A NONPARAMETRIC MULTICLASS PARTITIONING METHOD FOR CLASSIFICATION

by

SAUL BRIAN GELFAND

S.B. Physics, Massachusetts Institute of Technology

(1978)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREES OF

MASTER OF SCIENCE and ENGINEER in
ELECTRICAL ENGINEERING

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January 1982

© Massachusetts Institute of Technology 1982

Signature of Author _____
Department of Electrical Engineering
December 22, 1981

Certified By: _____
Sanjoy K. Mitter
Thesis Supervisor

Accepted By: _____
Arthur C. Smith
Chairman, Department Committee

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
MAY 20 1982
LIBRARIES
Archives

A NONPARAMETRIC MULTICLASS PARTITIONING METHOD FOR CLASSIFICATION

by

SAUL BRIAN GELFAND

Submitted to the Department of Electrical Engineering
on January 14, 1982 in partial fulfillment of the
requirements for the Degrees of Master of Science and Engineer in
Electrical Engineering

ABSTRACT

c classes are characterized by unknown probability distributions. A data sample containing labelled vectors from each of the c classes is available. The data sample is divided into test and training samples. A classifier is designed based on the training sample and evaluated with the test sample. The classifier is also evaluated based on its asymptotic properties as sample size increases.

A multiclass recursive partitioning algorithm which generates a single binary decision tree for classifying all classes is given. The algorithm has the same desirable statistical and computational properties as Friedman's (1977) 2-class algorithm. Prior probabilities and losses are accounted for. A tree termination algorithm which terminates binary decision trees in a statistically optimal manner is given. Gordon and Olshen's (1978) results on the asymptotic Bayes risk efficiency of 2-class recursive partitioning algorithms are extended to the c-class case and applied to the combined partitioning/termination algorithm. Asymptotic efficiency and consistent risk estimates are obtained with independent test and training sequences.

Thesis Supervisor: Dr. Sanjoy K. Mitter
Title: Professor of Electrical Engineering and Computer Science

ACKNOWLEDGEMENT

I would like to thank Professor Sanjoy Mitter for his intellectual and financial support throughout the course of this research. I would also like to acknowledge Don Gustafson whose ideas contributed heavily to the initial phase of the work.

CONTENTS

I. INTRODUCTION

In this chapter we give a statement of the nonparametric multiclass classification problem and briefly review previous work. We give a chapter-by-chapter summary and a list of the contributions of the thesis.

1.1 Statement of the Problem and Previous Work

We state the nonparametric multiclass classification problem as follows. c classes are characterized by unknown probability distribution functions. A data sample containing labelled vectors from each of the c classes is available. The data sample is divided into test and training samples. A classifier is designed based on the training sample and evaluated with the test sample. The classifier can also be evaluated based on its asymptotic properties, as sample size increases.

The best known approach to nonparametric classification is the k-nearest-neighbor rule introduced by Cover and Hart [1]. Let $\underline{\alpha} \in \mathbb{R}^d$ be the vector to be classified. The k-nearest-neighbor rule labels $\underline{\alpha}$ by plurality logic on the labels of the k-nearest vectors to $\underline{\alpha}$ (with respect to some metric) in the training sample. Advantages of the k-nearest-neighbor rule include:

(1) asymptotic Bayes risk efficiency is obtained if k is chosen to be a function of the training sample size n_1 such that

$$k(n_1) \to \infty \quad (\text{as } n_1 \to \infty) \tag{1.1}$$

$$\frac{k(n_1)}{n_1} \to 0 \quad (\text{as } n_1 \to \infty) \tag{1.2}$$

(2) valid for multiclass

Disadvantages include:

(1) computationally expensive (distance to all vectors in training sample must be computed for each $\underline{\alpha}$ to be classified)

(2) not invariant to coordinate-by-coordinate strictly monotone transformations, such as scaling

(3) not obvious how to introduce prior probabilities and losses

Friedman [2] has recently introduced a 2-class recursive partitioning algorithm, motivated in part by the work of Anderson [3], Henderson and Fu [4], and Meisel and Michalopoulos [5]. The algorithm has desirable statistical and computational properties, and the resulting classifier is a binary decision tree. We discuss Friedman's algorithm in detail in Chapter 2. Advantages of the Friedman algorithm include:

(1) asymptotic Bayes risk efficiency is obtained if the algorithm is appropriately modified (Gordon and Olshen [6])

(2) computationally efficient

(3) invariant to coordinate-by-coordinate strictly monotone transformations

(4) prior probabilities and losses are accounted for

The main disadvantage of Friedman's algorithm is that it is only applicable to the 2-class case. Friedman gives a multiclass modification but we point out several problems with his approach. A major thrust of this thesis is to generalize Friedman's algorithm to the c-class case ($c > 2$) in a way which maintains the advantages listed above.

1.2 Chapter-by-Chapter Summary

In Chapter 2, recursive partitioning is discussed. Data and binary decision tree notation is introduced. Friedman's 2-class algorithm is reviewed. Friedman's algorithm generates a binary decision tree by maximizing the Kolmogorov-Smirnov distance between marginal cumulative distribution functions at each node. In practice, an estimate of the Kolmogorov-Smirnov distance based on a training sample is maximized. Adaptive and transgenerated coordinates can be used in designing the tree. Friedman suggests that the c-class problem be solved by solving c 2-class problems. The resulting classifier has c binary decision trees. Several problems with this approach are pointed out. A multiclass recursive partitioning algorithm is given which generates a single binary decision tree for classifying all classes. A binary decision tree is generated by minimizing the Bayes risk at each node. In practice, an estimate of the Bayes risk based on a training sample is minimized.

In Chapter 3, termination of binary decision trees is discussed. An algorithm is given for optimally terminating a binary decision tree. The algorithm yields the unique tree with the fewest nodes which minimizes the Bayes risk. In practice an estimate of the Bayes risk based on a test sample is minimized. The algorithm is generalized to cost functions other than Bayes risk. Test and training sample division is discussed.

In Chapter 4, asymptotic results for the nonparametric multiclass classification problem are derived and applied to decision rules generated by the partitioning and termination algorithms of Chapters 2 and 3. Asymptotic Bayes risk efficiency of a decision rule is defined. Gordon

and Olshen's results for the 2-class case are briefly reviewed and modified for the multiclass problem. Gordon and Olshen's approach involves consistent density estimation, although their densities are with respect to a general dominating measure which need not be known. Their results apply to decision rules which partition a Euclidean observation space into boxes and are invariant to coordinate-by-coordinate strictly monotone transformation. No assumptions are made concerning the underlying cumulative distribution functions. For simplicity, we give modifications for our algorithms which obtain asymptotic efficiency only for continuous marginal cumulative distribution functions. However, it is shown in general that consistent density estimates (with respect to a general dominating measure) yield asymptotically efficient decision rules for the multiclass case. The proof of this result, which is quite simple for the 2-class case, is surprisingly difficult for the c-class problem ($c > 2$). Here, a simple graph-theoretic technique is used to simplify the problem. The results are applied to decision rules generated by the partitioning and termination algorithms of Chapters 2 and 3. Asymptotic efficiency is obtained with independent test and training sequences. Consistent risk estimates are obtained, even though the estimates are based on the same test sequence used for termination. Finally, it is shown that the rate at which the risk of a binary decision tree terminated by the Chapter 3 termination algorithm approaches the optimal Bayes risk is at least as fast as that of the tree terminated by optimizing a termination parameter, as Friedman suggests.

In Chapter 5 we draw the conclusion that Friedman's recursive partitioning algorithm can be extended to the multiclass case, with the

same desirable statistical and computational properties. However, we also conclude that certain issues arise in the c-class problem ($c > 2$) that did not exist or were obscured for the 2-class case. Suggestions are given for further work.

1.3 Contributions of Thesis

We list the major contributions of the thesis.

(1) A multiclass recursive partitioning algorithm which generates a single binary decision tree for classifying all classes is given. The algorithm has the same desirable statistical and computational properties as Friedman's 2-class algorithm. Prior probabilities and losses are accounted for.

(2) A tree termination algorithm which yields the unique tree with fewest nodes which minimizes the Bayes risk is given (applicable to 2-class case also).

(3) Gordon and Olshen's results on the asymptotic Bayes risk efficiency of 2-class recursive partitioning algorithms are extended to the multiclass case and applied to our algorithms. Asymptotic efficiency and consistent risk estimates are obtained with independent training and test sequences. Convergence rates for different termination criteria are compared.

II. TREE GENERATION

In Chapter 1 the nonparametric multiclass classification problem was stated and previous work on the subject was reviewed. In particular, Friedman [2] has recently introduced a 2-class recursive partitioning algorithm with desirable statistical and computational properties. The resulting classifier is a binary decision tree.

In this chapter, recursive partitioning is discussed. Data and binary decision tree notation is introduced. Friedman's 2-class algorithm is reviewed. Friedman's algorithm generates a binary decision tree by maximizing the Kolmogorov-Smirnov distance between marginal cumulative distribution functions at each node. In practice, an estimate of the Kolmogorov-Smirnov distance based on a training sample is maximized. Adaptive and transgenerated coordinates can be used in designing the tree. Friedman suggests that the c-class problem be solved by solving c 2-class problems. The resulting classifier has c binary decision trees. Several problems with this approach are pointed out. A multiclass recursive partitioning algorithm is given which generates a single binary decision tree for classifying all classes. A binary decision tree is generated by minimizing the Bayes risk at each node. In practice, an estimate of the Bayes risk based on a training sample is minimized.

2.1 Data and Binary Decision Tree Notation

In the sequel we denote a sequence by $x^{(1)}, x^{(2)}, \ldots$ or $[x^{(n)}]$ and reserve $\{x^{(n)}\}$ for the set which contains the single element $x^{(n)}$.

We shall often be dealing with $s > 1$ sequences but will only be interested in the $n_m^{\underline{th}}$ element of the $m^{\underline{th}}$ sequence, $m = 1,\ldots,s$, which we refer to as $x^{(n_m)}$ rather than $x_m^{(n_m)}$.

Let $[{}_k\underline{\alpha}^{(n)}]$ be a sequence of d-dimensional random vectors from the $k^{\underline{th}}$ class, $k = 1,\ldots,c$. Let $A_k^{(n)}$ denote the $k^{\underline{th}}$-class sequence ${}_k\underline{\alpha}^{(1)},\ldots,{}_k\underline{\alpha}^{(n(k))}$, $k = 1,\ldots,c$, and let $A^{(n)}$ denote the sequence $A_1^{(n)},\ldots,A_c^{(n)}$, where

$$n = \sum_{k=1}^{c} n(k) \tag{2.1}$$

Let $\#_{n(k)}(S)$ = the number of vectors in $A_k^{(n)}$ and $S \subset \mathbb{R}^d$, $k = 1,\ldots,c$, and

$$\#_n(S) = \sum_{k=1}^{c} \#_{n(k)}(S) \tag{2.2}$$

We assume that ${}_k\underline{\alpha}^{(1)}$, ${}_k\underline{\alpha}^{(2)}$, ... are independent identically distributed (i.i.d.) random vectors, $k = 1,\ldots,c$, and $A_1^{(n)},\ldots,A_c^{(n)}$ are jointly independent. $A^{(n)}$ will be referred to as the data sequence; a realization of $A^{(n)}$ will be referred to as the data sample.

Let $A_k^{(n_1)}$ denote an $n_1(k)$ element i.i.d. subsequence of $A_k^{(n)}$, $k = 1,\ldots,c$, and let $A^{(n_1)}$ denote the sequence $A_1^{(n_1)},\ldots,A_c^{(n_1)}$, where

$$n_1 = \sum_{k=1}^{c} n_1(k) \tag{2.3}$$

Let $\#_{n_1(k)}(S)$ = the number of vectors in $A_k^{(n_1)}$ and $S \subset \mathbb{R}^d$, $k = 1,\ldots,c$, and

$$\#_{n_1}(S) = \sum_{k=1}^{c} \#_{n_1(k)}(S) \tag{2.4}$$

Similarly, let $A_k^{(n_2)}$ denote an $n_2(k)$ element i.i.d. subsequence of $A_k^{(n)}$, $k = 1,\ldots,c$, and let $A^{(n_2)}$ denote the sequence $A_1^{(n_2)},\ldots,A_c^{(n_2)}$, where

$$n_2 = \sum_{k=1}^{c} n_2(k) \tag{2.5}$$

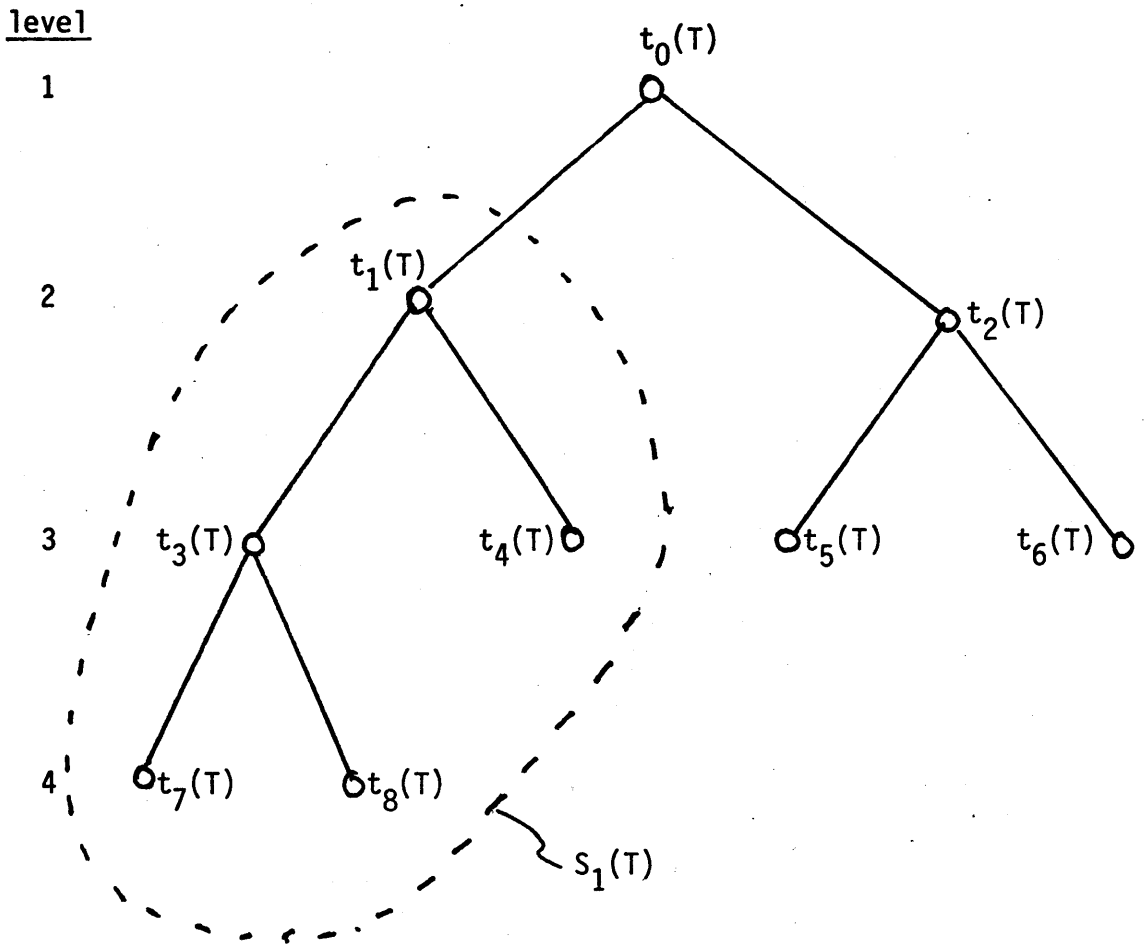
Let $\#_{n_2(k)}(S)$ = the number of vectors in $A_k^{(n_2)}$ and $S \subset \mathbb{R}^d$, $k = 1,\ldots,c$, and

$$\#_{n_2}(S) = \sum_{k=1}^{c} \#_{n_2(k)}(S) \tag{2.6}$$

Since $A_1^{(n)},\ldots,A_c^{(n)}$ are jointly independent, so are $A_1^{(n_1)},\ldots,A_c^{(n_1)}$ and $A_1^{(n_2)},\ldots,A_c^{(n_2)}$. We do <u>not</u> assume at this point that $A_k^{(n_1)}$, $A_k^{(n_2)}$ are independent, $k = 1,\ldots,c$. However, if $A_k^{(n_1)}$, $A_k^{(n_2)}$ are independent, $k = 1,\ldots,c$, then $A^{(n_1)}$, $A^{(n_2)}$ are independent. $A^{(n_1)}$, $A^{(n_2)}$ will be referred to as the training sequence and test sequence, respectively; a realization of $A^{(n_1)}$, $A^{(n_2)}$ will be referred to as the training sample and test sample, respectively. $A^{(n)}$, $A^{(n_1)}$, and $A^{(n_2)}$ are examples of the preliminary notational remarks.

Let $F_k(\underline{\alpha})$ be the joint cumulative distribution function of class k; $F_k(\alpha_i)$ be the marginal cumulative distribution function of class k for coordinate i; μ_k the probability measure of class k; π_k the prior probability of class k; ℓ_k the misclassification loss for class k. We assume there is no loss associated with correct classification.

A binary decision tree is shown in Figure 2.1 (cf. Meisel and



Note: node indices are monotonically increasing from left to right for any level, and from first to last level.

Figure 2.1 Binary Decision Tree T

Michalopoulos [5]). Let binary decision tree $T = \left\{\{t_0(T),\ldots,t_m(T)\},E(T)\right\}$, where $\{t_0(T),\ldots,t_m(T)\}$ are the nodes or decision points of T and E(T) are the edges of T; m(T) the number of nodes in T; o(T) the number of levels in T; $l_j(T)$, $r_j(T)$ pointers to the left and right subnodes of $t_j(T)$, respectively; $S_j(T)$ the subtree of T whose root node is $t_j(T)$. T is a finite binary decision tree if m(T) is finite. Let T_0 be a binary decision tree. $T \subset T_0$ if $t_0(T) = t_0(T_0)$, $\{t_0(T),\ldots,t_{m(T)}(T)\} \subset \{t_0(T_0),\ldots, t_{m(T_0)}(T_0)\}$, and $E(T) \subset E(T_0)$.

Example 2.1

For the (finite) binary decision tree T of Figure 2.1 we have $m(T) = 9$, $o(T) = 4$, $l_2(T) = 5$, $r_2(T) = 6$, and $S_1(T)$ as shown.

The decision parameters at node $t_j(T)$ are $i_j^*(T)$, $\alpha_{i_j^*}^*(T)$, $l_j(T)$, and $r_j(T)$, and are defined as follows. The root node $t_0(T)$ is the point at which the decision process begins. At node $t_j(T)$ the $i_j^{*\underline{th}}$ component of $\underline{\alpha}$ is used for discrimination. If $\alpha_{i_j^*} < \alpha_{i_j^*}^*$ the next decision will be made at $t_{l_j}(T)$. If $\alpha_{i_j^*} \geq \alpha_{i_j^*}^*$ the next decision will be made at $t_{r_j}(T)$. If $l_j(T) < 0$, $t_j(T)$ is a terminal node and $\underline{\alpha}$ is assigned to class $|l_j(T)|$. It is easily seen that a binary decision tree with these decision parameters can realize a decision rule that partitions $\mathbb{R}^d$ into boxes (rectangular parallelpipeds with sides parallel to the coordinate axes). The algorithms we discuss generate binary decision trees as the partitioning proceeds.

In Section 2.3 an algorithm is given which generates binary decision trees. In Chapter 3 an algorithm is given which optimally terminates binary decision trees. The tree termination algorithm requires all

nodes be labelled as if they were terminal nodes. This is most easily accomplished during partitioning. Thus the nodes of the binary decision tree before applying the tree termination algorithm actually have five decision parameters: $i_j^*(T)$, $\alpha_{i_j^*}^*(T)$, $l_j(T)$, $r_j(T)$, and $c_j(T)$, where class $c_j(T)$ is the label of $t_j(T)$ if $t_j(T)$ ultimately becomes a terminal node. After the tree termination algorithm is applied, $c_j(T)$ is no longer a decision parameter. The explicit dependence of quantities on trees will be dropped if the meaning is clear, e.g., $t_j \leftarrow t_j(T)$.

2.2 Friedman's Algorithm

Friedman's algorithm is based on a result of Stoller's [7] concerning univariate nonparametric 2-class classification. We assume

$$\ell_1\pi_1 = \ell_2\pi_2 \tag{2.7}$$

Consider the univariate case ($d = 2$). Stoller has solved the following problem: find α^* which minimizes the probability of error based on the decision rule:

$$\begin{aligned} &\alpha < \alpha^* && \text{decide class 1 or 2} \\ &\alpha \geqslant \alpha^* && \text{decide class 2 or 1} \end{aligned} \tag{2.8}$$

Let

$$D(\alpha) = |F_1(\alpha) - F_2(\alpha)| \tag{2.9}$$

be the Kolmogorov-Smirnov (K-S) distance between the two cumulative distribution functions. Stoller shows that

$$D(\alpha^*) = \max_{\alpha} D(\alpha) \tag{2.10}$$

If (2.8) does not provide sufficient discrimination, Stoller's procedure can be applied to $\{\alpha < \alpha^*\}$ and $\{\alpha \geq \alpha^*\}$ resulting in a decision rule with four intervals. In fact, Stoller's procedure can be applied recursively until all intervals in the decision rule meet a termination criterion. Terminal intervals are labelled as follows. Let [a,b) be a terminal interval which results from Stoller's procedure, and

$$\mu_{k^*}[a,b) = \max_{k=1,2} \mu_k[a,b) \tag{2.11}$$

Then class k* is the label of [a,b). Of course,

$$\mu_k[a,b) = F_k(b) - F_k(a) \tag{2.12}$$

Friedman extends Stoller's procedure to the multivariate case ($d \geq 2$) by solving the following problem: find α^*_{i*} and i* which minimize the probability of error based on the decision rule:

$$\begin{aligned} &\alpha_{i*} < \alpha^*_{i*} && \text{decide class 1 or 2} \\ &\alpha_{i*} \geq \alpha^*_{i*} && \text{decide class 2 or 1} \end{aligned} \tag{2.13}$$

Let

$$D(\alpha_i) = |F_1(\alpha_i) - F_2(\alpha_i)|, \tag{2.14}$$

the K-S distance between the two marginal cumulative distribution functions for coordinate i. Clearly,

$$D(\alpha^*_i) = \max_{\alpha_i} D(\alpha_i) \tag{2.15}$$

$$D(\alpha_{i^*}^*) = \max_{i=1,\ldots,d} D(\alpha_i^*) \tag{2.16}$$

As with the univariate case, Friedman's procedure can be applied recursively until all d-dimensional intervals or boxes in the decision rule meet a termination criterion. Terminal boxes are labelled as follows. Let B be a box which results from Friedman's procedure, and

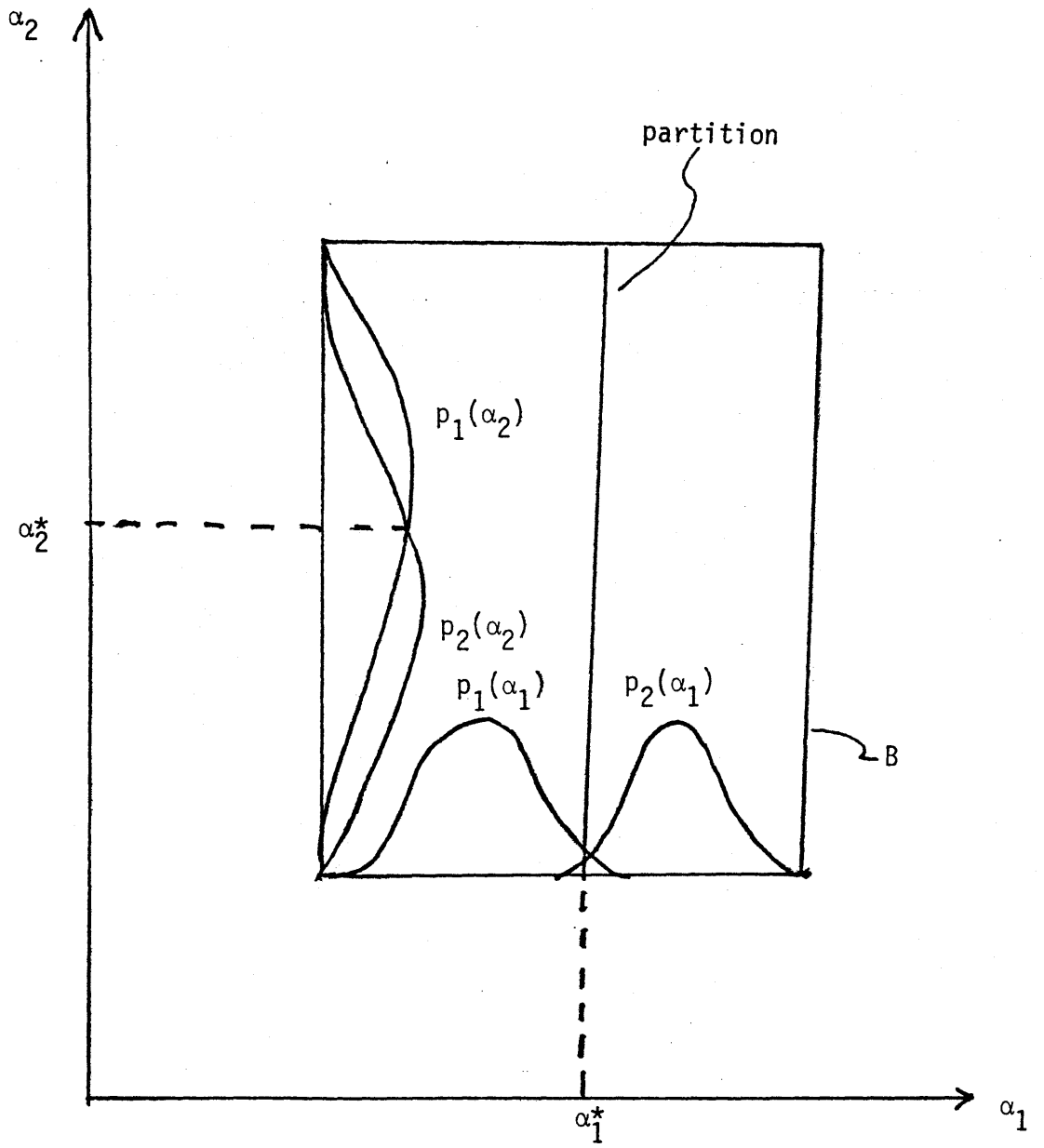
$$\mu_{k^*}(B) = \max_{k=1,2} \mu_k(B) \tag{2.17}$$

Then class k* is the label of B.

An example of Friedman's procedure for d = 2 is shown in Figure 2.2. A box $B \subset \mathbb{R}^2$ is to be partitioned, based on the within-box marginal cumulative distribution functions $F_k(\alpha_i)$ k,i = 1,2, or equivalently, the within-box marginal densities $p_k(\alpha_i) = \frac{dF_k(\alpha_i)}{d\alpha_i}$ k,i = 1,2 (the $p_k(\alpha_i)$ are shown). By inspection, the discrimination on coordinate 1 is greater than that on 2; consequently, i* = 1, $\alpha_{i^*}^* = \alpha_1^*$.

To apply Friedman's procedure to the nonparametric classification problem, $F_k(\alpha_i)$ and μ_k must be estimated from the training sample $A^{(n_1)}$. Let $\tilde{A}_k^{(n_1)}$ be a rearrangement of $A_k^{(n_1)}$ such that ${}_k\tilde{\alpha}_i^{(1)} \leq {}_k\tilde{\alpha}_i^{(2)} \leq \ldots \leq {}_k\tilde{\alpha}_i^{(n_1(k))}$, where ${}_k\tilde{\alpha}_i^{(j)}$ is the $i^{\underline{th}}$ component of ${}_k\underline{\tilde{\alpha}}^{(j)}$. An estimate of $F_k(\alpha_i)$ based on the training sample $A^{(n_1)}$ is:

$$F_k^{(n_1)}(\alpha_i) = \begin{cases} 0 & \alpha_i < {}_k\tilde{\alpha}_i^{(1)} \\ \dfrac{j}{n_1(k)} & {}_k\tilde{\alpha}_i^{(j)} \leq \alpha_i < {}_k\tilde{\alpha}_i^{(j+1)} \quad j = 1,\ldots,n_1(k)-1 \\ 1 & \alpha_i \geq {}_k\tilde{\alpha}_i^{(n_1(k))} \end{cases} \tag{2.18}$$



$i^* = 1$

$\alpha^*_{i^*} = \alpha_1^*$

Figure 2.2 Friedman's Algorithm (d = 2)

These (maximum likelihood) estimates are expected to work well with moderately large data bases and are pointwise consistent, i.e.,

$$F_k^{(n_1)}(\alpha_i) \xrightarrow{P} F_k(\alpha_i) \qquad (\text{as } n_1(k) \to \infty) \tag{2.19}$$

where $\xrightarrow{P}$ denotes convergence in probability. An estimate of $\mu_k(B)$ based on the training sample $A^{(n_1)}$ is

$$\hat{\mu}_k^{(n_1)}(B) = \frac{\#_{n_1(k)}(B)}{n_1(k)} \tag{2.20}$$

We note that (2.18) implies a preprocessing of data.

The partitions produced by Friedman's procedure can be associated with the nodes of a binary decision tree as described in Section 2.1. Termination criteria for Friedman's procedure are discussed in Chapter 3. Asymptotic properties are discussed in Chapter 4.

Adaptive and Transgenerated Coordinates

Adaptive and transgenerated coordinates are functions of the measured coordinates. They can be constructed as the partitioning proceeds, based on training subsamples. A great advantage of the Friedman algorithm is that many such coordinates can be added with little computational penalty.

Optimality

The Friedman algorithm is suboptimal in the sense that it only uses information from the marginal cumulative distribution functions. In certain pathological cases (cf. Gordon and Olshen [6]) this can result in poor performance. Gordon and Olshen modify Friedman's algorithm to

obtain asymptotic results. These modifications are discussed in Chapter 4. Their usefulness in the finite sample case appears to be highly data dependent.

Extension to Multiclass Problem

Friedman suggests that the c-class problem can be solved by solving c 2-class problems. In each 2-class problem, one of the classes is to be discriminated from all of the others taken as a group. A test vector is classified by directing it down all c trees and using plurality logic on the c terminal node training subsamples. There are two significant problems with this approach:

(1) Optimal labelling of decision regions is computationally expensive. This can be seen as follows. Let B_j be a terminal box which results from applying Friedman's 2-class algorithm to class j and classes 1,...,j-1,j+1,...,c taken as a group, and

$$\mu_{k^*}\left(\bigcap_{j=1}^{c} B_j\right) = \max_{k=1,\ldots,c} \mu_k\left(\bigcap_{j=1}^{c} B_j\right) \tag{2.21}$$

Then class k* is the label of $\bigcap_{j=1}^{c} B_j$. In practice, μ_k must be estimated. As estimate of μ_k based on the training sample $A^{(n_1)}$ is

$$\hat{\mu}_k^{(n_1)}\left(\bigcap_{j=1}^{c} B_j\right) = \frac{\#_{n_1(k)}\left(\bigcap_{j=1}^{c} B_j\right)}{n_1(k)} \tag{2.22}$$

Precomputation and storage of labels is expensive because of the number of $\bigcap_{j=1}^{c} B_j$. Online computation of labels is expensive because the training subsample at each node must be stored (not just $\#_{n_1(k)}(B_j)$, $k = 1,\ldots,c$),

and also because of the repeated computation to compute labels. Friedman appears to use a heuristic for labelling.

(2) It is unlikely that desirable asymptotic properties can be found. Since the c trees are generated independently, $\#_{n_1(k)}(\bigcap_{j=1}^{c} B_j)$ cannot be easily restricted. This property is crucial to Gordon and Olshen's results.

In the next section, a multiclass recursive partitioning algorithm is given which generates a single binary decision tree for classifying all classes. This circumvents the problems described above.

2.3 Multiclass Recursive Partitioning Algorithm

Friedman's procedure can be extended to the c-class case ($c > 2$) by solving the following problems: find α_{i*}^*, i*, m*, and n* which minimize the probability of error based on the decision rule

$$\begin{aligned} &\alpha_{i*} < \alpha_{i*}^* && \text{decide class m* or n*} \\ &\alpha_{i*} \geq \alpha_{i*}^* && \text{decide class n* or m*} \end{aligned} \tag{2.23}$$

Let

$$D_{m,n}(\alpha_i) = |F_m(\alpha_i) - F_n(\alpha_i)|, \tag{2.24}$$

the K-S distance between the marginal cumulative distributions of classes m and n for coordinate i. Clearly,

$$D_{m,n}(\alpha_i^*) = \max_{\alpha_i} D_{m,n}(\alpha_i) \tag{2.25}$$

$$D_{m,n}(\alpha_{i*}^*) = \max_{i=1,\ldots,d} D_{m,n}(\alpha_i^*) \tag{2.26}$$

$$D_{m^*,n^*}(\alpha^*_{i*}) = \max_{\substack{m=1,\ldots,c \\ n=1,\ldots,c \\ m\neq n}} D_{m,n}(\alpha^*_{i*}) \tag{2.27}$$

(2.24) - (2.27) replace (2.14) - (2.16) in the Friedman procedure. Instead of (2.17) we have

$$\mu_{k^*}(B) = \max_{k=1,\ldots,c} \mu_k(B) \tag{2.28}$$

Otherwise the procedures are the same. Note that m* and n* are not decision parameters.

To this point, it has been assumed that

$$\ell_1\pi_1 = \ldots = \ell_c\pi_c \tag{2.29}$$

To remove this restriction, we solve the following problems: find α^*_{i*}, i*, m* and n* which minimize the Bayes risk based on the decision rule

$$\begin{aligned} &\alpha_{i*} < \alpha^*_{i*} && \text{decide class } m^* \\ &\alpha_{i*} \geqslant \alpha^*_{i*} && \text{decide class } n^* \end{aligned} \tag{2.30}$$

First we solve: find α^*_i which minimizes the Bayes risk based on the decision rule

$$\begin{aligned} &\alpha_i < \alpha^*_i && \text{decide class } m \\ &\alpha_i \geqslant \alpha^*_i && \text{decide class } n \end{aligned} \tag{2.31}$$

The Bayes risk of decision rule (2.31) for $\alpha^*_i = \alpha_i$ is

$$R_{m,n}(\alpha_i) = \sum_{k=1}^{c} \ell_k \pi_k \sum_{\substack{j=1 \\ j\neq k}}^{c} \Pr\{\text{decide } j \mid k\}$$

$$= \ell_m \pi_m \mu_m[\alpha_i, \infty) + \ell_n \pi_n \mu_n(-\infty, \alpha_i) + \sum_{\substack{k=1 \\ k\neq m,n}}^{c} \ell_k \pi_k$$

$$= \ell_m \pi_m (1 - F_m(\alpha_i)) + \ell_n \pi_n F_n(\alpha_i) + \sum_{\substack{k=1 \\ k\neq m,n}}^{c} \ell_k \pi_k \tag{2.32}$$

Thus

$$R_{m,n}(\alpha_i^*) = \min_{\alpha_i} R_{m,n}(\alpha_i) \tag{2.33}$$

It follows that

$$R_{m,n}(\alpha_{i*}^*) = \min_{i=1,\ldots,d} R_{m,n}(\alpha_i^*) \tag{2.34}$$

$$R_{m*,n*}(\alpha_{i*}^*) = \min_{\substack{m=1,\ldots,c \\ n=1,\ldots,c \\ m\neq n}} R_{m,n}(\alpha_{i*}^*) \tag{2.35}$$

When this procedure is applied recursively, one or more classes may have zero measure on a box to be partitioned. Clearly, the sum in (2.32) and the minimization in (2.35) should only be over classes with positive measure. (2.32) - (2.35) replace (2.14) - (2.16) in the Friedman procedure. Instead of (2.17) we have

$$\ell_{k*} \pi_{k*} \mu_{k*}(B) = \max_{k=1,\ldots,c} \ell_k \pi_k \mu_k(B) \tag{2.36}$$

Otherwise, the procedures are the same.

In Chapter 3 an algorithm is given for optimally terminating a binary decision tree. The test sample is used both to terminate the tree and to estimate the risk of the terminated tree. This adds constraints to the problem of test and training sample division.

III. TREE TERMINATION

In Chapter 2, a multiclass recursive partitioning algorithm was given based upon the ideas of Friedman [2]. The resulting classifier is a binary decision tree. A binary decision tree is generated by minimizing the Bayes risk at each node. In practice, an estimate of the Bayes risk based on a training sample is minimized.

In this chapter, termination of binary decision trees is discussed. An algorithm is given for optimally terminating a binary decision tree. The algorithm yields the unique tree with the fewest nodes which minimizes the Bayes risk. In practice an estimate of the Bayes risk based on a test sample is minimized. The algorithm is generalized to cost functions other than Bayes risk. Test and training sample division is discussed.

3.1 Termination Criteria

Let $B_j(T)$ be the box associated with node $t_j(T)$. The Bayes risk of binary decision tree T is given by

$$R(T) = \sum_{k=1}^{C} \ell_k \pi_k \sum_{\substack{j=1 \\ j \neq k}}^{C} \Pr\{\text{decide } j | k\}$$

$$= \sum_{k=1}^{C} \ell_k \pi_k \sum_{j: t_j \in T,\ l_j < 0} \mu_k(B_j) I(|l_j| \neq k) \qquad (3.1)$$

where

$$I(|l_j| \neq k) = \begin{cases} 1 & |l_j| \neq k \\ 0 & |l_j| = k \end{cases} \tag{3.2}$$

An estimate of R(T) based on data sample $A^{(n)}$ is

$$\hat{R}^{(n)}(T) = \sum_{k=1}^{c} \ell_k \pi_k \sum_{j: t_j \epsilon T,\ l_j<0} \hat{\mu}_k^{(n)}(B_j) I(|l_j| \neq k)$$

$$= \sum_{k=1}^{c} \frac{\ell_k \pi_k}{n(k)} \sum_{j: t_j \epsilon T,\ l_j<0} \#_{n(k)}(B_j) I(|l_j| \neq k) \tag{3.3}$$

Similarly, $\hat{R}^{(n_1)}(T)$, $\hat{R}^{(n_2)}(T)$ are estimates of R(T) based on the training sample $A^{(n_1)}$ and the test sample $A^{(n_2)}$, respectively.

Let $T_0^{(n_1)}$ be the binary decision tree generated by applying the partitioning algorithm of Section 2.3 to the training sample $A^{(n_1)}$, with termination criteria that terminal nodes contain vectors only from a single class. Thus

$$\hat{R}^{(n_1)}(T_0^{(n_1)}) = \sum_{k=1}^{c} \frac{\ell_k \pi_k}{n_1(k)} \sum_{j: t_j \epsilon T_0^{(n_1)},\ l_j<0} \#_{n_1(k)}(B_j) I(|l_j| \neq k)$$

$$= 0 \tag{3.4}$$

i.e., the entire training sample is correctly classified. But if class distributions overlap then the optimal Bayes rule should <u>not</u> correctly classify the entire training sample. Thus we are led to examine termination criteria other than terminal nodes contain vectors from only a single class.

Friedman suggests that the number of training vectors at terminal nodes should be large enough to provide good estimates of the class

measures. Friedman introduces a termination parameter k = minimum number of training vectors at a terminal node, which is determined by minimizing an estimate of the Bayes risk based on the test sample $A^{(n_2)}$. But for large c and fixed k there are many possible terminal node populations. Thus the optimum k might be expected to vary from node to node.

In Section 3.2 an algorithm is given for optimally terminating a binary decision tree. The algorithm yields the unique tree with fewest nodes which minimizes the Bayes risk.

3.2 Optimal Tree Termination

Let T_0 be a finite binary decision tree. We want to solve the following problem: find $T_* \subset T_0$ such that

$$R(T_*) = \min_{T \subset T_0} R(T) \tag{3.5}$$

Consider the following tree termination algorithm:

Tree Termination Algorithm

$j = 1$

(i) if (Bayes risk does not increase when the descendents of $t_{m(T_0)-j}$ are deleted and $t_{m(T_0)-j}$ becomes a terminal node)

{delete descendents of $t_{m(T_0)-j}$ and make $t_{m(T_0)-j}$ a terminal node}

$j \leftarrow j+1$

if ($j \leq m(T_0)$) go to (i)

end

Theorem 3.1

Let $T_* \subset T_0$ be the binary decision tree which results from applying the tree termination algorithm to T_0. Then T_* is the unique tree with fewest nodes such that

$$R(T_*) = \min_{T \subset T_0} R(T) \tag{3.6}$$

Proof

We first derive a simplified deletion rule for deleting a node's descendents. Let T_b be the tree before the descendents of t_i are deleted and t_i becomes a terminal node; T_a the tree after the descendents of t_i are deleted and t_i becomes a terminal node. Expanding (3.1) gives

$$R(T_b) = \sum_{k=1}^{C} \ell_k \pi_k \Big(\sum_{j: t_j \in S_i,\ l_j<0} \mu_k(B_j) I(|l_j| \neq k) + \sum_{j: t_j \notin S_i,\ l_j<0} \mu_k(B_j) I(|l_j| \neq k) \Big) \tag{3.7}$$

and

$$R(T_a) = \sum_{k=1}^{C} \ell_k \pi_k \Big(\mu_k(B_i) I(c_i \neq k) + \sum_{j: t_j \notin S_i,\ l_j<0} \mu_k(B_j) I(|l_j| \neq k) \Big) \tag{3.8}$$

The descendents of t_i are deleted and t_i becomes a terminal node if

$$R(T_a) - R(T_b) = \sum_{k=1}^{C} \ell_k \pi_k \Big(\mu_k(B_i) I(c_i \neq k) - \sum_{j: t_j \in S_i,\ l_j<0} \mu_k(B_j) I(|l_j| \neq k) \Big) \tag{3.9}$$

$$\leqslant 0$$

The interpretation of (3.9) is that the decision to delete descendents of t_i and make t_i a terminal node depends only on t_i and S_i.

Given $T \subset T_0$ we construct $T' \subset T_0$ such that $R(T') \leq R(T)$. Let $T_0, T_1, \ldots, T_{o(T_0)-1} = T_*$ be the sequence of trees generated by applying the termination algorithm to T_0, where T_i is the tree after terminating the $o(T_0)-i^{\underline{th}}$ level of T_0; $t_{i_1},\ldots,t_{i_{z(i)}}$ the level i terminal nodes of $T \subset T_0$. $T' \subset T_0$ is constructed from T by the following algorithm:

$i = 1$

(i) $j = 1$

(ii) if (there exists a nonterminal node $t_k(T_{o(T_0)-i})$ such that $t_{i_j}(T) = t_k(T_{o(T_0)-i})$)

{replace $t_{i_j}(T)$ by $S_k(T_{o(T_0)-i})$}

$j \leftarrow j+1$

if $(j \leq z(i))$ go to (ii)

$i \leftarrow i+1$

if $(i \leq o(T))$ go to (i)

end

An example of the construction of T' is shown in Figure 3.1. Since $T \subset T_0$ and $S_k(T_{o(T_0)-i})$ is a subtree of $T_{o(T_0)-i} \subset T_0$, it follows that $T' \subset T_0$. Now consider a $t_{i_j}(T)$ which was replaced by $S_k(T_{o(T_0)-i})$. Since the descendents of $t_k(T_{o(T_0)-i})$ were not deleted by the termination algorithm, we have from (3.9) that $R(T') < R(T)$ $(T_b = T', T_a = T)$. If we allow that no $t_{i_j}(T)$ was replaced, we have $R(T') \leq R(T)$.

Observe that T_* results from applying the termination algorithm to T'. This follows from (3.9) and induction on the nodes of T'. Thus $R(T_*) \leq R(T')$ which implies $R(T_*) \leq R(T)$. Since this is true for any $T \subset T_0$ we have $R(T_*) = \min_{T \subset T_0} R(T)$. Now suppose there exists $T \subset T_0$ such

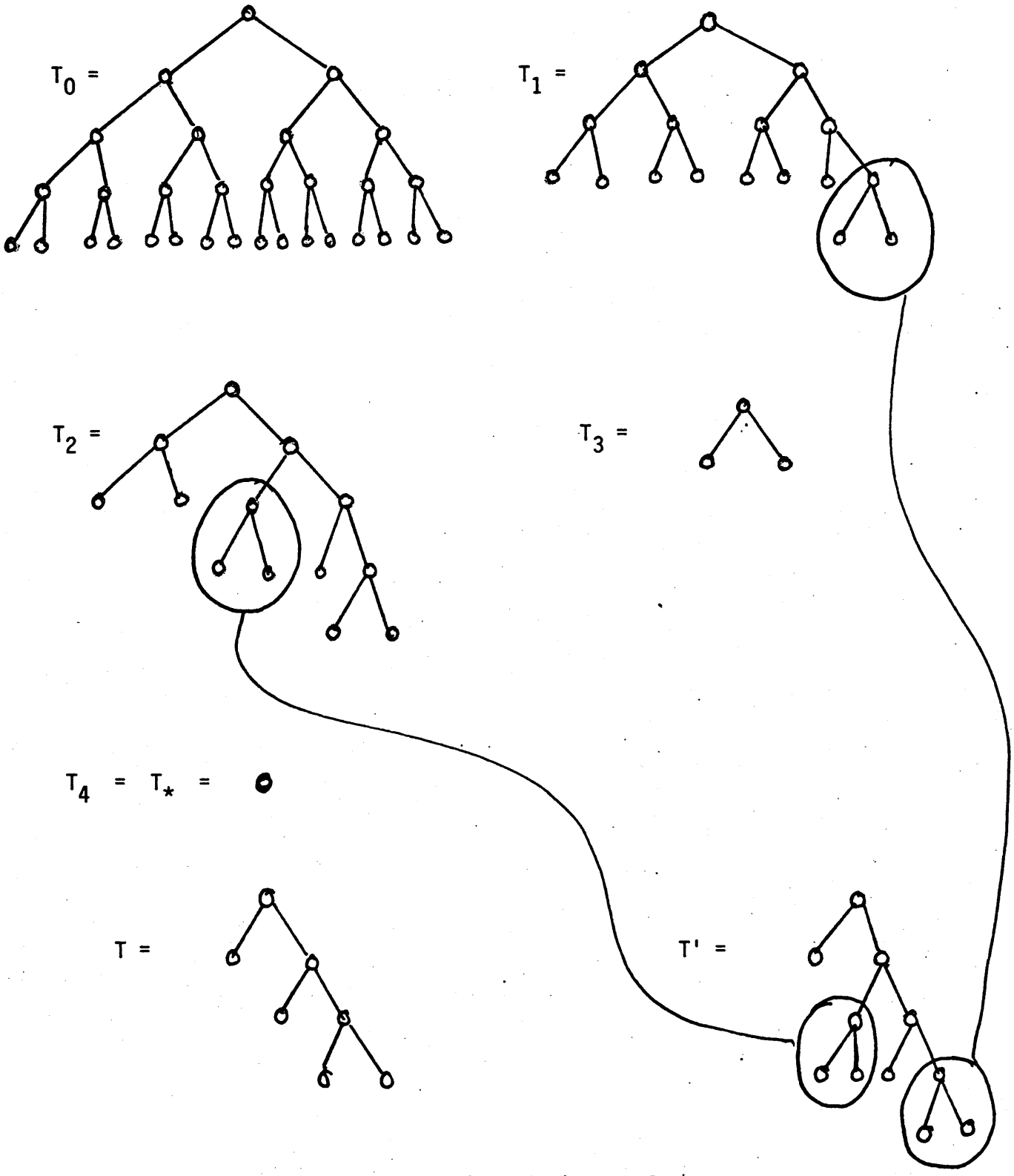


Figure 3.1 Construction of $T' \subset T_0$ from $T \subset T_0$

that $T \neq T_*$, $m(T) \leq m(T_*)$, and $R(T) = R(T_*)$. Then there are nonterminal nodes of T_* which are terminal nodes of T. Let $t_i(T)$ be a terminal node of T such that $t_j(T_*) = t_i(T)$ is a nonterminal node of T_*; T' be T with $t_i(T)$ replaced by $S_j(T_*)$. Since the descendents of $t_j(T_*)$ were not deleted by the termination algorithm, we have from (3.9) that $R(T') < R(T)$ ($T_b = T'$, $T_a = T$). But $R(T_*) \leq R(T')$ implies that $R(T_*) < R(T)$, a contradiction.|

In practice $T_0 = T_0^{(n_1)}$ and an estimate of the Bayes risk based on the test sample $A^{(n_2)}$ is minimized. Let $T_*^{(n_1,n_2)}$ be the binary decision tree which results from applying the termination algorithm to $T_0^{(n_1)}$ based on the test sample $A^{(n_2)}$. Then

$$\hat{R}^{(n_2)}(T_*^{(n_1,n_2)}) = \min_{T \subset T_0^{(n_1)}} \hat{R}^{(n_2)}(T) \tag{3.10}$$

Finally we give the simplified deletion rule based on the test sample $A^{(n_2)}$. The descendents of t_i are deleted and t_i becomes a terminal node if:

$$\hat{R}^{(n_2)}(T_a) - \hat{R}^{(n_2)}(T_b) = \sum_{k=1}^{c} \frac{\ell_k \pi_k}{n_2(k)} \left(\#_{n_2(k)}(B_i) I(c_i \neq k) - \sum_{j: t_j \in S_i,\ l_j < 0} \#_{n_2(k)}(B_j) I(|l_j| \neq k) \right)$$

$$\leq 0 \tag{3.11}$$

Cost Functions Other Than Bayes Risk

Inspection of the proof of Theorem 3.1 shows that cost functions $Q(T)$ of the form

$$Q(T) = \sum_{j: t_j \in T,\ l_j < 0} q(t_j) \tag{3.12}$$

can be optimized by the termination algorithm.

3.3 Test and Training Sample Division

$T_*^{(n_1,n_2)}$ is generated by applying the termination algorithm to $T_0^{(n_1)}$ based on the test sample $A^{(n_2)}$. $\hat{R}^{(n_2)}(T_*^{(n_1,n_2)})$ is an estimate of $R(T_*^{(n_1,n_2)})$ based on the same test sample $A^{(n_2)}$. The asymptotic implications of this procedure are discussed in Chapter 4. We mention here that $A^{(n_1)}$, $A^{(n_2)}$ must be independent and n_1, n_2 must be increased in a prescribed manner to obtain desirable asymptotic properties. Since we want to use the entire data sample it follows that

$$n(k) = n_1(k) + n_2(k) \qquad k = 1,\ldots,c \tag{3.13}$$

which implies

$$n = n_1 + n_2 \tag{3.14}$$

In addition, common sense indicates that we must have

$$n_1(k) \simeq n_2(k) \qquad k = 1,\ldots,c \tag{3.15}$$

which implies

$$n_1 \simeq n_2 \tag{3.16}$$

We complete the discussion of tree termination by drawing the following analogy. The 2-step procedure of tree generation and termination is similar to the solution of a general regression problem if tree generation is associated with generating models of different order, and tree

termination with determining the optimal order.

In Chapter 4 we investigate asymptotic properties for multiclass classification algorithms in general, and for the algorithms given in Chapters 2 and 3 in particular.

IV. ASYMPTOTIC RESULTS

In Chapters 2 and 3 a 2-step procedure was given for solving the nonparametric multiclass classification problem. A multiclass recursive partitioning algorithm generates a binary decision tree by minimizing the Bayes risk at each node. In practice, an estimate of the Bayes risk based on a training sample is minimized. A termination algorithm yields the unique tree with fewest nodes which minimizes the Bayes risk. In practice, an estimate of the Bayes risk based on a test sample is minimized.

In this chapter, asymptotic results for the nonparametric multiclass classification problem are derived and applied to decision rules generated by the partitioning and termination algorithms of Chapters 2 and 3. Asymptotic Bayes risk efficiency of a decision rule is defined. Gordon and Olshen's [6] results for the 2-class case are briefly reviewed and modified for the multiclass problem. Gordon and Olshen's approach involves consistent density estimation, although their densities are with respect to a general dominating measure which need not be known. Their results apply to decision rules which partition a Euclidean observation space into boxes and are invariant to coordinate-by-coordinate strictly monotone transformations. No assumptions are made concerning the underlying cumulative distribution functions. For simplicity, we give modifications for our algorithms which obtain asymptotic efficiency only for continuous marginal cumulative distribution functions. However, it is shown in general that consistent density estimates (with respect to a general dominating measure) yield asymptotically efficient decision

rules for the multiclass case. The proof of this result, which is quite simple for the 2-class case, is surprisingly difficult for the c-class problem ($c > 2$). Here, a simple graph-theoretic technique is used to simplify the problem. The results are applied to decision rules generated by the partitioning and termination algorithms of Chapters 2 and 3. Asymptotic efficiency is obtained with independent test and training sequences. Consistent risk estimates are obtained, even though the estimates are based on the same test sequence used for termination. Finally, it is shown that the rate at which the risk of a binary decision tree terminated by the Chapter 3 termination algorithm approaches the optimal Bayes risk is at least as fast as that of the tree terminated by optimizing a termination parameter, as Friedman suggests.

4.1 Measure-Consistent Density Estimation

Let $\mathfrak{a}^{(n)}$ be a decision rule based on the data sequence $A^{(n)}$. Note that $R(\mathfrak{a}^{(n)})$ is a random variable; by convention, the expectation has not been taken over the data sequence. We say that $\mathfrak{a}^{(n)}$ is asymptotically Bayes risk efficient if

$$R(\mathfrak{a}^{(n)}) \xrightarrow{P} R(\mathfrak{a}_B) = \inf_{\mathfrak{a}} R(\mathfrak{a}) \qquad (\text{as } n \to \infty) \tag{4.1}$$

Many approaches to showing asymptotic efficiency of decision rules have involved consistent density estimation. In general, these results have shown that if the underlying cumulative distribution functions are Lebesgue absolutely continuous, then pointwise consistent density estimates yield asymptotically efficient decision rules (cf. Fix and Hodges [8], Van Ryzin [9]). If a function is Lebesgue absolutely continuous then it is

continuous and has a derivative almost-everywhere. Thus, if a density is singular or even if it is discontinuous on a set of Lebesgue measure >0 then the corresponding cumulative distribution function is not Lebesgue absolutely continuous and the results do not apply.

Let $\nu = \sum_{i=1}^{c} \phi_i \mu_i$, where $\sum_{i=1}^{c} \phi_i = 1$, $\phi_i > 0$, $i = 1,\ldots,c$. Then $\mu_1,\mu_2,\ldots,\mu_c$ are absolutely continuous with respect to ν, i.e., $\nu(S) = 0$ implies $\mu_i(S) = 0$, $i = 1,\ldots,c$, where S is any measurable set. From the Radon-Nikodym theorem, there exists measurable functions $\frac{d\mu_1}{d\nu}, \ldots, \frac{d\mu_c}{d\nu}$ such that

$$\mu_i(S) = \int_S \frac{d\mu_i}{d\nu}\, d\nu, \qquad i = 1,\ldots,c \tag{4.2}$$

The $\left[\frac{d\mu_i}{d\nu}\right]$ are Radon-Nikodym derivatives and have the interpretation of densities, but with respect to the measure ν.

Let $\frac{\hat{d\mu}_1^{(n)}}{d\nu}, \frac{\hat{d\mu}_2^{(n)}}{d\nu}, \ldots, \frac{\hat{d\mu}_c^{(n)}}{d\nu}$ be measurable functions such that

$$\nu\left\{\underline{\alpha} : \left|\frac{\hat{d\mu}_i^{(n)}}{d\nu}(\underline{\alpha}) - \frac{d\mu_i}{d\nu}(\underline{\alpha})\right| > \varepsilon\right\} \xrightarrow{P} 0 \qquad (\text{as } n \to \infty) \tag{4.3}$$

for all $\varepsilon > 0$. We say that $\frac{\hat{d\mu}_i^{(n)}}{d\nu}$ is a measure-consistent estimate of $\frac{d\mu_i}{d\nu}$ based on the data sequence $A^{(n)}$. Gordon and Olshen [6] have shown that measure-consistent density estimates yield asymptotically efficient decision rules for the 2-class case (Section 4.2). They give modifications which can be applied to decision rules which partition $\mathbb{R}^d$ into boxes and are invariant to coordinate-by-coordinate strictly monotone

transformations. Their modified rules yield measure-consistent density estimates and consequently are asymptotically efficient. No assumptions are made concerning the underlying cumulative distribution functions. For simplicity, we give modifications for our algorithms which yield measure-consistent density estimates and consequently asymptotically efficient decision rules only for continuous marginal cumulative distribution functions. We refer the reader to Gordon and Olshen's paper for the general case.

Gordon and Olshen introduce the idea of a p-quantile cut. We only consider the case of continuous marginal cumulative distribution functions. Given a box B, a p-quantile cut on the $i^{\underline{th}}$ coordinate has been achieved at α_i^* if

$$\max\left\{\#_n(B \cap \{\alpha_i < \alpha_i^*\}),\ \#_n(B \cap \{\alpha_i \geq \alpha_i^*\})\right\} \leq p \cdot \#_n(B) \tag{4.4}$$

i.e., if at most a fraction p of the vectors in B land in either daughter box. Note that it is unimportant how vectors with $\alpha_i = \alpha_i^*$ are assigned to the daughter boxes since continuous marginal cumulative distribution functions imply $\nu\{\alpha_i^*\} = 0$.

Let $\hat{u}_{GO}^{(n)}$ be a decision rule which partitions $\mathbb{R}^d$ into a finite set of boxes and is invariant to coordinate-by-coordinate strictly monotone transformations, and let $B^{(n)}(\underline{\alpha})$ be the unique box which contains $\underline{\alpha}$. Let $\hat{\nu}^{(n)}$, $\hat{\mu}_1^{(n)}$, $\hat{\mu}_2^{(n)}$, ..., $\hat{\mu}_c^{(n)}$ be the usual set-wise consistent estimates of ν, μ_1, μ_2, ..., μ_c based on the data sequence $A^{(n)}$, i.e.,

$$\hat{\mu}_i^{(n)}(S) = \frac{\#_{n(i)}(S)}{n(i)} \qquad i = 1,\ldots,c \tag{4.5}$$

$$\hat{\nu}^{(n)}(S) = \sum_{i=1}^{c} \phi_i \hat{\mu}_i^{(n)}(S) \tag{4.6}$$

From simple properties of measurable sets and functions $\frac{\hat{\mu}_i^{(n)}(B^{(n)}(\underline{\alpha}))}{\hat{\nu}^{(n)}(B^{(n)}(\underline{\alpha}))}$, $i = 1,\ldots,c$ are measurable functions. The following theorem follows from Gordon and Olshen's results.

Theorem 4.1

Let $p \in [\frac{1}{2},1)$. If

(1) there exists fixed positive θ such that for n large enough

$$\frac{n(i)}{n} \in (\theta, 1-\theta), \qquad i = 1,2 \tag{4.7}$$

(2) there exists k(n) such that

$$\frac{k(n)}{\sqrt{n}} \to \infty \qquad (\text{as } n \to \infty) \tag{4.8}$$

$$\frac{k(n)}{\sqrt{n}} \to 0 \qquad (\text{as } n \to \infty) \tag{4.9}$$

(3) $$\hat{\nu}^{(n)}\{\underline{\alpha} : \#_n(B^{(n)}(\underline{\alpha})) > k(n)\} \xrightarrow{P} 1 \qquad (\text{as } n \to \infty) \tag{4.10}$$

(4) an increasingly large number of p-quantile cuts are made on every coordinate

then

$$\nu\left\{\underline{\alpha} : \left|\frac{\hat{\mu}_i^{(n)}(B^{(n)}(\underline{\alpha}))}{\hat{\nu}^{(n)}(B^{(n)}(\underline{\alpha}))} - \frac{d\mu_i}{d\nu}(\underline{\alpha})\right| > \varepsilon\right\} \xrightarrow{P} 0 \qquad (\text{as } n \to \infty) \tag{4.11}$$

for all $\varepsilon > 0$, $i = 1,2$.

Proof

See Gordon and Olshen [6]. |

Theorem 4.1a says that under the stated conditions, $\frac{\hat{d\mu_i}}{d\nu}(\underline{\alpha}) = \frac{\hat{\mu}_i^{(n)}(B^{(n)}(\underline{\alpha}))}{\hat{\nu}^{(n)}(B^{(n)}(\underline{\alpha}))}$ is a measure-consistent estimate of $\frac{d\mu_i}{d\nu}(\underline{\alpha})$, $i = 1,2$. Theorem 4.1a can be used to modify Friedman's 2-class algorithm to obtain measure-consistent density estimates. We call this modified algorithm Gordon and Olshen's 2-class algorithm. Since Gordon and Olshen are only concerned with asymptotic results, the algorithm is applied to the entire data sequence $A^{(n)}$ rather than the training sequence $A^{(n_1)}$. We shall have more to say about this in Section 4.3. Let $p \in [\frac{1}{2},1)$, $k(n) = n^{5/8}$, and ω a large integer. We are given a box B.

Gordon and Olshen's 2-class Algorithm

```
if (coordinate i has not been partitioned in the ω most recent par-
    titionings which led to B)

   {i* ← i
    α*_i* ← median α_i for data vectors in B}

else

   {compute i*, α*_i* from (2.14) - (2.16)
    α*_i* ← max{min{p quantile, α*_i*},
                min{1-p quantile, α*_i*},
                  min{p quantile, 1-p quantile}}}

if (termination criteria not satisfied)

   {partition B on coordinate i* at α*_i*}

else

   {do not partition B:  B is a terminal box}

end
```

The termination criteria are:

(1) $\#_n(B) = \#_{n(i)}(B)$ for some $i = 1,2$ OR (4.12)

(2) $\min\left\{\#_n(B \cap \{\alpha_{i*} < \alpha^*_{i*}\}), \#_n(B \cap \{\alpha_{i*} \geqslant \alpha^*_{i*}\})\right\} \leqslant k(n)$ (4.13)

We now consider the multiclass case. Inspection of Gordon and Olshen's results indicate that the generalization of Theorem 4.1a for $c > 2$ is true.

Theorem 4.1

Let $p \in [\frac{1}{2},1)$. If there exists fixed positive θ such that for n large enough

$$\frac{n(i)}{n} \in (\theta, 1-\theta), \qquad i = 1,\ldots,c \tag{4.14}$$

(2), (3), and (4) as in Theorem 4.1a, then

$$\nu\left\{\underline{\alpha} : \left|\frac{\hat{\mu}_i^{(n)}(B^{(n)}(\underline{\alpha}))}{\hat{\nu}^{(n)}(B^{(n)}(\underline{\alpha}))} - \frac{d\mu_i}{d\nu}(\underline{\alpha})\right| > \varepsilon\right\} \xrightarrow{P} 0 \qquad (\text{as } n\to\infty) \tag{4.15}$$

for all $\varepsilon > 0$, $i = 1,\ldots,c$

Proof

See Gordon and Olshen [6].

Theorem 4.1 can be used to modify the multiclass partitioning algorithm of Section 2.3 to obtain measure-consistent density estimates by substituting (2.32) - (2.35) for (2.14) - (2.16) in the Gordon and Olshen 2-class algorithm and changing (4.12) in an obvious manner.

In Section 4.2 we review Gordon and Olshen's proof that measure-consistent density estimates yield asymptotically efficient decision rules for the 2-class case. No assumptions are made concerning the underlying cumulative distribution functions. The proof of this result, which is quite simple for the 2-class case, is surprisingly difficult for the c-class problem. Here, a simple graph-theoretic technique is used to simplify the problem.

4.2 Measure-Consistent Density Estimation and Asymptotic Efficiency

We want to show that measure-consistent density estimates yield asymptotically efficient decision rules, with no assumptions on the underlying cumulative distribution functions. We start with the 2-class case and follow Gordon and Olshen.

Let

$$I_{ij}(\underline{\alpha}) = I(\ell_i \pi_i \frac{d\mu_i}{d\nu}(\underline{\alpha}) > \ell_j \pi_j \frac{d\mu_j}{d\nu}(\underline{\alpha})) \tag{4.16}$$

$$I_{ij}^{(n)}(\underline{\alpha}) = I(\ell_i \pi_i \frac{\hat{d\mu}_i^{(n)}}{d\nu}(\underline{\alpha}) > \ell_j \pi_j \frac{\hat{d\mu}_j^{(n)}}{d\nu}(\underline{\alpha})) \tag{4.17}$$

For the 2-class case we have

$$\begin{aligned} R(\mathfrak{a}_B) &= \int [(1-I_{12})\ell_1\pi_1 \frac{d\mu_1}{d\nu} + I_{12}\ell_2\pi_2 \frac{d\mu_2}{d\nu}] \, d\nu \\ &= \ell_1\pi_1 - \int I_{12}(\ell_1\pi_1 \frac{d\mu_1}{d\nu} - \ell_2\pi_2 \frac{d\mu_2}{d\nu}) \, d\nu \end{aligned} \tag{4.18}$$

$$R(\mathfrak{a}^{(n)}) = \ell_1\pi_1 - \int I_{12}^{(n)}(\ell_1\pi_1 \frac{d\mu_1}{d\nu} - \ell_2\pi_2 \frac{d\mu_2}{d\nu}) \, d\nu \tag{4.19}$$

where the dependence on $\underline{\alpha}$ has been suppressed for notational convenience. Note that $R(\hat{a}^{(n)})$ is a random variable; by convention, the expectation has not been taken over the data sequence. We have the following theorem.

Theorem 4.2a

Let $\frac{d\hat{\mu}_1^{(n)}}{d\nu}(\underline{\alpha})$, $\frac{d\hat{\mu}_2^{(n)}}{d\nu}(\underline{\alpha})$ be measurable functions such that

$$\nu\left\{\underline{\alpha} : \left|\frac{d\hat{\mu}_i^{(n)}}{d\nu}(\underline{\alpha}) - \frac{d\mu_i}{d\nu}(\underline{\alpha})\right| > \varepsilon\right\} \xrightarrow{P} 0 \qquad (\text{as } n\to\infty) \tag{4.20}$$

for all $\varepsilon > 0$, $i = 1,2$. Then

$$R(\hat{a}^{(n)}) \xrightarrow{P} R(\hat{a}_B) \qquad (\text{as } n\to\infty) \tag{4.21}$$

Proof

For $\varepsilon > 0$ let

$$W_1 = \left\{\underline{\alpha} : \left|\ell_1\pi_1\frac{d\mu_1}{d\nu}(\underline{\alpha}) - \ell_2\pi_2\frac{d\mu_2}{d\nu}(\underline{\alpha})\right| \leqslant \varepsilon\right\} \tag{4.22}$$

$$W_2 = \left\{\underline{\alpha} : \left|\ell_1\pi_1\frac{d\mu_1}{d\nu}(\underline{\alpha}) - \ell_2\pi_2\frac{d\mu_2}{d\nu}(\underline{\alpha})\right| > \varepsilon\right\} = W_1^c \tag{4.23}$$

From the Radon-Nikodym theorem, $\frac{d\mu_1}{d\nu}$, $\frac{d\mu_2}{d\nu}$ are measurable functions. Using simple properties of measurable sets and functions we have W_1, W_2 are measurable and $I_{12}(\ell_1\pi_1\frac{d\mu_1}{d\nu} - \ell_2\pi_2\frac{d\mu_2}{d\nu})$, $I_{12}^{(n)}(\ell_1\pi_1\frac{d\mu_1}{d\nu} - \ell_2\pi_2\frac{d\mu_2}{d\nu})$ are measurable on W_1, W_2. Thus

$$R(\hat{a}_B) = \ell_1\pi_1 - \sum_{k=1}^{2}\int_{W_k} I_{12}(\ell_1\pi_1\frac{d\mu_1}{d\nu} - \ell_2\pi_2\frac{d\mu_2}{d\nu})\,d\nu \tag{4.24}$$

$$R(\underline{a}^{(n)}) = \ell_1\pi_1 - \sum_{k=1}^{2} \int_{W_k} I_{12}^{(n)}(\ell_1\pi_1\frac{d\mu_1}{d\nu} - \ell_2\pi_2\frac{d\mu_2}{d\nu})d\nu \tag{4.25}$$

We show

$$\int_{W_k} I_{12}^{(n)}(\ell_1\pi_1\frac{d\mu_1}{d\nu} - \ell_2\pi_2\frac{d\mu_2}{d\nu})d\nu \xrightarrow{P} \int_{W_k} I_{12}(\ell_1\pi_1\frac{d\mu_1}{d\nu} - \ell_2\pi_2\frac{d\mu_2}{d\nu})d\nu$$

$$(\text{as } n\to\infty),\ k = 1,2 \tag{4.26}$$

Consider W_1. We have

$$\left|\int_{W_1} I_{12}^{(n)}(\ell_1\pi_1\frac{d\mu_1}{d\nu} - \ell_2\pi_2\frac{d\mu_2}{d\nu})d\nu - \int_{W_1} I_{12}(\ell_1\pi_1\frac{d\mu_1}{d\nu} - \ell_2\pi_2\frac{d\mu_2}{d\nu})d\nu\right|$$

$$\leqslant \int_{W_1} \left|I_{12}^{(n)} - I_{12}\right|\left|\ell_1\pi_1\frac{d\mu_1}{d\nu} - \ell_2\pi_2\frac{d\mu_2}{d\nu}\right|d\nu$$

$$\leqslant \int_{W_1} 1\cdot\varepsilon d\nu \leqslant \varepsilon \tag{4.27}$$

Now consider W_2. First,

$$\left|I_{12}^{(n)}(\ell_1\pi_1\frac{d\mu_1}{d\nu} - \ell_2\pi_2\frac{d\mu_2}{d\nu})\right| \leqslant \ell_1\pi_1\frac{d\mu_1}{d\nu} + \ell_2\pi_2\frac{d\mu_2}{d\nu},$$

which is integrable over W_2. Second,

$$\nu\left\{\underline{\alpha}\in W_2 : \left|I_{12}^{(n)}(\ell_1\pi_1\frac{d\mu_1}{d\nu}(\underline{\alpha}) - \ell_2\pi_2\frac{d\mu_2}{d\nu}(\underline{\alpha})) - I_{12}(\ell_1\pi_1\frac{d\mu_1}{d\nu}(\underline{\alpha}) - \ell_2\pi_2\frac{d\mu_2}{d\nu}(\underline{\alpha}))\right| > \varepsilon'\right\} \xrightarrow{P} 0$$

$$(\text{as } n\to\infty) \tag{4.28}$$

for all $\varepsilon' > 0$. We see this as follows. For $\varepsilon' > 0$,

$$\left| I_{12}^{(n)}(\ell_1\pi_1\frac{d\mu_1}{d\nu}(\underline{\alpha}) - \ell_2\pi_2\frac{d\mu_2}{d\nu}(\underline{\alpha})) - I_{12}(\ell_1\pi_1\frac{d\mu_1}{d\nu}(\underline{\alpha}) - \ell_2\pi_2\frac{d\mu_2}{d\nu}(\underline{\alpha}))\right| = 0 < \varepsilon'$$

whenever

$$\left|\frac{\hat{d\mu}_k^{(n)}}{d\nu}(\underline{\alpha}) - \frac{d\mu_k}{d\nu}(\underline{\alpha})\right| < \frac{\varepsilon}{2\ell_k\pi_k}, \quad k = 1,2, \quad \text{for all } \underline{\alpha} \in W_2$$

Thus

$$\nu\left\{\underline{\alpha}\in W_2 : \left| I_{12}^{(n)}(\ell_1\pi_1\frac{d\mu_1}{d\nu}(\underline{\alpha}) - \ell_2\pi_2\frac{d\mu_2}{d\nu}(\underline{\alpha})) - I_{12}(\ell_1\pi_1\frac{d\mu_1}{d\nu}(\underline{\alpha}) - \ell_2\pi_2\frac{d\mu_2}{d\nu}(\underline{\alpha}))\right| \geqslant \varepsilon'\right\}$$

$$\leqslant \nu\left\{\underline{\alpha} : \left|\frac{d\mu_k^{(n)}}{d\nu}(\underline{\alpha}) - \frac{d\mu_k}{d\nu}(\underline{\alpha})\right| \geqslant \frac{\varepsilon}{2\ell_k\pi_k} \text{ for some } k = 1,2\right\}$$

$$\leqslant \sum_{k=1}^{2} \nu\left\{\underline{\alpha} : \left|\frac{d\mu_k^{(n)}}{d\nu}(\underline{\alpha}) - \frac{d\mu_k}{d\nu}(\underline{\alpha})\right| \geqslant \frac{\varepsilon}{2\ell_k\pi_k}\right\} \tag{4.29}$$

Taking the limit in probability (as $n \to \infty$) of both sides of the above inequality gives (4.28). Finally, apply the Lebesgue Dominated Convergence Theorem.

The proof of Theorem 4.2a is quite simple. However, the proof for $c > 2$ is considerably more complicated.

Let $\underline{x}_i \in \mathbb{R}^e$, $i = 1,\ldots,c$, and $\|\cdot\|$ be a norm on $\mathbb{R}^e$. Given $\varepsilon > 0$, we recursively partition $\{1,2,\ldots,c-1\}$ into disjoint sets $I_1,\ldots,I_q$ as follows:

$$I_m^{(1)} = \{i\} \qquad \text{for some } i = 1,\ldots,c-1;\ i \notin I_1,\ldots,I_{m-1}$$

$$I_m^{(k)} = \{i \neq c : \|\underline{x}_i - \underline{x}_j\| \leq \varepsilon \text{ for some } j \in I_m^{(k-1)}\}$$

$$I_m \triangleq I_m^{(c-1)}$$

Note that $I_m = I_m^{(\ell)}$ $\ell \geq c-1$.

Example 4.1

In Figure 4.1, $e = 2$, $c = 10$. Let $\|\underline{x}\| = (\underline{x} \cdot \underline{x})^{\frac{1}{2}}$, $\varepsilon = \sqrt{2}$. Then $I_1 = \{1,2,3\}$, $I_2 = \{4,5\}$, $I_3 = \{6,7,8,9\}$.

The proof of the following lemma uses some basic definitions and theorems from graph theory (cf. Harary [10]).

Lemma 4.1

Let $a_i \in \mathbb{R}$, $i = 1,\ldots,c$, such that $\sum_{i=1}^{c} a_i = 0$. Then there exists b_{ij}, $i = 1,\ldots,c$, $j = 1,\ldots,c$, such that

(1) $$\sum_{i=1}^{c} \sum_{j=1}^{c} b_{ij}(\underline{x}_i - \underline{x}_j) = \sum_{i=1}^{c} a_i \underline{x}_i \tag{4.30}$$

(2) for each $m = 1,\ldots,q$ and any $i \in I_m$

$$b_{ic} = \sum_{k \in I_m} a_k \tag{4.31}$$

$$b_{jc} = 0 \qquad j \in I_m;\ j \neq i \tag{4.32}$$

(3) $$b_{ci} = 0 \qquad i = 1,\ldots,c \tag{4.33}$$

(4) $$b_{ij} = 0 \qquad \|\underline{x}_i - \underline{x}_j\| > \varepsilon;\ j \neq c \tag{4.34}$$

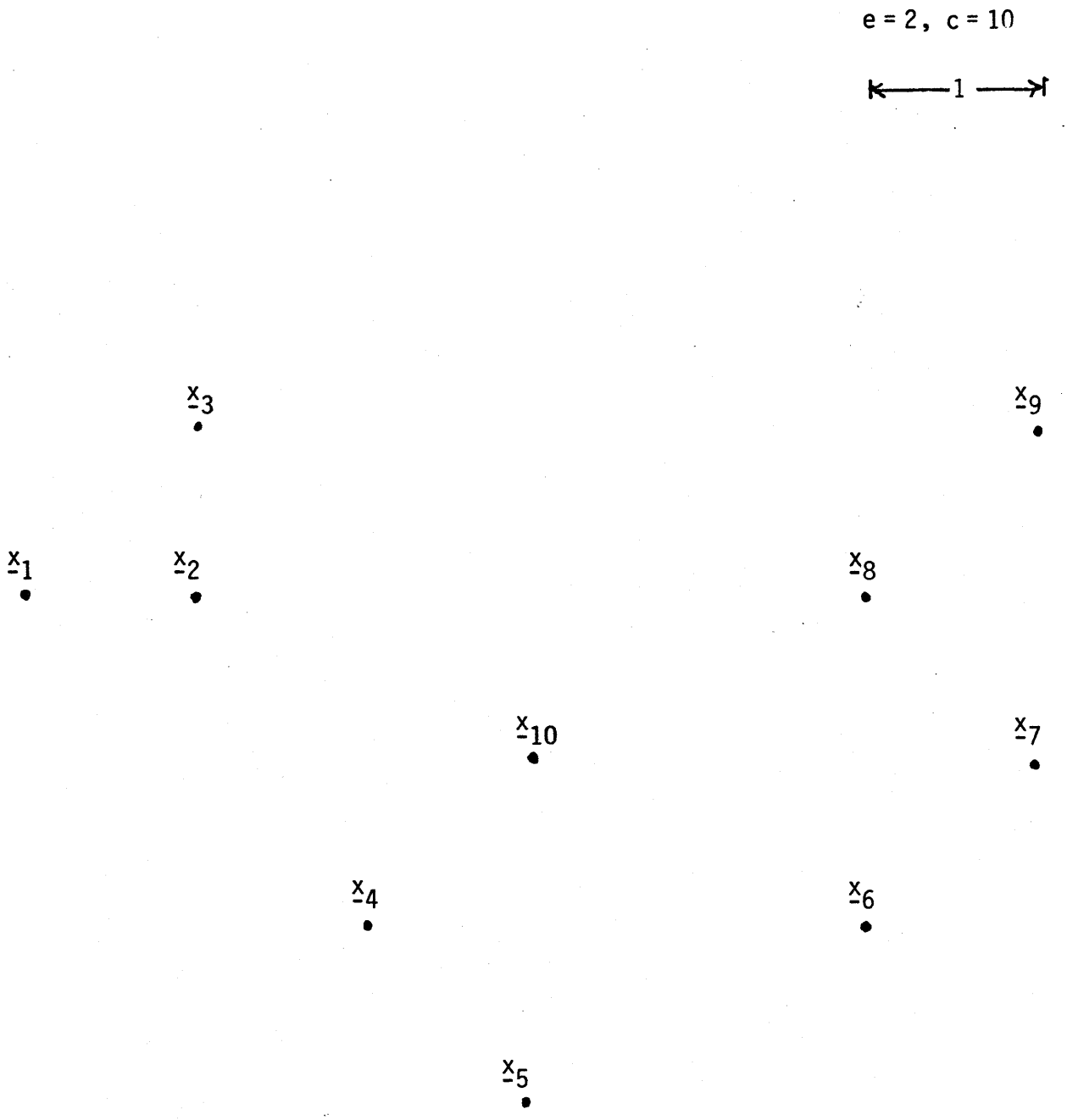


Figure 4.1 Point Distribution in $\mathbb{R}^2$

(5) for each $m = 1,\ldots,q$, $i \in I_m$, $j \in I_m$

$$b_{ij} = \sum_{k \in I_m(i,j) \subset I_m} a_k \tag{4.35}$$

Proof

We first associate each $\underline{x}_i \in \mathbb{R}^e$ with a node labelled i. A directed spanning tree $\mathcal{F}'$ is constructed as follows (see Figure 4.2):

(a) Construct graphs $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_q$ corresponding to $I_1, I_2, \ldots, I_q$ by equating the indices in I_m with nodes in $\mathcal{G}_m$. If $i \in I_m$, $j \in I_m$ and $\|\underline{x}_i - \underline{x}_j\| \leq \varepsilon$, include edge (i,j) in $\mathcal{G}_m$. Note that $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_q$ are connected graphs.

(b) Construct graph $\mathcal{G}$ by adding node c to $\bigcup_{m=1}^{q} \mathcal{G}_m$. For each $m = 1,\ldots,q$, choose $i \in I_m$ and add edge (c,i) to $\mathcal{G}$. Note that $\mathcal{G}$ is a connected graph.

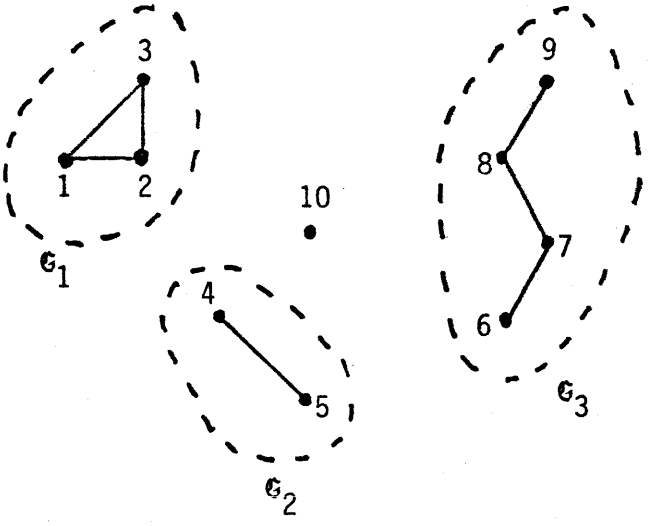
(c) Since $\mathcal{G}$ is a connected graph, it must have a subgraph which is a spanning tree. Let $\mathcal{F}$ be a spanning tree in $\mathcal{G}$. Construct $\mathcal{F}'$, a directed spanning tree, by directing all of the edges in $\mathcal{F}$ away from node c.

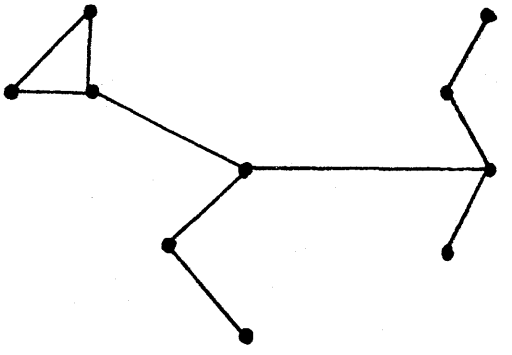
Let $P_k = (c = t_1, t_2, \ldots, t_n = k)$ be the path from node c to node k in $\mathcal{F}'$. Since $\mathcal{F}'$ is directed, (t_k, t_{k+1}) is a directed edge along P_k (denoted $(t_k, t_{k+1}) \in P_k$), but (t_{k+1}, t_k) is not. Thus

$$\sum_{(j,i) \in P_k} (\underline{x}_i - \underline{x}_j) = \underline{x}_k - \underline{x}_c \tag{4.36}$$

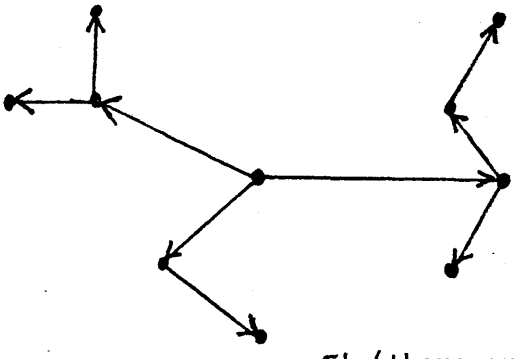
which implies

$$\sum_{k=1}^{c} a_k \sum_{(j,i) \in P_k} (\underline{x}_i - \underline{x}_j) = \sum_{k=1}^{c} a_k (\underline{x}_k - \underline{x}_c) = \sum_{k=1}^{c} a_k \underline{x}_k \tag{4.37}$$





$\mathcal{G}$ (there are other possibilities)



$\mathcal{F}'$ (there are other possibilities)

Figure 4.2 Construction of Directed Spanning Tree $\mathcal{F}'$

since $\sum_{k=1}^{c} a_k = 0$ by assumption. Interchanging the order of summation we have:

$$\sum_{i=1}^{c} \sum_{j=1}^{c} \left(\sum_{k:(j,i)\in P_k} a_k \right)(\underline{x}_i - \underline{x}_j) = \sum_{k=1}^{c} a_k \underline{x}_k \tag{4.38}$$

Thus, if

$$b_{ij} = \sum_{k:(j,i)\in P_k} a_k \tag{4.39}$$

then

$$\sum_{i=1}^{c} \sum_{j=1}^{c} b_{ij}(\underline{x}_i - \underline{x}_j) = \sum_{i=1}^{c} a_k \underline{x}_k \tag{4.40}$$

which verifies (1). For each $m = 1,\ldots,q$, the $i \in I_m$ which was (arbitrarily) chosen in the construction of $\mathcal{G}$ will satisfy (2). (3) and (4) are obvious. For each $m = 1,\ldots,q$, $i \in I_m$, $j \in I_m$, $I_m(i,j) = \{k : (j,i) \in P_k\}$ will satisfy (5).

Example 4.2

For $\mathcal{F}'$ in Figure 4.2 we have

$$b_{2,10} = a_1 + a_2 + a_3$$

$$b_{4,10} = a_4 + a_5$$

$$b_{7,10} = a_6 + a_7 + a_8 + a_9$$

$$b_{12} = a_1$$

$$b_{32} = a_3$$

$$b_{54} = a_5$$

$$b_{67} = a_6$$

$$b_{87} = a_8 + a_9$$

$$b_{98} = a_9$$

$$b_{ij} = 0 \qquad \text{other}$$

For the c-class case we have

$$R(\mathfrak{d}_B) = \int[(1 - I_{12}I_{13}\cdots I_{1c})\ell_1\pi_1\frac{d\mu_1}{d\nu}$$

$$+ (1-(1-I_{12})I_{23}\cdots I_{2c})\ell_2\pi_2\frac{d\mu_2}{d\nu}$$

$$\vdots$$

$$+ (1-(1-I_{1c})(1-I_{2c})\cdots(1-I_{c-1,c}))\ell_c\pi_c\frac{d\mu_c}{d\nu}]\ d\nu$$

$$= \sum_{i=1}^{c-1} \ell_i\pi_i - \int[I_{12}I_{13}\cdots I_{1c}\ell_1\pi_1\frac{d\mu_1}{d\nu}$$

$$+ (1-I_{12})I_{23}\cdots I_{2c}\ell_2\pi_2\frac{d\mu_2}{d\nu}$$

$$\vdots$$

$$+ (1-I_{1,c-1})(1-I_{2,c-1})\cdots(1-I_{c-2,c-1})I_{c-1,c}\ell_{c-1}\pi_{c-1}\frac{d\mu_{c-1}}{d\nu}$$

$$- (1-(1-I_{1c})(1-I_{2c})\cdots(1-I_{c-1,c}))\ell_c\pi_c\frac{d\mu_c}{d\nu}]\ d\nu \qquad (4.41)$$

$$R(\mathfrak{d}^{(n)}) = \sum_{i=1}^{c-1} \ell_i\pi_i - \int[I_{12}^{(n)}I_{13}^{(n)}\cdots I_{1c}^{(n)}\ell_1\pi_1\frac{d\mu_1}{d\nu}$$

$$+ (1-I_{12}^{(n)})I_{23}^{(n)}\cdots I_{2c}^{(n)}\ell_2\pi_2\frac{d\mu_2}{d\nu}$$

$$\vdots$$

$$+ (1-I^{(n)}_{1,c-1})(1-I^{(n)}_{2,c-1})\cdots(1-I^{(n)}_{c-2,c-1})I^{(n)}_{c-1,c}\ell_{c-1}\pi_{c-1}\frac{d\mu_{c-1}}{d\nu}$$

$$- (1-(1-I^{(n)}_{1c})(1-I^{(n)}_{2c})\cdots(1-I^{(n)}_{c-1,c}))\ell_c\pi_c\frac{d\mu_c}{d\nu}]\ d\nu \qquad (4.42)$$

Let

$$a_1 = I_{12}I_{13}\cdots I_{1c} \qquad (4.43.1)$$

$$a_2 = (1-I_{12})I_{23}\cdots I_{2c} \qquad (4.43.2)$$

$$\vdots \qquad\qquad \vdots$$

$$a_{c-1} = (1-I_{1,c-1})(1-I_{2,c-1})\cdots(1-I_{c-2,c-1})I_{c-1,c} \qquad (4.43.c-1)$$

$$a_c = -(1-(1-I_{1c})(1-I_{2c})\cdots(1-I_{c-1,c})) \qquad (4.43.c)$$

and

$$a^{(n)}_1 = I^{(n)}_{12}I^{(n)}_{13}\cdots I^{(n)}_{1c} \qquad (4.44.1)$$

$$a^{(n)}_2 = (1-I^{(n)}_{12})I^{(n)}_{23}\cdots I^{(n)}_{2c} \qquad (4.44.2)$$

$$\vdots$$

$$a^{(n)}_{c-1} = (1-I^{(n)}_{1,c-1})(1-I^{(n)}_{2,c-1})\cdots(1-I^{(n)}_{c-2,c-1})I^{(n)}_{c-1,c} \qquad (4.44.c-1)$$

$$a^{(n)}_c = -(1-(1-I^{(n)}_{1c})(1-I^{(n)}_{2c})\cdots(1-I^{(n)}_{c-1,c})) \qquad (4.44.c)$$

Then

$$R(\underline{a}_B) = \sum_{i=1}^{c-1} \ell_i\pi_i - \int \sum_{i=1}^{c} a_i\ell_i\pi_i\frac{d\mu_i}{d\nu}\ d\nu \qquad (4.45)$$

$$R(\underline{a}^{(n)}) = \sum_{i=1}^{c-1} \ell_i\pi_i - \int \sum_{i=1}^{c} a^{(n)}_i\ell_i\pi_i\frac{d\mu_i}{d\nu}\ d\nu \qquad (4.46)$$

Note that $R(\hat{a}^{(n)})$ is a random variable; by convention the expectation has not been taken over the data sequence. Now $\sum_{i=1}^{c} a_i = \sum_{i=1}^{c} a_i^{(n)} = 0$ for all $\underline{\alpha} \in \mathbb{R}^d$. Thus (1) of Lemma 4.1 with $e = 1$ and $\underline{x}_i = x_i = \ell_i \pi_i \frac{d\mu_i}{d\nu}$, $i = 1,\ldots,c$, gives

$$R(\hat{a}_B) = \sum_{i=1}^{c-1} \ell_i \pi_i - \int \sum_{i=1}^{c} \sum_{j=1}^{c} b_{ij} \left(\ell_i \pi_i \frac{d\mu_i}{d\nu} - \ell_j \pi_j \frac{d\mu_j}{d\nu}\right) d\nu \qquad (4.47)$$

$$R(\hat{a}^{(n)}) = \sum_{i=1}^{c-1} \ell_i \pi_i - \int \sum_{i=1}^{c} \sum_{j=1}^{c} b_{ij}^{(n)} \left(\ell_i \pi_i \frac{d\mu_i}{d\nu} - \ell_j \pi_j \frac{d\mu_j}{d\nu}\right) d\nu \qquad (4.48)$$

We now state and prove the multiclass extension of Theorem 4.2a.

Theorem 4.2

Let $\frac{\hat{d\mu}_1^{(n)}}{d\nu}(\underline{\alpha})$, $\frac{\hat{d\mu}_2^{(n)}}{d\nu}(\underline{\alpha})$, ..., $\frac{\hat{d\mu}_c^{(n)}}{d\nu}(\underline{\alpha})$ be measurable functions such that

$$\nu\left\{\underline{\alpha} : \left|\frac{\hat{d\mu}_i^{(n)}}{d\nu}(\underline{\alpha}) - \frac{d\mu_i}{d\nu}(\underline{\alpha})\right| > \varepsilon\right\} \xrightarrow{P} 0 \qquad (\text{as } n \to \infty) \qquad (4.49)$$

for all $\varepsilon > 0$, $i = 1,\ldots,c$. Then

$$R(\hat{a}^{(n)}) \xrightarrow{P} R(\hat{a}_B) \qquad (\text{as } n \to \infty) \qquad (4.50)$$

Proof

For $\varepsilon > 0$ let

$$W_{ij} = \left\{\underline{\alpha} : \left|\ell_i \pi_i \frac{d\mu_i}{d\nu}(\underline{\alpha}) - \ell_j \pi_j \frac{d\mu_j}{d\nu}(\underline{\alpha})\right| \leq \varepsilon\right\}, \; i = 1,\ldots,c, \; j = 1,\ldots,c, \qquad (4.51)$$

$$\mathbb{W} = \{W_{12}, W_{12}^c\} \times \{W_{13}, W_{13}^c\} \times \ldots \times \{W_{1c}, W_{1c}^c\} \times \{W_{23}, W_{23}^c\} \times \{W_{24}, W_{24}^c\} \times \ldots \times \{W_{2c}, W_{2c}^c\} \times \ldots$$

$$\times \{W_{c-1,c}, W_{c-1,c}^c\}$$

$$= \{\underline{W}_1, \underline{W}_2, \ldots, \underline{W}_r\}, \tag{4.52}$$

$$\underline{W}_k = (W_k^{12}, W_k^{13}, \ldots, W_k^{1c}, W_k^{23}, W_k^{24}, \ldots, W_k^{2c}, \ldots, W_k^{c-1,c}), \quad k = 1, \ldots, r, \tag{4.53}$$

and

$$W_k = \bigcap_{i=1}^{c} \bigcap_{j=1}^{c} W_k^{ij}, \qquad k = 1, \ldots, r \tag{4.54}$$

It is clear that $W_1, W_2, \ldots, W_r$ are disjoint and $\bigcup_{k=1}^{r} W_k = \mathbb{R}^d$. As in the 2-class case, $W_1, W_2, \ldots, W_r$ are measurable and $b_{ij}(\ell_i \pi_i \frac{d\mu_i}{d\nu} - \ell_j \pi_j \frac{d\mu_j}{d\nu})$, $b_{ij}^{(n)}(\ell_i \pi_i \frac{d\mu_i}{d\nu} - \ell_j \pi_j \frac{d\mu_j}{d\nu})$ are measurable on W_k, $i = 1, \ldots, c$, $j = 1, \ldots, c$, $k = 1, \ldots, r$. Thus

$$R(\mathfrak{a}_B) = \sum_{i=1}^{c-1} \ell_i \pi_i - \sum_{k=1}^{r} \int_{W_k} \sum_{i=1}^{c} \sum_{j=1}^{c} b_{ij}(\ell_i \pi_i \frac{d\mu_i}{d\nu} - \ell_j \pi_j \frac{d\mu_j}{d\nu})\, d\nu$$

$$= \sum_{i=1}^{c-1} \ell_i \pi_i - \sum_{k=1}^{r} \sum_{i=1}^{c} \sum_{j=1}^{c} \int_{W_k} b_{ij}(\ell_i \pi_i \frac{d\mu_i}{d\nu} - \ell_j \pi_j \frac{d\mu_j}{d\nu})\, d\nu \tag{4.55}$$

$$R(\mathfrak{a}^{(n)}) = \sum_{i=1}^{c-1} \ell_i \pi_i - \sum_{k=1}^{r} \sum_{i=1}^{c} \sum_{j=1}^{c} \int_{W_k} b_{ij}^{(n)}(\ell_i \pi_i \frac{d\mu_i}{d\nu} - \ell_j \pi_j \frac{d\mu_j}{d\nu})\, d\nu \tag{4.56}$$

We show

$$\int_{W_k} b_{ij}^{(n)}(\ell_i\pi_i\frac{d\mu_i}{d\nu} - \ell_j\pi_j\frac{d\mu_j}{d\nu})d\nu \xrightarrow{P} \int_{W_k} b_{ij}(\ell_i\pi_i\frac{d\mu_i}{d\nu} - \ell_j\pi_j\frac{d\mu_j}{d\nu})d\nu$$

$$(\text{as } n\to\infty), \quad i=1,\ldots,c,\ j=1,\ldots,c,\ k=1,\ldots,r \qquad (4.57)$$

If we choose $\|\cdot\| = |\cdot|$, and the same ε to generate $I_1, I_2, \ldots, I_q$ and $W_1, W_2, \ldots, W_r$, then $[I_m]$ is <u>fixed</u> on W_k. We apply (2) of Lemma 4.1 to each W_k as follows. For each $m = 1,\ldots,q$, if $\left|\frac{d\mu_i}{d\nu} - \frac{d\mu_c}{d\nu}\right| \leq \varepsilon$ for some $i \in I_m$, choose $i_0 = i$; otherwise for any $i \in I_m$, choose $i_0 = i$. Then take $b_{i_0 c} = \sum_{k\in I_m} a_k$. Note that $i_0 = i_0(k,m)$.

Consider the $i,j^{\underline{th}}$ integral on W_k. Suppose $\left|\ell_i\pi_i\frac{d\mu_i}{d\nu} - \ell_j\pi_j\frac{d\mu_j}{d\nu}\right| \leq \varepsilon$ on W_k. We have the following cases:

$$i \in I_m,\ j = c,\ i = i_0(k,m) \Rightarrow b_{ij} = \sum_{k\in I_m} a_k,\ b_{ij}^{(n)} = \sum_{k\in I_m} a_k^{(n)} \qquad \text{(Lemma 4.1(2))}$$

$$i \in I_m,\ j = c,\ i \neq i_0(k,m) \Rightarrow b_{ij} = 0,\ b_{ij}^{(n)} = 0 \qquad \text{(Lemma 4.1(2))}$$

$$i = c \Rightarrow b_{ij} = 0,\ b_{ij}^{(n)} = 0 \qquad \text{(Lemma 4.1(3))}$$

$$i \in I_m,\ j \in I_m \Rightarrow b_{ij} = \sum_{k\in I_m(i,j)\subset I_m} a_k,\ b_{ij}^{(n)} = \sum_{k\in I_m(i,j)\subset I_m} a_k^{(n)} \qquad \text{(Lemma 4.1(5))}$$

In any case

$$\left|\int_{W_k} b_{ij}^{(n)}(\ell_i\pi_i\frac{d\mu_i}{d\nu} - \ell_j\pi_j\frac{d\mu_j}{d\nu})d\nu - \int_{W_k} b_{ij}(\ell_i\pi_i\frac{d\mu_i}{d\nu} - \ell_j\pi_j\frac{d\mu_j}{d\nu})d\nu\right|$$

$$\leq \int_{W_k} |b_{ij}^{(n)} - b_{ij}|\left|\ell_i\pi_i\frac{d\mu_i}{d\nu} - \ell_j\pi_j\frac{d\mu_j}{d\nu}\right|d\nu$$

$$\leq \int_{W_k} 1 \cdot \varepsilon d\nu \leq \varepsilon \tag{4.58}$$

since for any $I \neq c$ $\sum_{k \in I} a_k = 0$ or 1, $\sum_{k \in I} a_k^{(n)} = 0$ or 1. Now suppose $\left| \ell_i \pi_i \frac{d\mu_i}{d\nu} - \ell_j \pi_j \frac{d\mu_j}{d\nu} \right| > \varepsilon$ on W_k. We have the following cases:

$$i \in I_m,\ j = c,\ i = i_0(k,m) \Rightarrow b_{ij} = \sum_{k \in I_m} a_k,\ b_{ij}^{(n)} = \sum_{k \in I_m} a_k^{(n)} \qquad \text{(Lemma 4.1(2))}$$

$$i \in I_m,\ j = c,\ i \neq i_0(k,m) \Rightarrow b_{ij} = 0,\ b_{ij}^{(n)} = 0 \qquad \text{(Lemma 4.1(2))}$$

$$i = c \Rightarrow b_{ij} = 0,\ b_{ij}^{(n)} = 0 \qquad \text{(Lemma 4.1(3))}$$

$$j \neq c \Rightarrow b_{ij} = 0,\ b_{ij}^{(n)} = 0 \qquad \text{(Lemma 4.1(4))}$$

We need only show (!)

$$\int_{W_k} b_{i_0 c}^{(n)} (\ell_{i_0} \pi_{i_0} \frac{d\mu_{i_0}}{d\nu} - \ell_c \pi_c \frac{d\mu_c}{d\nu}) d\nu \xrightarrow{P} \int_{W_k} b_{i_0 c} (\ell_{i_0} \pi_{i_0} \frac{d\mu_{i_0}}{d\nu} - \ell_c \pi_c \frac{d\mu_c}{d\nu}) d\nu \quad (\text{as } n \to \infty) \tag{4.59}$$

where the dependence of i_0 on k,m has been suppressed. First,

$$\left| b_{i_0 c}^{(n)} (\ell_{i_0} \pi_{i_0} \frac{d\mu_{i_0}}{d\nu} - \ell_c \pi_c \frac{d\mu_c}{d\nu}) \right| \leq \ell_{i_0} \pi_{i_0} \frac{d\mu_{i_0}}{d\nu} + \ell_c \pi_c \frac{d\mu_c}{d\nu}$$

which is integrable over W_k. Second,

$$\nu\left\{\underline{\alpha} \in W_k : \left| b_{i_0 c}^{(n)} (\ell_{i_0} \pi_{i_0} \frac{d\mu_{i_0}}{d\nu}(\underline{\alpha}) - \ell_c \pi_c \frac{d\mu_c}{d\nu}(\underline{\alpha})) - b_{i_0 c} (\ell_{i_0} \pi_{i_0} \frac{d\mu_{i_0}}{d\nu}(\underline{\alpha}) - \ell_c \pi_c \frac{d\mu_c}{d\nu}(\underline{\alpha})) \right| > \varepsilon' \right\} \xrightarrow{P} 0$$

$$(\text{as } n \to \infty) \tag{4.60}$$

for all $\varepsilon' > 0$. We see this as follows. We have

$$b_{i_0 c} = \sum_{k \in I_m} a_k = -\sum_{k \in I_m^c} a_k - a_c \qquad (I_m^c = \{1,\ldots,c-1\} - I_m) \tag{4.61}$$

Since $-\sum_{k \in I_m^c} a_k - a_c$ does not depend on I_{ij} for $i \in I_m$, $j \in I_m$, it follows that $\sum_{k \in I_m} a_k$ depends on I_{ij} only for $i \in I_m$, $j \in I_n$, $n \neq m$, and $i \in I_m$, $j = c$.

From the choice of $i_0(k,m)$, $\left|\ell_{i_0}\pi_{i_0}\frac{d\mu_{i_0}}{d\nu} - \ell_c\pi_c\frac{d\mu_c}{d\nu}\right| > \varepsilon$ implies $\left|\ell_i\pi_i\frac{d\mu_i}{d\nu} - \ell_c\pi_c\frac{d\mu_c}{d\nu}\right| > \varepsilon$ for all $i \in I_m$. Thus $b_{i_0 c}$ depends on I_{ij} only for $\left|\ell_i\pi_i\frac{d\mu_i}{d\nu} - \ell_j\pi_j\frac{d\mu_j}{d\nu}\right| > \varepsilon$. Similarly, $b^{(n)}_{i_0 c} = \sum_{k \in I_m} a_k^{(n)}$ depends on $I_{ij}^{(n)}$ only for $\left|\ell_i\pi_i\frac{d\mu_i}{d\nu} - \ell_j\pi_j\frac{d\mu_j}{d\nu}\right| > \varepsilon$. Reasoning as in the proof of Theorem 4.2a, (4.60) must be true. Finally, apply the Lebesgue Dominated Convergence Theorem. |

The interested reader might work through the 3-class case in detail, which begins to reveal the structure of the problem. The 2-class case lacks this structure almost entirely. We remark that there is a much simpler proof of Theorem 4.2, which generalizes directly from the 2-class case, if it is assumed that $\frac{d\mu_1}{d\nu}, \frac{d\mu_2}{d\nu}, \ldots, \frac{d\mu_c}{d\nu}$ are ν-almost-everywhere continuous.

4.3 Asymptotic Efficiency for Multiclass Partitioning and Termination Algorithms

We now apply the results of Sections 4.1 and 4.2 to prove asymptotic

efficiency for decision rules generated by the multiclass partitioning and termination algorithms of Chapters 2 and 3. Since all the decision rules to be discussed can be realized as binary decision trees, we refer to decision trees rather than rules.

Let $T_{GO}^{(n)}$ be the binary decision tree generated by applying the multiclass partitioning algorithm of Section 2.3 to the data sequence $A^{(n)}$, modified as in Section 4.1.

Corollary 4.1

$$R(T_{GO}^{(n)}) \xrightarrow{P} R(\mathring{a}_B) \qquad (\text{as } n \to \infty) \tag{4.62}$$

Proof

Let $\mathring{a}^{(n)} = \mathring{a}_{GO}^{(n)} = T_{GO}^{(n)}$, $\frac{d\hat{\mu}_i^{(n)}}{d\nu}(\underline{\alpha}) = \frac{\hat{\mu}_i^{(n)}(B^{(n)}(\underline{\alpha}))}{\hat{\nu}^{(n)}(B^{(n)}(\underline{\alpha}))}$ $i = 1,\ldots,c$, and combine Theorems 4.1 and 4.2. |

Now let $T_0^{(n_1)}$ be the binary decision tree generated by applying the multiclass partitioning algorithm of Section 2.3 to the training sequence $A^{(n_1)}$, modified as in Section 4.1 but with only (4.12) as the termination criteria. Thus the terminal nodes of $T_0^{(n_1)}$ contain vectors only from a single class and furthermore $T_{GO}^{(n_1)} \subset T_0^{(n_1)}$. As in Chapter 3, let $T_*^{(n_1)}$ be the tree generated by applying the termination algorithm to $T_0^{(n_1)}$ based on the class distributions; $T_*^{(n_1,n_2)}$ the tree generated by applying the termination algorithm to $T_0^{(n_1)}$ based on the test sequence $A^{(n_2)}$. From Theorem 3.1

$$R(T_*^{(n_1)}) = \min_{T \subset T_0^{(n_1)}} R(T) \tag{4.63}$$

$$\hat{R}^{(n_2)}(T_*^{(n_1,n_2)}) = \min_{T \subset T_0^{(n_1)}} \hat{R}^{(n_2)}(T) \tag{4.64}$$

Lemma 4.2

Let $[x_1^{(n)}]$, $[x_2^{(n)}]$, ..., $[x_s^{(n)}]$ be sequences of random variables. If $x_m^{(n)}$ is bounded for all n, m = 1,...,s, and

$$x_m^{(n)} \xrightarrow{P} c_m, \text{ constant} \quad (\text{as } n \to \infty),\ m = 1,\ldots,s \tag{4.65}$$

then

$$\min_{m=1,\ldots,s} x_m^{(n)} \xrightarrow{P} \min_{m=1,\ldots,s} c_m \quad (\text{as } n \to \infty) \tag{4.66}$$

Proof

Since $x_m^{(n)}$ is bounded for all n, m = 1,...,s, $\min_{m=1,\ldots,s} x_m^{(n)}$ exists for all n. Since $x_m^{(n)} \xrightarrow{P} c_m$, m = 1,...,s, c_m is bounded, m = 1,...,s, and $\min_{m=1,\ldots,s} c_m$ exists. Let $\underline{x} = (x_1,x_2,\ldots,x_s)$, $\underline{x}^{(n)} = (x_1^{(n)},x_2^{(n)},\ldots,x_s^{(n)})$, $\underline{c} = (c_1,c_2,\ldots,c_s)$, and $f(\underline{x}) = \min_{m=1,\ldots,s} x_m$. We use the following result: if $\underline{x}^{(n)} \xrightarrow{P} \underline{c}$ and $f(\underline{x})$ is continuous then $f(\underline{x}^{(n)}) \xrightarrow{P} f(\underline{c})$. $x_m^{(n)} \xrightarrow{P} c_m$, m = 1,...,s, implies $\underline{x}^{(n)} \xrightarrow{P} \underline{c}$. We now show $f(\underline{x}) = \min_{m=1,\ldots,s} x_m$ is continuous. Let $|\underline{x} - \underline{y}| < \delta = \varepsilon$. Then $|x_m - y_m| < \varepsilon$, m = 1,...,s. Suppose $x_i = \min_{m=1,\ldots,s} x_m$ and $y_j = \min_{m=1,\ldots,s} y_m$. If i = j then

$$\left|\min_{m=1,\ldots,s} x_m - \min_{m=1,\ldots,s} y_m\right| = |x_i - y_i| = |x_j - y_j| < \varepsilon$$

If $i \neq j$ proceed as follows. If $y_j \leq x_i - \varepsilon$ then $|x_j - y_j| < \varepsilon$ implies $x_j < x_i$, a contradiction. If $y_j \geq x_i + \varepsilon$ then $|x_i - y_i| < \varepsilon$ implies $y_i < y_j$, a

contradiction. Thus

$$\left| \min_{m=1,\ldots,s} x_m - \min_{m=1,\ldots,s} y_m \right| = |x_i - y_j| < \varepsilon$$

which completes the proof. |

We have another corollary to Theorem 4.2.

Corollary 4.2

If $A^{(n_1)}$, $A^{(n_2)}$ are independent and $n_2(i) \to \infty$ as $n_2 \to \infty$, $i = 1,\ldots,c$, then

$$\hat{R}^{(n_2)}(T_*^{(n_1,n_2)}) \xrightarrow{P} R(T_*^{(n_1,n_2)}) \quad (\text{as } n_2 \to \infty) \tag{4.67}$$

$$\hat{R}^{(n_2)}(T_*^{(n_1,n_2)}) \xrightarrow{P} R(T_*^{(n_1)}) \quad (\text{as } n_2 \to \infty) \tag{4.68}$$

$$R(T_*^{(n_1,n_2)}) \xrightarrow{P} R(T_*^{(n_1)}) \quad (\text{as } n_2 \to \infty) \tag{4.69}$$

$$R(T_*^{(n_1)}) \xrightarrow{P} R(\grave{a}_B) \quad (\text{as } n_1 \to \infty) \tag{4.70}$$

Proof

Since $T_0^{(n_1)}$ is a finite binary decision tree, $\{T \subset T_0^{(n_1)}\}$ is finite. Let $\{T \subset T_0^{(n_1)}\} = \{T_1, T_2, \ldots, T_s\}$. Since $A^{(n_1)}$, $A^{(n_2)}$ are independent the Weak Law of Large Numbers gives

$$\hat{R}^{(n_2)}(T_m) \xrightarrow{P} R(T_m) \quad (\text{as } n_2(i) \to \infty \;\; i = 1,\ldots,c) \; m = 1,\ldots,s \tag{4.71}$$

Thus

$$\hat{R}^{(n_2)}(T_m) \xrightarrow{P} R(T_m) \quad (\text{as } n_2 \to \infty) \;\; m = 1,\ldots,s \tag{4.72}$$

since $n_2(i) \to \infty$ as $n_2 \to \infty$, $i = 1,\ldots,c$, by assumption.

(1) For $\varepsilon > 0$, $|\hat{R}^{(n_2)}(T_*^{(n_1,n_2)}) - R(T_*^{(n_1,n_2)})| < \varepsilon$ whenever $|\hat{R}^{(n_2)}(T_m) - R(T_m)| < \varepsilon$, $m = 1,\ldots,s$, since $T_*^{(n_1,n_2)} \subset T_0^{(n_1)}$. Thus

$$\Pr\{|\hat{R}^{(n_2)}(T_*^{(n_1,n_2)}) - R(T_*^{(n_1,n_2)})| \geqslant \varepsilon\}$$

$$\leqslant \Pr\{|\hat{R}^{(n_2)}(T_m) - R(T_m)| \geqslant \varepsilon \text{ for some } m = 1,\ldots,s\}$$

$$\leqslant \sum_{m=1}^{s} \Pr\{|\hat{R}^{(n_2)}(T_m) - R(T_m)| \geqslant \varepsilon\} \tag{4.73}$$

Taking the limit (as $n_2 \to \infty$) of both sides of the above inequality and using (4.72) gives the desired result.

(2) Let $x_m^{(n_2)} = \hat{R}^{(n_2)}(T_m)$ and $c_m = R(T_m)$, $m = 1,\ldots,s$. Since $\hat{R}^{(n_2)}(T_m)$ is bounded for all n_2, $m = 1,\ldots,s$, (4.72) implies the conditions of Lemma 4.2 are satisfied. Thus

$$\min_{m=1,\ldots,s} \hat{R}^{(n_2)}(T_m) \xrightarrow{P} \min_{m=1,\ldots,s} R(T_m) \quad (\text{as } n_2 \to \infty) \tag{4.74}$$

From Theorem 3.1 it follows that

$$\hat{R}^{(n_2)}(T_*^{(n_1,n_2)}) \xrightarrow{P} R(T_*^{(n_1)}) \quad (\text{as } n_2 \to \infty) \tag{4.75}$$

(3) If $x^{(n)} \xrightarrow{P} y^{(n)}$ and $x^{(n)} \xrightarrow{P} z^{(n)}$ then $y^{(n)} \xrightarrow{P} z^{(n)}$. Thus (3) follows from (1), (2).

(4) Since $T_{GO}^{(n_1)} \subset T_0^{(n_1)}$, Theorem 3.1 gives $R(T_*^{(n_1)}) \leqslant R(T_{GO}^{(n_1)})$. Thus, for $\varepsilon > 0$, $|R(T_*^{(n_1)}) - R(\hat{a}_B)| = R(T_*^{(n_1)}) - R(\hat{a}_B) < \varepsilon$ whenever

$|R(T_{GO}^{(n_1)}) - R(\mathfrak{d}_B)| = R(T_{GO}^{(n_1)}) - R(\mathfrak{d}_B) < \varepsilon$. Thus

$$\Pr\{|R(T_*^{(n_1)}) - R(\mathfrak{d}_B)| \geq \varepsilon\} \leq \Pr\{|R(T_{GO}^{(n_1)}) - R(\mathfrak{d}_B)| \geq \varepsilon\} \tag{4.76}$$

Taking the limit (as $n_1 \to \infty$) of both sides of the above inequality and using Corollary 4.1 gives the desired result.|

Corollary 4.2(2) shows that $\hat{R}^{(n_2)}(T_*^{(n_1,n_2)})$ is a consistent estimate of $R(T_*^{(n_1)})$. (2) and (4) taken together show that $\hat{R}^{(n_2)}(T_*^{(n_1,n_2)})$ is also a consistent estimate of $R(\mathfrak{d}_B)$. These results are true even though the same test sequence $A^{(n_2)}$ is used for estimation and termination. However, the following example shows that if $A^{(n_1)}$, $A^{(n_2)}$ are not independent, and the iterated limit is not respected, then $\hat{R}^{(n_2)}(T_*^{(n_1,n_2)})$ need not be a consistent estimate of $R(\mathfrak{d}_B)$.

Example 4.3

Let $A^{(n_1)} = A^{(n_2)} = A^{(n)}$. Then $\hat{R}^{(n_2)}(T_0^{(n_1)}) = \hat{R}^{(n)}(T_0^{(n)}) = 0$ which implies $\hat{R}^{(n_2)}(T_*^{(n_1,n_2)}) = \hat{R}^{(n)}(T_*^{(n,n)}) = 0$, for all n. If class distributions overlap then $R(\mathfrak{d}_B) > 0$ and $\hat{R}^{(n)}(T_*^{(n,n)})$ cannot be a consistent estimate of $R(\mathfrak{d}_B)$.|

From Corollary 4.1, $T_{GO}^{(n)}$ is an asymptotically efficient decision rule. Since $R(T_{GO}^{(n)})$ is bounded for all n we have

$$E_n(R(T_{GO}^{(n)})) \to R(\mathfrak{d}_B) \qquad (\text{as } n \to \infty) \tag{4.77}$$

when E_n is expectation over the data sequence $A^{(n)}$. This can in fact be taken as the definition of asymptotic Bayes risk efficiency instead of (4.1). Corollary 4.2, (3) and (4) taken together show that $T_*^{(n_1,n_2)}$ is

also an asymptotically efficient decision rule. Since $R(T_*^{(n_1,n_2)})$, $R(T_*^{(n_1)})$ are bounded for all n_1, n_2 we have

$$E_{n_1}E_{n_2}(R(T_*^{(n_1,n_2)})) \to E_{n_1}(R(T_*^{(n_1)})) \quad (\text{as } n_2 \to \infty) \tag{4.78}$$

$$E_{n_1}(R(T_*^{(n_1)})) \to R(\dot{a}_B) \quad (\text{as } n_1 \to \infty) \tag{4.79}$$

where E_{n_1}, E_{n_2} are expectations over the training sequence $A^{(n_1)}$ and the test sequence $A^{(n_2)}$, respectively. It is not possible to directly compare the asymptotic properties of $T_{GO}^{(n)}$ and $T_*^{(n_1,n_2)}$ because of the iterated limit in (4.69), (4.70) and (4.78), (4.79). Number sequences $[n_1^{(n)}]$, $[n_2^{(n)}] = [n - n_1^{(n)}]$ must be found such that

$$R(T_*^{(n_1^{(n)},n-n_1^{(n)})}) \xrightarrow{P} R(\dot{a}_B) \quad (\text{as } n \to \infty) \tag{4.80}$$

or

$$E_n(R(T_*^{(n_1^{(n)},n-n_1^{(n)})})) \to R(\dot{a}_B) \quad (\text{as } n \to \infty) \tag{4.81}$$

Some work has been done on this problem but no results are available.

Since Gordon and Olshen are only concerned with asymptotic results, they choose the termination parameter k(n) to satisfy (4.8), (4.9). For n fixed, any $k = 1,\ldots,n$ is acceptable. Friedman suggests k be determined by minimizing an estimate of the Bayes risk based on the test sequence $A^{(n_2)}$. We now investigate the asymptotic properties of such a rule.

Let $T_1^{(n_1)}, T_2^{(n_1)}, \ldots, T_{n_1}^{(n_1)}$ be binary decision trees generated by applying the multiclass partitioning algorithm of Section 2.3 to the training sequence $A^{(n_1)}$, modified as in Section 4.1, but with $k = i$ for $T_i^{(n_1)}$. Let

$$R(T_F^{(n_1)}) = \min_{i=1,\ldots,n_1} R(T_i^{(n_1)}) \tag{4.82}$$

$$\hat{R}^{(n_2)}(T_F^{(n_1,n_2)}) = \min_{i=1,\ldots,n_1} \hat{R}^{(n_2)}(T_i^{(n_1)}) \tag{4.83}$$

By comparing $T_F^{(n_1)}$ to $T_*^{(n_1)}$ and $T_F^{(n_1,n_2)}$ to $T_*^{(n_1,n_2)}$, it is clear that Corollary 4.2 holds with $T_*^{(n_1)}$, $T_*^{(n_1,n_2)}$ replaced by $T_F^{(n_1)}$, $T_F^{(n_1,n_2)}$, respectively. Unlike $T_{GO}^{(n)}$ and $T_*^{(n_1,n_2)}$, it is easy to compare $T_F^{(n_1,n_2)}$ and $T_*^{(n_1,n_2)}$. The following corollary shows that $R(T_*^{(n_1,n_2)})$ converges to $R(\mathfrak{u}_B)$ at least as fast as does $R(T_F^{(n_1,n_2)})$ (in the indicated sense).

Corollary 4.3

For all $\varepsilon > 0$, $\delta > 0$, n_1, there exists $n_2^0(\varepsilon,\delta,n_1)$ such that

$$\Pr\{R(T_*^{(n_1,n_2)}) \geq R(T_F^{(n_1,n_2)}) + \varepsilon\} < \delta \quad n_2 \geq n_2^0(\varepsilon,\delta,n_1) \tag{4.84}$$

Proof

Since $T_F^{(n_1)} \subset T_*^{(n_1)}$ we have $R(T_*^{(n_1)}) \leq R(T_F^{(n_1)})$. Thus, for $\varepsilon > 0$, $R(T_*^{(n_1,n_2)}) < R(T_F^{(n_1,n_2)}) + \varepsilon$ whenever $|R(T_*^{(n_1,n_2)}) - R(T_*^{(n_1)})| < \frac{\varepsilon}{2}$, $|R(T_F^{(n_1,n_2)}) - R(T_F^{(n_1)})| < \frac{\varepsilon}{2}$. Thus,

$$\Pr\{R(T_*^{(n_1,n_2)}) \geq R(T_F^{(n_1,n_2)}) + \varepsilon\}$$

$$\leq \Pr\{|R(T_*^{(n_1,n_2)}) - R(T_*^{(n_1)})| \geq \frac{\varepsilon}{2} \text{ or } |R(T_F^{(n_1,n_2)}) - R(T_F^{(n_1,n_2)})| \geq \frac{\varepsilon}{2}\}$$

$$\leq \Pr\{|R(T_*^{(n_1,n_2)}) - R(T_*^{(n_1)})| \geq \frac{\varepsilon}{2}\} + \Pr\{|R(T_F^{(n_1,n_2)}) - R(T_F^{(n_1)})| \geq \frac{\varepsilon}{2}\} \tag{4.85}$$

From Corollary 4.2(3) there exists $n_2^0(\varepsilon,\delta,n_1)$ such that

$$\Pr\{|R(T_*^{(n_1,n_2)}) - R(T_*^{(n_1)})| \geq \frac{\varepsilon}{2}\} < \frac{\delta}{2} \qquad n_2 \geq n_2^0(\varepsilon,\delta,n_1) \tag{4.86}$$

$$\Pr\{|R(T_F^{(n_1,n_2)}) - R(T_F^{(n_1)})| \geq \frac{\varepsilon}{2}\} < \frac{\delta}{2} \qquad n_2 \geq n_2^0(\varepsilon,\delta,n_1) \tag{4.87}$$

(4.85), (4.86), and (4.87) complete the proof.

As a final comment, the analysis/measure-theoretic results used in this chapter (properties of measurable sets and functions; absolute continuity; Radon-Nikodym and Lebesgue Dominated Convergence theorems) can be found in Royden [11] and Fleming [12]. Probability/statistical results (consistency and efficiency of estimators; convergence of random sequences and functions) can be found in Rao [13]. A discussion of iterated limits is given in Bartle [14].

V. CONCLUSIONS AND SUGGESTIONS FOR FURTHER WORK

In this chapter we draw the conclusion that Friedman's recursive partitioning algorithm can be extended to the multiclass case, with the same desirable statistical and computational properties. However, we also conclude that certain issues arise in the c-class problem ($c > 2$) that did not exist or were obscured for the 2-class case. Suggestions are given for further work.

5.1 Conclusions

We have seen that Friedman's [2] 2-class recursive partitioning algorithm can be extended to the multiclass case, with the same desirable statistical and computational properties. However, we have also seen that certain issues arise in the c-class problem ($c > 2$) that did not exist or were obscured in the 2-class case. Consider Friedman's suggestion that the c-class problem be solved by solving c 2-class problems. This appears to be a satisfactory solution. In fact, we were able to account for prior probabilities and losses by considering mixture marginal cumulative distribution functions for a group of classes, although we do not give this result here. But a solution was not found for the computational/storage problem of optimally labelling decision regions, or for the problem of restricting the number of training vectors in a decision region. This led to the conclusion that a single decision tree was needed for classifying all classes, and consequently to the multiclass recursive partitioning algorithm of Section 2.3. Similarly, Friedman

suggests that the number of training vectors in a terminal node be large enough to obtain good estimates of the within-node class measures. Friedman introduces the termination parameter k = minimum number of training vectors in a terminal node. But for large c and fixed k there are many possible terminal node populations. This led to the conclusion that the optimal k might vary from node to node and consequently to the tree termination algorithm of Chapter 3. Finally, the proof of Theorem 4.2, that measure-consistent density estimates yield asymptotically efficient decision rules for the c-class case revealed a structure that was almost entirely lacking for only 2 classes.

5.2 Suggestions for Further Work

Sufficient numerical work should be done to confirm our results. We note that both Friedman's suggestion for solving the c-class problem and the multiclass recursive partitioning algorithm of Section 2.3 have been implemented and tested by Monte Carlo procedure on a problem with $c = 5$ Gaussian classes in a $d = 2$ dimensional space. The c binary decision trees generated by Friedman's multiclass algorithm were terminated by a nonoptimized value of the termination parameter k, and the decision regions were labelled in the manner he suggests. For simplicity, the Section 2.3 algorithm was also terminated by a nonoptimized value of k. Results are given in [15]. The results indicate that the Section 2.3 algorithm has a lower average class error rate for a given complexity (number of terminal nodes). However, much more thought should be given to the problem of finding suitable test cases, the tree-termination algorithm should be used, and other parametric and nonparametric methods should also be compared.

GLOSSARY OF SYMBOLS

(~ in order introduced)

c	number of classes
d	dimension of observation space
$x^{(1)}, x^{(2)}, \ldots$ or $[x^{(n)}]$	a sequence
${}_k\underline{\alpha}^{(j)}$	$j^{\underline{th}}$ vector in $k^{\underline{th}}$ class data sequence (sample)
${}_k\alpha_i^{(j)}$	$i^{\underline{th}}$ component of ${}_k\underline{\alpha}^{(j)}$
$A_k^{(n)}$	$k^{\underline{th}}$ class data sequence (sample)
$A_k^{(n_1)}$	$k^{\underline{th}}$ class training sequence (sample)
$A_k^{(n_2)}$	$k^{\underline{th}}$ class test sequence (sample)
$A^{(n)}$	data sequence (sample)
$A^{(n_1)}$	training sequence (sample)
$A^{(n_2)}$	test sequence (sample)
$n(k)$	number of vectors in $k^{\underline{th}}$ class data sequence (sample) $A_k^{(n)}$
$n_1(k)$	number of vectors in $k^{\underline{th}}$ class training sequence (sample) $A_k^{(n_1)}$
$n_2(k)$	number of vectors in $k^{\underline{th}}$ class test sequence (sample) $A_k^{(n_2)}$
n	number of vectors in data sequence (sample) $A^{(n)}$

n_1	number of vectors in training sequence (sample) $A^{(n_1)}$
n_2	number of vectors in test sequence (sample) $A^{(n_2)}$
$\#_{n(k)}(S)$	number of vectors in $k^{\underline{th}}$ class data sequence (sample) $A_k^{(n)}$ and $S \subset \mathbb{R}^d$
$\#_{n_1(k)}(S)$	number of vectors in $k^{\underline{th}}$ class training sequence (sample) $A_k^{(n_1)}$ and $S \subset \mathbb{R}^d$
$\#_{n_2(k)}(S)$	number of vectors in $k^{\underline{th}}$ class test sequence (sample) $A_k^{(n_2)}$ and $S \subset \mathbb{R}^d$
$\#_n(S)$	number of vectors in data sequence (sample) $A^{(n)}$ and $S \subset \mathbb{R}^d$
$\#_{n_1}(S)$	number of vectors in training sequence (sample) $A^{(n_1)}$ and $S \subset \mathbb{R}^d$
$\#_{n_2}(S)$	number of vectors in test sequence (sample) $A^{(n_2)}$ and $S \subset \mathbb{R}^d$
$F_k(\underline{\alpha})$	joint cumulative distribution function of class k
$F_k(\alpha_i)$	marginal cumulative distribution function of class k for coordinate i
μ_k	probability measure of class k
π_k	prior probability of class k
ℓ_k	misclassification loss for class k
T	a binary decision tree
$t_j(T)$	$j^{\underline{th}}$ node or decision point of binary decision tree T
$t_0(T)$	root node of binary decision tree T
E(T)	edges of binary decision tree T
m(T)	number of nodes in binary decision tree T

$o(T)$	number of levels in binary decision tree T
$l_j(T)$	pointer to left subnode of $t_j(T)$
$r_j(T)$	pointer to right subnode of $t_j(T)$
$S_j(T)$	subtree of binary decision tree T with root node $t_j(T)$
$i_j^*(T)$	partitioned coordinate at $t_j(T)$
$\alpha_{i_j^*}^*(T)$	value of partitioned coordinate $i_j^*(T)$ at $t_j(T)$
$\|l_j(T)\|$	label of $t_j(T)$ if $t_j(T)$ terminal node ($l_j(T) < 0$)
$c_j(T)$	label of $t_j(T)$ if $t_j(T)$ ultimately becomes a terminal node
$D(\alpha_i)$	Kolmogorov-Smirnov distance between $F_1(\alpha_i)$, $F_2(\alpha_i)$
B	box (rectangular parallelpiped with sides parallel to coordinate axes)
$\tilde{A}_k^{(n_1)}$	rearrangement of $k^{\underline{th}}$ class training sequence (sample) $A_k^{(n_1)}$
${}_k\underline{\tilde{\alpha}}^{(j)}$	$j^{\underline{th}}$ vector in $\tilde{A}_k^{(n_1)}$
${}_k\tilde{\alpha}_i^{(j)}$	$i^{\underline{th}}$ component of ${}_k\underline{\tilde{\alpha}}^{(j)}$
$\hat{F}_k^{(n)}(\alpha_i)$	estimate of $F_k(\alpha_i)$ based on data sequence (sample) $A^{(n)}$
$\hat{F}_k^{(n_1)}(\alpha_i)$	estimate of $F_k(\alpha_i)$ based on training sequence (sample) $A^{(n_1)}$
$\hat{F}_k^{(n_2)}(\alpha_i)$	estimate of $F_k(\alpha_i)$ based on test sequence (sample) $A^{(n_2)}$
$\hat{\mu}_k^{(n)}$	estimate of μ_k based on data sequence (sample) $A^{(n)}$
$\hat{\mu}_k^{(n_1)}$	estimate of μ_k based on training sequence (sample) $A^{(n_1)}$

$\hat{\mu}_k^{(n_2)}$	estimate of μ_k based on test sequence (sample) $A^{(n_2)}$
B_j	in Chapter 2, a terminal box which results from applying Friedman's 2-class algorithm to class j and classes 1,...,j-1,j+1,...,c taken as a group. In Chapter 3, the box associated with $t_j(T)$.
$D_{m,n}(\alpha_i)$	Kolmogorov-Smirnov distance between $F_m(\alpha_i)$, $F_n(\alpha_i)$
$R_{m,n}(\alpha_i)$	Bayes risk of partitioning coordinate i at α_i
$R(T)$	Bayes risk of binary decision tree T
$I(\text{l.e.})$	1 if l.e. (logical expression) is true, 0 otherwise
$\hat{R}^{(n)}(T)$	estimate of R(T) based on data sequence (sample) $A^{(n)}$
$\hat{R}^{(n_1)}(T)$	estimate of R(T) based on training sequence (sample) $A^{(n_1)}$
$\hat{R}^{(n_2)}(T)$	estimate of R(T) based on test sequence (sample) $A^{(n_2)}$
$T_0^{(n_1)}$	binary decision tree generated by applying the multi-class partitioning algorithm of Section 2.3 to the training sequence (sample) $A^{(n_1)}$ with termination criterion that terminal nodes only contain vectors from a single class (in Chapter 4, modified as in Section 4.1 but with only (4.12) as the termination criteria)
k	minimum number of vectors at a terminal node (Friedman's termination parameter)
T_0	a finite binary decision tree
T_*	binary decision tree generated by applying the tree termination algorithm to T_0 based on the actual class distributions
T_b	binary decision tree before descendents of some t_i are deleted and t_i becomes a terminal node
T_a	binary decision tree after descendents of some t_i are deleted and t_i becomes a terminal node
T'	binary decision tree constructed from $T \subset T_0$ such that $T' \subset T_0$, $R(T') \leq R(T)$

$T_*^{(n_1)}$	binary decision tree generated by applying the tree termination algorithm to $T_0^{(n_1)}$ based on the actual class distributions
$T_*^{(n_1,n_2)}$	binary decision tree generated by applying the tree termination algorithm to $T_0^{(n_1)}$ based on the test sequence (sample) $A^{(n_2)}$
$Q(T)$	a cost function which can be optimized by the tree termination algorithm
$\hat{a}$	a decision rule
$R(\hat{a})$	the Bayes risk of decision rule $\hat{a}$
$\hat{a}^{(n)}$	a decision rule based on the data sequence (sample) $A^{(n)}$
$\hat{a}_B$	the optimal Bayes decision rule
ν	a convex combination of $\mu_1,\mu_2,\ldots,\mu_c$
$\hat{\nu}^{(n)}$	an estimate of ν based on the data sequence (sample) $A^{(n)}$
$\frac{d\mu_k}{d\nu}(\underline{\alpha})$	Radon-Nikodym derivative of measure μ_k with respect to measure ν
$\widehat{\frac{d\mu_k}{d\nu}}^{(n)}(\underline{\alpha})$	estimate of $\frac{d\mu_k}{d\nu}$ based on the data sequence (sample) $A^{(n)}$
$\hat{a}_{GO}^{(n)}$	a decision rule which partitions $\mathbb{R}^d$ into a finite set of boxes and is invariant to coordinate-by-coordinate strictly monotone transformations
$B^{(n)}(\underline{\alpha})$	the unique box in $\hat{a}_{GO}^{(n)}$ which contains $\underline{\alpha}$
$\|\cdot\|$	a norm on $\mathbb{R}^e$
$I_m^{(k)}$	index set used to recursively define I_m
$I_1,I_2,\ldots,I_q$	index sets $\subset\{1,2,\ldots,c-1\}$
$I_m(i,j)$	index set $\subset I_m$

$\mathcal{F}'$	a directed spanning tree ($\mathcal{F}$ with edges directed away from node c)
$\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_q$	connected graphs corresponding to $I_1, I_2, \ldots, I_q$
$\mathcal{G}$	connected graph which has spanning tree $\mathcal{F}$ as a subgraph
$\mathcal{F}$	a spanning tree which is a subgraph of $\mathcal{G}$
P_k	path from node c to node k in $\mathcal{F}'$
$I_{ij}(\underline{\alpha})$	an indicator function
$I_{ij}^{(n)}(\underline{\alpha})$	an indicator function based on data sequence (sample) $A^{(n)}$
W_{ij}	a set $\subset \mathbb{R}^d$
$\mathbb{W}$	Cartesian product of sets of sets $\subset R^d$
$\underline{W}_1, \underline{W}_2, \ldots, \underline{W}_r$	?-tuples of sets $\subset \mathbb{R}^d$
$W_1, W_2, \ldots, W_r$	sets $\subset \mathbb{R}^d$
$T_{GO}^{(n)}$	binary decision tree generated by applying the multi-class partitioning algorithm of Section 2.3 to the data sequence (sample) $A^{(n)}$, modified as in Section 4.1
$T_{GO}^{(n_1)}$	same as $T_{GO}^{(n)}$ except uses training sequence (sample) $A^{(n_1)}$
E_n	expectation over data sequence $A^{(n)}$
E_{n_1}	expectation over training sequence $A^{(n_1)}$
E_{n_2}	expectation over test sequence $A^{(n_2)}$
$[n_1^{(n)}]$, $[n_2^{(n)}]$	number sequences

$T_1^{(n_1)}, T_2^{(n_1)}, \ldots, T_{n_1}^{(n_1)}$ binary decision trees generated by applying the multiclass partitioning algorithm of Section 2.3 to the training sequence $A^{(n_1)}$, modified as in Section 4.1 but with $k = i$ for $T_i^{(n_1)}$

$T_F^{(n_1)}$ $T_i^{(n_1)}$ which minimizes Bayes risk based on the actual class distributions

$T_F^{(n_1,n_2)}$ $T_i^{(n_1)}$ which minimizes Bayes risk based on the test sequence (sample) $A^{(n_2)}$

REFERENCES

[1] T. M. Cover and P. E. Hart (1967), "Nearest Neighbor Pattern Classification," IEEE Trans. Information Theory, Vol. IT-13, pp. 21-27.

[2] J. H. Friedman (1977), "A Recursive Partitioning Decision Rule for Non-parametric Classification," IEEE Trans. Computers, Vol. C-26, pp. 404-408.

[3] T. W. Anderson (1966), "Some Nonparametric Multivariate Procedures Based on Statistically Equivalent Blocks," in Multivariate Analysis (ed. P. R. Krishnaiah), New York: Academic Press.

[4] E. G. Henrichon and K.-S. Fu (1969), "A Nonparametric Partitioning Procedure for Pattern Classification," IEEE Trans. Computers, Vol. C-18, pp. 614-624.

[5] W. S. Meisel and D. R. Michalopoulos (1973), "A Partitioning Algorithm With Application in Pattern Classification and the Optimization of Decision Trees," IEEE Trans. Computers, Vol. C-22, pp. 93-103.

[6] L. Gordon and R. Olshen (1978), "Asymptotically Efficient Solutions to the Classification Problem," Annals of Statistics, Vol. 6, pp. 515-533.

[7] D. C. Stoller (1954), "Univariate Two-Population Distribution-Free Discrimination," J. Amer. Stat. Assoc., Vol. 49, pp. 770-775.

[8] E. Fix and J. L. Hodges (1951), "Discriminatory Analysis, Nonparametric Classifications," USAF School of Aviat. Med., Randolph Field, Texas, Project 21-49-004, Report 4, Contract AF41 (128)-31.

[9] J. VanRyzin (1966), "Bayes Risk Consistency of Classification Procedures Using Density Estimation," Sankhyā, Series A, Vol. 28, pp. 261-270.

[10] F. Harary (1969), Graph Theory, Reading, Mass.: Addison-Wesley.

[11] H. L. Royden (1968), Real Analysis, New York: Macmillan.

[12] W. Fleming (1977), Functions of Several Variables, New York: Springer-Verlag.

[13] C. R. Rao (1973), Linear Statistical Inference and its Applications, New York: John Wiley and Sons.

[14] R. Bartle (1975), The Elements of Real Analysis, New York: John Wiley and Sons.

[15] D. E. Gustafson, S. Gelfand and S. K. Mitter (1980), "A Non-parametric Multiclass Partitioning Method for Classification," Proc. Fifth International Conf. on Pattern Recognition, Miami Beach, pp. 654-659.