JOURNAL OF COMPUTER AND SYSTEM SCIENCES 50, 20–31 (1995)

# On the Complexity of Teaching*

SALLY A. GOLDMAN†

*Department of Computer Science, Washington University, St. Louis, Missouri 63130*

AND

MICHAEL J. KEARNS‡

*AT&T Bell Laboratories, Murray Hill, New Jersey 07974*

Received August 11, 1992; revised February 1, 1994

While most theoretical work in machine learning has focused on the complexity of learning, recently there has been increasing interest in formally studying the complexity of *teaching*. In this paper we study the complexity of teaching by considering a variant of the on-line learning model in which a helpful teacher selects the instances. We measure the complexity of teaching a concept from a given concept class by a combinatorial measure we call the *teaching dimension*. Informally, the teaching dimension of a concept class is the minimum number of instances a teacher must reveal to uniquely identify *any* target concept chosen from the class. 

## 1. INTRODUCTION

While most theoretical work in machine learning has focused on the complexity of learning, recently there has been some work on the complexity of *teaching* [9, 10, 15, 18, 19]. Our teaching model is a variation of the standard on-line learning model in which the learning session is divided into a set of trials where in each trial the learner is asked to make a prediction for some unknown instance from the domain. After the prediction is made, the learner is told whether the prediction is correct and is then able to use polynomial time before proceeding to the next trial. In this paper we study the complexity of teaching by considering a variant of this model in which a helpful teacher selects the instances (this is the teacher-directed learning model of Goldman, Rivest, and Schapire [9]). We measure the complexity of teaching a concept from a given concept class by a combinatorial measure we call the *teaching dimension*. Informally, the teaching dimension of a concept class is the minimum number of instances a teacher must reveal to uniquely identify *any* target concept chosen from the class.

We show that this new dimension measure is fundamentally different from the Vapnik–Chervonenkis dimension [3, 13, 22] and the dimension measure of Natarajan [15]. While we show that there is a concept class $C$ for which the teaching dimension is $|C| - 1$, we prove that in such cases there is one "hard-to-teach" concept that when removed yields a concept class that has a teaching dimension of one. More generally, when the teaching dimension of $C$ is $|C| - k$ then by removing a single concept one obtains a class with a teaching dimension of $k$.

We then explore the computational problem of finding an optimal teaching sequence for a given target concept. We show that given a concept class $C$ and a target concept $c \in C$ the problem of finding an optimal teaching sequence for $c$ is equivalent to finding a minimum set covering. While the problem of finding a minimum set covering is fairly well understood, there is an important distinction between the set cover problem and the problem of computing optimal teaching sequences. In the set covering problem, no assumptions are made about the structure of the sets, whereas for our problem we are assuming that there is a short (polynomial-sized) representation of the objects contained in each set. We also describe a straightforward relation between the teaching dimension and the number of membership queries needed for exact identification.

We then study the teaching dimension for the following concept classes. For the Boolean classes considered (i.e., all but the orthogonal rectangles) each concept is constructed from the set $V_n = \{v_1, v_2, ..., v_n\}$ of $n$ Boolean variables.

* A preliminary version of this paper appeared in the "Proceedings Fourth Annual Workshop on Computational Learning Theory, August 1991," pp. 303–314. Most of this research was carried out while both authors were at MIT Laboratory for Computer Science with support provided by ARO Grant DAAL03-86-K-0171, DARPA Contract N00014-89-J-1988, NSF Grant CCR-88914428, and a grant from the Siemens Corporation. S. Goldman was also supported in part by a G.E. Foundation Junior Faculty Grant and NSF Grant CCR-9110108.

† E-mail: sg@cs.wustl.edu.

‡ E-mail: mkearns@research.att.com.

0022-0000/95 $6.00

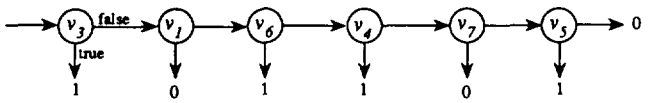


FIG. 1. A monotone decision list : $\langle (v_3, 1), (v_1, 0), (v_6, 1), (v_4, 1), (v_7, 0), (v_5, 1) \rangle$.

- A *monotone monomial* is a conjuction of a subset of the variables in $V_n$, where each variable selected must appear unnegated.
- A *(arbitrary) monomial* is the conjuction of a subset of the variables in $V_n$ where each variable selected may appear negated or unnegated.
- A *monotone decision list* [16] is a list $L = \langle (y_1, b_1), ..., (y_l, b_l) \rangle$, where each $y_i \in V_n$ and each $b_i \in \{0, 1\}$. For an $x \in \{0, 1\}^n$ we define $L(x)$, the output from $L$ on input $x$, as follows: $L(x) = b_j$, where $1 \leqslant j \leqslant l$ is the least value such that $y_j$ is 1 in $x$; $L(x) = 0$ if there is no such $j$. Let $b(v_i)$ denote the bit associated with $v_i$. One may think of a decision list as an extended "**if–then–elseif**–...–, **else**" rule. See Fig. 1 for an example monotone decision list on $v_1, ..., v_7$.
- Let $[i, j]$ denote the set $\{m \in \mathbf{N} | i \leqslant m \leqslant j\}$. An *orthogonal rectangle in* $\{0, 1, ..., n-1\}^d$ is defined by

$$\left\{ \bigtimes_{k=1}^{d} [i_k, j_k] \;\middle|\; 1 \leqslant i_k \leqslant j_k \leqslant n \right\},$$

where $i_k$ and $j_k$ are the minimum and maximum positive values in dimension $k$. We note that this class is the same as the class $\mathrm{BOX}_n^d$ of Maass and Turàn [14].

- A *monotone k-term DNF formula* is a disjunction of at most $k$ monotone monomials over the variable set $V_n$, where $k$ is any constant.
- A *k-term μ-DNF formula* is a disjunction of at most $k$ arbitrary monomials over the variable set $V_n$, where $k$ is any constant. Furthermore, the $\mu$ condition requires that each variable in $V_n$ can appear in only one of the monomials in the disjunction.

We start by giving tight bounds on the teaching dimension for the class of monomials and then extend this result to the more complicated concept classes of monotone $k$-term DNF formulas and $k$-term $\mu$-DNF formulas. We also compute bounds on the teaching dimension for the class of monotone decision lists and orthogonal rectangles in $\{0, 1, ..., n-1\}^d$. In computing the teaching dimension for these classes we also provide some general results that will aid in computing the teaching dimension for other concept classes. Finally, we prove that for concept classes closed under exclusive-or the teaching dimension is at most logarithmic in the size of the concept class.

## 2. THE TEACHING DIMENSION

In this section we formally define the teaching dimension. A *concept* $c$ is a Boolean function over some domain of *instances*. Let $X$ denote the set of instances, and let $C \subseteq 2^X$ be a concept class over $X$. Often $X$ and $C$ are parameterized according to some complexity measure $n$. For a concept $c \in C$ and instance $x \in X$, $c(x)$ denotes the classification of $c$ on instance $x$.

The basic goal of the teacher is to teach the learner to perfectly predict whether any given instance is a positive or negative instance of the *target concept*. Thus, the learner must achieve *exact identification* of the target concept. Of course, the teacher would like to achieve this goal with the fewest number of examples possible. In order to preclude unnatural "collusion" between the teacher and the learner (such as agreed-upon coding schemes to communicate the name of the target via the instances selected without regard for the labels), which could trivialize the teaching dimension measure, we simply ask that the teaching sequence chosen induces *any consistent algorithm* to exactly identify the target, where we define a learning algorithm to be *consistent* if its hypotheses are always consistent with all previously seen examples.

We formalize this as follows. Let an *instance sequence* denote a sequence of *unlabeled* instances, and an *example sequence* denote a sequence of *labeled instances*. For concept class $C$ and target concept $c \in C$, we say $T$ is a *teaching sequence for* $c$ *(in* $C$*)* if $T$ is an example sequence that uniquely specifies $c$ in $C$—that is, $c$ is the only concept in $C$ consistent with $T$. Let $T(c)$ be the set of all teaching sequences for $c$. We define the *teaching dimension* $\mathrm{TD}(C)$ of a concept class $C$ as

$$\mathrm{TD}(C) = \max_{c \in C} \left( \min_{\tau \in T(c)} |\tau| \right).$$

In other words, the teaching dimension of a concept class is the minimum number of examples a teacher must reveal to uniquely identify *any* concept in the class. Note, however, that the optimal teaching sequence is allowed to vary with the target concept, as opposed to the *universal identification sequences* of Goldman, Kearns, and Schapire [8].

Finally, we define the Vapnik–Chervonenkis dimension [22]. Let $X$ be the instance space, and let $C$ be a concept class over $X$. A finite set $Y \subseteq X$ is *shattered* by $C$ if $\{c \cap Y \mid c \in C\} = 2^Y$. In other words, $Y \subseteq X$ is shattered by $C$ if for each subset $Y' \subseteq Y$, there is a concept $c \in C$ which contains all of $Y'$, but none of the instances in $Y - Y'$. The *Vapnik–Chervonenkis dimension* of $C$, denoted $\mathrm{VCD}(C)$, is defined to be the smallest $d$ for which no set of $d+1$ points is shattered by $C$. Blumer *et al.* [3] have shown that this combinatorial measure of a concept class characterizes the number of examples required for learning any concept in the

class under the distribution-free or PAC model of Valiant [21].

In the next section we briefly discuss some related work. In Section 4 we compare the teaching dimension to both the Vapnik–Chervonenkis dimension and Natarajan's dimension measure. Furthermore, we show that when the teaching dimension of a class $C$ is $|C|-k$ then by removing a single concept one obtains a class with a teaching dimension of $k$. In Section 5 we first explore the problem of computing an optimal teaching sequence for a given target concept. Next we consider the problem of computing the teaching dimension for various concept classes. In particular, we give both upper and lower bounds on the teaching dimension for several well-studied concept classes. We also describe some general techniques that will aid in computing the teaching dimension for other concept classes. Finally, we prove that for concept classes closed under exclusive-or the teaching dimension is at most logarithmic in the size of the concept class. In Section 6 we summarize our results and discuss some open problems.

## 3. RELATED WORK

Our work is directly motivated by the teacher-directed learning model of Goldman *et al.* [9]. The number of mistakes made by the learner in an on-line learning model depends on the sequence of instances presented. Goldman, *et al.* [9] extended the mistake bound model to include several methods for the selection of instances. A *query sequence* is a permutation $\pi = \langle x_1, x_2, \ldots, x_{|X_n|} \rangle$ of $X_n$, where $x_t$ is the instance presented to the learner at the $t$th trial. They refer to the agent selecting the query sequence as the *director*. One of the directors that they considered was a helpful teacher who knows the target concept and wants to minimize the learner's mistakes. To select $x_t$, the teacher uses knowledge of the target concept, $x_1, \ldots, x_{t-1}$, and the learner's predictions on $x_1, \ldots, x_{t-1}$. To avoid allowing the learner and teacher to have a coordinated strategy, in this scenario they define the mistake bound as the maximum number of mistakes (i.e., incorrect predictions) made by the learner where the maximum is taken over all consistent learning algorithms. It is easily shown that the teaching dimension of a concept class is equal to the optimal mistake bound under teacher-directed learning. Clearly the teaching dimension is a lower bound for the optimal mistake bound—unless the target concept has been uniquely identified a mistake could be made on the next prediction. Furthermore, the teaching dimension is an upper bound for the optimal mistake bound since no mistakes could be made after the target concept has been exactly identified. Thus Goldman *et al.* [9] have computed exact bounds for the teaching dimension for binary relations and total orders.

Independently, Shinohara and Miyano [20] introduced a notion of teachability that shares the same basic framework as our work. In particular, they consider a notion of teachability in which a concept class is *teachable by examples* if there exists a polynomial size sample under which all consistent learners will exactly identify the target concept. So a concept class is teachable by examples if the teaching dimension is polynomially bounded. The primary focus of their work is to establish a relationship between learnability and teachability.[1]

Jackson and Tomkins [12] considered a variation of our teaching model in which they can study teacher/learner pairs in which the teacher chooses examples tailored to a particular learner. To avoid encoding schemes between the teacher and learner, they consider the interaction between the teacher and learner as a modified prover–verifier session [11] in which the learner and teacher can collude, but no adversarial teacher can cause the learner to output an hypothesis that is inconsistent with the sample. While it appears that the teacher's knowledge of the learner in this model would be useful, they have shown that under their model the teacher must still produce a teaching sequence that eliminates all but one concept from the concept class. They also introduced the notion of a small amount of *trusted information* that the teacher can provide the learner. This trusted information can be used by the teacher to provide the learner with the size complexity of the target (e.g., the size parameter $k$ of a $k$-term DNF formula). In our results presented here, the learner and teacher sometimes share such information, but we do not provide a formal characterization of what kind of information can be shared in this manner. More recently, Goldman and Mathias [7] have presented a formal model of teaching in which the teacher is tailored to a particular learner, yet the teaching protocol is designed so that no collusion is possible.

The work of Romanik and Smith [18, 17] on testing geometric objects shares similarities with our work. They propose a testing problem that involves specifying for a given target concept a set of test points that can be used to determine if a tested object is equivalent to the target. However, their primary concern is to determine for which concept classes there exists a finite set of instances such that any concept in the class which is consistent on the test set is "close" to the target in a probabilistic sense. Their work suggests an interesting variation on our teaching model in which the teaching sequence is only required to eliminate those hypotheses that are not "close" to the target. Such a model would use a PAC-style [21] success criterion for the learner versus the exact-identification-style [1] criterion that we have used here.

In other related work, Anthony, Brightwell, Cohen, and Shawe-Taylor [2] define the *specification number* of a

[1] A few results presented here were independently discovered by Shinohara and Miyano. We shall note such areas of independent discovery.

concept $c \in C$ to be the cardinality of the smallest sample for which only $c$ is consistent with the sample. Thus the specification number is just the length of the optimal teaching sequence for $c$. Their paper studies several aspects of the specification number with an emphasis on determining the specification numbers of hypotheses in the set of linearly separable Boolean functions.

Salzberg, Delcher, Health, and Kasif [19] have also considered a model of learning with a helpful teacher. Their model requires the teacher to present the shortest example sequence so that any learner using a particular algorithm (namely, the nearest-neightbor algorithm) learns the target concept. The fundamental philosophical difference between their work and ours is that we do not assume that the teacher knows the algorithm used by the learner.

The work of Goldman, *et al.* [8], in which they described a technique for exactly identifying certain classes of read-once formulas from random examples, is also related to our work. They defined a *universal identification sequence* for a concept class $C$ as a *single* instance sequence that distinguishes *every* concept $c \in C$. Furthermore, they proved the existence of polynomial-length universal identification sequences for the concept classes of logarithmic-depth read-once majority formulas and logarithmic-depth read-once positive NOR formulas. Observe that a universal identification sequence is always a teaching sequence (modulo different labelings) for any concept in the class, and thus their work provides an upper bound on the teaching dimension for the concept classes they considered.

Finally, Natarajan [15] defines a dimension measure for concept classes of Boolean functions that measures the complexity of a concept class by the length of the shortest example sequence for which the target concept is the unique most specific concept consistent with the sample. Thus, like the work of Salzberg *et al.* [19], Natarajan places more stringent requirements on the learner. We discuss the relation between Natarajan's dimension measure and the teaching dimension in Section 4.2.

## 4. COMPARISON TO OTHER DIMENSION MEASURES

In this section we compare the teaching dimension to two other dimension measures that have been used to describe the complexity of concept classes.

### 4.1. *Vapnik–Chervonenkis Dimension*

We now show that the teaching dimension is fundamentally different from the well-studied Vapnik–Chervonenkis dimension. Blumer *et al.* [3] have shown that this combinatorial measure of a concept class exactly characterizes (modulo dependencies on the accuracy and confidence parameters) the number of examples required for learning under the distribution-free or PAC model of Valiant [21].

| concept | instances |
|---|---|
| $c_*$ | + + + + ⋯ + + + |
| $c_1$ | − + + + ⋯ + + + |
| $c_2$ | + − + + ⋯ + + + |
| $c_3$ | + + − + ⋯ + + + |
| ⋮ | |
| $c_{\|X\|-1}$ | + + + + ⋯ + − + |
| $c_{\|X\|}$ | + + + + ⋯ + + − |

FIG. 2. A concept class $C$ for which $\mathrm{TD}(C) = |C| - 1$.

We now compare the teaching dimension to the VC dimension. We begin by showing that neither of these dimension measures dominates the other: in some cases, $\mathrm{TD}(C) \gg \mathrm{VCD}(C)$, and in other cases, $\mathrm{TD}(C) \ll \mathrm{VCD}(C)$.

LEMMA 1.[2] *There is a concept class $C$ for which* $\mathrm{TD}(C) = |C| - 1$ *and* $\mathrm{VCD}(C) = 1$.

*Proof.* Consider the concept class $C$ in which a distinguished concept $c_*$ classifies all instances as positive. In addition, for each $x \in X$ there is a concept $c_x \in C$ that classifies all instances but $x$ as positive (Fig. 2). No concept in $C$ has two negative instances, so clearly no set of two points is shattered. Since any singleton set is shattered, $\mathrm{VDC}(C) = 1$. Finally, since each instance distinguishes only one of the concepts $c_1, \ldots, c_{|X|}$ from $c_*$, $\mathrm{TD}(C) = |C| - 1$. ∎

Although the concept class $C$ from Lemma 1 appears to be quite simple, since the teacher must succeed for any consistent learner, it is hard to teach. However, in Theorem 4 we shall see that with slight modification, $C$ is in fact easy to teach. Also observe that one can obtain the result of Lemma 1 using the class of singletons and the empty set.

Thus the VC dimension can be arbitrarily smaller than the teaching dimension. Furthermore, we note that the concept class used in the proof of Lemma 1 has the largest possible teaching dimension.

OBSERVATION 2. *For any concept class* $C$, $\mathrm{TD}(C) \leqslant |C| - 1$.

*Proof.* Each concept in $C$ must differ from all other concepts for at least one instance. Thus for each concept that is not the target concept there must be an example that it and the target concept classify differently. ∎

We now show that the teaching dimension may be smaller than the VC dimension.

LEMMA 3. *There is a concept class $C$ for which* $\mathrm{TD}(C) < \mathrm{VCD}(C)$.

[2] Independently, Shinohara and Miyano [20] give a construction of a class of concepts that is PAC-learnable but not teachable by examples.

| | $x_0$ | $x_1$ | $x_2$ | $\cdots$ | $x_{n-2}$ | $x_{n-1}$ | $x_n$ | $x_{n+1}$ | $\cdots$ | $x_{n+\lg n-1}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $c_0$ | + | − | − | $\cdots$ | − | − | − | − | $\cdots$ | − |
| $c_1$ | − | + | − | $\cdots$ | − | − | + | − | $\cdots$ | − |
| $c_2$ | − | − | + | $\cdots$ | − | − | − | + | $\cdots$ | − |
| ⋮ | | | | | | | | | | |
| $c_{n-2}$ | − | − | − | $\cdots$ | + | − | + | + | $\cdots$ | − |
| $c_{n-1}$ | − | − | − | $\cdots$ | − | + | + | + | $\cdots$ | + |

FIG. 3. A concept class $C_n$ for which $\mathrm{VCD}(C_n) > \mathrm{TD}(C_n)$.

*Proof.* Let $C_n = \{c_0, c_2, ..., c_{n-1}\}$ and $X_n = \{x_0, x_1, ..., x_{x+\lg n-1}\}$ where $n = 2^k$ for some constant $k$. The concept $c_i$ classifies instance $x_i$ as positive and the rest of $x_0, ..., x_{n-1}$ as negative. The remaining $\lg n$ instances for $c_i$ are classified according to the binary representation of $i$ (Fig. 3). Clearly $\{x_n, ..., x_{n+\lg n-1}\}$ is a shattered set and thus $\mathrm{VCD}(C_n) = \lg n = \lg |C_n|$. However, $\mathrm{TD}(C_n) = 1$, since instance $x_i$ uniquely defines concept $c_i$. ∎

For finite $C$, $\mathrm{VCD}(C) \leqslant \lg |C|$ and $\mathrm{TD}(C) \geqslant 1$, and thus $\mathrm{VCD}(C) \leqslant \lg |C| \cdot \mathrm{TD}(C)$. So the concept class of Lemma 3 provides the maximum factor by which the VC dimension can exceed the teaching dimension for a finite concept class. Combined with Lemma 1 we have a complete characterization of how the teaching dimension relates to the VC dimension.

We now uncover another key difference between the VC dimension and the teaching dimension: the potential effect of removing a single concept from the concept class. Let $C$ be a concept class with $\mathrm{VCD}(C) = d$, and let $C' = C - \{f\}$ for some $f \in C$. Regardless of the choice of $f$, clearly $\mathrm{VCD}(C') \geqslant d - 1$. In contrast to this we have the following result.

THEOREM 4. *Let $C$ be a concept class for which $\mathrm{TD}(C) \geqslant |C| - k$. Then there exists an $f \in C$ such that for $C' = C - \{f\}$, $\mathrm{TD}(C') \leqslant k$.*

*Proof.* Let $f$ be a concept from $C$ that requires a teaching sequence of length $\mathrm{TD}(C)$. Let $T$ be an *optimal* teaching sequence for $f$ (i.e., $|T| = \mathrm{TD}(C)$). We let $C' = C - \{f\}$, and prove that $\mathrm{TD}(C') \leqslant k$. To achieve this goal we must show that for each concept $c \in C'$ there exists a teaching sequence for $c$ of length at most $k$. Without loss of generality, let $f' \in C'$ be the target concept that requires the longest teaching sequence. We now prove that there is a teaching sequences for $f'$ of length at most $k$.

The intuition for the remainder of this proof is as follows. Since $T$ is an optimal teaching sequence, for each instance $x_i \in T$ there must be a concept $f_i \in C'$ such that $f(x_i) \neq f_i(x_i)$. However, it may be that some $x_i \in T$ distinguishes $f$ from *many* concepts in $C'$. We say that all concepts in $C' - \{f_i\}$ that $x_i$ distinguishes from $f$ are eliminated as the possible target "*for free.*" Intuitively, since the teaching dimension is large, few concepts can be eliminated "for free." We then use this observation to show that $\mathrm{TD}(C')$ is small.

We now formalize this intuition. Let $x_* \in T$ be an instance that $f$ and $f'$ classify differently. We now define a set $F$ of concepts that are distinguished from $f$ "for free." For ease of exposition we shall also create a set $S$ that will contain $C - F - \{f\}$. We build the sets $F$ and $S$ as follows. Initially let $S = \{f'\}$ and let $F = \{c \in C' - \{f'\} | c(x_*) = F'(x_*)\}$. That is $F$ initially contains the concepts that are also distinguished from $f$ by $x_*$. Then for each $x \in T - \{x_*\}$ of all concepts in $C' - F - S$ that disagree with $f$ on $x$ place one of these concepts (choose arbitrarily) in $S$ and place the rest in $F$. Since $T$ is a teaching sequence at the end of this process $|S| = \mathrm{TD}(C)$ and $C = F \cup S \cup \{f\}$. Furthermore, since $\mathrm{TD}(C) \geqslant |C| - k$ we get that $|F| = |C| - \mathrm{TD}(C) - 1 \leqslant k - 1$. That is, at most $k - 1$ concepts are eliminated "for free."

We now generate a teaching sequence for $f'$ of length at most $k$. By the definition of $F$ and $S$ any concept in $C' - \{f'\}$ that classifies $x_*$ as $f'$ classifies it must be in $F$. That is, *all* concepts in $S$ are distinguished from $f'$ by $x_*$. Furthermore, since $|F| \leqslant k - 1$ at most $k - 1$ additional instances are needed to distinguish $f'$ from the concepts in $F$. Finally since $C' = F \cup S$ it follows that there is a teaching sequence for $f'$ of length at most $1 + (k - 1) = k$. This completes the proof of the theorem. ∎

So when $k = 1$, Theorem 4 implies that for any concept class $C$ which $\mathrm{TD}(C) = |C| - 1$, there exists a concept whose removal causes the teaching dimension of the remaining class to be reduced to 1. We briefly mention an interesting consequence of this result. Although it appears that for a concept class $C$ with $\mathrm{TD}(C) = |C| - 1$ there is little the teacher can do, this result suggest the following strategy to teach target concept $f \in C$: first teach some concept $f'$ in $C - \{f\}$ and then (if possible) list the instances that $f$ and $f'$ classify differently.

While we have shown that the teaching dimension and the VC dimension are fundamentally different, there are some relations between them. We now derive an upper bound for the teaching dimension that is based on the VC dimension.

THEOREM 5. *For any concept class $C$,*

$$\mathrm{TD}(C) \leqslant \mathrm{VCD}(C) + |C| - 2^{\mathrm{VCD}(C)}.$$

*Proof.* Let $x_1, ..., x_d$ be a shattered set of size $d$ for $d = \mathrm{VCD}(C)$. By the definition of a shattered set, in an example sequence consisting of these $d$ instances, all but one of a set of $2^d$ concepts are eliminated. Thus after placing $x_1, ..., x_d$ in the teaching sequence, there are at most $|C| - 2^d + 1$ concepts remaining. Finally, we use the naive

algorithm of Observation 2 to eliminate the remaining functions using at most $|C| - 2^d$ additional examples. ▮

### 4.2. *Natarajan's Dimension Measure*

In this section we compare the teaching dimension to the following dimension measure defined by Natarajan [15] for concept classes of Boolean functions:

$$\mathrm{ND}(C) = \min_d \left\{ \begin{array}{l} \text{For all } c \in C, \text{ there exists a} \\ \text{labeled sample } S_c \text{ of cardinality} \\ d \text{ such that } c \text{ is consistent with} \\ S_c \text{ and for all } c' \in C \text{ that are} \\ \text{consistent w*ith } S_c,\ c \subseteq c' \end{array} \right\}.$$

Natarajan shows that this dimension measure characterizes the class of Boolean functions that are learnable with one-sided error from polynomially many examples. He also gives the following result relating this dimension measure to the VC dimension.

THEOREM (Natarajan), *For concept classes $C_n$ chosen from the domain of Boolean functions over $n$ variables* $\mathrm{VCD}(C_n) \leq n \cdot \mathrm{ND}(C_n)$.

Note that the definition of Natarajan's dimension measure is similar to the teaching dimension, except that if there is more than one concept consistent with the given set of examples, the learner is required to pick the most specific one. Using this correspondence we obtain the following result.

LEMMA 6. *For concept classes $C_n$ chosen from the domain of Boolean functions over $n$ variables,* $\mathrm{TD}(C_n) \geq \mathrm{ND}(C_n)$.

*Proof.* Suppose there exists a concept class for which $\mathrm{TD}(C_n) < \mathrm{ND}(C_n)$. By the definition of the teaching dimension, there exists a sequence of $\mathrm{TD}(C_n)$ examples that uniquely specifies the target concept from all concepts in $C_n$. Let this sequence of $\mathrm{TD}(C_n)$ examples be the labeled sample $S_c$ used in the definition of $\mathrm{ND}(C_n)$. Since $S_c$ uniquely specifies that $c \in C_n$ there are no $c' \in C_n$ for $c' \neq c$ that are consistent with $S_c$. This gives a contradiction, thus proving that $\mathrm{TD}(C_n) \geq \mathrm{ND}(C_n)$. ▮

Combining Lemma 6 with the theorem of Natarajan gives the following result.

COROLLARY 7. *For concept classes $C_n$ chosen from the domain of Boolean functions over $n$ variables,* $\mathrm{TD}(C_n) \geq \mathrm{VCD}(C_n)/n$.

We know derive a lower bound for the teaching dimension that applies in all domains.

THEOREM 8. *For any concept class $C$ over domain $X$,* $\mathrm{TD}(C) \geq (\lg |C| - 1)/\lg |X|$.

*Proof.* We begin by observing that if $\mathrm{TD}(C) \leq k$ then $|C| \leq 2^k \binom{|X|}{k} \leq 2 \cdot |X|^k$. Thus,

$$\mathrm{TD}(C) \geq \min\{k \text{ such that } 2 \cdot |X|^k \geq |C|\}.$$

Solving for minimum such $k$ yields the desired result. ▮

By applying this theorem to the Boolean domain we get the following corollary.

COROLLARY 9. *For concept classes $C_n$ chosen from the domain of Boolean functions over $n$ variables* $\mathrm{TD}(C_n) \geq (\lg C_n - 1)/n$.

## 5. COMPUTING THE TEACHING DIMENSION

In this section we compute the teaching dimension for several concept classes and provide general techniques to aid in computing the teaching dimension for other concept classes.

Before considering the problem of computing the teaching dimension, we first briefly discuss the computational problem of finding an optimal teaching sequence for a given target concept. For the concept classes considered below, not only do we compute the teaching dimension, we also give efficient algorithms to find the optimal teaching sequence for any given target. However, in general, what can we say about the problem of finding an optimal teaching sequence?

We define the *optimal teaching sequence* problem as follows. The input contains a list of the positive examples for each concept in $C$ in some standard encoding. In addition, the input contains a concept $c_* \in C$ to teach and an integer $k$. The question is then: Is there a teaching sequence for $c_*$ of length $k$ or less? We now show that this problem is equivalent to the minimum cover problem. (See Garey and Johnson [6] for a formal description of the minimum cover problem).

THEOREM 10. *The optimal teaching sequence problem is equivalent to a minimum cover problem in which there are $|C| - 1$ objects to be covered and $|X|$ sets from which to form the covering.*

*Proof.* To see that these are equivalent problems, we associate the concepts from $C - \{c_*\}$ with the objects in the minimum cover problem. Similarly, we associate the sets for the minimum cover problem with the instances of the teaching problem as follows: An object associated with $c \in C - \{c_*\}$ is placed in the set associated with instance $x \in X$ if and only if $c(x) \neq c_*(x)$ (i.e., $x$ distinguishes $c$ from $c_*$). Observe that the sets in which a given instance is placed are distinct from the concepts for which the instance is

positive. It is easily seen that an optimal teaching sequence directly corresponds to an optimal set covering with $|C| - 1$ objects and $|X|$ sets. ∎

We note that Shinohara and Miyano [20] independently obtained the similar result that computing an optimal teaching sequence (what they call the minimum key problem) is $\mathcal{NP}$-complete by giving a reduction from the hitting set problem. Also a similar result was independently proven by Cherniavsky, Valuathapillai, and Statman [4] in the context of their "helpful game" formulation of inference. More recently, Anthony *et al.* [2] have considered the problem of computing an optimal teaching sequence when it is know that every instance is a positive example of *exactly* three concept from $C$. By giving a reduction to exact cover by 3-sets, they show that even in this restricted situation the problem of computing an optimal teaching sequence is $\mathcal{NP}$-hard. Since, the set covering problem is known to be $\mathcal{NP}$-complete even when all sets have size at most three [6], the following corollary to Theorem 10 immediately follows.

COROLLARY 11. *The optimal teaching sequence problem is $\mathcal{NP}$-hard even if it is known that each instance in $X$ is a positive example for at most three concepts from $C$.*

While it is $\mathcal{NP}$-complete to compute a minimum set covering, Chvatal [5] proves that the greedy algorithm (which is a polynomial-time algorithm) computes a cover that is within a logarithmic factor of the minimum cover. While this problem appears to be well understood, there is one very important distinction between the minimum cover problem and the problem of computing an optimal teaching sequence. In the set-covering problem, no assumptions are made about the structure of the sets, and thus they are input as lists of positive examples in some standard encoding. For the problem of computing an optimal teaching sequence, we are usually assuming that there is a short (polynomial-sized) representation of the objects contained in each set, such as a simple monomial. Thus we suggest the following interesting research question: What is the complexity of the set-covering problem when the sets have some natural and concise description?

Although the problem of computing the optimal teaching sequence for teaching a given concept is interesting, we now focus on computing the teaching dimension for various concept classes. We start by describing a straightforward relation between the teaching dimension and the number of membership queries needed to achieve exact identification. (A *membership query* is a call to an oracle that, on input $x$ for any $x \in X$, classifies $x$ as either a positive or negative instance according to the target concept $c \in C$.)

OBSERVATION 12. *The number of membership queries needed to exactly identify any given $c \in C$ is at least* $\mathrm{TD}(C)$.

*Proof.* Suppose that $\mathrm{TD}(C)$ is greater than the number of membership queries needed for exact identification. By the definition of exact identification, the sequence of membership queries used must be a teaching sequence that is shorter than the claimed shortest teaching sequence. This gives a contradiction. ∎

Thus an algorithm that achieves exact identification using membership queries provides an upper bound on the teaching dimension.

As noted earlier, since the teaching dimension is equivalent to the optimal mistake bound under teacher-directed learning, the results of Goldman *et al.* [9] give tight bounds on the teaching dimension for binary relations and total orders. We now compute both upper and lower bounds on the teaching dimension for the concept classes of monotone monomials, arbitrary monomials, monotone decision lists, orthogonal rectangles in $\{0, 1, ..., n-1\}^d$, monotone $k$-term DNF formulas, and $k$-term $\mu$-DNF formulas.

### 5.1. *Monomials*

We now prove a tight bound on the teaching dimension for the class of monotone monomials, and then we generalize this result for arbitrary monomials.

THEOREM 13.[3] *For the concept class $C_n$ of monotone monomials over $n$ variables*

$$\mathrm{TD}(C_n) = \min(r + 1, n),$$

*where $r$ is the number of relevant variables.*

*Proof.* We begin by exhibiting a teaching sequence of length $\min(r + 1, n)$. First we present a positive example in which all the relevant variables are 1 and the rest are 0. This example ensures that no irrelevant variables are in the target monomial. (If all the variables are in the monomial then this positive example can be eliminated).

Next we present $r$ negative examples to prove that the $r$ relevant variables are in the monomial. To achieve this goal, we take the positive example from above and flip each relevant bit, one at a time. So for each relevant variable $v$ there is a positive example and a negative example that differ only in the value of $v$, thus proving that $v$ is relevant. Thus this sequence is a teaching sequence.

We now prove that no shorter sequence of examples suffices. If any variable is not in the monomial, a positive example is required to rule out the monomial containing all variables. We now show that $r$ negative examples are required. At best, each negative example proves that at least one variable, from those that are 0, must be in the target. Suppose that a set of $r - 1$ negative examples (and any num-

[3] Shinohara and Miyano [20] independently showed that the teaching dimension of monotone monomials over $n$ variables is at most $n$.

ber of positive examples) proved that all $r$ relevant variables must be in the target. We construct a monomial, missing a relevant variable, that is consistent with this example sequence: for each negative example select one of the relevant variables that is 0 and place it in the monomial. This procedure clearly creates a consistent monotone monomial with at most $r-1$ literals. ∎

We now give a simple extension of these ideas to give a tight bound on the teaching dimension for the class of arbitrary monomials. The key modification is that the positive examples not only prove which variables are relevant, but they also provide the sign of the relevant variables. (By the *sign* of a variable we simply mean whether or not the variable is negated.)

THEOREM 14. *For the concept class $C_n$ of monomials over $n$ variables*

$$\mathrm{TD}(C_n) = \min(r+2, n+1),$$

*where $r$ is the number of relevant variables.*

*Proof.* First we exhibit a teaching sequence of length $\min(r+2, n+1)$. We present two positive examples—in each make all the literals in the target monomial true and reverse the setting of all the irrelevant variables. For example, if there are five variables and the target monomial, is $\bar{v}_1 v_2 v_5$ then present "01001, +" and "01111, +." Next, $r$ negative examples are used to prove that each remaining literal is in the monomial; take the first positive example and negate each relevant variable, one at a time. For the example above, the remainder of the teaching sequence is "11001, −", "00001, −," and "01000, −."

We now prove that the above example sequence is a teaching sequence for the target monomial. We use the following facts.

*Fact* 1. Let $C_n$ be the class of monomials and let $c \in C_n$ be the target concept. If some variable $v$ is 0 in a positive example then $v$ cannot be in $c$. Likewise, if $v$ is 1 in a positive example then $\bar{v}$ cannot be in $c$.

*Fact* 2. Let $C_n$ be the class of monomials, and let $c \in C_n$ be the target concept. Let $x^+ \in X_n$ be a positive example and let $x^- \in X_n$ be a negative example. If $x^+$ and $x^-$ are identical, except that some variable $v$ is 1 in $x^+$ and 0 in $x^-$, then $v$ must appear in $c$. Likewise, if $x^+$ and $x^-$ are identical, except that some variable $v$ is 0 in $x^+$ and 1 in $x^-$, then $\bar{v}$ must appear in $c$.

We first prove that no irrelevant variables are in a monomial consistent with the positive examples. Since each irrelevant variable is set to both 0 and 1 in a positive example, it follows from Fact 1 that none of these variables could appear in any consistent monomial. (Each relevant variable has the same value in both positive instances.) From Fact 1 it also follows that the positive examples provide the signs of the relevant variables. (If all variables are in the monomial, then only one positive example revealing the sign of each variable is needed.)

We now show that any monomial that is consistent with the negative examples must contain all the relevant variables. For each relevant variable $v_r$ the teaching sequence contains a positive example and a negative example that differ only in the assignment to $v_r$; so by Fact 2, $v_r$ must be relevant. Thus the above example sequence is in fact a teaching sequence.

We now prove that no shorter sequence suffices. If any variable, say $v_i$ is irrelevant then at least two positive examples are required in a valid teaching sequence since the teacher must prove that both $v_i$ and $\bar{v}_i$ are not in the target monomial. Finally, the argument used in Theorem 13 proves that $r$ negative examples are needed. ∎

Observe that Theorems 13 and 14 can easily be modified to give the dual result for the classes of monotone and arbitrary 1-DNF formulas, where a 1-DNF formula is simply a disjunction of the variables in $V_n$.

### 5.2. *Monotone Decision Lists*

Next we consider the concept class of monotone decision lists.

THEOREM 15. *For the concept class $C_n$ of monotone decision lists over $n$ variables*

$$\mathrm{TD}(C_n) \leqslant 2n-1.$$

*Proof.* We construct a teaching sequence of length at most $2n-1$. For each variable $v_i$ (assume that all the irrelevant variables are at the end of the list with an associated bit of 0), we first teach $b(v_i)$ and then we teach the ordering of the nodes.

To teach $b(v_i)$ present the instance $x$ in which $v_i$ is 1 and all other variables are 0. Then $b(v_i)$ is 1 if and only if $x$ is positive. Thus using $n$ examples, we teach the bit associated with each variable. Next we teach the ordering of the nodes. Observe that for consecutive nodes $y_i$ and $y_{i+1}$ for which $b(y_i) = b(y_{i+1})$, reversing the order of these nodes produces an equivalent concept. Thus, the learner can order them arbitrarily. For each $1 \leqslant i \leqslant n-1$ we present the example in which all the variables are 0 except for $y_i$ and all $y_j$, where $j > i$ and $b(y_j) \neq b(y_i)$. (Figure 4 shows the teaching sequence for the target concept of Fig. 1.) Observe that the $i$th example in this portion of the teaching sequence proves that $y_i$ precedes all nodes $y_j$ for which $j > i$ and $b(y_i) \neq b(y_j)$. This ordering information is sufficient to reconstruct the ordering of the nodes. ∎

We now show that the upper bound of Theorem 15 is asymptotically tight by proving that the teaching dimension

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | , | − |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | , | − |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | , | + |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | , | + |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | , | + |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | , | + |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | , | − |
| 1 | 1 | 1 | 0 | 0 | 0 | 1 | , | + |
| 1 | 0 | 0 | 1 | 1 | 1 | 0 | , | − |
| 0 | 1 | 0 | 0 | 0 | 1 | 1 | , | + |
| 0 | 1 | 0 | 1 | 0 | 0 | 1 | , | + |
| 0 | 0 | 0 | 0 | 1 | 0 | 1 | , | − |
| 0 | 1 | 0 | 0 | 1 | 0 | 0 | , | + |

FIG. 4. Teaching sequence for the target concept shown in Fig. 1.

of monotone decision lists is at least $n$. Unless the learner knows $b(v_i)$ for all $i$, he could not possibly know the target concept. However, to teach $b(v_i)$ (for $1 \leqslant i \leqslant n$), $n$ examples are needed; any single example only teaches $b_j$ for the smallest $j$ for which $y_j$ is 1.

### 5.3. *Orthogonal Rectangles in* $\{0, 1, ..., n-1\}^d$

We now consider the concept class of orthogonal rectangles in $\{0, 1, ..., n-1\}^d$.

THEOREM 16.[4] *For the concept class* $C_d$ *of orthogonal rectangles in* $\{0, 1, ..., n-1\}^d$:

$$\mathrm{TDS}(C_d) = 2 + 2d.$$

*Proof.* We build the following teaching sequence $T$. Select any two opposing corners of the box and show those points as positive instances. Now for each of these points show the following $d$ negative instances: for each dimension, give the neighboring point (unless the given point is on the border of the space $\{0, 1, ..., n-1\}^d$ in the given dimension) just outside the box in that dimension as a negative instance. (See Fig. 5 for an example teaching sequence.) Clearly the target concept is consistent with $T$, thus to prove that $T$ is a teaching sequence we need only show that it is the only concept that is consistent with $T$. Let $b$ be the target box and suppose that there is some other box $b'$ that is consistent with $T$. Since $b \neq b'$ either $b'$ makes a false positive or false negative error. However, the two positive examples ensure that $b' \supseteq b$ and the negative examples ensure that $b \supseteq b'$. Thus such a $b'$ could not exist.

[4] Romanik and Smith [18, 17] independently obtained this result for the special case that $d = 2$.

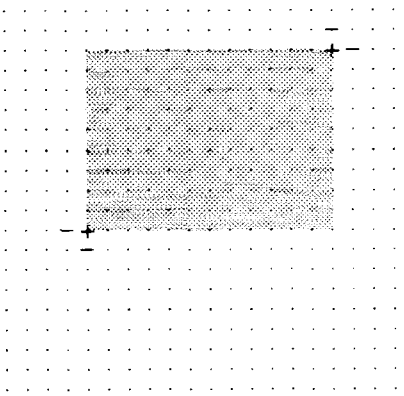

FIG. 5. The teaching sequence for a concept selected from the class of orthogonal rectangles in $\{0, 1, \cdots, n-1\}^d$ for $n = 20$ and $d = 2$.

We now show that no sequence $T'$ of less than $2 + 2d$ examples could be a teaching sequence. Suppose that $n \geqslant 4$ and let the target concept be a box defined by opposing corners $1^d$ and $(n-2)^d$. We first argue that $T'$ must contain two positive points—if $T'$ contained a single positive point then the box containing only that point would be consistent with $T'$. Thus any teaching sequence must contain at least two positive points. Finally, to prevent a hypothesis $B'$ from making a false positive error there must be a negative example that eliminates any hypothesis that moves any face out by even one unit. Clearly a single point can only serve this purpose for one face. Since a $d$-dimensional box has $2d$ faces, $2d$ negative examples are needed. ∎

### 5.4. *Monotone k-Term DNF Formulas*

We now describe how bounds on the teaching dimension for simple concept classes can be used to derive bounds on the teaching dimension for more complex classes. We begin by using the result of Theorem 13 to upper bound the teaching dimension for monotone $k$-term DNF formulas. We note that it is crucial to this result that the learner knows $k$. While the teacher can force the learner to create new terms, there is no way for the teacher to enforce an upper bound on the number of terms in the learner's formula.

THEOREM 17. *For the class* $C_n$ *of monotone k-term DNF formulas over n variables,*

$$\mathrm{TD}(C_n) \leqslant l + k,$$

*where l is the number of literals in the target formula.*

*Proof.* Let $f = t_1 \vee t_2 \vee \cdots \vee t_k$ be the target formula, and let $f(x)$ denote the value of $f$ on input $x$. We assume, without loss of generality, that $f$ is *reduced*, meaning that $f$ is not equivalent to any formula obtained by removing one of its terms. The approach we use is to independently teach each term of the target formula.

For all $i$ we build the teaching sequence $T_i$ for term $t_i$ (as if $t_i$ were a concept from the class of monotone monomials) as described in Theorem 13. We now prove that $T = T_1 T_2 \cdots T_k$ is a teaching sequence for $f$. The key property we use is:

$$\text{for any } x \in T_i, f(x) = + \text{ if and only if } t_i(x) = +. \quad (1)$$

To prove that property (1) holds we prove that, for all $x \in T_i$ all terms, except for possibly $t_i$, are negative on $x$. Recall that all variables not in $t_i$ are 0 in every $x \in T_i$. Thus we need just to prove that each term of $f$, except for $t_i$, must contain *some* variable that is not in $t_i$. Suppose that for term $t_j$ no such variables exists. Then $t_j$ would contain a subset of the variables in $t_i$. However, this violates the assumption that $f$ is reduced. Thus property (1) holds.

We now use property (1) to prove that $T$ is a teaching sequence for $f$. We first show that $f$ is consistent with $T$. From property (1) we know that $f$ is positive on $x \in T_i$ if and only if $t_i(x) = +$ and, since $T_i$ is a teaching sequence for $t_i$, it follows that $f(x) = +$, if and only if $x$ is positive. Thus $f$ is consistent with $T$. Finally, we prove by contradiction that $T$ uniquely specifies $f$. For monotone monomials $g_1, g_2, \ldots, g_k$ suppose the $g = g_1 \vee \cdots \vee g_k$ is consistent with $T$, yet $g \neq f$. Then there must exist some term $t_i$ from $f$ that is not equal to any term in $g$. Without loss of generality suppose that $g_j(x) = +$ for a positive point $x \in T_i$. (Some term in $g$ must be true on $x$ since $g$ is assumed to be consistent with $T_i$.) By property (1), this implies that $g_j$ can only contain those variables that are in $t_i$. Finally, since $g_j$ must reply correctly on all the negative points in $T_i$, it follows that $g_j = t_i$, giving the desired contradiction.

To complete the proof of the theorem, we need to compute the size of $T$. From Theorem 13 we know that for each $i$, $|T_i| \leqslant r + 1$, where $r$ is the number of literals in $t_i$. Thus it follows that $|T| \leqslant l + k$, where $l$ is the number of literals in $f$. ∎

We note that by using the teaching sequence from monotone 1-DNF formulas as the "building blocks" we can prove a dual result for monotone $k$-term CNF formulas. Also we obtain the lower bound of

$$\mathrm{TD}(C_n) \geqslant l,$$

where $C_n$ is the class of monotone $k$-term DNF formulas over $n$ variables and $l$ is the number of literals in the target formula, by just using the lower bound on the teaching dimension of the monomials.

### 5.5. *k-Term μ-DNF Formulas*

We now extend the idea of Theorem 17 to use teaching sequences for monomials to build a good teaching sequence for a $k$-term $\mu$-DNF formula.

THEOREM 18. *For the class $C_n$ of $k$-term $\mu$-DNF formulas over $n$ variables,*

$$\mathrm{TD}(C_n) \leqslant n + 2k.$$

*Proof.* As in Theorem 17 we assume that the target formula $f = t_1 \vee t_2 \vee \cdots t_k$ is reduced. Once again, the key idea here is to independently teach each term of $f$.

For each term $t_i$ of $f$ we construct a portion $T_i$ of the final teaching sequence $T = T_1 \cdots T_k$ using the basic strategy from our method of teaching monomials described in the proof of Theorem 14. Namely,

1. $T_i$ should consist of two positive examples in which all the relevant variables are 1 and each irrelevant variable is set differently in the two examples (except when overriden by (3) below).

2. $T_i$ should consist of $r$ negative examples that are obtained by taking one of the two positive examples and negating each relevant variable, one at a time.

In addition we add the requirements that:

3. If $t_i$ is a singleton term (e.g., $t_i = v_j$ or $t_i = \bar{v}_j$) then *both* positive examples (and thus, also, each negative example) in $T_1, \ldots, T_{i-1}, T_{i+1}, \ldots, T_k$ should be selected so that $t_i$ is false.

4. If $t_i$ is not a singleton term, then each positive example (and thus, also, each negative example) in $T_1, \ldots, T_{i-1}, T_{i+1}, \ldots, T_k$ should have at least one literal from term $t_i$ set to false. Since $f$ is a $\mu$-formula and all the remaining terms contain at least two literals, this goal is easily achieved.

We now prove that $T$ is a teaching sequence for $f$. From properties (3) and (4) above it follows that for any $x \in T_i$, all the terms in the formula, except for possibly $t_i$, are false on input $x$. We now prove that $T_i (1 \leqslant i \leqslant k)$ is a teaching sequence for $t_i$. The key observation is to first consider the portion of the teaching sequence associated with the singleton terms. It follows from properties (1) and (2) and the proof of Theorem 14 that each singleton is a term in the formula. Finally, since each variable can only appear in one term, by properties (1) and (2) the portion of the teaching sequence associated with each remaining term is a teaching sequence for that term. Thus the technique of Theorem 14 can be used to prove that $T$ is a teaching sequence for $f$. ∎

Note that we obtain the lower bound

$$\mathrm{TD}(C_n) \geqslant n,$$

where $C_n$ is the class of $k$-term $\mu$-DNF formulas over $n$ variables by just using the lower bound on the teaching dimension of monomials.

### 5.6. *Classes Closed under XOR*

We now discuss a situation in which one can generate a teaching sequence that has length logarithmic in the size of the concept class. For $c_1, c_2 \in C$ we define $c = c_1 \text{XOR } c_2$ as follows: for each instance $x \in X$, $c(x)$ is the exclusive-or of $c_1(x)$ and $c_2(x)$. We say a concept class $C$ is *closed under* XOR if the concept $c$, obtained by taking the bitwise exclusive-or of any pair of concepts $c_i, c_j \in C$ (for all $i, j$), is also in $C$.

THEOREM 19. *If $C$ is closed under* XOR *then there exists a teaching sequence of size at most* $\lfloor \lg(|C| - 1) \rfloor + 1$.

*Proof.* We construct a teaching sequence using the following algorithm.

*Build-teaching-sequence* $(f_*, C)$

1. Repeat until $f_*$ is uniquely determined
2. Find instance, $x$, for which $f_*$ disagrees with a non-eliminated function from $C$
3. Use $x$ as the next example

We now show that each instance selected by *Build-teaching-sequence* removes at least half of the non-eliminated functions from $C - \{f_*\}$. Consider the example $x$ added in step 2 of *Build-teaching-sequence*. Let $\tau$ contain the examples that have already been presented to the learner, and let $\mathcal{V}$ contain the concepts from $C - \{f_*\}$ that are consistent with $\tau$. We now show that at least half of the concepts in $\mathcal{V}$ must disagree with $x$. Suppose that $x$ is a positive example. If all the concepts in $\mathcal{V}$ predict that $x$ is negative, then $x$ eliminates all remaining concepts in $\mathcal{V}$ besides the target. Otherwise, there exists some set $\mathcal{V}_x^+$ of concepts that predict $x$ is positive. (Let $\mathcal{V}_x^- = \mathcal{V} - \mathcal{V}_x^+$ be the concepts from $\mathcal{V}$ that predict $x$ is negative.) By the choice of $x$, it must be that $|\mathcal{V}_x^-| \geq 1$, so let $g_1$ be a concept in $\mathcal{V}_x^-$. We now use the fact that $C$ is closed under XOR to prove that for each concept $g_2 \in \mathcal{V}_x^+$, there is a one-to-one mapping to a concept in $\mathcal{V}_x^-$. First consider the result of taking $f_* \oplus g_1$. Since all the elements in $\mathcal{V}$ are consistent with examples in $\tau$, for these instances $f_* \oplus g_1$ is 0. For the instance $x$, $f_* \oplus g_1$ is 1. Now consider taking $f_* \oplus g_1 \oplus g_2$ for $g_2 \in \mathcal{V}_x^+$. Since the instances in $\tau$ are 0 in $f_* \oplus g_1$, in $f_* \oplus g_1 \oplus g_2$ the instances in $\tau$ are the same as in $f_*$. For $x$, $f_* \oplus g_1 \oplus g_2$ is 0. Finally, since all concepts in $\mathcal{V}_x^+$ must disagree with each other on some instance in $X - \tau - \{x\}$, the concepts $f_* \oplus g_1 \oplus g_2 \in \mathcal{V}_x^-$ form a one-to-one correspondence with the concepts $g_2 \in \mathcal{V}_x^+$ (Fig. 6).

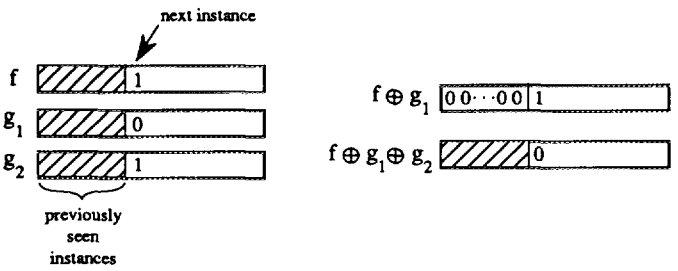


FIG. 6. Result of combining with XOR.

Repeating this same argument when $x$ is a negative instance, we conclude that on each instance at least half of the concepts in $\mathcal{V}$ are eliminated. Since *Build-teaching-sequence* removes at least half of the non-eliminated concepts with each example, after at most $\lfloor \lg(|C| - 1) \rfloor$ examples $\mathcal{V}$ contains at most one concept from $C - \{f\}$. Finally, one additional example is used to distinguish this remaining concept from the target concept. ∎

## 6. CONCLUSIONS AND OPEN PROBLEMS

In this paper we have studied the complexity of teaching a concept class by considering the minimum number of instances that a teacher must reveal to uniquely identify any target concept chosen from the class. A summary of our specific results are given in Table I. Observe that, unlike universal identification sequences, for these concept classes the teaching sequence selected is highly dependent on the target concept.

While this model of teaching provides a nice initial model, it clearly has some undesirable features. By restricting the teacher to teach all consistent learners, one side effect is that some concepts, such as that given in Lemma 1 that intuitively are easy to teach, are extremely difficult to teach

**TABLE I**

**Summary of Results**

| Concept class | Lower bound | Upper bound |
|---|---|---|
| Monotone monomials | $\min(r+1, n)$ | $\min(r+1, n)$ |
| Monomials | $\min(r+2, n+1)$ | $\min(r+2, n+1)$ |
| Monotone $k$-term DNF formulas | $l+1$ | $l+k$ |
| $k$-term $\mu$-DNF formulas | $n$ | $n+2k$ |
| Monotone decision lists | $n$ | $2n-1$ |
| Orthogonal rectangles in $\{0, 1, \ldots, n-1\}^d$ | $2+2d$ | $2+2d$ |

in our model. We are currently exploring variations of our model in which the teacher is more powerful, yet collusion is still forbidden. Also, a variation of this model in which the teaching sequence is only required to eliminate $\epsilon$-bad hypotheses is an interesting direction to pursue.

Finally, we suggest the following open problems. It would be quite informative to determine whether large and powerful classes (such as polynomial-sized monotone circuits) have polynomial teaching dimensions. Potentially the technique of Goldman *et al.* [8] may be useful in solving this problem. Another good area of research is to study the time complexity of computing optimal teaching sequences. In Section 5 we not only prove that there are small teaching dimensions for many classes, but we actually give efficient algorithms for computing the optimal teaching sequence. Are there natural classes for which computing the optimal teaching sequence is hard?

## ACKNOWLEDGMENTS

We are indebted to Manfred Warmuth for suggesting this line of research and providing the definition of the teaching dimension. We thank Ron Rivest for many useful conversations and his comments on an earlier draft of this paper. We thank Tibor Hegedus for providing the proof for Theorem 8. Finally, we thank an anonymous referee for many useful suggestions and comments.

## REFERENCES

1. D. Angluin, Queries and concept learning, *Mach. Learning* **2**, No. 4 (1988), 319–342.
2. M. Anthony, G. Brightwell, D. Cohen, and J. Shawe-Taylor, On exact specification by examples, *in* "Proceedings, Fifth Annual Workshop on Computational Learning Theory," pp. 311–318, ACM Press, July 1992.
3. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, Learnability and the Vapnik–Chervonenkis dimension, *J. Assoc. Comput. Mach.* **36**, No. 4 (1989), 929–965.
4. J. C. Cherniavsky, M. Velauthapillai, and R. Statman, Inductive inference: An abstract approach, *in* "Proceedings, 1988 Workshop on Computational Learning Theory," pp. 251–266, Morgan Kaufmann, San Mateo, CA, 1988.
5. V. Chvatal, A greedy heuristic for the set covering problem, *Math. Oper. Res.* **4**, No. 3 (1979), 233–235.
6. M. R. Garey and D. S. Johnson, "Computers and Intractability: A guide to the Theory of NP-Completeness," Freeman, San Francisco, 1979.
7. S. Goldman and D. Mathias, Teaching a smarter learner, *in* "Proceedings, Sixth Annual ACM Conference on Computational Learning Theory," pp. 67–76, ACM Press, New York, 1993.
8. S. A. Goldman, M. J. Kearns, and R. E. Schapire, Exact identification of circuits using fixed points of amplification functions, *SIAM J. Comput.* **22**, No. 4 (1993), 705–726.
9. S. A. Goldman, R. L. Rivest, and R. E. Schapire, Learning binary relations and total orders, *SIAM J. Comput.* **22**, No. 5 (1993), 1006–1034.
10. S. A. Goldman, "Learning Binary Relations, Total Orders, and Read-once Formulas," Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, MIT, July 1990.
11. S. Goldwasser, S. Goldwasser, and C. Rackoff, The knowledge complexity of interactive proofs, *in* "26th Annual Symposium on Foundations of Computer Science, October 1985," pp. 291–304.
12. J. Jackson and A. Tomkins, A computational model of teaching, *in* "Proceedings, Fifth Annual Workshop on Computational Learning Theory," pp. 319–326, ACM Press, New York, 1992.
13. N. Linial, Y. Mansour, and R. L. Rivest, Results on learnability and the Vapnik–Chervonenkis dimension, *Inform. and Comput.* **90**, No. 1 (1991), 33–49.
14. W. Maass and G. Turán, On the complexity of learning from counterexamples, *in* "30th Annual Symposium on Foundations of Computer Science, October 1989," pp. 262–267.
15. B. K. Natarajan, On learning Boolean functions, *in* "Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing," pp. 296–304, May 1987.
16. R. L. Rivest, Learning decision lists, *Machine Learning* **2** (3) (1987), 229–246.
17. K. Romanik, Approximate testing and learnability, *in* "Proceedings of the Fifth Annual Workshop on Computational Learning Theory," pp. 327–332, ACM Press, July 1992.
18. K. Romanik and C. Smith, Testing geometric objects, Technical Report UMIACS-TR-90-69, University of Maryland, College Park, Department of Computer Science, 1990.
19. S. Salzberg, A. Delcher, D. Heath, and S. Kasif, Learning with a helpful teacher, *in* "12th International Joint Conference on Artificial Intelligence, August 1991," pp. 705–711.
20. A. Shinohara and S. Miyano, Teachability in computational learning, *New Generation Comput.* **8** (1991), 337–347.
21. L. Valiant, A theory of learnable, *Comm. ACM* **27**, No. 11 (1984), 1134–1142.
22. V. N. Vapnik and A. Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theory Probab.* **16**, No. 2 (1971), 264–280.