NONPARAMETRIC PROCEDURES FOR LEARNING WITH AN IMPERFECT TEACHER

Thesis for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
RONALD JOSEPH RICHTER
1972

This is to certify that the

thesis entitled

NONPARAMETRIC PROCEDURES FOR

LEARNING WITH AN IMPERFECT TEACHER

presented by

Ronald Joseph Richter

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Electrical Engineering

Major professor

Date November 8, 1972

ABSTRACT

NONPARAMETRIC PROCEDURES FOR LEARNING
WITH AN IMPERFECT TEACHER

By

Ronald Joseph Richter

In this dissertation a general pattern recognition problem is investigated in which the classification of an observed phenomenon or pattern is inferred from a set of training patterns. The statistical approach to pattern recognition is taken. The pattern classes are assumed to be characterized by probability distributions that are inadequately known. In order to form decision rules the unknowns must be "learned" from a set of training patterns. The training patterns are classified by an imperfect teacher that makes errors in its classifications. This type of learning, called learning with an imperfect teacher, lies in between supervised learning and unsupervised learning.

In the first part of the thesis, a probabilistic model of an imperfect teacher is proposed. Expressions are developed relating a perfect teacher to an imperfect teacher. An example is studied to show the asymptotic effects of using misclassified training patterns in an algorithm designed for supervised learning. The example illustrates the need for developing algorithms specifically for use with an imperfect teacher.

A class of nonparametric learning procedures is proposed for learning to recognize patterns with an imperfect teacher. The procedures require prior knowledge only of the nonsingular matrix of error probabilities characterizing the teacher and of whether the patterns are discrete or continuous random variables. Formal proofs are given showing

that the procedures are asymptotically optimal in the sense that they have expected risks which converge with increasing number of training patterns to the optimal (Bayes) risk. Theorems on rates of convergence are also obtained.

In the latter part of the thesis, the two-class recognition problem is investigated in detail with the objective being to study the quantitative and qualitative effects of using an imperfect teacher rather than a perfect teacher. Large-sample approximations are developed for evaluating the expected risk of the estimated decision rules. The performance of the learning procedures is studied in several examples involving normal, triangular, and binomial distributions. Various large sample properties of the expected risk are also investigated.

Finally, a measure of relative performance is proposed for quantitatively evaluating the effects of an imperfect teacher. This measure is evaluated for the important case of a zero-one loss function. The measure is then used along with a cost of training to establish an overall cost for an imperfect teacher. Conditions are established under which an imperfect teacher is more cost effective than a perfect teacher.

NONPARAMETRIC PROCEDURES FOR LEARNING WITH AN IMPERFECT TEACHER

By

RONALD JOSEPH RICHTER

A THESIS

Submitted to

Michigan State University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

Department of Electrical Engineering

and Systems Science

1972

ACKNOWLEDGEMENTS

The author wishes to thank his thesis advisor, Dr. Richard C. Dubes, for the guidance and encouragement given throughout this research.

Thanks are also due to Dr. Dennis C. Gilliland for his helpful suggestions. Appreciation is also expressed to Dr. G. L. Park, Dr. R. O. Barr, and Dr. J. H. Stapleton for their interest in this work.

The principal support for this research was a United States Steel Foundation Fellowship. The cooperation of The Magnavox Company is also gratefully acknowledged.

Special thanks go to Mrs. Jean Emig for her excellence in typing.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Pattern recognition is the study of ways in which machines, usually meaning digital computers and associated equipment, can observe an environment, learn to distinguish relevant details from background trivia, and make sound and reasonable decisions. The task of designing machines to recognize patterns appears in many different forms in a variety of disciplines. The problems encountered range from the practical to the profound, from engineering design and economics to the theories of artificial intelligence and human learning. But the central problem of pattern recognition is to develop procedures or algorithms that effectively classify an observed phenomenon as resulting from one of a set of sources.

In this thesis a general pattern recognition problem is investigated in which the classification of an observed phenomenon or pattern is inferred from a set of training patterns. A statistical approach to pattern recognition is taken. In this approach the sources or pattern classes are assumed to be characterized by probability distributions, and statistical decision theory is used as the mathematical tool for deriving classification procedures.

When the probability distributions that describe the patterns are inadequately known, the unknowns must be "learned" from a set of training patterns. Learning with a teacher (supervised learning) refers to the

situation in which all of the training patterns have been correctly classified as to origin. When the true classifications of the training patterns are unknown, the learning is said to occur without a teacher (unsupervised learning). This thesis investigates a third type of learning, learning with an imperfect teacher, which lies somewhere between supervised and unsupervised learning. For this type of learning the training patterns are classified by an imperfect teacher that makes errors in its classifications.

1.1 SURVEY OF THE PATTERN RECOGNITION PROBLEM

Surveys of early work in pattern recognition have been written by Nagy [N-1] and Ho and Agrawala [H-1]. Nilsson [N-2] presented some of the early work on the theory of "learning machines," or machines that can be trained to recognize patterns. The idea of finding clustering transformations for designing pattern recognizers was developed by Sebestyen [S-3]. A book edited by Kanal [K-1] describes applications of pattern recognition to the problem of character recognition, and a book edited by Watanabe [W-1] is a very good collection of papers emphasizing the philosophy of various approaches. Sequential methods in statistical pattern recognition have been presented by Fu [F-4].

Ho and Agrawala [H-1] identify three fundamental problems associated with pattern recognition: characterization, abstraction, and generalization. Characterization is concerned with the problem of selecting the measurements which should be taken on the objects and of developing methods for reducing these measurements to a set of real variables which effectively characterize the objects and are amenable to

automatic data processing. The set of real numbers obtained from the measurements on an object is called a pattern, and each element of the pattern is called a feature. Each of the states of nature or sources of patterns is said to be a pattern class. At present there is no unifying theory for the selection of features [I-2]. Much of the problem of feature selection is left to the ingenuity of the system designer [I-1].

After the features have been selected, the designer, using all available information, must devise a decision procedure for classifying a new pattern of unknown origin. The process of deriving a decision rule from the available information is called abstraction. The ability of the resulting decision rule to correctly classify new patterns is termed generalization. This quality is best stated in probabilistic terms, such as the probability of correct classification.

These three aspects of the pattern recognition problem are not entirely independent of each other. Clearly, the choice of improper features manifests itself in unduly complex forms for the decision rule with poor generalization ability. Similarly, the ability to generalize may be the criterion for choosing the features. In this thesis it will be assumed that the features have been judiciously chosen. This research is concerned with a particular abstraction problem and the generalization ability of the resulting algorithms.

Two primary approaches have been taken to the abstraction problem in the literature [B-1]: namely, a deterministic approach and a statistical one. In the deterministic approach, the goal of the recognition system is to partition the feature space into regions such that each region can be identified with a pattern class. A functional form is

often assumed for the decision function, and unknown parameters are derived from a set of classified training patterns. Many of the deterministic procedures take the form of error correction algorithms or gradient descent algorithms. Blaydon [B-1], Nilsson [N-2], and Ho and Kashyap [H-2] have described such methods.

This thesis is concerned with the statistical approach to the abstraction problem. In this approach the mathematical tools of statistical decision theory [F-2] are applied to the problem of designing the classifier. The pattern features are assumed to be described by probability distributions; and optimal decision rules are obtained to satisfy certain classification criteria; for example, minimum average risk. A recent book by Fukunaga [F-5] presents this statistical approach.

The problem of learning arises during the abstraction process when the probability distributions characterizing the pattern classes are inadequately known. The unknowns of the class distributions must be estimated or "learned" from training patterns drawn from each of the pattern classes. When the origins of the training patterns are known, the learning is said to take place with a perfect teacher (supervised learning); but when the classification of the training patterns is unknown, the learning is called unsupervised (without a teacher).

Abramson [A-1] and Spragins [S-9] have studied the convergence question in supervised learning when only a few unknown parameters need to be learned. Aizerman, et. al. [A-2] developed the method of potential functions for supervised learning, and Van Ryzin [V-1], [V-2] has shown convergence of a procedure that estimates complete density functions via Parzen type [P-1], [C-1] density estimators.

Most of the practical learning procedures that have been developed use supervised learning. Theoretical foundations for unsupervised learning have been established, but practical algorithms have been slow in coming. Fralick [F-3], Patrick and Hancock [P-2], and Yakowitz [Y-1] have discussed various theoretical aspects of unsupervised learning.

A third type of learning, which has received little attention in the literature, is studied in this thesis. This type may be called "learning with an imperfect teacher" and lies somewhere in between supervised and unsupervised learning.

1.2 THE IMPERFECT TEACHER

In many practical situations it is unreasonable to assume that all of the training patterns supplied to a learning system by a "teacher" are correctly classified. The teacher often classifies the training patterns using past experience and additional information not available to the learning system. But even this additional information may not be adequate to ensure that the teacher's classifications are all correct.

As an example, a case has been reported in the literature [M-3] in which two groups of electrocardiographers reading the same set of 561 EKG's disagreed in over 40 percent of the normal-abnormal classifications. If one were to attempt to use this set of EKG's for supervised learning, he is confronted with a dilemma. Which group's classifications should be used? It is clearly questionable to attempt to perform supervised learning with such unreliable data. One might solve the problem by using only those EKG's for which both groups agreed. But this is a wasteful procedure since time and money are involved in recording

and reading each EKG. Another solution might be to use further tests and clinical records from each patient as an aid for classifying each EKG. Again this may be a costly and time consuming undertaking with no guarantee of success. Thus it is reasonable to talk about an "imperfect teacher" and consider situations in which the training patterns may not all be correctly classified.

Estimation and decision making with misclassified data has received some attention in the literature. Bross [B-2] considered the effects of misclassified data on estimators and significance tests involving binomial distributions. Tenenbein [T-1] extended this work to investigate the effects of estimation with misclassified multinomial data. He presented a double sampling scheme that minimized a measurement cost.

Linear classifiers have been used in many pattern recognition problems. Lachenbruch [L-1] looked at the effects of misclassification when learning a linear discriminant function for a Gaussian classification problem. He exhibited the asymptotic effects of estimating the mean and variance of the normal distribution from training samples that had a constant probability of being misclassified. Whitney and Dwyer [W-2] derived the asymptotic performance of the k-nearest neighbor rule [C-2] which used training patterns classified by an imperfect teacher.

Most of the research involving imperfect teachers has been concerned with analyzing the performance of existing supervised learning algorithms when some of the training samples are misclassified. Recently Shanmugam [S-4], [S-5] developed an error correction procedure for learning with an imperfect teacher. Shanmugam studied a nonparametric learning scheme that was asymptotically optimal in the sense that it had

an average risk which converged to the Bayes risk. He considered only a two-class problem, used a Bayes decision rule with a zero-one loss function, and assumed that the teacher had the same rate of misclassification for each pattern class. He proposed a threshold feedback scheme for gradually phasing out the teacher in the case when the class densities had disjoint support.

The work in this thesis removes some of the assumptions needed for the convergence of Shanmugam's algorithms. An alternative and more general approach to learning with an imperfect teacher is developed herein. The procedures considered are analogous to the type considered by Van Ryzin [V-1], [V-2] for supervised learning. These procedures are nonparametric in the sense that minimal knowledge about the class distributions is assumed to be available to the system designer.

1.3 THESIS CONTRIBUTIONS

The emphasis of this research is on developing and evaluating a procedure for learning with an imperfect teacher. The first contribution of this thesis appears in Chapter II which contains a mathematical model of the learning problem. A probabilistic model for an imperfect teacher is proposed and expressions are developed relating the imperfect teacher to a perfect teacher. These basic relations provide the key for studying the class of learning procedures proposed in Chapter III. An example is also presented in Chapter II to illustrate the effects of using misclassified training patterns in an algorithm designed for supervised learning. The example provides motivation for developing algorithms specifically for use with an imperfect teacher.

The second major contribution appears in Chapter III. A specific class of nonparametric procedures is proposed for learning to recognize patterns with an imperfect teacher. Formal proofs are given showing that the procedures are asymptotically optimal in the sense that they have expected risks which converge with increasing number of training patterns to the optimal (Bayes) risk. The conditions under which the convergence holds are very general for the two cases considered: i.) the case of discrete valued patterns and ii.) the case of patterns with a.e. continuous densities. Theorems on rates of convergence are also obtained. These rate theorems serve as guidelines for selecting the parameters for the estimators.

The final contributions of the thesis are presented in Chapter IV where attention is restricted to the two-class recognition problem. The purpose of Chapter IV is to investigate the quantitative and qualitative effects of using an imperfect teacher rather than a perfect one. Two large-sample approximations are developed for evaluating the expected risk of the estimated decision rules. Qualitative effects of the imperfect teacher are studied in section 4.3 by considering examples involving normal, triangular, and binomial distributions.

In the latter part of Chapter IV, a measure of relative performance is proposed for quantitatively evaluating the effects of the imperfect teacher. This measure, which can be evaluated for the case of a zero-one loss function, is used along with a cost of training to evaluate an overall cost of an imperfect teacher. The results in Chapter IV are original and represent a contribution to the study of learning theory.

CHAPTER II

THE LEARNING PROBLEM

This chapter presents the mathematical model to be used to study the problem of learning with an imperfect teacher. The problem is formulated in terms of statistical decision theory. A Bayesian decision strategy is employed throughout.

In Section 2.1, the basic model for statistical pattern recognition is defined. The Bayes decision rule and associated Bayes risk are given for the M-class problem. In Section 2.2, a model is proposed for an imperfect teacher. Section 2.3 then develops some fundamental relations between the imperfect teacher and a perfect teacher. Finally, Section 2.4 investigates the asymptotic effects of using an imperfect teacher with a procedure designed for supervised learning.

2.1 MATHEMATICAL MODEL FOR PATTERN RECOGNITION

The major objective of pattern recognition is to derive decision rules for classifying a pattern X as coming from one of M possible sources or pattern classes. In the statistical approach to pattern recognition the sources are characterized by probability distributions, and decision theoretic strategies are used for decision making.

The mathematical model for pattern classification to be used in this research will now be defined. The pattern X is defined to be a

vector random variable taking values in a feature space T, a subset of Euclidean n-space. The set of pattern classes is denoted as $\Omega = \{\omega_1, \omega_2, \ldots, \omega_M\}$ with ω_i representing the ith pattern class.

The pattern classes are assumed to be characterized by a prior probability distribution $P = (P_1, P_2, \ldots, P_M)^T$ where P_i is the prior probability of occurrence of pattern class ω_i. Let Λ be an identification random variable defined over Ω such that $\Lambda(\omega_i) = i$. The probability distribution of Λ is then just the prior distribution P. The probability density function of the pattern X given that X is from pattern class ω_i is denoted by $f(\cdot|\Lambda = i)$. This is a density with respect to some measure ν on T. Denote the vector of conditional densities by $f(\cdot) = (f(\cdot|\Lambda = 1), f(\cdot|\Lambda = 2), \ldots, f(\cdot|\Lambda = M))^T$.

After observing a value of the pattern random variable X, the pattern recognizer is required to classify X as having come from one of the M possible pattern classes ω_i with density $f(x|\Lambda = i)$. The random variable Λ indicating the class that produced X is, however, unobservable. The general elements of this statistical decision making problem are outlined in Appendix A. A Bayesian decision strategy is used to obtain optimal decision rules.

The Bayes decision rule, which follows from (A.5) of Appendix A, is any randomized decision rule $\delta_B(x) = (\delta_{B1}(x), \delta_{B2}(x), \ldots, \delta_{BM}(x))$ satisfying

$$\delta_{Bj}(x) = \begin{cases} 1 & \text{if } D_j(x) < \min\limits_{k \neq j} D_k(x) \\ 0 & \text{if } D_j(x) > \min\limits_{k \neq j} D_k(x), \\ \gamma_j & \text{if } D_j(x) = \min\limits_{k \neq j} D_k(x) \end{cases} \qquad j = 1, 2, \ldots, M \qquad (2.1)$$

where the Bayes discriminant function $D_k(x)$ is given by

$$D_k(x) = \sum_{\lambda=1}^{M} L_{k\lambda} P_\lambda f(x|\Lambda = \lambda); \qquad k = 1, 2, \ldots, M. \tag{2.2}$$

The function $\delta_{Bj}(x)$ is interpreted as the conditional probability of classifying the pattern X as coming from pattern class ω_j given that the observed value of the pattern is X = x. The loss function $L_{k\lambda}$ represents numerical loss incurred by deciding X is from pattern class ω_k when the true pattern class is ω_λ. The loss function is a nonnegative, real-valued function that is assumed to satisfy the condition $L_{ij} > L_{jj} \geq 0$, $i \neq j$.

The Bayes rule minimizes the average loss of misclassification. This minimum average loss, known as the Bayes risk, follows from (A.7) and is given by

$$\begin{aligned} R_B(P, f) &= \sum_{\lambda=1}^{M} P_\lambda \int_T \sum_{j=1}^{M} L_{j\lambda} f(x|\Lambda = \lambda) \delta_{Bj}(x)\, d\nu(x) \\ &= \sum_{j=1}^{M} \int_T D_j(x) \delta_{Bj}(x)\, d\nu(x). \end{aligned} \tag{2.3}$$

The notation for R_B in (2.3) emphasizes the dependence of the Bayes risk on the prior distribution P and the class-conditional densities f. This notation will be of use in later chapters.

The optimal Bayes rule can be used for decision making only if one has exact knowledge of the class-conditional densities and of the prior distribution. When these distributions are known, the abstraction and generalization problems are essentially solved. The system designer needs only to implement the decision rule of (2.1). The resulting system performance will be given by the Bayes risk in (2.3).

But in practice, the required distributions are usually only partially known. The unknowns of the class distributions must be learned (estimated) from training patterns drawn from each of the pattern classes. The pattern recognition system must use the training patterns provided by a (imperfect) teacher to abstract a decision rule for classifying new patterns of unknown origin. One would hope that the learned decision rules adapt or converge with increasing number of training patterns to what the optimal rule would be if the true probability distributions were known.

2.2 MODEL OF THE IMPERFECT TEACHER

The training patterns to be used for learning a decision rule are represented by a sequence of independent, identically distributed random variables $Y^N = (Y_1, Y_2, \ldots, Y_N)$. Each random variable $Y_i = (X_i, \Lambda_i', \Lambda_i)$ consists of a training pattern X_i, a label Λ_i' provided by an imperfect teacher, and an identification random variable Λ_i representing the true classification of X_i. Only X_i and Λ_i' are available to the learning system; Λ_i remains unknown. The pattern random variable X_i is assumed to be distributed as $f(\cdot|\Lambda = \lambda)$ if the correct classification of X_i is $\Lambda_i = \lambda$. The identification random variables are identically distributed according to the prior distribution P.

The labels provided by the imperfect teacher are modeled as identically distributed random variables on Ω having a probability distribution defined by

$$\Pr[\Lambda_i' = j | \Lambda_i = k] = \beta_{jk} \tag{2.4}$$

with $\beta_{jk} \geq 0$, $\sum_{j=1}^{M} \beta_{jk} = 1$. The probabilities β_{jk} are assumed not to be

a function of i or of the value of the training pattern X_i. An M × M matrix of probabilities characterizing the imperfect teacher may be defined by

$$\underline{\beta} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1M} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2M} \\ \vdots & & & \\ \beta_{M1} & \beta_{M2} & \cdots & \beta_{MM} \end{bmatrix}. \tag{2.5}$$

This simple probabilistic model of an imperfect teacher is one that is justifiable in many practical situations. The model is an extension of one used by Whitney and Dwyer [W-2] and by Shanmugam [S-5]. When $\beta_{ii} = 1$ for all i, $\Lambda_i' = \Lambda_i$ with probability one, in which case the model reduces to one of a perfect teacher. When $\beta_{ij} = 1/M$ for all i and j, the model corresponds to unsupervised learning since having a teacher that supplies equally likely labels for all pattern classes is equivalent to having no teacher at all. There is no classification information available in the labels.

2.3 SOME FUNDAMENTAL RELATIONS

Several expressions can be derived to relate the probability distributions associated with an imperfect teacher to those for a perfect teacher. The labels of the perfect teacher are distributed as the identification random variable Λ, and the class-conditional densities with the perfect teacher are $f(\cdot|\Lambda)$. With an imperfect teacher, the density of X conditioned on Λ' is given for j = 1, 2, ..., M by

$$f(x|\Lambda' = j) = \sum_{k=1}^{M} f(x|\Lambda' = j, \Lambda = k)\Pr[\Lambda = k|\Lambda' = j]. \tag{2.6}$$

But

$$f(x|\Lambda' = j, \Lambda = k) = \frac{Pr[\Lambda' = j|x, \Lambda = k]f(x|\Lambda = k)}{Pr[\Lambda' = j|\Lambda = k]}$$

$$= f(x|\Lambda = k) \tag{2.7}$$

where the last equality follows from the assumption that β_{jk} is not a function of the value of X. Also Bayes rule gives

$$Pr[\Lambda = k|\Lambda' = j] = \frac{Pr[\Lambda' = j|\Lambda = k]Pr[\Lambda = k]}{\sum_{\lambda=1}^{M} Pr[\Lambda' = j|\Lambda = \lambda]Pr[\Lambda = \lambda]}$$

$$= \beta_{jk}P_k/P'_j, \qquad j, k = 1, 2, \ldots, M. \tag{2.8}$$

Here P'_j is the probability that Λ' equals j,

$$P'_j = Pr[\Lambda' = j]$$

$$= \sum_{i=1}^{M} \beta_{ji}P_i, \qquad j = 1, 2, \ldots, M. \tag{2.9}$$

Substitution of (2.7) and (2.8) into (2.6) then gives

$$P'_j f(x|\Lambda' = j) = \sum_{k=1}^{M} \beta_{jk}P_k f(x|\Lambda = k), \qquad j = 1, 2, \ldots, M. \tag{2.10}$$

Equations (2.9) and (2.10) can be expressed in a convenient vector form by defining the following vectors and matrices:

$$f'(x) = (f(x|\Lambda' = 1), \ldots, f(x|\Lambda' = M))^T \tag{2.11a}$$

$$P' = (P'_1, P'_2, \ldots, P'_M)^T \tag{2.11b}$$

$$\underline{P} = \text{diag}\ (P_1, P_2, \ldots, P_M) \tag{2.11c}$$

and

$$\underline{P}' = \text{diag}\ (P'_1, P'_2, \ldots, P'_M). \tag{2.11d}$$

Equation (2.10) then becomes

$$\underline{P}'f'(x) = \underline{\beta}\ \underline{P}\ f(x) \tag{2.12}$$

and (2.9) has the form

$$P' = \underline{\beta}\, P. \tag{2.13}$$

When $\underline{\beta}$ is nonsingular, it may be inverted to obtain from (2.12) and (2.13)

$$P = \underline{\beta}^{-1} P' \tag{2.14}$$

and

$$\underline{P} f(x) = \underline{\beta}^{-1} \underline{P}' f'(x). \tag{2.15}$$

The expressions (2.12) thru (2.15) describe the relations between the probability distributions for the perfect teacher and those for the imperfect teacher. Equations (2.14) and (2.15) provide the key for developing and studying in Chapter III and Chapter IV an algorithm for learning with an imperfect teacher. Previous studies with imperfect teachers [W-2], [S-4], [S-5] have used (2.10) for the case of two pattern classes (M = 2), but none of these studies have made use of the inverse relations in (2.14) and (2.15). Shanmugam [S-4], [S-5] avoided the need to use the inverse relations by restricting the density functions to having disjoint supports and by using a zero-one loss function along with a known prior distribution. The inverse relations remove the need for such restrictive assumptions.

2.4 EFFECTS OF MISCLASSIFICATIONS ON SUPERVISED LEARNING PROCEDURES

Many algorithms have been developed for supervised learning. Several of these algorithms have also been used with an imperfect teacher [S-5], [W-2], [L-1]. For the problem considered in [S-5], it was shown that misclassified training patterns could be used in the supervised learning procedure and that the resulting decision rule would

still converge to the Bayes rule. The use of misclassified data did not prevent the algorithm from converging to the desired decision rule. If this were always the case, one would have little motivation for developing learning procedures which take into account the imperfect teacher. One could just ignore the fact that the data was misclassified and use existing algorithms which were designed for supervised learning.

An example presented in this section shows that, as one would expect, misclassified training data can significantly effect the convergence of a supervised learning algorithm. This example provides motivation for developing learning procedures for use with an imperfect teacher.

The learning procedures discussed in [S-5] and [V-1], as well as those proposed in Chapter III, are based on nonparametric estimators of the class-conditional densities and prior distribution. In [V-1], the training patterns with label k were used to form estimates $\hat{f}(x|\Lambda = k)$ and $\hat{P}_k$ of $f(x|\Lambda = k)$ and P_k, respectively. An estimate of the Bayes discriminant function D_k was then defined as

$$\hat{D}_k(x) = \sum_{j=1}^{M} L_{kj}\, \hat{P}_j \hat{f}(x|\Lambda = j). \tag{2.16}$$

A decision rule $\hat{\delta}$ was formed according to (2.1) by replacing D_k with the estimate $\hat{D}_k$. It was shown in [V-1] that, under suitable conditions, $\hat{\delta}$ converged to the Bayes decision rule δ_B.

Now if the training patterns were misclassified, one would be estimating $f(x|\Lambda' = k)$ and P'_K, not $f(x|\Lambda = k)$ and P_k, when using the patterns with label k. The resulting decision rule would, at best, converge to a decision rule δ' having the form of (2.1) but with

$$D_k'(x) = \sum_{j=1}^{M} L_{kj}\, P_j' f(x|\Lambda' = j), \qquad k = 1, 2, \ldots, M \tag{2.17}$$

as the discriminant functions. This decision rule would be equivalent to the Bayes rule only if the risk (see equation (A.2))

$$R(P, f, \delta') = \sum_{j=1}^{M} \int D_j(x)\, \delta_j'(x)\, d\nu(x) \tag{2.18}$$

were equal to the Bayes risk $R_B(P, f)$.

For the problem considered in [S-4], [S-5], $R(P, f, \delta')$ and $R_B(P, f)$ were equal. Hence for that problem, the misclassified data did not alter the convergence of the learning procedure. Shanmugam [S-5] has shown that for the M-class problem, δ' and δ_B are equivalent if the following conditions are satisfied:

a.) A zero-one loss function is used; i.e.,

$L_{ii} = 0$ and $L_{ij} = 1, \quad i \neq j.$

b.) The probabilities characterizing the imperfect teacher are

$$\beta_{ii} = \beta > 1/M, \qquad i = 1, 2, \ldots, M \tag{2.19}$$

and

$$\beta_{ij} = \frac{1 - \beta}{M - 1}, \qquad i, j = 1, 2, \ldots, M, \quad i \neq j. \tag{2.20}$$

Under these conditions, supervised learning algorithms based on density estimators as in [V-1] will perform asymptotically just as well with an imperfect teacher as with a perfect teacher. The following example shows what happens when either condition (a.) or (b.) is not satisfied.

Suppose that there are two pattern classes ω_1 and ω_2, and assume that under pattern class ω_j the pattern X has a normal distribution with mean μ_j and variance σ^2,

$$f(x|\Lambda = j) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\,(-(x - \mu_j)^2/2\sigma^2). \tag{2.21}$$

For convenience assume that $\mu_2 > \mu_1$.

The decision rule $\delta' = (\delta_1', \delta_2')$ obtained with the imperfect teacher is given by (2.1) using the discriminant functions D_k' in place of D_k. The rule is $\delta_1'(x) = 1$ if

$$L_{11}P_1'f(x|\Lambda' = 1) + L_{12}P_2'f(x|\Lambda' = 2) \leq L_{21}P_1'f(x|\Lambda' = 1) + L_{22}P_2'f(x|\Lambda' = 2), \tag{2.22}$$

and $\delta_2'(x) = 1 - \delta_1'(x)$. Making use of (2.12) one can show that (2.22) simplifies to $\delta_1'(x) = 1$ if either $x \geq \tau$, $\rho > 0$ or $x \leq \tau$, $\rho < 0$, where

$$\rho = -(L_{21} - L_{11})\beta_{12} + (L_{12} - L_{22})\beta_{22} \tag{2.23}$$

and

$$\tau = \frac{\sigma^2}{\mu_2 - \mu_1} \ln\left[\frac{P_1}{P_2}\,\frac{(L_{21} - L_{11})\beta_{11} - (L_{12} - L_{22})\beta_{21}}{(L_{12} - L_{22})\beta_{22} - (L_{21} - L_{11})\beta_{12}}\right] + \frac{\mu_1 + \mu_2}{2}. \tag{2.24}$$

The risk incurred in using δ' follows from (2.18). One can show that (2.18) can be evaluated as

$$R(P, f, \delta') = \begin{cases} P_1L_{21} + P_2L_{22} - A & \text{if} \quad \rho > 0 \\ P_1L_{11} + P_2L_{12} + A & \text{if} \quad \rho < 0 \end{cases} \tag{2.25}$$

where

$$A = (L_{21} - L_{11})P_1\Phi(\gamma_1) + (L_{12} - L_{22})P_2\Phi(\gamma_2) \tag{2.26}$$

$$\gamma_1 = \frac{\tau - \mu_1}{\sigma} = \frac{1}{S} \ln\left[\frac{P_1}{P_2}\,\frac{(L_{21} - L_{11})\beta_{11} - (L_{12} - L_{22})\beta_{21}}{(L_{12} - L_{22})\beta_{22} - (L_{21} - L_{11})\beta_{12}}\right] + \frac{S}{2} \tag{2.27}$$

$$\gamma_2 = (\tau - \mu_2)/\sigma = \gamma_1 - S \tag{2.28}$$

$$S = (\mu_2 - \mu_1)/\sigma \tag{2.29}$$

and

$$\Phi(a) = \int_{-\infty}^{a} (2\pi)^{-\frac{1}{2}} \exp\ (-t^2/2)\ dt. \qquad (2.30)$$

The Bayes decision rule and the Bayes risk may be found for this example by setting $\beta_{11} = \beta_{22} = 1$ in the above equations. It is clear from the above equations that in general $R(P, f, \delta')$ will vary with β_{11} and β_{22}. Hence the rule δ' is not always equivalent to the Bayes rule.

In Figure 2.1 the risk $R(P, f, \delta')$ is plotted as a function of β_{22} for the case of a zero-one loss function and equal prior probabilities. The figure shows that for small β_{22}, the risk exceeds the Bayes risk by a significant amount. In this case, estimating the density functions without compensating for the misclassified data leads to a decision rule which is not a Bayes rule. Thus, one is motivated to develop a learning procedure for use with an imperfect teacher.

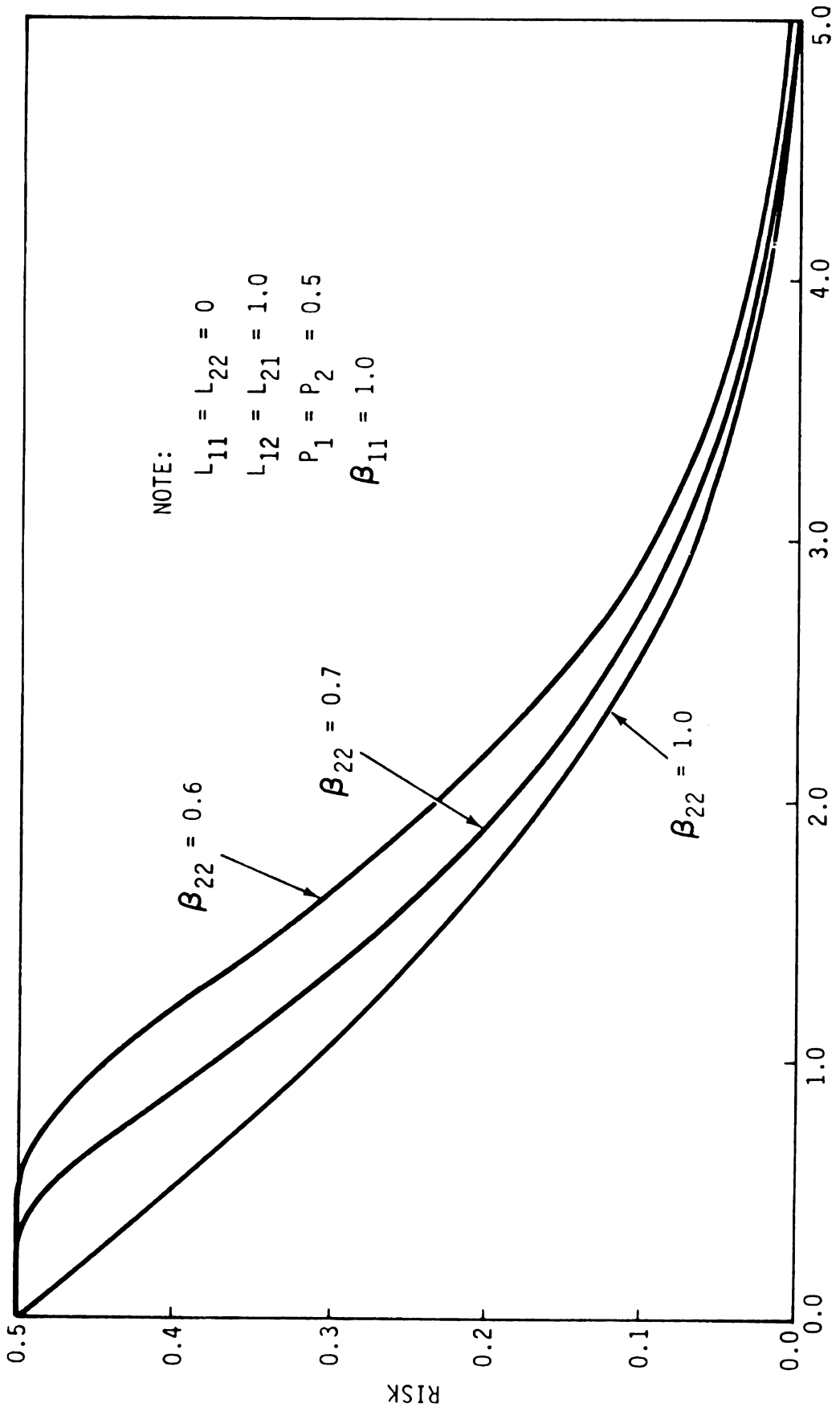

Figure 2.1. Asymptotic Risk with an Imperfect Teacher

CHAPTER III

A NONPARAMETRIC LEARNING PROCEDURE

In this chapter, a class of nonparametric procedures is developed for learning with an imperfect teacher. The specific problem considered is that of classifying a pattern pair (X, Λ), X observable and Λ unobservable, based on a set of training data. The training patterns are assumed to be generated by an imperfect teacher as described in Section 2.2.

The method used to learn the decision rule is essentially one of estimating unknown density functions with Parzen type estimators. These density estimators, which are summarized in Appendix B, were first proposed by Parzen [P-1] and later studied by Cacoullos [C-1], Murthy [M-4], Van Ryzin [V-3], and others. Van Ryzin [V-1], [V-2] used these estimators to develop a nonparametric scheme for supervised learning. Procedures analogous to those given by Van Ryzin for supervised learning are developed in this chapter for learning with an imperfect teacher.

The general form of the learning procedures and the resulting decision rules are presented in Section 3.1. Convergence criteria are discussed in Section 3.2. In Section 3.3 a preliminary lemma is presented for proving convergence of the classification procedures. The exact conditions under which convergence holds are examined in two cases:

a.) the discrete case (Section 3.4) in which the feature space T is countable and ν is counting measure; and

b.) the continuous case (Section 3.5) in which T is Euclidean n-space, ν is n-dimensional Lebesgue measure, and the class-conditional densities $f(\cdot|\Lambda = \lambda)$ are continuous a.e.ν.

In both cases, theorems are proved that give rates of convergence for the algorithms.

The learning procedures proposed in this chapter are nonparametric in the sense that they only require knowledge of whether the densities are of case (a.) or case (b.). The prior probabilities are also taken to be unknown. The $\underline{\beta}$ matrix describing the imperfect teacher is assumed to be nonsingular and known. The methods of proof follow those of Van Ryzin [V-1].

3.1 DECISION RULES

Let $\{(X_1, \Lambda_1'), (X_2, \Lambda_2'), \ldots, (X_N, \Lambda_N')\}$ be a sequence of training data from an imperfect teacher as defined in Section 2.2. Using these training patterns, one is required to form a decision rule for classifying a pattern X of unknown origin.

The Bayes discriminant functions in (2.2) may be rewritten as

$$D_j(x; v) = \sum_{k=1}^{M} L_{jk} v_k(x), \quad j = 1, 2, \ldots, M \tag{3.1}$$

where

$$v_k(x) = P_k f(x|\Lambda = k) \tag{3.2}$$

with v being the row vector

$$v = (v_1, v_2, \ldots, v_M). \tag{3.3}$$

From (2.1), the Bayes decision rule may be written as

$$\delta_B(x; v) = (\delta_{B1}(x; v), \delta_{B2}(x; v), \ldots, \delta_{BM}(x; v)) \tag{3.4}$$

where

$$\delta_{Bj}(x;\, v) = \begin{cases} 1 & \text{if} \quad D_j(x;\, v) < \min\limits_{k \neq j} D_k(x;\, v) \\ 0 & \text{if} \quad D_j(x;\, v) > \min\limits_{k \neq j} D_k(x;\, v) \\ \gamma_j & \text{if} \quad D_j(x;\, v) = \min\limits_{k \neq j} D_k(x;\, v) \end{cases} \tag{3.5}$$

The notation in (3.4) and (3.5) displays the dependence of the Bayes rule on the vector v (or, alternatively, on P and f).

If the training patterns are used to obtain an estimate $\hat{v}_N(x)$ of $v(x)$ and if $\hat{v}_N$ is a good approximation to v, then a reasonable decision rule may be formed by using $\hat{v}_N$ in (3.4) as if it were the true v. The resulting decision rule is

$$\hat{\delta}_N(x) = \delta_B(x;\, \hat{v}_N) = (\hat{\delta}_{N1}(x),\, \ldots,\, \hat{\delta}_{NM}(x)) \tag{3.6}$$

with

$$\hat{\delta}_{Nj}(x) = \begin{cases} 1 & \text{if} \quad \hat{D}_{Nj}(x) > \min\limits_{k \neq j} \hat{D}_{Nk}(x) \\ 0 & \text{if} \quad \hat{D}_{Nj}(x) < \min\limits_{k \neq j} \hat{D}_{Nk}(x) \\ \gamma_j & \text{if} \quad \hat{D}_{Nj}(x) = \min\limits_{k \neq j} \hat{D}_{Nk}(x) \end{cases} \tag{3.7}$$

and with the estimates of the Bayes discriminant functions given by

$$\begin{aligned} \hat{D}_{Nj}(x) &= D_j(x;\, \hat{v}_N) \\ &= \sum_{i=1}^{M} L_{ji} \hat{v}_{Ni}(x). \end{aligned} \tag{3.8}$$

The class of estimators of $v(x)$ proposed here is defined as follows. Let $\{g_N(x, y)\}$ be a sequence of nonnegative real-valued functions on $T \times T$ such that

$$\int_T g_N(x, y)\, dv(x) = 1 \quad \text{a.e.} v \tag{3.9}$$

and

$$\lim_{N\to\infty} E[g_N(x, X)|\Lambda = \lambda] = f(x|\Lambda = \lambda) \quad \text{a.e.}\nu. \tag{3.10}$$

Let

$$\Delta(i, j) = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j \end{cases} \tag{3.11}$$

be the Kronecker delta, and denote the elements of $\underline{\beta}^{-1}$ by b_{ij}, $\underline{\beta}^{-1} = [b_{ij}]$. Then the estimator $\hat{v}_N = (\hat{v}_{N1}, \hat{v}_{N2}, \ldots, \hat{v}_{NM})$ is defined by

$$\hat{v}_{Nj}(x) = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{M} b_{ji}\Delta(\Lambda_k', i)g_N(x, X_k). \tag{3.12}$$

Explicit expressions for the g_N functions will be exhibited in Section 3.4 for the discrete case and in Section 3.5 for the continuous case.

The form of (3.12) may be obtained from the following considerations. Suppose that N_i is the number of training patterns classified by the imperfect teacher as coming from pattern class ω_i. Then an estimate of $f(x|\Lambda' = i)$ is given by the Parzen estimator described in Appendix B,

$$\hat{f}(x|\Lambda' = i) = \frac{1}{N_i} \sum_{k=1}^{N} \Delta(\Lambda_k', i)g_{N_i}(x, X_k). \tag{3.13}$$

An estimate of the probability P_i is given by N_i/N. Combining these estimates according to the inverse relation (2.15) results in an estimator for $v(x)$ having the form in (3.12).

The decision rule in (3.6), along with the estimator in (3.12), forms the class of procedures for learning with an imperfect teacher to be studied in the remainder of this thesis. When $\beta_{ii} = 1$ for all i, these procedures reduce to those of Van Ryzin [V-1] for supervised learning.

3.2 CONVERGENCE CRITERIA

The risk incurred in using decision rule $\hat{\delta}_N$ is given by (A.2) as

$$R(v, \hat{\delta}_N) = \sum_{j=1}^{M} \int_T \left[\sum_{i=1}^{M} L_{ji} P_i f(x|\Lambda = i) \right] \hat{\delta}_{Nj}(x)\, d\nu(x)$$

$$= \sum_{j=1}^{M} \int_T D_j(x; v)\, \hat{\delta}_{Nj}(x)\, d\nu(x) \tag{3.14}$$

and the Bayes risk is given by (2.3) as

$$R_B(v) = \sum_{j=1}^{M} \int_T D_j(x; v)\, \delta_{Bj}(x; v)\, d\nu(x). \tag{3.15}$$

The dependence of the risks on the vector v is displayed by the notation in (3.14) and (3.15).

One would hope that as $N \to \infty$, the estimated rule $\hat{\delta}_N$ would in some sense converge to the Bayes rule δ_B. Now the risk $R(v, \hat{\delta}_N)$ is a random variable since it is a function of the training set. So the decision rule $\hat{\delta}_N$ is said to converge to the Bayes rule δ_B if the risk $R(v, \hat{\delta}_N)$ converges in some sense to the Bayes risk $R_B(v)$. The following definitions of convergence are due to Van Ryzin [V-2].

DEFINITION 3.1. A decision rule $\hat{\delta}_N$ is said to be Bayes Risk Consistent if for every $\varepsilon > 0$,

$$\Pr[R(v, \hat{\delta}_N) - R_B(v) \geq \varepsilon] \to 0 \quad \text{as} \quad N \to \infty; \tag{3.16}$$

i.e., if $R(v, \hat{\delta}_N)$ converges in probability to $R_B(v)$ as $N \to \infty$.

Let $E_N[\cdot]$ denote the expectation with respect to the distribution of the training data, $\{(X_1, \Lambda_1', \Lambda_1), \ldots, (X_N, \Lambda_N', \Lambda_N)\}$. Then define the expected risk for the learning procedure as

$$R_N(v) \triangleq E_N[R(v, \hat{\delta}_N)]. \tag{3.17}$$

R_N is a nonrandom variable which depends upon the vector v.

DEFINITION 3.2. A decision rule $\hat{\delta}_N$ is said to be Mean Risk Consistent if

$$\lim_{N\to\infty} R_N(v) = R_B(v). \tag{3.18}$$

Definition 3.2 is a special case of the following definition:

DEFINITION 3.3. A decision rule $\hat{\delta}_N$ is said to be Mean Risk Consistent of order $r > 0$ if

$$\lim_{N\to\infty} E_N[|R(v, \hat{\delta}_N) - R_B(v)|^r] = 0. \tag{3.19}$$

The following notation is defined for use in later sections:

$$\Delta R(v, \hat{\delta}_N) \triangleq R(v, \hat{\delta}_N) - R_B(v) \tag{3.20}$$

and

$$\Delta R_N(v) \triangleq R_N(v) - R_B(v). \tag{3.21}$$

So a decision rule $\hat{\delta}_N$ is Bayes Risk Consistent if $\Delta R(v, \hat{\delta}_N)$ converges to zero in probability, and the rule is Mean Risk Consistent if ΔR_N converges to zero in the ordinary sense of a limit.

Since convergence in the mean implies convergence in probability [L-2], Mean Risk Consistency implies Bayes Risk Consistency. A stronger relation between the various types of convergence is established by the following property.

Property 3.1. If a decision rule $\hat{\delta}_N$ is Mean Risk Consistent, then it is Mean Risk Consistent of order r for any $r > 0$.

PROOF: From (3.14), it follows that

$$|R(v, \hat{\delta}_N)| \leq \sum_{j=1}^{M} \sum_{i=1}^{M} \int_T |L_{ji} P_i f(x|\Lambda = i) \hat{\delta}_{Nj}(x)| \, d\nu(x)$$

$$\leq \sum_{j=1}^{M} \sum_{i=1}^{M} L_{ji} P_i \int_T f(x|\Lambda = i) \, d\nu(x)$$

$$= \sum_{j=1}^{M} \sum_{i=1}^{M} L_{ji} P_i < \infty. \tag{3.22}$$

So the sequence of random variables $\{R(v, \hat{\delta}_N)\}$ is a.s. uniformly bounded. Hence $R(v, \hat{\delta}_N)$ converges to $R_B(v)$ in the rth mean if it converges for r = 1 [L-2, p. 158].

End of proof.

In the following sections the estimated decision rules are shown to be Mean Risk Consistent. The above property then establishes that the rules are Mean Risk Consistent of order r for any r > 0.

3.3 PRELIMINARY LEMMAS

The two lemmas presented in this section are used to establish convergence theorems and rate theorems for the decision rule $\hat{\delta}_N$ in both the discrete case (Section 3.4) and the continuous case (Section 3.5).

Let $E[\cdot]$ denote expection with respect to the mixture distribution

$$p(x) = \sum_{i=1}^{M} P_i f(x|\Lambda = i)$$

$$= \sum_{i=1}^{M} P_i' f(x|\Lambda' = i). \tag{3.23}$$

The estimator $\hat{v}_N$ satisfies the following lemma:

LEMMA 3.1. Let $\hat{v}_{Nj}(x)$, j = 1, 2, ..., M, be defined by (3.12). Then

$$E_N[\hat{v}_{Nj}(x)] = P_j E[g_N(x, X)|\Lambda = j]$$

$$= P_j \int_T g_N(x, z) f(z|\Lambda = j)\, d\nu(z). \tag{3.24}$$

PROOF: Taking expectations in (3.12) gives

$$E_N[\hat{v}_{Nj}(x)] = \frac{1}{N} \sum_{k=1}^{M} \sum_{i=1}^{M} b_{ji} E_N[\Delta(\Lambda_k', i) g_N(x, X_k)]. \tag{3.25}$$

Since the training patterns are identically distributed according to the mixture distribution (3.23), equation (3.25) may be written as

$$\begin{aligned} E_N[\hat{v}_{Nj}(x)] &= \sum_{i=1}^{M} b_{ji} E[\Delta(\Lambda', i) g_N(x, X)] \\ &= \sum_{i=1}^{M} b_{ji} P_i' E[g_N(x, X) | \Lambda' = i] \\ &= P_j E[g_N(x, X) | \Lambda = j] \end{aligned} \tag{3.26}$$

where the last equality follows from (2.15).

End of proof.

The decision rule $\hat{\delta}_N$ satisfies the following lemma:

LEMMA 3.2. The average risk for the decision rule $\hat{\delta}_N$ defined by (3.7), (3.8), and (3.12) satisfies

$$\begin{aligned} 0 \leq \Delta R_N \leq & \sum_{i=1}^{M} \overline{L}_i \int E_N[|\hat{v}_{Ni}(x) - P_i E[g_N(x, X) | \Lambda = i]|] \, d\nu(x) \\ & + \sum_{j=1}^{M} \int \{E_N[\hat{D}_{Nj}(x)] - D_j(x)\} \cdot \{\delta_{Bj}(x; v) - E_N[\hat{\delta}_{Nj}(x)]\} \, d\nu(x) \end{aligned} \tag{3.27}$$

where

$$\overline{L}_i = \max_{1 \leq j \leq M} L_{ji} \tag{3.28}$$

and the integrals are over T.

PROOF: By the definition of Bayes risk

$$R_B(v) \leq R(v, \hat{\delta}_N) \tag{3.29}$$

so that $\Delta R_N(v) \geq 0$. Also

$$R(\hat{v}_N, \delta_B(x; v)) \geq R_B(\hat{v}_N). \tag{3.30}$$

So

$$\Delta R_N \leq E_N[R(v, \hat{\delta}_N)] - R_B(v) + E_N[R(\hat{v}_N, \delta_B(x; v)) - R_B(\hat{v}_N)]$$

$$= E_N[R(v, \hat{\delta}_N) - R_B(\hat{v}_N)] + E_N[R(\hat{v}_N, \delta_B(x; v)) - R_B(v)]. \tag{3.31}$$

Consider the first expectation in (3.31) and use (3.14) to get

$$E_N[R(v, \hat{\delta}_N) - R_B(\hat{v}_N)] = E_N[\sum_{j=1}^{M} \int \{D_j(x) - \hat{D}_{Nj}(x)\}\hat{\delta}_{Nj}(x)\, d\nu(x)]$$

$$= \sum_{j=1}^{M} \int E_N[(\hat{D}_{Nj}(x) - E_N[\hat{D}_{Nj}(x)])(-\hat{\delta}_{Nj}(x)]\, d\nu(x)$$

$$+ \sum_{j=1}^{M} \int (D_j(x) - E_N[\hat{D}_{Nj}(x)])\, E_N[\hat{\delta}_{Nj}(x)]\, d\nu(x). \tag{3.32}$$

The first integral in (3.32) can be bounded as follows using Lemma 3.1:

$$\sum_{j=1}^{M} \int E_N[(\hat{D}_{Nj}(x) - E_N[\hat{D}_{Nj}(x)])(-\hat{\delta}_{Nj}(x))]\, d\nu(x)$$

$$= \sum_{i=1}^{M} \int E_N[\{\sum_{j=1}^{M} L_{ji}\hat{\delta}_{Nj}(x)\}\{P_i E[g_N(x, X)|\Lambda = i] - \hat{v}_{Ni}(x)\}]\, d\nu(x)$$

$$\leq \sum_{i=1}^{M} \int E_N[|\cdot|]\, d\nu(x)$$

$$\leq \sum_{i=1}^{M} \int E_N[\overline{L}_i |P_i E[g_N(x, X)|\Lambda = i] - \hat{v}_{Ni}(x)|]\, d\nu(x) \tag{3.33}$$

where the last inequality follows from the fact that

$$\sum_{j=1}^{M} L_{ji}\hat{\delta}_{Nj}(x) \leq \overline{L}_i \sum_{j=1}^{M} \hat{\delta}_{Nj}(x) = \overline{L}_i. \tag{3.34}$$

So then substituting from (3.33) and (3.32) into (3.31) gives

$$\Delta R_N \leq \sum_{i=1}^{M} \overline{L}_i \int E_N[|P_i E[g_N(x, X)|\Lambda = i] - \hat{v}_{Ni}(x)|] \, d\nu(x)$$

$$+ \sum_{j=1}^{M} \int (D_j(x) - E_N[\hat{D}_{Nj}(x)]) \, E_N[\hat{\delta}_{Nj}(x)] \, d\nu(x)$$

$$+ \sum_{j=1}^{M} \int E_N[\hat{D}_{Nj}(x) - D_j(x)] \, \delta_{Bj}(x; v) \, d\nu(x)$$

$$= \sum_{i=1}^{M} \overline{L}_i \int E_N[|P_i E[g_N(x, X)|\Lambda = i] - \hat{v}_{Ni}(x)|] \, d\nu(x)$$

$$+ \sum_{j=1}^{M} \int \{E_N[\hat{D}_{Nj}(x)] - D_j(x)\}\{\delta_{Bj}(x; v) - E_N[\hat{\delta}_{Nj}(x)]\} \, d\nu(x). \qquad (3.35)$$

End of proof.

3.4 THE DISCRETE CASE

For the discrete case the observed random variable X takes values in a countable set T, and ν is counting measure. The class-conditional densities are probability distributions on T,

$$f(x|\Lambda = j) = Pr[X = x|\Lambda = j], \quad j = 1, 2, \ldots, M. \qquad (3.36)$$

The g_N functions for this case are defined by

$$g_N(x, y) = g(x, y) = \begin{cases} 1 & \text{if} \quad x = y \\ 0 & \text{if} \quad x \neq y \end{cases} \qquad (3.37)$$

so that

$$E[g_N(x, y)|\Lambda = j] = f(x|\Lambda = j). \qquad (3.38)$$

Note that with this g_N function, the estimator $\hat{f}(x|\Lambda' = i)$ in (3.13) is merely the emperical distribution of those X_k's for which $\Lambda_k' = i$.

The convergence of the decision rule $\hat{\delta}_N$ is established by the following theorem.

THEOREM 3.2. In the discrete case, if the decision rule $\hat{\delta}_N$ is defined by (3.7), (3.8), and (3.12) with g_N given by (3.37), then $\hat{\delta}_N$ is Mean Risk Consistent.

PROOF: Note that

$$\hat{v}_{Nj}(x) = \frac{1}{N} \sum_{k=1}^{N} Z_j(x, X_k) \tag{3.39}$$

is the average of a sequence of independent, identically distributed random variables, $\{Z_j(x, X_k)\}$, that are distributed as the random variable

$$Z_j(x, X) = \sum_{i=1}^{M} b_{ji} \Delta(\Lambda', i) g(x, X). \tag{3.40}$$

Since

$$E[Z_j^2(x, X)] = \sum_{i=1}^{M} b_{ji}^2 P_i' f(x|\Lambda' = i) < \infty, \tag{3.41}$$

the Kolmogorov strong law of large numbers [L-2] implies that for each $x \in T$

$$\hat{v}_{Nj}(x) \to E[Z_j(x, X)] \text{ w.p.1} \quad \text{as} \quad N \to \infty. \tag{3.42}$$

But from (3.40) and (2.15) it follows that

$$\begin{aligned} E[Z_j(x, X)] &= \sum_{i=1}^{M} b_{ji} P_i' f(x|\Lambda' = i) \\ &= P_j f(x|\Lambda = j) = v_j(x). \end{aligned} \tag{3.43}$$

Thus $\hat{v}_{Nj}(x)$ converges to $v_j(x)$ with probability one. Also $\hat{v}_{Nj}(x)$ is bounded as follows:

$$|\hat{v}_{Nj}(x)| = |\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{M} b_{ji}\Delta(\Lambda_k', i)g(x, X_k)|$$

$$\leq \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{M} |b_{ji}|\Delta(\Lambda_k', i)g(x, X_k)$$

$$\leq \max_{1\leq i\leq M} |b_{ji}| < \infty. \tag{3.44}$$

The w.p.1 convergence and boundedness guarantee that $\hat{v}_{Nj}(x) \to v_j(x)$ in the mean [L-2, p. 158]; i.e.,

$$\lim_{N\to\infty} E[|\hat{v}_{Nj}(x) - v_j(x)|] = 0 \quad \text{for all } x\varepsilon T. \tag{3.45}$$

Now

$$E_N[|\hat{v}_{Nj}(x) - v_j(x)|] \leq E_N[|\hat{v}_{Nj}(x)|] + v_j(x)$$

$$\leq \frac{1}{N}\sum_{k=1}^{N} E_N[|\sum_{i=1}^{M} b_{ji}\Delta(\Lambda_k', i)g(x, X_k)|] + v_j(x)$$

$$= \sum_{i=1}^{M} |b_{ji}|P_i'f(x|\Lambda' = i) + v_j(x)$$

$$\leq (\max_{1\leq i\leq M} |b_{ji}|)\sum_{i=1}^{M} P_i'f(x|\Lambda' = i) + v_j(x)$$

$$= (\max_{1\leq i\leq M} |b_{ji}|)\sum_{i=1}^{M} v_i(x) + v_j(x) = s(x). \tag{3.46}$$

Since $v_j(x)$ is integrable for all j, $s(x)$ is an integrable function. So the Legesgue dominated convergence theorem along with the above inequality implies that

$$\lim_{N\to\infty} \int E_N[|\hat{v}_{Nj}(x) - v_j(x)|]\, d\nu(x) = 0, \quad j = 1, 2, \ldots, M. \tag{3.47}$$

Thus the first term in the bound of Lemma 3.2 converges to zero. By (3.38), the second term of the bound is zero for all N. Hence

$$\lim_{N\to\infty} \Delta R_N = 0. \tag{3.48}$$

End of proof.

A rate of convergence of ΔR_N to zero is given by the following theorem.

THEOREM 3.3. Let

$$\rho = \max_{1\le j\le M} \int_{x\in T} [\sum_{i=1}^{M} (\sum_{\ell=1}^{M} b_{j\ell}^2 \beta_{\ell i}) v_i(x) - v_j^2(x)]^{\frac{1}{2}}. \tag{3.49}$$

If $\rho < \infty$, there exists a constant $H(v)$ such that

$$0 \le \Delta R_N \le H(v)/\sqrt{N}. \tag{3.50}$$

Furthermore,

$$H(v) \le \rho M^{\frac{1}{2}} ||\overline{L}||; \quad ||\overline{L}|| = (\sum_{i=1}^{M} \overline{L}_i^2)^{\frac{1}{2}}. \tag{3.51}$$

PROOF: Consider the bound given by Lemma 3.2. The second term of the bound is zero for the discrete case. So by (3.27) and (3.38) it follows that ΔR_N is bounded as

$$\Delta R_N \le \sum_{j=1}^{M} \overline{L}_j \int E_N[|\hat{v}_{Nj}(x) - v_j(x)|]\, d\nu(x) \tag{3.52}$$

The Schwarz inequality gives

$$E_N[|\hat{v}_{Nj}(x) - v_j(x)|] \le (E_N[(\hat{v}_{Nj}(x) - v_j(x))^2])^{\frac{1}{2}}. \tag{3.53}$$

Lemma 3.1 and (3.38) imply that

$$E_N[(\hat{v}_{Nj}(x) - v_j(x))^2] = E_N[\hat{v}_{Nj}^2(x)] - v_j^2(x). \tag{3.54}$$

From (3.12) it follows that

$$E_N[\hat{v}_{Nj}^2(x)] = \frac{1}{N^2} \sum_{k=1}^{N} \sum_{\ell=1}^{N} \sum_{m=1}^{M} \sum_{i=1}^{M} b_{jm} b_{ji}\, E_N[\Delta(\Lambda_k', m)\Delta(\Lambda_\ell', i)\, g(x, X_k) g(x, X_\ell)]. \tag{3.55}$$

When $k \neq \ell$, (X_k, Λ_k') and (X_ℓ, Λ_ℓ') are statistically independent. So

$$E_N[\hat{v}_{Nj}^2(x)] = \frac{1}{N^2} \sum_{k=1}^{N} \sum_{i=1}^{M} \sum_{m=1}^{M} b_{jm} b_{ji} \, E_N[\Delta(\Lambda_k', m)\Delta(\Lambda_k', i) \, g^2(x, X_k)]$$

$$+ \frac{1}{N^2} \sum_{\substack{k=1 \\ k \neq \ell}}^{N} \sum_{\ell=1}^{N} \sum_{i=1}^{M} \sum_{m=1}^{M} E_N[b_{jm}\Delta(\Lambda_k', m)g(x, X_k)] \cdot E_N[b_{ji}\Delta(\Lambda_\ell', i)g(x, X_\ell)]$$

$$= \frac{1}{N} \sum_{i=1}^{M} b_{ji}^2 P_i' f(x|\Lambda' = i)$$

$$+ \frac{(N-1)N}{N^2} \left(\sum_{i=1}^{M} b_{ji} P_i' f(x|\Lambda' = i) \right)^2 \qquad (3.56)$$

Substituting (2.15) into (3.56) gives

$$E[\hat{v}_{Nj}^2(x)] = \frac{1}{N} \sum_{i=1}^{M} b_{ji}^2 P_i' f(x|\Lambda' = i) + (1 - \frac{1}{N}) v_j^2(x) \qquad (3.57)$$

So

$$E_N[|\hat{v}_{Nj}(x) - v_j(x)|] \leq \frac{1}{\sqrt{N}} \left[\sum_{i=1}^{M} b_{ji}^2 P_i' f(x|\Lambda' = i) - v_j^2(x) \right]^{\frac{1}{2}}. \qquad (3.58)$$

Recalling that ν is counting measure, one obtains upon substitution of (3-58) into (3-52)

$$\Delta R_N \leq \sum_{j=1}^{M} \bar{L}_j \sum_{x \in T} \frac{1}{\sqrt{N}} \left[\sum_{i=1}^{M} b_{ji}^2 P_i' f(x|\Lambda' = i) - v_j^2(x) \right]^{\frac{1}{2}}. \qquad (3.59)$$

Define

$$H(v) \triangleq \sum_{j=1}^{M} \bar{L}_j \sum_{x \in T} \left[\sum_{i=1}^{M} b_{ji}^2 P_i' f(x|\Lambda' = i) - v_j^2(x) \right]^{\frac{1}{2}}$$

$$= \sum_{j=1}^{M} \bar{L}_j \sum_{x \in T} \left[\sum_{i=1}^{M} b_{ji}^2 \left(\sum_{\ell=1}^{M} \beta_{i\ell} P_\ell f(x|\Lambda = \ell) \right) - v_j^2(x) \right]^{\frac{1}{2}}$$

$$= \sum_{j=1}^{M} \bar{L}_j \sum_{x \in T} \left[\sum_{\ell=1}^{M} \left(\sum_{i=1}^{M} b_{ji}^2 \beta_{i\ell} \right) v_\ell(x) - v_j^2(x) \right]^{\frac{1}{2}}. \qquad (3.60)$$

Then

$$\Delta R_N \leq H(v)/\sqrt{N}. \tag{3.61}$$

The second conclusion of the theorem results by noting that

$$H(v) \leq \rho \sum \overline{L}_j \leq \rho M^{\frac{1}{2}} ||\overline{L}||. \tag{3.62}$$

End of proof.

3.5 THE CONTINUOUS CASE

For the continuous case the pattern vector X is taken to be a continuous random vector. Specifically, let the feature space T be Euclidean n-space and let ν be n-dimensional Lebesgue measure. Assume that the class-conditional densities $f(x|\Lambda = j)$ are continuous a.e.ν for $j = 1, 2, \ldots, M$.

The g_N functions to be used to form the estimator $\hat{v}_N$ have the form described in Appendix B. The g_N function is taken to be defined by

$$g_N(x, y) = \frac{1}{h^n} K\left(\frac{x - y}{h_N}\right) \tag{3.63}$$

such that

$$\lim_{N\to\infty} h_N = 0 \tag{3.64}$$

and

$$\lim_{N\to\infty} Nh_N^n = \infty. \tag{3.65}$$

The kernel $K(\cdot)$ is assumed to satisfy the conditions of (B.1). Examples of univariate kernels are given in Table B.1. One way of obtaining multivariate kernels is to use products of univariate kernels.

Lemma B.1 guarantees that

$$E[g_N(x, X)|\Lambda = j] \to f(x|\Lambda = j) \text{ a.e.}\nu \quad \text{as} \quad N \to \infty \tag{3.66}$$

and

$$h_N^n E[g_N^2(x, X)|\Lambda = j] \to Qf(x|\Lambda = j) \text{ a.e.}\nu \quad \text{as} \quad N \to \infty \tag{3.67}$$

where

$$Q = \int K^2(y)\, d\nu(y). \tag{3.68}$$

The asymptotic unbiasedness of (3.66) replaces the unbiased property (3.38) of the discrete densities.

The following lemma due to Van Ryzin [V-1] will be used to prove convergence of $\hat{\delta}_N$ for continuous densities.

LEMMA 3.3. Let (T', G', ν') be a measure space upon which is defined two sequences of integrable functions $\{q_N(x)\}$ and $\{q_N'(x)\}$ such that $|q_N(x)| \leq |q_N'(x)|$ a.e.ν'. If $q_N(x) \to q(x)$ in measure, $q_N'(x) \to q_N'(x)$ in the mean, and $q(x)$ and $q'(x)$ are integrable, then $q_N(x) \to q(x)$ in the mean.

Theorem 3.4 establishes the convergence of the decision procedures for continuous densities.

THEOREM 3.4. In the continuous case, if the decision rule $\hat{\delta}_N$ is defined by (3.7), (3.8), and (3.12) with g_N given by (3.63), then the decision rule is Mean Risk Consistent.

PROOF: The second term in the bound of Lemma 3.2 is given by

$$T_2 \triangleq \sum_{j=1}^{M} \int \{E_N[\hat{D}_{Nj}(x)] - D_j(x)\}\{\delta_{Bj}(x; v) - E_N[\hat{\delta}_{Nj}(x)]\}\, d\nu(x)$$

$$= \sum_{j=1}^{M} \sum_{i=1}^{M} L_{ji} P_i \int \{E[g_N(x, X)|\Lambda = i] - f(x|\Lambda = i)\}$$

$$\cdot \{\delta_{Bj}(x; v) - E_N[\hat{\delta}_{Nj}(x)]\}\, d\nu(x). \tag{3.69}$$

But for all x and all N

$$|\delta_{Bj}(x; v) - E_N[\hat{\delta}_{Nj}(x)]| \leq 1. \tag{3.70}$$

Taking the absolute value of the integrand in (3.69) gives

$$T_2 \leq \sum_{j=1}^{M} \sum_{i=1}^{M} L_{ji} P_i \int |E[g_N(x, X)|\Lambda = i] - f(x|\Lambda = i)|\, d\nu(x). \tag{3.71}$$

Fubini's Theorem together with (3.9) implies that

$$\int E[g_N(x, X)|\Lambda = i]\, d\nu(x) = E[\int g_N(x, X)\, d\nu(x)|\Lambda = i] = 1. \tag{3.72}$$

So $E[g_N(x, X)|\Lambda = i]$ is a density function. Since (3.66) holds, Scheffe's theorem [P-3, p.22] implies that

$$\lim_{N\to\infty} \int |E[g_N(x, X)|\Lambda = i] - f(x|\Lambda = i)|\, d\nu(x) = 0. \tag{3.73}$$

Hence $T_2 \to 0$ as $N \to \infty$.

The first term in the bound of Lemma 3.2 is given by

$$T_1 \triangleq \sum_{j=1}^{M} \overline{L}_j \int E_N[|\hat{v}_{Nj}(x) - P_j\, E[g_N(x, X)|\Lambda = j]|]\, d\nu(x). \tag{3.74}$$

Schwarz inequality gives

$$\begin{aligned} q_{Nj}(x) &\triangleq E_N[|\hat{v}_{Nj}(x) - P_j\, E[g_N(x, X)|\Lambda = j]|] \\ &\le \{E_N[(\hat{v}_{Nj}(x) - P_j\, E[g_N(x, X)|\Lambda = j])^2]\}^{\frac{1}{2}} \\ &= \{E_N[\hat{v}^2_{Nj}(x)] - P_j^2\, E^2[g_N(x, X)|\Lambda = j]\}^{\frac{1}{2}} \end{aligned} \tag{3.75}$$

where the last equality follows from expanding the square and applying Lemma 3.1. Now (3.35) together with (2.12) gives

$$\begin{aligned} E_N[\hat{v}^2_{Nj}(x)] &= \frac{1}{N} \sum_{i=1}^{M} b^2_{ji} P'_i\, E[g^2_N(x, X)|\Lambda' = i] \\ &\quad + \frac{N-1}{N} \Big(\sum_{i=1}^{M} b_{ji} P'_i\, E[g_N(x, X)|\Lambda' = i]\Big)^2 \\ &= \frac{1}{N} \sum_{i=1}^{M} b^2_{ji} \sum_{\ell=1}^{M} \beta_{i\ell} P_\ell\, E[g^2_N(x, X)|\Lambda = \ell] \\ &\quad + (1 - \tfrac{1}{N})(P_j\, E[g_N(x, X)|\Lambda = j])^2. \end{aligned} \tag{3.76}$$

Thus

$$q_{Nj}(x) \leq (Nh_N^n)^{-\frac{1}{2}} \{ \sum_{\ell=1}^{M} (\sum_{i=1}^{M} b_{ji}^2 \beta_{i\ell}) P_\ell h_N^n \, E[g_N^2(x, X) | \Lambda = \ell] - h_N^n p_j^2 \, E^2[g_N(x, X) | \Lambda = j] \}^{\frac{1}{2}}. \tag{3.77}$$

Next note that (3.66) and (3.67) imply that the right hand side of (3.77) converges to zero a.e.ν as $N \to \infty$. Hence $q_{Nj}(x) \to 0$ a.e.ν as $N \to \infty$, and in particular $q_{Nj}(x)$ converges to zero in measure for each $j = 1, 2, \ldots, M$.

Alternately, $q_{Nj}(x)$ may be bounded in a manner similar to (3.46)

$$\begin{aligned} q_{Nj}(x) &\leq E[|\hat{v}_{Nj}(x)|] + P_j \, E[g_N(x, X) | \Lambda = j] \\ &\leq \max_{1 \leq i \leq M} |b_{ji}| \sum_{k=1}^{M} P_k' \, E[g_N(x, X) | \Lambda' = k] \\ &\quad + P_j \, E[g_N(x, X) | \Lambda = j] \\ &= \max_{1 \leq i \leq M} |b_{ji}| \sum_{k=1}^{M} P_k \, E[g_N(x, X) | \Lambda = k] \\ &\quad + P_j \, E[g_N(x, X) | \Lambda = j] \\ &\triangleq q_{Nj}'(x). \end{aligned} \tag{3.78}$$

Equation (3.73) shows that $E[g_N(x, X) | \Lambda = j]$ converges in the mean to $f(x | \Lambda = j)$ for each j. Clearly then $q_{Nj}'(x)$ converges in the mean to the integrable function

$$q_j'(x) \triangleq \max_{1 \leq i \leq M} |b_{ji}| \sum_{i=1}^{M} v_i(x) + v_j(x). \tag{3.79}$$

Hence Lemma 3.3 implies that

$$\lim_{N \to \infty} \int q_{Nj}(x) \, d\nu(x) = 0, \quad j = 1, 2, \ldots, M. \tag{3.80}$$

Thus $T_1 \to 0$ as $N \to \infty$.

End of proof.

A rate of convergence similar to that in Theorem 3.3 can also be established for the case of continuous densities. For any density function $f(x)$, define a translate function $\tau(\cdot)$ as

$$\tau(y, f) \triangleq \int |f(x + y) - f(x)| \, d\nu(x). \tag{3.81}$$

The rate theorem is as follows:

THEOREM 3.5. Let the decision rule $\hat{\delta}_N$ be defined according to (3.7), (3.8), and (3.12) with g_N given by (3.63). In addition suppose that the following conditions are satisfied:

$$\text{i)} \quad \int ||y|| \, K(y) \, d\nu(y) < \infty, \tag{3.82a}$$

ii) for each $\lambda = 1, 2, \ldots, M$

$$E\Big[\prod_{i=1}^{n} |X_i|^{1+\xi'} \,\Big|\, \Lambda = \lambda\Big] < \infty \quad \text{for some } \xi' > 0, \tag{3.82b}$$

iii) for some $0 < \gamma \leq 1$ and some constant C

$$\max_{1 \leq \lambda \leq M} \tau(y; f(\cdot|\Lambda = \lambda)) \leq C||y||^{\gamma}. \tag{3.82c}$$

Then choosing $h_N = 0(N^{-1/(n+\alpha)})$ implies that there exist constants Q_1 and Q_2 such that for large N

$$\Delta R_N \leq \begin{cases} Q_1 \, N^{-\gamma/(n+\alpha)} & \text{if} \quad \alpha > 2\gamma \\ Q_2 \, N^{-\alpha/2(n+\alpha)} & \text{if} \quad \alpha < 2\gamma \\ (Q_1 + Q_2) N^{-\alpha/(n+2\alpha)} & \text{if} \quad \alpha = 2\gamma \end{cases} \tag{3.83}$$

The proof of this theorem is given in Appendix C.

The conditions in (3.82) necessary for proving Theorem 3.5 are very general. Condition (3.82a) is satisfied by all of the kernels in

Table B.1 except for the Cauchy kernel (entry 5). Condition (3.82b) is a weak moment condition on the density functions. Condition (3.82c) covers an extensive class of densities as discussed by Van Ryzin [V-1].

Van Ryzin has shown that when the density functions $f(x|\Lambda = k)$, $k = 1, 2, \ldots, M$, are absolutely continuous and have first partial derivatives which are integrable, then (3.82c) is satisfied with $\gamma = 1$. It is easy to shown that when $\gamma = 1$ the bound in (3.83) is smallest if α is chosen to equal 2. In this case the rate of convergence is $\Delta R_N = 0(N^{-1/(n+2)})$. So a good rule of thumb would be to choose $h_N = 0(N^{-1/(n+2)})$. Of course if α were known, a better rate of convergence could be obtained by properly choosing α.

CHAPTER IV

EFFECTS OF THE IMPERFECT TEACHER

In Chapter II, a model was proposed for an imperfect teacher. An example (Section 2.4) was presented to show that the use of misclassified training data can significantly affect the convergence of a supervised learning procedure. A nonparametric algorithm for learning with an imperfect teacher was then proposed in Chapter III. The algorithm took into account the misclassifications so that the estimated decision rules converged to the Bayes rule as the number of training patterns became large. Rates of convergence were also given by establishing bounds on ΔR_N that were functions of the number of training patterns used to learn the decision rules. Since these bounds hold for a large class of probability distributions, one would not expect the bounds to be very tight; they only give a loose measure of rate of convergence.

In this chapter the two-class learning problem is investigated in some detail. The objective is to study the qualitative and quantitative effects of an imperfect teacher on learning. This objective is achieved by restricting attention to the two-class problem and assuming specific forms for the underlying class-conditional densities. Measures of performance for comparing the perfect and imperfect teacher are proposed and studied.

A large sample approximation is developed in Section 4.2 to evaluate the expected risk in the two-class problem. The approximation

is then used in Section 4.3 to study three examples of learning with an imperfect teacher. In Section 4.4 large sample properties of the expected risk are investigated. A quantitative measure of performance is proposed in Section 4.5. The measure is evaluated for a zero-one loss function and used in Section 4.6 to compute a cost associated with training. This cost provides a second quantitative measure for comparing an imperfect teacher with a perfect one.

4.1 EXPECTED RISK FOR THE TWO-CLASS PROBLEM

The Bayes decision rule for the two-class problem may be written from (3.1) and (3.5) as

$$\begin{aligned} \delta_{B1}(x) &= 1 \quad \text{if} \quad D(x) \geq 0 \\ &= 0 \quad \text{if} \quad D(x) < 0 , \end{aligned} \tag{4.1}$$

$$\delta_{B2}(x) = 1 - \delta_{B1}(x) \tag{4.2}$$

where the Bayes discriminant function has the form

$$\begin{aligned} D(x) &= (L_{21} - L_{11})P_1 f(x|\Lambda = 1) - (L_{12} - L_{22})P_2 f(x|\Lambda = 2) \\ &= (L_{21} - L_{11})v_1(x) - (L_{12} - L_{22})v_2(x) . \end{aligned} \tag{4.3}$$

It follows from (4.3), (3.8) and (3.12) that the estimator of D(x) has the form

$$\begin{aligned} \hat{D}_N(x) &= (L_{21} - L_{11})\hat{v}_1(x) - (L_{12} - L_{22})\hat{v}_2(x) \\ &= \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{M} [(L_{21} - L_{11})b_{1i} - (L_{12} - L_{22})b_{2i}]\Delta(\Lambda_k', i) g_N(x, X_k) \\ &= \frac{1}{N} \sum_{k=1}^{N} W_{Nk}(x) \end{aligned} \tag{4.4}$$

where $\{W_{Nk}\}$ is a sequence of independent random variables that are identically distributed as the random variable

$$W_N(x) = \sum_{i=1}^{2} [(L_{21} - L_{11})b_{1i} - (L_{12} - L_{22})b_{2i}]\Delta(\Lambda', i)g_N(x, X)$$

$$= [G_1\Delta(\Lambda', 1) + G_2\Delta(\Lambda', 2)]g_N(x, X) \tag{4.5}$$

with

$$G_i = (L_{21} - L_{11})b_{1i} - (L_{12} - L_{22})b_{2i}. \tag{4.6}$$

The estimated decision rule is then

$$\hat{\delta}_{N1}(x) = 1 \quad \text{if} \quad \hat{D}_N(x) \geq 0$$
$$= 0 \quad \text{if} \quad \hat{D}_N(x) < 0, \tag{4.7}$$

$$\hat{\delta}_{N2}(x) = 1 - \hat{\delta}_{N1}(x). \tag{4.8}$$

The risk for any decision rule $\delta = (\delta_1, \delta_2)$ can be expressed in a convenient form for the two-class problem as follows. From (A.4) the risk is given by

$$R(P, f, \delta) = P_1 \int [L_{11}\delta_1(x) + L_{21}\delta_2(x)]f(x|\Lambda = 1)\, d\nu(x)$$
$$+ P_2 \int [L_{12}\delta_1(x) + L_{22}\delta_2(x)]f(x|\Lambda = 2)\, d\nu(x). \tag{4.9}$$

Using the relation

$$\delta_2(x) = 1 - \delta_1(x) \tag{4.10}$$

in (4.9) gives

$$R(P, f, \delta) = P_1L_{21} + P_2L_{22} + \int [P_1(L_{11} - L_{21})f(x|\Lambda = 1)$$
$$+ P_2(L_{12} - L_{22})f(x|\Lambda = 2)]\delta_1(x)\, d\nu(x)$$
$$= P_1L_{21} + P_2L_{22} - \int D(x)\delta_1(x)\, d\nu(x). \tag{4.11}$$

This expression shows that the risk for any decision rule in the two-class problem can be expressed as a sum of two constant terms that are

not functions of the decision rule plus an integral involving the decision rule $\delta_1(\cdot)$ and the Bayes discriminant function $D(\cdot)$.

The risk for the estimated decision rule $\hat{\delta}_N$ is then

$$R(P, f, \hat{\delta}_N) = P_1 L_{21} + P_2 L_{22} - \int D(x)\hat{\delta}_{N1}(x)\, d\nu(x), \tag{4.12}$$

and the expected risk with N training patterns is

$$\begin{aligned} R_N(P, f) &= P_1 L_{21} + P_2 L_{22} - \int D(x) E_N[\hat{\delta}_{N1}(x)]\, d\nu(x) \\ &= P_1 L_{21} + P_2 L_{22} - \int D(x) \Pr[\hat{D}_N(x) \geq 0]\, d\nu(x). \end{aligned} \tag{4.13}$$

To evaluate the expected risk, one must be able to compute the probability that the estimated discriminant function is nonnegative. Since $\hat{D}_N$ is the average of a sequence of independent, identically distributed random variables, the calculation of $\Pr[\hat{D}_N(x) \geq 0]$ is generally an extremely difficult problem, although a normal approximation for $\hat{D}_N$ may be used when N is large. This approximation, which may be used to evaluate the expected risk given by (4.13), is developed in the next section. It is assumed throughout this chapter that $1/2 < \beta_{ii} \leq 1$, $i = 1, 2$.

4.2 NORMAL APPROXIMATION

The first two moments of the random variable $W_N(x)$ defined in (4.5) will be used in the normal approximation. From (4.5) it follows that the expected value of W_N is given by

$$\begin{aligned} E[W_N(x)] &= \sum_{j=1}^{2} G_j E[\Delta(\Lambda', j) g_N(x, X)] = \sum_{j=1}^{2} G_j P_j' E[g_N(x, X) | \Lambda' = j] \\ &= \sum_{j=1}^{2} G_j \sum_{i=1}^{2} \beta_{ji} P_i E[g_N(x, X) | \Lambda = i] \end{aligned} \tag{4.14}$$

where the last equality follows from (2.10). Interchanging summations results in

$$E[W_N(x)] = (G_1\beta_{11} + G_2\beta_{21})P_1E[g_N(x, X)|\Lambda = 1] + (G_1\beta_{12} + G_2\beta_{22})P_2E[g_N(x, X)|\Lambda = 2]. \quad (4.15)$$

For the two-class problem the $\underline{\beta}$ matrix is

$$\underline{\beta} = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} \quad (4.16)$$

so that the inverse matrix is

$$\underline{\beta}^{-1} = \frac{1}{\beta_{11} + \beta_{22} - 1} \begin{bmatrix} \beta_{22} & -\beta_{12} \\ -\beta_{21} & \beta_{11} \end{bmatrix}. \quad (4.17)$$

Substituting (4.17) and (4.6) into (4.15) and simplifying gives

$$\begin{aligned} E[W_N(x)] &= (L_{21} - L_{11})P_1E[g_N(x, X)|\Lambda = 1] \\ &\quad - (L_{12} - L_{22})P_2E[g_N(x, X)|\Lambda = 2] \\ &= \sum_{j=1}^{2} (L_{2j} - L_{1j})P_jE[g_N(x, X)|\Lambda = j]. \end{aligned} \quad (4.18)$$

Note that the expected value of $W_N(x)$ is not a function of the teacher; that is, it does not depend upon the $\underline{\beta}$ matrix. The second moment of W_N is

$$\begin{aligned} E[W_N^2(x)] &= E[(G_1\Delta(\Lambda', 1) + G_2\Delta(\Lambda', 2))^2 g_N^2(x, X)] \\ &= G_1^2P_1'E[g_N^2(x, X)|\Lambda' = 1] + G_2^2P_2'E[g_N^2(x, X)|\Lambda' = 2]. \end{aligned} \quad (4.19)$$

Proceeding as in (4.14), the second moment may be written as

$$E[W_N^2(x)] = H_1 P_1 E[g_N^2(x, X) | \Lambda = 1] + H_2 P_2 E[g_N^2(x, X) | \Lambda = 2] \quad (4.20)$$

where

$$H_1 = G_1^2 \beta_{11} + G_2^2 \beta_{21} \quad (4.21)$$

and

$$H_2 = G_1^2 \beta_{12} + G_2^2 \beta_{22}. \quad (4.22)$$

Substitution of (3.38) into (4.18) shows that when X is a discrete random variable, $E[W_N(x)] = D(x)$. Thus $\hat{D}_N$ is an unbiased estimator of D in the discrete case. Also for discrete X, (4.20) becomes

$$E[W_N^2(x)] = H_1 P_1 f(x|\Lambda = 1) + H_2 P_2 f(x|\Lambda = 2) \quad (4.23)$$

which is finite for all x. It then follows that $\hat{D}_N$ is asymptotically normal as defined by the following theorem:

THEOREM 4.1. Let the training patterns be represented by discrete random variables. Then $\hat{D}_N(x)$ is asymptotically normal in the sense that, for every real number a,

$$\lim_{N \to \infty} \Pr \left[\frac{\hat{D}_N(x) - D(x)}{(VAR[W_N(x)]/N)^{\frac{1}{2}}} \leq a \right] = \Phi(a) \quad (4.24)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution,

$$\Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt. \quad (4.25)$$

PROOF: Since $\hat{D}_N$ is the average of a sequence of independent, identically distributed random variables with finite mean and variance, the conclusion follows immediately from the Lindberg-Levy Central Limit Theorem [F-1, p. 256].

End of proof.

Now consider the case when the patterns are continuous random variables and the g_N functions have the form described in Section 3.5.

Proving that $\hat{D}_N$ is asymptotically normal in this case is not as direct as in Theorem 4.1 because the mean and variance of W_N are now both functions of N. The asymptotic normality of $\hat{D}_N$ is given by the following theorem.

THEOREM 4.2. If the estimator $\hat{D}_N$ is defined by (4.4) with g_N defined as in (3.63), then at every continuity point of $f(x|\Lambda = k)$, $k = 1, 2$, $\hat{D}_N$ is asymptotically normal in the sense that, for every real number a,

$$\lim_{N\to\infty} \Pr\left[\frac{\hat{D}_N(x) - E[W_N(x)]}{(VAR[W_N(x)]/N)^{\frac{1}{2}}} \leq a\right] = \Phi(a). \tag{4.26}$$

PROOF: Let x be a continuity point of $f(x|\Lambda = k)$, $k = 1, 2$. From Parzen [P-1, p. 1069] and Loève [L-2, p. 316], it follows that a sufficient condition for (4.26) to hold is that for some $\xi > 0$,

$$q_N(x) \triangleq \frac{N\, E[|W_N(x) - E[W_N(x)]|^{2+\xi}]}{(N\, VAR[W_N(x)])^{1+\xi/2}} \to 0 \quad \text{as} \quad N \to \infty. \tag{4.27}$$

Now from (4.5) it follows that

$$\begin{aligned} E[|W_N|^{2+\xi}] = {} & |G_1|^{2+\xi} P_1' E[g_N^{2+\xi}(x, X)|\Lambda' = 1] \\ & + |G_2|^{2+\xi} P_2' E[g_N^{2+\xi}(x, X)|\Lambda' = 2]. \end{aligned} \tag{4.28}$$

Using Lemma B.1, it follows from (4.28) that

$$\begin{aligned} \lim_{N\to\infty} h_N^{n(1+\xi)} E[|W_N|^{2+\xi}] = {} & \{|G_1|^{2+\xi} P_1' f(x|\Lambda' = 1) \\ & + |G_2|^{2+\xi} P_2' f(x|\Lambda' = 2)\} \int K^{2+\xi}(y)\, d\nu(y). \end{aligned} \tag{4.29}$$

Since the conditions on the kernel K guarantee that [C-1, p. 185] $\int K^{2+\xi}(y)\, d\nu(y) < \infty$, (4.29) implies that

$$0 < \lim_{N\to\infty} h_N^{n(1+\xi)} E[|W_N - E[W_N]|^{2+\xi}] < \infty \tag{4.30}$$

and

$$0 < \lim_{N\to\infty} h_N^n \, VAR[W_N] < \infty. \tag{4.31}$$

But q_N may be written as

$$q_N(x) = \frac{h_N^{n(1+\xi/2)}}{N^{\xi/2} h_N^{n(1+\xi)}} \frac{h_N^{n(1+\xi)} E[|W_N - E[W_N]|^{2+\xi}]}{(h_N^n \, VAR[W_N])^{1+\xi/2}}. \tag{4.32}$$

Thus $q_N \to 0$ as $N \to \infty$ because the first term of (4.32) goes to zero as $N \to \infty$ while the numerator and denominator of the second term remain finite.

End of proof.

Theorems 4.1 and 4.2 provide justification for using a normal approximation to $Pr[\hat{D}_N(x) \geq 0]$ in (4.13). Specifically, for large N

$$Pr[\hat{D}_N(x) \geq 0] \approx 1 - \Phi\left(\frac{-N^{\frac{1}{2}} E[W_N(x)]}{(VAR[W_N(x)])^{\frac{1}{2}}}\right)$$

$$= \Phi\left(\frac{N^{\frac{1}{2}} E[W_N(x)]}{(VAR[W_N(x)])^{\frac{1}{2}}}\right). \tag{4.33}$$

Thus for large N the expected risk may be approximated by

$$R_N(P, f) \approx P_1 L_{21} + P_2 L_{22} - \int D(x) \, \Phi\left(\frac{N^{\frac{1}{2}} E[W_N(x)]}{(VAR[W_N(x)])^{\frac{1}{2}}}\right) d\nu(x). \tag{4.34}$$

Even with the large sample approximation, the integral in (4.33) is a formidable expression that cannot be easily evaluated. In the next section, three examples are studied in an attempt to evaluate the effects of an imperfect teacher. For these examples, (4.34) is evaluated numerically to study the learning algorithm.

4.3 EXAMPLES OF LEARNING

Example 1 - Normal Distributions

In the first example to be studied, the patterns are assumed to have univariate normal distributions. Under pattern class ω_i the patterns are normally distributed with mean μ_i and variance σ^2. Since the density functions are continuous, the class of estimators described in Section 3.5 will be used for learning a decision rule. The kernel for the estimator is taken to be

$$K(y) = (2\pi)^{-\frac{1}{2}} \exp\ (-y^2/2). \tag{4.35}$$

The g_N function is then

$$g_N(x, y) = \frac{1}{h_N(2\pi)^{\frac{1}{2}}} \exp\ \left(-\frac{(x - y)^2}{2h_N^2}\right). \tag{4.36}$$

The first and second moments of g_N are needed in order to evaluate the approximation for the expected risk. The conditional mean is given by

$$E[g_N(x, X) \mid \Lambda = i] = \int_{-\infty}^{\infty} h_N^{-1}(2\pi)^{-\frac{1}{2}} \exp\ \{-(x - z)^2/2h_N^2\}$$

$$\sigma^{-1}(2\pi)^{-\frac{1}{2}} \exp\ \{-(z - \mu_i)^2/2\sigma^2\}\ dz. \tag{4.37}$$

Completing the squares in the exponentials and integrating the resulting expression gives

$$E[g_N(x, X) \mid \Lambda = i] = [2\pi(\sigma^2 + h_N^2)]^{-\frac{1}{2}} \exp\ \{-\tfrac{1}{2}(x - \mu_i)^2/(\sigma^2 + h_N^2)\}. \tag{4.38}$$

A similar evaluation for the second moment shows that

$$E[g_N^2(x, X) \mid \Lambda = i] = (2\pi h_N)^{-1}(2\sigma^2 + h_N^2)^{-\frac{1}{2}} \exp\ \{-\tfrac{1}{2}(x - \mu_i)^2/(\sigma^2 + h_N^2/2)\}. \tag{4.39}$$

The Bayes discriminant function for this example is found from

(4.3) to be

$$D(x) = P_1(L_{21} - L_{11})(2\sigma^2)^{-\frac{1}{2}} \exp\{-(x - \mu_1)^2/2\sigma^2\}$$
$$- P_2(L_{12} - L_{22})(2\sigma^2)^{-\frac{1}{2}} \exp\{-(x - \mu_2)^2/2\sigma^2\}. \qquad (4.40)$$

The Bayes risk for this example was derived in (2.25).

The above relations in conjunction with (4.18) and (4.19) can be used to evaluate, via numerical integration, the approximation for the expected risk given by (4.34). The results of evaluating this approximation are presented in Figures 4.1 thru 4.4. For this example the following parameter values were used: $\mu_1 = 1.5$, $\mu_2 = -1.5$, and $\sigma^2 = 1.0$. A zero-one loss function was used for all cases shown. The parameter h_N was chosen to have the form $h_N = N^{-\alpha}$. Following the rule of thumb suggested in Section 3.5, α was chosen to be 1/3.

Figures 4.1 and 4.2 present the expected risk plotted as a function of $\beta(\beta_{11} = \beta_{22} = \beta)$ for the prior probability of pattern class ω_1 equal to 0.5 and 0.1 respectively. Figure 4.3 shows the same case as Figure 4.1 except that $\beta_{11} = 1.0$. Figure 4.4 shows the expected risk plotted as a function of the number of training patterns. These figures illustrate the effects of an imperfect teacher. Basically, the imperfect teacher slows down the rate of convergence. For $\beta > 0.9$ the effect of the imperfect teacher is small, but for $\beta < 0.7$, say, convergence of the algorithm is much slower.

The dotted line in Figures 4.1, 4.2, and 4.3 is the teacher's risk. This risk is just the teacher's probability of error which is equal to $1 - \beta$. The figures show that for small β, the learning algorithm becomes better than the teacher as the number of training patterns increases.

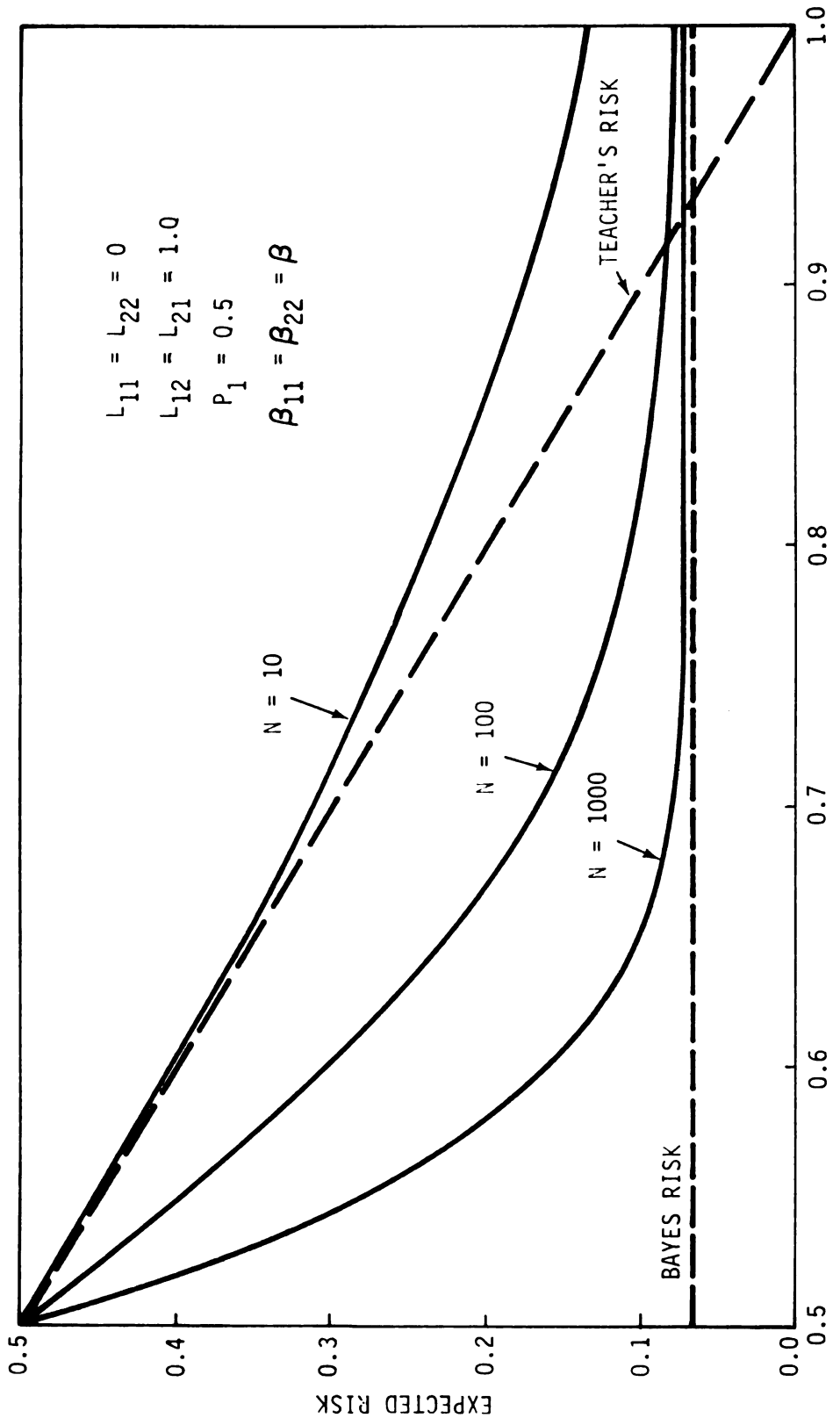

Figure 4.1. Expected Risk with Normal Distributions, $P_1 = 0.5$

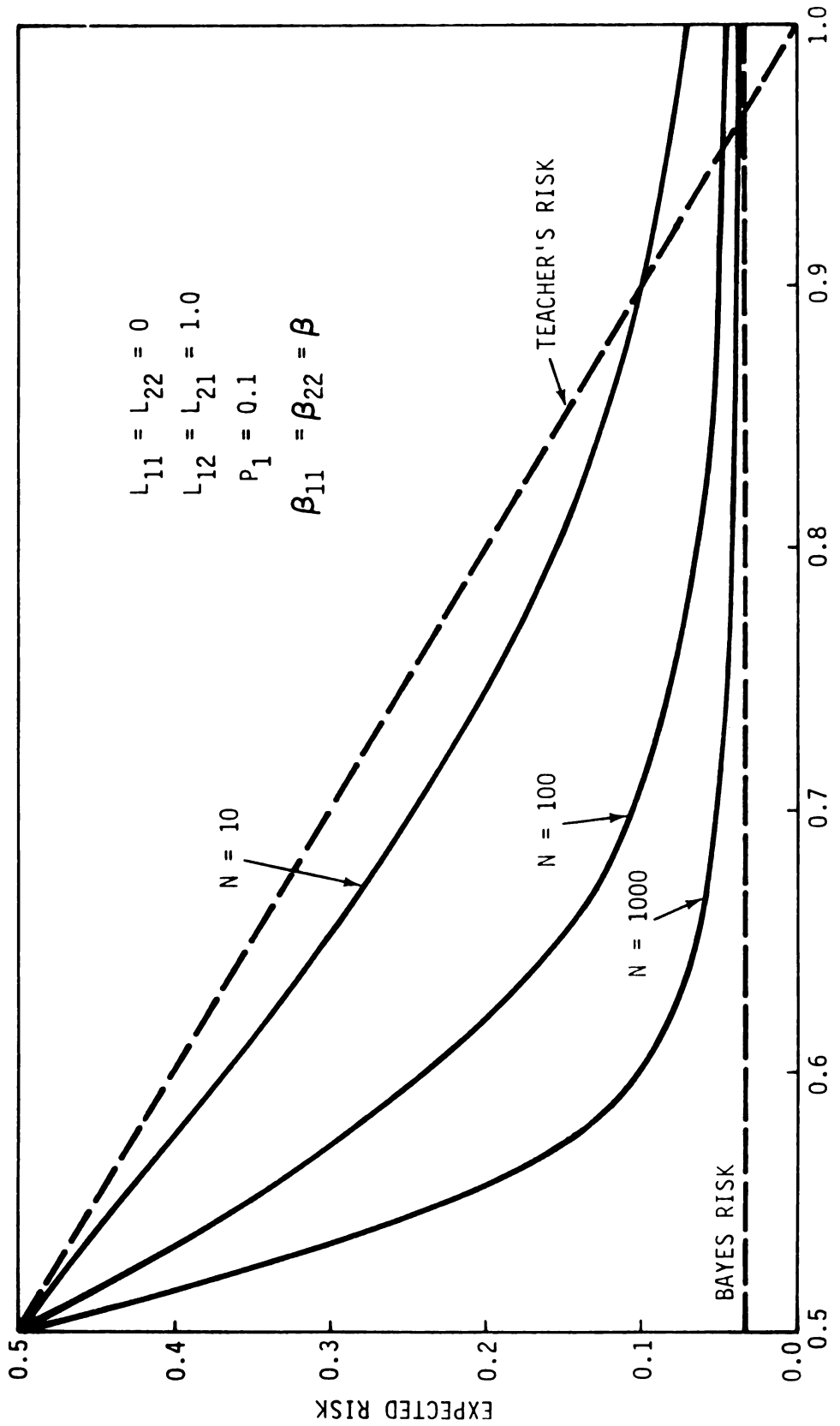

Figure 4.2. Expected Risk with Normal Distributions, $P_1 = 0.1$

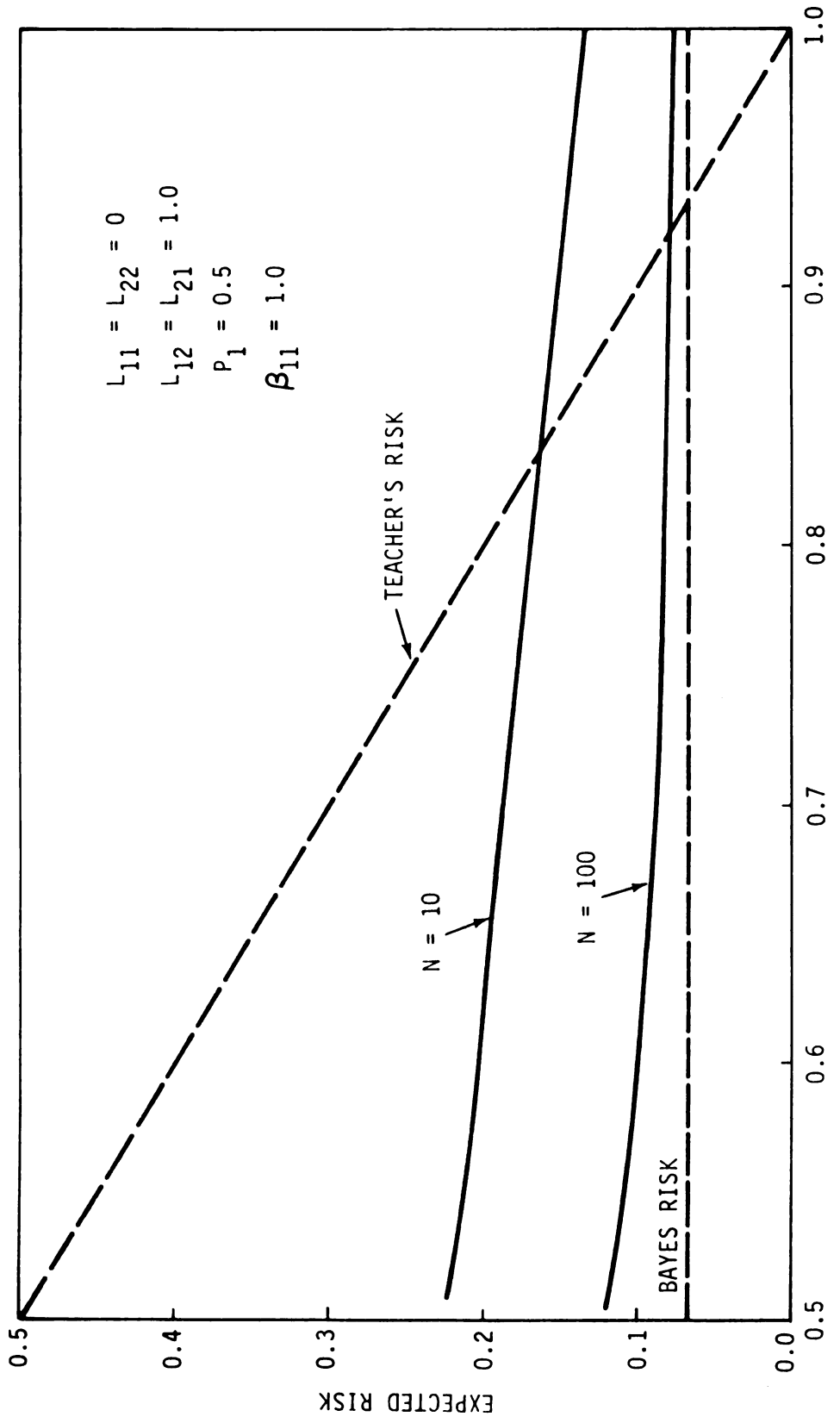

Figure 4.3. Expected Risk with Normal Distributions, $\beta_{11} \neq \beta_{22}$

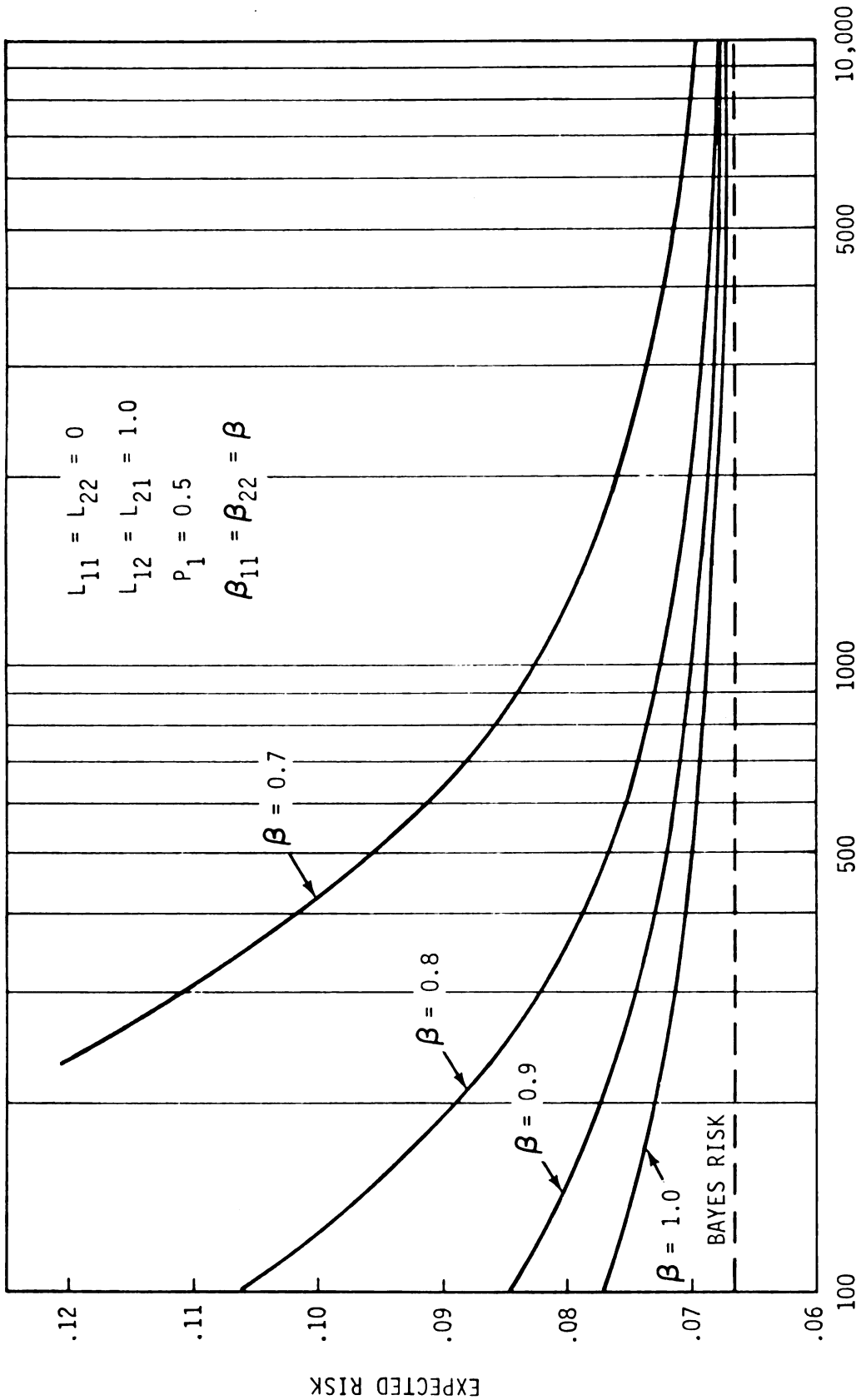

Figure 4.4. Expected Risk for Large N

Example 2 - Triangular Distributions

For the second example let the class-conditional densities be disjoint triangular densities as shown in Figure 4.5. These densities were used in [S-5] for a simulation of an imperfect teacher. They are considered here to show that one can evaluate the expected risk without resorting to simulations.

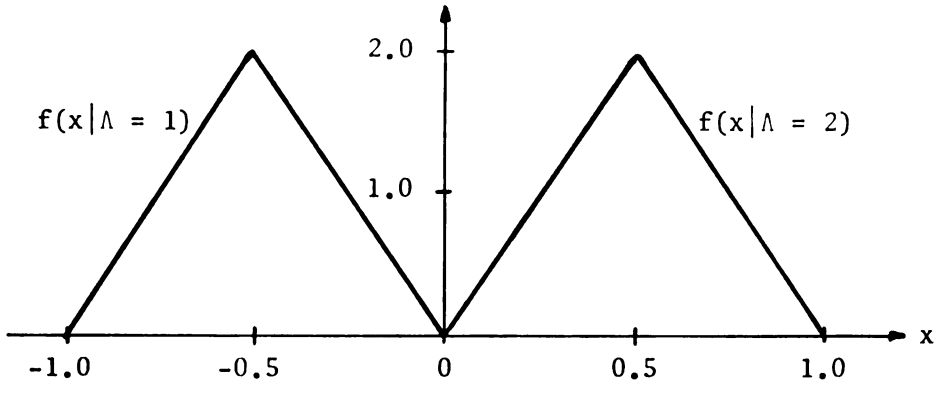

Figure 4.5 Triangular Density Functions

The normal kernel (4.35) will also be used in this example with the parameter α chosen to be $\frac{1}{2}$ in order to conform with the example in [S-5]. After some lengthy algebra, one can show that the moments of the g_N function are the following:

$$E[g_N(x, X)|\Lambda = 1] = 4x\Phi(\frac{x}{h_N}) - 4(2x - 1)\Phi(\frac{x - 0.5}{h_N}) + 4(x - 1)\Phi(\frac{x - 1.0}{h_N})$$

$$\frac{+4h_N}{(2\pi)^{\frac{1}{2}}}\left[\exp\left\{\frac{-x^2}{2h_N^2}\right\} - 2\exp\left\{\frac{-(x - 0.5)^2}{2h_N^2}\right\} + \exp\left\{\frac{-(x - 1.0)^2}{2h_N^2}\right\}\right] \qquad (4.41)$$

$$E[g_N(x, X)|\Lambda = 2] = 4x\Phi(\frac{x}{h_N}) - 4(2x + 1)\Phi(\frac{x + 0.5}{h_N}) + 4(x + 1)\Phi(\frac{x + 1.0}{h_N})$$

$$\frac{+4h_N}{(2\pi)^{\frac{1}{2}}}\left[\exp\left\{\frac{-x^2}{2h_N^2}\right\} - 2\exp\left\{\frac{-(x + 0.5)^2}{2h_N^2}\right\} + \exp\left\{\frac{-(x + 1.0)^2}{2h_N^2}\right\}\right] \qquad (4.42)$$

$$E[g_N^2(x, X) | \Lambda = 1] = \frac{2}{h_N\sqrt{\pi}} [x\Phi(\frac{\sqrt{2}x}{h_N}) - (2x - 1)\Phi(\frac{\sqrt{2}(x - 0.5)}{h_N})$$

$$+ (x - 1.0)\Phi(\frac{\sqrt{2}(x - 1.0)}{h_N})] + \frac{1}{\pi}[\exp\{\frac{-x^2}{h_N^2}\}$$

$$- 2 \exp\{\frac{-(x - 0.5)^2}{h_N^2}\} + \exp\{\frac{-(x - 1.0)^2}{h_N^2}\}] \quad (4.43)$$

$$E[g_N^2(x, X) | \Lambda = 2] = \frac{2}{h_N\sqrt{\pi}} [x\Phi(\frac{\sqrt{2}x}{h_N}) - (2x + 1)\Phi(\frac{\sqrt{2}(x + 0.5)}{h_N})$$

$$+ (x + 1.0)\Phi(\frac{\sqrt{2}(x + 1.0)}{h_N})] + \frac{1}{\pi}[\exp\{\frac{-x^2}{h_N^2}\}$$

$$- 2 \exp\{\frac{-(x + 0.5)^2}{h_N^2}\} + \exp\{\frac{-(x + 1.0)^2}{h_N^2}\}]. \quad (4.44)$$

Equations (4.41) thru (4.44) can now be used along with (4.18) and (4.20) to evaluate numerically the expected risk. Figures 4.6 and 4.7 present the results of this evaluation using a zero-one loss function and equal prior probabilities. The results of the simulation in [S-5] are also shown on Figure 4.7. There is reasonable agreement between the simulation and the large sample approximation.

Example 3 - Binomial Distributions

For the final example, suppose that the patterns are discrete random variables that have binomial distributions under both pattern classes,

$$f(x|\Lambda = j) = \binom{C}{x} \theta_j^x (1 - \theta_j)^{C-x}, \quad x = 0, 1, \ldots, C, \quad j = 1, 2. \quad (4.45)$$

For this discrete case, the estimator $\hat{D}_N$ is formed according to (4.4) with the g_N function defined by (3.37).

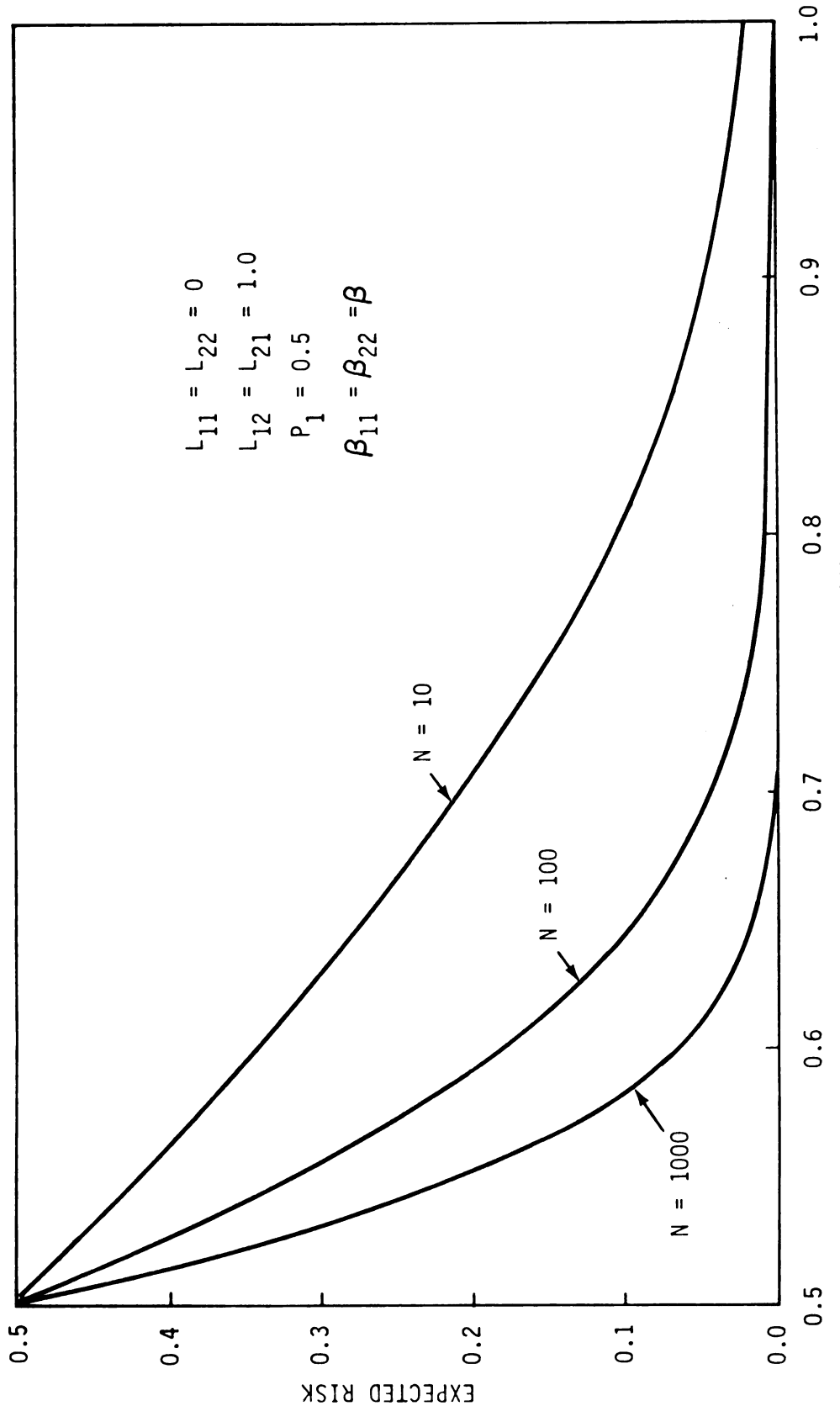

Figure 4.6. Expected Risk with Triangular Densities

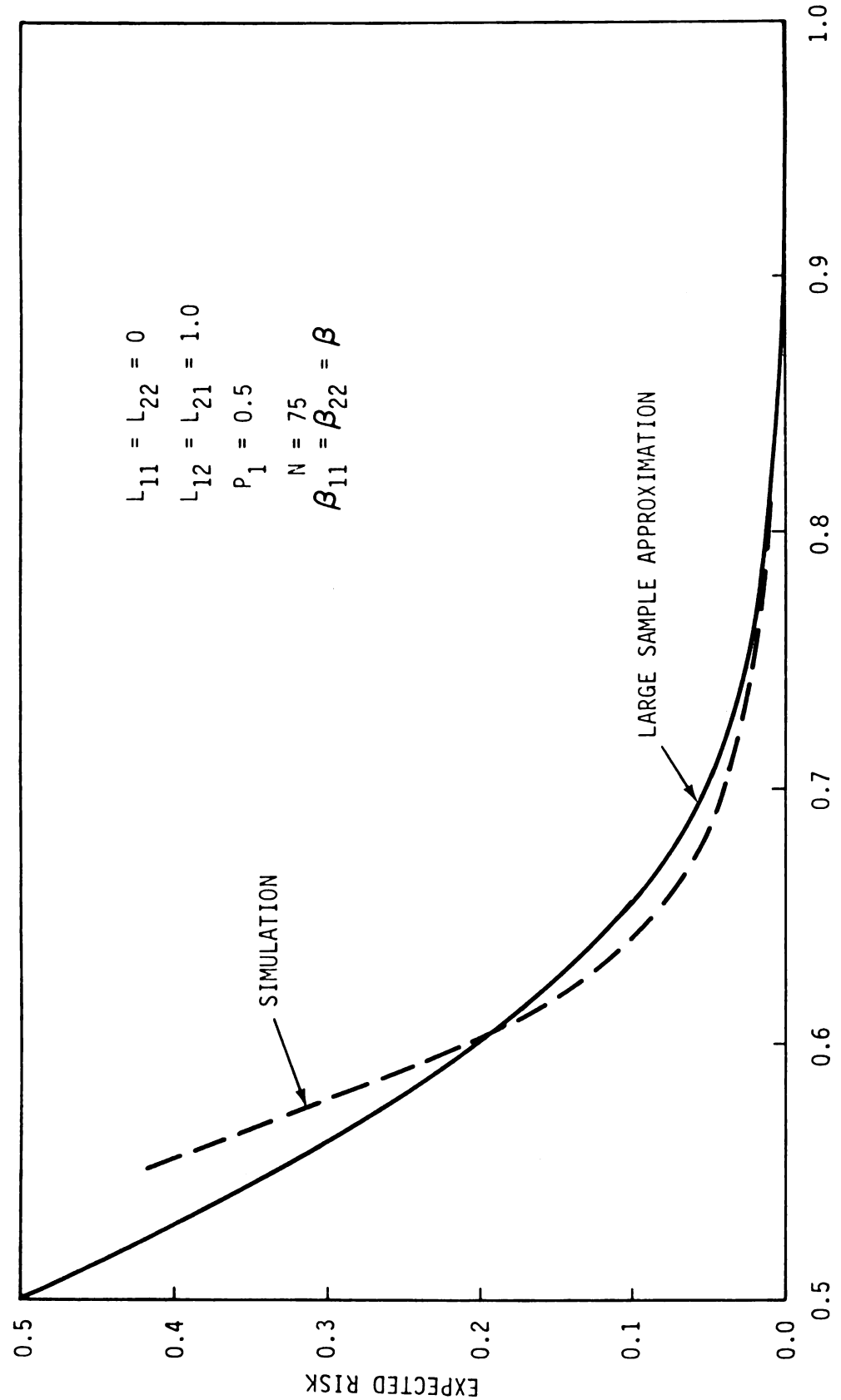

Figure 4.7. Comparison of Large Sample Approximation and Simulation

From (4.1) and (4.3) it follows that the Bayes decision rule for this example may be written as $\delta_{B1}(x) = 1$ if

$$f(x|\Lambda = 1)/f(x|\Lambda = 2) \geq P_2(L_{12} - L_{22})/P_1(L_{21} - L_{11}). \tag{4.46}$$

Substituting (4.45) into (4.46) and simplifying, results in the condition that $\delta_{B1}(x) = 1$ if

$$Ax \geq \tau \tag{4.47}$$

where

$$A = \ln(\theta_1(1 - \theta_2)/(1 - \theta_1)\theta_2) \tag{4.48}$$

and

$$\tau = \ln((L_{12} - L_{22})/(L_{21} - L_{11})) + \ln(P_2/P_1) - C \ln(1 - \theta_1)/(1 - \theta_2)). \tag{4.49}$$

The Bayes risk for this example is then

$$R_B = P_1L_{21} + P_2L_{22} - \sum_{Ax \geq \tau} D(x). \tag{4.50}$$

The large sample approximation for the expected risk is just

$$R_N \approx P_1L_{21} + P_2L_{22} - \sum_{x=0}^{C} D(x)\Phi(\sqrt{N}\, D(x)/(VAR[W_N(x)])^{\frac{1}{2}}) \tag{4.51}$$

where

$$VAR[W_N(x)] = H_1P_1f(x|\Lambda = 1) + H_2P_2f(x|\Lambda = 2) - D^2(x) \tag{4.52}$$

and the H_i are defined according to (4.21) and (4.22).

Figure 4.8 shows an evaluation of (4.51) for a typical set of parameters: $C = 10$, $\theta_1 = 0.25$, and $\theta_2 = 0.75$. A zero-one loss function and equal prior probabilities are assumed. The behavior of the learning algorithm is similar to that in the previous examples.

In all three of these examples, the expected risk rapidly becomes large as β approaches 0.5. But for β greater than 0.8, only a moderate number, say $N < 1000$, of samples are necessary for the expected risk to

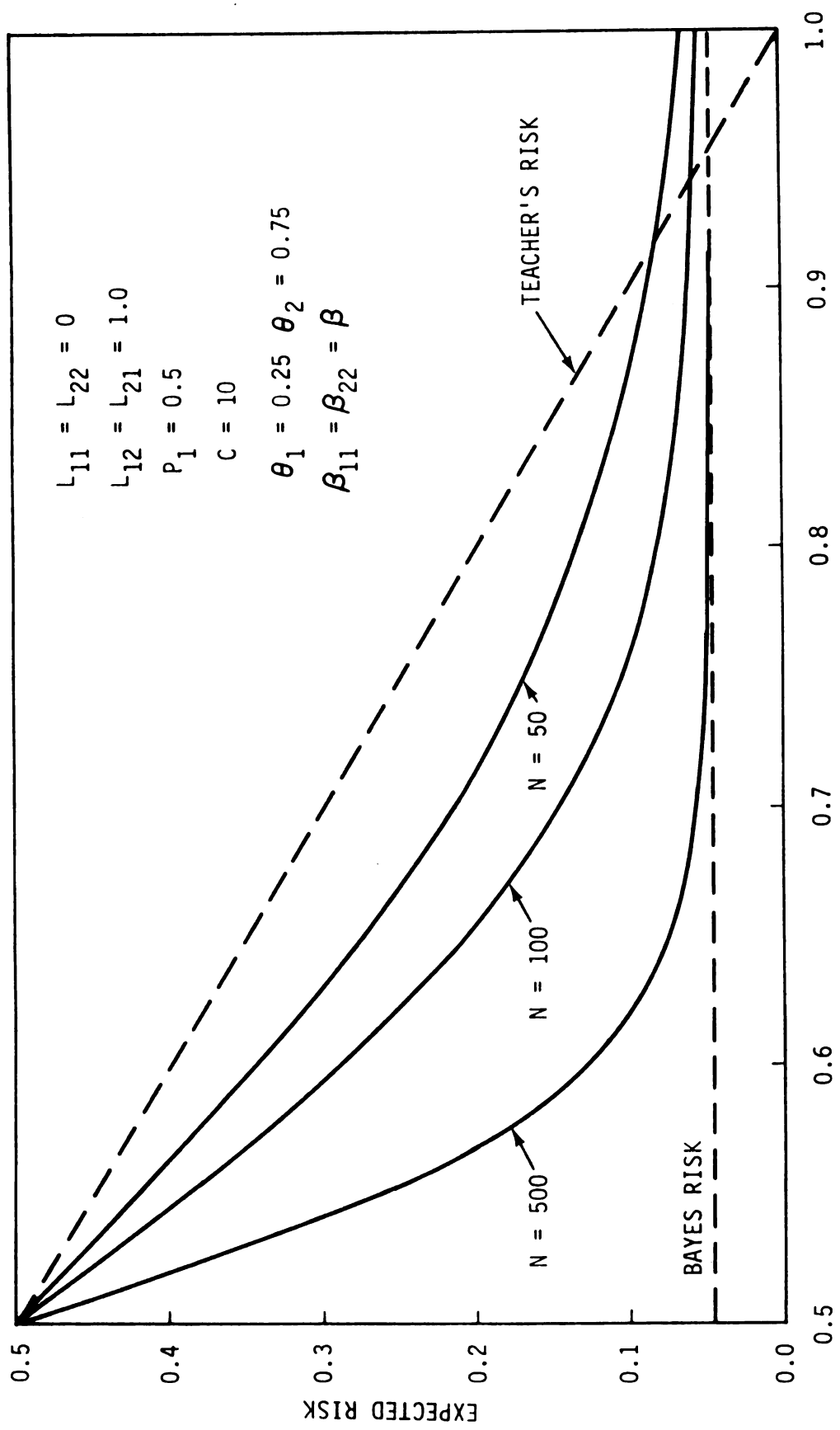

Figure 4.8. Expected Risk with Binomial Distributions

be close to the Bayes risk. In all cases shown, the expected risk is a decreasing function of N and of β_{ii}. In the next section it is shown that these characteristics hold in general.

Even though these examples provide a qualitative feel for the behavior of the learning algorithms, one would like to have quantitative measures of performance that characterize the effects of the imperfect teacher. The remaining sections of this chapter address the problem of developing such measures of performance.

4.4 LARGE SAMPLE PROPERTIES OF THE EXPECTED RISK

In Theorems 4.1 and 4.2 the estimator $\hat{D}_N$ was shown to be asymptotically normal with a certain mean and variance. For the continuous case the mean and variance of W_N were functions of N. In order to obtain performance measures, it is necessary to develop in this section a normal approximation in which the mean and variance are not functions of N. Only the continuous case will be considered in this section.

The following two lemmas establish the asymptotic properties of the mean and variance of $\hat{D}_N$ in the continuous case.

LEMMA 4.1. At every continuity point of $f(x|\Lambda = k)$, $k = 1, 2$,

$$\lim_{N\to\infty} E[\hat{D}_N(x)] = D(x); \tag{4.53}$$

i.e., $\hat{D}_N$ is an asymptotically unbiased estimator of D.

PROOF: Lemma B.1 along with (4.4) and (4.18) imply that at every continuity point of the densities $f(x|\Lambda = k)$

$$\lim_{N\to\infty} E[\hat{D}_N(x)] = \lim_{N\to\infty} E[W_N(x)] = D(x). \tag{4.54}$$

End of proof.

LEMMA 4.2. At every continuity point of $f(x|\Lambda = k)$, $k = 1, 2$,

$$\lim_{N\to\infty} Nh_N^n \, VAR[\hat{D}_N(x)] = Q[H_1P_1f(x|\Lambda = 1) + H_2P_2f(x|\Lambda = 2)] \tag{4.55}$$

with $Q = \int K^2(y)\, d\nu(y)$ and H_1 and H_2 defined by (4.21) and (4.22) respectively.

PROOF: Lemma B.1 implies that at every continuity point x of the density functions

$$\lim_{N\to\infty} h_N^n \, E[g_N^2(x, X)|\Lambda = k] = Qf(x|\Lambda = k). \tag{4.56}$$

Now from (4.4)

$$Nh_N^n \, VAR[\hat{D}_N(x)] = h_N^n \, E[W_N^2(x)] - h_N^n \, E^2[W_N(x)]. \tag{4.57}$$

Equation (4.54) and the fact that $h_N^n \to 0$ as $N \to \infty$ imply that the second term of (4.57) converges to zero as $N \to \infty$. Thus substituting (4.20) into (4.57) and using (4.56) gives the desired conclusion.

End of proof.

In view of Lemmas 4.1 and 4.2, it would seem reasonable to expect that $\hat{D}_N(x)$ is asymptotically normal with mean $D(x)$ and variance

$$\sigma_N^2 = [H_1P_1f(x|\Lambda = 1) + H_2P_2f(x|\Lambda = 2)]Q/Nh_N^n. \tag{4.58}$$

Showing that such a condition holds requires some further restrictions on the rate of convergence of h_N to zero. Lemma B.2 leads to the necessary restrictions. In view of Theorem 4.2, it is clear that

$$\lim_{N\to\infty} Pr[(\hat{D}_N(x) - D(x))/\sigma_N \leq a] = \Phi(a) \tag{4.59}$$

provided

$$VAR[\hat{D}_N(x)]/\sigma_N^2 \to 1 \quad \text{as} \quad N \to \infty \tag{4.60}$$

and

$$(E[\hat{D}_N(x)] - D(x))/\sigma_N \to 0 \quad \text{as} \quad N \to \infty. \tag{4.61}$$

Lemma 4.2 guarantees that (4.60) is satisfied. Also

$$\frac{E[\hat{D}_N(x) - D(x)]}{\sigma_N} = (Nh_N^{n+4})^{\frac{1}{2}} \frac{h_N^{-2}\{E[\hat{D}_N(x)] - D(x)\}Q^{-\frac{1}{2}}}{[H_1P_1f(x|\Lambda = 1) + H_2P_2f(x|\Lambda = 2)]^{\frac{1}{2}}}. \tag{4.62}$$

So from (4.4), (4.18) and Lemma B.2, it follows that condition (4.61) will be satisfied if $Nh_N^{n+4} \to 0$ as $N \to \infty$. This is then the additional condition on h_N required for proving that $\hat{D}_N$ is asymptotically normal in the sense of the following theorem.

THEOREM 4.3. In the continuous case if $\hat{D}_N$ is defined by (4.4), if $f(x|\Lambda = k)$, $k = 1, 2$, have continuous partial derivatives of third order a.e., and if h_N satisfies the conditions

$$Nh_N^{n+4} \to 0, \; Nh_N^n \to \infty \quad \text{as} \quad N \to \infty, \tag{4.63}$$

then $\hat{D}_N(x)$ is asymptotically normal in the sense that, for every real number a,

$$\lim_{N\to\infty} \Pr[(\hat{D}_N(x) - D(x))/\sigma_N(x) \leq a] = \Phi(a) \tag{4.64}$$

with $\sigma_N(x)$ defined by (4.58).

In the remainder of this chapter, the conditions of Theorem 4.3 are assumed to be satisfied. In particular h_N is taken to be $h_N = N^{-\alpha}$, $1/(n + 4) < \alpha < 1/n$. For sufficiently large N, the expected risk may be approximated by

$$R_N(P, f) \approx P_1L_{21} + P_2L_{22} - \int D(x)\Phi(t(x)) \, d\nu(x) \tag{4.65}$$

with

$$t(x) = N^{(1-n\alpha)/2}D(x)Q^{-\frac{1}{2}}[H_1P_1f(x|\Lambda = 1) + H_2P_2f(x|\Lambda = 2)]^{-\frac{1}{2}}. \tag{4.66}$$

In the three examples in Section 4.3 it was observed that the expected risks were decreasing functions of N and β. The following two theorems show that, as one would expect for large N, R_N is always a decreasing function of N, β_{11}, and β_{22}.

THEOREM 4.4. For large N, the expected risk R_N is a strictly decreasing function of N.

PROOF: Treat N as a real variable. Then from (4.65) it follows that

$$\partial R_N/\partial N = -\int (2\pi)^{-\frac{1}{2}} D(x) \exp\{-t^2(x)/2\}\partial t(x)/\partial N \, d\nu(x)$$

$$= -\tfrac{1}{2}(2\pi)^{-\frac{1}{2}}(1 - n\alpha)N^{-n\alpha/2} \int D^2(x) \exp\{-t^2(x)/2\}Q^{-\frac{1}{2}}$$

$$\cdot [H_1P_1f(x|\Lambda = 1) + H_2P_2f(x|\Lambda = 2)]^{-\frac{1}{2}} \, d\nu(x) < 0 \qquad (4.67)$$

since $(1 - n\alpha) > 0$ for $\alpha < 1/n$ and the integrand is always nonnegative.

End of proof.

THEOREM 4.5. For large N, the expected risk R_N is a strictly decreasing function of β_{11} and β_{22}.

PROOF: For large N, (4.65) implies that

$$\partial R_N/\partial \beta_{11} = -\int (2\pi)^{-\frac{1}{2}} D(x) \exp\{-t^2(x)/2\}\partial t(x)/\partial \beta_{11} \, d\nu(x). \qquad (4.68)$$

But

$$\frac{\partial t(x)}{\partial \beta_{11}} = \frac{-\frac{1}{2}N^{(1-n\alpha)/2}D(x)\{\partial H_1/\partial \beta_{11}P_1f(x|\Lambda = 1) + \partial H_2/\partial \beta_{11}P_2f(x|\Lambda = 2)\}}{Q^{\frac{1}{2}}[H_1P_1f(x|\Lambda = 1) + H_2P_2f(x|\Lambda = 2)]^{3/2}} \qquad (4.69)$$

By differentiating (4.21) and (4.22) with respect to β_{11}, one can show that

$$\frac{\partial H_1}{\partial \beta_{11}} = \frac{-(1 - \beta_{11} + \beta_{22}(2\beta_{11} - 1))(L_{21} - L_{11} + L_{12} - L_{22})^2}{(\beta_{11} + \beta_{22} - 1)^2} < 0 \qquad (4.70)$$

$$\frac{\partial H_2}{\partial \beta_{11}} = \frac{-[(1 - \beta_{11})(1 - \beta_{22}) + \beta_{11}\beta_{22}](L_{21} - L_{11} + L_{12} - L_{22})^2}{(\beta_{11} + \beta_{22} - 1)^2} < 0. \qquad (4.71)$$

Substitution of (4.69), (4.70), and (4.71) into (4.68) then shows that $\partial R_N/\partial \beta_{11} < 0$. Thus R_N is a strictly decreasing function of β_{11}. A similar argument holds for β_{22}.

End of proof.

One should note that even though Theorems 4.4 and 4.5 were stated and proved for the continuous case, they also hold for the discrete case. This follows immediately by comparing (4.64) with (4.23) and (4.24).

4.5 A MEASURE OF PERFORMANCE

The problem that one encounters in trying to develop any quantitative measure of the effects of the imperfect teacher is that the behavior of the learning algorithm depends upon several factors:

a.) the class conditional densities,

b.) the prior probabilities,

c.) the loss function,

d.) the $\underline{\beta}$ matrix, and

e.) the form of the g_N function.

Any measure of performance which one defines will generally be a function of all of these parameters. What one would like for comparing a perfect teacher with an imperfect teacher is a measure which is insensitive to all of the parameters except the $\underline{\beta}$ matrix.

A second problem that one encounters is that an analytic evaluation of the expected risk is virtually impossible except in special cases. Even then, numerical methods and the large sample approximations developed in the previous sections must be used. In view of these considerations, the measure of performance proposed in this section will also be evaluated from the large sample approximations.

The criterion proposed here for comparing an imperfect teacher to a perfect teacher is defined as follows. For a given number N_o of training patterns, the nonparametric learning procedure with a perfect teacher has some expected risk, call it R_{N_o}. Define $N(\underline{\beta}, N_o)$ to be the number of training patterns required for the learning algorithm with an imperfect teacher to have the same expected risk. Then a measure of relative performance may be defined as

$$\eta(\underline{\beta}, N_o) \triangleq N(\underline{\beta}, N_o)/N_o. \tag{4.72}$$

In general this measure of relative performance will be a function of N_o as well as a function of the five factors listed earlier. But η does provide a quantitative measure of the effect of the imperfect teacher in the sense that it measures the additional number of training patterns required to compensate for the imperfect teacher.

The dependence of (4.72) on N_o may be removed by defining an asymptotic measure of performance according to

$$\eta(\underline{\beta}) = \lim_{N_o \to \infty} \eta(\underline{\beta}, N_o) \tag{4.73}$$

provided the limit exists. In general the evaluation of (4.73) is an intractable problem. But for one very important case, evaluation of (4.73) is rather direct.

Consider the case when a zero-one loss function is used and $\beta_{11} = \beta_{22} = \beta$. Then for large N, the expected risk for the continuous case is given by (4.65). Now suppose an N_o is chosen and (4.65) is evaluated for the perfect teacher ($\beta = 1$) to obtain R_{N_o}. Let $R_N(\beta)$ denote the expected risk when using N training patterns with an imperfect teacher that has a probability β of correctly classifying a training pattern. To find $N(\beta, N_o)$, one must then solve the equation

$$R_N(\beta) = R_{N_o}(\beta) \tag{4.74}$$

for N. One way in which (4.74) is satisfied is for the function $t(\cdot)$ of (4.66) to be the same for both R_N and R_{N_o}. But this will occur if

$$(2\beta - 1)N^{(1-n\alpha)/2} = N_o^{(1-n\alpha)/2}. \tag{4.75}$$

Thus

$$N(\beta, N_o) = N_o/(2\beta - 1)^{2/(1-n\alpha)} \tag{4.76}$$

Since by Theorems 4.4 and 4.5 R_N is a strictly decreasing function of N and β, (4.76) is the unique solution of (4.74). Hence in this case the measure of relative performance is

$$\eta(\beta) = (2\beta - 1)^{-2/(1-n\alpha)}. \tag{4.77}$$

This measure of performance has the pleasant property that it is not a function of the class-conditional densities or prior probabilities, and it depends upon the g_N function only through the parameter α. Thus for a zero-one loss function, $\eta(\beta)$ in (4.77) is interpreted as the proportionate number of training patterns required for an imperfect teacher to yield the same expected risk as a perfect teacher.

Even when $\beta_{11} \neq \beta_{22}$ but a zero-one loss function is still used,

(4.77) provides a meaningful bound. Since R_N is a decreasing function of β_{11}, β_{22}, and N,

$$\eta(\max(\beta_{11}, \beta_{22})) \leq \eta(\beta_{11}, \beta_{22}) \leq \eta(\min(\beta_{11}, \beta_{22})). \qquad (4.78)$$

Thus using $\min(\beta_{11}, \beta_{22})$ in (4.77) yields an upper bound on the actual measure of relative performance, and using $\max(\beta_{11}, \beta_{22})$ gives a lower bound.

In Figure 4.9 the measure $\eta(\beta)$ is plotted as a function of β for $n = 1$ and $\alpha = 0.2$, 0.33, 0.5, and 0.65. The figure shows, for example, that with $\alpha = \frac{1}{2}$ the algorithm requires roughly 2.5 times as many training patterns when $\beta = 0.9$ as compared to a perfect teacher. When $\beta = 0.8$, 7.6 times as many training patterns are required; and when $\beta = 0.7$, nearly 40 times as many training patterns are required for the imperfect teacher to match the performance of the perfect teacher. Thus a poor teacher is costly in the sense that many more training patterns are required to achieve a performance equal to that achieved with a perfect teacher.

4.6 COST OF TRAINING

The idea of a cost associated with an imperfect teacher can be further developed by assuming that the teacher incurs a cost in classifying each training pattern. Assume that the cost of classifying a training pattern is an increasing function of β_{11} and β_{22}. Let $T(\underline{\beta})$ denote this cost function.

Suppose that one wishes to attain a given expected risk R^* with minimum cost. If a perfect teacher is used, N^* training patterns will be required to achieve R^* at a total cost of $N^*T(I)$, I being the unit

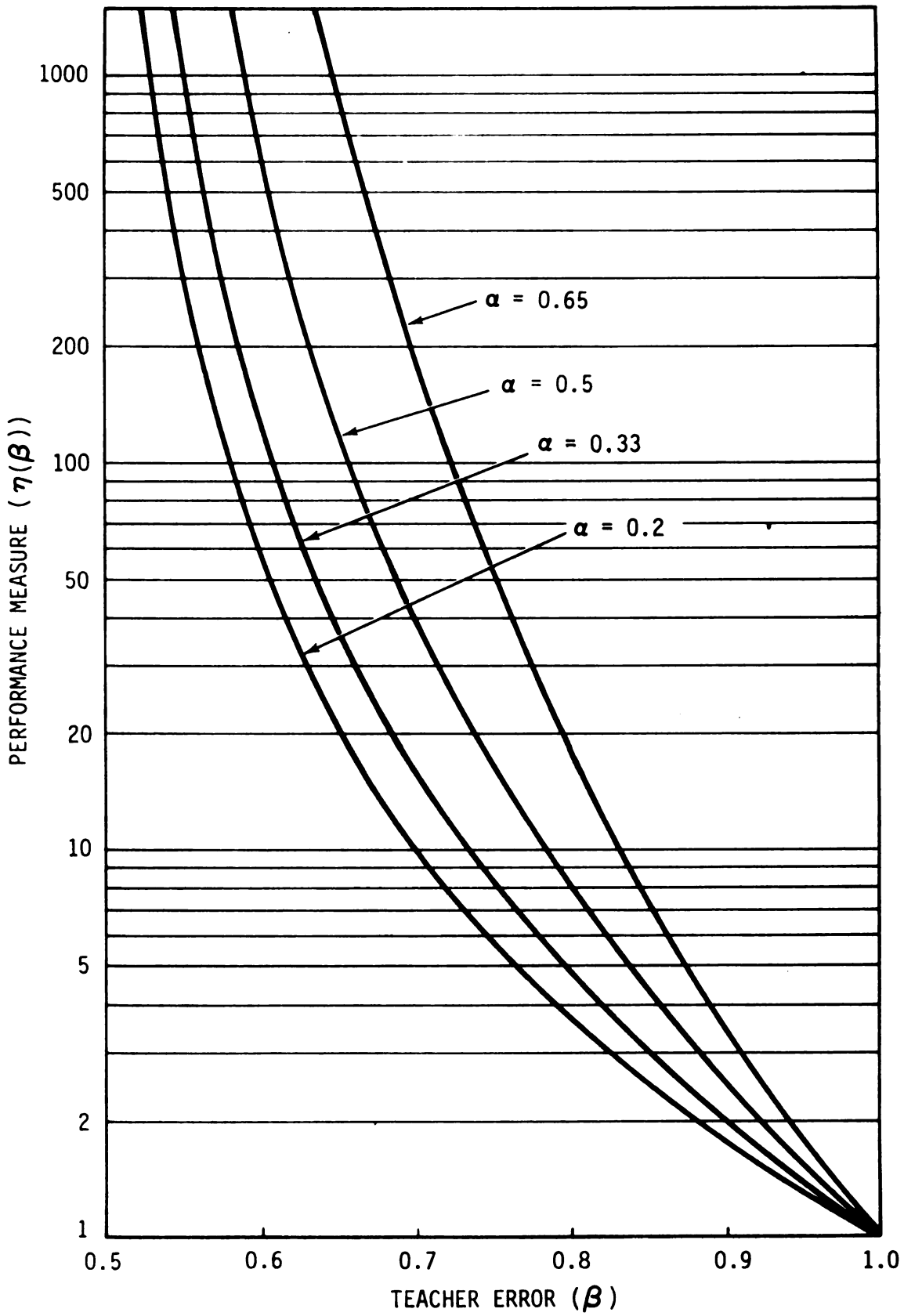

Figure 4.9. Measure of Relative Performance

matrix. For an imperfect teacher, the number of training patterns required to achieve R^* is approximately $N^*\eta(\underline{\beta}, N^*)$ and the cost incurred is

$$C = N^*\eta(\underline{\beta}, N^*)T(\underline{\beta}). \tag{4.79}$$

Differentiating (4.79) with respect to β_{ii} and equating to zero gives

$$N^*[\frac{\partial \eta}{\partial \beta_{ii}} T(\underline{\beta}) + \eta(\underline{\beta}, N^*) \frac{\partial T(\beta)}{\partial \beta_{ii}}] = 0, \quad i = 1, 2. \tag{4.80}$$

Solving this set of simultaneous equations for β_{11} and β_{22} gives the $\underline{\beta}$ matrix that minimizes C whenever the solution satisfies $\frac{1}{2} < \beta_{ii} < 1$.

As an example, suppose that a zero-one loss function is used and that $\beta_{11} = \beta_{22} = \beta$. Then $\eta(\beta)$ is given by (4.78). Let the cost of classification have the form (see Figure 4.10)

$$T(\beta) = a_1 + a_2(2\beta - 1)^{\xi} \tag{4.81}$$

Then

$$\frac{\partial C}{\partial \beta} = 2N^*(2\beta - 1)^{-1-2/(1-n\alpha)}[\frac{-2a_1}{1 - n\alpha} + a_2(\xi - \frac{2}{1 - n\alpha})(2\beta - 1)^{\xi}]. \tag{4.82}$$

and the minimum cost occurs at

$$\beta^* = \frac{1}{2} + \frac{1}{2}\left[\frac{2a_1}{a_2} \frac{1}{\xi(1 - n\alpha) - 2}\right]^{1/\xi} \tag{4.83}$$

provided $\frac{1}{2} < \beta^* < 1$. Otherwise the minimum occurs for a perfect teacher. From (4.83) it follows that a perfect teacher is best whenever

$$\xi \leq 2(1 + a_1/a_2)/(1 - n\alpha). \tag{4.84}$$

Figure 4.11 shows the relative cost plotted as a function of β for several parameter values. For this figure $n = 1$ and $\alpha = 1/3$.

Figure 4.11 indicates that when there is a cost associated with classifying the training patterns, a perfect teacher is not necessarily

the best. An imperfect teacher may provide an acceptable level of learning at a lower cost than a perfect teacher.

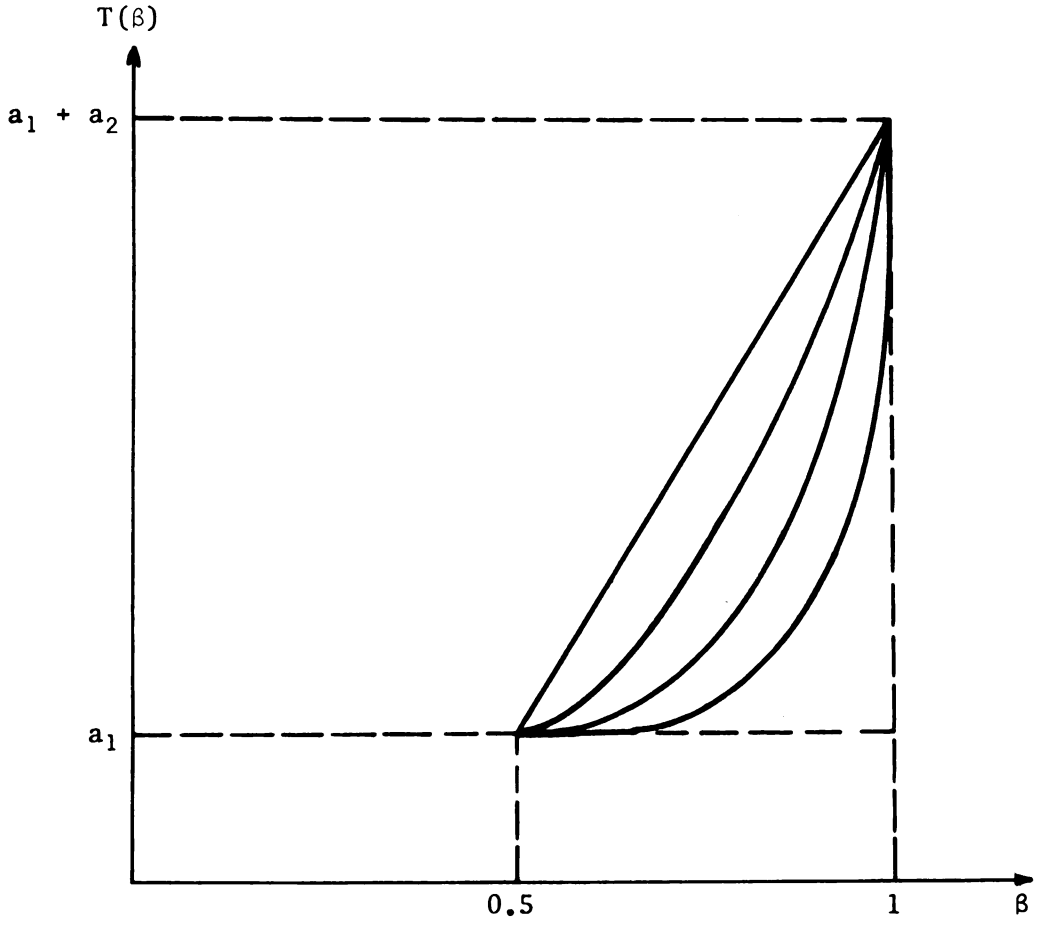

Figure 4.10. Teachers Cost for Classification

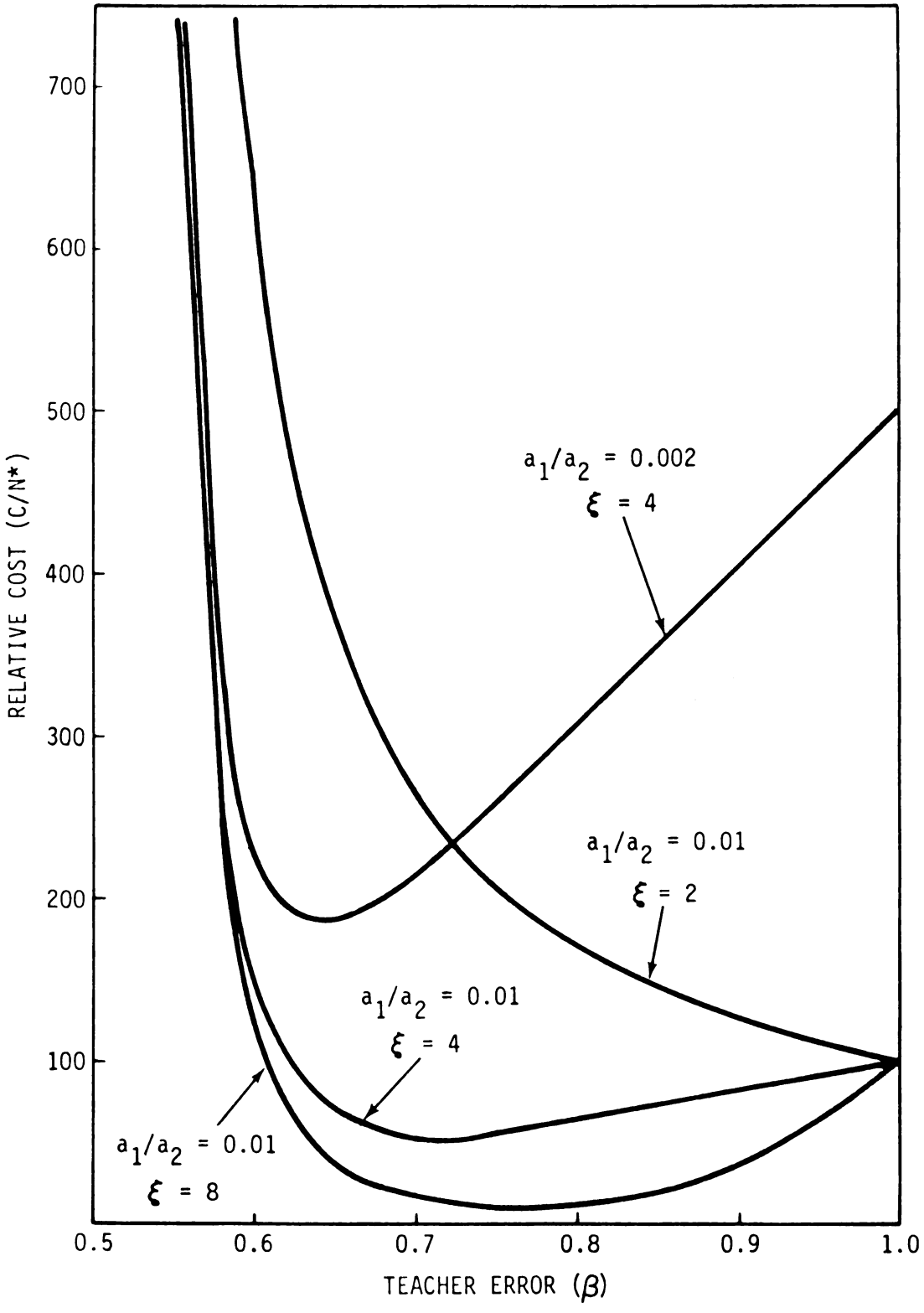

Figure 4.11. Relative Cost of Training

CHAPTER V

CONCLUSIONS

This chapter summarizes the main results of the thesis and discusses possibilities for future research.

5.1 SUMMARY

This thesis has been concerned with studying nonparametric procedures for learning to recognize patterns with an imperfect teacher. The statistical approach to pattern recognition was taken, and statistical decision theory was used as the tool for developing and evaluating learning algorithms. The objective was to study procedures for learning a Bayes decision rule using training patterns some of which were misclassified by an imperfect teacher.

In Chapter II, the general M-class statistical pattern recognition problem was outlined and a model for an imperfect teacher was proposed. The imperfect teacher was characterized by a matrix of conditional error probabilities. The main result of Chapter II was the development of a set of expressions relating the probability distributions of the perfect teacher with those of the imperfect teacher. These relations were one of the key factors used for developing in Chapter III learning procedures that required less restrictive assumptions than in previous studies. An example was also presented to show the asymptotic effects

of using an imperfect teacher with an algorithm designed for supervised learning. It was concluded that misclassified training patterns could prevent the convergence of supervised learning algorithms and that procedures are needed for use specifically with imperfect teachers.

In Chapter III a class of nonparametric procedures was proposed for learning with an imperfect teacher. The procedures required prior knowledge only of the nonsingular matrix of error probabilities characterizing the teacher and of whether the pattern random variable was discrete or continuous. The main result of the chapter was formal proofs of the convergence of the estimated decision rules to the Bayes rule in both the discrete case and the continuous case. The procedures were asymptotically optimal in the sense that the expected risk of the decision procedures converged to the Bayes risk with increasing number of training patterns. Theorems were also given establishing rates of convergence of the expected risk. These rate theorems provided guidelines for choosing the parameters in the estimators.

The two-class problem was studied in more detail in Chapter IV. The expected risk was expressed in a convenient form as a function of the Bayes discriminant function and of the probability that the estimated discriminant function was nonnegative. A large sample approximation was then developed to evaluate the expression for the expected risk. This approximation was used to study three examples of learning. The examples indicated that an imperfect teacher could have a significant effect on the rate of learning. When the teacher's error rate was high, learning occurred at a much slower rate than when a perfect teacher was used. The examples also indicated that the learning procedure would eventually perform better than a poor teacher.

A second large sample approximation was derived and used to establish several large sample properties of the expected risk. It was shown that the expected risk was a strictly decreasing function of the number of training patterns and of the teacher's error rate, β_{ii}.

A measure of relative performance was then proposed for quantitatively comparing an imperfect teacher with a perfect teacher. This performance measure was a measure of the additional number of training patterns required to compensate for an imperfect teacher. The measure was readily evaluated for the case of a zero-one loss function. For this important case, the measure was not a function of the underlying probability distributions. Thus it provided a very useful criterion for evaluating the effects of the imperfect teacher. It was concluded that a poor teacher was costly in the sense that many more training patterns were required to achieve the same expected risk as obtained with a perfect teacher.

By assigning a cost for classifying the training patterns, an overall cost of an imperfect teacher was derived. It was shown that when the cost of training is proportional to the quality of the teacher, an imperfect teacher may be preferred. By using an imperfect teacher, a learning system may be able to achieve a given level of performance at less overall cost than with a perfect teacher.

5.2 EXTENSIONS

The work in this thesis suggests several possibilities for future research. The concept of an imperfect teacher can be extended to many areas of pattern recognition.

This thesis has been concerned with one general class of learning

procedures. One can also look at several other types of supervised learning algorithms and attempt to develop analogous algorithms for learning with an imperfect teacher. For example, one might develop algorithms based on stochastic approximation, the method of potential functions, the k-nearest neighbor rule, or the various mean-square estimation methods for discriminant functions. All of these techniques for supervised learning will be affected by an imperfect teacher, and consequently they should be altered for use with an imperfect teacher.

The model proposed in Section 2.2 for the imperfect teacher could also be developed further. A natural extension would be to assume that the teacher's error probabilities are a function of the observed values of the training patterns. One might also consider a situation in which the teacher improves with time and experience.

It was assumed throughout this thesis that the matrix of error probabilities characterizing the teacher was known. An interesting problem would be to investigate conditions under which learning could occur without knowledge of the matrix of error probabilities. More structure would necessarily have to be assumed known about the probability distributions. This problem would be very close to one of unsupervised learning.

In Sections 4.5 and 4.6 measures of performance were proposed for quantitatively evaluating the effects of an imperfect teacher. These measures were chosen because they were intuitively appealing and mathematically tractable. Further work is needed in developing other measures of performance. Also the idea of a cost for the teacher could be pursued further.

Finally, one might consider ways of gradually phasing the teacher

out of the learning process. It was observed in the examples of Section 4.3 that the learning procedures eventually perform better than a poor teacher. One wonders whether the learned decision rules could eventually be used to classify the training patterns and thus eliminate the teacher. This might speed up the learning process since the learning procedure eventually has a smaller expected risk than the teacher. Eliminating the teacher would also be desirable when there is a cost of training associated with the teacher as in Section 4.6.

BIBLIOGRAPHY

BIBLIOGRAPHY

[A-1] N. Abramson and D. Braverman, "Learning to recognize patterns in a random environment," IRE Trans. Inform. Theory, vol. IT-8, pp. 58-63, 1962.

[A-2] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "Method of potential functions in the problem of restoration of functional converter characteristic by means of points observed randomly," Autom. and Remote Control, vol. 25, pp. 1705-1714, 1964.

[B-1] C. C. Blaydon, "Recursive algorithms for pattern classification," Rept. No. 520, Division of Engineering and Applied Physics, Harvard Univ., Cambridge, Mass., 1967.

[B-2] I. Bross, "Misclassification in 2 × 2 tables," Biometrics, vol. 10, pp. 478-486, 1954.

[C-1] T. Cacoullos, "Estimation of a multivariate density," Ann. Inst. Statist. Math., vol. 18, pp. 179-189, 1966.

[C-2] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inform. Theory, vol. IT-13, pp. 21-27, Jan. 1967.

[F-1] W. Feller, An Introduction to Probability Theory and Its Applications, vol. II. New York: Wiley, 1966.

[F-2] T. S. Ferguson, Mathematical Statistics: A Decision Theoretic Approach. New York: Academic Press, 1967.

[F-3] S. C. Fralick, "Learning to recognize patterns without a teacher," IEEE Trans. Inform. Theory, vol. IT-13, pp. 57-65, Jan. 1967.

[F-4] K. S. Fu, Sequential Methods in Pattern Recognition and Machine Learning. New York: Academic Press, 1968.

[F-5] K. Fukunaga, Introduction to Statistical Pattern Recognition. New York: Academic Press, 1972.

[H-1] Y. C. Ho and A. K. Agrawala, "On pattern classification algorithms, introduction and survey," Proc. IEEE, vol. 56, pp. 2101-2114, Dec. 1968.

[H-2] Y. C. Ho and R. L. Kashyap, "An algorithm for linear inequalities and its applications," IEEE Trans. Electronic Computers, vol. EC-14, pp. 683-688, Oct. 1965.

[I-1] IEEE Transactions on Computers, vol. C-20, Sept. 1971, (Special issue on feature extraction).

[I-2] B. G. Ishaku, "Feature extraction and Φ-function selection in Φ-systems," Ph.D. dissertation, Mich. State Univ., East Lansing, Mich., Oct. 1971.

[K-1] L. N. Kanal, Ed., Pattern Recognition. Wash., D. C.: Thompson, 1968.

[L-1] P. A. Lachenbruch, "Discriminant analysis when the initial samples are misclassified," Technometrics, vol. 8, pp. 657-662, Nov. 1966.

[L-2] M. Loeve, Probability Theory, 2nd ed. Princeton, N. J.: Van Nostrand, 1955.

[M-1] W. S. Meisel, "Potential functions in mathematical pattern recognition," IEEE Trans. Comput., vol. C-18, pp. 911-918, Oct. 1969.

[M-2] J. M. Mendel and K. S. Fu, Ed., Adaptive, Learning, and Pattern Recognition Systems: Theory and Applications. New York: Academic Press, 1970.

[M-3] A. M. Mucciardi and E. E. Gose, "A comparison of seven techniques for choosing subsets of pattern recognition properties," IEEE Trans. Comput., vol. C-20, pp. 1023-1031, Sept. 1971.

[M-4] V. K. Murthy, "Nonparametric estimation of multivariate densities with applications," in Proc. Int. Symp. on Multivariate Analysis, P. Krishnaih, Ed. New York: Academic Press, 1966, pp. 43-56.

[N-1] G. Nagy, "State of the art in pattern recognition," Proc. IEEE, vol. 56, pp. 836-862, May 1968.

[N-2] N. J. Nilsson, Learning Machines. New York: McGraw-Hill, 1965.

[P-1] E. Parzen, "An estimation of a probability density function and mode," Ann. Math. Statist., vol. 33, pp. 1065-1076, 1962.

[P-2] E. A. Patrick and J. C. Hancock, "Nonsupervised sequential classification and recognition of patterns," IEEE Trans. Inform. Theory, vol. IT-12, pp. 362-372, July 1966.

[P-3] M. L. Puri and P. K. Sen, Nonparametric Methods in Multivariate Analysis. New York: Wiley, 1971.

[R-1] H. Robbins, "The empirical Bayes approach to statistical decision problems," Ann. Math. Statist., vol. 35, pp. 1-20, 1964.

[R-2] H. L. Royden, Real Analysis, 2nd ed. New York: Macmillan, 1968.

[S-1] S. C. Schwartz, "Convergence of risk in adaptive pattern recognition procedures," Proc. 5th Allerton Conf. on Circuits and Systems Theory, pp. 800-806, 1967.

[S-2] S. C. Schwartz, "Estimation of probability density by an orthogonal series," Ann. Math. Statist., vol. 38, pp. 1261-1265, 1967.

[S-3] G. S. Sebestyen, Decision-Making Processes in Pattern Recognition. New York: Macmillan, 1962.

[S-4] K. Shanmugam and A. M. Breipohl, "An error correcting procedure for learning with an imperfect teacher," IEEE Trans. Systems, Man, and Cybernetics, vol. SMC-1, pp. 223-229, July 1971.

[S-5] K. Shanmugam, "Learning to recognize patterns with an imperfect teacher," Ph.D. dissertation, Oklahoma State Univ., Stillwater, Okla., 1970.

[S-6] K. Shanmugam, "A parametric procedure for learning with an imperfect teacher," IEEE Trans. Inform. Theory (Corresp.), vol. IT-18, pp. 300-302, March 1972.

[S-7] D. F. Specht, "Generation of polynomial discriminant functions for pattern recognition," IEEE Trans. Elect. Comput., vol. EC-16, pp. 308-319, June 1967.

[S-8] D. F. Specht, "Series estimation of a probability density function," Technometrics, vol. 13, pp. 409-424, May 1971.

[S-9] J. Spragins, "Learning without a teacher," IEEE Trans. Inform. Theory, vol. IT-11, pp. 544-549, April 1966.

[T-1] A. Tenenbein, "A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection," Technometrics, vol. 14, pp. 187-202, Feb. 1972.

[V-1] J. Van Ryzin, "Non-parametric Bayesian decision procedures for (pattern) classification with stochastic learning," Fourth Prague Conf. on Inform. Thy., Statistical Decision Functions, and Random Processes, 1965.

[V-2] J. Van Ryzin, "Bayes risk consistency of classification procedures using density estimation," Sankhya, Series A, vol. 28, pp. 261-270, 1966.

[V-3] J. Van Ryzin, "On strong consistency of density estimators," Ann. Math. Statist., vol. 40, pp. 1765-1772, 1969.

[W-1] S. Watanabe, Ed., Methodologies of Pattern Recognition. New York: Academic Press, 1969.

[W-2] A. W. Whitney and S. J. Dwyer, "Performance and implementation of the k-nearest neighbor rule with incorrectly identified training patterns," Proc. 4th Allerton Conf. on Circuit and System Theory, pp. 96-106, 1966.

[W-3] C. T. Wolverton and T. J. Wagner, "Asymptotically optimal discriminant functions for pattern classification," IEEE Trans. Inform. Theory, vol. IT-15, pp. 258-265, March 1969.

[Y-1] S. J. Yakowitz, "Unsupervised learning and the identification of finite mixtures," IEEE Trans. Inform. Theory, vol. IT-16, pp. 330-338, May 1970.

APPENDICES

APPENDIX A

OPTIMAL DECISION RULES

This appendix outlines the elements of statistical decision theory that are used in this thesis. The results, which are taken from the literature [F-2], [R-1], are presented from the viewpoint of pattern classification.

A problem in statistical decision theory consists of the following basic elements:

a.) A state space Ω with generic element ω representing a "state of nature."

b.) An action space A with generic element α representing an action available to the decision maker.

c.) A loss function $L(\alpha, \omega)$ defined on $A \times \Omega$ and representing the loss incurred in taking action α when the true state of nature is ω.

d.) An observable random variable X belonging to a space T on which a σ-finite measure ν is defined. When the true state of nature is ω, X has a specified probability density $f(\cdot|\omega)$ with respect to ν.

The type of decision making problem of concern here is the statistical classification problem. In classification problems the state space is a finite set, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_M\}$. The action space is $A = \{\alpha_1, \alpha_2, \ldots, \alpha_M\}$ where action α_i is "say ω_i was active to produce

the observed X."

The decision rules that are convenient for classification problems are the so called behavioral decision rules [F-2, pp. 198]. Let $\delta(x) = (\delta_1(x), \ldots, \delta_M(x))$ be a vector of real-valued measurable functions on T such that $\sum_{j=1}^{M} \delta_j(x) = 1$, $\delta_j(x) \geq 0$. Denote the class of all such functions by Δ. Then any $\delta \epsilon \Delta$ is a behavioral decision rule with $\delta_j(x) = Pr[\alpha_j | x]$ identified as the conditional probability of classifying X as coming from ω_j given that X = x is observed. In other words, a behavioral decision rule assigns to each $x \epsilon T$ a probability distribution on A.

A Bayesian strategy for classification involves the notion of a prior distribution on the state space Ω. This distribution is denoted by $P = (P_1, P_2, \ldots, P_M)$ with P_i being the prior probability assigned to state ω_i.

For any decision rule $\delta \epsilon \Delta$, the average risk associated with state of nature $\omega_i \epsilon \Omega$ is defined as

$$r(\delta, \omega_i) = E[\sum_{j=1}^{M} L(\alpha_j, \omega_i)\delta_j(x) | \omega_i] \tag{A.1}$$

where $E[\cdot | \omega_i]$ denotes expectation with respect to the conditional density $f(\cdot | \omega_i)$. The overall risk of misclassification with decision rule δ under the prior distribution P and conditional densities $f = (f(\cdot | \omega_1), \ldots, f(\cdot | \omega_M))$ is defined by

$$R(P, f, \delta) = \sum_{i=1}^{M} P_i r(\delta, \omega_i). \tag{A.2}$$

A decision rule $\delta_B \epsilon \Delta$ is said to be a Bayes rule with respect to the prior distribution P if and only if it satisfies

$$R(P, f, \delta_B) = \inf_{\delta \epsilon \Delta} R(P, f, \delta). \tag{A.3}$$

Thus a Bayes decision rule minimizes the overall risk of misclassification.

Substituting (A.1) into (A.2) gives

$$R(P, f, \delta) = \sum_{i=1}^{M} P_i \int_T \sum_{j=1}^{M} L(\alpha_j, \omega_i)\delta_j(x)f(x|\omega_i)\, d\nu(x)$$

$$= \sum_{j=1}^{M} \int_T \left[\sum_{i=1}^{M} P_i\, L(\alpha_j, \omega_i)f(x|\omega_i)\right]\delta_j(x)\, d\nu(x). \quad (A.4)$$

A choice of δ minimizing the risk in (A.4) is seen to be any behavioral decision rule $\delta_B = (\delta_{B1}, \delta_{B2}, \ldots, \delta_{BM})$ satisfying

$$\delta_{Bj}(x) = \begin{cases} 1 & \text{if} \quad D_j(x) < \min\limits_{k\neq j} D_k(x) \\ 0 & \text{if} \quad D_j(x) > \min\limits_{k\neq j} D_k(x) \\ \gamma_j & \text{if} \quad D_j(x) = \min\limits_{k\neq j} D_k(x) \end{cases} \quad (A.5)$$

where

$$D_k(x) = \sum_{i=1}^{M} P_i\, L(\alpha_k, \omega_i)f(x|\omega_i), \qquad k = 1, 2, \ldots, M \quad (A.6)$$

and where $\{\gamma_j\}$ is such that δ_B is a measurable function and $\gamma_j \geq 0$ and $\sum_{j=1}^{M} \delta_{Bj} = 1$.

The Bayes risk is the risk incurred by a Bayes decision rule. Substituting (A.6) into (A.4) results in the following expression for the Bayes risk:

$$R_B(P, f) = \sum_{j=1} \int_T D_j(x)\delta_{Bj}(x)\, d\nu(x). \quad (A.7)$$

The notation in (A.7) emphasizes the dependence of R_B on the prior distribution P and conditional densities f.

The identification of the above statistical classification problem with pattern recognition is immediate. The state space corresponds to the set of pattern classes, T to the feature space, and X to the pattern

to be classified. When the pattern is discrete valued, ν is taken to be counting measure; when X is real valued, ν is taken to be Lebesgue measure on Euclidean n-space. The optimal decision rules for pattern recognition are then the Bayes rules given by (A.5).

APPENDIX B

NONPARAMETRIC ESTIMATION OF DENSITY FUNCTIONS

This appendix presents a nonparametric method of estimating a probability density function. The method was first proposed by Parzen [P-1] for estimating a univariate density and later extended to multivariate densities by Cacoullos [C-1] and Murthy [M-4].

Let $X_1, X_2, \ldots, X_N$ be N independent observations on an n-dimensional random variable X with density function f(x). Let $K(y) = K(y_1, y_2, \ldots, y_n)$ be a Borel scalar function on Euclidean n-space E_n such that

$$K(y) \geq 0 \tag{B.1a}$$

$$\int_{E_n} K(y)\, dy = 1 \tag{B.1b}$$

$$\sup_{y \in E_n} K(y) < \infty \tag{B.1c}$$

and

$$||y||^n K(y) \to 0 \quad \text{as} \quad ||y|| \to \infty. \tag{B.1d}$$

The function $K(\cdot)$ is called the kernel of the estimator.

Define a function g_N on $E_n \times E_n$ by

$$g_N(x, y) = \frac{1}{h_N^n} K\left(\frac{x - y}{h_N}\right) \tag{B.2}$$

with $\{h_N\}$ a sequence of positive constants satisfying

$$h_N \to 0 \quad \text{as} \quad N \to \infty \tag{B.3a}$$

and

$$N h_N^n \to \infty \quad \text{as} \quad N \to \infty. \tag{B.3b}$$

Then the nonparametric estimator for the density function f(x) is defined as

$$\hat{f}_N(x) = \frac{1}{N} \sum_{i=1}^{N} g_N(x, X_i). \tag{B.4}$$

Cacoullos [C-1] has shown that the estimator $\hat{f}_N$ is asymptotically unbiased and consistent at all points of continuity of f.

There are many kernels which satisfy (B.1). Typical examples of univariate kernels are shown in Table B.1. Specht [S-7], [S-8] has investigated methods for approximating the Gaussian kernel (entry 3 of Table B.1) to obtain estimators which have fixed storage requirements.

In this dissertation the density estimator of (B.4) is not directly used; instead, linear combinations of the g_N functions are used to form estimators of discriminant functions. The following two lemmas concerning the g_N function are proven in [C-1]:

LEMMA B.1. Let g_N be a function on $E_n \times E_n$ defined by (B.1) thru (B.3). Then at every continuity point of a density function f(x),

$$\lim_{N\to\infty} h_N^{n(r-1)} E[g_N^r(x, X)] = f(x) \int_{E_n} K^r(y)\, dy. \tag{B.5}$$

LEMMA B.2. Let g_N be defined as above. If a probability density function f(x) has continuous partial derivatives of third order in a neighborhood of x, then

$$\lim_{N\to\infty} h_N^{-2}\{E[g_N(x, X)] - f(x)\} = I/2 \tag{B.6}$$

where

$$I = \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \int y_i y_j K(y)\, dy. \tag{B.7}$$

TABLE B.1. UNIVARIATE KERNELS

$K(y)$	$\int_{-\infty}^{\infty} K^2(y)\,dy$
$\frac{1}{2}$ for $\lvert y\rvert \leq 1$ 0 for $\lvert y\rvert > 1$	$\frac{1}{2}$
$1 - \lvert y\rvert$ for $\lvert y\rvert \leq 1$ 0 for $\lvert y\rvert > 1$	$\frac{2}{3}$
$(2\pi)^{-\frac{1}{2}} \exp(-y^2/2)$	$\frac{1}{2\sqrt{\pi}}$
$\frac{1}{2}\exp(-\lvert y\rvert)$	$\frac{1}{2}$
$\frac{1}{\pi}\frac{1}{1+y^2}$	$\frac{1}{\pi}$
$\frac{1}{\pi}\left(\frac{\sin y}{y}\right)^2$	$\frac{1}{3\pi}$

APPENDIX C

PROOF OF THEOREM 3.5

The proof of Theorem 3.5 is given in this appendix. The proof is as follows.

The first term in the bound of Lemma 3.2 is

$$T_1 = \sum_{j=1}^{N} \bar{L}_j \int q_{Nj}(x)\, d\nu(x) \tag{C.1}$$

where

$$q_{Nj}(x) = E_N[|\hat{v}_{Nj}(x) - P_j\, E[g_N(x, X)|\Lambda = j]|]. \tag{C.2}$$

Now from (3.77) it follows that

$$q_{Nj}(x) \leq N^{-\frac{1}{2}} \Big(\sum_{\ell=1}^{M} \Big(\sum_{i=1}^{M} b_{ji}^2 \beta_{i\ell} \Big) P_\ell\, E[g_N^2(x, X)|\Lambda = \ell] \Big)^{\frac{1}{2}} \tag{C.3}$$

So

$$T_1 \leq N^{-\frac{1}{2}} \sum_{j=1}^{M} \bar{L}_j \int \Big(\sum_{\ell=1}^{M} \Big(\sum_{i=1}^{M} b_{ji}^2 \beta_{i\ell} \Big) P_\ell\, E[g_N^2(x, X)|\Lambda = \ell] \Big)^{\frac{1}{2}} d\nu(x)$$

$$\leq N^{-\frac{1}{2}} \sum_{j=1}^{M} \bar{L}_j \Big\{ \int \prod_{s=1}^{n} (1 + |x_s|^{1+\xi}) \sum_{\ell=1}^{M} \Big(\sum_{i=1}^{M} b_{ji}^2 \beta_{i\ell} \Big) P_\ell$$

$$E[g_N^2(x, X)|\Lambda = \ell])\, d\nu(x) \Big\}^{\frac{1}{2}} \cdot \Big\{ \int \prod_{s=1}^{n} (1 + |x_s|^{1+\xi})^{-1}\, d\nu(x) \Big\}^{\frac{1}{2}}$$

$$= N^{-\frac{1}{2}} \sum_{j=1}^{M} \bar{L}_j \Big\{ \sum_{\ell=1}^{M} \Big(\sum_{i=1}^{M} b_{ji}^2 \beta_{i\ell} \Big) P_\ell \int \prod_{s=1}^{n} (1 + |x_s|^{1+\xi})$$

$$E[g_N^2(x, X)|\Lambda = \ell]\, d\nu(x) \Big\}^{\frac{1}{2}} \cdot \Big\{ \prod_{s=1}^{n} \int (1 + |x_s|^{1+\xi})^{-1}\, dx_s \Big\}^{\frac{1}{2}} \tag{C.4}$$

where the second inequality follows from Schwarz inequality. Van Ryzin [V-1] has shown that if (3.82b) is satisfied, then

$$\lim_{N\to\infty} h_N^n \int \Big[\prod_{s=1}^{n} (1 + |x_s|^{1+\xi})\Big] E[g_N^2(x, X)|\Lambda = \ell]\, d\nu(x)$$

$$= E\Big[\prod_{s=1}^{n} (1 + |x_s|^{1+\xi})|\Lambda = \ell\Big] \int K^2(y)\, d\nu(y) < \infty \qquad (C.5)$$

where $\xi = \min(\xi', 1/n)$. Since

$$\int (1 + |x_s|^{1+\xi})^{-1}\, dx_s < \infty, \qquad (C.6)$$

(C.5) along with (C.4) implies that

$$\lim_{N\to\infty} (Nh_N^n)^{\frac{1}{2}} T_1 \le \sum_{j=1}^{M} \overline{L}_j \Big\{\sum_{\ell=1}^{M} \Big(\sum_{i=1}^{M} b_{ji}^2 \beta_{i\ell}\Big) P_\ell$$

$$\cdot E\Big[\prod_{s=1}^{n} (1 + |x_s|^{1+\xi})|\Lambda = \ell\Big] \int K^2(y)\, d\nu(y)\Big\}^{\frac{1}{2}} \Big\{\prod_{s=1}^{n} \int (1 + |x_s|^{1+\xi})^{-1}\, dx_s\Big\}^{\frac{1}{2}}$$

$$\triangleq Q_1. \qquad (C.7)$$

Thus for large N

$$T_1 \le Q_1 (Nh_N^n)^{-\frac{1}{2}}. \qquad (C.8)$$

Now consider the second term in the bound of Lemma 3.2 (see equation (3.69)),

$$T_2 = \sum_{i=1}^{M} P_i \int \{E[g_N(x, X)|\Lambda = i] - f(x|\Lambda = i)\}$$

$$\cdot \Big\{\sum_{j=1}^{M} L_{ji} (\delta_{Bj}(x, v) - E_N[\hat{\delta}_{Nj}(x)]\Big\}\, d\nu(x) \qquad C.9)$$

Using (3.63) and the definition of the expectation operator gives

$$T_2 = \sum_{i=1}^{M} P_i \int \{ \int \frac{1}{h_N^n} K(\frac{x - y}{h_N}) \, f(y|\Lambda = i) \, d\nu(y) - f(x|\Lambda = i) \}$$

$$\cdot \{ \sum_{j=1}^{M} L_{ji} (\delta_{Bj}(x; \nu) - E_N[\hat{\delta}_{Nj}(x)] \} \, d\nu(x). \qquad (C.10)$$

Making a change of variable and applying Fubini's Theorem gives

$$T_2 = \sum_{i=1}^{M} P_i \int \int K(z)[f(x - h_N z|\Lambda = i) - f(x|\Lambda = i)]$$

$$\cdot \sum_{j=1}^{M} L_{ji} (\delta_{Bj}(x; \nu) - E_N[\hat{\delta}_{Nj}(x)]) \, d\nu(x) \, d\nu(z)$$

$$\leq \sum_{i=1}^{M} P_i \iint K(z) |f(x - h_N z|\Lambda = i) - f(x|\Lambda = i)|$$

$$\cdot \left| \sum_{j=1}^{M} L_{ji} (\delta_{Bj}(x; \nu) - E_N[\hat{\delta}_{Nj}(x)]) \right| d\nu(x) \, d\nu(z)$$

$$\leq 2 \sum_{i=1}^{M} \bar{L}_i P_i \iint K(z) \, \tau(h_N(z); f(\cdot|\Lambda = i)) \, d\nu(z) \qquad (C.11)$$

where the function $\tau(.;.)$ is defined by (3.81) and the last inequality follows from the fact that the magnitude of the second term in the integrand is bounded by $2\bar{L}_i$. Conditions (3.82a) and (3.82c) then imply that

$$h_N^{-\gamma} \, T_2 \leq 2C \int K(z) ||z||^{\gamma} \, d\nu(z) \sum_{i=1}^{M} P_i \bar{L}_i \triangleq Q_2. \qquad (C.12)$$

Thus for large N

$$\Delta R_N \leq T_1 + T_2 \leq Q_1 (N h_N^n)^{-\frac{1}{2}} + Q_2 \, h_N^{\gamma}. \qquad (C.13)$$

If h_N is then chosen as $h_N = 0(N^{-1/(n+\alpha)})$, it follows from (C.13) that

$$\Delta R_N \leq \begin{cases} Q_1 N^{-\gamma/(n+\alpha)} & \text{if} \quad \alpha > 2\gamma \\ Q_2 N^{-\alpha/2(n+\alpha)} & \text{if} \quad \alpha < 2\gamma \\ (Q_1 + Q_2) N^{-\alpha/(n+2\alpha)} & \text{if} \quad \alpha = 2\gamma. \end{cases} \tag{C.14}$$

This completes the proof of Theorem 3.5.