On the Limits of Efficient Teachability

Rocco A. Servedio*

Division of Engineering and Applied Sciences

Harvard University

Cambridge, MA 02138

rocco@deas.harvard.edu

Keywords: computational learning theory, machine learning, teaching dimension

1 Introduction

In recent years a number of researchers in learning theory have developed and analyzed computational models of teaching [8, 9, 10, 13, 16, 17, 18]. One motivation for this work is the hope that stronger and more realistic positive results might be achievable in a setting where labeled examples are provided by a helpful teacher rather than an omniscient adversary or a worst-case probability distribution as is usually assumed in learning theory. However, as discussed in [8, 9, 13, 16], any reasonable teaching model must disallow teacher-learner collusion which trivializes the learning process – for instance, it should not be possible for the teacher to simply encode a representation of the target concept in the instances it chooses for the learner according to a prearranged coding scheme.

Motivated by these considerations, Goldman and Kearns [8] have studied a model in which the teacher must provide the learner with a *teaching sequence,* i.e. a sequence of labeled examples which is consistent with the target concept and with no other concept

*Supported in part by an NSF Graduate Fellowship and by NSF grant CCR-95-04436.

from the class being taught. Goldman and Kearns also defined a combinatorial measure called the *teaching dimension* of a concept class, which measures the minimum number of labeled examples required to uniquely specify any concept in the class (precise definitions are given in Section 1.1). This measure takes into account the idea that a helpful teacher will choose the most useful sequence of examples possible, but disallows teacher-learner collusion in that the examples chosen must provide enough information to identify the target concept to any learning algorithm, not just a learner who is "in cahoots" with the teacher.

In this note we answer two open questions posed by Goldman and Kearns [8] about this teaching model. Goldman and Kearns established polynomial upper bounds on the teaching dimension of monotone monomials, arbitrary monomials, monotone decision lists, monotone DNF, and read-once DNF, and asked whether the class of polynomial-size monotone circuits also has polynomial teaching dimension. In Section 2 we give a simple construction which yields a strong negative answer. The construction in fact shows that

- The class of polynomial-sized monotone formulae has exponential teaching dimension;
- The class of polylogarithmic-sized monotone formulae has superpolynomial teaching dimension;
- The class of subpolynomial-sized depth-3 monotone formulae has superpolynomial teaching dimension.

Goldman and Kearns also posed the question of whether there are natural concept classes for which computing optimal (i.e. shortest) teaching sequences is computationally hard. We provide a positive answer in Section 3 by showing that the problem of computing the length of an optimal teaching sequence is NP-hard for the well-studied class of intersections of two halfspaces. Our results indicate strong limits on efficient teachability in the Goldman-Kearns model.

1.1 Definitions

Let X be a finite set which is the space of all possible instances. A *concept* c is a Boolean function mapping X to $\{0,1\}$ and a *concept class* $C \subseteq 2^X$ is a collection of concepts. We can equivalently view a concept c as a subset of X, so "$x \in c$" and "$c(x) = 1$" mean the same thing. A *labeled instance* is an ordered pair $\langle x, y \rangle$ where $x \in X$ and $y \in \{0, 1\}$. We say that a sequence S of labeled instances is *consistent* with a concept c if $c(x) = y$ for every pair $\langle x, y \rangle$ belonging to S. Such a sequence S is a *teaching sequence* for $c \in C$ if S is consistent with no other concept in C. The name "teaching sequence" is appropriate because a helpful teacher can use such a sequence to ensure that a learner can uniquely identify the target concept. A teaching sequence for $c \in C$ is *optimal* if no other teaching sequence for c contains fewer examples. We write $|S|$ to denote the number of examples in a teaching sequence S.

Goldman and Kearns [8] defined the *teaching dimension* of a concept class C to be

$$TD(C) = \max_{c \in C} \{|S| \ : \ S \text{ is an optimal teaching sequence for } c \in C\}.$$

The teaching dimension of C is thus the minimum number of labeled examples which suffice to uniquely specify any concept in C. A closely related notion was defined and studied by Shinohara and Miyano [18], who refer to the length of an optimal teaching sequence for $c \in C$ as the *specification number* of c and say that a concept class C is *teachable by examples* if $TD(C)$ is bounded by a polynomial.

2 Large Teaching Dimension for Small Monotone Formulae

2.1 A hard-to-teach class

For $x \in X$ the singleton concept c_x is the concept which takes value 1 on the point x and value 0 on all other points. Let C be the concept class of cardinality $|X| + 1$ which contains the $|X|$ different singleton concepts and the null concept $c_\emptyset$ which takes value 0 on all points. Goldman and Kearns [8] have observed that the concept class C has teaching dimension $|C| - 1 = |X|$. This is because the shortest teaching sequence for $c_\emptyset$ must contain every instance in X (if an instance x is not included then the sequence is consistent with both $c_\emptyset$ and c_x). Jackson and Tomkins [13] have used this observation to show that the teaching dimension of the class of 1-decision lists is 2^n, and Anthony et. al. [1] have used it to show that the teaching dimension of the class of linear threshold functions over $\{0,1\}^n$ is 2^n.

We will use the following simple variant of this observation: let A, B be disjoint subsets of X, let c_A be the characteristic function of A, and for $b \in B$ let c_b be the characteristic function of $A \cup \{b\}$. Then the concept class C' of cardinality $|B| + 1$ which consists of the $|B|$ different concepts c_b and the concept c_A has teaching dimension $|B|$, since a teaching sequence for c_A must include every instance in B.

2.2 The lower bound

We consider the instance space $X = \{0,1\}^n$ with n even. For $x \in \{0,1\}^n$ we write $|x|$ to denote $x_1 + \ldots + x_n$, i.e. the number of 1s in x. Let $A = \{x \in \{0,1\}^n : |x| \geq n/2\}$ and let $B = \{x \in \{0,1\}^n : |x| = n/2 - 1.\}$. The majority function MAJ_n is the characteristic function of A, and for $b \in B$ the characteristic function of $A \cup \{b\}$ is $MAJ_n \vee T_b$, where T_b is the monotone monomial

$$\bigwedge_{b_i=1} x_i.$$

Valiant [19] has established the existence of an $O(n^{5.3})$-size monotone formula for MAJ_n. It follows that each concept c_A, c_b can be expressed as a monotone formula of size $O(n^{5.3})$. Since $|B| = \binom{n}{n/2-1} = \Theta(2^n/\sqrt{n})$, the observation of Section 2.1 implies

Theorem 1 *Let C_1 be the class of monotone formulae of size $O(n^{5.3})$ over variables $x_1, \ldots, x_n$. Then $TD(C_1)$ is exponential in n.*

By specializing this construction to a small subset of variables we can show that smaller concept classes also have superpolynomial teaching dimension. For $1 \leq k \leq n$ we define A_k and B_k by

$$A_k = \{x \in \{0,1\}^n : x_1 + \cdots + x_k \geq k/2\}$$

$$B_k = \{x \in \{0,1\}^n : x_1 + \cdots + x_k = k/2 - 1\}.$$

The concept class $\{c_{A_k}\} \cup \{c_{b_k}\}_{b_k \in B_k}$ has teaching dimension at least $\binom{k}{k/2-1}$. This is because for every $(k/2-1)$-element subset of the first k variables, any teaching sequence for c_{A_k} must include some negative example in which the first k variables have 1s in exactly the positions corresponding to that subset. These negative examples must be distinct for different subsets, so the teaching dimension is at least $\binom{k}{k/2-1} = \Theta(2^k/\sqrt{k})$.

Using Valiant's construction, each concept c_{A_k}, c_{b_k} can be expressed as a monotone formula of size $O(k^{5.3})$. Taking $k = \omega(\log n)$ yields $\binom{k}{k/2-1} = n^{\omega(1)}$, and we obtain

Theorem 2 *Let $k(n)$ be a function which is $\omega(\log n)$ and let C_2 be the class of monotone formulae of size $O(k(n)^{5.3})$ over variables $x_1, \ldots, x_n$. Then $TD(C_2)$ is superpolynomial in n.*

A different construction yields a lower bound on the teaching dimension of small constant-depth circuits. A folklore theorem in circuit complexity states that the function MAJ_k can be expressed as a monotone depth-3 OR-AND-OR formula of size $2^{O(\sqrt{k}\log k)}$. The function c_{A_k} is precisely MAJ_k; for each c_{b_k} we need only OR the above formula with a monotone conjunction of $k/2-1$ variables. This can be done without increasing the depth and hence each c_{b_k} can also be expressed as a depth-3 formula of size $2^{O(\sqrt{k}\log k)}$. We thus have

Theorem 3 *Let $k(n)$ be a function which is $\omega(\log n)$ and let C_3 be the class of monotone formulae of size $2^{O(\sqrt{k(n)}\log k(n))}$ and depth 3 over $x_1, \ldots, x_n$ (note that this size bound is $n^{o(1)}$ for $k(n) = o((\log n/\log\log n)^2)$). Then $TD(C_3)$ is superpolynomial in n.*

Using different methods Jeff Jackson [12] has given another proof that the class of subpolynomial-sized constant-depth monotone formulae has superpolynomial teaching dimension.

2.3 Discussion

Goldman and Kearns [8] observed that the number of membership queries needed to exactly learn any concept class C is at least $TD(C)$. Our results in Section 2.2 thus imply that the concept classes C_1, C_2 and C_3 are not learnable from a polynomial number of membership queries. Hellerstein et al. [11] essentially gave a converse to the Goldman-Kearns observation by proving that for any projection-closed Boolean concept class C, if C has polynomial teaching dimension then C is exactly learnable from a polynomial number of membership queries.

As an application of their result, Hellerstein et al. showed that the class of $\log n$-CNF $\cap$ $\log n$-DNF is learnable from a polynomial number of membership queries (this is the class of

Boolean functions which can be expressed both as a CNF where each clause contains at most $\log n$ literals and as a DNF where each term contains at most $\log n$ literals). Our techniques can be used to give a complementary result showing that $\Theta(\log n)$ is best possible in the above theorem. It is easy to see that the concepts c_{A_k}, c_{b_k} defined above belong to the class k-CNF $\cap$ k-DNF (since each such concept depends on only the first k variables). Together with the Goldman-Kearns bound relating the number of membership queries required for learning to the teaching dimension, our results thus imply

Theorem 4 *For any function $k(n) = \omega(\log n)$, the class $k(n)$-CNF $\cap$ $k(n)$-DNF is not learnable from a polynomial number of membership queries.*

Finally, we also note that various augmentations of the Goldman-Kearns model have been proposed in the learning theory literature. Jackson and Tomkins [13] allow the teacher to pass the learner a small amount of "trusted information," while Goldman and Mathias [9] study coordinated teacher-learner pairs which are specially tailored to each other, and Mathias [16] considers an interactive model in which the teacher and learner communicate over a series of rounds. For each of these more sophisticated teaching models it is known that any concept class which is learnable by an efficient deterministic algorithm using membership and equivalence queries can be taught to a polynomial-time learner. Since the class C' of Section 2.1 is efficiently learnable from a single equivalence query, our simple lower bound technique does not apply.

3 Hardness of Computing Optimal Teaching Sequences

We now consider the computational complexity of computing optimal teaching sequences. The following problem was shown to be NP-hard via reduction from HITTING SET by Shinohara and Miyano [18] (see also [7]):

SPECIFICATION NUMBER

Instance: A class C of concepts over an n-element instance space (given as a $|C| \times n$ zero/one matrix), a concept $c \in C$, and an integer $k \leq |C|$.

Question: Is $|S| \leq k$, where S is an optimal teaching sequence for $c \in C$?

Goldman and Kearns [8] gave an alternate proof of hardness by reducing from MINIMUM COVER, and Anthony et. al [1] showed that the similar problem TEACHING DIMENSION is NP-hard:

TEACHING DIMENSION

Instance: A class C of concepts over an n-element instance space (given as a $|C| \times n$ zero/one matrix) and an integer $k \leq |C|$.

Question: Is $TD(C) \leq k$?

These results demonstrate that computing optimal teaching sequences can be computationally hard. However, in each of the above problems the concept class is part of the problem instance, and the reductions which establish hardness work by constructing specialized concept classes which correspond to instances of difficult combinatorial problems.

Goldman and Kearns [8] posed the question of whether there are fixed, natural concept classes for which computing the optimal teaching sequence is hard. In this section we give a positive answer by showing that the familiar class of intersections of two halfspaces has this property.

3.1 Intersections of Two Halfspaces

A Boolean function h on $\{0,1\}^n$ is a *halfspace* if there is a weight vector $w \in R^n$ and a threshold $\theta \in R$ such that $h(x) = 1$ iff $w \cdot x \geq \theta$. A function g is an *intersection of two halfspaces* if there are halfspaces h_1, h_2 such that $g(x) = 1$ iff $h_1(x) = h_2(x) = 1$ (i.e. $g = h_1 \cap h_2$). The concept class of intersections of halfspaces is a natural one which has been widely studied in computational learning theory [2, 3, 4, 5, 6, 14, 15, 20].

We consider the following problem:

SPECIFICATION NUMBER FOR INTERSECTION OF HALFSPACES (SNIH)

Instance: Two halfspaces h_1 and h_2 over $\{0,1\}^d$ (each of which is represented by a $(d+1)$-tuple $(w_1, \ldots, w_d, \theta) \in R^{d+1}$) and an integer $k \leq 2^d$.

Question: Is $|S| \leq k$, where S is an optimal teaching sequence for the concept $h_1 \cap h_2$?

Our hardness proof for SNIH uses the following lemma which is illustrated in Figure 1.

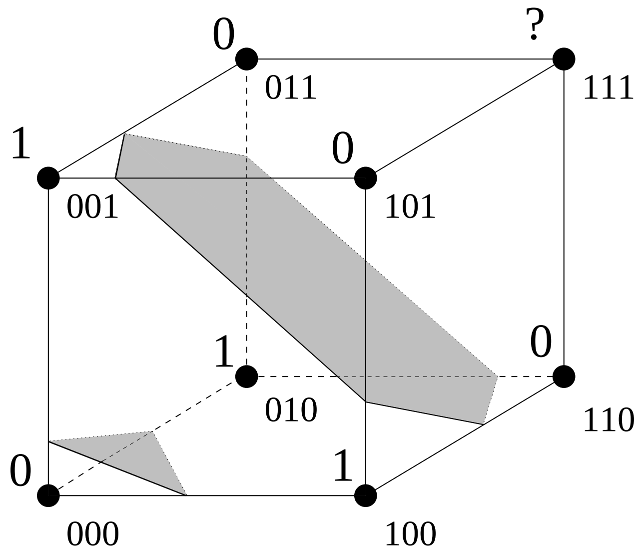

Figure 1: Any intersection of two halfspaces which induces this labelling on the three-dimensional Boolean cube must label 111 with a 0.

Lemma 5 *Let $g = h_1 \cap h_2$ be an intersection of halfspaces over $\{0,1\}^3$. If*

$$g(100) = g(010) = g(001) = 1 \qquad \text{and} \qquad g(000) = g(110) = g(101) = g(011) = 0$$

then $g(111) = 0$.

Proof: Let g be defined by the halfspaces $u \cdot x \geq \theta$ and $v \cdot x \geq \tau$. The positive examples $100, 010, 001$ imply that $u_1, u_2, u_3 \geq \theta$ and $v_1, v_2, v_3 \geq \tau$. One of the two halfspaces must contain at most one of the negative examples $110, 101, 011$, so without loss of generality we suppose that $u \cdot x \geq \theta$ does not contain 110 and 101, i.e. $u_1 + u_2 < \theta$ and $u_1 + u_3 < \theta$. This implies that $u_1, u_2, u_3 < 0$ and thus $u_1 + u_2 + u_3 < u_1 + u_2 < \theta$, so $g(111) = 0$. ■

The main result of this section is the following:

Theorem 6 *SNIH is NP-hard.*

Proof: We reduce from the following well-known NP-hard problem:

PARTITION

Instance: A finite sequence $a_1, \ldots, a_n$ of positive integers.

Question: Is there a subset $S \subseteq \{1, \ldots, n\}$ such that $\sum_{i \in S} a_i = \sum_{i \notin S} a_i$?

The reduction maps instances of PARTITION to instances of SNIH in the following way: if $a_1, \ldots, a_n$ is an instance of PARTITION and $A = \sum_{i=1}^{n} a_n$, then the corresponding $d = n+3$-dimensional SNIH instance is

$$\left(w \cdot x \geq \frac{3A}{2},\ -w \cdot x \geq -\frac{3A}{2},\ k = 2^{n+3} - 1\right)$$

where the vector $w \in R^{n+3}$ is defined as $w_i = a_i$ for $1 \leq i \leq n$ and $w_{n+1} = w_{n+2} = w_{n+3} = A$. Note that the target concept is equivalent to the hyperplane $w \cdot x = 3A/2$.

If $a_1, \ldots, a_n$ is a negative instance of PARTITION, then there is no $z \in \{0,1\}^n$ such that $a \cdot z = A/2$, so $w \cdot x = 3A/2$ contains no points of $\{0,1\}^{n+3}$. Since the class of intersections of two halfspaces over $\{0,1\}^{n+3}$ contains all 2^{n+3} singleton concepts as well as the null concept, as in Section 2.1 we have that the optimal teaching sequence for the concept $w \cdot x = 3A/2$ is of length 2^{n+3}, so the SNIH instance is negative.

Now suppose that $a_1, \ldots, a_n$ is a positive instance of PARTITION. Let $S \subseteq \{1, \ldots, n\}$ be the desired partition and let $z \in \{0,1\}^n$ be the characteristic vector of S. If we write (z, b_1, b_2, b_3) to denote $(z_1, \ldots, z_n, b_1, b_2, b_3) \in \{0,1\}^{n+3}$, then it is clear that $(z,1,0,0)$, $(z,0,1,0)$, $(z,0,0,1)$ satisfy $w \cdot x = 3A/2$ and that $(z,0,0,0)$, $(z,1,1,0)$, $(z,1,0,1)$, $(z,0,1,1)$ do not. By fixing the first n coordinates to be z, any intersection of two halfspaces over $\{0,1\}^{n+3}$ induces an intersection of two halfspaces over $\{0,1\}^3$. Lemma 5 now implies that any intersection of halfspaces which labels $(z,1,0,0)$, $(z,0,1,0)$, $(z,0,0,1)$ positively and $(z,0,0,0)$, $(z,1,1,0)$, $(z,1,0,1)$, $(z,0,1,1)$ negatively must label $(z,1,1,1)$ negatively. Thus the optimal teaching sequence for the concept $w \cdot x = 3A/2$ is of length at most $2^{n+3} - 1$, since it need not include the point $(z,1,1,1)$. ■

Theorem 6 shows that the problem of computing the length of an optimal teaching sequence can be computationally hard even for fixed natural concept classes. We note in this context that the teaching models proposed in [9, 13, 16] address this issue by explicitly considering the time complexity of the teaching algorithm as a model parameter. However, we also observe that the length of an optimal teaching sequence can be exponential for the class of intersections of two halfspaces (or even for the simpler class of single halfspaces), and thus exponential time could be required for a teacher to explicitly transmit such a teaching sequence. The following question is thus of interest: is there a natural concept class for which $TD(C)$ is polynomially bounded but the problem of constructing a teaching sequence is computationally hard? A related question, which has also been considered by Anthony et al. [1], is whether computing the specification number of a single halfspace is NP-hard.

4 Acknowledgements

We thank Les Valiant for his advice and encouragement, Jeff Jackson for helpful questions and for alerting us to [12], and the anonymous referees for several useful comments and suggestions.

References

[1] M. Anthony, G. Brightwell, and J. Shawe-Taylor. On specifying boolean functions using labelled examples, *Discrete Applied Math.,* 61(1), 1995, 1-25.

[2] E. Baum. On learning a union of halfspaces, *Journal of Complexity,* 6(1), 1990, 67-101.

[3] E. Baum. Polynomial time algorithms for learning neural nets, *IEEE Trans. Neural Networks,* 2, 1991, 5-19.

[4] A. Blum, P. Chalasani, S. Goldman and D. Slonim. Learning with unreliable boundary queries, *J. Comput. Syst. Sci.,* 56, 1998, 209-222.

[5] A. Blum and R. Kannan. Learning an intersection of a constant number of halfspaces over a uniform distribution, *J. Comput. Syst. Sci.,* 54, 1997, 371-380.

[6] A. Blum and R. Rivest. Training a 3-node neural net is NP-complete, *in* D. Touretzky, ed., "Advances in Neural Information Processing Systems I," 1989, 494-501.

[7] J. Cherniavsky, R. Statman and M. Velauthapillai. Inductive inference – an abstract approach, *in* "Proc. 1988 Workshop on Comp. Learning Theory," 1988, 251-266.

[8] S.A. Goldman and M.J. Kearns. On the complexity of teaching, *J. Comput. Syst. Sci.,* 50, 1995, 20-31.

[9] S.A. Goldman and H.D. Mathias. Teaching a smarter learner, *J. Comput. Syst. Sci.,* 52, 1996, 255-267.

[10] T. Hegedus. Generalized teaching dimensions and the query complexity of learning, *in* "Proc. Eighth Conf. on Comp. Learning Theory," 1995, 108-117.

[11] L. Hellerstein, K. Pillaipakkamnatt, V. Raghavan and D. Wilkins. How many queries are needed to learn? *J. ACM,* 43(5), 1996, 840-862.

[12] J. Jackson, personal communication, 2000.

[13] J. Jackson and A. Tomkins. A computational model of teaching, *in* "Proc. Fifth Ann. Workshop on Comp. Learning Theory," 1992, 319-326.

[14] S. Kwek and L. Pitt. PAC learning intersections of halfspaces with membership queries, *Algorithmica,* 22, 1998, 53-75.

[15] W. Maass and G. Turán. Algorithms and lower bounds for on-line learning of geometrical concepts, *Machine Learning,* 14, 1994, 251-269.

[16] H.D. Mathias. A model of interactive teaching, *J. Comput. Syst. Sci.,* 54, 1997, 487-501.

[17] S. Salzberg, A. Delcher, D. Heath and S. Kasif. Learning with a helpful teacher, *in* "Proc. 12th Int. Joint Conf. on Artif. Intel.," 1991, 705-711.

[18] A. Shinohara and S. Miyano. Teachability in computational learning, *New Generation Computing,* 8, 1991, 337-347.

[19] L. G. Valiant. Short monotone formulae for the majority function, *J. Algorithms,* 5, 1984, 363-366.

[20] S. Vempala. A random sampling based algorithm for learning the intersection of halfspaces, *in* "Proc. 38th Annual Symp. on Found. of Comp. Sci.," 1997, 508-513.