Bounds in query learning

Hunter Chase HCHASE2@UIC.EDU and **James Freitag** FREITAGJ@GMAIL.COM
Department of Mathematics, University of Illinois at Chicago, Chicago, IL

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We introduce new combinatorial quantities for concept classes, and prove lower and upper bounds for learning complexity in several models of learning in terms of various combinatorial quantities. In the setting of equivalence plus membership queries, we give an algorithm which learns a class in polynomially many queries whenever any such algorithm exists. Our approach is flexible and powerful enough to give new and very short proofs of the efficient learnability of several prominent examples (e.g. regular languages and regular ω-languages), in some cases also producing new bounds on the number of queries.

Keywords: Equivalence query learning, proper equivalence queries, membership queries, exact concept learning, regular languages, omega-regular languages, finite automata, model theory

1. Introduction

Fix a set X and denote by $\mathcal{P}(X)$ the collection of all subsets of X. A *concept class* $\mathcal{C}$ on X is a subset of $\mathcal{P}(X)$. Neither X nor $\mathcal{C}$ are assumed to be finite, though this case is of particular interest. In the equivalence query (EQ) learning model, a learner attempts to identify a target set $A \in \mathcal{C}$ by means of a series of data requests called *equivalence queries*. The learner has full knowledge of $\mathcal{C}$, as well as a hypothesis class $\mathcal{H}$ with $\mathcal{C} \subseteq \mathcal{H} \subseteq \mathcal{P}(X)$. An *equivalence query* consists of the learner submitting a hypothesis $B \in \mathcal{H}$ to a teacher, who either returns *yes* if $A = B$, or a counterexample $x \in A \triangle B$. In the former case, the learner has succeeded, and in the latter case, the learner uses the new information to update and submit a new hypothesis. In this manuscript, we are interested in the worst case number of required queries. We will also consider learning with equivalence and membership queries (EQ+MQ). In a membership query, a learner submits a single element x from the base set X to the teacher, who returns the value $A(x)$, where A is the target concept. In this setting, the learner may choose to make either type of query at any stage, submitting any $x \in X$ for a membership query or submitting any $B \in \mathcal{H}$ for an equivalence query. The learner succeeds when they submit the the target concept A as an equivalence query.

In addition to applications, we consider several fundamental problems in these settings:

1. Give a characterization in terms of some simple combinatorial quantities of $(\mathcal{C}, \mathcal{H})$ for when there is a bound on the number of required queries in the EQ or EQ+MQ model.
2. Determine simple combinatorial quantities in the EQ and EQ+MQ models which characterize *efficient* learnability—that is, learnability in a polynomial number of queries.
3. Given a class $\mathcal{C}$ for which learning is possible when $\mathcal{H} = \mathcal{P}(X)$, determine the class $\mathcal{H}$ of minimal complexity which makes this possible, if one exists.

Versions of problems 1) and 2) have been considered in a variety of models of learning; for instance, finite Littlestone dimension characterizes learnability in online learning Littlestone (1988) Ben-David et al. (2009) and finite VC-dimension characterizes learnability in the PAC model Blumer et al. (1989). More recently, Alon et al. (2019) show that finite Littlestone dimension is required for approximately differentially private learning (though the converse is open). Our main result, Theorem 2.24, gives a complete characterization of when the classes $(\mathcal{C}, \mathcal{H})$ can be *efficiently* learned in the EQ+MQ model in terms of simple combinatorial quantities associated with the classes, answering the problem 2) in that model. This result involves establishing several new upper and lower bounds for learning complexity in terms of our combinatorial quantities. These bounds, which we describe next, turn out to be sufficient to answer the problem 1) in both the EQ and EQ+MQ models.

With Theorems 2.6 and 2.24, we give upper bounds for the number of queries required for EQ and EQ+MQ learning a class $\mathcal{C}$ with hypotheses $\mathcal{H}$ in terms of the *Littlestone dimension of* $\mathcal{C}$, denoted $\mathrm{Ldim}(\mathcal{C})$, and the *consistency dimension of* $\mathcal{C}$ *with respect to* $\mathcal{H}$, denoted $\mathrm{C}(\mathcal{C}, \mathcal{H})$. We also give lower bounds for the number of required queries in terms of these quantities. Littlestone dimension is well-known in learning theory (Littlestone, 1988) and model theory.[1]

Consistency dimension is a more subtle invariant, which we detail in section 2. When $\mathcal{H}$ is taken to be the power set $\mathcal{P}(X)$, $\mathrm{C}(\mathcal{C}, \mathcal{H}) = 1$. For various examples of set systems with $\mathcal{H} = \mathcal{C}$, one has $\mathrm{C}(\mathcal{C}, \mathcal{H}) = \infty$. In 2.2, in solving problem 3), we define a new invariant, the consistency threshold of $\mathcal{C}$, and provide a construction (for arbitrary $\mathcal{C}$) of a hypothesis class $\mathcal{H}$ which is not much more complicated than $\mathcal{C}$ (e.g. it is of the same Littlestone dimension as $\mathcal{C}$) such that $\mathrm{C}(\mathcal{C}, \mathcal{H}) \leq \mathrm{Ldim}(\mathcal{C}) + 1$. This provides a complete answer to problem 3) in the EQ+MQ model for both learnability and efficient learnability and for learnability in the EQ model. In 2.3, we compare our bounds and invariants to those previously appearing in the literature.

Consistency dimension has been used to study query learning, but had not been previously used in conjunction with Littlestone dimension. In the EQ+MQ setting, Theorem 2.24 considers both together and gives an upper bound of $\mathrm{C}(\mathcal{C}, \mathcal{H})\,\mathrm{Ldim}(\mathcal{C})$ on the number of queries, improving the upper bound of $\lceil \mathrm{C}(\mathcal{C}, \mathcal{H}) \log_2 |\mathcal{C}| \rceil$ in Hellerstein et al. (1996) and Balcázar et al. (2002) and generalizing to infinite classes. Moreover, together with appropriate lower bounds, Theorem 2.24 also identifies consistency dimension and Littlestone dimension as the relevant quantities in classifying efficient learnability in this setting.

In section 3 we demonstrate the practicality of our results by providing simple and fast proofs of the efficient learnability of regular languages and certain ω-languages, reproving results of Angluin (1987); Angluin and Fisman (2016); Fisman et al. (2018); Fisman (2018). Besides the conceptual simplicity of the approach, the bounds in learning complexity resulting from our algorithm have some novel aspects. For instance, our bounds have no dependence on the length of the strings provided to the learner as counterexamples, in contrast to existing algorithms.

2. A combinatorial characterization of EQ-learnability

Often, one assumes that X is finite, and the emphasis is placed on finding bounds on the number of queries it may take to learn any $A \in \mathcal{C}$. We also consider the case where X is infinite, for which we give the following definition.

1. In model theory, Littlestone dimension is called Shelah 2-rank, see Chase and Freitag (2019) for additional details.

Definition 2.1 *Let $\mathcal{C}$ and $\mathcal{H}$ be set systems on a set X. $\mathcal{C}$ is* learnable with equivalence queries *from $\mathcal{H}$ if there exists some algorithm for the learner to submit hypotheses from $\mathcal{H}$ and some $n < \omega$ such that any concept $A \in \mathcal{C}$ is learnable in at most n equivalence queries, given any teacher returning counterexamples. Let $\mathrm{LC}^{EQ}(\mathcal{C}, \mathcal{H})$ be the least such n if $\mathcal{C}$ is learnable with equivalence queries from $\mathcal{H}$, and $\mathrm{LC}^{EQ}(\mathcal{C}, \mathcal{H}) = \infty$ otherwise.*

$\mathrm{LC}^{EQ}(\mathcal{C}, \mathcal{H})$ is called the learning complexity*, representing the optimal number of queries needed in the worst-case scenario.*

Learning complexity for learning a concept class $\mathcal{C}$ with membership queries from the base set X or equivalence queries from the hypothesis class $\mathcal{H}$ is defined in the same manner and is denoted by $\mathrm{LC}^{EQ+MQ}(\mathcal{C}, \mathcal{H})$.

2.1. EQ-learnability from Littlestone and consistency dimension

The first key property is the Littlestone dimension of $\mathcal{C}$, denoted $\mathrm{Ldim}(\mathcal{C})$.[2] Its relevance to query learning was identified by Littlestone himself.

Proposition 2.2 *(Littlestone, 1988, Theorems 5 and 6) If $\mathrm{LC}^{EQ}(\mathcal{C}, \mathcal{H}) \leq d+1$, then $\mathrm{Ldim}(\mathcal{C}) \leq d$. If $\mathcal{H} = \mathcal{P}(X)$, then the converse holds.*

Notice in particular that if $\mathrm{Ldim}(\mathcal{C}) = \infty$, then $\mathcal{C}$ cannot be learned with equivalence queries, even with $\mathcal{H} = \mathcal{P}(X)$. The assumption that $\mathcal{H} = \mathcal{P}(X)$ makes learning straightforward, but this may be too strong for many settings. However, without some additional hypotheses on $\mathcal{H}$, learnability may already be hopeless, even for *very simple* set systems. For instance, let $\mathcal{C}$ be the set of singletons of the set X. If $\mathcal{H} = \mathcal{C}$, then we may take as long as $|X|$ to learn if X is finite, or never learn at all if X is infinite. However, if the learner is allowed to guess $\emptyset$, this forces the teacher to identify the target singleton immediately.

The strategy of Proposition 2.2 permeates both learnability and non-learnability proofs; identifying a specific set amounts to reducing the Littlestone dimension of the family of possible concepts to 0; actually submitting the target concept before the Littlestone dimension reaches 0 can be thought of as a best-case scenario that we cannot rely on. Non-learnability then amounts to an inability to reduce the Littlestone dimension of the family of possible concepts to 0 through a series of finitely many equivalence queries. The main purpose of this section is to give precise conditions on $\mathcal{H}$ and $\mathcal{C}$ which *characterize* learnability.

Definition 2.3 *Given a set X, a* partially specified subset *A of X is a partial function $A : X \to \{0, 1\}$.*

- *Say $x \in A$ if $A(x) = 1$, $x \notin A$ if $A(X) = 0$, and membership of x is unspecified otherwise. The* domain *of A, $\mathrm{dom}(A)$, is $A^{-1}(\{0, 1\})$. Call A* total *if $\mathrm{dom}(A) = X$. We identify subsets $A \subseteq X$ with total partially specified subsets. The* size *of A, $|A|$, is the cardinality of $\mathrm{dom}(A)$.*
- *Given two partially specified subsets A and B, write $A \preceq B$ if A and B agree on $\mathrm{dom}(A)$; call A a* restriction *of B and B an* extension *of A.*
- *Given a set $Y \subseteq \mathrm{dom}(A)$, the restriction $A|_Y$ to A to Y is the partial function where $A|_Y(x) = A(x)$ for all $x \in Y$, and is unspecified otherwise.*

2. A definition of Littlestone dimension appears in the appendices.

- *Given a set system $\mathcal{C}$ on X, A is* n-consistent *with $\mathcal{C}$ if every size n restriction of A has an extension in $\mathcal{C}$. Otherwise, say A is* n-inconsistent.[3] *A is* finitely consistent *with $\mathcal{C}$ if every restriction of A of finite size has an extension in $\mathcal{C}$—that is, A is n-consistent with $\mathcal{C}$ for all $n < \omega$.*

- *Given a set system $\mathcal{C}$ on X, $x \in X$, and $j \in \{0,1\}$, let $\mathcal{C}^{(x,j)} = \{A \in \mathcal{C} \mid A(x) = j\}$. That is, $\mathcal{C}^{(x,0)} = \{A \in \mathcal{C} \mid x \notin A\}$ and $\mathcal{C}^{(x,1)} = \{A \in \mathcal{C} \mid x \in A\}$*

The following definition is a translation into set systems of a definition that first appeared in Balcázar et al. (2002).

Definition 2.4 *The* consistency dimension *of $\mathcal{C}$ with respect to $\mathcal{H}$, denoted $\mathrm{C}(\mathcal{C},\mathcal{H})$, is the least integer n such that for every subset $A \subseteq X$ (viewed as a total partially specified subset), if A is n-consistent with $\mathcal{C}$, then $A \in \mathcal{H}$. If no such n exists, then say $\mathrm{C}(\mathcal{C},\mathcal{H}) = \infty$.*

Observe that $\mathrm{C}(\mathcal{C},\mathcal{H}) = 1$ iff $\mathcal{H}$ shatters[4] the set of all elements $x \in X$ such that there are A_0 and A_1 in $\mathcal{C}$ such that $x \notin A_0$ but $x \in A_1$. In this case, it is possible to learn any concept in $\mathcal{C}$ in at most $\mathrm{Ldim}(\mathcal{C}) + 1$ equivalence queries, using the method of Proposition 2.2. So we may assume that $\mathrm{C}(\mathcal{C},\mathcal{H}) > 1$.

The following simple but useful lemma states that the number of queries needed to learn a finite union of classes is at most the sum of the number of queries needed to learn each class on its own.

Lemma 2.5 *Suppose that for each $i < n$, $\mathcal{C}_i$ is a concept class on X and $\mathcal{H}$ is a hypothesis class on X. Suppose that $\mathrm{LC}^{EQ}(\mathcal{C}_i,\mathcal{H}_i) = m_i$. Then $\mathrm{LC}^{EQ}(\mathcal{C},\mathcal{H}) \leq \sum_{i<n} m_i$, where $\mathcal{C} := \cup_{i<n}\mathcal{C}_i$ and $\mathcal{H} := \cup_{i<n}\mathcal{H}_i$.*

We can now give an upper bound for the learning complexity in terms of Littlestone dimension and consistency dimension.

Theorem 2.6 *Suppose $\mathrm{Ldim}(\mathcal{C}) = d < \infty$ and $1 < \mathrm{C}(\mathcal{C},\mathcal{H}) = c < \infty$. Then $\mathrm{LC}^{EQ}(\mathcal{C},\mathcal{H}) \leq c^d$.*

Proof We proceed by induction on d. The base case, $d = 0$, is trivial, as then $\mathcal{C}$ is a singleton.

Suppose $\mathrm{Ldim}(\mathcal{C}) = d + 1$. Suppose there is some element x such that $\mathrm{Ldim}(\mathcal{C}^{(x,0)}) < d + 1$ and $\mathrm{Ldim}(\mathcal{C}^{(x,1)}) < d + 1$. Then by induction, any concept in $\mathcal{C}^{(x,0)}$ can be learned in at most c^d queries with guesses from $\mathcal{H}$, and the same is true for $\mathcal{C}^{(x,1)}$. Then by Lemma 2.5, any concept in $\mathcal{C}$ can be learned in at most $2c^d \leq c^{d+1}$ equivalence queries.

If no such x exists, then for all x, either $\mathrm{Ldim}(\mathcal{C}^{(x,0)}) = d + 1$ or $\mathrm{Ldim}(\mathcal{C}^{(x,1)}) = d + 1$. Let B be such that $x \in B$ iff $\mathrm{Ldim}(\mathcal{C}^{(x,1)}) = d + 1$.

If $B \in \mathcal{H}$, then we submit B as our query. If we are incorrect, then by choice of B, the class $\mathcal{C}'$ of concepts consistent with the counterexample x_0 will have Littlestone dimension $\leq d$. By induction, any concept in $\mathcal{C}'$ can be learned in at most c^d many queries, and so we learn a in at most $c^d + 1 \leq c^{d+1}$ queries.

3. We emphasize that, in this context, being n-inconsistent means only that there is some size n restriction that has no extension in $\mathcal{C}$. We do not mean that *all* size n restrictions have no extension in $\mathcal{C}$.
4. Recall that a set system $\mathcal{C}$ shatters a set A if, for all $B \subseteq A$, there is $C \in \mathcal{C}$ such that $C \cap A = B$.

If $B \notin H$, then, since $\mathrm{C}(\mathcal{C}, \mathcal{H}) = c$, there are some $x_0, \ldots, x_{c-1}$ such that there is no $A \in \mathcal{C}$ such that $B|_{\{x_0,\ldots,x_{c-1}\}} \preceq A$. Then

$$\mathcal{C} = (\mathcal{C}^{(x_0,1-B(x_0))}) \cup \ldots \cup (\mathcal{C}^{(x_{c-1},1-B(x_{c-1}))}),$$

and $\mathrm{Ldim}(\mathcal{C}^{(x_i,1-B(x_i))}) \leq d$ for each i. Then, by induction, for each i, any concept in $\mathcal{C}^{(x_i,1-B(x_i))}$ can be learned in at most c^d many queries with guesses from $\mathcal{H}$. By Lemma 2.5, any concept in $\mathcal{C}$ can be learned in at most c^{d+1} many queries with guesses from $\mathcal{H}$. ■

On the other hand, Proposition 2.2 gives a lower bound of $\mathrm{Ldim}(\mathcal{C}) + 1 \leq \mathrm{LC}^{EQ}(\mathcal{C}, \mathcal{H})$. There is also a known lower bound for learning complexity in terms of consistency dimension:

Proposition 2.7 *(Balcázar et al., 2002, Theorem 2) Suppose there is some partially specified subset A which is n-consistent with $\mathcal{C}$ but which does not have a total extension in $\mathcal{H}$. Then $n < \mathrm{LC}^{EQ}(\mathcal{C}, \mathcal{H})$.*

In particular, if $\mathrm{C}(\mathcal{C}, \mathcal{H}) \geq c$, then there is some subset A which is $(c-1)$-consistent with $\mathcal{C}$ but which does not belong to $\mathcal{H}$. Then $c \leq \mathrm{LC}^{EQ}(\mathcal{C}, \mathcal{H})$. So $\mathrm{C}(\mathcal{C}, \mathcal{H}) \leq \mathrm{LC}^{EQ}(\mathcal{C}, \mathcal{H})$. In fact, the proposition is stronger, and we will obtain a stronger bound in the form of strong consistency dimension in section 2.3.

Furthermore, if $\mathrm{C}(\mathcal{C}, \mathcal{H}) = \infty$, then $\mathcal{C}$ cannot be learned with equivalence queries from $\mathcal{H}$. Combining Theorem 2.6 and Propositions 2.2 and 2.7, we obtain the following:

Theorem 2.8 *$\mathcal{C}$ is learnable with equivalence queries from $\mathcal{H}$ iff $\mathrm{Ldim}(\mathcal{C}) < \infty$ and $\mathrm{C}(\mathcal{C}, \mathcal{H}) < \infty$.*

2.2. Obtaining finite consistency dimension

We have established that finite consistency dimension is essential for EQ-learning. The central question we answer in this subsection is: given $\mathcal{C}$, can one obtain a hypothesis class $\mathcal{H}$ which is not much more complicated than $\mathcal{C}$ with the property that $\mathrm{C}(\mathcal{C}, \mathcal{H})$ is finite?

Definition 2.9 *Fix a set system $\mathcal{C}$ on a set X. $\mathcal{C}$ has* consistency threshold $n < \infty$ *if, given any hypothesis class $\mathcal{H} \supseteq \mathcal{C}$, we have that*

$$\mathrm{C}(\mathcal{C}, \mathcal{H}) < \infty \quad \textit{iff} \quad \mathrm{C}(\mathcal{C}, \mathcal{H}) \leq n.$$

Lemma 2.10 *Suppose A is a partially specified subset finitely consistent with $\mathcal{C}$. Then there is a total extension $A' \succeq A$ finitely consistent with $\mathcal{C}$.*

Proposition 2.11 *Let $\mathcal{C}, \mathcal{H}$ be set systems and let A be a partially specified subset. The following are equivalent:*

(i) *A is finitely consistent with $\mathcal{C}$.*

(ii) *If $\mathrm{C}(\mathcal{C}, \mathcal{H}) < \infty$, then there is a total extension $A' \succeq A$ in $\mathcal{H}$.*

In particular, if $\mathrm{C}(\mathcal{C}, \mathcal{H}) < \infty$, then $\mathcal{H}$ contains all finitely consistent subsets. That is, extensions of all finitely consistent partially specified subsets (equivalently, by Lemma 2.10, all finitely consistent total partially specified subsets) are necessary to obtain $\mathrm{C}(\mathcal{C}, \mathcal{H}) < \infty$. Consistency threshold classifies when this is a sufficient condition.

Proposition 2.12 *The following are equivalent:*

(i) *$\mathcal{C}$ has consistency threshold $\leq n < \infty$.*

(ii) *For all (total partially specified) subsets A, if A is n-consistent with $\mathcal{C}$, then A is finitely consistent with $\mathcal{C}$.*

(iii) *If $\mathcal{H}$ contains all finitely consistent (total partially specified) subsets, then $\mathrm{C}(\mathcal{C}, \mathcal{H}) \leq n$.*

In particular, if $\mathcal{C}$ has finite consistency threshold, then $\mathrm{C}(\mathcal{C}, \mathcal{H}) < \infty$ iff $\mathcal{H}$ contains all finitely consistent subsets.

Corollary 2.13 *Suppose $\mathcal{C}$ does not have finite consistency threshold. Then for arbitrarily large n, there is some total subset A_n which is n-consistent but not $(n+1)$-consistent with $\mathcal{C}$.*

Finite consistency threshold is not strictly necessary to provide a positive answer to the central question of this subsection; nevertheless, it does identify a clear qualitative dividing line. When $\mathcal{C}$ has finite consistency threshold, $\mathcal{H}$ only needs to contain all finitely consistent subsets; letting $\mathcal{H}_\infty$ be the set of all finitely consistent subsets, we obtain a minimum hypothesis class such that learning is possible.

Where $\mathcal{C}$ does not have finite consistency threshold, more is required; we must add some hypotheses which are inconsistent with the concepts in $\mathcal{C}$, and there is no minimal $\mathcal{H}$ such that learning is possible. However, for each m, we can replace "finitely consistent" with "m-consistent" to obtain a class $\mathcal{H}_m$ such that $\mathrm{C}(\mathcal{C}, \mathcal{H}_m) \leq m$—let $\mathcal{H}_m$ be the collection of all subsets which are m-consistent with $\mathcal{C}$. Note that $\mathcal{H}_m$ is clearly the minimum hypothesis class such that $\mathrm{C}(\mathcal{C}, \mathcal{H}) \leq m$.

Note that for all m, $\mathcal{H}_\infty \subseteq \mathcal{H}_m$. By Proposition 2.12, if $\mathcal{C}$ has consistency threshold n, then for all $m \geq n$, $\mathcal{H}_m = \mathcal{H}_n = \mathcal{H}_\infty$. If $\mathcal{C}$ does not have finite consistency threshold, there is no minimal $\mathcal{H}$ such that $\mathrm{C}(\mathcal{C}, \mathcal{H}) < \infty$; by Corollary 2.13, if $\mathrm{C}(\mathcal{C}, \mathcal{H}) = m$, then there is $m' \geq m$ such that $\mathcal{H}_{m'} \subsetneq \mathcal{H}$.

By choosing m appropriately, given any $\mathcal{C}$, we can find a hypothesis class such that $\mathrm{C}(\mathcal{C}, \mathcal{H}) < \infty$ without increasing the Littlestone dimension; that is, $\mathrm{Ldim}(\mathcal{H}) = \mathrm{Ldim}(\mathcal{C})$.

Theorem 2.14 *Suppose $\mathrm{Ldim}(\mathcal{C}) = d < \infty$. Then there is $\mathcal{H}$ such that $\mathrm{C}(\mathcal{C}, \mathcal{H}) < \infty$ and $\mathrm{Ldim}(\mathcal{H}) = \mathrm{Ldim}(\mathcal{C})$. Furthermore, we can find such an $\mathcal{H}$ such that $\mathrm{C}(\mathcal{C}, \mathcal{H}) \leq \mathrm{Ldim}(\mathcal{C}) + 1$.*

2.3. From consistency to strong consistency

From an algorithms perspective, the result of Theorem 2.6 is unsatisfactory, since it is exponential in $\mathrm{Ldim}(\mathcal{C})$. We give an example to show that, without modification, we cannot expect a significant improvement.

Example 2.15 *Fix $c > 2$ and d. Let $\{a_\tau \mid \tau \in [c]^i, 1 \leq i \leq d\}$ be distinct elements indexed by finite nonempty sequences of length at most d from $[c]$. For $\sigma \in [c]^d$, let $B_\sigma = \{a_\tau \mid \tau \subseteq \sigma\}$. Let $\mathcal{C} = \{B_\sigma \mid \sigma \in [c]^d\}$. Then $\mathrm{Ldim}(\mathcal{C}) = d$.*

If we take $\mathcal{C}$ to also be our hypothesis class, then $\mathrm{C}(\mathcal{C}, \mathcal{C}) = c + 1$. Indeed, the (total partially specified) subset $A = \{a_0\}$ is c-consistent but not $(c + 1)$ consistent with $\mathcal{C}$, witnessed by the restriction of A to $\{a_0, a_{0,0}, \ldots, a_{0,c-1}\}$, so $\mathrm{C}(\mathcal{C}, \mathcal{C}) \geq c + 1$. On the other hand, if A is a subset $(c + 1)$-consistent with $\mathcal{C}$, then, by induction on the length of τ, for each $1 \leq i \leq d$, A contains exactly one a_τ with $\tau = i$, so $A \in \mathcal{C}$.

However, it may take as long as c^d many equivalence queries to learn; if the teacher returns a_σ as a counterexample to hypothesis A_σ, then the learner can only eliminate A_σ.

The most promising modification is the following variant of consistency dimension, which also appeared in Balcázar et al. (2002) in a slightly different form.

Definition 2.16 *The* strong consistency dimension *of $\mathcal{C}$ with respect to $\mathcal{H}$, denoted $\mathrm{SC}(\mathcal{C}, \mathcal{H})$, is the least integer n such that for every partially specified subset A, if A is n-consistent with $\mathcal{C}$, then A has an extension in $\mathcal{H}$. If no such n exists, then say $\mathrm{SC}(\mathcal{C}, \mathcal{H}) = \infty$.*

We therefore make the stronger requirement that all partially specified subsets that are n-consistent be consistent, rather than just all totally partially specified subsets. It is immediate from the definition that $\mathrm{C}(\mathcal{C}, \mathcal{H}) \leq \mathrm{SC}(\mathcal{C}, \mathcal{H})$. At the smallest levels, consistency dimension and strong consistency dimension are equal.

Proposition 2.17 *If $\mathrm{C}(\mathcal{C}, \mathcal{H}) = 1$, then $\mathrm{SC}(\mathcal{C}, \mathcal{H}) = 1$. If $\mathrm{C}(\mathcal{C}, \mathcal{H}) = 2$, then $\mathrm{SC}(\mathcal{C}, \mathcal{H}) = 2$.*

As the following examples show, consistency dimension and strong consistency dimension may differ when $\mathrm{C}(\mathcal{C}, \mathcal{H}) \geq 3$.

Example 2.18 *Let $X = \{a, b, c, d, e\}$. Let*

$$\mathcal{C} = \mathcal{H} = \{\{a, b, c\}, \{a, b, d\}, \{a, c, d, e\}, \{b, c, d, e\}\}.$$

One can verify that $\mathrm{C}(\mathcal{C}, \mathcal{H}) = 3$, but the partially specified subset $\{a, b, c, d\}$ with e unspecified witnesses that $\mathrm{SC}(\mathcal{C}, \mathcal{H}) > 3$.

Example 2.19 *Continuing Example 2.15, observe that $\mathrm{SC}(\mathcal{C}, \mathcal{C}) = c^d$. In particular, the partially specified subset A' given by*

$$A'(a_\tau) = \begin{cases} 0 & |\tau| = d \\ \textit{undefined} & \textit{otherwise} \end{cases}$$

witnesses that $\mathrm{SC}(\mathcal{C}, \mathcal{C}) > c^d - 1$. Then we learn in at most $SC(\mathcal{C}, \mathcal{C})$ many queries. Moreover, this demonstrates that consistency dimension and strong consistency dimension can differ by an arbitrarily large amount (allowing $\mathrm{Ldim}(\mathcal{C})$ to vary), and that strong consistency dimension may even be exponentially larger than consistency dimension.

Strong consistency dimension, like consistency dimension, categorizes equivalence query learning:

Theorem 2.20 *$\mathcal{C}$ is learnable with equivalence queries from $\mathcal{H}$ iff* $\mathrm{Ldim}(\mathcal{C}) \leq \infty$ *and* $\mathrm{SC}(\mathcal{C}, \mathcal{H}) < \infty$. *In particular,* $\mathrm{SC}(\mathcal{C}, \mathcal{H}) \leq \mathrm{LC}^{EQ}(\mathcal{C}, \mathcal{H})$.

Proof For the reverse direction, use Theorem 2.6 and the observation that $\mathrm{C}(\mathcal{C}, \mathcal{H}) \leq \mathrm{SC}(\mathcal{C}, \mathcal{H})$.

For the forward direction, use Propositions 2.2 and 2.7. In particular, if $\mathrm{SC}(\mathcal{C}) \geq c$, then there is a partially specified subset A that is $(c-1)$-consistent with $\mathcal{C}$ but which has no total extension in $\mathcal{H}$. Then, by Proposition 2.7, $c \leq \mathrm{LC}^{EQ}(\mathcal{C}, \mathcal{H})$. ■

Corollary 2.21 *Suppose* $\mathrm{Ldim}(\mathcal{C}) < \infty$. *Then* $\mathrm{C}(\mathcal{C}, \mathcal{H}) < \infty$ *iff* $\mathrm{SC}(\mathcal{C}, \mathcal{H}) < \infty$.

The distinction between consistency dimension and strong consistency dimension is subtle, and many previous results hold with little to no modification if one replaces consistency dimension with strong consistency dimension. On the other hand, our work in section 3 will reveal the practical difficulties associated with strong consistency dimension in complicated concept classes.

We have already seen in Theorem 2.20 that strong consistency dimension provides a better lower bound for learning complexity. It is also known in the finite case that strong consistency dimension also gives a stronger upper bound for learning complexity:

Theorem 2.22 *(Balcázar et al., 2002, Theorem 2) Suppose $\mathcal{C}$ is finite. Then* $\mathrm{LC}^{EQ}(\mathcal{C}, \mathcal{H}) \leq \lceil SC(\mathcal{C}, \mathcal{H}) \cdot \ln |\mathcal{C}| \rceil$.

In light of Example 2.19, one hopes that improved bounds on learning can be found in terms of strong consistency dimension and Littlestone dimension when $\mathcal{C}$ is infinite. We are unable to show this presently, but offer some evidence in this direction:

Proposition 2.23 *Suppose* $\mathrm{Ldim}(\mathcal{C}) = d < \infty$ *and* $\mathrm{SC}(\mathcal{C}, \mathcal{H}) = 2 < \infty$. *Then* $\mathrm{LC}^{EQ}(\mathcal{C}, \mathcal{H}) = d + 1$.

The proof of Proposition 2.23 uses strong consistency in a key way, as the hypothesis is generated by extending a certain partially specified subset. Nevertheless, the conclusion holds under the assumption that $C(\mathcal{C}, \mathcal{H}) = 2$, due to Proposition 2.17.

2.4. Adding membership queries and efficient learning of finite classes

Consistency dimension was originally derived from the notion of polynomial certificates, which was used to characterize learning with equivalence and membership queries in the finite case by Hellerstein et al. (1996). The following is an improvement of the upper bound on EQ+MQ learning complexity of $\lceil \mathrm{C}(\mathcal{C}, \mathcal{H}) \log_2 |\mathcal{C}| \rceil$ implicit in the proof of Theorem 3.1.1 in Hellerstein et al. (1996) (stated explicitly in Balcázar et al. (2002)). Our bound replaces $\log_2 |\mathcal{C}|$ with $\mathrm{Ldim}(\mathcal{C})$.

Theorem 2.24 *Suppose* $\mathrm{Ldim}(\mathcal{C}) = d < \infty$ *and* $\mathrm{C}(\mathcal{C}, \mathcal{H}) = c < \infty$. *Then* $\mathrm{LC}^{EQ+MQ}(\mathcal{C}, \mathcal{H}) \leq c'd + 1$, *where* $c' = \max\{1, c-1\}$.

Proof [5]

5. The algorithm is similar to that of Theorem 2.6. However, the applications of Lemma 2.5 are replaced with membership queries.

We proceed by induction on d. The base case, $d = 0$, is trivial, as then $\mathcal{C}$ is a singleton.

Suppose $\mathrm{Ldim}(\mathcal{C}) = d + 1$. Suppose there is some element x such that $\mathrm{Ldim}(\mathcal{C}^{(x,0)}) < d + 1$ and $\mathrm{Ldim}(\mathcal{C}^{(x,1)}) < d + 1$. Then by induction, any concept in $\mathcal{C}^{(x,0)}$ can be learned in at most $c'd + 1$ queries with guesses from $\mathcal{H}$, and the same is true for $\mathcal{C}^{(x,1)}$. Submit x as a membership query. This tells us whether the target concept lies in $\mathcal{C}^{(x,0)}$ or $\mathcal{C}^{(x,1)}$, and then we require at most $c'd + 1$ many queries, for a total of $c'd + 2 \leq c'(d + 1) + 1$ many queries.

If no such x exists, then for all x, either $\mathrm{Ldim}(\mathcal{C}^{(x,0)}) = d + 1$ or $\mathrm{Ldim}(\mathcal{C}^{(x,1)}) = d + 1$. Let B be such that $x \in B$ iff $\mathrm{Ldim}(\mathcal{C}^{(x,1)}) = d + 1$.

If $B \in \mathcal{H}$, then we submit B as our query. If we are incorrect, then by choice of B, the class $\mathcal{C}'$ of concepts consistent with the counterexample x_0 will have Littlestone dimension $\leq d$. By induction, any concept in $\mathcal{C}'$ can be learned in at most $c'd + 1$ many queries, and so we learn the target in at most $c'd + 2 \leq c'(d + 1) + 1$ queries.

If $B \notin H$, then, since $\mathrm{C}(\mathcal{C}, \mathcal{H}) = c$, there are some $x_0, \ldots, x_{c-1}$ such that there is no $A \in \mathcal{C}$ such that $B|_{\{x_0,\ldots,x_{c-1}\}} \preceq A$. (Observe that this cannot happen when $c = 1$. In fact, Proposition 2.17 and the proof of Proposition 2.23 imply that this cannot even happen when $c = 2$. In particular, $c' = c - 1$.) Then

$$\mathcal{C} = (\mathcal{C}^{(x_0,1-B(x_0))}) \cup \ldots \cup (\mathcal{C}^{(x_{c-1},1-B(x_{c-1}))}),$$

and $\mathrm{Ldim}(\mathcal{C}^{(x_i,1-B(x_i))}) \leq d$ for each i. By induction, any concept in each $\mathcal{C}^{(x_i,1-B(x_i))}$ can be learned in at most $c'd + 1$ many queries. By submitting $x_0, \ldots, x_{c-2}$ as membership queries, we can determine some i such that the target belongs to $\mathcal{C}^{(x_i,1-B(x_i))}$ (if the result of each membership query on x_j is $B(x_j)$, then we know that $i = c-1$). We therefore learn in at most $c'd+1+(c-1) = c'(d + 1) + 1$ many queries. ■

We have a lower bound on learning complexity in terms of consistency dimension in this setting analogous to Proposition 2.7:

Proposition 2.25 *Suppose there is some (total) subset A which is n-consistent with $\mathcal{C}$ but which does not have a total extension in $\mathcal{H}$. Then $n < \mathrm{LC}^{EQ+MQ}(\mathcal{C}, \mathcal{H})$. In particular, $\mathrm{C}(\mathcal{C}, \mathcal{H}) \leq \mathrm{LC}^{EQ+MQ}(\mathcal{C}, \mathcal{H})$.*

Finally, putting together the various upper and lower bounds from this section we give a characterization of those problems efficiently learnable by equivalence and membership queries:

Theorem 2.26 *Let $(\mathcal{C}_n, \mathcal{H}_n)$ for $n \in \mathbb{N}$ be concept classes and hypothesis classes, respectively. Let $c_n = \mathrm{C}(\mathcal{C}_n, \mathcal{H}_n)$. Let $d_n = \mathrm{Ldim}(\mathcal{C}_n)$. The classes $\mathcal{C}_n$ are learnable with at most polynomially in n many equivalence queries from $\mathcal{H}_n$ and membership queries if and only if c_n and d_n are bounded by a polynomial in n. If there is any algorithm for learning an arbitrary concept of $\mathcal{C}_n$ using at most polynomially in n many membership queries and equivalence queries in $\mathcal{H}_n$, then the algorithm from Theorem 2.24 also learns $\mathcal{C}_n$ using at most polynomially many membership queries and equivalence queries in $\mathcal{H}_n$.*

Proof In Theorem 2.24, we proved that $\mathrm{LC}^{EQ+MQ}(\mathcal{C}, \mathcal{H}) \leq c'd+1$, where $c' = \max\{1, \mathrm{C}(\mathcal{C}, \mathcal{H})-1\}$ and $d = \mathrm{Ldim}(\mathcal{C})$. So, if c_n and d_n are polynomially bounded, then so is $\mathrm{LC}^{EQ+MQ}(\mathcal{C}_n, \mathcal{H}_n)$.

In Proposition 2.25, we showed that $\mathrm{LC}^{EQ+MQ}(\mathcal{C}, \mathcal{H}) \geq \mathrm{C}(\mathcal{C}, \mathcal{H})$, so it follows that if c_n is not polynomially bounded then neither is $\mathrm{LC}^{EQ+MQ}(\mathcal{C}_n, \mathcal{H}_n)$.

Now suppose that d_n is not polynomially bounded. By (Auer and Long, 1994, Theorem 2.1) [6] we have

$$\mathrm{LC}^{EQ+MQ}(\mathcal{C},\mathcal{H}) \geq \mathrm{LC}^{EQ+MQ}(\mathcal{C},\mathcal{P}(X)) \geq \log\left(\frac{4}{3}\right)\cdot \mathrm{LC}^{EQ}(\mathcal{C},\mathcal{P}(X)).$$

By (Littlestone, 1988, Theorems 5 and 6), we can replace $\mathrm{LC}^{EQ}(\mathcal{C},\mathcal{P}(X))$ with $\mathrm{Ldim}(\mathcal{C})$. Thus:

$$\mathrm{LC}^{EQ+MQ}(\mathcal{C}_n,\mathcal{H}_n) \geq \log\left(\frac{4}{3}\right)\cdot d_n,$$

and it follows that $\mathrm{LC}^{EQ+MQ}(\mathcal{C}_n,\mathcal{H}_n)$ is not polynomially bounded, completing the proof. ■

Finally, the upper and lower bounds of this section also yield a characterization of which infinite classes are learnable in finitely many equivalence and membership queries.

Corollary 2.27 *$\mathcal{C}$ is learnable with membership queries and equivalence queries from $\mathcal{H}$ iff* $\mathrm{Ldim}(\mathcal{C}) < \infty$ *and* $\mathrm{C}(\mathcal{C},\mathcal{H}) < \infty$.

3. Efficient learnability of regular languages

In a seminal paper, Angluin (1987) showed that regular languages are efficiently learnable with equivalence queries plus membership queries, and in this subsection, we will use Theorem 2.24 to give an alternate short proof of this fact.[7] Let $\mathcal{L}_{n,m}$ be the class of binary regular languages on strings of length at most m specified by a deterministic finite automaton on at most n nodes. The $\mathcal{L}^*$ algorithm of Angluin (1987) specifically uses $\mathcal{O}(n)$ equivalence queries and $\mathcal{O}(mn^2)$ membership queries. We let $\mathrm{DFA}_2(n)$ denote the collection of (equivalence classes of) deterministic finite automata accepting binary strings and having at most n nodes. The proof of the next proposition is straightforward.

Proposition 3.1 *The Littlestone dimension of DFA*$_2(n)$ *is at most* $\mathcal{O}(n\log n)$.

The proof of the following proposition reveals the connection between consistency and the Myhill-Nerode theorem.

Proposition 3.2 $\mathrm{C}(DFA_2(n)) \leq 2^{\binom{n+1}{2}} = n(n+1)$.

Proof Fix a subset C of binary strings and x, y binary strings. We say that z is a (C-) distinguishing extension of x and y if $xz \in C$ but $yz \notin C$ or vice versa. If x and y have no distinguishing extension, then we say x and y are C-equivalent, and write $x \sim_C y$. The Myhill-Nerode theorem (Nerode, 1958) says that a subset of binary strings of length m is the accept set of a finite automaton with at most n nodes if and only if the number of $\sim_C$ classes is at most n. Thus, given any subset C

6. The inequality of Auer and Long (1994) gives a lower bound for LC^{EQ+MQ} which improved on the lower bound of $\frac{\mathrm{LC}^{EQ}(\mathcal{C},\mathcal{P}(X))}{\log(1+\mathrm{LC}^{EQ}(\mathcal{C},\mathcal{P}(X)))}$ from (Maass and Turán, 1990, Theorem 3). In fact, Theorem 3 of Maass and Turán (1990) actually suffices for our purposes.
7. In the following sections, we only make use of *proper equivalence queries*, that is, $\mathcal{H} = \mathcal{C}$. We shall therefore let $\mathrm{C}(\mathcal{C}) := \mathrm{C}(\mathcal{C},\mathcal{C})$, which we will call the consistency dimension of $\mathcal{C}$ (with analogous notation for strong consistency dimension).

of the binary strings of length m which is not a regular language recognized by an automaton with at most n nodes, there are at least $n+1$ $\sim_C$-classes of elements. Pick representatives $x_0, \ldots, x_n$ from $n+1$ classes, and for each $i < j$, pick some z_{ij} that is a distinguishing extension of x_i and x_j. Then restricting C to the partial assignment on $\{x_k z_{ij} \mid i < j,\ k = i, j\}$, a domain of size $2\binom{n+1}{2} = n(n+1)$ that witnesses that $x_i \not\sim_C x_j$ for all $i \neq j$, we can see that this restriction is inconsistent with the class of regular languages recognized by automata with at most n nodes. Therefore $\mathrm{C}(\mathrm{DFA}_2(n)) \leq n(n+1)$. [8] ■

Now, by Theorem 2.24 and the previous two results, it follows that:

Theorem 3.3 *The class $\mathcal{L}_{n,m}$ is learnable in at most $\mathcal{O}(n \log n)$ equivalence queries and at most $O\left(n^3 \log n\right)$ membership queries.*

It is interesting to note that contrary to $\mathcal{L}^*$, when using the algorithm from Theorem 2.24, there is no dependence on m, the length of the binary strings which the teacher is allowed to provide as counterexamples[9].

Theorem 2.6 now implies that $\mathcal{L}_{n,m}$ is learnable in at most $(n(n+1))^{\mathcal{O}(n \log n)}$ equivalence queries. Theorem 2.22 shows that a finite class $\mathcal{C}$ is learnable in at most $\lceil \mathrm{SC}(\mathcal{C}) \cdot \ln |\mathcal{C}| \rceil$ equivalence queries. Since Angluin (1990) showed that $\mathcal{L}_{n,m}$ is not learnable in polynomially many equivalence queries, it follows that $\mathrm{SC}(\mathcal{L}_{n,m})$ cannot be polynomial in n, m.

3.1. Learning ω-languages

In this section, we consider the natural extension to languages on infinite strings indexed by ω, called ω-languages. For an alphabet Σ, we denote by Σ^ω, the strings of symbols from Σ of order type ω. Similar to the previous section, we consider an automaton, which consists of the collection $\mathbb{A} = (\Sigma, Q, q_0, \delta)$, where Q is a finite collection of states, q_0 is the initial state, and $\delta : Q \times \Sigma \to 2^Q$ is a transition rule. To form a language, an automaton is equipped with an acceptance criterion.[10] Fix a subset $F \subseteq Q$. A run of a *Büchi automaton* is accepting if and only if it visits the set F infinitely often. An ω-language is ω-regular if it is recognized by a non-deterministic Büchi automaton. A run of a *co-Büchi automaton* is accepting if and only if it visits F only finitely often. Let $\psi : Q \to \{1, \ldots, k\}$ be a function, which we think of as a coloring of the states of the automaton. Let c be the minimum color which is visited infinitely often. A run of a *parity automaton* is accepting if and only if c is odd.

Two ω-regular languages are equivalent if they agree on the set of periodic words (McNaughton, 1966), which allows for the possibility of recognizing the ω-language using finitary automata. This is the approach of Angluin and Fisman (2016); Fisman et al. (2018), whose notation we follow closely. A *family of DFAs* (FDFA) $\mathcal{F}$ is a pair (Q, P) where Q is a DFA with $|Q|$ states and P is a collection of $|Q|$ many DFAs, which we refer to as *progress DFAs* - one DFA P_q for each state q of Q. Given a pair of finite words, (u, v), a run of our family of DFAs consists of running Q on u, then running $P_{Q(u)}$ on v where $Q(u)$ is the ending state of Q on u. The pair (u, v) can be used to represent an infinite periodic word uv^ω.

8. Note that the same proof shows that the consistency dimension of $\mathrm{DFA}_m(n)$ is also at most $n(n+1)$.
9. We should also note that $\mathcal{L}^*$ was improved by Schapire to give a better bound on membership queries (still depending on m). Schapire (1991).
10. Numerous acceptance criteria have been extensively studied in the literature, and we refer the reader to Angluin and Fisman (2016); Fisman et al. (2018); Fisman (2018) for overviews.

Let FDFA(n, m) be the class of families of deterministic finite automata where the leading automaton has at most n nodes and the progress automata each have at most m nodes. It is *not* quite true that once an ω-regular language has been reduced to an FDFA that one can use $\mathcal{L}^*$ directly to learn the various DFAs in the family (see Angluin and Fisman (2016, section 4)). It is also not completely obvious what the bounds for Littlestone and consistency dimension are in terms of the DFAs in the family, but the next two results give such bounds which imply the efficient learnability of ω-regular languages.

Proposition 3.4 *The class FDFA(n, m) has Littlestone dimension at most $\mathcal{O}(n \log n + nm \log m)$.*

Proposition 3.5 $\mathrm{C}(\textit{FDFA}(n, m)) \leq 2^{\binom{n(m+1)}{2}} = \mathcal{O}(n^2 m^2)$.

Using the previous two results together with Theorem 2.24, one can deduce the efficient learnability of FDFA(n, m):

Theorem 3.6 *The class FDFA(n, m) is learnable in at most $\mathcal{O}(n \log n + nm \log m)$ equivalence queries and at most $\mathcal{O}((\log n + m \log m) \cdot (n^3 m^2))$ membership queries.*

We have formulated our bounds in terms of the number of states in the FDFA corresponding to a given ω-language. In Angluin and Fisman (2016); Fisman et al. (2018) bounds on the number of states of FDFAs in terms of the number of states of automata for ω-languages with various acceptors are given. Specifically, the following bounds hold:

1. When $\mathcal{A}$ is a deterministic Büchi (DBA) or co-Büchi (DCA) automaton with n states, there is an equivalent FDFA of size at most $(n, 2n)$ (Fisman et al., 2018, 5.3).
2. When $\mathcal{A}$ is a deterministic partiy automaton (DPA) with n states and k colors, there is an equivalent FDFA of size at most (n, kn) (Fisman et al., 2018, 5.4).
3. When $\mathcal{A}$ is an nondeterministic Büchi automaton (NBA) with n states, there is an equivalent FDFA of size at most $(2^{\mathcal{O}(n \log n)}, 2^{\mathcal{O}(n \log n)})$.

Any NBA can be translated into a DPA, and so 2) yields the efficient learnability of ω-regular languages *in terms of the number of states in a DPA* (this translation also yields 3). However, the translation from NBA to DPA is known to require an exponential increase in the number of states in general (Piterman, 2006). From an FDFA of size at most (n, k) there is a translation into an NBA with at most $\mathcal{O}(n^2 k^3)$ states (Fisman et al., 2018, Theorem 5.8), and so it follows that the exponential increase in states in moving from NBAs to FDFAs is necessary (Fisman et al., 2018, Theorem 5.6).

Finally, we mention that Angluin and Fisman (2018) define restricted classes of ω-languages for which right-congruence is *fully informative*, and isolate numerous classes (e.g. for each type of acceptor from the previous subsection) of ω-languages for which an infinitary invariant of the Myhill-Nerode theorem holds. This variant of Myhill-Nerode is sufficient to bound the consistency dimension (and thus establish the learnability) of the classes in terms of the number of of right equivalence classes of $\sim_{\mathcal{L}}$ similar to the proof of Proposition 3.2.

Acknowledgments

We would like to thank Lev Reyzin and György Turán for much advice and useful discussion about query learning as well as pointing us towards many useful references. Thanks are also owed to the participants of Dagstuhl Seminar 19361 "Logic and Learning", especially Dana Fisman. Hunter Chase was supported by NSF grant 2002165. James Freitag was supported by NSF award no. 1700095 and CAREER award 1945251.

References

Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private pac learning implies finite littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 852–860, 2019.

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

Dana Angluin. Negative results for equivalence queries. *Machine Learning*, 5(2):121–150, 1990.

Dana Angluin and Dana Fisman. Learning regular omega languages. *Theoretical Computer Science*, 650:57–72, 2016.

Dana Angluin and Dana Fisman. Regular omega-languages with an informative right congruence. *Electronic Proceedings in Theoretical Computer Science*, 277:265–279, 09 2018. doi: 10.4204/EPTCS.277.19.

Peter Auer and Philip M Long. Simulating access to hidden information while learning. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 263–272. ACM, 1994.

José L. Balcázar, Jorge Castro, David Guijarro, and Hans-Ulrich Simon. The consistency dimension and distribution-dependent learning from queries. *Theoretical Computer Science*, 288(2):197–215, 2002.

Shai Ben-David, Dávid Pál, and Shai Shalev-shwartz. Agnostic online learning. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT*, 2009.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Hunter Chase and James Freitag. Model theory and machine learning. *Bulletin of Symbolic Logic*, 25(3):319–332, 2019.

Dana Fisman. Inferring regular languages and ω-languages. *Journal of Logical and Algebraic Methods in Programming*, 98:27–49, 2018.

Dana Fisman, Udi Boker, and Dana Angluin. Families of DFAs as acceptors of ω-regular languages. *Logical Methods in Computer Science*, 14, 2018.

Lisa Hellerstein, Krishnan Pillaipakkamnatt, Vijay Raghavan, and Dawn Wilkins. How many queries are needed to learn? *Journal of the ACM*, 43(5):840–862, 1996.

Yoshiyasu Ishigami and Sei'ichi Tani. VC-dimensions of finite automata and commutative finite automata with k letters and n states. *Discrete Applied Mathematics*, 74(2):123–134, 1997.

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

Wolfgang Maass and György Turán. On the complexity of learning from counterexamples and membership queries. In *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*, pages 203–210. IEEE, 1990.

Robert McNaughton. Testing and and generating infinite sequences by a finite automaton. *Information and Control*, 9(5):521–530, 1966.

Anil Nerode. Linear automaton transformations. *Proceedings of the American Mathematical Society*, 9(4):541–544, 1958.

Nir Piterman. From nondeterministic Büchi and Streett automata to deterministic parity automata. In *Logic in Computer Science, 2006 21st Annual IEEE Symposium on*, pages 255–264. IEEE, 2006.

Robert E Schapire. The design and analysis of efficient learning algorithms. Technical report, Massachusetts Institute of Technology Lab for Computer Science, 1991.

Appendix A. Littlestone dimension

Let $\mathcal{C}$ be a concept class on a set X.

Definition A.1 *A* binary element tree of height h *is a complete binary tree of height h whose non-leaf nodes are labeled by elements of X and whose leaves are labeled by sets in $\mathcal{C}$ (see Figure 1). The height of the tree is the length of the path from the root to any leaf.*

Definition A.2 *Given a binary element tree, a node v_1 is* below *a node v_2 if v_2 lies on the (unique) path from v_1 to the root of the tree. We say that v_1 is* left-below *v_2 if v_1 is below v_2 and the first edge along the path from v_2 to v_1 goes down and to the left. The notion of* right-below *is defined analogously. When a node labeled by b is left-below a node labeled by a, we write $a <_L b$. Similarly, when a node labeled by b is right-below a node labeled by a, we write $a <_R b$.*

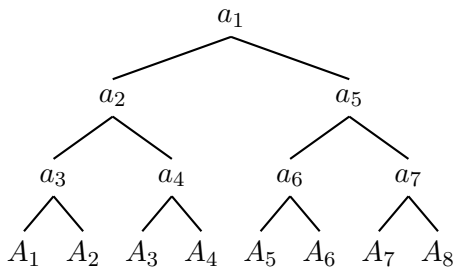

Figure 1: A binary element tree of height three. Here $a_i \in X$ and $A_i \in \mathcal{C}$. The leaf labeled with A_4 is well-labeled if and only if $a_1 \notin A_4$ and $a_2, a_4 \in A_4$. For all other a_i, there is no requirement about membership in A_4.

Definition A.3 *A leaf labeled by $A \in \mathcal{C}$ is* properly labeled *if, for each node a that A is below, we have*

$$a \in A \text{ if and only if } a <_R A.$$

Definition A.4 *The* Littlestone dimension *of a set system $\mathcal{C}$, $\mathrm{Ldim}(\mathcal{C})$, is the maximum integer n such that there exists a binary element tree of height n where all leaves can be properly labeled by sets in $\mathcal{C}$. If there is no maximum n, we write $\mathrm{Ldim}(\mathcal{C}) = \infty$.*

Appendix B. Proofs of results from section 2

B.1. Proof of Proposition 2.2

Suppose $\mathrm{Ldim}(\mathcal{C}) \geq d+1$. We show that we can force the learner to use at least $d+2$ equivalence queries. Construct a binary element tree of height $d+1$ with proper labels from $\mathcal{C}$ witnessing $\mathrm{Ldim}(\mathcal{C}) \geq d+1$. Given the first hypothesis H_0 from the learner, return the element on the 0th level on the tree as a counterexample. Continue this, returning the element on the ith level along the path consistent with previous counterexamples as the counterexample to hypothesis H_i. We will return $d+1$ counterexamples, and the learner still requires one more hypothesis to identify the concept. Since this will occur for one of the proper labels A of the binary element tree, we have forced the learner to use at least $d+2$ equivalence queries for some $A \in \mathcal{C}$.

Suppose $\mathrm{Ldim}(\mathcal{C}) = d < \infty$. Let $\mathcal{C}_0 = \mathcal{C}$. Inductively define $\mathcal{C}_i$, $i = 1, \ldots, d$ as follows. Given $\mathcal{C}_i$, for any $x \in X$ and $j \in \{0,1\}$, let

$$\mathcal{C}_i^{(x,j)} := \{A \in \mathcal{C}_i \mid A(x) = j\},$$

and let

$$B_i := \{x \in X \mid \mathrm{Ldim}(\mathcal{C}_i^{(x,1)}) \geq \mathrm{Ldim}(\mathcal{C}_i^{(x,0)})\}.$$

Submit B_i as the hypothesis. If B_i is correct, we are done. Otherwise, we receive a counterexample x_i. Set

$$\mathcal{C}_{i+1} := \{A \in V_i \mid A(x_i) \neq B_i(x_i)\}$$

to be the concepts which have the correct label for x_i. Observe that at each stage, $\mathrm{Ldim}(\mathcal{C}_{i+1}) < \mathrm{Ldim}(\mathcal{C}_i)$. Therefore, if we make d queries without correctly identifying the target, then we must have $\mathrm{Ldim}(\mathcal{C}_d) = 0$. Then V_d is a singleton, which must be the target concept.

B.2. Proof of Lemma 2.5

We give the proof for $n = 2$; then the result for $n > 2$ follows easily by induction.

To learn a target concept $A \in \mathcal{C} = \mathcal{C}_0 \cup \mathcal{C}_1$ with hypotheses from $\mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1$, begin by assuming that $A \in \mathcal{C}_0$. Attempt to learn A by making guesses from $\mathcal{H}_0$, according to the procedure by which any concept in $\mathcal{C}_0$ is learnable in at most m_0 many queries. If, after making m_0 many queries, we have failed to learn A, then we conclude that $A \notin \mathcal{C}_0$, whence $A \in \mathcal{C}_1$. We can then learn A in at most m_1 many additional queries with guesses from $\mathcal{H}_1$.

B.3. Proof of Proposition 2.7

By hypothesis, given any equivalence query H, the teacher can find some $x \in \mathrm{dom}(A)$ such that $H(x) \neq A(x)$. Moreover, since A is n-consistent with $\mathcal{C}$, the teacher is able to return a counterexample of this form for the first n equivalence queries. Thus $\mathcal{C}$ cannot be learned with fewer than $n+1$ equivalence queries from $\mathcal{H}$.

B.4. Proof of Lemma 2.10

Let $X = \{x_\alpha \mid \alpha < |X|\}$ be a well-ordering of X. Let $A_0 = A$. We inductively define a $\preceq$-chain of partially specified subsets A_α, where each A_α is defined on $\mathrm{dom}(A) \cup \{x_\xi \mid \xi < \alpha\}$ and is finitely consistent with $\mathcal{C}$. For α a limit ordinal, set $A_\alpha = \cup_{\xi<\alpha} A_\xi$. It is clear that A_α is finitely consistent with $\mathcal{C}$ if all A_ξ for $\xi < \alpha$ are.

At any successor stage $\alpha + 1$, if $x_\alpha \in \mathrm{dom}(A_\alpha)$, set $A_{\alpha+1} = A_\alpha$. Otherwise, we must extend A_α to x_α while remaining finitely consistent with $\mathcal{C}$. Assume for contradiction that neither $B_0 := A_\alpha \cup \{x_\alpha \mapsto 0\}$ nor $B_1 := A_\alpha \cup \{x_\alpha \mapsto 1\}$ are finitely consistent with $\mathcal{C}$. Then there are finite sets $Y_0, Y_1 \subseteq \mathrm{dom}(A_\alpha)$ such that $B_0|_{Y_0 \cup \{a_\alpha\}}$ and $B_1|_{Y_1 \cup \{a_\alpha\}}$ have no extension in $\mathcal{C}$. But $A_\alpha|_{Y_0 \cup Y_1}$ has an extension B in $\mathcal{C}$, and B must be an extension of either $B_0|_{Y_0 \cup \{a_\alpha\}}$ or $B_1|_{Y_1 \cup \{a_\alpha\}}$, a contradiction. So A_α has a finitely consistent extension to x_α, and we set $A_{\alpha+1}$ to be such an extension.

We then take $A' = \cup_{\xi<|X|} A_\xi$.

B.5. Proof of Proposition 2.11

(i) $\Rightarrow$ (ii): Let $A' \succeq A$ be a total extension finitely consistent with $\mathcal{C}$. If $\mathrm{C}(\mathcal{C}, \mathcal{H}) < \infty$, then $A' \in \mathcal{H}$.

(ii) $\Rightarrow$ (i): We show the contrapositive. Suppose that A is not finitely consistent with $\mathcal{C}$, witnessed by some size n restriction A_0, which is a $\preceq$-minimal such restriction. We find some $\mathcal{H}$ such that $\mathrm{C}(\mathcal{C}, \mathcal{H}) < \infty$ but $\mathcal{H}$ contains no total extension of A. Let $\mathcal{H}$ be the collection of all (total partially specified) subsets which are not extensions of A_0. So A has no total extension in $\mathcal{H}$. We claim that $\mathrm{C}(\mathcal{C}, \mathcal{H}) \leq n$. Indeed, observe that given any (total partially specified) subset B that is n-consistent with $\mathcal{C}$, we have $A_0 \not\preceq B$, and then $B \in \mathcal{H}$.

B.6. Proof of Proposition 2.12

(i) $\Rightarrow$ (ii) Assume for contradiction that there is some total A which is n-consistent but not finitely consistent. Let m be minimal such that A is m-inconsistent. Then there is a size m restriction $A' \preceq A$ that has no extension in $\mathcal{C}$. Then let $\mathcal{H}$ contain all subsets which do not extend A'.

We claim that $\mathrm{C}(\mathcal{C}, \mathcal{H}) = m$. Note that A witnesses that $\mathrm{C}(\mathcal{C}, \mathcal{H}) \geq m$. On the other hand, observe that given any partially specified subset B that is m-consistent with $\mathcal{C}$, we have $A' \not\preceq B$, and then it is easy to see that B has a total extension in $\mathcal{H}$.

(ii) $\Rightarrow$ (iii): If $\mathcal{H}$ contains all finitely consistent subsets, and all n-consistent subsets are finitely consistent, then $\mathrm{C}(\mathcal{C}, \mathcal{H}) \leq n$ holds immediately.

(iii) $\Rightarrow$ (i): By Proposition 2.11, if $\mathrm{C}(\mathcal{C}, \mathcal{H}) < \infty$, then $\mathcal{H}$ already has all finitely consistent subsets. Then $\mathrm{C}(\mathcal{C}, \mathcal{H}) \leq n$.

B.7. Proof of Theorem 2.14

Fix some $m > d = \mathrm{Ldim}(\mathcal{C})$. Let $\mathcal{H}_m$ be the collection of all subsets which are m-consistent with $\mathcal{C}$. It is immediate that $\mathrm{C}(\mathcal{C}, \mathcal{H}_m) \leq m < \infty$.

Assume for contradiction that $\mathrm{Ldim}(\mathcal{H}_m) > \mathrm{Ldim}(\mathcal{C})$. Consider a binary element tree of height $\mathrm{Ldim}(\mathcal{H}_m)$ that can be properly labeled with elements of $\mathcal{H}_m$; in particular, there is some leaf which cannot be labeled with an element of $\mathcal{C}$. Consider such a leaf. The path through the binary element tree to this leaf defines a partially specified subset A that is $(d+1)$-inconsistent with $\mathcal{C}$. In particular, any total extension is $(d+1)$-inconsistent, so m-inconsistent, and so does not belong to $\mathcal{H}_m$. This contradicts our ability to label the leaf with an element of $\mathcal{H}$.

In particular, recall that when $\mathcal{C}$ has finite consistency threshold n, A is n-consistent with $\mathcal{C}$ iff it is finitely consistent with $\mathcal{C}$. So setting $\mathcal{H}_m$ as above with m at least the finite consistency threshold amounts to setting $\mathcal{H}_m$ to be the collection of all finitely consistent partially specified subsets. In this case, $\mathrm{Ldim}(\mathcal{H}_m) = \mathrm{Ldim}(\mathcal{C})$ even if $m \leq d$, as increasing the Littlestone dimension requires adding something inconsistent with $\mathcal{C}$.

Regardless of whether $\mathcal{C}$ has finite consistency dimension, we can let $m = d + 1$. Then $\mathrm{C}(\mathcal{C}, \mathcal{H}_m) \leq m = d + 1$.

B.8. Proof of Proposition 2.17

Observe that $\mathrm{C}(\mathcal{C}, \mathcal{H}) = 1$ iff $\mathrm{SC}(\mathcal{C}, \mathcal{H}) = 1$ iff $\mathcal{H}$ shatters the set of all elements $x \in X$ such that there are A_0 and A_1 in $\mathcal{C}$ such that $x \notin A_0$ but $x \in A_1$.

Suppose that $\mathrm{C}(\mathcal{C}, \mathcal{H}) = 2$. Let A be a partially specified subset that is 2-consistent with $\mathcal{C}$. We wish to find a total extension of A in $\mathcal{H}$. It suffices to find a total extension $B \succeq A$ that is 2-consistent with $\mathcal{C}$.

Let $X = \{x_\alpha \,|\, \alpha < |X|\}$ be a well-ordering of X. Let $A_0 = A$. We inductively define a $\preceq$-chain of partially specified subsets A_α, where each A_α is defined on $\mathrm{dom}(A) \cup \{x_\xi \,|\, \xi < \alpha\}$ and is 2-consistent with $\mathcal{C}$. For α a limit ordinal, set $A_\alpha = \cup_{\xi<\alpha} A_\xi$. It is clear that A_α is 2-consistent with $\mathcal{C}$ if all A_ξ for $\xi < \alpha$ are.

At any successor stage $\alpha+1$, if $x_\alpha \in \mathrm{dom}(A_\alpha)$, set $A_{\alpha+1} = A_\alpha$. Otherwise, we must extend A_α to x_α while remaining 2-consistent with $\mathcal{C}$. Assume for contradiction that neither $B_0 := A_\alpha \cup \{x_\alpha \mapsto 0\}$ nor $B_1 := A_\alpha \cup \{x_\alpha \mapsto 1\}$ are 2-consistent with $\mathcal{C}$. Then there are $y_0, y_1 \in \mathrm{dom}(A_\alpha)$ such that $B_0|_{\{y_0, x_\alpha\}}$ and $B_1|_{\{y_1, x_\alpha\}}$ have no extension in $\mathcal{C}$. But $A_\alpha|_{\{y_0, y_1\}}$ has an extension B in $\mathcal{C}$, and B must be an extension of either $B_0|_{\{y_0, x_\alpha\}}$ or $B_1|_{\{y_1, x_\alpha\}}$, a contradiction. So A_α has a 2-consistent extension to x_α, and we set $A_{\alpha+1}$ to be such an extension.

We then take $\cup_{\xi<|X|} A_\xi$ to be our total extension.

B.9. Proof of Theorem 2.22

As this was originally framed in the setting where concepts were represented by strings, we give an abbreviated translation of the original proof into the language of set systems. This proof demonstrates the utility of constructing a partial hypothesis and taking some complete extension.

Let $c = \mathrm{SC}(\mathcal{C}, \mathcal{H})$. At stage i, let $\mathcal{C}_i \subseteq \mathcal{C}$ be the set of remaining possible target concepts. Let A_i be the partially specified subset given by

$$A(x) = \begin{cases} 1 & x \text{ belongs to more than } \frac{c-1}{c}|\mathcal{C}_i| \text{ many } C \in \mathcal{C}_i \\ 0 & x \text{ belongs to less than } \frac{1}{c}|\mathcal{C}_i| \text{ many } C \in \mathcal{C}_i \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Observe that A is c-consistent with $\mathcal{C}$—given any $Y := \{x_0, \ldots, x_{c-1}\} \subseteq \mathrm{dom}(A)$, for each j, less than $\frac{1}{c}|\mathcal{C}_i|$ many remaining concepts disagree with A on x_j, so less than $c\frac{1}{c}|\mathcal{C}_i| = |\mathcal{C}_i|$ many concepts disagree with A on some x_j. So some concept agrees with A on Y. So A is c-consistent.

So we can find some $B \in \mathcal{H}$ such that $B \succeq A$, and we submit B as our hypothesis. By choice of A, if we receive a counterexample, we will have $|\mathcal{C}_{i+1}| \leq \frac{c-1}{c}|\mathcal{C}_i|$. Repeating this $\lceil c \cdot \ln |\mathcal{C}| \rceil$ many times is enough to identify and submit the target concept.

B.10. Proof of Proposition 2.23

We know by Proposition 2.2 that $d + 1$ is a lower bound. We show that it is also an upper bound.

Let $\mathcal{C}_0 = \mathcal{C}$. Inductively define $\mathcal{C}_i$, $i = 1, \ldots, d$ as follows. Construct the partially specified subset A_i where

$$A_i(x) = \begin{cases} 0 & \mathrm{Ldim}(\mathcal{C}_i^{(x,0)}) = \mathrm{Ldim}(\mathcal{C}_i) \\ 1 & \mathrm{Ldim}(\mathcal{C}_i^{(x,1)}) = \mathrm{Ldim}(\mathcal{C}_i) \\ \text{undefined} & \text{otherwise.} \end{cases} \tag{B.1}$$

We claim that A_i has an extension in H. By our assumption that $\mathrm{SC}(\mathcal{C}, \mathcal{H}) = 2$, it suffices to check that A is 2-consistent with $\mathcal{C}_i$. Suppose for contradiction that there are $a_0, a_1 \in \mathrm{dom}(A_i)$ such that, without loss of generality, $A_i(a_0) = A_i(a_1) = 0$, but there is no extension of $A_i|_{\{a_0,a_1\}}$ in $\mathcal{C}_i$. Then observe that $\mathcal{C}_i^{(x_0,0)} \subseteq \mathcal{C}_i^{(x_1,1)}$, whence

$$\mathrm{Ldim}(\mathcal{C}_i) \geq \mathrm{Ldim}(\mathcal{C}_i^{(x_1,1)}) \geq \mathrm{Ldim}(\mathcal{C}_i^{(x_0,0)}) = \mathrm{Ldim}(\mathcal{C}_i),$$

so $\mathrm{Ldim}(\mathcal{C}_i^{(x_1,1)}) = \mathrm{Ldim}(\mathcal{C}_i)$. But we also have $\mathrm{Ldim}(\mathcal{C}_i^{(x_1,0)}) = \mathrm{Ldim}(\mathcal{C}_i)$, a contradiction, as we could then construct a binary element tree with proper labels from $\mathcal{C}_i$ of height $\mathrm{Ldim}(\mathcal{C}_i) + 1$ with x_1 at the root.

Let $B_i \in \mathcal{H}$ be a total extension of A_i. Submit B_i as the hypothesis. If B_i is correct, we are done. Otherwise, we receive a counterexample x_i. Set

$$\mathcal{C}_{i+1} := \{B \in \mathcal{C}_i \mid B(x_i) \neq B_i(x_i)\}.$$

Observe that at each stage, $\mathrm{Ldim}(\mathcal{C}_{i+1}) < \mathrm{Ldim}(\mathcal{C}_i)$. Therefore, if we make d queries without correctly identifying the target, then we must have $\mathrm{Ldim}(\mathcal{C}_d) = 0$. Then $\mathcal{C}_d$ is a singleton, which must be the target concept.

B.11. Proof of Proposition 2.25

We first show that $n < \mathrm{LC}^{EQ+MQ}(\mathcal{C}, \mathcal{H})$. If the learner submits x as a membership query, the teacher returns $A(x)$ if possible, that is, if there is a concept $B \in \mathcal{C}$ which agrees with the previous data and satisfies $B(x) = A(x)$.

By hypothesis, given any equivalence query H, the teacher can find some $x \in \mathrm{dom}(A)$ such that $H(x) \neq A(x)$, and the teacher returns a counterexample of this form if possible, that is, if there is a concept $B \in \mathcal{C}$ which agrees with the previous data and satisfies $B(x) = A(x)$.

Moreover, since A is n-consistent with $\mathcal{C}$, the teacher is able to return data of this form for the first n queries. Thus $\mathcal{C}$ cannot be learned with fewer than $n+1$ equivalence queries from $\mathcal{H}$.

From this, it follows that $\mathrm{C}(\mathcal{C}, \mathcal{H}) \leq \mathrm{LC}^{EQ+MQ}(\mathcal{C}, \mathcal{H})$.

Appendix C. Proofs from section 3

C.1. Proof of Proposition 3.1

In Ishigami and Tani (1997, Proposition 1), it is shown that $|DFA_2(n)| \leq \frac{n^{2n}2^n n}{n!} \leq 2^{\mathcal{O}(n \log n)}$. From this, it follows that the Littlestone dimension of $DFA_2(n)$ is at most $\mathcal{O}(n \log n)$.

C.2. Proof of Proposition 3.4

The number of FDFAs of size (n, m) is clearly at most $|DFA_2(n)| \cdot |DFA_2(m)|^n$. That is

$$|FDFA(n, m)| \leq |DFA_2(n)| \cdot |DFA_2(m)|^n.$$

It follows that

$$\mathrm{Ldim}(FDFA(n, m)) \leq \log(|DFA_2(n)| \cdot |DFA_2(m)|^n)$$

and using Ishigami and Tani (1997, Proposition 1), the desired bound follows.

C.3. Proof of Proposition 3.5

A run of an FDFA on (u, v) can be simulated by the run of an appropriate automaton in the class $DFA_3(n \cdot (m+1))$. To see this, input word $u\$v$ where $\$$ is a new symbol (recall we are assuming u, v are binary) to a DFA which has the same diagram as the FDFA but with an edge labeled with $\$$ from each state of the leading automaton to the initial state of the corresponding progress DFA. Now it follows by Proposition 3.2 that the consistency dimension of $FDFA(n, m)$ is at most $2^{\binom{n(m+1)}{2}}$.