Recursive Teaching Dimension, Learning Complexity, and Maximum Classes*

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

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

Abstract. This paper is concerned with the combinatorial structure of concept classes that can be learned from a small number of examples. We show that the recently introduced notion of recursive teaching dimension (RTD, reflecting the complexity of teaching a concept class) is a relevant parameter in this context. Comparing the RTD to self-directed learning, we establish new lower bounds on the query complexity for a variety of query learning models and thus *connect teaching to query learning*.

For many general cases, the RTD is upper-bounded by the VC-dimension, e.g., classes of VC-dimension 1, (nested differences of) intersection-closed classes, "standard" boolean function classes, and finite maximum classes. The RTD thus is the first model to *connect teaching to the VC-dimension*.

The combinatorial structure defined by the RTD has a remarkable resemblance to the structure exploited by sample compression schemes and hence *connects teaching to sample compression*. Sequences of teaching sets defining the RTD coincide with unlabeled compression schemes both (i) resulting from Rubinstein and Rubinstein's corner-peeling and (ii) resulting from Kuzmin and Warmuth's Tail Matching algorithm.

1 Introduction

The complexity of the problem of learning a concept C in a given concept class $\mathcal{C}$ can be measured in different ways. If A is a learning algorithm of a particular type, one measures for instance how much information A must process, how many prediction errors A will make on single attributes of C, or how expensive the computation executed by A is, when identifying C. The worst-case behavior of A is given by the highest such amount measured over all concepts $\mathcal{C}$. The complexity value assigned to $\mathcal{C}$ with respect to the underlying learning model is then defined as the best possible worst-case behavior of any learning algorithm.

While run-time and memory complexity are important aspects of machine learning problems, the aspect of "information complexity" (e.g., how many labeled data points are needed for learning) has at least equally important status.

* This work was supported by the Deutsche Forschungsgemeinschaft Grant SI 498/8-1 and by the Natural Sciences and Engineering Research Council of Canada (NSERC).

M. Hutter et al. (Eds.): ALT 2010, LNAI 6331, pp. 209–223, 2010.

From an application point of view, the cost of a machine learning process is often dominated by the amount of data needed. From a theoretical point of view, the study of information complexity yields formal guarantees concerning the amount of data that needs to be processed to solve a learning problem. Moreover, analyzing information complexity often helps to understand the structure of a given class of target concepts. In addition, the theoretical study of information complexity helps to identify parallels between various formal models of learning.

The reason for these parallels is that algorithms used in a number of different formal learning models reflect principles related to sample compression schemes, i.e., 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 $\mathcal{C}$, see [14, 5]. Here the size of a scheme is the size of the largest subset resulting from compression of any sample consistent with some concept in $\mathcal{C}$. Similarly teachers, which provide examples to the learner in models of co-operative learning, compress concepts to subsets of particularly "helpful" examples, cf. [6, 19, 10, 2].

In the past two decades, several learning models were defined with the aim of achieving low information complexity in a non-trivial way. One such model is learning from partial equivalence queries [15], which subsumes all types of queries for which negative answers are witnessed by counterexamples, e.g., membership, equivalence, subset, superset, and disjointness queries [1]. As lower bounds on the information complexity in this model (here called query complexity) hold for numerous learning models, they are particularly interesting objects of study.

In the query model of self-directed learning [7], a query is a prediction of a label for an instance of the learner's choice and the learner "pays" only for wrong predictions. Self-directed learners are very powerful; they yield a query complexity lower-bounding the one obtained from partial equivalence queries [8]. Even though the self-directed learning complexity can exceed the VC-dimension, existing results show some connection between these two complexity measures.

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 $\mathcal{C}$, see [22]. The nesting is defined by (i) choosing 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 $\mathcal{C}$ and (ii) recursively repeating this process with the remaining concepts. The largest number of examples required at any stage is the *recursive teaching dimension (RTD)* of $\mathcal{C}$. The RTD significantly improves on bounds for previous teaching models. It lower-bounds not only the complexity of the "classical" teaching model [6, 19] but also the complexity of iterated optimal teaching [2], which is often significantly below the classical teaching dimension.

Using the RTD, this paper is the first one to establish a relation between teaching complexity and complexity of query learning, between teaching complexity and the VC-dimension, as well as between teaching complexity and sample compression, in particular revealing a surprisingly strong connection to unlabeled sample compression, cf. [4, 12]. No such relations are exhibited by the classical teaching models. Our main contributions are the following.

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

(ii) We establish a 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 and nested differences thereof, 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). It remains open whether every class of VC-dimension d has an RTD linear in d.

(iii) We establish a connection between the RTD and *unlabeled compression schemes.* To prove that the RTD of a finite maximum class equals its VC-dimension, we use a recent result from [17]. Rubinstein and Rubinstein show that, for all maximum classes, a technique called corner-peeling defines unlabeled compression schemes whose size equals the VC-dimension. Corner-peeling is a particular way of recursively removing concepts from the given concept class, while representing every such peeling step by a small subset of the underlying instance space, i.e., an unlabeled sample. Firstly, the recursive nesting of concept classes is common to both peeling and RTD. Secondly, and more importantly, we observe that every maximum class 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. A closer look reveals the following two facts for any finite maximum class $\mathcal{C}$ of VC-dimension d:

- Both Rubinstein and Rubinstein's corner-peeling *and* Kuzmin and Warmuth's Tail Matching [12] construct unlabeled compression schemes for $\mathcal{C}$ that map to samples exactly coinciding with those used in the RTD model for $\mathcal{C}$. All samples are of size at most d.
- The RTD model allows for a nesting of $\mathcal{C}$ that uses samples of size at most d whose unlabeled versions form an unlabeled compression scheme of size d.

The correspondence between RTD and compression schemes is quite remarkable, because these models arose in different branches of Learning Theory and, for that reason, differ in several respects:

- The RTD-model has comparatively restrictive rules for producing teaching sets (which is a kind of compression).
- It does not explicitly address the issue of *sample* compression (but rather compresses the concept as a function on the whole domain).

Despite these differences, sample compression schemes lead to RTD-nestings for a wide variety of classes (including linear arrangements and halfspaces). Consequently, the question of whether or not the RTD is linear in the VC-dimension appears to be related to the long-standing open question of whether or not the sample compression complexity is linear in the VC-dimension, cf. [14]. We believe that studying the RTD will continue to provide new insights into the combinatorial structure of concept classes that possess small compression schemes.

2 Definitions, Notations and Facts

Throughout this paper, X denotes a finite set and $\mathcal{C}$ denotes a concept class over the 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, $[n] := \{1,\dots,n\}$. 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* [15] 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. LC-PARTIAL$(\mathcal{C})$ denotes the smallest number q such that there is some learning algorithm that exactly identifies each 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* [7, 8] 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 SDC$(\mathcal{C})$, is defined as the minimum worst-case number of mistakes that a learning algorithm A can achieve on $\mathcal{C}$, if A exactly identifies every $C^* \in \mathcal{C}$.

The *mistake bound* [13] of a particular learning algorithm A for concept class $\mathcal{C}$, denoted $M_A(\mathcal{C})$, is the worst-case number of 0,1-prediction mistakes made by A on any given sequence of instances labeled consistently according to some target concept from $\mathcal{C}$. The *optimal mistake bound* for a concept class $\mathcal{C}$, denoted $M_{opt}(\mathcal{C})$, is the minimum of $M_A(\mathcal{C})$ over all learning algorithms A.

Clearly, LC-PARTIAL and SDC are monotonic, and M_{opt} is twofold monotonic. The following chain of inequalities is well-known [8, 15]:

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

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

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

The quantity $\mathrm{TD}(\mathcal{C}) := \mathrm{TS}_{max}(\mathcal{C})$ is called the *teaching dimension* of $\mathcal{C}$ [6]. Note that TD is monotonic. A concept class $\mathcal{C}$ consisting of exactly one concept C has teaching dimension 0 because $\emptyset \in \mathcal{TS}(C, \{C\})$.

Definition 2 (see [22]). *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 $\mathrm{ord}(P) := \max_{t=1,\ldots,N-1} |S_t|$ *is called the* order of the teaching plan P*. Finally, we define*

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

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.) As observed in [22], 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 $|S_t| = \mathrm{TS}_{min}(\{C_t, \ldots, C_N\})$ holds for all $t \in \{1, \ldots, N-1\}$.

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 for 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 [16, 9] 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 each 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 examined thoroughly at an early stage of research on computational learning theory [9].

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 [20, 18]. 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 [21, 5] that a class over a domain X is maximum iff Sauer's inequality holds with equality for $X' = X$.

The following definition is from [12]:

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:*

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

(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 [12] that a representation mapping with the non-clashing property guarantees that, for every sample S labeled according to a concept from $\mathcal{C}$, 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 viewed as having color i).

The following definitions are from [17] (although, stylistically, we are stressing here the similarities to 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.*

$\mathcal{C}$ 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 $H(C, C')$ that connects C and C', where H denotes the Hamming distance. [17] showed the following:

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

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

Although it was known before [12] that any maximum class has an unlabeled compression scheme, the scheme resulting from corner-peeling has some very special and nice features. We will see an application in Section 5, where we show that $\mathrm{RTD}(\mathcal{C}) = \mathrm{VCD}(\mathcal{C})$ for every maximum class $\mathcal{C}$.

3 Recursive Teaching and Query Learning

Kuhlmann proved the following result:

Lemma 7 (see [11]). *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. $\mathrm{RTD}(\mathcal{C}) \leq \mathrm{SDC}(\mathcal{C}) \leq$ *LC-PARTIAL*$(\mathcal{C}) \leq M_{opt}(\mathcal{C})$.
2. $\mathrm{RTD}^*(\mathcal{C}) \leq M_{opt}(\mathcal{C})$.

As demonstrated in [8], 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 [8, 22], illustrates this:

Corollary 9. *1. If* $\mathrm{VCD}(\mathcal{C}) = 1$, *then* $\mathrm{RTD}(\mathcal{C}) = \mathrm{SDC}(\mathcal{C}) = 1$.
2. RTD(*Monotone Monomials*) = SDC(*Monotone Monomials*) = 1.
3. RTD(*Monomials*) = SDC(*Monomials*) = 2.
4. $\mathrm{RTD}(\mathrm{BOX}_n^d) = \mathrm{SDC}(\mathrm{BOX}_n^d) = 2$.
5. RTD(*m-Term Monotone DNF*) $\leq$ SDC(*m-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 are from [8] and, as mentioned above, they apply to RTD as well. Lower bound 1 on RTD (for concept classes with at most two distinct concepts) is trivial. RTD(Monomials) = 2 is shown in [22]. 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 zeros (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
- $l_1, \dots, l_m$, 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 $l_1, \dots, l_m$. 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. [15] 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, $\emptyset \subset \{1\} \subset \{1,2\} \subset \cdots \subset \{1,2,\dots,n\}$, has length $n+1$, but its self-directed learning complexity is 1. Theorem 8 in [3] implies that SDC is lower-bounded by $\log|\mathcal{C}|/\log|X|$ if $\mathrm{SDC}(\mathcal{C}) \geq 2$. A similar bound applies to RTD:

Lemma 10. *Suppose* $\mathrm{RTD}(\mathcal{C}) \geq 2$. *Then,* $\mathrm{RTD}(\mathcal{C}) \geq \frac{\log|\mathcal{C}|}{1+\log|X|}$ *and repetition-free teaching plans for* $\mathcal{C}$ *are of order at least* $\frac{\log|\mathcal{C}|}{\log|X|}$.

Proof. Let $k := \mathrm{RTD}(\mathcal{C})$, and let P be a teaching plan of order k for $\mathcal{C}$. Clearly, P contains $|\mathcal{C}|$ pairwise different teaching sets, and every teaching set is a labeled subset of X of size at most k. Thus,

$$|\mathcal{C}| \leq \sum_{i=1}^{k} \binom{|X|}{i} 2^i \leq 2^k \Phi_k(|X|) \leq (2|X|)^k . \tag{6}$$

Solving for k yields the desired lower bound on $\mathrm{RTD}(\mathcal{C})$. In a similar calculation for repetition-free teaching plans, a factor 2^i (and later 2^k) is missing in (6). □

[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.

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 $M = [M'|\mathbf{0}]$ where M' is a lower-triangular square-matrix with ones on the main-diagonal and $\mathbf{0}$ denotes the all-zeros 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 $(x_1, 0), \ldots, (x_{N-1}, 0)$, 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. □

Note that concept classes of recursive teaching dimension 1 can have arbitrarily large VC-dimension. However, [11] presents a family $(\mathcal{C}_m)_{m \geq 1}$ of concept classes such that $\mathrm{VCD}(\mathcal{C}_m) = 2m$ but $\mathrm{RTD}(\mathcal{C}_m) \geq \mathrm{TD}_{min}(\mathcal{C}_m) = 3m$. This shows that RTD cannot generally be upper-bounded by the VC-dimension (but leaves open the possibility of an upper bound of the form $O(\mathrm{VCD}(\mathcal{C}))$).

As we have seen in this section, the gap between $\mathrm{SDC}(\mathcal{C})$ and 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 of the fact that both measures assign value 1 to the same family of concept classes).

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

4 Recursive Teaching and Intersection-Closed Classes

As shown by Kuhlmann [11], $\mathrm{TS}_{min}(\mathcal{C}) \le I(\mathcal{C})$ holds for every intersection-closed concept class $\mathcal{C}$. Kuhlmann's central argument (which occurred first in a proof of a related result in [8]) can be applied recursively so that the following is obtained:

Lemma 12. *For every intersection-closed class $\mathcal{C}$, $\mathrm{RTD}(\mathcal{C}) \le I(\mathcal{C})$.*

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:

- $|S_i| \le k$ 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 $I(\mathcal{C}) \le \mathrm{VCD}(\mathcal{C})$, we get

Corollary 13. *For every intersection-closed class $\mathcal{C}$, $\mathrm{RTD}(\mathcal{C}) \le \mathrm{VCD}(\mathcal{C})$.*

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

For every fixed constant d (e.g., $d = 2$), [11] presents a family $(\mathcal{C}_m)_{m \ge 1}$ of intersection-closed concept classes such that the following holds:[3]

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

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 12 generalizes to nested differences:

Theorem 14. *If $\mathcal{C}$ is intersection-closed then $\mathrm{RTD}(\mathrm{DIFF}^{\le d}(\mathcal{C})) \le d \cdot I(\mathcal{C})$.*

Proof. Any concept $C \in \mathrm{DIFF}^{\le d}(\mathcal{C})$ 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 $\mathrm{DIFF}^{\le d}(\mathcal{C})$ as follows. Let C be a concept with a minimal representation of the form (8), and let the

[3] A family satisfying (7) but *not* being intersection-closed was presented previously [3].

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 $\mathrm{DIFF}^{\le d}(\mathcal{C})$. Therein, the concepts are in lexicographic order so that, when teaching concept C with minimal representation (8), the concepts preceding C w.r.t. $\sqsupset$ are discarded already. A teaching set T for C is then obtained as follows:

– For every $j = 1, \dots, d$, 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 $\mathrm{DIFF}^{\le d}(\mathcal{C})$ that is consistent with T. Since concepts being lexicographically larger than C are discarded already, T is a teaching set for C. □

Corollary 15. *Let $\mathcal{C}_1, \dots, \mathcal{C}_r$ be intersection-closed classes over the domain X. Assume that the "universal concept" X belongs to each of these classes.*[4] *Then,*

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

Proof. Consider the concept class $\mathcal{C} := \mathcal{C}_1 \wedge \dots \wedge \mathcal{C}_r := \{C_1 \cap \dots \cap C_r \mid C_i \in \mathcal{C}_i \text{ for } i = 1, \dots, r\}$. According to [9], we have:

1. $\mathcal{C}_1 \cup \dots \cup \mathcal{C}_r$ is a subclass of $\mathcal{C}$.
2. $\mathcal{C}$ is intersection-closed.
3. Let $C = C_1 \cap \dots \cap C_r \in \mathcal{C}$. 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 $S_1 \cup \dots \cup S_r$ is a spanning set for C w.r.t. $\mathcal{C}$.

Thus $I(\mathcal{C}) \le I(\mathcal{C}_1) + \dots + I(\mathcal{C}_r)$. The corollary follows from Theorem 14. □

5 Recursive Teaching Dimension and 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 $k \le \mathrm{VCD}(\mathcal{C})$ instances is shattered, which implies $\mathrm{RTD}(\mathcal{C}) \ge \mathrm{TS}_{min}(\mathcal{C}) \ge \mathrm{VCD}(\mathcal{C})$. Thus, we can focus on the reverse direction and pursue the question whether $\mathrm{RTD}(\mathcal{C}) \le \mathrm{VCD}(\mathcal{C})$. We shall answer this question to 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 $t = 1, \dots, N$, C'_t is a cube in $\{C_t, \dots, C_N\}$ which contains C_t and whose colors (augmented by their labels according to C_t) form a teaching set for $C_t \in \{C_t, \dots, C_N\}$.

[4] This assumption is not restrictive: adding the universal concept to an intersection-closed class does not destroy the intersection-closedness.

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 16. *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 17. *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 18. *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 [12], we obtain:

Corollary 19. *Every corner-peeling plan for a maximum class is strong, and therefore induces a teaching plan of the same order.*

Since [17] showed that every maximum class $\mathcal{C}$ can be VCD($\mathcal{C}$)-corner-peeled, we may conclude that $\mathrm{RTD}(\mathcal{C}) \le \mathrm{VCD}(\mathcal{C})$. As mentioned above, $\mathrm{RTD}(\mathcal{C}) \ge \mathrm{VCD}(\mathcal{C})$ for every maximum class $\mathcal{C}$, which implies

Corollary 20. *For every maximum class $\mathcal{C}$,* $\mathrm{RTD}(\mathcal{C}) = \mathrm{VCD}(\mathcal{C})$*.*

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 close this section by establishing a connection between repetition-free teaching plans and representations having the acyclic non-clashing property:

Lemma 21. *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 $\{C_{t+1}, \ldots, C_N\}$. Repetition-freeness is obvious since r is injective. □

Corollary 22. *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.*

An inspection of [17] 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 $\mathcal{C}$). An inspection of [12] 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 $\mathcal{C}$.

6 Conclusions

This paper relates the RTD, a recent teaching complexity notion, to classical learning complexity parameters. One of these parameters is SDC, the complexity of self-directed learning—the most information-efficient query model known to date. Our result 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.

Another important parameter in our comparison is the VC-dimension. Although the VC-dimension can be arbitrarily large for classes of recursive teaching dimension 1 (see Theorem 11 and the remark thereafter) and arbitrarily smaller than SDC [3, 11], it does not generally lie 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 is possible for the RTD.

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 refers to a combinatorial structure that is of high relevance for the complexity of information-efficient learning and sample compression. Analyzing the question whether teaching plans defining the RTD can in general 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 Malte Darnstädt and Michael Kallweit for helpful and inspiring discussions. Moreover, we would like to thank three anonymous referees for many helpful suggestions and for pointing out a gap in the proof of Theorem 14.

References

[1] Angluin, D.: Queries and concept learning. Mach. Learn. 2, 319–342 (1988)

[2] Balbach, F.: Measuring teachability using variants of the teaching dimension. Theoret. Comput. Sci. 397, 94–113 (2008)

[3] Ben-David, S., Eiron, N.: Self-directed learning and its relation to the VC-dimension and to teacher-directed learning. Mach. Learn. 33, 87–104 (1998)

[4] Ben-David, S., Litman, A.: Combinatorial variability of Vapnik-Chervonenkis classes with applications to sample compression schemes. Discrete Appl. Math. 86(1), 3–25 (1998)

[5] Floyd, S., Warmuth, M.: Sample compression, learnability, and the vapnik-chervonenkis dimension. Mach. Learn. 21(3), 269–304 (1995)

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

[7] Goldman, S., Rivest, R., Schapire, R.: Learning binary relations and total orders. SIAM J. Comput. 22(5), 1006–1034 (1993)

[8] Goldman, S., Sloan, R.: The power of self-directed learning. Mach. Learn. 14(1), 271–294 (1994)

[9] Helmbold, D., Sloan, R., Warmuth, M.: Learning nested differences of intersection-closed concept classes. Mach. Learn. 5, 165–196 (1990)

[10] Jackson, J., Tomkins, A.: A computational model of teaching. In: 5th Annl. Workshop on Computational Learning Theory, pp. 319–326 (1992)

[11] Kuhlmann, C.: On teaching and learning intersection-closed concept classes. In: Fischer, P., Simon, H.U. (eds.) EuroCOLT 1999. LNCS (LNAI), vol. 1572, pp. 168–182. Springer, Heidelberg (1999)

[12] Kuzmin, D., Warmuth, M.: Unlabeled compression schemes for maximum classes. J. Mach. Learn. Research 8, 2047–2081 (2007)

[13] Littlestone, N.: Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. Mach. Learn. 2(4), 285–318 (1988)

[14] Littlestone, N., Warmuth, M.: Relating data compression and learnability. Technical report, UC Santa Cruz (1986)

[15] Maass, W., Turán, G.: Lower bound methods and separation results for on-line learning models. Mach. Learn. 9, 107–145 (1992)

[16] Natarajan, B.: On learning boolean functions. In: 19th Annl. Symp. Theory of Computing, pp. 296–304 (1987)

[17] Rubinstein, B., Rubinstein, J.: A geometric approach to sample compression (2009) (unpublished manuscript)

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

[19] Shinohara, A., Miyano, S.: Teachability in computational learning. New Generat. Comput. 8, 337–348 (1991)

[20] Vapnik, V., Chervonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. Theor. Probability and Appl. 16, 264–280 (1971)

[21] Welzl, E.: Complete range spaces (1987) (unpublished notes)

[22] Zilles, S., Lange, S., Holte, R., Zinkevich, M.: Teaching dimensions based on cooperative learning. In: 21st Annl. Conf. Learning Theory, pp. 135–146 (2008)