Order Compression Schemes

Malte Darnstädt[a], Thorsten Kiss[a], Hans Ulrich Simon[a], Sandra Zilles[b]

[a] *Fakultät für Mathematik, Ruhr-Universität Bochum, D-44780 Bochum, Germany*
[b] *Department of Computer Science, University of Regina, Regina, SK, Canada S4S 0A2*

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.

To obtain small compressed sets, order compression schemes for a concept class $\mathcal{C}$ must often use a proper superset $\mathcal{H} \supset \mathcal{C}$ as a hypothesis space. We thus further compare order compression schemes for $\mathcal{C}$ to order compression schemes for such hypothesis spaces, leading to a study of a number of mutually related combinatorial parameters specifying compressibility.

Keywords: learning theory, sample compression schemes, 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 $\mathcal{C}$ [1, 2]. 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 $\mathcal{C}$.

Littlestone and Warmuth [1] pointed out that their sample bounds for PAC-learning, stated in terms of the sample compression number, could sometimes 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 of great interest to the learning theory community, as witnessed by a continuing series of publications on the topic, see, e.g., [4, 5, 6, 2, 7, 1, 8, 9]. Floyd and Warmuth [2] 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[1] $\mathcal{C}$ 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, and it is shown to exceed the VC-dimension in some cases. 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.
- Many finite concept classes $\mathcal{C}$ for which the sample compression conjecture is proven, i.e., that are known to 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 [10, 11]), for intersection-closed classes, for Dudley classes [12], 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 order compression scheme is not linear in the VC-dimension. First, due to the simplicity of the construction of order compression schemes, this should be less challenging than disproving the general conjecture. Second, 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 [13]. 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 exactly to the acyclic graphs in this definition.

We further provide a number of properties of order compression that may potentially be helpful in future studies. For this purpose, we introduce a combinatorial parameter called *order compression number* of a concept class $\mathcal{C}$, which refers to the smallest possible order compression scheme for $\mathcal{C}$. This parameter turns out to satisfy helpful monotonicity and sub-additivity properties.

For many concept classes $\mathcal{C}$, smallest possible order compression schemes necessarily involve a hypothesis space $\mathcal{H}$ (i.e., a class of concepts used as decompressions of compressed sample sets) that strictly comprises $\mathcal{C}$. Interestingly, the order compression number for such $\mathcal{H}$ can be larger than that of $\mathcal{C}$, i.e., order-compressing

[1] 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 [4].

only the subclass $\mathcal{C}$ of $\mathcal{H}$ using the hypothesis space $\mathcal{H}$ can be "easier" than order-compressing the whole class $\mathcal{H}$. Our results on the relationship between the order compression number of $\mathcal{C}$ and that of $\mathcal{H}$ include a provably smallest class for which these two parameters differ.

It remains open whether or not the size of a smallest possible order compression scheme is linear in the VC-dimension. We demonstrate that the former can exceed the latter by a factor of $3/2$, but so far no example of a larger discrepancy between the two parameters is known. Interestingly, the factor $3/2$ is also the largest known discrepancy between the recursive teaching dimension and the VC-dimension [5], which suggests a further relationship between teaching and sample compression.

This article is an extension of a published conference paper [14].

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 classes and 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 and every $X' \subseteq X$, we define $X(S) := \{x \in X |\ (x, 0) \in S \text{ or } (x, 1) \in S\}$ and $S_{|X'} := \{(x, b) \in S |\ x \in X'\}$. Further, $\mathrm{Cons}(S, \mathcal{H})$ is the set of all hypotheses in $\mathcal{H}$ that are consistent with S, i.e., $\mathrm{Cons}(S, \mathcal{H}) = \{H \in \mathcal{H} \mid H(x) = l \text{ for all } (x, l) \in S\}$. S is called *$\mathcal{C}$-realizable* if $\mathrm{Cons}(S, \mathcal{C}) \neq \emptyset$. The *VC-dimension* of the class $\mathcal{C}$, denoted by $\mathrm{VCD}(\mathcal{C})$, 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* [1] 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 for "sample compression scheme for $\mathcal{C}$ using hypotheses from $\mathcal{H}$".

Definition 1. *The* compression number *of a pair $(\mathcal{C}, \mathcal{H})$, denoted by* $\mathrm{CN}(\mathcal{C}, \mathcal{H})$*, called the* compression number *of $(\mathcal{C}, \mathcal{H})$, is the minimum size of an an arbitrary $(\mathcal{C}, \mathcal{H})$-scheme.*

Example 2. *The following compression scheme for the class $\mathcal{P}_2$ of all subsets of $\{1, 2\}$ is of size 1, and thus $\mathrm{CN}(\mathcal{P}_2, \mathcal{P}_2) = 1$.*

- *Any sample without a positive example is compressed to $\emptyset$ (the empty sample), which is decompressed to $\emptyset$ (the empty hypothesis).*
- *Any sample containing $(1, +)$ but not $(2, +)$ is compressed to $\{(1, +)\}$, which is decompressed to $\{1\}$.*
- *Any sample containing $(2, +)$ but not $(1, -)$ is compressed to $\{(2, +)\}$, which is decompressed to $\{1, 2\}$.*
- *The sample $\{(1, -), (2, +)\}$ is compressed to $\{(1, -)\}$, which is decompressed to $\{2\}$.*

Now let $\mathcal{P}_n$ be the power set over n elements. For $n \geq 4$, the following compression scheme shows that $\mathrm{CN}(\mathcal{P}_n, \mathcal{P}_n) \leq \lfloor n/2 \rfloor$:

- *As before, any sample containing only negative examples is compressed to $\emptyset$, which is again decompressed to $\emptyset$, i.e. the hypothesis assigning negative labels to all elements.*

- *Any sample containing only positive examples is compressed to an arbitrary single positive example, which is decompressed to a hypothesis assigning positive labels to all elements.*
- *Any sample containing exactly one positive and one or more negative examples is compressed to the positive and an arbitrary negative example; the decompressed hypothesis assigns negative labels to all elements missing in the compressed sample.*
- *All remaining samples are compressed to their examples with the minority label. The resulting sets can be distinguished from the ones of the afore-mentioned cases (if the minority label is positive, the resulting set will contain at least two examples, separating it from the second case), and the decompressed hypothesis assigns the opposite label to all elements missing in the compressed sample.*

Since a compression scheme of size k uses at most $\sum_{i=0}^{k} 2^i \binom{n}{i}$ compression sets, no class of more than $\sum_{i=0}^{k} 2^i \binom{n}{i}$ concepts can have such a scheme. Since $\binom{n}{0} + 2 \cdot \binom{n}{1} = 2n + 1 < 2^n$ for $n \geq 3$, this implies that $\mathrm{CN}(\mathcal{P}_n, \mathcal{H}) > 1$ for all $n \geq 3$ and all $\mathcal{H} \supseteq \mathcal{P}_n$, and thus $\mathrm{CN}(\mathcal{P}_3, \mathcal{P}_3) = \mathrm{CN}(\mathcal{P}_4, \mathcal{P}_4) = \mathrm{CN}(\mathcal{P}_5, \mathcal{P}_5) = 2$.

Let (f, g) be a sample compression 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 every class $\mathcal{C}$ of finite VC-dimension d has a sample compression scheme of size linear in d, or even equal to d [2].

2.2. Teaching

Following Goldman and Kearns [15] 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 $\mathcal{C}$. $\mathrm{TD}(C, \mathcal{C}) = \min\{|S| \mid S \in \mathcal{TS}(C, \mathcal{C})\}$, called the *teaching dimension* of C with respect to $\mathcal{C}$, denotes the smallest possible size of a teaching set for C with respect to $\mathcal{C}$. According to Zilles et al. [13], 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 $\mathrm{ord}(P) = \max\{|S_t| \mid 1 \leq t \leq m\}$. 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_t is chosen from $\mathcal{C}_t = \mathcal{C} \setminus \{C_1, \ldots, C_{t-1}\}$ 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. [6].

3. Properties of Order Compression Schemes

We will now introduce *order compression schemes*, a notion that is inspired by Fan's work [17] and that is central to this paper:

Definition 3. *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 the 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 3 must be the same number, shows that any such scheme is indeed a compression scheme:

Proposition 4. *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 for $(\mathcal{C}, \mathcal{H})$ is a $(\mathcal{C}, \mathcal{H})$-scheme.*

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

Definition 5. *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 orders over $\mathcal{H}$).*

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 6. *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 order $\{0\} < \ldots < \{n-1\} < \emptyset$ 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 $\mathrm{OCN}(\mathcal{H}, \mathcal{H}) = \mathrm{OCN}(\mathcal{C}, \mathcal{H}) = 1$.

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 order. 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 indices).*

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 $\mathcal{H} \supseteq \mathcal{C}$ when considering order compression schemes. 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 7. *The* order compression number *of $\mathcal{C}$, denoted $\mathrm{OCN}(\mathcal{C})$ (resp. $\mathrm{OCN}'(\mathcal{C})$), is the minimum of $\mathrm{OCN}(\mathcal{C}, \mathcal{H})$ (resp. $\mathrm{OCN}(\mathcal{H}, \mathcal{H})$) over the choice of $\mathcal{H} \supseteq \mathcal{C}$. We say that $\mathcal{C}$ is* OC-closed *if $\mathrm{OCN}(\mathcal{C}, \mathcal{C}) = \mathrm{OCN}(\mathcal{C})$, i.e., if the size of the smallest proper order scheme for $\mathcal{C}$ coincides with the size of the smallest (not necessarily proper) order scheme for $\mathcal{C}$. The notations $\mathrm{CN}(\mathcal{C})$ and $\mathrm{CN}'(\mathcal{C})$, referring to arbitrary compression schemes, should be understood analogously. We say that $\mathcal{C}$ is* AC-closed *if $\mathrm{CN}(\mathcal{C}, \mathcal{C}) = \mathrm{CN}(\mathcal{C})$.*

Note that

$$\mathrm{OCN}(\mathcal{C}) \leq \mathrm{OCN}'(\mathcal{C}) \leq \mathrm{OCN}(\mathcal{C},\mathcal{C}) \;\text{ and }\; \mathrm{CN}(\mathcal{C}) \leq \mathrm{CN}'(\mathcal{C}) \leq \mathrm{CN}(\mathcal{C},\mathcal{C}) \tag{1}$$

with equality for OC-closed (resp. AC-closed) concept classes. The concept class $\mathcal{H}$ from Example 6 (singletons augmented by the empty set) is OC-closed, whereas the class $\mathcal{C}$ from Example 6 (singletons) is not. The concept class $\mathcal{P}_2$ from Example 2 is AC-closed because, clearly, there is no compression scheme for $\mathcal{P}_2$ of size 0.

Definition 8. *Let $X_1,\ldots,X_s$ be pairwise disjoint domains. For $i=1,\ldots,s$, let $\mathcal{C}_i$ be a concept class over X_i. The* direct sum of $\mathcal{C}_1,\ldots,\mathcal{C}_s$ *is defined as follows:*

$$\biguplus_{r=1}^{s} \mathcal{C}_r := \{\cup_{r=1}^{s} C_r |\; C_r \in \mathcal{C}_r\}$$

Let K be a function that assigns a real number to each concept class $\mathcal{C}$. We say that K is sub-additive on $\mathcal{C}_1,\ldots,\mathcal{C}_s$ *if*

$$K\left(\biguplus_{r=1}^{s} \mathcal{C}_r\right) \leq \sum_{r=1}^{s} K(\mathcal{C}_r) \;. \tag{2}$$

If (2) holds with equality, we say that K is additive on $\mathcal{C}_1,\ldots,\mathcal{C}_s$. *We say that K* (sub-)additive *if it is (sub-)additive on every finite sequence of concept classes over pairwise disjoint domains.*

In the sequel, we apply the operation "$\uplus$" to arbitrary sequences of concept classes with the understanding that, after renaming instances if necessary, the underlying domains are pairwise disjoint. If K is additive (sub-additive) on finite sequences consisting of "disjoint duplicates" of the same concept class $\mathcal{C}$, we say that K is *additive (sub-additive, resp.) on* $\mathcal{C}$. We say that $K \leq K'$ if $K(\mathcal{C}) \leq K'(\mathcal{C})$ holds for every concept class $\mathcal{C}$.

Lemma 9. *Suppose that K is additive on $\mathcal{C}$, K' is sub-additive on $\mathcal{C}$, $K \leq K'$, and $K(\mathcal{C}) = K'(\mathcal{C})$. Then, K' is additive on $\mathcal{C}$.*

Proof. Let $\mathcal{C} \uplus \ldots \uplus \mathcal{C}$ be the direct sum of s disjoint duplicates of $\mathcal{C}$. Then,

$$K'(\mathcal{C} \uplus \ldots \uplus \mathcal{C}) \leq s \cdot K'(\mathcal{C}) = s \cdot K(\mathcal{C}) = K(\mathcal{C} \uplus \ldots \uplus \mathcal{C}) \leq K'(\mathcal{C} \uplus \ldots \uplus \mathcal{C}) \;.$$

Thus all terms in this chain of equalities/inequalities are actually equal. □

It is well known that VCD and RTD are additive [5]. As we show now, OCN and CN are sub-additive.

Lemma 10. *Let $\mathcal{C}' \subseteq \mathcal{H}'$ (resp. $\mathcal{C}'' \subseteq \mathcal{H}''$) be concept classes over the same domain X' (resp. X'') where $X' \cap X'' = \emptyset$. Then the following holds:*

$$\mathrm{OCN}(\mathcal{C}' \uplus \mathcal{C}'', \mathcal{H}' \uplus \mathcal{H}'') \;\leq\; \mathrm{OCN}(\mathcal{C}', \mathcal{H}') + \mathrm{OCN}(\mathcal{C}'', \mathcal{H}'') \tag{3}$$

$$\mathrm{CN}(\mathcal{C}' \uplus \mathcal{C}'', \mathcal{H}' \uplus \mathcal{H}'') \;\leq\; \mathrm{CN}(\mathcal{C}', \mathcal{H}') + \mathrm{CN}(\mathcal{C}'', \mathcal{H}'') \tag{4}$$

Proof. Let $\mathcal{C} = \mathcal{C}' \uplus \mathcal{C}''$, $\mathcal{H} = \mathcal{H}' \uplus \mathcal{H}''$ and $X = X' \cup X''$. Inequality (4) is pretty obvious because any $\mathcal{C}$-realizable sample S decomposes into $S' := S_{|X'}$ and $S'' := S_{|X''}$, and these sets can be compressed and decompressed separately by means of the optimal $(\mathcal{C}', \mathcal{H}')$-scheme and the optimal $(\mathcal{C}'', \mathcal{H}'')$-scheme.
The reasoning for inequality (3) is similar but slightly more complicated because we have to commit to an order over the hypotheses in $\mathcal{H}$. Details follow.
Let $m' = |\mathcal{H}'|$ and let $H'_1,\ldots,H'_{m'}$ be an order over the hypotheses from $\mathcal{H}'$ that leads to a $(\mathcal{C}', \mathcal{H}')$-scheme, say (f', g'), of size $d' := \mathrm{OCN}(\mathcal{C}', \mathcal{H}')$. Let $m'', H''_1,\ldots,H''_{m''}, f'', g'', d''$ be the analogous quantities for the classes $\mathcal{C}''$ and $\mathcal{H}''$. We choose the following order for the $m'm''$ hypotheses from $\mathcal{H}$:

$$H'_1 \cup H''_1, H'_1 \cup H''_2, \ldots, H'_1 \cup H''_{m''} \;,\; \ldots \;,\; H'_{m'} \cup H''_1, H'_{m'} \cup H''_2, \ldots, H'_{m'} \cup H''_{m''} \tag{5}$$

Let S be a $\mathcal{C}$-realizable sample. S decomposes into $S' := S_{|X'}$ and $S'' := S_{|X''}$. Note that S' is $\mathcal{C}'$-realizable and S'' is $\mathcal{C}''$-realizable. Suppose that $H'_{t'} \cup H''_{t''}$ is the rightmost hypothesis in (5) that is consistent with S. Then t' (resp. t'') is the largest number such that $H'_{t'}$ (resp. $H''_{t''}$) is consistent with S' (resp. S''). $f'(S')$ (resp. $f''(S'')$) is a smallest subset of S' (resp. S'') that is a teaching set for $H'_{t'}$ with respect to $\{H'_{t'}, \ldots, H'_{m'}\}$ (resp. for $H''_{t''}$ with respect to $\{H''_{t''}, \ldots, H''_{m''}\}$). It follows that $f'(S') \cup f''(S'')$ is a teaching set for $H'_{t'} \cup H''_{t''}$ with respect to $\{H'_{t'} \cup H''_{t''}, \ldots H'_{m'} \cup H''_{m''}\}$. Moreover, $|f'(S') \cup f''(S'')| \leq d' + d''$. It follows that the ordering (5) induces a $(\mathcal{C}, \mathcal{H})$-scheme of size at most $d' + d''$, as required. □

Corollary 11. $\mathrm{OCN}, \mathrm{OCN}', \mathrm{CN}, \mathrm{CN}'$ *are sub-additive.*

Proof. Let us consider OCN. It clearly suffices to verify (2) for $s = 2$ (with OCN in the role of K). The sub-additivity of OCN is obtained from Lemma 10 and the following calculation:

$$\begin{aligned}
\mathrm{OCN}(\mathcal{C}_1 \uplus \mathcal{C}_2) &= \min_{\mathcal{H} \supseteq \mathcal{C}_1 \uplus \mathcal{C}_2} \mathrm{OCN}(\mathcal{C}_1 \uplus \mathcal{C}_2, \mathcal{H}) \\
&\leq \min_{\mathcal{H}_1 \supseteq \mathcal{C}_1, \mathcal{H}_2 \supseteq \mathcal{C}_2} \mathrm{OCN}(\mathcal{C}_1 \uplus \mathcal{C}_2, \mathcal{H}_1 \uplus \mathcal{H}_2) \\
&\leq \min_{\mathcal{H}_1 \supseteq \mathcal{C}_1} \mathrm{OCN}(\mathcal{C}_1, \mathcal{H}_1) + \min_{\mathcal{H}_2 \supseteq \mathcal{C}_2} \mathrm{OCN}(\mathcal{C}_2, \mathcal{H}_2) \\
&= \mathrm{OCN}(\mathcal{C}_1) + \mathrm{OCN}(\mathcal{C}_2)
\end{aligned}$$

The first inequality makes use of the fact that every pair of classes $\mathcal{H}_1, \mathcal{H}_2$ over X_1, X_2, respectively, can be merged to $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$. The second inequality makes use of Lemma 10. The sub-additivity of $\mathrm{OCN}', \mathrm{CN}, \mathrm{CN}'$ can be verified analogously. □

Lemma 12. *Let $\mathcal{P}_n$ be the power set over n elements. Then $\mathrm{CN}(\mathcal{P}_{5k}) \leq 2k$ for all $k \geq 1$.*

Proof. The result follows directly from the fact that, by Example 2, $\mathrm{CN}(\mathcal{P}_5) = 2$ and from the sub-additivity of the compression number. □

We obtain an almost matching lower bound of $\mathrm{CN}(\mathcal{P}_{5k}) > k$ from the following result:

Lemma 13 (Floyd, Warmuth [2]). *For any class $\mathcal{C}$ of VC-dimension d it holds that $\mathrm{CN}(\mathcal{C}) > d/5$.*

We also observe certain monotonicity properties of OCN, OCN′, CN, and CN′, which will prove useful later.

Theorem 14. *Let X denote the domain of the classes $\mathcal{H}$ and $\mathcal{C} \subseteq \mathcal{H}$, and let $X' \subseteq X$. Then, $\mathrm{CN}(\mathcal{C}, \mathcal{H}) \geq \mathrm{CN}(\mathcal{C}_{|X'}, \mathcal{H}_{|X'})$ and $\mathrm{OCN}(\mathcal{C}, \mathcal{H}) \geq \mathrm{OCN}(\mathcal{C}_{|X'}, \mathcal{H}_{|X'})$.*

Proof. Let $\mathcal{C}' := \mathcal{C}_{|X'}$ and $\mathcal{H}' := \mathcal{H}_{|X'}$. For any hypothesis $H \in \mathcal{H}$, let H' denote its restriction to X'. The inequality for CN is pretty obvious. A $\mathcal{C}'$-realizable sample S' such that $X(S') \subseteq X'$ is $\mathcal{C}$-realizable, and $H \in \mathcal{H}$ is consistent with S' iff H' is consistent with S'. We may therefore compress S' to $S'' \subseteq S'$ and decompress S'' to some $H_t \in \mathcal{H}$ according to an optimal $(\mathcal{C}, \mathcal{H})$-scheme, and then consider H'_t as the decompression of S'' within the hypothesis class $\mathcal{H}'$. Since H_t is consistent with S', H'_t is consistent with S' too.

Next we verify the inequality for OCN. Let $H_1 < \ldots < H_m$ be the order over $\mathcal{H} = \{H_1, \ldots, H_m\}$ 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 order. Let S be a $\mathcal{C}$-realizable sample over the restricted domain X', and let $t \in \{1, \ldots, m\}$ be the largest index such that $H_t \in \mathrm{Cons}(S, \mathcal{H})$. The

definition of order compression schemes implies that $f(S) \subseteq S$ 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 $\tau \in \{1, \ldots, m'\}$. Note that $X(f(S)) \subseteq X(S) \subseteq X'$ so that a hypothesis H_i is consistent with $f(S)$ iff H'_i is consistent with $f(S)$. It follows that $\tau \in \{1, \ldots, m'\}$ is the largest index such that $H'_t = H'_{i(\tau)} \in \mathrm{Cons}(S, \mathcal{H}_{|X'})$. Moreover, $f(S) \subseteq S$ 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 $|f(S)| \leq \mathrm{OCN}(\mathcal{C}, \mathcal{H})$, which concludes the proof. □

The following example shows that it is essential that, when fixing the order over the hypotheses in $\mathcal{H}' = \mathcal{H}_{|X'}$ in the proof of Theorem 14, we have chosen the *latest* representative for $H' \in \mathcal{H}'$ in the sequence $H_1, \ldots, H_m$.

Example 15. *Let $X = \{1, \ldots, n\}$ and $\mathcal{C} = \mathcal{H} = \{\{1\}, \{2\}, \ldots, \{n\}, \emptyset\}$. We fix the following order over $\mathcal{H}$:*

$$\{1\},\ \{2\},\ \ldots,\ \{n\},\ \emptyset \tag{6}$$

We know from Example 6 that the corresponding proper order scheme has size 1. It follows that $\mathcal{C}$ is OC*-closed and the above order induces an optimal order scheme.*

Let $X' = X \setminus \{1\}$. Note that the empty concept over X' is represented twice in (6): by $\{1\}$ and by the empty set. Suppose we pick $\{1\}$ as the representative instead of $\emptyset$. Then we arrive at the following order over $\mathcal{H}' = \mathcal{H}_{|X'}$:

$$\emptyset,\ \{2\},\ \ldots,\ \{n\} \tag{7}$$

Consider now the sample $S' = \{(2, 0), \ldots, (n, 0)\}$. Similar as in the proper order compression scheme from Example 6, S' cannot be compressed to a proper subsample and thus the order scheme based on order (7) has size $n - 1$.

Corollary 16. *For every $X' \subseteq X$: $\mathrm{OCN}(\mathcal{C}) \geq \mathrm{OCN}(\mathcal{C}_{|X'})$, $\mathrm{OCN}'(\mathcal{C}) \geq \mathrm{OCN}'(\mathcal{C}_{|X'})$, $\mathrm{CN}(\mathcal{C}) \geq \mathrm{CN}(\mathcal{C}_{|X'})$, $\mathrm{CN}'(\mathcal{C}) \geq \mathrm{CN}'(\mathcal{C}_{|X'})$.*

Proof. The result for OCN is obtained from Theorem 14 as follows:

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

The proof for the other complexity measures is analogous. □

Definition 17. *Let K be a function that assigns a "complexity" $K(\mathcal{C})$ to each concept class $\mathcal{C}$. We say that K is* monotonic *if $\mathcal{C}' \subseteq \mathcal{C}$ implies that $K(\mathcal{C}') \leq K(\mathcal{C})$. We say that K is* twofold monotonic *if K is monotonic and, for every concept class $\mathcal{C}$ over X and every $X' \subseteq X$, it holds that $K(\mathcal{C}_{|X'}) \leq K(\mathcal{C})$.*

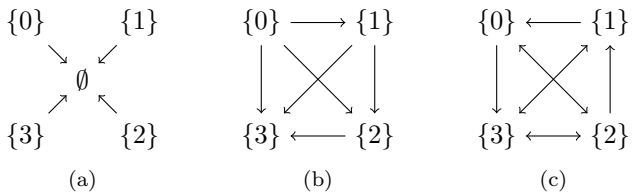
It is easy to see that OCN and OCN′ are monotonic. For $\mathcal{C}' \subseteq \mathcal{C}$, we get:

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

The analogous reasoning applies to CN and CN′. In combination with Corollary 16, we get:

Corollary 18. $\mathrm{OCN}, \mathrm{OCN}', \mathrm{CN}, \mathrm{CN}'$ *are twofold monotonic.*

Lemma 19. $\mathrm{OCN}'(\mathcal{C}) \leq |X| - 1$ *unless* $\mathcal{C} = 2^X$.



(a) (b) (c)

Figure 1: The compression graphs of the compression schemes from Example 6 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 the 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 $\{3\}$.)

PROOF. Choose the order $C_1, \ldots, C_N$ of the concepts C_i in $\mathcal{C}$ such that for all i there is some $x_i \in X$ with either $C_i \cup \{x_i\} \notin \{C_i, \ldots, C_N\}$ or $C_i \setminus \{x_i\} \notin \{C_i, \ldots, C_N\}$. From a sample consistent with C_i but not consistent with $C_{i+1}, \ldots, C_N$, we may remove a labeled example for x_i and obtain a subset of S of size at most $|X| - 1$ (possibly equal to S) which still is a teaching set for C_i with respect to $\{C_i, \ldots, C_N\}$. □

A useful tool for analyzing compression schemes in general and order compression schemes in particular is the "compression graph":

Definition 20. *Let (f, g) be a $(\mathcal{C}, \mathcal{H})$-scheme. The digraph $\mathcal{G}_{\text{comp}}(f, g) = (V, E)$, called the* 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 6 for the case $n = 4$.

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

Theorem 21. *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 order over 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 order 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

$$j = \max\{i : \ H_i \in \text{Cons}(S', \mathcal{H})\} \ . \tag{8}$$

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 $t \in \{1, \ldots, m\}$ be maximum such that $H_t \in \text{Cons}(S, \mathcal{H})$. 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 (8). 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 normalized 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 21 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 orders of V.

4. Order Compression Schemes and Teaching

The definition of order compression schemes bears some similarity to the model of *recursive teaching* [13], 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 $t \in \{1, \ldots, m\}$, 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 $t \in \{1, \ldots, m\}$, S_t is of largest cardinality among all $\mathcal{C}$-realizable inclusion-minimal teaching sets for H_t with respect to $\{H_t, H_{t+1}, \ldots, H_m\}$.

We say that a $(\mathcal{C}, \mathcal{H})$-scheme (f, g) is *$\mathcal{H}$-exhaustive* iff for all $H \in \mathcal{H}$ there exists a $\mathcal{C}$-realizable labeled sample S such that $g(f(S)) = H$.

Theorem 22. *There is an $\mathcal{H}$-exhaustive 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 order 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 $f(S_t)$ with $g(f(S_t)) = H_t$ 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 order. 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 (f, g) induced by the order $H_1 < \ldots < H_m$ is of size k. Note that each hypothesis in $\mathcal{H}$ is mapped to, because, for all t, $H_t = g(f(S_t))$. □

Theorem 22 leads to the following lower bound on OCN.

Lemma 23. *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 22, 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 23 can be strengthened as follows:

Theorem 24. $\mathrm{OCN}(\mathcal{C}) \geq \mathrm{RTD}^*(\mathcal{C})$.

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:

$$\begin{aligned} \mathrm{OCN}(\mathcal{C}) = \mathrm{OCN}(\mathcal{C}, \mathcal{H}) &\overset{Th.14}{\geq} \mathrm{OCN}(\mathcal{C}_{|X'}, \mathcal{H}_{|X'}) \\ &\geq \mathrm{OCN}(\mathcal{C}_{|X'}) \\ &\overset{L.23}{\geq} \mathrm{RTD}(\mathcal{C}_{|X'}) = \mathrm{RTD}^*(\mathcal{C}) \end{aligned}$$

This proves the theorem. □

Since $\mathrm{RTD}^*(\mathcal{C}) \geq \mathrm{VCD}(\mathcal{C})$ (see [6]), we immediately obtain the following corollary from Theorem 24:

Corollary 25. $\mathrm{OCN}(\mathcal{C}) \geq \mathrm{VCD}(\mathcal{C})$.

It is known from previous work that VCD can exceed RTD by an arbitrary amount [6]. Thus, Corollary 25 implies that also OCN can exceed RTD by an arbitrary amount.

Example 26 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, for the time being, it is not known whether the gap can be made larger than a factor of 1.5.

Example 26. *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 *[5]. In this particular example,* $\mathrm{RTD}(\mathcal{C}_{MW}) = 3$*, while* $\mathrm{VCD}(\mathcal{C}_{MW}) = 2$.

A proper sample compression scheme (f, g) of size 2 for $\mathcal{C}_{MW}$ can be designed as follows. Let a, b, c, d, e be any cyclic permutation of x_5, x_4, x_3, x_2, x_1. Thus, in the following description of a compression scheme, (a, b) can represent any solid edge in Fig. 2 (in clockwise direction). Similarly, (a, c) can represent any dashed edge.

Since samples of size 2 or less can be compressed and decompressed in an obvious fashion, we may assume a sample size of at least 3. The rules for compression given below will map S to a subsample S' of size precisely 2 such that S' corresponds either to a solid edge (a, b) or to a dashed edge (a, c) in Fig. 2. With these notations, the rules for decompression are as follows:

$$(a,1),(b,1) \mapsto \{a,b\} \quad \textit{and} \quad (a,0),(c,0) \mapsto \overline{\{a,c\}} = \{b,d,e\} \tag{9}$$

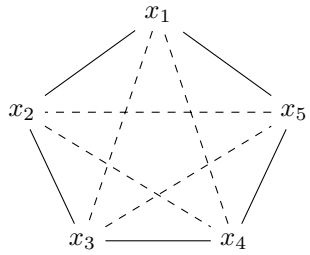
$$(a,1),(c,1) \mapsto \{a,c,e\} \quad \textit{and} \quad (a,0),(b,0) \mapsto \overline{\{a,b,c\}} = \{d,e\} \tag{10}$$

$$(a,0),(c,1) \mapsto \{c,d\} \quad \textit{and} \quad (a,1),(e,0) \mapsto \overline{\{b,e\}} = \{a,c,d\} \tag{11}$$

Let S be a sample of size at least 3. Let S_+ (resp. S_-) denote the set of 1-labeled (resp. 0-labeled) samples in S. The rules for compressing S to a subsample $S' = f(S)$ (with rules higher up having higher priority to apply) are as follows:

$\mathcal{C}_{MW}$	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.

Figure 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.

1. *If S_+ is of the form $\{(a,1),(b,1)\}$, then $S' = S_+$. Otherwise, if S_- is of the form $\{(a,0),(c,0)\}$, then $S' = S_-$.*
2. *If S contains $(a,0),(b,0)$ for some solid edge (a,b) but does not contain $(c,1)$, then $S' = \{(a,0),(b,0)\}$. Similarly, if S contains $(a,1),(c,1)$ for some dashed edge (a,c) but does not contain $(e,0)$, then $S' = \{(a,1),(c,1)\}$.*
3. *If S contains $(a,0),(b,0),(c,1)$ for some solid edge (a,b), then $S' = \{(a,0),(c,1)\}$. Similarly, if S contains $(a,1),(c,1),(e,0)$ for some dashed edge (a,c), then $S' = \{(a,1),(e,0)\}$.*

Note that compression rule 1 applies to any S of size 5 and to any S such that $|S_+| = |S_-| = 2$. Rule 2 applies to any S such that $|S_+| \geq 3$ or $|S_-| \geq 3$. Thus the first two rules cover completely the cases where $|S| \geq 4$. It is easy to check, that for any S of size 3, exactly one of the three compression rules applies. Thus the three rules completely determine how a sample of size at least 3 is actually compressed. Furthermore, it is easy to check that the three compression rules match with the aforementioned three decompression rules so as to form a valid compression scheme of size 2.

Note that the compression graph $\mathcal{G} = (V, E)$ for this scheme has cycles. For instance, there is a loop between the two hypotheses $C_1 = \{x_1, x_2\}$ and $C_{10} = \{x_1, x_2, x_4\}$; the edge $(C_{10}, C_1) \in E$ is witnessed by the sample $\{(x_1,1),(x_2,1)\}$, while $(C_1, C_{10}) \in E$ is witnessed by $\{(x_3,0),(x_5,0)\}$.

In fact, we already know that no sample compression scheme of size 2 for $\mathcal{C}_{MW}$ can be acyclic since $\mathrm{OCN}'(\mathcal{C}_{MW}) \geq \mathrm{OCN}(\mathcal{C}_{MW}) \geq \mathrm{RTD}(\mathcal{C}_{MW}) = 3$. The question whether $\mathrm{OCN}'(\mathcal{C}_{MW}) = 3$ can be answered affirmatively because the following reasoning 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.

Corollary 27. *There exists a family $(\mathcal{C}_d)_{d\geq 1}$ of concept classes such that $\mathrm{VCD}(\mathcal{C}_d) = 2d$ and $\mathrm{RTD}(\mathcal{C}_d) = \mathrm{OCN}(\mathcal{C}_d) = \mathrm{OCN}'(\mathcal{C}_d) = 3d$.*

Proof. Let $\mathcal{C}_d = \mathcal{C}_{MW} \uplus \ldots \uplus \mathcal{C}_{MW}$ be the direct sum of d disjoint duplicates of $\mathcal{C}_{MW}$. Since $\mathrm{VCD}(\mathcal{C}_{MW}) = 2$ and VCD is additive, it follows that $\mathrm{VCD}(\mathcal{C}_d) = 2d$. We know from Example 26 that $\mathrm{RTD}(\mathcal{C}_{MW}) = \mathrm{OCN}'(\mathcal{C}_{MW}) = 3$. Since RTD is additive, OCN' is sub-additive and $\mathrm{RTD} \leq \mathrm{OCN}'$, we may conclude from Lemma 9 that OCN' is additive on $\mathcal{C}_{MW}$, which yields $\mathrm{OCN}'(\mathcal{C}_d) = 3d$. □

Example 26 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 21, 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 $\mathcal{C}$. For an intersection-closed class $\mathcal{C}$ and $T \subseteq X$, $\langle T \rangle_\mathcal{C}$ 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 $\mathrm{VCD}(\mathcal{C})$ [18, 19].

A class $\mathcal{C}$ of VC-dimension d over a domain X of cardinality n is called *maximum* if $|\mathcal{C}| = \sum_{i=0}^{d} \binom{n}{i}$ [10, 11]. The following definition was introduced by Kuzmin and Warmuth [7]. 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:

$$\forall C \neq C' \in \mathcal{C}, \exists x \in r(C) \cup r(C') : C(x) \neq C'(x) \tag{12}$$

As shown in [7], 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 order $C_1 < \ldots < C_m$ of the concepts in $\mathcal{C} = \{C_1, \ldots, C_m\}$ is the following modification of (12):

$$\forall 1 \leq i < j \leq m, \exists x \in r(C_i) : C_i(x) \neq C_j(x) \tag{13}$$

For instance, the representation function resulting from the Tail Matching Algorithm [7] has the acyclic non-clashing property.

Theorem 28. *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 21 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 (13). 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 indices so that G is acyclic. □

The following result is immediate from Theorem 28 and Corollary 25:

Corollary 29. *Every intersection-closed or maximum class $\mathcal{C}$ is* OC*-closed and therefore satisfies*

$$\mathrm{OCN}'(\mathcal{C}) = \mathrm{OCN}(\mathcal{C}) = \mathrm{OCN}(\mathcal{C}, \mathcal{C}) = \mathrm{VCD}(\mathcal{C}) \ .$$

The proper order schemes for the classes mentioned in Theorem 28 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 30 (Dudley [12]). *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 $D_{\mathcal{F},h} = \{C_f \mid f \in \mathcal{F}\}$ is called a Dudley class*.*

We briefly note that the VC-dimension of $D_{\mathcal{F},h}$ is equal to the dimension of the vector space $\mathcal{F}$ [20]. 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., [21],
- unions of at most k intervals over $\mathbb{R}$,
- n-dimensional balls.

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

Lemma 31 (Ben-David and Litman [4]). *Dudley classes of VC-dimension k are embeddable in maximum classes of VC-dimension k.*

Corollary 32. *Let $\mathcal{C}$ be a Dudley class of VC-dimension k. Then,*

$$\mathrm{OCN}'(\mathcal{C}) = \mathrm{OCN}(\mathcal{C}) = \mathrm{VCD}(\mathcal{C}) \ . \tag{14}$$

Proof. Since $\mathrm{OCN}'(\mathcal{C}) \geq \mathrm{OCN}(\mathcal{C}) \geq \mathrm{VCD}(\mathcal{C})$ holds for any concept class, it suffices to show that $\mathrm{VCD}(\mathcal{C}) \geq \mathrm{OCN}'(\mathcal{C})$. Let $\mathcal{C}_{max} \supseteq \mathcal{C}$ be a maximum class of the same VC-dimension as $\mathcal{C}$ (in accordance with Lemma 31). Corollary 29 now yields

$$\mathrm{VCD}(\mathcal{C}) = \mathrm{VCD}(\mathcal{C}_{max}) = \mathrm{OCN}(\mathcal{C}_{max}, \mathcal{C}_{max}) \geq \min_{H \supseteq \mathcal{C}} \mathrm{OCN}(\mathcal{H}, \mathcal{H}) = \mathrm{OCN}'(\mathcal{C}) \ ,$$

and the corollary follows. □

As the proof of Corollary 32 shows, equality (14) is obtained whenever a class $\mathcal{C}$ is embeddable in a maximum class of the same VC-dimension. Another family which fits into this pattern consists of all classes of VC-dimension 1. Such classes are known to be contained in maximum classes of VC-dimension 1 [22]. Thus:

Corollary 33. *Let $\mathcal{C}$ be a concept class of VC-dimension 1. Then,*

$$\mathrm{OCN}'(\mathcal{C}) = \mathrm{OCN}(\mathcal{C}) = \mathrm{VCD}(\mathcal{C}) \ .$$

In combination, we obtain:

Corollary 34. $\mathrm{OCN}'(\mathcal{C}) = \mathrm{OCN}(\mathcal{C}) = \mathrm{VCD}(\mathcal{C})$ *provided that* $\mathcal{C}$ *belongs to at least one of the families* $\mathcal{F}_\cap$, $\mathcal{F}_{max}$, $\mathcal{F}_{Dudley}$ *or* $\mathcal{F}_1$.

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 35. *Suppose that* $\mathcal{C}$ *is intersection-closed. Then,*

$$\mathrm{OCN}'(\mathrm{DIFF}^{\leq d}(\mathcal{C})) \leq \mathrm{OCN}(\mathrm{DIFF}^{\leq d}(\mathcal{C}), \mathrm{DIFF}^{\leq d}(\mathcal{C})) \leq d \cdot \mathrm{VCD}(\mathcal{C}) \ .$$

Proof. As already noted in (1), the first inequality holds for any concept class. As for the second inequality, we have to design a proper order scheme for $\mathcal{H} := \mathrm{DIFF}^{\leq d}(\mathcal{C})$ of size less than or equal to $d \cdot \mathrm{VCD}(\mathcal{C})$.

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 [6]). We can represent $C \in \mathcal{H}$ as

$$C = C_1 \setminus \overbrace{(C_2 \setminus (\cdots (C_{d-1} \setminus C_d) \cdots))}^{=:D_1} \tag{15}$$

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 (15) is minimal in the sense that $C_j = \langle C_j \setminus D_j \rangle_\mathcal{C}$ holds for all $1 \leq j \leq d$.

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 $S_1 = \{x \mid (x, 1) \in S\}$ 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 $S_2 = \{x \in C_1 \mid (x, 0) \in S\}$ 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 $S_3 = \{x \in C_2 \mid (x, 1) \in S\}$ and $C_3 = \langle S_3 \rangle_\mathcal{C}$. Proceed iteratively 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 with respect to $\{H_t, \ldots, H_m\}$. 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. Separating OCN′ from OCN

In this section, we will focus on concept classes over the domain $\{w, x, y, z\}$ of size 4. As visualized in Figure 3, we can represent concepts over $\{w, x, y, z\}$ as nodes in the 4-dimensional Boolean cube. The latter, denoted W^4 in what follows, is drawn as a pair of 3-dimensional cubes, one for $w = 0$ and the other one for $w = 1$. The edges between corresponding nodes in the subcubes $w = 0$ and $w = 1$, respectively, are (in most cases) not explicitly shown in order to avoid clutter.

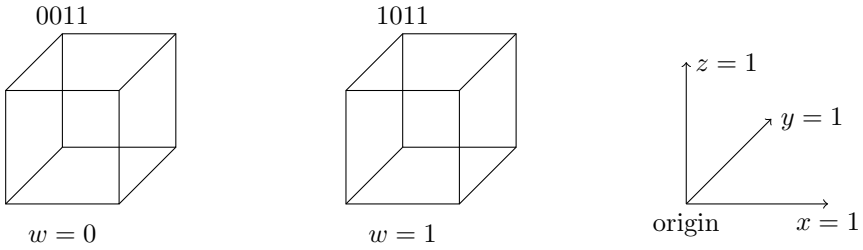


Figure 3: A coordinate system that lets us graphically represent concepts over $\{w, x, y, z\}$ as nodes in the 4-dimensional Boolean cube W^4.

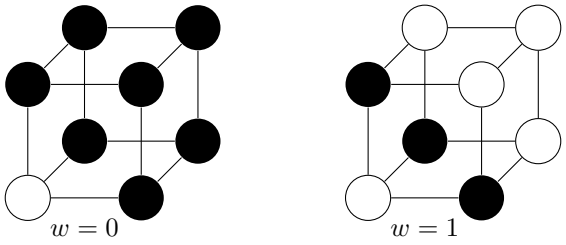


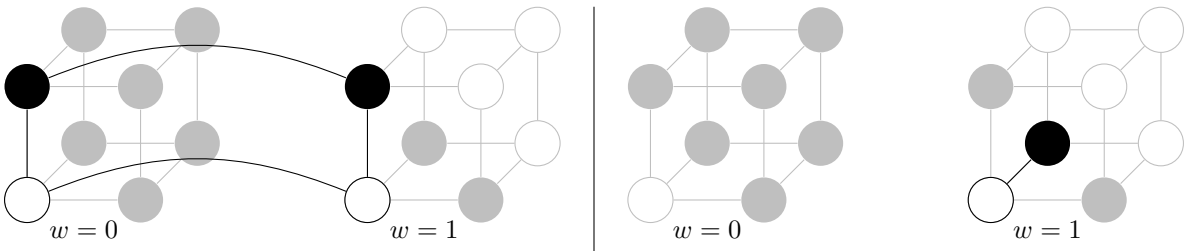
Figure 4: The concept class $\mathcal{C}_1$. The edges between the subcubes $w = 0$ and $w = 1$ are not shown.

In what follows, the nodes in W^4 that represent the concepts of a given class are drawn as black disks. White disks represent concepts outside the class. An example, the concept class $\mathcal{C}_1$, is shown in Figure 4.

Each labeled sample S induces a subcube in W^4 that includes all nodes that are consistent with S. Vice versa, each subcube is induced by a sample. In the following we will identify subcubes (especially edges) with the samples that induce them. For example, using the concept class $\mathcal{C}_1$, after seeing the sample $S = \{(x, 0), (y, 0)\}$ we know that all consistent nodes lie in the subcube from Figure 5a. Since this subcube contains more than one black node, S is not a teaching set. On the other hand, $S' = \{(w, 1), (x, 0), (z, 0)\}$ is a teaching set for the concept $(1, 0, 1, 0)$ with respect to $\mathcal{C}_1$ as demonstrated by Figure 5b. Note that a sample with k elements induces a subcube of dimension $4 - k$.

A class $\mathcal{C}$ has VC-dimension 2 (or less) iff the following holds: for every collection of 3 instances, there exists a 3-bit pattern which cannot be realized by any concept in $\mathcal{C}$. As for W^4 this means that, for every $u \in \{w, x, y, z\}$, there exists an edge along dimension u whose endpoints are both white. An inspection of Figure 4 shows that $\mathcal{C}_1$ has VC-dimension 2.

Suppose that the class $\mathcal{C}$ over domain $\{w, x, y, z\}$ results from W^4 by the deletion of a minimal set of nodes such that one edge along each of the 4 dimensions is missing (like, for instance, the class $\mathcal{C}_1$ from Figure 4). It follows that $\mathcal{C}$ is a maximal class of VC-dimension 2, i.e., adding any concept to $\mathcal{C}$ would increase the VC-dimension.



(a) Subcube induced by $S = \{(x, 0), (y, 0)\}$

(b) Subcube induced by $S' = \{(w, 1), (x, 0), (z, 0)\}$

Figure 5: The subcubes induced by labeled samples S and S' in the class $\mathcal{C}_1$.

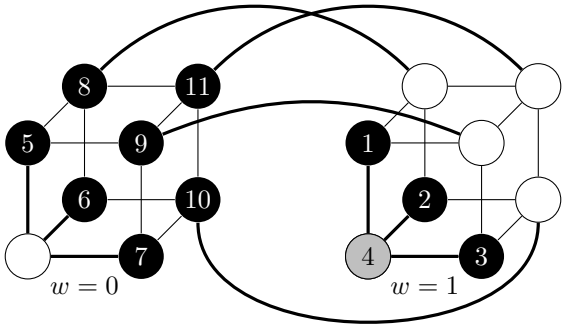


Figure 6: The concept class $\mathcal{C}_1$ (the black nodes) and its extension $\mathcal{H}_1$ (the black nodes and the gray node). The critical edges are shown as thick lines.

Since $\mathrm{OCN}'(\mathcal{C}) = \mathrm{OCN}(\mathcal{H}, \mathcal{H})$ (for a suitably chosen $\mathcal{H} \supseteq \mathcal{C}$) is lower-bounded by $\mathrm{VCD}(\mathcal{H})$, we draw the following conclusion: if $\mathrm{OCN}'(\mathcal{C}) = 2$, then $\mathrm{OCN}(\mathcal{C}, \mathcal{C}) = 2$, i.e., there must be a proper OCS of size 2 for $\mathcal{C}$. As for this proper scheme, the following must hold: there must exist a concept $C \in \mathcal{C}$ (namely the leftmost concept in the order inducing the scheme) such that every teaching set of size 4 or 3 contains a teaching set of size 2 as a subset. In other words: every edge (= a subcube of dimension 1) with precisely one black node must be part of a square (= a subcube of dimension 2) that still does not contain any other black node. Edges violating this condition will henceforth be called *critical edges*.

From Figure 6 (where the critical edges are highlighted) one can see that $\mathrm{OCN}'(\mathcal{C}_1) > 2$ and, by Lemma 19, $\mathrm{OCN}'(\mathcal{C}_1) = 3$. However, as illustrated in Figure 6, we get $\mathrm{OCN}(\mathcal{C}_1) = 2$. In other words, $\mathcal{C}_1$ can be extended to a hypothesis class, say $\mathcal{H}_1 \supset \mathcal{C}_1$, such that $\mathrm{OCN}(\mathcal{C}_1, \mathcal{H}_1) = 2$. As shown in Figure 6, $\mathcal{H}_1$ consists of all black nodes (corresponding to the concepts from $\mathcal{C}_1$) plus an additional node that is shown in gray. The figure also shows the critical edges (drawn as thick edges) which allowed us to argue that $\mathrm{OCN}(\mathcal{C}_1, \mathcal{C}_1) \geq 3$. The main difference between $\mathcal{C}_1$ and $\mathcal{H}_1$ is that the gray node neutralizes the three critical edges in the subcube $w = 1$. To see that the order indicated by the numbering in the figure leads to a $(\mathcal{C}_1, \mathcal{H}_1)$-OCS of size 2, one argues as follows:

- Let S be any $\mathcal{C}_1$-realizable sample set that is inconsistent with nodes 5 through 11. S must contain $(w, 1)$. The nodes 1, 2, 3, 4 correspond to $(0, 0, 1)$, $(0, 1, 0)$, $(1, 0, 0)$, $(0, 0, 0)$ in (x, y, z)-coordinates, i.e., to the singletons and the empty set over $\{x, y, z\}$. For this subclass, the given order yields an order compression scheme of size 1, cf. Example 6. Thus, S will be compressed to a set of up to two elements, one of which is $(w, 1)$.
- Let S be any $\mathcal{C}_1$-realizable sample set that is consistent with at least one of the nodes 5 through 11, which form a proper subset of the boolean cube over $\{x, y, z\}$. By Lemma 19, S is compressed to a set of size at most 2.

The preceding discussion can be summarized as follows:

Theorem 36. *The concept class $\mathcal{C}_1$ from Figure 4 separates* OCN' *from* OCN. *More precisely:* $\mathrm{OCN}(\mathcal{C}_1) = 2$ *and* $\mathrm{OCN}'(\mathcal{C}_1) = 3$.

Note that Theorem 36 also shows that the VC-dimension of the optimal hypothesis class $\mathcal{H}$ is not a lower bound for $\mathrm{OCN}(\mathcal{C})$, while it is a lower bound for $\mathrm{OCN}'(\mathcal{C}) = \mathrm{OCN}(\mathcal{H}, \mathcal{H})$ by Corollary 25.

6.1. Proper Sub-Additivity of OCN'

We observed in Corollary 11 that CN is sub-additive. Since $\mathrm{CN}(\mathcal{P}_{5k}) \leq 2k$ by Lemma 12, and $\mathcal{P}_{5k}$ is the direct sum of $5k$ disjoint copies of $\mathcal{P}_1$, the sub-additivity of CN is proper. In particular, we have shown CN to be "below additivity by a factor of 2/5."

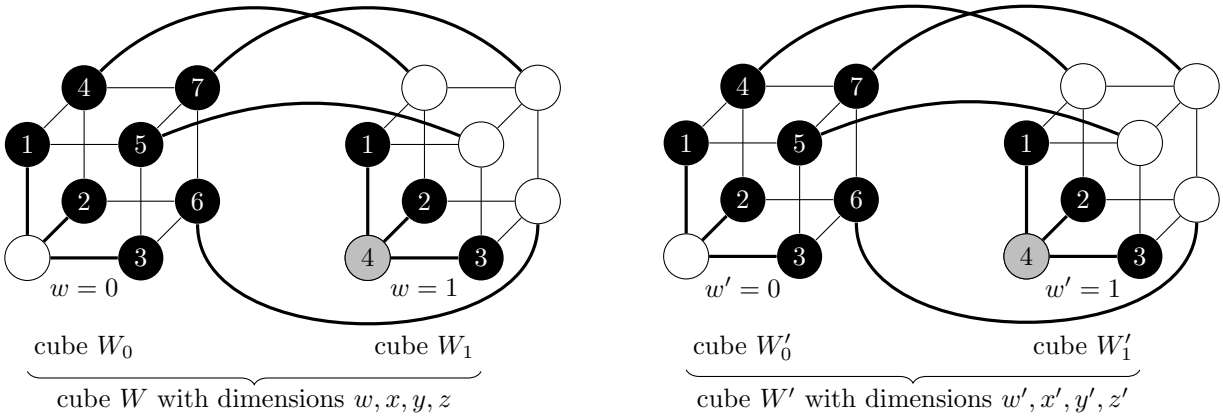


Figure 7: The concept class $\mathcal{C}_2$ (pairs of black nodes) and its extension $\mathcal{H}_2$ (pairs of black nodes plus the pair of gray nodes).

By contrast, the power sets $\mathcal{P}_n$ do not suffice to show that OCN or OCN$'$ are properly sub-additive. If at all, the latter can only be witnessed by a pair of concept classes of which at least one satisfies OCN $>$ VCD (OCN$'$ $>$ VCD, respectively). This is an immediate consequence of the following three facts: (i) OCN and OCN$'$ are sub-additive by Corollary 11, (ii) VCD is additive, and (iii) OCN$'$ $\geq$ OCN $\geq$ VCD. Note that the VC-dimension does not bound the compression number CN from below, which is why proper sub-additivity is easier to verify for CN than for OCN or OCN$'$. Thus the problem of quantifying the sub-additivity of OCN and OCN$'$ is closely related to the problem of determining by which factor the order compression number can exceed the VC-dimension.

Since OCN$'(\mathcal{C}_1) = 3$, the following result implies that OCN$'$ is not additive on $\mathcal{C}_1$.

Theorem 37. *Let $\mathcal{C}_1$ be the concept class from Figure 4. Let $\mathcal{C}_2$ be the direct sum of two disjoint duplicates of $\mathcal{C}_1$. Then,* OCN$'(\mathcal{C}_2) = 5$.

Proof. First we show the upper bound OCN$'(\mathcal{C}_2) \leq 5$.

We may visualize $\mathcal{C}_2$ as the set of pairs of black nodes in Figure 7 where the first black node is taken from the cube W with dimensions w, x, y, z and the second one is taken from the cube W' with dimensions w', x', y', z'. As shown in Figure 7, W (resp. W') decomposes into the subcubes W_0 and W_1 (resp. W'_0 and W'_1). Let H_0 be the hypothesis that is represented by the pair of gray nodes in Figure 7. Let $\mathcal{H}_2 := \mathcal{C}_2 \cup \{H_0\}$. It suffices to show that OCN$(\mathcal{H}_2, \mathcal{H}_2) \leq 5$. To this end, we fix an order on $\mathcal{H}_2$ as follows:

Group 1: The smallest elements are those pairs that consist of one black node from W_1 and one black node from W'_1. We can fix an arbitrary order of these pairs.

Group 2: The next element in the order is H_0.

Group 3: Group 3a is formed by pairs with one black node from W_0 (in the order that is indicated in Figure 7) and one black node from W'_1 (in an arbitrary order). Group 3b is analogously formed by pairs with one black node from W'_0 and one black node from W_1.

Group 4: Largest in the order are the remaining hypotheses, i.e., the pairs with one black node from W_0 and one black node from W'_0.

It follows from Lemma 19 and (3) that the hypotheses in Group 4 have a proper order compression scheme of size at most 4. Thus, it suffices to show that the minimal teaching sets for the hypotheses from the first three groups (always relative to the given order over $\mathcal{H}_2$) have size at most 5. We make use of the fact that any teaching set decomposes into a set S of labeled examples from $X \times \{0, 1\}$ for $X := \{w, x, y, z\}$ (so that S can be viewed as a subcube in W) and a set S' of labeled examples from $X' \times \{0, 1\}$ for $X' := \{w', x', y', z'\}$

(so that S' can be viewed as a subcube in W'). $|S|$ and $|S'|$ are both clearly bounded by $|X| = |X'| = 4$. However, it is easy to see that minimal teaching sets even satisfy $|S|, |S'| \leq 3$. In the following discussion of the first three groups, we suppose we are given a teaching set $S \cup S'$ for a hypothesis H in the group such that $|S| = |S'| = 3$. Thus, S induces an edge e in the cube W and S' induces an edge e' in the cube W'. Then, we argue that one of these edges can be extended to a square (or an even larger subcube) without ceasing to be a teaching set for H. Thus $S \cup S'$ contains a subset of size 5 that is a teaching set for H.

Group 1: It is easy to see that e must be an edge of the cube W_1. Similarly, e' must be an edge of the cube W_1'. It cannot be the case that both edges are critical[2] since, otherwise, $S \cup S'$ would be consistent with H_0 (= the pair of gray nodes) in Group 2. But any non-critical edge can be extended to a square without destroying the "teaching set" property.

Group 2: Group 2 consists just of H_0 (the pair of gray nodes). Then e is either an edge in W_1 (containing the gray node) or the edge that connects the gray node with the white node in W_0. An analogous statement applies to e'. If one of these edges contains a white node, then this edge already forms a teaching set (so that the other edge is superfluous) because $\mathcal{H}_2$ contains only one concept including a gray node. If none of e and e' contains a white node, then e can be replaced by the cube W_1 and e' can be replaced by the cube W_1' (because the black nodes from Group 1 are already out of the game, and $\mathcal{H}_2$ does not contain a combination of a black node with a gray node).

Group 3: We may focus on the hypotheses in the Group 3a because the reasoning for the hypotheses in Group 3b would be similar. The edge e must represent a teaching set for a black node w_0 in W_0. Let us consider first the case where e connects w_0 with a node in W_1. The order over the black nodes in W_0 (indicated by the numbers in Figure 7) is such that w_0 must have an "effectively white" neighbor v_0 within W_0 (where v_0 possibly is a formerly black node that precedes w_0 in the order so that all combinations of v_0 and a black node in W_1' are already out of the game). Now we can replace the edge e by the square formed by e and the edge with endpoint v_0 that is parallel to e. (Here we profit from the fact that the hypotheses from Groups 1 and 2 are out of the game already.) We finally have to consider the case where e is an edge in the cube W_0. But then we can readily extend it to the square formed by e and the corresponding edge in W_1.

It follows that $\mathrm{OCN}(\mathcal{H}_2, \mathcal{H}_2) \leq 5$, and we proceed with the proof of the lower bound $\mathrm{OCN}'(\mathcal{C}) \geq 5$.

If we use $\mathcal{C}_2$ as the hypothesis class, then there exists an inclusion-minimal teaching set of size 6 for every hypothesis from $\mathcal{H} = \mathcal{C}_2$, as witnessed by the "critical edges", which would yield $\mathrm{OCN}(\mathcal{H}, \mathcal{H}) = \mathrm{OCN}(\mathcal{C}_2, \mathcal{C}_2) = 6$. Thus $\mathcal{H} = \mathcal{C}_2$ is not the optimal choice. If $\mathcal{H}$ is chosen optimally, i.e. such that $\mathrm{OCN}'(\mathcal{C}_2) = \mathrm{OCN}(\mathcal{H}, \mathcal{H})$, then $\mathcal{H}$ contains at least one additional hypothesis that was not already in $\mathcal{C}_2$. Using Corollary 25 and the observation that $\mathcal{C}_2$ is a maximal class of VC-dimension 4, we conclude that $\mathrm{OCN}'(\mathcal{C}_2) = \mathrm{OCN}(\mathcal{H}, \mathcal{H}) \geq \mathrm{VCD}(\mathcal{H}) \geq 5$. □

Since OCN' is sub-additive according to Corollary 11, Theorem 37 implies the following result:

Corollary 38. *Let $\mathcal{C}_1$ be the concept class from Figure 4. Let $\mathcal{C}_k$ be the direct sum of k disjoint copies of $\mathcal{C}_1$. Then,* $\mathrm{OCN}'(\mathcal{C}_{2k}) \leq 5k$.

Since $2k \cdot \mathrm{OCN}'(\mathcal{C}_1) = 6k$, Corollary 38 shows that OCN' is "separated away" from additivity on $\mathcal{C}_1$ by a factor of 5/6 even when the number of disjoint duplicates becomes arbitrarily large.

6.2. Minimality of the Class Separating OCN' *from* OCN

It is well known that RTD and VCD coincide on all classes over a domain of size 4 or less [5]. The next result shows that no class over a domain of size 3 or less separates OCN' from VCD (implying that the complexity measures $\mathrm{VCD}, \mathrm{RTD}, \mathrm{RTD}^*, \mathrm{OCN}, \mathrm{OCN}'$ coincide on such classes):

[2] = drawn as a thick edge in Figure 7

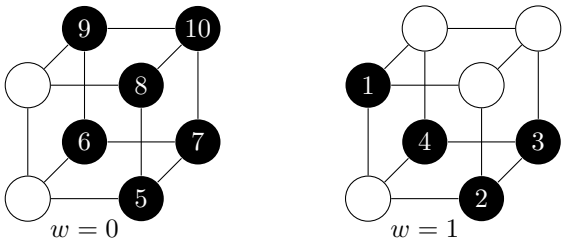


Figure 8: A white subgraph V of W^4 with two connected components of size 3.

Lemma 39. *Let $\mathcal{C}$ be a concept class over a domain of size 3 or less. Then,* $\mathrm{OCN}'(\mathcal{C}) = \mathrm{VCD}(\mathcal{C})$.

Proof. If $\mathrm{VCD}(\mathcal{C}) = 1$, then, according to Corollary 33, $\mathrm{OCN}'(\mathcal{C}) = 1$. If $\mathrm{VCD}(\mathcal{C}) = 3$, then $\mathrm{OCN}'(\mathcal{C}) = 3$ (because we assumed that $|X| \leq 3$). Suppose now that $\mathrm{VCD}(\mathcal{C}) = 2$. It follows that either $|X| = 2$ and $\mathcal{C} = 2^X$ or $|X| = 3$ and $\mathcal{C} \neq 2^X$. In the former case it is obvious that $\mathrm{OCN}'(\mathcal{C}) = 2$. In the latter case, $\mathrm{OCN}'(\mathcal{C}) = 2$ is obtained from Lemma 19. □

So far we know that 4 is the smallest domain size allowing for a separation of OCN' and VCD. One may ask whether there is a smaller concept class than $\mathcal{C}_1$, which has size 10, over a domain of size 4 that still separates OCN' from VCD. In the sequel, we will answer this question to the negative.

Let $\mathcal{C}$ be a class over domain $\{w, x, y, z\}$. A simple reasoning (similar to the one in the proof of Lemma 39) shows that $\mathcal{C}$ can possibly separate OCN' from VCD only if $\mathrm{VCD}(\mathcal{C}) = 2$. So suppose that this is the case and that $\mathcal{C}$ is a maximal class, i.e., it results from W^4 by the deletion of a minimal set of nodes such that one edge along each of the 4 dimensions is missing.

Our goal is to show the following

Claim: Either $\mathcal{C} = \mathcal{C}_1$ (up to automorphisms of W^4) or $\mathrm{OCN}'(\mathcal{C}) = 2$.

It is easy to see that every proper subclass of $\mathcal{C}_1$ has a proper order compression scheme of size 2. Thus, the claim implies that at least 10 concepts over a domain of size 4 are required for the separation of OCN' from VCD. We still have to prove the claim. To this end, let $V = W^4 \setminus \mathcal{C}$ denote the set of white nodes. In the sequel, we will identify V with its induced subgraph of W^4. Let s denote the size of the largest connected component of V. It is easy to see that $s \geq 3$. In other words, V must contain a path of length 2.

Suppose first that $s = 3$, i.e., a largest connected component L_1 in V forms a path of length 2. It is easy to see (simply because W^4 does not leave much space) that there can be only one other connected component in V, say L_2. L_2 must contain the remaining 2 of the 4 deleted edges thereby forming a path of length 2 as well. As illustrated in Figure 8, there exists (up to automorphisms of W^4) only one choice for V.[3] An inspection of the concept class $\mathcal{C} = W^4 \setminus V$ in Figure 8 shows that there exists an order over the black nodes (indicated by the numbers in Figure 8) such that every edge, whose endpoints are a black node numbered i and either a white node or another black node numbered $j < i$, can be extended to a square not containing any black nodes with numbers larger than i. As explained earlier in this section, this means that $\mathrm{OCN}'(\mathcal{C}) = 2$.

Second, consider the case $s \geq 4$. A connected component of size 4 must either form a path P_3 of length 3 or a star S_3.

Suppose first that it forms a star S_3. But then, up to automorphisms of W^4, $\mathcal{C}$ must coincide with the class $\mathcal{C}_1$ in Figure 4.[4]

Suppose now that a component of V contains P_3. It is easy to see that P_3, in combination with an edge along each of the 4 dimensions, leads to a component of a size exceeding 4. It is furthermore easy to see that one of the following cases must occur:

[3] The middle nodes of the two connected components, the nodes $(0, 0, 0, 0)$ and $(1, 1, 1, 1)$ in Figure 8, must necessarily be located at opposing ends of W^4 because, otherwise, a connected component of size exceeding 3 will result.

[4] The edge along the dimension not covered by S_3 must involve a node that is located opposite to the center of S_3 because, otherwise, a larger connected component will result.

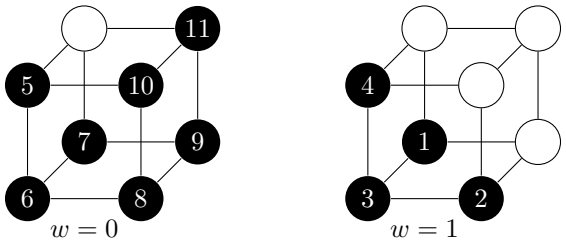


Figure 9: A connected white subgraph whose complement is a concept class of order compression number 2.

Case 1 V has precisely one connected component that forms a path of length 4 with one edge along each of the 4 dimensions.
It follows that there are $11 = \binom{4}{0} + \binom{4}{1} + \binom{4}{2}$ nodes in $\mathcal{C} = W^4 \setminus P_4$ so that $\mathcal{C}$ is a maximum class of VC-dimension 2. Thus, $\mathrm{OCN}'(\mathcal{C}) = 2$.

Case 2 One of the connected components of V contains S_3 as a partial subgraph.
In this case, $\mathcal{C}$ looks (up to automorphisms of W^4) like the concept class shown in Figure 9. As the numbering of the black nodes in this figure indicates, there is an order over these nodes that witnesses $\mathrm{OCN}'(\mathcal{C}) = 2$.

The discussion shows that the above claim is true. Since $\mathrm{OCN}'(\mathcal{C}) = 2$ for each proper subclass of $\mathcal{C}_1$, we now arrive at the following result:

Theorem 40. 1. *Let $\mathcal{C}$ be a concept class over a domain of size 4 that separates* OCN' *from* VCD. *Then $\mathcal{C}$ equals $\mathcal{C}_1$ up to automorphisms of W^4. In particular, $|\mathcal{C}| = 10$.*
2. *Let $\mathcal{C}$ be a concept class over a domain of size 4 or less. Then,* $\mathrm{OCN}(\mathcal{C}) = \mathrm{VCD}(\mathcal{C})$.

7. Conclusions

Order compression schemes obey a very simple structure and exhibit interesting connections to teaching and graph theory. Furthermore, in many 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 $\mathrm{OCN}(\mathcal{C}, \mathcal{H}) = \mathrm{OCN}(\mathcal{C})$. Further interesting questions are:

- Is there a concept class for which OCN exceeds VCD by more than a factor of 1.5?
- Are there any further interesting conditions on concept classes under which OCN coincides with any of the parameters VCD, CN, and RTD?
- If $\mathcal{C}$ and $\mathcal{H}$ are given, how to find an order over $\mathcal{H}$ that induces the best possible order compression scheme for $\mathcal{C}$, without trying out all possible orders?
- For which families of concept classes are OCN, OCN', CN, CN' additive?
- Are there any interesting bounds on the amount of subadditivity of OCN and OCN'?

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

References

[1] N. Littlestone, M. K. Warmuth, Relating data compression and learnability, Tech. rep., University of California at Santa Cruz (1986).

[2] S. Floyd, M. K. Warmuth, Sample compression, learnability, and the Vapnik-Chervonenkis dimension, Machine Learning 21 (3) (1995) 269–304.

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, M. K. Warmuth, Learnability and the Vapnik-Chervonenkis dimension, J. ACM 36 (4) (1989) 929–965.

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

[5] T. Doliwa, G. Fan, H. U. Simon, S. Zilles, Recursive teaching dimension, VC-dimension, and sample compression, To appear in: Journal of Machine Learning Research.

[6] T. Doliwa, H.-U. Simon, S. Zilles, Recursive teaching dimension, learning complexity, and maximum classes, in: ALT, 2010, pp. 209–223.

[7] D. Kuzmin, M. K. Warmuth, Unlabeled compression schemes for maximum classes, Journal of Machine Learning Research 8 (2007) 2047–2081.

[8] B. I. Rubinstein, J. H. Rubinstein, A geometric approach to sample compression, Journal of Machine Learning Research 13 (2012) 1221–1261.

[9] R. Livni, P. Simon, Honest compressions and their application to compression schemes, in: Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013), JMLR Workshop and Conference Proceedings, 2013, pp. 77–92.

[10] N. Sauer, On the density of families of sets, J. Comb. Theory, Ser. A 13 (1) (1972) 145–147.

[11] E. Welzl, Complete range spaces, Unpublished manuscript (1987).

[12] R. M. Dudley, A course on empirical processes, Lecture Notes in Mathematics 1097 (1984) 1–142.

[13] S. Zilles, S. Lange, R. Holte, M. Zinkevich, Models of cooperative teaching and learning, Journal of Machine Learning Research 12 (2011) 349–384.

[14] M. Darnstädt, T. Doliwa, H. U. Simon, S. Zilles, Order compression schemes, in: Proceedings of the 24th International Conference on Algorithmic Learning Theory (ALT 2013), Vol. 8139 of Lecture Notes in Computer Science, Springer, 2013, pp. 173–187.

[15] S. A. Goldman, M. J. Kearns, On the complexity of teaching, J. Comput. Syst. Sci. 50 (1) (1995) 20–31.

[16] A. Shinohara, S. Miyano, Teachability in computational learning, New Generation Computing 8 (4) (1991) 337–347.

[17] G. Fan, A graph-theoretic view of teaching, M.Sc. Thesis, University of Regina (2012).

[18] B. K. Natarajan, On learning boolean functions, in: Proceedings of the 19th Annual Symposium on Theory of Computing, 1987, pp. 296–304.

[19] D. P. Helmbold, R. H. Sloan, M. K. Warmuth, Learning nested differences of intersection-closed concept classes, Machine Learning 5 (1990) 165–196.

[20] R. Wenocur, R. Dudley, Some special Vapnik-Chervonenkis classes, Discrete Mathematics 33 (3) (1981) 313 – 318.

[21] E. Alpaydin, Introduction to Machine Learning, 2nd ed., MIT Press, 2010.

[22] E. Welzl, G. Wöginger, On Vapnik-Chervonenkis dimension one, unpublished manuscript (1987).