Learning by Distances *

Shai Ben-David † Alon Itai ‡ Eyal Kushilevitz §

Dept. of Computer Science
Technion, Haifa 32000, Israel

*An early version of this paper appeared in the proceedings of *the 3rd Workshop on Computational Learning Theory*, August 1990, pp. 232-245.

†e-mail: shai@techsel.bitnet, shai@cs.technion.ac.il

‡e-mail: itai@techsel.bitnet. Supported by the Fund for the Promotion of Research at the Technion.

§e-mail: eyalk@techunix.bitnet

A running head: **Learning By Distances**
Mailing address for proofs:

Shai Ben-David
Computer Science Department
Technion, Haifa 32000, Israel
e-mail: shai@cs.technion.ac.il, shai@techsel.bitnet

Abstract

A model of *learning by distances* is presented. In this model a concept is a point in a metric space. At each step of the learning process the student guesses a hypothesis and receives from the teacher an approximation of its distance to the target.

A notion of a distance, measuring the proximity of a hypothesis to the correct answer, is common to many models of learnability. By focusing on this fundamental aspect we discover some general and simple tools for the analysis of learnability tasks.

As a corollary we present new learning algorithms for Valiant's PAC scenario with any given distribution. These algorithms can learn any PAC–learnable class and, in some cases, settle for significantly less information than the usual labeled examples.

Insight gained by the new model is applied to show that every class of subsets $\mathcal{C}$ that has a finite VC–dimension is PAC–learnable with respect to any fixed distribution. Previously known results of this nature were subject to complicated measurability constraints.

1 Introduction

The notion of a metric, quantifying the proximity between objects, plays a role whenever approximations are considered. Most models of computational learnability involve a process in which some "target" is approximated by "hypotheses". Furthermore, in many cases the information supplied by the "teacher" gives some indication of the degree of the "student's" success and can, therefore, be interpreted as approximating the distance between a current hypothesis and the target. We focus on this basic aspect by presenting a model of learnability in a metric space. Our model disregards all the attributes of concepts except the distances between them. We establish a close connection between the metric structure of a concept class and its learnability features.

Consider the following example: A police investigator is trying to come up with a drawing of a criminal. After getting some basic information from a witness (who has seen the criminal), the investigator presents the witness with sketches. The witness indicates to what degree the pictures resemble the criminal. This process can be regarded as a learning process where the investigator (the student) offers guesses, and the witness (the teacher) responds with approximations of their distances to the target. Two basic parameters of such a process are the *precision* of the teacher (how elaborate and faithful the witness is) and the *accuracy* required from the student (how exact should the drawing get to guarantee the police does not suspect an innocent citizen). We call such a process *learning by distances* (LBD).

We characterize learnability by distances of concept classes in terms of Kolmogorov's ε-entropy of a metric space [18]. We present several learning algorithms, and consider their complexity both in terms of the number of queries they require, and in terms of their computational complexity.

Apart from its significance in modeling some important aspects of "real" learning processes, the notion of LBD-learnability sheds new light on the more familiar models of learnability. It identifies some basic motives that re-appear in various learning problems. We relate previously discussed models of learnability (PAC-learnability model [21], and variants of Angluin's equivalence queries model [2]) to the LBD-model.

We show that the learnability of a concept class in the PAC sense is tightly related to the learnability-by-distances of that class in the metric spaces induced by probability distributions. Given a concept class $\mathcal{C}$ over a sample space $\mathcal{Y}$, every probability distribution P induces a natural (pseudo) metric on $\mathcal{C}$ - the distance between two concepts is the probability of their symmetric difference. Any LBD-student for the induced metric problem provides a heuristic for PAC-learning with respect to the distribution P: The examples presented by a PAC-teacher may be used to approximate the distance between the hypothesis and the target. The LBD-student may use these estimates to produce a hypothesis which is close to the target in the pseudo-metric and therefore is an approximation in the PAC sense.

Our model emphasizes the distinction between the "probably" and the "approximately" in the "Probably Approximately Correct" (PAC) scenario; The probability of success is determined by the precision of our estimates of the distances and is, therefore, controlled by the size of sample we request at each stage. Independently, the quality of approximation obtained by our final hypothesis is determined by the accuracy parameter of the LBD student.

As a matter of fact, the algorithm of [19] for learning AC^0 functions under the (fixed) uniform distribution is an LBD based algorithm. They approximate the target function by estimating its Fourier coefficients (relative to some basis for the space of real functions on $\{0,1\}^n$). Each such coefficient can be viewed as a distance between the target function and the appropriate basis function. Their main lemma shows that AC^0 functions can be approximated on the basis of "few" (quasi-polynomial) such distances. To approximate these distances they consider random examples as described above.

An insightful consequence of our analysis concerns the amount of information required by a student for successful learning. A PAC student that follows the steps of an LBD student, as described above, needs very little information. Rather than requiring a teacher's example $< y, b >$ (where $y \in \mathcal{Y}$ and b tells if $y \in t$) he can settle for just one bit b^c indicating whether the y drawn by the teacher is in $c \triangle t$ – the symmetric difference between the student's current hypothesis c and the target t. We show that for any distribution P, any concept class that can be learnt on the basis of P-random labeled examples (as in the PAC model) can be learnt on the basis of 0–1 information as above. In many cases "real life" learning is indeed based on this type of limited information. As an example, consider a biologist who is trying to identify the best vaccine for some disease of mice. He keeps coming up with candidate variations and tries them on random samples of sick mice. It is most likely that he cares about the rate of recovery of his mice rather than the exact identity of each recovered mouse.

Blumer et. al. [10] characterize the concept classes that are learnable by every *consistent student* (a student that always picks hypothesis consistent with the examples it has seen). Let us call such classes *consistently learnable*. They show that if some complicated measure theoretic constraints hold, then a class is consistently learnable iff it has a finite VC-dimension. Under these constraints learnability (by any student) implies consistent learnability. Although we do not know of any "natural" concept class for which these constraints are not met, [7, 10] show examples of classes of Borel subsets of the interval $[0,1]$ that have VC-dimension 1 but still are not consistently learnable (even in the uniform distribution on $[0,1]$). The algorithms we derive for PAC-learnability imply that the finite VC-dimension criteria does characterize PAC-learnability relative to any fixed distribution without any measurability constraints. In particular, these algorithms learn the classes mentioned above and thus strictly exceed the learning capability of some consistent students.

The paper is roughly divided into two parts. In the next three sections we define and investigate the basic properties of the LBD-model and the learning heuristics it offers. In the second part of the paper (section 5) we discuss the connections between our model and Valiant's PAC-model and we apply the insight gained by the LBD-model to get new results in the theory of PAC-learnability.

2 Preliminaries

In this section we give the formal definition of the *learning by distances (LBD)* model.

Let $(\mathcal{X}, d)$ be a pseudo-metric space. I.e., d is a function from $\mathcal{X}^2$ to non-negative reals satisfying (for every $x, y, z \in \mathcal{X}$):

- $d(x, x) = 0$.
- $d(x, y) = d(y, x)$.
- $d(x, z) \leq d(x, y) + d(y, z)$.

A *concept class* $\mathcal{C}$ is a subset of $\mathcal{X}$. Note that we depart from the common terminology; here a concept is just a point in the space $\mathcal{X}$ while in the common scenario concepts are subsets of an underlying domain.

Definition 1: *Let ϵ, γ be positive reals and let t be a member of a pseudo-metric space $(\mathcal{X}, d)$.*

- *A function $f : \mathcal{X} \to \boldsymbol{R}^+ \cup \{SUCCESS\}$ is an (ϵ, γ)-*approximation of the distance *to t if, for every x, $f(x) = SUCCESS$ iff $d(x, t) \leq \epsilon$, and whenever $d(x, t) > \epsilon$, $f(x) \stackrel{\gamma}{=} d(x, t)$[1].*
- *An (ε, γ)-*learning sequence *for t is a sequence*

$$(c_1, r(c_1)), (c_2, r(c_2)), \ldots, (c_k, r(c_k))$$

where, for all $i \leq k$, $c_i \in \mathcal{X}$ and $r(c_i)$ is an (ϵ, γ)-approximation of the distance (of c_i) to t.

In our model, learning sequences are generated by an interaction between a student and a teacher, the c_i's are the student's queries (or hypotheses) and the $r(c_i)$'s are the teacher's responses.

Formally, a *student* is a function S from learning sequences to $\mathcal{C}$.

S *produces* $(c_1, r(c_1)), (c_2, r(c_2)), \ldots, (c_k, r(c_k))$ if for all $i < k$

$$S(\mathcal{C}, \varepsilon, \gamma, (c_1, r(c_1)), \ldots, (c_i, r(c_i))) = c_{i+1}.$$

Namely, S is the strategy of the student for choosing his hypotheses.

[1] $a \stackrel{\gamma}{=} b$ denotes $|a - b| \leq \gamma$.

A finite learning sequence is *successful* if $r(c_k) = SUCCESS$, where c_k is the last hypothesis in the sequence.
As can be seen from the definitions above, γ is a precision parameter for the teacher's answers and ε defines the accuracy required from the student.

A concept class $\mathcal{C}$ is (ε, γ)-*learnable* if there exists a student S and a function $\ell : \mathcal{C} \to \mathbb{N}$ such that for every target $t \in \mathcal{C}$, every (ε, γ)-learning sequence for t,

$$(c_1, r(c_1)), (c_2, r(c_2)), \ldots, (c_{\ell(t)}, r(c_{\ell(t)}))$$

produced by S, is successful. Note that $\ell(t, \varepsilon, \gamma)$ determines the number of queries the student is allowed to ask before reaching a conclusion.

A concept class $\mathcal{C}$ is ε-*learnable* if it is (ε, γ)-learnable for all $\gamma \leq \frac{\varepsilon}{2}$.

A concept class $\mathcal{C}$ is *learnable* if it is ε-learnable for all $\varepsilon > 0$.

Each of these learnability notions is called *uniform* if the number of queries ℓ does not depend on t (it may depend on ε, γ and $\mathcal{C}$).

Note that this notion of learnability is purely qualitative and is not concerned with the resources required by the learning process. We postpone the quantitative consideration (sample and computational complexity) to Section 4.

3 A Characterization of Learnable Concept Classes

In this section we give a complete characterization of LBD learnable concept classes in terms of their metric entropy. We also present a game theoretic characterization of learnability by ℓ-many queries.

Consider the discrete metric space $(\mathcal{X}_u, d_u)$ in which for every $x \neq y \in \mathcal{X}_u$, the distance $d_u(x, y)$ is 1. It is quite evident that for $\varepsilon < 1$, a concept class in such a metric space cannot be learnt unless the student exhaustively searches for the target. On the other hand, let $\mathcal{C} = \mathcal{X} = \{[a, a + 0.1] \subseteq [0, 1]\}$ and the distance between two intervals I and J is defined by $d(I, J) \stackrel{def}{=} \mu(I \triangle J)$, where $\triangle$ denotes the symmetric difference and μ is the usual Lesbegue measure. To learn this concept class the student can guess one by one the elements in $A = \left\{[a, a + 0.1] \in \mathcal{C} | a = \frac{k}{4} \cdot \varepsilon,\ k \in \mathbb{N}\right\}$. Note that, although A is finite ($|A| \leq \frac{4}{\varepsilon}$), and $\mathcal{C}$ is uncountable, for any target $t \in \mathcal{C}$, there exists an element of A which is ε-close to t.
The following notion of capacity is therefore essential for analyzing learnability.

Definition 2: *A set $A \subseteq \mathcal{X}$ is an ε–*approximation *of a set $B \subseteq \mathcal{C}$ if for every $b \in B$ there exists $a \in A$ such that $d(a, b) \leq \varepsilon$. The ε–*capacity *of a set $B \subseteq \mathcal{C}$ is defined as the size of a minimal ε–approximation for B. I.e.,*

$$CAP(B, \varepsilon) = \min \{|A| \ : A \textit{ is an } \varepsilon\textit{-approximation of } B \ \}.$$

(Note that $CAP(B, \varepsilon)$ may be an infinite cardinal).

Kolmogorov and Tihomirov [18] discuss this notion and define the ε–*entropy* of a set B in a metric space as $\log_2 CAP(B, \varepsilon)$. If one regards B as, say, a family of typographical symbols and d is a measure of their distinctness, then, the ε–entropy of B measures the amount of information that can be unambiguously represented by B when there is an ε amount of "noise" in the system. In the context of learning, the analogous notion is usually referred to as an ε-net [22]

Example 1: *Let B be the unit cube in $\boldsymbol{R}^n$, then $CAP(B,\varepsilon) \leq (\frac{\sqrt{n}}{\varepsilon})^n$. On the other hand in the discrete metric space $(\mathcal{X}_u, d_u)$ for every set B and every $\varepsilon < 1$, $CAP(B,\varepsilon) = |B|$, where $|B|$ denotes the cardinality of B.*

The following lemma is a simple property of capacities:

Lemma 1: *Let $B = \cup_{i=1}^{k} B_i$ then $CAP(B,\varepsilon) \leq \sum_{i=1}^{k} CAP(B_i,\varepsilon)$.*

Proof: The claim follows from the fact that if for every i the set A_i is an ε–approximation of B_i, then $\cup_{i=1}^{k} A_i$ is an ε–approximation of $\cup_{i=1}^{k} B_i$. □

Definition 3: *A (pseudo-)metric space, $\mathcal{X}$, is called* bounded *if there exists b such that $d(x,y) \leq b$ for every $x, y \in \mathcal{X}$.*

Note that if for some ε, $CAP(\mathcal{X},\varepsilon) < \infty$, then $\mathcal{X}$ is bounded.

The following lemma provides an information theoretic lower-bound on the number of rounds needed for successful learning, in terms of the capacity of the concept class and the number of different possible teacher responses. We formulate the lemma in terms of our LBD model but it is applicable to any deterministic learning scenario.

Lemma 2: *For every $\varepsilon > 0$ and every $q > 0$, if, for some γ, a concept class $\mathcal{C}$ in a metric space $(\mathcal{X}, d)$ is uniformly (ε,γ)–learnable in at most ℓ rounds and a teacher that gives at most q many possible responses in each round, then*

$$CAP(\mathcal{C},\varepsilon) \leq \frac{q^\ell - 1}{q - 1}.$$

Proof: Let S be a student that (ε,γ)–learns $\mathcal{C}$ in ℓ steps. Let T_ℓ be the tree of all possible query sequences of length at most ℓ, producible by the student S on the basis of the teacher responses. By the assumption, the number of nodes in T_ℓ is at most

$$n_\ell \stackrel{def}{=} \frac{q^\ell - 1}{q - 1}.$$

The fact that S is a successful student implies that for any target $t \in \mathcal{C}$ there exists a node in this tree which is ε–close to t. Thus the nodes of the tree form an ε–approximation for $\mathcal{C}$ of size n_ℓ. □

The following theorem relates the learnability of a concept class to its capacity.

Theorem 1: *For every $\varepsilon > 0$ and every $\gamma > 0$, a concept class $\mathcal{C}$ in a bounded metric space $(\mathcal{X}, d)$ is uniformly (ε, γ)–learnable if and only if*

$$CAP(\mathcal{C}, \varepsilon) < \infty.$$

(Note that the condition is independent of γ).

Proof: Assume that $CAP(\mathcal{C}, \varepsilon) = n$. We prove that $\mathcal{C}$ is uniformly (ε, γ)–learnable by presenting the $STUDENT$ algorithm which learns this concept class within n queries. This algorithm works in an oblivious way: It chooses a set $A = \{a_1, a_2, \ldots, a_n\}$ which is an ε–approximation of $\mathcal{C}$ (such an A exists by the definition of CAP) and exhaustively searches through A. I.e, it presents the elements of A as queries to the teacher until it receives the answer $SUCCESS$. (Note that for this direction the restriction that $\gamma > 0$ is irrelevant.)

For the reverse implication, the first observation to note is that a successful student should be able to learn with *any* legal teacher. There exists a legal teacher which always answers with r_i's in $\{SUCCESS, \varepsilon + \gamma, \varepsilon + 2\gamma, \varepsilon + 3\gamma, \ldots, b\}$. Let q be the size of this set ($q \cong \frac{b-\varepsilon}{\gamma}$). Applying Lemma 2 we get that

$$CAP(\mathcal{C}, \varepsilon) \leq \frac{q^\ell - 1}{q - 1} < \infty.$$

□

Corollary 1: *If a concept class $\mathcal{C}$ in a bounded metric space $(\mathcal{X}, d)$ is uniformly (ε, γ)–learnable, for some $\gamma > 0$, then $CAP(\mathcal{C}, \varepsilon)$ is finite and $\mathcal{C}$ is learnable using at most $CAP(\mathcal{C}, \varepsilon)$ many queries.*

Proof: Follows from the proof of theorem 1. □

It is worthwhile to note that $STUDENT$ is very succinct in terms of the information it requires from the teacher. It operates upon the binary responses $SUCCESS/OTHER$. $STUDENT$ is also efficient in the following sense: Its queries do not depend on the responses it gets, and therefore they can all be calculated a-priori, on the basis of knowing just the space $(\mathcal{X}, d)$, the concept class $\mathcal{C}$, and the parameters ε and γ.

The above theorem shows that as far as the question of learnability is concerned, any class that is learnable – is learnable by a non-adaptive algorithm receiving only $FAIL/SUCCESS$ teacher responses (rather than approximated distances). In particular, as far as the question of learnability is discussed, the parameter γ is irrelevant. On the other hand, with respect to quantitative complexity measures, such as the number of needed queries or the computational complexity, other algorithms that exploit the metric structure of the space $\mathcal{X}$ can do better (see section 4 and [8]).

Definition 4: *A set $B \subseteq \mathcal{X}$ is* totally bounded *if $CAP(B, \varepsilon)$ is finite for every $\varepsilon > 0$.*

Corollary 2: *A concept class $\mathcal{C}$ in a bounded metric space $(\mathcal{X}, d)$ is uniformly learnable if and only if it is totally bounded.*

The following theorem characterizes non-uniformly learnable concept classes. Recall that a concept class $\mathcal{C}$ is non-uniformly learnable if for some function $\ell : \mathcal{C} \to \mathbb{N}$ there exists a student that for any $t \in \mathcal{C}$ will produce a successful learning sequence of length at most $\ell(t)$.

Theorem 2: *For every ε and every $\gamma > 0$, a concept class $\mathcal{C}$ in a metric space $(\mathcal{X}, d)$ is (non-uniformly) (ε, γ)–learnable if and only if $CAP(\mathcal{C}, \varepsilon)$ is countable (possibly finite).*

Proof: If $A \subseteq \mathcal{C}$ is a countable ε-approximation then algorithm $STUDENT$ has just to enumerate the elements of $A = \{a_i\}_{i \in \mathbb{N}}$ and present them to the teacher in that order, until the teacher responses with $SUCCESS$. On the other hand, assume $\mathcal{C}$ is (ε, γ)–learnable for some function $\ell : \mathcal{C} \to \mathbb{N}$. We can now proceed as in the proof of theorem 1. Without loss of generality the teacher's responses are in the countable set $\{k \cdot \gamma | k \in \mathbb{N}\} \cup \{SUCCESS\}$. Let T be the tree of all finite sequences generated by S on the basis of such responses. T is a countable ε–approximation of $\mathcal{C}$. $\square$

Definition 5: *A set B in a metric space $(\mathcal{X}, d)$ is* separable *if there exists a countable $A \subseteq \mathcal{X}$ such that for every $b \in B$ and every $\varepsilon > 0$ there exists $a \in A$ such that $d(a, b) < \varepsilon$.*

Corollary 3: *A concept class $\mathcal{C}$ in a metric space $(\mathcal{X}, d)$ is learnable if and only if it is separable.*

3.1 A Game Theoretic Characterization

Let us conclude this section with a game theoretic characterization of learnability in ℓ many queries. We define a game $G(\mathcal{C}, \varepsilon, \gamma)$ for two players. The game proceeds as follows: In each step Player I picks some $c_i \in \mathcal{X}$. Player II picks r_i ($r_i > \varepsilon$). Let $B_i = \left\{ b \in \mathcal{C} | \forall j \leq i,\ d(c_j, b) \stackrel{\gamma}{=} r_j \text{ and } d(c_j, b) > \varepsilon \right\}$. We denote by G_ℓ the game that runs for ℓ many steps. Player II wins the game G_ℓ if $CAP(B_\ell, \varepsilon) \geq 2$. Otherwise, Player I wins.

Claim 1: *Player I has a winning strategy in the game $G_\ell(\mathcal{C}, \varepsilon, \gamma)$ if and only if the concept class $\mathcal{C}$ is (ε, γ) –learnable within $\ell + 1$ queries.*

Proof: If there exists a winning strategy for Player I then the student regards the teacher as a simulator for Player II, and picks its queries according to Player's I strategy. The only possible response which is not in the vocabulary of Player II is $SUCCESS$, but once the student receives such a response the learning process is over. Let $(c_1, r(c_1)), \ldots, (c_\ell, r(c_\ell))$ be the run produced in the game (note that for all i the target belongs to B_i). We know that $CAP(B_\ell, \varepsilon) = 1$, thus there exists a point $c_{\ell+1}$ which is at distance at most ε from any point in B_ℓ, and in particular from t.
For the other direction, suppose that Player I does not have a winning strategy, it follows (see e.g. [13]) that Player II has a winning strategy. Let the teacher regard the student as a simulator for Player I, and choose the r_i's according to the winning strategy of Player II. Note that, in any step, the set B_i is the set of all "possible targets" consistent with the answers given during the first i steps. After ℓ steps the set B_ℓ satisfies $CAP(B_\ell, \varepsilon) \geq 2$. This implies that, for any $c_{\ell+1}$ the student may choose, there exists a point in B_ℓ which is ε-far from $c_{\ell+1}$

and may serve as the target. In other words, a winning strategy for Player II shows how the teacher can, by an appropriate choice of responses $r(c_i)$, guarantee that for any learning sequence $(c_1, r(c_1)), \ldots, (c_{\ell+1}, r(c_{\ell+1}))$ produced by a student (on the basis of the teacher responses), there exists some target t such that this sequence is a learning sequence for t but $c_{\ell+1}$ is not ε-close to t. It follows that the concept class $\mathcal{C}$ is not (ε, γ)-learnable in $\ell+1$ steps. $\square$

4 The Complexity of The Student

Let us now turn to the quantitative aspects of the learning process. We have shown (in the proof of theorem 1) that any learnable class (in the sense of the LBD-model) is learnable by the naive algorithm $STUDENT$. We shall now see examples showing that in many occasions other learning algorithms have superior performance.

The first question we address is the number of queries an algorithm uses. Lemma 2 and Corollary 1 of the previous section immediately imply the following bounds:

Claim 2: *Let $\mathcal{C}$ be a bounded concept class and let D denote its diameter (i.e. $D = \sup\{d(s,t) : s, t \in \mathcal{C}\}$). Let ℓ be the number of queries needed for (ϵ, γ)-learning the class $\mathcal{C}$, finally, let q denote the minimal number of possible teacher replies ($q \cong \frac{D}{\gamma}$).*

$$\log_q CAP(\mathcal{C}, \varepsilon) \leq \ell \leq CAP(\mathcal{C}, \varepsilon)$$

We shall show that these bounds are best possible in the general case. The lower bound of the above claim indicates that the query-complexity is affected by the type of replies supplied by the teacher. The next pair of examples demonstrates this phenomenon; For some concept classes, for any algorithm based upon binary responses, the number of needed queries necessarily grows to infinity as ε goes to 0, whereas in other cases, a smart learner can exploit approximated distances responses to guarantee success within a constant number of queries.

Example 2: *Let $\mathcal{C}$ be a concept class in a Euclidean space (i.e. $\mathcal{C} \subseteq \boldsymbol{R}^n$). If $\mathcal{C}$ is unbounded then, for every ϵ, $CAP(\mathcal{C}, \epsilon) = \infty$. Just the same such a class is learnable through a finite number of queries; For every target concept, $t \in \mathcal{C}$, and any set $\{x_1, \ldots, x_{n+1}\}$ of $n+1$- many points in $\boldsymbol{R}^n$, the set of (exact) distances $\{d(x_1, t), \ldots, d(x_{n+1}, t)\}$ determines t uniquely. Even when the distances are only approximated, it is not hard to see that 2^n such teacher's responses suffice (e.g. the distances of the target from the corners of an n-dimensional cube containing it).*

The point we wish to note is that such a learning strategy depends upon the infinity of the range of possible teacher replies. In other words, the above lower bound is meaningless as the parameter q there becomes ∞.

A similar phenomena arises in the bounded case, $\mathcal{C} \subseteq [0,1]^n$. In this case $CAP(\mathcal{C}, \epsilon)$ is finite but grows to infinity as ϵ goes to zero. The number of needed queries can be kept 2^n by using the same algorithm while demanding the teacher's accuracy to grow with ϵ. The point to note is that, as the accuracy of the approximations grows, so does the cardinality q of the range of possible teacher replies.

The next example shows that, once q is kept constant, the same classes demand unboundedly many queries, if one wishes to learn them for arbitrarily small ϵ's.

Example 3*:* *Let $\mathcal{C}$ be the unit ball in $\boldsymbol{R}^n$. By Lemma 2, any student receiving only $FAIL/SUCCESS$ responses will require $O(\varepsilon^{-n})$ queries to learn $\mathcal{C}$.*

This advantage does not always exist, in the next example no algorithm can beat the algorithm $STUDENT$ with respect to the number of needed queries.

Example 4*:* *Consider the discrete space, where all distances are 1, then even exact responses $d(c_i, t)$ ($\gamma = 0$) do not shorten the learning process – they are just as informative as $FAIL/SUCCESS$.*

Next, we address the question of the computational complexity of the algorithms. We call a student *consistent* if it has the property that its i-th hypothesis behaves like the target with respect to all previous queries. Formally, for all $j < i \leq k$ its learning sequence

$$(c_1, r(c_1)), (c_2, r(c_2)), \ldots, (c_k, r(c_k))$$

satisfies

$$d(c_i, c_j) - \gamma \leq r_j \leq d(c_i, c_j) + \gamma.$$

In each step a consistent student decreases the set of consistent hypotheses. The target belongs to all these sets, and the i-th hypothesis belongs to the consistent set of step i. Although such an algorithm is very natural, the following example shows that it may be intractable:

Example 5*:* *In order to prove lower bounds on the computational complexity we consider an infinite family of finite metric spaces $\{(\mathcal{X}_G, d)\}_G$, where $\mathcal{X}_G$ is the set of all simple paths (and cycles) of the graph $G = (V, E)$.*
The distance between two paths $c_1, c_2 \subset E$ is

$$d(c_1, c_2) = |c_1 \Delta c_2|.$$

It is easy to verify that this is indeed a metric.
Finding a hypothesis at the same distance from c_1 as the target may be NP-hard. Let $\varepsilon = \gamma = 1/3$ and G a Hamiltonian graph with n vertices and m edges. If the target is a Hamiltonian cycle of G and c_1 is the empty cycle, then $n - 1/3 \leq r(c_1) \leq n + 1/3$. Since, all subsequent hypotheses of a consistent student must be at distance $\approx n$ from c_1, they must also be Hamiltonian cycles. However finding a Hamiltonian cycle in a Hamiltonian graph is NP-hard in the sense that the existence of a polynomial algorithm implies P=NP. (The fact that a Hamiltonian cycle is known to exist doesn't make the task of finding one easy: Had there been an n^k time algorithm we could have run it on any graph, and if after n^k steps it did not output a cycle we would have known that the graph was not Hamiltonian.)
It should be noted, however, that the learning problem is not difficult: There exists a simple algorithm to learn any cycle in $|E| + 1$ queries, each requiring constant computation: If $E = \{e_1, \ldots, e_m\}$ then let $c_i = \{e_i\}$. If $e_i \notin t$ then $r(c_i) \stackrel{\frac{1}{3}}{=} |t| + 1$, otherwise, $e_i \in t$ and $r(c_i) \stackrel{\frac{1}{3}}{=} |t| - 1$. Thus, it is easy to find $c_{m+1} = t$.

5 The Relation to PAC Learnability

As discussed in the introduction, a notion of a metric space, reflecting the proximity relation among concepts, is common to many models of computational learnability. In this section we apply our metric approach by demonstrating how various learnability problems can be formulated in the LBD model. We concentrate on Valiant's PAC learnability and show how the metric view point yields some improvements to basic results concerning that model, we conclude with a sketched representation of some 'Membership Queries' models in the new framework.

Let $\mathcal{Y}$ be some universe set. Let $\mathcal{X}$ be a σ-algebra of subsets of $\mathcal{Y}$ (i.e., $\mathcal{X} \subseteq 2^{\mathcal{Y}}$), and let $\mathcal{C} \subseteq \mathcal{X}$. Every probability measure P over $\mathcal{Y}$, under which all members of $\mathcal{X}$ are measurable, induces a natural pseudo-metric on $\mathcal{X}$: For every $a, b \in \mathcal{X}$ let $d_P(a,b)$ be $P(a \triangle b)$. In our analysis we assume that the metric is known to the student. This corresponds to learnability with respect to fixed distributions in the PAC–model. [2]

The (distribution-free) PAC-learnability of a class $\mathcal{C}$ depends upon its Vapnik-Chervonenkis dimension (VC–dimension) [23, 10]. The following result of Dudley [11, Theorem 9.3.1] relates the VC–dimension of a class $\mathcal{C}$ with the capacity CAP (as defined in Section 3) in such induced pseudo-metrics d_P.

Theorem 3*:* ***(Dudley)*** *For a family of sets $\mathcal{C}$ let*

$$s(\mathcal{C}) = \inf \left\{ w \mid \exists K, \forall \text{ probability distribution } P, \forall \varepsilon > 0,\ CAP_{d_P}(\mathcal{C}, \varepsilon) \leq K \cdot \left(\frac{1}{\varepsilon}\right)^{w} \right\}.$$

For every such $\mathcal{C}$

1. *$s(\mathcal{C}) \leq VCDim(\mathcal{C})$.*
2. *$s(\mathcal{C}) = \infty$ if and only if $VCDim(\mathcal{C}) = \infty$.*

The theorem relates the rate of growth of $CAP_{d_P}(\mathcal{C}, \varepsilon)$, to the VC–dimension of $\mathcal{C}$. If $\mathcal{C}$ has a finite VC–dimension, v, then $CAP_{d_P}(\mathcal{C}, \varepsilon)$ is bounded by $K \cdot \left(\frac{1}{\varepsilon}\right)^{v}$, for all P. The VC–dimension is ∞ if and only if no such polynomial bound exists.
As an immediate corollary of the above theorem we get:

Theorem 4*:* *If $\mathcal{C}$ is a class of sets that has a finite VC–dimension then for every distribution P, the concept class $\mathcal{C}$ is uniformly (ε, γ)-learnable-by-distances in the metric d_P, for any $\varepsilon > 0$ and $\gamma \geq 0$. Furthermore, the needed number of queries is $O\left(\left(\frac{1}{\varepsilon}\right)^{VCDim(\mathcal{C})}\right)$.*

Proof: Theorem 3 guarantees that for such $\mathcal{C}$, $CAP_{d_P}(\mathcal{C}, \varepsilon)$ is finite for all ε (more precisely, $O\left(\left(\frac{1}{\varepsilon}\right)^{v}\right)$), and now Theorem 1 ensures the learnability by distances of $\mathcal{C}$ using $O\left(\left(\frac{1}{\varepsilon}\right)^{v}\right)$ many queries. $\square$

[2] An appealing aspect of the PAC–model is its "distribution freeness" – the ability to learn even when the underlying distribution is not known. The LBD–model can be adapted to such a scenario by adding an initial segment of queries to each learning session. The new segment is meant to provide the student with enough (approximated) information about the metric. We can prove the existence of such metric-free learning processes for, e.g., concept classes that are well-behaved (see [7]).

Corollary 4: *If $\mathcal{C}$ is a class of sets that is (distribution-free) PAC-learnable then for every distribution P, the concept class $\mathcal{C}$ is uniformly learnable-by-distances in the metric d_P.*

Proof: By [10], if $\mathcal{C}$ is (distribution-free) PAC-learnable then it has a finite VC–dimension. □

Benedek and Itai [9] have studied PAC-learnability for fixed distributions. They show the following characterization of such learnability:

Theorem 5: ***(Benedek-Itai)*** *A concept class $\mathcal{C} \subseteq 2^{\mathcal{Y}}$ is PAC-learnable relative to a distribution P on $\mathcal{Y}$ if and only if for every $\varepsilon > 0$ there exists a finite set $A \subseteq 2^{\mathcal{Y}}$ that ε-approximates $\mathcal{C}$.*

Theorem 6: *$\mathcal{C} \subseteq 2^{\mathcal{Y}}$ is PAC-learnable relative to a distribution P if and only if it is uniformly learnable by distances in the metric d_P.*

Proof: Apply Theorems 1 and 5. □

We now turn to a closer examination of these two types of learning processes. The PAC and the LBD models have some evident common features. In both learning procedures there is a concept class $\mathcal{C}$ known to the student, a target $t \in \mathcal{C}$ known to the teacher, and the student is trying to come up with a hypothesis h that is a sufficiently good approximation to t. The difference between the models lie in the nature of the communication between the teacher and the student. In the PAC model, the student receives from the teacher pairs $< y, b >$, where y is drawn at random according to the probability distribution P on $\mathcal{Y}$, and b is a bit indicating whether $y \in t$. In the LBD model, the student presents a hypothesis h and receives from the teacher an approximation of the distance between h and the target t.

It turns out that these differences can be easily bridged. In Theorems 4 and 5 we have seen that the PAC-learnability of $\mathcal{C}$ in $(\mathcal{Y}, P)$ implies its LBD-learnability in $(\mathcal{X}, d_P)$. The reverse direction is true as well (for every fixed distribution). We prove it by showing how any LBD-student for $\mathcal{C}$ in $(\mathcal{X}, d_P)$ can be transformed into a PAC-student for $\mathcal{C}$ in $(\mathcal{Y}, P)$.

Theorem 7: *Let $\mathcal{C} \subseteq \mathcal{X} \subseteq 2^{\mathcal{Y}}$, P, and d_P be as above. For every $\varepsilon, \delta > 0$, if there exists an LBD-student, S_{LBD}, that $(\frac{\varepsilon}{2}, \frac{\varepsilon}{4})$ learns $\mathcal{C}$ (in $(\mathcal{X}, d_P)$) uniformly in $\ell \stackrel{def}{=} \ell(\frac{\varepsilon}{2}, \frac{\varepsilon}{4})$ steps, then there exists a PAC-student S_{PAC} that (ε, δ) learns $\mathcal{C}$ (in $(\mathcal{Y}, P)$) after seeing $\ell \cdot \frac{\log 2\ell + \log \frac{1}{\delta}}{\frac{\varepsilon^2}{8} \cdot \log e}$ examples. (See Theorem 8 for a better bound on the number of examples.)*

For the proof we need the following inequality (due to Hoeffding [16]):

Lemma 3: ***(Hoeffding)*** *Let X_i $(1 \leq i \leq n)$, be n random variables each of which equals 1 with probability p and equals 0 otherwise. Then*

$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i \;-\; p\right| > \beta\right] \leq 2 \cdot e^{-2n\beta^2}.$$

(Intuitively, this means that if you have enough samples then you are close to the expected value).

Proof: **(of the theorem)** Let $S_{LBD}(\frac{\varepsilon}{2}, \frac{\varepsilon}{4})$ be a learning-by-distance student for the concept class $\mathcal{C}$. We construct a student S_{PAC} that PAC-learns $\mathcal{C}$ (with respect to P), and with probability $1-\delta$ finds an ε-approximation for the target. The idea of S_{PAC} is to simulate S_{LBD}. Any time S_{LBD} asks a query, c_i, the student S_{PAC} will try to approximate $d(c_i, t)$ using the examples given by the PAC-teacher. S_{PAC} works as follows:

- Simulate S_{LBD} for at most ℓ queries. Any time S_{LBD} presents a query c_i, ask the teacher for n examples $(y_1, b_1), \ldots, (y_n, b_n)$ (the value of n will be determined later). Define X_k to be 1 if $y_k \in t \triangle c_i$, and 0 otherwise (this can be determined using the bit b_k). Use $\frac{1}{n}\sum_{k=1}^{n} X_k$ as an approximation for $d(c_i, t)$. Output the first hypothesis c_i that yields approximated distance $d(c_i, t) < \frac{3\varepsilon}{4}$ (if within ℓ queries no such c_i is found then output arbitrary hypothesis. E.g., the c_i with the smallest known approximated distance to the target).

By the assumption, S_{LBD} $(\frac{\varepsilon}{2}, \frac{\varepsilon}{4})$ learns $\mathcal{C}$ uniformly in ℓ steps. This means that if the teacher gives the student the distances $d(c_i, t)$ with an error smaller than $\frac{\varepsilon}{4}$ then the student, within ℓ steps, finds an hypothesis in distance at most $\frac{\varepsilon}{2}$ of the target. Therefore, assuming that in our simulation all the approximated distances are within $\frac{\varepsilon}{4}$ from the real distances, then one of the ℓ queries is guaranteed to be at (real) distance $\frac{\varepsilon}{2}$ from the target. This distance will be approximated to at most $\frac{3\varepsilon}{4}$. On the other hand, every query with distance more than ε will be approximated to more than $\frac{3\varepsilon}{4}$. What we still have to show is how to choose n so that this condition holds with probability at least $1-\delta$. The probability that S_{PAC} fails, is not larger than ℓ times the probability that it fails in a single approximation. Using Lemma 3 we can bound the failure probability by:

$$\ell \cdot 2 \cdot e^{-2n\left(\frac{\varepsilon}{4}\right)^2}.$$

In order to make the above term be less than δ, it suffices to choose

$$n \geq \frac{\log 2\ell + \log \frac{1}{\delta}}{\frac{\varepsilon^2}{8} \cdot \log e}.$$

□

The PAC learning strategy of applying an LBD-student, as in the proof above, provides a clear separation between the accuracy and the confidence parameters of the PAC-model (the accuracy parameter ε and the confidence parameter δ). The success probability of S_{PAC} is fully determined by the number of examples, n, he requests for evaluating each distance $r(c_i)$. Independently, the accuracy of the approximation provided by S_{PAC}, is determined by the choice of the accuracy parameter ε for the LBD-strategy.

5.1 An Interactive Variant of PAC Learning

One way of viewing the above results is that they allow learnability through a protocol whose information-exchange parameters differ from those of the usual PAC protocol.

In the usual PAC scenario the communication is one-way – from the teacher to the student. (This is in accordance with viewing the PAC teacher as a way of modeling 'nature' or the ‘environment’ of an observing student). The LBD framework models a different scenario – its learning process is inherently interactive; the teacher’s responses are a function of the student’s queries.

The PAC protocol used in the proof of the above theorem, S_{PAC}, offers an interactive variant of PAC learnability. In that model, the *Interactive PAC model* defined below, the responsiveness of the teacher helps her save in the size of the messages she has to transmit.

For $0 \leq p \leq 1$, let X_p denote a random variable that equals 1 with probability p and equals 0 otherwise.

Definition 6*: (Interactive PAC Model)* *Let $\mathcal{Y}$ be a set and $\mathcal{C} \subseteq 2^{\mathcal{Y}}$. Let P be a probability distribution over $\mathcal{Y}$. Given a target concept $t \in \mathcal{C}$, the communication between the teacher and the student is carried out in rounds. In each round the student transmits a hypothesis query $h \in \mathcal{C}$ and the teacher responses by a bit value of the random variable $X_{P(h \triangle t)}$.*

Note that for $y \subseteq \mathcal{Y}$ the random variable $X_{P(y)}$ can be viewed as an indication of the membership in y of a point in $\mathcal{Y}$ drawn randomly according to P.

Corollary 5*: (to Theorem 7)* *For every concept class $\mathcal{C} \subseteq 2^{\mathcal{Y}}$, a probability distribution P on $\mathcal{Y}$, ε and δ , if $\mathcal{C}$ is (ε, δ) PAC-learnable (by randomly drawn labeled examples), then $\mathcal{C}$ is (ε, δ)-learnable in the Interactive-PAC model (by binary random variables of the form $X_{P(y)}$).*

Proof: By Corollary 4, PAC learnability implies LBD. Now invoke the proof of Theorem 7. The student S_{PAC} (in that proof) is actually an Interactive-PAC student - it does not make use of the identity of the sampled points y, the only information it needs is the values of the random variables X_i. □

In other words, by allowing a reactive teacher, the amount of information a PAC-student has to receive can sometimes be drastically reduced (relative to the student in Valiant’s model [21]). For example, in the learning process of geometric shapes in the Euclidean space $[0, 1]^n$ (relative to the uniform distribution) the family of learnable classes loses nothing if the information given by the teacher is obscured by erasing the real vector $\overline{x}$ in every labeled example $< \overline{x}, b >$ and replacing the bit b by a bit b^c which is 1 iff $\overline{x} \in c \triangle t$ (where c is the current student’s query and t is the target concept).

From the communication complexity point of view, the Interactive-PAC model may, in some cases, be more efficient than the PAC model (even when the overall communication is considered). In the usual PAC scenario, the messages transmitted by the teacher are pairs consisting of a point in the underlying space $\mathcal{Y}$ and a membership (in the target) bit. Such messages are $\log |\mathcal{Y}| + 1$ bits long. On the other hand, an Interactive-PAC teacher transmits just a single binary bit in every communication round. In each such round the student transmits the identity of a hypothesis concept – $\log |C|$ many bits. Consequently, the relative communication complexity of these models depends upon the ratio between $|\mathcal{Y}|$ and $|\mathcal{C}|$.

Although for many learning problems $|\mathcal{Y}| > |\mathcal{C}|$, favoring the PAC communication protocol, this is not always the case. In particular, when the underlying space and the concept class are both infinite, if the cardinality of the concept class is smaller than that of the underlying space, (e.g. when considering a countable concept class over a Euclidean space R^n), then the Interactive-PAC approach offers a better overall communication complexity.

The above considerations can also be applied to distribution-free PAC-learnability:

Corollary 6: *If $C \subseteq 2^{\mathcal{Y}}$ is distribution-free PAC-learnable then, for every distribution P over $\mathcal{Y}$, C is Interactive-PAC-learnable by a student that has access to the metric function d_P on C.*

The following theorem shows that if one wishes to give up the information saving then, allowing the student access to labeled examples (as in the PAC model), the number of examples needed in the last theorem can be reduced.

Theorem 8: *Let $\mathcal{C} \subseteq \mathcal{X} \subseteq 2^{\mathcal{Y}}$, P, and d_P be as above. For every $\varepsilon, \delta > 0$, if there exists an LBD-student, S_{LBD}, that $(\frac{\varepsilon}{2}, \frac{\varepsilon}{4})$ learns $\mathcal{C}$ (in $(\mathcal{X}, d_P)$) uniformly in ℓ steps, then there exists a PAC-student S_{PAC} that (ε, δ) learns $\mathcal{C}$ (in $(\mathcal{Y}, P)$) after seeing $\frac{\log 2\ell + \log \frac{1}{\delta}}{\frac{\varepsilon^2}{8} \cdot \log e}$ examples.*

Proof: The PAC student S_{PAC} operates as in the proof of Theorem 7, except for the way he estimates distances. Rather than viewing new examples for every needed distance, S_{PAC} will ask for $n = \frac{\log 2\ell + \log \frac{1}{\delta}}{\frac{\varepsilon^2}{8} \cdot \log e}$ many examples at the very beginning of the process. He then uses this sample to estimate each of the ℓ many distances needed.
The same calculation as in the proof of Theorem 7 shows that this number of examples suffices for ensuring that with probability $\geq 1 - \delta$ all distances are within $\frac{\varepsilon}{4}$ from their true value. $\square$

It might be interesting to compare the number of examples needed by our student S_{PAC}, to known lower and upper bounds on PAC-learnability. For the sake of this comparison recall that ℓ (the number of steps in Theorem 7) is at most $CAP(\mathcal{C}, \frac{\varepsilon}{2})$. It follows that S_{PAC} can learn with $O\left(\frac{1}{\varepsilon^2} \cdot (\log 2CAP(\mathcal{C}, \varepsilon) + \log \frac{1}{\delta}\right)$.

Benedek and Itai [9] investigate the number of examples as a function of $n(\mathcal{C}, \varepsilon)$ which is the size of a maximal set of concepts that are pairwise ε-far. They get a lower bound of $\Omega\left(\log n(\mathcal{C}, 2\varepsilon) + \log(1 - \delta)\right)$. Note that $n(\mathcal{C}, \varepsilon)$ is closely related to $CAP(\mathcal{C}, \varepsilon)$, it is easy to see that $CAP(\mathcal{C}, \frac{\varepsilon}{2}) \geq n(\varepsilon) \geq CAP(\mathcal{C}, \varepsilon)$. By Theorem 3, in pseudo-metrics induced by PAC problems $CAP(\mathcal{C}, \varepsilon) = O\left(\left(\frac{1}{\varepsilon}\right)^{VCDim(\mathcal{C})}\right)$.

Ehrenfeucht et. al. [12] investigate the number of examples needed for distribution free PAC learning as a function of the VC-dimension. They prove a lower bound of $\Omega\left(\frac{1}{\varepsilon} \cdot (d + \log \frac{1}{\delta})\right)$, where d denotes the VC-dimension of the concept class to be learnt. (As a matter of fact they prove a slightly stronger result. Namely, for every concept class there exists a distribution such that any algorithm learning the class with respect to this *fixed* distribution needs the above number of examples). Using Theorem 3 again, our upper bound becomes $O\left(\frac{1}{\varepsilon^2} \cdot (d \cdot \log \frac{2}{\varepsilon} + \log \frac{1}{\delta})\right)$. This bound holds for every fixed distribution.

One should note that, by [10], any consistent algorithm receiving labeled examples learns within $max\{\frac{8d}{\varepsilon} \cdot \log \frac{13}{\varepsilon},\ \frac{4}{\varepsilon} \cdot \log \frac{2}{\varepsilon}\}$ many examples. For the cost of extra examples we gain, on top of the information saving discussed above, a relaxation in the needed measurability constraints. This is the focus of the next subsection.

5.2 Relaxing the PAC Measurability Constraints

The last issue we would like to address in light of the tight connection between these models is the measurability problem. A fundamental theorem of PAC-learnability is the following Blumer et al. [10] characterization of distribution-free learnability in terms of the VC–dimension.

Theorem 9: ***(Blumer et al.)*** *For a* well behaved *concept class $\mathcal{C}$ the following are equivalent:*

1. *$\mathcal{C}$ has a finite VC–dimension.*
2. *$\mathcal{C}$ is uniformly learnable by any consistent student (a student who picks any hypothesis consistent with its input examples).*
3. *$\mathcal{C}$ is uniformly learnable (i.e. there exists a successful student for $\mathcal{C}$).*

The well-behavior condition is a complicated measurability constraint which is practically impossible to verify. (Luckily, it is usually satisfied by concept classes of interest). This condition is necessary for the implication from (1) to (2). There are examples of concept classes of Borel subsets of $[0,1]$ that have VC–dimension 1 and yet a naive consistent algorithm fails to learn them even with respect to the fixed uniform distribution on the unit interval (see [7] or the appendix of [10]).

The following consequence of our analysis shows that in the context of fixed distributions the implication from (1) to (3) is robust against any measurability difficulties.

Theorem 10: *A concept class $\mathcal{C} \subseteq 2^{\mathcal{Y}}$ has a finite VC–dimension if and only if it is learnable with respect to every fixed distribution on $\mathcal{Y}$ (i.e. for every distribution on $\mathcal{Y}$ there exists a uniform PAC-learning algorithm for $\mathcal{C}$).*

Proof: If $\mathcal{C}$ has a finite VC–dimension then by Theorem 4 for every distribution P on $\mathcal{Y}$, $\mathcal{C}$ is learnable in (X, d_P). Theorem 7 shows that $\mathcal{C}$ is PAC-learnable with respect to P. For the other direction one can apply the proof of this implication (from (3) to (1) of Theorem 9) in [10], it does not depend upon any measurability assumptions. □

Let us mention again that the improvement here over the results of [10] and [9] is that we assume no well-behavior condition. On the other hand, our theorem applies to fixed distributions i.e. the learning algorithm depends upon the underlying distribution whereas the consistent student in [10] is a fixed algorithm that handles every distribution.

The equivalence between conditions 2 and 3 in Theorem 9 is disappointing in the sense that it rules out the possibility that one algorithm may have better learning capabilities than another - all consistent algorithms can learn exactly the same classes. Theorem 10 implies that, in the context of fixed-distribution-learning (at least), this equivalence is broken. There exist learning algorithms (e.g., those based on an LBD-student) that can learn classes that are not learnable by other consistent PAC-students.

5.3 Relations with Other Models

In this subsection we exhibit the possibility of viewing various other learnability models, as special cases of the LBD model. More specifically, we consider the framework for learning a concept class using certain types of *queries*, presented by Angluin [1]. We show that some types of queries can be considered as queries for a "distance" information, under the appropriate definition of a metric space.

One of the interesting types of queries is the *equivalence queries*. In this model, the student guesses a concept and gets a $SUCCESS$ if he guessed the target. If he did not guess the target he gets a $FAIL$ together with a counterexample to the student's hypothesis (see e.g. [1, 2, 3, 5, 4, 20, 17], for examples of work that use this model). For example, consider the problem of learning classes of languages (e.g. regular languages, context-free languages). The model allows the student to ask queries which are languages (using some fixed representation). The teacher answers with $SUCCESS$ or with a counterexample (e.g., a string which is in the symmetric difference of the target language and the hypothesis).

This model has several variants, depending on the way in which the teacher chooses the counterexample. Consider the following two variants:

1. The teacher's counterexample is the first word in the lexicographical order on which the target language and the query-language do not agree [20, 17],

2. The teacher's counterexample is only the *length* of the first word on which the target language and the query-language do not agree [17].

One can formulate the known results in these two variants of the model using the terminology of the LBD-model by applying the following definitions: Let $\mathcal{X} = \{L | L \subseteq \{0,1\}^*,\ L \text{ is regular}\}$. Denote by $w_1, w_2, w_3, \ldots$ the words in $\{0,1\}^*$ in a lexicographical order. Let L_1, L_2 be any two regular languages, and let i be the minimal index such that $w_i \in L_1 \triangle L_2$ (i.e. w_i belongs to exactly one of the two languages). Finally, define $d_1(L_1, L_2) = \frac{1}{i}$, and $d_2(L_1, L_2) = \frac{1}{|w_i|+1}$. The main results in these two models become:

- In the metric space $(\mathcal{X}, d_1)$ there exists a polynomial-time LBD-student that exactly identifies the target [20, 17] (the polynomial is in the number of states of an automata accepting the target language, which is part of the student's input).

- In the metric space $(\mathcal{X}, d_2)$ no such polynomial-time student exists [2, 3, 17]. (Actually, the overall result of these papers put together is stronger).

A different variant of the learning by equivalence queries model is called in [1] the *restricted* model. In this variant the student, when coming up with a wrong hypothesis, gets only $FAIL$ but does not get any counterexample with it. This model can be viewed as an LBD-problem by endowing the concept class with the discrete metric. (As already mentioned in this metric space there is no better strategy than exhaustively searching for the target.) Angluin shows how a student learning in this model can be simulated by a PAC-student. Her simulation [1, Section 2.4] of ℓ restricted equivalence queries takes $O(\ell^2)$ random examples (if we ignore ε and δ). Our simulation in Theorem 7 is similar to hers but improves the size of the needed sample to $O(\ell \log \ell)$. It should be noted however, that

Angluin's simulation works also for the unrestricted model, and that the number of equivalence queries required for learning certain concept classes in the unrestricted model may be exponentially smaller than the number of restricted equivalence queries [1, section 3.1].

6 Acknowledgment

We would like to express our gratitude to Hugo Krawczyk for his help in clarifying the statistical issues related to this work.

References

[1] Angluin D., "Queries and Concept Learning" *Machine Learning*, 2(4), pp. 319-342, 1988.

[2] Angluin D., "Equivalence Queries and Approximate Fingerprints" *Proc. of 2nd COLT* pp. 134-145, 1989.

[3] Angluin D., "Negative Results for Equivalence Queries", *YALE-DCS-RR-648*, 1988.

[4] Angluin D., "Learning Regular Sets from Queries and Counterexamples", *Information and Computation*, 75, pp. 87-106, 1987.

[5] Angluin D., "Types of Queries for Concept Learning", *YALE-DCS-TR-479*, 1986.

[6] Ben-David S., G. M. Benedek, and Y. Mansour, "A Parameterization Scheme for Classifying Models of Learnability", *Proc. of 2nd COLT* pp. 285-302, 1989.

[7] Ben-David S., and G. M. Benedek, "Measurability Constraints on *PAC* Learnability", Preprint, 1989.

[8] Ben-David S., and R. Fraiman, "Algorithms for Learning by Distances", Technion Technical Report ♯750, 1992.

[9] Benedek G. M., and A. Itai, "Learnability by Fixed Distributions", *Proc. of 1st COLT* pp. 80-90, 1988, to appear in Theoretical Computer Science.

[10] Blumer A., A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and The Vapnik-Chervonenkis Dimension", *JACM*, 36(4), pp. 929-965, 1989.

[11] Dudley R. M., "A Course on Empirical Processes", *Lecture Notes in Mathematics, 1097*, 1984.

[12] Ehrenfeucht A., D. Haussler, M. Kearns, and L. Valiant, "A General Lower Bound on the Number of Examples Needed for Learning", *Information and Computation*, 82, pp. 247-261, 1989.

[13] Gale D., and F. M. Stewart, "Infinite Games with Perfect Information", *Annals of Mathematics*, Vol. 28, pp. 245-266, 1953.

[14] Garey M. R., and D. S. Johnson, "Computers and Intractability: A Guide to The Theory of NP-Completeness", 1979.

[15] Haussler D., "Generalizing the PAC Model: Sample Size Bounds From Metric Dimension-Based Uniform Convergence Results", *Proc. of 29th FOCS* pp. 40-45, 1988.

[16] Hoeffding W., "Probability Inequalities for Sums of Bounded Random Variables", *Journal of the American Statistical Association*, 58, pp. 13-30, 1963.

[17] Ibarra O., and T. Jiang, "Learning Regular Languages from Counterexamples", *Proc. of 1st COLT* pp. 371-385, 1988.

[18] Kolmogorov A. N., and V. M. Tihomirov, *AMS Translations*, Series 2, Vol. 17, 277-364, 1961.

[19] Linial N., Y. Mansour, and N. Nisan, "Constant Depth Circuits, Fourier Transform, and Learnability", *Proc. of 30th FOCS* pp. 574-579, 1989.

[20] Porat S., and J. Feldman, "Learning Automata from Ordered Examples", *Proc. of 1st COLT* pp. 386-396, 1988.

[21] Valiant L. G., "A Theory of The Learnable", *CACM*, 27(11), pp. 1134-1142, 1984.

[22] Vapnik V. N., "Estimation of Dependences Based on Empirical Data", Springer-Verlag, New York, 1982.

[23] Vapnik V. N., and A. J. Chervonenkis, "On Uniform Convergence of Relative Frequencies of Events to Their Probabilities", *Theory of Prob. and Appl.*, 16, pp. 264-280, 1971.