The consistency dimension and distribution-dependent learning from queries (Extended abstract with appendices)

José L. Balcázar, Jorge Castro, David Guijarro
Dept. Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Campus Nord, 08034 Barcelona, Spain
Email: balqui|castro|david@lsi.upc.es *

Hans-Ulrich Simon
Fakultät für Informatik
Lehrstuhl Mathematik und Informatik
Ruhr-Universitaet Bochum
D-44780 Bochum
Email: simon@lmi.ruhr-uni-bochum.de

Abstract

We prove a new combinatorial characterization of polynomial learnability from equivalence queries, and state some of its consequences relating the learnability of a class with the learnability via equivalence and membership queries of its subclasses obtained by restricting the instance space. Then we propose and study two models of query learning in which there is a probability distribution on the instance space, both as an application of the tools developed from the combinatorial characterization and as models of independent interest.

*Work supported in part by the EC through the Esprit Program EU BRA program under project 20244 (ALCOM-IT) and the EC Working Group EP27150 (NeuroColt II) and by the spanish DGES (PB95-0787).

1 Introduction

The learning via queries model was introduced by Angluin [2, 3]. In this model, the learning algorithm obtains information about the target concept asking queries to a teacher or expert. The algorithm has to output an exact representation of the target concept in polynomial time. Target concepts are formalized as languages over a binary alphabet. Frequently, it is assumed that the teacher can answer correctly two kinds of questions from the learner: membership queries and equivalence queries[1]. Unless otherwise specified, all our discussions are in the "proper learning" framework where the hypotheses come from the same class as the target. A combinatorial notion, called approximate fingerprints, turned out to characterize precisely those concept classes that can be learned from polynomially many equivalence queries of polynomial size [4, 6].

Hellerstein et al. gave a beautiful characterization of polynomially (EQ,MQ)-learnable representation classes [7]. They introduced the notion of polynomial certificates for a representation class $\mathcal{R}$ and proved that $\mathcal{R}$ is polynomially learnable from equivalence and membership queries iff it has polynomial certificates.

The first main contribution of this paper is to propose a new combinatorial characterization of learnability from equivalence queries, surprisingly close to certificates, and quite different (and also simpler to handle) than the approximate fingerprints: the consistency dimension, that one can see as the analog of the VC dimension for query models.

Angluin [2, 3] showed that, when only approximate identification is required, equivalence queries can be substituted by a random sample. Thus, a PAC learning algorithm can be obtained from an exact learning algorithm that makes equivalence queries. In PAC learning, introduced by Valiant [10], one has to learn a target concept with high probability, in polynomial time (and, a fortiori, from a polynomial number of examples), within a certain error, under all probability distributions on the examples. Because of this last requirement, to learn under all distributions, PAC learning is also called distribution-free, or distribution-independent, learning. Distribution-independent learning is a strong requirement, but it can be relaxed to define PAC learning under specific distributions, or families of distributions. Indeed, several concept classes that are not known to be polynomially learnable, or known not to be polynomially learnable if RP $\neq$ NP, turn out to be polynomially learnable under some fixed distributions or families of distributions.

In comparison to PAC learning, one drawback of the query models is that they do not have this added flexibility of relaxing the "distribution-free" condition. The standard transformation sets them automatically at the "distribution-free" level. The second main contribution of this paper is the proposal of two learning models in which counterexamples are not provided adaptatively by a (helpful or treacherous) teacher, but instead are nonadaptatively sampled according to a probability distribution.

We prove that the distribution-free form of one of these models exactly coincides with standard learning from equivalence queries, while the other model is captured by the randomized version of the standard model. This allows us to extend, in a natural way, the query learning model to an explicit "distribution-free' setting where this restrictive condition can be naturally relaxed. Appendix B describes one example of a case in which this relaxation is useful, by exhibiting a query learning algorithm that works for a specific distribution, and points to others in the references. Some of the facts that we prove of these new models make use of the consistency dimension characterization proved earlier as the first contribution of the paper.

Our notation and terminology is standard. Most definitions will be given in the same section where they are needed. Generally, let X be a set, the *sample space*. A *concept* is a subset of X. A

[1]Such a teacher is called sometimes "minimally adequate".

concept class is a set $\mathcal{C} \subseteq 2^X$ of concepts. A *representation class* is a four-tuple $\mathcal{R} = (R, \Gamma, \Phi, \Sigma)$. Σ and Γ are finite alphabets. Strings of characters in Σ are used to describe elements of the sample space X, and strings of characters in Γ are used to encode concepts. $R \subseteq \Gamma^*$ is the set of strings that are concept encodings or *representations*. Let $\Phi : R \longrightarrow 2^{\Sigma^*}$ be a function that maps these representations into concepts over Σ. R may be thought of as a collection of names of concepts, and for any $r \in R$, $\Phi(r) \subseteq \Sigma^n$, for some n, is the concept named by r. The set $\{\Phi(r) : r \in R\}$ is the concept class associated with $\mathcal{R}$. For a concept r, we denote by $l(r)$ the minimum length of the representations of r. Thus, to each concept r we associate the length of the instances that make it up, n, and its own representation length, m, and we speak of a "doubly parameterized concept class". For such a class $\mathcal{C}$, we denote $\mathcal{C}_{n,m}$ the concepts corresponding to instance size n and with concept size at most m.

2 The Strong Consistency Dimension and its Applications

The proof, as it was given in [7], of the characterization of (EQ,MQ)-learning in terms of polynomial certificates implicitly contains concrete lower and upper bounds on the number of queries needed to learn $\mathcal{R}$. In Subsection 2.1, we make these bounds more explicit by introducing the so-called consistency dimension of $\mathcal{R}$ and writing the bounds in terms of this dimension (and some other parameters associated with $\mathcal{R}$). In Subsection 2.2, we define the notions of a"strong certificate" and of the"strong consistency dimension" and show that they fit the same purpose for EQ-learning as the former notions did for (EQ,MQ)-learning: we derive lower and upper bounds on the number of EQs needed to learn $\mathcal{R}$ in terms of the strong consistency dimension and conclude that $\mathcal{R}$ is polynomially EQ-learnable iff it has a polynomial strong certificate. In Subsection 2.3, we prove that the strong consistency dimension of a class equals the maximum of the consistency dimensions taken over all subclasses (induced by a restriction of the domain). This implies that the number of EQs needed to learn a concept class roughly equals the total number of EQs and MQs needed to learn the hardest subclass.

For ease of technical exposition, we need the following definitions. A *partially defined concept* C on domain Σ^n is a function from Σ^n to $\{0, 1, *\}$, where "$*$" stands for "undefined". Since partially defined concepts and samples can be identified in the obvious manner, we use the terms "partially defined concept" and "sample" interchangeably in the sequel. The *support* of C is defined as $\mathrm{supp}(C) = \{x \in \Sigma^n : C(x) \in \{0,1\}\}$. The *breadth* of C is defined as the cardinality of its support and denoted as $|C|$. The *size* of C is defined as the smallest size of a concept that is consistent with C. It is denoted as $|C|_{\mathcal{R}}$. Note that this definition coincides with the previous definition of size when C has full support Σ^n. Sample Q is called *subsample* of sample C (denoted as $Q \sqsubseteq C$) if $\mathrm{supp}(Q) \subseteq \mathrm{supp}(C)$ and Q, C coincide on $\mathrm{supp}(Q)$. Throughout this section, $\mathcal{R} = (\Sigma, \Delta, R, \mu)$ denotes a representation class defining a doubly parameterized concept class $\mathcal{C}$.

2.1 Certificates and Consistency Dimension

$\mathcal{R}$ *has polynomial certificates* if there exist two-variable polynomials p and q, such that for all $m, n > 0$, and for all $C : \Sigma^n \to \{0, 1\}$ the following condition is valid:

$$|C|_{\mathcal{R}} > p(n, m) \Rightarrow (\exists Q \sqsubseteq C : |Q| \leq q(m, n) \wedge |Q|_{\mathcal{R}} > m) \tag{1}$$

The *consistency dimension* of $\mathcal{R}$ is the following three-variable function: $\mathrm{cdim}_{\mathcal{R}}(n, m, M)$, where $M \geq m > 0$ and $n > 0$, is the smallest number d such that for all $C : \Sigma^n \to \{0, 1\}$ the following condition is valid:

$$|C|_{\mathcal{R}} > M \Rightarrow (\exists Q \sqsubseteq C : |Q| \le d \wedge |Q|_{\mathcal{R}} > m) \tag{2}$$

An obviously equivalent but quite useful reformulation of Condition (2) is

$$(\forall Q \sqsubseteq C : |Q| \le d \Rightarrow |Q|_{\mathcal{R}} \le m) \Rightarrow |C|_{\mathcal{R}} \le M. \tag{3}$$

In words: if each subsample of $C : \Sigma^n \to \{0,1\}$ of breadth at most d has a consistent representation of size at most m, then C has a consistent representation of size at most M.

The following result is (more or less) implicit in [7].

Theorem 2.1 *$cdim_{\mathcal{R}}(n,m,M) \le LC_{\mathcal{R}}^{EQ,MQ}(n,m,M) \le \lceil cdim_{\mathcal{R}}(n,m,M) \cdot \log |\mathcal{C}_{n,m}| \rceil$.*

Note that the lower and the upper bound are polynomially related because

$$\log |\mathcal{C}_{n,m}| \le m \cdot \log(1 + |\Delta|). \tag{4}$$

Clearly, Theorem 2.1 implies that $\mathcal{R}$ is polynomially (EQ,MQ)-learnable iff it has a polynomial certificate. We omit the proof of Theorem 2.1 in this brief abstract.

2.2 Strong Certificates and Strong Consistency Dimension

We want to adapt the notions "certificate" and "consistency dimension" to the framework of EQ-learning. Surprisingly, we can use syntactically almost the same notions, except for a subtle but striking difference: the universe of C will be extended from the set of all concepts over domain Σ^n to the corresponding set of partially defined concepts. This leads to the following definitions.

$\mathcal{R}$ has polynomial strong certificates if there exist two-variable polynomials p and q, such that for all $m, n > 0$, and for all $C : \Sigma^n \to \{0,1,*\}$ Condition (1) is valid.

The *strong consistency dimension* of $\mathcal{R}$ is the following three-variable function: $\mathrm{scdim}_{\mathcal{R}}(n,m,M)$, where $M \ge m > 0$ and $n > 0$, is the smallest number d such that for all $C : \Sigma^n \to \{0,1,*\}$ Condition (2) is valid. Again, instead of Condition (2), we can use the equivalent Condition (3). In words: if each subsample of $C : \Sigma^n \to \{0,1,*\}$ of breadth at most d has a consistent representation of size at most m, then C has a consistent representation of size at most M.

Theorem 2.2 *$scdim_{\mathcal{R}}(n,m,M) \le LC_{\mathcal{R}}^{EQ}(n,m,M) \le \lceil scdim_{\mathcal{R}}(n,m,M) \cdot \ln |\mathcal{C}_{n,m}| \rceil$.*

Proof. For the sake of brevity, let $q = \mathrm{LC}_{\mathcal{R}}^{EQ}(n,m,M)$ and $d = \mathrm{scdim}_{\mathcal{R}}(n,m,M)$.

We prove the first inequality by exhibiting an adversary that forces any learner to spend as many queries as given by the strong consistency dimension. The minimality of d implies that there is a sample C such that $(\forall Q \sqsubseteq C : |Q| \le d-1 \Rightarrow |Q|_{\mathcal{R}} \le m)$, but still $|C|_{\mathcal{R}} > M$. Thus, any learner issuing as equivalence queries hypotheses of size at most M fails to be consistent with C, and a counterexample from C can be safely provided; and up to the $d-1$ queries can be processed in this way, and still be consistent with some target of size at most m. Hence, at least d queries go by until an affirmative answer is obtained.

In order to prove $q \le \lceil d \ln |\mathcal{C}_{n,m}| \rceil$, we describe an appropriate EQ-learner A. A keeps track of the current version space $\mathcal{V}$ (which is $\mathcal{C}_{n,m}$ initially). For $i = 0, 1$, let

$$S_{\mathcal{V}}^i = \{x \in \Sigma^n : \text{the fraction of concepts } C \in \mathcal{V} \text{ with } C(x) = 1 - i \text{ is smaller than } 1/d\}.$$

In other words, a large fraction (at least $1 - 1/d$) of the concepts in $\mathcal{V}$ votes for output label i on instances from $S_{\mathcal{V}}^i$. Let $C_{\mathcal{V}}$ be the sample assigning label $i \in \{0,1\}$ to all instances from $S_{\mathcal{V}}^i$

and label "$*$" to all remaining instances (those without a clear majority). Let Q be an arbitrary but fixed subsample of $C_{\mathcal{V}}$ such that $|Q| \le d$. The definition of $S^i_{\mathcal{V}}$ implies (through some easy-to-check counting) that there exists a concept $C \in \mathcal{V} \subseteq \mathcal{C}_{n,m}$ that is consistent with Q. Applying Condition (3), we conclude that $|C_{\mathcal{V}}|_{\mathcal{R}} \le M$, i.e., there exists an $H \in \mathcal{C}_{n,M}$ that is consistent with $C_{\mathcal{V}}$. The punchline of this discussion is: if A issues the EQ with hypothesis H, then the next counterexample will shrink the current version space by a factor $1 - 1/d$ (or by a smaller factor). Since the initial version space contains $|\mathcal{C}_{n,m}|$ concepts and since A is done as soon as $|\mathcal{V}| \le 1$, a sufficiently large number of EQs is obtained by solving

$$(1 - 1/d)^q |\mathcal{C}_{n,m}| < e^{-q/d} |\mathcal{C}_{n,m}| \le 1$$

for q. Clearly, $q = \lceil d \ln |\mathcal{C}_{n,m}| \rceil$ is sufficiently large. •

Since the lower and the upper bound in Theorem 2.2 are polynomially related according to Inequality(4), we obtain

Corollary 2.3 *$\mathcal{R}$ is polynomially EQ-learnable iff it has a polynomial strong certificate.*

2.3 EQs Alone versus EQs and MQs

The goal of this subsection is to show that the number of EQs needed to learn a concept class is closely related to the total number of EQs and MQs needed to learn the hardest subclass. The formal statement of the main results requires the following definitions.

Let $\mathcal{S} = (S_n)_{n \ge 1}$ with $S_n \subseteq \Sigma^n$ be a family of subdomains. The *restriction of a concept* $C : \Sigma^n \to \{0, 1\}$ *to* S_n is the partially defined concept (sample) with support S_n which coincides with C on its support. The class containing all restrictions of concepts from $\mathcal{C}$ to the corresponding subdomain from $\mathcal{S}$ is called the *subclass of* $\mathcal{C}$ *induced by* $\mathcal{S}$ and denoted as $\mathcal{C}|\mathcal{S}$.

The notions polynomial certificate, consistency dimension, and learning complexity are adapted to the subclass of $\mathcal{C}$ induced by $\mathcal{S}$ in the obvious way. $\mathcal{R}|\mathcal{S}$ (in words: $\mathcal{R}$ restricted to $\mathcal{S}$) *has polynomial certificates* if there exist two-variable polynomials p and q, such that for all $m, n > 0$, and for all $C : \Sigma^n \to \{0, 1, *\}$ such that $\mathrm{supp}(C) = S_n$, Condition (1) is valid. The *consistency dimension* of $\mathcal{R}|\mathcal{S}$ is the following three-variable function: $\mathrm{cdim}_{\mathcal{R}}(S_n, m, M)$ is the smallest number d such that for all $M \ge m > 0, n > 0$, and for all $C : \Sigma^n \to \{0, 1, *\}$ such that $\mathrm{supp}(C) = S_n$, Condition (2) is valid. Again, instead of Condition (2), we can use the equivalent Condition (3).

Quantity $\mathrm{LC}^{EQ,MQ}_{\mathcal{R}}(S_n, m, M)$ is defined as the smallest total number of EQs and MQs needed to learn the class of concepts from $\mathcal{C}_{n,m}$ restricted to S_n with hypotheses from $\mathcal{C}_{n,M}$ restricted to S_n. Quantity $\mathrm{LC}^{EQ}_{\mathcal{R}}(S_n, m, M)$ is understood analogously. Note that

$$\mathrm{LC}^{EQ}_{\mathcal{R}}(S_n, m, M) \le \mathrm{LC}^{EQ}_{\mathcal{R}}(n, m, M) \tag{5}$$

is valid in general, because EQs become more powerful (as opposed to MQs which become less powerful) when we pass from the full domain to a subdomain (for the obvious reasons). We have the analogous inequality for the strong consistency dimension, but no such statement can be made for $\mathrm{LC}^{EQ,MQ}_{\mathcal{R}}$ or the consistency dimension.

The following result is a straightforward generalization of Theorem 2.1.

Theorem 2.4 $\mathit{cdim}_{\mathcal{R}}(S_n, m, M) \le \mathit{LC}^{EQ,MQ}_{\mathcal{R}}(S_n, m, M) \le \lceil \mathit{cdim}_{\mathcal{R}}(S_n, m, M) \cdot \log |(\mathcal{C}|\mathcal{S})_{n,m}| \rceil$.

We now turn to the main results of this section. The first one states that the strong consistency dimension of a class is the maximum of the consistency dimensions taken over all induced subclasses:

Theorem 2.5 *$scdim_{\mathcal{R}}(n,m,M) = \max_{S \subseteq \Sigma^n} cdim_{\mathcal{R}}(S,m,M)$.*

Proof. Let d_* be the smallest d which makes Condition (2) valid for all $C : \Sigma^n \to \{0,1,*\}$. Let $d_*(S)$ be the corresponding quantity when C ranges only over all samples with support S. It is evident that $d_* = \max_{S \subseteq \Sigma^n} d_*(S)$. The theorem now follows, because by definition $d_* = \mathrm{scdim}_{\mathcal{R}}(n,m,M)$ and $d_*(S) = \mathrm{cdim}_{\mathcal{R}}(S,m,M)$. •

Corollary 2.6 *1. A representation class $\mathcal{R}$ has a polynomial strong certificate iff all its induced subclasses have a polynomial certificate.*

2. A representation class is polynomially EQ-learnable iff all its induced subclasses are polynomially (EQ,MQ)-learnable.

The third result states that the number of EQs needed to learn a class equals roughly the total number of EQs and MQs needed to learn the hardest induced subclass.

Corollary 2.7
$\max_{S \subseteq \Sigma^n} LC_{\mathcal{R}}^{EQ,MQ}(S,m,M) \le LC_{\mathcal{R}}^{EQ}(n,m,M) \le \left\lceil \ln|\mathcal{C}_{n,m}| \cdot \max_{S \subseteq \Sigma^n} LC_{\mathcal{R}}^{EQ,MQ}(S,m,M) \right\rceil$.

Remember that gap $\ln|\mathcal{C}_{n,m}|$ is bounded above by $m \cdot \ln(1+|\Delta|)$.

3 Equivalence queries with a probability distribution

Let now $\mathcal{D}$ denote a class of probability distributions on X, the instance space for a computational learning framework. The two subsections of this section introduce respective variants of equivalence query learning that somehow take this distribution into account.

We briefly describe now the first one. In the ordinary model of EQ-learning $\mathcal{C}$, with hypotheses from $\mathcal{H}$, the counterexamples for incorrect hypotheses are arbitrarily chosen, and we can think of an intelligent adversary making these choices. *EQ-learning $\mathcal{C}$ from $\mathcal{D}$-teachers* (still with hypotheses from $\mathcal{H}$) proceeds as ordinary EQ-learning, except for the following important differences:

1. Each run of the learning algorithm refers to an arbitrary but fixed pair (C, D) such that $C \in \mathcal{C}$ and $D \in \mathcal{D}$.

2. The goal is *to learn C from the D-teacher*, i.e., C is considered as target concept (as usual), and the counterexample to an incorrect hypothesis H is randomly chosen according to the conditional distribution $D(\cdot, C \Delta H)$, where Δ denotes the symmetric difference of sets. Success is defined when this symmetric difference has zero probability.

Clearly, the more restricted the class $\mathcal{D}$ of probability distributions, the easier the task for the learner. In this extended abstract, we focus on the following three choices of $\mathcal{D}$.

- $\mathcal{D}_{all}$ denotes the class of all probability distributions on X. This is the most general case.

- $\mathcal{D}_{unif}$ denotes the class of distributions that are uniform on a subdomain $S \subseteq X$ and assign zero probability to instances from $X \setminus S$. This case will be relevant in a later section.

- $\mathcal{D} = \{D\}$ is the most specific case, where $\mathcal{D}$ constains only a single probability distribution D. We use it only briefly in the last section, and also in appendix B where we consider the universal distribution.

Loosely speaking, the main results of this section are as follows:

- The next subsection proves that, for $\mathcal{D} = \mathcal{D}_{all}$, EQ-learning from $\mathcal{D}$-teachers is exactly as hard (same number of queries) as the standard model. (This result is only established for deterministic learners.) Thus, we are not actually introducing yet one more learning model, but characterizing an existing, widely accepted, one in a manner that provides the additional flexibility of the probability distribution parameter. Thus we obtain a sensible definition of distribution-dependent equivalence-query learning. Appendix B gives a specific learning algorithm that shows that the model is somewhat useful.

- Then, the following section introduces a natural variant of it, that actually captures existing algorithms, and discusses the relationships between both models and with the standard EQ model.

3.1 Random versus Arbitrary Counterexamples

We use upper index $EQ[\mathcal{D}]$ to indicate that the D-teacher for some $D \in \mathcal{D}$ plays the role of the EQ-oracle.

Theorem 3.1 $LC^{EQ[\mathcal{D}_{all}]}(\mathcal{C}, \mathcal{H}) = LC^{EQ}(\mathcal{C}, \mathcal{H})$.

Proof. Let A be an algorithm which EQ-learns $\mathcal{C}$ from $\mathcal{D}$-teachers. with hypotheses from $\mathcal{H}$. Let $l \geq \mathrm{LC}^{EQ}(\mathcal{C}, \mathcal{H})$ be the largest number of EQs needed by A when we allow an adversary to return arbitrary counterexamples to hypotheses.[2] There must exist a concept $C \in \mathcal{C}$, hypotheses $H_0, \ldots, H_{l-2} \in \mathcal{C}$ and instances $x_0, \ldots, x_{l-2} \in X$, such that the learner issues the $l-2$ incorrect hypotheses H_i when learning target concept C, and the x_i are the counterexamples returned to these hypotheses by the adversary, respectively. We claim that there exists a distribution D such that, with probability $1-\delta$, the D-teacher returns the same counterexamples. This is technically achieved by setting $D(x_i) = (1-\alpha)\alpha^i$, for $i = 0, \ldots, l-3$, and $D(x_{l-2}) = \alpha^{l-2}$. An easy computation shows that the probability that the D-teacher presents another sequence of counterexample as the adversary equals $(l-2)\alpha$. Setting $\alpha = \delta/(l-2)$, the proof is complete. •

Therefore, the distribution-free case of our model coincides with standard EQ-learning.

Corollary 3.2 *Let $\mathcal{R} = (\Sigma, \Delta, R, \mu)$ be a representation class defining a doubly parameterized concept class $\mathcal{C}$. Then $LC_{\mathcal{R}}^{EQ[\mathcal{D}_{all}]}(n, m, M) = LC_{\mathcal{R}}^{EQ}(n, m, M)$ for all $M \geq m > 0, n > 0$.*

This obviously implies that learners for the distribution-free equivalence model can be transformed, through the standard EQ model, into distribution-free PAC learners. We note in passing that, applying the standard techniques directly on our model, we can prove the somewhat stronger fact that, for each individual distribution $\mathcal{D}$, a learner from $\mathcal{D}$-teachers can be transformed into an algorithm that PAC-learns over $\mathcal{D}$. A second side remark, that is relevant to the case discussed in appendix B, is that we can assume knowledge of a bound on the size of the target, by applying the usual trick of guessing it and increasing the guess whenever necessary (some necessary details will be provided in the full version).

[2] For the time being, there is no guarantee that A succeeds at all, because it expects the counterexamples to be given from a $\mathcal{D}$-teacher. We will however see subsequently that there exists a distribution which sort of simulates the adversary.

3.2 EQ-Learning from Random Samples

In this subsection, we discuss another relaxation of the ordinary EQ-learning model. *EQ-learning of $\mathcal{C}$ from $\mathcal{D}$-samples of size p and with hypotheses from $\mathcal{H}$* proceeds as ordinary EQ-learning, except for the following differences:

1. Each run of the learning algorithm refers to an arbitrary but fixed pair (C, D) such that $C \in \mathcal{C}$ and $D \in \mathcal{D}$.

2. The goal of the learner is to learn C from (ordinary) EQs and a sample P consisting of p examples drawn independently at random according to D and labeled correctly according to C. In other words, instead of EQ-learning C from scratch, the learner gets P as additional input.

Again the goal is to output a hypothesis for which the probability of disagreement with the target is zero; this time, the information about the distribution does not come from the counterexamples, but rather from the initial additional sample. We will show in this section that, for certain distributions, this model is strictly weaker than the model of EQ-learning from $\mathcal{D}$-teachers. However, in the distribution-free sense, it corresponds to the randomized version of the model described previously.

We first show that each algorithm for EQ-learning from $\mathcal{D}$-samples can be converted into a randomized algorithm for EQ-learning from $\mathcal{D}$-teachers such that the total number of EQs of the latter algorithm is not too large.

Theorem 3.3 *Let q be the number of EQs needed to learn $\mathcal{C}$ from $\mathcal{D}$-samples of size p and with hypothesis from $\mathcal{H}$. It holds, $LC^{EQ[\mathcal{D}]}(\mathcal{C}, \mathcal{H}) \leq (p+1)(p+q)$.* (Proof in Appendix A)

We show next an example that has an identification learning algorithm in the EQ from D-teachers learning model, but does not have such algorithm in the EQ learning from D-samples model.

A DNF$_n$ formula is any sum $t_1+t_2+\cdots+t_k$ of monomials, where each monomial t_i is the product of some literals chosen from $\{x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_n\}$. Let DNF $= \cup_n$ DNF$_n$ be the representation class of disjunctive normal form formulas. Let us consider the class $\mathcal{D}$ of distributions D defined in the following way. For each $n \geq 1$ two different words x_n and y_n are chosen. Distribution D is

$$\begin{aligned} D(x_n) &= 6/\pi^2(1/n^2 - 1/2^n) \\ D(y_n) &= 6/(\pi^2 2^n) \\ D(z_n) &= 0 \quad \text{for any word } z_n \text{ of length } n \text{ different from } x_n \text{ and } y_n. \end{aligned}$$

$\mathcal{D}$ is obtained by letting x_n and y_n run over all pairs of different words of length n.

A very easy EQ from $\mathcal{D}$-teachers algorithm identifies DNF. The algorithm has to do at most two trivial equivalence queries to know the value of the target formula f on x_n and y_n. First, it asks whether f is zero. If a counterexample e is given —e must be x_n or y_n— it will make a second query $f = t_e$?, where t_e is a monomial that only evaluates to one on e. When x_n, y_n and the values $f(x_n)$ and $f(y_n)$ are known the target formula is identified —it does not matter the value of the formula on other points: they have zero probability—.

We show now that there is a distribution $D \in \mathcal{D}$ such that DNF formulas are not identifiable in the EQ learning from D-samples model. Here we refer to learning DNF's of size polynomial in n from polynomially many equivalence queries of polynomial size, and with an extra initial sample of polynomial breadth. First we note that sampling according to $D \in \mathcal{D}$ there is a non-negligible probability of obtaining a sample that only contains copies of x_n.

Lemma 3.4 *For any polynomial q and $0 < \delta < 1$, there exists an integer k_0 such that for all $n \geq k_0$ the probability that a D-sample S of size $q(n, 1/\delta)$ does not contain y_n is greater than δ.*

The negative result for DNF formulas in the EQ learning from $\mathcal{D}$-samples model follows from next theorem.

Theorem 3.5 *There exist a distribution D in $\mathcal{D}$ such that DNF is not EQ learnable from D-samples.*

Proof. (Sketch: more details in final version) Angluin [4] showed there is no equivalence queries algorithm for DNF formulas. The proof takes advantage of Angluin's result. The idea is the following. For each equivalence queries algorithm we choose a word where, for some target formula and some consistent teacher, the hypothesis returned by the algorithm assigns a wrong value. Then, we define a probability distribution D in $\mathcal{D}$ which gives little probability, but greater than zero, to these words. Drawing examples according to such D there is a non-negligible probability that the chosen words do not appear in the sample (lemma 3.4). So, any equivalence queries candidate to learn DNF from D-samples has to output a right hypothesis without knowing the value of the target concept on the nonzero low probability words. It is not hard to see that this is impossible. •

4 Tighter queries

The remainder of the paper uses the machinery developed in section 2 to obtain stronger results relating the models of the previous section, under one more technical condition: that the learning algorithm knows the size of the target, and never queries hypotheses longer than that. Some important learning algorithms do not have this property, but there are still quite a few (among the exact learners from equivalence queries only) that work in sort of an incremental fashion that leads to this property. The results become interesting because they lead to a precise characterization of randomized learners from $\mathcal{D}$-teachers.

We first rewrite our combinatorial material of the previous section in an extremely useful, geometrically intuitive form (1-spheres), and prove that for $m = M$ these structures capture clearly the strong consistency dimension. Applications follow in the next subsection.

4.1 Strong Consistency Dimension and 1-Spheres

In this subsection, we prove that the strong consistency dimension also equals the size of the largest 1-sphere (to be defined subsequently) that can be represented by concepts chosen from $\mathcal{C}$. This implies that the number of EQs needed to learn a class roughly equals the size of the largest 1-sphere that is representable.

A popular method for getting lower bounds on the number of queries is to show that the class of target concepts contains a basic "hard-to-learn" combinatorial structure. Sometimes, a particularly easy case can be used: if the empty set is not representable but a number of singletons are, then a lower bound follows. In this Subsection, we consider a conceptually very simple structure: the so-called 1-spheres. These are actually a disguised (read isomorphic) version of sets of singletons, with the empty set simultaneously forbidden. Then we show that the strong consistency dimension equals the size of a largest 1-sphere that can be represented by $\mathcal{C}$.

To make the last statement precise, we need several definitions. Let S be a finite set, and $S_0 \subseteq S$. The *1-sphere with support S around center S_0*, denoted as $H^1_S(S_0)$ in the sequel, is the collection of sets $S_1 \subseteq S$ such that $|S_0 \Delta S_1| = 1$, where Δ denotes the symmetric difference of sets. In other words, $S_1 \subseteq S$ belongs to $H_S^1(S_0)$ if the Hamming distance between S_0 and S_1 is 1. Thus, it is formed by all the points at distance (radius) 1 from the center in Hamming space.

Let us now assume that $S \subseteq \Sigma^n$. Let S' be an arbitrary subset of S. The sample $C' : \Sigma^n \to \{0, 1, *\}$ *which represents S' (as a subset of S)* is the sample with support S that assigns label 1 to all instances from S', and label 0 to all instances from $S \setminus S'$. We say that $H_S^1(S_0)$ *is representable by* $\mathcal{C}_{n,[m:M]}$ if the following two conditions are valid:

(A) Let C_0 be the sample with support S which represents S_0. Then, $|C_0|_{\mathcal{R}} > M$.

(B) Each sample C_1 with support S, which represents a set $S_1 \in H_S^1(S_0)$, satisfies $|C_1|_{\mathcal{R}} \leq m$.

Thus, for the particular case of $M = m$, all points in Hamming space on the surface of the sphere are representable within size m but the center is not; just as the above-mentioned use of singletons, which form the 1-sphere centered on the empty set. The *size* of $H_S^1(S_0)$ is defined as $|S|$. We define the three-variable function $\mathrm{sph}_{\mathcal{R}}(n, m, M)$ as the size of the largest 1-sphere which is representable by $\mathcal{C}_{n,[m:M]}$. We only prove facts for the case $m = M$ here.

We now turn to the main result of this subsection.

Theorem 4.1 *$sph_{\mathcal{R}}(n, m, m) = scdim_{\mathcal{R}}(n, m, m)$.* (Proof in Appendix A)

4.2 Applications of the sphere number

In this subsection, $\mathcal{C}$ denotes a concept class. The main results of this section are derived without referring to a representation class $\mathcal{R}$. We will however sometimes apply a general theorem to the special case where the concept class consists of concepts with a representation of size at most m.

It will be convenient to adapt some of our notations accordingly. For instance, we say that *1-sphere $H_S^1(S_0)$ is representable by $\mathcal{C}$* if $S \subseteq X$ and the following two conditions are valid:

(A) $\mathcal{C}$ does not contain a hypothesis H that assigns label 1 to all instances in S_0 and label 0 to all instances in $S \setminus S_0$.

(B) For each $S' \in H_S^1(S_0)$, there exists a concept $C' \in \mathcal{C}$ that assigns label 1 to all instances in S' and label 0 to all insatnces in $S \setminus S'$.

The following notation will be used in the sequel. If $S = \{x_1, \ldots, x_s\}$, then $S_i = S_0 \Delta \{x_i\}$ for $i = 1, \ldots, s$. Thus, $S_1, \ldots, S_s$ are the sets belonging to $H_S^1(S_0)$. The concept from $\mathcal{C}$ which represents S_i in the sense of Condition (B) is denoted as C_i.

The *sphere number associated with $\mathcal{C}$*, denoted as $\mathrm{sph}(\mathcal{C})$, is the size of the largest 1-sphere that is representable by $\mathcal{C}$. Similar conventions are made for the learning complexity measure LC.

Theorem 4.2 *Let $\mathcal{C} = H_S^1(S_0)$ be a 1-sphere and D an arbitrary but fixed distribution on S. Then, $LC^{EQ[D]}(\mathcal{C}) \leq 1 + \lceil \log(1/\delta) \rceil$.* (Proof in Appendix A)

As the number of EQs needed to learn 1-spheres from arbitrary counterexamples equals the size s of the 1-sphere, and the upper bound in Theorem 4.2 does not depend on s at all, the model of EQ-learning from the D-teacher for a fixed distribution D is, in general, more powerful than the ordinary model. The gap between the number of EQs needed in both models can be made arbitrarily large.

Recall that $\mathcal{D}_{unif}$ denotes the class of distributions that are uniform on a subdomain $S \subseteq X$ and assign zero probability to instances from $X \setminus S$.

Theorem 4.3 *The following lower bound even holds for randomized learners:*

$$LC^{EQ}(\mathcal{C}) \geq LC^{EQ[\mathcal{D}_{unif}]}(\mathcal{C}) \geq (1-\delta)sph(\mathcal{C}).$$

Proof. The first inequality is trivial. We prove the second one. Let $S = \{x_1, \ldots, x_s\}$, and let $C_1, \ldots, C_s$ be the concepts from $\mathcal{C}$ used to represent $S_1, \ldots, S_s \in H^1_S(S_0)$, respectively. For $j = 1, \ldots, s$, let D_j be the probability distribution that assigns zero probability to x_j and is uniform on the remaining instances from S. Clearly, $D_j \in \mathcal{D}_{unif}$.

A learner must receive answer YES with probability at least $1-\delta$ of success for each pair (C, D), where $C \in \mathcal{C}$ is the target concept, and counterexamples are returned randomly according to $D \in \mathcal{D}$. It follows that, if target concept C_j is drawn uniformly at random from $\{C_1, \ldots, C_s\}$, and counterexamples are subsequently returned according to D_j, answer YES is still obtained with probability at least $1-\delta$ of success. Note that we randomize over the uniform distribution on the 1-sphere (random selection of the target concept), over the drawings of distribution D_j conditioned to the current sets of counterexamples, respectively, and over the internal coin tosses of the learner.

Assume w.l.o.g. that all hypotheses are consistent with the counterexamples received so far. Let C' be the next hypothesis, and $S' \subseteq S$ the subset of instances from S being labeled 1 by C'. Because $H^1_S(S_0)$ is representable by $\mathcal{C}$, S' must differ from S_0 on at least one element of S. If $S' = S_j$, then the learner receives answer YES. Otherwise, the set $U = (S' \Delta S_j) \setminus \{x_j\}$ is not empty. Note that the counterexample x_i to C' is picked from U uniformly at random. This leads to the removal of only C_i from the current version space $\mathcal{V}$.

The punchline of this discussion is that the following holds after the returnal of q counterexamples:

1. The current version space $\mathcal{V}$ contains $s-q$ candidate concepts from $\{C_1, \ldots, C_s\}$. They are (by symmetry) statistically indistinguishable to the learner.

2. The next hypothesis is essentially a random guess in $\mathcal{V}$, that is, the chance to receive answer YES is exactly $1/|\mathcal{V}|$. The reason is that, from the perspective of the learner, all candidate target concepts in $\mathcal{V}$ are equally likely.[3]

If answer YES is received before s EQs were issued, then only because it was guessed within $\mathcal{V}$ by chance. We can illustrate this by thinking of two players. Player 1 determines at random a number between 1 and s (the hidden target concept). Player 2 starts random guesses. The probability that the target number was determined after q guesses is exactly q/s. Thus, at least $(1-\delta)s$ guesses are required to achieve probability $1-\delta$ of success. •

Corollary 4.4 *Let $\mathcal{R} = (\Sigma, \Delta, R, \mu)$ be a representation class defining a doubly parameterized concept class $\mathcal{C}$. The following lower bound holds for all m and n, even for randomized learners:*

$$LC^{EQ}_{\mathcal{R}}(n, m, m) \geq LC^{EQ[\mathcal{D}_{unif}]}_{\mathcal{R}}(n, m, m) \geq (1-\delta)sph_{\mathcal{R}}(n, m, m)$$

This means that, considering learning algorithms that do not make queries longer than the size of the target, the information-theoretic barrier for EQ-learning from arbitrary counterexamples is still a barrier for EQ-learning from $\mathcal{D}_{unif}$-teachers. This negative result even holds when the learner is randomized, so that this implies that it applies as well to the model of EQ-learning from $\mathcal{D}$-samples, which has been proved earlier to be subsumed by randomized learners from $\mathcal{D}$-teachers.

[3]This might look unintuitive at first glance, because the learner does not necessarily draw the next hypothesis at random from $\mathcal{V}$ according to the uniform distribution. But notice that a random bit cannot be guessed with a probability of success larger than 1/2 no matter which procedure for "guessing" is applied. This is the kind of argument that we used.

References

[1] D. Angluin. Inference of reversible languages. *J. Assoc. Comput. Mach.*, 29:741–765, 1982.

[2] D. Angluin. Learning regular sets from queries and counterexamples. *Information and Computation*, 75:87–106, 1987.

[3] D. Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 1988.

[4] D. Angluin. Negative results for equivalence queries. *Machine Learning*, 5:121–150, 1990.

[5] J. Castro and J. L. Balcázar. Simple pac learning of simple decision lists. In *Sixth Internatinal Workshop, ALT'95*, pages 239–248, Fukuoka, Japan, 1995. LNAI. Springer.

[6] R. Gavaldà. On the power of equivalence queries. In *EUROCOLT'93*, pages 193–203. LNAI. Springer, 1993.

[7] L. Hellerstein, K. Pillaipakkamnatt, V. Raghavan, and D. Wilkins. How many queries are needed to learn? *Journal of the ACM*, 43(5):840–862, 1996.

[8] M. Li and P. Vitányi. Learning simple concepts under simple distributions. *SIAM Journal on Computing*, 20(5):911–935, 1991.

[9] H. Simon. Learning decision lists an trees with equivalence queries. In *Second European Conference, EuroCOLT'95*, pages 322–336, Barcelona, Spain, 1995. Lecture Notes in Artificial Intelligence.

[10] L. Valiant. A theory of the learnable. *Comm. ACM*, 27:1134–1142, 1984.

Appendix A: Some proofs

Proof. (of Theorem 3.3) Let A be a learning from $\mathcal{D}$-samples of size p algorithm that shows that q EQs queries are enough to learn $\mathcal{C}$ with hypothesis from $\mathcal{H}$. Let us consider a randomized learning from $\mathcal{D}$-teachers algorithm B that simulates A in the way explained below.

First, B builds samples $S_0, \ldots, S_p$, doing repetitively equivalence queries with the empty and total concepts and after that, it simulates the computation of A on these samples. Sample S_i is constructed asking for i counterexamples to the empty concept and $p-i$ counterexamples to the total concept. So, S_i contains exactly i positive examples. The order of the examples in S_i is defined by the choice of i random positions between 1 and p where positive examples are located. The relative order of positive (respectively negative) examples is the order they were obtained.

Let C be the target concept and let D be an arbitrary but fixed distribution in $\mathcal{D}$. Let $\langle x_1, \ldots, x_p \rangle$ be a sample with i positive examples. It will be generate by algorithm B with probability

$$\mathrm{Prob}_B(S_i = \langle x_1, \ldots, x_p \rangle) = \frac{D(x_1) \cdots D(x_p)}{D(C)^i (1-D(C))^{p-i} \binom{p}{i}}.$$

In the denominator $D(C)$ and $1 - D(C)$ are respectively the normalization factors of the positive and negative counterexamples, and the combinatorial factor comes from the randomized process of B that defines the order in S_i. We note that this number is exactly the probability of obtaining $\langle x_1, \ldots, x_p \rangle$ when a sample with i positive examples is drawn according to D. In other words, if Π_i denotes de event formed by the samples of size p with i positive examples,

$$\mathrm{Prob}_B(S_i = \langle x_1, \ldots, x_p \rangle) = \mathrm{Prob}_A(S_i = \langle x_1, \ldots, x_p \rangle / \Pi_i).$$

The simulation carried out by B does not work only when $S_0, S_1, \ldots, S_p$ are all of them samples where algorithm A fails. We can write the probability of failure of B as the product

$$\prod_{i=0}^{p} \mathrm{Prob}_B(A \text{ fails on } S_i).$$

By the discussion above, this product can be rewrite as

$$\prod_{i=0}^{p} \mathrm{Prob}_A(A \text{ fails on } S_i / \Pi_i).$$

By lemmma 4.5 this product can be bounded by the following sum

$$\sum_{i=0}^{p} \mathrm{Prob}_A(A \text{ fails on } S_i/\Pi_i)\mathrm{Prob}_A(\Pi_i) = \sum_{i=0}^{p} \mathrm{Prob}_A(A \text{ fails on } \Pi_i) = \mathrm{Prob}_A(A \text{ fails}).$$

As we wanted to show, this probability is, by hypothesis, less than δ. •

The following lemma used in the proof states a well known property of real numbers.

Lemma 4.5 *Let $x_1, \ldots, x_n$ and $\lambda_1, \ldots, \lambda_n$ be real numbers in $[0,1]$ with $\lambda_1 + \cdots + \lambda_n = 1$. Then,*

$$\prod_{i=1}^{n} x_i \leq \sum_{i=1}^{n} \lambda_i x_i$$

Proof. (of Theorem 4.1) For the sake of brevity, let $d = \mathrm{scdim}_{\mathcal{R}}(n,m,m)$ and $s = \mathrm{sph}_{\mathcal{R}}(n,m,m)$.

Let $H_S^1(S_0)$ be a largest 1-sphere that is representable by $\mathcal{C}_{n,[m:m]}$. Thus, $|S| = s$. In order to prove $d \geq s$, we assume for the sake of contradiction $d < s$. Consider the sample C_0 with support S that represents S_0. By Condition (A), $|C_0|_{\mathcal{R}} > m$. According to Condition (2) applied to C_0, there exists a subsample $Q \sqsubseteq C_0$ such that $|Q| \leq d < s$ and $|Q|_{\mathcal{R}} > m$. Let $S_Q = \mathrm{supp}(Q) \subset S$. Let Q_1 be a sample with support S that totally coincides with Q on S_Q, and coincides with C_0 on $S_0 \setminus S_Q$ except for one instance. Clearly, Q_1 represents a set $S_1 \in H_S^1(S_0)$. By Condition (B), $|Q_1|_{\mathcal{R}} \leq m$. Since $|Q|_{\mathcal{R}} \leq |Q_1|_{\mathcal{R}}$, we arrived at a contradiction. Note that this inequality works as well for $M > m$.

Finally, we prove $s \geq d$. It follows from the minimality of d and Condition (2) that there exists a sample $C : \Sigma^n \to \{0, 1, *\}$ such that the following holds:

1. $|C|_{\mathcal{R}} > m$.

2. $\forall Q \sqsubseteq C : (|Q| \leq d - 1 \Rightarrow |Q|_{\mathcal{R}} \leq m)$.

A sample satisfying these conditions is called *core* in the sequel. Let C_0 be a *minimal core*, i.e., a core without a proper subsample that is still a core. Since the second condition remains true when we substitute a subsample $C \sqsubseteq C_0$ for C_0, the first condition is violated for all proper subsamples of C_0. Let S denote the support of C_0. Note that $|S| \geq d$ (because otherwise the two conditions for core C_0 become contradictory). Let $S_0 \subseteq S$ be the set represented by C_0. We claim that $H_S^1(S_0)$ is representable by $\mathcal{C}_{n,[m:m]}$ (which would conclude the proof). Condition (A) is obvious because $|C_0|_{\mathcal{R}} > m$. Condition (B) can be seen as follows. For each $x \in S$, define C_x as the proper subsample of C with support $S \setminus \{x\}$, and C'_x as the sample with support S that coincides with C on $S \setminus \{x\}$, but disagrees on x. Because each C_x is a proper subsample of C, it follows that $|C_x|_{\mathcal{R}} \leq m$ for all $x \in S$. We conclude that the same remark applies to samples C'_x, since a concept that is consistent with C_x, but inconsistent with C, must be consistent with C'_x. Finally note that the samples C'_x, $x \in S$, are exactly the representations of the sets in $H_S^1(S_0)$, respectively. •

Proof. (of Theorem 4.2) Let $S = \{x_1, \ldots, x_s\}$, and let $C_1, \ldots, C_s$ be the concepts from $\mathcal{C}$ used to represent $S_1, \ldots, S_s \in H_S^1(S_0)$, respectively. Let $H_1, \ldots, H_s$ be a permutation of $C_1, \ldots, C_s$ sorted according to increasing values of $D(x_i)$. Consider the EQ-learner which issues its hypotheses in this order. It follows that as long as there exist counterexamples of a strictly positive probability, the probability that the teacher returns the counterexample x_j associated with the target concept C_j is at least 1/2 per query. Thus, the probability that the target is not known after $\lceil \log(1/\delta) \rceil$ EQs is at most δ. Thus, with probability at least $1 - \delta$, one more query suffices to receive answer YES. •

Appendix B

5 Reversible automata and the universal distribution

The purpose of this section is to present an example of a new learning algorithm for the distribution-dependent teacher model introduced in this paper. It works for the universal distribution, and can be seen as a reverse-engineering of a known simple-PAC algorithm that gets improved into (probably) exact learning.

It can be proved that there exists an universal enumerable distribution, denoted by $\mathbf{m}$, that multiplicatively dominates each enumerable distribution. This universal distribution has many important properties. One of them is that

$$\mathbf{m}(x) = 2^{-K(x)+O(1)},$$

where $K(x)$ denotes the Kolmogorov complexity of x (see [8]). Thus, this property can be interpreted as follows: Under $\mathbf{m}$, easily describable objects have high probability, and complex objects have low probability.

Specifically, we prove that reversible languages are learnable from $\mathbf{m}$-teachers. We follow [1] and [8] for definitions and basic results. Another example of learning from universal distribution-dependent teachers (actually in the stronger model of distributed initialization) appears in [5], based on a query learning algorithm that is essentially that of [9]).

Reversible languages were defined by Angluin in [1]. Let $A = (Q, I, F, \Sigma, \delta)$ be a finite automaton, where I is a set of initial states. The reverse of δ, denoted δ^r, is defined by:

$$\delta^r(q, a) = \{q' : q \in \delta(q', a)\} \quad \text{for all } a \in \Sigma, q \in Q.$$

The reverse of the automaton A is $A^r = (Q, F, I, \Sigma, \delta^r)$. So, we obtain A^r from A by interchanging the initial and the final states and reversing each transition. It is easy to see that $L(A^r) = L(A)^r$.

We say that the automaton A is 0-*reversible* iff both A and A^r are deterministic. Alternatively, A is 0-reversible iff A is deterministic, has at most one final state, and $\delta(q_1, a) = \delta(q_2, a)$ implies $q_1 = q_2$. A language L is said to be 0-reversible iff there exists a 0-reversible automaton A such that $L = L(A)$. For example, the set of strings containing an even number of zeroes is a 0-reversible language.

The notion of 0-reversibility can be generalized as follows. We say that the deterministic automaton $A = (Q, q_0, F, \Sigma, \delta)$ is k-*reversible* iff in the possibly nondeterministic automaton A^r, if two different states q_1, q_2 are both initial states, or if both $q_1, q_2 \in \delta^r(q_3, a)$ for some $a \in \Sigma$, then no string x of length k satisfies simultaneously $\delta^r(q_1, x) \neq \emptyset$ and $\delta^r(q_2, x) \neq \emptyset$. This guarantees that any nondeterministic choice in the operation of A^r can be resolved by looking ahead k symbols past the current one. A language is k-reversible iff it is accepted by a k-reversible automaton. For each k, the language $1^{k+1}1^\star$ is $(k+1)$-reversible but not k-reversible (see [1]). In general, a language L is said to be reversible iff L is k-reversible for some k.

Angluin [1] defines a *characteristic sample* of a k-reversible language L to be a finite subset S_k of L with the property that L is the smallest k-reversible language that contains S_k. It holds

Theorem 5.1 ([1]) *Let L be any k-reversible language. Then there exists a characteristic sample S_k for L.*

In the same paper, Angluin describes, for each k, an algorithm, denoted k-*RI*, to infer k-reversible languages from finite sets of strings. Roughly speaking, algorithm k-*RI* proceeds as follows. Let S be a nonempty finite set of strings. On input S, k-RI constructs the trivial prefix tree deterministic automaton A_0 that recognizes S, and finds the finest partition π_f of the states of A_0 such that A_0/π_f is k-reversible. It then outputs a canonical acceptor for $L(A_0/\pi_f)$ and halts.

To find π_f, k-RI begins with the trivial partition and repeatedly merges any two classes that violate the conditions for k-reversibility in the quotient automaton A_0/π. When no pair of classes is found to violate these conditions, the resulting partition is π_f.

Theorem 5.2 ([1]) *Let S be a nonempty finite set of strings, and let A_f be the automaton output by k-RI on input S. Then $L(A_f)$ is the smallest k-reversible language containing S. The running time of k-RI is $O(kn^3)$, where n is one more than the sum of the lengths of the strings in S.*

From theorems 5.1 and 5.2 Angluin shows that reversible languages can be identified *in the limit*: for any sequence that eventually contains all possible examples, the process produces a sequence of guesses that eventually converges to a correct one.

It is known that even fixed length 0-reversible languages do have approximate fingerprints [4], so that they are not learnable from polynomially many polynomial size equivalence queries. We define next the class of simple reversible languages. Li and Vitányi showed in [8] that this class is learnable in their simple-PAC model. We will show that in fact, simple reversible languages are EQ-learnable from $\mathbf{m}$-teachers.

Let A be a deterministic finite automaton and let $l(A)$ denote the description length of A. Let $c > 0$ be an integer. A path from the initial state to a final state is c-simple iff it has Kolmogorov complexity bounded by $c \log l(A)$. We say that the automaton A is c-simple iff each state of A lies on a c-simple path. The class of c-simple automata properly includes the class of deterministic automata that have Kolmogorov complexity descriptions of size $(c-1)\log l(A)$ (see [8]).

We denote $\mathcal{SR}_c^k$ the class of c-simple k-reversible automata A such that $k \le \log l(A)$, and $\mathcal{SR}_c$ the union of the classes $\mathcal{SR}_c^k$, for $k \ge 0$. We say that the language L is c-simple k-reversible (respectively c-simple reversible) iff L is accepted by an automaton in $\mathcal{SR}_c^k$ ($\mathcal{SR}_c$). For example, for any $k \ge 0$, the languages $L^1_{k,n} = \{0^j 1^n : j \ge k\}$ and $L^2_{k,n} = \{1^k u : |u|_0 \bmod n = 0\}$, where $|u|_0$ denotes the number of zeroes in string u, can be shown to be c-simple k-reversibles but not $(k-1)$-reversibles. It is not difficult to see that $\mathcal{SR}_0^c$ contains automata with Kolmogorov complexity descriptions much larger than $\log l(A)$.

With some little modifications in the original proof of theorem 5.1, one can show that simple reversible languages have characteristic samples of low Kolmogorov complexity. The following result is more-or-less implicit in [8] (through their references to [1]), but there is no single consistent source for it. We will provide a complete proof in the final version.

Theorem 5.3 *Let L be a c-simple k-reversible language and let $A = (Q, q_0, F, \Sigma, \delta)$ be an acceptor for L in $\mathcal{SR}_c^k$. There exists a characteristic sample S_k for L such that all strings in S_k have Kolmogorov complexity $(2c+3)\log l(A)$.*

Now we are ready to prove that the class $\mathcal{SR}_c$ is EQ learnable from $\mathbf{m}$-teachers. First, we show a learning algorithm for $\mathcal{SR}_c^k$. We assume that the learner has access to an upper bound n on the size $l(A)$ of the target. Since the algorithm will actually halt only upon an accepting equivalence query, the standard technique of growing guesses at the size allows us to dispose of this assumption to exactly fit our original definition.

Theorem 5.4 *The class $\mathcal{SR}_c^k$ is EQ learnable from $\mathbf{m}$-teachers.* **Proof.** *We consider the following learning algorithm. Here A denotes the target automaton, E is the set of counterexamples, δ has* *the usual meaning, and* k-RI *is the Angluin's inference algorithm for* k*-reversible languages. The target is assumed to be a* k*-reversible language, with* $2^k \le l(A) \le n$*.*

Learn_SR (c, k, δ) **returns** B : *automaton*
 $E := \emptyset$; $B := \emptyset$; $i := 0$;
 Ask the equivalence query $B = A$*?*
 while *the teacher does not reply "yes"* **and** $i \le n^{2c+5}/\delta$ **do**
  Let e *be the new counterexample*
  $E := E \cup \{e\}$
  $B := k\text{-RI}(E)$
  $i := i + 1$
  Ask the equivalence query $B = A$*?*
 endwhile
 return B
endfunction

By the previously recalled results, all counterexamples received by the algorithm are positive. Its analysis is quite straightforward; essentially, it consists in proving by standard methods that all necessary examples appear with probability greater than $1 - \delta$ *along the way, and that after* n^{2c+5}/δ *queries this would ensure that* E *contains a characteristic sample for* $L(A)$*.*

We consider now the class $\mathcal{SR}_c$: simple k-reversible automata A where $k \le \log l(A)$. The following algorithm, where negative examples can be provided by the teacher, tries to learn c-simple reversible automata. Here, P denotes the set of positive counterexamples given by the teacher and p is the cardinality of P.

Learn_ESR (c, δ) **returns** B : automaton
 $k := 0$; $P := \emptyset$; $B := \emptyset$; $p := 0$;
 Ask the equivalence query $B = A$?
 while the teacher does not reply "yes" **and** $p \le l(A)^{2c+5}/\delta$ **do**
  Let e be the new counterexample
  if e is positive
   then $P := P \cup \{e\}$; $p := p + 1$
   else $k := k + 1$
  endif
  $B :=$ *k-RI*(P)
  Ask the equivalence query $B = A$?
 endwhile
 return B
endfunction

The algorithm *Learn_ESR* runs algorithms *k-RI* repeatedly on the set P of positive counterexamples till the target automaton has been learned or the bound for p has been reached. When the teacher provides a positive counterexample e, *Learn_ESR* puts e in P and increases the value of p. If a negative counterexample is given, *Learn_ESR* only increases the value of the variable k; indeed, *k-RI* constructs the smallest k-reversible superlanguage of the positive examples, so for the correct k only positive examples can be ever seen: a negative one points out that k is too small. Standard arguments show now that, if counterexamples are provided according to the universal distribution, with probability at least $1 - \delta$, *Learn_SR* outputs an exact representation of A.

Theorem 5.5 *The class $\mathcal{SR}_c$ is learnable from EQ to* $\mathbf{m}$*-teachers.*

Thus, "technologically speaking", our combinatorial material here is essentially from [8], but their proofs repeatedly need to resort to the details of the proofs in [1], and therefore are only partially described (and some necessary hypothesis only informally stated, if at all). Our distribution-dependent teachers introduced here make it easier to design an algorithm learning from equivalence queries to $\mathbf{m}$-teachers than a standard simple PAC one. In fact, most examples of simple PAC learning given in [8] have easier algorithms for EQ learning from $\mathbf{m}$-teachers.

This can be interpreted as follows: the model based on distribution-dependent teachers incorporates in a single concept both the active interaction of query learning and the universal distribution sampling, and thus allows for the direct use of query learning algorithms as subroutines; so, interaction issues and sampling issues are dealt with in a more structured fashion. Then interaction becomes unnecessary via the standard query-to-PAC transformation. On the other hand, the direct construction of simple PAC learning algorithms frequently requires to embody into the same proof both the arguments regarding the sampling and the arguments that prove the necessary nonadaptiveness to bypass the interaction.

Of course, algorithms in our models give somewhat stronger theorems since, on the one hand, by the transformation outlined in the previous section, from these algorithms one immediately obtains simple PAC algorithms; but, in principle, the converse needs to specifically argument each case, since simple PAC algorithms might not be exact, as requested by our models.