sample compression scheme for $\mathcal{C}$ [12, 8]. The size of a sample compression scheme is the size of the largest subset resulting from compressing any sample consistent with some concept in $\mathcal{C}$. In what follows, we will use the term *sample compression number* of a concept class $\mathcal{C}$ to refer to the smallest possible size of a sample compression scheme for $\mathcal{C}$.

* This work was partly supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Littlestone and Warmuth [12] pointed out that their sample bounds for PAC-learning, stated in terms of the sample compression number, would improve on the classical bounds stated in terms of the VC-dimension (see [3]), if one could show that the sample compression number is linear in the VC-dimension. The precise relationship between these two combinatorial parameters is hence of great interest to the learning theory community, as witnessed by a continuing series of publications on the topic, see, e.g., [2, 4, 5, 8, 11, 12, 14]. Floyd and Warmuth [8] conjectured that any concept class $\mathcal{C}$ of VC-dimension d (abbreviated by $\mathrm{VCD}(\mathcal{C}) = d$) possesses a sample compression scheme of size d, but to date this conjecture (called the sample compression conjecture) remains open. It is even open whether or not the sample compression number is linear in the VC-dimension.

While this paper does not resolve the sample compression conjecture, it offers a potential new stepping stone for its solution. We demonstrate that any total order over a finite concept class[3] $\mathcal{C}$ induces a sample compression scheme for $\mathcal{C}$; we call such schemes *order compression schemes*. The principle behind the construction of this special kind of compression scheme is very simple; in particular, it does not fully exploit the available options for "coding tricks". Consequently, order compression schemes are not in general optimal in terms of size, i.e., for some concept classes the smallest possible size for an order compression scheme exceeds the sample compression number. Despite their "suboptimality", order compression schemes exhibit properties that make them a very promising object of study for future research on the sample compression conjecture:

– Every finite concept class has order compression schemes, and thus our notion is generally applicable.

– Almost all finite concept classes $\mathcal{C}$ for which the sample compression conjecture is proven, i.e., that have a sample compression scheme of size $\mathrm{VCD}(\mathcal{C})$, have an order compression scheme of size $\mathrm{VCD}(\mathcal{C})$. In particular, we prove the existence of order compression schemes of size VC-dimension for maximum classes (classes of the largest possible size for a given VC-dimension [15, 17]), for intersection-closed classes, for Dudley classes [6], and for classes of VC-dimension 1. This suggests that there might be further well-structured concept classes for which order compression schemes witness the correctness of the sample compression conjecture.

– If the sample compression conjecture is true, then a helpful step towards its proof could be to gain a deeper understanding of the circumstances under which order compression schemes are not optimal in terms of size. In particular, it could be worth analyzing which "coding tricks" beyond those systematically exploited by order compression schemes can make general compression schemes more powerful.

– If the sample compression conjecture is false, then a helpful step towards disproving it could be an attempt to show that the size of the smallest possible

[3] In this paper, we focus on finite concept classes. This is not a restriction, since it is proven that the sample compression conjecture is true if and only if it is true for all finite concept classes [2].

order compression scheme is not linear in the VC-dimension. Firstly, due to the simplicity of the construction of order compression schemes, this should be less challenging than disproving the general conjecture. Secondly, the results obtained and the proof techniques used could potentially yield new insights into the theory behind the general conjecture.
– The definition of order compression schemes exhibits similarities to a model of learning from helpful teachers, namely the model of recursive teaching [19]. The central complexity parameter delimiting the number of examples required for learning in this model is the *recursive teaching dimension*. By proving that the recursive teaching dimension lower-bounds the smallest possible size of order compression schemes, we establish a new connection between teaching and sample compression. Hence, known results on the recursive teaching model may provide useful tools for the study of order compression schemes.
– Order compression schemes also exhibit interesting graph-theoretic properties. We define a graph representation of general sample compression schemes and prove that order compression schemes correspond to exactly the acyclic graphs in this definition.

2 Preliminaries

Throughout this paper, $X = \{x_1, \ldots, x_n\}$ is a set of cardinality $n \in \mathbb{N}$, called the instance space, and $\mathcal{C}, \mathcal{H} \subseteq \mathcal{P}(X)$ denote concept resp. hypothesis classes over the domain X. For $X' \subseteq X$, we define $\mathcal{C}_{|X'} := \{C \cap X' |\ C \in \mathcal{C}\}$. We treat concepts and hypotheses interchangeably as subsets of X and as $0,1$-valued functions on X. A labeled example is a pair (x, l) with $x \in X$ and $l \in \{0, 1\}$. A *labeled sample* is a set of labeled examples. For every labeled sample S, we define $X(S) = \{x \in X \mid (x, 0) \in S \text{ or } (x, 1) \in S\}$. Further, $\text{Cons}(S, \mathcal{H})$ is the set of all hypotheses in $\mathcal{H}$ that are consistent with S, i.e., $\text{Cons}(S, \mathcal{H}) = \{H \in \mathcal{H} \mid H(x) = l \text{ for all } (x, l) \in S\}$. S is called *$\mathcal{C}$-realizable* if $\text{Cons}(S, \mathcal{C}) \neq \emptyset$. The VC-dimension of the class $\mathcal{C}$, denoted by $\text{VCD}(\mathcal{C})$, is defined as the cardinality of the largest $X' \subseteq X$ such that $\mathcal{C}_{|X'}$ is the power set of X'.

2.1 Sample Compression

A *sample compression scheme* [12] of size k for the class $\mathcal{C}$ consists of a compression function f and a reconstruction function g. The compression function f maps every $\mathcal{C}$-realizable sample S to a subset of size at most k, called compression set. The reconstruction function g maps any compression set $f(S)$ to a hypothesis $g(f(S)) \subseteq X$, where

$$\forall\ (x, l) \in S :\ g(f(S))(x) = l\ \ ,$$

i.e., $g(f(S))$ is consistent with the original sample set S. The set $\mathcal{H} = \{g(f(S)) \mid S \text{ is a } \mathcal{C}\text{-realizable sample}\}$ consists of all hypotheses that are used by the compression scheme. Obviously, the hypothesis class $\mathcal{H}$ must be at least as powerful as $\mathcal{C}$, i.e., $\mathcal{C} \subseteq \mathcal{H}$. A sample compression scheme fulfilling $\mathcal{H} = \mathcal{C}$ is called

proper. We often write "$(\mathcal{C},\mathcal{H})$-scheme" as an abbreviation of "sample compression scheme for $\mathcal{C}$ using hypotheses from $\mathcal{H}$".

Let (f,g) represent a scheme. We say that two labeled samples S', S'' are *g-equivalent* if $g(S') = g(S'')$. We can certainly normalize schemes according to the following policy: $S' = f(S)$ is always chosen as a subset of S of minimal size among all subsets of S that are g-equivalent to S'. We will henceforth implicitly assume that a scheme is normalized in this sense.

An important open question is whether any class $\mathcal{C}$ has a sample compression scheme of size linear in $\mathrm{VCD}(\mathcal{C})$, or even of size equal to $\mathrm{VCD}(\mathcal{C})$ [8].

2.2 Teaching

Following Goldman and Kearns [9] and Shinohara and Miyano [16], a set S of labeled examples is called a *teaching set* for a concept $C \in \mathcal{C}$ (with respect to $\mathcal{C}$) if C is the only concept in $\mathcal{C}$ that is consistent with S, i.e., if $\mathrm{Cons}(S,\mathcal{C}) = \{C\}$. By $\mathcal{TS}(C,\mathcal{C})$ we denote the set of all teaching sets for C with respect to $\mathcal{C}$. $\mathrm{TD}(C,\mathcal{C}) = \min\{|S| \mid S \in \mathcal{TS}(C,\mathcal{C})\}$ denotes the smallest possible size of a teaching set for C with respect to $\mathcal{C}$. According to Zilles et al. [19], a *teaching plan* for a class $\mathcal{C} = \{C_1,\ldots,C_m\}$ of cardinality m is a sequence

$$P = ((C_1,S_1),\ldots,(C_m,S_m))$$

in which $S_t \in \mathcal{TS}(C_t,\{C_t,\ldots,C_m\})$ for all $t = 1,\ldots,m$. The *order of P* is defined as $\mathrm{ord}(P) = \max\{|S_t| \mid 1 \le t \le m\}$. The *recursive teaching dimension* of $\mathcal{C}$, denoted by $\mathrm{RTD}(\mathcal{C})$, is the minimum order over all teaching plans of $\mathcal{C}$, and is always witnessed by a teaching plan $P = ((C_1,S_1),\ldots,(C_m,S_m))$ for $\mathcal{C}$ in which, for all $t \in \{1,\ldots,m\}$,

- C_i is chosen from $\mathcal{C}_t = \{C_t, C_{t+1},\ldots,C_m\}$ such that $\mathrm{TD}(C_t,\mathcal{C}_t) = \min_{C\in\mathcal{C}_t}(\mathrm{TD}(C,\mathcal{C}_t))$.
- S_t is a smallest possible teaching set for C_t with respect to $\mathcal{C}_t$.

The notion

$$\mathrm{RTD}^*(\mathcal{C}) = \max_{X'\subseteq X}(\mathrm{RTD}(\mathcal{C}_{|X'}))$$

was introduced by Doliwa et al. [5].

3 Properties of Order Compression Schemes

We will now introduce *order compression schemes*, a notion that is inspired by Fan's work [7] and that is central to this paper:

Definition 1. *Let $\mathcal{H} = \{H_1,\ldots,H_m\}$ be a hypothesis class over X and let $\mathcal{C} \subseteq \mathcal{H}$. Let $<$ be a total order on $\mathcal{H}$, say $H_1 < H_2 < \ldots < H_m$. An* order compression scheme *for $(\mathcal{C},\mathcal{H})$ with respect to $<$ is a pair (f,g) of mappings that satisfies the following properties for all $\mathcal{C}$-realizable samples S:*

1. *Let t be largest number such that H_t is consistent with S. Then $f(S)$ is a smallest subset of S that is a teaching set for H_t with respect to $\{H_t, \ldots, H_m\}$.*
2. *Let t be the largest number such that H_t is consistent with $f(S)$. Then $g(f(S)) = H_t$.*

Because the definition of f and g is constructive it is clear that any order over a hypothesis class induces an order compression scheme.

The similarly obvious observation, that the t in the first and second part of Definition 1 must be the same number, shows that any such scheme is indeed a compression scheme:

Proposition 1. *Let $\mathcal{H} = \{H_1, \ldots, H_m\}$ be any hypothesis class over X and $\mathcal{C} \subseteq \mathcal{H}$. Then any order compression scheme using hypotheses from $\mathcal{H}$ is a $(\mathcal{C}, \mathcal{H})$-scheme.*

Since order compression schemes are compression schemes, we are particularly interested in their size:

Definition 2. *The order compression number of a pair $(\mathcal{C}, \mathcal{H})$, denoted by $\mathrm{OCN}(\mathcal{C}, \mathcal{H})$, is the minimum size of an order $(\mathcal{C}, \mathcal{H})$-scheme (where the minimum is taken over all total orderings of $\mathcal{H}$).*

The following example shows that non-proper order compression schemes can be greatly superior to proper ones. In particular, there are concept classes for which the smallest possible non-proper order compression scheme is of size 1, while the smallest possible proper schemes can be arbitrarily large:

Example 1. Let $\mathcal{C} = \{\{0\}, \ldots, \{n-1\}\}$ be the class of singletons, and let $\mathcal{H} = \mathcal{C} \cup \{\emptyset\}$. The improper order compression scheme with respect to the ordering $\{0\} < \ldots < \{n-1\} < \emptyset$ is easily seen to be of size 1:

- A $\mathcal{C}$-realizable sample S including a positive example, say $(k, 1)$, is compressed to $\{(k, 1)\}$.
- All other samples are compressed to $\emptyset$.

Note that $\{(k, 1)\}$ is a teaching set for $\{k\}$ with respect to $\mathcal{H}$, and $\emptyset$ is a teaching set for $\emptyset$ with respect to $\{\emptyset\}$. Thus, the described compression mapping respects the policy of order compression schemes. The compressed sets are of size at most 1. We thus obtain $\mathrm{OCN}(\mathcal{C}, \mathcal{H}) = 1$.

By contrast, consider proper order compression schemes for $\mathcal{C}$. For reasons of symmetry, we may assume that $\{0\} < \ldots < \{n-1\}$ is the underlying ordering. The sample $S = \{(1, 0), \ldots, (n-1, 0)\}$ is a teaching set for $\{0\}$. However, any compression to a proper subsample would be decompressed to some $\{i\}$ such that $i > 0$, which is an inconsistency to $(i, 0) \in S$. It follows that $\mathrm{OCN}(\mathcal{C}, \mathcal{C}) = n - 1$.

It should be noted that general compression schemes for the class $\mathcal{C}$ of singletons can be made proper and of size 1 at the same time:

- A sample S including a positive example is compressed as described above for order compression schemes.

– A non-empty $\mathcal{C}$-realizable sample S not including positive examples can be of size at most $n-1$. Let $k \in \{0,\ldots,n-1\}$ be an index such that k occurs in S but $k+1 \bmod n$ does not. Then S is compressed to $\{(k,0)\}$ (resolving ambiguities in favor of smaller indexes).

Clearly, $\{(k,1)\}$ is decompressed to $\{k\}$, and $\{(k,0)\}$ is decompressed to $\{k+1 \bmod n\}$. We obtain a proper compression scheme of size 1 for the class of singletons.

This example raises the question of what is the optimal choice for the hypothesis class $\mathcal{H} \supseteq \mathcal{C}$. The best choice for $\mathcal{H}$ leads us to the order compression number of a class $\mathcal{C}$ which is formally defined as follows:

Definition 3. *The order compression number of $\mathcal{C}$, denoted by $\mathrm{OCN}(\mathcal{C})$, is the minimum of $\mathrm{OCN}(\mathcal{C},\mathcal{H})$ over the choice of $\mathcal{H} \supseteq \mathcal{C}$.*

Theorem 1. *Let X denote the domain of the classes $\mathcal{H}$ and $\mathcal{C} \subseteq \mathcal{H}$, and let $X' \subseteq X$. Then, $\mathrm{OCN}(\mathcal{C},\mathcal{H}) \geq \mathrm{OCN}(\mathcal{C}_{|X'},\mathcal{H}_{|X'})$.*

Proof. Let $H_1 < \ldots < H_m$ be the ordering of $\mathcal{H} = \{H_1,\ldots,H_m\}$ such that the corresponding order compression scheme has size $\mathrm{OCN}(\mathcal{C},\mathcal{H})$. For $i = 1,\ldots,m$, let H'_i denote the restriction of H_i to X'. Note that $i \neq j$ does not necessarily imply $H'_i \neq H'_j$ since different hypotheses might coincide on X'. Let $m' \leq m$ denote the number of distinct restrictions. Pick indices $i(1) < \ldots < i(m')$ such that the sequence $H'_{i(1)},\ldots,H'_{i(m')}$ contains every restriction exactly once and, subject to this constraint, the indices $i(j)$ are chosen as large as possible, i.e., for every hypothesis from $\mathcal{H}_{|X'}$, we select the latest representative in the sequence $H_1,\ldots,H_m$. Consider now the order compression scheme for $(\mathcal{C}_{|X'},\mathcal{H}_{|X'})$ with $H'_{i(1)} < \ldots < H'_{i(m')}$ as the underlying ordering. Let S' be a $\mathcal{C}$-realizable sample over the restricted domain X', and let $t \in [m]$ be the largest index such that $H_t \in \mathrm{Cons}(S,\mathcal{H})$. The definition of order compression schemes implies that $f(S) \subseteq S \subseteq X'$ is a smallest teaching set for H_t with respect to $\{H_t,\ldots,H_m\}$. By the maximality of t, H_t is the latest representative of H'_t in the sequence $H_1,\ldots,H_m$ so that $t = i(\tau)$ for some $\tau \in [m']$. Thus, $\tau \in [m']$ is the largest index such that $H'_t = H'_{i(\tau)} \in \mathrm{Cons}(S,\mathcal{H}_{|X'})$. Clearly, $f(S) \subseteq S \subseteq X'$ is a teaching set for H'_t with respect to $\mathcal{H}'_\tau := \{H'_{i(\tau)},\ldots,H'_{i(m')}\}$. It follows that the size of the smallest teaching set for H'_t with respect to $\mathcal{H}'_\tau$ is bounded by $|S_t|$. We obtain an order $(\mathcal{C}_{|X'},\mathcal{H}_{|X'})$-scheme whose size is bounded from above by $\mathrm{OCN}(\mathcal{C},\mathcal{H})$. □

Corollary 1. *For every $X' \subseteq X$: $\mathrm{OCN}(\mathcal{C}) \geq \mathrm{OCN}(\mathcal{C}_{|X'})$.*

Proof. The result is obtained from Theorem 1 as follows:

$$\begin{aligned}\mathrm{OCN}(\mathcal{C}) &= \min_{\mathcal{H}} \mathrm{OCN}(\mathcal{C},\mathcal{H}) \geq \min_{\mathcal{H}} \mathrm{OCN}(\mathcal{C}_{|X'},\mathcal{H}_{|X'}) \\ &\geq \min_{\mathcal{H}'} \mathrm{OCN}(\mathcal{C}_{|X'},\mathcal{H}') = \mathrm{OCN}(\mathcal{C}_{|X'})\end{aligned}$$

□

A useful tool for analyzing order compression schemes in particular and compression schemes (f, g) in general is the "compression graph", a digraph that we introduce in

Definition 4. *Let (f, g) be a $(\mathcal{C}, \mathcal{H})$-scheme. The digraph $\mathcal{G}_{\text{comp}}(f, g) = (V, E)$, called* compression graph associated with (f, g)*, is given as follows:*

1. *V equals the set of hypotheses $\mathcal{H}$.*
2. *For any $H_1, H_2 \in \mathcal{H}$, $(H_1, H_2) \in E$ if there exists a $\mathcal{C}$-realizable labeled sample S such that both H_1 and H_2 are consistent with $f(S)$ and $g(f(S)) = H_2$.*

A compression scheme (f, g) is called acyclic if the induced compression graph is acyclic.

For illustration, Figure 1 shows the compression graphs of the compression schemes from Example 1 for the case $n = 4$.

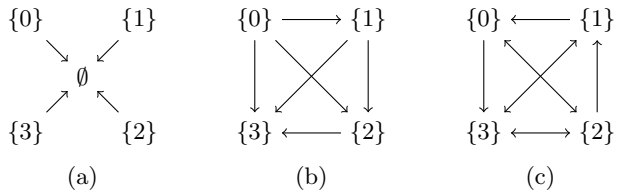


Fig. 1. The compression graphs of the compression schemes from Example 1 for the class $\mathcal{C} = \{\{0\}, \{1\}, \{2\}, \{3\}\}$: (a) results from the acyclic improper scheme of size 1 with $\mathcal{H} = \mathcal{C} \cup \{\emptyset\}$, (b) results from an acyclic proper scheme of size 3, and (c) results from the cyclic proper scheme of size 1 (assuming that the empty set is decompressed to $\{3\}$, which yields an additional edge from $\{2\}$ to $\{3\}$.)

The following result presents a useful characterization of order compression schemes, which we will exploit several times in Section 5:

Theorem 2. *For any $(\mathcal{C}, \mathcal{H})$-scheme (f, g), the following holds: the compression scheme (f, g) is acyclic iff (f, g) is an order compression scheme.*

Proof. Assume first that (f, g) is an order compression scheme. Let $H_1 < \ldots < H_m$ be the underlying ordering of the hypotheses from $\mathcal{H}$. Pick an arbitrary edge (H_i, H_j) of the compression graph G associated with (f, g). The definition of compression graphs implies that there exists a sample S such that S is realizable by $\mathcal{C}$, $H_i, H_j \in \text{Cons}(f(S), \mathcal{H})$ and $g(f(S)) = H_j$. The definition of order compression schemes implies that $j > i$. Thus, G is acyclic.
Assume now that the compression graph $G = (V, E)$ associated with (f, g) is acyclic. Let $H_1 < \ldots < H_m$ be a topological ordering of $V = \mathcal{H}$. Let S be an arbitrary sample that is realizable by $\mathcal{C}$, let $S' = f(S)$ and let $H_j = g(S')$. The definition of compression graphs implies that, for every $H_i \in \text{Cons}(S', \mathcal{H})$, the

edge (H_i, H_j) belongs to E. Since the hypotheses are ordered topologically, we may conclude that

$$j = \max\{i : \ H_i \in \mathrm{Cons}(S', \mathcal{H})\} \ . \tag{1}$$

This is precisely how decompression proceeds in an order compression scheme. It now suffices to show that the compression function f agrees with the definition of an order compression scheme too. To this end, let S be a sample that is realizable by $\mathcal{C}$, and let $t \in [m]$ be maximum such that $H_t \in \mathrm{Cons}(S, \mathcal{H})$. In particular, $H_{t+1}, \ldots, H_m$ are not consistent with S. Let $S' = f(S)$ and $H_j = g(S')$. According to the definition of schemes, H_j actually is consistent with S so that $H_j \in \{H_1, \ldots, H_t\}$. As already mentioned earlier in this proof, the definition of compression graphs implies that j satisfies (1). Since H_t is consistent with S, it is certainly consistent with $S' \subseteq S$ too. Since, as mentioned before, $H_j \in \{H_1, \ldots, H_t\}$, we may conclude that $g(f(S)) = H_j = H_t$. We remind the reader that we implicitly assume all compression functions f to pick subsets of S of minimal size among all g-equivalent ones. It follows that $f(S) = S'$ is a smallest subset of S whose g-image is H_t. Furthermore, since g acts like a decompression function of an order compression scheme, it follows that S', among all subsets of S, is a smallest teaching set for H_t with respect to $\{H_t, H_{t+1}, \ldots, H_m\}$. Since this is precisely how compression should proceed in order compression schemes, we are done. □

Note that the proof of Theorem 2 implies the following: the total orders on $\mathcal{H}$ that induce order compression schemes with an acyclic compression graph $G = (V, E)$ are precisely the topological orderings of V.

4 Order Compression Schemes and Teaching

The definition of order compression schemes bears some similarity to the model of *recursive teaching* [19], and the notion of order compression number hence is related to the complexity parameter of this teaching model, namely the recursive teaching dimension.

Let $\mathcal{C} \subseteq \mathcal{H} = \{H_1, \ldots, H_m\}$, and let $P = ((H_1, S_1), \ldots, (H_m, S_m))$ be a teaching plan for $\mathcal{H}$. Then P is called *realizable by* $\mathcal{C}$ if the samples $S_1, \ldots, S_m$ are realizable by $\mathcal{C}$. P is called *inclusion-minimal with respect to* $\mathcal{C}$ if P is realizable by $\mathcal{C}$ and, for every $t \in [m]$, there is no proper subset of S_t that is a teaching set for H_t with respect to $\{H_t, H_{t+1}, \ldots, H_m\}$. P is called a *maximal* $(\mathcal{C}, \mathcal{H})$*-plan (among the inclusion-minimal ones)* if, for every $t \in [m]$, S_t is of largest cardinality among all $\mathcal{C}$-realizable inclusion-minimal teaching sets for H_t with respect to $\{H_t, H_{t+1}, \ldots, H_m\}$.

Theorem 3. *There is an order* $(\mathcal{C}, \mathcal{H})$*-scheme of size* k *iff there is a maximal* $(\mathcal{C}, \mathcal{H})$*-plan of order* k*.*

Proof. Let (f, g) represent an order $(\mathcal{C}, \mathcal{H})$-scheme of size k, and let $H_1 < \ldots < H_m$ be the underlying ordering of $\mathcal{H} = \{H_1, \ldots, H_m\}$. According to the definition

of $(\mathcal{C}, \mathcal{H})$-schemes, there must exist a sample S_t such that S_t is realizable by $\mathcal{C}$ and $g(f(S_t)) = H_t$. According to the definition of order compression schemes, t is maximum subject to $H_t \in \mathrm{Cons}(f(S_t), \mathcal{H})$. Thus, $f(S_t)$ is a teaching set for H_t with respect to $\{H_t, H_{t+1}, \ldots, H_m\}$ (and $f(S_t)$ is inclusion-minimal with this property by our implicit assumption of dealing with normalized schemes only). Since $f(S_t) \subseteq S_t$ and S_t is realizable by $\mathcal{C}$, $f(S_t)$ is realizable by $\mathcal{C}$ too. It follows from this discussion that the teaching sets $f(S_t)$ such that $g(f(S_t)) = H_t$ represent a teaching plan that is inclusion-minimal with respect to $\mathcal{C}$. In order to get a maximal $(\mathcal{C}, \mathcal{H})$-plan, we proceed as described above, with the following exception: S_t such that $g(f(S_t)) = H_t$ is not chosen arbitrarily but as a set of maximal size among all $\mathcal{C}$-realizable inclusion-minimal teaching sets for H_t with respect to $\{H_t, H_{t+1}, \ldots, H_m\}$. In this case, $f(S_t) = S_t$, and we obtain a maximal $(\mathcal{C}, \mathcal{H})$-plan of order k.

Suppose now that $P = ((H_1, S_1), \ldots, (H_m, S_m))$ is a maximal $(\mathcal{C}, \mathcal{H})$-plan of order k. Consider the order $(\mathcal{C}, \mathcal{H})$-scheme with $H_1 < \ldots < H_m$ as the underlying ordering. Let S be a $\mathcal{C}$-realizable sample and let t be maximum subject to $H_t \in \mathrm{Cons}(S, \mathcal{H})$. Then S is a $\mathcal{C}$-realizable teaching set for H_t with respect to $\{H_t, H_{t+1}, \ldots, H_m\}$. Recall that order compression maps S to a smallest $S' \subseteq S$ that is still a teaching set for H_t with respect to $\{H_t, H_{t+1}, \ldots, H_m\}$. Then clearly $|S'| \leq |S_t| \leq k$ because S_t is the teaching set for H_t with respect to $\{H_t, H_{t+1}, \ldots, H_m\}$ taken from a maximal $(\mathcal{C}, \mathcal{H})$-plan P. The discussion shows that the order $(\mathcal{C}, \mathcal{H})$-scheme with $H_1 < \ldots < H_m$ as the underlying ordering is of size k. □

Theorem 3 leads to the following lower bound on OCN.

Lemma 1. *For every concept class $\mathcal{C}$:* $\mathrm{OCN}(\mathcal{C}) \geq \mathrm{RTD}(\mathcal{C})$.

Proof. Choose $\mathcal{H}$ such that $\mathrm{OCN}(\mathcal{C}) = \mathrm{OCN}(\mathcal{C}, \mathcal{H})$. Let P be a teaching plan for $\mathcal{H}$ that is realizable by $\mathcal{C} \subseteq \mathcal{H}$ (e.g., a maximal $(\mathcal{C}, \mathcal{H})$-plan). According to Theorem 3, it suffices to show that the order of P is lower-bounded by $\mathrm{RTD}(\mathcal{C})$. To this end, we define the plan $P_{\mathcal{C}}$, called the *projection of P on $\mathcal{C}$*, as follows: $P_{\mathcal{C}}$ is obtained from P by deletion of all items (H_i, S_i) such that $H_i \notin \mathcal{C}$. It is obvious that $P_{\mathcal{C}}$ is a valid teaching plan for $\mathcal{C}$, and the order of $P_{\mathcal{C}}$ is smaller than or equal to the order of P. Thus, $\mathrm{OCN}(\mathcal{C}) \geq \mathrm{RTD}(\mathcal{C})$. □

The definition of RTD^* implies that $\mathrm{RTD}^*(\mathcal{C}) \geq \mathrm{RTD}(\mathcal{C})$, and it is easy to find classes for which RTD^* is considerably larger than RTD. Thus it is remarkable that Lemma 1 can be strengthened as follows:

Theorem 4. $\mathrm{OCN}(\mathcal{C}) \geq \mathrm{RTD}^*(\mathcal{C})$.

Proof. Let $\mathcal{H} = \{H_1, \ldots, H_m\}$ be a hypothesis class such that $\mathrm{OCN}(\mathcal{C}) = \mathrm{OCN}(\mathcal{C}, \mathcal{H})$. Let X be the domain of $\mathcal{C}$ and of $\mathcal{H}$. Let $X' \subseteq X$ such that $\mathrm{RTD}^*(\mathcal{C}) = \mathrm{RTD}(\mathcal{C}_{|X'})$. With this notation, the following holds:

$$\begin{aligned} \mathrm{OCN}(\mathcal{C}) = \mathrm{OCN}(\mathcal{C}, \mathcal{H}) &\overset{Th.1}{\geq} \mathrm{OCN}(\mathcal{C}_{|X'}, \mathcal{H}_{|X'}) \\ &\overset{L.1}{\geq} \mathrm{RTD}(\mathcal{C}_{|X'}, \mathcal{H}_{|X'}) = \mathrm{RTD}^*(\mathcal{C}) \end{aligned}$$

This proves the theorem. □

Since $\mathrm{RTD}^*(\mathcal{C}) \geq \mathrm{VCD}(\mathcal{C})$ (see [5]), we immediately obtain the following corollary from Theorem 4:

Corollary 2. $\mathrm{OCN}(\mathcal{C}) \geq \mathrm{VCD}(\mathcal{C})$.

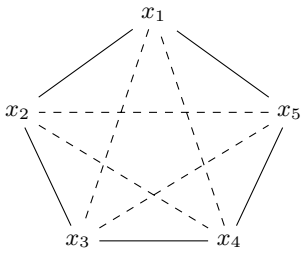
It is known from previous work that VCD can exceed RTD by an arbitrary amount [5]. Thus, Corollary 2 implies that also OCN can exceed RTD by an arbitrary amount.

Example 2 below presents a concept class $\mathcal{C}_{MW}$ such that $\mathrm{VCD}(\mathcal{C}_{MW}) = 2$ and $\mathrm{OCN}(\mathcal{C}_{MW}) = 3$, thereby showing that the inequality $\mathrm{OCN}(\mathcal{C}) \geq \mathrm{VCD}(\mathcal{C})$ can be strict occasionally. By means of padding, it is easy to find classes $\mathcal{C}$ of arbitrarily large VC-dimension such that $\mathrm{OCN}(\mathcal{C}) = 1.5 \cdot \mathrm{VCD}(\mathcal{C})$. However, at the time being, it is not known whether the gap can be made larger than a factor of 1.5.

Example 2. Consider the class $\mathcal{C}_{MW}$ in Figure 2, which was found by Manfred Warmuth (personal communication). It is the smallest concept class for which RTD exceeds VCD [4]. In this particular example, $\mathrm{RTD}(\mathcal{C}_{MW}) = 3$, while $\mathrm{VCD}(\mathcal{C}_{MW}) = 2$.

$\mathcal{C}_{MW}$	x_1	x_2	x_3	x_4	x_5
C_1	1	1	0	0	0
C_2	0	1	1	0	0
C_3	0	0	1	1	0
C_4	0	0	0	1	1
C_5	1	0	0	0	1
C_6	0	1	0	1	1
C_7	0	1	1	0	1
C_8	1	0	1	1	0
C_9	1	0	1	0	1
C_{10}	1	1	0	1	0

(a) $\mathcal{C}_{MW}$ given as a table of concepts.



(b) All concepts in $\mathcal{C}_{MW}$ are given either by the vertices of the solid edges or by the complements of the vertices of the dashed edges.

Fig. 2. The smallest compression schemes for the concept class $\mathcal{C}_{MW}$ are always cyclic. Part (b) is a nice visualization of this concept class.

An improper sample compression scheme (f, g) for $\mathcal{C}_{MW}$ of size 2 can be defined as follows (there also exists a proper, but more involved scheme of the same size): any set S that is homogeneously labeled is compressed to the empty set (in the case of label 1) or a single example with label 0 (in the other case). If S has mixed labels we consider the following cases: sets that contain exactly one or two examples with label 1 are compressed to these one or two examples. Sets that contain three examples with label 1 and two with label 0 are compressed to

the two 0-labeled examples. Since X consists of five elements, only the following case is left: S contains three or four 1-labeled examples and exactly one 0-labeled example. In that case S is compressed to any pair with mixed labels. The decompression of $f(S)$ proceeds in the obvious way: if $f(S)$ consists exactly of one or two 1-labeled examples or of a single 0-labeled example, g chooses the label 0 for all points outside of $f(S)$. Otherwise g assigns the label 1 to these points.

Note that the compression graph $\mathcal{G} = (V, E)$ for this scheme has cycles. For instance, there is a loop between the two hypotheses $H = \{x_2\}$ and $H' = \{x_2, x_3, x_5\}$; the edge $(H', H) \in E$ is witnessed by the sample $\{(x_1, 0), (x_2, 1)\}$ (this also shows that indeed $H \in \mathcal{H}$), while $(H, H') \in E$ is witnessed by $\{(x_1, 0), (x_2, 1), (x_3, 1), (x_4, 0), (x_5, 1)\}$.

In fact, we already know that no sample compression scheme of size 2 for $\mathcal{C}_{MW}$ can be acyclic since $\mathrm{OCN}(\mathcal{C}_{MW}) \geq \mathrm{RTD}(\mathcal{C}_{MW}) = 3$. The reverse direction $\mathrm{OCN}(\mathcal{C}_{MW}) \leq 3$ is obtained by the following claim which provides us with a proper order scheme of size 3:

Every $\mathcal{C}_{MW}$-realizable sample S contains a subsample S' of size at most 3 such that every concept from $\mathcal{C}_{MW}$ which is consistent with S' is consistent with S too. The claim is obvious if $|S| \leq 3$. It is obvious if $|S| = 5$ because every concept has a teaching set of size 3 (consisting either of three positive or of three negative examples). Let now $|S| = 4$. If S still contains one of the teaching sets of size 3, we are done. Otherwise we may assume for reasons of symmetry that $S = \{(x_1, 0), (x_2, 1), (x_3, 1), (x_4, 0)\}$. But then $S' = \{(x_1, 0), (x_2, 1), (x_3, 1)$ fits our purpose.

Example 2 demonstrates that the size of the best order scheme (with acyclic compression graph) can occasionally be larger than the size of the best arbitrary scheme (with a non-acyclic compression graph).

5 Order Schemes for Special Classes

The following families of concept classes $\mathcal{C}$ are known to have sample compression schemes of size $\mathrm{VCD}(\mathcal{C})$:

- the family $\mathcal{F}_\cap$ of intersection-closed classes,
- the family $\mathcal{F}_{max}$ of maximum classes,
- the family $\mathcal{F}_{Dudley}$ of Dudley classes,
- the family $\mathcal{F}_1$ of classes of VC-dimension 1.

In the sequel, we show that (some of) the standard sample compression schemes for classes from these families induce an acyclic compression graph so that, according to Theorem 2, they actually are order schemes. Before starting our investigation with intersection-closed and maximum classes, we briefly remind the reader of some standard definitions and facts. A class $\mathcal{C}$ is called *intersection-closed* if the intersection of any two concepts from $\mathcal{C}$ is itself a concept in $\mathcal{C}$ as well. For $T \subseteq X$, $\langle T \rangle_\mathcal{C}$ denotes the unique smallest concept in $\mathcal{C}$ containing T. A *spanning set* for a set $T \subseteq X$ is a set $T' \subseteq T$ such that $\langle T' \rangle_\mathcal{C} = \langle T \rangle_\mathcal{C}$.

It is called *minimal* if no proper subset T'' of T' satisfies $\langle T'' \rangle_{\mathcal{C}} = \langle T' \rangle_{\mathcal{C}}$. It is well known that the size of any minimal spanning set is bounded from above by $\mathrm{VCD}(\mathcal{C})$ [13, 10].

A class $\mathcal{C}$ of VC-dimension d over a domain X of cardinality n is called *maximum* if $|\mathcal{C}| = \sum_{i=0}^{d} \binom{n}{i}$ [15, 17]. The following definition was introduced by Kuzmin and Warmuth [11]. An *unlabeled sample compression scheme* for a maximum class $\mathcal{C}$ of VC-dimension d is given by a bijective mapping r that assigns to every concept $C \in \mathcal{C}$ a set $r(C) \subseteq X$ of size at most d such that the following condition, referred to as the *non-clashing property*, is satisfied:

$$\forall C \neq C' \in \mathcal{C}, \exists x \in r(C) \cup r(C') : C(x) \neq C'(x) \tag{2}$$

As shown in [11], the non-clashing property guarantees that, for every $\mathcal{C}$-realizable sample S, there is exactly one concept $C \in \mathcal{C}$ that is consistent with S and satisfies $r(C) \subseteq X(S)$. This allows to compress S by $f(S) = r(C)$ and to decompress $r(C)$ by $g(r(C)) = C$, i.e., the decompression function g is the inverse of the bijective function r. The *acyclic non-clashing property* with respect to an ordering $C_1 < \ldots < C_m$ of the concepts in $\mathcal{C} = \{C_1, \ldots, C_m\}$ is the following modification of (2):

$$\forall 1 \leq i < j \leq m, \exists x \in r(C_i) : C_i(x) \neq C_j(x) \tag{3}$$

For instance, the representation function resulting from the Tail Matching Algorithm [11] has the acyclic non-clashing property.

Theorem 5. *There are proper order schemes for $\mathcal{C}$ of size $\mathrm{VCD}(\mathcal{C})$ provided that $\mathcal{C}$ is intersection-closed or maximum.*

Proof. First, suppose that $\mathcal{C}$ is intersection-closed. Let S be a $\mathcal{C}$-realizable sample, and let S_+ be the subsample consisting precisely of all positive examples in S. Then the standard scheme (known to be of size $\mathrm{VCD}(\mathcal{C})$) compresses S to a minimal spanning set $S' \subseteq S_+$ for S_+. A sample $S' = f(S)$ (always consisting of positive examples only) is decompressed to the smallest set in $\mathcal{C}$ that contains S', i.e., $g(S') = \langle X(S'_+) \rangle_{\mathcal{C}}$. Consider now the compression graph G associated with (f, g). Every sample $S' = f(S)$ induces edges leading from concepts properly containing $\langle X(S'_+) \rangle_{\mathcal{C}}$ to $\langle X(S'_+) \rangle_{\mathcal{C}}$. Since edges always lead from sets to proper subsets, G is acyclic. We may therefore conclude from Theorem 2 that the scheme is an order scheme.

Second, suppose that $\mathcal{C}$ is a maximum class. We will argue that the scheme (f, g) induced by a representation function r is an order scheme provided that r satisfies (3). Again it suffices to show that the compression graph G associated with (f, g) is acyclic. To this end, let (C_i, C_j) be an edge in G. Thus there exists a sample S such that $C_i, C_j \in \mathrm{Cons}(S, \mathcal{C})$ and $C_j = g(f(S))$. The latter condition is equivalent to C_j being the unique concept that is consistent with S and satisfies $r(C_j) \subseteq X(S)$. Since both of C_i, C_j are consistent with S, they do not disagree on $r(C_j)$. According to the acyclic non-clashing property, they disagree on $r(C_i)$ and $i < j$. Thus edges in G always go from smaller to larger indexes so that G is acyclic. □

The proper order schemes for the classes mentioned in Theorem 5 can be used as (non-proper) order-schemes for subclasses. The family of subclasses of maximum classes is very rich and comprises the so-called Dudley classes.

Definition 5 (Dudley [6]). *Let $\mathcal{F}$ be a vector space of real-valued functions over some domain X and $h : X \to \mathbb{R}$. For every $f \in \mathcal{F}$, let*

$$C_f(x) := \begin{cases} 1, & \text{if } f(x) + h(x) \geq 0 \\ 0, & \text{else} \end{cases} .$$

Then $D_{\mathcal{F},h} = \{C_f | f \in \mathcal{F}\}$ is called a Dudley class. *The* dimension *of $D_{\mathcal{F},h}$ is equal to the dimension of the vector space $\mathcal{F}$.*

Some popular examples of Dudley classes include:

- collections of half spaces over $\mathbb{R}^n$, which are very common objects of study in machine learning, such as in artificial neural networks and support vector machines, see, e.g., [1],
- unions of at most k intervals over $\mathbb{R}$,
- n-dimensional balls.

Now, the following well-known result comes into play:

Lemma 2 (Ben-David and Litman [2]). *Dudley classes of dimension k are embeddable in maximum classes of VC-dimension k.*

Lemma 2 combined with Theorem 5 yields

Corollary 3. *Let $\mathcal{C}$ be a Dudley class. Then $\mathcal{C}$ has a (possibly improper) order scheme of size* $\mathrm{VCD}(\mathcal{C})$.

Another family for which we obtain order schemes of size VC-dimension is the one consisting of all classes of VC-dimension 1. Such classes are known to be contained in maximum classes of VC-dimension 1 [18].

Corollary 4. *Let $\mathcal{C}$ be a concept class of VC-dimension 1. Then $\mathcal{C}$ has a (possibly improper) order scheme of size 1.*

In combination, we obtain:

Corollary 5. $\mathrm{OCN}(\mathcal{C}) = \mathrm{VCD}(\mathcal{C})$ *provided that $\mathcal{C}$ belongs to at least one of the families $\mathcal{F}_\cap, \mathcal{F}_{max}, \mathcal{F}_{Dudley}, \mathcal{F}_1$.*

Finally, we can generalize our result on intersection-closed classes to the case of nested differences of such classes. A *nested difference* of depth d over $\mathcal{C}$ is a concept $C_1 \setminus (C_2 \setminus (\ldots (C_{d-1} \setminus C_d) \ldots))$ where each C_i belongs to $\mathcal{C}$. The class of nested differences of depth at most d over $\mathcal{C}$ is denoted by $\mathrm{DIFF}^{\leq d}(\mathcal{C})$. Our generalization of the result on intersection-closed classes is the following.

Theorem 6. $\mathrm{OCN}(\mathrm{DIFF}^{\leq d}(\mathcal{C})) \leq d \cdot \mathrm{VCD}(\mathcal{C})$ *provided that $\mathcal{C}$ is intersection-closed.*

In the proof of this theorem, we assume that any concept in $\mathrm{DIFF}^{\leq d}(\mathcal{C})$ is given in a normal form as follows (see [5]). We can represent $C \in \mathcal{H}$ as

$$C = C_1 \setminus \overbrace{(C_2 \setminus (\cdots (C_{d-1} \setminus C_d) \cdots))}^{=:D_1} \tag{4}$$

such that for every j it holds that $C_j \in \mathcal{C} \cup \{\emptyset\}$ and, unless $C_j = \emptyset$, C_{j+1} is a proper subset of C_j. Then, for $D_j = C_{j+1} \setminus (C_{j+2} \setminus (\ldots (C_{d-1} \setminus C_d) \ldots))$, we can assume that the representation of the form (4) is minimal in the sense that $C_j = \langle C_j \setminus D_j \rangle_{\mathcal{C}}$ holds for all $1 \leq j \leq d$.

Proof. Let $\mathcal{H} = \mathrm{DIFF}^{\leq d}(\mathcal{C})$. We define a partial order $\sqsupset$ on $\mathcal{H}$. Given two concepts $C, C' \in \mathcal{H}$ let $C = C_1 \setminus D_1$ and $C = C'_1 \setminus D'_1$ be their normalized representations. Then $C \sqsupset C'$ iff $C_1 \supset C'_1$ or $C_1 = C'_1 \wedge D_1 \sqsupset D'_1$.

Let $(H_1, \ldots, H_m)$ be any order over $\mathcal{H}$ such that $j < i$ if $H_j \sqsupset H_i$ and let (f, g) be the corresponding proper order scheme.

Recall that, given a $\mathcal{H}$-realizable sample S, the compression function f finds the largest t such that $H_t \in \mathrm{Cons}(S, \mathcal{H})$ and then compresses S to a teaching set for H_t with respect to $\{H_t, \ldots, H_m\}$. We will now describe a method for constructing this hypothesis H_t: let $S_1 = \{x | (x, 1) \in S\}$ and $C_1 = \langle S_1 \rangle_{\mathcal{C}}$, i.e., C_1 is the smallest concept in $\mathcal{C}$ that is consistent with all examples in S that are labeled with 1. Note that C_1 can be inconsistent with some of the examples in S that are labeled with 0 – hence, let $S_2 = \{x \in C_1 | (x, 0) \in S\}$ and $C_2 = \langle S_2 \rangle_{\mathcal{C}}$. Then $C_1 \setminus C_2$ itself can disagree with some S on some examples contained with label 1 in S. Again, let $S_3 = \{x \in C_2 | (x, 1) \in S\}$ and $C_3 = \langle S_3 \rangle_{\mathcal{C}}$. Proceed inductively in this manner until the nested difference $H_S = C_1 \setminus (C_2 \setminus (\ldots (C_{d-1} \setminus C_d) \ldots))$ is consistent with S. This procedure will find a concept consistent with S in at most d steps, because of the normal form assumption on all the underlying concepts in $\mathcal{H}$. By construction, every $H \in \mathrm{Cons}(S, \mathcal{H})$ fulfills $H \sqsupseteq H_S$, and thus H_S is the last concept in the underlying order that is consistent with S. Hence, H_S equals the desired concept H_t. Now $f(S)$ is a smallest teaching set for H_t with respect to $\{H_t, \ldots, H_m\}$, among the subsets of S.

We can now give an upper bound on the size of the order compression scheme defined above: for any i, let $S'_i \subseteq S_i$ be smallest such that $\langle S'_i \rangle_{\mathcal{C}} = \langle S_i \rangle_{\mathcal{C}}$, i.e. S'_i is a minimal spanning set. Augment the instances of S'_i by the label 1 if i is odd and by 0 otherwise. Then let S' be the union over all S'_i for $1 \leq i \leq d$. It follows that S' is a (not necessarily minimal) teaching set for H_t in $\{H_t, \ldots, H_m\}$. Thus $|f(S)| \leq |S'|$ and, because $|S'_i| \leq \mathrm{VCD}(\mathcal{C})$, we obtain $|f(S)| \leq d \cdot \mathrm{VCD}(\mathcal{C})$. □

6 Conclusions

Order compression schemes obey a very simple structure and exhibit interesting connections to teaching and graph theory. Furthermore, in most of the cases where the sample compression conjecture is known to be true, it can already be verified using order compression schemes. We hence believe that order compression schemes provide a useful notion for studying sample compression schemes in general.

While we presented a number of important fundamental properties of order compression schemes, several questions remain open, most notably the question of how VCD and OCN relate in general. One of many challenges in this context could be to devise a method for finding a best possible hypothesis space $\mathcal{H}$ for $\mathcal{C}$, so that an order compression scheme for $\mathcal{H}$ induces the best possible order compression scheme for $\mathcal{C}$, i.e., so that $\mathrm{OCN}(\mathcal{C}, \mathcal{H}) = \mathrm{OCN}(\mathcal{C})$.

Acknowledgements. We would like to thank three anonymous referees for their insightful and inspiring comments and suggestions.

References

1. Alpaydin, E.: Introduction to Machine Learning, 2nd ed. MIT Press (2010)
2. Ben-David, S., Litman, A.: Combinatorial variability of Vapnik-Chervonenkis classes with applications to sample compression schemes. Discrete Applied Mathematics 86, 3–25 (1998)
3. Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K.: Learnability and the Vapnik-Chervonenkis dimension. J. ACM 36(4), 929–965 (1989)
4. Doliwa, T., Fan, G., Simon, H.U., Zilles, S.: Recursive teaching dimension, VC-dimension, and sample compression (January 2013), submitted
5. Doliwa, T., Simon, H.U., Zilles, S.: Recursive teaching dimension, learning complexity, and maximum classes. In: ALT. pp. 209–223 (2010)
6. Dudley, R.M.: A course on empirical processes. Lecture Notes in Mathematics 1097, 1–142 (1984)
7. Fan, G.: A graph-theoretic view of teaching, M.Sc. Thesis, University of Regina (2012)
8. Floyd, S., Warmuth, M.K.: Sample compression, learnability, and the Vapnik-Chervonenkis dimension. Machine Learning 21(3), 269–304 (1995)
9. Goldman, S.A., Kearns, M.J.: On the complexity of teaching. J. Comput. Syst. Sci. 50(1), 20–31 (1995)
10. Helmbold, D.P., Sloan, R.H., Warmuth, M.K.: Learning nested differences of intersection-closed concept classes. Machine Learning 5, 165–196 (1990)
11. Kuzmin, D., Warmuth, M.K.: Unlabeled compression schemes for maximum classes. Journal of Machine Learning Research 8, 2047–2081 (2007)
12. Littlestone, N., Warmuth, M.K.: Relating data compression and learnability. Tech. rep., University of California at Santa Cruz (1986)
13. Natarajan, B.K.: On learning boolean functions. In: Proceedings of the 19th Annual Symposium on Theory of Computing. pp. 296–304 (1987)
14. Rubinstein, B.I., Rubinstein, J.H.: A geometric approach to sample compression. Journal of Machine Learning Research 13, 1221–1261 (2012)
15. Sauer, N.: On the density of families of sets. J. Comb. Theory, Ser. A 13(1), 145–147 (1972)
16. Shinohara, A., Miyano, S.: Teachability in computational learning. New Generation Computing 8(4), 337–347 (1991)
17. Welzl, E.: Complete range spaces. Unpublished manuscript (1987)
18. Welzl, E., Wöginger, G.: On Vapnik-Chervonenkis dimension one (1987), unpublished manuscript
19. Zilles, S., Lange, S., Holte, R., Zinkevich, M.: Models of cooperative teaching and learning. Journal of Machine Learning Research 12, 349–384 (2011)