On the Complexity of Learning from Counterexamples and Membership Queries

Wolfgang Maass*,** and György Turán*,***

ABSTRACT.

We show that for any concept class $\mathcal{C}$ the number of equivalence and membership queries that are needed to learn $\mathcal{C}$ is bounded from below by $\Omega(\text{VC-dimension}(\mathcal{C}))$. Furthermore we show that the required number of equivalence and membership queries is also bounded from below by $\Omega(\text{LC - ARB}(\mathcal{C})/\log(1+\text{LC - ARB}(\mathcal{C})))$, where LC - ARB($\mathcal{C}$) is the required number of steps in a different model where no membership queries but equivalence queries with arbitrary subsets of the domain are permitted. These two relationships are the only relationships between the learning complexities of the common on-line learning models and the related combinatorial parameters that have remained open (see section 3 of [MT1]).

As an application of the first lower bound we determine the number of equivalence and membership queries that are needed to learn monomials of k out of n variables.

In the last section we examine learning algorithms for threshold gates that are based on equivalence queries. We show that a threshold gate can not only learn concepts but also non-decreasing functions in polynomially many steps. On the other hand we show that all distributed learning algorithms for threshold gates that are of a similar type as the Δ-rule or the WINNOW-algorithm are inherently slow.

*Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL. 60680. E-mail: U45381 @ UICVM.BITNET.

**Written under partial support by NSF-Grant CCR 8903398.

***Automata Theory Research Group of the Hungarian Academy of Sciences, Szeged, Hungary. Partially supported by OTKA-433. E-mail: U11557 @ UICVM.BITNET.

1. Introduction.

We continue in this paper our investigation [MT1] of the complexity of learning in the on-line learning models of Angluin [A1]. In the most basic one of these models (which may be viewed as a generalization of the classical learning models for perceptrons [R], [MP] and neural networks [N], [RM]) the learner proposes "hypotheses" H from a fixed "concept class" $\mathcal{C} \subseteq 2^X$ over a finite domain X. The goal of the learner is to "learn" an unknown "target concept" $C_T \in \mathcal{C}$ that has been fixed by the "environment". Whenever the learner proposes some hypothesis $H \in \mathcal{C}$, with $H \neq C_T$, the environment responds with some "counterexample" $x \in H\Delta C_T := (C_T - H) \cup (H - C_T)$. x is called a "positive counterexample" if $x \in C_T - H$, and x is called a "negative counterexample" if $x \in H - C_T$. A learning algorithm for $\mathcal{C}$ is any algorithm A that produces new hypotheses

$$H_{i+1}^A := A(x_1, \dots, x_i;\ H_1^A, \dots, H_i^A)$$

in dependence of counterexamples $x_j \in H_j^A \Delta C_T$ for the preceding hypotheses H_j^A. (One also refers to these hypotheses as "equivalence queries" [A1].)

The "learning complexity" LC(A) of such a learning algorithm A is defined by

$$\begin{aligned}\mathrm{LC}(A) := \max\{i \in \mathbf{N} \mid\ & \text{there is some } C_T \in \mathcal{C} \text{ and some} \\ & \text{choice of counterexamples} \\ & x_j \in H_j^A \Delta C_T \text{ for } j = 1, \dots, i-1 \\ & \text{such that } H_i^A \neq C_T\}.\end{aligned}$$

The "learning complexity" LC($\mathcal{C}$) of a concept class $\mathcal{C}$ is defined by

$$\mathrm{LC}(\mathcal{C}) := \min\{\mathrm{LC}(A) \mid A \text{ is a learning algorithm for } \mathcal{C}\}.$$

In a variation of this model one considers a more active learner A that can also "carry out experiments", i.e. he may also ask "membership queries" $x \in C_T$? for $x \in X$ (where X is the domain of the concept class $\mathcal{C} \subseteq 2^X$). For any learning algorithm A for $\mathcal{C}$ that uses equivalence and membership queries one defines LC(A) as the maximal number of queries that A needs to identify some target concept $C_T \in \mathcal{C}$ (for some choice of counterexamples to its equivalence queries). We set

$$\text{LC - MEMB}(\mathcal{C}) := \min\{\text{LC}(A) \mid A \text{ is a learning algorithm for } \mathcal{C} \text{ that may use equivalence queries with hypotheses } H \in \mathcal{C} \text{ and membership queries}\}.$$

It has turned out that there are several important concept classes for which learning with equivalence and membership queries is much easier than learning either with equivalence queries only, or with membership queries only: for example DFA [A2], [A4], k-term DNF [A3], [PV] and read-once formulas [AHK]. (In some of these results the model is somewhat different as it takes into account the amount of computation performed by the learning algorithm, resp. the length of the counterexamples received). However the exact power of the LC - MEMB model has remained somewhat elusive, because there has been no method available for proving nontrivial lower bounds for LC - MEMB.

We show in section 2 of this paper that LC - MEMB($\mathcal{C}$) = Ω(VC - dim($\mathcal{C}$)) for every concept class $\mathcal{C}$ (VC - dim($\mathcal{C}$) is the Vapnik-Chervonenkis dimension of $\mathcal{C}$, see [BEHW], [MT1]). As an application of this lower bound we determine LC - MEMB($\mathcal{C}_{k,n}$) for the class $\mathcal{C}_{k,n}$ of conjunctions of k literals from n variables.

In section 3 we establish a somewhat unexpected relationship between LC - MEMB($\mathcal{C}$) and LC - ARB($\mathcal{C}$). LC - ARB($\mathcal{C}$) is the learning complexity of $\mathcal{C}$ in another learning model where the learner may ask no membership queries, but equivalence queries with arbitrary hypotheses $H \subseteq X$ (i.e. it is not required that $H \in \mathcal{C}$). We will demonstrate in section 3 that this relationship can also be used to derive significant lower bounds for LC - MEMB($\mathcal{C}$) for various concrete $\mathcal{C}$.

In section 4 we address several questions about the complexity of learning algorithms for threshold gates (in this section we examine learning algorithms that use only equivalence queries). The class of concepts $C \subseteq \{1, \ldots, n\}^d$ that can be computed by a threshold gate of fan-in d (where the weights and the threshold of the gate may be arbitrary real numbers, or equivalently arbitrary integers) can be characterized as follows:

$$\text{HALFSPACE}_n^d := \{C \subseteq \{1, \ldots, n\}^d \mid \exists \text{ halfspace } H \subseteq R^d \text{ with } C = H \cap \{1, \ldots, n\}^d\}.$$

It has been shown that LC(HALFSPACE$_n^d$) is polynomial in d and $\log n$ [MT1].

We examine in section 4 of this paper the question whether this positive result can be extended to the arguably simplest threshold circuit with more than one threshold gate (Theorem 4), or to the learning of functions on a single threshold gate with several adaptive thresholds (Theorem 5). Furthermore we show for a large class of distributed learning algorithms for a threshold gate (where each weight is controlled by a different processor) that they are all substantially slower than the fastest unrestricted learning algorithm for a threshold gate (Theorem 6).

2. Learning With Equivalence and Membership Queries and the Vapnik-Chervonenkis Dimension.

It is obvious that LC($\mathcal{C}$) $\geq$ VC - dim($\mathcal{C}$) and MEMB($\mathcal{C}$) $\geq$ VC - dim($\mathcal{C}$) for any concept class $\mathcal{C}$ (MEMB($\mathcal{C}$) results from a restriction of the model for LC - MEMB where the learner may ask only membership queries). But there are concept classes $\mathcal{C}$ for which LC - MEMB($\mathcal{C}$) < VC - dim($\mathcal{C}$). For example for

$$\text{MAJORITY} := \left\{ C \subseteq \{0, 1, \ldots, n\} \mid 0 \in C \Leftrightarrow |C - \{0\}| > \frac{n}{2} \right\}$$

one has VC - dim(MAJORITY) = n, but LC - MEMB(MAJORITY) $\leq \frac{n}{2} + 1$ (ask first whether $0 \in C_T$, approximate C_T "from above" if the answer is yes, otherwise "from below"). Other examples show that

$$\text{LC - MEMB}(\mathcal{C}_n) \leq (2 - \log 3) \cdot \text{ VC - dim}(\mathcal{C}_n) + o(n)$$

for suitable concept classes $\mathcal{C}_n$ over $\{0,1\}^n$ with VC - dim($\mathcal{C}_n$) = n (see [MT1]).

The following result shows that nevertheless the VC - dimension provides for any $\mathcal{C}$ a lower bound for LC - MEMB($\mathcal{C}$).

Theorem 1. LC - MEMB($\mathcal{C}$) $\geq \frac{1}{7} \cdot$ VC - dim($\mathcal{C}$) for every concept class $\mathcal{C}$.

Remarks.

1. The proof shows that in fact LC - ARB - MEMB($\mathcal{C}$) = Ω(VC - dim($\mathcal{C}$)) for any $\mathcal{C}$, where LC - ARB - MEMB is the learning model that allows both membership queries and equivalence queries with arbitrary hypotheses.
2. Theorem 1 in combination with Theorem 2.1 of [BEHW] implies that for any finite $\mathcal{C}$ $O(\max(\frac{1}{\epsilon}\log\frac{2}{\delta}, \frac{\text{LC - MEMB}(\mathcal{C})}{\epsilon}\log\frac{13}{\epsilon}))$ samples are sufficient for pac-learning. This appears to be the first result which indicates that learning in the LC - MEMB model cannot be substantially faster than pac-learning (if one ignores possible differences in the computational complexity).

Proof of Theorem 1. Let $S \subseteq X$ be a set of maximal size that is shattered by $\mathcal{C}$ (i.e. $\mathcal{C} \cap S = 2^S$), where X is the domain of $\mathcal{C}$. Consider the following adversary strategy (we write CAND for the class of those $C \in \mathcal{C}$ that are still candidates for C_T at the beginning of the considered learning step; we write CAND′ for the class of $C \in \mathcal{C}$ that are still candidates after the adversary has given his response in the considered step):

I. For a membership query $x \in C_T$? reply "yes" iff $|\{C \cap S \mid C \in \text{CAND and } x \in C\}| \geq |\{C \cap S \mid C \in \text{CAND and } x \notin C\}|$.

II. For a hypothesis $H \subseteq X$ choose some $g \in S$ as counterexample such that $M_g \geq M_x$ for all $x \in S$, where $M_x := |\{C \cap S \mid C \in \text{CAND and } H(x) \neq C(x)\}|$.

It is obvious that $|\{C \cap S \mid C \in \text{CAND}'\}| \geq \frac{1}{2}|\{C \cap S \mid C \in \text{CAND}\}|$ in case I. However an estimate of the type $|\{C \cap S \mid C \in \text{CAND}'\}| = \Omega(|\{C \cap S \mid C \in \text{CAND}\}|)$ is in general false for case II (e.g. consider the case where $|C_T \cap S| = 1$ and $\{C \cap S \mid C \in \text{CAND}\} = \{M \subseteq S \mid |M| \leq 1\}$ and $H = \emptyset$; we have here $1 = |\{C \cap S \mid C \in \text{CAND}'\}| \leq \frac{1}{|S|} \cdot |\{C \cap S \mid C \in \text{CAND}\}|$ for any possible counterexample from S). However the following Lemma (which is of some interest on its own) implies that as long as $|\{C \cap S \mid C \in \text{CAND}\}|$ is still "relatively large" an estimate $|\{C \cap S \mid C \in \text{CAND}'\}| = \Omega(|\{C \cap S \mid C \in \text{CAND}\}|)$ can also be established for case II.

Lemma. Let $f : (0,1) \to \mathbf{R}$ be defined by $f(z) = -z\log z - (1-z)\log(1-z)$. Then for any $\alpha \in (0,1)$, $Y \neq \emptyset$, and $\mathcal{E} \subseteq 2^Y$ with $|\mathcal{E}| \geq 2^{\alpha \cdot |Y|}$ one has for the unique $\beta \in (0, \frac{1}{2}]$ with $f(\beta) = \alpha$ that

$$\exists y \in Y\left(\beta \leq \frac{|\{E \in \mathcal{E} \mid y \in E\}|}{|\mathcal{E}|} \leq 1 - \beta\right).$$

Proof of the Lemma. Let R be a random variable with uniform distribution over $\mathcal{E}$. For $y \in Y$ let R_y be the induced random variable with

$$R_y = \begin{cases} 1, & \text{if } y \in R \\ 0, & \text{if } y \notin R. \end{cases}$$

Then $\Pr[R_y = 1] = |\{E \in \mathcal{E} \mid y \in E\}|/|\mathcal{E}|$. The entropy $H(R)$ of the random variable R satisfies $H(R) = \log|\mathcal{E}| \geq \alpha \cdot |Y|$. Since one can identify R with the vector $\langle R_y\rangle_{y \in Y}$ of the random variables R_y, one has $\sum_{y \in Y} H(R_y) \geq H(R)$ (this follows from the fact that $H(U,V) = H(U) + H(V \mid U)$ and $H(V \mid U) \leq H(V)$ for arbitrary random variables U, V; see [CK]). Therefore there exists some $y_0 \in Y$ with $H(R_{y_0}) \geq \alpha$.

One has $\alpha \leq H(R_{y_0}) = f(\Pr[R_{y_0} = 1]) = f(1 - \Pr[R_{y_0} = 1])$. Fix $\beta \in (0, \frac{1}{2}]$ so that $f(\beta) = \alpha$ (such β exists because $f[(0,1)] = (0,1]$ and $f(z) = f(1-z)$). Since f is non-decreasing on $(0, \frac{1}{2}]$ one gets from $f(\Pr[R_{y_0} = 1]) \geq \alpha$ that $\Pr[R_{y_0} = 1] \geq \beta$. Similarly $f(1 - \Pr[R_{y_0} = 1]) \geq \alpha$ implies that $1 - \Pr[R_{y_0} = 1] \geq \beta$, thus $\Pr[R_{y_0} = 1] \leq 1 - \beta$. □

In order to finish the proof of Theorem 1 one sets $\alpha := \frac{1}{3}$ and $Y := S$ in the Lemma. For $\beta \in (0, \frac{1}{2}]$ with $f(\beta) = \alpha$ one has then $\beta \geq \beta_0 := 0.0615$. The Lemma guarantees that as long as $|\{C \cap S \mid C \in \text{CAND}\}| \geq 2^{\frac{1}{3}|S|}$, after the response of the adversary it holds that

$$|\{C \cap S \mid C \in \text{CAND}'\}| \geq \beta_0 |\{C \cap S \mid C \in \text{CAND}\}|.$$

Hence if for some i it holds that $\beta_0^i \geq 2^{-\frac{2}{3}|S|}$ then after i responses of the adversary

$$|\{C \cap S \mid C \in \text{CAND}\}| \geq 2^{|S|} \cdot \beta_0^i \geq 2^{\frac{1}{3}|S|} > 1$$

and the learning algorithm cannot reach a conclusion yet. The proof is completed by noting that i can be chosen to be $\frac{1}{7}|S|$. □

As an application of Theorem 1 we determine LC - MEMB for the concept class

$$\mathcal{C}_{k,n} := \{C \subseteq \{0,1\}^n \mid C \text{ is definable by an AND of up to } k \text{ literals from } \{x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_n\}\}.$$

This concept class is of interest not only in the case $k = n$ but also in the case $k \ll n$ (see [L]): many practically occurring concepts are defined as an AND of very few literals from a very large reservoir of potentially relevant attributes. However if the number n of potentially relevant attributes is so large that n learning steps are not feasible, then $\mathcal{C}_{k,n}$ is not learnable from equivalence queries only since $\text{LC}(\mathcal{C}_{k,n}) \geq \binom{n}{k}$. Thus it is of interest to examine whether $\mathcal{C}_{k,n}$ can be learned substantially faster if the learner can make equivalence and membership queries.

Theorem 2. LC - MEMB$(\mathcal{C}_{k,n}) = \Theta(k(1+\log\frac{n}{k}))$.

Remark.

1. This result shows in particular that LC$(\mathcal{C}_{n,n}) =$ LC - MEMB$(\mathcal{C}_{n,n}) = \Theta(n)$. To our knowledge no significant lower bound was previously known for LC - MEMB$(\mathcal{C}_{n,n})$ (or for LC - MEMB$(\mathcal{C}_{k,n})$ in the case $k < n$).
2. Littlestone [L] has shown that $\mathcal{C}_{k,n}$ can be learned equally fast if one uses equivalence queries with arbitrary hypotheses. Remark 1 after Theorem 1 implies that no further speed-up results if one allows both arbitrary hypotheses and membership queries.

Idea of the proof of Theorem 2. For the upper bound one starts with $H_1 := \emptyset$ (which is defined by $x_1 \wedge \bar{x}_1$). If $H_1 \neq C_T$ and $p \in C_T - H_1$ is a positive counterexample then one can use membership queries to find the up to k "relevant" variables in p by binary search. The lower bound follows from Littlestone's result [L] that VC - dim$(\mathcal{C}_{k,n}) = \Omega(k(1+\log\frac{n}{k}))$ together with Theorem 1. □

3. Membership Queries Versus Equivalence Queries with Arbitrary Hypotheses.

It has been shown in [MT1] that there are concept classes $\mathcal{C}$ for which LC - MEMB$(\mathcal{C}) < \frac{1}{2}\cdot$ LC - ARB$(\mathcal{C})$. However it has remained unknown whether there are concept classes $\mathcal{C}$ for which LC - MEMB$(\mathcal{C})$ is substantially smaller than LC - ARB$(\mathcal{C})$ (in fact, apart from the relationship that has been settled in the preceding section, this remains the only open problem about the relationship between the learning complexities and combinatorial invariants that were considered in Fig. 1 of [MT1]).

We write LC - ARB - MEMB for the learning model where the learner may ask both membership queries and equivalence queries with arbitrary hypotheses $H \subseteq X$.

Theorem 3. LC - MEMB$(\mathcal{C}) \geq$ LC - ARB - MEMB$(\mathcal{C}) \geq$
$$\frac{\text{LC - ARB}(\mathcal{C})}{\log(1+\text{LC - ARB}(\mathcal{C}))} \geq \frac{\text{LC - ARB}(\mathcal{C})}{\log(1+\log|\mathcal{C}|)}$$
for any concept class $\mathcal{C}$ with $|\mathcal{C}| > 1$.

Remarks.

1. It is open whether this lower bound can be improved to LC - MEMB$(\mathcal{C}) = \Omega($LC - ARB$(\mathcal{C}))$.
2. The lower bound for LC - ARB - MEMB implies that a learner cannot learn substantially faster if in addition to the ability to ask arbitrary equivalence queries he can also ask membership queries.

Proof of Theorem 3. Consider a decision tree T for $\mathcal{C}$ (where each node is labeled by some $x \in X$, each non-leaf has two outgoing edges with labels 0,1, the leaves are labeled by the concepts from $\mathcal{C}$) so that every leaf has depth $\geq d :=$ LC - ARB$(\mathcal{C})$. Such T exists by a result of Littlestone [L] (see also [MT1]).

We use T to define an adversary strategy such that after any i membership queries and any j equivalence queries there exists a set R of $\geq \frac{2^d}{2^i\cdot(d+1)^j}$ nodes on level d such that below each $r \in R$ there exists a leaf that is labeled by some $C \in$ CAND (we define CAND as in the proof of Theorem 1). Initially one has $|R| = 2^d$. We consider now an arbitrary step in the learning process where $|R| > 1$.

Case I. The learner makes a membership query "$x \in C_T$?".

The adversary responds with "yes" iff $|\{v \in R \mid$ there is a leaf below v with label $C \in$ CAND such that $x \in C\}| \geq \frac{|R|}{2}$.

Case II. The learner makes an equivalence query with hypothesis $H \subseteq X$.

H need not occur as a leaf of T, but it defines a unique path from the root of T to some node v_H on level d. Assume for contradiction that for every node v on this path the immediate subtree of v which does not contain v_H contains fewer than $\frac{|R|}{d+1}$ nodes of R. Then R is contained in the union of $\{v_H\}$ and d sets of size $< \frac{|R|}{d+1}$. Thus one gets
$$|R| < 1 + d\cdot\frac{|R|}{d+1} \leq \frac{|R|}{d+1} + d\cdot\frac{|R|}{d+1} = 1,$$
a contradiction (we use for the second inequality the fact that $|R| \geq d+1$, this follows from the preceding observation together with the assumption that $|R| > 1$). Thus there exists some node $\tilde{v}$ on the path from the root to v_H so that the immediate subtree of $\tilde{v}$ that does not contain v_H has $\geq \frac{|R|}{d+1}$ nodes of R. The adversary gives the label of this node $\tilde{v}$ as a counterexample to H. Thus $\geq \frac{|R|}{d+1}$ nodes of R remain "alive" after this response. □

Consider the concept classes
$$\begin{aligned}\text{BALL}^d_n :=& \{C \subseteq \{1,\dots,n\}^d \mid \exists \text{ ball } B \subseteq \mathbf{R}^d \\ & \text{with } C = B \cap \{1,\dots,n\}^d\} \text{ and} \\ \text{HALFSPACE}^d_n :=& \{C \subseteq \{1,\dots,n\}^d \mid \exists \text{ halfspace } H \subseteq \mathbf{R}^d \\ & \text{with } C = H \cap \{1,\dots,n\}^d\}.\end{aligned}$$

The following result provides the first lower bound for LC - MEMB(HALFSPACE^d_n) and LC - MEMB(BALL^d_n) that is superlinear in d.

Corollary. LC - MEMB($\mathcal{C}$) $\geq$ LC - ARB - MEMB($\mathcal{C}$) = $\Omega(\frac{d^2 \log n}{\log d + \log\log n})$ for $\mathcal{C}$ = BALL$_n^d$, HALFSPACE$_n^d$.

Remark. The best known upper bounds for these concept classes $\mathcal{C}$ are: LC - ARB - MEMB($\mathcal{C}$) $\leq$ LC - ARB($\mathcal{C}$) $\leq \log|\mathcal{C}| = O(d^2 \log n)$ and LC - MEMB($\mathcal{C}$) $\leq$ LC($\mathcal{C}$) = $O(d^2(\log d + \log n))$; see [MT2].

The proof of the corollary uses Theorem 9 in combination with the fact that LC - ARB($\mathcal{C}$) = $\Omega(d^2 \log n)$ (Corollary 9 of [MT1]). □

4. On the Complexity of Learning Algorithms for Threshold Gates.

It has been shown [MT1] that LC(HALFSPACE$_n^d$) = $O(d^3(\log d + \log n))$ with a learning algorithm that requires only polynomially in d, $\log n$ many computation steps. In [MT2] this is improved to $O(d^2(\log d + \log n))$. We examine in this section possible extensions of these positive results to the learning of circuits with several threshold gates, to the learning of functions, and to distributed learning models.

Let 2 - HALFSPACE$_n^2$:= $\{C \cap C' \mid C, C' \in$ HALFSPACE$_n^2\}$. This concept class is of interest since it contains exactly those concepts that can be computed by an AND of two threshold gates of fan-in 2 (with input variables ranging over $\{1, \ldots, n\}$). A threshold circuit of this type appears to be the simplest type of threshold circuit that consists of more than just a single threshold gate.

Theorem 4. LC(2 - HALFSPACE$_n^2$) = $\Omega(n)$ (thus LC(2 - HALFSPACE$_n^2$) is not polynomial in LC(HALFSPACE$_n^2$)).

Proof. Let $Q := \{\langle i,j \rangle \in \{1, \ldots, n\}^2 \mid i \in \{1,n\}$ or $j \in \{1,n\}\}$ be the perimeter of the considered domain $X_n := \{1, \ldots, n\}^2$. Let B' be a ball of radius 1 with centerpoint $\langle \lceil \frac{n}{2} \rceil, \lceil \frac{n}{2} \rceil \rangle$, and set $B := B' \cap X_n$.

Consider the following adversary strategy. If $H \in$ 2 - HALFSPACE$_n^2$ is a hypothesis with $B - H \neq \emptyset$ one responds with a point from $B - H$ as positive counterexample. If $H \in$ 2 -HALFSPACE$_n^2$ is a hypothesis with $B \subseteq H$ then one responds with an arbitrary point from $H \cap Q$ as negative counterexample (it is easy to see that $B \subseteq H$ implies that $H \cap Q \neq \emptyset$).

Obviously there is a collection $\mathcal{C}$ of $\Theta(n)$ sets $H \in$ 2 - HALFSPACE$_n^2$ that are of the form $S \cap X_n$ for some strip S with $B \subseteq S$ which is bounded by two parallel lines with distance 2 and so that the sets $H \cap Q$ for $H \in \mathcal{C}$ are pairwise different. The preceding adversary strategy forces the learner to make $\Omega(n)$ equivalence queries because a positive counterexample from B eliminates no $C \in \mathcal{C}$ as possible target concept, and a negative counterexample from Q eliminates at most constantly many $C \in \mathcal{C}$ as possible target concepts. □

Remarks.

1. Theorem 4 complements the lower bound result of Blum and Rivest [BR] (which uses $P \neq NP$) for pac-learning on a circuit that is similar, except that in their circuit the two threshold gates on level 1 have fan-in n with binary input variables). Note also that their lower bound for the required computation time for pac-learning does not imply any lower bound for LC.
2. With a similar argument as in the proof of Theorem 4 (replace B by a circle of radius $\frac{n}{4}$, choose as negative counterexamples corner points of squares of minimal size that contain B) one can also show that LC(GENERAL - POSITION - BOX$_n^2$) = $\Omega(n)$, where GENERAL - POSITION - BOX$_n^2$:= $\{R \cap \{1, \ldots, n\}^2 \mid R$ is a rectangle (not necessarily axis-parallel)$\}$. This complements the result of [MT1], where it is shown that LC(BOX$_n^2$) = $O(\log n)$ for BOX$_n^2$:= $\{R \cap \{1, \ldots, n\}^2 \mid R$ is an axis-parallel rectangle$\}$.

The following result is motivated by the fact that a threshold gate with binary output is a rather unsatisfactory model for the computational abilities of a neuron. One usually views the "firing rate" of a neuron as its "output". This firing rate of a neuron is reported to change between a few and several hundred spikes per second [CA]. Therefore the usual type of (discrete) threshold gate with outputs from $\{0, 1\}$ provides only a very crude model for a neuron. In order to achieve a better approximation we consider instead a "multi-threshold gate" G that has in addition to its weights $w_1, \ldots, w_d \in \mathbf{R}$ s thresholds $t_1 \leq \ldots \leq t_s$ ($t_j \in \mathbf{R}$, $s \in \mathbf{N}$). We assume that such a gate G computes the following function $f_G : \{1, \ldots, n\}^d \to \{0, \ldots, s\}$:

$$f_G(x_1, \ldots, x_d) = \begin{cases} \max\left\{ j \mid \sum_{i=1}^{d} w_i x_i \geq t_j \right\}, & \text{if this set is not empty} \\ 0, & \text{otherwise.} \end{cases}$$

We write $\mathcal{F}_{n,d,s}$ for the class of all functions $f_G : \{1, \ldots, n\}^d \to \{0, \ldots, s\}$ that are computable by such multi-threshold gates G (with arbitrary weights and thresholds from $\mathbf{R}$). Note that $\mathcal{F}_{n,d,s}$ contains various discrete approximations to the frequently considered "sigmoid" continuous threshold functions [RM] (see also [OA] for results on non-monotone multilevel threshold functions).

In order to analyze the complexity of learning an arbitrary target function $f_T \in \mathcal{F}_{n,d,s}$ (through an exchange of hypotheses $f \in \mathcal{F}_{n,d,s}$ and counterexamples $\underline{x} \in \{1,\dots,n\}^d$ with $f(\underline{x}) \neq f_T(\underline{x})$) one has to specify what information the learner will receive about the counterexample $\underline{x}$:

- the correct value $f_T(\underline{x})$
- only the information whether $f_T(\underline{x}) > f(\underline{x})$ or $f_T(\underline{x}) < f(\underline{x})$
- just the information that $f_T(\underline{x}) \neq f(\underline{x})$.

The argument from the proof of Theorem 4 can be used to show that with the third type of feedback there is no learning algorithm that can learn arbitrary $f_T \in \mathcal{F}_{n,d,s}$ in polynomial in $d, s, \log n$ many steps. On the other hand the following result shows that there exists such a feasible learning algorithm for the second type of feedback. It is rather interesting that the first type of feedback (which appears to be quite unrealistic in the case of a neuron) is not required for fast learning of functions on a threshold gate.

Theorem 5. There is a learning algorithm A for $\mathcal{F}_{n,d,s}$ that learns any $f_T \in \mathcal{F}_{n,d,s}$ from at most $O((d+s)^2(\log(d+s)+\log n))$ counterexamples to hypotheses $f \in \mathcal{F}_{n,d,s}$ (we assume here that any counterexample to f is a pair $\langle \underline{x}, b\rangle$ from $\{1,\dots,n\}^d \times \{0,1\}$ with $f_T(\underline{x}) \neq f(\underline{x})$ and $b = 1$ iff $f_T(\underline{x}) > f(\underline{x})$). This learning algorithm A uses altogether at most polynomially in $\log n, d, s$ many computation steps.

The proof of Theorem 5 uses a reduction to learning a half-space in $\{1,\dots,n\}^{d+s}$. $\square$

The existence of a fast learning algorithm for HALFSPACE$_2^d$ (that requires only polynomially in d many equivalence queries and computation steps [MT1], [MT2]) gives rise to the question whether there exists a similarly fast learning algorithm for threshold gates that is distributed in the sense that each weight w_i ($i = 1,\dots,d$) is controlled by a separate processor (with some bound on the communication among these processors). Examples for distributed learning algorithms for threshold gates are the Δ-rule (also called Hebb's rule) [R], [MP], [RM] and Littlestone's algorithms WINNOW 1 and WINNOW 2 [L].

We will prove a lower bound for all learning algorithms that are K-bounded in the following sense.

Definition. We call a learning algorithm A for HALFSPACE$_2^d$ K-bounded (for some $K \in \mathbf{N}$) if the following conditions are satisfied.

a) There are $d+1$ sets $S_1,\dots,S_{d+1}$, where each S_i consists of up to K functions $h : \mathbf{R} \to \mathbf{R}$. We demand that $h \circ h' = h' \circ h$ for any $h, h' \in S_j$, $j = 1,\dots,d+1$.

b) Let $H_s := \{\underline{x} \in \{0,1\}^d \mid \sum_{i=1}^{d} w_i(s) \cdot x_i \geq t(s)\}$ be the s-th hypothesis of the learning algorithm A (in an arbitrary learning process). If H_s is not equal to the target concept then the learning algorithm A produces the next hypothesis H_{s+1} by choosing for each $i \in \{1,\dots,d+1\}$ some $h_i \in S_i$ and by setting $w_i(s+1) = h_i(w_i(s))$ for $i = 1,\dots,d$ and $t(s+1) = h_{d+1}(t(s))$ (there is no limitation on the way in which the operations $h_i \in S_i$ are chosen at each learning step).

Remark.

The definition of a K-bounded learning algorithm does not attempt to capture the intuitive notion of a "distributed learning algorithm". However a distributed learning algorithm where each processor can receive at any learning step only one of K possible signals from its environment (i.e. from the part of the input to which it has access, from other processors, and from the feedback-device) is likely to be K-bounded (provided that the weight-change operations of each processor are commutative). In particular it is easy to see that the Δ-rule, WINNOW 1, and WINNOW 2 are all K-bounded for $K := 3$.

Theorem 6. Let A be an arbitrary K-bounded learning algorithm that can learn any concept from some concept class $\mathcal{C} \subseteq$ HALFSPACE$_2^d$ with $\leq T$ equivalence queries (A is allowed to produce also hypotheses from HALFSPACE$_2^d$ that do not belong to $\mathcal{C}$). Then $T \geq |\mathcal{C}|^{1/K(d+1)} - 1$. In particular $T = 2^{\Omega(\frac{d}{K})}$ for $\mathcal{C} =$ HALFSPACE$_2^d$.

Proof. A can produce within T steps at most $(T+1)^{K\cdot(d+1)}$ different configurations of the parameters $w_1,\dots,w_d,t$ (since A is K-bounded the final configuration of the parameters only depends on how often each of the $\leq K$ operations $h \in S_i$ have been applied to w_i, respectively t, $i = 1,\dots,d+1$). This implies that $(T+1)^{K\cdot(d+1)} \geq |\mathcal{C}|$, because the final configurations are different for any two different target concepts from $\mathcal{C}$. Hence $T \geq |\mathcal{C}|^{1/K(d+1)} - 1$.

It is well known that $|\text{HALFSPACE}_2^d| \geq 2^{\frac{d(d-1)}{2}}$ (see [M]). $\square$

Remarks.

1. Its is obvious that the Δ-rule needs $2^{\Omega(d)}$ steps to learn certain target concepts $C \in$ HALFSPACE$_2^d$: at each step the Δ-rule can increase a weight by at most "+1"

(and certain $C \in \text{HALFSPACE}_2^d$ require exponential size weights). This argument can not be used to prove a lower bound for the WINNOW - rules, which can create exponential size weights in polynomially many steps. Furthermore no other argument is known that allows us to prove a lower bound for the speed of the WINNOW - algorithms which is superpolynomial in d.

2. The preceding theorem implies that neither WINNOW 1 nor WINNOW 2 can learn all monotone $C \in \text{HALFSPACE}_2^d$ in polynomially in d many steps (one uses here the fact that there are $\geq 2^{\frac{d(d-1)}{2}-d}$ different monotone $C \in \text{HALFSPACE}_2^d$). In fact, this result implies that if any WINNOW - algorithm learns any class $\mathcal{C}$ of monotone $C \in \text{HALFSPACE}_2^d$ in polynomially in d many steps, then $|\mathcal{C}| \leq 2^{O(d \log d)}$. On the other hand Littlestone [L] had exhibited classes $\mathcal{C}$ of monotone $C \in \text{HALFSPACE}_2^d$ with $|\mathcal{C}| = 2^d$ which can be learned by WINNOW in polynomially in d many steps (for example monotone disjunctions, and threshold functions with weights from $\{0,1\}$).

3. The preceding theorem implies that if any K-bounded learning algorithm for constant K can learn a class $\mathcal{C} \subseteq \text{HALFSPACE}_2^d$ in polynomially in d many steps, then $|\mathcal{C}| \leq 2^{O(d \log d)}$. This upper bound on the size of $\mathcal{C}$ is optimal (up to a constant factor in the exponent) since there exists in fact a class $\mathcal{C} \subseteq \text{HALFSPACE}_2^d$ of size $2^{\Theta(d \log d)}$ which can be learned by some K-bounded learning algorithm (with $K = 3$) in polynomially in d many steps: the familiar upper bound for the required number of learning steps of the Δ-rule (in terms of the separation δ of the target-hyperplane from 0) implies that for any fixed polynomial p the class

$$\mathcal{C}_p := \Big\{ C \subseteq \{0,1\}^d \mid \exists\, w_1, \ldots, w_d \in \{0,1,\ldots,p(d)\} \\ \exists t \in \mathbf{Z} \forall \underline{x} \in \{0,1\}^d \\ \Big(\underline{x} \in C \Leftrightarrow \sum_{i=1}^{d} w_i x_i \geq t \Big) \Big\}$$

can be learned with the Δ-rule in polynomially in d many steps (see Theorem 11.1 in [MP]). Furthermore it is easy to see that $|\mathcal{C}_p| = 2^{\Theta(d \log d)}$ for any fixed polynomial p with $p(d) \geq d$ (assign the numbers $1, \ldots, \frac{d}{2}$ in some permutation to $w_1, \ldots, w_{d/2}$, set $w_i := -1$ for $i > \frac{d}{2}$ and $t := 0$; any two permutations will define different concepts in $\mathcal{C}_p$).

This observation also implies that the Δ-rule is in a certain sense an optimal distributed learning algorithm for threshold gates: no other K-bounded learning algorithm with K constant can converge in polynomially in d many steps for a substantially larger class of target concepts from HALFSPACE_2^d. Furthermore, among subclasses of size $2^{O(d \log d)}$ the classes $\mathcal{C}_p \subseteq \text{HALFSPACE}_2^d$ (for which the Δ-rule converges fast) appear to be the most interesting ones. There are various examples of halfspaces that require integer weights of superpolynomial size, but they tend to be rather artificial from the point of view of learning (one such example is given in [M], it is easy to construct further examples by considering threshold gates that compare the size of two binary numbers).

REFERENCES

[A1] D. Angluin, Queries and concept learning, Machine Learning, **2**, (1988), 319-342.

[A2] D. Angluin, Learning regular sets from queries and counterexamples, Inf. and Contr., **75**, 1987, 87-106.

[A3] D. Angluin, Learning k-term DNF formulas using queries and counterexamples, Tech. Report YALEU / DCS / RR-559, Yale University, (1987).

[A4] D. Angluin, Equivalence queries and approximate fingerprints, Proc. of the 1989 Workshop on Computational Learning Theory (COLT '89), Morgan Kaufmann (San Mateo, 1989), 134-145.

[AHK] D. Angluin, L. Hellerstein, M. Karpinski, Learning read-once formulas with queries, Tech. Report TR-89-050, International Computer Science Institute (Berkeley, 1989).

[BR] A. Blum and R. L. Rivest, Training a 3-node neural network is NP-complete, Proc. of the 1988 Workshop on Computational Learning Theory (COLT '88), Morgan Kaufmann (San Mateo, 1988), 9-18.

[BEHW] A. Blumer, A. Ehrenfeucht, D. Haussler, M. K. Warmuth, Learnability and the Vapnik-Chervonenkis dimension, Tech. Report UCSC-CRL-87-20, University of California (Santa Cruz, 1988).

[CA] F. Crick, C. Asanuma, Certain aspects of the anatomy and physiology of the cerebral cortex, in: Parallel Distributed Processing vol. II, J. L. McClelland and D. E. Rumelhart, eds., MIT Press (Cambridge, 1986).

[CK] I. Csiszár and J. Körner, Information Theory, Academic Press (New York, 1981).

[Li] N. Littlestone, Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm, Machine Learning, **2** (4), 1987, 285-318.

[MT1] W. Maass and Gy. Turán, On the complexity of learning from counterexamples, Proc. of the 30th Annual IEEE FOCS 1989, 262-267.

[MT2] W. Maass and Gy. Turán, How fast can a threshold gate learn?, in preparation.

[MP] M. Minsky and S. Papert, Perceptrons: An Introduction to Computational Geometry, Expanded edition, MIT Press, 1988.

[M] S. Muroga, Threshold Logic and its Applications, Wiley (New York, 1971).

[OA] S. Olafsson and Y. S. Abu-Mostafa, The capacity of multilevel threshold functions, IEEE Transactions on Pattern Analysis and Machine Intelligence, **10** (2), (1988), 277-281.

[N] N. Nilsson, Learning Machines, McGraw-Hill (New York, 1965).

[PV] L. Pitt and L. G. Valiant, Computational limitations on learning from examples, J. of the ACM, **35**, (1988), 965-984.

[R] F. Rosenblatt, Principles of Neurodynamics, Spartan Books (New York, 1962).

[RM] D. E. Rumelhart and J. L. McClelland, Parallel Distributed Processing, MIT Press (Cambridge, 1986).