MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES

3 4456 0358218 4

ornl

ORNL/TM-11844
(CESAR-91/10)

OAK RIDGE
NATIONAL
LABORATORY

MARTIN MARIETTA

Analysis of Complexity Bounds for Pac-Learning with Random Sets

E. M. Oblow
V. R. R. Uppuluri

OAK RIDGE NATIONAL LABORATORY
CENTRAL RESEARCH LIBRARY
CIRCULATION SECTION
4500N ROOM 175
LIBRARY LOAN COPY
DO NOT TRANSFER TO ANOTHER PERSON
If you wish someone else to see this report, send in name with report and the library will arrange a loan.

MANAGED BY
MARTIN MARIETTA ENERGY SYSTEMS, INC.
FOR THE UNITED STATES
DEPARTMENT OF ENERGY

This report [illegible]

Available [illegible] Technical Information [illegible] 676 [illegible]

[illegible]

[illegible]

ORNL/TM-11844
CESAR-91/10

Engineering Physics and Mathematics Division

ANALYSIS OF COMPLEXITY BOUNDS FOR PAC-LEARNING WITH RANDOM SETS

E. M. Oblow
V. R. R. Uppuluri

DATE PUBLISHED — August 1991

Research sponsored by the
Office of Engineering Research Program
Basic Energy Sciences
U.S. Department of Energy
and
Office of Naval Research
U.S. Department of Defense

Prepared by the
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37831
managed by
MARTIN MARIETTA ENERGY SYSTEMS, INC.
for the
U.S. DEPARTMENT OF ENERGY
under contract DE-AC05-84OR21400

MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES
3 4456 0358218 4

CONTENTS

ACKNOWLEDGMENTS v
ABSTRACT vii
1. INTRODUCTION 1
2. LEARNABILITY THEORY 3
2.1 PAC-LEARNING FORMALISM 3
2.2 RANDOM SETS 4
2.3 ERROR MEASURE MOMENTS 5
3. MEASURES OF RANDOM SETS 7
3.1 RANDOM SET THEOREMS 7
3.2 ILLUSTRATIVE FINITE-SPACE PROBLEM 8
4. ANALYSIS OF COMPLEXITY BOUNDS 11
4.1 PERFECT LEARNING 11
4.1.1 Semi-Uniform Distribution Case 14
4.1.2 Worst-Case Distribution 15
4.1.3 Test Results 16
4.2 APPROXIMATE LEARNING 16
4.2.1 Positive Or Negative Examples Only 17
4.2.2 Positive and negative examples together 19
5. CONCLUSIONS 22
REFERENCES 23

LIST OF FIGURES

Fig.		Page
1	Partition probabilities	9
2	Perfect learning partition probabilities	11
3	Generic perfect learning with negative examples	12
4	Approximate learning partition probabilities	17
5	Approximate learning empirical partition results	20

LIST OF TABLES

Table		Page
1	Bounds for several problems with $\epsilon = 0.1$ and $\delta = 0.5$.	16

ACKNOWLEDGMENTS

The authors wish to thank Oscar Manley of the Department of Energy's Basic Energy Science Section and Alan Meyrowitz of the Office of Naval Research for their continuing financial support of random set research. The suggestions and comments of my colleagues Chuck Glover and N. Rao were also invaluable.

ABSTRACT

Learnability in Valiant's pac-learning formalism is reformulated in terms of expected (average) error instead of confidence and error parameters. A finite-domain, random set formalism is introduced to develop algorithm-dependent, distribution-specific analytic error estimates. Two random set theorems for finite concept-spaces are presented to facilitate these developments. Analyses are carried out for several illustrative problems with worst-case and semi-uniform distributions of learning examples. Analytic bounds on the sample size needed to achieve a specified average error are established. Useful approximations for these bounds and a worst-case distribution are also derived. Conclusions are drawn about the potential value of average-error bounds in improving the stated efficiency of pac-learning algorithms.

1. INTRODUCTION

This paper takes an alternative view of Valiant's theory of learnability (Valiant 1984, 1985) and the generalizations to it published by Blumer, et al., (1989). Valiant's original approach defines a probabilistic basis for approximating concepts from a class of concepts that in many cases might otherwise be intractable to learn. For this reason, the theory has been termed pac-learning, an acronym for *probably approximately correct* learning. While this seminal theory has renewed interest in the field of machine learning, much of the work to date has been theoretical (see e. g. Haussler & Pitt 1988; Rivest, et al., 1989; Haussler 1988a,b). Such efforts have focused primarily on the existential issue of establishing complexity bounds for algorithms which learn concepts from a class under worst-case scenarios.

The major strength of the pac-learning formalism is in establishing algorithm-independent, distribution-free complexity bounds. This is also, we feel, one of its weakness. Since existential bounds are not algorithm-dependent or distribution-specific, pac-research results tend lack tight constructive proofs and can not be used to generate optimal algorithms for specific problem classes. In addition, the use of both error and confidence parameters to define learnability appears to be unnecessary and difficult to use in practice. Other, simpler performance measures are often the only ones that can be readily evaluated. The impact of some of these limitations on practical problem solving have been recognized and discussed in Haussler 1990.

To address these problems, we take another look at the pac-learning formalism with an eye towards simplifying its basis. The broad perspective of information-based complexity theory (Traub, et al., 1988) has shown that many levels of approximation exist for "solving" computationally intractable problems. These levels are not just restricted to worst-case analyses. Often average-case scenarios make a deterministically intractable problem amenable to practical low-order algorithmic solution. The strong interest computer scientists have in distinguishing between P and NP problem classes has limited the impact these other methods might have on machine learning algorithm development. We suggest that better analytic tools and tighter bounds on sample complexity can be found more easily by studying other levels of approximation and distribution-specific methods (see e. g. Benedict & Itai 1990 or Haussler 1990 for a first step in this direction). In addition, we feel the usefulness of pac-methods would also be improved by redefining learnability in a more practical fashion.

For these reasons, we propose a definition of pac-learnability in terms of moments of the probabilistic error distribution. In particular, a definition which uses only the first moment or average error has distinct advantages over the conventional pac-learnability definition. The average error is both easier to deal with analytically and can be readily estimated in practice. This simpler approach could be termed paca-learning (probabilistically approximately correct on average). Valiant's definition, which uses aconfidence measure approach to the distribution of errors, is somewhat cumbersome and unneeded in practice (i. e. most problems appear to be insensitive to the confidence parameter the theory requires). To make maximum use of this reformulation we will adopt a finite, random set framework for our analysis.

This approach allows some powerful, yet simple, theorems to be used to generate useful algorithm-dependent, distribution-specific, analytic results.

In the following sections of this paper, Valiant's theory will first be reviewed and then redefined in random set terms. Two theorems for random set measures on finite concept-spaces will then be presented. These theorems will be used to determine analytic sample complexity bounds in some illustrative cases. Several sampling scenarios (i. e.semi-uniform and worst-case distributions) will be dealt with to highlight the practical applicability of this approach. All the developments will preserve the distribution-specific nature of the learning examples so that easier learning scenarios (i. e. *not* worst-case, distribution-free) can be readily analyzed.

2. LEARNABILITY THEORY

To begin, we briefly review Valiant's pac-learning theory. Since this formalism has been generalized considerably in the years since its introduction, we use the more recent notation of Blumer, et al., (1989). In this exposition we will explicitly highlight the random set aspects of pac-learning theory. For the sake of clarity, the discussions will be restricted to finite concept-classes and sample-spaces. In this framework, we can more easily discuss learnability in set-theoretic terms.

2.1 PAC-LEARNING FORMALISM

In Valiant's approach, the goal is to learn a target concept t which is a member of a class of target concepts T. For our purposes, T will be a finite class and its members will be defined in terms of sets of points x in a finite space of learning examples X. In this set framework, T is a subset of the power set of X (i. e. $T \subseteq 2^X$). Any $t \in T$ will, therefore, be a particular subset of points $t \subseteq X$. The complement of $t \in X$, denoted by $\tilde{t}$, will then be $\tilde{t} \equiv X - t$.

In order to analyze learnability probabilistically, a probability distribution $p(x)$ is assumed to be defined over X. A sample of m instances, denoted by $\bar{\eta} = (x_1, \ldots, x_m)$, is randomly drawn from X using this distribution to learn t or its complement $\tilde{t}$. Each element of this sample is assumed to be labelled t or $\tilde{t}$ [i. e. for any $x \in X$, label $l(x) = 1$ if $x \in t$, otherwise $l(x) = 0$]. This random, labelled m-sample, denoted by $\bar{x} = ([x_1, l(x_1)], \ldots, [x_m, l(x_m)])$, is an m-sample in the Cartesian product space $\bar{X}^m \equiv X^m \times \{0,1\}^m$.

A pac-learning algorithm $\mathcal{A}$ is defined as a function $\mathcal{A}(\bar{x})$

$$\mathcal{A}(\bar{x}): \quad \left[\bar{x} \in \bar{X}^m\right] \rightarrow \left[h \in H_m \subseteq H\right] \quad , \tag{1}$$

which takes as its input the m-sample $\bar{x}$ and produces as its output a hypothesis $h \in H$. For our purposes, H is defined to be the class of sets $H \subseteq 2^X$ that can be generated from any $\mathcal{A}$ using any finite or countably infinite sample. H_m will be used to denote the subset of H that can be generated from $\mathcal{A}$ with any fixed sample size m. In learning theory, any $h \in H$ is meant to be an approximation to any $t \in T$.

In more specific terms, h is said to probabilistically approximate t with error κ. This error is defined by the probability measure associated with the error set $r \in X$, defined as the symmetric difference between h and t. That is

$$\kappa \equiv p(r) = p(h \triangle t) \quad , \tag{2}$$

where the symmetric difference is defined by $r = h \triangle t = (t - h) \cup (h - t)$.

The error thus represents the probability of erroneously classifying a randomly drawn example $x \in X$ using hypothesis h as an approximation to t. The particular $r \in X$ in this expression is clearly just a single instantiation of a class of random error sets that can generated from $\mathcal{A}(\bar{x})$. Other members of this class are generated by the other random m-samples $\bar{x} \in \bar{X}^m$.

The number of sampled instances m is usually specified to be $m(\epsilon, \delta)$, a function of two parameters ϵ and δ, called the approximation *error* and *confidence*, respectively. Both of these parameters have values in the range $(0, 1)$. Given fixed values

for ϵ and δ, learnability is defined in Valiant's formalism to be the existence of an algorithm $\mathcal{A}(\bar{x})$ which produces a hypothesis $h \in H$ that has an error $\kappa(\bar{x}) > \epsilon$ with probability $\leq \delta$. The smallest sample size $m(\epsilon, \delta)$ which satisfies this criterion in a worst-case scenario over all $t \in T$, is denoted by L_m and called the *sample complexity.*

For the finite concept classes and finite X we will deal with in this paper, Blumer, et al., (1989) have shown that for any consistent learning algorithm, L_m can be bounded from above with a sample size m_B given by

$$L_m < m_B = \frac{1}{\epsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right) \quad , \tag{3}$$

where $|H|$ is the cardinality of the hypothesis class H and $T \subseteq H$.

2.2 RANDOM SETS

In order to discuss learnability in random set terms, we must first reinterpret the pac-learning approach in random set terms. As a first step, it is sufficient at this point to note the stochastic nature of two fundamental sets in Valiant's theory, the hypothesis set h and the error set r. Both are clearly random sets since they arise from a deterministic mapping $\mathcal{A}(\bar{x})$ of a random variable $\bar{x}$. That is, for a given value of m and a particular target set t, each probabilistic realization of $\bar{x}$ gives rise to a particular $h \in H_m$ and $r \in X$ as a result of the mapping given in Eq. (1). The probability of sampling any $\bar{x}$ is determined by the probability $\xi(\bar{x})$ of picking any random m-sample of points x from the probability distribution $p(x)$.

In random set theory (see e. g. Kendall 1977 and Goodman & Nguyen 1985) each set has two measures of importance. The first measure is associated with the range and the second with the domain of a deterministic map. The traditional issue addressed in this framework is that usually many measures are available for formulating solutions to a particular problem. Invariance arguments are then used to unify these measures. This approach is the focal point of a field like stochastic geometry.

In the learning case, invariance arguments are not needed. Both the domain and range measures are already defined uniquely. The domain measure, for example, is the probability $\xi(\bar{x})$ associated with each m-sample in $\bar{X}^m$ (i. e. a measure for each point $\bar{x} \in X^m$). The range measure is clearly $p(x)$. We can, therefore, develop expressions for the moments of the measures of a random set (discussed later in the paper). The central point for us here is, that since both h and r are sets in X, they both have probability measures in X. These measures, $p(x \in h)$ and $p(x \in r)$, respectively, are both generated from $p(x)$. Having measures defined over both the domain and the range of $\mathcal{A}(\bar{x})$ allows us to develop a random set basis for pac-learning.

As far as practical benefits are concerned, we feel that an analysis of the moments of a random set (specifically the error-set) is more useful than the standard pac-learning approach. In fact, average-error learnability is directly related to the problem of estimating learning curves. Ideally, a learning curve is a measure of an algorithm's performance error at any stage in a learning procedure. In practice, this curve is usually constructed by keeping track of the number of errors made in a running window of examples. The difficulties of using the ϵ and δ parameters

in pac-learning theory to estimate such a learning curve are well noted by most practical problem solvers and acknowledged by Haussler (1990). A moments-based approach to Valiant's theory is one way to remedy this situation. Several random set theorems will be developed to make this moment approach easier to work with both analytically and in a algorithm-dependent, distribution-specific framework.

2.3 ERROR MEASURE MOMENTS

To make practical use of the measure properties of the random sets h and r, we propose a redefinition of learnability in the pac-formalism. Instead of using both error and confidence parameters, we drop the confidence (since it plays only a minor role in most analyses anyway) and work with moments of the measures of the random error sets. Specifically, we will be most interested in the first and second moments, the average error and its variance, respectively. Since the probability of making a misclassification error is defined in Eq. (2), the expectation (average) of the measure of the random set r is the average misclassification error resulting from hypothesis h. This expectation, for a given target t and a given sample size m, is defined in general as

$$E\left[p(r)\right] \equiv \overline{\kappa} = \sum_{h \in H_m} p(h \triangle t)\xi(h) \quad , \tag{4}$$

where $\xi(h)$, called the ξ-measure of h, is the probability of generating the random set h from $\mathcal{A}$. It is defined as

$$\xi(h) \equiv \sum_{\mathcal{A}(\bar{x})=h} p(\bar{x}) = \sum_{\mathcal{A}(\bar{x})=h} p(x_1)p(x_2)\cdots p(x_m) \quad . \tag{5}$$

That is, $\xi(h)$ is the sum of the probabilities of all the m-samples, $\bar{x}$, from which the algorithm generates h.

For fixed m, $\mathcal{A}$ can generate hypotheses only in the class H_m, so the distribution over all m-samples of $\bar{x}$ is normalized by

$$\sum_{\bar{x} \in \bar{X}^m} p(\bar{x}) = \sum_{h \in H_m} \xi(h) = 1 \quad . \tag{6}$$

The higher moments of the error distribution for fixed m and t are likewise defined by

$$E\left[p^k(r)\right] \equiv \overline{\kappa^k} = \sum_{h \in H_m} p^k(h \triangle t)\xi(h) \quad . \tag{7}$$

Using these definition we can formally redefine learnability as the existence of an algorithm $\mathcal{A}(\bar{x})$ which produces a hypothesis h that has an *expected error* $E[\kappa(\bar{x})] < \epsilon$. Higher moments can also be included in this definition, but we will forgo this added complication for the sake of simplicity. The smallest sample size m which satisfies this criterion in a worst-average error scenario for all targets $t \in T$, will be denoted as $L_{\overline{m}}$ and called the *average sample-complexity* for learning this class.

The usefulness of this redefinition will become clearer after we present some theorems for measures of random sets in certain problem classes. At this point, however, it is important to point out that in some cases there may be little difference between the traditional Valiant definition of learnability and the proposed random set version. The two may be comparable when $\delta = 0.5$. In this case, Valiant's definition produces a median error while the random set approach produces an average error. While these two error measures can, in general, be very different, in some cases the differences are minimal. In any event, where the differences are great, the random set analysis can be carried to higher moments of the error-set measures to get results to compare with the traditional pac-approach.

3. MEASURES OF RANDOM SETS

The practical value of redefining learnability as we did in the last section is based to a large extent on the availability of several powerful random set theorems. These theorems go back at least to Robbins (1944) who derived general expressions for the moments of the measure of a random set. They provide the basis for developing analytic expressions for sample complexity in a number of important cases. Since we are primarily interested in finite-space learning, we will make use of Robbins' results only for this more specialized class of problems. More general continuous-space analyses, however, are possible with the integral forms that are the basis of Robbins' original theorems.

3.1 RANDOM SET THEOREMS

The two primary theorems presented by Robbins (1944) deal with the moments of the measures of a random set in continuous measure spaces. The finite-space counterparts of these theorems suitable for our learning work can be developed along the following lines:

Definition 1. Let r be a random ξ-measurable set in the finite space X. Then for every possible $r \subseteq X$, there is a ξ-measure $\xi(r)$. If $R \subseteq 2^X$ is the space of all possible sets r, then this means that the probability that r belongs to any ξ-measurable subset $S \subseteq R$ is

$$\xi(S) = \xi(r \in S) \equiv \sum_{r \in R} g(r, S)\xi(r) \quad , \tag{8}$$

where $g(r, S)$ is the indicator function

$$g(r, S) = \begin{cases} 1 & \text{if } r \in S \quad ; \\ 0 & \text{if } r \notin S \quad . \end{cases} \tag{9}$$

Definition 2. Let x be a random p-measurable point in X. Then for every $x \in X$ there is a p-measure $p(x)$. The probability that x belongs to any p-measurable subset $r \subseteq X$ is then

$$p(r) = p(x \in r) \equiv \sum_{x \in X} g(x, r)p(x) \quad . \tag{10}$$

Definition 3. Given the product space $X \times R$ and a p, ξ-measure for every pair (x, r), let S_x be the space of all $r \in R$ that contain point x. Then, if $x \in r$ we have $r \in S_x$ and if $x \notin r$ we have $r \notin S_x$. Thus $g(r, S_x) = g(x, r)$, $\forall (x, r)$ pairs.

Robbins (1944) uses continuous versions of these definitions together with Fubini's theorem to prove two integral theorems. For our purposes, we can use the simple relationship

$$\sum_{r \in R} \sum_{x \in X} g(x, r)p(x)\xi(r) = \sum_{x \in X} \sum_{r \in R} g(r, S_x)\xi(r)p(x) \quad , \tag{11}$$

to state finite-space counterparts of these integral theorems.

Theorem I. Let r be a random measurable subset of a finite space X with ξ-measure $\xi(r)$. Let x be a random measurable point in X with p-measure $p(r)$. For any possible measurable pair $(x, r) \in X \times R$, let $\xi(x \in r)$ denote the ξ-measure of any point $x \in X$ being a member of the random set r. Then the expected p-measure of r will be given by

$$E\left[p(r)\right] = \sum_{x \in X} p(x)\xi(x \in r) \quad , \tag{12}$$

where,

$$\xi(x \in r) = \sum_{r \in R} g(r, S_x)\xi(r) \quad . \tag{13}$$

If we restrict ourselves to probability measures, then $\xi(x \in r)$ is the probability that x is a member of a random set r. This probability is clearly equivalent to the ξ-measure of those random sets $r \in X$ that contain x.

A suitable extension of this theorem allows higher moments of the p-measure of r to be calculated.

Theorem II. Let r be a random measurable subset of the finite space X with ξ-measure $\xi(r)$. Let x be a random measurable point in X with p-measure $p(r)$. For any possible measurable $k+1$-tuple $(x_1, \ldots, x_k, r) \in X^k \times R$, let $\xi(x_1, \ldots, x_k \in r) \equiv \xi(x_1, \mathbf{and} x_2, \ldots \mathbf{and} x_k \in r)$ for a point $(x_1, \ldots, x_k) \in X^k$. Then the expected k-th moment of the p-measure of r will be given by,

$$E\left[p^k(r)\right] \equiv \sum_{x_1 \in X} \cdots \sum_{x_k \in X} p(x_1) \cdots p(x_k)\xi(x_1, \ldots, x_k \in r) \quad , \tag{14}$$

where

$$\xi(x_1, \ldots, x_k \in r) = \sum_{r \in R} g\left(r, S_{x_1}\right) \cdots g\left(r, S_{x_k}\right) \xi(r) \quad . \tag{15}$$

Here, when restricted to probability measures $\xi(x_1, \ldots, x_k \in r)$ is the probability that (x_1 **and** x_2 ... **and** x_k) are members of a random set r.

The use of the relationship in Eq. (11) and the definitions provided makes the proofs of these theorems trivial in the finite-space case. For continuous-space problems, Robbins (1944) provides simple, yet elegant proofs of both of these theorems. These latter proofs rely on the linear properties of the measures involved and Fubini's theorem to decouple and reverse the integration orders over the p, ξ-measure space. The key advantage of using these theorems in a pac-learning context is being able to more easily estimate the ξ-measures. In learning (as we will soon see), these measures represent the probability of any example(s) being a member of a random error set r. They allows analytic estimates of average errors and variances to be made in some interesting classes of learning problems.

3.2 ILLUSTRATIVE FINITE-SPACE PROBLEM

To make practical use of the theorems just stated, we will focus our attention on one particular class of random hypothesis sets – disjoint covers of X. For illustrative

purposes, therefore, we assume that the hypothesis class H will be constructed from a fixed, finite partition of X into s disjoint sets which completely cover X. Each subset of this partition will be denoted by b_i, $i = 1, \ldots, s$. Any specific $t \in T$ will then give rise to two component subsets of each b_i, these being $b_i^+ \subseteq t$ and $b_i^- \subseteq \tilde{t}$. Here, we adopt the common convention of calling samples drawn from t as *positive* examples (i. e. $x \in t \equiv x^+$) and samples drawn from $\tilde{t}$ as *negative* examples (i. e. $x \in \tilde{t} \equiv x^-$).

Using this notation, the points x in each subset b_i will then have a probability of classification as t of p_i^+ and as $\tilde{t}$ of p_i^- with

$$p_i^+ = \sum_{x \in b_i^+} p(x) \quad \text{and} \quad p_i^- = \sum_{x \in b_i^-} p(x) \quad . \tag{16}$$

These probabilities are suitably normalized so that

$$\sum_{i=1}^{s} (p_i^+ + p_i^-) = 1 \quad , \tag{17}$$

and

$$p^+ \equiv p(x \in t) = \sum_{i=1}^{s} p_i^+ \quad \text{and} \quad p^- \equiv p\left(x \in \tilde{t}\right) = \sum_{i=1}^{s} p_i^- \quad . \tag{18}$$

A schematic table representing this partitioning is

	b_1	b_2	b_3	$\cdots$	$\cdots$	b_s
t	p_1^+	p_2^+	p_3^+	$\cdots$	$\cdots$	p_s^+
$\tilde{t}$	p_1^-	p_2^-	p_3^-	$\cdots$	$\cdots$	p_s^-

Fig. 1. Partition probabilities.

We now assume that the space of all possible hypotheses H generated by a learning algorithm $\mathcal{A}$ is the power set of these partition subsets. That is, if $B \equiv \{b_1, \cdots, b_s\}$ then $H = 2^B$ and $|H| = 2^s$. In general, the target space will still be defined as $T \subseteq 2^X$, but in one case to be discussed, we will let $H = T$. This latter class of problems is the counterpart of an infinite hypothesis class shattered by s points, making s the VC-dimension of the hypothesis class (see Blumer, et al., 1989 and Vapnik 1982).

With the general partition structure just defined, two subproblems will be discussed in more detail. The first, is a problem class in which perfect learning is possible. In this case $H = T$, so that $\forall b_i$ either $p_i^- = 0$ or $p_i^+ = 0$. The second, is a problem class in which only approximate learning is possible for some $t \in T$. That is, $t \notin H$ for some $t \in T$, so that for this t both $p_i^+ \neq 0$ and $p_i^- \neq 0$ in at least one subset of the partition.

These two cases are usually distinguished by the asymptotic behavior of the errors made in performance. In the perfect learning case, the hypothesis space contains a perfect representation for t (i. e. $t \in 2^B$) and the asymptotic error, therefore, approaches zero. For this case, we will, in addition, assume that once an example is seen it will be classified correctly in all future selections (i. e. consistent learning). In the approximate learning case, the representation is inadequate and the error cannot be reduced to zero even if every example is seen. Depending on the representation and the stage of the learning process, errors are, therefore, possible for both seen and unseen examples. Both of these cases are discussed more fully in Blumer, et al., (1989).

4. ANALYSIS OF COMPLEXITY BOUNDS

Using the random set formalism described in the last section, we can now derive analytic sample complexity bounds for several illustrative perfect and approximate learning problems. For each problem, two distribution-specific scenarios will be analyzed: (1) a worst-case scenario, in which the distribution of learning examples is such that a maximum average-error bound is achieved and (2) a semi-uniform scenario, in which learning examples will be uniformly distributed at the subset level in the partition structure. The first case, because it is a worst-case scenario, is equivalent to a distribution-free case. In the second case, we are dealing with a clearly distribution-specific problem. For this latter case, we will be assuming that what we call a semi-uniform distribution is uniform at the subset level. That is,

$$p_i^- = \alpha/s \quad \text{and} \quad p_i^+ = (1-\alpha)/s \quad , \quad \forall i \quad , \tag{19}$$

where α is the fraction of negative examples in the sample space.

Since this partition structure is simple, distribution-free bounds on sample complexity in the pac-learning definition of learnability can easily be stated. For any consistent learning algorithm, the sample complexity bound is given in Eq. (3) with $|H| = 2^s$. Though useful from an existential point of view, such a result provides few clues as to how to actually construct a worst-case distribution or how to take advantage of an easier learning scenarios (such as the semi-uniform case). We address these limitations in the analyses which follow.

4.1 PERFECT LEARNING

The first problem we will study is the perfect learning case. This problem can be solved using either positive or negative examples alone by analogy with Valiant's consistent learning algorithm for k-DNF's and k-CNF's (Valiant 1984). The approach we present is discussed in detail in Oblow (1990) and Oblow (1991). Using the random set framework given in this latter work, a representative problem for the partition structure given in Eqs. 16-18 and Fig. 1 can be defined as follows

	b_1	b_2	b_3	$\cdots$	$\cdots$	b_s
t	p_1^+	0	0	$\cdots$	$\cdots$	p_s^+
$\tilde{t}$	0	p_2^-	p_3^-	$\cdots$	$\cdots$	0

Fig. 2. Perfect learning partition probabilities.

Since uniform learnability over the whole target class $t \in 2^B$ is our concern, we only have to choose the hardest target in the class to learn. In Valiant's Boolean framework, clearly the elimination of all hypotheses to some probabilistic tolerance is the hardest learning problem. Thus, the target that is hardest to learn starting from the full hypothesis set is the null set. If we restrict our attention to algorithms that learn using only positive or negative examples alone (see e. g. Valiant 1984 or Oblow 1991), then a representative problem is generated by assuming that the target or its complement is the null set and it must be learned using only counterexamples. That is, if $t = \phi$, we have to learn t from negative examples starting with $h = X \approx t$; if $\tilde{t} = \phi$, we have to learn $\tilde{t}$ from positive examples starting with $h = X \approx \tilde{t}$. Without loss of generality, the first of these two cases will be used as our generic illustrative example. In this case, we will drop the negative example superscript (i. e. $p_i^- \equiv p_i$) and let the partition table be:

	b_1	b_2	b_3	$\cdots$	$\cdots$	b_s
t	0	0	0	$\cdots$	$\cdots$	0
$\tilde{t}$	p_1	p_2	p_3	$\cdots$	$\cdots$	p_s

Fig. 3. Generic perfect learning with negative examples.

With this representation, a *consistent* learning algorithm in which the hypothesis set always correctly classifies any examples that have already been seen can easily be developed. This algorithm starts with the default hypothesis $h = X \approx t$ and eliminates as inconsistent any subset b_i in which counterexamples (negative examples) are found. After seeing m examples, the algorithm returns an h which is the union of all subsets in which no counterexamples are found. If at least one counterexample is seen in each of the b_i, the null set will be learned perfectly and the performance error for this algorithm will be zero.

In such an algorithm, it is easy to see that the probability of any example being a member of the error set $r = h \triangle t = h$, is the same as the probability that it is a member of a partition subset with no observed counterexamples. The union of all such subsets will be called the *unseen* set and be denoted by u. Since all subsets in the partition which have been sampled will by definition be correctly classified in this case, the only mistakes that can be made by this algorithm are in selections from u (i. e. from subsets with no observed counterexamples).

In this way, the probability that a particular b_i contains no counterexamples after m samples is clearly $(1 - p_i)^m$. This is also the probability that no $x \in b_i$ will be selected in m random samples from $p(x)$ and, therefore, b_i will be a subset of u.

In random set terms, the probability that any arbitrary $x \in b_i$ is a member of the random set u after m random trials is then

$$\xi(h) = \xi(x \in u) = (1 - p_i)^m \quad . \tag{20}$$

The simplicity of Eq. (20), coupled with the generality of Theorem I allows us now to produce a problem-specific, analytic expression for the average error in this case.

Thus, for the hardest target case (i. e $t = \phi$), the expected measure of the random set $r = h = u$ is given by

$$\begin{aligned} E\left[p(r)\right] = \overline{\kappa} &= \sum_{x \in X} p(x)\xi\left(x \in r\right) \\ &= \sum_{i=1}^{s} \sum_{x \in b_i} p(x)\left(1 - p_i\right)^m \\ &= \sum_{i=1}^{s} p_i\left(1 - p_i\right)^m \quad . \end{aligned} \tag{21}$$

This expression represents the average performance error for such an algorithm. It, therefore, also represents a learning curve for this case as a function of m. Clearly, the error uniformly decreases to zero as examples in each previously unseen subset b_i are sampled. Since this expression holds for any probability distribution of examples $p(x)$, it also provides an analytic average-error estimate for any problem-specific distribution of examples in a worst-target scenario.

Again, because of the simplicity of Eq. (20), we can extend this analysis to include higher moments of the error measure. In this manner we can completely characterize the probabilistic distribution of error measures through its moments. In this perfect learning case, Theorem II can be used to give higher moments as

$$\overline{\kappa^k} = \sum_{x_{i_1} \in X} \cdots \sum_{x_{i_k} \in X} p(x_{i_1}) \cdots p(x_{i_k}) \xi\left(x_{i_1}, \ldots, x_{i_k} \in r\right) \quad , \tag{22}$$

where

$$\xi\left(x_{i_1}, \ldots, x_{i_k} \in r\right) = \left[\, 1 - p\left(x_{i_1} \in r \text{ or } \cdots \text{ or } x_{i_k} \in r\right) \,\right]^m \quad . \tag{23}$$

Simplifying this expression, gives the final result that

$$\begin{aligned} \overline{\kappa^k} &= \sum_{j_1=1}^{s} \sum_{x_{i_1} \in b_{j_1}} p(x_{i_1} \in b_{j_1}) \cdots \sum_{j_k=1}^{s} \sum_{x_{i_k} \in b_{j_k}} p(x_{i_k} \in b_{j_k}) \\ &\quad \times \left[\, 1 - p(x_{i_1} \in b_{j_1} \text{ or } \cdots \text{ or } x_{i_k} \in b_{j_k}) \,\right]^m \\ &= \sum_{j_1=1}^{s} p_{j_1} \cdots \sum_{j_k=1}^{s} p_{j_k} \left[1 - (p_{j_1} + \cdots + p_{j_k})\right]^m \quad , \end{aligned} \tag{24}$$

where, because of the *or* operation, the probabilities can only be summed if the j's are distinct. We thus artificially define

$$p_j + p_{j'} = p_j \quad \text{if} \quad j = j' \quad , \quad \forall j, j' \quad . \tag{25}$$

This expression completely characterizes the distribution of error measures that arises from randomly sampling examples from $p(x)$. The fact that this result is simple, allows some interesting analyses to be carried out for specific sampling distributions.

4.1.1 Semi-Uniform Distribution Case

For the semi-uniform case defined as $p_i = 1/s, \forall i$, it is easily seen from Eq. (21) that

$$\overline{\kappa} = \left(1 - \frac{1}{s}\right)^m \quad . \tag{26}$$

The average performance error at any stage in this consistent learning procedure is thus $O(e^{-m/s})$.

Solving for sample size using this result, we see that the sample complexity bound for average-error learnability to within ϵ is

$$L_{\overline{m}} = \frac{\ln \epsilon}{\ln(1 - \frac{1}{s})} \approx s \ln \frac{1}{\epsilon} \quad . \tag{27}$$

This expression is linear in s, as is Blumer's result in Eq. (3) with $|H| = 2^s$, but it is logarithmic in $1/\epsilon$ as opposed to linear in Blumer's case.

The second moment of this distribution can likewise be evaluated. Using Theorem II we see that

$$\begin{aligned} \overline{\kappa^2} &= \sum_{i_1=1}^{s} \sum_{i_2=1}^{s} p_{i_1} p_{i_2} \xi\left(x_1, x_2 \in r\right) \\ &= \frac{1}{s}\left(1 - \frac{1}{s}\right)^m + \frac{s-1}{s}\left(1 - \frac{2}{s}\right)^m \quad . \end{aligned} \tag{28}$$

This expression allows the variance in the error distribution to be expressed as

$$\begin{aligned} V(\kappa) &= \overline{\kappa^2} - \overline{\kappa}^2 \\ &= \frac{1}{s}\left(1 - \frac{1}{s}\right)^m + \frac{s-1}{s}\left(1 - \frac{2}{s}\right)^m - \left(1 - \frac{1}{s}\right)^{2m} \quad , \end{aligned} \tag{29}$$

which for large s is

$$V(\kappa) \approx \frac{1}{s}\left(1 - \frac{1}{s}\right)^m \quad . \tag{30}$$

In this case, we see that the variance is clearly $O(\overline{\kappa}/s)$, yielding high confidence in $\overline{\kappa}$ for large s.

4.1.2 Worst-Case Distribution

The more interesting scenario for perfect learning is the worst-average distribution problem, which is equivalent to a distribution-free case by definition. For this case we can also get analytic results and bounds on sample complexity. To obtain such results, we first have to find the probability distribution that yields a worst-average error for any arbitrary sample size m.

We first observe that since the individual maximum for any term in the sum over i in Eq. (21) is given by

$$p_i = \frac{1}{m+1} \quad , \tag{31}$$

the maximum average-error can be bounded by

$$\overline{\kappa}_{\max} \leq \frac{s}{(m+1)} \left(\frac{m}{m+1} \right)^m < \frac{s}{me} \quad . \tag{32}$$

To determine the distribution that actually produces the maximum average-error, we let p_1 be a normalization constant and differentiate Eq. (21) with respect to each p_i, $i = 2, \ldots, s$. Setting all these derivatives equal to zero, we get

$$(1-p_i)^{m-1}\left[1-(m+1)p_i\right] = (1-p_1)^{m-1}\left[1-(m+1)p_1\right] \quad \forall i > 1 \quad , \tag{33}$$

subject to the normalization condition

$$p_1 = 1 - \sum_{i=2}^{s} p_i \quad . \tag{34}$$

Since each p_i term is equal to the same constant, the worst-average distribution can be seen to be a member of a family of distributions with a single parameter β

$$p_i = \begin{cases} 1-\beta & \text{if } i = 1 \quad ; \\ \beta/(s-1) & \text{if } i > 1 \quad . \end{cases} \tag{35}$$

For the case of $m > s$, the system of equations given in Eq. (33) can be solved with an approximation $O[(s/m)^m]$ to find the β giving the maximum average-error. This result is

$$\beta \approx e\epsilon \quad . \tag{36}$$

The worst-average error is thus approximately

$$\overline{\kappa}_{\max} \approx \frac{s-1}{em} \quad , \tag{37}$$

and the worst-average sample complexity is, therefore, given by

$$L_{\overline{m}} \approx \frac{s-1}{e\epsilon} \quad . \tag{38}$$

This expression should be compared to Blumer's bound with $\delta = 0.5$ given in Eq. (3) (i. e. the worst-median sample complexity).

4.1.3 Test Results

To illustrate some distribution-specific results, the specific family of distributions given in Eq. (35) was used to test the analytic results given in Eq. (38). For comparison purposes, a simple computational algorithm was written to stochastically estimate a pac-learning complexity bound L_m for the β-family, given s, ϵ, δ. The estimation program uses samples from any member of the β-distribution family to find the sample size m_β which satisfies the condition that the error be less than ϵ with probability less than δ. It steps through m and integrates the actual error distribution to arrive at its results. For each β value, each m-sample is repeated several thousand times to get adequate statistics (i. e. accuracy to within a few percent). L_m is then defined as the smallest value of m_β over the whole β-family. Comparisons were then made with the analytic values of $L_{\overline{m}}$ which we derived and Blumer's bound m_B.

Results for several selected cases are given in Tables 1. The table covers a range of s values from 16 to 256 and a large case of 2048. Bounds are given for an error value of $\epsilon = 0.1$ and (for pac-results) a confidence value of $\delta = 0.5$. Blumer's bound m_B and $L_{\overline{m}}$ are given for each case for comparison purposes.

s	$L_{\overline{m}}$	L_m	m_B
16	55	60	112
32	114	119	217
64	232	237	428
128	468	472	849
256	939	945	1,690
2048	7,532	7,557	13,476

Table 1. Bounds for several problems with $\epsilon = 0.1$ and $\delta = 0.5$.

The results in this table show clearly that the worst-average errors computed from Eq. (38) are a good approximation to the pac-learning worst-median sample complexity for the β-family of distributions used. At worst (for $s = 16$), the largest disagreement between the worst-median sample complexity L_m and the worst-average value $L_{\overline{m}}$ is only 10%. The analytic expression for $L_{\overline{m}}$ should, therefore, provide useful information in practical problem solving. Since Blumer's bound m_B is about a factor of two higher than the $L_{\overline{m}}$ bound in all of these cases, $L_{\overline{m}}$ appears to be a useful lower bound to sample complexity even in the pac-learning formalism.

4.2 APPROXIMATE LEARNING

The more difficult problem we address now is approximate learning. This case can arise in practice in several different ways, the two most important being: 1)

perfect learning with noise and 2) learning with an inadequate representation (i. e. $t \notin H$ for some $t \in T$). The inadequate representation case will be assumed in the developments to follow simply because it is slightly more general. For our illustrative example, this means that the partition structure will be inadequate to exactly represent the target. A probabilistic approximation to any target $t \in T$ is then the best solution possible. A schematic way of looking at this general approximate learning problem is given in Fig. 4.

	b_1	b_2	b_3	$\cdots$	$\cdots$	b_s
t	p_1^+	p_2^+	p_3^+	$\cdots$	$\cdots$	p_s^+
$\tilde{t}$	p_1^-	p_2^-	p_3^-	$\cdots$	$\cdots$	p_s^-

Fig. 4. Approximate learning partition probabilities.

This is the same format as given in Fig. 1, except for the stipulation that because $t \notin H$ for some $t \in T$, we will have both $p_i^+ > 0$ and $p_i^- > 0$ in at least one subset b_i. Under these conditions, the best any learning algorithm can do with this representation is produce an approximate error (i. e. $\overline{\kappa} > 0$). To discuss this case in a useful manner, we note that in each partition subset b_i, the smallest asymptotic error for any target can be generated by choosing the subset label (i. e. t or $\tilde{t}$) that has the highest probability. That is, the best approximate label in each b_i is the larger of p_i^+ or p_i^-. This does not necessarily produce the best set representation of the target, only the smallest performance error. Since these probabilities are not known beforehand, the learning problem will be to empirically approximate this best asymptotic solution.

Several useful algorithms can be constructed to approximate the solution of this class of problems. Learning t from negative examples or learning $\tilde{t}$ from positive examples will be discussed first. An algorithm using both positive and negative examples will then be presented to produce the best approximation possible. Using either positive or negative examples alone will only approach this best solution under special conditions. As for scenarios in this case, it is not useful to analyze a worst-case scenario if both positive and negative examples are used in learning. It is obvious that a worst-case error of $\overline{\kappa}_{\max} = 0.5$ is achieved for a target and distribution in which $p_i^+ = p_i^- = 0.5\,, \forall i$. For this reason we will just analyze the semi-uniform distribution scenario and bound all errors by 0.5.

4.2.1 Positive Or Negative Examples Only

To illustrate learning with positive or negative examples alone, we use the same problem developed for perfect learning. That is, the target will be such that all

subset probabilities are non-zero and we will learn t from negative examples alone. For this approximate learning scenario, however, we let $0 < p_i^+ \leq p_i^-$, $\forall i$, to insure an upper bound to the error of 0.5 for comparison purposes. In this fashion the positive examples can be viewed as noise or a poor representation of the null set in the perfect learning framework. While only negative examples will be used for learning, we will measure performance error using both positive and negative examples.

To solve this problem we can use the same consistent learning algorithm discussed before. That is, we start with $h = X \approx t$ and each subset is labelled t until a negative (counterexample) is chosen, at which time the subset is labelled $\tilde{t}$. The error set under these conditions is composed of two subsets. The first subset consists of all $x^+ \in b_i$, $\forall b_i$ that *have* counterexamples. The second subset consists of all $x^+ \in b_i$, $\forall b_i$ with *no* counterexamples. Because of the complementary nature of the b_i with and without counterexamples, the results of interest can be written in terms of the random set u, the set with no counterexamples.

In this case, the probability of any $x \in b_i$ is a member of the random error set r after m selections from $p(x^-)$ is composed of the probabilities of x_i^- being a member of u

$$\xi(x_i^- \in r) = \xi(x_i^- \in u) = \left(1 - p_i^-\right)^m \quad , \tag{39}$$

and x_i^+ not being a member of u

$$\xi(x_i^+ \in r) = \xi(x_i^+ \notin u) = 1 - \left(1 - p_i^-\right)^m \quad . \tag{40}$$

Applying Theorem I to this problem we get

$$\begin{aligned} \overline{\kappa} &= \sum_{i=1}^{s} \sum_{x \in b_i} p(x)\xi(x \in r) \\ &= \sum_{i=1}^{s} \left[\sum_{x^- \in b_i} p(x^-)\xi(x^- \in r) + \sum_{x^+ \in b_i} p(x^+)\xi(x^+ \in r) \right] \quad , \end{aligned} \tag{41}$$

and the final result is, therefore,

$$\overline{\kappa} = \sum_{i=1}^{s} \left[\, p_i^- \left(1 - p_i^-\right)^m + p_i^+ \left[1 - \left(1 - p_i^-\right)^m\right] \, \right] \quad . \tag{42}$$

This expression is similar to the one given for perfect learning in Eq. (21), except for the additional error term arising from the approximation inherent in the subsets already having counterexamples. That is, even after seeing a counterexample from a subset in this approximate learning case, errors will still be made in the fraction of the samples in the subset labelled t .

We now specialize the result for a semi-uniform distribution of positive and negative examples (only the negative ones being used for learning) of the form

$$p_i^- = \frac{\alpha}{s} \quad \text{and} \quad p_i^+ = \frac{1-\alpha}{s} \quad , \quad \forall i \quad , \tag{43}$$

with $\alpha > 0.5$.

The average error for this case is then given by

$$\overline{\kappa} = 1 - \alpha + (2\alpha - 1)\left(1 - \frac{\alpha}{s}\right)^m \quad . \tag{44}$$

The asymptotic value for this average error is thus $1 - \alpha$, the fraction of the total number of examples that are positive. The initial ($m = 0$) average error is clearly α. For $\alpha = 1$, we get the perfect learning result given in Eq. (21), as expected.

If a prescribed error ϵ is in the range $1 - \alpha < \epsilon \leq \alpha$, then a useful sample complexity can be defined for this case. This result is

$$L_{\overline{m}} = \frac{\ln\left(\frac{2\alpha-1}{\epsilon+\alpha-1}\right)}{\ln\left(\frac{s}{s-\alpha}\right)} \approx \frac{s}{\alpha}\ln\left(\frac{2\alpha-1}{\epsilon+\alpha-1}\right) \quad . \tag{45}$$

To be comparable to the perfect learning results given in Eq. (27), it is instructive to rework this sample complexity in terms of probabilistic convergence of the average error to its asymptotic value. This special case is useful if decisions need to be made about the rate at which an approximate algorithm learns. Thus, if we subtract the asymptotic value from the average error and define this shifted error as

$$\overline{\kappa}_a \equiv \overline{\kappa} - (1 - \alpha) = (2\alpha - 1)\left(1 - \frac{\alpha}{s}\right)^m \quad , \tag{46}$$

we then can find the sample complexity for which $\overline{\kappa}_a = \epsilon$ as

$$L_{\overline{m}_a} \approx \frac{s}{\alpha}\left[\ln\frac{1}{\epsilon} + \ln(2\alpha - 1)\right] \quad , \tag{47}$$

which is now comparable to Eq. (27).

4.2.2 Positive and negative examples together

The most general approximate learning problem we will discuss is the case where both positive and negative examples are used in learning. The essential aspects of this problem are captured in an algorithm that learns the best classification label for each subset of the partition structure separately. We will again assume, for the sake of comparison, that all subset probabilities are nonzero and that $p_i^- > p_i^+$, $\forall i$. Clearly, the best that can be done empirically in any subset, is to select as the hypothesis, thelabel that has the highest empirical number of examples.

Given that the partition results for this empirical problem are

	b_1	b_2	b_3	$\cdots$	$\cdots$	b_s
t	n_1^+	n_2^+	n_3^+	$\cdots$	$\cdots$	n_s^+
$\tilde{t}$	n_1^-	n_2^-	n_3^-	$\cdots$	$\cdots$	n_s^-

Fig. 5. Approximate learning empirical partition results.

an optimal algorithm for each subset b_i of the partition is

$$h_i = \begin{cases} t & \text{if} \quad n_i^+ \geq n_i^- \quad ; \\ \tilde{t} & \text{if} \quad n_i^- > n_i^+ \quad . \end{cases} \tag{48}$$

Here, n_i^+ and n_i^- are the observed numbers of positive and negative examples recorded in the ith-subset b_i, respectively, and h_i is the hypothesis label to be given to that subset. In this case, to be consistent with the previous illustrative cases presented, we have explicitly defaulted to a t label if the observed numbers of examples in any b_i are equal.

In the asymptotic limit, the empirical frequencies of examples in each subset will approach the probabilities p_i^+ and p_i^-. The worst error in performance, for the stated condition that $p_i^- > p_i^+$, will then be given by

$$\overline{\kappa}_{\max} = \sum_{i=1}^{s} p_i^- \quad . \tag{49}$$

Since the empirical results in each subset are probabilistic, Theorem I can again be used in this case to determine the average error on a subset-by-subset basis. The error set r_i for each b_i is again composed of two subsets r_i^- and r_i^+. The first subset r_i^- is the set of all $x^- \in b_i$, if a t label is chosen for b_i. This error set is realized when $n_i^+ \geq n_i^-$. The second subset r_i^+ is the set of all $x^+ \in b_i$, if a $\tilde{t}$ label is chosen for b_i. This error set is realized when $n_i^- > n_i^+$.

The complementary nature of the two components of the error subset r_i in each b_i, allows us to write the resulting overall error in terms of only the conditions which give rise to r_i^-. This result is

$$\begin{aligned}\overline{\kappa} &= \sum_{i=1}^{s}\left[\sum_{x^-\in b_i} p(x^-)\xi(x^-\in r) + \sum_{x^+\in b_i} p(x^+)\xi(x^+\in r)\right] \\ &= \sum_{i=1}^{s}\left[p_i^- G_i + p_i^+(1-G_i)\right] \quad ,\end{aligned} \tag{50}$$

where G_i is the probability of an example b_i being labelled t thus realizing random error set r_i^-. That is, in any m-sample

$$G_i \equiv \xi(n_i^+ \geq n_i^-) \quad . \tag{51}$$

For the finite-space problems we are dealing with, the function G_i can be evaluated analytically from the multinomial distribution describing the probabilistic m-sample outcomes. This evaluation can be broken up into two components: 1) the probability of picking n_i examples in b_i and 2) the probability that, given n_i examples in b_i, we then get $n_i^+ \geq n_i^-$. The result is that G_i can be written as

$$G_i = \sum_{j=0}^{m}\binom{m}{j}(p_i)^j(1-p_i)^{m-j}\sum_{k=j/2}^{j}\binom{j}{k}\left(p_i^+\right)^k\left(p_i-p_i^+\right)^{j-k} \quad . \tag{52}$$

The simplest case that can be analyzed in closed form in this framework is again the semi-uniform distribution scenario given in Eq. (19). Under such conditions, the worst-average error is given by

$$\overline{\kappa} = 1-\alpha+(2\alpha-1)\,G_i \quad , \tag{53}$$

where

$$G_i = \left(\frac{s-1}{s}\right)^m \sum_{j=0}^{m}\binom{m}{j}\left(\frac{\alpha}{s(s-1)}\right)^j \sum_{k=j/2}^{j}\binom{j}{k}\left(\frac{1-\alpha}{\alpha}\right)^k \quad . \tag{54}$$

Any further analysis of this class of problems requires numerical methods.

5. CONCLUSIONS

In summary, we were able to develop improved analytical tools for studying pac-learnability using a random set formalism. The major advantage of this approach was found to be the availability of two powerful random set theorems for analyzing performance errors and sample complexity in a algorithm-dependent, distribution-specific framework. To use these theorems we first had to redefine learnability in terms of moments of the measures of the error set, a departure from the error and confidence approach first proposed by Valiant. The moment approach allowed analytic expressions to be developed for performance errors in several important problem classes. It also provided tools for studying learning in a problem-specific environment as opposed to Valiant's worst-case, distribution-free approach.

Several illustrative examples were worked out in detail. These included perfect and approximate learning problems in a disjoint-partition concept space. For most of the cases studied, analytic expressions were presented for the average error in learning performance and an associated sample complexity. In one case the second moment of the error distribution was presented in analytic form as well. The expressions derived were algorithm-dependent and distribution-specific, in that the results were explicitly dependent on the algorithm analyzed and the probability distribution of learning examples.

The major conclusions that can be drawn from this work are that the random set approach makes pac-learnability analysis more amenable to practical problem solving. The problem-specific nature of the results show that for many problems, worst-case analyses can greatly overestimate the number of random examples needed to meet a given performance criterion. The average-error expressions developed using this approach are also more amenable to experimental verification while learning is taking place. That is, analytic average-error expressions can be directly verified by a learning curve estimated incrementally while learning.

REFERENCES

1. Benedict, G.M. & Itai, A. (1988). Learnability by fixed distributions. *Proceedings of the 1988 Workshop on Computational Learning Theory.* Boston, MA: Morgan Kaufmann.

2. Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M.K. (1989). Learnability and the Vapnik-Chervonenkis dimension *JACM, 6,* 929-965.

3. Haussler, D. (1988a). Quantifying inductive bias. *Artificial Intelligence, 36,* 177-221.

4. Haussler, D. (1988b). Generalizing the PAC model: Sample size bounds from metric dimension-based uniform convergence results. *Proceedings of the 1987 Symposium on Foundations of Computer Science.* Boston, MA: Morgan Kaufmann.

5. Haussler, D. (1990). *Probably approximately correct learning.* Technical Report UCSC-CRL-90-16, University of California, Santa Cruz CA.

6. Haussler, D. & Pitt, L. (Eds.) (1988). *Proceedings of the 1988 Workshop on Computational Learning Theory.* Boston, MA: Morgan Kaufmann.

7. Kendall, D.G. (1977). Foundations of a theory of random sets. In Harding, E.F. & Kendall, D.G. (Eds.), *Stochastic Geometry.* New York: John Wiley & Sons.

8. Goodman, I.R. & Nguyen H.T. (1985), *Uncertainty Models For Knowledge-Based Systems.* New York: North-Holland.

9. Oblow, E.M. (1990). *Implementing Valiant's learnability theory using random sets.* Technical Report ORNL/TM-11512R, Oak Ridge National Laboratory, Oak Ridge TN.

10. Oblow, E.M. (1991). Implementing Valiant's learnability theory using random sets. *Machine Learning, 6,* in press.

11. Rivest, R., Haussler, D., & Warmuth, M.K. (Eds.) (1989). *Proceedings of the 1989 Workshop on Computational Learning Theory.* Santa Cruz, CA: Morgan Kaufmann.

12. Robbins, H.E. (1944). On the measure of a random set. *Ann. Math. Stat., 15* 70-75.

13. Traub, J.F., Wasilkowski, G.W., & Wozniakowski, H. (1988). *Information-Based Complexity.* Boston, MA: Academic Press.

14. Valiant, L.G. (1984). A theory of the learnable. *Comm. ACM, 27,* 1134-1142.

15. Valiant, L.G. (1985). Learning disjunctions and conjunctions. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (pp. 560-566). Los Angeles, CA: Morgan Kaufmann.

16. Vapnik, V. (1982). *Estimation of Dependencies Based on Empirical Data.* New York, NY: Springer-Verlag.

INTERNAL DISTRIBUTION

1. B. R. Appleton
2. M. Beckerman
3. K. Fujimura
4. C. W. Glover
5. J. P. Jones
6–10. H. E. Knee
11. G. E. Liepins
12–16. R. C. Ward
17–21. R. C. Mann
22–26. E. M. Oblow
27. F. G. Pin
28. V. Protopopescu
29. D. B. Reister
30. J. C. Schryver
31. P. Spelt
32. R. W. Brockett (Consultant)
33. J. J. Dorning (Consultant)
34. J. E. Leiss (Consultant)
35. N. Moray (Consultant)
36. M. F. Wheeler (Consultant)
37. EPMD Reports Office
38–39. Laboratory Records Department
40. Laboratory Records, ORNL-RC
41. Document Reference Section
42. Central Research Library
43. ORNL Patent Section

EXTERNAL DISTRIBUTION

44. Dana Angluin, Computer Science Department, Yale University, P. O. Box 2158, New Haven, Conecticut 06520
45. Peter Allen, Department of Computer Science, 450 Computer Science, Columbia University, New York, N.Y.
46. J. A. Barhen, Jet Propulsion Laboratory, MS 198/330, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California 91109-8099
47. Anselm Blumer, Department of Mathematics & Computer Science, Tufts University, Medford, Massachusetts 02155
48. Wayne Book, Department of Mechanical Engineering, J S Coon Bldg, Room 306, Georgia Institute of Technology, Atlanta, GA 30332
49. Peter Cheeseman, NASA Ames Research Center, Mail Stop 244-17, Moffett Field, California 94035
50. Jie Cheng, Department of Electrical Engineering & Computer Science, The University of Michigan, Ann Arbor, Michigan 48109-2122
51. Steve Dubowsky, MIT, Bldg 3, Rm 469A, 77 Massachusetts Ave, Cambridge, MA 02139
52. Andrzej Ehrenfeucht, University of Colorado, Boulder, Colorado 80309
53. Irwin. R. Goodman, Naval Ocean Systems Center, San Diego, CA 92152
54. John J. Grefenstette, Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, D. C. 20375-5000
55. David Haussler, Department of Computer & Information Sciences, University of California, Santa Cruz, California 95064
56. Haym Hirsh, Computer Science Department, Stanford University, Stanford, California 94305
57. Avi Kak, Department of Electrical Engineering, Engineering Mall, Purdue University, Lafeyette, IN 47907

58. Michael Kearns, Aiken Computation Laboratory, Harvard University, Cambridge, Massachusetts 02138
59. M. H. Kalos, Courant Institute of Math Science, New York University, 251 Mercer Street, New York, NY 10012
60. Philip D. Laird, NASA Ames Research Center, Moffett Field, California 94035
61. Pat Langley, Department of Information & Computer Science, University of California, Irvine, California 92717
62. Nick Littlestone, Department of Computer & Information Sciences, University of California, Santa Cruz, California 95064
63. Oscar P. Manley, U.S. Department of Energy, Division of Engineering, Mathematical and Geosciences, Office of Basic Energy Sciences, Germantown, MD 20545
64. Alan Meyrowitz, Office of Naval Research, Code 1133, 800 N. Quincy St., Arlington, VA 22217
65. B. K. Natarajan, The Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
66. Judea Pearl, Computer Science Department, University of California, 405 Hilgard Ave., Los Angeles, CA 90024
67. J. R. Quinlan, Basser Department of Computer Science, University of Sydney, Sydney, New South Wales 2006, AUSTRALIA
68. N. S. V. Rao, Department of Computer Science, Old Dominion University, Norfolk, VA 23529-0162
69. Zbigniew W. Ras, Department of Computer Science, University of North Carolina, Charlotte, North Carolina 28223
70. Ronald L. Rivest, MIT Laboratory for Computer Science, Cambridge, Massachusetts 02139
71. George Shackelford, Department of Information & Computer Science, University of California, Irvine, California 92717
72. Robert Sloan, MIT Laboratory for Computer Science, Cambridge, Massachusetts 02139
73. Wes Snyder, Department of Radiology, Bowman Gray School of Medicine, 300 South Hawthorne Drive, Winston Salem, NC 27103
74. Richard S. Sutton, GTE Laboratories, Inc., 40 Sylvan Road, Waltham, Massachusetts 02254
75. V. R. R. Uppuluri, 130 Indian Lane, Oak Ridge, TN 37830
76. Paul E. Utgoff, Department of Computer & Information Science, University of Massachusetts, Amherst, Massachusetts 01003
77. L. G. Valiant, Aiken Computation Laboratory, Harvard University, Cambridge, Massachusetts 02138
78. Jeffrey S. Vitter, Department of Computer Science, Brown University, Providence, Rhode Island 02912
79. Manfred K. Warmuth, Department of Computer & Information Sciences, University of California, Santa Cruz, California 95064
80. C. R. Weisbin, Jet Propulsion Laboratory, MS 198/330, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California 91109-8099
81. Lotfi A. Zadeh, Computer Science Division, University of California, Berkeley, CA 94720
82. Maria Zemankova, Department of Computer Science, University of Tennessee, Knoxville, Tennessee 37996-1301

83. Office of Assistant Manager, Energy Research and Development, DOE-ORO, P.O. Box 2001, Oak Ridge, TN 37831

84–93. Office of Scientific and Technical Information, Department of Energy, Oak Ridge, TN 37831