MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES

3 4456 0310635 0

ORNL/TM-11904
(CESAR-91/23)

ornl

OAK RIDGE
NATIONAL
LABORATORY

MARTIN MARIETTA

N-Learners Problem: Fusion of Concepts

Nageswara S. V. Rao
E. M. Oblow
Charles W. Glover
Gunar E. Liepins

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 has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from the Office of Scientific and Technical Information, P.O. Box 62, Oak Ridge, TN 37831; prices available from (615) 576-8401, FTS 626-8401.

Available to the public from the National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161.

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ORNL/TM-11904
CESAR-91/23

Engineering Physics and Mathematics Division

N-LEARNERS PROBLEM: FUSION OF CONCEPTS

Nageswara S.V. Rao
Department of Computer Science
Old Dominion University
Norfolk, VA 23529-0162

E. M. Oblow
Charles W. Glover
Gunar E. Liepins

DATE PUBLISHED — September 1991

Research sponsored by the
Office of Engineering Research Program
Basic Energy Sciences
U.S. Department of Energy
and
Office of Naval Research
Intelligent Systems Program
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 0310635 0

CONTENTS

ACKNOWLEDGMENTS v
ABSTRACT vii
1. INTRODUCTION 1
2. N-LEARNERS PROBLEM 3
3. SIMPLE FUSERS 5
3.1 ONE-SIDED LEARNERS 5
3.2 COMBINATION OF LEARNERS 9
4. COMPARISON OF LEARNERS 10
5. FUSION AS LEARNING 15
6. FUSION BY LINEAR THRESHOLD FUNCTIONS 22
7. CONCLUSIONS 27
APPENDIX A 28
REFERENCES 29

LIST OF TABLES

Table		Page
1	Leaning Boolean formulae.	19
2	Concept learning.	20

ACKNOWLEDGMENTS

The authors gratefully acknowledge the continuing financial support of this learning research by Oscar Manley of the Basic Energy Sciences Program in the Department of Energy and Alan Meyrowitz in the Intelligent Systems Program of the Office of Naval Research in the Department of Defense. In addition, N.S.V. Rao is partially funded by National Science Foundation under grant #IRI-9108610, Old Dominion University Summer Faculty Fellowship for 1991 and Virginia's Center for Innovative Technology under contract # INF-90-015. The authors would also like to thank Larry Wilson and Rao Chaganti for comments on earlier drafts of this paper.

ABSTRACT

We are given N learners each capable of learning concepts (subsets) of a domain set X in the sense of Valiant, i.e. for any $c \in C \subseteq 2^X$, given a finite set of examples of the form $< x_1, M_c(x_1) >; < x_2, M_c(x_2) >; \ldots; < x_l, M_c(x_l) >$ generated according to an unknown probability distribution P_X on X, each learner produces a close approximation to c with a high probability. We are interested in combining the N learners using a single *fuser* or *consolidator*. We consider the paradigm of *passive fusion*, where each learner is first trained with the sample without the influence of the consolidator. The *composite system* is constituted by the fuser and the individual learners. We consider two cases: open and closed fusion. In *open fusion* the fuser is given the sample and the hypotheses of the individual learners; we show that the fusion rule can be obtained by formulating this problem as another learning problem. For the case all individual learners are trained with the same sample, we show sufficiency conditions that ensure the composite system to be better than the best of the individual: the hypothesis space of the consolidator (a) satisfies the *isolation property* of degree at least N, and (b) has Vapnik-Chervonenkis dimension less than or equal to that of every individual learner. If individual learners are trained by independently generated samples, we obtain a much weaker bound on the VC-dimension of the hypothesis space of the fuser. Second, in *closed fusion* the fuser does not have an access to either the training sample or the hypotheses of the individual learners. By suitably designing a linear threshold function of the outputs of individual learners, we show that the composite system can be made better than the best of the learners.

Keywords and Phrases: N-learners problem, computational learnability, passive fusion, open fusion, closed fusion.

1. INTRODUCTION

In several practical applications, we are faced with the problem of combining information from several sources. Such examples abound among humans and machines alike. For example, a judge is required to examine the evidences from attorneys and deliver a judgement — in this case the information sources are "competitive". In the example of a robot equipped with a number of sensors, the problem is to combine the sensory information and form a description of the environment — in this case the information could be (a) "cooperative", e.g. two sensors indicating an obstacle in the way; (b) "competitive", e.g. a faulty sensor giving a reading different from a non-faulty one; or (c) "complementary", e.g. one sensor giving the shape and another giving the color.

The *N-Learners Problem* is a special (abstracted) case of data fusion: we are given multiple learners that infer concepts, and the problem is to design a *consolidator* or *fuser* that combines the outputs of the individual learners. In the paradigm of *passive fusion* the individual learners are not supervised by the fuser. In *active fusion* the individual learners are controlled by the fuser. Note that a passive fuser is a special type of active fuser that chooses not to control the learning process of the individual learners. In this paper, we discuss only passive fusion when the individual learners are as described under the framework of Valiant [28]. The learning problems (i.e., design of individual learners) under this framework have been extensively studied over the past five to six years; a small selection of results is presented in Section 4.

A variant of the N-learners problem has been first discussed in [25] in the context of sensor fusion in a hybrid system. Potential applications of the N-learners problem include sensor fusion [11,16], hybrid systems [13,25], information pooling and group decision models [14,20], and majority systems [8].

Consider a system of N learners $L_1, L_2, ..., L_N$, where each L_i learns concepts (subsets) of a domain X in the sense of Valiant [28], i.e., given a sufficiently large sample of examples of $c \in C \subseteq 2^X$, a hypothesis h close to c will be produced with a high probability. The closeness of the hypothesis (learned concept) h to c is specified by a *precision* parameter ϵ, and the probability that this closeness is achieved is specified by a *confidence* parameter $1 - \delta$. Given two learners trained by the same number of examples, the one with higher or equal confidence for the same value of precision is considered *better* (this notion is more precisely defined in Section 4). In this paper, we only consider the problem of designing a *fuser* (or *consolidator*) such that the *composite system*, of the fuser with the N learners, can be made better than the best of the learners.

There are other interesting criteria for designing a fuser. For example, we might be interested in making the composite system learn concepts that are not learnable by the individual learners. In [26] a system capable of learning Boolean combinations of halfspaces by utilizing a system of perceptrons is described; note that a single perceptron is incapable of learning such concepts [23].

We first illustrate some simple cases where the composite system can be easily seen to be better than each of the learners (Section 3), and then consider more general cases.

2 *Introduction*

We consider two paradigms:

(a) **Open Fusion:** In open fusion, the fuser is given the training examples and the hypotheses of the individual learners. We introduce a property called the *isolation*, and present sufficiency conditions that ensure the composite system to be better than the best of the learners. We show that the problem of designing the fuser can be solved by casting it as another learning problem that can be solved using known methods if the suitable isolation property is satisfied. We consider the two cases:

(i) all learners are trained with the same sample, and

(ii) each learner is individually trained with a separate random sample.

We derive sufficiency conditions for several formulations of the learnability problem such that the composite system has higher confidence than the best of the learners. In both cases, the hypothesis class of the fuser must satisfy the isolation property of degree N, where N is the number of individual learners; additionally, the condition in the first case is that the Vapnik and Chervonenkis dimension [7] (VC dimension) of the fuser be smaller than or equal to that of every learner. And in the second case the fuser can have much larger VC dimension (the exact bound is specific to the formulation of the learning mechanism of L_i's). In other formulations that do not use VC dimension (e.g., learnability under fixed distributions [6], learning under metric spaces [15]), we use the corresponding parameters to express sufficiency conditions.

(b) **Closed Fusion:** In closed fusion, the fuser does not have access to either the examples or the hypotheses of the individual learners. We show that a linear threshold fuser can be designed such that the composite system is better than the best of the learners. This result shows that that even if all individual learners are completely consistent with the sample (i.e. all of them have zero empirical error), we can still make the performance of the composite system better than that of any individual learner.

The organization of this paper is as follows: A precise formulation of the present version of the N-learners problem is presented in Section 2. Specialized examples where a suitable fusion rule makes the overall system better than the best of the learners are given in Section 3. A selection of existing learning formulations, and an approach to compare the learners are outlined in Section 4. The general problem is solved in Section 5 for the case of open fusion. Closed fusion using linear threshold functions (of the outputs of the learners) is addressed in Section 6.

2. N-LEARNERS PROBLEM

A *concept* is a subset of a domain set X; for a concept $c \subseteq X$ we define a membership function $M_c : X \rightarrow \{0,1\}$ such that for $x \in X$: $M_c(x) = 1$ if $x \in c$ and $M_c(x) = 0$ if $x \notin c$ (for ease of presentation we abuse the notation by interchangeably using c and M_c when the reference is clear from the context). A set of concepts is called the *concept class*. An *example* of a concept $c \subseteq X$ is any pair $< x, M_c(x) >$ for $x \in X$, and an *l-sample* of c is a sequence of l examples of c. An example $< x, M_c(x) >$ is called *positive* if $M_c(x) = 1$ and *negative* if $M_c(x) = 0$. We assume that an example is randomly produced according to a distribution P_X on X, and the examples of a sample are produced independently. Each learner for a concept class $C \subseteq 2^X$ has a *hypothesis class* $H \subseteq 2^X$. We mainly consider the cases where P_X is unknown; we consider one case where P_X is known.

The concept class C is *learnable* by a hypothesis class H if for every P_X, for any concept $c \in C$, there exists $l < \infty$ such that: given an l-sample of c (i.e., $< x_1 M_c(x_1) >, < x_2, M_c(x_2) >, \ldots, < x_l, M_c(x_l) >$) and ϵ and δ, ($0 < \epsilon, \delta < 1$) an *approximation* $h \in H$ can be produced such that

$$Prob[\mu(c\Delta h) > \epsilon] < \delta \tag{2.1}$$

where $c\Delta h = (c-h)\cup(h-c)$ and $\mu(c\Delta h) = \int\limits_{M_c(x)\neq M_h(x)} dP_X(x)$, i.e. the integration is over all x that precisely belong to one of c and h but not to both [7,28]. Note that $\mu(c\Delta h)$ is the probability that a randomly chosen example (with respect to P_X) will yield different values under M_c and M_h. Here c is often called the *target concept* and h is called the *hypothesis* of the learner. Informally, the equation (2.1) means that with an arbitrarily specified confidence $1 - \delta$, we must be able to produce a hypothesis that approximates the target concept within ϵ which is also arbitrarily specified.

The above formulation is popularly referred to as the *Probably Approximately Correct* (PAC) learning [7,28]. Several variants of this basic problem have been studied by a number of researchers (see the references). We refer to condition in Eq (2.1) as the (ϵ, δ)-*condition*. This condition is also often expressed as

$$Prob[\mu(c\Delta h) \leq \epsilon] \geq 1 - \delta. \tag{2.2}$$

We say that C is *polynomially learnable* by H if the number of examples, often called the *sample size*, needed to ensure the (ϵ, δ)-condition is a polynomial in $1/\epsilon$, $1/\delta$ and some appropriate parameters of H such as VC dimension, cardinality, etc. [7,28]. In some formulations, the sample size has to be polynomial in the complexity of the target complexity (e.g. [2,22]). Also, in some cases only a single parameter is used to express the (ϵ, δ)-condition; for example, $h = \frac{1}{\epsilon} = \frac{1}{\delta}$ is used in [22,28].

Now consider N *learners* such that each learner has the same concept class $C \subseteq 2^X$ and the same output space of $\{0,1\}$. The ith learner L_i has a hypothesis space of $H_i \subseteq 2^X$. Each learner has its own way of producing a hypothesis. In other words, the individual learners can differ in their hypothesis classes, and/or in the methods used to produce the hypotheses. The *N-learners problem* deals with designing a *fuser* or *consolidator* that learns a map from the outputs of the N learners to $\{0,1\}$.

4 *N-Learners Problem*

Let H_F denote the hypothesis space of the fuser. One example for the hypothesis class of fuser is a set of Boolean combinations of at most N variables. We consider more general hypothesis classes by embedding the N-dimensional Boolean cube into $\Re^N$, and letting H_F to be a subset of $2^{\Re^N}$.

Our main objective is to make the composite system of the fuser and the learners better than the best of the individual learners. The notion of "better" is formalized in Section 4. The problem of designing the fuser critically depends on the information available to the fuser such as the type of learners, the examples given to the individual learners, etc. We are intersected in identifying the cases in which a fuser can be designed to outperform the individual learners.

3. SIMPLE FUSERS

To illustrate the basic ideas of passive fusion, we consider a very simple set of N learners; we show that in these cases a fuser can be very easily designed. Our examples consist of learners, called *learners with one sided error*, that are allowed to make mistakes on only positive examples or only on negative examples of a concept.

3.1 ONE-SIDED LEARNERS

Consider a set of learners such that each H_i makes only *one-sided errors* as in Valiant [28], i.e., corresponding to a target concept f, L_i learns a hypothesis f_i such that $f \subseteq f_i$; we denote this by $f \Rightarrow f_i$. In other words, L_i is not allowed to misclassify members of f, but can misclassify non-members of f. Valiant [29] discusses learning algorithms for such learners for several cases of Boolean formulae (see also Natarajan [22] for some additional work).

Now consider that each L_i has been trained with a sample that ensures (ϵ, δ)-condition, and $x \in X$ is given to be classified. We define a fuser F in this case to yield a value of 1 if and only if all L_is yield 1s, i.e., the hypothesis of the consolidator is $h = \bigcap_i f_i$. Clearly, we have $f \Rightarrow \bigcap_i f_i$. Properties of the fuser are given in the following theorem.

Theorem 1. *For the system of N statistically independent learners with one-sided errors such that $f \Rightarrow f_i$ for each learner L_i, let the hypothesis of the fuser be $h = \bigcap_i f_i$. Then we have:*

(i) With the probability at least $1 - \delta^N$ we have $\mu(f \Delta h) \leq \epsilon$.

(ii) With the probability at least $(1-\delta)^N$, we have

$$\mu(f \Delta h) \leq \epsilon - \max_{ij}\{\rho_{ij}\}$$

where $\rho_{ij} = \mu(f_i \Delta (f_i \cap f_j))$.

(iii) With the probability at least $(1-\delta)^N$, we have for any $f_0 \in \{f_1, f_2, \ldots, f_N\}$:

$$\mu(f \Delta h) \leq \epsilon - \sum_i \rho_i + \sum_{i,j} \rho_{i,j} + \cdots + (-1)^N \rho_{i_1, i_2, \ldots, i_N}$$

where $\rho_i = \mu(f \Delta (f_0 \cap f_i))$, $\rho_{i,j} = \mu(f \Delta (f_0 \cap f_i \cap f_j))$, for distinct $i, j \in \{1, 2, \ldots, N\}$ and

$$\rho_{i_1, i_2, \ldots, i_j} = \mu(f \Delta (f_0 \cap f_{i_1} \cap \ldots \cap f_{i_j}))$$

for distinct $i_1, i_2, \ldots i_j \in \{1, 2, \ldots, N\}$ and $1 \leq j \leq N$.

(iv) Let $\mathcal{F} = \{f_1, f_2, \ldots, f_N\}$ and $\mathcal{F}_k = \mathcal{F} - \{f_k\}$. For any f_k, let

$$\lambda_{f_k} = Prob\left[f_k \bigcap \left(\bigcup_{f \in \mathcal{F}_k} f\right)\right].$$

Then we have

$$\lambda_{f_k} = -\sum_{f_i \in \mathcal{F}_k} Prob[f_k \cap f_i] + \sum_{f_i, f_j \in \mathcal{F}_k} Prob[f_k \cap f_i \cap f_j] + \cdots$$
$$+ (-1)^{N-1} Prob\left[f_k \bigcap \left(\bigcap_{f_i \in \mathcal{F}_k} f_i\right)\right].$$

We also have

$$\mu(f \Delta h) \leq \epsilon - \max_{f_k \in \mathcal{F}} \{Prob[f_k] - \lambda_{f_k}\}.$$

Proof: Since each L_i is a Valiant's learner, we have $\mu(f \Delta f_i) \geq \epsilon$ with a probability of at most δ. Thus with a probability of at least $1 - \delta^N$, there exists L_i with $\mu(f \Delta f_i) \leq \epsilon$. Since $h \Rightarrow f_i$ and $f \Rightarrow h$, every x in h but not in f, will definitely be in f_i. Thus every x that contributes a non-zero value to $\mu(f \Delta h)$ will also contribute the same value to $\mu(f \Delta f_i)$. Hence we have (i). For the second part, with a probability of $(1 - \delta)^N$, we have $\mu(f \Delta f_i \leq \epsilon)$ for each L_i. By taking AND of f_i and f_j we reduce ϵ by ρ_{ij}.

A tighter bound on the total amount of reduction in ϵ can be estimated by using the the Inclusion-Exclusion Principle [12,18]; this bound is given in (iii) and (iv) parts. Consider a set Y. For any $A_1, A_2, \ldots, A_N \subseteq Y$, and a function $\phi : 2^Y \mapsto [0, M]$ for $M < \infty$, such that $\phi(A \cup B) = \phi(A) + \phi(B)$ for any two disjoint subsets $A, B \subseteq Y$, we have

$$\phi(\bar{A}_1 \cap \bar{A}_2 \cap \ldots \cap \bar{A}_r) = \phi(Y) - \sum_i \phi(A_i) + \sum_{i,j} \phi(A_i \cap A_j) + \ldots +$$
$$(-1)^r \phi(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_r}),$$

where $\bar{A}_i = Y - A_i$.

We identify ϕ with the probability measure [12]. Let $Y = f_0 - f$, thus $Prob[Y] = \mu(f_0 \Delta f) \leq \epsilon$; and let $A_{i_1, i_2, \ldots, i_j} = (f_0 \cap f_1 \cap \ldots f_j) - f$, thus $\rho_{i_1, i_2, \ldots, i_j} = \mu(f_0 \cap f_{i_1} \cap \ldots f_{i_j}) \Delta f)$. Thus the part (iii) directly follows. Part (iv) follows along the lines of (iii) by simple algebraic manipulation. QED.

Now consider the implications of this theorem. First, we are able to ensure that the composite system has a higher confidence factor of $1 - \delta^N$ of bounding the precision by ϵ; also, the ratio of the corresponding confidence factors increases with N since $\frac{(1-\delta^N)}{1-\delta} = 1 + \delta + \delta^2 + \ldots + \delta^{N-1}$. Second, with lesser confidence, we can reduce the ϵ by $\rho_{i,j}$. Note that $\rho_{i,j} \cup \rho_{j,i} = \mu(f_i \Delta f_j)$; thus the reduction in ϵ is proportional to the biggest amount of "dissimilarity" between two hypotheses of the individual learners. An expression for more precise reduction in ϵ is given in (iii) and (iv).

Example 3.1. *Finite Hypotheses Classes:*

Consider the case of finite number of hypotheses for each learner, i.e. $|H_i| < \infty$. Assume that $C \subseteq H_i$ for $i = 1, 2, \ldots, N$. Then the number of examples needed to ensure learnability is given by $m = \frac{1}{\epsilon} \ln\left(\frac{|H_i|}{\delta}\right)$ as in Natarajan [22] (also see Blumer et al [7]). A learning algorithm for L_i in this case simply produces a hypothesis that is consistent with the sample, i.e. any $h \in H_i$ that contains all positive examples and does not contain any negative examples is guaranteed to satisfy the (ϵ, δ)-condition. In our present case, we assume that the individual learners differ in the way they pick their hypotheses.

Now consider the condition (i) of Theorem 1. If a single learner of the above kind has to satisfy the (ϵ, δ^N) condition, then the number of examples required is given by $m_1 = \frac{1}{\epsilon} \ln\left(\frac{|H_i|}{\delta^N}\right)$. Thus the effect of composite syetem is as if the number of examples to a single learner has been increased by $\frac{1}{\epsilon} \ln\left(\frac{1}{\delta^{N-1}}\right)$.

Similar computations can also be carried out using the other parts of Theorem 1. By using part (ii) we have

$$m_1 = \frac{1}{\epsilon - \rho_{\max}} \ln\left(\frac{|H_i|}{(1-(1-\delta)^N}\right)$$

where $\rho_{\max} = \max_{ij}\{\rho_{ij}\}$.

Since

$$[1-(1-\delta)^N]^{\frac{1}{\epsilon-\rho_{\max}}} \leq [1-(1-\delta)^N]^{\frac{1}{\epsilon}}$$

the effective increase in the number of examples is at least

$$\frac{\rho_{\max}}{\epsilon(\epsilon - \rho_{\max})} \ln(|H_i|) - \frac{1}{\epsilon} \ln\left(\frac{1}{\delta} - \frac{(1-\delta)^N}{\delta}\right).$$

It is interesting to note that if $\epsilon = \rho_{\max}$, then the effective increment in the examples is ∞; this is because this condition is tantamount to recovering f completely. Also the more the number of learners, the more will the effective increment in the number of examples.

We now consider the counterpart case where each learner L_i satisfies the property: any $x \in X$ in f_i must be in f, i.e. L_i is allowed to misclassify elements of f, but is not allowed to misclassify non-members of f. i.e., L_i learns concept f_i such that $f_i \subseteq f$; we denote this condition by $f \Leftarrow f_i$. In this case, the hypothesis of the fuser is $h = \bigcup_i f_i$. We have $f \Leftarrow \bigcup_i f_i$.

Theorem 2. *For the system of N statistically independent learners with one-sided errors such that $f \Leftarrow f_i$ for each learner L_i, let the hypothesis of the fuser be $h = \bigcup_i f_i$. Then we have:*

(i) With probability at least $1 - \delta^N$ we have $\mu(f \Delta h) \leq \epsilon$,

(ii) With probability of at least $(1-\delta)^N$ *we have*

$$\mu(f\Delta h) \leq \epsilon - \max_{i,j}\{\rho_{i,j}\}$$

where $\rho_{i,j} = \mu(f_i\Delta(f_i \cap f_j))$.

(iii) With probability at least $1-\delta^N$ *we have*

$$\mu(f\Delta h) \leq N\epsilon - (N-1)\gamma_f + \sum_{i,j}\gamma_{i,j} + \ldots + (-1)^N\gamma_{i_1,i_2,\ldots,i_N}$$

where $\gamma_f = Prob[f]$, $\gamma_{i,j} = Prob[f_i \cap f_j]$, *for distinct* $i,j \in \{1,2,\ldots,N\}$ *and*

$$\gamma_{i_1,i_2,\ldots,i_j} = Prob[f_{i_1} \cap f_{i_2} \cap \ldots \cap f_{i_j}]$$

for distinct $i_1, i_2, \ldots i_j \in \{1,2,\ldots,N\}$ *and* $1 \leq j \leq N$.

(iv) With probability $(1-\delta)^N$, *for*

$$Prob[f] \geq \epsilon + \frac{\kappa}{N-1} + \frac{2^{N-1}-\frac{1}{2}}{N-1}(\epsilon_U - \epsilon_L) + \frac{N\epsilon_L}{N-1}$$

for some κ, *we have*

$$\mu(f\Delta h) < \epsilon - \kappa$$

where $\epsilon_U = \max_j \max_{i_1,i_2,\ldots,i_j} \rho_{i_1,i_2,\ldots,i_j}$ *and where* $\epsilon_L = \min_j \min_{i_1,i_2,\ldots,i_j} \rho_{i_1,i_2,\ldots,i_j}$

Proof: The proofs for parts (i) and (ii) are similar to those in Theorem 1. For (iii), we first note, for $f = Y$

$$\mu(f\Delta h) = Prob(\bar{f}_1 \cap \bar{f}_2 \cap \ldots \cap \bar{f}_N).$$

Then using the Inclusion-Exclusion principle, and $Prob(Y) = Prob(f)$, and $Prob(f_i) \geq \gamma_f - \epsilon$, we have (iii). Then

$$Prob(\bar{f}_1 \cap \bar{f}_2 \cap \ldots \cap \bar{f}_{r-1}) = Prob(f) - \sum_i \gamma_i + \sum_{i,j}\gamma_{i,j} + \ldots + (-1)^{r-1}\gamma_{i_1,i_2,\ldots,i_{r-1}}.$$

Let us denote

$$\lambda = \sum_{i,j}\gamma_{i,j} + \sum_{i,j,k}\gamma_{i,j,k} + \ldots + (-1)^N\gamma_{i_1,i_2,\ldots,i_N}.$$

Then the condition $\mu(f\Delta h) \leq \epsilon - \kappa$, where $\kappa \geq 0$, is implied by

$$Prob[f] \geq \epsilon + \frac{\lambda}{N-1} + \frac{\kappa}{N-1}$$

We now derive an upper bound on λ as follows:

$$\lambda \leq \binom{N}{2}\epsilon_U - \binom{N}{3}\epsilon_L + \ldots$$

Now we have

$$\binom{N}{1} + \binom{N}{2} + \ldots + \binom{n}{n} = 2^N - 1$$

and

$$\binom{N}{1} + \binom{N}{3} + \ldots = \binom{N}{2} + \binom{N}{4} + \ldots = 2^{N-1} - \frac{1}{2}.$$

Thus $\lambda \leq [2^{N-1} - \frac{1}{2}](\epsilon_U - \epsilon_L) + N\epsilon_L$. QED.

If $\epsilon_U = \epsilon_L$, then we can reduce the ϵ by $\partial\epsilon = (Prob[f] - \epsilon)(N-1) - \frac{N}{N-1}\epsilon_L$. The value of ϵ can be reduced by employing a suitable number of learners. In a general case, such reduction may not be achievable.

3.2 COMBINATION OF LEARNERS

Now we consider the case where we have N independent learners of the type in Theorem 1 and N independent learners of the type in Theorem 2. The hypotheses of the first type of leaners are denoted by $f_1, f_2, \ldots, f_N$, and those of the second type are denoted by $g_1, g_2, \ldots, g_N$. Let $\mathcal{F} = \{f_1, f_2, \ldots, f_N\}$ and $\mathcal{G} = \{g_1, g_2, \ldots, g_N\}$ Then we obtain a static fusion rule given by (for $x \in X$):

if $x \in \bigcup_{i=1}^{N} g_i$, **then** output 1

else if $x \notin \bigcap_{i=1}^{N} f_i$, **then** output 0

else flip a fair coin, and output 1 if heads or output 0 otherwise;

Consider that $N = 1$, then with probability $(1-\delta)^2$ we have $f\Delta h \leq \epsilon$. To see this, note that with probability $1-\delta$ both f_1 and g_1 will guarantee a precision of ϵ. Using the above algorithm, the region of error is $(f_1 - f) \cup (f - g_1)$ and the probability of error for any point in this region is $\frac{1}{2}$. Thus the total probability of error is $\frac{1}{2}[\mu(f_1\Delta f) + \mu(f\Delta g_1)] \leq \epsilon$. In the general case the following are direct consequences of the discussion of last subsections.

(i) With probability at least $(1-\delta^N)^2$ we have $\mu(f\Delta h) \leq \epsilon$,

(ii) With probability of at least $(1-\delta)^{2N}$ we have

$$\mu(f\Delta h) \leq \epsilon - \frac{1}{2}\left[\max_{i,j}\{\mu(f_i\Delta(f_i \cap f_j))\} + \max_{i,j}\{\mu(g_i\Delta(g_i \cap g_j))\}\right]$$

In view of (i) this algorithm fares worse than the best of the individual learners for the particular case when $N = 1$. In a general case the condition $(1-\delta^N)^2 \geq 1-\delta$ is implied by $1-\delta^N \geq 1 - \frac{\delta}{2}$ [1]. This condition in turn is satisfied if $N \geq 1 + \frac{\ln 2}{(1/\delta)}$; thus, we can make the system have confidence higher than $1-\delta$ by suitably choosing the number of learners. In particular if $\delta < 1/2$, i.e. each learner performs better than a fair coin, it suffices to ensure $N \geq 2$. Also note that this algorithm can exploit the diversity of members of both $\mathcal{F}$ and $\mathcal{G}$ as in (ii). More precise bounds on $\mu(f\Delta h)$ can also be worked out along the lines of last subsections.

4. COMPARISON OF LEARNERS

In judging the performance of various learners, it is often necessary to make comparisons between the learners. We characterize a learner by the *parameter pair* (ϵ, δ); recall that the learner L_i and the fuser F are characterized by the pairs (ϵ_i, δ_i) and (ϵ_F, δ_F) respectively. In general, for many formulations of the learnability problem, there is a functional relation between ϵ_i and δ_i of a learner as will be subsequently illustrated. Also, some authors use a single parameter to characterize the learners (e. g. [22,28]). We compute the *adjusted* δ_i, denoted by $\hat{\delta}_i$, corresponding to the mean value of $\bar{\epsilon} = \sum_{i=1}^{N} \epsilon_i$; the subsequent discussion is, however, valid for any other value for $\bar{\epsilon}$. In the case all learners are trained by the same number of examples we define $\delta_{min} = \min_i \hat{\delta}_i$. The parameter $\hat{\delta}$ is used for comparing the various learners. L_i is considered *better* than L_j if $\hat{\delta}_i \leq \hat{\delta}_j$, i.e. $1 - \hat{\delta}_i \geq 1 - \hat{\delta}_j$, for the same sample size; this definition is extensively used subsequently in this paper.

We now present a list of examples for different existing formulations of the individual learners; in each formulation we describe some interesting features and underlying formulae of the learnability problem. Our intention here is to illustrate some important existing formulations, and also to derive formulae for adjusted $\hat{\delta}$'s for these cases. These formulae are used in obtaining some sufficiency conditions in Section 5. In the rest of this section we assume that each learner is trained with a m-sample.

Example 4.1. *Finite Hypotheses Classes:*

From Example 3.1, we have, $\delta_i = |H_i|e^{-m\epsilon_i}$, and the adjusted δ is given by $\hat{\delta}_i = |H_i|e^{-m\bar{\epsilon}}$, where $\bar{\epsilon} = \frac{1}{N}\sum_{i=1}^{N} \epsilon_i$. Note that if all learners are trained by the same set of examples (or the same number of examples), the learner with least number of hypotheses will yield highest adjusted confidence. Stated in another way, the learner that can explain the sample using only a smaller set of hypotheses will a better predictor of the target concept.

The formulation of Example 4.1 can be adapted to take into the account the *complexity* of the target concept, denoted by n, as illustrated in the next example.

Example 4.2. *Occam's Razor:*

Consider the case H_i is countably infinite. We assume that $C \subseteq H_i$. Let the *complexity* of a hypothesis $h \in H_i$ denote the number of bits needed to specify h in some fixed encoding [7]. An *Occam-algorithm* for H_i with constant parameters $c_i \geq 1$ and $0 \leq \alpha_i < 1$ is a learning algorithm that:

(i) produces a hypothesis of complexity at most $n^{c_i}m^{\alpha_i}$ when given a sample of size m of any concept of C of complexity at most n, and

(ii) runs in time polynomial in the length of the sample.

It has been shown that the existence of an Occam-algorithm for H implies polynomial learnability [7]. Then given a sample of $c \in C$ of complexity at most n, the number of examples needed to ensure polynomial learnability is given by

$$m = k\ln(1/\delta) + k(n^{c_i}/\epsilon)^{1/(1-\alpha_i)}$$

where k is a constant. The value of the adjusted δ is given by

$$\hat{\delta}_i = e^{x - \frac{\epsilon m}{k}}$$

where $x = n^{\frac{c_i}{1-\alpha_i}} \epsilon^{\frac{-\alpha_i}{1-\alpha_i}}$. Note that in this case each learner is characterized by the parameters of its Occam-algorithm.

We now consider more general cases of X and H_i as discussed in [7]. A family $H \subseteq 2^X$ *shatters* a set $X_l = \{x_1, x_2, \ldots, x_l\} \subseteq X$, if $\{h \cap X_l | h \in H\} = 2^{X_l}$, i.e. for every subset of X_l there exists $h \in H$ that contains this subset but not its complement. The *Vapnik and Chervonenkis dimension* of H, denoted by $VCdim(H)$ is the maximum size such that every subset of X of this size is shattered by H. The $VCdim(.)$ plays a very critical role in learnability in that $C \subseteq H$ is learnable if and only if $VCdim(H) < \infty$ [7].

Example 4.3. *Infinite domain and infinite hypotheses classes:*

We now consider the cases when X is finite, finitely enumerable, a Boolean cube or a vector space and $C \subseteq H_i$; let $VCdim(H_i) = d_i < \infty$. From Blumer et al [7], the number of examples needed for learnability using hypothesis space H, $VCdim(H) = d$, is given by

$$\max\left(\frac{4}{\epsilon}\log\frac{2}{\delta}, \frac{8d}{\epsilon}\log\frac{13}{\epsilon}\right).$$

Here, the learning algorithm simply outputs any hypothesis that is consistent with the given sample. Thus we take a sufficient value of the required number of examples given by:

$$m = \frac{4}{\epsilon}\log\frac{2}{\delta} + \frac{8d}{\epsilon}\log\frac{13}{\epsilon}.$$

Then adjusted δ_i corresponding to a m-sample is given by

$$\hat{\delta}_i = 2e^{2d_i \log\left(\frac{13}{\epsilon}\right)} e^{\frac{-m\epsilon}{4}}.$$

Further assume that each example is subjected to misclassification with a probability of $1 - \gamma$. Then the number of examples needed to ensure (ϵ, δ)-condition is given by (Blumer et al [7, Theorem A3.1]):

$$\max\left(\frac{8}{\gamma^2\epsilon}\ln\frac{8}{\delta}, \frac{16d}{\gamma^2\epsilon}\ln\frac{16}{\gamma^2\epsilon}\right).$$

Then adjusted δ_i is given by

$$\hat{\delta}_i = 8e^{2d_i \ln\left(\frac{16}{\gamma^2\varepsilon}\right)} e^{\frac{-m\gamma^2\varepsilon}{8}}.$$

Now if all learners are trained by the same set of examples, the leaner with least VC-dimension for the set of hypotheses will yield highest adjusted confidence.

In some formulations of learnability, the notion of *cover* plays a very important role; one such instance occurs when P_X is known. For $\epsilon \geq 0$, a set $C_\epsilon \subseteq 2^X$ is an *ϵ-cover* of $C \subseteq 2^X$ under P_X if for every $c \in C$ there is $\hat{c} \in C_\epsilon$ such that $\mu(c\Delta\hat{c}) \leq \epsilon$. C is *finitely coverable* with respect to P_X if for every $\epsilon \geq 0$ there is a finite ϵ-cover C_ϵ of C. It has been shown in [6] that C is learnable with respect to P_X if and only if C is finitely coverable with respect to P_X.

Example 4.4. *Learnability under fixed distribution:*

We will now consider the case where P_X is known, which is known as *learnability under fixed distributions* [6]. In this case, the number of examples needed to ensure (ϵ, δ) is given by

$$m = \frac{32}{\epsilon} \ln\left(\frac{N_\epsilon}{\delta}\right)$$

where $N_\epsilon = N_\epsilon(C, P_X, \epsilon)$ is the cardinality of a finite $\frac{\epsilon}{2}$-cover of hypothesis space H with respect to P_X. In this case the adjusted δ_i is given by

$$\hat{\delta}_i = N_i e^{-\frac{m\varepsilon}{32}}$$

where N_i is the $\frac{\bar{\epsilon}}{2}$-cover of H_i. If each example is subjected to misclassification with a probability of ρ, the required number of examples is given by

$$m = \frac{12(\epsilon + \rho - 2\epsilon\rho)}{\epsilon^2(1/2 - \rho)^2} \ln \frac{N}{\delta}$$

The adjusted δ is given by

$$\hat{\delta}_i = N_i e^{-\left(\frac{m\varepsilon^2(1/2-\rho)^2}{12(\varepsilon+\rho-2\varepsilon\rho)}\right)}.$$

More details on these aspects can be found in [6].

Example 4.5. *Learning under malicious errors:*

Kearns and Li [17] study the problem of learning in presence of classification error. The probability distribution P_X is represented as a combination of two distributions P_X^+ and P_X^- based on concept c; P_X^+ and P_X^- are distributions on c and $X - c$ respectively. Here a learning algorithm can make calls to oracle POS_M^β (NEG_M^β) which produces $x \in X$ such that

(a) with probability at least $1 - \beta$, x belongs to c ($X - c$); and

(b) with probability β, x's classification may or may not be correct.

If $\beta \leq \epsilon/2$, a learning algorithm, called *β-robust Occam Algorithm*, exists (for hypothesis space H) with the sample size given by

$$m = \frac{24}{\epsilon}\left(\ln|H| + \ln\frac{2}{\delta}\right)$$

such that ϵ bounds both the errors due to P_X^+ and P_X^-. Let the m examples returned by POS_M^β and NEG_M^β be $u_1, u_2, ..., u_m \in X$, and $v_1, v_2, ..., v_m \in X$ respectively. Here, the β-robust Occam Algorithm returns, with probability at least $1-\delta$, $h \in H$ such that

(i) $\frac{|\{u_i | u_i \notin h\}|}{m} \leq \frac{\epsilon}{2}$

(ii) $\frac{|\{v_i | v_i \in h\}|}{m} \leq \frac{\epsilon}{2}$.

Thus, with high probability h agrees with a fraction exceeding $1 - \frac{\epsilon}{2}$ of the sample. In this case the adjusted δ is given by $\hat{\delta}_i = 2|H_i|e^{-\frac{m\epsilon}{24}}$.

Example 4.6. *Learning subsets of metric spaces:*

We now consider a special case of Haussler [15] who discusses the problem of learning functions of the form $f : X \mapsto Y$; the following discussion deals with case $Y = [0, M] \subseteq R$ (note that for concept learning we have $Y = \{0, 1\}$). Here P_X is chosen from a set of distributions $\mathcal{D}$; if $\mathcal{D}$ denotes set of all probability distributions we have the case of distribution-free learning of Valiant [28] and if $|\mathcal{D}| = 1$ we have the learnability by fixed distribution by Benedek and Itai [6].

Consider a set S with a metric $d : S \times S \mapsto R^+$. For any $\epsilon \geq 0$, an *ϵ-cover* for $T \subseteq S$ is a finite set $N \subseteq S$ such that for all $x \in T$ there exists $y \in N$ such that $d(x, y) \leq \epsilon$. Note that this cover is different from that in Example 4.4.

Consider a probability space (S, D), for $D \in \mathcal{D}$, and let F denote a set of real-valued random variables on S. For any fixed sequence $\bar{\xi} = (\xi_1, \ldots, \xi_m) \in S^m$ and $f \in F$, let $\hat{E}_{\bar{\xi}}(f) = \frac{1}{m}\sum_{i=1}^{m} f(\xi_i)$ which is the empirical estimate of the mean of f based on $\bar{\xi}$. Further for $\bar{\xi} \in S^m$, define

$$F_{|\bar{\xi}} = \{(f(\xi_1), \ldots, f(\xi_m)) | f \in F\}$$

Let d_{L^1} denote the L^1 metric. Then let $N(\epsilon, F_{|\bar{\xi}}, d_{L^1})$ be the size of the smallest ϵ-cover of $F_{|\xi}$ by arbitrary points in R^m under metric L^1. Let us define a metric $d_v(r, s) = \frac{|r-s|}{v+r+s}$, for ant real $v > 0$ and non-negative reals r and s. The sample size needed to solve the learnability problem for any $\mathcal{D}$ in this case is given by

$$\frac{16}{v\epsilon^2}\left[\ln[4E(N(\frac{\epsilon v}{8}, F_{|\bar{\xi}}, d_{L^1}))] + \ln(\frac{1}{\delta})\right].$$

The adjusted δ_i in this case is given by

$$4E[N(\frac{\epsilon v}{8}, F^i_{|\xi}, d_{L^1})]e^{-\left[\frac{\epsilon^2 vm}{16}\right]}$$

where $F^i(.)$ corresponds to the hypothesis class of learner L_i.

5. FUSION AS LEARNING

Consider that the product space of the outputs of the learner $\{0,1\}^N$ be embedded in $\Re^N$. Let H_F be a family of subsets of $\Re^N$; if H_F consists of Boolean functions they can be embedded in $\Re^N$ and viewed as subsets of $\Re^N$. Let $\bar{0}$ denote the origin, and $\bar{1}_i$ denote the N-dimensional vector with 1 in the ith component and 0 in all other components. We say that H_F satisfies *isolation property of degree* N if for $i = 1, 2, \cdots, N$, there exists $h_i \in H_F$ such that $h_i \cap \{\bar{0}, \bar{1}\} = \{\bar{0}, \bar{1}\}$, and $\sum_{j \neq i} w_j \bar{1}_j \notin h_i$ for $w_j \in \{0,1\}$ and $\sum_{j \neq i} w_j \geq 1$. The isolation degree of H_F, denoted by $I_dim(H_F)$ is the maximum value of N such that H_F, embedded in $\Re^N$, has the isolation property of degree N. This property, as simple as it is, is sufficient to guarantee that the passive fusion through learning yields a system that is at least as good as the best of the learners.

Example 5.1.

Consider that H_F is the set of all Boolean functions of N variables. This set trivially satisfies the isolation property of degree N.

Example 5.2.

Let H_F correspond to set of all hyperplanes of $\Re^N$. It is trivial to note that $I_dim(H_F) = N$. In the next example we show that various subclasses of one-dimensional hyperplanes will have $I_dim(.)$ of N.

Example 5.3.

Let H_F correspond to set of all line segments of the form $\{xt + (1-t)y|$ for some $x, y \in \Re^N$ and $|x - y| \geq 1\}$. Here $I_dim(H_F) = N$. Note that the same value is retained if we restrict the line segments to orthonormal or iso-oriented, i.e. parallel to one of the coordinate axes. If the condition $|x - y| \geq 1$ is changed to $|x - y| < 1$, then $I_dim(H_F) = 0$.

Example 5.4.

Consider the class of iso-oriented boxes of dimensionality more than one such that the length of each side is greater than 1. For this set $I_dim(.) = 0$.

Using $I_dim()$, we show the following theorem.

Theorem 3. *Let the domain be either finite, finitely enumerable, a vector space or a Boolean lattice. Consider that same set of examples are used in training the individual learners, and each learner is statistically independent. Let δ_F and δ_i denote the adjusted δ's of the fuser and the learner L_i respectively. If H_F satisfies*

isolation property of degree N, and has a VC-dimension less than or equal to that of the smallest of learners, then composite system can be made such that

$$\delta_F \leq \min_i \delta_i.$$

Proof: We solve the passive fusion problem as another learning problem - we pick the hypothesis $h \in H_F$ that has the least amount of empirical error. Thus by the isolation property of degree N, we are guaranteed to pick a hypothesis whose error is less than or equal to the least among the f_is. Note that by the isolation property, the fuser can "mimic" any of the individual learners. For all the four cases of X - namely, vector space, Boolean lattice, finitely enumerable and finite - the required number of examples to ensure (ϵ, δ)-condition is monotonically related to the VC-dimension [7]. Thus for the same number of examples, the consolidator will have least adjusted δ because its VC-dimension is at most as large as the least of the learners. QED.

We now consider the case when each of the learners is trained by an independently generated m-sample.

Theorem 4. *Consider that each learners is trained by an independently generated m-sample. Let $\bar{\epsilon} = \frac{1}{N}\sum_{i=1}^{N} \epsilon_i$, and δ_i be the adjusted δ of L_i. Then a sufficiency condition for making confidence of the composite system higher than or equal to that of the best of the (statistically independent) learners is given by:*

(i) If X is finite, then

$$|H_F| \leq \min_i\{|H_i|\}e^{(N-1)m\bar{\epsilon}}.$$

(ii) If X is a vector space, Boolean lattice, or finitely enumerable, then

$$VCdim(H_F) \leq \min_i\{VCdim(H_i)\} + \frac{(N-1)m\bar{\epsilon}}{8\log\frac{13}{\bar{\epsilon}}}.$$

(iii) If the classification of the examples is subjects to an error with probability $1-\rho$, then

$$VCdim\,(H_F) \leq \min_i\{VCdim(H_i)\} + \frac{(N-1)\rho^2 m\epsilon}{16\ln\frac{16}{\rho^2\epsilon}}.$$

Proof: The critical point is that the fuser has an (Nm)-sample to pick its hypothesis, whereas each L_i has m examples. First consider the case of finite set. The adjusted δ_i for this case can be obtained from Natarajan [22] as $\delta_i = |H_i|e^{-m\bar{\epsilon}}$ (from Example 4.1). Then $\delta_F = |H_F|e^{-Nm\bar{\epsilon}}$. The condition $\delta_F \leq \min_i \delta_i$ establishes the claim (i). From Blumer et al [7], the number of examples needed for ensuring (ϵ, δ) condition in the case (ii) is given by

$$\max\left(\frac{4}{\epsilon}\log\frac{2}{\delta}, \frac{8d}{\epsilon}\log\frac{13}{\epsilon}\right).$$

The adjusted δ_i is given by (as in Example 4.3.)

$$\delta_i = 2e^{2d_i \log\left(\frac{13}{\epsilon}\right)} e^{\frac{-m\epsilon}{4}}.$$

Now a simple algebraic manipulation of the condition $\delta_F \leq \min_i \delta_i$ establishes (ii). Case (iii) follows similarly (using Example 4.3). QED.

We now take several formulations of the learnability problem and obtain sufficiency conditions similar to those in Theorem 4.

Example 5.5. *Learning Boolean Formulae:*

We consider problem of learning certain disjunctive normal form (DNF) discussed in Valiant [29]. Let M_0 denote the set of all monomials on a set of t predicates, where each monomial is a product obtained by a subset of the predicates. A DNF expression f is of the form $\sum_{m_i \in M} m_i$, where $M \subseteq M_0$. A set $M_1 \subseteq M_0$ is *polynomial generable deterministically* if and only if there is a deterministic algorithm that runs in time polynomial in t and generates descriptions of all members of M_1. A learning algorithm for some M_1 that guarantees (ϵ, δ)- condition such that $h = \frac{1}{\epsilon} = \frac{1}{\delta}$ requires $L(h, |M_1|) \leq 2h(|M_1| + \ln h)$ negative examples. Notice that in this case there is only a single parameter h that characterizes the learner.

Let M_i and h_i denote the monomial set and the parameter of L_i respectively. Now let h_F denote the parameter of the fuser, and let $h_{\min}$ be the parameter of the learner with minimum M_i. Now we show that a sufficiency condition for $h_F \geq h_{\min}$ is $|M_F| \leq N|M_{\min}|$ for $N \geq \frac{h_F \log h_F}{h_{\min} \log h_{\min}}$.
Using $2h_{\min}(|M_{\min}| + \ln h_{\min}) = m$ and $2h_F(|M_F| + \ln h_F) = Nm$, and by equating the two values for m we obtain

$$|M_F| = \frac{N h_{\min} |M_{\min}|}{h_F} + \frac{1}{h_F} \ln\left(\frac{{h_{\min}}^{h_{\min} N}}{{h_F}^{h_F}}\right). \tag{5.5.1}$$

Under the condition $N \geq \frac{h_F \log h_F}{h_{\min} \log h_{\min}}$, we get the second term on the right hand side to be positive. This implies $|M_F| \geq N|M_{\min}|(h_{\min}/h_F)$, which yields $h_F \geq h_{\min}$ if $|M_F| \leq N|M_{\min}|$. Now consider the condition for the general case. Eq (5.5.1) can be rewritten as

$$\frac{|M_F|}{N|M_{\min}|} = \frac{h_{\min}}{h_F} + \frac{1}{N|M_{\min}|} \ln\left(\frac{{h_{\min}}^{\frac{N h_{\min}}{h_F}}}{h_F}\right).$$

The the following condition

$$\frac{|M_F|}{N|M_{\min}|} \leq 1 + \frac{1}{N|M_{\min}|} \ln\left(\frac{{h_{\min}}^{\frac{N h_{\min}}{h_F}}}{h_F}\right). \tag{5.5.2}$$

implies $h_F \geq h_{\min}$. Also, since $h_{\min} > 1$, the condition

$$\frac{|M_F|}{N|M_{\min}} \leq 1 + \frac{1}{N|M_{\min}|}\ln(1/h_{\min})$$

or equivalently $|M_F| \leq N|M_{\min}| + \ln(1/h_F)$ implies Eq (5.5.2) which in turn implies $h_F \geq h_{\min}$.

On the other hand, by noting that $h_F > 1$ we have

$$\frac{h_{\min}^{\frac{Nh_{\min}}{h_F}}}{h_F} \leq h^{\frac{Nh_{\min}}{h_F}} \leq h^{Nh_{\min}}.$$

Thus the following condition

$$\frac{|M_F|}{N|M_{\min}|} \geq 1 + \frac{Nh_{\min}}{N|M_{\min}|}\ln(h_{\min})$$

or equivalently $|M_F| \geq N|M_{\min}| + Nh_{\min}log(h_{\min})$ in turn implies $h_F \leq h_{\min}$.

Several other cases of learning Boolean formulae are presented in Table 1. For the case of learning DNF under error probability of $\rho = \frac{1}{4h|M_1|}$, by equating the value of m we obtain (M_F is the monomial set of the fuser and see Table 1 for definition of other terms)

$$\frac{|M_F|}{N|M_{\min}|} = \frac{h_{\min}}{h_F}\frac{\log(|M_{\min}|h_{\min})}{\log(|M_F|h_F)}. \tag{5.5.3}$$

Now under the condition $|M_F| \leq |M_{\min}|\min(h_{\min}/h_F, N)$, we have $|M_F|h_F \leq |M_{\min}|h_{\min}$ and $\frac{|M_F|}{|M_{\min}|} \leq N$. The first condition implies that $1 \leq \frac{\log(|M_{\min}|h_{\min})}{\log(|M_F|h_F)}$, which reduces Eq (5.5.3) to $\frac{|M_F|}{|M_{\min}|} \geq \frac{Nh_{\min}}{h_F}$. The second condition then implies that $h_F \geq h_{\min}$.

Now consider the third case in Table 1. In this case the number examples needed to ensure the (ϵ, δ)- condition is given by $h(nD(n) + \log n)$ where $D(n)$ is called the dimension of the hypothesis class (see [16] for the precise definition). Then the condition $h_F \geq h_{\min}$ is equivalent to

$$\frac{Nm}{nD_F(n) + \log n} \geq \frac{m}{nD_{\min}(n) + logn}$$

such that $D_{\min}(n) = \min_i\{D_i(n)\}$, where $D_i(n)$ is the dimension of H_i. The above equation is equivalent to

$$D_F(n) \geq ND_{\min}(n) + \frac{N-1}{n}\log n.$$

Case	Formula for m	Condition for $\delta_F \leq \min_i \delta_i$
Learning DNF	$2h(\lvert M_{\min}\rvert + \ln h)$	$\lvert M_F\rvert \leq N\lvert M_{\min}\rvert + \log(1/h_{\min})$
DNF with errors Valiant [21] $\rho = \frac{1}{4h\lvert M_1\rvert}$:error probability M_1: subset of monomials $h = 1/\epsilon = 1/\delta$	$36h\lvert M_1\rvert\log(\lvert M_1\rvert h)$ M_1^i: monomial set of L_i $M_{\min} = \min_i M_1^i$	$\lvert M_F\rvert \leq \lvert M_{\min}\rvert \min(h_{\min}/h_F, N)$
Boolean functions Natarajan [16] n: number of variables $D(n)$: dimension of hypotheses family $h = 1/\epsilon = 1/\delta$	$h(nD(n) + \log n)$	$D_F(n) < ND_{\min}(n) + \frac{(N-1)}{n} logn$

Table 1. Leaning Boolean formulae.

Example 5.6.

We now present the results along the lines of Theorem 4 for Example 4.1 through 4.6 in Table 2; the derivations for all, except the Occam's algorithm, are direct. For Occam's algorithm we have

$$m = k\ln(1/\delta) + k(n^{c_i}/\bar{\epsilon})^{1/(1-\alpha_i)}.$$

Since each Occam's algorithm is characterized by (c_i, α_i), we normalize with respect to $\bar{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i$. Thus normalized value of c_i, denoted by $\bar{c}_i$ is given by $n^{c_i}m\alpha_i = n^{\bar{c}_i}m^{\bar{\alpha}}$; thus $\bar{c}_i = \frac{n\log c_i + (\alpha_i - \bar{\alpha})\log m}{\log n}$. Then the condition $\delta_F \leq \min_i \delta_i$ is equivalent to

$$\left(\frac{n^{c_F}}{\epsilon}\right)^{1/(1-\bar{\alpha})} \leq \left(\frac{n^{c_{\min}}}{\epsilon}\right)^{1/(1-\bar{\alpha})} + \frac{m(N-1)}{k} \qquad (4.6.1).$$

Since $0 < \bar{\alpha} < 1$, we have $\frac{1}{1-\bar{\alpha}} \geq 1$. Thus the condition (4.6.1) is definitely implied by

$$n^{c_F} \leq n^{c_{\min}} + \frac{m(N-1)\bar{\epsilon}}{k}.$$

Since $n \geq 1$, this condition in turn is implied by $c_F \leq c_{\min} + \frac{m(N-1)\bar{\epsilon}}{k}$.

Ex. No.	Case	Condition for $\delta_F \leq \min_i \delta_i$
4.1.	Finite hypotheses classes	$\lvert H\rvert \leq \lvert H_{\min}\rvert e^{(N-1)m\bar{\epsilon}}$
4.2.	Occam's algorithm	$c_F \leq c_{\min} + \frac{m(N-1)\bar{\epsilon}}{k}$
4.3.	Infinite hypothesis classes	$d_F \leq d_{\min} + \frac{(N-1)m\bar{\epsilon}}{8\ln(13/\epsilon)}$
4.4.	Fixed distribution - no error	$N_\epsilon^F \leq N_\epsilon^{\min} e^{(N-1)\frac{m\bar{\epsilon}}{32}}$
	Fixed distribution with error	$N_\epsilon^F \leq N_\epsilon^{\min} e^{\frac{(N-1)\bar{\epsilon}^2(1/2-\rho)^2}{12(\bar{\epsilon}+\rho-2\bar{\epsilon}\rho)}}$
4.5.	Learning under malicious error	$\lvert H\rvert \leq \lvert H_{\min}\rvert e^{(N-1)m\bar{\epsilon}/24}$
4.6.	Learning in metric spaces	$E[N_F(\frac{\epsilon v}{8}, F^F_{\lvert\bar{\xi}}, d_{L^1})]$ $\leq E[N_{\min}(\frac{\epsilon v}{8}, F^i_{\lvert bar\xi}, d_{L^1})]e^{\frac{(N-1)\bar{\epsilon}^2 vm}{16}}$

Table 2. Concept learning.

We now discuss some more formulations of the learnability problem that have not been covered earlier.

Example 5.7. *Learning under noise:*

We consider the case where the classification is prone to an error with probability $\eta \leq \eta_b$. This case has been studied by Angluin and Laird [3]; this noise model is more general than that of Valiant [29] and more benign than that of Kearns and Li [17]. In this case the sample size required to ensure (ϵ, δ) condition is given by

$$m = \frac{2}{\epsilon^2(1-2\eta_b)^2} \ln\left(\frac{2|H_i|}{\delta)}\right).$$

The condition of Theorem 4 yields

$$|H| \leq |H_{\min}| e^{\frac{(N-1)(1-2\eta_b)^2\epsilon^2 m}{2}}$$

where $H_{\min}$ corresponds to the learner with minimum adjusted δ_i.

Example 5.8. *Heuristic learnability:*

Consider that we wish to learn a concept class C_2 by using a hypothesis class C_1 when concepts of latter can only approximate those of C_2. This form of learning has been studied by Amsterdam [2]. A concept class C_1 is said to be *h-dense* in the concept class C_2 if and only if for every $c_2 \in C_2$ and probability P_X, there exists $c_1 \in C_1$ such that $\mu(c_1 \Delta c_2) \leq \epsilon$. Then the concept class C_1 is *h-heuristically learnable* by a concept C_2 if there exists an algorithm A that

(i) runs in time polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$ and the size of $c_1 \in C_1$, and

(ii) outputs a concept $c_2 \in C_2$ such that with probability $1 - \delta$, we have $\mu(c_1 \Delta c_2) < h + \epsilon$.

The number of examples needed to ensure (ϵ, δ)-condition is given by $m = \frac{2}{\epsilon^2} \ln\left(\frac{4|C_1|}{\delta}\right)$. The condition $\delta_F \leq \min_i \delta_i$ yields the following condition

$$|H| = |H_{\min}| e^{(N-1)\frac{m\epsilon^2}{2}}$$

where each H_i is h-dense in C, and $H_{\min}$ corresponds to the learner with minimum adjusted δ_i.

Example 5.9. *Concepts on Strings:*

Now we consider that X is a set of strings on a fixed alphabet $\sum$; this case has been studied by Natarajan [22]. Here X can be naturally partitioned into finitely enumerable classes using the length of the strings, i.e., $X = \bigcup_{i=1}^{\infty} X_m$, where X_m denotes the set of all strings of length m on the alphabet $\sum$. For L_i now let H_i^n denote the set of all strings of H_i of length at most n. The *dimension* of H_i is defined to be $d_i(n) = \log(H_i^n)$. The number of examples needed to ensure the (ϵ, δ)-condition is given by $m = h(d_i(n) + \log(h))$. In this case let $h_F(n)$ denote the parameter of the fuser and $h_{\min}(n)$ denote the parameter of the learner with minimum dimension $d_{\min}(n)$ for particular n. A sufficiency condition for $h_F(n) \leq h_{\min}(n)$ is $d_F(n) \leq N d_{\min}(n)$ for $N > \frac{h_F(n) \log h_F(n)}{h_{\min}(n) \log h_{\min}(n)}$ as in the case of Example 5.5. We obtain a more refined condition as follows: note that $m = h_{\min}(n)[d_{\min}(n) + \log h_{\min}(n)]$ and $Nm = h_F(n)[d_F(n) + \log h_F(n)]$. Thus we obtain (by using value of m from the former in the latter)

$$d_F(n) = N\left(\frac{h_{\min}(n)}{h_F(n)}\right) d_{\min}(n) + \frac{N h_{\min}(n)}{h_F(n)} \log(h_{\min}(n)) - \log(h_F(n)). \quad (5.9.1)$$

The rest of the derivation is similar to that of Example 5.5. Also by noting that $h_{\min}(n) > 1$, the condition of Eq (5.9.1) is implied by

$$d_F(n) \leq N d_{\min}(n) + \log(1/h_{\min}).$$

---

6. FUSION BY LINEAR THRESHOLD FUNCTIONS

The method of the last section requires the knowledge of the examples and the hypotheses of the individual learners. We now consider the case where such knowledge is not readily available. This section also covers the case where the empirical error achieved by the fuser is greater than or equal to that of some of the individual learners. This situation can happen when H_F does not have the isolation property of degree N and also when the sufficiency conditions of the last section are not satisfied.

We consider all learners with minimum $\hat{\delta}_i$, denoted by δ. Further we assume that $\delta < 1/2$. Thus, each of these learners is guaranteed to ensure that the precision is greater than ϵ with a probability of at most delta δ. Further assume that we have N such learners in a given set of $M \geq N$ learners.

We now discuss a specific class of fusion rules obtained by taking the linear combinations of the outputs of the learners and comparing with a threshold, i.e. functions of the form $\sum_{i=1}^{N} y_i \geq rN$. Note that the example of Theorem 1 (Section 3) corresponds to $r = 1$, and the example of Theorem 2 corresponds to $r = 1/N$. We first show the result for the special case $r = 1/2$, and then generalize it.

The basic intuition for expecting that such fusers are possible is provided by the following *Jury Theorem* formulated by Condorcet in 1785.

Condorcet Jury Theorem. *[14] Given a group of N (N is odd) voters each capable of making the right choice in a set of two alternatives with probability p independent of others, the probability that the majority (the rule that chooses the decision of the majority of the group) makes the right choice is given by*

$$P_N = \sum_{i=m}^{N} \binom{N}{i} p^i (1-p)^{N-i}$$

where $m = (N+1)/2$.

Further

(i) if $1 > p > 1/2$, then P_N is monotonically increasing in N and $\lim_{N\to\infty} P_N \to 1$;

(ii) if $1/2 > p > 0$, then P_N is monotonically decreasing in N and $\lim_{N\to\infty} P_N \to 0$;

(iii) if $p = 1/2$, then $P_N = 1/2$ for all N.

Several variations and extensions of this theorem have been studied in various contexts such as information pooling group decision making models (see Grofman and Owen [14] and Miller [20] for some recent surveys), majority systems [8,27], etc.

In the present context, if N is sufficiently large, then the composite system can be made to have confidence arbitrarily close to 1; hence its δ can be made smaller than that of the best learner (namely, δ). Here, we are interested in more exact conditions that guarantee that the composite system is better than the best of the learners. We first show that a majority system ensures this condition if the number of learners is larger than $\frac{k}{2(\frac{1}{2}-\delta)}$ and $0 < \delta < 1/2$, such that $k \leq \min\left(1+\sqrt{2}, \frac{1}{(1/2-\delta)}\right)$

in Theorem 5; this condition does not guarantee that majority system can be used as a fuser for smaller values of N. Then, we show that the composite system can be guaranteed to be better than the best of the learners by suitably choosing r of the threshold function as specified in Theorem 6.

Theorem 5. *If the number of learners is larger than* $\frac{k}{2(\frac{1}{2}-\delta)}$ *and* $0 < \delta < 1/2$, *such that* $0 < k \leq \min\left(1+\sqrt{2}, \frac{1}{(1/2-\delta)}\right)$ *then the fusion rule* $\sum_{i=1}^{N} y_i \geq N/2$ *will ensure that* $\delta_F \leq \delta$, *where* δ *corresponds to the lowest of the individual learners trained with an* l*-sample. For example, by choosing* $k = 2(1/2-\delta)$ *we have* $\delta_F \leq \delta$.

Proof: Let event E_i denote the fact that $\mu(f\Delta f_i) \leq \epsilon$. E_i occurs with a probability of δ. Assuming all events E_i are independent, the probability that there will be at most $N/2$ successes of the events from a total of N Bernoulli trials is given by (Angluin and Laird [3]):

$$L(\delta, N, 1/2) = \sum_{k=0}^{N/2} \binom{N}{k} \delta^k (1-\delta)^{N-k} \leq e^{-2(1/2-\delta)^2 N}$$

This is the probability with which this composite rule generates an error greater than ϵ. Thus, we have the probability that $\mu(f\Delta g) \geq \epsilon$ is less than $e^{-2(1/2-\delta)^2 N}$. By using $N \geq \frac{k}{2(1/2-\delta)}$ we have $e^{-2(1/2-\delta)^2 N} \leq e^{-(1/2-\delta)k}$. Now we show that under the hypotheses of the theorem $\delta \geq e^{-(1/2-\delta)k}$. In the view of Lemma A1, under the condition $(1/2-\delta)k \leq 1$ it suffices to show that

$$\delta \geq 1 - k(1/2-\delta) + \frac{k^2}{2}(1/2-\delta)^2.$$

Since $\delta < 1/2$ and $k > 0$ this condition is implied by

$$\delta \geq 1 + k(1/2-\delta) + k^2(1/2-\delta)^2.$$

This condition in turn yields a quadratic equation

$$k^2(\delta - 1/2)^2 - (k+1)(\delta - 1/2) + 1/2 \leq 0$$

whose solution is given by

$$(\delta - 1/2) = \frac{(k+1)}{2k^2} \pm \frac{1}{2k^2}\sqrt{(k+1)^2 - 2k^2}. \tag{6.1}$$

In order that the roots of the above equation are real, the quantity under square root sign must be positive. This yields a quadratic equation $k^2 - 2k - 1 \leq 0$ whose roots are $1 \pm \sqrt{2}$; this condition yields $1-\sqrt{2} \leq k \leq 1+\sqrt{2}$. This condition coupled with the condition $(1/2-\delta)k < 1$ is satisfied by the hypothesis

$$0 < k \leq \min\left(1+\sqrt{(2)}, \frac{1}{(1/2-\delta)}\right). \tag{6.2}$$

Note that the range of k always exists since the lower bound is always negative and the upper bound is positive by hypothesis.

Now going back to Eq (6.1), we have $(k+1)^2 - 2k^2 = (k+1-\sqrt{2}k)(k+1+\sqrt{2}k)$. Let $X = k(1-\sqrt{2})+1$. Now the condition $X \geq 0$ yields $k \leq \frac{1}{\sqrt{2}-1} = \sqrt{2}+1$; this condition is subsumed by Eq (6.2). Then we have

$$[(1+\sqrt{2})k+1]^2 \geq (k+1)^2 - 2k^2 \geq [(1-\sqrt{2})k+1]^2$$

Now the condition of Eq (6.1) can be expanded as

$$\frac{(k+1)}{2k^2} - \frac{1}{2k^2}\sqrt{(k+1)^2 - 2k^2} \leq (\delta - 1/2) \leq \frac{(k+1)}{2k^2} + \frac{1}{2k^2}\sqrt{(k+1)^2 - 2k^2}.$$

Since $X^2 \leq [(k+1)^2 - 2k^2]$ and $X \geq 0$ this condition is implied by

$$\frac{(k+1)-X}{2k^2} \leq (\delta - 1/2) \leq \frac{(k+1)+X}{2k^2}.$$

Using this, we obtain the bounds on δ as follows:

$$\frac{1}{\sqrt{2}k} + \frac{1}{2k^2} \leq \delta \leq \frac{1}{k}(1 - \frac{1}{\sqrt{2}}) + \frac{3}{2k^2}.$$

Now in order that the range is feasible for δ we have

$$\frac{k}{\sqrt{2}} \leq k(1 - \frac{1}{\sqrt{2}}) + 1$$

which yields $k \leq \frac{1}{\sqrt{2}-1}$ which is equivalent to $k \leq 1+\sqrt{2}$.

To show the second part choose $k = 2(1/2 - \delta)$ which translates to condition on δ as $\frac{1-\sqrt{2}}{2} \leq \delta \leq \frac{1+\sqrt{2}}{2}$, which is implied by the hypothesis $0 < \delta < 1/2$. Hence the Theorem. QED

Note that $r = 1/2$ corresponds to the majority rule; it is possible to reduce the confidence even under a non-majority rule by choosing a suitable value for r as shown in the following theorem.

Theorem 6. *For $0 < \delta < 1/2$ and $N \geq \frac{k^3}{2(k-1)}$ for any $r = \delta + \frac{k-1}{k^2}$ and $k > 1$, δ_F is smaller than that of the lowest adjusted δ of the individual learners.*

Proof: Along the lines of proof of Theorem 4 and from Angluin and Laird [3], we have

$$LE(\delta, N, r) \leq e^{-2(r-\delta)^2 N}$$

By choosing $N \geq \frac{k}{2(r-\delta)}$ the condition for reducing the adjusted δ of the entire system, we have the condition

$$\delta \geq 1 - k(r-\delta) + \frac{k^2(r-\delta)^2}{2}$$

This yields a quadratic equation in $(r - \delta)$ whose solution yields the following condition:

$$\frac{(k-1) - \sqrt{(1-k)^2 - 2k^2(1-r)}}{k^2} \leq (r - \delta) \leq \frac{(k-1) + \sqrt{(1-k)^2 - 2k^2(1-r)}}{k^2}.$$

Note that, since $k > 1$ and $k\sqrt{2(1-r)} > 0$

$$(1-k)^2 - 2k^2(1-r) \geq [(k-1) - k\sqrt{2(1-r)}]^2,$$

and

$$[(k-1) - k\sqrt{2(1-r)}]^2 = [k\sqrt{2(1-r)} - k + 1]^2 = [k(\sqrt{2(1-r)} - 1) + 1]^2.$$

Now let $X = k(\sqrt{2(1-r)} - 1) + 1$. First consider the case $X \geq 0$; this condition implies

$$\sqrt{2(1-r)} \geq 1 - 1/k.$$

The above condition on $r - \delta$ is implied by the following condition (similar to the proof of Theorem 5)

$$\frac{k-1-X}{k^2} \leq (r - \delta) \leq \frac{k-1+X}{k^2}.$$

Thus the lower bound is evaluated as follows:

$$\frac{2k - 2 - k\sqrt{2(1-r)}}{k^2} \geq \frac{k-1}{k^2}.$$

Then the upper bound can be evaluated as

$$\frac{k\sqrt{2(1-r)}}{k^2} \geq \frac{k-1}{k^2}.$$

Thus the choice of $r = \delta + \frac{k-1}{k^2}$ will definitely imply the required condition on $r - \delta$. To ensure $0 < r < 1$ we need to ensure that $0 < \frac{k-1}{k^2} < 1/2$; this condition can be easily shown to be satisfied if $k > 1$. In the second case when $X \leq 0$, we use $Y = -X$ and the rest of the derivation is essentially the same. The condition $N \geq \frac{k^3}{2(k-1)}$ is straight forward. Hence the theorem. QED

As an example of the conditions of this theorem, for $k = 2$, we obtain $N \geq 4$ and $r = \delta + 0.25$.

Consider that all learners are consistent, i.e. they correctly classify each example. In this case, the method of last section stands the chance of picking one of the learners which may or may not be the best (i.e. has lowest adjusted δ); whereas, using Theorem 6 we are guaranteed to have the performance of the best of the learner in the worst case.

One of the natural questions to ask is: can we minimize the expected error of misclassification of the fuser in the class of functions of the form $\sum_{i=1}^{N} w_i y_i \geq r$, for $w_i \in \Re$? If the probabilities with which the examples are chosen from the target concept and its complement, and $Prob[f_i - f]$ and $Prob[f - f_i]$ are known, then we can compute w_i's and r of the required fuser by using the result of Chow [9]; this method cannot be directly applied in distribution-free formulations, or in the fixed distribution formulation of [6]. It would be interesting to see if some estimated values can be used to compute w_i's and r in order to guarantee close to optimal performance.

7. CONCLUSIONS

We have addressed the N-learners problem where each learner is capable of learning subsets of a domain set X in the sense of Valiant [28]. That is for each learner and for any $c \in C \subseteq 2^X$, given a finite set of examples of the form $< x_1, M_c(x_1) >; \ldots ; < x_l, M_c(x_l) >$ generated according to an unknown distribution on X, each learner produces a close approximation to c with high a probability. The N learners problem requires a combination of the outputs of the N learners using a single consolidator. We considered the paradigm of *passive fusion*, where each learner is trained with the sample, and then consolidator is allowed to use the sample and the functions of the learners. We inferred the fusion rule by formulating this problem as a basic learning problem. A sufficiency condition to make this composite system better than the best of the individual learners is: the hypothesis space of the consolidator (a) satisfies the *isolation property* of degree at least N, and (b) has Vapnik-Chervonenkis dimension less than or equal to that of the individual learners. Then we considered the case where the fuser does not have access to the training sample or the hypotheses of the individual learners. Then by suitably designing a linear threshold function of the outputs of individual learners, we showed that the confidence parameter of the entire system can be made greater than or equal to that of a best learner.

This work can be used as a basis for applications involving sensor fusion, information pooling and majority systems; to this end, however, some extensions and adaptations of the present formulation would be needed. Future work can be focussed in several directions. Obvious extensions to the present work include fusion of functions and relations, and active fusion. Another topic of interest deals with the cases where the composite system is capable of achieving tasks that are beyond the capabilities of the individual learners. Some preliminary work on this topic has been presented in [26].

APPENDIX A

Lemma A.1.

(a) For $0 < x < 1$,

$$e^{-x} \leq 1 - \frac{x}{1!} + \frac{x^2}{2!}$$

(b) For $M \leq x \leq M+1$, for positive integer M

$$1 - \frac{x}{1!} + \frac{x^2}{2!} + \ldots - \frac{x^M}{M!} \leq e^{-x}\ leq 1 - \frac{x}{1!} + \frac{x^2}{2!} + \ldots - \frac{x^{M-1}}{(M-1)!} \quad \textit{for } M \textit{ odd}$$

$$1 - \frac{x}{1!} + \frac{x^2}{2!} + \ldots - \frac{x^{M-1}}{(M-1)!} \leq e^{-x} \leq 1 - \frac{x}{1!} + \frac{x^2}{2!} + \ldots + \frac{x^M}{M!} \quad \textit{for } M \textit{ even}$$

Proof: We express $e^{-x} = 1 - \frac{x}{1!} + \frac{x}{2!} - \ldots + (-1)^i \frac{x^i}{i!} + \ldots$ into even and odd forms respectively as follows:

$$e^{-x} = \left(1 - \frac{x}{1!}\right) + \frac{x^2}{2!}\left(1 - \frac{x}{3}\right) + \ldots + \frac{x^i}{i!}\left(1 - \frac{x}{i+1}\right) + \ldots$$

$$e^{-x} = 1 - \frac{x}{1!}\left(1 - \frac{x}{2}\right) - \frac{x^3}{3!}\left(1 - \frac{x}{4}\right) - \ldots - \frac{x^i}{i!}\left(1 - \frac{x}{i+1}\right) - \ldots$$

Simple algebraic manipulation of these equations will establish the required result. QED

REFERENCES

1. M. Abramowitz and I. A. Stegun (eds), *Handbook of Mathematical Functions*, Dover Pub. Inc., New York, 1965.
2. J. Amsterdam, Extending the Valiant learning model, *Proc. 5th Int. Conf. on Machine Learning*, 1988, 381-394.
3. D. Angluin and P. D. Laird, Learning from Noisy examples, *Machine Learning*, vol. 2, 1988, 343-370.
4. E. B. Baum, On learning a union of half spaces, *Journal of Complexity*, vol.6, 1990, pp. 67-101.
5. G. Benedek and A. Itai, Nonuniform learnability, *Proc. of 15th Int. Conf. on Automata, Languages and Programming*, 1988, 82-92.
6. G. Benedek and A. Itai, Learnability by fixed distributions, *proc. 1988 Workshop on Computational Learning Theory*, 1988, 80-90.
7. A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth, Learnability and the Vapnik-Chervonenkis Dimension, *J. Assoc. Comput. Mach.*, vol 36(4), 1989.
8. P. J. Boland, F. Proschan and Y. L. Tong, Modelling dependence in simple and indirect majority systems, *J. Applied Probability*, vol. 26, 1989, pp. 81-88.
9. C. K. Chow, Statistical independence and threshold functions, *IEEE Trans. on Electronic Computers*, vol. EC-16, 1965, pp. 66-68.
10. R. M. Dudley, A course on empirical processes, in *Lecture Notes in Mathematics*, vol. 1097, Springer-Verlag, New York, 1984.
11. H. F. Durrant-Whyte, Consistent integration and propagation of disparate sensor observations, *Int. J. Robot. Res.*, vol. 6, no.3, 1987, pp. 3-24.
12. W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 1, John Wiley, New York, 1957.
13. C. W. Glover, M. Silliman, M. Walker, P. Spelt, N. S. V. Rao, Hybrid neural network and rule-based pattern recognition system capable of self-modification, *Proc. SPIE Conf. on Applications of Artificial Intelligence*, vol. VIII, ed. M. Trivedi, 1990.
14. B. Grofman and G. Owen, Condorcet models, avenues for future research, in *Information Pooling and Group Decision Making: Proceedings of the Second University of California, Irvine, Conference on Political Economy*, eds. B. Grofman and G. Owen, Jai Press Inc., Greenwich, Connecticut, 1986, pp. 173-192.
15. D. Haussler, Generalizing the PAC model: sample size bounds from metric dimension-based uniform convergence, *Proc. 3rd Symp. on Foundations of Computer Science*, 1989, 40-45.
16. A. Kak and S. Chen (eds.), *Spatial Reasoning and Multi-Sensor Fusion*, Morgan Kaufman Pub. Inc., 1987.

17. M. Kearns and M. Li, Learning in the presence of malicious errors, *Proc. 1988 Symposium on Theory of Computing*, 1988, 267-280.

18. C. L. Liu, *Introduction to Combinatorial Mathematics*, McGraw-Hill Book co, New York, 1968.

19. N. Littlestone, Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm, *Machine Learning*, vol. 2, 1988, pp. 285-318.

20. N. R. Miller, Information, electorates, and democracy: Some extensions and interpretations of the Condorcet Jury Theorem, in *Information Pooling and Group Decision Making: Proceedings of the Second University of California, Irvine, Conference on Political Economy*, eds. B. Grofman and G. Owen, Jai Press Inc., Greenwich, Connecticut, 1986, pp. 173-192.

21. M. L. Minsky and S. A. Papert, *Perceptrons*, expanded edition, MIT Press, 1988.

22. B. K. Natarajan, On learning sets and functions, *Machine Learning*,vol 4, 1989, 67-97.

23. N. J. Nilsson, *Learning Machines*, McGraw-Hill Book Co., 1965.

24. Y. Pao, *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley Pub. Co., 1989.

25. N. S. V. Rao and C. W. Glover, Hybrid intelligent perception system, *Proc. Southeastcon 91*, Williamsburg, Virginia, April 1991, pp. 832-836.

26. N. S. V. Rao, E. M. Oblow, C. W. Glover, Learning Boolean concepts of hyperplanes, ORNL Tech. Rep., under preparation.

27. S. N. Srihari, Reliability analysis of majority vote systems, *Information Sciences*, vol. 26, 1982, pp. 243-256.

28. L. G. Valiant, A theory of the learnable, *Communications of the ACM*, 27(11):1134-1142.

29. L. G. Valiant, Learning disjunctions of conjunctions, *Proc. 9th International Joint Conf. Artificial Intelligence*, 1985, pp. 560-566.

30. V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, New York, 1982.

INTERNAL DISTRIBUTION

1. B. R. Appleton
2. M. Beckerman
3–7. C. W. Glover
8. J. P. Jones
9–13. H. E. Knee
14–18. G. E. Liepins
19–23. R. C. Ward
24–28. R. C. Mann
29–33. E. M. Oblow
34. F. G. Pin
35. V. Protopopescu
36. D. B. Reister
37. J. C. Schryver
38. P. Spelt
39. R. W. Brockett (Consultant)
40. J. J. Dorning (Consultant)
41. J. E. Leiss (Consultant)
42. N. Moray (Consultant)
43. M. F. Wheeler (Consultant)
44. EPMD Reports Office
45–46. Laboratory Records Department
47. Laboratory Records, ORNL–RC
48. Document Reference Section
49. Central Research Library
50. ORNL Patent Section

EXTERNAL DISTRIBUTION

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

65. Michael Kearns, Aiken Computation Laboratory, Harvard University, Cambridge, Massachusetts 02138
66. M. H. Kalos, Courant Institute of Math Science, New York University, 251 Mercer Street, New York, NY 10012
67. Philip D. Laird, NASA Ames Research Center, Moffett Field, California 94035
68. Pat Langley, Department of Information & Computer Science, University of California, Irvine, California 92717
69. Nick Littlestone, Department of Computer & Information Sciences, University of California, Santa Cruz, California 95064
70. Oscar P. Manley, U.S. Department of Energy, Division of Engineering, Mathematical and Geosciences, Office of Basic Energy Sciences, Germantown, MD 20545
71. Alan Meyrowitz, Office of Naval Research, Code 1133, 800 N. Quincy St., Arlington, VA 22217
72. B. K. Natarajan, The Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
73. Judea Pearl, Computer Science Department, University of California, 405 Hilgard Ave., Los Angeles, CA 90024
74. J. R. Quinlan, Basser Department of Computer Science, University of Sydney, Sydney, New South Wales 2006, AUSTRALIA

75–79. N. S. V. Rao, Department of Computer Science, Old Dominion University, Norfolk, VA 23529-0162

80. Zbigniew W. Ras, Department of Computer Science, University of North Carolina, Charlotte, North Carolina 28223
81. Ronald L. Rivest, MIT Laboratory for Computer Science, Cambridge, Massachusetts 02139
82. George Shackelford, Department of Information & Computer Science, University of California, Irvine, California 92717
83. Robert Sloan, MIT Laboratory for Computer Science, Cambridge, Massachusetts 02139
84. Wes Snyder, Department of Radiology, Bowman Gray School of Medicine, 300 South Hawthorne Drive, Winston Salem, NC 27103
85. Richard S. Sutton, GTE Laboratories, Inc., 40 Sylvan Road, Waltham, Massachusetts 02254
86. V. R. R. Uppuluri, 130 Indian Lane, Oak Ridge, TN 37830
87. Paul E. Utgoff, Department of Computer & Information Science, University of Massachusetts, Amherst, Massachusetts 01003
88. L. G. Valiant, Aiken Computation Laboratory, Harvard University, Cambridge, Massachusetts 02138
89. Jeffrey S. Vitter, Department of Computer Science, Brown University, Providence, Rhode Island 02912
90. Manfred K. Warmuth, Department of Computer & Information Sciences, University of California, Santa Cruz, California 95064
91. C. R. Weisbin, Jet Propulsion Laboratory, MS 198/330, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California 91109-8099
92. Lotfi A. Zadeh, Computer Science Division, University of California, Berkeley, CA 94720
93. Maria Zemankova, Department of Computer Science, University of Tennessee, Knoxville, Tennessee 37996-1301

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

95–104. Office of Scientific and Technical Information, Department of Energy, Oak Ridge, TN 37831