Teaching Dimensions Based on Cooperative Learning

Sandra Zilles[1]**, Steffen Lange**[2]**, Robert Holte**[1]**, and Martin Zinkevich**[3]

[1] University of Alberta, Dept. of Computing Science, Edmonton, AB, Canada, {zilles,holte}@cs.ualberta.ca
[2] Darmstadt University of Applied Sciences, Dept. of Computer Science, Darmstadt, Germany, s.lange@fbi.h-da.de
[3] Yahoo! Research, Mission College, CA, USA, maz@yahoo-inc.com

Abstract

The problem of how a teacher and a learner can cooperate in the process of learning concepts from examples in order to minimize the required sample size without "coding tricks" has been widely addressed, yet without achieving teaching and learning protocols that meet what seems intuitively an optimal choice for selecting samples in teaching.

We introduce the model of subset teaching sets, based on the idea that both teacher and learner can exploit the assumption that the partner is cooperative. We show how this can reduce the sample size drastically without using coding tricks. For instance, monomials can be taught with only two examples independent of the number of variables.

The corresponding variant of the teaching dimension (STD) turns out to be nonmonotonic with respect to subclasses of concept classes. We discuss why this nonmonotonicity might be inherent in optimal cooperative teaching scenarios. Nevertheless, trying to overcome nonmonotonicity, we introduce a second variant, the recursive teaching dimension (RTD), which is monotonic and yields the same positive results for some concept classes, such as the class of all monomials, yet can be arbitrarily worse than the STD.

1 Introduction

1.1 Motivation and approach

One major branch of learning theory and machine learning is the theory and practice of learning concepts from examples. Considering a finite instance space and a class of (thus finite) concepts over that space, it is obvious that each concept can be uniquely determined if enough examples are known. Much less obvious is how to minimize the number of examples required to identify a concept, and with this aim in mind models of *cooperative learning* and learning from *good examples* were designed and analyzed. The selection of good examples to be presented to a learner is often modeled using a teaching device (teacher) that is assumed to be benevolent by selecting examples expediting the learning process (see for instance [AK97, JT92, GM96, Mat97]).

Throughout this paper we assume that teaching/learning proceeds stepwise; in each step the teacher presents an example (that is, an instance paired with a label 1 or 0, according to whether or not the instance belongs to the target concept) to the learner and the learner returns a concept it believes to be the target concept. If the learner's conjecture is right the process ends, otherwise both proceed to the next step. This process will terminate successfully for any concept c in a given concept class C if the following three conditions hold: (1) the teacher never presents any example twice, (2) the teacher labels the examples correctly according to the current target concept, and (3) the learner always returns a concept consistent with the examples seen so far. The sample size, i.e., the number of examples the teacher presents to the learner enroute to termination, is the object of optimization; in particular we are concerned with the worst case sample size measured over all concepts in C. Other than that, computational complexity issues are not the focus of this paper.

A typical question is *How can a teacher and a learner cooperatively minimize the worst case sample size without using coding tricks?*—a coding trick being, e.g., any *a priori* agreement on encoding concepts in examples, depending on the concept class C. For instance, if teacher and learner agreed on a specific order for the concept representations and the instances and agreed to use the j^{th} instance in this ordering to teach the j^{th} concept, that would be a coding trick.[1]

A considerable amount of the learning theory literature deals with the teaching dimension of concept classes (and variants thereof, see, e.g., [SM91, GK95, ABCS92]). The teaching dimension of a concept $c \in C$ is the size of the minimum sample that is consistent with c but not with any other concept in C. Obviously teacher and learner can succeed with such a sample without coding tricks.

The teaching dimension however does not always seem to capture the intuitive idea of cooperation in teaching and learning. Consider the following simple example. Let C_0 consist of the empty concept and all singleton concepts over a given instance space $X = \{x_1, \dots, x_n\}$. Each singleton concept $\{x_i\}$ has a teaching dimension of 1, since the single positive example $(x_i, 1)$ is sufficient for determining

[1] There is so far no generally accepted definition of what a coding trick (sometimes also called "collusion") in general is. The reader is referred to [AK97, OS02, GM96] for a treatment of this question in different learning models.

$\{x_i\}$. In contrast to that, the empty concept has a teaching dimension of n—every example has to be presented. However, if the learner assumed the teacher was cooperative—and would therefore present a positive example if the target concept was non-empty—the learner could confidently conjecture the empty concept upon seeing just one negative example.

Let us extend this reasoning to a slightly more complex example, the class of all boolean functions that can be represented as a monomial over m variables ($m = 4$ in this example). Imagine yourself in the role of a learner knowing your teacher will present helpful examples. If the teacher sent you the examples

$$(0100, 1), (0111, 1)\,,$$

what would be your conjecture? Presumably most people would conjecture the monomial $M \equiv \overline{v_1} \wedge v_2$, as does for instance the algorithm proposed in [Val84]. Note that this choice is not uniquely determined by the data: the empty monomial and the monomials $\overline{v_1}$ and v_2 are also consistent with these examples. And yet M seems the best choice, because we'd think the teacher would not have kept any bit in the two examples constant if it was not in the position of a relevant variable. In this example, the natural conjecture is the most specific concept consistent with the sample, but that does not, in general, capture the intuitive idea of cooperative learning. For example, consider the concept class consisting of just the three concepts $\{\beta\}$, $\{\alpha, \beta\}$, $\{\alpha, \gamma\}$. If the teacher presented $(\alpha, 1)$ as an example, there would be two most specific consistent concepts. But a learner that assumed the teacher was cooperative could confidently guess $\{\alpha, \beta\}$ to be the target concept, because a cooperative teacher would have presented the unambiguous $(\gamma, 1)$ if $\{\alpha, \gamma\}$ was the target concept.

Could the learner's reasoning about the teacher's behavior in these examples be implemented without a coding trick? We will show below that no coding trick is necessary to achieve exactly this behavior of teacher and learner; there is a general principle that teachers and learners can independently implement to cooperatively learn any finite concept class. When applied to the class of monomials this principle enables any monomial to be learned from just two examples, regardless of the number m of variables.

Our approach is to define a new model of cooperation in learning, based on the idea that each partner in the cooperation tries to reduce the sample size by exploiting the assumption that the other partner does so. If this idea is iteratively propagated by both partners, one can refine teaching sets iteratively ending up with a framework for highly efficient teaching and learning without any coding tricks. It is important to note that teacher and learner do not agree on any order of the concept class or any order of the instances. All they know about each others' strategies is a general assumption about how cooperation should work independent of the concept class or its representation.

We show that the resulting variant of the teaching dimension—called the *subset teaching dimension (STD)*—is not only a uniform lower bound of the teaching dimension but can be constant where the original teaching dimension is exponential, even in cases where only one iteration is needed. For example, as illustrated above, the STD of the class of monomials over m variables is 2, in contrast to its original teaching dimension of 2^m.

Some examples however will reveal a nonmonotonicity of the subset teaching dimension: some classes possess subclasses with a higher subset teaching dimension, which is at first glance not very intuitive. We will explain below why in a cooperative model such a nonmonotonicity does not have to contradict intuition; additionally we introduce a second model of cooperative teaching and learning, that results in a monotonic dimension, called the *recursive teaching dimension (RTD)*. Comparing our complexity notions in terms of the sample size required for teaching and learning shows that achieving monotonicity here results in a loss in terms of sample efficiency; however, even though the RTD has some deficiencies compared to the STD, it still significantly improves on previously studied variants of the teaching dimension.

1.2 Related work

The problem of defining good or helpful examples in learning has been studied in different fields of learning theory. Various learning models that involve one particular teacher can be found in [AK97, JT92, GM96, Mat97]; these mostly focus on learning boolean functions.

The teaching dimension has been analyzed in the context of online learning [BE98, RY95] and in the model of learning from queries, e.g., in [Heg95] and in [Han07], with a focus on active learning in the PAC framework. In contrast to these models, in inductive inference the learning process is not necessarily considered to be finite. Approaches to defining learning infinite concepts from good examples [FKW93, LNW98] do not focus on the size of a finite sample of good examples, but rather on characterizing the cases in which learners can identify concepts from only finitely many examples.

The approach we present in this paper is mainly based on an idea by Balbach [Bal08]. He defined and analyzed a model in which, under the premise that the teacher uses a minimal teaching set as a sample, a learner can reduce the size of a required sample by eliminating concepts which possess a teaching set smaller than the number of examples provided by the teacher so far. Iterating this idea, the size of the teaching sets might be gradually reduced significantly. Though our approach is syntactically quite similar to Balbach's, the underlying idea is a different one (we do not consider elimination by the sample size but elimination by the sample content as compared to all possible teaching sets). The resulting variant of the teaching dimension in general yields a much better performance in terms of sample size than Balbach's model does.

2 Preliminaries

Let $\mathbb{N}$ denote the set of all non-negative integers, $\emptyset$ denote the empty set, and $|A|$ denote the cardinality of a finite set A. Concerning the teaching framework, we will mostly follow the notation used in [Bal08].

In the models of teaching and learning to be defined below, we will always assume that the goal in an interaction between a teacher and a learner is to make the learner identify a (finite) concept c over a (finite) instance space X. To formalize this, let $n > 0$ be a natural number and let $X =$

$\{x_1, \ldots, x_n\}$ be an instance space. A *concept* c is a subset of X and a *concept class* C is a set of concepts. Consequently, concepts and concept classes considered below will always be finite. As a special case we sometimes consider boolean functions over variables $v_1, \ldots, v_m$ as concepts, which just means to represent the instance space X by $\{0,1\}^m$.

We identify every concept c with its membership function given by $c(x_i) = 1$ if $x_i \in c$, and $c(x_i) = 0$ if $x_i \notin c$, where $1 \leq i \leq n$. Given a *sample*, i.e., a set $S = \{(y_1, b_1), \ldots, (y_j, b_j)\} \subseteq X \times \{0,1\}$ of labeled *examples*, we say that c is consistent with S if $c(y_i) = b_i$ for all $i \in \{1, \ldots, j\}$. If C is a concept class then we define

$$Cons(S, C) = \{c \in C \mid c \text{ is consistent with } S\} .$$

The sample S is called a *teaching set* for c with respect to C if $Cons(S, C) = \{c\}$. A teaching set allows a learning algorithm to uniquely identify a concept in the concept class C. Striving for sample efficiency, one is particularly interested in teaching sets of minimal size, called *minimal teaching sets*. The *teaching dimension* of c in C is the size of such a minimal teaching set, i.e., $TD(c, C) = \min\{|S| \mid Cons(S, C) = \{c\}\}$, the worst case of which defines the teaching dimension of C, i.e., $TD(C) = \max\{TD(c, C) \mid c \in C\}$. To refer to the set of all minimal teaching sets of c with respect to C, we use

$$TS(c, C) = \{S \mid Cons(S, C) = \{c\} \text{ and } |S| = TD(c, C)\} .$$

The reader is referred to [GK95, SM91] for original studies on teaching sets.

Recall our assumptions concerning the learning process: it proceeds stepwise; in each step the teacher presents a single example to the learner and the learner returns a conjecture about the target concept. The process stops when and only when a correct conjecture is made by the learner. Our minimal requirements on cooperative partners here is that teachers never present any example twice and always label the examples correctly according to the target concept, and that every conjecture a learner returns is consistent with the information seen up to that step.

The teaching dimension [GK95] then gives a measure of the worst case sample size needed by a learner if the teacher uses only minimal teaching sets for teaching. The reason is that a teaching set eliminates all but one concept due to inconsistency. However, if the learner knows $TD(c, C)$ for every $c \in C$ then sometimes concepts could also be eliminated by the mere number of examples presented to the learner. For instance, assume a learner knows that all but one concept $c \in C$ have a teaching set of size one and that the teacher will teach using teaching sets. After having seen 2 examples, no matter what they are, the learner could eliminate all concepts but c. This idea, referred to as elimination by sample size, was introduced in [Bal08]. If a teacher knew that a learner eliminates by consistency and by sample size then the teacher could consequently reduce some teaching sets (e.g, here, if $TD(c, C) \geq 3$, a new "teaching set" for c could be built consisting of only 2 examples).

More than that—this idea is iterated by Balbach [Bal08]: if the learner knew that the teacher uses such reduced "teaching sets" then the learner could adapt his assumption on the size of the samples to be expected for each concept, which could in turn result in a further reduction of the "teaching sets" by the teacher and so on. The following definition captures this idea formally.

Definition 1 (Balbach teaching dimension [Bal08]) *Let C be a concept class, $c \in C$, and S a sample. Let $BTD^0(c, C) = TD(c, C)$. We define iterated dimensions for all $k \in \mathbb{N}$ as follows.*

- $Cons_{size}(S, C, k)$ $= \{c \in Cons(S, C) \mid BTD^k(c, C) \geq |S|\}$.
- $BTD^{k+1}(c, C)$ $= \min\{|S| \mid Cons_{size}(S, C, k) = \{c\}\}$

Let z be minimal such that $BTD^{z+1}(c, C) = BTD^z(c, C)$ for all $c \in C$. The iterated Balbach teaching dimension of c in C is defined by $BTD(c, C) = BTD^z(c, C)$ and the iterated Balbach teaching dimension of the class C is $BTD(C) = \max\{BTD(c, C) \mid c \in C\}$.[2]

Obviously, $BTD(C) \leq TD(C)$ for every concept class C. How much the sample complexity can actually be reduced by a cooperative teacher/learner pair according to this "elimination by sample size" principle, is illustrated by the concept class C_0 consisting of the empty concept and all singleton concepts over X. The teaching dimension of this class is n, whereas the BTD is 2. A more interesting example is the class of monomials, which contains only one concept for which the BTD-iteration yields an improvement.

Theorem 2 (Balbach [Bal08]) *Let $m \in \mathbb{N}$ and C the class of all boolean functions over $m \geq 2$ variables that can be represented by a monomial. Let $c_0 = \emptyset$ be the concept represented by a contradictory monomial.*

1. $BTD(c_0, C) = m + 2 < 2^m = TD(c_0, C)$.
2. $BTD(c, C) = TD(c, C)$ *for all* $c \in C$ *with* $c \neq c_0$.

The intuitive reason why $BTD(c_0, C) = m+2$ in Theorem 2 is that samples for c_0 of size $m+1$ or smaller are consistent also with monomials different from c_0. These other monomials hence cannot be eliminated—neither by size nor by inconsistency.

3 Teaching and learning using subset teaching sets

3.1 The model

The approach studied by Balbach [Bal08] does not fully meet the intuitive idea of teacher and learner exploiting the knowledge that either partner behaves cooperatively. Consider for instance one more time the class C_0 containing the empty concept and all singletons over $X = \{x_1, \ldots, x_n\}$. Each concept $\{x_i\}$ has the unique minimal teaching set $\{(x_i, 1)\}$ in this class, whereas the empty concept only has a teaching set of size n, namely $\{(x_1, 0), \ldots, (x_n, 0)\}$. The idea of elimination by size allows a learner to conjecture the empty

[2] [Bal08] denotes this by *IOTTD*, called iterated optimal teacher teaching dimension; we deviate from this notation for the sake of convenience.

concept as soon as two examples have been provided, due to the fact that all other concepts possess a teaching set of size one. This is why the empty concept has an BTD equal to 2 in this example.

However, as we have argued in the introduction, it would also make sense to devise a learner in a way to conjecture the empty concept as soon as a first example for that concept is provided—knowing that the teacher would not use a negative example for any other concept in the class. In terms of teaching sets this means to reduce the teaching sets to their minimal subsets that are not contained in minimal teaching sets for other concepts in the given concept class.

Formally, we define this refinement operator and its iteration as follows.

Definition 3 *Let C be a concept class, $c \in C$, and S a sample. Let $STD^0(c, C) = TD(c, C)$, $STS^0(c, C) = TS(c, C)$. We define iterated sets for all $k \in \mathbb{N}$ as follows.*

- $Cons_{sub}(S, C, k) = \{c \in C \mid S \subseteq S' \text{ for some } S' \in STS^k(c, C)\}$.
- $STD^{k+1}(c, C) = \min\{|S| \mid Cons_{sub}(S, C, k) = \{c\}\}$
- $STS^{k+1}(c, C) = \{S \mid Cons_{sub}(S, C, k) = \{c\}, |S| = STD^{k+1}(c, C)\}$.

Let z be minimal such that $STS^{z+1}(c, C) = STS^z(c, C)$ for all $c \in C$.[3]

A sample S with $Cons_{sub}(S, C, z) = \{c\}$ is called a subset teaching set for c in C. The subset teaching dimension of c in C is defined as $STD(c, C) = STD^z(c, C)$ and we denote by $STS(c, C) = STS^z(c, C)$ the set of all minimal subset teaching sets for c in C. The subset teaching dimension of C is $STD(C) = \max\{STD(c, C) \mid c \in C\}$.

For illustration, consider again the concept class C_0, i.e., $C_0 = \{c_i \mid 0 \le i \le n\}$, where $c_0 = \emptyset$ and $c_i = \{x_i\}$ for all $i \in \{1, \ldots, n\}$. Obviously, for $k \ge 1$,

$$STS^k(c_i) = \{\{(x_i, 1)\}\} \text{ for all } i \in \{1, \ldots, n\}$$

and

$$STS^k(c_0) = \{\{(x_i, 0)\} \mid 1 \le i \le n\}\,.$$

Hence $STD(C_0) = 1$.

The definition of $STS(c, C)$ induces a protocol for teaching and learning: for a target concept c, a teacher presents the examples in a subset teaching set for c to the learner. The learner will also be able to pre-compute all subset teaching sets for all concepts and determine the target concept from the sample provided by the teacher.[4]

Protocol 4 *Let C be a concept class.*

0. *Teacher and learner both compute $STS(c, C)$ for all $c \in C$.*

Let $c \in C$ be a target concept known to the teacher.

1. *The teacher chooses a set $S \in STS(c, C)$ at random.*
2. *The teacher presents S to the learner (stepwise/batch).*
3. *The learner looks up and identifies the unique concept $c \in C$ for which $S \in STS(c, C)$.*

It is important to note at this point that Definition 3 as such is independent of the particular shape or structure of the concept class. It does not presume any special order of the concept representations or of the instances, i.e., teacher and learner do not have to agree on any such order to make use of the teaching and learning protocol. That means, given a special concept class C, the computation of its subset teaching sets does not involve any special coding trick depending on C—it just follows a general rule.

3.2 Comparison to the Balbach teaching dimension

Obviously, Protocol 4 based on the subset teaching dimension never requires a sample larger than a teaching set; often a smaller sample is sufficient. Similarly, the subset teaching dimension compares to the Balbach teaching dimension as follows.

Proposition 5 1. *$STD(C) \le BTD(C)$ for every concept class C.*

2. *There is a concept class C with $STD(C) < BTD(C)$.*

Proof. Assertion (1) immediately follows from the definitions. Informally, if a (Balbach) teaching set S in one iteration for a concept c is going to be reduced according to the BTD-rule (see Definition 1), then $|S| \ge |S'| + 2$ for every (Balbach) teaching set S' on the current state of iteration for some concept $c' \neq c$ consistent with S. In particular, if the Balbach teaching dimension of c is reduced to some value $u < |S|$, then S has got a subset of size u (or even smaller) that is not contained in any teaching set for any concept $c' \neq c$ in C. The minimal such subset has cardinality at most u and is at least as big as a minimal subset teaching set for c.

Assertion (2) is witnessed by the class C_0 containing the empty concept and all singletons over X. ■

The second assertion of this proposition even holds in a stronger form, see Theorem 6.

Theorem 6 *For each $u \in \mathbb{N}$ there is a concept class C such that $STD(C) = 1$ and $BTD(C) = u$.*

Proof. Let $n = 2^u + u$ be the number of instances in X. Define a concept class $C = C^u_{0/1}$ as follows. For every $s = (s_1, \ldots, s_u) \in \{0, 1\}^u$, C contains the concepts $c_{s,0} = \{x_i \mid 1 \le i \le u \text{ and } s_i = 1\}$ and $c_{s,1} = c_{s,0} \cup \{x_{u+1+int(s)}\}$. Here $int(s) \in \mathbb{N}$ is defined by $int(s) = \sum_{i=0}^{u-1} s_{i+1} \cdot 2^i$. We claim that $STD(C) = 1$ and $BTD(C) = u$.

Let $s = (s_1, \ldots, s_u) \in \{0, 1\}^u$. Then

$$\begin{aligned} TS(c_{s,0}, C) &= \{\{(x_i, s_i) \mid 1 \le i \le u\} \\ &\quad \cup \{(x_{u+1+int(s)}, 0)\}\} \\ TS(c_{s,1}, C) &= \{\{(x_{u+1+int(s)}, 1)\}\} \end{aligned}$$

Since for each $c \in C$ the minimal teaching set for c with respect to C contains an example that does not occur in the

[3] Such a z exists because $STD^0(c, C)$ is finite and can hence be reduced only finitely often.

[4] Note that we focus on sample size here, but neglect efficiency issues arising from the pre-computation of all subset teaching sets.

minimal teaching set for any other concept $c' \in C$, one obtains $STD(C) = 1$ in just one iteration. See Table 1 for the case $u = 2$.

In contrast to that, we obtain $BTD^0(c_{s,0}, C) = u + 1$, $BTD^1(c_{s,0}, C) = u$, and $BTD^0(c_{s,1}, C) = 1$ for all $s \in \{0,1\}^u$. Consider any $s \in \{0,1\}^u$ and any sample $S \subseteq \{(x, c_{s,0}(x)) \mid x \in X\}$ with $|S| = u - 1$. Clearly there is some $s' \in \{0,1\}^u$ with $s' \neq s$ such that $c_{s',0} \in Cons(S, C)$. So $|Cons(S, C, 1)| > 1$ and in particular $Cons(S, C, 1) \neq \{c_{s,0}\}$. Hence $BTD^2(c_{s,0}, C) = BTD^1(c_{s,0}, C)$, which finally implies $BTD(C) = u$. ■

concept	STS^0	STS^1
$\emptyset$	$\{(x_1,0),(x_2,0),(x_3,0)\}$	$\{(x_3,0)\}$
$\{x_3\}$	$\{(x_3,1)\}$	$\{(x_3,1)\}$
$\{x_2\}$	$\{(x_1,0),(x_2,1),(x_4,0)\}$	$\{(x_4,0)\}$
$\{x_2,x_4\}$	$\{(x_4,1)\}$	$\{(x_4,1)\}$
$\{x_1\}$	$\{(x_1,1),(x_2,0),(x_5,0)\}$	$\{(x_5,0)\}$
$\{x_1,x_5\}$	$\{(x_5,1)\}$	$\{(x_5,1)\}$
$\{x_1,x_2\}$	$\{(x_1,1),(x_2,1),(x_6,0)\}$	$\{(x_6,0)\}$
$\{x_1,x_2,x_6\}$	$\{(x_6,1)\}$	$\{(x_6,1)\}$

Table 1: Iterated subset teaching sets for the class $C^u_{0/1}$ with $u = 2$, where $C^u_{0/1} = \{c_{00,0}, c_{00,1} \ldots, c_{11,0}, c_{11,1}\}$ with $c_{00,0} = \emptyset$, $c_{00,1} = \{x_3\}$, $c_{01,0} = \{x_2\}$, $c_{01,1} = \{x_2, x_4\}$, $c_{10,0} = \{x_1\}$, $c_{10,1} = \{x_1, x_5\}$, $c_{11,0} = \{x_1, x_2\}$, $c_{11,1} = \{x_1, x_2, x_6\}$.

3.3 Teaching monomials

This section provides an analysis of the STD for a more natural example, the monomials, showing that the very intuitive example given in the introduction is indeed what a cooperative teacher and learner in our model would come up with. The main result is that the STD of the class of all monomials is 2, independent on the number m of variables, whereas its teaching dimension is exponential in m and its BTD is linear in m, cf. [Bal08].

Theorem 7 *Let $m \in \mathbb{N}$ and C the class of all boolean functions over m variables that can be represented by a monomial. Then $STD(C) = 2$.*

Proof. Let $m \in \mathbb{N}$ and $s = (s_1, \ldots, s_m)$, $s' = (s'_1, \ldots, s'_m)$ elements in $\{0,1\}^m$. Let $\triangle(s, s')$ denote the Hamming distance of s and s', i.e., $\triangle(s, s') = \sum_{1 \leq i \leq m} |s(i) - s'(i)|$.

We distinguish the following types of monomials M over m variables.

Type 1: M is the empty monomial.

Type 2: M has got m variables, $M \not\equiv v_1 \wedge \overline{v_1}$.

Type 3: M has got k variables, $1 \leq k < m$, $M \not\equiv v_1 \wedge \overline{v_1}$.

Type 4: M is contradictory, i.e., $M \equiv v_1 \wedge \overline{v_1}$.

The following facts state some properties of the corresponding minimal teaching sets.

Fact 1: If M is of type 1 and $S \in STS^0(M, C)$, then S contains two positive examples of Hamming distance m.

Fact 2: If M is of type 2 and $S \in STS^0(M, C)$, then S contains (i) one positive example and (ii) m negative examples, where the Hamming distance between two negative examples is less than m.

Fact 3: If M is of type 3 and $S \in STS^0(M, C)$, then S contains (i) two positive examples of Hamming distance $m - k$ and (ii) k negative examples, where the Hamming distance between each two negative examples is less than m.

Fact 4: If M is of type 4 and $S \in STS^0(M, C)$, then $S = \{(s, 0) \mid s \in \{0,1\}^m\}$.

Fact 5: For every $s \in \{0,1\}^m$ there are two different monomials M, M' of type 3 such that $(s, 1) \in S \cap S'$ for some $S \in STS^0(M, C)$ and some $S' \in STS^0(M', C)$.

Fact 6: For every $s \in \{0,1\}^m$ there are two different monomials M, M' of type 3 such that $(s, 0) \in S \cap S'$ for some $S \in STS^0(M, C)$ and some $S' \in STS^0(M', C)$.

Fact 7: For every $s \in \{0,1\}^m$ there are two different monomials M, M' of type 2 such that $(s, 0) \in S \cap S'$ for some $S \in STS^0(M, C)$ and some $S' \in STS^0(M', C)$.

Fact 8: If M is of type 2, $S \in STS^0(M, C)$ and $S' \subset S$, then there is a monomial M_3 of type 3 such that $S' \subseteq S_3$ for some $S_3 \in STS^0(M_3, C)$.

After the first iteration we obtain the following facts.

Fact 9: If M is of type 1 and $S \in STS^1(M, C)$, then $S \in STS^0(M, C)$.

Fact 10: If M is of type 2 and $S \in STS^1(M, C)$, then $S \in STS^0(M, C)$.

Fact 11: If M is of type 3 and $S \in STS^1(M, C)$, then S contains two positive examples.

Fact 12: If M is of type 4 and $S \in STS^1(M, C)$, then S contains two negative examples of Hamming distance m.

After the second iteration we obtain the following facts.

Fact 13: If M is of type 1 and $S \in STS^2(M, C)$, then $S \in STS^1(M, C)$.

Fact 14: If M is of type 2 and $S \in STS^2(M, C)$, then S contains one positive and one negative example. Moreover, for every $s \in \{0,1\}^m$, there is a monomial M of type 2 such that $(s, 0) \in S$ for some $S \in STS^2(M, C)$.

Fact 15: If M is of type 3 and $S \in STS^1(M, C)$, then $S \in STS^2(M, C)$.

Fact 16: If M is of type 4 and $S \in STS^2(M, C)$, then $S \in STS^1(M, C)$.

Combining the insights achieved so far, it is easily seen that $STD^3(M, C) = STD^2(M, C) = 2$ for all $M \in C$. ■

For illustration of this proof in case $m = 2$ see Table 2.

A further simple example showing that the STD can be constant as compared to an exponential teaching dimension, this time with an STD of 1, is the following.

Let $C^m_{\vee DNF}$ contain all boolean functions over m variables that can be represented by a 2-term DNF of the form $v_1 \vee M$, where M is a monomial that contains, for each i with $2 \leq i \leq m$, either the literal v_i or the literal $\overline{v_i}$. Moreover, $C^m_{\vee DNF}$ contains the boolean function that can be represented by the monomial $M' \equiv v_1$.

Theorem 8 *Let $m \in \mathbb{N}$.*

1. $TD(C^m_{\vee DNF}) = 2^{m-1}$.

2. $STD(C^m_{\vee DNF}) = 1$.

	STS^0	STS^1

v_1	{(10,1),(11,1),(00,0)}
	{(10,1),(11,1),(01,0)}
		{(10,1),(11,1)}

$\overline{v_1}$	{(00,1),(01,1),(10,0)}
	{(00,1),(01,1),(11,0)}
		{(00,1),(01,1)}

v_2	{(01,1),(11,1),(00,0)}
	{(01,1),(11,1),(10,0)}
		{(01,1),(11,1)}

$\overline{v_2}$	{(00,1),(10,1),(01,0)}
	{(00,1),(10,1),(11,0)}
		{(00,1),(10,1)}

$v_1 \wedge v_2$	{(11,1),(01,0),(10,0)}	{(11,1),(01,0),(10,0)}
$v_1 \wedge \overline{v_2}$	{(10,1),(00,0),(11,0)}	{(10,1),(00,0),(11,0)}
$\overline{v_1} \wedge v_2$	{(01,1),(00,0),(11,0)}	{(01,1),(00,0),(11,0)}
$\overline{v_1} \wedge \overline{v_2}$	{(00,1),(01,0),(10,0)}	{(00,1),(01,0),(10,0)}

$v_1 \wedge \overline{v_1}$	{(00,0),(01,0),(10,0),(11,0)}	{(00,0),(01,0)}
		{(00,0),(10,0)}
		{(01,0),(11,0)}
		{(10,0),(11,0)}

λ	{(00,1),(11,1)}
	{(01,1),(10,1)}
		{(00,1),(11,1)}
		{(01,1),(10,1)}

	STS^2	STS^3
v_1	{(10,1),(11,1)}	{(10,1),(11,1)}
$\overline{v_1}$	{(00,1),(01,1)}	{(00,1),(01,1)}
v_2	{(01,1),(11,1)}	{(01,1),(11,1)}
$\overline{v_2}$	{(00,1),(10,1)}	{(00,1),(10,1)}

$v_1 \wedge v_2$	{(11,1),(01,0)}
	{(11,1),(10,0)}
		{(11,1),(01,0)}
		{(11,1),(10,0)}

$v_1 \wedge \overline{v_2}$	{(10,1),(00,0)}
	{(10,1),(11,0)}
		{(10,1),(00,0)}
		{(10,1),(11,0)}

$\overline{v_1} \wedge v_2$	{(01,1),(00,0)}
	{(01,1),(11,0)}
		{(01,1),(00,0)}
		{(01,1),(11,0)}

$\overline{v_1} \wedge \overline{v_2}$	{(00,1),(01,0)}
	{(00,1),(10,0)}
		{(00,1),(01,0)}
		{(00,1),(10,0)}

$v_1 \wedge \overline{v_1}$	{(00,0),(01,0)}
	{(00,0),(10,0)}
	{(01,0),(11,0)}
	{(10,0),(11,0)}
		{(00,0),(01,0)}
		{(00,0),(10,0)}
		{(01,0),(11,0)}
		{(10,0),(11,0)}

λ	{(00,1),(11,1)}
	{(01,1),(10,1)}
		{(00,1),(11,1)}
		{(01,1),(10,1)}

Table 2: Iterated subset teaching sets for the class of all monomials over $m = 2$ variables. Here λ denotes the empty monomial.

Proof. The straightforward details concerning the proof of Assertion (2) are omitted; Assertion (1) can be verified as follows.

Let S be a sample that is consistent with M'. Assume that for some $s \in \{0,1\}^m$, the sample S does not contain the negative example $(s, 0)$. Obviously, there is a 2-term DNF $D \equiv v_1 \vee M$ such that D is consistent with $S \cup \{(s, 1)\}$. Hence S is not a teaching set for M'. Since there are exactly 2^{m-1} 2-term DNFs that represent different functions in C, a teaching set for M' must contain at least 2^{m-1} examples. ■

4 Nonmonotonicity and the recursive teaching dimension

4.1 Nonmonotonicity versus redundancy of variables

Interpreting the subset teaching dimension as a measure of complexity of a concept class in terms of cooperative teaching and learning, we observe a fact that is worth discussing, namely the nonmonotonicity of this complexity notion, as stated by the following theorem.

Theorem 9 *There is a concept class C with $STD(C') > STD(C)$ for some subclass $C' \subset C$.*

Sketch of proof. This is witnessed by the concept classes $C = C^u_{0/1}$ and their subclasses $C' = \{c_{s,0} \mid s \in \{0,1\}^u\}$ used in the proof of Theorem 6 (see Table 1 for $u = 2$). ■

Note that this nonmonotonicity result holds with a fixed number of instances n. In fact, if n was not considered fixed then every concept class C' would have a superset C (via addition of instances) of lower subset teaching dimension. However, the same even holds for the teaching dimension itself which we yet consider monotonic since it is monotonic given fixed n. So whenever we speak of monotonicity we assume a fixed instance space X.

Of course such an instance space X might contain *redundant* instances the removal of which would not affect the subset teaching dimension and would retain a non-redundant subset of the set of all subset teaching sets. In the following subsection, where we discuss a possible intuition behind the nonmonotonicity of the STD, redundancy conditions on instances will actually play an important role and show the usefulness of the following technical discussion. However, it is not straightforward to impose a suitable redundancy condition characterizing when an instance can be removed.

We derive such a condition starting with a redundancy condition for the original variant of teaching sets. For that purpose we introduce the notion C^{-x} for the concept class resulting from C after removing the instance x from the instance space X. Here C is any concept class over X and $x \in X$ is any instance. For example, if $X = \{x_1, x_2, x_3\}$ and $C = \{\{x_1\}, \{x_1, x_2\}, \{x_2, x_3\}\}$ then

$$C^{-x_3} = \{\{x_1\}, \{x_1, x_2\}, \{x_2\}\}$$

considered over the instance space $\{x_1, x_2\}$.

Lemma 10 *Let C be a concept class over X and $x \in X$. If for all $c \in C$ and for all $S \in TS(c, C)$*

$$(x, c(x)) \in S \Rightarrow$$
$$\exists y \neq x \; [(S \setminus \{(x, c(x))\}) \cup \{(y, c(y))\} \in TS(c, C)] \, ,$$

then for all $c \in C$ and for all samples S

$$S \in TS(c, C^{-x}) \iff [S \in TS(c, C) \; \wedge \; (x, c(x)) \notin S] \, .$$

Proof. Note that $|C^{-x}| = |C|$. Let $c \in C$ be an arbitrary concept and let S be any sample over X.

First assume $S \in TS(c, C)$ and $(x, c(x)) \notin S$. Since obviously $TD(c, C^{-x}) \geq TD(c, C)$ we immediately obtain $S \in TS(c, C^{-x})$.

Second assume $S \in TS(c, C^{-x})$. By definition, we have $(x, c(x)) \notin S$. Hence it remains to prove that $S \in TS(c, C)$. If $S \notin TS(c, C)$ then there exists some $T \in TS(c, C)$ with $|T| < |S|$. We distinguish two cases.

Case 1. $(x, c(x)) \notin T$.

Then $T \in TS(c, C^{-x})$ in contradiction to the facts $S \in TS(c, C^{-x})$ and $|S| \neq |T|$.

Case 2. $(x, c(x)) \in T$.

Then by the premise of the lemma there exists a $y \neq x$ such that

$$A \overset{\text{def}}{=} (S \setminus \{(x, c(x))\}) \cup \{(y, c(y))\} \in TS(c, C)\,.$$

Since $(x, c(x)) \notin A$ we have $A \in TS(c, C^{-x})$ and $|A| = |T| \neq |S|$. This again contradicts $S \in TS(c, C^{-x})$.

Since both cases reveal a contradiction, we obtain $S \in TS(c, C)$. ■

For illustration see Table 3. In this example the instances x_4 and x_5 meet the redundancy condition. After eliminating x_5, x_4 still meets the condition and can be removed as well. The new representation of the concept class then involves only the instances x_1, x_2, x_3.

concept in C	TS
$\emptyset$	$\{(x_1,0),(x_3,0)\}$, $\{(x_1,0),(x_4,0)\}$, $\{(x_1,0),(x_5,0)\}$
$\{x_1\}$	$\{(x_1,1),(x_2,0)\}$, $\{(x_1,1),(x_5,0)\}$
$\{x_3,x_4,x_5\}$	$\{(x_2,0),(x_3,1)\}$, $\{(x_2,0),(x_4,1)\}$, $\{(x_2,0),(x_5,1)\}$
$\{x_2,x_3,x_4,x_5\}$	$\{(x_1,0),(x_2,1)\}$, $\{(x_2,1),(x_4,1)\}$
$\{x_1,x_2,x_5\}$	$\{(x_2,1),(x_3,0)\}$, $\{(x_3,0),(x_5,1)\}$
$\{x_1,x_2,x_3,x_5\}$	$\{(x_1,1),(x_3,1)\}$, $\{(x_3,1),(x_4,1)\}$

concept in $(C^{-x_5})^{-x_4}$	TS
$\emptyset$	$\{(x_1,0),(x_3,0)\}$
$\{x_1\}$	$\{(x_1,1),(x_2,0)\}$
$\{x_3\}$	$\{(x_2,0),(x_3,1)\}$
$\{x_2,x_3\}$	$\{(x_1,0),(x_2,1)\}$
$\{x_1,x_2\}$	$\{(x_2,1),(x_3,0)\}$
$\{x_1,x_2,x_3\}$	$\{(x_1,1),(x_3,1)\}$

Table 3: Teaching sets for a class C before and after elimination of two redundant instances.

Lemma 10 provides a condition on an instance x. If that instance is eliminated from the instance space then the resulting concept class C^{-x} does not only have the same teaching dimension as C but, even more, for each of its concepts c the teaching sets are exactly those that are teaching sets for c in C and do not contain an example involving the eliminated instance x. Note that even though several instances might meet that condition at the same time, only one at a time may be removed. For the remaining instances it has to be checked whether the condition still holds after elimination of the first redundant instance.

So one legitimate redundancy condition for instances—considering teaching sets—is the one given in the premise of Lemma 10.

This condition can be extended to a redundancy condition with respect to subset teaching sets.

Theorem 11 *Let C be a concept class over X and $x \in X$. If for all $k \in \mathbb{N}$, for all $c \in C$, and for all $S \in STS^k(c, C)$*

$$(x, c(x)) \in S \Rightarrow$$
$$\exists y \neq x\ [(S \setminus \{(x, c(x))\}) \cup \{(y, c(y))\} \in STS^k(c, C)]\,,$$

then for all $k \in \mathbb{N}$, for all $c \in C$, and for all samples S

$$S \in STS^k(c, C^{-x})$$
$$\Longleftrightarrow$$
$$[S \in STS^k(c, C)\ \wedge\ (x, c(x)) \notin S]\,.$$

Proof. Note that $|C^{-x}| = |C|$. We prove the theorem by induction on k.

For $k = 0$ this follows immediately from Lemma 10. So assume that the claim is proven for some k (induction hypothesis). It remains to show that it then also holds for $k + 1$.

For that purpose note that

$$\forall c \in C\ \forall A \in STS^k(c, C)\ \exists B \in STS^k(c, C^{-x})$$
$$[|A| = |B|\ \wedge\ A \setminus \{(x, c(x))\} \subseteq B]\ \ (*)$$

by combination of the induction hypothesis with the premise of the theorem.

Choose an arbitrary $c \in C$.

First assume $S \in STS^{k+1}(c, C)$ and $(x, c(x)) \notin S$. By the definition of subset teaching sets, there is an $S' \in STS^k(c, C)$ with

$$S \subseteq S'\,. \tag{1}$$

Using $(*)$ we can assume without loss of generality that

$$S' \in STS^k(c, C^{-x})\,. \tag{2}$$

Moreover, again by the definition of subset teaching sets, one obtains $S \not\subseteq S''$ for every $S'' \in STS^k(c', C)$ with $c' \neq c$. The induction hypothesis then implies

$$S \not\subseteq S'' \text{ for every } S'' \in STS^k(c', C^{-x}) \text{ with } c' \neq c\,. \tag{3}$$

Due to (1), (2), (3) we get either $S \in STS^{k+1}(c, C^{-x})$ or $|S| > STD^{k+1}(c, C^{-x})$. In the latter case there would be a set $T \in STS^{k+1}(c, C^{-x})$ with $|T| < |S|$. T is a subset of some set in $STS^k(c, C^{-x})$ and thus also of some set in $STS^k(c, C)$ by induction hypothesis. If T was contained in some $T' \in STS^k(c', C)$ for some $c' \neq c$ then we could again assume without loss of generality, using $(*)$ and $(x, c(x)) \notin T$, that T is contained in some set in $STS^k(c', C^{-x})$—in contradiction to $T \in STS^{k+1}(c, C^{-x})$. Therefore $T \in STS^{k+1}(c, C)$ and so $|T| = |S|$—a contradiction. This implies $S \in STS^{k+1}(c, C^{-x})$.

Second assume that $S \in STS^{k+1}(c, C^{-x})$. Obviously, $(x, c(x)) \notin S$, so that it remains to show $S \in STS^{k+1}(c, C)$.

Because of $S \in STS^{k+1}(c, C^{-x})$ there exists some set $S' \in STS^k(c, C^{-x})$ such that

$$S \subseteq S'\,. \tag{4}$$

The induction hypothesis implies

$$S' \in STS^k(c, C)\,. \tag{5}$$

Moreover, by the definition of subset teaching sets, one obtains $S \not\subseteq S''$ for every $S'' \in STS^k(c', C^{-x})$ with $c' \neq c$. If there was a set $S'' \in STS^k(c', C)$ with $c' \neq c$ and $S \subseteq S''$ then $(*)$ would imply that without loss of generality $S'' \in STS^k(c', C^{-x})$. So we have

$$S \not\subseteq S'' \text{ for every } S'' \in STS^k(c', C) \text{ with } c' \neq c\,. \tag{6}$$

Combining (4), (5), (6) we get either $S \in STS^{k+1}(c, C)$ or $|S| > STD^{k+1}(c, C)$. In the latter case there would be a set $T \in STS^{k+1}(c, C)$ with $|T| < |S|$. T is a subset of some set $T' \in STS^{k}(c, C)$. We can assume without loss of generality, using $(*)$, that $T' \in STS^{k}(c, C^{-x})$. If T was contained in some set in $STS^{k}(c', C^{-x})$ for some $c' \neq c$ then by induction hypothesis T would be contained in some set in $STS^{k}(c', C)$ for some $c' \neq c$. This is a contradiction to $T \in STS^{k+1}(c, C)$. So $T \in STS^{k+1}(c, C^{-x})$ and hence $|T| = |S|$—a contradiction. Thus $S \in STS^{k+1}(c, C)$. ■

4.2 The reason for nonmonotonicity

The idea about why the teaching dimension can decrease when a concept class increases is best illustrated by an example in which the addition of a single concept has this effect. In a simple such example, the instance space consists of three elements α, β, γ. First, consider the four distinct concepts that all contain γ, $c_{001} = \{\gamma\}$, $c_{011} = \{\beta, \gamma\}$, $c_{101} = \{\alpha, \gamma\}$, $c_{111} = \{\alpha, \beta, \gamma\}$. When these four concepts are the only ones in the class the teaching sets for them all are necessarily size two—elements α and β and their respective labels—because γ is a member of all of them, it cannot be part of any teaching set. If one more concept is added to the class the subset teaching sets all become size 1. Table 4 shows the computation when $c_{000} = \emptyset$ is added.

concept	STS^0	STS^1	STS^2	STS^3
$\emptyset$	$\{(\gamma,0)\}$	$\{(\gamma,0)\}$	$\{(\gamma,0)\}$	$\{(\gamma,0)\}$
$\{\gamma\}$	$\{(\alpha,0),(\beta,0),(\gamma,1)\}$	$\{(\gamma,1)\}$	$\{(\gamma,1)\}$	$\{(\gamma,1)\}$
$\{\beta,\gamma\}$	$\{(\alpha,0),(\beta,1)\}$	$\{(\alpha,0),(\beta,1)\}$	$\{(\alpha,0)\}$	$\{(\alpha,0)\}$
$\{\alpha,\gamma\}$	$\{(\alpha,1),(\beta,0)\}$	$\{(\alpha,1),(\beta,0)\}$	$\{(\beta,0)\}$	$\{(\beta,0)\}$
$\{\alpha,\beta,\gamma\}$	$\{(\alpha,1),(\beta,1)\}$	$\{(\alpha,1),(\beta,1)\}$	$\{(\alpha,1),(\beta,1)\}$	$\{(\beta,1)\}$

Table 4: Illustration of the nonmonotonicity of STD.

From a more general point of view, it is not obvious how to explain why a teaching dimension resulting from a cooperative model should be nonmonotonic.

First of all, this is a counter-intuitive observation when considering STD as a notion of complexity—intuitively any subclass of C should be at most as complex for teaching and learning as C.

However, there is in fact an intuitive explanation for the nonmonotonicity of the complexity in cooperative teaching and learning: when teaching $c \in C$, instead of providing examples that eliminate all concepts in $C \setminus \{c\}$ (as is the idea underlying minimal teaching sets) cooperative teachers would rather pick only those examples that distinguish c from its "most similar" concepts in C. Similarity here is measured by the number of instances on which two concepts agree (i.e., dissimilarity is given by the Hamming distance between the concepts, where a concept c is represented as a bit vector $(c(x_1), \ldots, c(x_n))$). This is reflected in the subset teaching sets in all illustrative examples considered above.

Considering a class $C = C^{u}_{0/1}$, one observes that a subset teaching set for a concept $c_{s,0}$ contains only the negative example $(x_{u+1+int(s)}, 0)$ distinguishing it from $c_{s,1}$ (its nearest neighbor in terms of Hamming distance). A learner will recognize this example as the one that separates only that one pair $(c_{s,0}, c_{s,1})$ of nearest neighbors. In contrast to that, if we consider only the subclass $C' = \{c_{s,0} \mid s \in \{0,1\}^u\}$, the nearest neighbors of each $c_{s,0}$ are different ones, and every single example separating one nearest neighbor pair also separates other nearest neighbor pairs. Thus no single example can be recognized by the learner as a separating example for one unique pair of concepts.

This intuitive idea of subset teaching sets being used for distinguishing a concept from its nearest neighbors has to be treated with care though. The reason is that the concept class may contain "redundant" instances, i.e., instances that could be removed from the instance space according to Theorem 11.

Such redundant instances might on the other hand affect Hamming distances and nearest neighbor relations. Only after their elimination the notion of nearest neighbors in terms of Hamming distance becomes well-defined. Consider for instance Table 3. In the concept class C over 5 instances the only nearest neighbor of $\emptyset$ is $\{x_1\}$ and an example distinguishing $\emptyset$ from $\{x_1\}$ would be $(x_1, 0)$. Moreover, no other concept is distinguished from its nearest neighbors by the instance x_1. According to the intuition explained here, this would suggest $\{(x_1, 0)\}$ being a subset teaching set for $\emptyset$ although the subset teaching sets here equal the teaching sets and are all of cardinality 2.

After instance elimination of x_4, x_5 there is only one subset teaching set for $\emptyset$, namely $\{(x_1, 0), (x_3, 0)\}$. This is still of cardinality 2 but note that now $\emptyset$ has two nearest neighbors, namely $\{x_1\}$ and $\{x_3\}$. The two examples in the subset teaching set are those that distinguish $\emptyset$ from its nearest neighbors. Note that either one of these two examples is not unique as an example used for distinguishing a concept from its nearest neighbors: $(x_1, 0)$ would be used by $\{x_2, x_3\}$ for distinguishing itself from its nearest neighbor $\{x_1, x_2, x_3\}$; $(x_3, 0)$ would be used by $\{x_1, x_2\}$ for distinguishing itself from its nearest neighbor $\{x_1, x_2, x_3\}$. So the subset teaching set for $\emptyset$ has to contain both examples.

This shows that in general a subclass of a class C can have a higher complexity than C if crucial nearest neighbors of some concepts are missing.

To summarize,

- nonmonotonicity has an intuitive reason and is not an indication for an ill-defined version of the teaching dimension,
- nonmonotonicity is in fact required if we want to capture the idea that the existence of specific concepts to distinguish a target concept from is beneficial for teaching and learning.

So, the STD captures certain intuitions about teaching and learning that monotonic dimensions *cannot* capture; at the same time monotonicity might in other respects itself be an intuitive property of teaching and learning which then the STD cannot capture.

In particular there are two underlying intuitive properties that seem to not be satisfiable by a single variant of the teaching dimension.

So in contrast one may wish to have a cooperative teaching and learning model going along with a monotonic complexity measure. It is not hard to show that BTD in fact is monotonic, see Theorem 12.

Theorem 12 *If C is a concept class and $C' \subseteq C$ a subclass of C, then $BTD(C') \leq BTD(C)$.*

Proof. Fix C and $C' \subseteq C$. We will prove by induction on k that

$$BTD^k(c, C') \leq BTD^k(c, C) \text{ for all } c \in C \qquad (7)$$

for all $k \in \mathbb{N}$.

$k = 0$: Property (7) holds because of $BTD^0(c, C') = TD(c, C') \leq TD(c, C) = BTD^0(c, C)$ for all $c \in C$.

Induction hypothesis: assume (7) holds for a fixed k.

$k \rightsquigarrow k+1$: First, observe that

$$\begin{aligned} & Cons_{size}(S, C', k) \\ = \quad & \{c \in Cons(S, C') \mid BTD^k(c, C') \geq |S|\} \\ \subseteq \quad & \{c \in Cons(S, C') \mid BTD^k(c, C) \geq |S|\} \text{ (ind. hyp.)} \\ \subseteq \quad & \{c \in Cons(S, C) \mid BTD^k(c, C) \geq |S|\} \\ = \quad & Cons_{size}(S, C, k) \end{aligned}$$

Second, for all $c \in C$ we obtain

$$\begin{aligned} & BTD^{k+1}(c, C') \\ & \quad = \quad \min\{|S| \mid Cons_{size}(S, C', k) = \{c\}\ \} \\ & \quad \leq \quad \min\{|S| \mid Cons_{size}(S, C, k) = \{c\}\ \} \\ & \quad \leq \quad BTD^{k+1}(c, C) \end{aligned}$$

This completes the proof. ■

So, on the one hand, we have the teaching framework based on the subset teaching dimension which results in a nonmonotonic dimension, and on the other hand we have a monotonic dimension in the BTD framework, which unfortunately does not always meet our idea of a best possible cooperative teaching and learning protocol. That raises the question whether nonmonotonicity is necessary to achieve certain positive results. In fact, the nonmonotonicity concerning the class $C^u_{0/1}$ is not counter-intuitive, but would a dimension that is monotonic also result in a worse sample complexity than the STD in general, such as, e.g., for the monomials?

In other words, is there a teaching/learning framework

- resulting in a monotonic variant of a teaching dimension and
- achieving similarly good results as the subset teaching dimension?

At this point of course it is difficult to define what "similarly good" means. However, we would like to have a constant dimension for the class of all monomials, as well as, e.g., a teaching set of size 1 for the empty concept in our often used concept class C_0.

We will now via several steps introduce at least a monotonic variant of the teaching dimension and show that for most of the examples studied above, it is as low as the subset teaching dimension. General comparisons will be made in Section 5, in particular in order to show that this new framework is uniformly at least as efficient as the BTD framework (or better), while sometimes being less efficient than the STD framework. This reflects to a certain extent that monotonicity constraints might affect sample efficiency.

4.3 The teaching plan model

We will first define the notion for our variant of teaching dimension and show its monotonicity. The nonmonotonicity of STD is caused by considering every STS^k-set for every concept when computing an STS^{k+1}-set for a single concept. Hence the idea in the following approach is to impose an order onto the concept class, in terms of the "teaching complexity" of the concepts. This is what the teaching dimension does as well, but our design principle is a recursive one. After selecting a concept which is "easy to teach" because of possessing a small minimal teaching set, we eliminate this concept from our concept class and consider only the remaining concepts. Again we determine the one with the lowest teaching dimension, now however measured with respect to the class of remaining concepts, and so on. The resulting notion of dimension is therefore called the *recursive teaching dimension*.

Definition 13 *Let C be a concept class, $|C| = N$. A teaching plan for C is a sequence $p = ((c_1, S_1), \ldots, (c_N, S_N)) \in (C \times 2^{X \times \{0,1\}})^N$ such that*

1. *$C = \{c_1, \ldots, c_N\}$.*
2. *$S_j \in TS(c_j, \{c_j, \ldots, c_N\})$ for $1 \leq j \leq N-1$.*
3. *$S_N = \{(x, 1-b) \mid (x, b) \in S_{N-1}\}$.*[5]

The order of p is given by $ord(p) = \max\{|S_j| \mid 1 \leq j \leq N\}$. The recursive teaching dimension of C is defined by $RTD(C) = \min\{ord(p) \mid p$ is a teaching plan for $C\}$.

The desired monotonicity property, see Proposition 14, follows immediately from the definition.

Proposition 14 *If C is a concept class and $C' \subseteq C$ is a subclass of C, then $RTD(C') \leq RTD(C)$.*

We can define a set of canonical teaching plans for any finite concept class C. As it will turn out, their order always equals $RTD(C)$.

Definition 15 *Let C be a concept class, $p = ((c_1, S_1), \ldots, (c_N, S_N))$ a teaching plan for C. p is called a canonical teaching plan for C, if for any $i, j \in \{1, \ldots, N\}$:*

$$i < j \Rightarrow TD(c_i, \{c_i, \ldots, c_N\}) \leq TD(c_j, \{c_i, \ldots, c_N\})\,.$$

Theorem 16 *Let C be a concept class and p a canonical teaching plan for C. Then $ord(p) = RTD(C)$.*

[5] Note that the cardinality of both S_{N-1} and S_N must be 1.

Proof. Let C and p as in the theorem be given, $p = ((c_1, S_1), \ldots, (c_N, S_N))$. Let $p' = ((c'_1, S'_1), \ldots, (c'_N, S'_N))$ be any teaching plan for C. It remains to prove that $ord(p) \le ord(p')$.

For that purpose choose the minimal $j \in \{1, \ldots, N\}$ such that $|S_j| = ord(p)$. By definition of a teaching plan, $TD(c_j, \{c_j, \ldots, c_N\}) = ord(p)$. Let $i \in \{1, \ldots, N\}$ be minimal such that $c'_i \in \{c_j, \ldots, c_N\}$. Let $k \in \{1, \ldots, N\}$ fulfill $c_k = c'_i$. By definition of a canonical teaching plan, $TD(c_k, \{c_j, \ldots, c_N\}) \ge TD(c_j, \{c_j, \ldots, c_N\}) = ord(p)$. This obviously yields $ord(p') \ge TD(c'_i, \{c'_i, \ldots, c'_N\}) \ge TD(c_k, \{c_j, \ldots, c_N\}) \ge ord(p)$. ■

To summarize briefly, the recursive teaching dimension is a monotonic complexity notion which in fact has got some of the properties we desired; e.g., it is easily verified that $RTD(C_0) = 1$ (by any teaching plan in which the empty concept occurs last) and that the RTD of the class of all monomials equals 2 (see below). Thus the RTD overcomes some of the weaknesses of BTD, while at the same time preserving monotonicity.

As it will turn out later, there are some interesting relations between BTD, STD, and RTD.

A property that might be relevant for establishing these relations is based on the following definition.

Definition 17 *Let C be a concept class, $|C| = N$. A TS-teaching plan for C is a sequence*

$$p = ((c_1, S_1^1), \ldots, (c_N, S_1^N, \ldots, S_N^N))$$

such that

1. $C = \{c_1, \ldots, c_N\}$.
2. $S_k^j \in TS(c_j, \{c_k, \ldots, c_N\})$ *for* $1 \le k \le j \le N$.
3. $S_k^j \subseteq S_{k-1}^j$ *for* $1 < k \le j \le N$.

The order of p is given by $ord(p) = \max\{|S_j^j| \mid 1 \le j \le N\}$. The recursive TS-teaching dimension of C is defined by $RTTD(C) = \min\{ord(p) \mid p$ is a TS-teaching plan for $C\}$.

TS-teaching plans differ from original teaching plans in that they require their sets being built up in stages as subsets of those in previous stages, starting from teaching sets.

However, as it turns out, concerning the RTD it suffices to consider this restricted form of teaching plans.

Lemma 18 *Let C be a concept class. Then $RTTD(C) = RTD(C)$. In particular, there is a TS-teaching plan $p = ((c_1, S_1^1), \ldots, (c_N, S_1^N, \ldots, S_N^N))$ for C such that $ord(p) = RTD(C)$ and $((c_1, S_1^1), \ldots, (c_N, S_N^N))$ is a canonical teaching plan for C.*

The proof is omitted.

4.4 Monomials revisited

In this subsection, we will pick up the two examples from Subsection 3.3 again, this time to determine the recursive teaching dimension.

Theorem 19 *Let $m \in \mathbb{N}$ and C the class of all boolean functions over m variables that can be represented by a monomial. Then $RTD(C) = 2$.*

Proof. Fix m and C. For all $i \in \{0, \ldots, m\}$ let C^i be the subclass of all $c \in C$ that can be represented by a non-contradictory monomial M that has got i variables. There is exactly one concept in C not belonging to any subclass C^i of C, namely the concept c^* representable by a contradictory monomial.

The proof is based on the following observation.

Observation. For any $i \in \{0, \ldots, m\}$ and any $c \in C^i$: $TD(c, C' \cup \{c^*\}) \le 2$, where $C' = \bigcup_{i \le j \le m} C^j$.

Now it is easily seen that $ord(p) \le 2$ for every teaching plan $p = ((c_1, S_1), \ldots, (c_N, S_N))$ for C that meets the following requirements:

(a) $c_1 \in C^0$ and $c_N = c^*$.

(b) For any $k, k' \in \{0, \ldots, N-1\}$: If $k < k'$, then $c_k \in C^i$ and $c_{k'} \in C^j$ for some $i, j \in \{0, \ldots, m\}$ with $i \le j$.

Since obviously $TD(c, C) \ge 2$ for all $c \in C$, we obtain $RTD(C) = 2$.

For illustration of the case $m = 2$ see Table 5. ■

		TS
λ	C^0	{(00,1),(11,1)}
v_1	C^1	{(10,1),(11,1)}
$\overline{v_1}$	C^1	{(00,1),(01,1)}
v_2	C^1	{(01,1),(11,1)}
$\overline{v_2}$	C^1	{(00,1),(10,1)}
$v_1 \wedge v_2$	C^2	{(11,1)}
$v_1 \wedge \overline{v_2}$	C^2	{(10,1)}
$\overline{v_1} \wedge v_2$	C^2	{(01,1)}
$\overline{v_1} \wedge \overline{v_2}$	C^2	{(00,1)}
$v_1 \wedge \overline{v_1}$		{(00,0)}

Table 5: Recursive teaching sets in a teaching plan of order 2 for the class of all monomials over $m = 2$ variables. λ denotes the empty monomial.

For the sake of completeness, note $RTD(C^m_{\vee DNF}) = 1$ where $C^m_{\vee DNF}$ is the class of boolean functions over m variables as defined in Subsection 3.3.

Theorem 20 $RTD(C^m_{\vee DNF}) = 1$ *for all* $m \in \mathbb{N}$.

Sketch of proof. This follows straightforwardly from the fact that $TD(c, C^m_{\vee DNF}) = 1$ for every concept c corresponding to a 2-term DNF of form $v_1 \vee M$.

For illustration see Table 2. ■

5 Comparison of teaching dimension notions

This section provides an analysis of the relationships between RTD, BTD, and STD.

Theorem 21 1. *If C is a concept class then $RTD(C) \le BTD(C)$.*

2. *There is a concept class C with $RTD(C) < BTD(C)$.*

Proof. Assertion (2) is witnessed by the concept class C_0 containing the empty concept and all singletons. Obviously, $RTD(C_0) = 1$ and $BTD(C_0) = 2$.

To prove Assertion (1), let C be a concept class with $RTD(C) = u$. By Theorem 16 there is a canonical teaching plan $p = ((c_1, S_1), \dots, (c_N, S_N))$ for C with $\mathit{ord}(p) = u$. Fix $j \leq \mathbb{N}$ minimal such that $|S_j| = u$ and define $C' = \{c_j \dots, c_N\}$. Obviously, $RTD(C') = u$. Moreover, using Theorem 12, $BTD(C') \leq BTD(C)$. Thus it suffices to prove $u \leq BTD(C')$.

To achieve this, we will prove by induction on k that $u \leq BTD^k(c, C')$ for all $k \in \mathbb{N}$ for all $c \in C'$.

$k = 0$: $BTD^0(c, C') = TD(c, C') \geq u$ for all $c \in C'$.

Induction hypothesis: assume $u \leq BTD^k(c, C')$ for all $c \in C'$ holds for a fixed k.

$k \rightsquigarrow k + 1$: Suppose by way of contradiction that there is a concept $c^* \in C'$ with $u > BTD^{k+1}(c^*, C')$. In particular, there exists a sample S^* such that $|S^*| < u$ and $\mathit{Cons}_{size}(S^*, C', k) = \{c^*\}$.

By induction hypothesis, the set $\mathit{Cons}_{size}(S^*, C', k)$ defined by $\{c \in \mathit{Cons}(S^*, C') \mid BTD^k(c, C') \geq |S^*|\}$ is equal to $\mathit{Cons}(S^*, C')$. Note that $TD(c, C') \geq u$ for all $c \in C'$ implies either $|\mathit{Cons}(S^*, C')| \geq 2$ or $\mathit{Cons}(S^*, C') = \emptyset$. We obtain a contradiction to $\mathit{Cons}_{size}(S^*, C', k) = \{c^*\}$.

This completes the proof. ■

Comparing the STD to the RTD turns out to be a bit more complex. We can show that the recursive teaching dimension can be arbitrarily larger than the subset teaching dimension; it can even be larger than the maximal STD computed over all subsets of the concept class.

Theorem 22 *1. For each $u \in \mathbb{N}$ there is a concept class C such that $STD(C) = 1$ and $RTD(C) = u$.*

2. There is a concept class C such that $\max\{STD(C') \mid C' \subseteq C\} < RTD(C)$.

Sketch of proof. Assertion (1) is witnessed by the classes $C^u_{0/1}$ defined in the proof of Theorem 6.

To verify Assertion (2), consider the concept class $C = \{c_1, \dots, c_6\}$ given by $c_1 = \emptyset$, $c_2 = \{x_1\}$, $c_3 = \{x_1, x_2\}$, $c_4 = \{x_2, x_3\}$, $c_5 = \{x_2, x_4\}$, $c_6 = \{x_2, x_3, x_4\}$. It is not hard to verify that $TD(c, C) = 2$ for all $c \in C$ and thus $\mathit{ord}(p) = 2$ for every teaching plan p for C. Therefore $RTD(C) = 2$. Moreover $STD(C') = 1$ for all $C' \subseteq C$ (the computation of $STD(C)$ is shown in Table 6; further details are omitted). ■

concept	STS^0	STS^1	STS^2
$\emptyset$	$\{(x_1, 0), (x_2, 0)\}$	$\{(x_1, 0)\}$	$\{(x_1, 0)\}$
$\{x_1\}$	$\{(x_1, 1), (x_2, 0)\}$	$\{(x_1, 1), (x_2, 0)\}$	$\{(x_1, 1)\}$ $\{(x_2, 0)\}$
$\{x_1, x_2\}$	$\{(x_1, 1), (x_2, 1)\}$	$\{(x_2, 1)\}$	$\{(x_2, 1)\}$
$\{x_2, x_3\}$	$\{(x_3, 1), (x_4, 0)\}$	$\{(x_4, 0)\}$	$\{(x_4, 0)\}$
$\{x_2, x_4\}$	$\{(x_3, 0), (x_4, 1)\}$	$\{(x_3, 0)\}$	$\{(x_3, 0)\}$
$\{x_2, x_3, x_4\}$	$\{(x_3, 1), (x_4, 1)\}$	$\{(x_3, 1), (x_4, 1)\}$	$\{(x_3, 1)\}$ $\{(x_4, 1)\}$

Table 6: Iterated subset teaching sets for the class $C = \{c_1, \dots, c_6\}$ given by $c_1 = \emptyset$, $c_2 = \{x_1\}$, $c_3 = \{x_1, x_2\}$, $c_4 = \{x_2, x_3\}$, $c_5 = \{x_2, x_4\}$, $c_6 = \{x_2, x_3, x_4\}$.

We conjecture moreover that $STD(C) \leq RTD(C)$ for all concept classes C, however, we cannot prove that at the time of writing. However, we can provide a general proof idea that solely relies on a lemma that we conjecture.

Lemma 23 (Conjecture) *Let C be a concept class and $p = ((c_1, S_1), \dots, (c_N, S_N))$ a teaching plan for C. Let j fulfill $\mathit{ord}(p) = |S_j|$ and $STD(c_j, C) \geq \mathit{ord}(p)$. Then there is a teaching plan*

$$p = ((c_1, S'_1), \dots, (c_N, S'_N))$$

for C and a sample $S \in STS(c_j, C)$ such that $S'_j \subseteq S$.

The proof of the following theorem, which helps to summarize the relations between our different variants of teaching dimensions, relies on this lemma—hence in fact the theorem is also a conjecture at the time of writing. Note that its correctness, together with Theorem 21 and Lemma 18, would imply

$$STD(C) \leq RTD(C) = RTTD(C) \leq BTD(C)$$

for all concept classes C. Here all inequalities are necessary since proven to not be equalities.

Theorem 24 (Based on conjecture Lemma 23) *Let C be a concept class. Then $STD(C) \leq RTD(C)$.*

Sketch of proof (relying on Lemma 23). Prove property (P_j) by induction for all $j \geq 1$.

> (P_j):
> If C is a concept class of at least j concepts and p is any teaching plan for C (not necessarily canonical), then $STD(c_j, C) \leq \mathit{ord}(p)$ where c_j is the j^{th} concept in the teaching plan p.

For $j = 1$ this is obvious, because

$$STD(c_1, C) \leq TD(c_1, C) \leq \mathit{ord}(p)\,.$$

The induction hypothesis is that (P_i) holds for all $i \leq j$, j fixed.

To prove (P_{j+1}), choose a concept class C and a teaching plan $p = ((c_1, S_1), \dots, (c_N, S_N))$ for C. Consider the $j + 1^{st}$ concept c_{j+1} in p.

Case 1. $|S_{j+1}| < \mathit{ord}(p)$.

If $|S_{j+1}| < \mathit{ord}(p)$, then we swap c_j and c_{j+1} and get a new teaching plan

$$\begin{aligned} p \;=\; & ((c_1, S_1), \dots, (c_{j-1}, S_{j-1}), \\ & (c_{j+1}, T), (c_j, T'), \dots, (c_n, S_N)) \end{aligned}$$

for C. Note that $|T'| \leq |S_j|$. Now c_{j+1} is in j^{th} position and its corresponding set T, due to the swap, fulfills $|T| \leq |S_{j+1}| + 1 \leq \mathit{ord}(p)$. By induction hypothesis we get $STD(c_{j+1}, C) \leq \mathit{ord}(p)$.

Case 2. $|S_{j+1}| = \mathit{ord}(p)$.

This is the more difficult case. Using Lemma 18 we can prove that S_{j+1} is a subset of a teaching set of c_{j+1} with respect to any of the classes $\{c_i, \dots, c_N\}$ where $i \leq j + 1$.

But in fact we would need Lemma 23 to tell us that S_{j+1} is a subset of a subset teaching set of c_{j+1} with respect to C.

Assume that $STD(c_{j+1}, C) > \mathit{ord}(p)$. This implies that S_{j+1} is a subset of some subset teaching set for c_{j+1} without being contained in any other subset teaching set for any other concept. Then S_{j+1} would itself be a subset teaching set for c_{j+1} in contradiction to its size being smaller than $STD(c_{j+1}, C)$.

To see why S_{j+1} couldn't be contained in any subset teaching set for any $c \neq c_{j+1}$, $c \in C$, note that c_{j+2}, ..., c_N are not consistent with S_{j+1} and the concepts c_1, ..., c_j by induction hypothesis have a too low subset teaching dimension in C. ■

6 Conclusions and open problems

We have introduced a new model of teaching and learning, based on what we call subset teaching sets. This model captures the idea of a teacher and a learner cooperating in order to learn concepts in finite classes from small samples.

This model avoids coding tricks and provides a generally applicable procedure for a uniform protocol of cooperative learning. It achieves results that are, for a specific concept class, such as the monomials, no less efficient than known algorithms that are designed especially for that one concept class (and perform inefficiently in terms of sample size on others).

The resulting subset teaching dimension turns out to be nonmonotonic—a fact that is illustrated and explained by the nature of the underlying definition.

In order to compare this subset teaching dimension to monotonic variants of teaching dimensions related to cooperation in learning, we introduced two equivalent notions of "recursive teaching dimensions", being monotonic by definition. They turn out to be very helpful in providing bounds for previous notions (they are significantly better than the original teaching dimension and variants thereof). However, even though they behave so well, the nonmonotonic subset teaching dimension in general seems to be better.

Examples have shown that even the recursive teaching dimensions cannot always compete with the subset teaching dimension, though our conjecture that the recursive teaching dimension can never be lower than the subset teaching dimension is still open.

We plan to close this gap in our proof, to find characterizations for these teaching dimensions, and to provide evidence to another conjecture, namely that, for reasonable definitions of the term "coding trick", there is no teaching and learning model that avoids coding tricks and is better than the model based on the subset teaching dimension.

Acknowledgments

We gratefully acknowledge the support of Laura Zilles and Michael Geilke who developed and provided a software tool for computing subset teaching sets and teaching plans.

Many thanks are due to the anonymous referees for their helpful comments.

This work was partly funded by the Alberta Ingenuity Centre for Machine Learning.

References

[ABCS92] M. Anthony, G. Brightwell, D.A. Cohen, and J. Shawe-Taylor. On exact specification by examples. In *Proc. of 5th Annual Workshop on Computational Learning Theory (COLT'92)*, pages 311–318. ACM, New York, 1992.

[AK97] D. Angluin and M. Krikis. Teachers, learners and black boxes. In *Proc. of the 10th Annual Conference on Computational Learning Theory (COLT'97)*, pages 285–297. ACM, New York, 1997.

[Bal08] F. Balbach. Measuring teachability using variants of the teaching dimension. *Theoret. Comput. Sci.*, 397(1-3):94–113, 2008.

[BE98] S. Ben-David and N. Eiron. Self-directed learning and its relation to the VC-dimension and to teacher-directed learning. *Machine Learning*, 33(1):87–104, 1998.

[FKW93] R. Freivalds, E.B. Kinber, and R. Wiehagen. On the power of inductive inference from good examples. *Theoret. Comput. Sci.*, 110(1):131–144, 1993.

[GK95] S.A. Goldman and M.J. Kearns. On the complexity of teaching. *J. Comput. Syst. Sci.*, 50(1):20–31, 1995.

[GM96] S.A. Goldman and H.D. Mathias. Teaching a smarter learner. *J. Comput. Syst. Sci.*, 52(2):255–267, 1996.

[Han07] S. Hanneke. Teaching dimension and the complexity of active learning. In *Proc. of the 20th Annual Conference on Learning Theory (COLT 2007)*, pages 66–81. LNCS 4539, Springer, Berlin, 2007.

[Heg95] T. Hegedüs. Generalized teaching dimensions and the query complexity of learning. In *Proc. of the 8th Annual Conference on Computational Learning Theory (COLT'95)*, pages 108–117. ACM, New York, 1995.

[JT92] J. Jackson and A. Tomkins. A computational model of teaching. In *Proc. of 5th Annual Workshop on Computational Learning Theory (COLT'92)*, pages 319–326. ACM, New York, 1992.

[LNW98] S. Lange, J. Nessel, and R. Wiehagen. Learning recursive languages from good examples. *Ann. Math. Artif. Intell.*, 23(1-2):27–52, 1998.

[Mat97] H.D. Mathias. A model of interactive teaching. *J. Comput. Syst. Sci.*, 54(3):487–501, 1997.

[OS02] Matthias Ott and Frank Stephan. Avoiding coding tricks by hyperrobust learning. *Theoret. Comput. Sci.*, 284(1):161–180, 2002.

[RY95] R.L. Rivest and Y.L. Yin. Being taught can be faster than asking questions. In *Proc. of the 8th Annual Conference on Computational Learning Theory (COLT'95)*, pages 144–151. ACM, New York, 1995.

[SM91] A. Shinohara and S. Miyano. Teachability in computational learning. *New Generation Comput.*, 8(4):337–348, 1991.

[Val84] L.G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.