The Power of Random Counterexamples

Dana Angluin DANA.ANGLUIN@YALE.EDU and **Tyler Dohrn** TYLER.DOHRN@YALE.EDU
Department of Computer Science, Yale University

Editors: Steve Hanneke and Lev Reyzin

Abstract

Learning a target concept from a finite $n \times m$ concept space requires $\Omega(n)$ proper equivalence queries in the worst case. We propose a variation of the usual equivalence query in which the teacher is constrained to choose counterexamples randomly from a known probability distribution on examples. We present and analyze the Max-Min learning algorithm, which identifies an arbitrary target concept in an arbitrary finite $n \times m$ concept space using at most an expected $\log_2 n$ proper equivalence queries with random counterexamples.

Keywords: Proper Equivalence Query, Exact Concept Learning, Model Learning

1. Introduction

Query learning may be viewed as a game between a *learner* and a *teacher*. The learner attempts to learn a *target concept* chosen by the teacher from a known *concept space* by asking structured queries that are answered by the teacher.

We define a finite concept space as an $n \times m$ boolean matrix C with no duplicate rows or columns. We may think of the columns as examples and each row as a concept consisting of a subset of the examples, where $C_{ij} = 1$ if and only if example j is an element of concept i. In this boolean matrix representation of a concept space, the target concept is thus given by a target row.

The queries we consider are proper equivalence queries as defined by Angluin (1987, 1988). The learner chooses a row h and makes an equivalence query with h. The teacher responds either with "yes", indicating that row h is the target concept, or a *counterexample* k, which is an example on which the learner's hypothesis h and the target concept t disagree, that is, $C_{hk} \neq C_{tk}$. Note that because the values are boolean, the learner also learns the value of C_{tk} from the counterexample k.

The queries are *proper* in the sense that the learner's hypothesis must be drawn from the given concept space. For *improper* equivalence queries, the learner may hypothesize any subset of examples, represented by an arbitrary boolean m-vector. Concepts expressed by proper equivalence queries rather than improper equivalence queries may be preferred as more explanatory.

After making a number of equivalence queries and seeing a number of counterexamples, the *consistent* rows are those whose values agree with the target concept for all the counterexamples seen so far. Inconsistent rows cannot be the target concept and may be eliminated. The learner *exactly identifies* the target concept when either an equivalence query is answered "yes" or there is only one remaining consistent row, which must be the target concept. The goal of the learner is exact identification of any target concept using as few equivalence queries as possible. The strategy of exhaustive search, querying each row in turn, accomplishes exact identification in at most $n - 1$ equivalence queries for any target concept from any concept space.

Note that in general the teacher has a choice of which counterexample to return to the learner. In order to prove worst-case upper bounds, the usual assumption has been that the teacher answers adversarially, that is, chooses counterexamples to maximize the number of equivalence queries used by the learner. We consider an alternative assumption: there is a known probability distribution on examples, and in choosing a counterexample, the teacher selects at random from this distribution, conditioned on the example being a counterexample to the learner's hypothesis. We call this new setting *equivalence queries with random counterexamples*. We propose the Max-Min learning algorithm and show that it achieves exact identification of any target concept in any concept space using at most an expected $\log_2 n$ proper equivalence queries with random counterexamples.

For a motivating example, consider the Identity Concept Space, I_n, which is just the $n \times n$ identity matrix. Suppose that the learner poses an equivalence query with hypothesis row h, but the target row is actually row $t \neq h$. In the identity matrix, each pair of distinct row vectors i, j have Hamming distance 2. This means that the teacher has two choices of counterexample: column h or column t. The teacher could answer unhelpfully, selecting column h of the hypothesis row as a counterexample. The $n-1$ other rows of I_n have a 0 in column h, so only h is eliminated. The teacher could also answer helpfully, selecting column t as a counterexample. Row t is the only row that differs from row h in column t, so this would allow the learner to conclude with certainty that the target row is row t.

If we assume the teacher answers adversarially, learning a target row in the identity matrix requires $n-1$ equivalence queries in the worst case. This is undesirable, because many query learning tasks use a concept representation for which n is exponential in a task length parameter s. We would thus prefer to be able to learn a target row in $O(\log n)$ equivalence queries.

In the case of I_n with a uniform distribution over examples, consider a learning algorithm that repeatedly selects an arbitrary consistent row h to query. If h is not the target row, the equivalence query with random counterexamples for h has a $\frac{1}{2}$ probability of being answered unhelpfully, which eliminates just row h, and a $\frac{1}{2}$ chance of being answered helpfully, which immediately identifies the target concept. Thus, the expected number of equivalence queries used is bounded above by the expected number of fair coin flips until the first heads, which is just 2.

We survey related work in the next section, give fundamental definitions and two lower bounds in Section 3, prove the key property of elimination graphs in Section 4, give the Max-Min learning algorithm and its analysis in Section 5, explore relationships with the VC-dimension in Section 6, and conclude with a summary and an open problem in Section 7.

2. Related work

This work is in the area of exact learning with queries, described by Angluin (1988, 2004), which primarily concerns worst-case bounds for exact identification of a target concept assuming adversarial selection of counterexamples by the teacher. Our choice to consider randomly selected counterexamples is new. Vaandrager (2017) reviews a variety of methods and practical applications for model learning with queries.

Littlestone (1988) introduced the setting of online prediction of the labels of examples in which the key measure is the worst-case total number of mistakes of prediction made by the learner. Using the observation that a prediction algorithm has an implicit hypothesis, prediction algorithms and learning algorithms using (possibly improper) equivalence queries can be converted to each other, with bounds that differ by at most 1.

Littlestone defined the Halving Algorithm, an online prediction algorithm that on input x predicts the majority vote of the remaining consistent hypotheses on input x. For the Halving Algorithm the implicit hypothesis is the majority vote hypothesis as defined in Angluin (1988). A mistake of prediction (or, equivalently, a counterexample to the majority vote hypothesis) eliminates at least half the remaining consistent concepts, which implies a bound of $\log_2 n$ on the worst-case number of mistakes of prediction (or, equivalently, on the number of (possibly improper) equivalence queries until exact identification is achieved). Littlestone also defined the Standard Optimal Algorithm and gave a concept class on which the Standard Optimal Algorithm makes at most 2 mistakes, while the Halving Algorithm makes 3 mistakes for some target concept.

Though the strategy of choosing the majority vote hypothesis achieves at most $\log_2 n$ queries in the worst case, the hypotheses it uses are not necessarily drawn from the given concept class, that is, they may be improper. The worst-case lower bound on the number of proper equivalence queries that may be required for exact identification of a concept is $(n-1)$, as shown by the Identity Concept Space. This phenomenon occurs in practice: Angluin (1990) proved non-polynomial information theoretic lower bounds on learning finite automata and DNF formulas using proper equivalence queries. Our current work shows that by relaxing the selection of counterexamples to be random, a worst-case expected $\log_2 n$ proper equivalence queries suffice for exact identification in any finite concept class.

Valiant (1984) introduced the PAC model, in which the learner attempts to learn a concept with access to labeled examples drawn iid from an unknown probability distribution. In the PAC setting, the learner is only required to find a concept ϵ-approximately equal to the target concept, rather than exactly equal. Blumer et al. (1989) showed the close relationship between the VC-dimension of a concept class and the number of random examples required to learn concepts from it.

Haussler et al. (1994) considered a variant of the online prediction model in which the examples to predict are drawn iid from a fixed probability distribution. The focus of their work was to give an algorithm that has an asymptotically optimal bound of d/t on the probability of a mistake at trial t, where d is the VC-dimension of the concept class.

Another related area is research on the worst-case number of examples needed by a helpful teacher to teach a concept from a given concept class to a learner. The classic notion of the Teaching Dimension of a concept class was introduced by Shinohara and Miyano (1991) and Goldman and Kearns (1995). Zilles et al. (2011) introduced the Recursive Teaching Dimension of a concept class, which has very recently been shown to have a close relation to the VC-dimension of the class by Hu et al. (2017). In Section 6 we show that the VC-dimension gives a lower bound but not an upper bound on the worst case expected number of equivalence queries used by the Max-Min algorithm.

One important question is the computational feasibility of the Max-Min algorithm. For example, if the concept class consists of the deterministic finite acceptors (DFAs) over an alphabet of two symbols with s states, and examples are the strings of length l, then $n = 2^s s^{2s}$ and $m = 2^l$, which implies that $\log n$ is bounded by a polynomial in s. However, Gold (1978) showed that even the problem of determining whether there exists a DFA of s states consistent with a given set of labeled examples is NP-complete. Angluin (1987) showed that a learning algorithm using equivalence queries can be polynomially transformed to a PAC learning algorithm. Kearns and Valiant (1994) showed that under generally accepted cryptographic assumptions, concepts represented by DFAs are not polynomial-time PAC-learnable. Thus, we do not expect the Max-Min algorithm to run in time polynomial in $\log n$ in general.

3. Preliminaries

We introduce definitions and notation, and prove two lower bounds that provide context for the positive results of the paper.

3.1. Definitions and notation

For any positive integer n, $[n]$ denotes the set $\{0, 1, \ldots, n-1\}$. For any positive real x, $\log x$ denotes the base 2 logarithm of x.

Definition 1 *For positive integers n and m, denote the set of all $n \times m$ boolean matrices by $\mathcal{M}_{n\times m}$.*

We assume 0-based indexing of matrices, with row indices $[n]$ and column indices $[m]$.

Definition 2 *A matrix C is a **concept space** if $C \in \mathcal{M}_{n\times m}$ with no duplicated rows or columns.*

Because we assume a probability distribution over examples, the effect of duplicated columns can be achieved by adjusting the probabilities. Because no rows are duplicated, we know that rows i and j with $i \neq j$ will differ in at least one column, which simplifies the presentation.

Definition 3 *For a concept space $C \in \mathcal{M}_{n\times m}$, define $\Delta(i, j)$ to be the set of column indices in which the column-entries of row i and row j are unequal. That is,*

$$\Delta(i, j) := \{k : C_{ik} \neq C_{jk}\}.$$

Thus $\Delta(h, t)$ is the set of possible counterexamples when the hypothesis row is h and the target row is t. Next we specify the selection of counterexamples for equivalence queries with random counterexamples.

Definition 4 *Given a concept space C, an **example distribution** π is a probability distribution over the columns of C such that $\pi(k) > 0$ for every column $k \in [m]$. We assume that both the concept space C and the fixed example distribution π are known to the learner. For rows $i \neq j$, we define a random variable $K_{i,j}$ that selects a column k from the distribution π conditioned on the event $k \in \Delta(i, j)$. That is,*

$$\mathbb{P}[K_{i,j} = k] = \frac{\pi(k)}{\sum_{\ell\in\Delta(i,j)} \pi(\ell)}.$$

When the target concept is t and the learner makes an equivalence query with random counterexamples for hypothesis row $h \neq t$, the teacher selects a random counterexample given by $K_{h,t}$.

The next definition specifies for a given row i and column k, the fraction of rows of C that agree with the value of row i in column k. These rows will be eliminated as inconsistent if k is returned as a counterexample to hypothesis row i.

Definition 5 *In a concept space $C \in \mathcal{M}_{n\times m}$, the **column vote** of row i in column k is denoted $V(i, k)$ and is defined as the fraction of rows whose entry in column k is equal to that of row i. That is,*

$$V(i, k) = \frac{|\{j : C_{jk} = C_{ik}\}|}{n}.$$

Because every row is either 0 or 1 in column k, we have the following.

Lemma 6 *If C is a concept space and $C_{ik} \neq C_{jk}$ then $V(i,k) + V(j,k) = 1$.*

Next we introduce the elimination graph, which is the primary mathematical object of study in this paper.

Definition 7 *The **elimination graph** for a concept space C and an example distribution π, denoted by $G_{elim}(C, \pi)$, is a weighted directed graph with vertex set $V = [n]$ and real-valued edge weights $E(i,j) \in [0,1]$. For any pair of distinct vertices i and j, $E(i,j)$ is given by the expected fraction of rows eliminated from C if an equivalence query is posed with hypothesis row i and target row j. That is,*

$$E(i,j) = \sum_{k \in \Delta(i,j)} \mathbb{P}(K_{i,j} = k) \cdot V(i,k).$$

As one example of an elimination graph, consider the Identity Concept Space I_n, with π_n the uniform distribution on $[n]$. Then $G_{\text{elim}}(I_n, \pi_n) = (V, E)$, where $V = [n]$, and $E(i,j) = \frac{1}{2}$ for all rows $i \neq j$. As another example, consider the following 5×5 concept space and its elimination graph with respect to the uniform distribution on examples.

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} & 8/15 & 3/5 & 11/20 & 1/2 \\ 7/15 & & 8/15 & 8/15 & 7/15 \\ 2/5 & 7/15 & & 1/2 & 9/20 \\ 9/20 & 7/15 & 1/2 & & 2/5 \\ 1/2 & 8/15 & 11/20 & 3/5 & \end{pmatrix} \tag{1}$$

To see that $E(0,1) = 8/15$, note that if the first equivalence query is with row 0 and the actual target is row 1, then there are three possible counterexamples: column 0, which eliminates $3/5$ of the hypotheses, column 2, which eliminates $2/5$ of the hypotheses, and column 3, which eliminates $3/5$ of the hypotheses, for an expected fraction of $8/15 = (1/3)(3/5 + 2/5 + 3/5)$ of hypotheses eliminated.

We next define a measure that indicates the worst-case expected fraction of rows that will be eliminated if the learner makes an equivalence query with row i.

Definition 8 *Row i of a concept space $C \in \mathcal{M}_{n \times m}$ is called **α-informative** if $\min_{j \neq i} E(i,j) \geq \alpha$.*

If row i is α-informative, then an equivalence query with random counterexamples specifying hypothesis row i will eliminate at least a fraction α of the rows in C in expectation for any target concept. Intuitively, the notion of α-informativity allows the learner to compare the utility of various equivalence queries.

The main technical result of this paper, proved in Section 4, is that in every concept space C, there exists a row that is $\frac{1}{2}$-informative. This leads to the Max-Min learning algorithm, that achieves at most an expected $\log n$ equivalence queries to identify any target concept in any concept space of n rows. In the example in equation (1), the first and last rows are $\frac{1}{2}$-informative.

3.2. Random consistent hypotheses are not enough

The example of learning an arbitrary target row in the Identity Concept Space I_n using equivalence queries with random counterexamples suggests that random counterexamples are quite powerful. For the I_n example, we saw that target row identification can be achieved with $O(1)$ expected

equivalence queries for any consistent querying strategy. It might seem that random counterexamples are so powerful that the learner can propose randomly selected consistent hypothesis rows and still identify the target row in a sub-linear expected number of equivalence queries. We now show that this is not true in general, even in a space consisting of I_n with a single row of all zeros appended, which we denote by J_n.

Definition 9 *For any positive integer n, J_n is the concept space consisting of $n+1$ rows and n columns in which the first n rows are the same as in I_n and the last row consists of n zeroes.*

Theorem 10 *Consider the concept space J_n with the uniform distribution on examples. Suppose the learner repeatedly selects a hypothesis row chosen uniformly at random from the currently consistent rows. When the target is row n (the row of all zeroes), the learner requires an expected $\Omega(n)$ equivalence queries with random counterexamples for exact identification of the target row.*

Proof For all rows $i \neq n$, $\Delta(i, n) = \{i\}$. Thus the counterexamples presented to the learner are chosen deterministically. Moreover, the only row eliminated when column i is returned as a counterexample is row i, because $C_{ii} = 1$ while $C_{ji} = 0$ for all $j \neq i$. The learning task reduces to randomly sampling a search space without replacement in search of a target element, which requires an expected $\Omega(n)$ equivalence queries with random counterexamples. ■

Note that if the learner had proposed the row of all zeroes as a hypothesis row, this would have identified any target row in just one equivalence query. However, it is not obvious that there will *always* exist a suitable hypothesis row. In Section 4, we show that there will always be a hypothesis row that eliminates at least half of the remaining consistent hypotheses in expectation.

3.3. A lower bound

We now show a worst-case lower bound for any randomized learning algorithm of $\lfloor \log n \rfloor - 1$ on the expected number of equivalence queries with random counterexamples, where n is the number of rows in the concept space. We first define the Complete Concept Space, which contains all subsets of examples.

Definition 11 *Let m be a positive integer and let $n = 2^m$. The **Complete Concept Space** of size n is a matrix $C_n \in \mathcal{M}_{n \times m}$. The rows of C_n are the set of all unique binary vectors of length m in increasing order when considered as binary representations of integers.*

For example,

$$C_4 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Theorem 12 *Let m be a positive integer and let $n = 2^m$. Any randomized learning algorithm that exactly identifies any target concept from the Complete Concept Space C_n with respect to the uniform distribution on examples requires at least an expected $m - 1$ equivalence queries with random counterexamples for some target concept.*

Proof Let A be any deterministic learning algorithm for C_n, and X_m be the expected number of equivalence queries with random counterexamples made by A in learning C_n, where $n = 2^m$ and the target concept is chosen uniformly at random from C_n. By Yao's minimax principle, it suffices to show that X_m is bounded below by $m - 1$.

Clearly $X_1 \geq 1$, because C_2 has two concepts. Because the distribution on target concepts is uniform, the first query made by A on C_n has a probability of $1/2^m$ of being the target concept. Otherwise, the counterexample is a random column, which eliminates exactly half of C_n. The remaining concept space (with the constant column eliminated) is isomorphic to $C_{n'}$ for $n' = 2^{m-1}$. The conditional distribution of the target concept is uniform in the remaining space. Thus, for $m > 1$,

$$X_m \geq 1 + \left(1 - \frac{1}{2^m}\right) X_{m-1}.$$

Considering the recurrence $Y_1 = 1$ and $Y_m = 1 + (1 - (1/2^m))X_{m-1}$, we have

$$\begin{aligned} Y_m &= 1 + \sum_{j=2}^{m} \prod_{k=j}^{m} \left(1 - \frac{1}{2^k}\right) \\ &\geq 1 + \sum_{j=2}^{m} \left(1 - \sum_{k=j}^{m} \frac{1}{2^k}\right) \\ &\geq m - \sum_{j=2}^{m} \frac{1}{2^{j-1}} \\ &> m - 1. \end{aligned}$$

Because $X_m \geq Y_m$ for all $m \geq 1$, the result follows. ■

4. Elimination graphs

We define deficient ℓ-cycles and use them to prove results for the elimination graph $G_{\text{elim}}(C, \pi)$. Theorem 16 shows that there are no deficient 3-cycles in any elimination graph. Theorem 17 generalizes Theorem 16 to show that there are no deficient ℓ-cycles in any elimination graph for any $\ell \geq 2$. These general results on $G_{\text{elim}}(C, \pi)$ are used in Theorem 18 to show that in any concept space with any example distribution, there exists a $\frac{1}{2}$-informative row.

Lemma 13 *For any concept space C, any example distribution π and any pair $i \neq j$ in $[n]$,*

$$E(i, j) + E(j, i) = 1.$$

Proof From the definition of $E(i, j)$,

$$E(i, j) + E(j, i) = \sum_{k \in \Delta(i,j)} \mathbb{P}(K_{i,j} = k) \cdot V(i, k) + \sum_{k \in \Delta(j,i)} \mathbb{P}(K_{i,j} = k) \cdot V(j, k).$$

By the definition of Symmetric Row-Difference, $\Delta(i, j) = \Delta(j, i)$. Therefore,

$$E(i, j) + E(j, i) = \sum_{k \in \Delta(i,j)} \mathbb{P}(K_{i,j} = k) \cdot \big(V(i, k) + V(j, k)\big).$$

By Lemma 6, $V(i,k) + V(j,k) = 1$ for $k \in \Delta(i,j)$. Therefore,

$$\begin{aligned} E(i,j) + E(j,i) &= \sum_{k \in \Delta(i,j)} \mathbb{P}(K_{i,j} = k) \\ &= 1. \end{aligned}$$

■

Definition 14 *Let $G = (V, E)$ be a weighted directed graph and $\ell > 1$ an integer. A **deficient ℓ-cycle** in G is defined as a sequence $v_0, \ldots, v_{\ell-1}$ of distinct vertices such that for all $i \in [\ell]$, $E(v_i, v_{(i+1) \pmod \ell}) \leq \frac{1}{2}$, with strict inequality for at least one $i \in [\ell]$.*

Corollary 15 *For any concept space C and any counterexample distribution π, there are no deficient 2-cycles in $G_{elim}(C, \pi)$.*

Proof Suppose by way of contradiction that there exists a deficient 2-cycle i_0, i_1. Then $i_0 \neq i_1$, $E(i_0, i_1) \leq \frac{1}{2}$, and $E(i_1, i_0) \leq \frac{1}{2}$, and at least one of these inequalities is strict. Then $E(i_0, i_1) + E(i_1, i_0) < 1$. This contradicts Lemma 13. ■

The proof that there are no deficient 3-cycles is somewhat more involved, but will enable an inductive proof that there are no deficient ℓ-cycles for any $\ell \geq 2$.

Theorem 16 *For any concept space $C \in \mathcal{M}_{n \times m}$ and any counterexample distribution π, there are no deficient 3-cycles in $G_{elim}(C, \pi)$.*

Proof Assume for the sake of contradiction that there is a deficient 3-cycle i_0, i_1, i_2 in $G_{\text{elim}}(C, \pi)$. Then $E(i_0, i_1)$, $E(i_1, i_2)$ and $E(i_2, i_0)$ are all less than or equal to $\frac{1}{2}$, with at least one of them strictly less than $\frac{1}{2}$. By the definition of $E(i_0, i_1)$,

$$\sum_{k \in \Delta(i_0, i_1)} \mathbb{P}(K_{i_0,i_1} = k) \cdot V(i_0, k) \leq \frac{1}{2}.$$

Expanding the definition of K_{i_0,i_1},

$$\sum_{k \in \Delta(i_0, i_1)} \left(\frac{\pi(k)}{\sum_{\ell \in \Delta(i_0,i_1)} \pi(\ell)} \right) \cdot V(i_0, k) \leq \frac{1}{2}.$$

Multiplying through by $\sum_{\ell \in \Delta(i_0,i_1)} \pi(\ell)$ and changing the bound variable ℓ to k,

$$\sum_{k \in \Delta(i_0, i_1)} \pi(k) \cdot V(i_0, k) \leq \frac{1}{2} \sum_{k \in \Delta(i_0, i_1)} \pi(k). \tag{2}$$

We partition the set of columns k in $\Delta(i_0, i_1)$ into two sets depending on the whether i_2 agrees with i_0 or i_1 in column k. Define $D(0, 2)$ to be the set of columns k such that the values of i_0 and i_2 agree in column k, and the value of i_1 is different. Similarly, define $D(1, 2)$ to be the set of columns

k such that the values of i_1 and i_2 agree in column k and the value of i_0 is different. These two sets are disjoint and

$$\Delta(i_0, i_1) = D(0,2) \cup D(1,2).$$

Similarly, we can define $D(0,1)$ to be the set of columns k such that the values of i_0 and i_1 agree in column k and the value of i_2 is different. Then $D(0,1)$ is disjoint from both $D(0,2)$ and $D(1,2)$ and we have

$$\Delta(i_1, i_2) = D(0,1) \cup D(0,2)$$

and

$$\Delta(i_2, i_0) = D(1,2) \cup D(0,1).$$

Using the partition of $\Delta(i_0, i_1)$ and equation (2), we have

$$\sum_{k \in D(0,2)} \pi(k) \cdot V(i_0, k) + \sum_{k \in D(1,2)} \pi(k) \cdot V(i_0, k) \leq \frac{1}{2}\Big(\sum_{k \in D(0,2)} \pi(k) + \sum_{k \in D(1,2)} \pi(k) \Big). \quad (3)$$

The analogous inequalities for edges $E(i_1, i_2)$ and $E(i_2, i_0)$ are as follows:

$$\sum_{k \in D(0,1)} \pi(k) \cdot V(i_1, k) + \sum_{k \in D(0,2)} \pi(k) \cdot V(i_1, k) \leq \frac{1}{2}\Big(\sum_{k \in D(0,1)} \pi(k) + \sum_{k \in D(0,2)} \pi(k) \Big), \quad (4)$$

$$\sum_{k \in D(1,2)} \pi(k) \cdot V(i_2, k) + \sum_{k \in D(0,1)} \pi(k) \cdot V(i_2, k) \leq \frac{1}{2}\Big(\sum_{k \in D(1,2)} \pi(k) + \sum_{k \in D(0,1)} \pi(k) \Big). \quad (5)$$

Adding the left hand sides of the inequalities (3, 4, 5) yields the following:

$$\sum_{k \in D(0,2)} \pi(k) \cdot (V(i_0, k) + V(i_1, k)) + \sum_{k \in D(1,2)} \pi(k) \cdot (V(i_0, k) + V(i_2, k)) + \sum_{k \in D(0,1)} \pi(k) \cdot (V(i_1, k) + V(i_2, k)).$$

By Lemma 6, this sum reduces to the following:

$$\sum_{k \in D(0,2)} \pi(k) + \sum_{k \in D(1,2)} \pi(k) + \sum_{k \in D(0,1)} \pi(k).$$

Adding the right hand sides of the inequalities (3, 4, 5) yields the following:

$$\frac{1}{2} \cdot 2\Big(\sum_{k \in D(0,2)} \pi(k) + \sum_{k \in D(1,2)} \pi(k) + \sum_{k \in D(0,1)} \pi(k) \Big).$$

Because at least one of the inequalities (3, 4, 5) is strict, the sum of the left hand sides is strictly less than the sum of the right hand sides, that is,

$$\sum_{k \in D(0,2)} \pi(k) + \sum_{k \in D(1,2)} \pi(k) + \sum_{k \in D(0,1)} \pi(k) < \sum_{k \in D(0,2)} \pi(k) + \sum_{k \in D(1,2)} \pi(k) + \sum_{k \in D(0,1)} \pi(k).$$

This contradiction concludes the proof. ■

Theorem 17 *For any concept space C, any example distribution π, and any integer $\ell \geq 2$, there are no deficient ℓ-cycles in $G_{elim}(C, \pi)$.*

Proof We proceed by induction, where the predicate $P(\ell)$ indicates that there exist no deficient ℓ-cycles in $G_{\text{elim}}(C, \pi)$. Corollary 15 and Theorem 16 imply $P(2)$ and $P(3)$. We will now show that $P(\ell) \rightarrow P(\ell + 1)$ for all $\ell \geq 3$. Suppose there exists a deficient $(\ell + 1)$-cycle: $i_0, i_1, \ldots, i_\ell$. Rotate the cycle as necessary so that $E(i_0, i_1) < \frac{1}{2}$ and the rest of the edges have weights less than or equal to $\frac{1}{2}$. Consider $E(i_2, i_0)$. If $E(i_2, i_0) \leq \frac{1}{2}$, then i_0, i_1, i_2 is a deficient 3-cycle, which cannot exist by $P(3)$. Otherwise, $E(i_2, i_0) > \frac{1}{2}$ and therefore $E(i_0, i_2) < \frac{1}{2}$ by Lemma 13. Then $i_0, i_2, i_3, \ldots, i_\ell$ is a deficient ℓ-cycle, which cannot exist by the inductive hypothesis $P(\ell)$. ■

Theorem 18 *For any concept space C and any example distribution π there exists at least one $\frac{1}{2}$-informative row in $G_{elim}(C, \pi)$.*

Proof Suppose by way of contradiction that there does not exist a row i such that $E(i, j) \geq \frac{1}{2}$ for all $j \neq i$. Then for every row i, there necessarily exists a row $j \neq i$ such that $E(i, j) < \frac{1}{2}$. Let $\phi(i)$ map each i to such a j. Define $\phi^d(i) = \phi(\phi^{d-1}(i))$ for each positive integer d, with $\phi^0(i) = i$. Beginning at an arbitrary row i_0, consider the sequence $i_0, \phi(i_0), \ldots, \phi^n(i_0)$, where n is the number of rows in C. By the Pigeonhole Principle, there exist distinct $a, b \in [n]$ such that $\phi^a(i_0) = \phi^b(i_0)$. Without loss of generality, take $a < b$. Because $\phi(i) \neq i$, $b \neq a + 1$, so $b - a \geq 2$. The subsequence $\phi^a(i_0), \phi^{a+1}(i_0), \ldots, \phi^{b-1}(i_0)$ is a deficient $(b - a)$-cycle, a contradiction by Theorem 17. ■

5. The Max-Min learning algorithm

We now show that for any target row in any concept space with n rows and any example distribution, if the learner always queries a $\frac{1}{2}$-informative row in the space of consistent rows, then the target row will be identified in $O(\log n)$ queries with high probability.

Definition 19 *At any point in the execution of a learning algorithm, the **learned information** is the set I of all pairs (k, b), where $k \in [m]$ is a column that has been returned as a counterexample to an equivalence query and $b \in \{0, 1\}$ is the value of the target row in column k.*

Definition 20 *Given concept space C and a set I of learned information, the set of **consistent rows**, denoted cons(C, I), is the set of rows i such that for each pair $(k, b) \in I$, $C_{ik} = b$.*

Define the Max-Min learning algorithm to be the strategy in which, at each opportunity to pose an equivalence query, a hypothesis row h is chosen such that

$$h = \underset{i \in \text{cons}(C,I)}{\arg\max} \left(\min_{j \in \text{cons}(C,I) \setminus \{i\}} E(i, j) \right). \tag{6}$$

By Theorem 18, $E(h, j) \geq \frac{1}{2}$ for all possible target rows $j \neq h$.

For any positive integer n, define $T(n)$ to be the maximum, over all concept spaces C of n rows, all example distributions π, and all possible target concepts t, of the expected number of equivalence queries with random counterexamples used by the Max-Min learning algorithm to identify the target concept t in C with respect to π.

Theorem 21 *For all positive integers n, $T(n) \leq \log n$.*

Proof We proceed by strong induction. Define $P(n)$ to be the predicate: $T(n) \leq \log n$. Then $P(1)$ is true because no queries are required in a trivial concept space. Moreover, $P(2)$ is true because only one query is necessary in any two-row concept space. Assume that $P(r)$ is true for all positive integers $r \leq n$ for some $n \geq 2$. Consider any concept space C with $n+1$ rows, any example distribution π, and any target concept t.

Define R to be the number of remaining consistent rows after the Max-Min algorithm makes one equivalence query with random counterexamples for row h, where h satisfies Equation (6). Then R is a random variable dependent on the teacher's random choice of counterexample. By the definition of h and Theorem 18, we know that $\mathbb{E}[R] \leq \frac{n+1}{2}$.

After the equivalence query the problem is reduced to learning the target row t in a concept space consisting of the remaining r consistent rows, where $1 \leq r \leq n$. The expected number of queries for this task is bounded above by $T(r)$, so we have the following:

$$T(n+1) \leq 1 + \sum_{r=1}^{n} \Big(\mathbb{P}[R = r] \cdot T(r) \Big).$$

Applying the inductive hypothesis,

$$T(n+1) \leq 1 + \sum_{r=1}^{n} \Big(\mathbb{P}[R = r] \cdot \log r \Big).$$

Applying Jensen's Inequality,

$$T(n+1) \leq 1 + \log \left(\mathbb{E}[R] \right).$$

Using the fact that $\mathbb{E}[R] \leq \frac{(n+1)}{2}$,

$$T(n+1) \leq 1 + \log \left(\frac{(n+1)}{2} \right) = \log \left(n+1 \right).$$

This concludes the inductive step. ■

Lemma 22 *Let C be any concept space with n rows, π any example distribution and t any target concept. Define R_ℓ to be the number of consistent rows remaining after ℓ equivalence queries, with queries chosen according to the Max-Min learning algorithm. Then $\mathbb{E}[R_\ell] \leq \frac{n}{2^\ell}$*

Proof By Theorem 18, we know that $\mathbb{E}[R_1] \leq \frac{n}{2}$. We use induction on ℓ:

$$\mathbb{E}[R_{\ell+1}] = \mathbb{E}[\mathbb{E}[R_{\ell+1} \mid R_\ell]] \leq \frac{\mathbb{E}[R_\ell]}{2} \leq \frac{n}{2^{\ell+1}}.$$

■

Theorem 23 *Let $\delta \in \mathbb{R} : 0 < \delta < 1$. Then, with probability at least $1 - \delta$, the Max-Min learning algorithm terminates in at most $O(\log(n\delta^{-1}))$ equivalence queries.*

Proof After ℓ equivalence queries, the identity of the target row is unknown only if there are at least two rows that are consistent with the answers to queries already asked by the learner, that is, $R_\ell \geq 2$. Applying Markov's inequality to R_ℓ, a non-negative random variable, we have

$$\mathbb{P}(R_\ell \geq 2) \leq \frac{\mathbb{E}[R_\ell]}{2} \leq \frac{n}{2^{\ell+1}}. \tag{7}$$

It suffices to take $\ell \geq \log(n\delta^{-1})$ to ensure that $\mathbb{P}(R_\ell < 2)$ is at least $1 - \delta$. ■

We note that the Max-Min learning algorithm does not necessarily achieve the optimal worst case expected value for every concept space. In particular, for the concept space in equation (1) with a uniform distribution over examples, the initial choice of the Max-Min algorithm could be the first row or the last row (which lead to bounds of 2.25 or 2.125, depending on how ties are broken), while there is a strategy that first queries the second row and then identifies any other target using just one more query, for a bound of 2 queries.

6. Comparison with the VC-dimension

As described in Section 2, several important quantities, for example, the number of examples required for PAC learning, the optimal error bound for predicting $\{0, 1\}$-sequences and the recursive teaching dimension, are closely related to the VC-dimension of a finite concept class. In this section we show that the VC-dimension gives a lower bound for the worst-case expected number of proper equivalence queries used by the Max-Min algorithm, but not an upper bound.

Definition 24 *Given a concept space C, a set $S = \{k_0, k_1, \ldots, k_{d-1}\}$ of d columns is **shattered** by C if and only if for every d-tuple $(b_0, b_1, \ldots, b_{d-1})$ of boolean values, there exists a row i such that $C_{i,k_l} = b_l$ for all $\ell \in [d]$. The VC-dimension of C is the largest d such that there exists a set of d columns shattered by C.*

If C has n rows then its VC-dimension is at most $\log n$. For all $n \geq 2$, the Identity Concept Space I_n has VC-dimension 1 and the Complete Concept Space C_n has VC-dimension $\log n$.

Theorem 25 *If C is a concept class of VC-dimension d, then any randomized learning algorithm to learn C must use at least an expected $\Omega(d)$ equivalence queries with random counterexamples for some target concept.*

Proof If C is a concept class of VC-dimension d, then we consider a shattered set S of d columns and a probability distribution on examples that is uniform on S and negligibly small elsewhere. This is essentially the situation of a Complete Concept Space with 2^d rows and the uniform distribution on examples. By Theorem 12, every randomized learning algorithm that exactly identifies any concept from C using equivalence queries with random counterexamples must use at least an expected $\Omega(d)$ queries in the worst case. ■

In the other direction, we show that there is a family of concept spaces such that the VC-dimension of each concept space in the family is 1, but the worst-case expected number of equivalence queries used by the Max-Min algorithm grows as $\Omega(\log n)$, where n is the number of rows

in the concept space. This example was given by Littlestone (1988), who showed that the Standard Optimal Algorithm makes $\Omega(\log n)$ mistakes of prediction in the worst case.

Theorem 26 *There exists a family of concept spaces D_n such that for each $n \geq 2$, D_n has n rows and VC-dimension 1, but on D_n the Max-Min algorithm uses an expected $\Omega(\log n)$ equivalence queries with random counterexamples in the worst case.*

Proof The class D_n contains n rows and $n-1$ columns. Row i consists of i 1's followed by $(n-1-i)$ 0's. Thus, row zero consists of $n-1$ 0's. We first show that for all $n \geq 2$, the VC-dimension of D_n is 1. The set containing just the first column is shattered (labeled 0 by row 0 and labeled 1 by row 1). However, no larger set of columns can be shattered. If $k_1 < k_2$ then in any row in which column k_1 is 0, column k_2 must also be 0.

To see that the Max-Min algorithm uses an expected $\Omega(\log n)$ equivalence queries in the worst case, we analyze the case when the target concept is row zero. With concept class D_n, the Max-Min algorithm will choose either the unique middle row (if n is odd) or one of the two middle rows (if n is even); for concreteness, assume that the row with smaller index is chosen. Then in either case, row $r(n) = \lfloor (n-1)/2 \rfloor$ is chosen. One of the $r(n)$ columns in which row $r(n)$ is 1 will be chosen as a counterexample, say k, and then the problem is reduced to learning the zero row in the reduced concept space D_k. This analysis leads to a recurrence $f(1) = 0$, $f(2) = 1$ and for $n \geq 3$,

$$f(n) = 1 + \frac{1}{r(n)} \sum_{k=1}^{r(n)} f(k),$$

where $f(n)$ is the expected number of equivalence queries for the Max-Min algorithm to learn the zero row in D_n. An inductive proof shows that $f(n) \geq c \ln n$ for some $c \geq 0.5$, establishing the lower bound. ■

7. Discussion

We have introduced a new setting for exact learning with proper equivalence queries in which the teacher returns a counterexample drawn from a known distribution on examples. We have proposed the Max-Min learning algorithm and shown that it uses at most an expected $\log_2 n$ proper equivalence queries with random counterexamples to identify any target concept in any concept space with n rows, with respect to any distribution on examples.

One intriguing open question is whether a much simpler learning algorithm might achieve the same bound if we consider expected performance with respect to a randomly drawn target concept, instead of a worst case target concept. In particular, we conjecture that a learning algorithm that hypothesizes a randomly chosen consistent row uses an expected $O(\log n)$ equivalence queries with random counterexamples to identify a randomly drawn target concept in a concept space of n rows.

8. Acknowledgements

The authors are grateful to the ALT reviewers for their careful attention to the paper. The bound in Theorem 12 was improved with the help of one of the reviewers and Stan Eisenstat.

References

Dana Angluin. Learning regular sets from queries and counterexamples. *Inf. Comput.*, 75(2):87–106, November 1987.

Dana Angluin. Queries and concept learning. *Mach. Learn.*, 2(4):319–342, 1988.

Dana Angluin. Negative results for equivalence queries. *Mach. Learn.*, 5(2):121–150, July 1990.

Dana Angluin. Queries revisited. *Theor. Comput. Sci.*, 313(2):175–194, February 2004.

Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, October 1989.

E Mark Gold. Complexity of automaton identification from given data. *Information and Control*, 37(3):302–320, 1978.

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

D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Inf. Comput.*, 115(2):248–292, 1994.

Lunjia Hu, Ruihan Wu, Tianhong Li, and Liwei Wang. Quadratic upper bound for recursive teaching dimension of finite VC classes. *PMLR*, 65:1147–1156, 2017. Presented at COLT 2017.

Michael Kearns and Leslie Valiant. Cryptographic limitations on learning boolean formulae and finite automata. *J. ACM*, 41(1):67–95, January 1994.

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

Ayumi Shinohara and Satoru Miyano. Teachability in computational learning. *New Generation Computing*, 8(4):337–347, 1991.

Frits Vaandrager. Model learning. *Commun. ACM*, 60(2):86–95, January 2017.

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

Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Models of cooperative teaching and learning. *J. Mach. Learn. Res.*, 12:349–384, February 2011.