LEARNING TO RECOGNIZE PATTERNS WITH AN IMPERFECT TEACHER

By

KUMARASAMY SHANMUGAM

Bachelor of Engineering
Madras University
Madras, India
1964

Master of Engineering
Indian Institute of Science
Bangalore, India
1966

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF PHILOSOPHY
May, 1970

LEARNING TO RECOGNIZE PATTERNS WITH AN IMPERFECT TEACHER

Thesis Approved:

Arthur M. Breipohl
Thesis Adviser

Bennett L. Basore

Ray Yarbrough

J. Leroy Folks

D. Durham
Dean of the Graduate College

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my thesis adviser, Professor A. M. Breiphol, for his assistance during this research. For the long hours of his time he so generously devoted on my behalf I am truly greatful.

I also wish to express my appreciation to the other members of my committee, Professor B. L. Basore, Professor Yarlagadda and Professor J. L. Folks for their advice and especially for their teaching excellence. I further wish to thank Dr. David E. Bee who served on my committee during the initial stages of my studies and to Dr. Gerald M. Funk for reading the final manuscript and offering valuable suggestions.

The financial support from the Department of the Army Electronics Command (Contract No. DAAB07-68-C-0083) is greatfully acknowledged. I also wish to thank the Department of Electrical Engineering, Oklahoma State University, for providing me the teaching assistantship and to Mr. Dwayne Wilson for his help in securing the same.

Finally, to my wife, Radha, I wish to express my appreciation for her patience and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Statement of the Problem. Pattern classification covers an extremely broad spectrum of problems ranging from the design and implementation of actual recognition devices to the philosophical question of learning and intelligence. A major step in pattern recognition common to all these problems consists of developing procedures which classify observations such that a particular strategy is optimized. This step can be formulated as follows. A set of N measurement pairs $(\underline{X}_1,\theta^1),(\underline{X}_2,\theta^2),\ldots,(\underline{X}_N,\theta^N)$ are given as sample (training) patterns. $\underline{X}_i, i = 1,2,\ldots,N$, are vector measurements drawn from one of the several possible categories of patterns $\theta_1,\theta_2,\ldots,\theta_R$. The θ^i's take the form $\hat{\theta}_k$, when the measurement $\underline{X}_i$ is identified as being from category θ_k. After development of a pattern classification procedure, only the measurement $\underline{X}$ is available and an estimate is desired of the category from which the measurement $\underline{X}$ was drawn.

For probabilistic patterns with known probability distribution functions, a Bayes' procedure can be used to arrive at an "optimum" decision rule. If all the information about the distribution functions is not known, then the unknown information must be estimated from the given sample patterns. This problem of estimating unknown information and using the estimators to arrive at a classification procedure is generally referred to as "learning to recognize patterns".

Usually a sample identification scheme furnishes the identification θ^i for the sample pattern X_i. In many practical situations the sample identification scheme may incorrectly identify some of the training samples. For example in attempts to classify severe storm patterns using electromagnetic data drom the storms, meteorological measurements such as reports of hail will be used as standards for identifying the sample patterns. Often these reports of hail are unreliable and hence there is a possibility of some incorrectly identified sample patterns. Another example where there is a possibility of incorrect identification of sample patterns occurs in attempts to classify lightning discharges into cloud to cloud or cloud to ground type based on electromagnetic data. Here the sample identification is based on visual sightings of the discharges. In many occasions, the discharge is only partly visible and hence the possibility of incorrect identification. These examples give the motivation for investigating pattern recognition schemes where the information in θ^i is known to be occasionally incorrect.

Depending on the nature of information available in the form of θ^i's, the problem of learning to recognize patterns can be subdivided into three classes. If θ^i contains identification of the true category from which $\underline{X}_i$ was drawn, then learning is said to take place with a "perfect teacher". If no identification of the true category, θ_i, is available then this class of problems is referred to as "learning without a teacher". In between these two is the problem of "learning with an imperfect teacher", where the imperfect teacher is characterized by

$$P(\hat{\theta}_i|\theta_i) = \beta_{ii}, \qquad i = 1,\ldots,R$$

and

$$\left.\begin{aligned} P(\hat{\theta}_j|\theta_i) &= \beta_{ji} \\ \beta_{ji} &< \beta_{ii} \end{aligned}\right\} \quad \begin{aligned} &i,j = 1,\ldots,R \\ &i \neq j \end{aligned} \qquad .$$

This dissertation is concerned with developing procedures for learning to recognize patterns with an imperfect teacher.

1.2 Existing Solutions. Various solutions to the problem of "learning with a perfect teacher" have been proposed in the literature. A recent survey of the literature on the solutions to the problem is given by Ho and Agarwala (1), and Nagy (2). These solutions can in general be divided into two major categories, parametric and nonparametric. In parametric methods a functional form of the conditional densities is assumed to be known except for a set of parameters. These parameters are estimated from the given set of identified sample patterns. Either simple point estimators or Bayesian estimators are used. A commonly used estimation procedure makes use of the recursive properties of Bayes' estimators. In addition to these, several sequential decision procedures are also available. Fu and his associates (3,4) have studied various aspects of sequential methods as applied to pattern recognition problems. The sequential probability ratio test used by Fu will require the smallest number of features to reach a classification decision on the average. Fu also suggested a time varying stopping boundary to assure that a classification decision is reached in a finite time.

Three of the most commonly used nonparametric methods are the nearest neighbor rule, threshold logic and the method of potential functions. The nearest neighbor rule, first proposed by Fix and Hidges (5) assigns $\underline{X}$ to the same category as that of its nearest neighbor among the identified sample patterns. The performance of the nearest neighbor

rule has been analyzed by Cover and Hart (6). The threshold logic unit is a linear categorizer which assigns $\underline{X}$ to category θ_1 if $\underline{X}^T\underline{W} > W_0$ and to θ_2 otherwise. The weights $\underline{W}$ and W_0 are determined iteratively through a "training" procedure. Proofs for convergence of these procedures can be found in Nilsson (7). The object in the method of potential functions, first developed by Aizerman (8) and Braverman (9), is to find a function $\psi(X) = \sum_{i=1}^{m} \Phi_i(\underline{X})W_i$ defined on the pattern space which is positive for all $\underline{X}\epsilon\theta_1$. The Φ_i's are a set of orthonormal functions specified ahead of time. Assuming certain normative conditions, the use of the theory of stochastic approximation leads to a sequence of weights which will converge to the optimum W_i's.

In addition to these methods there is another elegant method proposed by Sprecht (10), called the polynomial discriminant function method. In this method, the density functions $f_{\underline{X}|\theta_i}$ are estimated in a polynomial form and discriminant functions are formed using these polynomials. One of the main advantages of this method is that estimation can be done serially and this results in a reduction of storage requirements on the computer.

Although our society seems to abound with real life examples of "learning without a teacher" not much analytical work has been done towards solving this general problem. Fralick (11) first suggested a bounded scheme for learning to recognize the presence of signal in a noisy channel. Fralick's scheme makes use of the recursive properties of a Bayes' solution to realize a machine of finite size. Patrick and Hancock (12) extended this scheme to more general situations. There are many other methods suggested by various authors which are claimed to "work" in some sense. But very little analysis or computational

results have been reported. A summary of several of these procedures is given by Spragins (13).

Extention of methods like the one proposed by Fralick to the problem of learning with an imperfect teacher is extremely difficult. No finite size parametric schemes have been proposed. This is because of the fact that no finite dimensional sufficient statistics exist for the parameters of the densities involved and the reproducing properties of Bayesian procedures are lost due to additional terms introduced by the imperfect teacher. Imperfect identification also causes an overlap in the training patterns and this makes error correcting procedures, like that used with threshold logic, to fail to converge. Duda and Singleton (14) have shown that for orthogonal patterns the average weight vector converges to a solution vector even though the training patterns are incorrectly labelled. However this is not true for nonorthogonal patterns or for patterns with continuous components.

Whitney and Dwyer (15) have analyzed the performance of the nearest neighbor rule with an imperfect teacher and have shown that the expected risk, R_n, is bounded by

$$(1 - \beta) + (2\beta - 1)R^* \leq R_n \leq (1 - \beta) + (2\beta - 1)[2R^*(1 - R^*)] \quad .$$

Where β is the probability that the imperfect teacher correctly identifies a sample and R^* is the Bayes' risk. The above bounds are good when there are only two categories of patterns.

1.3 Present Contributions. In this dissertation a decision rule for dichotomizing patterns with an imperfect teacher is derived. Using a nonparametric estimator for the unknown densities appearing in the decision rule, a procedure for learning to dichotomize patterns with an

imperfect teacher is given. It is shown that the proposed learning scheme has an asymptotic average risk equal to the Bayes' minimum risk. For nonoverlapping densities and for densities with overlap less than $(1 - \beta)$ it is shown that the average asymptotic performance of the proposed learning scheme is better than the teacher and the nearest neighbor rule. It is also shown that for nonoverlapping densities the learning scheme performs better than the teacher on the average after looking at a finite number of sample patterns.

Using these results as motivation, feedback is considered as a means of improving the performance of the learning scheme. Several different feedback schemes are considered and their relative advantages and disadvantages are given. It is shown that a feedback learning scheme using a threshold in feedback provides an easy method for combining the learning scheme's own knowledge with that of the teacher. Expressions for the threshold are derived in terms of β and the sample size n using two different approaches. The idea of feedback is extended to the case of unequal sample size ($P(\theta_1) \gg P(\theta_2)$). Results of simulations of the proposed learning schemes with and without feedback are presented.

CHAPTER II

LEARNING TO RECOGNIZE PATTERNS WITH AN IMPERFECT TEACHER

2.1 Introduction. The main object of pattern recognition is to derive a decision rule for classifying a pattern $\underline{X}$ into one of the R possible categories $\theta_1,\ldots,\theta_R$ such that a particular strategy is optimized. Statistical decision theory can be used as means to establish discriminant functions for classifying probabilistic patterns. The strategy to be optimized is specified in terms of a loss function L_{ij}, defined for $i = 1,\ldots,R$ and $j = 1,\ldots,R$. The loss function L_{ij} represents the loss incurred when the machine or the student places a pattern actually belonging to category j into category i. If a machine classifies patterns such that the "average value" of L_{ij} is minimized, the machine is said to be optimum. Such a machine is also known as a Bayes' machine.

For symmetrical loss function of the form

$$L_{ij} = 1 - \delta_{ij} \quad , \tag{2.1.1}$$

where δ_{ij} is the kronecker delta function, it has been shown (7) that the Bayes' machine uses discriminant functions of the form

$$D_{\theta_i}(\underline{x}) = P(\theta_i) f_{\underline{X}|\theta_i}(\underline{x}|\theta_i) \quad . \tag{2.1.2}$$

$P(\theta_i)$ is the prior probability of occurance of category θ_i and $f_{\underline{X}|\theta_i}(\underline{x}|\theta_i)$ is the probability density function of pattern $\underline{X}$ given that

it belongs to θ_i. The machine assigns a given pattern $\underline{x}$ to category θ_i if

$$D_{\theta_i}(\underline{x}) > D_{\theta_j}(\underline{x}) \qquad j = 1,\ldots,R;\ j \neq i \quad . \qquad (2.1.3)$$

It is assumed in Equation 2.1.2 that all information relevant to the prior probabilities $P(\theta_i)$ and the conditional densities $f_{\underline{X}|\theta_i}$ were completely known. However, in practice, this information is only partially known and the unknown information must be learned (estimated) from the given set of labelled sample patterns. Several parametric (7) and nonparametric methods (6,10) are available for estimating the unknown information in the discriminant functions, the information associated with $D_{\theta_i}(\underline{x})$ being estimated from samples which are known to belong to category θ_i, i.e. from samples labelled as θ_i.

2.2 Decision Rule for Pattern Recognition with an Imperfect Teacher. A decision rule similar in form to the one described in Section 2.1 can be derived for learning with an imperfect teacher. The imperfect teacher labels the sample patterns as $\hat{\theta}_1,\ldots,\hat{\theta}_R$ with the probability of correct labeling given by

$$P(\hat{\theta}_i|\theta_i) = \beta_{ii}, \qquad i = 1,\ldots,R \quad , \qquad (2.2.1)$$

and the probability of incorrect labeling given by

$$P(\hat{\theta}_j|\theta_i) = \beta_{ji} < \beta_{ii}; \qquad i,j = 1,\ldots,R \quad . \qquad (2.2.2)$$
$$i \neq j$$

In this dissertation it will be assumed that

$$P(\hat{\theta}_i|\theta_i) = \beta > \frac{1}{R}; \qquad i = 1,\ldots,R \qquad (2.2.3a)$$

and

$$P(\hat{\theta}_j|\theta_i) = \{\frac{1 - \beta}{R - 1}\}; \qquad i,j = 1,\ldots,R;\ i \neq j \quad . \qquad (2.2.3b)$$

Also it will be assumed that the various probability density functions are independent of the label if the true categories were known, i.e.

$$f_{\underline{X}|\hat{\theta}_i,\theta_j}(\underline{x}|\hat{\theta}_i,\theta_j) = f_{\underline{X}|\theta_j}(\underline{x}|\theta_j); \qquad i,j = 1,\ldots,R \ . \qquad (2.2.4)$$

Due to the randomness of the labeling scheme of the teacher, characterized by Equations 2.2.3 and 2.2.4, the learning scheme does not know which of the sample patterns are correctly labelled. The only labeling information available to the student is one of $\hat{\theta}_1,\ldots,\hat{\theta}_R$. Hence in order to learn from these incorrectly labelled samples, it is necessary to derive a decision rule in terms the probabilities $P(\hat{\theta}_1),\ldots,P(\hat{\theta}_R)$ and the respective probability density functions $f_{\underline{X}|\hat{\theta}_1},\ldots,f_{\underline{X}|\hat{\theta}_R}$ rather than in terms of $P(\theta_1),\ldots,P(\theta_R)$ and $f_{\underline{X}|\theta_1},\ldots,f_{\underline{X}|\theta_R}$. Unless otherwise mentioned, the loss function used in the analysis in the following sections will be the one described in Equation 2.1.1.

Theorem 2.2.1. With an imperfect teacher characterized by Equations 2.2.3 and 2.2.4, and a loss function specified by Equation 2.1.1, a decision rule using discriminant functions of the form

$$D_{\hat{\theta}_i}(\underline{x}) = P(\hat{\theta}_i)f_{\underline{X}|\hat{\theta}_i}(\underline{x})$$

is equivalent to a Bayes' (optimum) decision rule using discriminant functions of the form

$$D_{\theta_i}(\underline{x}) = P(\theta_i)f_{\underline{X}|\theta_i}(\underline{x})$$

for classifying patterns.

Proof: The above theorem can be proved by showing that

$$D_{\hat{\theta}_i}(\underline{x}) \geq D_{\hat{\theta}_j}(\underline{x}) \text{ if and only if } D_{\theta_i}(\underline{x}) \geq D_{\theta_j}(\underline{x}) \quad ;$$

thus establishing that the two decision rules are equivalent.

Using Bayes' theorem and Equation 2.2.3, the prior probability of $\hat{\theta}_i$ occurring is

$$P(\hat{\theta}_i) = \sum_{k=1}^{R} P(\hat{\theta}_i|\theta_k)P(\theta_k) = P(\theta_i)\beta + \sum_{\substack{k=1 \\ k\neq i}}^{R} P(\theta_k)\left(\frac{1-\beta}{R-1}\right) \quad , \quad (2.2.5)$$

and the probability distribution function of $\underline{X}$ given that $\underline{X}$ is labelled as $\hat{\theta}_i$ is

$$F_{\underline{X}|\hat{\theta}_i}(\underline{x}|\hat{\theta}_i) = \sum_{k=1}^{R} F_{\underline{X}|\hat{\theta}_i,\theta_k}(\underline{x}|\hat{\theta}_i,\theta_k)P(\theta_k|\hat{\theta}_i)$$

Using the condition given in Equation 2.2.4, it follows that

$$F_{\underline{X}|\hat{\theta}_i}(\underline{x}|\hat{\theta}_i) = \sum_{k=1}^{R} F_{\underline{X}|\theta_k}(\underline{x}|\theta_k)P(\theta_k|\hat{\theta}_i) = \sum_{k=1}^{R} F_{\underline{X}|\theta_k}(\underline{x}|\theta_k)\frac{P(\hat{\theta}_i|\theta_k)P(\theta_k)}{P(\hat{\theta}_i)}$$

$$F_{\underline{X}|\hat{\theta}_i}(\underline{x}|\hat{\theta}_i) = \frac{1}{P(\hat{\theta}_i)}\left\{P(\theta_i)\beta F_{\underline{X}|\theta_i}(\underline{x}|\theta_i) + \sum_{\substack{k=1 \\ k\neq i}}^{R} P(\theta_k)\left(\frac{1-\beta}{R-1}\right)F_{\underline{X}|\theta_k}(\underline{x}|\theta_k)\right\} . \quad (2.2.6)$$

Since the existance of density functions is indirectly implied in assumption 2.2.4, the probability density function $f_{\underline{X}|\hat{\theta}_i}(\underline{x}|\hat{\theta}_i)$ can be obtained from Equation 2.2.6 as,

$$f_{\underline{X}|\hat{\theta}_i}(\underline{x}|\hat{\theta}_i) = \frac{1}{P(\hat{\theta}_i)} \{P(\theta_i)\beta f_{\underline{X}|\theta_i}(\underline{x}|\theta_i) + \sum_{\substack{k=1 \\ k\neq i}}^{R} P(\theta_k)(\frac{1-\beta}{R-1}) f_{\underline{X}|\theta_k}(\underline{x}|\theta_k)\} \quad . \tag{2.2.7}$$

Hence

$$D_{\hat{\theta}_i}(\underline{x}) = P(\theta_i)\beta f_{\underline{X}|\theta_i}(\underline{x}|\theta_i) + \sum_{\substack{k=1 \\ k\neq i}}^{R} P(\theta_k)(\frac{1-\beta}{R-1}) f_{\underline{X}|\theta_k}(\underline{x}|\theta_k) \quad .$$

A similar expression can be obtained for $D_{\hat{\theta}_j}(\underline{x})$ and from these two equations it follows that

$$\begin{aligned} D_{\hat{\theta}_i}(\underline{x}) - D_{\hat{\theta}_j}(\underline{x}) &= P(\theta_i)\beta f_{\underline{X}|\theta_i}(\underline{x}|\theta_i) - P(\theta_j)\beta f_{\underline{X}|\theta_j}(\underline{x}|\theta_j) \\ &\quad + P(\theta_j)(\frac{1-\beta}{R-1}) f_{\underline{X}|\theta_j}(\underline{x}|\theta_j) - P(\theta_i)(\frac{1-\beta}{R-1}) f_{\underline{X}|\theta_i}(\underline{x}|\theta_i) \\ &= [\beta - \frac{1-\beta}{R-1}] \{P(\theta_i) f_{\underline{X}|\theta_i}(\underline{x}|\theta_i) - P(\theta_j) f_{\underline{X}|\theta_j}(\underline{x}|\theta_j)\} \\ &= (\frac{\beta R - 1}{R-1}) \{D_{\theta_i}(\underline{x}) - D_{\theta_j}(\underline{x})\} \quad . \end{aligned}$$

Since $\beta > \frac{1}{R}$ by assumption, it follows from the above equation that

$$D_{\hat{\theta}_i}(\underline{x}) > D_{\hat{\theta}_j}(\underline{x}) \longleftrightarrow D_{\theta_i}(\underline{x}) > D_{\theta_j}(\underline{x})$$

and

$$D_{\hat{\theta}_i}(\underline{x}) = D_{\hat{\theta}_j}(\underline{x}) \longleftrightarrow D_{\theta_i}(\underline{x}) = D_{\theta_j}(\underline{x}) \quad . \tag{2.2.8}$$

The right hand side of Equation 2.2.8 defines the decision boundary between the domains of category θ_i and θ_j and it can be seen from the

above equation that $D_{\hat{\theta}_i}(\underline{x}) = D_{\hat{\theta}_j}(\underline{x})$ leads to the same boundary. Hence discriminant functions $D_{\theta_i}(\underline{x})$ and $D_{\hat{\theta}_i}(\underline{x})$ are equivalent.

Based on Theorem 2.2.1, the decision rule for classifying a given pattern $\underline{x}$ with an imperfect teacher is:

(1) compute $D_{\hat{\theta}_i}(\underline{x}) = P(\hat{\theta}_i) f_{\underline{X}|\hat{\theta}_i}(\underline{x}|\hat{\theta}_i)$ for $i = 1,\ldots,R$;

(2) assign $\underline{x}$ to category θ_i if $D_{\hat{\theta}_i}(\underline{x}) > D_{\hat{\theta}_j}(\underline{x})$ $j = 1,\ldots,R$; $j \neq i$.

(2.2.9)

A machine using the above decision rule will classify $\underline{x}$ into the same category as the Bayes' machine.

2.2.2 Special Cases of Theorem 2.2.1, The Two Category Problem. When there are only two categories of patterns θ_1 and θ_2, the imperfect teacher is characterized by

$$P(\hat{\theta}_i|\theta_i) = \beta > \frac{1}{2}\ ; \qquad i = 1,2$$

$$P(\hat{\theta}_j|\theta_i) = 1 - \beta; \qquad i,j = 1,2;\ i \neq j \tag{2.2.10}$$

$$f_{\underline{X}|\hat{\theta}_i,\theta_j}(\underline{x}|\hat{\theta}_i,\theta_j) = f_{\underline{X}|\theta_j}(\underline{x}|\theta_j) \qquad i,j = 1,2 \quad .$$

The second step in decision rule 2.2.9 can now be implemented by evaluating the sign of a single discriminant function

$$D_{\hat{\theta}}(\underline{x}) = P(\hat{\theta}_1) f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1) - P(\hat{\theta}_2) f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2) \tag{2.2.11}$$

$$= [2\beta - 1][P(\theta_1) f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) - P(\theta_2) f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)] \tag{2.2.12}$$

classification can be made according to the rule,

$$\text{assign } \underline{x} \text{ to } \theta_1 \text{ if } D_{\hat{\theta}}(\underline{x}) > 0$$

and

$$\text{assign } \underline{x} \text{ to } \theta_2 \text{ if } D_{\hat{\theta}}(\underline{x}) < 0 \quad . \qquad (2.2.13)$$

The following corollaries can be derived for the two category problem.

Corollary 2.2.1. If $\beta < \frac{1}{2}$, decision rule 2.2.13 still can be used for classifying patterns with minimum risk if $D_{\hat{\theta}}(\underline{x})$ in 2.2.11 is changed to

$$D_{\hat{\theta}}(\underline{x}) = P(\hat{\theta}_2) f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2) - P(\hat{\theta}_1) f_{\underline{X}|\theta_1}(\underline{x}|\hat{\theta}_1) \quad . \qquad (2.2.14)$$

Proof of this corollary follows directly from Equation 2.2.12.

Corollary 2.2.2. If $\beta = \frac{1}{2}$, then no classification will result from decision rule 2.2.13.

From Equation 2.2.12, it can be seen that $D_{\hat{\theta}}(\underline{x}) = 0$ for every $\underline{x}$ if $\beta = \frac{1}{2}$. Hence 2.2.13 does not give any classification.

In fact, when $\beta = \frac{1}{2}$ the probability density functions $f_{\underline{X}|\hat{\theta}_1}$ and functions $f_{\underline{X}|\hat{\theta}_2}$ become

$$f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1) = P(\theta_1) f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) + P(\theta_2) f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)$$

$$f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2) = P(\theta_1) f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) + P(\theta_2) f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) \quad .$$

The probability density function of $\underline{X}$, without any labels, is

$$f_{\underline{X}}(\underline{x}) = P(\theta_1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) + P(\theta_2)f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) \quad .$$

Hence when $\beta = \frac{1}{2}$,

$$f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1) = f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2) = f_{\underline{X}}(\underline{x}) \quad . \qquad (2.2.15)$$

Equation 2.2.15 implies that there is no information available in the labels for discrimination purposes. This is a problem of learning without a teacher and several methods are available for solving this problem (13).

Corollary 2.2.3. If the densities $f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)$ and $f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)$ do not overlap then

$$D_{\hat{\theta}}(\underline{x}) = f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2) \qquad (2.2.16)$$

is equivalent to the discriminant function used by a Bayes' machine.

Proof. The Bayes' machine uses a discriminant function of the form

$$P(\theta_1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) - P(\theta_2)f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) \quad .$$

For non overlapping densities, the above discriminant function is equivalent to

$$D_{\theta}(\underline{x}) = f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) - f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) \quad . \qquad (2.2.17)$$

Since, if $\underline{x} \epsilon \theta_1$, then $f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) = 0$; $f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) > 0$[1] and hence

$$f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) > f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) = 0 \quad .$$

---

[1] The set of $\underline{x}$ where both the densities are zero have a measure zero and hence ignored.

The above equation implies

$$P(\theta_1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) > f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) = 0 \quad ,$$

and hence

$$P(\theta_1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) > P(\theta_2)f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) = 0 \quad .$$

From Equation 2.2.7

$$\begin{aligned} f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2) &= \frac{1}{P(\hat{\theta}_1)} \{\beta P(\theta_1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) \\ &\quad + (1-\beta)P(\theta_2)f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)\} \\ &\quad - \frac{1}{P(\hat{\theta}_2)} \{(1-\beta)P(\theta_1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) \\ &\quad + \beta P(\theta_2)f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)\} \\ &= \frac{1}{P(\hat{\theta}_1)P(\hat{\theta}_2)} \{P(\theta_1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) \\ &\quad [\beta P(\hat{\theta}_2) - (1-\beta)P(\hat{\theta}_1)] \\ &\quad - P(\theta_2)f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) \\ &\quad [\beta P(\hat{\theta}_1) - (1-\beta)P(\hat{\theta}_2)]\} \\ &= \frac{1}{P(\hat{\theta}_1)P(\hat{\theta}_2)} \{P(\theta_1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)[(2\beta-1)P(\theta_2)] \\ &\quad - P(\theta_2)f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)[(2\beta-1)P(\theta_1)]\} \\ &= \frac{P(\theta_1)P(\theta_2)}{P(\hat{\theta}_1)P(\hat{\theta}_2)} (2\beta-1)\{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) - f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)\}. \end{aligned} \tag{2.2.18}$$

Hence

$$D_{\hat{\theta}}(\underline{x}) = \frac{P(\theta_1)P(\theta_2)}{P(\hat{\theta}_1)P(\hat{\theta}_2)} (2\beta - 1)D_{\theta}(\underline{x}) \quad . \qquad (2.2.19)$$

If $\beta > \frac{1}{2}$, Equation 2.2.19 shows that the discriminant functions 2.2.16 and 2.2.17 are equivalent. If $\beta < \frac{1}{2}$, instead of using $D_{\hat{\theta}}(\underline{x})$, $-D_{\hat{\theta}}(\underline{x})$ can be used to classify patterns optimally.

It can be seen from Equation 2.2.16 that $D_{\hat{\theta}}(\underline{x})$ does not involve the exact value of β. The only information the learning scheme needs to know is whether $\beta > \frac{1}{2}$ or $\beta < \frac{1}{2}$.

If the densities overlap, β is involved in the discriminant function $D_{\hat{\theta}}(\underline{x})$ given by Equation 2.2.11 through $P(\hat{\theta}_1)$ and $P(\hat{\theta}_2)$. If $P(\hat{\theta}_1)$ and $P(\hat{\theta}_2)$ were not known, they can be estimated from the number of times $\hat{\theta}_1$ and $\hat{\theta}_2$ occur in the labels (7). However estimation of $P(\theta_1)$ and $P(\theta_2)$, for use in discriminant function of the form

$$D_{\theta}(\underline{x}) = P(\theta_1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) - P(\theta_2)f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) \quad ,$$

from estimates of $P(\hat{\theta}_1)$ and $P(\hat{\theta}_2)$ is not possible without a knowledge of the value of β. The same is true for estimates of $f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)$ and $f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)$ from estimates of $f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)$ and $f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)$. Hence deriving the decision rule in terms of $P(\hat{\theta}_1)$, $P(\hat{\theta}_2)$, $f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)$ and $f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\theta_2)$ has the added advantage that the exact value of β need not be known for learning with an imperfect teacher.

2.3 Learning With an Imperfect Teacher. In the decision rule given in Equation 2.2.9, it was assumed that the prior probabilities

$P(\hat{\theta}_i)$, and probability density functions $f_{\underline{X}|\hat{\theta}_i}(\underline{x}|\hat{\theta}_i)$ were known. However, in practice, the information relevant to $P(\hat{\theta}_i)$ and $f_{\underline{X}|\hat{\theta}_i}(\underline{x}|\hat{\theta}_i)$ for the R categories is only partially known and the unknown information must be learned (estimated). In the remainder of this dissertation it will be assumed that R = 2 and that $P(\hat{\theta}_1)$ and $P(\hat{\theta}_2)$ are known. It will further be assumed that $\beta > \frac{1}{2}$. ($\beta < \frac{1}{2}$ can be taken care of, as explained in Corollary 2.2.1.) The distribution functions will be assumed to be absolutely continuous. No structural form for the density functions $f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)$ and $f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)$ will be assumed.

The densities can be estimated (learned) from a set of incorrectly labelled, independent sample patterns $\underline{X}_1,\ldots,\underline{X}_{n_1}$; $\underline{Y}_1,\ldots,\underline{Y}_{n_2}$. The $\underline{X}_i$'s are sample patterns with labels $\hat{\theta}_1$, the labeling done by an imperfect teacher characterized by Equation 2.2.3, and are identically distributed random vectors with a common probability density function $f_{\underline{X}|\hat{\theta}_1}$. The $\underline{Y}_i$'s are sample patterns with labels $\hat{\theta}_2$, and are identically distributed with a common probability density function $f_{\underline{X}|\hat{\theta}_2}$.

Parzen (16) has proposed and analyzed a class of non-parametric method of estimating univariable density functions. Murthy (17) extended this method to multivariable density functions. The propoerties of the above estimators are discussed in Appendix A. Using estimators of the form proposed by Parzen, an estimate of $f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)$ based on n_1 independent identically distributed sample patterns $\underline{X}_1,\ldots,\underline{X}_{n_1}$ is

$$\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) = \frac{1}{\sqrt{n_1}} \sum_{k=1}^{n_1} \frac{1}{(2\pi)^{\frac{p}{2}}} \operatorname{Exp}\left\{\frac{-[\underline{x} - \underline{X}_k]^T[\underline{x} - \underline{X}_k](n_1)^{\frac{1}{p}}}{2}\right\} \tag{2.3.1}$$

and an estimate of $f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)$ based on n_2 independent identically distributed sample patterns $\underline{Y}_1,\ldots,\underline{Y}_{n_2}$ is

$$\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) = \frac{1}{\sqrt{n_2}} \sum_{k=1}^{n_2} \frac{1}{(2\pi)^{\frac{p}{2}}} \text{Exp}\{\frac{-[\underline{x} - \underline{Y}_i]^T[\underline{x} - \underline{Y}_i](n_2)^{\frac{1}{p}}}{2}\} \quad . \tag{2.3.2}$$

In Equations 2.3.1 and 2.3.2 p is the dimension of the pattern space. Using the estimators $\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1)$ and $\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2)$ for $f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)$ and $f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)$ in 2.2.11 and 2.2.13, the procedure for learning to recognize patterns with an imperfect teacher is:

(1) Using the incorrectly identified sample patterns,

$\underline{X}_1,\ldots,\underline{X}_{n_1}$; $\underline{Y}_1,\ldots,\underline{Y}_{n_2}$, estimate $f_{\underline{X}|\hat{\theta}_1}$ and $f_{\underline{X}|\hat{\theta}_2}$;

(2) Using estimators $\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1)$ and $\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2)$, compute

$$\hat{D}_{\hat{\theta}}(\underline{x}) = P(\hat{\theta}_1)\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - P(\hat{\theta}_2)\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) \quad ; \tag{2.3.3}$$

(3) Assign $\underline{x}$ to

$$\theta_1 \text{ if } \hat{D}_{\hat{\theta}}(\underline{x}) > 0 \quad ,$$

and

$$\theta_2 \text{ if } \hat{D}_{\hat{\theta}}(\underline{x}) < 0 \quad . \tag{2.3.4}$$

2.4 Asymptotic Performance of the Learning Scheme. Using the consistency properties of the estimators $\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}$ and $\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}$, the asymptotic performance of the learning scheme proposed in the previous

paragraph can be analyzed.

Theorem 2.4.1. The estimate of the discriminant function

$$\hat{D}_{\hat{\theta}}(\underline{x}) = P(\hat{\theta}_1)\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - P(\hat{\theta}_2)\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2)$$

converges to

$$D_{\hat{\theta}}(\underline{x}) = P(\hat{\theta}_1)f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1) - P(\hat{\theta}_2)f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)$$

with probability one as $n_1, n_2 \longrightarrow \infty$.

Proof. From Equations A.2.8 and A.2.9, for every $\epsilon > 0$

$$\lim_{n_1 \to \infty} P\left\{\left| P(\hat{\theta}_1)[\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)]\right| < \frac{\epsilon}{2}\right\} = 1$$

$$\lim_{n_2 \to \infty} P\left\{\left| P(\hat{\theta}_2)[\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) - f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)]\right| < \frac{\epsilon}{2}\right\} = 1 \quad . \tag{2.4.1}$$

Also, if

$$\left| P(\hat{\theta}_1)[\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)]\right| < \frac{\epsilon}{2}$$

and

$$\left| P(\hat{\theta}_2)[\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) - f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)]\right| < \frac{\epsilon}{2} \quad , \tag{2.4.2}$$

then

$$\left|P(\hat{\theta}_1)[\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)] - P(\hat{\theta}_2)[\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) - f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)]\right| < \epsilon. \tag{2.4.3}$$

Since the estimators $\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1)$, $\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2)$ are independent, from (2.4.2) and (2.4.3),

$$P\{\Big| P(\hat{\theta}_1)[\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)]$$

$$- P(\hat{\theta}_2)[\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) - f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)]\Big| < \epsilon\}$$

$$\geq \{\Big| P(\hat{\theta}_1)[\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)]\Big| < \frac{\epsilon}{2}\}$$

$$P\{\Big| P(\hat{\theta}_2)[\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) - f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)]\Big| < \frac{\epsilon}{2}\}$$

As $n_1, n_2 \longrightarrow \infty$ each term on the right hand side of the above equation is equal to 1, and hence

$$P\{\Big| [P(\hat{\theta}_1)\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - P(\hat{\theta}_2)\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2)]$$

$$- [P(\hat{\theta}_1)f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1) - P(\hat{\theta}_2)f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)]\Big| < \epsilon\} = 1 \quad . \quad (2.4.4)$$

Equation 2.4.4. implies that

$$P(\hat{\theta}_1)\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - P(\hat{\theta}_2)\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}\ \hat{\theta}_2)$$

converges to

$$P(\hat{\theta}_1)f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1) - P(\hat{\theta}_2)f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)$$

with probability 1.

The result of Theorem 2.4.1 can be used to evaluate the asymptotic risk associated with learning with an imperfect teacher. Convergence

of $\hat{D}_{\hat{\theta}}(\underline{x}) = P(\hat{\theta}_1)\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - P(\hat{\theta}_2)\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2)$ to $D_{\hat{\theta}}(\underline{x})$ implies that, with probability one, $\underline{x}$ will be classified into the same category by decision rule 2.3.4 and 2.2.13. It has been shown in Theorem 2.2.1 that the decision rule 2.2.13 is equivalent to the decision rule of a Bayes' (optimum) machine. Therefore, with probability one, the learning scheme described in 2.3.4 classifies $\underline{x}$ into the same category as a Bayes' machine. Hence the conditional risk $r_s(\underline{x};n_1,n_2)$ associated with classifying $\underline{x}$ according to 2.3.4, converges to the Bayes' conditional risk, $r^*(\underline{x})$, with probability one, i.e.

$$r_s(\underline{x};n_1,n_2) \rightarrow r^*(\underline{x}) \text{ as } n_1,n_2 \rightarrow \infty \text{ with probability one.} \quad (2.4.5)$$

For a symmetrical loss function of the form given in 2.1.1 the Bayes' conditional risk is given by

$$r^*(\underline{x}) = \min\{P(\theta_1|\underline{X}),\ P(\theta_2|\underline{X})\} \quad . \quad (2.4.6)$$

As a consequence of 2.4.5,

$$E\{r_s(\underline{x};n_1,n_2)\} = r^*(\underline{x}) \text{ as } n_1,n_2 \rightarrow \infty \quad .$$

Taking the average on both sides with respect to $f_{\underline{X}}(\underline{x})$, the average risk associated with learning with an imperfect teacher is

$$R_s = \int r^*(\underline{x}) f_{\underline{X}}(\underline{x}) d\underline{x}$$

$$= P(\theta_1) \int_{D_2} f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) d\underline{x} + P(\theta_2) \int_{D_1} f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) d\underline{x} \quad (2.4.7a)$$

where

$$D_1 = \{\underline{x} : P(\theta_1) f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) > P(\theta_2) f_{\underline{X}|\theta_2}(\underline{x}|\theta_2), \text{ and}$$

$$D_2 = \{\underline{x} : P(\theta_2) f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) > P(\theta_1) f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) \quad .$$

The right hand side of Equation 2.4.7a is the Bayes' risk and hence

$$R_s = R^* \quad . \tag{2.4.7b}$$

Equation 2.4.7b states that the average asymptotic risk for learning with an imperfect teacher is equal to the Bayes', minimum, risk.

2.4.1 Comparison of Performance of the Learning Scheme With the Imperfect Teacher and the Nearest Neighbor Rule (NNR).

2.4.1a Non Overlapping Densities. If the conditional densities $f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)$ and $f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)$ do not overlap then the Bayes' risks, R^*, is

$$R^* = 0 \quad . \tag{2.4.8}$$

From Equation 2.4.7, the average asymptotic risk for the learning scheme is

$$R_s = R^* = 0 \quad . \tag{2.4.9}$$

From Whitney (15), the average asymptotic risk for the nearest neighbor rule R_n is given by

$$R_n = 1 - \beta \quad . \tag{2.4.10}$$

The average risk for the imperfect teacher is

$$R_t = 1 - \beta \quad . \tag{2.4.11}$$

From Equations 2.4.8, 2.4.9, 2.4.10, and 2.4.11, it can be seen that, on the average, the proposed learning scheme is better than the imperfect teacher and the nearest neighbor rule.

2.4.1b Overlapping Densities. If the conditional densities overlap, then the learning scheme better the performance of the imperfect teacher only if

$$R^* < (1 - \beta) \quad . \tag{2.4.12}$$

If $R^* > 1 - \beta^2$ then the optimum Bayes' scheme itself has a greater average risk than the imperfect teacher and hence the learning scheme whose asymptotic average risk is equal to R^* cannot be expected to do better than the imperfect teacher.

However, for the nearest neighbor rule Whitney (15) has shown that the average asymptotic risk is bounded by

$$R_n \geq (1 - \beta) + (2\beta - 1)R^* \quad . \tag{2.4.13}$$

From Equation 2.4.13 it can be seen that if R^* is less than $\frac{1}{2}$ then

$$R_n \geq R^* = R_s \quad .$$

Hence the learning scheme is better than the nearest neighbor rule if R^* is less than $\frac{1}{2}$.

2.5 Finite Sample Performance of the Learning Scheme. In the previous sections the asymptotic performance of the proposed learning scheme was analyzed and it was shown that if the density functions do not overlap then, on the average, the learning scheme performs better than the teacher. In this section it will be shown that the learning scheme performs better than the teacher, on the average, after being

---

[2] In order for the teacher to be this good, he must have extra information other than a complete knowledge of the density functions.

presented with a finite number of sample patterns $\underline{X}_1,\ldots,\underline{X}_{n_1}$; $\underline{Y}_1,\ldots,\underline{Y}_{n_2}$.

From Corollary 2.2.3, the learning scheme for classifying patterns when the density functions do not overlap is:

$$\text{Assign } \underline{x} \text{ to } \theta_1 \text{ if } \hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) > \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) \tag{2.5.1}$$

$$\text{Assign } \underline{x} \text{ to } \theta_2 \text{ if } \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) > \hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) \quad . \tag{2.5.2}$$

A given pattern $\underline{x}$ from category θ_1 is therefore classified correctly if (2.5.1) is satisfied and will be incorrectly classified if (2.5.2) holds. Assigning a value of +1 for correct classification and 0 for incorrect classification, the gain associated with classifying a pattern $\underline{x}$ form a category θ_1 is

$$g_s(\underline{x}|\theta_1;n_1,n_2) = P\{\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) > \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) \,\big|\, \underline{x}\epsilon\theta_1\} \tag{2.5.3}$$

Theorem 2.5.1

$$P\{\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) > \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) \,\big|\, \underline{x}\epsilon\theta_1\} \geq L(\beta,n_1,n_2,\underline{x})$$

where

$$L(\beta,n_1,n,\underline{x}) = \left[1 - \frac{C_1\beta}{\sqrt{n_1}\, f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}\right]\left[1 - \frac{C_2(1-\beta)}{\sqrt{n_2}\, f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}\right] \tag{2.5.4a}$$

and

$$C_1 = \frac{P(\theta_1)}{P(\hat{\theta}_1)}\left[\frac{P(\hat{\theta}_1)P(\hat{\theta}_2)}{P(\theta_1)P(\theta_2)(2\beta-1)}\right]^2 \frac{4}{(2\sqrt{\pi})^p}$$

$$C_2 = \frac{P(\theta_1)}{P(\hat{\theta}_2)}\left[\frac{P(\hat{\theta}_1)P(\hat{\theta}_2)}{P(\theta_1)P(\theta_2)(2\beta-1)}\right]^2 \frac{4}{(2\sqrt{\pi})^p} \quad . \tag{2.5.4b}$$

Proof. Let

$$D = f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2) > 0$$

and let

$$D_1 = \{\underline{x}: f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) > 0\} \quad .$$

$$P\{\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) > \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) \Big| \underline{x}\epsilon\theta_1\}$$

$$\geq P\{\Big| \hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)\Big| < \frac{D}{2} \cap$$

$$\Big|\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) - f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)\Big| < \frac{D}{2} \Big| \underline{x}\epsilon\theta_1\}.$$

Since the estimators $\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}$ and $\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}$ are independent, the right hand side of the inequality becomes

$$\geq P\{\Big|\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)\Big| < \frac{D}{2} \Big| \underline{x}\epsilon\theta_1\}$$

$$P\{\Big|\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) - f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)\Big| < \frac{D}{2} \Big| \underline{x}\epsilon\theta_1\}$$

$$\geq P\{\Big|\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)\Big|^2 < \frac{D^2}{4} \Big| \underline{x}\epsilon\theta_1\}$$

$$P\{\Big|\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) - f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)\Big|^2 < \frac{D^2}{4} \Big| \underline{x}\epsilon\theta_1\} \quad . \quad (2.5.6)$$

Let us consider

$$P\{\left|\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)\right|^2 < \frac{D^2}{4}\}$$

$$= 1 - P\{\left|\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)\right|^2 \geq \frac{D^2}{4}\}$$

$$\geq 1 - \frac{E\{\left|\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)\right|^2\}}{\frac{D^2}{4}} \qquad (2.5.7)$$

by Chebyshev's inequality. From Equation A.2.9,

$$E\{\left|\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)\right|^2\} \simeq \frac{1}{\sqrt{n_1}\ (2\sqrt{\pi})^p} f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1). \qquad (2.5.8)$$

Substituting (2.5.8) in (2.5.7), the right hand side of (2.5.7) becomes

$$\tilde{\geq} \left[1 - \frac{f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)}{\sqrt{n_1}\ (2\sqrt{\pi})^p} \cdot \frac{4}{D^2}\right] . \qquad (2.5.9)$$

Similarly

$$P\{\left|f_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) - f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)\right|^2 < \frac{D^2}{4}\} \geq \left[1 - \frac{f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)}{\sqrt{n_2}\ (2\sqrt{\pi})^p} \cdot \frac{4}{D^2}\right] . \qquad (2.5.10)$$

If $\underline{x}\epsilon\theta_1$ then $\underline{x}\epsilon D_1$, and on D_1

$$f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1) = \beta \frac{P(\theta_1)}{P(\hat{\theta}_1)} f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)$$

$$f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2) = (1 - \beta) \frac{P(\theta_1)}{P(\hat{\theta}_2)} f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)$$

and

$$D = \frac{P(\theta_1)P(\theta_2)}{P(\hat{\theta}_1)P(\hat{\theta}_2)} (2\beta - 1) f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) \quad . \qquad (2.5.11)$$

Substituting (2.5.9), (2.5.10) and (2.5.11) in (2.5.6), one obtains

$$P\{\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) > \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) \Big| \underline{x}\epsilon\theta_1\} \gtrsim$$

$$[1 - \frac{C_1\beta}{\sqrt{n_1}\, f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}][1 - \frac{C_2(1 - \beta)}{\sqrt{n_2}\, f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}]$$

where

$$C_1 = \frac{P(\theta_1)}{P(\hat{\theta}_1)} [\frac{P(\hat{\theta}_1)P(\hat{\theta}_2)}{P(\theta_1)P(\theta_2)(2\beta - 1)}]^2 \frac{4}{(2\sqrt{\pi})^p}$$

$$C_2 = \frac{P(\theta_1)}{P(\hat{\theta}_2)} [\frac{P(\hat{\theta}_1)P(\hat{\theta}_2)}{P(\theta_1)P(\theta_2)(2\beta - 1)}]^2 \frac{4}{(2\sqrt{\pi})^p}$$

and hence the proof of the theorem.

Substituting the results of Theorem 2.5.1 in Equation 2.5.3, the gain of the learning system associated with classifying a pattern $\underline{x}$ from category θ_1 becomes

$$g_s(\underline{x}\ \theta_1; n_1, n_2) \gtrsim [1 - \frac{C_1\beta}{\sqrt{n_1}\, f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}][1 - \frac{C_2(1 - \beta)}{\sqrt{n_2}\, f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)}] \quad .$$

A similar expression can be derived fro the gain associated with classifying a sample $\underline{x}$ from category θ_2. The gain of the teacher for classifying $\underline{x}$ from category θ_1 is

$$g_t(\underline{x}|\theta_1) = \beta \quad .$$

By setting

$$\left[1 - \frac{C_1\beta}{\sqrt{n_1}\, f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}\right]\left[1 - \frac{C_2(1-\beta)}{\sqrt{n_2}\, f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}\right] > \beta \quad ,$$

one can solve for n_1 and n_2, the sample size required by the learning scheme to better the performance of the teacher. Hence the justification for the claim that, in the case of non-overlapping densities, the learning scheme performs better than the teacher on the average after looking at a finite number of sample patterns. As an example, if $f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) = 1$, and equal prior probabilities $P(\theta_1)$ and $P(\theta_2)$, then on the average the learning scheme with a sample size $n_1, n_2 = 150$ will better an imperfect teacher with $\beta = 0.9$.

The sample size required by the learning scheme to better the performance of the teacher is given below in Table I. The densities used in these sample calculations are assumed to be uniformly distributed over non-overlapping intervals of unit length, with $P(\theta_1) = P(\theta_2) = \frac{1}{2}$.

TABLE I

SAMPLE SIZE REQUIRED BY THE LEARNING SCHEME TO PERFORM BETTER THAN THE TEACHER

β	0.60	0.70	0.80	0.90	0.95
Approximate Sample Size $n_1 + n_2$	5000	1800	500	300	1600

A rather surprising inference that can be derived from this example is that the learning scheme requires less samples to better the performance of a mediocre teacher than the number of samples it requires to better either a very bad or a very good teacher, i.e. it is easier to better a mediocre teacher.

In the following lemma, the dependency of the performance of the learning scheme on β is investigated.

Lemma 2.5.1. For sufficiently large sample size, $L(\beta,\underline{x},n_1,n_2)$ given in Equation 2.5.4 increases as β increases.

Proof. $L(\beta,\underline{x},n_1,n_2)$ can be written as

$$= [1 - L_1(\beta)][1 - L_2(\beta)]$$

where

$$L_1(\beta) = \frac{1}{\sqrt{n_1}} \frac{4}{(2\sqrt{\pi})^p} \left(\frac{P(\hat{\theta}_1)P(\hat{\theta}_2)}{P(\theta_1)P(\theta_2)(2\beta - 1)}\right)^2 \frac{P(\theta_1)\beta}{P(\hat{\theta}_1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}$$

$$L_2(\beta) = \frac{1}{\sqrt{n_2}} \frac{4}{(2\sqrt{\pi})^p} \left(\frac{P(\hat{\theta}_1)P(\hat{\theta}_2)}{P(\theta_1)P(\theta_2)(2\beta - 1)}\right)^2 \frac{P(\theta_1)(1 - \beta)}{P(\hat{\theta}_2)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}$$

$$\frac{d}{d\beta} L(\beta,\underline{x},n_1,n_2) = [1 - L_1(\beta)]\left[-\frac{dL_2(\beta)}{d\beta}\right] + [1 - L_2(\beta)]\left[-\frac{dL_1(\beta)}{d\beta}\right] \tag{2.5.12}$$

When n_1 and n_2 are large $[1 - L_1(\beta)]$ and $[1 - L_2(\beta)]$ are greater than zero; then if $\frac{dL_2(\beta)}{d\beta}$ and $\frac{dL_1(\beta)}{d\beta}$ are shown to be negative, the lemma is proved.

$$\frac{dL_1(\beta)}{d\beta} = a \frac{d}{d\beta} \left(\frac{P(\theta_1)P(\theta_2)^2}{(2\beta - 1)^2}\beta\right)$$ where a is a constant not involving β

$$\frac{1}{a}\frac{dL_1(\beta)}{d\beta} = P(\hat{\theta}_1)P(\hat{\theta}_2)^2\frac{(2\beta-1)[-2\beta-1]}{(2\beta-1)^4} + \frac{\beta}{(2\beta-1)^2}\frac{d}{d\beta}[P(\hat{\theta}_1)P(\hat{\theta}_2)^2]$$

$$= -\frac{P(\hat{\theta}_1)P(\hat{\theta}_2)^2(2\beta+1)}{(2\beta-1)^3} + \frac{\beta}{(2\beta-1)^2}\{P(\hat{\theta}_1)\frac{d}{d\beta}[P(\hat{\theta}_1)P(\hat{\theta}_2)]$$

$$+ P(\hat{\theta}_1)P(\hat{\theta}_2)\frac{d}{d\beta}[P(\hat{\theta}_1)]\}$$

$$= -\frac{P(\hat{\theta}_1)P(\hat{\theta}_2)^2(2\beta+1)}{(2\beta-1)^3} + \frac{\beta}{(2\beta-1)^2}\{P(\hat{\theta}_1)P(\hat{\theta}_2)[P(\theta_2)-P(\theta_1)]$$

$$+ P(\hat{\theta}_2)(1-2\beta)[P(\theta_1)-P(\theta_2)]^2\}$$

$$= -\frac{P(\hat{\theta}_1)P(\hat{\theta}_2)^2(2\beta+1)}{(2\beta-1)^3} + \frac{\beta}{(2\beta-1)^2}\{P(\hat{\theta}_1)P(\hat{\theta}_2)[\frac{P(\hat{\theta}_1)-P(\hat{\theta}_2)}{1-2\beta}]$$

$$+ P(\hat{\theta}_2)(1-2\beta)\frac{[P(\hat{\theta}_1)-P(\hat{\theta}_2)]^2}{(2\beta-1)^2}\}$$

$$= -\frac{P(\hat{\theta}_1)P(\hat{\theta}_2)^2(2\beta+1)}{(2\beta-1)^3} - \frac{\beta}{(2\beta-1)^3}P(\hat{\theta}_1)P(\hat{\theta}_2)[P(\hat{\theta}_1)-P(\hat{\theta}_2)]$$

$$- \frac{\beta}{(2\beta-1)^3}P(\hat{\theta}_2)[P(\hat{\theta}_1)-P(\hat{\theta}_2)]^2$$

$$= -\frac{P(\hat{\theta}_1)P(\hat{\theta}_2)}{(2\beta-1)^3}[(2\beta+1)P(\hat{\theta}_2)+\beta[P(\hat{\theta}_1)-P(\hat{\theta}_2)]] - \frac{\beta}{(2\beta-1)^3}$$

$$P(\hat{\theta}_2)[P(\hat{\theta}_1)-P(\hat{\theta}_2)]^2$$

$$= -\frac{P(\hat{\theta}_1)P(\hat{\theta}_2)}{(2\beta-1)^3}[(\beta+1)P(\hat{\theta}_2)+\beta P(\hat{\theta}_2)] - \frac{\beta}{(2\beta-1)^3}P(\hat{\theta}_2)[P(\hat{\theta}_1)-P(\hat{\theta}_2)]^2.$$

Hence

$$\frac{dL_1(\beta)}{d\beta} < 0 \quad .$$

Similarly it can be shown that

$$\frac{dL_2(\beta)}{d\beta} < 0 \quad .$$

Hence

$$\frac{d}{d\beta} \{L(\beta, \underline{x}, n_1, n_2)\} < 0$$

from Equation 2.5.12.

A similar result can be derived for a sample $\underline{x}$ from category θ_2. Lemma 2.5.1 implies that for a given sample size, a learning scheme with a better teacher acquires more knowledge than a scheme with a comparitively poor teacher.

The results derived in Sections 2.4 and 2.5 have been verified through simulations on the computer. The simulation results are discussed in the next section.

2.6 Simulations. The proposed learning scheme was simulated on the IBM-360 computer for both overlapping and non-overlapping density functions. The various density functions used in the simulations are shown in Figures 1 and 2. The prior probabilities for the categories were set equal to $\frac{1}{2}$. Samples were drawn from the two categories θ_1 and θ_2 and were labelled as $\hat{\theta}_1$ or $\hat{\theta}_2$ according to

$$P(\hat{\theta}_i | \theta_i) = \beta > \frac{1}{2} ; \qquad i = 1,2$$

$$P(\hat{\theta}_j | \theta_i) = 1 - \beta; \qquad i,j = 1,2;\ i \neq j$$

Using the incorrectly labelled samples, the densities $\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1)$ and $\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2)$ were estimated according to (2.3.1) and (2.3.2).

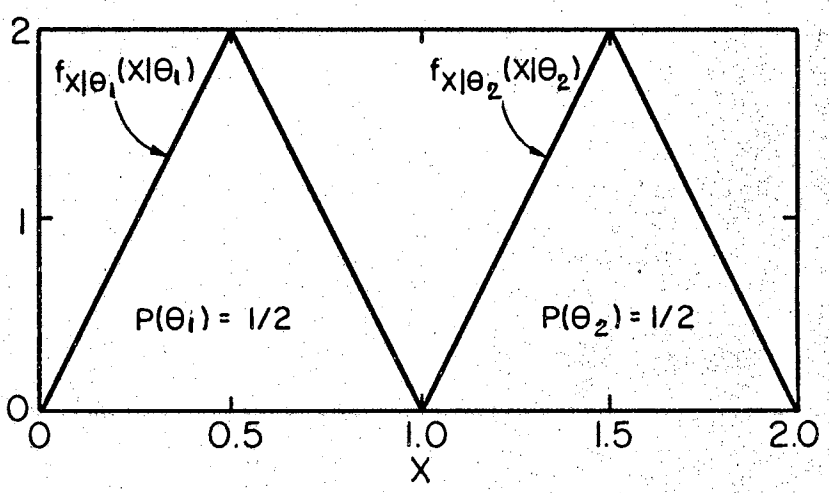

Figure 1. Non-overlapping Densities Used in Simulation

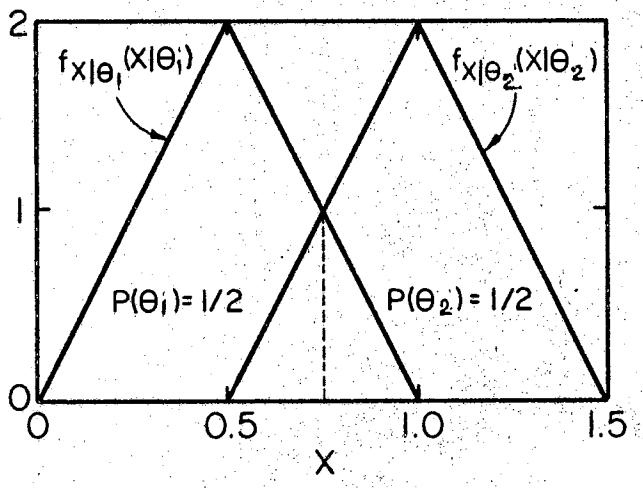

Figure 2. Overlapping Densities Used in Simulations

Fifty additional test samples were drawn from the two categories and the learning scheme was asked to classify these samples into θ_1 and θ_2 according to the decision rule:

classify the sample $\underline{x}$ as coming from θ_1 if

$$\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) > \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2)$$

classify the sample $\underline{x}$ as coming from θ_2 if

$$\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) > \hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) \quad .$$

The risk for the learning scheme was calculated based on the classification of fifty test samples, the loss function being +1 for incorrect classification and 0 for correct classification. For each value of β, ten runs were made with 75 and 100 training samples ($n_1 + n_2 = 75, 100$) and the average risk for the learning scheme was calculated. The results of the simulations are shown in Figures 3 and 4.

Figure 3 shows the plot of average risk versus β for the learning scheme for non-overlapping densities shown in Figure 1. The Bayes' risk R^* for non-overlapping densities is zero and the average risk for the imperfect teacher is $(1 - \beta)$. Figure 4 shows the same plot for overlapping densities shown in Figure 2. The Bayes' risk now is 0.125 and the average risk for the teacher is $(1 - \beta)$.

From Figures 3 and 4 the following theoretical results can be verified:

(1) The average asymptotic risk for the learning scheme converges to the Bayes' risk (as derived in Section 2.4);

(2) For non-overlapping densities the learning scheme betters the

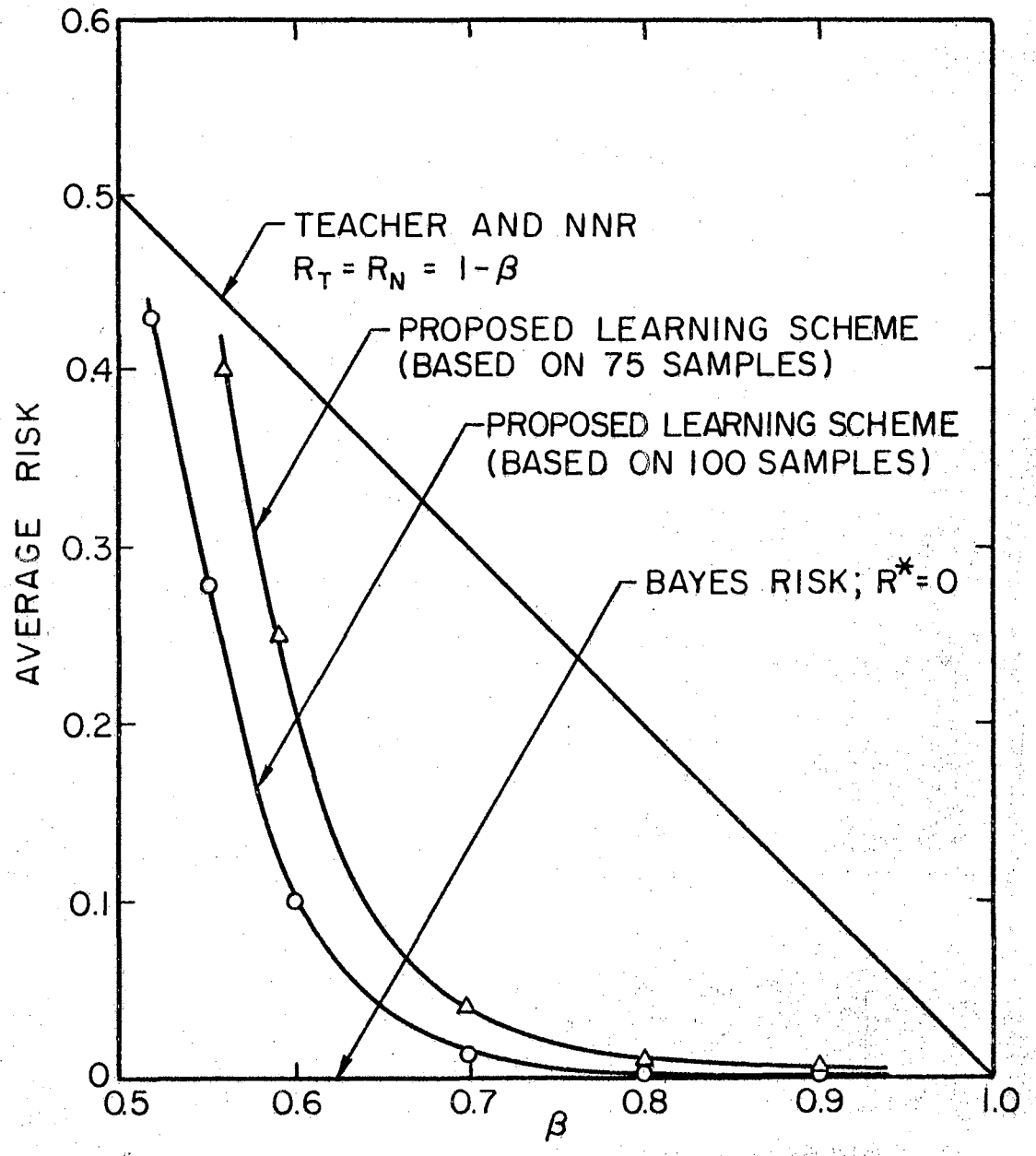

Figure 3. Average Risk vs β for Non-overlapping Densities

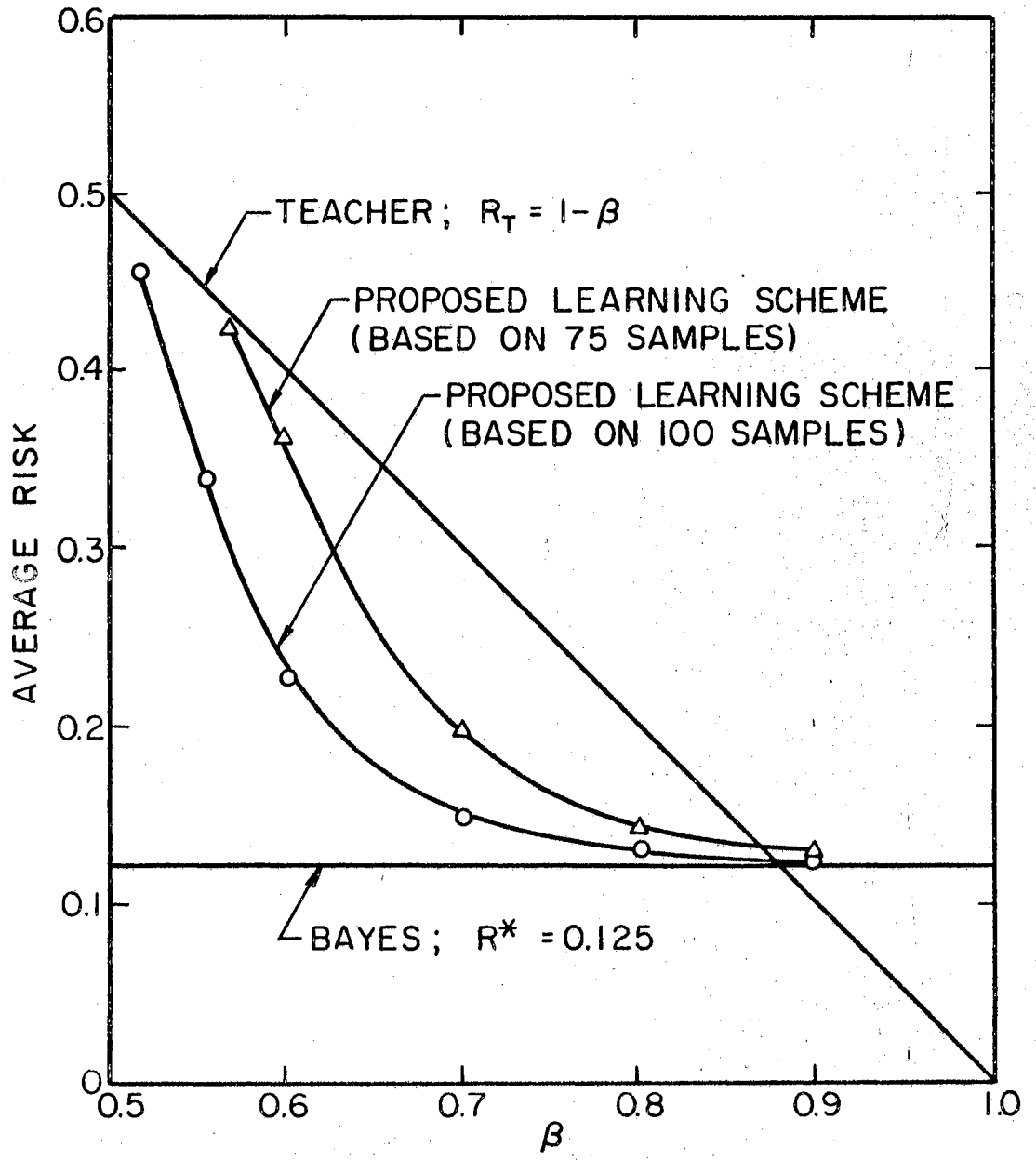

Figure 4. Average Risk vs β for Overlapping Densities

imperfect teacher, and the NNR after looking at a finite number of sample patterns (as derived in Section 2.5).

(3) For overlapping densities, the learning scheme is better than the teacher if $R^* < 1 - \beta$ (derived in Section 2.4). In Figure 4, this corresponds to $\beta < 0.875$.

(4) For a given number of training samples, the average risk decreases as β increases (as derived in Section 2.5).

It must be pointed out here that the plots represent only the "average" performance of the learning system. On an individual run, the performance of the learning scheme will depend on the number of correctly labelled samples. If a particular sequence of sample patterns had too many incorrect labels, then the performance of the learning scheme will be worse than the "average" performance. To illustrate this point, a summary of the performance of the learning scheme on individual runs is given in Table II for the non-overlapping densities shown in Figure 1.

TABLE II

SUMMARY OF PERFORMANCE[1]

Run # / β	1	2	3	4	5	6	7	8	9	10	Average Number of Errors
0.60	9	2	0	28	4	14	0	33	2	29	12.1
0.70	0	0	0	14	2	1	0	0	0	3	2
0.8	0	0	0	1	1	1	0	0	0	0	0.3

Densities: As in Figure 1

Sample Size: 75

Test Samples: 50

[1]Entries in the table denote the number of errors made by the learning scheme in classifying the test samples.

CHAPTER III

FEEDBACK LEARNING SCHEMES

3.1 Introduction. This chapter is concerned with investigating the possibility of "feedback" as a means of improving the performance of learning schemes with an imperfect teacher. The term "feedback learning scheme" is used here in conjunction with learning schemes which, instead of simply accepting the incorrectly labelled sample patterns provided by the imperfect teacher, attempt to question and possibly correct the labelling on some of these sample patterns. The questioning and relabelling at a particular stage of learning is based on the "knowledge" acquired by the learning scheme up to that stage. Since the learning scheme uses its own knowledge in an attempt to improve its performance the term "feedback learning" was thought to be appropriate.

The term "feedback" is not used here in the usual sense. However, the student detects an error in a probabilistic sense and initiates correction. This is analogous to feedback in control systems where an error detected in a deterministic sense leads to correction. Thus feedback broadly applies. Nevertheless, the reader may prefer other terms. For example, since the student is continuously changing his learning procedure as his knowledge increases the term adaptive learning or adaptive editing can be used. The term data refinement can also be used to describe this so called feedback.

For analysis purposes it will be assumed that there are only two

categories of patterns θ_1 and θ_2 with non-overlapping density functions $f_{\underline{X}|\theta_1}$ and $f_{\underline{X}|\theta_2}$. Equal prior probabilities, $P(\theta_1) = P(\theta_2) = \frac{1}{2}$, will be assumed. The case of unequal prior probabilities will be discussed at the end of Chapter IV. It will also be assumed that β is greater than $\frac{1}{2}$. If β is less than $\frac{1}{2}$, this can be taken care of as described in Corollary 2.2.1. Under assumptions just stated, justification for considering feedback as a means of improving the performance of learning schemes will be given. Several possible feedback schemes will be discussed and it will be shown that a thresholded feedback has many desireable properties over other schemes.

3.2 Justification for Feedback. Before going into theoretical justification for considering feedback, a rather philosophical motivation will be given based on an example that is of common occurrence in classrooms. Such an example is the attempt of a student (presumably with much less knowledge than his teacher) to question and possibly correct an inadvertent error made by his teacher. In spite of the fact that most of the student's knowledge is derived from his teacher he is still able to use this knowledge to occasionally correct his teacher. Even though correcting trivial errors do not necessarily mean an increase in knowledge, most of the student's learning takes place through questioning what is being said in an intelligent way. Hence it seems that such questioning and possibly correcting the labelling information supplied by the teacher in pattern recognition problems is all too relevant especially if the teacher is known to be imperfect.

Theoretical justification for considering the possibility of feedback in learning schemes is based on the results described in Chapter II. It was shown in Chapter II that the learning scheme performs better

than the teacher in the asymptotic case. It was also shown that the learning scheme performs better than the imperfect teacher after looking at a finite number of samples and that the finite sample performance of the learning scheme improves as β increases. In Appendix B, it is shown that the average rate of learning at initial stages increases as β increases. These results indicate that any improvement in β will result in an improvement in the performance of the learning scheme. Since the learning scheme performs better than the teacher after looking at a finite number of samples the student (the terms learning scheme and student are used in the same context) can verify the labelling given by the teacher based on his knowledge and correct some of the labels. Such correction if done successfully will result in a lesser number of incorrect labels in the sample patterns and hence lead to better performance.

A computer simulation will now be discussed to illustrate the effectiveness of this correction. An initial set of 75 sample patterns were drawn from the densities shown in Figure 1, with $P(\theta_1) = P(\theta_2) = \frac{1}{2}$. An imperfect teacher characterized by $\beta = 0.7$ labelled these samples as $\hat{\theta}_1$ and $\hat{\theta}_2$. Using these incorrectly labelled samples the student learned the densities $f_{\underline{X}|\hat{\theta}_1}$ and $f_{\underline{X}|\hat{\theta}_2}$. Fifty additional sample patterns were drawn and these samples were labelled by the imperfect teacher and the student. The student ignored the teacher's labelling and did his own labelling according to:

$$\begin{aligned} &\text{Label } \underline{x} \text{ as } \hat{\theta}_1 \text{ if } \hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) > \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) \\ &\text{Label } \underline{x} \text{ as } \hat{\theta}_2 \text{ if } \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) > \hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) \end{aligned} \quad . \tag{3.2.1}$$

The form of the estimators $\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}$ and $\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}$ is given in Section 2.3. A summary of the labelling for the teacher and the student is given in Table III. It can be seen from Table III that the performance of the student will be better if the additional samples were used with the labelling provided by the student himself since the student's labelling contains fewer incorrect labels than the teacher. However, it must be mentioned here that the student's labelling does not really improve β, but improves the ratio of the number of samples with correct labels to the total number of samples.

Example 3.2.1.

$$\beta = 0.70;\ P(\theta_1) = P(\theta_2) = \frac{1}{2}$$

$f_{X|\theta_1}$ and $f_{X|\theta_2}$ are same as those shown in Figure 1

$$n_1 + n_2 = 75$$

Total no. of test samples = 50

$$\beta_E = \frac{\text{Number of samples correctly labelled}}{\text{Total number of samples labelled}}$$

TABLE III

SUMMARY OF LABELLING: STUDENT VS TEACHER

Run No.	Teacher			Student		
	Number of Correct Labels	Number of Incorrect Labels	β_E $[E(\beta_E)=\beta]$	Number of Correct Labels	Number of Incorrect Labels	β_E
1	35	15	0.70	50	0	1.0
2	39	11	0.78	50	0	1.0
3	36	14	0.72	50	0	1.0
4	37	13	0.74	36	14	0.72
5	27	23	0.54	48	2	0.96
6	34	16	0.68	49	1	0.98
7	37	13	0.74	50	0	1.0
8	37	13	0.74	50	0	1.0
9	34	16	0.68	50	0	1.0
10	30	20	0.60	47	3	0.94
Ave	34.6	15.4	0.692	48	2	0.96

3.3 Types of Feedback. It has been shown through a simulation in the previous section that the learning scheme can use its own knowledge through feedback to correct the labelling provided by the teacher. Three different types of feedback will be considered in this section and their merits and demerits will be discussed.

The first feedback scheme to be considered uses a combination of the labelling provided by the teacher and the student. When the $(n + 1)^{st}$ sample $\underline{X}_{n+1}(n=n_1+n_2)$ is presented, the student provides his

own label for $\underline{X}_{n+1}$ based on his present knowledge. The label on $\underline{X}_{n+1}$ probided by the teacher is checked for agreement and the sample $\underline{X}_{n+1}$ is used by the student to update his knowledge if there is agreement. Otherwise $\underline{X}_{n+1}$ is discarded. The alogarithm for this feedback scheme is:

$$\begin{aligned}
&\text{Label } \underline{X}_{n+1} \text{ as } \hat{\theta}_1 \text{ and update } \hat{f}_{\underline{X}|\hat{\theta}_1;n_1} \text{ if } \hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{X}_{n+1}|\hat{\theta}_1) > \\
&\qquad \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{X}_{n+1}|\hat{\theta}_2) \text{ and if teacher says } \theta_1 \quad , \\
&\text{Label } \underline{X}_{n+1} \text{ as } \hat{\theta}_2 \text{ and update } \hat{f}_{\underline{X}|\hat{\theta}_2;n_2} \text{ if } \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{X}_{n+1}|\hat{\theta}_2) > \\
&\qquad \hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{X}_{n+1}|\hat{\theta}_1) \text{ and if teacher says } \theta_2 \quad , \\
&\text{Otherwise discard } \underline{X}_{n+1}.
\end{aligned} \tag{3.3.1}$$

Let us now look at N additional samples processed by the student according to (3.3.1). Out of these N samples the teacher will on the average label $N\beta$ samples correctly and $N(1 - \beta)$ samples incorrectly. Denoting the probability of correct labelling of the student by β_s, and assuming independence, the combined labelling scheme on the average labels $N\beta\beta s$ samples correctly, $N(1 - \beta)(1 - \beta_s)$ samples incorrectly and discards the remaining samples. On the average out of the $N\beta$ samples correctly labelled by the teacher, the combined labelling scheme throws away $(N\beta - N\beta\beta_s)$ samples and out of the $N(1 - \beta)$ samples incorrectly labelled by the teacher $[N(1 - \beta) - N(1 - \beta)(1 - \beta_s)]$ samples are thrown away. If β_s is greater than β, it can be shown that more incorrectly labelled samples are thrown away than correctly labelled samples. Hence throwing away samples does not seem to be bad.

However, the probability of correct classification for the student is dependent on the value of $\underline{X}_{n+1}$ and the particular string of samples thus far presented to the learning scheme. In this situation the argument given in the previous paragraph does not apply. Rather than throwing away samples it may be advantageous to correct the labels on the samples being thrown away and use these samples in the learning process. The feedback schemes discussed below are designed to make use of all the sample patterns.

In the second feedback scheme to be investigated no samples are thrown away. The student accepts the label provided by the teacher on the first N_F samples without questioning. On subsequent samples the student completely ignores the information supplied by the teacher and does his own labelling according to:

$$\text{Label } \underline{X}_{N+1} \text{ as } \theta_1 \text{ if } \hat{f}_{\underline{X}|\theta_1;n_1}(\underline{X}_{N+1}|\theta_1) > \hat{f}_{\underline{X}|\theta_2;n_2}(\underline{X}_{N+1}|\theta_2)$$

$$\text{Label } \underline{X}_{N+1} \text{ as } \theta_2 \text{ if } \hat{f}_{\underline{X}|\theta_2;n_2}(\underline{X}_{N+1}|\theta_2) > \hat{f}_{\underline{X}|\theta_1;n_1}(\underline{X}_{N+1}|\theta_1)$$

$$n_1 + n_2 = N \geq N_F \quad . \tag{3.3.2}$$

Depending on the label, $\underline{X}_{N+1}$ is used to update the estimate of the appropriate density function.

Two of the obvious disadvantages of this method are that there is no control over the amount of feedback and that the teacher is completely ignored. Even though the teacher is known to be imperfect, there is still useful information in the label supplied by the teacher if β is greater than $\frac{1}{2}$. The lack of control on the amount of feedback results in a large probability of incorrect feedback at the tails of the probability density functions. A numerical example is given in

Table IV.

Example 3.3.1.

$$\beta = 0.6;\ P(\theta_1) = P(\theta_2) = 0.5$$

$f_{X|\theta_1}$ and $f_{X|\theta_2}$ are same as in Figure 1

$$n_1 = n_2 = 16$$

$P\{\text{correct feedback} \mid x \epsilon \theta_1\}$ and $P\{\text{Incorrect feedback} \mid x \epsilon \theta_1\}$ were computed using normal approximations given in Appendix B

TABLE IV

PERFORMANCE OF FEEDBACK SCHEME 3.3.2

$f_{X\|\theta_1}(x\|\theta_1)$	$P(\text{correct feedback} \mid x\epsilon\theta_1) = p_1$	$P(\text{Incorrect feedback} \mid x\epsilon\theta_1) = p_2$	Ratio of $p_1:p_2$
2	0.85	0.15	5.6:1
1	0.775	0.225	3.0:1
0.5	0.70	0.30	2.5:1
0.25	0.64	0.36	1.8:1

Table IV contains the probability of correct feedback and probability of incorrect feedback associated with classifying a sample $\underline{x}$ from category θ_1 for various values of $f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)$. Normal approximations

discussed in Appendix B were used to calculate these probabilities. From the values listed in Table IV it can be seen that there is a large probability of incorrect feedback if $f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)$ is small. Also at these points the ratio of probability of correct feedback to probability of incorrect feedback is low.

The only way to improve the performance of this feedback scheme is to wait longer before starting feedback. This leads to the question, what is the "optimum" value of starting feedback N_F? To answer this question a complete knowledge of the prior probabilities and the density functions are required, besides a criteria to be "optimized". No attempt has been made towards obtaining an exact answer for this question. However two special cases of interest, $N_F = 0$ and $N_F = \infty$, have been considered. When $N_F = 0$, the student completely ignores the teacher from the beginning and the feedback scheme 3.3.2 is analogous to learning without a teacher. If $N_F = \infty$, the student would have acquired knowledge equivalent to that of a Bayes' scheme, and no further improvement in the student's knowledge is possible due to feedback since the limiting knowledge is independent of β. Hence neither starting feedback too early nor waiting till too late is good. A compromise is to delay feedback till the teacher has provided the learning scheme with more correctly labelled patterns than incorrectly labelled patterns; the probability of such an event can be used to determine the starting point N_F.

The thresholded feedback scheme considered in the next section implicitly provides an answer to the question of finding the "best" N_F.

3.4 Threshold in Feedback. Some of the disadvantages of the first two methods of feedback can be overcome by using a threshold in the feedback. When a sample pattern $\underline{X}_{N+1}$ is presented to the feedback learning scheme, the label on the sample supplied by the teacher is either accepted or changed by the student according to the following algorithm:

Accept the label provided by the teacher if

$$\left| \hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{X}_{N+1}|\hat{\theta}_1) - \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{X}_{N+1}|\hat{\theta}_2) \right| < T$$

and

change the label to $\hat{\theta}_1$ if $\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{X}_{N+1}|\hat{\theta}_1) > \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{X}_{N+1}|\hat{\theta}_2) + T$

change the label to $\hat{\theta}_2$ if $\hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{X}_{N+1}|\hat{\theta}_2) > \hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{X}_{N+1}|\hat{\theta}_1) + T$. (3.4.1)

In algorithm (3.4.1), $n_1, n_2 \geq 1$, $N = n_1 + n_2$ and T is the threshold. After the label is decided $\underline{X}_{N+1}$ is used to update the estimate of the density function corresponding to the accepted label.

It can be seen from (3.4.1) that the learning scheme ignores the teacher only if the density functions differ by more than T. Loosely stated the label provided by the teacher is questioned and changed only if the student is certain that the teacher is wrong and the student accepts whatever the teacher says if he is not sure of himself. This scheme does not throw away samples like in the first scheme described. Also feedback is done rather selectively unlike the second method where feedback was done on each sample.

The threshold feedback scheme has the following desirable properties:

1) Feedback starts at the modes of the density functions.

2) There is control over the amount of feedback.

3) By choosing T to be decreasing function of N, the teacher can be gradually phased out.

These properties can be established using the theorems proved in Appendix B.

It is shown in Theorem B.4.1 that the maximum probability of feedback occurs at the maximum value of $f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)$ if the given pattern $\underline{x}$ is from category θ_1 and that if $\underline{x}$ is from category θ_2 the maximum probability of feedback occurs at the maximum value of $f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)$. This implies that feedback starts where the density functions have large values. In regions where the density functions have large values, the densities are well separated and hence the student feels confident to challenge the teacher in these regions. Accordingly there is more feedback in these regions as desired. An illustrative example is given below.

Example 3.4.1.

$$\beta = 0.6;\ n_1, n_2 = 16$$

$f_{X|\theta_1}$ and $f_{X|\theta_2}$ are as shown in Figure 1

$$T = 0.1$$

$P(\text{correct feedback} | x \epsilon \theta_1)$ and $P(\text{Incorrect feedback} | x \epsilon \theta_1)$ were calculated according to Equation B.4.1

TABLE V

PERFORMANCE OF THRESHOLDED FEEDBACK SCHEME

$f_{\underline{X}\|\theta_1}(\underline{x}\|\theta_1)$	P(correct feedback$\|x\epsilon\theta_1$) = p_1	P(incorrect feedback$\|x\epsilon\theta_1$) = p_2	P(feedback$\|x\epsilon\theta_1$)	Ratio of $p_1:p_2$
2	0.788	0.03	0.818	26:1
1	.646	.13	0.776	4.9:1
0.5	.50	.15	0.65	3.3:1
0.25	.36	.13	0.49	2.8:1

From the example it can be seen that the probability of feedback gets lower and lower as the value of $f_{\underline{X}|\theta_1}$ decreases. Table V also gives the probability of correct feedback and the probability of incorrect feedback for several values of $f_{\underline{X}|\theta_1}$. Comparing these values with those listed in Table IV for a feedback scheme without threshold, it can be seen the use of a threshold lowers the probability of incorrect feedback at the tail end of the density function. Also the ratio of the probability of correct feedback to the probability of incorrect feedback is better with a threshold.

In Lemma B.4.1 it is shown that increasing T decreases the amount of feedback and vice versa. Hence by varying the threshold T the amount of feedback can be controlled, as opposed to the total lack of control on the amount of feedback in a feedback scheme without threshold. Lemma B.4.1 also shows that, by choosing T to be a decreasing function of N, the amount of feedback can be increased as the learning progresses.

This way the teacher can be gradually phased out as desired. The gradual phasing out of the teacher also provides an answer to the question of finding the optimum starting point for feedback, N_F, required in the second feedback scheme (3.3.2).

Based on the comparison given thus far, it is apparent that the feedback scheme with threshold is better than the other schemes considered. One of the problems associated with a feedback learning scheme with a threshold is the selection of the threshold T. Two methods are given in the next chapter for selection of threshold.

CHAPTER IV

SELECTION OF THRESHOLD

4.1 Introduction. This chapter is concerned with finding an expression for the threshold T in terms of β, and the sample size n. Before going into the actual derivation one might deduce the form of T as follows. It has been shown in Chapter II that for a given sample size n, the performance of the learning scheme improves as β increases. Hence for a given n, there should be more feedback in a learning scheme with a better teacher than in a learning scheme with a relatively bad teacher. This implies T should be a decreasing function of β since the amount of feedback increases as T decreases. Also it has been shown that T must $\rightarrow 0$ as $n \rightarrow \infty$. Since the variance of the discriminant function is a function of $\frac{1}{\sqrt{n}}$, it is intuitively obvious that T must also be decreasing as $\frac{1}{\sqrt{n}}$.

Two methods are given in the following sections for deriving an expression for T, for the equal sample size case. These approaches are later extended to the unequal sample size case. In all these derivations non-overlapping densities will be assumed. As has been mentioned earlier in Chapter II, with a large overlap in the densities the learning scheme can not perform better than the teacher and hence feedback is not good in these cases.

For purposes of analysis in this chapter the normal approximations of the estimators $\hat{f}_{\underline{X}|\theta_1;n}$ and $\hat{f}_{\underline{X}|\theta_2;n}$ described in Appendix B will be

used.

4.2 Minimax Approach. In this rather pessimistic approach the objective is to do feedback in such a way that the chances of incorrect feedback are minimum. Hence the quantity of interest is the probability of incorrect feedback. This probability is a function of β,n and the values of the density functions $f_{\underline{X}|\theta_1}$ and $f_{\underline{X}|\theta_2}$ at $\underline{x}$, the sample being fed back. The dependency on the density function is undesirable since these are quantities that we are trying to estimate. It is shown in this section that the dependency on the density functions can be removed by looking at the maximum value of the probability of incorrect feedback. An expression for T is derived by setting an upper bound on this maximum probability of incorrect feedback.

Theorem 4.2.1. P(Incorrect feedback$|\underline{x}\epsilon\theta_1$) is maximum at $\underline{x}_0$, where $\underline{x}_0\epsilon\theta_1$ is such that

$$(2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}_0|\theta_1) = T \quad .$$

Proof. From B.4.1

$$P(\text{Incorrect feedback}|\underline{x}\epsilon\theta_1) = \int_{-\infty}^{\frac{[-T-(2\beta-1)a]c}{\sqrt{a}} = L} N(0,1)d\xi$$

where

$$c = (2\sqrt{\pi})^{\frac{p}{2}} n^{\frac{1}{4}}$$

and

$$a = f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) \quad .$$

It will now be shown that L has a maximum at

$$a = \frac{T}{(2\beta-1)}$$

$$\frac{dL}{da} = -c\left[\frac{\sqrt{a}\,(2\beta-1) - \frac{[T + (2\beta-1)a]}{2\sqrt{a}}}{a}\right]$$

$$= -c\left[\frac{a(2\beta-1) - \frac{[T + (2\beta-1)a]}{2}}{a^{3/2}}\right]$$

$$= -c\left[\frac{a(2\beta-1) - T}{2a^{3/2}}\right]$$

$$= 0 \quad \text{at } \underline{x}_0 \text{ if } a = f_{\underline{X}|\theta_1}(\underline{x}_0|\theta_1) = \frac{T}{2\beta-1}$$

$$\frac{d^2L}{da} = -\frac{c}{2}\frac{d}{da}\left[\frac{a(2\beta-1) - T}{a^{3/2}}\right]$$

$$= -\frac{c}{2}\left[\frac{a^{\frac{3}{2}}(2\beta-1) - [a(2\beta-1)-T]3/2\sqrt{a}}{a^3}\right]$$

$$= -\frac{c}{2}\left[\frac{a(2\beta-1) - [a(2\beta-1)-T]3/2}{a^{5/2}}\right]$$

$$= -\frac{c}{4}\frac{[3T - a(2\beta-1)]}{a^{5/2}} \quad .$$

At

$$T = (2\beta-1)a,$$

$$\frac{d^2L}{da} < 0 \quad .$$

Hence L has a maximum at $T = (2\beta-1)a$. This implies $p(\text{incorrect feedback } \underline{x}\epsilon\theta_1|;n) = \int_{-\infty}^{L} N(0,1)d\xi$ has a maximum at $\underline{x}_0$ if $f_{\underline{X}|\theta_1}(\underline{x}_0|\theta_1) = \frac{T}{2\beta-1}$. Hence the proof the the theorem. The example discussed on page 48 illustrates this theorem. A similar result can be derived for $\underline{x}\epsilon\theta_2$.

Using Theorem 4.2.1 one can obtain the maximum value of $P(\text{incorrect feedback}|\underline{x}\epsilon\theta_1;n)$ as

$$= \int_{-\infty}^{-\frac{2T}{(T/2\beta-1)^{\frac{1}{2}}}c} N(0,1)d\xi \qquad (4.2.1)$$

where

$$c = n^{\frac{1}{4}}(2\sqrt{\pi})^{\frac{p}{2}} \quad .$$

Similarly it can be shown that the maximum $P(\text{incorrect feedback}|\underline{x}\epsilon\theta_2;n)$

$$= \int_{-\infty}^{-\frac{2T}{(T/2\beta-1)^{\frac{1}{2}}}c} N(0,1)d\xi \quad . \qquad (4.2.2)$$

From (4.2.1) and (4.2.2) it can be seen that the maximum value of $P(\text{incorrect feedback}|n)$

$$= \int_{-\infty}^{-2T^{\frac{1}{2}}c(2\beta-1)^{\frac{1}{2}}} N(0,1)d\xi \quad . \qquad (4.2.3)$$

Using Equation 4.2.3, the expression for the threshold can be obtained by setting maximum value of P(incorrect feedback|n) equal to α, i.e.

$$\int_{-\infty}^{-2T^{\frac{1}{2}}c(2\beta-1)^{\frac{1}{2}}} N(0,1)d\xi = \alpha \quad . \tag{4.2.4}$$

From the table of normal integrals, the value t_α can be determined such that

$$\int_{-\infty}^{-t_\alpha} N(0,1)d\xi = \alpha \quad . \tag{4.2.5}$$

Comparing Equation 4.2.5 and 4.2.4 it can be seen that

$$2T^{\frac{1}{2}}c(2\beta-1)^{\frac{1}{2}} = t_\alpha$$

$$T = \frac{t_\alpha^2}{4c^2(2\beta-1)}$$

$$= \frac{t_\alpha^2}{4(2\sqrt{\pi})^p \sqrt{n}\ (2\beta-1)}$$

$$T = \frac{c_\alpha}{\sqrt{n}\ (2\beta-1)} \tag{4.2.6}$$

where

$$c_\alpha = \frac{t_\alpha^2}{4(2\sqrt{\pi})^p} \quad .$$

Choosing T according to Equation 4.2.6 guarantees that on the average the maximum probability of incorrect feedback at any given stage of learning is equal to the desired value α. For any given pattern $\underline{x}$ the probability of incorrect feedback is less than or equal to α. The expression for T given in (4.2.6) has the desired properties that the teacher is gradually phased out, the amount of feedback at a given stage is more for a learning scheme with a good teacher compared to a learning scheme with a comparatively bad teacher and T decreases according to $\frac{1}{\sqrt{n}}$, the same rate of decrease as the variance of the estimator of discriminant function.

4.3 Decision Theory Approach. In the decision theory approach towards finding an expression for the threshold T, values are assigned to the possible out comes of feedback and T is chosen such that the "average value" is maximized. The decision tree for this problem is shown in Figure 5. The tree is drawn as if the true category of the pattern $\underline{x}$ to be fed back is known to the decision maker, the decision being the choice of threshold T. For each value of T there are three possible outcomes, namely, $\underline{x}$ is correctly fed back, no feedback or $\underline{x}$ is incorrectly fed back. The "values" associated with these outcomes are $a', 0$ and $-b'$ respectively, $a', b' > 0$. Using the probabilities of these outcomes the "average value" of feedback can be computed and T can be chosen such that the average value is maximized.

Theorem 4.3.1. The "average value" associated with feeding back a sample $\underline{x}\epsilon\theta_1$, at the n^{th} stage of learning,

$$E\{V_1(T)\} = a' \, P(\text{correct feedback} \mid \underline{x}\epsilon\theta_1; n) - b' \, P(\text{incorrect feedback} \mid \underline{x}\epsilon\theta_1; n), \qquad (4.3.1)$$

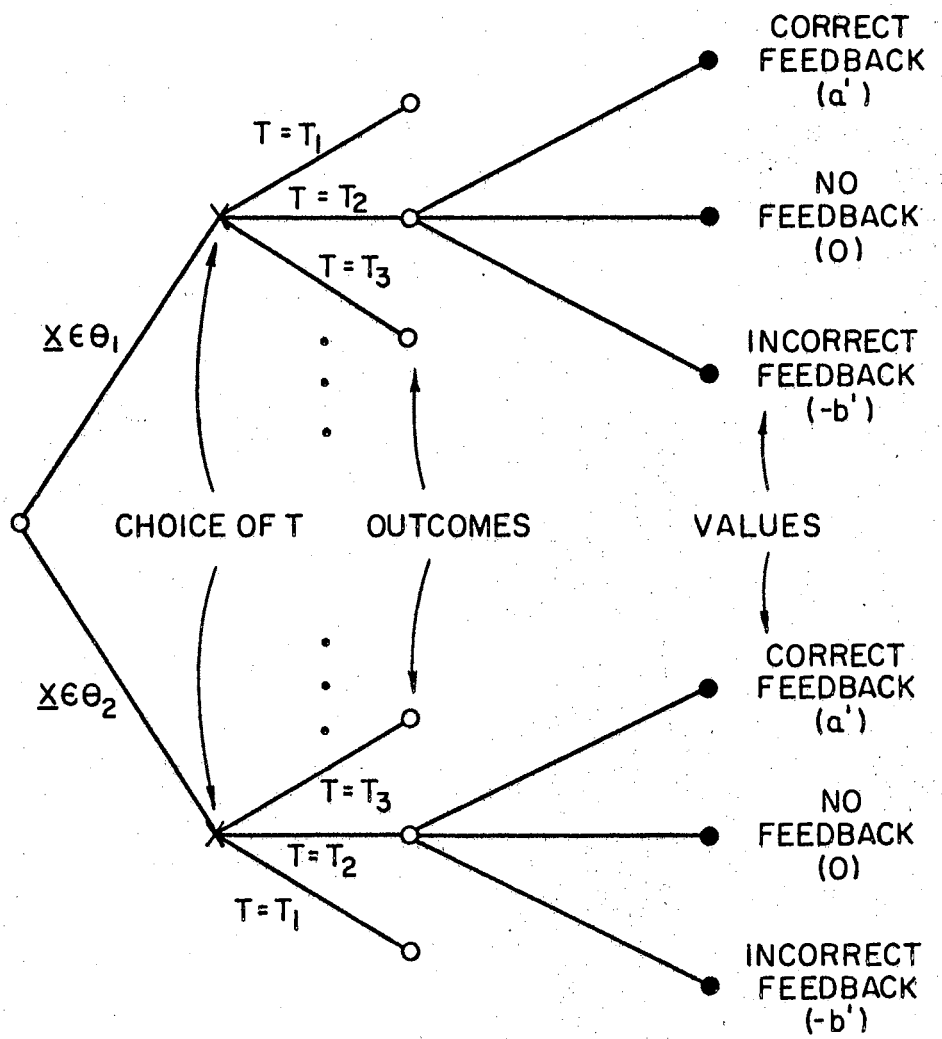

Figure 5. Decision Tree for Selection Threshold (Equal Sample Size)

is maximum if T is chosen to be

$$T = \frac{\log_e \frac{b'}{a'}}{2(2\sqrt{\pi})^p \sqrt{n}\ (2\beta-1)} \quad . \tag{4.3.2}$$

Proof.

$$P(\text{correct feedback}\,|\,\underline{x}\epsilon\theta_1) = \int_T^{\infty} N[(2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1),\ \frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}{(2\sqrt{\pi})^p \sqrt{n}}]d\xi$$

$$P(\text{incorrect feedback}\,|\,\underline{x}\epsilon\theta_1) = \int_{-\infty}^{-T} N[(2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1),\ \frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}{(2\sqrt{\pi})^p \sqrt{n}}]d\xi \quad .$$

Substituting these probabilities in (4.3.1)

$$E\{V_1(T)\} = a' \int_{\frac{[T-(2\beta-1)a]c}{\sqrt{a}}}^{-\infty} N(0,1)d\xi \quad -b' \int_{-\infty}^{\frac{[-T-(2\beta-1)a]c}{\sqrt{a}}} N(0,1)d\xi$$

where

$$a = f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)$$

and

$$c = (2\sqrt{\pi})^{\frac{p}{2}}\, n^{\frac{1}{4}} \quad .$$

Taking the derivative with respect to T, it follows that

$$\frac{d}{dT}\{E[V_1(T)]\} = -[\frac{a'}{\sqrt{2\pi}} \exp\{-\frac{c^2[T-(2\beta-1)a]^2}{2a}\} \frac{c}{\sqrt{a}}]$$

$$-[\frac{b'}{\sqrt{2\pi}} \exp\{-\frac{c^2[T+(2\beta-1)a]^2}{2a}\}(-\frac{c}{\sqrt{a}})$$

$$= \frac{c}{\sqrt{2\pi a}} [b'\exp\{-\frac{c^2[T+(2\beta-1)a]^2}{2a}\} - a'\exp\{-\frac{c^2[T-(2\beta-1)a]^2}{2a}\}].$$

Setting $\frac{d}{dT}\{E[V_1(T)]\} = 0$ gives

$$b'\exp\{-\frac{c^2[T+(2\beta-1)a]^2}{2a}\} = a'\exp\{-\frac{c^2[T-(2\beta-1)a]^2}{2a}\} \quad ,$$

since

$$\frac{c}{\sqrt{2\pi a}} \neq 0 \quad .$$

Taking logarithm on both sides

$$\log_e b' - \frac{c^2}{2a}[T + (2\beta-1)a]^2 = \log_e a' - \frac{c^2}{2a}[T - (2\beta-1)a]^2$$

$$\frac{c^2}{2a}\{[T + (2\beta-1)a]^2 - [T - (2\beta-1)a]^2\} = \log_e(\frac{b'}{a'})$$

$$\frac{c^2}{2a} \cdot 4aT(2\beta-1) = \log_e(\frac{b'}{a'})$$

$$T = \frac{\log_e(\frac{b'}{a'})}{2(2\sqrt{\pi})^p \sqrt{n}\ (2\beta-1)}$$

$$= \frac{c_0}{\sqrt{n}\ (2\beta-1)} \qquad (4.3.3)$$

where

$$c_0 = \frac{\log_e(\frac{b'}{a'})}{2(2\sqrt{\pi})^p} \quad .$$

It can be shown that $\frac{d^2}{dT^2}\{E[V_1(T)]\}$ is negative at $T = \frac{c_0}{\sqrt{n}\ (2\beta-1)}$. Hence $E\{V_1(T)\}$ is maximum at T given in Equation 4.3.3. Proceeding along the same lines it can be shown that $E\{V_2(T)\}$, the "average value" associated with feeding back a sample $\underline{x}\epsilon\theta_2$, is also maximum at T given in Equation 4.3.3. This implies that irrespective of whether $\underline{x}\epsilon\theta_1$ or $\underline{x}\epsilon\theta_2$, the "average value" of feedback is maximum if T is chosen according to (4.3.3).

The form of T given in (4.3.3) is the same as the one given in Equation 4.2.6 for the minimax approach. The only difference is in the constants appearing in the expression. These constants are determined by the choice of the maximum value of the probability of incorrect feedback, or the value function and hence are subjective in nature.

The algorithm for feedback now is:

Acccept the label provided by the teacher if

$$\left| \hat{f}_{\underline{X}|\hat{\theta}_1;n}(\underline{x}|\hat{\theta}_1) - \hat{f}_{\underline{X}|\hat{\theta}_2;n}(\underline{x}|\hat{\theta}_2) \right| < T \quad .$$

Relabel $\underline{x}$ as

$$\hat{\theta}_1 \text{ if } \hat{f}_{\underline{X}|\hat{\theta}_1;n}(\underline{x}|\hat{\theta}_1) > \hat{f}_{\underline{X}|\hat{\theta}_2;n}(\underline{x}|\hat{\theta}_2) + T$$

$$\hat{\theta}_2 \text{ if } \hat{f}_{\underline{X}|\hat{\theta}_2;n}(\underline{x}|\hat{\theta}_2) > \hat{f}_{\underline{X}|\hat{\theta}_1;n}(\underline{x}|\hat{\theta}_1) + T \qquad (4.3.4)$$

where T is the threshold given in Equation 4.2.6 for the minimax

approach and in Equation 4.3.3 for the decision theory approach.

4.4 Extension of Decision Theory Approach to the Unequal Sample Size Case. For the equal sample size case $n_1 = n_2 = n$, an expression for the threshold T was derived in the previous section and it was shown that the value of T given in Equation 4.3.3 maximizes both $E\{V_1(T)\}$ and $E\{V_2(T)\}$. With unequal sample size $n_1 \neq n_2$, it can be shown that

$$T_1 = \frac{\log_e(\frac{b'}{a'})}{2(2\sqrt{\pi})^p(2\beta-1)} \left(\frac{\beta}{\sqrt{n_1}} + \frac{1-\beta}{\sqrt{n_2}}\right) \tag{4.4.1}$$

maximizes $E\{V_1(T)\}$ and

$$T_2 = \frac{\log_e(\frac{b'}{a'})}{2(2\sqrt{\pi})^p(2\beta-1)} \left(\frac{\beta}{\sqrt{n_2}} + \frac{1-\beta}{\sqrt{n_1}}\right) \tag{4.4.2}$$

maximizes $E\{V_2(T)\}$. Since it is not known if $\underline{x}\epsilon\theta_1$ or $\underline{x}\epsilon\theta_2$, the above expressions are of no use. We need to further average $E\{V_1(T)\}$ and $E\{V_2(T)\}$ with respect to $P(\theta_1)f_{\underline{X}|\theta_1}$ and $P(\theta_2)f_{\underline{X}|\theta_2}$ respectively and find T that maximizes this average. The decision tree shown in Figure 5 is redrawn in Figure 6 for this purpose.

The quantity to be maximized now is

$$\begin{aligned} E\{V(T)\} = P(\theta_1) \int_{-\infty}^{\infty} f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)\{a'[P(\text{correct feedback}|\underline{x}\epsilon\theta_1;n_1,n_2)] \\ - b'[P(\text{incorrect feedback}|\underline{x}\epsilon\theta_1;n_1,n_2)]\}d\underline{x} \\ + P(\theta_2) \int_{-\infty}^{\infty} f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)\{a'[P(\text{correct feedback}|\underline{x}\epsilon\theta_2;n_1,n_2)] \\ - b'[P(\text{incorrect feedback}|\underline{x}\epsilon\theta_2;n_1,n_2)]\}d\underline{x}. \end{aligned} \tag{4.4.3}$$

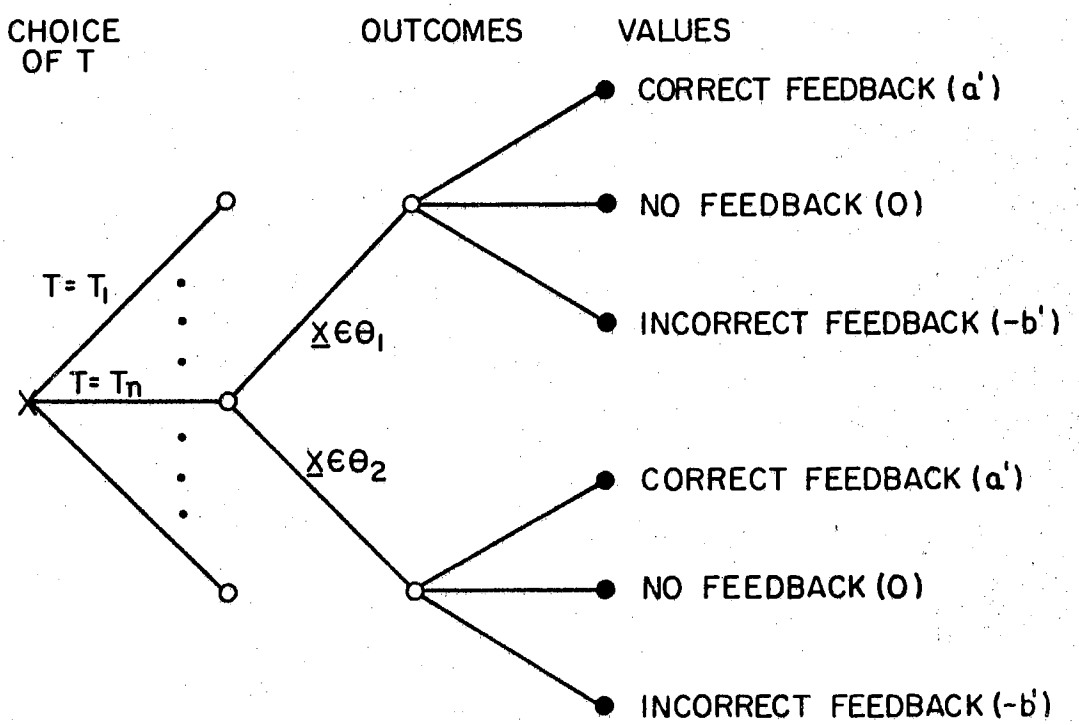

Figure 6. Decision Tree for Selection of Threshold (Unequal Sample Size)

In order to be able to maximize $E\{V(T)\}$ with respect to T, the density functions $f_{\underline{X}|\theta_1}$ and $f_{\underline{X}|\theta_2}$ must be specified completely. This is very unreasonable since if these densities were known then Bayes' procedure completely specifies the discriminant function and there is no need for any learning. Even if these densities were known there is still the problem of finding the value of T that maximizes $E\{V(T)\}$ given in Equation 4.4.3. An example will illustrate this difficulty.

Let us assume that $f_{\underline{X}|\theta_1}$ and $f_{\underline{X}|\theta_2}$ are two univariate density functions of random variables uniformly distributed over non-overlapping intervals of length 1. Substituting for the various probabilities on the right hand side of Equation 4.4.3 and taking derivative with respect to T,

$$\frac{d}{dT} E\{V(T)\} = P(\theta_1)\left[-\frac{a'}{\sqrt{2\pi}} c_1 \exp\left\{-\frac{c_1^2}{2}[T-(2\beta-1)]^2\right\} + \frac{b'}{\sqrt{2\pi}} c_1 \exp\left\{-\frac{c_1^2}{2}[T+(2\beta-1)]^2\right\}\right]$$
$$+ P(\theta_2)\left[-\frac{a'}{\sqrt{2\pi}} c_2 \exp\left\{-\frac{c_1^2}{2}[T-(2\beta-1)]^2\right\} + \frac{b'}{\sqrt{2\pi}} c_2 \exp\left\{-\frac{c_2^2}{2}[T+(2\beta-1)]^2\right\}\right] \quad (4.4.4)$$

where

$$c_1 = \frac{1}{(2\sqrt{\pi})^p}\left[\frac{\beta}{\sqrt{n_1}} + \frac{(1-\beta)}{\sqrt{n_2}}\right]^{\frac{1}{2}}$$

$$c_2 = \frac{1}{(2\sqrt{\pi})^p}\left[\frac{\beta}{\sqrt{n_2}} + \frac{(1-\beta)}{\sqrt{n_1}}\right]^{\frac{1}{2}} .$$

Setting $\frac{d}{dT}\{E[V(T)\} = 0$ in order to solve for extremum, it can be seen from (4.4.4) that even for very simple forms of $f_{\underline{X}|\theta_1}$ and $f_{\underline{X}|\theta_2}$ one has to solve a transcendental equation. For more general forms of the density functions, Equation 4.4.4 takes a more complicated integral form and no closed expression for T can be obtained by setting $\frac{d}{dT}\{E[V(T)]\} = 0$.

However an expression for T can be obtained by using approximations as explained below. Let us assume that the unequal sample size results from unequal prior probabilities $P(\theta_1)$ and $P(\theta_2)$ such that

$$\frac{P(\theta_1)}{P(\theta_2)} \gg \frac{\beta}{1-\beta} \quad . \tag{4.4.5}$$

Now,

$$P(\theta_1|\hat{\theta}_1) = \frac{P(\hat{\theta}_1|\theta_1)P(\theta_1)}{P(\hat{\theta}_1)} \quad ,$$

$$= \frac{P(\hat{\theta}_1|\theta_1)P(\theta_1)}{P(\hat{\theta}_1|\theta_1)P(\theta_1)+P(\hat{\theta}_1|\theta_2)P(\theta_2)} \quad ,$$

$$= \frac{\beta P(\theta_1)}{\beta P(\theta_1)+(1-\beta)P(\theta_2)} \quad .$$

Since $P(\theta_1) \gg P(\theta_2)$ from (4.4.5),

$$P(\theta_1)\beta \gg P(\theta_2)(1-\beta) \quad .$$

Hence

$$P(\theta_1|\hat{\theta}_1) \simeq \frac{\beta P(\theta_1)}{\beta P(\theta_1)} = 1 \quad . \tag{4.4.6}$$

Equation 4.4.6 states that the probability that a sample labelled as $\hat{\theta}_1$ by the teacher, being actually from category θ_1, is approximately equal to one. Hence there is no need for the student to question the samples labelled as $\hat{\theta}_1$.

Looking at the samples labelled as $\hat{\theta}_2$,

$$P(\theta_1|\hat{\theta}_2) = \frac{(1-\beta)P(\theta_1)}{P(\hat{\theta}_2)}$$

and

$$P(\theta_2|\hat{\theta}_2) = \frac{\beta P(\theta_2)}{P(\hat{\theta}_2)} \quad .$$

Since $P(\theta_1)(1-\beta) \gg P(\theta_2)\beta$,

$$P(\theta_1|\hat{\theta}_2) \gg P(\theta_2|\hat{\theta}_2) \quad . \qquad (4.4.7)$$

Inequality (4.4.7) implies that the probability that a sample labelled as $\hat{\theta}_2$ by the teacher is from category θ_1 is much larger than the probability that the sample is actually from category θ_2. Hence samples labelled as $\hat{\theta}_2$ by the teacher need to be checked, and reclassified if necessary.

For example if a total of N labelled patterns are given, then $N[P(\theta_1)(1-\beta) + P(\theta_2)\beta]$ patterns will, on the average, carry the label $\hat{\theta}_2$. But of these, $NP(\theta_1)(1-\beta)$ samples will, on the average, be from category θ_1. Since $NP(\theta_1)(1-\beta) \gg NP(\theta_2)\beta$, while feeding back samples labelled as $\hat{\theta}_2$ one needs to be concerned about these large numbers of samples from category θ_1 and hence maximize the function

$$E\{V_1(T)\} = a'P(\text{correct feedback}\,|\,\underline{x}\epsilon\theta_1;n_1,n_2) - b'P(\text{incorrect feedback}\,|\,\underline{x}\epsilon\theta_1;n_1,n_2) \quad .$$

The value of T that maximizes $E\{V_1(T)\}$ is given by

$$T = \frac{\log_e(\frac{b'}{a'})}{2(2\sqrt{\pi})^p(2\beta-1)}\left[\frac{\beta}{\sqrt{n_1}} + \frac{(1-\beta)}{\sqrt{n_2}}\right]$$

$$= \frac{c_0}{(2\beta-1)}\left[\frac{\beta}{\sqrt{n_1}} + \frac{(1-\beta)}{\sqrt{n_2}}\right] \qquad (4.4.8)$$

where

$$c_0 = \frac{\log_e(\frac{b'}{a'})}{2(2\sqrt{\pi})^p} \quad .$$

The algorithm for feedback now is:

Change the label on $\underline{X}$ only if teacher said $\hat{\theta}_2$ and

$$\hat{f}_{\underline{X}|\hat{\theta}_1;n_2}(\underline{x}|\hat{\theta}_1) > \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) + T \quad . \qquad (4.4.9)$$

4.5 Extension of Minimax Approach to the Unequal Sample Size Case.

The argument given in the previous section about relabelling only those samples with labels $\hat{\theta}_2$ can be used to obtain a value for the threshold T using the minimax approach. The quantity of interest now is:

$$P\left(\begin{matrix}\text{incorrect}\\ \text{feedback}\end{matrix}\middle|\, \underline{x}\epsilon\theta_1;n_1,n_2\right) = \int_{-\infty}^{-T} N\left[(2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1), \frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta)}{(2\sqrt{\pi})^p}\left(\frac{\beta}{\sqrt{n_1}} + \frac{1-\beta}{\sqrt{n_2}}\right)\right]d\xi$$

As in Theorem 4.2.1 it can be shown that the maximum $P(\text{incorrect feedback}|\, \underline{x}\epsilon\theta_1;n_1,n_2)$ occurs if

$$T = (2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)$$

and the maximum value of P(incorrect feedback$|\underline{x}\epsilon\theta_1;n_1,n_2$) is given by

$$\int_{-\infty}^{-2T^{\frac{1}{2}}(2\beta-1)^{\frac{1}{2}}c_1} N(0,1)d\xi$$

where

$$c_1 = \frac{1}{(2\sqrt{\pi})^{\frac{p}{2}}}\left[\frac{\beta}{\sqrt{n_1}} + \frac{1-\beta}{\sqrt{n_2}}\right] \quad .$$

By setting

$$\int_{-\infty}^{-2T^{\frac{1}{2}}(2\beta-1)^{\frac{1}{2}}c_1} N(0,1)d\xi = \int_{-\infty}^{-t_\alpha} N(0,1)d\xi = \alpha \quad ,$$

α being the desired maximum value of P(incorrect feedback$|\underline{x}\epsilon\theta_1;n_1,n_2$), the value of T can be obtained as

$$T = \frac{t_\alpha^2}{4(2\beta-1)(2\sqrt{\pi})^p}\left[\frac{\beta}{\sqrt{n_1}} + \frac{1-\beta}{\sqrt{n_2}}\right]$$

$$= \frac{c_\alpha}{(2\beta-1)}\left[\frac{\beta}{\sqrt{n_1}} + \frac{1-\beta}{\sqrt{n_2}}\right] \qquad (4.5.1)$$

where

$$c_\alpha = \frac{t_\alpha^2}{4(2\sqrt{\pi})^p} \quad .$$

The algorithm for feedback now is:

Change the label on $\underline{x}$ only if the teacher said $\hat{\theta}_2$ and

$$\hat{f}_{\underline{X}|\hat{\theta}_1;n_1}(\underline{x}|\hat{\theta}_1) > \hat{f}_{\underline{X}|\hat{\theta}_2;n_2}(\underline{x}|\hat{\theta}_2) + T \quad . \qquad (4.5.2)$$

4.6 Comments. Even though the expressions for the threshold T derived in this chapter display many desirable properties, no claim can be made about the "optimality" of these expressions. One of the short comings of the analysis presented in this chapter lies in treating β as constant. Due to feedback the "effective value of β", defined as the ratio of the number of sample patterns with correct labels to the total number of sample patterns, is changing. A formulation of this change is difficult, if at all possible. Even though probability statements can be made about correct feedback and incorrect feedback, these involve the unknown densities. Whereas for the teacher, β is independent of $\underline{x}$, the sample being labelled, the performance of the student will depend on the value of $\underline{x}$ and the performance of the student in the past. If the student has been incorrectly feeding back the initial samples, then subsequent samples will also be incorrectly fed back. This fundamental difference in the labelling procedure prevents an analysis of performance of the feedback learning scheme as was done in Chapter II for the learning scheme without feedback.

The performance of the proposed feedback learning scheme is evaluated through simulations in the next chapter.

CHAPTER V

SIMULATION RESULTS

5.1 Introduction. For reasons outlined in earlier chapters a complete theoretical analysis of the proposed feedback scheme is extremely difficult, if at all possible. This chapter is concerned with presenting the results of computer simulations of the feedback learning scheme. Three different learning situations, characterized by non-overlapping densities and equal prior probabilities, non-overlapping densities and unequal prior probabilities, and overlapping densities and equal prior probabilities were simulated and the results are presented below.

5.2 Simulations With Non-overlapping Densities; Equal Prior Probabilities. The densities $f_{\underline{X}|\theta_1}$ and $f_{\underline{X}|\theta_2}$ used in this part of the simulations are shown in Figure 1. A total of N_t samples were drawn from these densities with equal prior probabilities $P(\theta_1) = P(\theta_2) = \frac{1}{2}$. These samples were labelled as $\hat{\theta}_1$ and $\hat{\theta}_2$ by an imperfect teacher, characterized by Equation 2.2.10. The learning scheme without feedback accepted the label provided by the teacher and the densities $f_{X|\hat{\theta}_1}$ and $f_{X|\hat{\theta}_2}$ estimated according to (A.2.7). Based on the final estimates $\hat{f}_{X|\hat{\theta}_1;N_{t_1}}$ and $\hat{f}_{X|\hat{\theta}_2;N_{t_2}}$ ($N_t = N_{t_1} + N_{t_2}$) each test sample Z was classified as follows:

$$\theta_1 \text{ if } \hat{f}_{X|\hat{\theta}_1;N_{t_1}}(Z|\hat{\theta}_1) > \hat{f}_{X|\hat{\theta}_2;N_{t_2}}(Z|\hat{\theta}_2)$$

and as

$$\theta_2 \text{ if } \hat{f}_{X|\hat{\theta}_2;N_{t_2}}(Z|\hat{\theta}_2) > \hat{f}_{X|\hat{\theta}_1;N_{t_1}}(Z|\hat{\theta}_1) \quad . \tag{5.2.1}$$

The feedback learning scheme questioned the label on the sample patterns Z_{n+1} and modified the label according to the following rule:

Accept the label provided by the teacher if

$$\left| \hat{f}_{X|\hat{\theta}_1;n}(Z_{n+1}|\hat{\theta}_1) - \hat{f}_{X|\hat{\theta}_2;n}(Z_{n+1}|\theta_2) \right| < T$$

Label Z_{n+1} as $\hat{\theta}_1$ if $\hat{f}_{X|\hat{\theta}_1;n}(Z_{n+1}|\hat{\theta}_1) > \hat{f}_{X|\hat{\theta}_2;n}(Z_{n+1}|\hat{\theta}_2) + T$

Label Z_{n+1} as $\hat{\theta}_2$ if $\hat{f}_{X|\hat{\theta}_2;n}(Z_{n+1}|\hat{\theta}_2) > \hat{f}_{X|\hat{\theta}_1;n}(Z_{n+1}|\hat{\theta}_1) + T$

(5.2.2)

where

$$T = \frac{0.141\sqrt{2}}{\sqrt{n}\,(2\beta-1)} \quad . \tag{5.2.3}$$

In (5.2.2) $\hat{f}_{X|\hat{\theta}_1;n}$ and $\hat{f}_{X|\hat{\theta}_2;n}$ are the estimates of $f_{X|\hat{\theta}_1}$ and $f_{X|\hat{\theta}_2}$ based on a total of n sample patterns. Depending on the label on Z_{n+1} provided by (5.2.2), the corresponding estimate was updated and this procedure was repeated on all the sample patterns. Based on the final estimates $\hat{f}_{X|\hat{\theta}_1;N_{t_1}}$ and $\hat{f}_{X|\hat{\theta}_2;N_{t_2}}$, the test sample Z was classified

according to (5.2.1).

For each value of β ten runs were made with two different sample sizes, $N_t = 20$ and $N_t = 60$. In each run the performance was computed based on the classification of forty test samples, and the average risk was calculated by averaging the risk on the ten runs (the loss function is 1 for misclassification and 0 for correct classification). Figure 7 shows the plot of this simulation results for $N_t = 20$ and Figure 8 shows the plot of this simulation results for $N_t = 60$.

It can be seen from Figures 7 and 8 that feedback on the average improves the performance of the learning scheme. A rather interesting aspect of these plots is that feedback does not seem to improve the performance very much for both higher and lower values of β. At lower lower values of β, i.e. with a very bad teacher, the amount of feedback is small because the student does not learn enough to question his teacher very often. At higher values of β, i.e. with a very good teacher, the student acquires his limiting knowledge quickly and feedback does not help here since feedback does not increase the limiting knowledge. Hence it appears that feedback is most effective when the teacher is mediocre.

A summary of labelling on one of the computer runs for $\beta = 0.6$ is presented in Table VI to illustrate the relabelling of samples due to feedback and the gradual phasing out of the teacher. The second and third columns in Table VI contain samples labelled as $\hat{\theta}_1$ and $\hat{\theta}_2$ respectively by the teacher and columns four and five contain samples relabelled as $\hat{\theta}_1$ and $\hat{\theta}_2$ by the feedback scheme. N is the stage of learning. It can be seen from Table VI that at initial stages of learning the feedback scheme accepts the label provided by the teacher on most

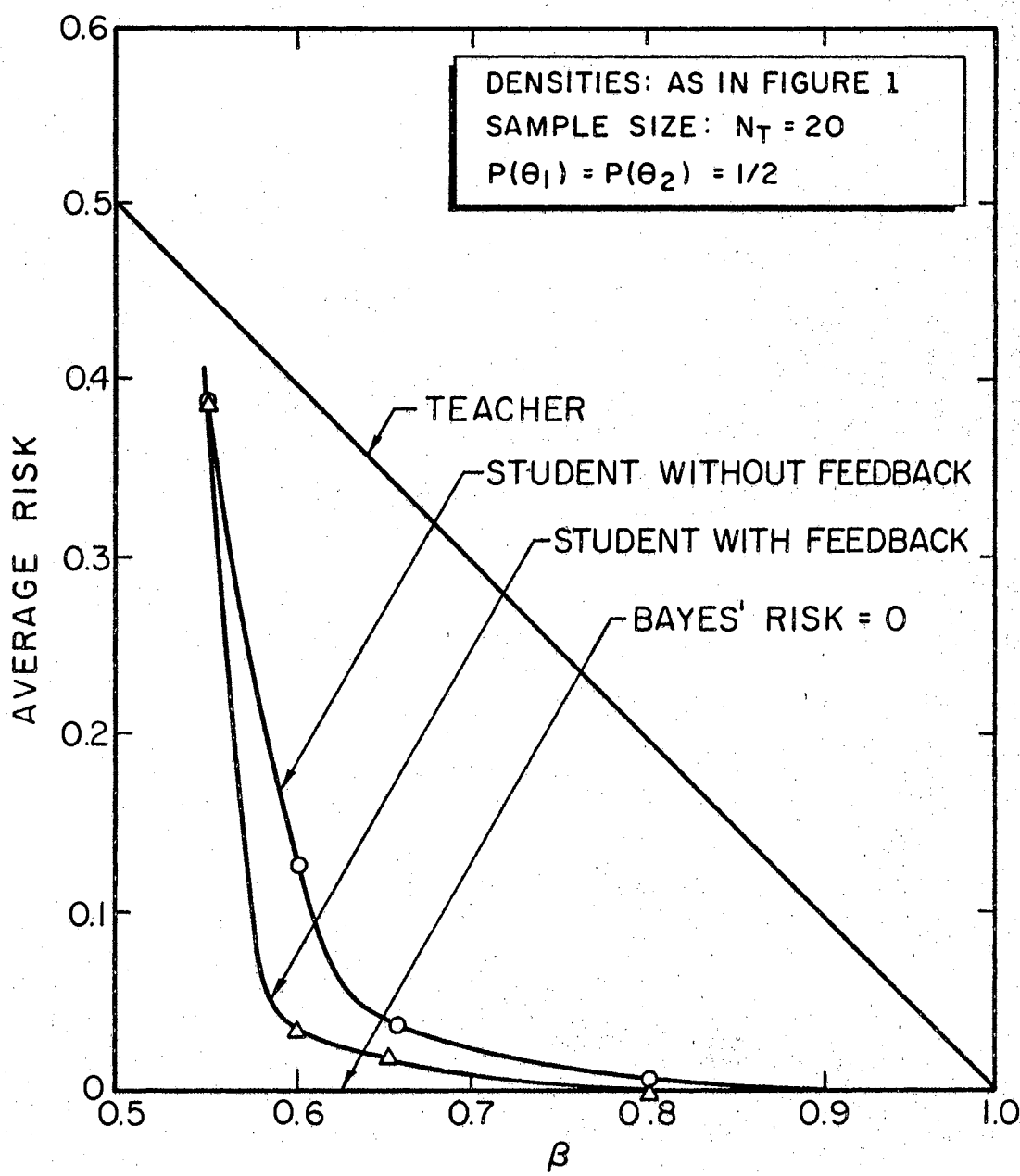

Figure 7. Effect of Feedback on Performance

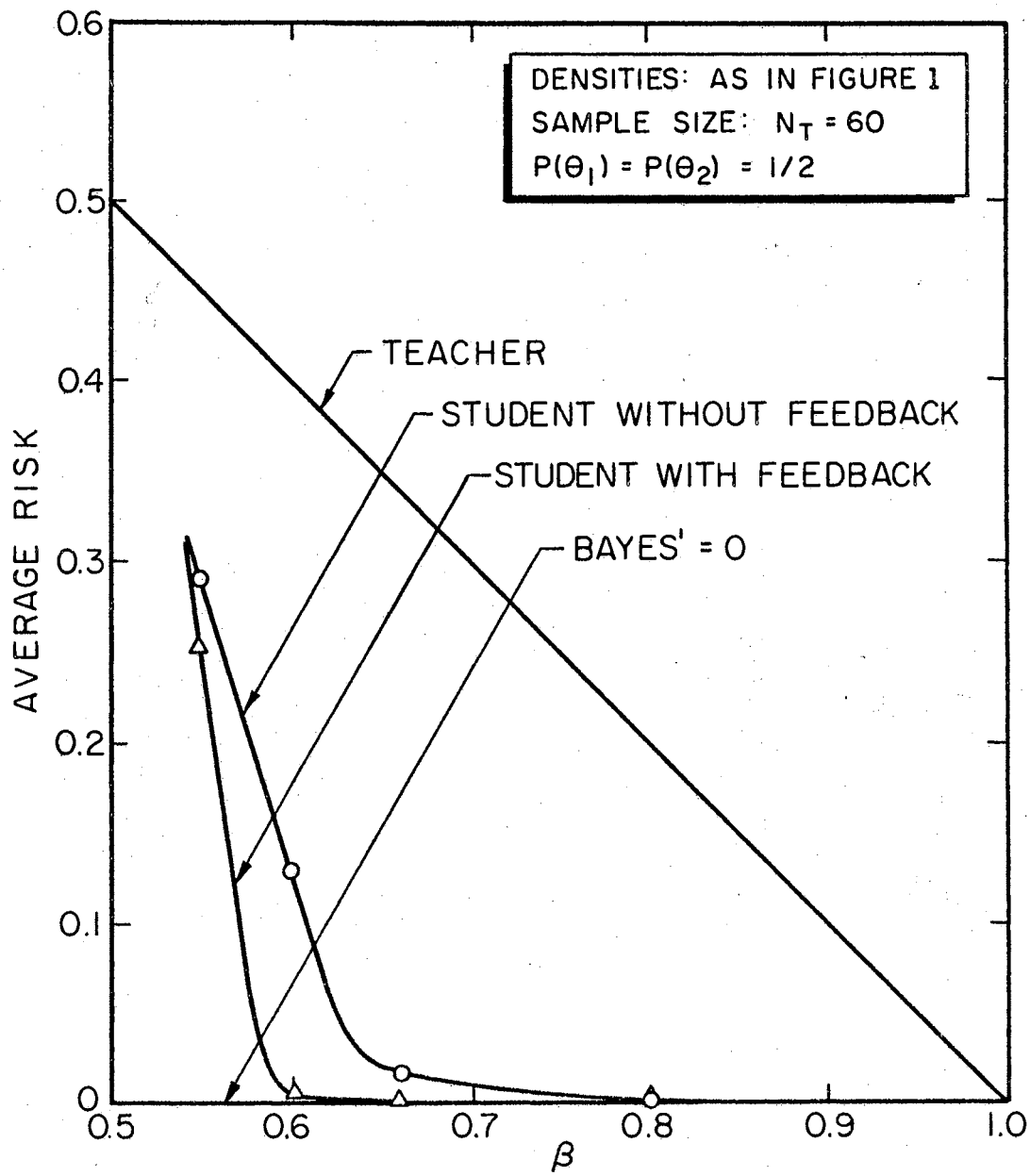

Figure 8. Effect of Feedback on Performance

TABLE VI

SUMMARY OF LABELLING WITH FEEDBACK

STAGE OF LEARNING	SAMPLES LABELLED BY TEACHER		SAMPLES RELABELLED BY FEEDBACK SCHEME	
	AS- $\hat{\theta}_1$	AS- $\hat{\theta}_2$	AS- $\hat{\theta}_1$	AS- $\hat{\theta}_2$
1	0.55020	0.48503	0.55020	0.48503
2	0.59779	0.69131	0.59779	0.69131
3	0.54465	1.44743	0.54465	1.44743
4	0.61472	1.51854	0.61472	1.51854
5	0.37303	1.62923	0.37303	1.62923
5	1.54007	0.45135	1.54007	0.45135
7	0.66136	0.68274	0.66136	0.68274
8	1.23579	1.51217	1.23579	1.51217
9	1.76557	0.46808	1.76557	0.46808
10	1.17509	1.72767	1.7509	1.72767
11	0.65530	0.41548	0.65530	0.41548
12	0.22466	1.74882	0.22466	1.74882
13	1.65705	1.59076	1.65705	1.59076
14	0.41132	0.54918	0.41132	0.54918
15	0.30575	1.71383	0.30675	1.71383
16	0.42291	0.45123	0.42291	0.45123
17	1.56166*	0.88403	0.62163	1.56166
18	1.59375*	1.73901	0.62428	0.88043
19	0.62163	1.23685	0.10980	1.59375
20	0.62428	1.55017	0.28518	1.73091
21	0.10980	1.85793	0.36264	1.23685
22	0.28518	1.46777	0.35588	1.55017
23	0.36264	1.57811	0.57933	1.85793
24	0.35588	1.77813	0.54739	1.46777
25	0.577933	0.02556	0.65951	1.57811
26	0.54739	1.42832	0.92683	1.77813
27	1.58060*	1.60593	0.72246	0.02556
28	0.65951	1.26315	0.84848	1.42832
29	0.92683	1.21079	0.64566	1.58060
30	0.72246	0.84848*	0.37283	1.60953
31	0.64566	1.82512	0.62532	1.26315
32	0.37283	1.62251	0.77855	1.21079
33	1.81883*	1.91005	0.41068	1.82512
34	1.71398*	0.62532*	0.66308	1.62251
35	0.77855	1.18466	0.76831	1.81883
36	0.41068	1.55648	0.41078	1.91005
37	0.66308	1.13248	0.58184	1.71398
38	0.76831	0.41078*	0.45755	1.18466
39	0.58184	0.45755*	0.54628	1.55648
40	0.54628	1.44212	0.13978	1.13248
41	0.13978	0.74364*	0.74364	1.44212
42	1.25586*	1.20546	0.57119	1.25586
43	0.57119	1.56648	0.69249	1.20546
44	0.69249	0.42697*	0.42697	1.56648
45	1.73285*	1.80929	0.45230	1.73285
46	1.45229*	0.45230*	0.76117	1.80929
47	1.20659*	1.43123	0.43915	1.45229
48	1.68309*	1.52286	0.81978	1.20659
49	0.76117	0.43915*	0.13040	1.43213
50	0.81978	0.13040*	------	1.68309
51	------	------	------	1.52286

*-DENOTES SAMPLES LABELLED INCORRECTLY BY TEACHER AND CORRECTLY RELABELLED BY THE FEEDBACK SCHEME

of the sample patterns. As learning progresses, more and more samples are correctly relabelled and for large values of N all the incorrectly labelled samples are relabelled correctly, i.e. the teacher is ignored completely.

It must be mentioned here that these plots only represent the performance on the average. The performance of the learning scheme on any given set of samples will depend on the teacher's labelling on these samples. A summary of performance of the feedback scheme on ten different sets of samples is given below in Table VII to illustrate this.

TABLE VII

SUMMARY OF PERFORMANCE WITH FEEDBACK[1]

β \ Run #	1	2	3	4	5	6	7	8	9	10	Average Risk
0.55	0.0	0.1	0.0	0.0	0.6	0.4	0.94	0.80	0.025	1.0	0.385
0.6	0.0	0.25	0.0	0.0	0.025	0.025	0.0	0.0	0.0	0.0	0.03
0.65	0.0	0.05	0.025	0.0	0.0	0.125	0.0	0.0	0.0	0.0	0.02
0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Densities: As in Figure 1; $P(\theta_1) = P(\theta_2) = \frac{1}{2}$

Sample Size $N_t = 20$

[1]Entries in table denote the risk for the feedback scheme.

5.3 Simulation With Non-overlapping Densities; Unequal Sample Size. For this part of the simulation the densities used are as shown in Figure 1. The prior probabilities were set at $P(\theta_1) = 0.9$ and $P(\theta_2) = 0.1$. The simulation procedure is similar to the one given in Section 5.2 except for the feedback part. As derived in Section 4.4 feedback was done only on samples with labels $\hat{\theta}_2$ as follows.

$$\text{Relabel } Z_{n+1} \text{ as } \hat{\theta}_1 \text{ if } \hat{f}_{X|\hat{\theta}_1;n_1}(Z_{n+1}|\hat{\theta}_1) > \hat{f}_{X|\hat{\theta}_2;n_2}(Z_{n+1}|\hat{\theta}_2) + T$$

where

$$T = \frac{0.141}{(2\beta-1)}\left[\frac{\beta}{\sqrt{n_1}} + \frac{1-\beta}{\sqrt{n_2}}\right]$$

where n_1, n_2 are the number of samples labelled as $\hat{\theta}_1$ and $\hat{\theta}_2$ respectively among the n samples.

The results of the simulation are shown in Figure 9. [Since the algorithm is derived for $P(\hat{\theta}_1) \gg P(\hat{\theta}_2)$, the lowest value of β used in this simulation was 0.65. This corresponds to $P(\hat{\theta}_1) = .62$ and $P(\hat{\theta}_2) = 0.38$.] From these plots it can be seen that once again the average risk of a feedback learning scheme is less than that of a learning scheme using no feedback.

5.4 Simulation With Overlapping Densities. It has been mentioned earlier that feedback is meaningful only if the student can perform better than the teacher. With a large overlap in the densities, to be precise if the Bayes' risk is greater than $(1-\beta)$, the student cannot perform better than his teacher. Hence feedback is not meaningful in these situations. For this reason, no analysis was done in previous

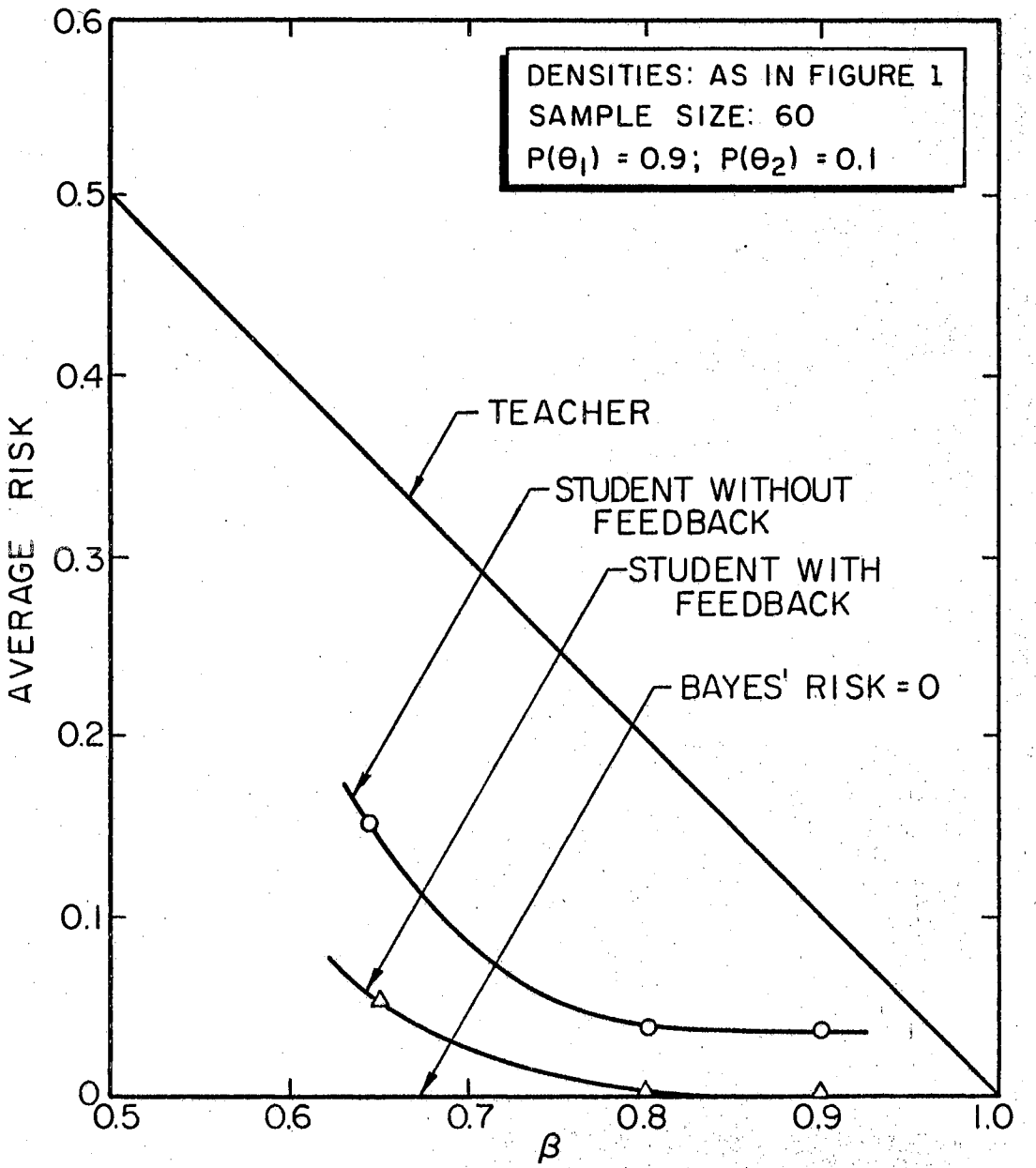

Figure 9. Effect of Feedback on Performance

chapters on feedback with overlapping densities. However, a simulation was done to investigate the performance of the feedback learning scheme with a small overlap in the density functions.

The density functions used in this simulation are shown in Figure 2. Other aspects of this simulation are identical to the one described in Section 5.2. The results of this simulation are shown plotted in Figure 10. It can be seen from these plots that feedback seems to improve the average performance for values of β up to 0.75. For higher values of β, because of overlap the student cannot perform better than the teacher and hence it is better to accept the label provided by the teacher rather than questioning it. But since overlap was not considered in deriving the expression for threshold, the feedback scheme simulated tries to relabel the examples and this results in a comparatively poor performance.

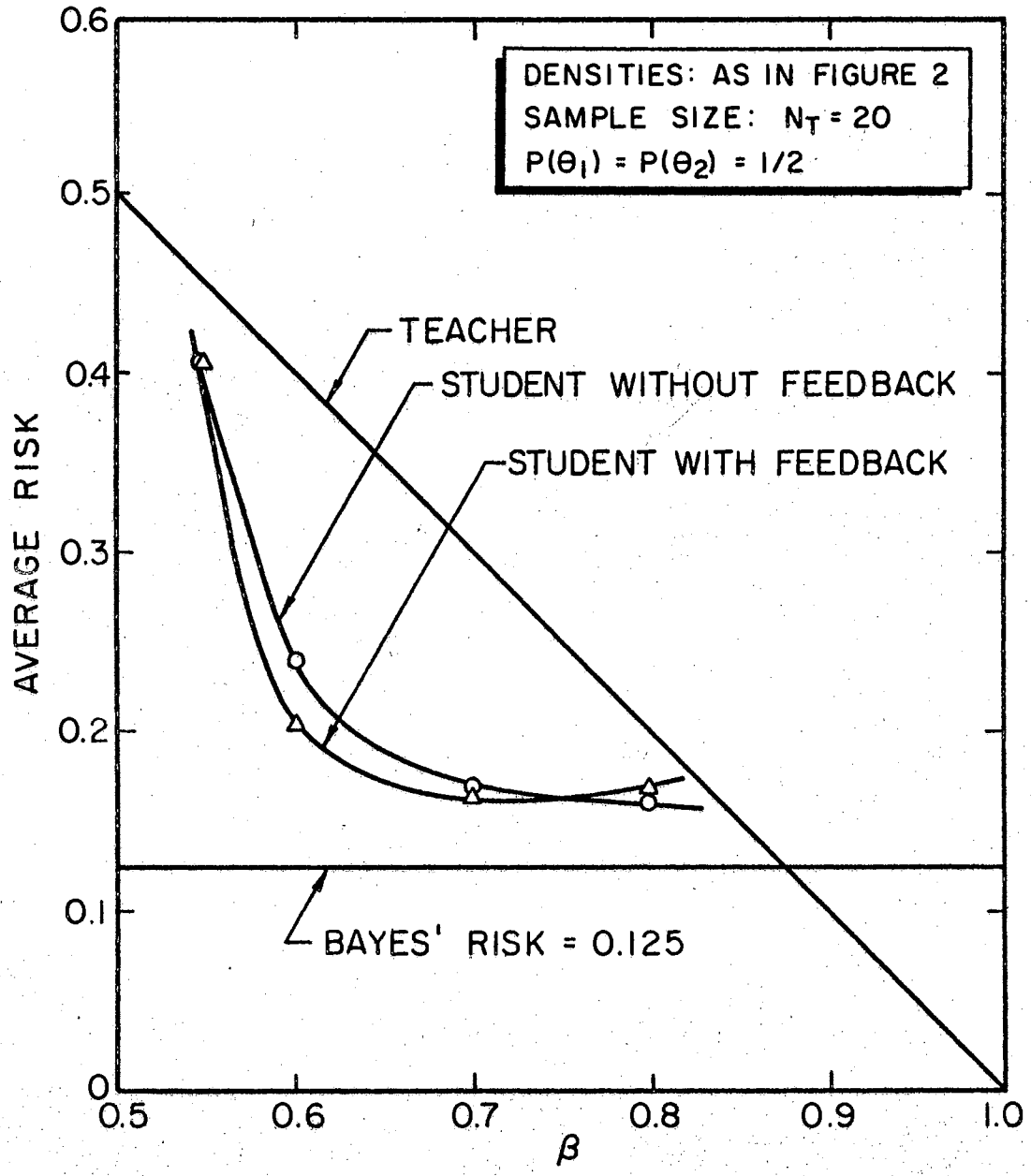

Figure 10. Effect of Feedback on Performance

CHAPTER VI

SUMMARY AND CONCLUSIONS

6.1 Summary. A decision rule was derived for classifying patterns into one of the two possible categories, with an imperfect teacher. A procedure for learning to recognize patterns with an imperfect teacher was developed using nonparametric estimators of the unknown density functions. The asymptotic performance of the proposed learning scheme was analyzed. The finite sample performance of the proposed learning scheme was analyzed under the assumptions that the density functions are smooth and non-overlapping. Based on the results of the analysis justification for considering feedback as a means of improving the performance of the learning scheme was given. Several schemes of feedback were considered and the properties of the feedback scheme using a threshold were derived using the normal approximations to the distribution of the estimators of the unknown density functions. Assuming equal sample size, methods of selecting the value of threshold were given and these methods are used to derive an approximate method of selecting the value of threshold for the unequal sample size case. The thresholded feedback scheme was simulated on the computer and the results were presented.

6.2 Conclusions. It has been shown that the learning scheme proposed in Chapter II has an average asymptotic risk equal to Bayes' (minimum) risk. For non-overlapping densities the performance of the

learning scheme is better than that of the imperfect teacher and the nearest neighbor rule. Also the learning scheme performs better than the teacher on the average even with a finite number of sample patterns. With overlapping densities, if the overlap is less than $(1-\beta)$, the average asymptotic performance of the proposed learning scheme is still better than that of the teacher and the nearest neighbor rule. The proposed learning scheme does not require the exact value of β, the probability of correct labelling by the teacher. The only knowledge required is if $\beta$is greater than or less than $\frac{1}{2}$. Also the learning scheme makes use of the incorrectly labelled sample patterns without requiring the correct label of the sample patterns.

The use of a threshold in the feedback learning scheme offers a simple, logical way to combine a student's knowledge with what is being given to him by the teacher. For non-overlapping densities the average performance of the feedback learning scheme seems to be better than that of a learning scheme not using feedback. For lower values of β, i.e. for a very bad teacher, feedback does not help much since the knowledge acquired by the student in the initial stages of learning is small and hence the amount of feedback is small too. For higher values of β, i.e. with a very good teacher no significant improvement in performance results due to feedback. Hence it appears that feedback is good where the teacher is mediocre.

Feedback results in an improvement in the performance in learning situations with unequal sample size. However, such a claim could not be made when the densities overlap. Further research needs to be done in this area.

6.3 Suggestions for Further Research. The concept of feedback in learning with an imperfect teacher can be applied to other pattern recognition methods such as the nearest neighbor rule and threshold logic. This concept may also be used in parametric pattern recognition methods using Bayesian recursive estimation procedures. Even though several different feedback schemes were considered in this dissertation and the thresholded feedback scheme was shown to be better than the other methods considered, this does not mean that the thresholded feedback scheme is the "best". Investigation needs to be done on the possibility of feedback schemes other than the ones mentioned in this dissertation. Also further research needs to be done on using feedback when the density functions overlap. Another area of research that needs to be explored is concerned with the estimators of density functions. A major problem in this area is the determination of a satisfactory "smoothing factor" to be used in Sprecht's approximation. Forms of estimators other than the ones suggested by Parzen (16) and Murthy (17) need to be investigated.

BIBLIOGRAPHY

1. Ho, Yu-Chi, and A. K. Agarwala. "On Pattern Classification Algorithms - Introduction and Survey." IEEE Transactions on Automatic Control. Vol. AC-13. (December, 1968) 676-690.

2. Nagy, George. "State of the art in Pattern Recognition." Proceedings of the IEEE. Vol. 5. (May, 1968) 836-862.

3. Fu, K. S., and Y. T. Chien. "Sequential Recognition Using a Nonparameter Ranking Procedure." IEEE Transactions on Information Theory. Vol. IT-11. (July, 1967) 484-492.

4. Fu, K. S., and Y. T. Chien. "On the Finite Stopping Rules and Nonparameter Techniques in a Feature Ordered Sequential Recognition System." Purdue University Report. TR-3366-16. (October, 1966).

5. Fix, E., and J. L. Hodges, Jr. "Discriminatory Analysis Nonparametric Discrimination." USAF School of Aviation Medicine, Project No. 21-49-004, Report No. 4. (1951).

6. Cover, T. M., and P. E. Hart. "Nearest Neighbor Classification." IEEE Transactions on Information Theory. Vol. IT-13. (January, 1967) 21-27.

7. Nilsson, Niles J. Learning Machine. New York: McGraw-Hill, 1964.

8. Aizerman. "The Theoretical Foundations of the Method of Potential Functions in the Problem of Teaching Automata to Classify Input Situations." Automation and Remote Control. Vol. 25. (June, 1964) 821-838.

9. Braverman. "The Method of Potential Functions in the Problem of Teaching Machines to Recognize Patterns." Automation and Remote Control. Vol. 27. (October, 1966) 174-177.

10. Sprecht. "Polynomial Discriminant Functions." IEEE Transactions on Electronic Computers. (June, 1967) 174-177.

11. Fralick, S. C. "Learning to Recognize Patterns Without a Teacher." IEEE Transactions on Information Theory. Vol. IT-13. (January, 1967) 57-65.

12. Patrick, E. A. and J. C. Hancock. "Nonsupervised Sequential Classification and Recognition of Patterns." IEEE Transactions on Information Theory. Vol. IT-12. (July, 1966) 366-372.

13. Spragins, J. "Learning Without a Teacher." IEEE Transactions on Information Theory. Vol. IT-12. (April, 1966) 223-229.

14. Duda, R. O., and R. C. Singleton. "Training a Threshold Logic Unit With Imperfectly Classified Patterns." Presented at the Western Joint Computer Conference. Los Angeles, California. (August, 1964).

15. Whitney, A. W., and S. J. Dwyer. "Performance and Implementation of the k-Nearest Neighbor Decision Rule With Incorrectly Identified Training Samples." Presented at the Fourth Annual Allerton Conference on Circuit and System Theory. (January, 1969).

16. Parzen, Emanuel. "On Estimation of a Probability Density Function and Mode." Annals of Mathematical Statistics. Vol. 33. (1962) 1065-1076.

17. Murthy, V. K. "Nonparametric Estimation of Multivariable Densities." Proceedings of an International Symposium on Multivariate Analysis. Dayton, Ohio. Academic Press. (1965).

APPENDIX A

ESTIMATION OF DENSITY FUNCTIONS

A.1 Parzen's Method. Let $X_1, X_2, \ldots, X_n$ be independent random variables identically distributed as a univariate random variable X whose distribution function $F_X(x)$ is absolutely continuous with probability density function $f_X(x)$. A class of estimators of the form

$$f_{X;n}(x) = \frac{1}{nh(n)} \sum_{j=1}^{n} K\{\frac{x-X_j}{h(n)}\} \tag{A.1.1}$$

have been proposed by Parzen for estimating $f_X(x)$. The estimate defined in Equation A.1.1 is asymptotically unbiased and consistent at all points x at which the probability density function is continuous if h(n) and K(y) satisfy the following conditions:

$$\left.\begin{array}{l} \lim_{n\to\infty} h(n) = 0 \\ \lim_{n\to\infty} \{\frac{1}{nh(n)}\} = 0 \end{array}\right\} \tag{A.1.2}$$

$$\left.\begin{array}{l} -\infty < \sup_{y} |K(y)| < \infty \\ \int_{-\infty}^{\infty} |K(y)|dy < \infty \\ \lim_{y\to\infty} |yK(y)| = 0 \\ \int_{-\infty}^{\infty} K(y)dy = 1 \end{array}\right\} . \tag{A.1.3}$$

A.1.1 Bias and Variance of Parzen's Estimate. If the transform of K(y) has a characteristic exponent two, then the bias of the estimate $\hat{f}_{X;n}(x)$ is given by

$$b[\hat{f}_{X;n}(x)] \simeq \frac{h^2}{2} f''_X(x) \int_{-\infty}^{\infty} y^2 K(y)dy \quad . \tag{A.1.4}$$

If the density function $f_X(x)$ is smooth, the second derivative $f''_X(x)$ is small and

$$b[\hat{f}_{X;n}(x)] \simeq 0 \quad . \tag{A.1.5}$$

The variance of $\hat{f}_{X;n}(x)$ may be computed by writing the estimator as an average,

$$\hat{f}_{X;n}(x) = \frac{1}{n} \sum_{k=1}^{n} V_{nk}$$

$$V_{nk} = \frac{1}{h(n)} K(\frac{[x-X_k]}{h(n)})$$

of independent random variables identically distributed as

$$V_n = \frac{1}{h(n)} K(\frac{[x-X]}{h(n)}) \quad .$$

Parzen has shown that the variance of V_n is

$$\sigma^2[V_n] \simeq \frac{1}{h(n)} f_X(x) \int_{-\infty}^{\infty} K^2(y)dy$$

and hence the variance of $\hat{f}_{X;n}(x)$ is

$$\sigma^2[\hat{f}_{X;n}(x)] \simeq \frac{1}{nh(n)} f_X(x) \int_{-\infty}^{\infty} K^2(y)dy \quad . \tag{A.1.6}$$

A.1.2 Consistency and Asymptotic Normality of Parzen's Estimate. Since h is chosen such that it satisfies Equation A.1.2, it can be seen

from Equations A.1.4 and A.1.6 that the bias and variance of $\hat{f}_{X;n}(x)$ tends to zero as n tends to infinity. Hence $\hat{f}_{X;n}(x)$ is a consistent estimate of $f_X(x)$. Parzen also shows that $\hat{f}_{X;n}(x)$ is asymptotically normally distributed, by showing that

$$P\{\frac{[\hat{f}_{X;n}(x) - f_X(x)]}{\sigma[\hat{f}_{X;n}(x)]} \leq C\} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{C} \exp[-\frac{1}{2} y^2]dy \text{ as } n \rightarrow \infty \quad .$$

Parzen also gives an idea of the closeness of the normal approximation from the Berry-Essen bound.

A.2 Extension of Parzen's Method to the Multivariate Case. Murthy (17) has extended Parzen's results to the multivariate case. The form of estimators proposed by Murthy for estimating a multivariate density function $f_{\underline{X}}(\underline{x})$,

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_p \end{bmatrix} \quad ,$$

based on a set of n independent identically distributed vectors $\underline{X}_1, \cdots, \underline{X}_n$,

$$\underline{X}_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \cdot \\ \cdot \\ \cdot \\ X_{ip} \end{bmatrix}$$

is given by

$$\hat{f}_{\underline{X};n}(\underline{x}) = \frac{1}{n} \sum_{j=1}^{n} \left(\prod_{i=1}^{p} B_{in} \right) K(B_{1n}[X_{j1}-x_1],\ldots,B_{pn}[X_{jp}-x_p]) \qquad (A.2.1)$$

Setting $p = 1$ and $B_{1n} = \frac{1}{h(n)}$ in Equation A.2.1, it can be seen that the above estimator reduces to that proposed by Parzen for the univariate case given in Equation A.1.1. The multidimensional "window" $K(x_1,x_2,\ldots,x_p)$ must satisfy

$$K(x_1,x_2,\ldots,x_p) \geq 0$$

$$K(x_1,x_2,\ldots,x_p) = K(\pm x_1,\pm x_2,\ldots,\pm x_p) \quad . \qquad (A.2.2)$$

For non-negative $x_{11},x_{21},\ldots,x_{p1}$ and $x_{12},x_{22},\ldots,x_{p2}$ such that

$$x_{i1} \geq x_{i2} \qquad i = 1,\ldots,p$$

$$K(x_{11},x_{21},\ldots,x_{p1}) \leq K(x_{12},x_{22},\ldots,x_{p2})$$

and

$$\int_{-\infty}^{\infty},\ldots,\int_{-\infty}^{\infty} K(x_1,x_2,\ldots,x_n)dx_1,\ldots,dx_n = 1 \quad . \qquad (A.2.3)$$

The sequence of non-negative constants B_{in} satisfy

$$B_{in} \longrightarrow \infty \text{ as } n \longrightarrow \infty \qquad i = 1,\ldots,p$$

$$\frac{\left(\prod_{i=1}^{p} B_{in} \right)}{n} \longrightarrow 0 \text{ as } n \longrightarrow \infty \quad . \qquad (A.2.4)$$

If the above conditions hold it has been shown that for large n the bias of the estimator

$$b[\hat{f}_{\underline{X};n}(\underline{x})] \cong 0 \qquad (A.2.5)$$

and the variance of the estimator

$$\sigma^2[\hat{f}_{\underline{X};n}(\underline{x})] \simeq \left(\frac{\prod_{i=1}^{p} B_{in}}{n}\right) f_{\underline{X}}(\underline{x}) \int_{-\infty}^{\infty},\ldots,\int_{-\infty}^{\infty} K^2(x_1,x_2,\ldots,x_p)dx_1,\ldots,dx_p \quad . \tag{A.2.6}$$

As derived by Parzen, Murthy has also derived the consistency and asymptotic normality of the estimators $f_{\underline{X};n}(\underline{x})$.

A major disadvantage of the estimators of Parzen and Murthy is that all the samples $\underline{X}_1,\ldots,\underline{X}_n$ must be permanently stored in the computer. This presents a problem when the samples $\underline{X}_i$ are of large dimensions and when the number of samples n is large, both of which are common in pattern recognition problems. Sprecht (10) has developed a method which would require a fixed storage capacity using a "window", $K(x_1,x_2,\ldots,x_p)$, similar in form to a multivariable normal density function. Hence in order to be able to make use of Sprecht's method, the following form of estimators has been used in this dissertation:

$$\hat{f}_{\underline{X};n}(\underline{x}) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \frac{1}{(2\pi)^{p/2}} \exp\left\{\frac{-[\underline{x}-\underline{X}_j]^T[\underline{x}-\underline{X}_j](n)^{\frac{1}{p}}}{2}\right\} \quad . \tag{A.2.7}$$

From Equations A.2.5 and A.2.6, the asymptotic bias and the variance of the estimator in Equation A.2.7 are

$$b[\hat{f}_{\underline{X};n}(\underline{x})] \simeq 0 \tag{A.2.8}$$

$$\sigma^2[\hat{f}_{\underline{X};n}(\underline{x})] \simeq \frac{f_{\underline{X}}(\underline{x})}{\sqrt{n}\ (2\sqrt{\pi})^p} \quad . \tag{A.2.9}$$

For the purpose of analysis of the feedback scheme discussed in this

dissertation the property of asymptotic normality of the estimator will be used and $\hat{f}_{\underline{X};n}(\underline{x})$ will be treated as a normally distributed random variable with the mean $f_{\underline{X}}(\underline{x})$ and the variance given in Equation A.2.9.

A.3 Sprecht's Approximation. The form of $f_{\underline{X};n}(\underline{x})$ suggested by Sprecht is

$$\hat{f}_{\underline{X};n}(\underline{x}) = \frac{1}{(2\pi)^{p/2}\,\sigma^{p}}\left[\frac{1}{n}\sum_{i=1}^{n}\exp\left\{\frac{-[\underline{X}_i-\underline{x}]^{T}[\underline{X}_i-\underline{x}]}{2\sigma^{2}}\right\}\right] \qquad (A.3.1)$$

where

$$\underline{X}_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \cdot \\ \cdot \\ \cdot \\ X_{ip} \end{bmatrix} \quad \text{and} \quad \underline{x} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \cdot \\ \cdot \\ \cdot \\ x_{ip} \end{bmatrix}$$

are p dimensional vectors. $\hat{f}_{\underline{X};n}(\underline{x})$ can be rewritten as

$$\hat{f}_{\underline{X};n}(\underline{x}) = \frac{1}{\sigma^{p}(2\pi)^{p/2}}\left[\exp\left(\frac{-\underline{x}^{T}\underline{x}}{2\sigma^{2}}\right)\right]\left\{\frac{1}{n}\sum_{i=1}^{n}\exp\left[\frac{x_1X_{i1}+\cdots+x_jX_{ij}+\cdots+x_pX_{ip}+B_i}{\sigma^{2}}\right]\right\}$$

where

$$B_i = -\frac{1}{2}\sum_{j=1}^{p}X_{ij}^{2} \quad .$$

Using Taylor series expansion and the multinomial theorem, the second term on the right hand side can be written as a polynomial $D^{n}(\underline{x})$ and $\hat{f}_{\underline{X};n}(\underline{x})$ can be written as

$$\hat{f}_{\underline{X};n}(\underline{x}) = \frac{1}{\sigma^p (2\pi)^{p/2}} \left[\exp\left(-\frac{\underline{x}^T \underline{x}}{2\sigma^2}\right)\right] D^n(\underline{x}) \qquad \text{(A.3.2)}$$

where

$$D^n(\underline{x}) = D^n_{00\ldots0} + D^n_{10\ldots0} x_1 + \cdots + D^n_{00\ldots1} x_p + D^n_{20\ldots0} x_1^2 + \cdots + D^n_{0\ldots2} x_p^2 + \cdots + D^n_{z_1,z_2\ldots z_p} x_1^{z_1} x_2^{z_2} \cdots x_p^{z_p} + \cdots \qquad \text{(A.3.3)}$$

The coefficients $D^n_{z_1 z_2 \ldots z_1}$ are given by

$$D^n_{z_1,z_2\ldots z_p} = \frac{1}{z_1! z_2! \ldots z_p! \sigma^{2h}} \frac{1}{n} \sum_{i=1}^{n} X_{i1}^{z_1} X_{i2}^{z_2} \cdots X_{ip}^{z_p} \exp\left(\frac{B_i}{\sigma^2}\right) ,$$

where

$$h = z_1 + z_2 + \cdots z_p \qquad \text{(A.3.4)}$$

Noting that

$$D^{n+1}_{z_1,z_2\ldots z_p} = \frac{n}{n+1} D^n_{z_1,z_2\ldots z_p} + \frac{X_{n+1\,1}^{z_1} X_{n+1\,2}^{z_2} \cdots X_{n+1\,p}^{z_p} \exp\left(\frac{B_{n+1}}{\sigma^2}\right)}{z_1! z_2! \ldots z_p!\, \sigma^{2h}\, (m+1)} , \qquad \text{(A.3.5)}$$

it can be seen that a recursive relationship exists for $D^n_{z_1,z_2\ldots z_p}$. Hence for a fixed number of terms M in the Taylor series, one needs to store only M coefficiencts of the polynomial to represent $\hat{f}_{\underline{X};n}(\underline{x})$ given in Equation A.3.2. These coefficients can simply be updated through the recursive Equation A.3.5 when additional samples become available. For problems involving large dimensionality and large number

of samples the saving in storage requirements on the computer is appreciable.

One of the major problems associated with Sprecht's method is the selection of a satisfactory "smoothing factor". Sprecht gave an expression for σ as a function of n, such that the expected mean square error of the estimate $\hat{f}_{\underline{X},n}(\underline{x})$ is minimum where $f_X(\underline{x})$ is a normal density function with mean zero and variance one. This however does not guarantee that the expected mean square error will be minimum for other forms of $f_{\underline{X}}(\underline{x})$. The author's advisor, Dr. Breiphol suggested that σ be taken as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{x})^2}{n(n-1)}} \; ; \qquad \bar{x} = \frac{\Sigma X_i}{n} \quad .$$

This leads to

$$E\{\int x \hat{f}_{X;n}(x)dx\} = E(X)$$

$$E\{\int [x - \bar{x}]^2 \hat{f}_{X;n}(x)dx\} = \text{Variance of } X \quad .$$

Further work needs to be done in this area.

APPENDIX B

ANALYSIS OF PROPOSED LEARNING SCHEME RELATED TO FEEDBACK

B.1 Introduction. This part of the appendix is concerned with analyzing the performance of the proposed learning scheme related to feedback. It will be assumed that there are only two categories of patterns with non-overlapping densities and equal prior probabilities. Normal approximations of the estimators of densities given by Parzen and Murthy will be used. β will be assumed to be greater than $\frac{1}{2}$, the case of $\beta < \frac{1}{2}$ can be taken care of through Corollary 2.2.1. The sample size n used in this appendix is assumed to be large enough to justify the use of asymptotic normal properties of the estimators.

B.2 Normal Approximation.

Lemma B.1. The estimator of the discriminant function

$$\hat{D}_{\hat{\theta};n}(\underline{x}) = \hat{f}_{\underline{X}|\hat{\theta}_1,n}(\underline{x}|\hat{\theta}_1) - \hat{f}_{\underline{X}|\hat{\theta}_2;n}(\underline{x}|\hat{\theta}_2)$$

is asymptotically normally distributed with a mean

$$D_{\hat{\theta}}(\underline{x}) = (2\beta-1)[f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) - f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)]$$

and variance

$$\frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) + f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)}{\sqrt{n}\ (2\sqrt{\pi})^p} .$$

Proof. From (A.2.8), and (A.2.9)

$$\hat{f}_{\underline{X}|\hat{\theta}_1;n}(\underline{x}|\hat{\theta}_1) \sim N(f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1), \frac{f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1)}{\sqrt{n}\ (2\sqrt{\pi})^p})$$

and

$$\hat{f}_{\underline{X}|\hat{\theta}_2;n}(\underline{x}|\hat{\theta}_2) \sim N(f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2), \frac{f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)}{\sqrt{n}\ (2\sqrt{\pi})^p}) \quad .$$

The estimators are independent because of the assumption of independent samples, and hence

$$\hat{f}_{\underline{X}|\hat{\theta}_1;n}(\underline{x}|\hat{\theta}_1) - \hat{f}_{\underline{X}|\hat{\theta}_2;n}(\underline{x}|\hat{\theta}_2) \sim N(f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1) - f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2),$$

$$\frac{f_{\underline{X}|\hat{\theta}_1}(\underline{x}|\hat{\theta}_1) + f_{\underline{X}|\hat{\theta}_2}(\underline{x}|\hat{\theta}_2)}{\sqrt{n}\ (2\sqrt{\pi})^p}).$$

Substituting

$$f_{\underline{X}|\hat{\theta}_1}(\underline{x}) = \beta f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) + (1-\beta) f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)$$

and

$$f_{\underline{X}|\hat{\theta}_2}(\underline{x}) = (1-\beta) f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) + \beta f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) \quad ,$$

it follows that

$$\hat{D}_{\hat{\theta};n}(\underline{x}) = \hat{f}_{\underline{X}|\hat{\theta}_1;n}(\underline{x}|\hat{\theta}_1) - \hat{f}_{\underline{X}|\hat{\theta}_2;n}(\underline{x}|\hat{\theta}_2) \sim N((2\beta-1)[f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) - f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)],$$

$$\frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) + f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)}{\sqrt{n}\,(2\sqrt{\pi})^p}) \quad . \qquad (B.2.1)$$

If $\underline{x}\epsilon\theta_1$, then $f_{\underline{X}|\theta_2}(\underline{x}|\theta_2) = 0$, and

$$\hat{f}_{\underline{X}|\hat{\theta}_1;n}(\underline{x}|\hat{\theta}_1) - \hat{f}_{\underline{X}|\hat{\theta}_2;n}(\underline{x}|\hat{\theta}_2) \sim N((2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1), \frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}{\sqrt{n}\,(2\sqrt{\pi})^p}) \quad . \qquad (B.2.2)$$

If $\underline{x}\epsilon\theta_2$, then $f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) = 0$, and

$$\hat{f}_{\underline{X}|\hat{\theta}_1;n}(\underline{x}|\hat{\theta}_1) - \hat{f}_{\underline{X}|\hat{\theta}_2;n}(\underline{x}|\hat{\theta}_2) \sim N(-(2\beta-1)f_{\underline{X}|\theta_2}(\underline{x}|\theta_2), \frac{f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)}{\sqrt{n}\,(2\sqrt{\pi})^p}) \quad . \qquad (B.2.3)$$

The approximations given in (B.2.1), (B.2.2), and (B.2.3) will be used in the following sections.

B.3 Analysis of Performance.

Lemma B.3.1. $P(\text{correct classification} \mid \underline{x}\epsilon\theta_1;n) \rightarrow 1$ as $n \rightarrow \infty$ where $n = n_1 = n_2$ is the sample size used in estimating the densities.

Proof.

$$P(\text{correct classification} \mid \underline{x}\epsilon\theta_1;n) = P[\hat{D}_{\hat{\theta};n}(\underline{x}) > 0 \mid \underline{x}\epsilon\theta_1] \quad . \qquad (B.3.1)$$

If $\underline{x}\epsilon\theta_1$, from (B.2.2)

$$\hat{D}_{\hat{\theta};n}(\underline{x}) \sim N((2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1), \frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}{\sqrt{n}\ (2\sqrt{\pi})^p}) \quad .$$

Hence

$$P[\hat{D}_{\hat{\theta};n}(\underline{x}) > 0 \,|\, \underline{x}\epsilon\theta_1] = \int_0^\infty N((2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1), \frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}{\sqrt{n}\ (2\sqrt{\pi})^p})d\xi \quad .$$

The notation $\int_a^b N(\mu,\sigma^2)d\xi$ denotes the integral

$$\int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}\frac{(\mu-\xi)^2}{\sigma^2}\}d\xi \quad .$$

Substituting the above integral in (B.3.1), it follows that

$$P(\text{correct classification}\,|\,\underline{x}\epsilon\theta_1;n)$$

$$= \int_0^\infty N((2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1), \frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}{\sqrt{n}\ (2\sqrt{\pi})^p})d\xi = \int_{L_1}^\infty N(0,1)d\xi \quad . \quad (B.3.2)$$

where

$$L_1 = \frac{-(2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}{[\frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}{\sqrt{n}\ (2\sqrt{\pi})^p}]^{\frac{1}{2}}} \quad .$$

As $n \longrightarrow \infty$, the lower limit of the above integral $\longrightarrow -\infty$ and hence $P(\text{correct classification}\,|\,\underline{x}\epsilon\theta_1;n) \longrightarrow 1$ as $n \longrightarrow \infty$.

A similar proof can be given for $P(\text{correct classification}\,|\,\underline{x}\epsilon\theta_2;n)$. Using these results the following theorem can be proved.

Theorem B.3.1. For a symmetrical loss function given in Equation 2.1.1 the average conditional risk associated with classifying a pattern $\underline{x} \longrightarrow 0$, as $n \longrightarrow \infty$.

Proof. The conditional risk is given by

$$\left.\begin{array}{l} (r_n(\underline{x}) = 0 \text{ if } \underline{x} \text{ is correctly classified} \\ \qquad\quad = 1 \text{ if } \underline{x} \text{ is misclassified} \end{array}\right\} \text{based on sample size } n$$

$$P(\text{correct classification}|\underline{x};n) = P(\theta_1)P(\text{correct classification}|\underline{x}\epsilon\theta_1;n) + P(\theta_2)P(\text{correct classification}|\underline{x}\epsilon\theta_2;n) .$$

From Lemma B.3.1, as $n \longrightarrow \infty$,

$$P(\text{correct classification}|\underline{x};n) \longrightarrow P(\theta_1) + P(\theta_2) = 1$$

and

$$P(\text{incorrect classification}|\underline{x};n) \longrightarrow 0 \quad . \qquad (B.3.3)$$

Hence, as $n \longrightarrow \infty$

$$E(r_n(\underline{x})) \longrightarrow 0 = r^*(\underline{x}), \text{ the conditional Bayes' risk.} \qquad (B.3.4)$$

Taking expectation with respect to $f_{\underline{X}}(\underline{x})$, it can be seen that the average asymptotic risk for the learning scheme is equal to the Bayes' risk as was shown in Chapter II.

Lemma B.3.2. For a given sample size $n_1 = n_2 = n$, n being large, $P(\text{correct classification}|\underline{x};n)$ increases as β increases.

Proof.

$$P(\text{correct classification}|\underline{x};n) = P(\theta_1)P(\text{correct classification}|\underline{x}\epsilon\theta_1;n) + P(\theta_2)P(\text{correct classification}|\underline{x}\epsilon\theta_2;n)$$

$$= P(\theta_1)\int_0^\infty N[(2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1), \frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}{\sqrt{n}\ (2\sqrt{\pi})^p}]d\xi$$

$$+ P(\theta_2)\int_{-\infty}^0 N[-(2\beta-1)f_{\underline{X}|\theta_2}(\underline{x}|\theta_2), \frac{f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)}{\sqrt{n}\ (2\sqrt{\pi})^p}]d\xi.$$

$$\frac{d}{d\beta}\{P(\text{correct classification}|\underline{x};n\} =$$

$$\frac{d}{d\beta}\{P(\theta_1)\int_{-(2\beta-1)[f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)(2\sqrt{\pi})^p\sqrt{n}]^{\frac{1}{2}}=L_1}^{\infty} N(0,1)d\xi$$

$$+ P(\theta_2)\int_{-\infty}^{(2\beta-1)[f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)(2\sqrt{\pi})^p\sqrt{n}]^{\frac{1}{2}}=L_2} N(0,1)d\xi\ \}$$

$$= P(\theta_1)2[f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)(2\sqrt{\pi})^p\sqrt{n}]^{\frac{1}{2}} \cdot N(0,1)\Big|_{L_1}$$

$$+ P(\theta_2)2[f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)(2\sqrt{\pi})^p\sqrt{n}]^{\frac{1}{2}} \cdot N(0,1)\Big|_{L_2}$$

(B.3.5)

where

$$N(0,1)\Big|_L = \frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}L^2\} \quad .$$

From (B.3.5) it can be seen that $\frac{d}{d\beta}\{P(\text{correct classification}|\underline{x};n)\}$ is always greater than zero at all points $\underline{x}$ where the densities are greater than zero. Hence as β increases the probability of correct classification also increases.

An expression for the rate of learning can be derived using the normal approximations as follows.

$$\text{Rate of learning} \triangleq \frac{d}{dn}^{1}\{P(\text{correct classification}|\underline{x};n)\}$$

$$P(\text{correct classification}|\underline{x};n) =$$

$$P(\theta_1)\int_{-(2\beta-1)[f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)(2\sqrt{\pi})^p\sqrt{n}]^{\frac{1}{2}}=L_1}^{\infty} N(0,1)d\xi$$

$$+ P(\theta_2)\int_{-\infty}^{(2\beta-1)[f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)(2\sqrt{\pi})^p\sqrt{n}]^{\frac{1}{2}}=L_2} N(0,1)d\xi$$

$$\frac{d}{dn}\{P(\text{correct classification}|\underline{x};n)\} =$$

$$P(\theta_1)(2\beta-1)[f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)(2\sqrt{\pi})^p]^{\frac{1}{2}}\frac{1}{4n^{3/4}}\cdot N(0,1)\Big|_{L_1}$$

$$+ P(\theta_2)(2\beta-1)[f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)(2\sqrt{\pi})^p]^{\frac{1}{2}}\frac{1}{4n^{3/4}}\cdot N(0,1)\Big|_{L_2}$$

$$= (2\beta-1)\frac{(2\sqrt{\pi})^{p/2}}{4n^{3/4}}\{P(\theta_1)[f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)]^{\frac{1}{2}}\cdot N(0,1)\Big|_{L_1}$$

$$+ P(\theta_2)[f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)]^{\frac{1}{2}}\cdot N(0,1)\Big|_{L_2}\}. \qquad \text{(B.3.6)}$$

---

[1] Even though n is discrete, it is treated as a continuous variable in this theorem.

The rate of learning given in (B.3.6) is good for large values of n only.

Another quantity of interest in studying the rate of learning is the average rate of learning defined as

$$\text{Average rate of learning} \triangleq \frac{P(\text{correct classification}|\underline{x};n)}{n} \quad . \quad \text{(B.3.7)}$$

Using the above definition the following lemma can be easily established.

Lemma B.3.3. The average rate of learning for a given sample size n increases as β increases.

Proof. In Lemma B.3.2 it was shown that $P(\text{correct classification}|\underline{x};n)$ increases as β increases. Substituting this result in (B.3.7) it immediately follows that the average rate of learning for a given sample size n increases as β increases.

For every value of $\beta > \frac{1}{2}$, the asymptotic learning approaches that of a Bayes machine independent of β. Hence the rate of learning does not depend on β, in fact the rate of learning $\to 0$ as $n \to \infty$ for every value of $\beta > \frac{1}{2}$.

B.4 Use of Threshold in Feedback. With a threshold T, the probability of correct classification and the probability of incorrect classification for $\underline{x} \epsilon \theta_1$ are given by

$$P\left(\begin{matrix}\text{correct}\\ \text{classification}\end{matrix}\middle|\, \underline{x}\epsilon\theta_1;n\right) = \int_T^{\infty} N\left((2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1), \frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}{\sqrt{n}\ (2\sqrt{\pi})^p}\right)d\xi$$

$$P\left(\begin{matrix}\text{incorrect}\\ \text{classification}\end{matrix}\middle|\, \underline{x}\epsilon\theta_1;n\right) = \int_{-\infty}^{-T} N\left((2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1), \frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}{\sqrt{n}\ (2\sqrt{\pi})^p}\right)d\xi \quad . \quad \text{(B.4.1)}$$

Similar expressions for $\underline{x}\epsilon\theta_2$ can be derived. Using these, the properties of the proposed feedback learning scheme can be analyzed.

Theorem B.4.1. Under assumptions stated in Section B.1,

$$P(\text{feedback}\,|\,\underline{x}\epsilon\theta_1;n) \text{ increases as } f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) \text{ increases} \quad .$$

Proof.

$$P(\text{feedback}\,|\,\underline{x}\epsilon\theta_1;n) = P\{\left| \hat{f}_{\underline{X}|\hat{\theta}_1;n}(\underline{x}|\hat{\theta}_1) - \hat{f}_{\underline{X}|\hat{\theta}_2;n}(\underline{x}|\hat{\theta}_2) \right| > T \,\Big|\, \underline{x}\epsilon\theta_1\}$$

$$= 1 - P\{\left| \hat{f}_{\underline{X}|\hat{\theta}_1;n}(\underline{x}|\hat{\theta}_1) - \hat{f}_{\underline{X}|\hat{\theta}_2;n}(\underline{x}|\hat{\theta}_2) \right| < T \,\Big|\, \underline{x}\epsilon\theta_1\} \quad .$$

From (B.2.2),

$$= 1 - \int_{-T}^{T} N[(2\beta-1)f_{\underline{X}|\theta_1}(\underline{x}|\theta_1), \frac{f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)}{\sqrt{n}\ (2\sqrt{\pi})^p}]d\xi \quad .$$

Denoting $f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)$ by a, and $[\sqrt{n}\ (2\sqrt{\pi})^p]^{\frac{1}{2}}$ by c, the above equation can be written as

$$= 1 - \int_{[\frac{-T-(2\beta-1)a}{\sqrt{a}}]c}^{[\frac{T-(2\beta-1)a}{\sqrt{a}}]c} N(0,1)d\xi \quad .$$

Taking derivative with respect to a

$$\frac{d}{da}\{P(\text{feedback}\,|\,\underline{x}\epsilon\theta_1;n)\} = -\frac{d}{da}\left\{\int_{[\frac{-T-(2\beta-1)a}{\sqrt{a}}]c}^{[\frac{T-(2\beta-1)a}{\sqrt{a}}]c} N(0,1)d\xi\right\}$$

$$= -\frac{1}{\sqrt{2\pi}} \exp\{-[\frac{T-(2\beta-1)a}{\sqrt{2a}}]^2 c^2\}[\frac{c\sqrt{a}[-(2\beta-1)] - \frac{[T-(2\beta-1)a]c}{2\sqrt{a}}}{a}]$$

$$+ \frac{1}{\sqrt{2\pi}} \exp\{-[\frac{T+(2\beta-1)a}{\sqrt{2a}}]^2 c^2\}[\frac{c\sqrt{a}[-(2\beta-1)] + \frac{[T+(2\beta-1)a]c}{2\sqrt{a}}}{a}]$$

$$= -\frac{1}{\sqrt{2\pi}} \exp\{-[\frac{T-(2\beta-1)a}{\sqrt{2a}}]^2 c^2\} \cdot \frac{c}{2} [\frac{-T-(2\beta-1)a}{a^{3/2}}]$$

$$+ \frac{1}{\sqrt{2\pi}} \exp\{-[\frac{T+(2\beta-1)a}{\sqrt{2a}}]^2 c^2\} \cdot \frac{c}{2} [\frac{-(2\beta-1)a+T}{a^{3/2}}]$$

$$= T \frac{1}{\sqrt{2\pi}} \frac{c}{2a^{3/2}} [\exp\{-[\frac{T-(2\beta-1)a}{\sqrt{2a}}]^2 c^2\} + \exp\{-[\frac{T+(2\beta-1)a}{\sqrt{2a}}]^2 c^2\}]$$

$$+ a \frac{1}{\sqrt{2\pi}} \frac{(2\beta-1)c}{2a^{3/2}} [\exp\{-[\frac{T-(2\beta-1)a}{\sqrt{2a}}]^2 c^2\} - \exp\{-[\frac{T+(2\beta-1)a}{\sqrt{2a}}]^2\}] \quad . \tag{B.4.2}$$

The first form of the above expression is greater than zero and in the second term, the exponential factors can be grouped as

$$\exp\{-[\frac{T-(2\beta-1)a}{\sqrt{2a}}]^2 c^2\}[1 - \exp\{-\frac{c^2}{\sqrt{2a}} [4Ta(2\beta-1)]\}] > 0 \quad ,$$

because $T > 0$, $a > 0$ and $(2\beta-1) > 0$. Hence from (B.4.2)

$$\frac{d}{da} \{P[\text{feedback} | \underline{x} \epsilon \theta_1; n]\} > 0 \quad ; \qquad a = f_{\underline{X}|\theta_1}(\underline{x}|\theta_1) \quad .$$

Hence the maximum value of feedback for a sample $\underline{x} \epsilon \theta_1$ occurs if $f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)$ is maximum.

Similarly it can be shown that the maximum value of feedback for a sample $\underline{x} \epsilon \theta_2$ occurs if $f_{\underline{X}|\theta_2}(\underline{x}|\theta_2)$ is maximum. This theorem is used in Section 3.4 to establish that with a threshold, the feedback starts where the densities are maximum.

Lemma B.4.1. $P\{\text{feedback}|\underline{x};n\}$ increases as the threshold T is decreased and vice versa.

Proof. Denoting $f_{\underline{X}|\theta_1}(\underline{x}|\theta_1)$ by a, and $[(2\sqrt{\pi})^p \sqrt{n}]^{\frac{1}{2}}$ by c

$$P\{\text{feedback}|\underline{x}\epsilon\theta_1;n\} = 1 - \int_{\frac{[-T-(2\beta-1)a]c}{\sqrt{a}} = L_1}^{\frac{[T-(2\beta-1)a]c}{\sqrt{a}} = L_2} N(0,1)d\xi$$

$$\frac{dP}{dT}\{\text{feedback}|\underline{x}\epsilon\theta_1;n\} = -\frac{c}{\sqrt{a}} N(0,1)\Big|_{L_2} - \frac{c}{\sqrt{a}} N(0,1)\Big|_{L_1} < 0 \quad .$$

Hence as T increases, $P(\text{feedback}|\underline{x}\epsilon\theta_1;n)$ decreases. A similar argument can be given for $P(\text{feedback}|\underline{x}\epsilon\theta_2;n)$. Combining the two it can be seen that $P\{\text{feedback}|\underline{x};n\}$ decreases as T increases and vice versa.

Lemma B.4.1 is used in Section 3.4 to establish that by varying T, the amount of feedback can be controlled.

An immediate consequence of Lemma B.4.1 is the possibility of a gradual phasing out of the teacher. If T is chosen as a decreasing function of n, the sample size, then in the initial stages of learning T will be large and the learning scheme can be made to depend more on the teacher than on his own knowledge. As $n \longrightarrow \infty$, $T \longrightarrow 0$ and

$$P(\text{feedback}|\underline{x}\epsilon\theta_1;n) = 1 - \int_{[\frac{T-(2\beta-1)a}{\sqrt{a}}]c}^{[\frac{-T-(2\beta-1)a}{\sqrt{a}}]c} N(0,1)d\xi$$

$$\longrightarrow 1 - \int_{\frac{-(2\beta-1)ac}{\sqrt{a}}}^{\frac{-(2\beta-1)ac}{\sqrt{a}}} N(0,1)d\xi$$

$$= 1 \qquad \text{as } n \longrightarrow \infty \quad .$$

Similarly $P(\text{feedback} \mid \underline{x} \epsilon \theta_2; n) \longrightarrow 1$ as $n \longrightarrow \infty$. Hence as $n \longrightarrow \infty$, every sample is fed back with proability 1 and the teacher is completely phased out.

VITA

2

Kumarasamy Shanmugam

Candidate for the Degree of

Doctor of Philosophy

Thesis: LEARNING TO RECOGNIZE PATTERNS WITH AN IMPERFECT TEACHER

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Kalvettuppalayam, Madras, India, January 6, 1943, the son of Kumarasamy and Arukkani.

Education: Graduated from S.S.V. High School Kodumudy in April, 1958; received the Bachelor of Engineering degree from Madras University, India, in April, 1964; received the Master of Engineering degree from the Indian Institute of Science, Bangalore, India, in August, 1966; completed the requirements for the Doctor of Philosophy degree at Oklahoma State University in May, 1970.

Professional Experience: Design Engineer, Subbiah Foundry, Coimbatore, India, January 1964, to August 1964; graduate research assistant, Oklahoma State University, September 1967 to May, 1970; graduate teaching assistant, Oklahoma State University, September 1968 to May, 1970.

Professional Organizations: Member of Eta Kappa Nu.