The consistency dimension and distribution-dependent learning from queries [*]

José L. Balcázar, Jorge Castro, David Guijarro
Dept. LSI, Universitat Politècnica de Catalunya,
Campus Nord, 08034 Barcelona, Spain, {balqui,castro,david}@lsi.upc.es

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

March 30, 2000

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 (Koala) and by the German-Israeli Foundation for Scientific Research and Development, grant I-403-001.06/95.

1 Introduction

The main models of learning via queries were introduced by Angluin [1, 2]. In these models, 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 an 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 concept. 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 [3, 5].

The essential intuition behind that fact is that the existence of queries that eliminate an inverse polynomial factor of the number of possibilities for the target concept at every step, is not only clearly sufficient, but also necessary to learn: if no such queries are available then adversaries can be designed that force any learner to spend too many queries in order to identify the target. This intuition can be fully formalized along the lines of the cited works; the formalization can be found in [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 strong consistency dimension.

Angluin [1, 2] showed that, when only approximate identification is required, equivalence queries can be replaced 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 [9], 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 distribution 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 adaptatively provided by a (helpful or treacherous) teacher, but instead are nonadaptatively

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

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. 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. We assume familiarity with the query-learning model. Most definitions will be given in the same section where they are needed. Generally, let X be a set, called *instance space* or *domain* in the sequel. A *concept* is a subset of X, where we prefer sometimes to regard C as a function from X to $\{0,1\}$. A *concept class* is a set $\mathcal{C} \subseteq 2^X$ of concepts. An element of X is called an *instance.* A pair (x,b), where $b \in \{0,1\}$ is a binary label, is called *example for concept* C if $C(x) = b$. A *sample* is a collection of labeled instances. Concept C is said to be *consistent with sample* S if $C(x) = b$ for all $(x,b) \in S$.

A *representation class* is a four-tuple $\mathcal{R} = (\Sigma, \Delta, R, \mu)$, where Σ and Δ are finite alphabets. Strings of characters in Σ are used to describe elements of the domain X, and strings of characters in Δ are used to encode representations of concepts. We denote by $R \subseteq \Delta^*$ the set of strings that are valid concept encodings or *representations.* Let $\mu : R \longrightarrow 2^{\Sigma^*}$ be a function that maps these representations into concepts over Σ. For ease of technical exposition, we assume that, for each $r \in R$ there exists some $n \geq 1$ such that $\mu(r) \subseteq \Sigma^n$. Thus each concept with a representation in R has a domain of the form Σ^n (as opposed to domain Σ^*).[2] The set $\mathcal{C} = \{\mu(r) : r \in R\}$ is the concept class associated with $\mathcal{R}$.

We define the *size of concept* $C : \Sigma^n \rightarrow \{0,1\}$ *w.r.t. representation class* $\mathcal{R}$ as the length of the shortest string $r \in R$ such that $C = \mu(r)$, or as ∞ if C is not representable within $\mathcal{R}$. This quantity is denoted by $|C|_{\mathcal{R}}$. With these definitions, $\mathcal{C}$ is a "doubly parameterized class", that is, it is partitioned into sets $\mathcal{C}_{n,m}$ containing all concepts from $\mathcal{C}$ with domain Σ^n and size at most m. The kind of query-learning considered in this paper is proper in the sense that concepts and hypotheses are picked from the same class $\mathcal{C}$. We will however allow that the size of an hypothesis exceeds the size of the target concept. The number of queries needed in the worst case to obtain an affirmative answer from the teacher, or "learning complexity", given that the target concept belongs to $\mathcal{C}_{n,m}$ and that the hypotheses of the learner may be picked from $\mathcal{C}_{n,M}$, is denoted by $\mathrm{LC}^{\mathcal{O}}_{\mathcal{R}}(n,m,M)$, where $\mathcal{O}$ specifies the allowed query types. In this paper, either $\mathcal{O} = EQ$ or $\mathcal{O} = (EQ, MQ)$. We speak of polynomial $\mathcal{O}$-learnability if $\mathrm{LC}^{\mathcal{O}}_{\mathcal{R}}(n,m,M)$ is polynomially bounded in n, m, M.

We close this section with the definition of a version space. At any intermediate stage of a query-learning process, the learner knows (from the teacher's answers received so far) a sample S for the target concept. The *current version space* $\mathcal{V}$ is the set of all concepts from $\mathcal{C}_{n,m}$ which are consistent with S. These are all concepts being still conceivable as target concepts. Therefore, a learning algorithm is a strategy that reduces the version space by stages until it becomes a singleton set.

[2] This is a purely technical restriction that allows us to present the main ideas in the most convincing way. It is easy to generalize the results in this paper to the case of domains with strings of varying length.

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 polynomial strong certificates. 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:

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

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

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

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 $C : \Sigma^n \rightarrow \{0,1\}$ 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 1 *$cdim_{\mathcal{R}}(n,m,M) \le LC_{\mathcal{R}}^{EQ,MQ}(n,m,M) \le \lceil cdim_{\mathcal{R}}(n,m,M) \cdot \log |\mathcal{C}_{n,m}| \rceil + 1$.*

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

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

Clearly, Theorem 1 implies that $\mathcal{R}$ is polynomially (EQ,MQ)-learnable iff it has polynomial certificates. We omit the proof of Theorem 1: it is quite straightforward after [7].

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 $C : \Sigma^n \rightarrow \{0,1,*\}$ Condition (1) is valid.

Accordingly, 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 $d > 0$ such that for all $C : \Sigma^n \rightarrow \{0,1,*\}$ Condition (2) is valid. Again, instead of Condition (2), we can use the equivalent Condition (3). In words: if each subsample of $C : \Sigma^n \rightarrow \{0,1,*\}$ 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 *$scdim_{\mathcal{R}}(n,m,M) \le LC_{\mathcal{R}}^{EQ}(n,m,M) \le \lceil scdim_{\mathcal{R}}(n,m,M) \cdot \ln |\mathcal{C}_{n,m}| \rceil + 1$.*

Proof. For the sake of brevity, let $q + 1 = \mathrm{LC}_{\mathcal{R}}^{EQ}(n,m,M)$ 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 up to $d-1$ equivalence queries with hypotheses of size at most M, fails to be consistent with C, and a counterexample from C can be provided such that there is still at least one consistent concept of size at most m (a potential target concept). 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 very 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 $S^i_{\mathcal{V}}$ and label "$*$" to all remaining instances (those without a so 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 the factor $1 - 1/d$ (or by a smaller factor). Since the initial version space contains $|\mathcal{C}_{n,m}|$ concepts, we will obtain a singleton version space ($|\mathcal{V}| = 1$) making q equivalence queries, by solving for q the following inequality,

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

Clearly, $q = \lceil d \ln |\mathcal{C}_{n,m}| \rceil$ is sufficiently large. Note that a single extra equivalence query will force an affirmative answer. •

Since the lower and the upper bound in Theorem 2 are polynomially related according to Inequality (4), we obtain

Corollary 3 *$\mathcal{R}$ is polynomially EQ-learnable iff it has polynomial strong certificates.*

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 result requires the following definitions.

Let $\mathcal{S} = (S_n)_{n\ge1}$ 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 of polynomial certificates, 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 $d > 0$ 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 $\mathrm{LC}_{\mathcal{R}}^{EQ,MQ}$ or the consistency dimension.

The following result is a straightforward generalization of Theorem 1.

Theorem 4

$$cdim_{\mathcal{R}}(S_n,m,M) \le LC_{\mathcal{R}}^{EQ,MQ}(S_n,m,M) \le \lceil cdim_{\mathcal{R}}(S_n,m,M)\cdot \log|(\mathcal{C}|\mathcal{S})_{n,m}|\rceil + 1.$$

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

Proof. Let d_* be the smallest $d>0$ 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 6 *1. A representation class $\mathcal{R}$ has polynomial strong certificates iff all its induced subclasses have polynomial certificates.*

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

The next 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 7

$$\max_{S\subseteq\Sigma^n} LC_{\mathcal{R}}^{EQ,MQ}(S,m,M) \le LC_{\mathcal{R}}^{EQ}(n,m,M) \le \left\lceil \ln|\mathcal{C}_{n,m}|\cdot \max_{S\subseteq\Sigma^n} LC_{\mathcal{R}}^{EQ,MQ}(S,m,M)\right\rceil + 1$$

Proof. The first inequality is obtained from (5) as follows:

$$\max_{S\subseteq\Sigma^n} \mathrm{LC}_{\mathcal{R}}^{EQ,MQ}(S,m,M) \le \max_{S\subseteq\Sigma^n} \mathrm{LC}_{\mathcal{R}}^{EQ}(S,m,M) \le \mathrm{LC}_{\mathcal{R}}^{EQ}(n,m,M).$$

Putting Theorems 2, 5, and 4 together, we get:

$$\begin{aligned}
\mathrm{LC}_{\mathcal{R}}^{EQ}(n,m,M) &\le \lceil \ln|\mathcal{C}_{n,m}|\cdot \mathrm{scdim}_{\mathcal{R}}(n,m,M)\rceil + 1\\
&= \lceil \ln|\mathcal{C}_{n,m}|\cdot \max_{S\subseteq\Sigma^n} \mathrm{cdim}_{\mathcal{R}}(S,m,M)\rceil + 1\\
&\le \lceil \ln|\mathcal{C}_{n,m}|\cdot \max_{S\subseteq\Sigma^n} \mathrm{LC}_{\mathcal{R}}^{EQ,MQ}(S,m,M)\rceil + 1.
\end{aligned}$$

●

Remember that the 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 such distributions 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 $D \in \mathcal{D}$, and to a given confidence parameter $0 < \delta < 1$.

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 $D(\cdot|C \oplus H)$, where $\oplus$ denotes the symmetric difference of sets. Success is defined when this symmetric difference has zero probability. The learner must achieve a success probability of at least $1 - \delta$.

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.

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.

- In the next section, we introduce a combinatorial quantity, called the sphere number, and show that it represents an information-theoretic barrier in the model of EQ-learning from $\mathcal{D}_{unif}$-teachers (even for randomized learning algorithms). However, this barrier is overcome for each fixed distribution D in the model of EQ-learning from the D-teacher.

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. For instance, $\mathrm{LC}^{EQ[\mathcal{D}]}(\mathcal{C}, \mathcal{H}, \delta)$ denotes the number of queries needed to achieve a success probability of at least $1 - \delta$ when EQ-learning $\mathcal{C}$ with hypotheses from $\mathcal{H}$ from $\mathcal{D}$-teachers.

Theorem 8 *For all* $0 < \delta < 1$, $LC^{EQ[\mathcal{D}_{all}]}(\mathcal{C}, \mathcal{H}, \delta) = LC^{EQ}(\mathcal{C}, \mathcal{H})$.

Proof. Direction $\leq$ is obvious. We prove the converse direction. 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.[3] Since $\mathrm{LC}^{EQ}(\mathcal{C})$ is defined taking all algorithms into account, we loose no generality in assuming that A always queries hypotheses that are consistent with previous counterexamples, so that all the counterexamples received along any run are different. 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 $l-1$ 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 at least $1-\delta$, 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 counterexamples as the adversary is at most $(l-2)\alpha$. Setting $\alpha = \delta/(l-2)$, the proof is complete. ●

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

Corollary 9 *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 $\mathcal{D}$.

3.2 EQ-Learning from Random Samples

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

[3] 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.

1. Each run of the learning algorithm refers to an arbitrary but fixed pair (C, D) such that $C \in \mathcal{C}$ and $D \in \mathcal{D}$, and to a given confidence parameter $0 < \delta < 1$.

2. The goal of the learner is to learn C from a sample P consisting of p examples drawn independently at random according to D and labeled correctly according to C, and using a special type of EQ-queries where the teacher can choose any counterexample only if the symmetric difference has positive D-probability. In other words, instead of EQ-learning C from scratch, the learner gets P as additional input and the teacher must give an affirmative answer when the set of counterexamples has zero probability under the distribution D. The learner must obtain an affirmative answer with a probability at least $1 - \delta$ of success.

Again the goal is to output a hypothesis for which the probability of disagreement with the target concept is zero; this time, the information about the distribution does not come from the counterexamples, but rather from the initial additional sample. Observe that the teacher can choose a zero probability counterexample as long as there is another counterexample with positive probability. One may wonder if this model is totally artificial; but we note that there are some learning algorithms in the literature that fit perfectly on it, for instance [4].

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 as those of the previous subsection, at the cost of a moderate overhead in the number of queries.

Theorem 10 *Let $q(\delta)$ be the number of EQs needed to learn $\mathcal{C}$ from $\mathcal{D}$-samples of size p and with hypothesis from $\mathcal{H}$ (and probability at least $1-\delta$ of success). It holds, $LC^{EQ[\mathcal{D}]}(\mathcal{C}, \mathcal{H}, \delta) \leq (p+1)(p+q(\delta))$.*

Proof. Let A be a learning from $\mathcal{D}$-samples of size p algorithm that shows that $q = q(\delta)$ EQs 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 in which 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 generated 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 the event formed by the samples of size p with i positive examples,

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

The simulation carried out by B fails only if $S_0, S_1, \ldots, S_p$ are all of them samples where algorithm A fails. We can write the probability of failure of B as the product

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

By the discussion above, this product can be rewritten as

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

By Lemma 11 below, this product can be bounded by the following sum

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

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 11 *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,*

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

We show next an example that has an identification learning algorithm in the EQ from $\mathcal{D}$-teachers learning model, but does not have such algorithm in the EQ learning from $\mathcal{D}$-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. Assume that two different words x_n and y_n have been chosen for each $n \geq 1$. Consider the associated distribution D defined by:

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

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

Let C be now any class able to represent concepts consisting of pairs $\{x_n, y_n\}$ within a reasonable size; for concreteness, pick DNF formulas consisting of complete minterms. A very easy algorithm learns them in our model of EQ from $\mathcal{D}$-teachers. The algorithm has to do at most two equivalence queries to know the value of the target formula f on x_n and y_n. First, it asks whether f is identically 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 the monomial that only evaluates to one on e (the minterm). Thus we find whether either or both of $f(x_n)$ and $f(y_n)$ are 1, and if so we also know x_n and/or y_n themselves. Now the target formula is identified: the value of the formula on other points does not matter because they have zero probability.

However, it is not difficult to see that there is a distribution $D \in \mathcal{D}$ such that DNF formulas are not identifiable in the model of learning from EQ and D-samples. 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 $D_n = D(\cdot|\Sigma^n), D \in \mathcal{D}$, there is a non-negligible probability of obtaining a sample that only contains copies of x_n.

Lemma 12 *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_n-sample S of size $q(n, 1/\delta)$ does not contain y_n is greater than δ.*

Proof. The probability that y_n does not appear in S is $(1 - n^2 2^{-n})^{q(n,1/\delta)}$. By using the inequality $1 - x \geq e^{x/(x-1)}$ for $x \leq 1$, this probability is at least

$$e^{\frac{q(n,1/\delta)}{1-2^n/n^2}}.$$

Fixed q and δ this quantity is close to one for large enough n. •

Then, the following negative result follows:

Theorem 13 *There exists a distribution D in $\mathcal{D}$ such that DNF is not EQ learnable from D-samples.*

Proof. The essential idea of the proof is that, after an initial sample revealing a single word, the algorithm is left with a task close enough to that of learning DNFs in the standard model with equivalence queries, which is impossible [3].

Formally, let us consider $M_1, M_2, \ldots$ an enumeration of the equivalence queries algorithms, where M_a has running time bounded by a polynomial p_a. Note that negative results for equivalence queries remain true if learning algorithms know the value of the target concept on a point, for example 0^n. As DNF is not identifiable by this kind of algorithms [3], for each algorithm M_a there exists an integer number $n_a > \max(n_{a-1}, k_0(p_a, \delta))$ —where $k_0(p_a, \delta)$ is as in lemma 12—, $f_a \in \mathrm{DNF}_{n_a}$ and a consistent teacher T_a such that M_a does not identify f_a when teacher T_a is considered. By the previous note, without loss of generality we can assume algorithm M_a knows the value of $f_a(0^{n_a})$. Let g_a be the hypothesis returned by (M_a, T_a) and y_{n_a} a word different from 0^{n_a} such that $g_a(y_{n_a}) \neq f_a(y_{n_a})$.

Now, we define the distribution $D \in \mathcal{D}$ as follows,

$$\begin{aligned} D(0^{n_a}) &= 6/\pi^2(1/n^2 - 1/2^n) \\ D(y_{n_a}) &= 6/(\pi^2 2^n) \\ D(z_{n_a}) &= 0 \quad \text{for any word of length } n_a \text{ different from } 0^{n_a} \text{ and } y_{n_a}, \end{aligned}$$

for the integer n_a as in the paragraph above. If n is an integer that does not correspond to any n_a, distribution D is defined in a similar way by interchanging y_n by 1^n.

We show that DNF is not EQ learnable from D-samples. By lemma 12 given a polynomial q and $0 < \delta < 1$, for any integer $n \geq k_0(q, \delta)$ and with probability greater than δ, it holds that a sample S of size $q(n, 1/\delta)$ drawn according to $D_n = D(\cdot|\Sigma^n)$ only contains copies of 0^n. If M is a polynomial time equivalence queries algorithm that tries to learn DNF from D-samples, then $M = M_a$ for some a. So, by construction, when the consistent teacher T_a for the target formula f_a is considered, M will output the wrong hypothesis g_a if a sample that only contains copies of 0^{n_a} is provided as input. As that kind of samples have probability greater than δ the error probability of M is greater than δ. •

4 The Sphere Number and its Applications

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 concept, 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

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. For instance, if the empty set is not representable but N singletons are, then the number of EQs, needed to identify a particular singleton, is at least N. In this Subsection, we consider a conceptually similarly simple structure: the so-called 1-spheres. They 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 is lower bounded by the size of the largest 1-sphere that can be represented by $\mathcal{C}$. Moreover, for $M = m$ both quantities coincide.

To make the last statements 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 $|S_0 \oplus S_1| = 1$, where $\oplus$ denotes the symmetric difference of sets. In other words, $S_1 \subseteq S$ belongs to $H^1_S(S_0)$ 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 $C' : \Sigma^n \rightarrow \{0, 1, *\}$ *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 $H^1_S(S_0)$ *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 $S_1 \in H^1_S(S_0)$, 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 $H^1_S(S_0)$ is defined as $|S|$. We define the three-variable function $\mathrm{sph}_{\mathcal{R}}(n, m, M)$, called *sphere number of $\mathcal{R}$* in the sequel, as the size of the largest 1-sphere which is representable by $\mathcal{C}_{n,[m:M]}$.

We now turn to the main result of this subsection, which implies that the sphere number is another lower bound on $\mathrm{LC}^{EQ}_{\mathcal{R}}(n, m, M)$.

Theorem 14 *$sph_{\mathcal{R}}(n, m, M) \leq scdim_{\mathcal{R}}(n, m, M)$ with equality for $M = m$.*

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

Let $H^1_S(S_0)$ 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 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 (and thus with C_0) on S_Q, and coincides with C_0 on $S \setminus S_Q$ except for one instance. Clearly, Q_1 represents a set $S_1 \in H^1_S(S_0)$. By Condition (B), $|Q_1|_{\mathcal{R}} \leq m$. Since $|Q|_{\mathcal{R}} \leq |Q_1|_{\mathcal{R}}$, we arrived at a contradiction.

We prove $s \geq d$ for the special case that $M = m$. It follows from the minimality of d and Condition (2) that there exists a sample $C : \Sigma^n \rightarrow \{0, 1, *\}$ such that the following holds:

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

2. $\exists Q_0 \sqsubseteq C : |Q_0| \leq d \wedge |Q_0|_{\mathcal{R}} > m$

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

Let S denote the support of Q_0. Note that $|S| = d$ (because otherwise the last two conditions become contradictory). Let $S_0 \subseteq S$ be the set represented by Q_0. We claim that $H^1_S(S_0)$ is representable by $\mathcal{C}_{n,[m:m]}$ (which would conclude the proof). Condition (A) is obvious because $|Q_0|_{\mathcal{R}} > m$. Condition (B) can be seen as follows. For each $x \in S$, define Q_x as the subsample of C with support $S \setminus \{x\}$, and Q'_x as the sample with support S that coincides with C on $S \setminus \{x\}$, but disagrees on x. Because each Q_x is a subsample of C of breadth

$d-1$, it follows that $|Q_x|_{\mathcal{R}} \leq m$ for all $x \in S$. We conclude that the same remark applies to samples Q'_x, since a concept that is consistent with Q_x, but inconsistent with Q_0, must be consistent with Q'_x. Finally note that the samples Q'_x, $x \in S$, are exactly the representations of the sets in $H^1_S(S_0)$, respectively.

●

It is possible to capture the strong consistency dimension, even when $M > m$, with the aid of a kind of structures that combines 1-spheres. We say that sample C is *k-singular* if the following two conditions hold:

1. $|C|_{\mathcal{R}} > k$.
2. $\forall Q \sqsubseteq C : Q \neq C \Rightarrow |Q|_{\mathcal{R}} \leq k$.

Note that $H^1_S(S_0)$ is representable by $\mathcal{C}_{n,[m:m]}$ iff the sample with support S that assigns label 1 to instances from S_0 and label 0 to instances from $S \setminus S_0$ is m-singular. We define the *singular number* $\mathrm{sing}_{\mathcal{R}}(n,m,M)$ as the following maximum.

$$\max_{C \text{ is } M\text{-singular}} \left\{ \min_{Q \text{ is } m\text{-singular}} \{|Q| \mid Q \sqsubseteq C\} \right\}$$

We show now that the singular number coincides with the strong consistency dimension.

Theorem 15 *$sing_{\mathcal{R}}(n,m,M) = scdim_{\mathcal{R}}(n,m,M)$.*

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

Let us assume $d < s$ and let C be a M-singular sample where the maximum s is achieved. Then, $|C|_{\mathcal{R}} > M$ and any m-singular subsample of C has size greater than d. Therefore, any sample Q, with $Q \sqsubseteq C$ and $|Q| \leq d$, has $|Q|_{\mathcal{R}} \leq m$ –otherwise C would contain a m-singular subsample of size at most d–. This contradicts the definition of d.

Now, we assume $d > s$. Let C be a minimal sample with the following properties,

1. $|C|_{\mathcal{R}} > M$.
2. $\forall Q \sqsubseteq C : |Q| \leq s \Rightarrow |Q|_{\mathcal{R}} \leq m$.

This minimal sample C exists by the definition of d. As any subsample of C satifies the second condition, by minimality, C must be M-singular. Moreover, by the second condition, all m-singular subsamples of C have size greater than s. This contradicts the definition of s. ●

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 instances in $S \setminus S'$.

The following notation will be used in the sequel. If $S = \{x_1, \ldots, x_s\}$, then $S_i = S_0 \oplus \{x_i\}$ 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 16 *Let $\mathcal{C} = H_S^1(S_0)$ be a 1-sphere and D an arbitrary but fixed distribution on S. Then, $LC^{EQ[D]}(\mathcal{C}, \delta) \leq 1 + \lceil \log(1/\delta) \rceil$.*

Proof. 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. •

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 16 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 17 *The following lower bound even holds for randomized learners:*

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

Proof. The first inequality is trivial. We prove the second one. Let $s = \mathrm{sph}(\mathcal{C})$ and $H_S^1(S_0)$ be the 1-sphere of size s that is representable by $\mathcal{C}$. 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. 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 $H_S^1(S_0)$ 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 $U = (S' \oplus S_j) \setminus \{x_j\}$ 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.[4]

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 18 *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:*

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

[4] 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.

Considering learning algorithms that do not make queries longer than the size of the target concept, Corollary 18 and Theorem 2 imply the following somewhat surprising result: A representation class is (determnistically) polynomially EQ-learnable (with answers given by an adversary) iff it is (probabilistically) polynomially learnable from $\mathcal{D}_{unif}$-teachers. Thus passing from deterministic to probabilistic learners and from the adversary-oracle to $\mathcal{D}_{unif}$-teachers does not significantly increase the learning power. This negative result applies as well to the model of EQ-learning from $\mathcal{D}_{unif}$-samples, which has been proved earlier to be subsumed by randomized learners from $\mathcal{D}_{unif}$-teachers.

It is an open problem whether the learning power significantly increases when $\mathcal{D}_{unif}$-teachers are combined with learners that do make queries longer than the size of the target concept.[5]

We finally would like to mention that the lower bound for randomized learners from arbitrary counterexamples in Corollary 18 is as good as the result from [8] (Theorem 3.3) which relates the learning complexity with deterministic and randomized algorithms.

References

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

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

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

[4] 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.

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

[6] Y. Hayashi, S. Matsumoto, A. Shinoara, and M. Takeda. Uniform characterization of polynomial-query learnabilities. In *First International Conference, DS'98*, pages 84–92, Fukuoka, Japan, 1998. LNAI. Springer.

[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] W. Maass. On-line learning with an oblivious environment and the power of randomization. In *Proc. 4th Annu. Workshop on Comput. Learning Theory*, pages 167–175, San Mateo, CA, 1991. Morgan Kaufmann.

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

[5] Although it is straightforward to show that $\mathrm{LC}^{EQ}_{\mathcal{R}}(n,m,M) \geq \mathrm{LC}^{EQ[\mathcal{D}_{unif}]}_{\mathcal{R}}(n,m,M,\delta) \geq (1-\delta)\mathrm{sph}_{\mathcal{R}}(n,m,M)$, quantity $\mathrm{sph}_{\mathcal{R}}(n,m,M)$ may in general be considerably smaller than $\mathrm{scdim}_{\mathcal{R}}(n,m,M)$. Thus, the lower bound becomes inferior to the strong consistency dimension.