Journal of Machine Learning Research 15 (2014) 3107-3131 Submitted 1/13; Revised 5/14; Published 10/14

Recursive Teaching Dimension, VC-Dimension and Sample Compression

Thorsten Doliwa THORSTEN.DOLIWA@RUB.DE
Faculty of Mathematics
Ruhr-Universität Bochum, D-44780 Bochum, Germany

Gaojian Fan GAOJIAN@UALBERTA.CA
Department of Computing Science
University of Alberta, Edmonton, AB, T6G 2E8, Canada

Hans Ulrich Simon HANS.SIMON@RUB.DE
Horst-Görtz Institute for IT Security and Faculty of Mathematics
Ruhr-Universität Bochum, D-44780 Bochum, Germany

Sandra Zilles ZILLES@CS.UREGINA.CA
Department of Computer Science
University of Regina, Regina, SK, S4S 0A2, Canada

Editor: Manfred Warmuth

Abstract

This paper is concerned with various combinatorial parameters of classes that can be learned from a small set of examples. We show that the recursive teaching dimension, recently introduced by Zilles et al. (2008), is strongly connected to known complexity notions in machine learning, e.g., the self-directed learning complexity and the VC-dimension. To the best of our knowledge these are the first results unveiling such relations between teaching and query learning as well as between teaching and the VC-dimension. It will turn out that for many natural classes the RTD is upper-bounded by the VCD, e.g., classes of VC-dimension 1, intersection-closed classes and finite maximum classes. However, we will also show that there are certain (but rare) classes for which the recursive teaching dimension exceeds the VC-dimension. Moreover, for maximum classes, the combinatorial structure induced by the RTD, called teaching plan, is highly similar to the structure of sample compression schemes. Indeed one can transform any *repetition-free teaching plan* for a maximum class $\mathcal{C}$ into an *unlabeled sample compression scheme* for $\mathcal{C}$ and vice versa, while the latter is produced by (i) the corner-peeling algorithm of Rubinstein and Rubinstein (2012) and (ii) the tail matching algorithm of Kuzmin and Warmuth (2007).

Keywords: recursive teaching, combinatorial parameters, Vapnik-Chervonenkis dimension, upper bounds, compression schemes, tail matching algorithm

1. Introduction

In the design and analysis of machine learning algorithms, the amount of training data that needs to be provided for the learning algorithm to be successful is an aspect of central

importance. In many applications, training data is expensive or difficult to obtain, and thus input-efficient learning algorithms are desirable. In computational learning theory therefore, one way of measuring the complexity of a concept class is to determine the worst-case number of input examples required by the best valid learning algorithm. What is a valid learning algorithm depends on the underlying model of learning. We refer to this kind of complexity measure as *information complexity*. For example, in PAC-learning (Valiant, 1984), the information complexity of a concept class $\mathcal{C}$ is the worst-case sample complexity a best possible PAC learner for $\mathcal{C}$ can achieve on all concepts in $\mathcal{C}$. In query learning (Angluin, 1988), it is the best worst-case number of queries a learner would have to ask to identify an arbitrary concept in $\mathcal{C}$. In the classical model of teaching (Goldman and Kearns, 1995; Shinohara and Miyano, 1991), the information complexity of $\mathcal{C}$ is given by its teaching dimension, i.e., the largest number of labeled examples that would have to be provided for distinguishing any concept in $\mathcal{C}$ from all other concepts in $\mathcal{C}$.

Besides the practical need to limit the required amount of training data, there are a number of reasons for formally studying information complexity. Firstly, a theoretical study of information complexity yields formal guarantees concerning the amount of data that needs to be processed to solve a learning problem. Secondly, analyzing information complexity often helps to understand the structural properties of concept classes that are particularly hard to learn or particularly easy to learn. Thirdly, the theoretical study of information complexity helps to identify connections between various formal models of learning, for example if it turns out that, for a certain type of concept class, the information complexity under learning model A is in some relationship with the information complexity under model B. This third aspect is the main motivation of our study.

In the past two decades, several learning models were defined with the aim of understanding in which way a low information complexity can be achieved. One such model is learning from partial equivalence queries (Maass and Turán, 1992), which subsume all types of queries for which negative answers are witnessed by counterexamples, e.g., membership, equivalence, subset, superset, and disjointness queries (Angluin, 1988). As lower bounds on the information complexity in this query model (here called query complexity) hold for numerous other query learning models, they are particularly interesting objects of study. Even more powerful are self-directed learners (Goldman et al., 1993). Each query of a self-directed learner is a prediction of a label for an instance of the learner's choice, and the learner gets charged only for wrong predictions. The query complexity in this model lower-bounds the one obtained from partial equivalence queries (Goldman and Sloan, 1994).

Dual to the models of query learning, in which the learner actively chooses the instances it wants information on, the literature proposes models of teaching (Goldman and Kearns, 1995; Shinohara and Miyano, 1991), in which a helpful teacher selects a set of examples and presents it to the learner, again aiming at a low information complexity. A recent model of teaching with low information complexity is recursive teaching, where a teacher chooses a sample based on a sequence of nested subclasses of the underlying concept class $\mathcal{C}$ (Zilles et al., 2008). The nesting is defined as follows. The outermost "layer" consists of all concepts in $\mathcal{C}$ that are easiest to teach, i.e., that have the smallest sets of examples distinguishing them from all other concepts in $\mathcal{C}$. The next layers are formed by recursively repeating this process with the remaining concepts. The largest number of examples required for teaching at any layer is the recursive teaching dimension (RTD) of $\mathcal{C}$. The RTD substantially

reduces the information complexity bounds obtained in previous teaching models. It lower bounds not only the teaching dimension—the measure of information complexity in the "classical" teaching model (Goldman and Kearns, 1995; Shinohara and Miyano, 1991)—but also the information complexity of iterated optimal teaching (Balbach, 2008), which is often substantially smaller than the classical teaching dimension.

A combinatorial parameter of central importance in learning theory is the VC-dimension (Vapnik and Chervonenkis, 1971). Among many relevant properties, it provides bounds on the sample complexity of PAC-learning (Blumer et al., 1989). Since the VC-dimension is the best-studied quantity related to information complexity in learning, it is a natural first parameter to compare to when it comes to identifying connections between information complexity notions across various models of learning. For example, even though the self-directed learning complexity can exceed the VC-dimension, existing results show some connection between these two complexity measures (Goldman and Sloan, 1994). However, the teaching dimension, i.e., the information complexity of the classical teaching model, does not exhibit any general relationship to the VC-dimension—the two parameters can be arbitrarily far apart in either direction (Goldman and Kearns, 1995). Similarly, there is no known connection between teaching dimension and query complexity.

In this paper, we establish the first known relationships between the information complexity of teaching and query complexity, as well as between the information complexity of teaching and the VC-dimension. All these relationships are exhibited by the RTD. Two of the main contributions of this work are the following:

- We show that the RTD is never higher (and often considerably lower) than the complexity of self-directed learning. Hence all lower bounds on the RTD hold likewise for self-directed learning, for learning from partial equivalence queries, and for a variety of other query learning models.

- We reveal a strong connection between the RTD and the VC-dimension. Though there are classes for which the RTD exceeds the VC-dimension, we present a number of quite general and natural cases in which the RTD is upper-bounded by the VC-dimension. These include classes of VC-dimension 1, intersection-closed classes, a variety of naturally structured Boolean function classes, and finite maximum classes in general (i.e., classes of maximum possible cardinality for a given VC-dimension and domain size). Many natural concept classes are maximum, e.g., the class of unions of up to k intervals, for any $k \in \mathbb{N}$, or the class of simple arrangements of positive halfspaces. It remains open whether every class of VC-dimension d has an RTD linear in d.

In proving that the RTD of a finite maximum class equals its VC-dimension, we also make a third contribution:

- We reveal a relationship between the RTD and *sample compression schemes* (Littlestone and Warmuth, 1996).

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 class $\mathcal{C}$ can be obtained from the size of a smallest sample compression scheme for $\mathcal{C}$ (Littlestone and Warmuth, 1996; Floyd and Warmuth, 1995). Here the size of a scheme is the size of the largest subset resulting from compression of any sample consistent with some concept in $\mathcal{C}$.

The relationship between RTD and unlabeled sample compression schemes (in which the compression sets consist only of instances without labels) is established via a recent result by Rubinstein and Rubinstein (2012). They show that, for any maximum class of VC-dimension d, a technique called corner-peeling defines unlabeled compression schemes of size d. Like the RTD, corner-peeling is associated with a nesting of subclasses of the underlying concept class. A crucial observation we make in this paper is that every maximum class of VC-dimension d allows corner-peeling with an additional property, which ensures that the resulting unlabeled samples contain exactly those instances a teacher following the RTD model would use. Similarly, we show that the unlabeled compression schemes constructed by Kuzmin and Warmuth's Tail Matching algorithm (Kuzmin and Warmuth, 2007) exactly coincide with the teaching sets used in the RTD model, all of which have size at most d.

This remarkable relationship between the RTD and sample compression suggests that the open question of whether or not the RTD is linear in the VC-dimension might be related to the long-standing open question of whether or not the best possible size of sample compression schemes is linear in the VC-dimension, cf. (Littlestone and Warmuth, 1996; Floyd and Warmuth, 1995). To this end, we observe that a negative answer to the former question would have implications on potential approaches to settling the second. In particular, if the RTD is not linear in the VC-dimension, then there is no mapping that maps every concept class of VC-dimension d to a superclass that is maximum of VC-dimension $O(d)$. Constructing such a mapping would be one way of proving that the best possible size of sample compression schemes is linear in the VC-dimension.

Note that sample compression schemes are not bound to any constraints as to how the compression sets have to be formed, other than that they be subsets of the set to be compressed. In particular, any kind of agreement on, say, an order over the instance space or an order over the concept class, can be exploited for creating the smallest possible compression scheme. As opposed to that, the RTD is defined following a strict "recipe" in which teaching sets are independent of orderings of the instance space or the concept class. These differences between the models make the relationship revealed in this paper even more remarkable. Further connections between teaching and sample compression can in fact be obtained when considering a variant of the RTD introduced by Darnstädt et al. (2013). This new teaching complexity parameter upper-bounds not only the RTD and the VC-dimension, but also the smallest possible size of a sample compression scheme for the underlying concept class. Darnstädt et al. (2013) dubbed this parameter *order compression number*, as it corresponds to the smallest possible size of a special form of compression scheme called *order compression scheme* of the class.

This paper is an extension of an earlier publication (Doliwa et al., 2010).

2. Definitions, Notation and Facts

Throughout this paper, X denotes a finite set and $\mathcal{C}$ denotes a concept class over domain X. For $X' \subseteq X$, we define $\mathcal{C}_{|X'} := \{C \cap X' |\ C \in \mathcal{C}\}$. We treat concepts 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 $l \in \{0,1\}$. If S is a set of labeled examples, we define $X(S) = \{x \in X \mid (x,0) \in S \text{ or } (x,1) \in S\}$. For brevity, $[n] := \{1,\ldots,n\}$. $\mathrm{VCD}(\mathcal{C})$ denotes the VC-dimension of a concept class $\mathcal{C}$.

Definition 1 *Let K be a function that assigns a "complexity" $K(\mathcal{C}) \in \mathbb{N}$ 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 $K(\mathcal{C}_{|X'}) \leq K(\mathcal{C})$.*

2.1 Learning Complexity

A *partial equivalence query* (Maass and Turán, 1992) of a learner is given by a function $h : X \to \{0,1,*\}$ that is passed to an oracle. The latter returns "YES" if the target concept C^* coincides with h on all $x \in X$ for which $h(x) \in \{0,1\}$; it returns a "witness of inequivalence" (i.e., an $x \in X$ such that $C^*(x) \neq h(x) \in \{0,1\}$) otherwise. $\text{LC-PARTIAL}(\mathcal{C})$ denotes the smallest number q such that there is some learning algorithm which can exactly identify any concept $C^* \in \mathcal{C}$ with up to q partial equivalence queries (regardless of the oracle's answering strategy).

A query in the model of *self-directed learning* (Goldman et al., 1993; Goldman and Sloan, 1994) consists of an instance $x \in X$ and a label $b \in \{0,1\}$, passed to an oracle. The latter returns the true label $C^*(x)$ assigned to x by the target concept C^*. We say the learner *made a mistake* if $C^*(x) \neq b$. The *self-directed learning complexity* of $\mathcal{C}$, denoted $\mathrm{SDC}(\mathcal{C})$, is defined as the smallest number q such that there is some self-directed learning algorithm which can exactly identify any concept $C^* \in \mathcal{C}$ without making more than q mistakes.

In the model of *online-learning*, the learner A makes a prediction $b_i \in \{0,1\}$ for a given instance x_i but, in contrast to self-directed learning, the sequence of instances $x_1, x_2, \ldots$ is chosen by an adversary of A that aims at maximizing A's number of mistakes. The *optimal mistake bound* for a concept class $\mathcal{C}$, denoted $M_{opt}(\mathcal{C})$, is the smallest number q such that there exists an online-learning algorithm which which can exactly identify any concept $C^* \in \mathcal{C}$ without making more than q mistakes (regardless of the ordering in which the instances are presented to A).

Clearly, LC-PARTIAL and SDC are monotonic, and M_{opt} is twofold monotonic. The following chain of inequalities is well-known (Goldman and Sloan, 1994; Maass and Turán, 1992; Littlestone, 1988):

$$\mathrm{SDC}(\mathcal{C}) \leq \text{LC-PARTIAL}(\mathcal{C}) \leq M_{opt}(\mathcal{C}) \leq \log|\mathcal{C}|\,. \tag{1}$$

2.2 Teaching Complexity

A *teaching set* for a concept $C \in \mathcal{C}$ is a set S of labeled examples such that C, but no other concept in $\mathcal{C}$, is consistent with S. Let $\mathcal{TS}(C,\mathcal{C})$ denote the family of teaching sets for

$C \in \mathcal{C}$, let $\mathrm{TS}(C;\mathcal{C})$ denote the size of the smallest teaching set for $C \in \mathcal{C}$, and let

$$\mathrm{TS}_{min}(\mathcal{C}) := \min_{C\in\mathcal{C}} \mathrm{TS}(C;\mathcal{C})\,,$$
$$\mathrm{TS}_{max}(\mathcal{C}) := \max_{C\in\mathcal{C}} \mathrm{TS}(C;\mathcal{C})\,,$$
$$\mathrm{TS}_{avg}(\mathcal{C}) := \frac{1}{|\mathcal{C}|}\sum_{C\in\mathcal{C}} \mathrm{TS}(C;\mathcal{C})\,.$$

The quantity $\mathrm{TD}(\mathcal{C}) := \mathrm{TS}_{max}(\mathcal{C})$ is called the *teaching dimension* of $\mathcal{C}$ (Goldman and Kearns, 1995). It refers to the concept in $\mathcal{C}$ that is hardest to teach. In the sequel, $\mathrm{TS}_{min}(\mathcal{C})$ is called the *best-case teaching dimension* of $\mathcal{C}$, and $\mathrm{TS}_{avg}(\mathcal{C})$ is called the *average-case teaching dimension* of $\mathcal{C}$. Obviously, $\mathrm{TS}_{min}(\mathcal{C}) \le \mathrm{TS}_{avg}(\mathcal{C}) \le \mathrm{TS}_{max}(\mathcal{C}) = \mathrm{TD}(\mathcal{C})$.

We briefly note that TD is monotonic, and that a concept class $\mathcal{C}$ consisting of exactly one concept C has teaching dimension 0 because $\emptyset \in \mathcal{TS}(C, \{C\})$.

Definition 2 (Zilles et al. 2011) *A* teaching plan for $\mathcal{C}$ *is a sequence*

$$P = ((C_1, S_1), \ldots, (C_N, S_N)) \tag{2}$$

with the following properties:

- $N = |\mathcal{C}|$ *and* $\mathcal{C} = \{C_1, \ldots, C_N\}$.
- *For all* $t = 1, \ldots, N$, $S_t \in \mathcal{TS}(C_t, \{C_t, \ldots, C_N\})$.

The quantity $\mathrm{ord}(P) := \max_{t=1,\ldots,N} |S_t|$ *is called the* order of the teaching plan P. *Finally, we define*

$$\begin{aligned} \mathrm{RTD}(\mathcal{C}) &:= \min\{\mathrm{ord}(P) \mid P \text{ is a teaching plan for } \mathcal{C}\}\,, \\ \mathrm{RTD}^*(\mathcal{C}) &:= \max_{X'\subseteq X} \mathrm{RTD}(\mathcal{C}_{|X'})\,. \end{aligned}$$

The quantity $\mathrm{RTD}(\mathcal{C})$ *is called the* recursive teaching dimension of $\mathcal{C}$.

A teaching plan (2) is said to be *repetition-free* if the sets $X(S_1), \ldots, X(S_N)$ are pairwise distinct. (Clearly, the corresponding labeled sets, $S_1, \ldots, S_N$, are always pairwise distinct.) Similar to the recursive teaching dimension we define

$$\mathrm{rfRTD}(\mathcal{C}) := \min\{\mathrm{ord}(P) \mid P \text{ is a repetition-free teaching plan for } \mathcal{C}\}\,.$$

One can show that every concept class possesses a repetition-free teaching plan. First, by induction on $|X| = m$, the full cube 2^X has a repetition-free teaching plan of order m: It results from a repetition-free plan for the $(m-1)$-dimensional subcube of concepts for which a fixed instance x is labeled 1, where each teaching set is supplemented by the example $(x, 1)$, followed by a repetition-free teaching plan for the $(m-1)$-dimensional subcube of concepts with $x = 0$. Second, "projecting" a (repetition-free) teaching plan for a concept class $\mathcal{C}$ onto the concepts in a subclass $\mathcal{C}' \subseteq \mathcal{C}$ yields a (repetition-free) teaching plan for $\mathcal{C}'$. Putting these two observations together, it follows that every class over instance set X has a repetition-free teaching plan of order $|X|$.

	x_1	x_2	x_3	x_4	x_5	TS_{min}	$\mathrm{TS}_{min}(C_i,\mathcal{C}\setminus\{C_1\})$	$\mathrm{TS}_{min}(C_i,\mathcal{C}\setminus\{C_2\})$	$\mathrm{TS}_{min}(C_i,\mathcal{C}\setminus\{C_{1/2}\})$
C_1	0	0	0	0	0	2	-	2	-
C_2	1	1	0	0	0	2	2	-	-
C_3	0	1	0	0	0	4	3	3	2
C_4	0	1	0	1	0	4	4	4	4
C_5	0	1	0	1	1	3	3	3	3
C_6	0	1	1	1	0	3	3	3	3
C_7	0	1	1	0	1	3	3	3	3
C_8	0	1	1	1	1	3	3	3	3
C_9	1	0	1	0	0	3	3	3	3
C_{10}	1	0	0	1	0	4	3	3	3
C_{11}	1	0	0	1	1	3	3	3	3
C_{12}	1	0	1	1	0	3	3	3	3
C_{13}	1	0	1	0	1	3	3	3	3

Table 1: A class with $\mathrm{RTD}(\mathcal{C}) = 2$ but $\mathrm{rfRTD}(\mathcal{C}) = 3$.

It should be noted though that $\mathrm{rfRTD}(\mathcal{C})$ may exceed $\mathrm{RTD}(\mathcal{C})$. For example, consider the class in Table 1, which is of RTD 2. In any teaching plan of order 2, both C_1 and C_2 have to be taught first with the same teaching set $\{x_1, x_2\}$ augmented by the appropriate labels. The best repetition free teaching plan for this class is of order 3.

As observed by Zilles et al. (2011), the following holds:

- RTD is monotonic.
- The recursive teaching dimension coincides with the order of any teaching plan that is *in canonical form*, i.e., a teaching plan $((C_1, S_1), \ldots, (C_N, S_N))$ such that for all $t = 1, \ldots, N$ it holds that $|S_t| = \mathrm{TS}_{min}(\{C_t, \ldots, C_N\})$.

Intuitively, a canonical teaching plan is a sequence that is recursively built by always picking an easiest-to-teach concept C_t in the class $\mathcal{C} \setminus \{C_1, \ldots, C_{t-1}\}$ together with an appropriate teaching set S_t.

The definition of teaching plans immediately yields the following result:

Lemma 3 *1. If K is monotonic and $\mathrm{TS}_{min}(\mathcal{C}) \leq K(\mathcal{C})$ for every concept class $\mathcal{C}$, then $\mathrm{RTD}(\mathcal{C}) \leq K(\mathcal{C})$ for every concept class $\mathcal{C}$.*

2. If K is twofold monotonic and $\mathrm{TS}_{min}(\mathcal{C}) \leq K(\mathcal{C})$ for every concept class $\mathcal{C}$, then $\mathrm{RTD}^*(\mathcal{C}) \leq K(\mathcal{C})$ for every concept class $\mathcal{C}$.

RTD and TS_{min} are related as follows:

Lemma 4 $\mathrm{RTD}(\mathcal{C}) = \max_{\mathcal{C}' \subseteq \mathcal{C}} \mathrm{TS}_{min}(\mathcal{C}')$.

Proof Let C_1 be the first concept in a canonical teaching plan P for $\mathcal{C}$ so that $\mathrm{TS}(C_1; \mathcal{C}) = \mathrm{TS}_{min}(\mathcal{C})$ and the order of P equals $\mathrm{RTD}(\mathcal{C})$. It follows that

$$\mathrm{RTD}(\mathcal{C}) = \max\{\mathrm{TS}(C_1; \mathcal{C}), \mathrm{RTD}(\mathcal{C} \setminus \{C_1\})\} = \max\{\mathrm{TS}_{min}(\mathcal{C}), \mathrm{RTD}(\mathcal{C} \setminus \{C_1\})\},$$

and $\mathrm{RTD}(\mathcal{C}) \leq \max_{\mathcal{C}' \subseteq \mathcal{C}} \mathrm{TS}_{min}(\mathcal{C}')$ follows inductively. As for the reverse direction, let $\mathcal{C}'_0 \subseteq \mathcal{C}$ be a maximizer of TS_{min}. Since RTD is monotonic, we get $\mathrm{RTD}(\mathcal{C}) \geq \mathrm{RTD}(\mathcal{C}'_0) \geq \mathrm{TS}_{min}(\mathcal{C}'_0) = \max_{\mathcal{C}' \subseteq \mathcal{C}} \mathrm{TS}_{min}(\mathcal{C}')$. ■

2.3 Intersection-closed Classes and Nested Differences

A concept class $\mathcal{C}$ is called *intersection-closed* if $C \cap C' \in \mathcal{C}$ for all $C, C' \in \mathcal{C}$. Among the standard examples of intersection-closed classes are the d-dimensional boxes over domain $[n]^d$:

$$\mathrm{BOX}_n^d := \{[a_1 : b_1] \times \cdots \times [a_d : b_d] \mid \ \forall i = 1, \ldots, d : \ 1 \leq a_i, b_i \leq n\}$$

Here, $[a : b]$ is an abbreviation for $\{a, a+1, \ldots, b\}$, where $[a : b]$ is the empty set if $a > b$. For the remainder of this section, $\mathcal{C}$ is assumed to be intersection-closed.

For $T \subseteq X$, we define $\langle T \rangle_{\mathcal{C}}$ as the smallest concept in $\mathcal{C}$ containing T, i.e.,

$$\langle T \rangle_{\mathcal{C}} := \bigcap_{T \subseteq C \in \mathcal{C}} C \,.$$

A *spanning set* for $T \subseteq X$ w.r.t. $\mathcal{C}$ is a set $S \subseteq T$ such that $\langle S \rangle_{\mathcal{C}} = \langle T \rangle_{\mathcal{C}}$. S is called a *minimal spanning set* w.r.t. $\mathcal{C}$ if, for every proper subset S' of S, $\langle S' \rangle_{\mathcal{C}} \neq \langle S \rangle_{\mathcal{C}}$. $I(\mathcal{C})$ denotes the size of the largest minimal spanning set w.r.t. $\mathcal{C}$. It is well-known (Natarajan, 1987; Helmbold et al., 1990) that every minimal spanning set w.r.t. $\mathcal{C}$ is shattered by $\mathcal{C}$. Thus, $I(\mathcal{C}) \leq \mathrm{VCD}(\mathcal{C})$. Note that, for every $C^\circ \in \mathcal{C}$, $I(\mathcal{C}_{|C^\circ}) \leq I(\mathcal{C})$, because every spanning set for a set $T \subseteq C^\circ$ w.r.t. $\mathcal{C}$ is also a spanning set for T w.r.t. $\mathcal{C}_{|C^\circ}$.

The class of *nested differences* of depth d (at most d) with concepts from $\mathcal{C}$, denoted $\mathrm{DIFF}^d(\mathcal{C})$ ($\mathrm{DIFF}^{\leq d}(\mathcal{C})$, resp.), is defined inductively as follows:

$$\begin{aligned} \mathrm{DIFF}^1(\mathcal{C}) &:= \mathcal{C} \,, \\ \mathrm{DIFF}^d(\mathcal{C}) &:= \{C \setminus D \mid C \in \mathcal{C}, D \in \mathrm{DIFF}^{d-1}(\mathcal{C})\} \,, \\ \mathrm{DIFF}^{\leq d}(\mathcal{C}) &:= \bigcup_{i=1}^{d} \mathrm{DIFF}^i(\mathcal{C}) \,. \end{aligned}$$

Expanding the recursive definition of $\mathrm{DIFF}^d(\mathcal{C})$ shows that, e.g., a set in $\mathrm{DIFF}^4(\mathcal{C})$ has the form $C_1 \setminus (C_2 \setminus (C_3 \setminus C_4))$ where $C_1, C_2, C_3, C_4 \in \mathcal{C}$. We may assume without loss of generality that $C_1 \supseteq C_2 \supseteq \cdots$ because $\mathcal{C}$ is intersection-closed.

Nested differences of intersection-closed classes were studied in depth at the early stages of research in computational learning theory (Helmbold et al., 1990).

2.4 Maximum Classes and Unlabeled Compression Schemes

Let $\Phi_d(n) := \sum_{i=0}^{d} \binom{n}{i}$. For $d = \mathrm{VCD}(\mathcal{C})$ and for any subset X' of X, we have $\left|\mathcal{C}_{|X'}\right| \leq \Phi_d(|X'|)$, according to Sauer's Lemma (Vapnik and Chervonenkis, 1971; Sauer, 1972). The concept class $\mathcal{C}$ is called a *maximum class* if Sauer's inequality holds with equality for every subset X' of X. It is well-known (Welzl, 1987; Floyd and Warmuth, 1995) that a class over a finite domain X is maximum iff Sauer's inequality holds with equality for $X' = X$.

The following definition was introduced by Kuzmin and Warmuth (2007):

Definition 5 *An* unlabeled compression scheme for a maximum class of VC-dimension d *is given by an injective mapping r that assigns to every concept C a set $r(C) \subseteq X$ of size at most d such that the following condition is satisfied:*

$$\forall C, C' \in \mathcal{C}\ (C \neq C'), \exists x \in r(C) \cup r(C') :\ C(x) \neq C'(x)\,. \tag{3}$$

(3) is referred to as the *non-clashing property*. In order to ease notation, we add the following technical definitions. A *representation mapping of order k for a (not necessarily maximum) class $\mathcal{C}$* is any injective mapping r that assigns to every concept C a set $r(C) \subseteq X$ of size at most k such that (3) holds. A representation-mapping r is said to have the *acyclic non-clashing property* if there is an ordering $C_1, \ldots, C_N$ of the concepts in $\mathcal{C}$ such that

$$\forall 1 \leq i < j \leq N, \exists x \in r(C_i) :\ C_i(x) \neq C_j(x)\,. \tag{4}$$

Considering maximum classes, it was shown by Kuzmin and Warmuth (2007) that a representation mapping with the non-clashing property guarantees that, for every sample S labeled according to a concept in $\mathcal{C}$, there is exactly one concept $C \in \mathcal{C}$ that is consistent with S and satisfies $r(C) \subseteq X(S)$. This allows to encode (compress) a labeled sample S by $r(C)$ and, since r is injective, to decode (decompress) $r(C)$ by C (so that the labels in S can be reconstructed). This coined the term "unlabeled compression scheme".

A concept class $\mathcal{C}$ over a domain X of size n is identified with a subset of $\{0,1\}^n$. The *one-inclusion-graph* $\mathcal{G}(\mathcal{C})$ associated with $\mathcal{C}$ is defined as follows:

- The nodes are the concepts from $\mathcal{C}$.
- Two concepts are connected by an edge if and only if they differ in exactly one coordinate (when viewed as nodes in the Boolean cube).

A *cube* $\mathcal{C}'$ in $\mathcal{C}$ is a subcube of $\{0,1\}^n$ such that every node in $\mathcal{C}'$ represents a concept from $\mathcal{C}$. In the context of the one-inclusion graph, the instances (corresponding to the dimensions in the Boolean cube) are usually called "colors" (and an edge along dimension i is said to have color i). For a concept $C \in \mathcal{C}$, $I(C; \mathcal{G}(\mathcal{C}))$ denotes the union of the instances associated with the colors of the incident edges of C in $\mathcal{G}(\mathcal{C})$, called *incident instances* of C. Recall that the *density* of a graph with m edges and n nodes is defined as m/n. As shown by Haussler et al. (1994, Lemma 2.4), the density of the 1-inclusion graph lower-bounds the VC-dimension, i.e., $\mathrm{dens}(\mathcal{G}(\mathcal{C})) < \mathrm{VCD}(\mathcal{C})$.

The following definitions were introduced by Rubinstein and Rubinstein (2012); we reformulate the notation in order to stress the similarities to the definition of teaching plans.

Definition 6 *A* corner-peeling plan for $\mathcal{C}$ *is a sequence*

$$P = ((C_1, \mathcal{C}'_1), \ldots, (C_N, \mathcal{C}'_N)) \tag{5}$$

with the following properties:

1. *$N = |\mathcal{C}|$ and $\mathcal{C} = \{C_1, \ldots, C_N\}$.*
2. *For all $t = 1, \ldots, N$, $\mathcal{C}'_t$ is a cube in $\{C_t, \ldots, C_N\}$ which contains C_t and all its neighbors in $\mathcal{G}(\{C_t, \ldots, C_N\})$. (Note that this uniquely specifies $\mathcal{C}'_t$.)*

The nodes C_t are called the corners of the cubes $\mathcal{C}'_t$, *respectively. The dimension of the largest cube among $\mathcal{C}'_1, \ldots, \mathcal{C}'_N$ is called the* order of the corner-peeling plan P. $\mathcal{C}$ can be d-corner-peeled *if there exists a corner-peeling plan of order d.*

A concept class $\mathcal{C}$ is called *shortest-path closed* if, for every pair of distinct concepts $C, C' \in \mathcal{C}$, $\mathcal{G}(\mathcal{C})$ contains a path of length $|C \triangle C'|$ (known as the Hamming distance) that connects C and C', where $\triangle$ denotes the symmetric difference. Note that every maximum class is shortest-path closed, but not vice versa. Rubinstein and Rubinstein (2012) showed the following:

1. If a maximum class $\mathcal{C}$ has a corner-peeling plan (5) of order $\mathrm{VCD}(\mathcal{C})$, then an unlabeled compression scheme for $\mathcal{C}$ is obtained by defining $r(C_t)$ to be the set of colors in cube $\mathcal{C}'_t$ for $t = 1, \ldots, N$.

2. Every maximum class $\mathcal{C}$ can be $\mathrm{VCD}(\mathcal{C})$-corner-peeled.

Although it had previously been proved (Kuzmin and Warmuth, 2007) that any maximum class of VC-dimension d has an unlabeled compression scheme of size d, the corner-peeling technique still provides very useful insights. We will see an application in Section 4.3, where we show that $\mathrm{RTD}(\mathcal{C}) = \mathrm{VCD}(\mathcal{C})$ for every maximum class $\mathcal{C}$.

3. Recursive Teaching Dimension and Query Learning

Kuhlmann proved the following result:

Lemma 7 (Kuhlmann 1999) *For every concept class $\mathcal{C}$:* $\mathrm{TS}_{min}(\mathcal{C}) \leq \mathrm{SDC}(\mathcal{C})$.

In view of (1), the monotonicity of LC-PARTIAL and SDC, the twofold monotonicity of M_{opt}, and in view of Lemma 3, we obtain:

Corollary 8 *For every concept class $\mathcal{C}$, the following holds:*

1. $\mathrm{RTD}(\mathcal{C}) \leq \mathrm{SDC}(\mathcal{C}) \leq \mathit{LC\text{-}PARTIAL}(\mathcal{C}) \leq M_{opt}(\mathcal{C})$.

2. $\mathrm{RTD}^*(\mathcal{C}) \leq M_{opt}(\mathcal{C})$.

As demonstrated by Goldman and Sloan (1994), the model of self-directed learning is extremely powerful. According to Corollary 8, recursive teaching is an even more powerful model so that upper bounds on SDC apply to RTD as well, and lower bounds on RTD apply to SDC and LC-PARTIAL as well. The following result, which is partially known from the work by Goldman and Sloan (1994) and Zilles et al. (2011), illustrates this:

Corollary 9

1. *If* $\mathrm{VCD}(\mathcal{C}) = 1$, *then* $\mathrm{RTD}(\mathcal{C}) = \mathrm{SDC}(\mathcal{C}) = 1$.

2. $\mathrm{RTD}(\mathit{Monotone\ Monomials}) = \mathrm{SDC}(\mathit{Monotone\ Monomials}) = 1$.

3. $\mathrm{RTD}(\mathit{Monomials}) = \mathrm{SDC}(\mathit{Monomials}) = 2$.

4. $\mathrm{RTD}(\mathrm{BOX}_n^d) = \mathrm{SDC}(\mathrm{BOX}_n^d) = 2$.

5. $\mathrm{RTD}(m\text{-}\mathit{Term\ Monotone\ DNF}) \leq \mathrm{SDC}(m\text{-}\mathit{Term\ Monotone\ DNF}) \leq m$.

6. SDC(m-*Term Monotone DNF*) $\geq$ RTD(m-*Term Monotone DNF*) $\geq m$ *provided that the number of Boolean variables is at least* $m^2 + 1$.

Proof All upper bounds on SDC were proved by Goldman and Sloan (1994) and, as mentioned above, they apply to RTD as well. The lower bound 1 on RTD (for concept classes with at most two distinct concepts) is trivial. RTD(Monomials) = 2 was shown by Zilles et al. (2011). As a lower bound, this carries over to BOX_n^d which contains Monomials as a subclass. Thus the first five assertions are obvious from known results in combination with Corollary 8.

As for the last assertion, we have to show that RTD(m-Term Monotone DNF) $\geq m$. To this end assume that there are $n \geq m^2 + 1$ Boolean variables. According to Lemma 4, it suffices to find a subclass $\mathcal{C}'$ of m-Term Monotone DNF such that $\mathrm{TS}_{min}(\mathcal{C}') \geq m$. Let $\mathcal{C}'$ be the class of all DNF formulas that contain precisely m pairwise variable-disjoint terms of length m each. Let F be an arbitrary but fixed formula in $\mathcal{C}'$. Without loss of generality, the teacher always picks either a minimal positive example (such that flipping any 1-bit to 0 turns it negative) or a maximal negative example (such that flipping any 0-bit to 1 turns it positive). By construction of $\mathcal{C}'$, the former example has precisely m ones (and reveals one of the m terms in F) and the latter example has precisely m zeroes (and reveals one variable in each term). We may assume that the teacher consistently uses a numbering of the m terms from 1 to m and augments any 0-component (component i say) of a negative example by the number of the term that contains the corresponding Boolean variable (the term containing variable x_i). Since adding information is to the advantage of the learner, this will not corrupt the lower-bound argument. We can measure the knowledge that is still missing after having seen a collection of labeled instances by the following parameters:

- m', the number of still unknown terms
- $l_1, \ldots, l_m$, where l_k is the number of still unknown variables in term k

The effect of a teaching set on these parameters is as follows: a positive example decrements m', and a negative example decrements some of $l_1, \ldots, l_m$. Note that n was chosen sufficiently large[1] so that the formula F is not uniquely specified as long as none of the parameters has reached level 0. Since all parameters are initially of value m, the size of any teaching set for F must be at least m. ■

In powerful learning models, techniques for proving lower bounds become an issue. One technique for proving a lower bound on RTD was applied already in the proof of Corollary 9: select a subclass $\mathcal{C}' \subseteq \mathcal{C}$ and derive a lower bound on $\mathrm{TS}_{min}(\mathcal{C}')$. We now turn to the question whether known lower bounds for LC-PARTIAL or SDC remain valid for RTD. Maass and Turán (1992) showed that LC-PARTIAL is lower-bounded by the logarithm of the length of a longest inclusion chain in $\mathcal{C}$. This bound does not even apply to SDC, which follows from an inspection of the class of half-intervals over domain $[n]$. The longest inclusion chain in this class, $\emptyset \subset \{1\} \subset \{1,2\} \subset \cdots \subset \{1,2,\ldots,n\}$, has length $n+1$, but its self-directed learning complexity is 1. Theorem 8 in the paper by Ben-David and Eiron (1998) implies

1. A slightly refined argument shows that requiring $n \geq (m-1)^2 + 1$ would be sufficient. But we made no serious attempt to make this assumption as weak as possible.

that SDC is lower-bounded by $\log|\mathcal{C}|/\log|X|$ if $\mathrm{SDC}(\mathcal{C}) \geq 2$. We next show that the same bound applies to RTD:

Lemma 10 *Suppose* $\mathrm{RTD}(\mathcal{C}) \geq 2$. *Then,* $\mathrm{RTD}(\mathcal{C}) \geq \frac{\log|\mathcal{C}|}{\log|X|}$.

Proof Samei et al. (2012) have shown that Sauer's bound holds with $\mathrm{RTD}(\mathcal{C})$ in the role of $\mathrm{VCD}(\mathcal{C})$, i.e., for $k = \mathrm{RTD}(\mathcal{C})$, the following holds:

$$|\mathcal{C}| \leq \sum_{i=1}^{k} \binom{|X|}{i} = \Phi_k(|X|) \leq |X|^k$$

Solving for k yields the desired lower bound on $\mathrm{RTD}(\mathcal{C})$. ■

A subset $X' \subseteq X$ is called $\mathcal{C}$-*distinguishing* if, for each pair of distinct concepts $C, C' \in \mathcal{C}$, there is some $x \in X'$ such that $C(x) \neq C'(x)$. The matrix associated with a concept class $\mathcal{C}$ over domain X is given by $M(x, C) = C(x) \in \{0, 1\}$. We call two concept classes $\mathcal{C}, \mathcal{C}'$ equivalent if their matrices are equal up to permutation of rows or columns, and up to flipping all bits of a subset of the rows.[2] The following result characterizes the classes of recursive teaching dimension 1:

Theorem 11 *The following statements are equivalent:*

1. $\mathrm{SDC}(\mathcal{C}) = 1$.
2. $\mathrm{RTD}(\mathcal{C}) = 1$.
3. *There exists a* $\mathcal{C}$*-distinguishing set* $X' \subseteq X$ *such that* $\mathcal{C}_{|X'}$ *is equivalent to a concept class whose matrix* M *is of the form* $M = [M'|\vec{0}]$ *where* M' *is a lower-triangular square-matrix with ones on the main-diagonal and* $\vec{0}$ *denotes the all-zeroes vector.*

Proof *1 implies 2.* If $\mathrm{SDC}(\mathcal{C}) = 1$, $\mathcal{C}$ contains at least two distinct concepts. Thus, $\mathrm{RTD}(\mathcal{C}) \geq 1$. According to Corollary 8, $\mathrm{RTD}(\mathcal{C}) \leq \mathrm{SDC}(\mathcal{C}) = 1$.

2 implies 3. Let P be a teaching plan of order 1 for $\mathcal{C}$, and let X' be the set of instances occurring in P (which clearly is $\mathcal{C}$-distinguishing). Let $(C_1, \{(x_1, b_1)\})$ be the first item of P. Let M be the matrix associated with $\mathcal{C}$ (up to equivalence). We make C_1 the first column and x_1 the first row of M. We may assume that $b_1 = 1$. (Otherwise flip all bits in row 1.) Since $\{(x_1, 1)\}$ is a teaching set for C_1, the first row of M is of the form $(1, 0, \ldots, 0)$. We may repeat this argument for every item in P so that the resulting matrix M is of the desired form. (The last zero-column represents the final concept in P with the empty teaching set.)

3 implies 1. Since X' is $\mathcal{C}$-distinguishing, exact identification of a concept $C \in \mathcal{C}$ is the same as exact identification of C restricted to X'. Let $x_1, \ldots, x_{N-1}$ denote the instances corresponding to the rows of M. Let $C_1, \ldots, C_N$ denote the concepts corresponding to the columns of M. A self-directed learner passes $(x_1, 0), (x_2, 0), \ldots$ to the oracle until it makes the first mistake (if any). If the first mistake (if any) happens for $(x_k, 0)$, the target concept must be C_k (because of the form of M). If no mistake has occurred on items

2. Reasonable complexity measures (including RTD, SDC, VCD) are invariant under these operations.

$(x_1, 0), \ldots, (x_{N-1}, 0)$, there is only one possible target concept left, namely C_N. Thus the self-directed learner exactly identifies the target concept at the expense of at most one mistake. ■

As we have seen in this section, the gap between $\mathrm{SDC}(\mathcal{C})$ and $\text{LC-PARTIAL}(\mathcal{C})$ can be arbitrarily large (e.g., the class of half-intervals over domain $[n]$). We will see below, that a similar statement applies to $\mathrm{RTD}(\mathcal{C})$ and $\mathrm{SDC}(\mathcal{C})$ (despite the fact that both measures assign value 1 to the same family of concept classes).

4. Recursive Teaching Dimension and VC-Dimension

The main open question that we pursue in this section is whether there is a universal constant k such that, for all concept classes $\mathcal{C}$, $\mathrm{RTD}(\mathcal{C}) \leq k \cdot \mathrm{VCD}(\mathcal{C})$. Clearly, $\mathrm{TS}_{min}(\mathcal{C}) \leq \mathrm{RTD}(\mathcal{C}) \leq \mathrm{RTD}^*(\mathcal{C})$, so that the implications from left to right in

$$\begin{aligned} \forall \mathcal{C} : \mathrm{RTD}^*(\mathcal{C}) \leq k \cdot \mathrm{VCD}(\mathcal{C}) \quad &\Leftrightarrow \quad \forall \mathcal{C} : \mathrm{RTD}(\mathcal{C}) \leq k \cdot \mathrm{VCD}(\mathcal{C}) \\ &\Leftrightarrow \quad \forall \mathcal{C} : \mathrm{TS}_{min}(\mathcal{C}) \leq k \cdot \mathrm{VCD}(\mathcal{C}) \end{aligned} \tag{6}$$

are obvious. But the implications from right to left hold as well as can be seen from the following calculations based on the assumption that $\mathrm{TS}_{min}(\cdot) \leq k \cdot \mathrm{VCD}(\cdot)$:

$$\mathrm{RTD}^*(\mathcal{C}) = \max_{X' \subseteq X} \max_{\mathcal{C}' \subseteq \mathcal{C}} \mathrm{TS}_{min}(\mathcal{C}'_{|X'}) \leq k \cdot \max_{X' \subseteq X} \max_{\mathcal{C}' \subseteq \mathcal{C}} \mathrm{VCD}(\mathcal{C}'_{|X'}) \leq k \cdot \mathrm{VCD}(\mathcal{C})$$

Here, the first equation expands the definition of RTD^* and applies Lemma 4. The final inequality makes use of the fact that VCD is twofold monotonic. As a consequence, the question of whether $\mathrm{RTD}(\cdot) \leq k \cdot \mathrm{VCD}(\cdot)$ for a universal constant k remains equivalent if RTD is replaced by TS_{min} or RTD^*.

4.1 Classes with RTD Exceeding VCD

In general the recursive teaching dimension can exceed the VC-dimension. Kuhlmann (1999) presents a family $(\mathcal{C}_m)_{m \geq 1}$ of concept classes for which $\mathrm{VCD}(\mathcal{C}_m) = 2m$ but $\mathrm{RTD}(\mathcal{C}_m) \geq \mathrm{TS}_{min}(\mathcal{C}_m) = 3m$. The smallest class in Kuhlmann's family, $\mathcal{C}_1$, consists of 24 concepts over a domain of size 16.

A smaller class $\mathcal{C}_W$ with $\mathrm{RTD}(\mathcal{C}_W) = \mathrm{TS}_{min}(\mathcal{C}_W) = 3$ and $\mathrm{VCD}(\mathcal{C}_W) = 2$ was communicated to us by Manfred Warmuth. It is shown in Figure 1.

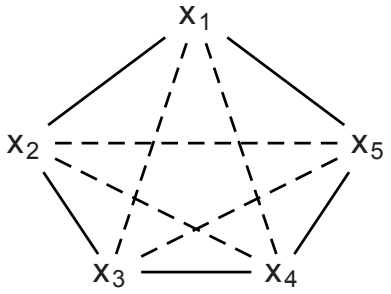
Brute-force enumeration shows that $\mathrm{RTD}(\mathcal{C}_W) = \mathrm{TS}_{min}(\mathcal{C}_W) = 3$ and $\mathrm{VCD}(\mathcal{C}_W) = 2$. Warmuth's class $\mathcal{C}_W$ is remarkable in the sense that it is the smallest concept class for which RTD exceeds VCD. In order to prove this, the following lemmas will be helpful.

Lemma 12 $\mathrm{RTD}(\mathcal{C}) \leq |X| - 1$ *unless* $\mathcal{C} = 2^X$.

Proof If $\mathcal{C} \neq 2^X$, then $\mathcal{C}$ must contain a concept C such that $C \triangle \{x\} \notin \mathcal{C}$ for some instance $x \in X$. Then, C can be uniquely identified within $\mathcal{C}$ using the instances from $X \setminus \{x\}$ and the corresponding labels. Iterative application of this argument leads to a teaching plan for $\mathcal{C}$ of order at most $|X| - 1$. ■

x_1	x_2	x_3	x_4	x_5
1	1	0	0	0
0	1	1	0	0
0	0	1	1	0
0	0	0	1	1
1	0	0	0	1
0	1	0	1	1
0	1	1	0	1
1	0	1	1	0
1	0	1	0	1
1	1	0	1	0

(a)



(b)

Figure 1: The smallest concept class $\mathcal{C}_W$ with $\mathrm{RTD}(\mathcal{C}_W) > \mathrm{VCD}(\mathcal{C}_W)$. The function table to the left can be extracted from the graph to the right by picking concept $\{x_i, x_j\}$ for every solid line and $\mathcal{X} \setminus \{x_i, x_j\}$ for every dashed line.

Note that Lemma 12 transfers to rfRTD, using an argument very similar to the one that implies the existence of a repetition-free teaching plan for every class (see the discussion just below Definition 2.)

For $x \in X$ and $\ell \in \{0,1\}$, $\mathcal{C}[x,\ell]$ is defined as the following subclass of $\mathcal{C}$:

$$\mathcal{C}[x,\ell] = \{C \in \mathcal{C} : \ C(x) = \ell\}$$

An instance x is called *redundant* if $\mathcal{C}[x,\ell] = \emptyset$ for some $\ell \in \{0,1\}$. Note that the label of a redundant instance does not contain any information about the underlying target concept from $\mathcal{C}$. With this notation, the following holds:

Lemma 13 *Let $\mathcal{C}$ be a concept class over domain X such that $\mathrm{TS}_{min}(\mathcal{C}) \geq 3$, and X does not contain redundant instances. Then, $\mathrm{VCD}(\mathcal{C}[x,\ell]) \geq 2$ for all $x \in X$ and $\ell \in \{0,1\}$.*

Proof By way of contradiction. Assume that $\mathrm{VCD}(\mathcal{C}[x,\ell]) \leq 1$ for some choice of x and ℓ. We will show that $\mathrm{TS}_{min}(\mathcal{C}) \leq 2$. According to Corollary 9, $\mathrm{VCD}(\mathcal{C}[x,\ell]) \leq 1$ implies that $\mathrm{TS}_{min}(\mathcal{C}[x,\ell]) \leq \mathrm{RTD}(\mathcal{C}[x,\ell]) \leq 1$. Now it can be seen that $\mathrm{TS}_{min}(\mathcal{C}) \leq 2$: choose (x,ℓ) as the first element in a teaching set and proceed with a teaching set of size $\mathrm{VCD}(\mathcal{C}[x,\ell]) \leq 1$ for the (non-empty) subclass $\mathcal{C}[x,\ell]$. ■

Lemma 14 *Let $\mathcal{C}$ be a concept class over domain X such that $\mathrm{VCD}(\mathcal{C}) = 2$, $\mathrm{TS}_{min}(\mathcal{C}) = 3$, and X does not contain redundant instances. Then $|X| \geq 5$ and, for all $x \in X$ and $\ell \in \{0,1\}$, $|C[x,\ell]| \geq 5$.*

Proof Let $x \in X$ and $\ell \in \{0,1\}$ be arbitrary but fixed. We first show that $|C[x,\ell]| \geq 5$. According to Lemma 13, $\mathrm{VCD}(\mathcal{C}[x,\ell]) \geq 2$. Since $\mathrm{VCD}(\mathcal{C}) = 2$, this implies that

$\mathrm{VCD}(\mathcal{C}[x,\ell]) = 2$. Let $C_1, C_2, C_3, C_4 \in \mathcal{C}[x,\ell]$ be concepts that shatter two points x', x'' in $X \setminus \{x\}$. For at least one of these four concepts, say for C_1, the neighboring concept $C_1 \triangle \{x\}$ does not belong to $\mathcal{C}$ (because otherwise the VC-dimension of $\mathcal{C}$ would be at least 3). If $C_1, \ldots, C_4$ were the only concepts in $\mathcal{C}[x,\ell]$, then $(x', C_1(x'))$ and $(x'', C_1(x''))$ would form a teaching set for C_1 in contradiction to $\mathrm{TS}_{min}(\mathcal{C}) = 3$. We conclude that C_1, C_2, C_3, C_4 are not the only concepts in $C[x,\ell]$ so that $|C[x,\ell]| \geq 5$.

We still have to show that $|X| \geq 5$. Clearly, $|X| \geq \mathrm{TS}_{min}(\mathcal{C}) = 3$. Let us assume by way of contradiction that $|X| = 4$, say $X = \{x, y, z, u\}$. We write concepts over X as 4-tuples $(C(x), C(y), C(z), C(u))$. The following considerations are illustrated in Figure 2. From Lemma 13 and from the assumption $\mathrm{VCD}(\mathcal{C}) = 2$, we may conclude that $\mathrm{VCD}(\mathcal{C}[u,0]) = 2 = \mathrm{VCD}(\mathcal{C}[u,1])$. The set of size 2 shattered by $\mathcal{C}[u,0]$ cannot coincide with the set of size 2 shattered by $\mathcal{C}[u,1]$ because, otherwise, the VC-dimension of $\mathcal{C}$ would be at least 3. Let's say, $\mathcal{C}[u,0]$ shatters $\{x,y\}$ but not $\{x,z\}$ and $\mathcal{C}[u,1]$ shatters $\{x,z\}$ but not $\{x,y\}$. By symmetry, we may assume that $\mathcal{C}[u,1]$ does not contain a concept that assigns label 1 to both x and y, i.e., the concepts $(1,1,0,1)$ and $(1,1,1,1)$ are missing in $\mathcal{C}[u,1]$. Since $\{x,z\}$ is shattered, $\mathcal{C}[u,1]$ must contain the concepts $(1,0,0,1)$ and $(1,0,1,1)$ so as to realize the label assignments $(1,0),(1,1)$ for (x,z). Recall from the first part of the proof that $|\mathcal{C}[u,\ell]| \geq 5$ for $\ell = 0,1$. Note that $|\mathcal{C}[u,\ell]| = 6$ would imply that $\{y,z\}$ is also shattered by $\mathcal{C}[u,\ell]$. Since $\mathrm{VCD}(\mathcal{C}) = 2$, this cannot occur for both subclasses $\mathcal{C}[u,1]$ and $\mathcal{C}[u,0]$ simultaneously. By symmetry, we may assume that $|\mathcal{C}[u,1]| = 5$. Thus, besides $(1,1,0,1)$ and $(1,1,1,1)$, exactly one more concept is missing in $\mathcal{C}[u,1]$. We proceed by case analysis:

Case 1: The additional missing concept in $\mathcal{C}[u,1]$, say C', has Hamming-distance 1 from one of $(1,1,0,1)$ and $(1,1,1,1)$. For reasons of symmetry, we may assume that $C' = (0,1,1,1)$. It follows that the concept $(0,1,0,1)$ belongs to $\mathcal{C}[u,1]$ and has the teaching set $\{(u,1),(y,1)\}$. This is a contradiction to $\mathrm{TS}_{min}(\mathcal{C}) = 3$.

Case 2: The additional missing concept in $\mathcal{C}[u,1]$ has Hamming-distance 2 from both of $(1,1,0,1)$ and $(1,1,1,1)$. Then $\mathcal{C}[u,1]$ contains $(0,1,1,1)$, $(0,1,0,1)$, $(1,0,1,1)$, and $(1,0,0,1)$. In particular, $\mathcal{C}[u,1]$ shatters $\{y,z\}$. In this case, it cannot happen that $\{y,z\}$ is shattered by $\mathcal{C}[u,0]$ too. Thus, $|\mathcal{C}[u,0]| = 5$. We may now expose $\mathcal{C}[u,0]$ to the same case analysis that we already applied to $\mathcal{C}[u,1]$. Since $\mathcal{C}[u,0]$ does not shatter $\{y,z\}$, Case 2 is excluded. As described above, Case 1 leads to a contradiction.

■

We are now ready to prove the minimality of Warmuth's class:

Theorem 15 *Let $\mathcal{C}$ be a concept class over domain X such that $\mathrm{RTD}(\mathcal{C}) > \mathrm{VCD}(\mathcal{C})$. Then $|\mathcal{C}| \geq 10$ and $|X| \geq 5$.*

Proof Obviously $\mathrm{VCD}(\mathcal{C}) = 0$ implies that $\mathrm{RTD}(\mathcal{C}) = 0$. According to Corollary 9, $\mathrm{VCD}(\mathcal{C}) = 1$ implies that $\mathrm{RTD}(\mathcal{C}) = 1$. So we may safely assume that $\mathrm{VCD}(\mathcal{C}) \geq 2$ and $\mathrm{RTD}(\mathcal{C}) \geq 3$. According to Lemma 4, we may assume that $\mathrm{RTD}(\mathcal{C}) = \mathrm{TS}_{min}(\mathcal{C})$ because, otherwise, our proof could proceed with the class $\mathcal{C}' \subseteq \mathcal{C}$ such that $\mathrm{RTD}(\mathcal{C}') = \mathrm{TS}_{min}(\mathcal{C}')$. We may furthermore assume that $\mathcal{C}[x,\ell] \neq \emptyset$ for all $x \in X$ and $\ell \in \{0,1\}$ because, otherwise, x is a redundant instance and the proof could proceed with the subdomain $X \setminus \{x\}$. We may

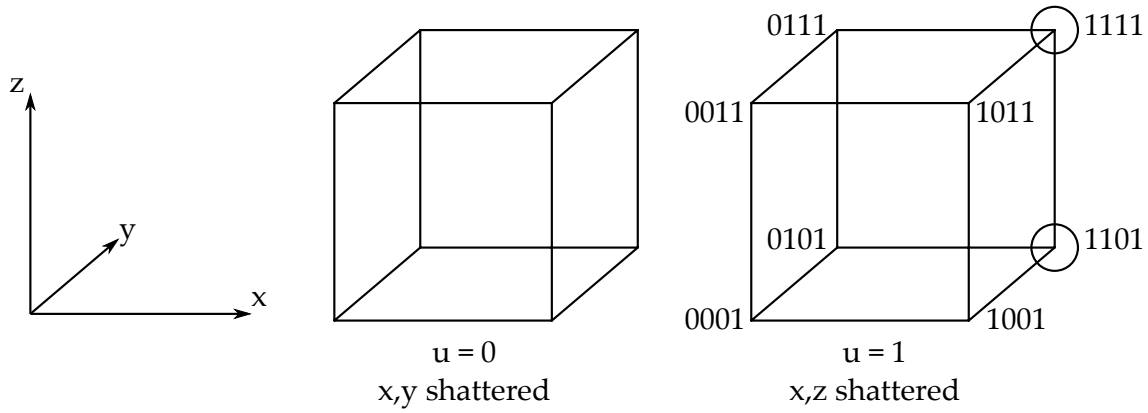


Figure 2: As indicated by circles, the concepts 1101 and 1111 are missing in $\mathcal{C}[u,1]$. There is exactly one additional concept C' which is missing. If $C' \in \{0101, 0111, 1001, 1011\}$, then C' has a teaching set of size 2. Otherwise, $\mathcal{C}[u,1]$ shatters y, z.

therefore apply Lemma 13 and conclude that $\mathrm{VCD}(\mathcal{C}[x,\ell]) \geq 2$ for all $x \in X$ and $\ell \in \{0,1\}$. Clearly $|X| \geq \mathrm{RTD}(\mathcal{C}) \geq 3$. We claim that $|X| \geq 5$, which can be seen as follows. First, note that $\mathcal{C} \neq 2^X$, because $\mathrm{RTD}(\mathcal{C}) > \mathrm{VCD}(\mathcal{C})$. Thus $\mathrm{RTD}(\mathcal{C}) \leq |X| - 1$ by Lemma 12 so that $|X| \geq \mathrm{RTD}(\mathcal{C}) + 1 \geq 4$. Assume $|X| = 4$ by way of contradiction. It follows that $\mathrm{RTD}(\mathcal{C}) \leq 3$ and $\mathrm{VCD}(\mathcal{C}) \leq 2$. Thus, $\mathrm{RTD}(\mathcal{C}) = 3$ and $\mathrm{VCD}(\mathcal{C}) = 2$. But then $|X| \geq 5$ by Lemma 14. Having established $|X| \geq 5$, it remains to prove that $|\mathcal{C}| \geq 10$. According to (1), $\mathrm{RTD}(\mathcal{C}) \leq \log|\mathcal{C}|$. $\mathrm{RTD}(\mathcal{C}) \geq 4$ would imply that $|\mathcal{C}| \geq 16 > 10$. We may therefore focus on the case $\mathrm{RTD}(\mathcal{C}) = 3$, which implies that $\mathrm{VCD}(\mathcal{C}) = 2$. But now it is immediate from Lemma 14 that $|\mathcal{C}| \geq 10$, as desired. ■

We close this section by showing that $\mathrm{RTD}(\mathcal{C}) - \mathrm{VCD}(\mathcal{C})$ can become arbitrarily large. This can be shown by a class whose concepts are disjoint unions of concepts taken from Warmuth's class $\mathcal{C}_W$. Details follow. Suppose that $\mathcal{C}_1$ and $\mathcal{C}_2$ are concept classes over domains X_1 and X_2, respectively, such that $X_1 \cap X_2 = \emptyset$. Then

$$\mathcal{C}_1 \uplus \mathcal{C}_2 := \{A \cup B |\ A \in \mathcal{C}_1, B \in \mathcal{C}_2\}\ .$$

We apply the same operation to arbitrary pairs of concept classes with the understanding that, after renaming instances if necessary, the underlying domains are disjoint. We claim that VCD, TS_{min} and RTD behave additively with respect to "$\uplus$", i.e., the following holds:

Lemma 16 *For all* $K \in \{\mathrm{VCD}, \mathrm{TS}_{min}, \mathrm{RTD}\}$*:* $K(\mathcal{C}_1 \uplus \mathcal{C}_2) = K(\mathcal{C}_1) + K(\mathcal{C}_2)$.

Proof The lemma is fairly obvious for $K = \mathrm{VCD}$ and $K = \mathrm{TS}_{min}$. Suppose that we have an optimal teaching plan that teaches the concepts from $\mathcal{C}_1$ in the order $A_1, \ldots, A_M$ (resp. the concepts from $\mathcal{C}_2$ in the order $B_1, \ldots, B_N$). Then, the teaching plan that proceeds in rounds and teaches $A_i \cup B_1, \ldots, A_i \cup B_N$ in round $i \in [M]$ witnesses that $\mathrm{RTD}(\mathcal{C}_1 \uplus \mathcal{C}_2) \leq$

$\mathrm{RTD}(\mathcal{C}_1) + \mathrm{RTD}(\mathcal{C}_2)$. The reverse direction is an easy application of Lemma 4. Choose $\mathcal{C}_1' \subseteq \mathcal{C}_1$ and $\mathcal{C}_2' \subseteq \mathcal{C}_2$ so that $\mathrm{RTD}(\mathcal{C}_1) = \mathrm{TS}_{min}(\mathcal{C}_1')$ and $\mathrm{RTD}(\mathcal{C}_2) = \mathrm{TS}_{min}(\mathcal{C}_2')$. Now it follows that

$$\mathrm{RTD}(\mathcal{C}_1 \uplus \mathcal{C}_2) \geq \mathrm{TS}_{min}(\mathcal{C}_1' \uplus \mathcal{C}_2') = \mathrm{TS}_{min}(\mathcal{C}_1') + \mathrm{TS}_{min}(\mathcal{C}_2') = \mathrm{RTD}(\mathcal{C}_1) + \mathrm{RTD}(\mathcal{C}_2) \ .$$

■

Setting $\mathcal{C}_W^n = \mathcal{C}_W \uplus \ldots \uplus \mathcal{C}_W$ with n duplicates of $\mathcal{C}_W$ on the right-hand side, we now obtain the following result as an immediate application of Lemma 16:

Theorem 17 $\mathrm{VCD}(\mathcal{C}_W^n) = 2n$ *and* $\mathrm{RTD}(\mathcal{C}_W^n) = 3n$.

We remark here that the same kind of reasoning cannot be applied to blow up rfRTD, because $\mathrm{rfRTD}(\mathcal{C} \uplus \mathcal{C})$ can in general be smaller than $2 \cdot \mathrm{rfRTD}(\mathcal{C})$: considering again the class $\mathcal{C}$ with $\mathrm{rfRTD}(\mathcal{C}) = 3$ from Table 1, simple brute-force computations show that $\mathrm{rfRTD}(\mathcal{C} \times \mathcal{C}) = 5$.

4.2 Intersection-closed Classes

As shown by Kuhlmann (1999), $\mathrm{TS}_{min}(\mathcal{C}) \leq I(\mathcal{C})$ holds for every intersection-closed concept class $\mathcal{C}$. Kuhlmann's central argument (which occurred first in a proof of a related result by Goldman and Sloan (1994)) can be applied recursively so that the following is obtained:

Lemma 18 *For every intersection-closed class* $\mathcal{C}$, $\mathrm{RTD}(\mathcal{C}) \leq I(\mathcal{C})$.

Proof Let $k := I(\mathcal{C})$. We present a teaching plan for $\mathcal{C}$ of order at most k. Let $C_1, \ldots, C_N$ be the concepts in $\mathcal{C}$ in topological order such that $C_i \supset C_j$ implies $i < j$. It follows that, for every $i \in [N]$, C_i is an inclusion-maximal concept in $\mathcal{C}_i := \{C_i, \ldots, C_N\}$. Let S_i denote a minimal spanning set for C_i w.r.t. $\mathcal{C}$. Then:

- $|S_i| \leq k$ and C_i is the unique minimal concept in $\mathcal{C}$ that contains S_i.
- As C_i is inclusion-maximal in $\mathcal{C}_i$, C_i is the only concept in $\mathcal{C}_i$ that contains S_i.

Thus $\{(x, 1) \mid x \in S_i\}$ is a teaching set of size at most k for C_i in $\mathcal{C}_i$. ■

Since $I(\mathcal{C}) \leq \mathrm{VCD}(\mathcal{C})$, we get

Corollary 19 *For every intersection-closed class* $\mathcal{C}$, $\mathrm{RTD}(\mathcal{C}) \leq \mathrm{VCD}(\mathcal{C})$.

This implies $\mathrm{RTD}^*(\mathcal{C}) \leq \mathrm{VCD}(\mathcal{C})$ for every intersection-closed class $\mathcal{C}$, since the property "intersection-closed" is preserved when reducing a class $\mathcal{C}$ to $\mathcal{C}_{|X'}$ for $X' \subseteq X$.

For every fixed constant d (e.g., $d = 2$), Kuhlmann (1999) presents a family $(\mathcal{C}_m)_{m \geq 1}$ of intersection-closed concept classes such that the following holds:[3]

$$\forall m \geq 1 : \ \mathrm{VCD}(\mathcal{C}_m) = d \text{ and } \mathrm{SDC}(\mathcal{C}_m) \geq m \, . \tag{7}$$

3. A family satisfying (7) but *not* being intersection-closed was presented previously by Ben-David and Eiron (1998).

This shows that $\mathrm{SDC}(\mathcal{C})$ can in general not be upper-bounded by $I(\mathcal{C})$ or $\mathrm{VCD}(\mathcal{C})$. It shows furthermore that the gap between $\mathrm{RTD}(\mathcal{C})$ and $\mathrm{SDC}(\mathcal{C})$ can be arbitrarily large (even for intersection-closed classes).

Lemma 18 generalizes to nested differences:

Theorem 20 *If $\mathcal{C}$ is intersection-closed then* $\mathrm{RTD}(\mathrm{DIFF}^{\leq d}(\mathcal{C})) \leq d \cdot I(\mathcal{C})$.

Proof Any concept $C \in \mathrm{DIFF}^{\leq d}(\mathcal{C})$ can be written in the form

$$C = C_1 \setminus \overbrace{(C_2 \setminus (\cdots (C_{d-1} \setminus C_d) \cdots))}^{=:D_1} \tag{8}$$

such that, for every j, $C_j \in \mathcal{C} \cup \{\emptyset\}$, $C_j \supseteq C_{j+1}$, and this inclusion is proper unless $C_j = \emptyset$. Let $D_j = C_{j+1} \setminus (C_{j+2} \setminus (\cdots (C_{d-1} \setminus C_d) \cdots))$. We may obviously assume that the representation (8) of C is *minimal* in the following sense:

$$\forall j = 1, \ldots, d : C_j = \langle C_j \setminus D_j \rangle_{\mathcal{C}} \tag{9}$$

We define a *lexicographic ordering*, $\sqsupset$, on concepts from $\mathrm{DIFF}^{\leq d}(\mathcal{C})$ as follows. Let C be a concept with a minimal representation of the form (8), and let the minimal representation of C' be given similarly in terms of C'_j, D'_j. Then, by definition, $C \sqsupset C'$ if $C_1 \supset C'_1$ or $C_1 = C'_1 \wedge D_1 \sqsupset D'_1$.

Let $k := I(\mathcal{C})$. We present a teaching plan of order at most dk for $\mathrm{DIFF}^{\leq d}(\mathcal{C})$. Therein, the concepts are in lexicographic order so that, when teaching concept C with minimal representation (8), the concepts preceding C w.r.t. $\sqsupset$ have been discarded already. A teaching set T for C is then obtained as follows:

- For every $j = 1, \ldots, d$, include in T a minimal spanning set for $C_j \setminus D_j$ w.r.t. $\mathcal{C}$. Augment its instances by label 1 if j is odd, and by label 0 otherwise.

By construction, C as given by (8) and (9) is the lexicographically smallest concept in $\mathrm{DIFF}^{\leq d}(\mathcal{C})$ that is consistent with T. Since concepts being lexicographically larger than C have been discarded already, T is a teaching set for C. ■

Corollary 21 *Let $\mathcal{C}_1, \ldots, \mathcal{C}_r$ be intersection-closed classes over the domain X. Assume that the "universal concept" X belongs to each of these classes.*[4] *Then,*

$$\mathrm{RTD}\left(\mathrm{DIFF}^{\leq d}(\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_r)\right) \leq d \cdot \sum_{i=1}^{r} I(\mathcal{C}_i)\,.$$

Proof Consider the concept class $\mathcal{C} := \mathcal{C}_1 \wedge \cdots \wedge \mathcal{C}_r := \{C_1 \cap \cdots \cap C_r \mid C_i \in \mathcal{C}_i \text{ for } i = 1, \ldots, r\}$. According to Helmbold et al. (1990), we have:

1. $\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_r$ is a subclass of $\mathcal{C}$.

4. This assumption is not restrictive: adding the universal concept to an intersection-closed class does not destroy the property of being intersection-closed.

2. $\mathcal{C}$ is intersection-closed.

3. Let $C = C_1 \cap \cdots \cap C_r \in \mathcal{C}$. For all i, let S_i be a spanning set for C w.r.t. $\mathcal{C}_i$, i.e., $S_i \subseteq C$ and $\langle S_i \rangle_{\mathcal{C}_i} = \langle C \rangle_{\mathcal{C}_i}$. Then $S_1 \cup \cdots \cup S_r$ is a spanning set for C w.r.t. $\mathcal{C}$.

Thus $I(\mathcal{C}) \leq I(\mathcal{C}_1) + \cdots + I(\mathcal{C}_r)$. The corollary follows from Theorem 20. ■

4.3 Maximum Classes

In this section, we show that the recursive teaching dimension coincides with the VC-dimension on the family of maximum classes. In a maximum class $\mathcal{C}$, every set of $k \leq \mathrm{VCD}(\mathcal{C})$ instances is shattered, which implies $\mathrm{RTD}(\mathcal{C}) \geq \mathrm{TS}_{min}(\mathcal{C}) \geq \mathrm{VCD}(\mathcal{C})$. Thus, we can focus on the reverse direction and pursue the question whether $\mathrm{RTD}(\mathcal{C}) \leq \mathrm{VCD}(\mathcal{C})$. We shall answer this question in the affirmative by establishing a connection between "teaching plans" and "corner-peeling plans".

We say that a corner-peeling plan (5) is *strong* if Condition 2 in Definition 6 is replaced as follows:

2'. For all $t = 1, \ldots, N$, $\mathcal{C}'_t$ is a cube in $\{C_t, \ldots, C_N\}$ which contains C_t and whose colors (augmented by their labels according to C_t) form a teaching set for $C_t \in \{C_t, \ldots, C_N\}$.

We denote the set of colors of $\mathcal{C}'_t$ as X_t and its augmentation by labels according to C_t as S_t in what follows. The following result is obvious:

Lemma 22 *A strong corner-peeling plan of the form (5) induces a teaching plan of the form (2) of the same order.*

The following result justifies the attribute "strong" of corner-peeling plans:

Lemma 23 *Every strong corner-peeling plan is a corner-peeling plan.*

Proof Assume that Condition 2 is violated. Then there is a color $x \in X \setminus X_t$ and a concept $C \in \{C_{t+1}, \ldots, C_N\}$ such that C coincides with C_t on all instances except x. But then C is consistent with set S_t so that S_t is *not* a teaching set for $C_t \in \{C_t, \ldots, C_N\}$, and Condition 2' is violated as well. ■

Lemma 24 *Let $\mathcal{C}$ be a shortest-path closed concept class. Then, every corner-peeling plan for $\mathcal{C}$ is strong.*

Proof Assume that Condition 2' is violated. Then some $C \in \{C_{t+1}, \ldots, C_N\}$ is consistent with S_t. Thus, the shortest path between C and C_t in $\mathcal{G}(\{C_t, \ldots, C_N\})$ does not enter the cube $\mathcal{C}'_t$. Hence there is a concept $C' \in \{C_{t+1}, \ldots, C_N\} \setminus \mathcal{C}'_t$ that is a neighbor of C_t in $\mathcal{G}(\{C_t, \ldots, C_N\})$, and Condition 2 is violated. ■

As maximum classes are shortest-path closed (Kuzmin and Warmuth, 2007), we obtain:

Corollary 25 *Every corner-peeling plan for a maximum class is strong, and therefore induces a teaching plan of the same order.*

Since Rubinstein and Rubinstein (2012) showed that every maximum class $\mathcal{C}$ can be $\mathrm{VCD}(\mathcal{C})$-corner-peeled, we may conclude that $\mathrm{RTD}(\mathcal{C}) \leq \mathrm{VCD}(\mathcal{C})$. As mentioned above, $\mathrm{RTD}(\mathcal{C}) \geq \mathrm{TS}_{min}(\mathcal{C}) \geq \mathrm{VCD}(\mathcal{C})$ for every maximum class $\mathcal{C}$. Thus the following holds:

Theorem 26 *For every maximum class $\mathcal{C}$, $\mathrm{RTD}(\mathcal{C}) = \mathrm{TS}_{min}(\mathcal{C}) = \mathrm{VCD}(\mathcal{C})$.*

The fact that, for every maximum class $\mathcal{C}$ and every $X' \subseteq X$, the class $\mathcal{C}_{|X'}$ is still maximum implies that $\mathrm{RTD}^*(\mathcal{C}) = \mathrm{VCD}(\mathcal{C})$ for every maximum class $\mathcal{C}$.

We establish a connection between repetition-free teaching plans and representations having the acyclic non-clashing property:

Lemma 27 *Let $\mathcal{C}$ be an arbitrary concept class. Then the following holds:*

1. *Every repetition-free teaching plan (2) of order d for $\mathcal{C}$ induces a representation mapping r of order d for $\mathcal{C}$ given by $r(C_t) = X(S_t)$ for $t = 1, \ldots, N$. Moreover, r has the acyclic non-clashing property.*

2. *Every representation mapping r of order d for $\mathcal{C}$ that has the acyclic non-clashing property (4) induces a teaching plan (2) given by $S_t = \{(x, C_t(x)) \mid x \in r(C_t)\}$ for $t = 1, \ldots, N$. Moreover, this plan is repetition-free.*

Proof

1. A clash between C_t and $C_{t'}$, $t < t'$, on $X(S_t)$ would contradict the fact that S_t is a teaching set for $C_t \in \{C_t, \ldots, C_N\}$.

2. Conversely, if $S_t = \{(x, C_t(x)) \mid x \in r(C_t)\}$ is not a teaching set for $C_t \in \{C_t, \ldots, C_N\}$, then there must be a clash on $X(S_t)$ between C_t and a concept from $\{C_{t+1}, \ldots, C_N\}$. The teaching plan induced by r is obviously repetition-free since r is injective.

■

Corollary 28 *Let $\mathcal{C}$ be maximum of VC-dimension d. Then, there is a one-one mapping between repetition-free teaching plans of order d for $\mathcal{C}$ and unlabeled compression schemes with the acyclic non-clashing property.*

A closer look at the work by Rubinstein and Rubinstein (2012) reveals that corner-peeling leads to an unlabeled compression scheme with the acyclic non-clashing property (again implying that $\mathrm{RTD}(\mathcal{C}) \leq \mathrm{VCD}(\mathcal{C})$ for maximum classes $\mathcal{C}$). Similarly, an inspection of the work by Kuzmin and Warmuth (2007) reveals that the unlabeled compression scheme obtained by the Tail Matching Algorithm has the acyclic non-clashing property, too. Thus, this algorithm too can be used to generate a recursive teaching plan of order $\mathrm{VCD}(\mathcal{C})$ for any maximum class $\mathcal{C}$.

It is not known to date whether every concept class $\mathcal{C}$ of VC-dimension d can be embedded into a maximum concept class $\mathcal{C}' \supseteq \mathcal{C}$ of VC-dimension $O(d)$. Indeed, finding such an

embedding is considered as a promising method for settling the sample compression conjecture. It is easy to see that a negative answer to our question "Is $\mathrm{RTD}(\mathcal{C}) \in O(\mathrm{VCD}(\mathcal{C}))$?" would deem this approach fruitless:

Theorem 29 *If* $\mathrm{RTD}(\mathcal{C})$ *is not linearly bounded in* $\mathrm{VCD}(\mathcal{C})$*, then there is no mapping* $\mathcal{C} \mapsto \mathcal{C}' \supseteq \mathcal{C}$ *such that* $\mathcal{C}'$ *is maximum and* $\mathrm{VCD}(\mathcal{C}')$ *is linearly bounded in* $\mathrm{VCD}(\mathcal{C})$*.*

Proof Suppose there is a universal constant k and a mapping MAXIMIZE that maps every concept class $\mathcal{C}$ to a concept class $\mathcal{C}' \supseteq \mathcal{C}$ such that $\mathcal{C}'$ is maximum and $\mathrm{VCD}(\mathcal{C}') \leq k \cdot \mathrm{VCD}(\mathcal{C})$. It follows that, for any concept class $\mathcal{C}$, the following holds:

$$\mathrm{RTD}(\mathcal{C}) \leq \mathrm{RTD}(\mathrm{MAXIMIZE}(\mathcal{C})) = \mathrm{VCD}(\mathrm{MAXIMIZE}(\mathcal{C})) \leq k \cdot \mathrm{VCD}(\mathcal{C}))$$

where the equation $\mathrm{RTD}(\mathrm{MAXIMIZE}(\mathcal{C})) = \mathrm{VCD}(\mathrm{MAXIMIZE}(\mathcal{C}))$ follows from Theorem 26. ■

According to (6), this theorem still holds if RTD is replaced by RTD^*.

4.4 Shortest-Path Closed Classes

In this section, we study the best-case teaching dimension, $\mathrm{TS}_{min}(\mathcal{C})$, and the average-case teaching-dimension, $\mathrm{TS}_{avg}(\mathcal{C})$, of a shortest-path closed concept class $\mathcal{C}$.

It is known that the instances of $I(C; \mathcal{G}(\mathcal{C}))$, augmented by their C-labels, form a unique minimal teaching set for C in $\mathcal{C}$ provided that $\mathcal{C}$ is a maximum class (Kuzmin and Warmuth, 2007). Lemma 30 slightly generalizes this observation.

Lemma 30 *Let* $\mathcal{C}$ *be any concept class. Then the following two statements are equivalent:*

1. $\mathcal{C}$ *is shortest-path closed.*
2. *Every* $C \in \mathcal{C}$ *has a unique minimum teaching set* S*, namely the set* S *such that* $X(S) = I(C; \mathcal{G}(\mathcal{C}))$*.*

Proof $1 \Rightarrow 2$ is easy to see. Let $\mathcal{C}$ be shortest-path closed, and let C be any concept in $\mathcal{C}$. Clearly, any teaching set S for C must satisfy $I(C; \mathcal{G}(\mathcal{C})) \subseteq X(S)$ because C must be distinguished from all its neighbors in $\mathcal{G}(\mathcal{C})$. Let $C' \neq C$ be any other concept in $\mathcal{C}$. Since C and C' are connected by a path P of length $|C \triangle C'|$, C and C' are distinguished by the color of the first edge in P, say by the color $x \in I(C; \mathcal{G}(\mathcal{C}))$. Thus, no other instances (=colors) besides $I(C; \mathcal{G}(\mathcal{C}))$ are needed to distinguish C from any other concept in $\mathcal{C}$.

To show $2 \Rightarrow 1$, we suppose 2 and prove by induction on k that any two concepts $C, C' \in \mathcal{C}$ with $k = |C \triangle C'|$ are connected by a path of length k in $\mathcal{G}(\mathcal{C})$. The case $k = 1$ is trivial. For a fixed k, assume all pairs of concepts of Hamming distance k are connected by a path of length k in $\mathcal{G}(\mathcal{C})$. Let $C, C' \in \mathcal{C}$ with $|C \triangle C'| = k+1 \geq 2$. Since $I(C; \mathcal{G}(\mathcal{C})) = X(S)$, there is an $x \in I(C; \mathcal{G}(\mathcal{C}))$ such that $C(x) \neq C'(x)$. Let C'' be the x-neighbor of C in $\mathcal{G}(\mathcal{C})$. Note that $C''(x) = C'(x)$ so that C'' and C' have Hamming-distance k. According to the inductive hypothesis, there is a path of length k from C'' to C' in $\mathcal{G}(\mathcal{C})$. It follows that C and C' are connected by a path of length $k + 1$. ■

Theorem 31 *Let $\mathcal{C}$ be a shortest-path closed concept class. Then,* $\mathrm{TS}_{avg}(\mathcal{C}) < 2\mathrm{VCD}(\mathcal{C})$.

Proof According to Lemma 30, the average-case teaching dimension of $\mathcal{C}$ coincides with the average vertex-degree in $\mathcal{G}(\mathcal{C})$, which is twice the density of $\mathcal{G}(\mathcal{C})$. As mentioned in Section 2.4 already, $\mathrm{dens}(\mathcal{G}(\mathcal{C})) < \mathrm{VCD}(\mathcal{C})$. ■

Theorem 31 generalizes a result by Kuhlmann (1999) who showed that the average-case teaching dimension of "d-balls" (sets of concepts of Hamming distance at most d from a center concept) is smaller than $2d$. It also simplifies Kuhlmann's proof substantially. In Theorem 4 of the same paper, Kuhlmann (1999) stated furthermore that $\mathrm{TS}_{avg}(\mathcal{C}) < 2$ if $\mathrm{VCD}(\mathcal{C}) = 1$, but his proof is flawed.[5] Despite the flawed proof, the claim itself is correct as we show now:

Theorem 32 *Let $\mathcal{C}$ be any concept class. If* $\mathrm{VCD}(\mathcal{C}) = 1$ *then* $\mathrm{TS}_{avg}(\mathcal{C}) < 2$.

Proof By Theorem 31, the average-case teaching dimension of a maximum class of VC-dimension 1 is less than 2. It thus suffices to show that any class $\mathcal{C}$ of VC-dimension 1 can be transformed into a maximum class $\mathcal{C}'$ of VC-dimension 1 without decreasing the average-case teaching dimension. Let $X' \subseteq X$ be a minimal set that is $\mathcal{C}$-distinguishing, i.e., for every pair of distinct concepts $C, C' \in \mathcal{C}$, there is some $x \in X'$ such that $C(x) \neq C(x')$. Let $m = |X|$ and $\mathcal{C}' = \mathcal{C}_{|X'}$. Obviously, $|\mathcal{C}'| = |\mathcal{C}|$ and $\mathrm{VCD}(\mathcal{C}') = 1$ so that $|\mathcal{C}'| \leq \binom{m}{0} + \binom{m}{1} = m + 1$. Now we prove that $\mathcal{C}'$ is maximum. Note that every $x \in X'$ occurs as a color in $\mathcal{G}(\mathcal{C}')$ because, otherwise, $X' \setminus \{x\}$ would still be $\mathcal{C}$-distinguishing (which would contradict the minimality of X'). As $\mathrm{VCD}(\mathcal{C}') = 1$, no color can occur twice. Thus $|E(\mathcal{G}(\mathcal{C}'))| = m$. Moreover, there is no cycle in $\mathcal{G}(\mathcal{C}')$ since a cycle would require at least one repeated color. As $\mathcal{G}(\mathcal{C}')$ is an acyclic graph of m edges, it has at least $m + 1$ vertices, i.e. $|\mathcal{C}'| \geq m + 1$. Thus, $|\mathcal{C}'| = m + 1$ and $\mathcal{C}'$ is maximum. This implies that $\mathrm{TS}_{avg}(\mathcal{C}') < 2\mathrm{VCD}(\mathcal{C}')$. Since $X' \subseteq X$ but X' is still $\mathcal{C}$-distinguishing, we obtain $\mathrm{TS}(C;\mathcal{C}) \leq \mathrm{TS}(C_{|X'}, \mathcal{C}')$ for all $C \in \mathcal{C}$. Thus, $\mathrm{TS}_{avg}(\mathcal{C}) \leq \mathrm{TS}_{avg}(\mathcal{C}') < 2\mathrm{VCD}(\mathcal{C}') = 2$, which concludes the proof. ■

We briefly note that $\mathrm{TS}_{avg}(\mathcal{C})$ cannot in general be bounded by $O(\mathrm{VCD}(\mathcal{C}))$. Kushilevitz et al. (1996) present a family $(\mathcal{C}_n)$ of concept classes such that $\mathrm{TS}_{avg}(\mathcal{C}_n) = \Omega(\sqrt{|\mathcal{C}_n|})$ but $\mathrm{VCD}(\mathcal{C}_n) \leq \log |\mathcal{C}_n|$.

We conclude this section by showing that there are shortest-path closed classes for which RTD exceeds VCD.

Lemma 33 *If* $\deg_{\mathcal{G}(\mathcal{C})}(C) \geq |X| - 1$ *for all* $C \in \mathcal{C}$, *then* $\mathcal{C}$ *is shortest-path closed.*

Proof Assume by way of contradiction that $\mathcal{C}$ is not shortest-path closed. Pick two concepts $C, C' \in \mathcal{C}$ of minimal Hamming-distance, say d, subject to the constraint of not being connected by a path of length d in $\mathcal{G}(\mathcal{C})$. It follows that $d \geq 2$. By the minimality of d, any

5. His Claim 2 states the following. *If* $\mathrm{VCD}(\mathcal{C}) = 1$, $C_1, C_2 \in \mathcal{C}$, $x \in X$, $x \notin C_1$, $C_2 = C_1 \cup \{x\}$, *then, for either* $(i,j) = (1,2)$ *or* $(i,j) = (2,1)$, *one obtains* $\mathrm{TS}(C_i;\mathcal{C}) = \mathrm{TS}(C_i - x; \mathcal{C} - x) + 1$ *and* $\mathrm{TS}(C_j;\mathcal{C}) = 1$. This is not correct, as can be shown by the class $\mathcal{C} = \{\{x_z : 1 \leq z \leq k\} : 0 \leq k \leq 5\}$ over $X = \{x_k : 1 \leq k \leq 5\}$, which has VC-dimension 1. For $C_1 = \{x_1, x_2\}$, $C_2 = \{x_1, x_2, x_3\}$, and $x = x_3$, we get $\mathrm{TS}(C_1;\mathcal{C}) = \mathrm{TS}(C_2;\mathcal{C}) = \mathrm{TS}(C_1 - x; \mathcal{C} - x) = 2$.

neighbor of C with Hamming-distance $d-1$ to C' does not belong to $\mathcal{C}$. Since there are d such missing neighbors, the degree of C in $\mathcal{G}(\mathcal{C})$ is bounded by $|X|-d \leq |X|-2$. This yields a contradiction. ■

Rubinstein et al. (2009) present a concept class $\mathcal{C}$ with $\mathrm{TS}_{min}(\mathcal{C}) > \mathrm{VCD}(\mathcal{C})$. An inspection of this class shows that the minimum vertex degree in its 1-inclusion graph is $|X|-1$. According to Lemma 33, this class must be shortest-path closed. Thus, the inequality $\mathrm{TS}_{min}(\mathcal{C}) \leq \mathrm{VCD}(\mathcal{C})$ does not generalize from maximum classes to shortest-path closed classes.

5. Conclusions

This paper relates the RTD, a recently introduced teaching complexity notion, to information complexity parameters of various classical learning models.

One of these parameters is SDC, the information complexity of self-directed learning, which constitutes the most information-efficient query learning model known to date. Our main result in this context, namely lower-bounding the SDC by the RTD, has implications for the analysis of information complexity in teaching and learning. In particular, every upper bound on SDC holds for RTD; every lower bound on RTD holds for SDC.

The central parameter in our comparison is the VC-dimension. Although the VC-dimension can be arbitrarily large for classes of recursive teaching dimension 1 (which is well-known and also evident from Theorem 11) and arbitrarily smaller than SDC (Ben-David and Eiron, 1998; Kuhlmann, 1999), it does not generally lie in between the two. However, while the SDC cannot be upper-bounded by any linear function of the VC-dimension, it is still open whether such a bound exists for the RTD. The existence of the latter would mean that the combinatorial properties that determine the information complexity of PAC-learning (i.e., of learning from randomly drawn examples) are essentially the same as those that determine the information complexity of teaching (i.e., of learning from helpfully selected examples), at least when using the recursive teaching model.

As a partial solution to this open question, we showed that the VC-dimension coincides with the RTD in the special case of maximum classes. Our results, and in particular the remarkable correspondence to unlabeled compression schemes, suggest that the RTD is based on a combinatorial structure that is of high relevance for the complexity of information-efficient learning and sample compression. Analyzing the circumstances under which teaching plans defining the RTD can be used to construct compression schemes (and to bound their size) seems to be a promising step towards new insights into the theory of sample compression.

Acknowledgments

We would like to thank Manfred Warmuth for the permission to include the concept class from Figure 1 in this paper. Moreover, we would like to thank Malte Darnstädt and Michael Kallweit for helpful and inspiring discussions, and the anonymous referees of both this article

and its earlier conference version for many helpful suggestions and for pointing out mistakes in the proofs of Theorems 20 and 32.

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

References

Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1988.

Frank Balbach. Measuring teachability using variants of the teaching dimension. *Theoretical Computer Science*, 397(1–3):94–113, 2008.

Shai Ben-David and Nadav Eiron. Self-directed learning and its relation to the VC-dimension and to teacher-directed learning. *Machine Learning*, 33(1):87–104, 1998.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association on Computing Machinery*, 36(4):929–965, 1989.

Malte Darnstädt, Thorsten Doliwa, Hans U. Simon, and Sandra Zilles. Order compression schemes. In *Proceedings of the 24th International Conference on Algorithmic Learning Theory*, pages 173–187, 2013.

Thorsten Doliwa, Hans U. Simon, and Sandra Zilles. Recursive teaching dimension, learning complexity, and maximum classes. In *Proceedings of the 21st International Conference on Algorithmic Learning Theory*, pages 209–223, 2010.

Sally Floyd and Manfred Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):1–36, 1995.

Sally A. Goldman and Michael J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995.

Sally A. Goldman and Robert H. Sloan. The power of self-directed learning. *Machine Learning*, 14(1):271–294, 1994.

Sally A. Goldman, Ronald L. Rivest, and Robert E. Schapire. Learning binary relations and total orders. *SIAM Journal on Computing*, 22(5):1006–1034, 1993.

David Haussler, Nick Littlestone, and Manfred K. Warmuth. Predicting $\{0,1\}$ functions on randomly drawn points. *Information and Computation*, 115(2):284–293, 1994.

David Helmbold, Robert Sloan, and Manfred K. Warmuth. Learning nested differences of intersection-closed concept classes. *Machine Learning*, 5:165–196, 1990.

Christian Kuhlmann. On teaching and learning intersection-closed concept classes. In *Proceedings of the 4th European Conference on Computational Learning Theory*, pages 168–182, 1999.

Eyal Kushilevitz, Nathan Linial, Yuri Rabinovich, and Michael E. Saks. Witness sets for families of binary vectors. *J. Comb. Theory, Ser. A*, 73(2):376–380, 1996.

Dima Kuzmin and Manfred K. Warmuth. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 2007.

Nick Littlestone. Learning quickly when irrelevant attributes abound: a new linear threshold algorithm. *Machine Learning*, 2(4):245–318, 1988.

Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability. Technical Report, UC Santa Cruz, 1996.

Wolfgang Maass and György Turán. Lower bound methods and separation results for on-line learning models. *Machine Learning*, 9:107–145, 1992.

Balas K. Natarajan. On learning boolean functions. In *Proceedings of the 19th Annual Symposium on Theory of Computing*, pages 296–304, 1987.

Benjamin I. P. Rubinstein and J. Hyam Rubinstein. A geometric approach to sample compression. *Journal of Machine Learning Research*, 13:1221–1261, 2012.

Benjamin I. P. Rubinstein, Peter L. Bartlett, and J. Hyam Rubinstein. Shifting: One-inclusion mistake bounds and sample compression. *Journal of Computer and System Sciences*, 75(1):37–59, 2009.

Rahim Samei, Pavel Semukhin, Boting Yang, and Sandra Zilles. Sauer's bound for a notion of teaching complexity. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, pages 96–110, 2012.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

Ayumi Shinohara and Satoru Miyano. Teachability in computational learning. *New Generation Computing*, 8(4):337–348, 1991.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theor. Probability and Appl.*, 16(2):264–280, 1971.

Emo Welzl. Complete range spaces, 1987. Unpublished Notes.

Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Teaching dimensions based on cooperative learning. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 135–146, 2008.

Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Models of cooperative teaching and learning. *Journal of Machine Learning Research*, 12:349–384, 2011.