LINEAR DECISION FUNCTIONS, WITH APPLICATION TO PATTERN RECOGNITION

DISSERTATION

Submitted in Partial Fulfillment

of the Requirements for the

Degree of

DOCTOR OF ELECTRICAL ENGINEERING

at the

POLYTECHNIC INSTITUTE OF BROOKLYN

by

Wilbur Hull Highleyman, II

June 1961

Approved:

May 23 1961

Head of Department

Approved by the Guidance Committee:

Major: Electrical Engineering

Arthur E. Laemmel
Research Associate Professor
of Electrical Engineering

Minor: Mathematics

Jules P. Russell
Professor of Mathematics

Minor: Physics

M. Birnbaum
Associate Professor of Physics

VITA

Wilbur H. Highleyman was born on August 20, 1933, in Kansas City, Missouri. He received the B.E.E. degree in 1955 from Rensselaer Polytechnic Institute, and the M.S. degree from the Massachusetts Institute of Technology in 1957.

From 1955 to 1956, he was a research assistant at Lincoln Laboratory developing transistor circuits for digital computers. The following year, Mr. Highleyman received a Fellowship from Melpar, Inc., and continued work in transistor circuits and thin ferromagnetic films.

From January, 1958, to the present he has been with the Bell Telephone Laboratories, Inc., as a member of the technical staff. From 1958 to 1960, he was concerned with the problem of character recognition. Since 1960, he has worked in the field of Data Communications. The work reported upon in this thesis was performed from 1959 to 1961 at the Bell Telephone Laboratories.

Mr. Highleyman is a member of Tau Beta Pi, Eta Kappa Nu, the Institute of Radio Engineers, and an associate member of Sigma Xi.

DEDICATION

to

PATIENCE

which so characterizes my wife

JOLINE

ACKNOWLEDGMENT

I would like to thank Professor A. E. Laemmel for his guidance in this work as thesis advisor, and to acknowledge his suggestion of the problems attacked in Chapter VIII. I would also like to thank Messrs. L. A. Kamentsky, W. H. Williams, F. W. Sinden, and R. Gnanadesikan for their helpful discussions and comments on this work. I am especially grateful to W. R. Cowell and E. Wolman for reviewing portions of the manuscript, as well as for the many other discussions into which they so willingly entered and contributed.

This work was made possible by the Bell Telephone Laboratories, Inc. The use of their electronic data processing facilities is gratefully acknowledged.

AN ABSTRACT

LINEAR DECISION FUNCTIONS, WITH APPLICATION TO PATTERN RECOGNITION

by

Wilbur Hull Highleyman, II

Advisor: Arthur E. Laemmel

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Electrical Engineering

A pattern recognition machine may be considered to consist of two principal parts, a receptor and a categorizer. The receptor makes certain measurements on the unknown pattern to be recognized; the categorizer determines from these measurements the particular allowable pattern class to which the unknown pattern belongs.

This paper is concerned with the study of a particular class of categorizers, the linear decision function. The optimum linear decision function is the best linear approximation to the optimum decision function in the following sense:

1. "Optimum" is taken to mean minimum loss (which includes minimum error systems).
2. "Linear" is taken to mean that each pair of pattern classes is separated by one and only one hyperplane in the measurement space.

This class of categorizers is of practical interest for two reasons:

1. It can be empirically designed without making any assumptions whatsoever about either the distribution of the receptor measurements or the a priori probabilities of occurrence of the pattern classes, providing an appropriate pattern source is available.
2. Its implementation is quite simple and inexpensive.

Various properties of linear decision functions are discussed. One such property is that a linear decision function is guaranteed to perform at least as well as a minimum distance categorizer.

Procedures are then developed for the estimation (or design) of the optimum linear decision function based upon an appropriate sampling from the pattern classes to be categorized. The design procedure allows one to eliminate certain redundancies in the resulting categorizer and also in the receptor. Rejection criteria may be included in the design if desired.

Some very general results are obtained in a discussion concerning the design and analysis of pattern recognition experiments. These results allow one to determine how a sample of fixed size should be partitioned between the design and test phases of a pattern recognition machine, and show one how to place confidence intervals on the resulting estimate of the performance of that machine.

A method of estimating performance bounds for a linear decision function which is applicable to the use of the design data is also discussed.

Finally, the concepts and procedures thus developed are applied for illustrative purposes to two examples of pattern recognition - the determination of the geographical source of radio signals based on measurements made at a monitor site, and the recognition of hand-printed numbers.

TABLE OF CONTENTS

TABLE OF FIGURES

CHAPTER I

INTRODUCTION

There has been an ever increasing interest for the past several years in the general problem of pattern recognition. Work in this field has ranged from commercial applications (such as the reading of machine-printed characters)[15, 45,48] to the study of adaptive networks ("learning machines") which have the capability of modifying themselves so as to perform a certain function better with experience.[11,16,33,34,43, 44,52,53]

For the purposes of this paper, the term "pattern recognition" will be taken in its broadest sense to refer to any discrete classification problem. That is, a pattern recognition machine may be described loosely as follows. It is a machine which is "aware" of a finite number of distinct classifications, or classes. These will hereafter be referred to as the allowable pattern classes. The machine is presented with an item upon which it makes certain measurements (or the measurements may have been made by other means and then given to the machine), after which it is required to make a decision. The decision usually falls into one of the two following categories:

a. The item belongs to a certain allowable pattern class.

b. The item cannot be identified with any certainty, and consequently the machine refuses to attempt identification. It is then said that the machine has rejected the item.

If the machine attempts a decision of the first type and is wrong, then it is said that the machine has made an error. Note that a rejection will not be considered as an error.

The following are some examples of pattern recognition problems which have either appeared in the literature or are known to the author:

a. The identification of machine-printed, hand-printed, and hand-written [21] alpha-numeric characters based on various geometrical measurements;

b. the identification of diseases from the symptoms;[13]

c. The identification of spoken words from frequency-time spectra;[32]

d. the geographical location of radio stations based on measurements of the fading characteristics of the received wave;[30]

e. the decoding of messages which have been encoded against noise;

f. the counting of permuted blood cells in a blood smear;

g. the identification of subatomic particles from cloud chamber qr bubble chamber tracks.

The various machines (or proposals therefore) resulting from efforts in these areas can, in general, be dichotomized according to their structure, i.e., determinate or indeterminate. A determinate machine is one which is predesigned according to some criterion or procedure, and which, when finally constructed, is left unchanged. Commercially available machines are all determinate; a good deal of the exploratory work in more sophisticated pattern recognition machines is also concerned with determinate structures (for example, see the works of Harmon and Frishkopf [21], Crumb and Rupe [13], Mathews and Denes [32], Bomba [5], and Grimsdale et al [24]).

An indeterminate machine is one in which some of the parameters of its internal structure are not specified at the time of construction; rather, the machine has the ability to adjust these parameters as it becomes more experienced in its assigned task. Hence it has the capability of "adapting to its environment", or of "learning". One of the outstanding examples of such a machine is Rosenblatt's Perceptron [43,44]; Widrow [52,53], Mattson [33,34], Clark and Farley [11,16], and Roberts [42] also discuss such machines.

Some of the so-called "adaptive machines" are simply modifications of basically determinate machines in which estimates of the parameters are improved through the accumulation of more and more data (for example, the proposals of Bledsoe and Browning [4], and Baran and Estrin [3]). Consequently, there is a somewhat hazy dividing line between determinate and indeterminate structures.

The rest of this paper will deal with determinate structures, although it is recognized that the sort of structure which is proposed could be made indeterminate through the above device of simply allowing the machine to estimate its own parameters through the long term accumulation of data.

Marill and Green [31] have described the general determinate pattern recognition system in a very clear manner. They note that it consists of two principal parts, a receptor and a categorizer:

a. "The receptor has as its input a physical sample to be recognized, and as its output a set ... of quantities which characterize the physical sample. These quantities will be called measurements of the sample ..."

b. "The output ... of the receptor constitutes the input to the categorizer. The categorizer is a device which assigns each of its ... inputs to one of a finite number ... of categories ..."

The measurements which a receptor makes on the input sample may be either continuous or discrete, and a given receptor may be required to make measurements of both types. For instance, a character recognition machine might have a receptor which makes the following measurements on an unknown character: the number of closures, cusps, and straight lines (discrete), and the length and direction of the straight lines (continuous).

The categorizer must apply some sort of decision criterion to the receptor output to decide to which of the allowable pattern classes, if any, the input pattern belongs. Or the categorizer may reject the pattern as being unrecognizable if the recognition decision is unreliable in some sense.

This paper will deal with the optimization of a certain class of categorizers. This class is characterized by a certain linearity of operation which will be described later, and which is of particular interest because of the economical implementation to which it leads. It will be assumed that the sort of measurements to be made have been decided a priori, and that a receptor has been constructed (or simulated) which will make the appropriate measurements. The design of the categorizer will be based upon the receptor measurements of samples taken from the real world of patterns belonging to the allowable pattern classes for the machine.

The following chapters will first deal with the definition of "optimum". Several theorems relating to and describing the class of categorizers under study will be given, after which algorithms for the actual design of the optimum categorizer based on the previously mentioned sample will be derived. A method to determine the "usefulness" of a given receptor measurement will be discussed, as well as a particular form of rejection criteria. A general discussion of the design and interpretation of pattern recognition experiments is followed by the results of some actual experiments in which categorizers of the above type were designed and tested.

CHAPTER II

THE DECISION-THEORETIC APPROACH TO PATTERN RECOGNITION

Before describing the particular class of categorizers of concern in this paper, it will be instructive to review some results of decision theory. This will not only lay the groundwork for later discussion, but will provide certain results needed in the development of rejection criteria.

Through the application of decision theory, one can actually state what the structure of the optimum categorizer should be for a particular pattern recognition problem. However, this optimum structure is often not realizable for two reasons. The first is a purely practical problem - an optimum categorizer is often too complex to be economically feasible. The second reason is somewhat more fundamental and restrictive. The design of the optimum categorizer, as will be seen, requires the complete knowledge of certain probability distributions which are usually not available to the designer. This is not simply a problem of parameter estimation, since usually even the form of the (often multi-dimensional) distributions is not known.

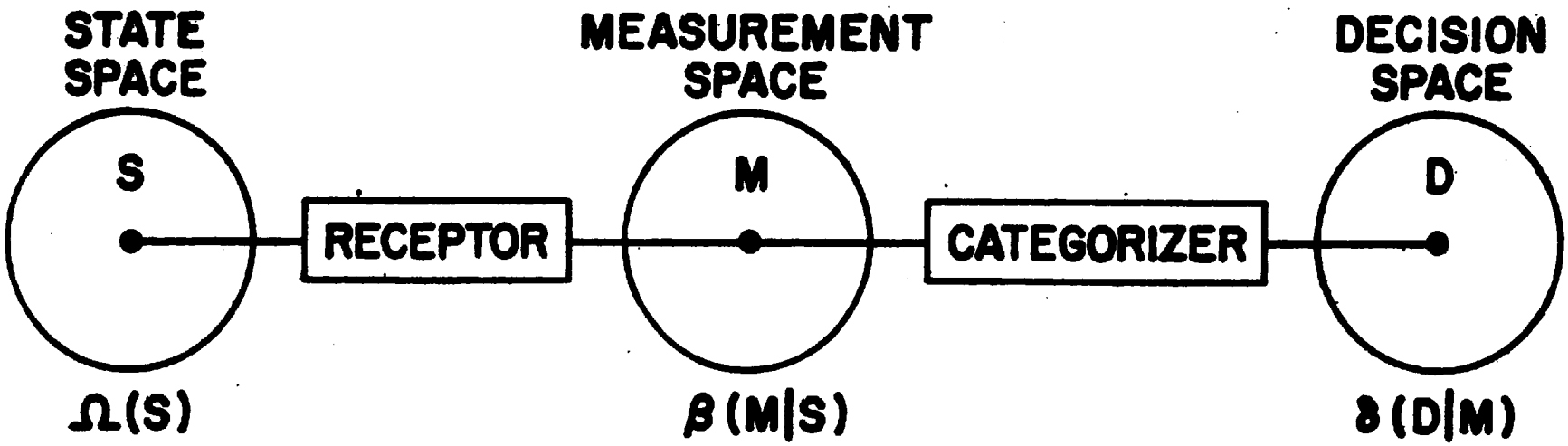
2.1 The Decision Theory Model of Pattern Recognition

The decision-theoretic model to be used is one described by Middleton and Van Meter [36] which has been modified to suit the pattern recognition problem. (See also

Chernoff and Moses [9] for a very clear and complete discussion of the fundamental concepts of decision theory.)

Let the various allowable pattern classes be described by a discrete state space, S, with probability distribution $\omega(s)$ (Figure 1). That is, a point s_i in S represents the i^{th} allowable pattern class ($1 \leq i \leq p$), and has an a priori probability of occurrence $\omega(s_i) = \omega_i$. Such a point in S will be called an input state.

When an unknown pattern is presented to the receptor, the receptor makes certain measurements on it. Let there be n (discrete or continuous) different measurements which the receptor makes on an input pattern. Then the output of the receptor, when operating upon a particular pattern, can be considered as a point in an n-dimensional measurement space, M. The concept of a measurement space is quite important to the development of later sections of this paper. Any possible input pattern results in a single point in measurement space, the coordinates of this point being determined by the receptor. That is, let the receptor output for a particular input pattern be the set of numbers $(m_1, m_2, \ldots, m_n) = m$. Then this set defines the coordinates of the point representing the input pattern in the space M. Each m_i, $1 \leq i \leq n$, represents a receptor measurement; m is the measurement vector and represents the set of receptor measurements. (For briefness, m will often be called merely a measurement when it is clear from the context whether a particular measurement or the set is meant.)



DECISION-THEORETIC MODEL

FIGURE 1

Let us assume the existence of a probability function (or density) over M, $\beta(M \mid S)$. Thus $\beta(m \mid s_1)$ is the conditional probability that a certain measurement m will be made, given a pattern from class i at the receptor; $\beta(M \mid S)$ is determined by the way the various patterns belonging to an allowable pattern class i (input state s_1) actually vary with respect to the measurements made by the receptor. It can then be said that the receptor maps each of the discrete points of S into a multitude of points, or into a continuum, in M depending on whether the measures are discrete or continuous. The various regions in M occupied by the mappings of the various points in S are usually overlapping; consequently the ideal recognizer (i.e., error free) is usually nonexistent. Some error will, in general, be made, since exactly the same set of receptor measures may occasionally be obtained for members from more than one allowable pattern class.

The categorizer then must make a decision as to which pattern class the measurement m belongs. This can be considered a mapping of the (discrete or continuous or mixed) space M onto a discrete decision space, D. There is a point in D for every allowable decision; the allowable decisions are usually

a. d_j: The pattern belongs to pattern class j $(1 \leq j \leq p)$.

b. d_o: The pattern is rejected as being unrecognizable. (The subscript zero will denote rejection.)

Let there also exist a probability function (or density) $\delta(D \mid M)$ over the space D, which is the probability that the categorizer will make the decision d_j, $0 \leq j \leq p$, given the measurement m. $\delta(D \mid M)$ is referred to as the decision function or decision criterion. If for any measurement m, $\delta(D \mid m)$ has a nonzero value for more than one point in D, then the decision function is said to be randomized. That is, a particular receptor output will not always be classified as belonging to the same pattern class. If, however, for any measurement m, $\delta(D \mid m)$ has a nonzero value (actually unity) for one and only one point in D, then $\delta(D \mid M)$ is said to be a nonrandomized decision function. It is found that the optimum decision functions discussed below are always nonrandomized. Note that the categorizer is nothing more than the implementation of the decision function $\delta(D \mid M)$.

Let a loss (or cost) function C(S,D) now be defined such that $C(s_i,d_j) = c_{ij}$ is the loss (cost) associated with making the decision d_j when the actual input state was s_i. The desired decision is d_i when the input state is s_i; therefore, the usual case requires that

$$c_{ij} > c_{io} > c_{ii} ,$$

where c_{io} is the loss associated with rejection when the input state is s_i. Note that c_{ij} is not necessarily equal to c_{ji}.

The probability of making a decision d_j when the input state is s_i is

$$p(d_j \mid s_i) = \int_M \beta(m \mid s_i)\delta(s_j \mid m)dm \ .$$

The loss when s_i is the input state (called the conditional loss) and when the decision function $\delta(D \mid M)$ is used is then

$$\tilde{C}(s_i,\delta) = \sum_{j=0}^{p} c_{ij} \int_M \beta(m \mid s_i)\delta(d_j \mid m)dm \ . \qquad (2.1)$$

Since the distribution of states is given by $\omega(S)$, the expected loss for the pattern recognition system is

$$C(\delta) = \sum_{i=1}^{p} \sum_{j=0}^{p} \int_M c_{ij}\omega_i\beta(m \mid s_i)\delta(d_j \mid m)dm \ . \qquad (2.2)$$

The optimum categorizer is defined as the implementation of that decision function, δ^*, which minimizes the expected loss, $C(\delta)$, under the appropriate <u>a priori</u> distribution $\omega(S)$ (Bayes strategy).

2.2 <u>The Solution of the Decision Theory Model</u>

The general solution to this problem has been given by Chow. [10] He shows that (2.2) is minimized by using the following decision criterion:

Let

$$Z_j(m) = \sum_{i=1}^{p} (c_{ij}-c_{io})\omega_i \beta(m \mid s_i), \qquad j = 0,1,\ldots,p.$$

Then

$$\delta^*(d_k \mid m) = 1$$

$$\delta^*(d_j \mid m) = 0 \quad \text{for all } j \neq k \qquad (2.3)$$

whenever

$$\min_j [Z_j(m)] = Z_k(m) \ .$$

Hence, given m, the optimum categorizer (in a minimum loss sense) computes a function $Z_j(m)$ for each of the allowable decisions, $j = 0,1,2,\ldots,p$, and chooses the smallest. (Note that $Z_o(m) = 0$ from (2.3).) In all decision functions of this sort, ties are arbitrarily resolved.

An interesting extension of Chow's result concerns the case when all losses due to misrecognition are equal, all losses due to rejection are equal, and all losses due to correct recognition are zero.[25] Let c be the loss due to misrecognition and c_o be the loss due to rejection, such that

$$c > c_o > 0 \ .$$

Then it is shown in Appendix I that the optimum decision rule is

$$\left.\begin{aligned} &\delta(d_k \mid m) = 1, \qquad k \neq 0 \\ &\delta(d_j \mid m) = 0, \qquad j \neq k \end{aligned}\right.$$

if

$$\omega_k \beta(m \mid s_k) \geq \omega_j \beta(m \mid s_j) \qquad \text{for all } 1 \leq j \leq p$$

and

$$\omega_k \beta(m \mid s_k) \geq \left(\frac{c-c_o}{c}\right) \sum_{i=1}^{p} \omega_i \beta(m \mid s_i) \; ;$$

however

$$\left.\begin{aligned} &\delta(d_o \mid m) = 1 \\ &\delta(d_j \mid m) = 0, \qquad j \neq 0 \end{aligned}\right.$$

if

$$\omega_j \beta(m \mid s_j) < \left(\frac{c-c_o}{c}\right) \sum_{i=1}^{p} \omega_i \beta(m \mid s_i)$$

$$\text{for all } 1 \leq j \leq p \; .$$

(2.4)

Chow has shown that decision criteria of this form correspond to minimizing the error rate for a given rejection rate. (The rejection rate is determined by the quantity $(c-c_o)/c$.) Therefore, minimizing the loss in the case of constant loss is equivalent to minimizing the error rate for a

certain rejection rate. This paper will treat the term "optimum" as meaning minimum loss; common practice has been to optimize systems with respect to error rate and rejection rate. The above development shows the simple relation between the two.

It can be seen that all of the decision functions just discussed depend, aside from the loss function, only upon comparisons between the a posteriori probabilities, $\xi(s_i \mid m)$, that the measurement m was caused by the input state s_i. That this is true can be shown by noting that

$$\xi(s_i \mid m) = \frac{\beta(m \mid s_i)\omega_i}{\varphi(m)} , \qquad (2.5)$$

where $\varphi(m)$ is the probability of occurrence of the measurement m. But once a measurement has been made, the various a posteriori probabilities differ only by the numerator of (2.5), and it is exactly this quantity which appears in the decision criteria (2.3) and (2.4).

Unfortunately, however, these a posteriori probabilities are usually unknown to the designer, and therefore categorizers based on the optimum decision function are not, in general, practically realizable. There are at least two ways around this difficulty:

a. Assume a certain form for the a posteriori distribution functions (or for the a priori function $\beta(M \mid S)$). Then by taking a random sampling from S, estimate the various parameters of these

distributions. For instance, a common assumption is that of normality and independence: given a certain pattern class, assume that the measurements made by the receptor are normally distributed, and that each measurement is independent of the others. Marill and Green [31], Chow [10], and Flores and Grey [18] have discussed this sort of approach.

b. Make no assumptions about the particular distributions involved, but rather make certain restrictions on the structure of the categorizer. Then search through all possible structures of this type to find the categorizer which is optimum with respect to a sampling of patterns from the real world. Decision trees [5,21] and adaptive machines [11,16,33,34,43,44,52,53] are examples of this sort of approach.

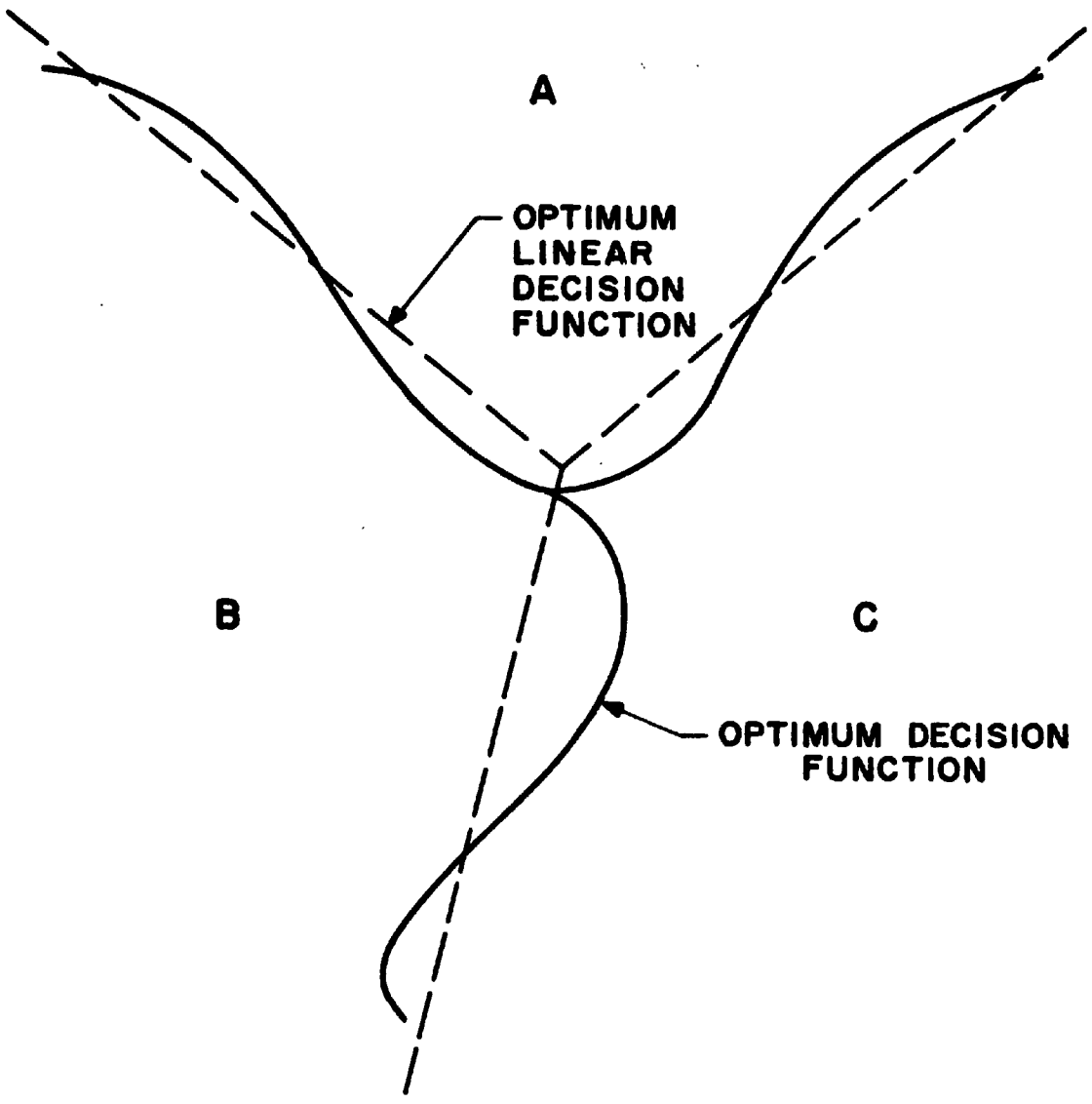
Clearly, neither of these approaches will yield a truly optimum categorizer, the first because of questionable assumptions, the second because of structural limitations. However, the use of either approach now makes the problem manageable, and optimum is reinterpreted to mean minimum loss within the framework of the approach.

2.3 Linear Decision Functions

There is another practical advantage that is realized by the second approach, namely one of economic feasibility. Even if the optimum decision function were known, its

implementation would require, in general, the use of a digital computer or other complex equipment. The cost of such equipment may, in many cases, outweigh the advantages of mechanized categorization. However, if the designer can limit his search to those structures which are economically feasible, and if the optimum structure in this class works well enough for the given purpose, then a technically feasible as well as an economically feasible solution has been found.

This paper is concerned with the study of just such a class of categorizers. To describe this class, consider a rephrasing of the optimum decision criterion (2.3). ((2.4) is simply a special case of (2.3).) Note that every point in the measurement space M is preassigned to a particular pattern class or to the rejection class by the decision criterion. Thus there is a subset M_i of M corresponding to each possible decision d_i, $0 \leq i \leq p$. Further, these subsets are nonoverlapping if the decision function is nonrandomized. The division of M into these subsets then uniquely identifies a certain decision function. We could equally well consider the decision function to be represented by the boundaries between the subsets. (Some liberty is taken here, since it will be assumed that a continuous boundary can be passed through a discrete space.) For instance, in Figure 2 is shown a two-dimensional measurement space (the receptor makes only two measurements on an input pattern) in which are shown the



DOMAINS OF THREE PATTERN CLASSES IN MEASUREMENT SPACE, AS DEFINED BY OPTIMUM AND OPTIMUM LINEAR DECISION FUNCTIONS

FIGURE 2

boundaries (the solid lines) between three different pattern classes, A, B, C. (Rejection regions are not included for simplicity.) A boundary will, in general, be some sort of curved surface. In fact, the domain of a particular pattern class may not even be singly connected.

The class of categorizers to be discussed herein may be loosely described as the optimum linear approximations to the true boundaries, under the further constraint of only one boundary per pair of pattern classes (such as those shown dotted in Figure 2). Optimum, as previously mentioned, is taken to mean minimum loss under the above constraints. Because of the linear properties of this decision criterion, a categorizer of this class will be said to implement a linear decision function. Although the primary purpose of the development is to study the synthesis of such a categorizer when the probability distributions are unknown, the problem of finding the optimum linear decision function when these distributions are known will also be discussed.

Of particular importance is the economical realization of a categorizer based upon a linear decision function. In an n-dimensional measurement space, a linear decision function will comprise a set of n-dimensional hyperplanes. An n-dimensional hyperplane is that set of all points $(x_1,\ldots,x_n)$ in M which satisfy a linear relation of the form

$$\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n + \alpha_o = 0$$

for a given set of α_i's. The fact that the actual boundaries are only portions of hyperplanes, i.e., each hyperplane usually terminates on other hyperplanes (Figure 2), is of little consequence. As will be shown in the next chapter, the representation of each boundary by a full hyperplane is equivalent.

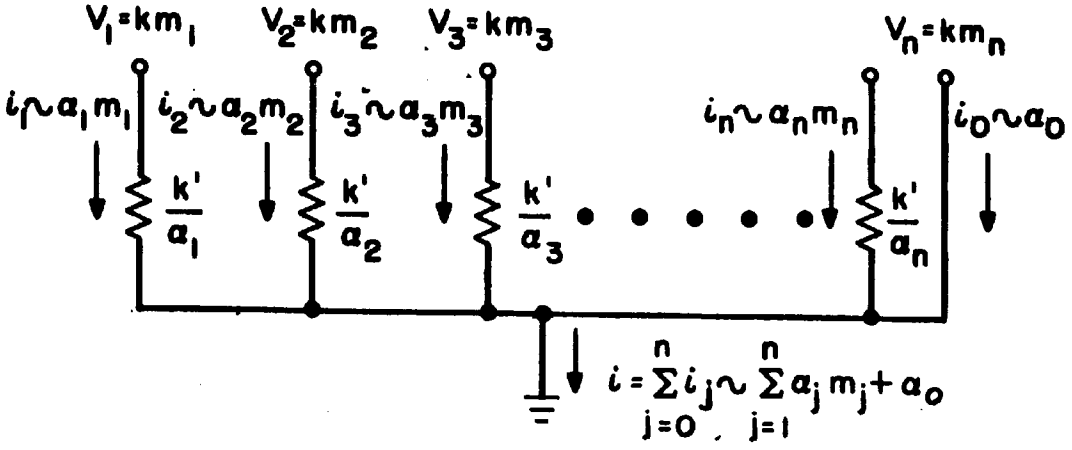
It will be shown later that, in order to classify a point m in M, it is only necessary to determine which side of each hyperplane this point is on. This is determined by the sign of the quantity*

$$\sum_{i=1}^{n} \alpha_i m_i + \alpha_o \ . \qquad (2.6)$$

In fact, the magnitude of this quantity is proportional to the distance of the point m from the hyperplane. Consequently, in order to classify a point m (that is, recognize an input pattern), it is only necessary to evaluate a set of quantities like (2.6). But such a calculation can be done with several varieties of very inexpensive networks, such as the resistive adder shown in Figure 3. This supports the statement of economy.

---

*Note that this linear form equated to zero defines a structure which is commonly found in the automata field. It goes by various names, such as artificial neuron [28], associative unit [43,44], and Adaline [52,53]. In this paper, it will simply be called by its already well-established name of "hyperplane".



IMPLEMENTATION OF A HYPERPLANE

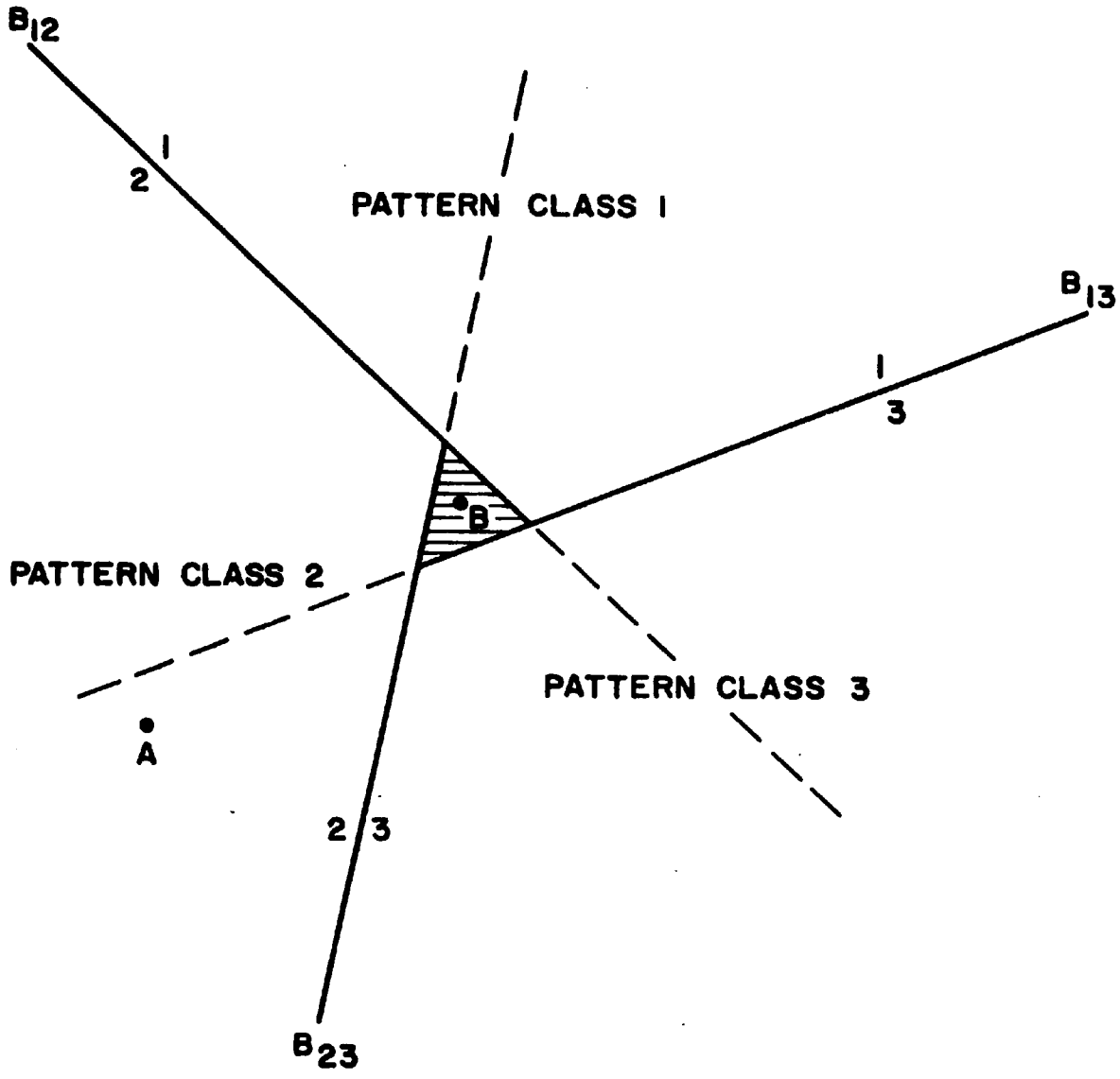
FIGURE 3

CHAPTER III

SOME PROPERTIES OF LINEAR DECISION FUNCTIONS

3.1 The Classifying Procedure

Before discussing some of the properties of linear decision functions, the classification procedure will first be discussed. Figure 4 illustrates a measurement space in which the domains of three pattern classes are shown, as determined by a linear decision function. The boundaries, which are really truncated hyperplanes, will be represented by the complete hyperplanes as indicated by the dotted lines. It will be seen that the truncation is automatically taken into account by the classifying procedure. Since there is one and only one boundary per pair of pattern classes, Figure 4 shows three boundaries separating the three classes. The boundary separating the i^{th} and j^{th} classes will be denoted B_{ij}. Further, in schematic representation as in Figure 4, each hyperplane B_{ij} will be identified by the pair of numbers, i,j, placed in such a way as to show which side of B_{ij} corresponds to class i, and which to class j. Since this is sufficient identification, the notation "B_{ij}" will usually be left off an illustration.



GEOMETRIC REPRESENTATION OF A
LINEAR DECISION FUNCTION

FIGURE 4

In order to classify a certain measurement, we note which side of each boundary it is on. If it is on the i^{th} side of B_{ij}, then it is known that the measurement is not to be identified with class j. Consider the measurement A shown in Figure 4.

a. It is on the 2 side of B_{12}; therefore it is not to be identified with class 1.

b. It is on the 3 side of B_{13}; therefore it is not to be identified with class 1. (Here the extension of B_{13} is used.)

c. It is on the 2 side of B_{23}; therefore it is not to be identified with class 3.

d. It is not to be identified with classes 1 or 3; therefore it will be identified with class 2.

This procedure may be represented by the following notation:

$$\not{1} - 2$$

$$\not{1} - 3$$

$$2 - \not{3} .$$

That is, $[\not{1} - 2]$ indicates that a point is on the 2 side of B_{12}, and therefore cannot be identified as a 1.

Note that, although A was in class 2, the boundary B_{13} was still used. In general, since it is not known in advance to which class a measurement belongs, all boundaries must be interrogated. The difficulty one might get into by

terminating the interrogation process early is illustrated by the measurement B in Figure 4:

$$\chi - 2$$

$$1 - \beta$$

$$\gamma - 3 \ .$$

Since B cannot belong to any of the three classes, it is rejected. This is not the normal sort of rejection due to an unreliable decision; rather it is a rejection inherent in a linear decision function. If the process had been terminated early, say after the second step, the measurement B would have been identified with class 2. In fact, it would be so identified with any of the three classes by taking the hyperplanes in the proper order. Therefore, one must reserve his decision until all boundaries have been interrogated (or until the measurement is definitely rejected). This suggests, then, that a categorizer based on a linear decision function ought to be a parallel rather than a sequential sort of machine, in that all boundaries might just as well be interrogated simultaneously.

One further comment is appropriate concerning the determination of which side of a hyperplane a point lies. Consider a hyperplane, B, given by

$$\sum_{i=1}^{n} \alpha_i x_i + \alpha_o = 0 \qquad (3.1)$$

where

$$\sum_{i=1}^{n} \alpha_i^2 = 1 \ . \tag{3.2}$$

Then the α_i, $1 \leq i \leq n$, are the direction cosines of the hyperplane, and α_o is its distance from the origin. We are interested in determining the distance, s, of a point m from B (Figure 5). Pass a hyperplane, say C, through m parallel to B. It will be represented by

$$\sum_{i=1}^{n} \alpha_i x_i + \beta_o = 0 \ , \tag{3.3}$$

where β_o is the distance of C from the origin. Hence the distance between B and C, which is also the distance between m and B, is

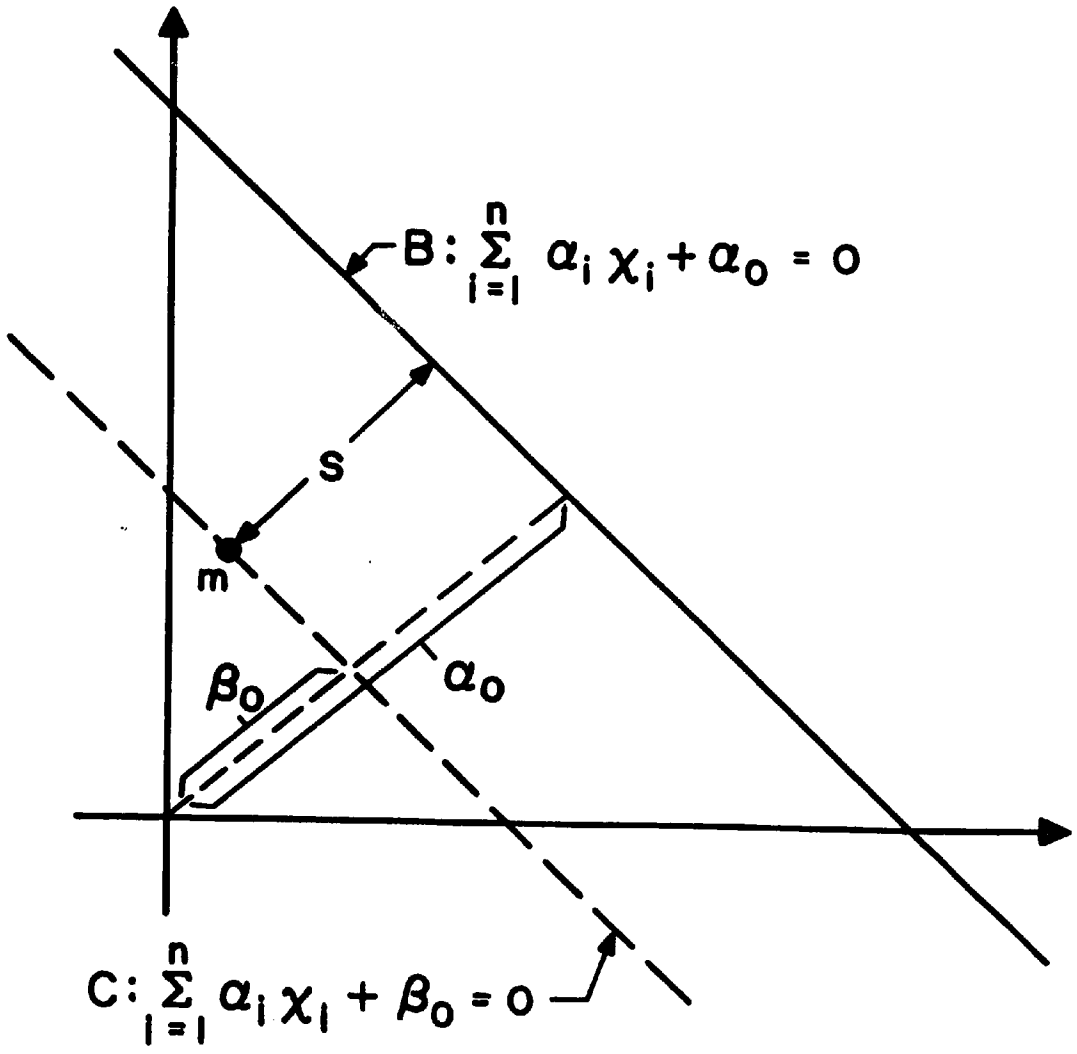
$$s = \alpha_o - \beta_o \ .$$

But C passes through m; therefore m must be a solution of (3.3), requiring that

$$\beta_o = - \sum_{i=1}^{n} \alpha_i m_i \ .$$

Therefore

$$s = \sum_{i=1}^{n} \alpha_i m_i + \alpha_o \ . \tag{3.4}$$



DETERMINATION OF THE DISTANCE OF
A POINT FROM A HYPERPLANE

FIGURE 5

Hence, the distance of a point to a hyperplane (3.1) is simply given by substituting the coordinates of the point into the expression for the hyperplane (as in (3.4)), providing the expression is in a normalized form, that is, that (3.2) holds. The point is on one side of the hyperplane if (3.4) is positive, and on the other if (3.4) is negative. Which side of the hyperplane is to be positive or negative is completely arbitrary, since multiplication of (3.1) by -1 changes the sign of (3.4), but does not change the hyperplane.

3.2 Some Theorems Pertaining to Linear Decision Functions

One may rightly ask just why he should consider a linear decision function. Is there any guarantee that it will work? In general, this question can only be answered by designing the categorizer, and then deciding whether the resulting system is good enough. However, some confidence in linear decision functions may be obtained from the following theorem.

Theorem 1: For any categorizer based upon minimizing a Euclidean distance* to a set of reference points, there exists a categorizer based on a linear decision function

---

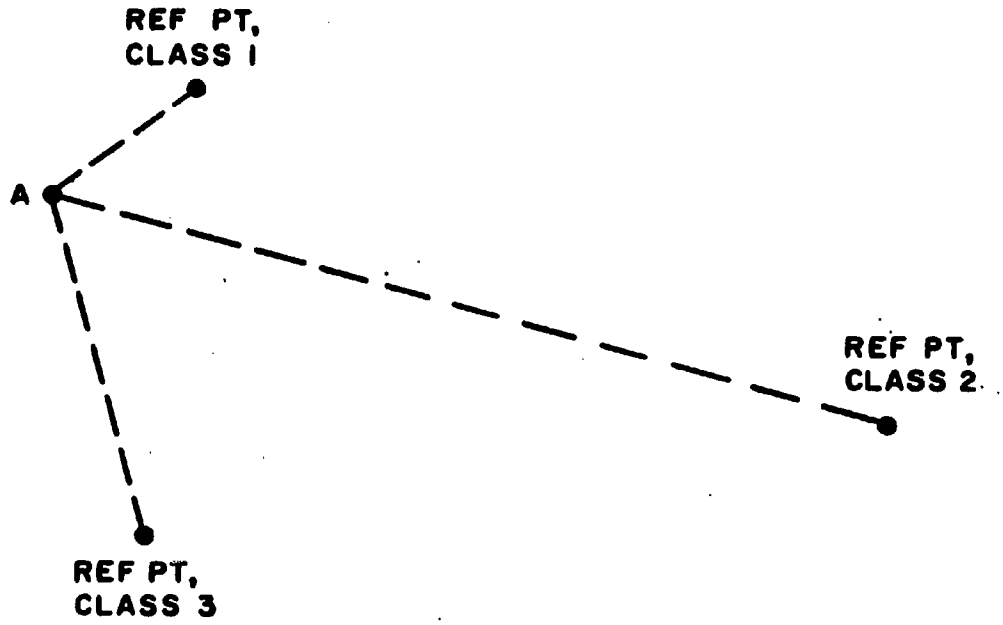
*If x and y are two points with coordinates x_i, y_i, $1 \leq i \leq n$, then the Euclidean distance s between x and y is

$$s = \sum_{i=1}^{n} (x_i - y_i)^2 .$$

which is at least as good. This includes categorizers which maximize a cross-correlation function, and those which minimize a Hamming distance.

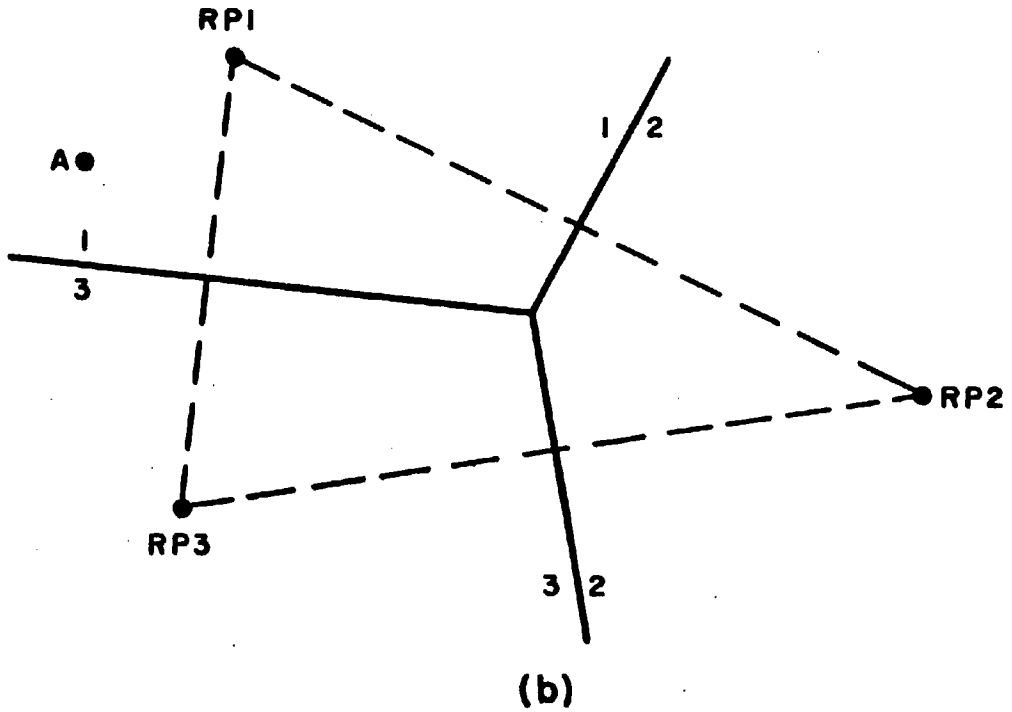
Proof: Figure 6a illustrates a minimum distance categorizer. A measurement A is identified with the class represented by that reference point to which it is closest in a Euclidean sense. Consider reference points 1 and 2 (RP1 and RP2) and the hyperplane B_{12} which is the perpendicular bisector of the line segment joining RP1 and RP2 (Figure 6b). Then the statement that a point A is closer to RP1 than to RP2 is equivalent to the statement that the point lies on the 1 side of B_{12}. By constructing such a hyperplane for every pair of reference points, a linear decision function equivalent to the minimum distance decision function is obtained. Therefore, minimum Euclidean distance decision functions are a subclass of linear decision functions. (Sebestyen [46,47] has considered non-Euclidean minimum distance decision functions, which are not a subclass of linear decision functions.)

Next it will be shown that maximizing a cross-correlation function is equivalent to minimizing a Euclidean distance. Consider an unknown measurement represented by a point m in an n-dimensional space, which is to be compared to a set of reference points $\{r_j\}$, $j = 1,2,\ldots,p$. Let $\bar{R}_i$ and



(a)

MINIMUM DISTANCE CATEGORIZER



(b)

LINEAR DECISION FUNCTION EQUIVALENT

THE RELATION OF A MINIMUM DISTANCE CATEGORIZER TO A LINEAR DECISION FUNCTION

FIGURE 6

$\overline{M}$ be the vectors originating from the origin and terminating on r_i and on m respectively. Let the magnitude of each reference vector be unity:

$$|\overline{R}_i| = \left[\sum_{i=1}^{n} r_i^2\right]^{1/2} = 1, \qquad 1 \leq i \leq p \ . \tag{3.5}$$

Then by the law of cosines, the distance of m from r_i, s_i, is

$$s_i^2 = |\overline{R}_i|^2 + |\overline{M}|^2 - 2|\overline{R}_i||\overline{M}| \cos \theta_i$$

$$= \overline{R}_i \cdot \overline{R}_i + \overline{M} \cdot \overline{M} - 2\overline{R}_i \cdot \overline{M} \ , \tag{3.6}$$

where θ_i is the angle between $\overline{M}$ and $\overline{R}_i$. A minimum distance categorizer will compute s_i for each $1 \leq i \leq p$ and base its decision on choosing the minimum s_i. Since $|\overline{M}|^2$ is common to each s_i, and $|\overline{R}_i|^2 = 1$ for all i, then minimizing s_i is equivalent to maximizing

$$\overline{R}_i \cdot \overline{M} = \sum_{i=1}^{n} r_i m_i \ .$$

But this is proportional to the cross-correlation φ_i between the measurement m and the reference pattern r_i:

$$\varphi_i = \frac{\overline{R}_i \cdot \overline{M}}{|\overline{M}||\overline{R}_i|} = \frac{\overline{R}_i \cdot \overline{M}}{|\overline{M}|} \ . \tag{3.7}$$

Since for a particular categorization trial, $|\bar{M}|$ is a constant common to all φ_i, then maximizing the cross-correlation (3.7) between a measurement m and a set of references $\{r_i\}$ is equivalent to minimizing the Euclidean distance (3.6) between m and the normalized reference points (normalized so that (3.5) holds).

That minimizing a Hamming distance is equivalent to minimizing a Euclidean distance is easily shown by noting that a Hamming distance is simply the square of a Euclidean distance. Let x and y be two n-bit binary numbers. The Hamming distance between x and y is the number of bit positions which differ in the two binary numbers. This may be written

$$D^2 = \sum_{i=1}^{n} (x_i - y_i)^2$$

(assuming each x_i and y_i take on the values 0 or 1) which is the Euclidean distance between the points x and y. This completes the proof to Theorem 1.

The upper bound on the number of hyperplanes required for a linear decision function is determined by noting that, for every pattern class, there will be one hyperplane separating it from every other pattern class. If there are n pattern classes, there will then be n(n-1) such hyperplanes. But this has counted each hyperplane twice. Therefore,

Theorem 2: For n pattern classes, a linear decision function comprises n(n-1)/2 hyperplanes.

It will be shown later that not all the hyperplanes are always needed, and techniques will be developed to detect unnecessary, or redundant, hyperplanes. Consequently we will have occasion to refer to complete linear decision functions, in which all the n(n-1)/2 hyperplanes are present, and incomplete linear decision functions in which some hyperplanes are not included.

Theorem 3: (Uniqueness) A complete linear decision function will classify any measurement into no more than one allowable pattern class.

Proof: Assume that a complete linear decision function has classified a measurement into both classes i and j. But because of the completeness criterion this linear decision function contains a hyperplane B_{ij} which will indicate that either the point cannot belong to class i or that it cannot belong class j (assuming that a point lying on a boundary is categorized according to some convention), thus contradicting the assumption. It has already been demonstrated that some measurements may not be classified into any of the allowable pattern classes by a linear decision function, complete or otherwise; these are the patterns which are rejected (see Figure 4). Of course, the classification determined by an incomplete linear decision function may not be unique.

Theorem 4: The points in a measurement space which are identified with a particular class by a linear decision function form a convex set.*

Proof: This is a standard proof [22] which is repeated here. The domain in measurement space corresponding to a particular pattern class is the set of all points satisfying a set of linear inequalities corresponding to the bounding hyperplanes for that class:

$$\begin{aligned} a_{11}x_1 + \ldots + a_{1n}x_n &< b_1 \\ &\vdots \\ a_{p1}x_1 + \ldots + a_{pn}x_n &< b_p \ . \end{aligned}$$

This can be written in matrix notation as

$$[a]\ x] < b] \ . \tag{3.8}$$

(Throughout this paper, when matrix notation is used, x] will correspond to a column vector, and $\underline{x}$, to its transpose, a row vector.)

Let $\ell]$ and m] be points satisfying (3.8), that is, they are members of the pattern class defined by (3.8):

$$[a]\ \ell] < b]$$
$$[a]\ m] < b] \ .$$

- -

*A convex set is one in which a line segment joining any two points belonging to the set is contained within the set.

Let q] be a point on the line segment joining ℓ] and m]:

$$q] = w\ \ell] + (1-w)\ m]$$

where $0 \leq w \leq 1$. Then

$$[a]\ q] = w[a]\ \ell] + (1-w)[a]\ m] < w\ b] + (1-w)\ b] = b]\ .$$

Therefore q] also lies in the domain defined by (3.8). Since q] may be any point on the line segment connecting ℓ] and m], then this line segment also lies in that domain. Therefore, the line segment joining any two points in the domain defined by (3.8) lies completely within that domain, and the domain is thus convex.

The suggestion is sometimes made that perhaps a linear transformation on the measurement space may group like patterns closer together and separate unlike patterns, so that a linear decision function may perform better under the transformation than otherwise. That this is an invalid suggestion is demonstrated by the next theorem. We will first need to prove the following lemma.

Lemma: The relative distances of any set of points to a linear boundary are invariant under any nonsingular affine transformation.*

---

*A nonsingular affine transformation is a nonsingular linear transformation followed by a translation.

That is, if s_1 and s_2 are the distances of points m_1 and m_2 to a hyperplane, and s_1' and s_2' are the distances of the images of these points, respectively, to the image of this hyperplane under a nonsingular affine transformation, then

$$\frac{S_1}{S_2} = \frac{S_1'}{S_2'} .$$

Proof of lemma: Let M be the measurement space, and let B be a hyperplane in M. B is given by the equation

$$\langle \alpha, x] + \alpha_o = 0 .$$

Consider the affine transformation

$$x'] = [U]\, x] + t]$$

or

$$x] = [U]^{-1}x'] - [U]^{-1}t]$$

where [U] is nonsingular, and t] represents a translation of origin. Then the equation for the image plane B′ in its normalized form is

$$\frac{\langle \alpha, [U]^{-1}x'] - \langle \alpha, [U]^{-1}t] + \alpha_o}{k} = 0 ,$$

where k is the magnitude of the vector $\langle \alpha, [U]^{-1}$.

The image of a point m] is

$$m'] = [U]\, m] + t] .$$

The distance s of m] from B is

$$s = \underline{\alpha}\, m] + \alpha_o \ .$$

The distance s' of m'] from B' is

$$s' = \frac{\underline{\alpha}\, [U]^{-1}[U]\ m] + \underline{\alpha}\, [U]^{-1}t] - \underline{\alpha}\, [U]^{-1}t] + \alpha_o}{k} \ ,$$

or

$$s' = \frac{\underline{\alpha}\, m] + \alpha_o}{k} = \frac{s}{k} \ .$$

Since k is a constant independent of the point m], the lemma is proved.

Theorem 5: The categorization defined by a linear decision function is invariant under a nonsingular affine transformation on the measurement space.

Proof: From the preceding lemma, any two points which were initially on the same side of a hyperplane before the transformation will still be so after the transformation, thus proving the theorem.

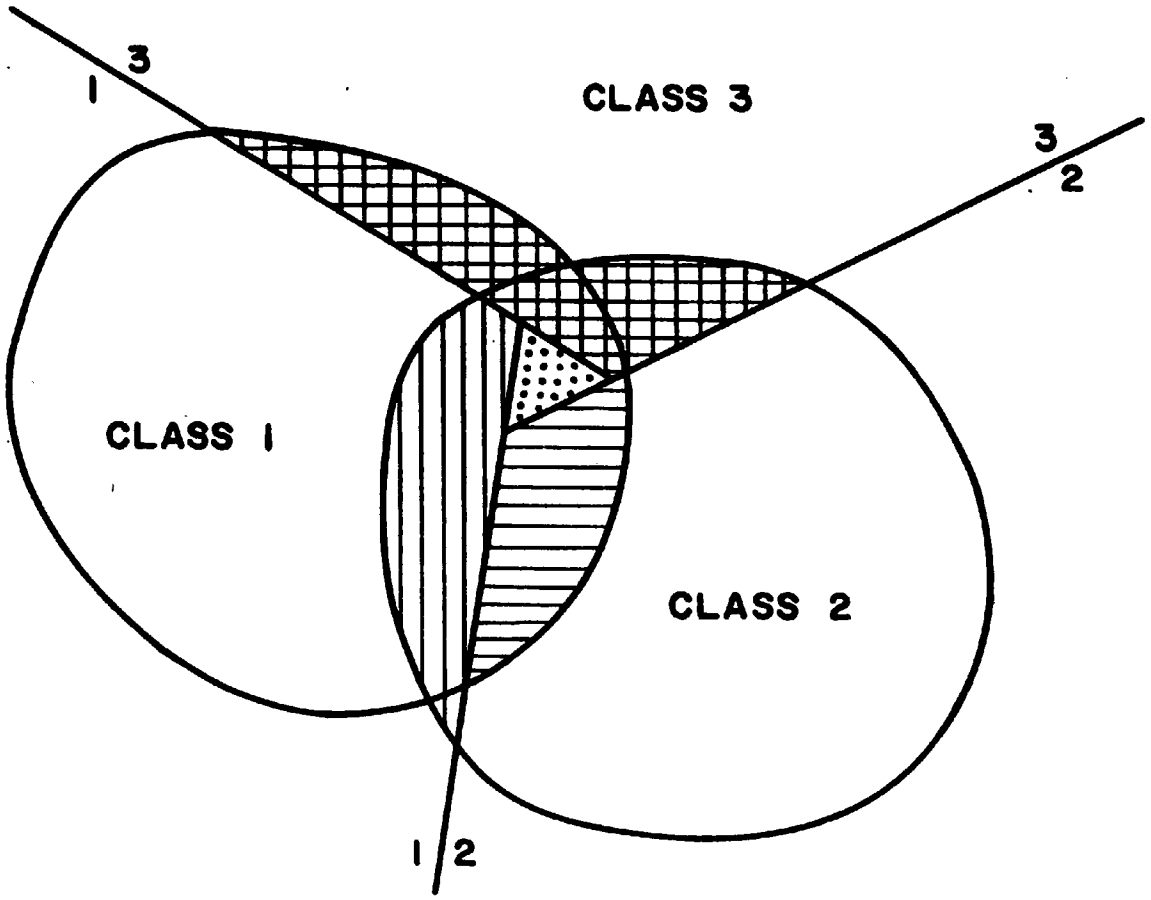
CHAPTER IV

THE SEQUENTIAL SYNTHESIS OF A LINEAR DECISION FUNCTION

4.1 Justification of Sequential Synthesis

The complete and accurate determination of a linear decision function requires the simultaneous determination of the several hyperplanes defining it. To see this more clearly, consider Figure 7 in which a linear decision function categorizing three classes in a measurement space is illustrated. Let the closed curves shown in this figure represent, for purposes of discussion, the domains in measurement space of classes 1 and 2. In general, the losses associated with the various possibilities for misrecognition or rejection are different. Therefore the boundary B_{12}, for instance, must be chosen so as to minimize the loss (given by equation (2.2)) associated with various factors, such as:

a. The misclassification of members of class 1 into class 2 (the horizontally hatched area);
b. the misclassification of members of class 2 into class 1 (the vertically hatched area);
c. the misclassification of members of other classes into class 1;
d. the misclassification of members of other classes into class 2;



ILLUSTRATING THE REQUIREMENT
OF SIMULTANEOUS SYNTHESIS

FIGURE 7

e. the rejection of members of various classes (the dotted area).

Note that the members of classes 1 and 2 which are already misclassified into other classes (in this case, into class 3 as illustrated by the cross-hatched area in Figure 7) are not to be considered in the determination of the optimum B_{12}; these are members which are going to be misclassified anyway, regardless of the position of B_{12}. Therefore, in order to optimize B_{12}, the other boundaries, B_{13} and B_{23} in this case, must be known. But their determination also depends on B_{12}, by the same argument. Therefore, all of the boundaries comprising an optimum linear decision function must be determined simultaneously.

However, for a moderate number of allowable pattern classes, n, the number of hyperplanes, $n(n-1)/2$, comprising a complete linear decision function becomes large and the problem might easily become unmanageable. It would certainly be a more palatable procedure if each hyperplane could be determined independently of the others. In particular, consider a suboptimum linear decision function defined by a set of hyperplanes, one for each pair of the allowable pattern classes, in which each hyperplane is determined by minimizing the loss associated with the total confusion between the two particular classes which it separates. That this is usually a good approximation to the optimum linear decision function is shown by the following argument.

Consider an optimum complete linear decision function L*, defined by boundaries denoted by B^*_{ij}, which is based on a loss function such that all losses due to misrecognition are equal to c, and all losses due to rejection are equal to $c_o < c$. L* has an expected loss C(L*).

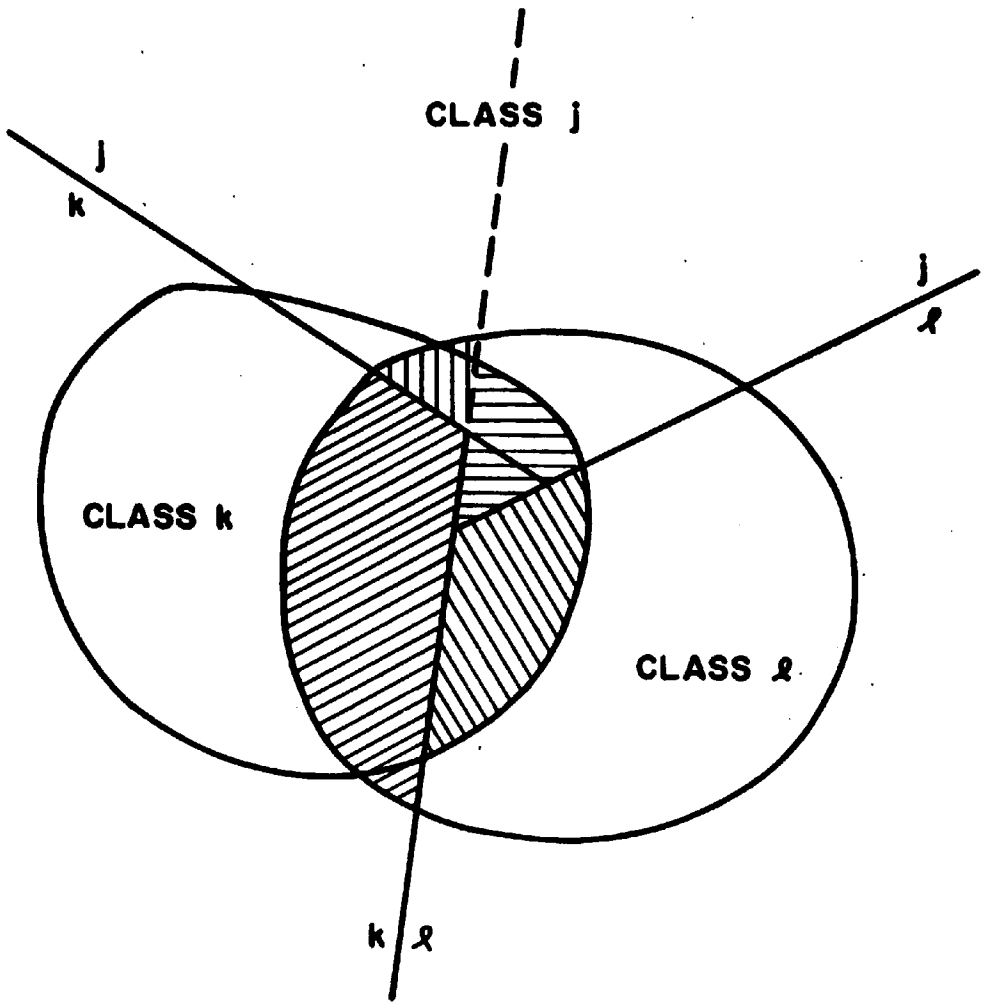
Let one of the boundaries of L*, say $B^*_{k\ell}$, be replaced by a suboptimum boundary $B_{k\ell}$ which has been determined such that it minimizes the expected loss associated with the total confusion between classes k and ℓ. This suboptimum linear decision function will be designated L, and will have an expected loss $C(L) \geq C(L^*)$.

Let us define some special quantities for L; similar definitions will hold for L*. First, let $L_j = 1$ for all points in M that are identified with class j, $0 \leq j \leq p$; $L_j = 0$ otherwise.

The expected loss associated with only the confusion between classes k and ℓ under L may be denoted $C_{k\ell}$ and is

$$C_{k\ell} = c \int_M [\omega_k L_\ell \beta(m \mid s_k) + \omega_\ell L_k \beta(m \mid s_\ell)]dm \ .$$

This is illustrated in Figure 8, in which is shown a linear decision function similar to that of Figure 7. This may be considered as representing either L or L*. The expected loss, $C_{k\ell}$, is that associated with the actual confusion between classes k and ℓ as illustrated by the diagonal hatching in Figure 8.



THE VARIOUS LOSSES ASSOCIATED
WITH TWO CLASSES

FIGURE 8

Note that, if $B_{k\ell}$ were to be extended through all space, as shown by the dotted line in Figure 8, then one could talk about the expected loss $P_{k\ell}$ associated with the confusion between the classes k and ℓ outside of the regions $L_k = 1$ and $L_\ell = 1$, as indicated by the horizontal and vertical hatching in Figure 8:

$$P_{k\ell} = c \int_M \sum_{\substack{j=0 \\ j \neq k, \ell}}^{p} L_j [\omega_k \beta(m \mid s_k) + \omega_\ell \beta(m \mid s_\ell)] dm \ .$$

Let us also define a quantity $R_{k\ell}$ which is the loss under L associated with all rejection regions for which $B_{k\ell}$ is one of the boundaries:

$$R_{k\ell} = c_o \int_M L_o(k\ell) [\omega_k \beta(m \mid s_k) + \omega_\ell \beta(m \mid s_\ell)] dm \ ,$$

where $L_o(k\ell) = 1$ for all rejection regions for which $B_{k\ell}$ is a boundary, and is zero otherwise. In Figure 8, this integration is over the one and only rejection region.

Now, the suboptimum boundary $B_{k\ell}$ was determined by minimizing the total expected loss, $C_{k\ell} + P_{k\ell}$, associated with the confusion between classes k and ℓ. Hence,

$$C_{k\ell} + P_{k\ell} \leq C^*_{k\ell} + P^*_{k\ell} \ . \qquad (4.1)$$

The changes in the expected loss for the system when $B_{k\ell}$ is used in place of $B^*_{k\ell}$ are due to the additional confusion between the classes k and ℓ, and to changes in rejection regions bounded by $B_{k\ell}$. Therefore

$$C(L) - C(L^*) = (C_{k\ell} - C^*_{k\ell}) + (R_{k\ell} - R^*_{k\ell}) \geq 0 .$$

Then, from (4.1),

$$C(L) - C(L^*) \leq (P^*_{k\ell} - R^*_{k\ell}) - (P_{k\ell} - R_{k\ell}) \geq 0 . \quad (4.2)$$

Note that $P_{k\ell} > R_{k\ell}$ and $P^*_{k\ell} > R^*_{k\ell}$ since the subspace over which $R_{k\ell}$ is integrated is contained within the subspace over which $P_{k\ell}$ is integrated.

$P_{k\ell}$ (or $P^*_{k\ell}$) represents only a portion of the sum of the expected losses for classes k and ℓ under L (or L*). It does not include $C_{k\ell}$ (or $C^*_{k\ell}$), nor does it include the misidentified members of classes k and ℓ which are on the correct side of the extended boundary $B_{k\ell}$ (or $B^*_{k\ell}$). $P_{k\ell}$ (or $P^*_{k\ell}$) is further reduced in (4.2) by $R_{k\ell}$ (or $R^*_{k\ell}$). Therefore, from (4.2), the degradation C(L) - C(L*) is less than the difference between two numbers, each of which are smaller than the sum of the expected losses for the classes k and ℓ under L*.

Hence one can expect, with reasonable confidence, that the increase in the expected loss when an optimum hyperplane is replaced by a suboptimum hyperplane will be very much

less than the sum of the expected losses of the two classes in question under the optimum linear decision function.

Consequently, it may be concluded that the independent (or sequential) determination of the hyperplanes rather than their simultaneous determination is a useful approximation to the optimum linear decision function. The degradation in system performance will be small, the savings in computational effort great.

Even if it were deemed that this approximation is not good enough, the concept of sequential determination is still valid, for the approximation may be made better by an iterative process. First determine the hyperplanes independently, giving an initial linear decision function L_1. Then only those members of each class which are correctly recognized by L_1 are used to recompute independently the hyperplanes, giving another linear decision function L_2. This process can be repeated until no significant improvement in performance is observed.

It will be noted that this argument assumed constant costs for misrecognition and rejection. If this restriction is dropped, the situation becomes more complicated, since all the sources of loss as mentioned earlier in the chapter must now be considered. However, the general result will still obtain - that is, the approximation is still a useful one. Therefore, the remainder of this paper will have its emphasis placed upon this sort of approximation to the optimum linear decision function.

4.2 Upper Bound on the Expected System Loss, as Determined from the Constituent Hyperplanes

When one has determined a hyperplane, B_{ij}, one can associate with it an expected loss, $C_{ij}(B_{ij})$, depending upon its performance in separating the two classes i and j, upon the loss coefficients c_{ij} and c_{ji} associated respectively with confusing the i^{th} class with the j^{th} class and vice versa, and upon the a priori probabilities ω_i and ω_j of occurrence of the classes i and j:

$$C_{ij}(B_{ij}) = \omega_i c_{ij} \int_{H_j(B_{ij})} \beta(m \mid s_i)dm + \omega_j c_{ji} \int_{H_i(B_{ij})} \beta(m \mid s_j)dm , \tag{4.3}$$

where $\int_{H_i(B_{ij})} \ldots dm$ indicates integration over the half space which includes all points identified as class i by B_{ij}. It will be of interest later to relate the expected loss for the hyperplanes to the expected loss for the system; this relation is given by Theorem 6.

Theorem 6: The expected loss associated with a linear decision function is not greater than the sum of the expected losses associated with its constituent hyperplanes.

Proof: Consider a particular class, say i. The expected loss associated with members of this class under the linear decision function L is

$$C_i(L) = \omega_i \sum_{j=0}^{p} c_{ij} \int_{R_j} \beta(m \mid s_i)dm \ ,$$

where $c_{ii} = 0$, and $\int_{R_j} \ldots dm$ indicates integration over that region in measurement space for which $L_j = 1$. R_o denotes the rejection regions. The expected loss for the system is

$$C(L) = \sum_{i=1}^{p} \sum_{j=0}^{p} \omega_i c_{ij} \int_{R_j} \beta(m \mid s_i)dm \ . \qquad (4.4)$$

Denote the sum of the expected losses for each of the constituent hyperplanes of L by $C_B(L)$; it may be written, from (4.3),

$$C_B(L) = \sum_{i=1}^{p} \sum_{j=i}^{p} \omega_i c_{ij} \int_{H_j(B_{ij})} \beta(m \mid s_i)dm$$

$$+ \sum_{i=1}^{p} \sum_{j=i}^{p} \omega_j c_{ji} \int_{H_i(B_{ij})} \beta(m \mid s_j)dm$$

$$= \sum_{i=1}^{p} \sum_{j=1}^{p} \omega_i c_{ij} \int_{H_j(B_{ij})} \beta(m \mid s_i)dm \ .$$

Denote by $\int_{S_i(B_{ij})} \ldots dm$, $1 \leq j \leq p$, the integral over that subspace which includes all points which are identified as class i by B_{ij}, but which are not rejected by L. Denote

by $\int_{S_o(B_{io})} \ldots dm$ the integral over that subspace which includes all points rejected by L. Then if $c_{ij} > c_{io}$, one can write

$$C_B(L) \geq \sum_{i=1}^{p} \sum_{j=0}^{p} \omega_i c_{ij} \int_{S_j(B_{ij})} \beta(m \mid s_i) dm \ . \qquad (4.5)$$

But the subspace R_j, $j \neq 0$, is contained within the subspace $S_j(B_{ij})$, since R_j is bounded by all the hyperplanes B_{kj}, $1 \leq k \leq p$, and does not contain any rejection regions, whereas $S_j(B_{ij})$ is bounded only by the hyperplane B_{ij} and also contains no rejection regions. Also $R_o = S_o(B_{io})$ by definition. Therefore

$$\int_{R_j} \beta(m \mid s_i) dm \leq \int_{S_j(B_{ij})} \beta(m \mid s_i) dm \ , \qquad 1 \leq j \leq p$$

$$\int_{R_o} \beta(m \mid s_i) dm = \int_{S_o(B_{io})} \beta(m \mid s_i) dm \ .$$

Hence, C(L), given by (4.4) is not greater than the right-hand side of (4.5), thus proving the theorem.

A useful corollary is immediately obvious.

Corollary: If the expected loss for each of the constituent hyperplanes of a linear decision function L is zero, then the expected loss for L is also zero.

4.3 Some Special Cases of Optimum Hyperplanes

Before discussing some general procedures useful for determining the optimum hyperplane separating two classes, it is interesting to note some special cases.

Theorem 7: The optimum decision function δ^*, not containing a rejection decision, for the case of two classes which

a. are equally probable *a priori*,

b. have identical losses associated with misrecognition,

c. have probability distributions over the measurement space which are unimodal, spherically symmetrical, and identical except for a displacement of modes,

is a linear decision function comprising a hyperplane which is the perpendicular bisector of the line segment joining the two modes.

Proof: Let the two classes be denoted class 1 and class 2. Let the probability distribution for class 1 over the measurement space be designated $\beta(r_1)$, where r_1 is the distance of a point r from the mode of the distribution for class 1. Similarly for class 2. From the optimum decision rule given by equation (2.4), δ^* is the following:

$$\delta^*(d_1 \mid m) = 1$$

$$\delta^*(d_2 \mid m) = 0$$

if

$$\beta(r_1) > \beta(r_2) \ ;$$

$$\delta^*(d_1 \mid m) = 0$$

$$\delta^*(d_2 \mid m) = 1$$

if

$$\beta(r_1) < \beta(r_2) \ .$$

Therefore, the boundary equivalent to δ^* in the measurement space is that locus of points satisfying

$$\beta(r_1) = \beta(r_2) \ .$$

Since β is a monotone descending function of r, this implies

$$r_1 = r_2 \ .$$

That is, the boundary equivalent to δ^* is the locus of those points which are equidistant from the modes of the distributions of classes 1 and 2. But this locus is the hyperplane which is the perpendicular bisector of the line segment joining the modes.

An example of such a case is when the two classes are described by Gaussian distributions which have the same variance for each variate, and which have zero covariances. The general solution for the optimum decision function for the Gaussian case is well known, [2,31] and the results are repeated here in terms of the equivalent boundary. (Equal costs of misrecognition and equal <u>a priori</u> probabilities are assumed.)

Let two classes, i and j, be described in measurement space by Gaussian distribution functions. Let $u_i]$ be the (n-dimensional) mean of the distribution of class i, and $[V_i]$ be its covariance matrix. Then the boundary equivalent to the optimum linear decision function is the set of points $\underline{x}$ such that

$$- \frac{1}{2} \underline{x} \left([V_i]^{-1} - [V_j]^{-1} \right) x] + \underline{x} \left([V_i]^{-1} u_i] - [V_j]^{-1} u_j] \right)$$

$$- \frac{1}{2} \underline{u_i} [V_i]^{-1} u_i] + \frac{1}{2} \underline{u_j} [V_j]^{-1} u_j] + \ln k_{ij} = 0 , \qquad (4.6)$$

where

$$k_{ij}^2 = \frac{|V_j|}{|V_i|} .$$

Note that this is a quadratic function. However, if the two covariance matrices are equal, then the boundary becomes linear. This is stated in the following theorem:

Theorem 8: The optimum decision function δ^*, not containing a rejection decision, for the case of two classes which

a. are equally probable *a priori*,

b. have identical losses associated with misrecognition,

c. have probability distributions over the measurement space which are Gaussian and which have equal covariance matrices [V],

is a linear decision function comprising a hyperplane given by the set of all points $\underline{x}$ satisfying

$$\underline{x}\,[V]^{-1}(u_i-u_j)] - \frac{1}{2}\,\underline{(u_i+u_j)}\,[V]^{-1}(u_i-u_j)] = 0 \ ,$$

where u_i] is the mean of the distribution for class i.

In a good part of the work to follow, we will be interested in estimating the optimum linear boundary between two classes based upon a sampling from the two classes. Geometrically, this can be interpreted as passing a hyperplane through two sets of points so as to optimally separate them in some sense. We will say that two sets of points in measurement space are linearly separable if they can be separated perfectly by a hyperplane. Following are two theorems dealing with linear separability. (The problem of linear separability of points described by Boolean functions has received significant attention by other workers.[33,34, 52,53,35])

Theorem 9: Two sets of points are linearly separable if and only if their convex hulls* are nonintersecting.

- -

*The convex hull of a set of points S is the smallest convex set containing S. It is the set of all convex combinations of sets of points from S.[22]

Proof: It is proved in set theory that two non-intersecting convex sets may be separated by a hyperplane.[49] This proves the "if" part of the theorem. To prove the "only if", assume that the convex hulls of the sets S_i and S_j intersect. There then exists at least one point p which is a convex combination of sets of points from S_i and also from S_j. It was shown in the proof of Theorem 4 that if a set of points satisfied a family of linear inequalities (in this case, only one such inequality), then all convex combinations of these points also obeyed that family of linear inequalities. If there exists a hyperplane, B_{ij}, separating S_i and S_j, then, since p is a convex combination of a set of points from S_i, p must be on the i side of B_{ij}. By a similar argument, it must also be on the j side of B_{ij}. This is contradictory, and therefore S_i and S_j are not linearly separable.

The following corollary is immediately obvious from the proof of Theorem 9.

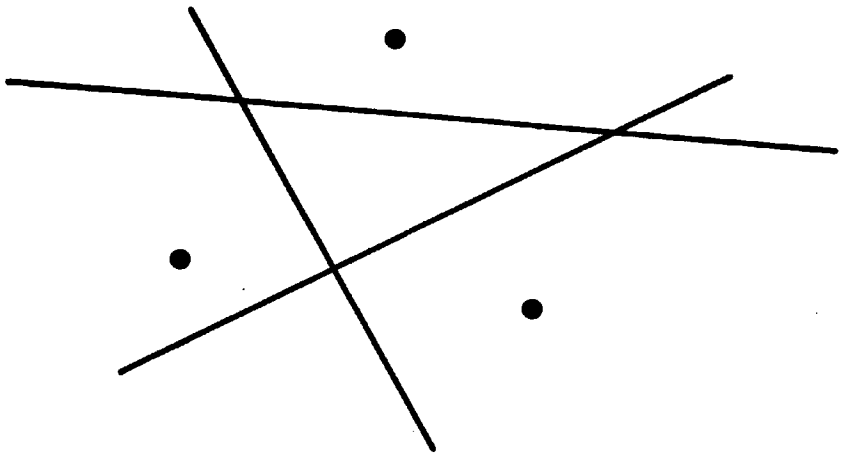
Corollary: If there exists at least one point which is a convex combination of points from the set S_i and which is also a convex combination of points from S_j, then S_i and S_j are not linearly separable.

An algebraic test for linear separability can be obtained from this corollary which leads to a set of n linear equations in p unknowns, where n is the dimensionality of the space and p is the total number of points in S_i and S_j. If these equations have a solution, then the sets of points are

not linearly separable. However, the difficulty of determining the existence of a solution to this set is in general more difficult than actually trying to find a hyperplane which will separate the points (the procedure for which will be discussed in Section 5.3). Furthermore, if the sets are not linearly separable, this corollary will not indicate to what degree this is true; i.e., perhaps the sets can be separated with only small error. The empirically determined hyperplane will give an estimate of the degree of separation of the pattern classes from which these sample points were drawn (as will be discussed in Section 8.2). Therefore this corollary will be pursued no further.

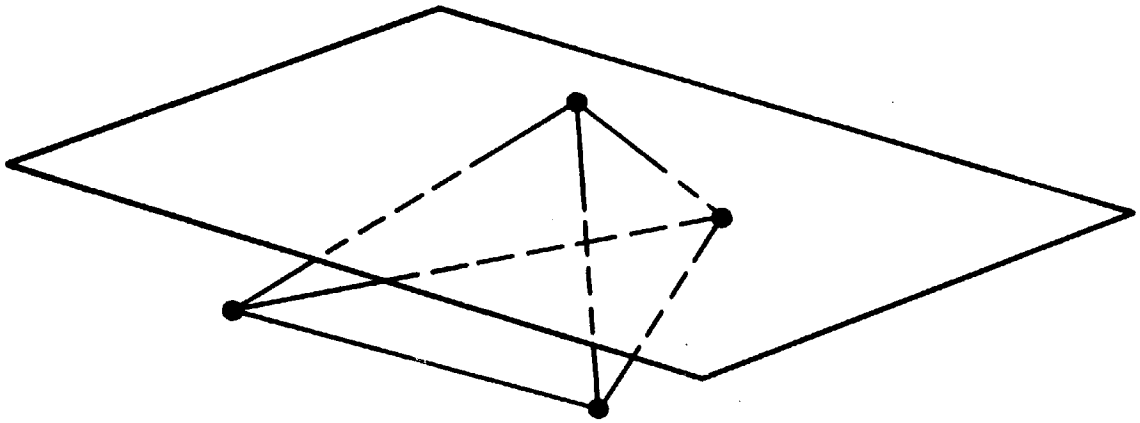
Let us say that a set of q points in a space of n dimensions, $q \leq n+1$, is nondegenerate if the points cannot be contained in a linear subspace of q-1 dimensions. In Figure 9 are shown 3 nondegenerate points in two dimensions, and 4 nondegenerate points in 3 dimensions. Note that, in each case, the points can be separated into any two categories desired by an n-dimensional hyperplane. This is generalized in the next theorem.

Theorem 10: Let S be a set of q nondegenerate points in an n-dimensional space, $q \leq n+1$. Let S_1 consist of any k of these points, and S_2 consist of the remaining q-k points. Then S_1 and S_2 are linearly separable.



TWO DIMENSIONS

(a)



THREE DIMENSIONS

(b)

LINEAR SEPARABILITY OF NONDEGENERATE POINTS

FIGURE 9

Proof: It is obvious from Figure 9 that the theorem is true for n = 1,2, and 3. Let us assume that it is true for n-1 dimensions and prove that it then holds for n-dimensions. If $q \leq n$, the theorem follows immediately, since the n-dimensional hyperplane need only contain the (n-1) dimensional hyperplane which properly separated the two sets S_1 and S_2. The extreme case then consists of a set S of n nondegenerate points in n-1 dimensions which is separated into two sets, S_1 and S_2, by a hyperplane, say B_{n-1}. Let us now add one dimension to the space and one point, p, to S, such that $S' = S \cup p$ is still nondegenerate, i.e., p is not contained in the n-dimensional hyperplane containing S and B_{n-1}. Then an n-dimensional hyperplane B_n may be passed through B_{n-1} such that p falls on either side of it. Since it was assumed that B_{n-1} could be chosen to separate S arbitrarily, then B_n can separate S' arbitrarily, thus proving the theorem.

This theorem and the two following corollaries will be important later in the discussion of the practical interpretation and use of linear decision functions.

Corollary 1: Let S be a set of q points in an n-dimensional space such that any subset of S containing no more than n+1 points is nondegenerate. Let $q \leq n+m-1$. Then S can be separated into m nonempty sets by a linear decision function.

Proof: By the corollary to Theorem 6, it is only necessary to show that each of the m sets is linearly separable. Let S be separated into the sets $S_1,\ldots,S_m$. Consider the case in which $q = n+m-1$, and the sets $S_1,\ldots,S_{m-1}$ each contain one point from S, leaving n points from S to comprise S_m. Then S_m and S_k, $1 \leq k \leq m-1$, are linearly separable by Theorem 10, since their union contains n+1 points. In any other possible case, the number of points contained in the union of any two sets S_i and S_j will be less than n+1, thus proving the corollary.

Corollary 2: Let S be a set of q points in an n-dimensional space such that any subset of S containing no more than n+1 points is nondegenerate. Let

$$q \leq \frac{mn}{2}, \qquad n \text{ even}$$

$$q \leq \frac{m(n+1)}{2}, \qquad n \text{ odd}.$$

Then S can be separated into m subsets of equal size by a linear decision function.

Proof: The union of any two subsets will contain n nondegenerate points if n is even, n+1 nondegenerate points if n is odd. Therefore, each pair of subsets is linearly separable by Theorem 10, and the corollary is then proved by invoking the corollary to Theorem 6.

CHAPTER V

DETERMINATION OF THE OPTIMUM LINEAR BOUNDARY SEPARATING TWO CLASSES

This chapter will deal with the problem of determining the optimum (minimum loss) hyperplane which separates a pair of classes. In the general case, which is treated here, the loss associated with misrecognition of a member from one class is not the same as that loss for the other class. Recall from Chapter II, however, that when the losses are equal, then minimum loss corresponds to minimum error.

Two cases will be discussed. In the first, it is assumed that the pertinent conditional probability functions over measurement space, $\beta(m \mid s_i)$, are continuous, and that these probabilities and the *a priori* probabilities of occurrence, ω_i, are known. In the second case, which is the case of practical interest, it is assumed that nothing is known about the probabilities $\beta(m \mid s_i)$, and that the *a priori* probabilities ω_i may or may not be known. The determination of the optimum hyperplane is then based upon an appropriate sampling from the pattern classes.

5.1 The Optimum Hyperplane for the Case of Known Distributions

Let $\beta(m \mid s_i)$ be the probability density function of class i over the measurement space, ω_i be the *a priori* probability of occurrence of class i, and c_{ij} be the loss

associated with misidentifying a member of class i with class j. Denote a hyperplane which separates the classes i and j by B_{ij}, and let it be defined in the coordinate system $(x_1,\ldots,x_n)$ by the equation

$$x_1 = \sum_{k=2}^{n} \alpha_k x_k + \alpha_o \ . \tag{5.1}$$

Let $\nu(m \mid s_i, B_{ij})$ be the conditional probability density function of class i over the boundary B_{ij}:

$$\nu(m \mid s_i, B_{ij}) = \frac{\beta(m \mid s_i)}{\int_{B_{ij}} \beta(m \mid s_i)dm}$$

for all m satisfying (5.1). $\int_{B_{ij}} \ldots dm$ denotes integration over the boundary B_{ij}. Define the weighted conditional probability density function of class i over the boundary B_{ij} by

$$\tau(m \mid s_i, B_{ij}) = c_{ij}\omega_i \nu(m \mid s_i, B_{ij}) \ .$$

Theorem 11: The optimum linear boundary B_{ij}, separating two classes i and j which have weighted conditional probability density functions over B_{ij} given by

$$\tau_i = \tau(m \mid s_i, B_{ij})$$

$$\tau_j = \tau(m \mid s_j, B_{ij}) \ ,$$

must satisfy the following conditions:

a. The integrals of τ_i and τ_j over B_{ij} must be equal.

b. The means of τ_i and τ_j must be equal.

Proof: Let B_{ij} be oriented such that the half-space identified as class i corresponds to

$$x_1 < \sum_{k=2}^{n} \alpha_k x_k + \alpha_o \ .$$

The expected loss is then

$$C(B_{ij}) = c_{ij}\omega_i \int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{\infty} dx_2 \int_{-\infty}^{\sum_{k=2}^{n} \alpha_k x_k + \alpha_o} \beta(m \mid s_i) dx_1$$

$$+ c_{ji}\omega_j \int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{\infty} dx_2 \int_{\sum_{k=2}^{n} \alpha_k x_k + \alpha_o}^{\infty} \beta(m \mid s_j) dx_1 \ . \tag{5.2}$$

We wish to find the coefficients of the hyperplane B_{ij} which correspond to extreme points of (5.2). First differentiate (5.2) with respect to α_o:

$$\frac{\partial C(B_{ij})}{\partial \alpha_o} = - \int_{B_{ij}} \tau_i dm + \int_{B_{ij}} \tau_j dm = 0 \ ,$$

which is condition a. of the theorem. Next, differentiate (5.2) with respect to α_k, $2 \leq k \leq n$:

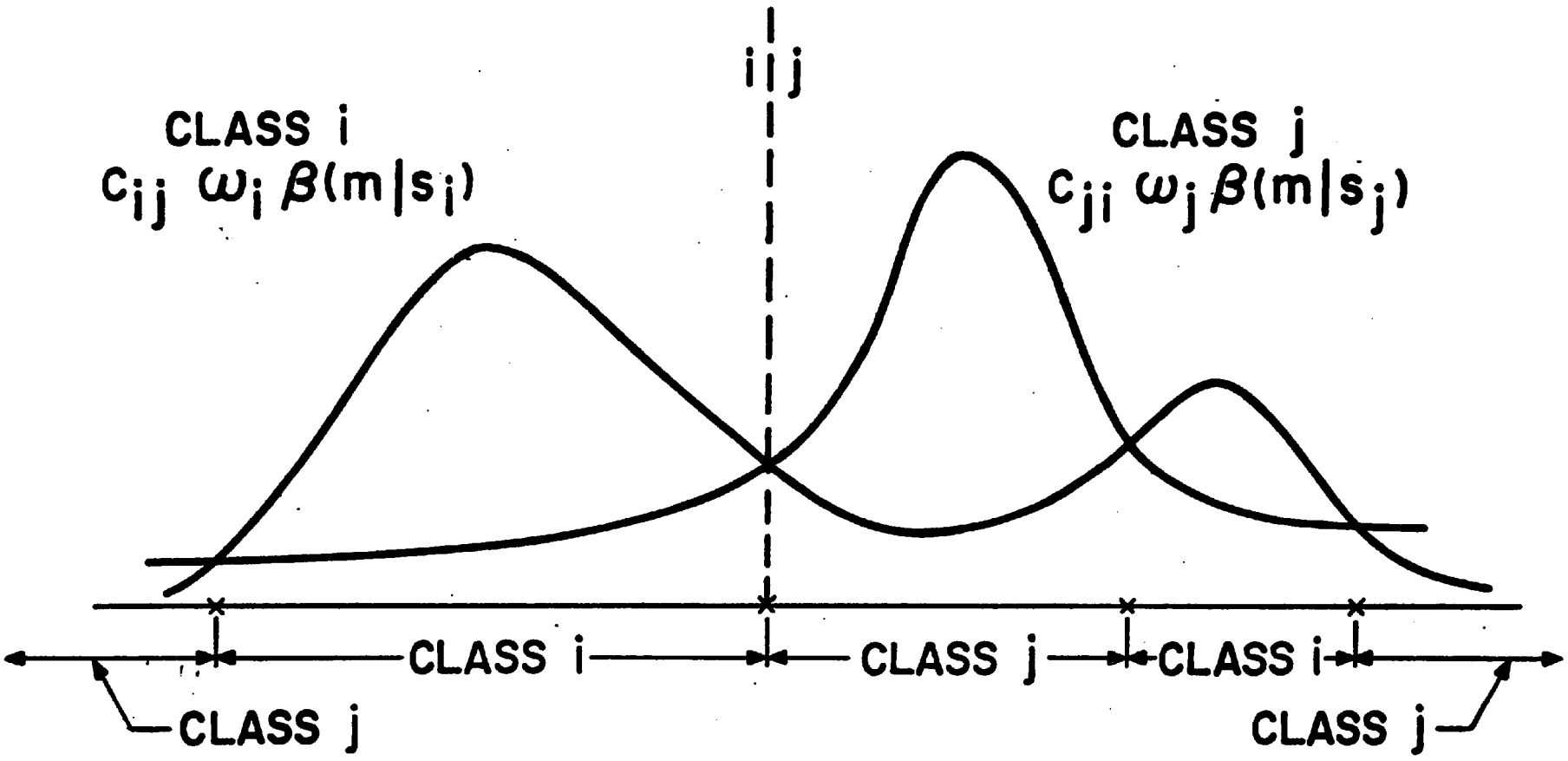
$$\frac{\partial C(B_{ij})}{\partial \alpha_k} = - \int_{B_{ij}} x_k \tau_i dm + \int_{B_{ij}} x_k \tau_j dm = 0,$$

$$2 \leq k \leq n \ .$$

A similar expression may be obtained for $k = 1$ by rewriting (5.1) in terms of some other coordinate. This set of conditions, i.e., for $1 \leq k \leq n$, corresponds to condition b. of the theorem.

In general, there will be several hyperplanes satisfying the conditions of Theorem 11. Some of these will correspond to maxima of $C(B_{ij})$, others to minima. These must then be searched to determine which corresponds to the absolute minimum of $C(B_{ij})$.

It might also be pointed out that this theorem allows an approach to linear decision functions using more than one hyperplane per pair of classes by combining, with the appropriate logic, the various hyperplanes satisfying the conditions of Theorem 11. As an example of this, consider Figure 10 in which a pair of one-dimensional functions is shown corresponding



MULTIPLE BOUNDARY LINEAR DECISION FUNCTION

FIGURE 10

to the weighted probability density functions $c_{ij}\omega_i\beta(m \mid s_i)$. The boundaries satisfying the conditions of Theorem 11 in one dimension are those points in Figure 10 where the distributions are equal (denoted by an x). Let all of these boundaries be combined, logically, into a linear decision function such that the categorization is as shown by the arrows. Then this also corresponds to the optimum decision function, δ^*, given by (2.3). The linear decision function which would have been used under the constraint of one boundary per pair of classes is shown by the dotted line; its loss is significantly greater than the multiple boundary decision function.

5.2 The Estimated Optimum Hyperplane for the Case of Unknown Distributions

We will now assume that the designer has no knowledge concerning the form of the probability function $\beta(m \mid s_i)$, but he may or may not know the *a priori* probabilities, ω_i. We will assume the existence of all such probabilities and probability functions, whether known or not.

If a hyperplane B_{ij} is passed through M, such as to divide classes i and j in some fashion, then a certain portion of the members of classes i and j will be misidentified by B_{ij}. Let p_i be the probability of misidentification of a member from class i, given B_{ij} and a member of class i (p_i is the integral of $\beta(m \mid s_i)$ over the half-space on the

j side of B_{ij}). Then the conditional loss associated with B_{ij} (see equation (2.1)) is

$$\tilde{C}(B_{ij}) = c_{ij}\omega_i' p_i + c_{ji}\omega_j' p_j$$
$$= c_{ij}e_i + c_{ji}e_j \,, \qquad (5.3)$$

where $\omega_i' = \frac{\omega_i}{\omega_i+\omega_j}$ and $\omega_j' = \frac{\omega_j}{\omega_i+\omega_j}$, and $e_i = \omega_i' p_i$ is the probability of misrecognition, given B_{ij}, of a member from class i when patterns are chosen randomly from classes i and j according to ω_i' and ω_j'.

Theorem 12: Construct a hyperplane B_{ij} in measurement space M which divides M into two half spaces, all the points in one being identified as class i, the points in the other being identified as class j. Consider two sampling procedures designed to estimate the conditional cost $\tilde{C}(B_{ij})$:

a. The a priori probabilities ω_k are unknown. Let it be assumed that there exists a pattern source which will generate patterns from classes i and j randomly according to ω_i' and ω_j'. Draw a pattern from this source, identify it, and then determine the identification according to B_{ij}. This latter identification will either be in error or will be correct. Repeat this experiment n times. Let m_i be the number of samples from class i which are misidentified by B_{ij} as class j, and likewise for m_j.

b. The <u>a priori</u> probabilities ω_k are known. Take n_i samples from class i and n_j samples from class j such that

$$n_i = \omega_i' n \ ,$$
$$n_j = \omega_j' n \ , \qquad (5.4)$$

(It will be assumed that ω_i' and ω_j' are such that (5.4) can be met exactly.) Identify each of these n samples according to B_{ij}. Let m_i be the number of samples from class i misidentified by B_{ij} as class j, and likewise for m_j.

Then the maximum likelihood estimate in either case* for the conditional loss $\hat{C}(B_{ij})$ is

$$\hat{C}(B_{ij}) = \frac{c_{ij}m_i + c_{ji}m_j}{n} \qquad (5.5)$$

Proof: Take case a. first. Three events of interest can occur:

1. A pattern from class i is misidentified with probability e_i.
2. A pattern from class j is misidentified with probability e_j.
3. A pattern is recognized correctly with probability $1-e_i-e_j$.

- -

*We make no further distinction here between these two cases of sampling. They are compared further in Chapter VIII.

These events are discrete, independent, and mutually exclusive. Therefore, the probability distribution for m_i and m_j is multinomial:

$$P(m_i,m_j) = \binom{n}{m_i\ m_j} e_i^{m_i} e_j^{m_j} (1-e_i-e_j)^{n-m_i-m_j} .$$

The maximum likelihood estimate for e_i is [19]

$$\hat{e}_i = \frac{m_i}{n} ,$$

and likewise for e_j. Since the maximum likelihood estimate of a sum of independent variables is the sum of the maximum likelihood estimates, then from (5.3)

$$\hat{\tilde{C}}(B_{ij}) = \frac{c_{ij}m_i + c_{ji}m_j}{n} .$$

For the second case, the number of samples taken from each class is fixed by (5.4). By an argument similar to the preceding argument, m_i is binomially distributed with parameters n_i and p_i, and likewise for m_j. Then the probability function for m_i and m_j is given by

$$P(m_i,m_j) = \binom{n_i}{m_i} p_i^{m_i} (1-p_i)^{n_i-m_i} \binom{n_j}{m_j} p_j^{m_j} (1-p_j)^{n_j-m_j} .$$

Again, the maximum likelihood estimate for p_i is

$$\hat{p}_i = \frac{m_i}{n_i} ,$$

and likewise for p_j. Using the linear summation rule again with (5.3), one obtains

$$\hat{\tilde{C}}(B_{ij}) = \frac{c_{ij}\omega_i' m_i}{n_i} + \frac{c_{ji}\omega_j' m_j}{n_j} .$$

Making use of (5.4), this reduces to

$$\hat{\tilde{C}}(B_{ij}) = \frac{c_{ij}m_i + c_{ji}m_j}{n} .$$

If we take samples from a pair of classes according to either criterion, there will be a set of hyperplanes (infinite in number) which will minimize the maximum likelihood estimate of the conditional loss (5.5). It is quite reasonable, then, to choose one of the hyperplanes from this set as the estimate of the optimum hyperplane separating the two classes. That is, it is clear from (5.5) that we will search for a hyperplane which will minimize the loss associated with the sample points. This is also intuitively quite reasonable.

Note that Theorem 12 and the resulting procedure is independent of the probability functions over the measurement space. Hence one need make no assumptions concerning the form of these functions, nor need he concern himself with the dependencies between the various measurements.*

---

*Another criterion for determining the optimum hyperplane is discussed in Appendix III. It makes use of a rather weak assumption concerning the probability functions over the measurement space, and so is not completely nonparametric.

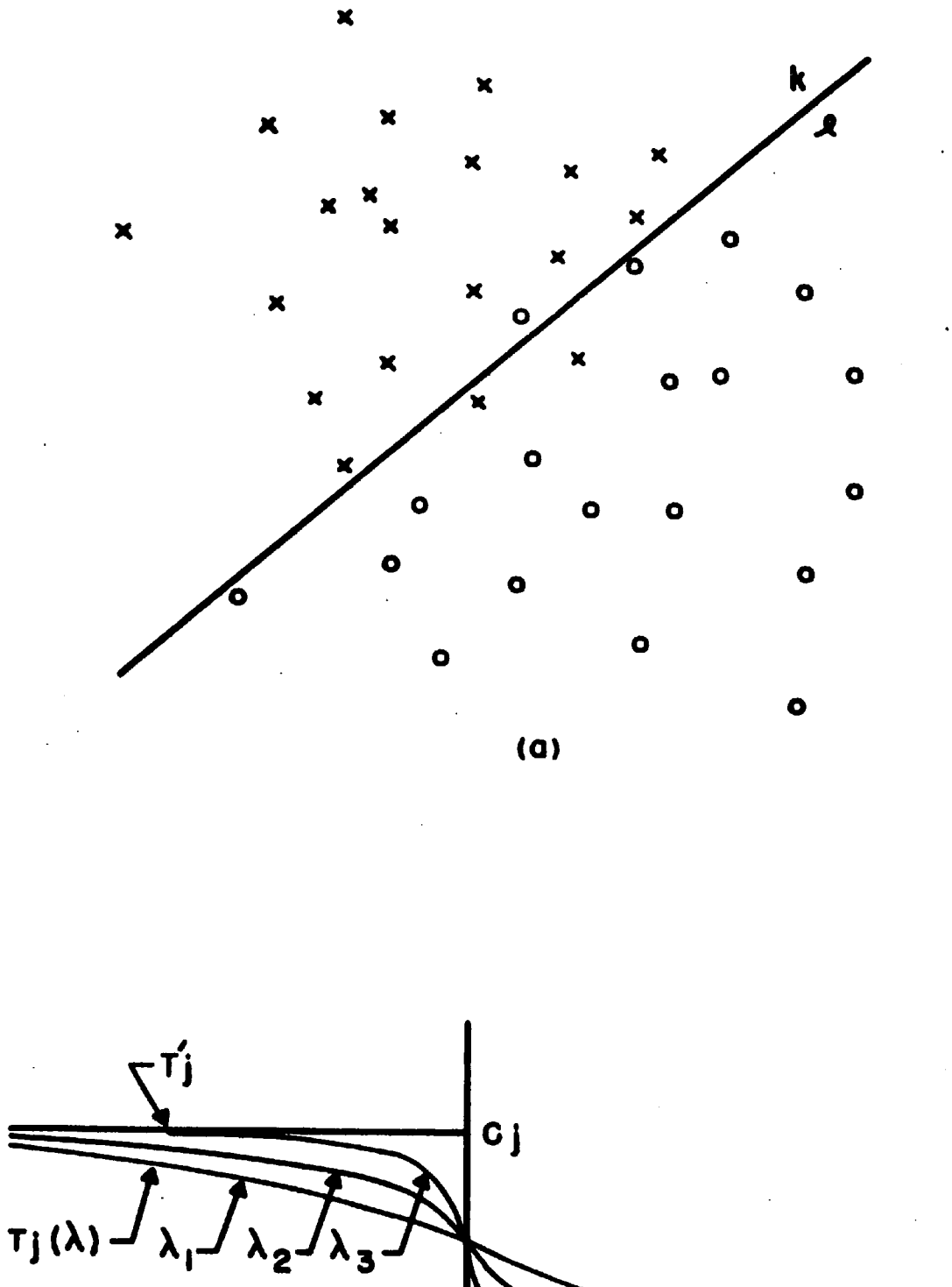
5.3 A Computation Algorithm for the Case of Unknown Distributions

In this section will be developed an iteration algorithm which will be useful for determining that boundary which minimizes the maximum likelihood estimate of the conditional loss for the boundary. There has been some work by others concerning similar boundaries when the measurement space is a binary space [33,34,52,53], or when the classes are Gaussian distributed in measurement space (discriminant functions [2,17,50] yield a good approximation for this case).

Figure 11a illustrates this problem for two classes, k and ℓ. Samples from class k are shown by crosses, from class ℓ by circles. A boundary, $B_{k\ell}$, is indicated. Let us number these samples from 1 to m, there being a total of m samples, and define a weight, T'_j, for the j^{th} sample point, $1 \leq j \leq m$, such that

$T'_j = 0$ if the point is on the correct side of $B_{k\ell}$;
$T'_j = c_{k\ell}$ if the point represents a sample from class k on the ℓ side of $B_{k\ell}$;
$T'_j = c_{\ell k}$ if the point represents a sample from class ℓ on the k side of $B_{k\ell}$.

It is clear then that minimizing the estimate of the conditional loss (5.5) is equivalent to minimizing



ILLUSTRATING THE ITERATIVE ALGORITHM

FIGURE 11

$$T'(\alpha_i) = \sum_{j=1}^{m} T_j' \qquad (5.6)$$

where the α_i, $0 \leq i \leq n$, are the coefficients of the hyperplane $B_{k\ell}$ defined by (3.1). Since there are (n+1) coefficients for an n-dimensional hyperplane, $T'(\alpha_i)$ can be interpreted as an (n+1)-dimensional function for which we want to find the absolute minimum. One powerful technique for doing this is the method of steepest descent [7], by which one makes an initial guess as to the solution, and then computes the gradient of the function at that point. A small step is taken in the direction of the negative gradient (i.e., "downhill"), and the gradient is recomputed at that point. This process is repeated, the steps being made smaller, until one is arbitrarily close to the minimum point.

One problem inherent in all gradient methods is that the minimum which is finally reached is not necessarily the absolute minimum. If one is in doubt as to whether the solution obtained actually represents an absolute minimum, he can only try other initial starting points and accept the smallest minimum which he obtains.

The problem which immediately arises in trying to apply this method to (5.6) is illustrated by Figure 11b, where the form of T_j' is shown. s_j is the distance of the j^{th} point from the boundary $B_{k\ell}$, and will be considered

to be positive if the point j is on the correct side of the boundary. That is, points representing members of class k will be a distance greater than zero from the boundary if they are on the k side of $B_{k\ell}$, and will be a distance less than zero from the boundary if they are on the ℓ side. (A technique for handling this mathematically will be introduced later.) The quantity c_j in Figure 11b is equal to $c_{k\ell}$ if the j^{th} point represents a member from class k; $c_j = c_{\ell k}$ otherwise. It is the cost of misrecognition for the j^{th} sample. As the α_i are adjusted, the hyperplane moves around in the measurement space, and $T'(\alpha_i)$ makes a discrete jump of magnitude c_j every time the hyperplane passes through a sample point. Otherwise $T'(\alpha_i)$ remains constant. Consequently it has no meaningful derivatives and no meaningful gradient at a point, and gradient methods such as the method of steepest descent are not applicable.

However, it is possible to approximate T_j' by some function $T_j(s_j,\lambda)$ which is continuous everywhere, and which has the property

$$\lim_{\lambda\to\infty} T_j(s_j,\lambda) = T_j' ,$$

where T_j is written as a function of the distance of the j^{th} point from the hyperplane to emphasize this continuous dependence. Such a function is shown in Figure 11b. If the function

$$T(\alpha_i,\lambda) = \sum_{j=1}^{m} T_j(s_j,\lambda) \qquad (5.7)$$

were to be minimized for some finite λ with respect to the α_i, and then λ increased and (5.7) minimized again, and this process repeated, one would expect the hyperplane to converge to one of the set of hyperplanes minimizing (5.6). This minimization process can now make use of the method of steepest descent, and will be developed in more detail below.

There are many functions which would be suitable for $T_j(\lambda)$. One convenient one is the cumulative Gaussian distribution with zero mean and standard deviation $1/\sqrt{2}\lambda$; it will be denoted $G(\lambda s_j)$. Then

$$T_j(s_j,\lambda) = c_j[1 - G(\lambda s_j)] ,$$

where

$$\frac{\partial[G(\lambda s_j)]}{\partial s_j} = \frac{\lambda}{\sqrt{\pi}} e^{-(\lambda s_j)^2} . \qquad (5.8)$$

Then

$$T(\alpha_i,\lambda) = \sum_{j=1}^{m} c_j[1 - G(\lambda s_j)].$$

The gradient of $T(\alpha_i,\lambda)$ is determined by the derivatives

$$\frac{\partial T(\alpha_i,\lambda)}{\partial \alpha_i} = - \sum_{j=1}^{m} c_j \frac{\partial [G(\lambda s_i)]}{\partial s_i} \frac{\partial s_j}{\partial \alpha_i} , \qquad 0 \leq i \leq n . \tag{5.9}$$

Before proceeding further, we must sidetrack and discuss the problem raised by the fact that s_j is defined as being positive if the j^{th} point is on the correct side of the hyperplane. Let m_{ij} be the i^{th} coordinate of the j^{th} point. Then, as developed in Chapter III, s_j is given by

$$s_j = \sum_{i=1}^{n} \alpha_i m_{ij} + \alpha_o .$$

Now the sign s_j can be chosen arbitrarily, since the coefficients of the hyperplane $B_{k\ell}$ may be all multiplied by (-1) without affecting the location of $B_{k\ell}$. Alternatively, s_j may be written

$$s_j = \sum_{i=0}^{n} \alpha_i m_{ij} , \tag{5.10}$$

where m_{oj} is called an <u>artificial coordinate</u> and is defined to be +1. Now it is seen that the sign of s_j may also be reversed by reversing the signs of all of the coordinates, m_{ij}, of the j^{th} point. This is the device that will be used here. It will be decided <u>a priori</u> which class, say k, will be positive, and the coordinates (including the artificial coordinate) of each sample from k will be entered

into the equation with true signs. All of the coordinates, including the artificial coordinate, for each sample from the other class will be entered with reverse sign. This will satisfy the condition on the sign of s_j, and allow the equations to be written more simply.

Keeping this artifice in mind, we now return to (5.9). From (5.10), we wish to compute $\partial s_j/\partial \alpha_i$. However, (5.10) holds only if the coefficients α_i are in normalized form, that is, that (3.2) holds:

$$\sum_{i=1}^{n} \alpha_i^2 = 1 \ . \tag{3.2}$$

Assume that (3.2) holds and write (5.10) as

$$s_j = \frac{\sum_{i=0}^{n} \alpha_i m_{ij}}{\left[\sum_{k=1}^{n} \alpha_k^2\right]^{1/2}} \ ,$$

which will guarantee that (3.2) will hold even if the α_i's are incremented. Then

$$\frac{\partial s_j}{\partial \alpha_i} = m_{ij}\left[\sum_{k=1}^{n} \alpha_k^2\right]^{-1/2} - \frac{1}{2}\left[\sum_{k=1}^{n} \alpha_k^2\right]^{-3/2}(2\alpha_i)\left[\sum_{i=0}^{n} \alpha_i m_{ij}\right], \quad 1 \le i \le n$$

$$\frac{\partial s_j}{\partial \alpha_o} = m_{oj}\left[\sum_{k=1}^{n} \alpha_k^2\right]^{-1/2} \ .$$

Applying (3.2) and (5.10),

$$\frac{\partial s_j}{\partial \alpha_i} = m_{ij} - \alpha_i s_j, \qquad 1 \leq i \leq n$$
$$\frac{\partial s_j}{\partial \alpha_o} = m_{oj}. \tag{5.11}$$

Consequently, from (5.8) and (5.11), (5.9) may be written

$$\frac{\partial T(\alpha_i,\lambda)}{\partial \alpha_i} = -\frac{\lambda}{\sqrt{\pi}} \sum_{j=1}^{m} (m_{ij} - \alpha_i s_j) e^{-(\lambda s_j)^2}, \qquad 1 \leq i \leq n$$
$$\frac{\partial T(\alpha_i,\lambda)}{\partial \alpha_o} = -\frac{\lambda}{\sqrt{\pi}} \sum_{j=1}^{m} m_{oj} e^{-(\lambda s_j)^2}. \tag{5.12}$$

The equations (5.12) give the gradient of $T(\alpha_i,\lambda)$. To make a step along the direction of the negative gradient ("downhill"), select a new set of coefficients α_i', $0 \leq i \leq n$, such that

$$\alpha_i' = \alpha_i = \theta \frac{\partial T}{\partial \alpha_i}, \tag{5.13}$$

where θ is some constant which governs the size of the step.

According to the iterative procedure, a value of λ and an initial set of coefficients are chosen, and the gradient (5.12) computed. These coefficients are incremented according to (5.13), and the process repeated. When

the process reaches a minimum, a larger value of λ is chosen and the iteration procedure repeated. This goes on until some termination criterion is met.

In order to determine when a minimum for any given λ has been passed, the value of $\partial T/\partial\theta$ for the old and new coefficients is computed. It will of course be negative for the old coefficients, since the gradient is being followed in the negative direction. If a minimum is passed, however, $\partial T/\partial\theta$ will be positive for the new coefficients. In this case, the minimum is approximated by choosing a new θ based upon a linear interpolation between the two values of $\partial T/\partial\theta$ for the old and new coefficients.

The expression for $\partial T/\partial\theta$ is developed below.

$$\frac{\partial T}{\partial \theta} = \sum_j \frac{\partial T_j}{\partial s_j} \frac{\partial s_j}{\partial \theta}$$

where

$$s_j = \left[\sum_{i=0}^{n} \left(\alpha_i - \theta \frac{\partial T}{\partial \alpha_i}\right) m_{ij}\right] \left[\sum_{k=1}^{n} \left(\alpha_k - \theta \frac{\partial T}{\partial \alpha_k}\right)^2\right]^{-1/2} .$$

Then

$$\frac{\partial s_j}{\partial \theta} = \left[- \sum_{i=0}^{n} \frac{\partial T}{\partial \alpha_i} m_{ij} \right] \left[\sum_{k=1}^{n} \left(\alpha_k - \theta \frac{\partial T}{\partial \alpha_k} \right)^2 \right]^{-1/2}$$

$$- \frac{1}{2} \left[\sum_{i=0}^{n} \left(\alpha_i - \theta \frac{\partial T}{\partial \alpha_i} \right) m_{ij} \right] \left[\sum_{k=1}^{n} \left(\alpha_k - \theta \frac{\partial T}{\partial \alpha_k} \right)^2 \right]^{-3/2}$$

$$\left[-2 \sum_{k=1}^{n} \alpha_k \frac{\partial T}{\partial \alpha_k} + 2\theta \sum_{k=1}^{n} \left(\frac{\partial T}{\partial \alpha_k} \right)^2 \right] . \qquad (5.14)$$

For $\theta = 0$, if (3.2) holds,

$$\left. \frac{\partial s_j}{\partial \theta} \right|_{\theta=0} = - \sum_{i=0}^{n} \frac{\partial T}{\partial \alpha_i} m_{ij} + s_j \sum_{i=1}^{n} \alpha_i \frac{\partial T}{\partial \alpha_i} . \qquad (5.15)$$

Then

$$\frac{\partial T}{\partial \theta} = - \frac{\lambda}{\sqrt{\pi}} \sum_{j=1}^{m} \frac{\partial s_j}{\partial \theta} e^{-(\lambda s_j)^2} . \qquad (5.16)$$

The evaluation of (5.16) with (5.15) will give the value of $\partial T/\partial \theta$ for the original coefficients α_i. For the new $\alpha_i' = \alpha_i - \theta \, \partial T/\alpha_i$, the complete expression (5.14) must be used, and the s_j appearing in the exponent of (5.16) must be the distance of the j^{th} point from the new hyperplane.

The value of θ which is chosen is rather arbitrary. If it is chosen too large the boundary may oscillate about or diverge from the minimum. If it is chosen too small, the iteration process will take longer than necessary. The following value of θ was found by experience to yield reasonable results

$$\theta = \frac{1}{15\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial T}{\partial \alpha_i}\right)^2\right]^{1/2}} .$$

The value of λ is also rather arbitrary. The initial value, λ_o, was chosen so that

$$\lambda_o = 1/s_{j_{max}} ,$$

where $s_{j_{max}}$ is the distance of the furthest point from the initial hyperplane. After the interpolated values of α_i which minimize $T(\alpha_i,\lambda)$ are determined, λ is doubled. This process continues until $\lambda s_{j_{min}} \geq 3$, at which point the iteration terminates. $s_{j_{min}}$ is the distance of the closest point to the hyperplane. When $\lambda s_{j_{min}} \geq 3$, then $e^{-(\lambda s_j)^2} < .0001$ for all points, and the gradient becomes extremely small.

The initial hyperplane was usually chosen to be the perpendicular bisector of the line segment joining the sample means of the two classes. As indicated earlier,

several initial choices ought to be made and the best result used. It will be seen in Chapter X that this choice of an initial hyperplane is not always the best.

This iterative procedure is summarized in the flow chart of Figure 12.

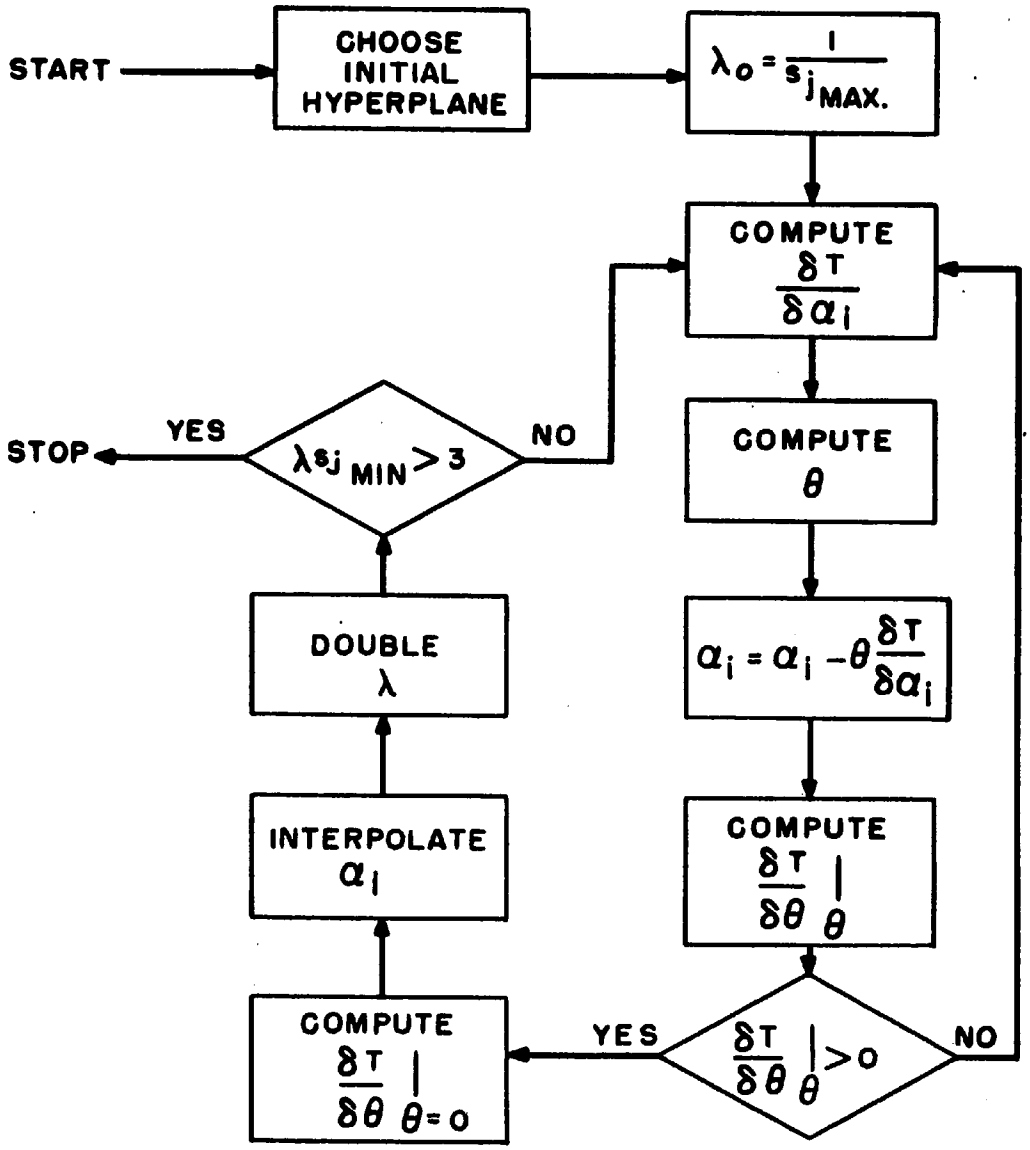
5.4 An Example of Categorization

To show the relation between these various approaches to the problem of categorization (the optimum decision function, the optimum linear decision function based on knowledge of the distributions, and the optimum linear decision function based on sampling), the following two-class problem was solved using each technique.

Problem: There are two pattern classes, 1 and 2, upon which two measurements, x and y, are made. The measurements are independent and normally distributed with the following parameters:

Class 1:	$\sigma_{1x} = 1$	$\mu_{1x} = 1$
	$\sigma_{1y} = .5$	$\mu_{1y} = 1$
Class 2:	$\sigma_{2x} = .1$	$\mu_{2x} = 2$
	$\sigma_{2y} = 2$	$\mu_{2y} = 0$.

The a priori probabilities of occurrence and the misrecognition losses are the same for each class. Determine the boundaries between the classes in the measurement space x,y.



FLOW CHART OF THE ITERATIVE PROCESS

FIGURE 12

Solution 1: Optimum decision function. The boundary corresponding to the optimum decision function is given by (4.6) which yields

$$-99x^2 + 3.75y^2 + 398x - 8y - 393 = 0 .$$

This is a hyperbolic boundary, and is shown in Figure 13. In this illustration, the 1σ contours of the classes 1 and 2 are also shown. The region identified as class 2 is of course that region between the two curves of the hyperbola.

Solution 2: Optimum linear decision function based on knowledge of the distributions. Let the optimum linear boundary to be given by

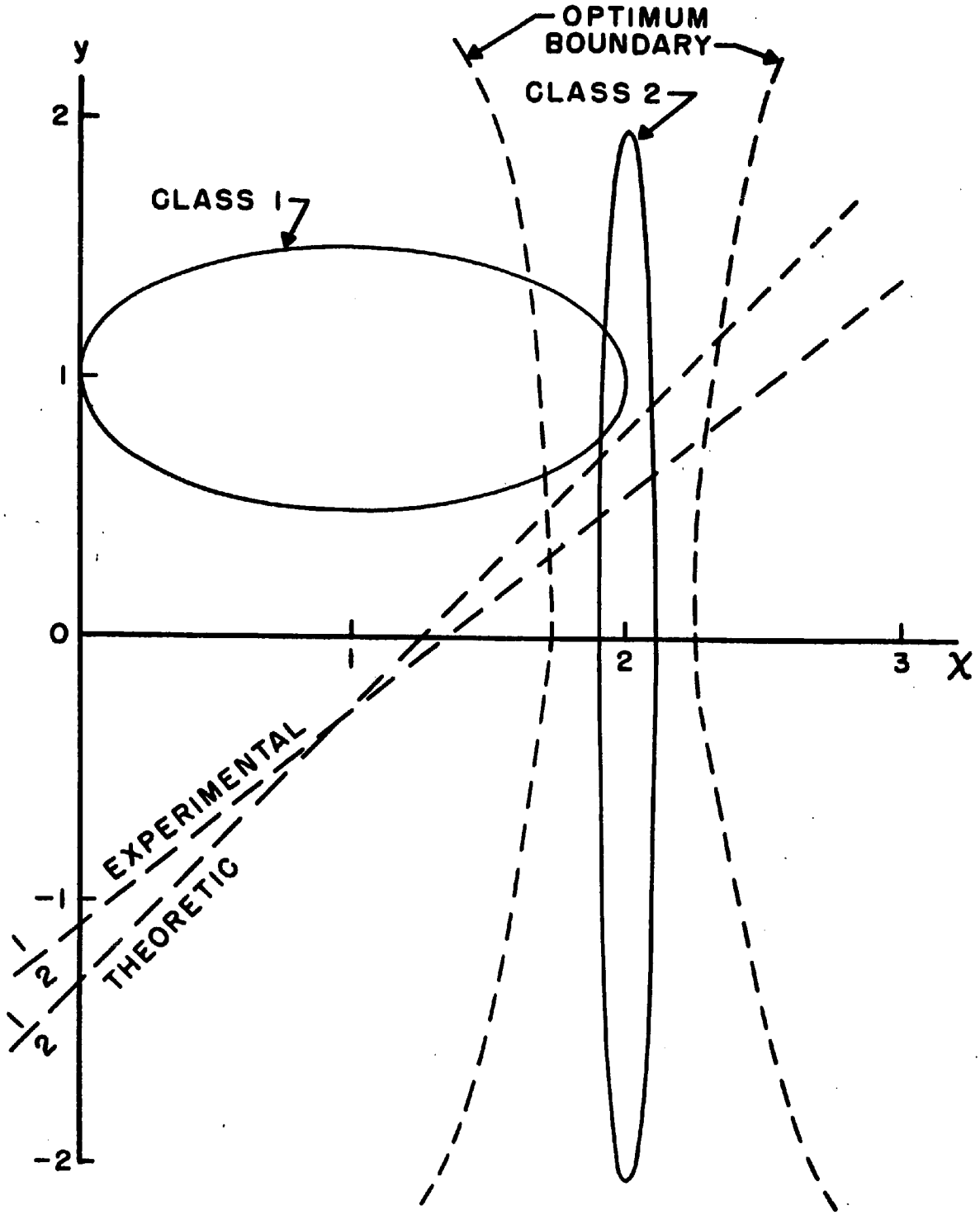
$$y = ax + b .$$

By substituting the Gaussian forms into the conditions given by Theorem 11, one obtains the following implicit equations for a and b (the details will not be included):

$$\left(\sigma_{1y}^2 + a^2\sigma_{1x}^2\right)e^{\gamma_1} = \left(\sigma_{2y}^2 + a^2\sigma_{2x}^2\right)e^{\gamma_2} ,$$

and

$$\frac{\left(\sigma_{1y}^2 + a^2\sigma_{1x}^2\right)}{\sigma_{1y}^2\mu_{1x} + \sigma_{1x}^2 a(\mu 1_y - b)} = \frac{\left(\sigma_{2y}^2 + a^2\sigma_{2x}^2\right)}{\sigma_{2y}^2\mu_{2x} + \sigma_{2x}^2 a(\mu_{2y} - b)} ,$$



AN EXAMPLE OF SOME OF THE
APPROACHES TO CATEGORIZATION

FIGURE 13

where

$$\gamma_i = \frac{[a\mu_{ix} - (\mu_{iy}-b)]^2}{\sigma_{iy}^2 + a^2\sigma_{ix}^2} , \qquad i = 1,2 .$$

An iterative solution of these equations yields the expression for the optimum linear hyperplane:

$$y = 1.04x - 1.32 . \qquad (5.17)$$

This is shown plotted in Figure 13 as the "theoretical" linear boundary.

Solution 3: Optimum linear decision function based on sampling from the classes. This solution was obtained on the IBM 7090 digital computer. The iteration algorithm of the previous section was programmed, as well as a "pattern source", a random number generator which generated numbers according to the particular normal distributions of the problem.

One hundred sample points were taken from each class. Various initial boundaries were tried:

$$x = 1$$
$$x = 2.5$$
$$y = x - 1$$
$$y = 4.2x - 6.8 .$$

Each of the final boundaries were slightly different, but the important point is that each one categorized the points

in exactly the same manner. (Thirty-nine points were always misclassified.) An example of one of the final boundaries is

$$y = .816x - 1.11 \qquad (5.18)$$

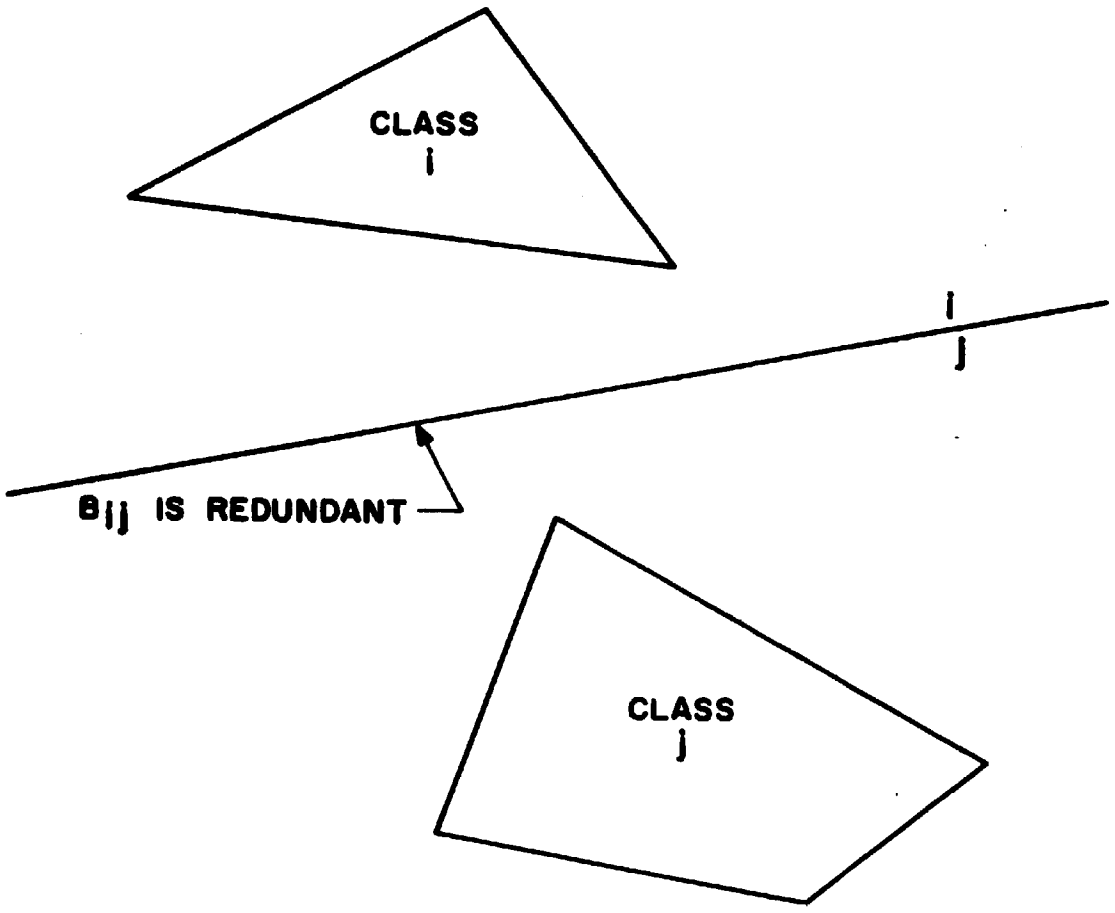
which is plotted in Figure 13 as the linear boundary marked "experimental". Compare (5.17) to (5.18); the difference illustrates the sampling error.

CHAPTER VI

ELIMINATION OF REDUNDANCIES

There are at least two possible types of redundancies which may occur in machines as described so far. One of the redundancy types has to do with the receptor, and hence may occur in other types of pattern recognition machines as well. This is the redundancy of certain measurements made by the receptor; that is, there is a possibility that certain measurements will contain no, or perhaps only a little, information concerning the proper categorization of the allowable pattern classes. Redundant measurements are often eliminated by intuition before the design of the receptor, but this intuitive elimination is not always complete, and sometimes is not even possible. It is an interesting property of linear decision functions that they can help the designer locate redundant measurements.

The other type of redundancy is contained in the linear decision function itself. It is quite possible that a particular boundary may be completely unnecessary to the segmentation of measurement space, and therefore this boundary need not be instrumented in the machine. As an example, Figure 14 shows the convex polytopes bounding two classes, i and j. Each face of a polytope enclosing, say, class i is a section of a hyperplane separating class i from one of the



BOUNDARY REDUNDANCY

FIGURE 14

other classes. In Figure 14, the boundary B_{ij} is also shown. It is not a face of either of the polytopes containing class i or class j. It adds nothing to the categorization (it cannot, by definition of the categorization process, be a face of any other polytope), and can therefore be eliminated.

This chapter will deal with the detection of these two types of redundancies.

6.1 Detection of Redundant Measurements

The detection of a redundant measurement is a simple process. Consider a particular hyperplane given by

$$\sum_{i=1}^{n} \alpha_i x_i + \alpha_o = 0 \ . \qquad (6.1)$$

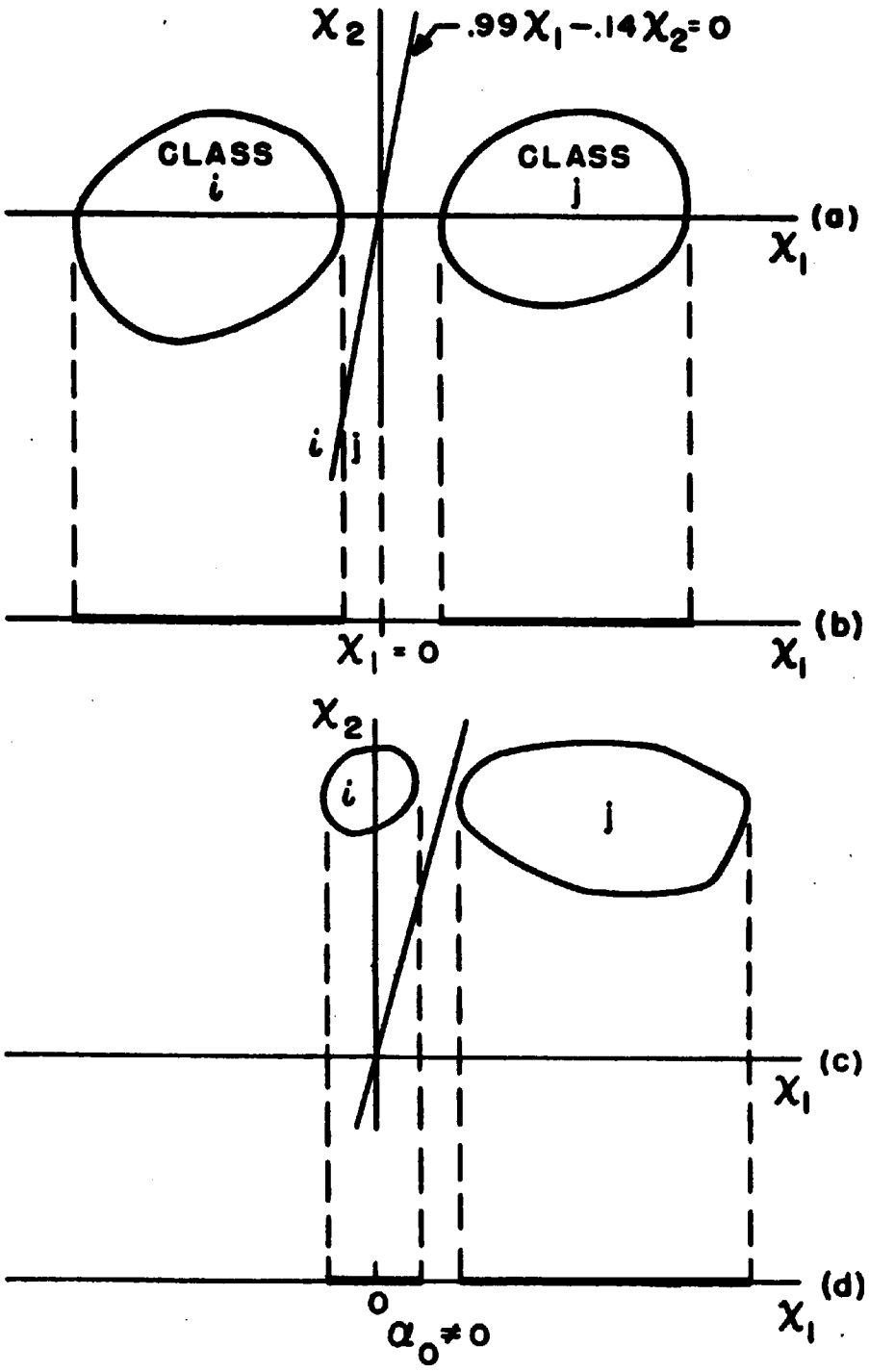
Each α_i, $1 \leq i \leq n$, is the direction cosine of the hyperplane with respect to the i^{th} coordinate, providing the normalization (3.2) holds. If α_i is zero, then the hyperplane is parallel to the x_i axis. If the measurement represented by x_i were not made, then this would correspond to projecting the measurement space onto the n-dimensional hyperplane defined by the remaining n-1 coordinate axes. The projection of each point into this new measurement space is in a direction parallel to the x_i coordinate axis. If the original hyperplane (6.1) were parallel to this axis (α_i=0), then the projected n-1-dimensional bounding hyperplane would separate the projected measurement space in exactly the same manner

as in the original n-dimensional measurement space. This can also be seen from (6.1); if an α_i is zero, then the measurement x_i has no effect on the categorization, and is redundant with respect to this hyperplane. If an α_i is small, it might then be expected that the measurement will have little effect, and might still be deemed redundant.

These ideas are illustrated graphically in Figure 15a, where the closed curves represent the domains of class i and class j. A bounding hyperplane is shown, in which α_2 is small with respect to α_1. If x_2 were then eliminated, the new measurement space is as shown in Figure 15b. In some cases, α_o may have to be recomputed to obtain a more accurate categorization in the new space, as shown in Figures 15c,d.

Now if a measurement x_i is redundant (according to some preset criterion pertaining to the increase in loss when x_i is removed) for <u>all</u> of the constituent hyperplanes comprising a linear decision function, then it is reasonable to remove that measurement from the receptor.

These concepts now give the designer some direction in looking for redundant measurements. He should consider first those measurements for which the maximum absolute value of the direction cosine over the set of hyperplanes comprising the decision function is smallest. The maximum absolute value of the direction cosine associated with a measurement



REDUNDANT MEASUREMENTS

FIGURE 15

may be said to be an indication of the redundancy of that measurement. The smaller this value, the more redundant the measurement. In particular, if all the direction cosines associated with a measurement are zero, then that measurement is absolutely redundant and may be removed without affecting the performance of the system at all.

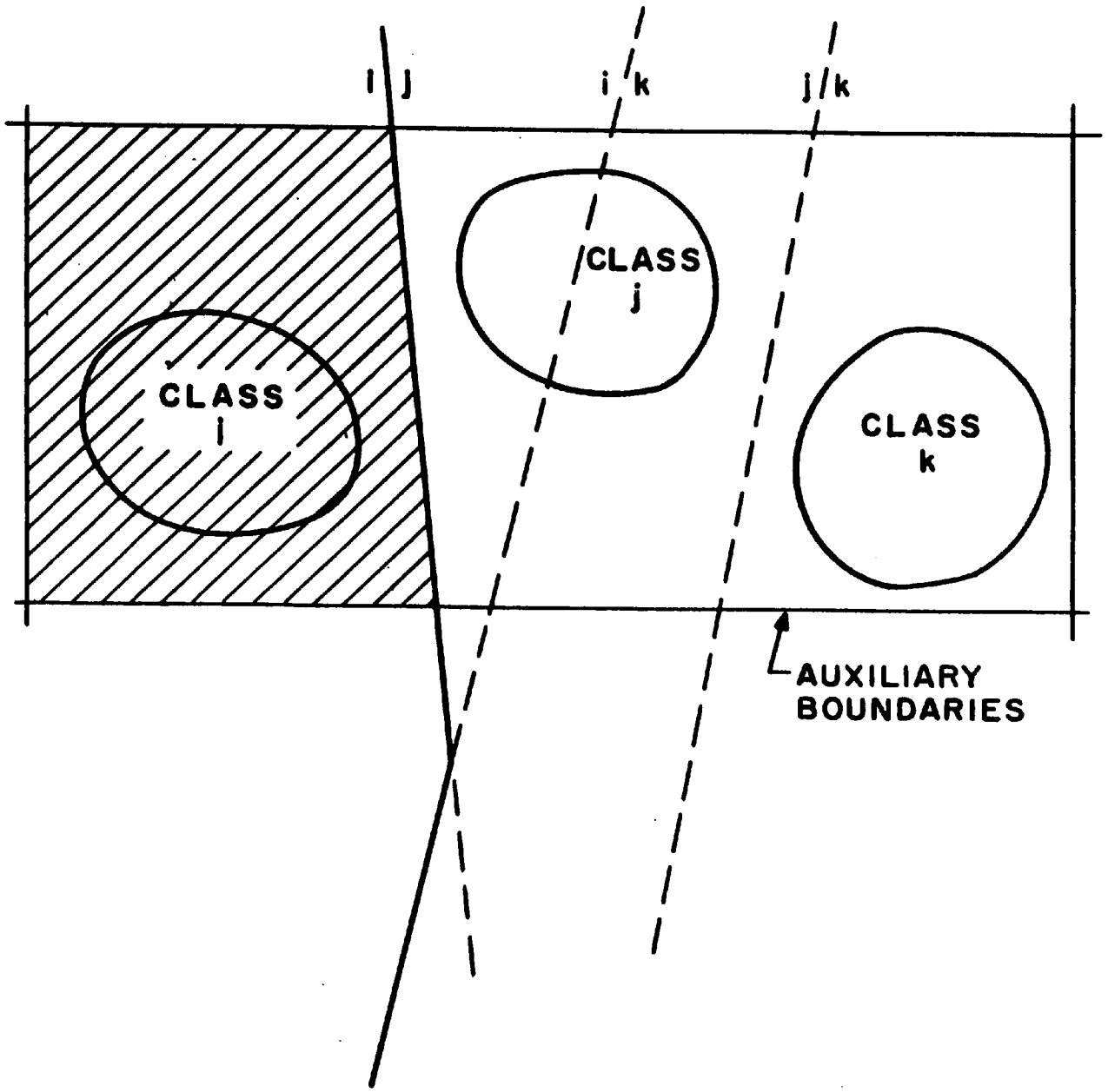
6.2 Detection of Redundant Boundaries

The true definition of a redundant boundary was given in the introduction to this chapter. It is a boundary which is not effective in the segmentation of measurement space into the various categories. That is, B_{ij} is redundant if it is not a face of at least one of the convex polytopes bounding the classes i and j.

Let us modify this definition slightly. In Figure 16 is shown the range of three pattern classes in measurement space (the closed curves). A possible linear decision function is also shown. Note that, according to the previous definition, B_{ik} is nonredundant. But it intersects the polytope containing class i in a region well outside the range of the measurements, and therefore adds nothing to the practical categorization. If, however, we enclose the set of pattern classes with another set of boundaries indicating the range of measurements, and consider these auxiliary boundaries* along with the boundaries comprising the linear

---

*This is still perfectly general, since the auxiliary boundaries may be at infinity.



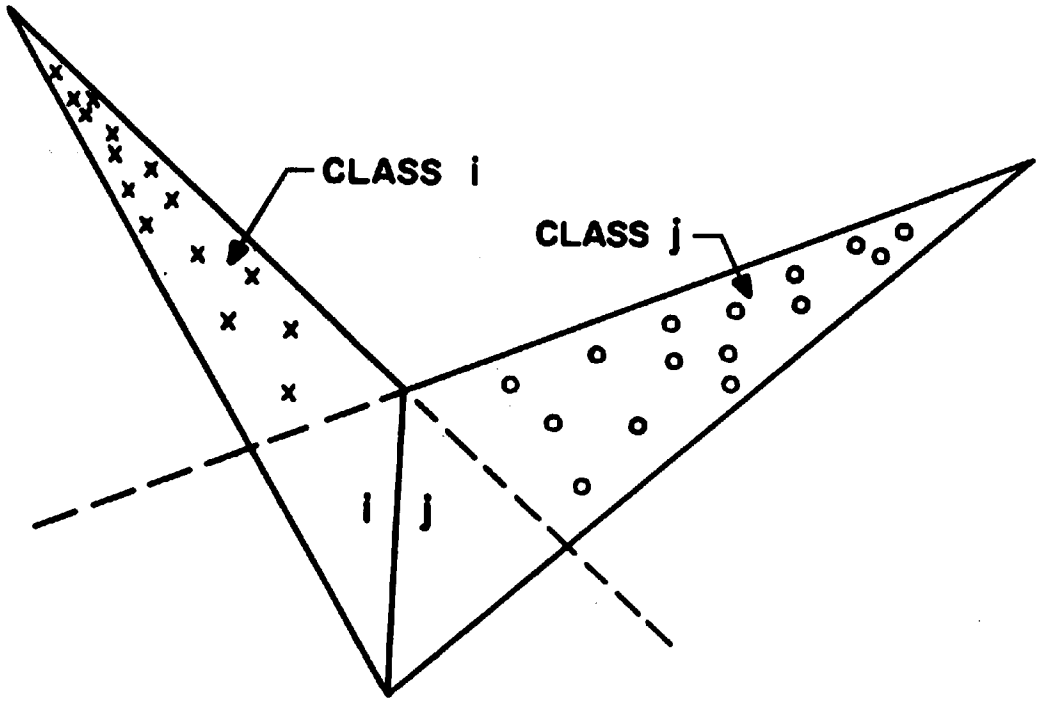
GEOMETRICAL REDUNDANCY

FIGURE 16

decision function, the boundaries such as B_{ik} will now be redundant. For instance, in Figure 16, with the auxiliary boundaries included, the polytope enclosing class i is indicated by the hatched region. B_{ik} is not a face of this polytope (nor of the polytope enclosing class k), and is therefore redundant. This sort of redundancy will be called geometrical redundancy. Note that the auxiliary boundaries are neither tested for redundancy nor built into the machine. They are included only as an artifice to make this definition more practical.

Methods will be given here for determining geometrical redundancy. However, it will be seen that these methods will usually involve a great deal of computation. A less accurate, but more practical definition of redundancy is therefore also considered. A boundary will be said to be redundant in a sample sense if its removal does not affect the classification of a set of sample patterns. These sample patterns may very well be the ones which were originally used to determine the linear decision function.

Clearly, a boundary which is geometrically redundant is also redundant in a sample sense. That the converse is not true is shown in Figure 17. The boundary B_{ij} is not geometrically redundant (the auxiliary boundaries are assumed to contain the two polytopes shown). But its removal will not affect the classification of the sample points shown by the crosses and circles. It is therefore redundant in a sample sense.



B_{ij} IS REDUNDANT IN A SAMPLE SENSE,
BUT IS NOT GEOMETRICALLY REDUNDANT

FIGURE 17

This example also points up another difference between the two definitions. A linear decision function which is complete except for the elimination of gecmetrically redundant boundaries still obeys the uniqueness property of Theorem 3 for all points contained within the auxiliary boundaries. However, when a boundary which is redundant in a sample sense but which is not geometrically redundant is removed, the linear decision function is no longer unique, i.e., there will be regions in measurement space which will be assigned to more than one class. For instance, in Figure 17, points in the area below the dotted extensions of the two boundaries will be identified as belonging to both classes i and j. This will be no problem providing that no patterns can ever fall in this area, or that the probability of such an occurrence is small and such patterns rejected. If the sample size is large, one can be quite confident that the probability of such an occurrence (a multiple recognition) is small.

6.2.1 Geometrical Redundancy

The determination of geometrically redundant boundaries is carried out by testing the boundaries one at a time. Each such test, as will be seen, requires a good deal of computation. The following theorem aids in this regard by giving a simple test which will identify some of the geometrically nonredundant boundaries by using a set of

correctly categorized samples. (These samples may be taken, for instance, from the samples which were used to design the linear decision.) These boundaries then need not be tested for redundancy.

Theorem 13: If a sample point, which represents a member from pattern class i and which is correctly categorized by the linear decision function, is closer to the boundary B_{ij} than to any other boundary B_{ik} or to any auxiliary boundary, then B_{ij} is geometrically nonredundant.

Proof: Let P_i be the convex polytope in measurement space containing those points identified as class i by the linear decision function. Let P_i' be that polytope contained in P_i which contains those points which are not only identified as class i, but which are also contained within the auxiliary boundaries. (It is assumed impossible to obtain a set of measurements lying outside of the auxiliary boundaries.) Suppose a sample point p, which represents a member of class i, and which lies within the convex polytope P_i', is closer to B_{ij} than to any other boundary B_{ik} or to any auxiliary boundary. Assume B_{ij} is geometrically redundant; it therefore is not a face of P_i', but rather lies outside of P_i'. Then any line segment joining p to B_{ij} must pass through one face of P_i, which is one of the other boundaries B_{ik} or one of the auxiliary boundaries. Therefore p cannot be closer to B_{ij} than to any other boundary B_{ik} or to any auxiliary boundary. But this contradicts the original supposition; therefore B_{ij} is geometrically nonredundant.

Boundaries which are not found to be nonredundant by the above procedure must now be tested for geometrical redundancy. The basic problem may be stated as follows. If there are m allowable pattern classes, then each polytope P_i containing a pattern class i is defined according to a linear decision function by m-1 linear inequalities (the bounding hyperplanes). If there are in addition b auxiliary boundaries, then the modified polytope P_i' is defined by (m-1+b) linear inequalities. This set may be written

$$\begin{aligned}
&\alpha_{11}x_1 + \ldots + \alpha_{1n}x_n < \alpha_{10} \\
&\quad\vdots \qquad\qquad\quad \vdots \\
&\alpha_{m-1,1}x_1 + \ldots + \alpha_{m-1,n}x_n < \alpha_{m-1,0} \\
&\alpha_{m,1}x_1 + \ldots + \alpha_{m,n}x_n < \alpha_{mb} \\
&\quad\vdots \qquad\qquad\quad \vdots \\
&\alpha_{m-1+b,1}x_1 + \ldots + \alpha_{m-1+b,n}x_n < \alpha_{m-1+b,n}
\end{aligned} \tag{6.2}$$

The last b inequalities of (6.2) correspond to the auxiliary boundaries. We wish to test a hyperplane represented by one of the first m-1 inequalities for geometrical redundancy. We will say that a boundary B_{ij} is geometrically redundant with respect to pattern class i if it is not a face of the polytope P_i'. If it is geometrically redundant with respect to class i, then it must be tested for such redundancy with respect to class j. A boundary can only be eliminated if it is redundant with respect to both of the classes which it separates.

Following are two methods for testing for geometrical redundancy with respect to one pattern class.

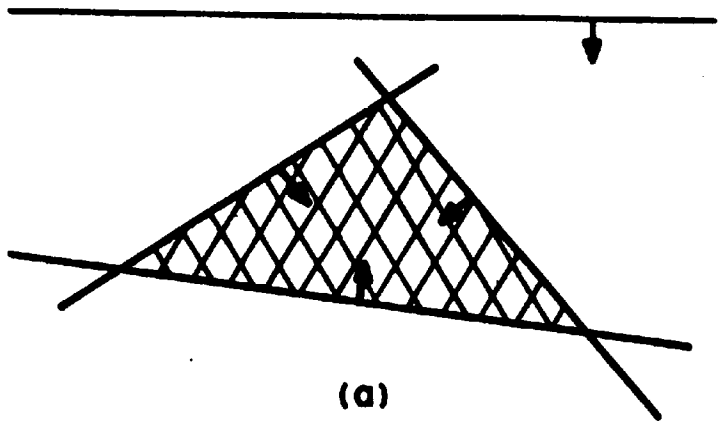
Method 1: Boundary Inversion*

We assume first that the set of inequalities (6.2), defining class i, has a solution (otherwise no patterns would ever be categorized into class i). Note that if a particular hyperplane is geometrically redundant with respect to class i, then reversing its sense (reversing its inequality sign in (6.2)) will cause the set (6.2) to have no solution. If the hyperplane is not geometrically redundant with respect to class i, then the set (6.2) will still have a solution when the inequality sign corresponding to that hyperplane is reversed. This is illustrated in Figure 18. The arrows indicate the half space for each hyperplane which satisfies the inequality; the cross-hatching indicates the solution space, if any. There is a problem if a redundant hyperplane happens to pass through a vertex or a higher order edge of P_i'; we assume the likelihood of this to be very small.

Consequently, in order to determine the redundancy of a hyperplane with respect to a class, one need only to reverse its sense in (6.2) and see whether the resulting set of linear inequalities has a solution. Methods of finding a solution to a set of linear inequalities, if one exists, or of indicating that a solution does not exist have been developed in the field of linear programming [22]

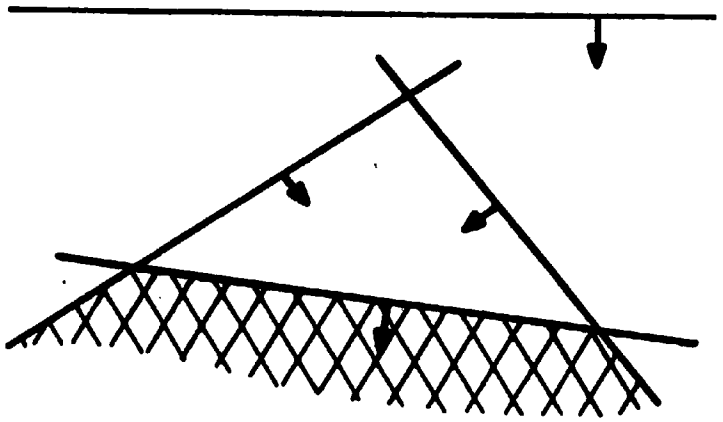
---

*Suggested by F. W. Sinden.



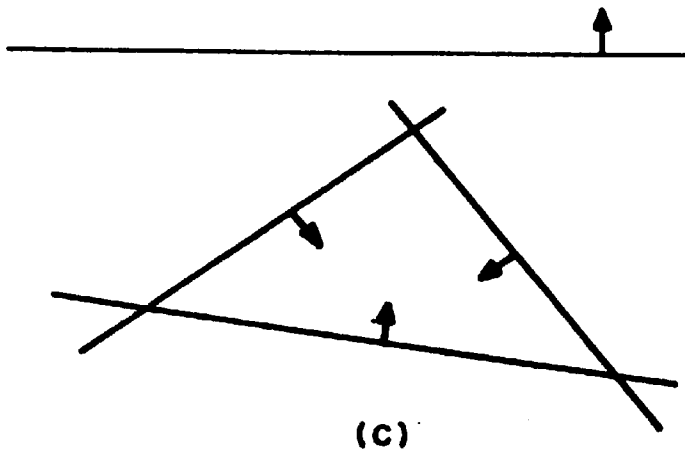
(a)

A SET OF LINEAR BOUNDARIES



(b)

INVERSION OF A GEOMETRICALLY NONREDUNDANT BOUNDARY



(c)

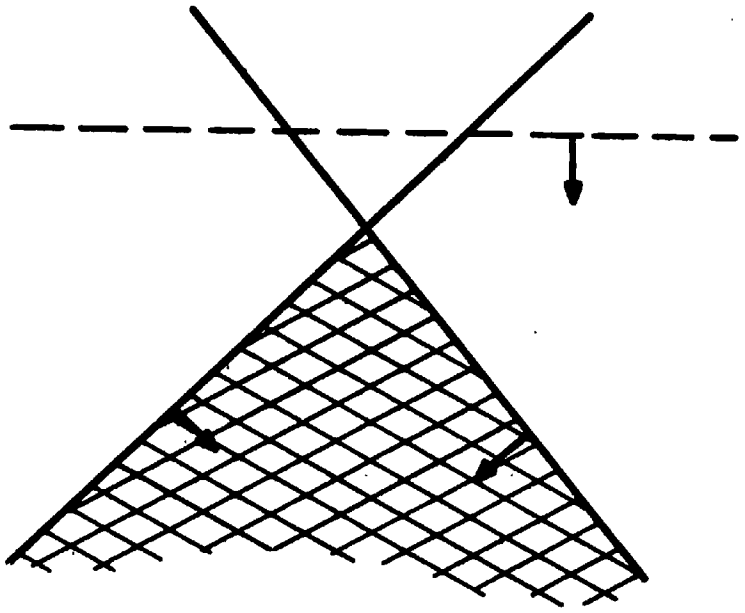
INVERSION OF A GEOMETRICALLY REDUNDANT BOUNDARY

and are applicable here. Relaxation methods [1,37] for solving a set of linear inequalities should be avoided, since they break down if a solution does not exist.

Method 2: Boundary Projection

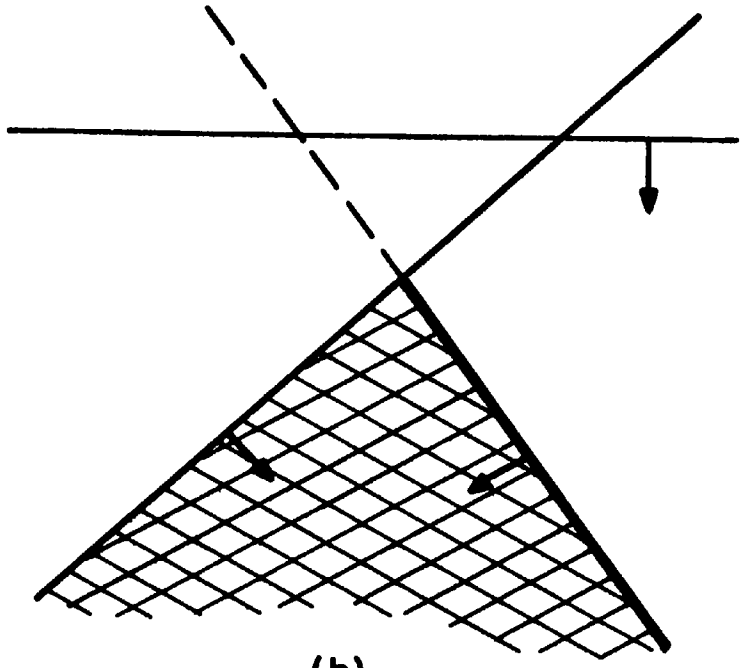
For discussional purposes, denote the boundary to be tested as B_k, $1 \leq k \leq m-1$, where the first m-1 boundaries are the ones associated with the linear decision function. If one considers the intersection of all the other (m+b-2) boundaries with B_k as a new set of (m+b-2) linear inequalities in an (n-1-dimensional space (the surface defined by B_k), then if B_k is geometrically nonredundant with respect to class i, this new set of linear inequalities will have a solution. In fact, the solution space is just that portion of B_k which is a face of the polytope P_i'. This is illustrated in Figure 19, where the solution space for the linear decision function is illustrated by the cross-hatched area. The solution space on the boundary being tested is denoted by a heavy line, whereas the rest of that boundary is drawn dashed.

Mathematically, this can be accomplished by eliminating the inequality corresponding to B_k from (6.2) by eliminating one of the variables x_i, $1 \leq i \leq n$, which has a nonzero coefficient in that inequality. This gives an augmented set of (m+b-2) linear inequalities in (n-1) dimensions which may then be tested for the existence of a solution as in Method 1. This will be a little simpler than



(a)

PROJECTION ONTO A GEOMETRICALLY REDUNDANT BOUNDARY



(b)

PROJECTION ONTO A GEOMETRICALLY NONREDUNDANT BOUNDARY

FIGURE 19

in Method 1 since the number of boundaries and the dimensionality are each reduced by one.

Geometrically, the above mathematical procedure corresponds to projecting the intersections of the (m+b-2) hyperplanes with the hyperplane B_k onto a hyperplane which is perpendicular to the eliminated coordinate axis, x_i. The nondestruction of the solution space by this second projection is guaranteed by the fact that the coefficient of x_i in the linear inequality representing B_k is nonzero.

This leads one to another simple test for geometrical nonredundancy.

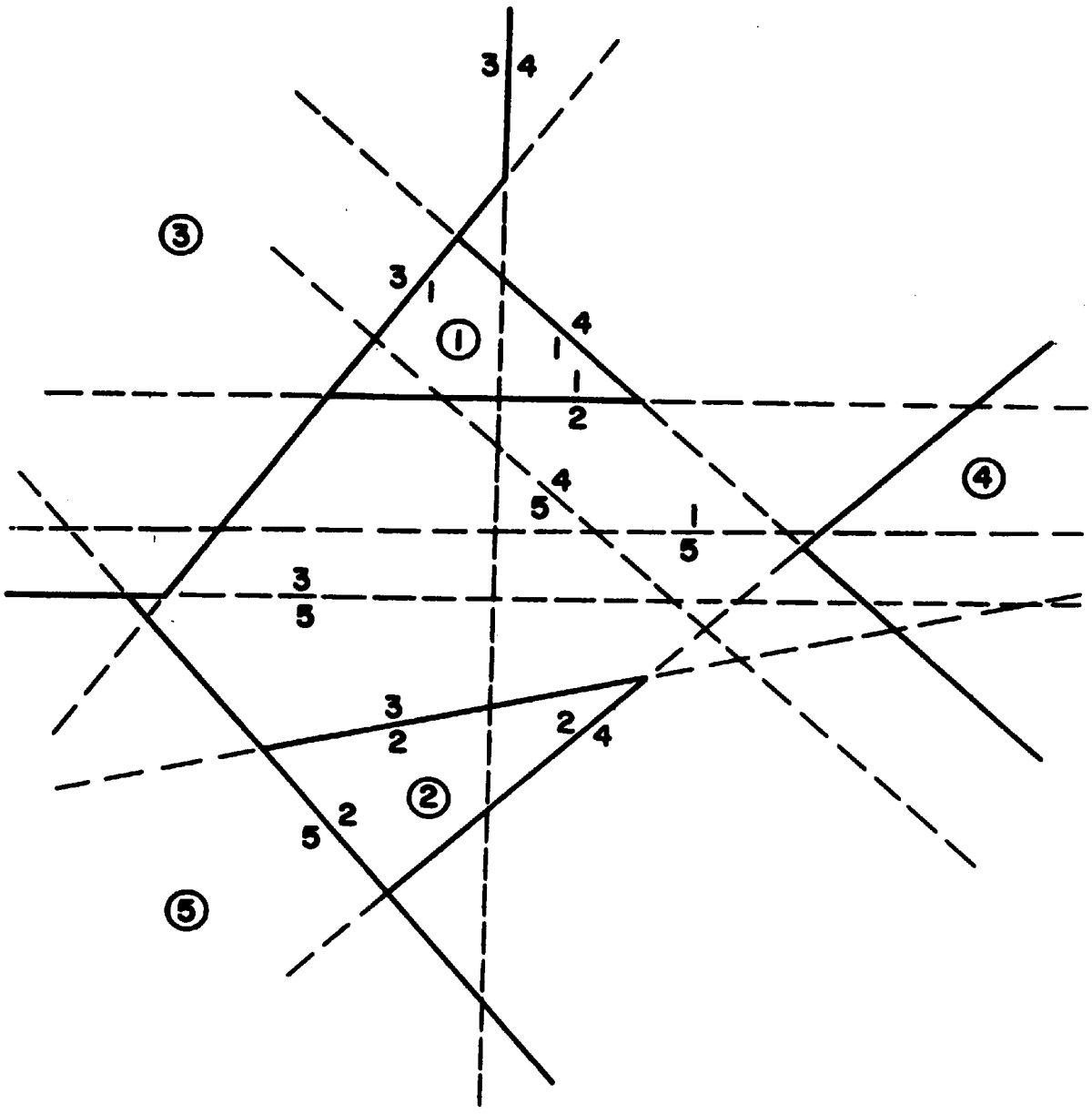
Theorem 14: Consider the hyperplanes (including the auxiliary boundaries) bounding a particular pattern class 1. If one of these hyperplanes, say B_k, has a nonzero coefficient associated with a particular coordinate, and all the other hyperplanes have zero coefficients associated with that coordinate, then B_k is geometrically nonredundant with respect to pattern class 1.

Proof: Let the coordinate in question be x_j. Eliminate B_k from (6.2) by eliminating x_j. The new set is then exactly like the original set (6.2), except that B_k has been removed. This is due to the zero coefficients of x_j contained in all hyperplanes but B_k. Since (6.2) had a solution space, then removing B_k can only enlarge that solution space. Therefore B_k is geometrically nonredundant with respect to class 1.

6.2.2 Redundancy in a Sample Sense

Recall that a boundary is redundant in a sample sense if its removal does not incorrectly classify any member of a set of sample points, each of which was originally classified correctly. Unfortunately, each boundary cannot be tested separately as in the determination of geometric redundancy. Figure 20 illustrates this problem. Five pattern classes are shown, along with the ten boundaries comprising the complete linear decision function. The regions which are associated with each of the pattern classes by the linear decision function are shown bounded by heavy lines. It is clear from Figure 20 that boundaries B_{15} and B_{45} are geometrically redundant.

So far as redundancy in a sample sense is concerned, note that if only B_{15} were to be removed, there would be no change in the classification of any possible sample point which was correctly recognized with B_{15} included. Likewise, if only B_{12} were to be removed, there would be no change in classification of any such point (even though B_{12} is geometrically nonredundant). Therefore either B_{12} or B_{15} may be redundant in a sample sense. But if both B_{12} and B_{15} were to be removed, then there would be points in the regions associated with classes 2 and 5 which would be also classified as class 1; that is, there would be multiple recognition errors. This illustrates that redundancy in a sample sense



THE NONUNIQUENESS OF BOUNDARIES
REDUNDANT IN A SAMPLE SENSE

FIGURE 20

cannot be determined one hyperplane at a time, and furthermore that such an elimination is not unique, as in geometrical redundancy.

An iterative algorithm will be described which has been found to be useful in determining a set of boundaries redundant in a sample sense. First, let us make some definitions.

Def. 1: If the single removal of a boundary from a complete linear decision function causes confusion between some of the samples, it is said to be an unconditionally significant boundary.

Def. 2: If the single removal of a boundary from a complete linear decision function causes no confusion between the samples, then it is said to be a conditionally redundant boundary.

Def. 3: If B_{ij} is the only boundary of class i and class j which is conditionally redundant, then it is an unconditionally redundant boundary of the first kind.

Def. 4: If B_{ij} is a conditionally redundant boundary, and if, after the removal of all conditionally redundant boundaries and unconditionally redundant boundaries of the first kind from the complete linear decision function, samples from classes i and j are not confused, then B_{ij} is said to be an unconditionally redundant boundary of the second kind.

The algorithm is as follows

Step 1: Remove each boundary one at a time from the complete linear decision function and determine whether it is unconditionally significant or conditionally redundant. If it is unconditionally significant, it need never be tested again, for it must remain in the linear decision function. If a boundary is an unconditionally redundant boundary of the first kind, it also need never be tested again, since it can be eliminated permanently. That is, if B_{ij} is the only conditionally redundant boundary associated with either class i or j, then it is the only boundary of either class i or j about which there might be a question of elimination. Since eliminating B_{ij} causes no confusion between classes i and j, and since the removal of other conditionally redundant boundaries cannot affect either class i or j, then B_{ij} can be eliminated safely.

Step 2: Remove all conditionally redundant boundaries and unconditionally redundant boundaries of the first kind, and thus determine which are unconditionally redundant boundaries of the second kind. These need never be tested again since they can also be permanently removed.

Step 3: Reinsert the boundaries which are still conditionally redundant into the linear decision function and repeat step one with only these conditionally redundant boundaries and the unconditionally significant boundaries.

Step 4: Repeat steps two and three until a steady-state has been reached, i.e., the looping of this cycle does not change the number of conditionally redundant boundaries.

Step 5: After the above iteration converges, one will usually be left with only a few, if any, conditionally redundant boundaries. This remaining set will have the property that any one of the conditionally redundant boundaries may be removed without confusing any sample points, but, if all the conditionally redundant boundaries are removed, then there will be confusion. Clearly, then, at least one of these could be eliminated; perhaps more than one could be eliminated. If the number of conditionally redundant boundaries is small, the maximum number that might be eliminated may be determined by trial and error. Alternatively, certain symmetries might be noted in these boundaries which will allow an intelligent choice of those boundaries which can finally be eliminated. Since this problem will arise fairly infrequently, and since the number of various symmetries is large, suffice it to say that such symmetries exist and can be effectively utilized. Their further discussion hardly seems warranted here.

This algorithm will allow the designer to determine those boundaries which are redundant in a sample sense. Usually, the computational effort involved is significantly less than that required in the determination of geometrical redundancies, although the methods and results are not nearly so clear cut.

CHAPTER VII

REJECTION CRITERIA

Rejection regions which are inherent to linear decision functions have already been discussed. In addition, it is desirable to be able to introduce additional rejection regions which will prevent recognition of patterns whose classification is doubtful. Regardless of the sort of rejection criterion used, it should be compatible with the economy of implementation of linear decision functions; otherwise its inclusion would hardly seem worthwhile.

Consequently we will describe, in rather general terms, various "linear" rejection criteria, and then discuss one of these in more detail. Throughout this discussion, it will be assumed that the regions of doubtful recognition are near the boundaries themselves. This seems intuitively reasonable, and a use of the Central Limit Theorem later in the chapter will give one even more confidence in this assumption.

7.1 Linear Rejection Criteria

There are several ways in which one might interject a rejection criterion into a linear decision function. Perhaps the most general sort of linear rejection criteria, within the framework presented so far, would be to attempt to find two hyperplanes separating each pair of classes.

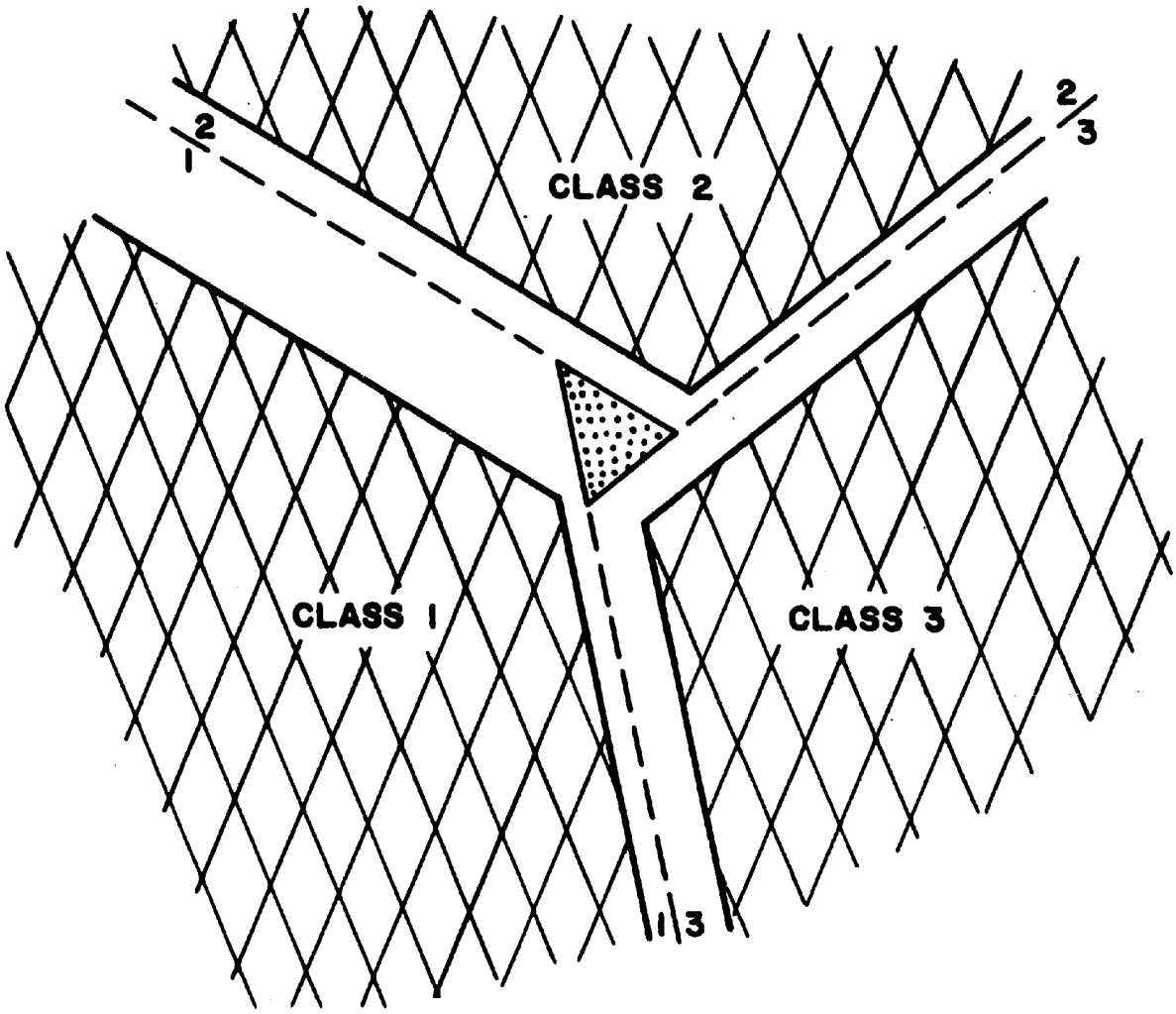
These hyperplanes would in general not be parallel. The regions to either side of the two hyperplanes taken together would be the regions for the two classes; the region between the two hyperplanes would be a rejection region. The problem would then be to determine these two hyperplanes simultaneously so that they are optimum in some sense - for instance, to meet a certain error rate with the minimum possible rejection rate, or perhaps to minimize the expected loss when rejection criteria are included.

A somewhat less general rejection criterion would be to constrain these two hyperplanes so that they are parallel to each other. The direction and separation of the hyperplanes would then be chosen in some optimum manner.

An even more restrictive case will be considered here. We will concern ourselves with rejection criteria which consist of two planes per pair of pattern classes, each parallel to and on opposite sides of the optimum linear boundary as determined by the methods of the previous chapters. The problem then reduces to simply finding the optimum separation between each optimum boundary and its corresponding rejection boundary on either side of it. We will continue to define optimum as meaning minimum expected loss under the constraints.

7.2 Optimization of a Type of Linear Rejection Criterion

Figure 21 illustrates the sort of rejection cirterion to be discussed. The dashed lines indicate the optimum linear



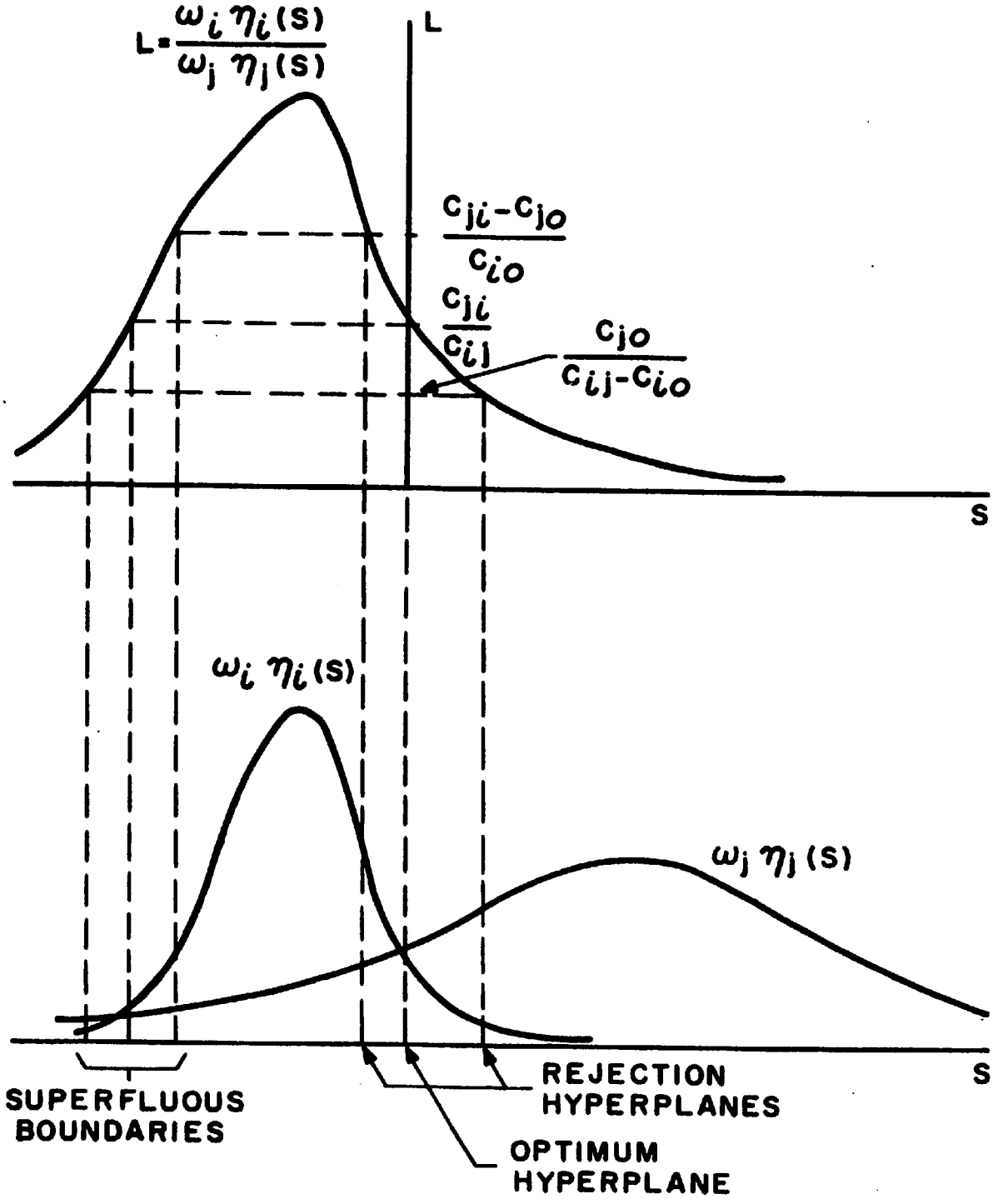
A LINEAR REJECTION CRITERION

FIGURE 21

decision function separating three classes when rejection criteria are not considered. The dotted region corresponds to the rejection inherent in a linear decision function. The actual regions corresponding to those points which are classified with a certain pattern class are shown cross-hatched; the regions in the neighborhood of each of the optimum boundaries are the rejection regions to be considered.

It might first be pointed out that the implementation of this rejection criterion is very simple. The device which decides whether the distance of a point from a plane is positive or negative (see Figure 3) will simply have a positive and negative threshold built into it corresponding to the distances of the two rejection hyperplanes from the optimum hyperplane. If the distance of a point is between these two thresholds, then some sort of rejection signal is delivered to the logical circuitry combining the hyperplane outputs.

When considering one optimum boundary at a time, the problem of the placement of the rejection hyperplanes can be conveniently reduced exactly to a one-dimensional problem. The one dimension is the distance from the optimum hyperplane. In Figure 22 is shown a plot of the density of distances from the hyperplane B_{ij} of members from classes i and j. Denote such a density function for class i by $\eta_i(s)$. If $\eta_i(s)$ and $\eta_j(s)$ are known, then the results of Chapter II,



LINEAR REJECTION

FIGURE 22

and in particular the decision function (2.3), may be used to determine the optimum position of the two rejection hyperplanes with respect to the optimum hyperplane.

If the sample size is large, $\eta_i(s)$ may be estimated graphically. However, the following argument gives a very powerful alternative for approximating $\eta_i(s)$. Recall that the distance of a point m from a hyperplane is given by

$$s = \sum_{i=1}^{n} \alpha_i m_i + \alpha_o,$$

where m_i is the i^{th} coordinate of the point, and the α_i are the normalized coefficients of the hyperplane. But m_i is a random variable, and hence, if n is large, s is a weighted sum of a large number of random variables. If the dependencies between the random variables are weak, one may then reasonably expect from the Central Limit Theorem [19] that the distribution of s is approximated by a normal distribution. Of course, if the measurements m_i are independent and normally distributed, then the normality of s follows immediately for any n.

Consequently, $\eta_i(s)$ is, to a good approximation in many cases, a normal density function. Its mean and variance can be easily estimated from the samples which were used to design the linear decision function.

The optimum decision function (2.3) can now be solved for this case. Let the two classes be i and j; c_{ij} is the cost of misrecognizing a member of class i as class j; c_{io} is the cost of rejecting a member of class i; $c_{ii} = 0$; ω_i is the a priori probability of occurrence of class i. From (2.3) we write

$$Z_i = (c_{ji}-c_{jo})\omega_j\eta_j(s) - c_{io}\omega_i\eta_i(s) ,$$

$$Z_j = (c_{ij}-c_{io})\omega_i\eta_i(s) - c_{jo}\omega_j\eta_j(s) .$$

Class i is chosen if $Z_i < Z_j$ and $Z_i < 0$. Class j is chosen if $Z_j < Z_i$ and $Z_j < 0$. The point is rejected if $Z_i > 0$ and $Z_j > 0$. Denote by L the likelihood ratio

$$L = \frac{\omega_i\eta_i(s)}{\omega_j\eta_j(s)} .$$

Clearly we would expect the optimum boundary (without rejection) to occur at $Z_i = Z_j$, or the point in Figure 22 for which

$$L = \frac{c_{ji}}{c_{ij}} .$$

Whether or not the estimated optimum linear boundary actually falls near this position is a function of the sampling error and the degree of approximation of the normality assumption.

The rejection region on either side of the optimum hyperplane is given by the condition that Z_i and Z_j be positive. This yields the region

$$\frac{c_{jo}}{c_{ij}-c_{io}} \leq L \leq \frac{c_{ji}-c_{jo}}{c_{io}} , \qquad (7.1)$$

which is illustrated in Figure 22. The distance of the two rejection hyperplanes from the optimum hyperplane is determined by the equality signs of (7.1).

If one is interested in minimizing the error rate for a given rejection rate, then he has only to set $c_{ij} = c_{ji} = c$, and $c_{io} = c_{jo} = c_o < c$. Letting

$$k = \frac{c}{c_o} - 1 ,$$

then the rejection region is that region for which

$$\frac{1}{k} \leq L \leq k , \qquad (7.2)$$

where the value of k will set the rejection rate. (Note that if $c \leq 2c_o$, $k \leq 1$, and there will be no rejection.)

Assuming that the distances of the sample points representing classes i and j from the hyperplane B_{ij} are normally distributed with means μ_i and $\mu_j < \mu_i$, and standard deviations σ_i and σ_j, then

$$\ln L = \ln \frac{\omega_i \sigma_j}{\omega_j \sigma_i} + \frac{1}{2}\left[\left(\frac{s-\mu_j}{\sigma_j}\right)^2 - \left(\frac{s-\mu_i}{\sigma_i}\right)^2\right] .$$

The two rejection hyperplanes are determined by the equality signs of (7.2) (or (7.1)); hence the separation

between the optimum hyperplane and the rejection hyperplanes are those values of s satisfying the following equations:

$$as^2 + 2bs + c = \delta_m, \qquad m = i,j \tag{7.3}$$

where

$$a = \frac{\sigma_i^2 - \sigma_j^2}{\sigma_i^2\sigma_j^2},$$

$$b = \frac{\mu_i\sigma_j^2 - \mu_j\sigma_i^2}{\sigma_i^2\sigma_j^2},$$

$$c = \frac{\mu_j^2\sigma_i^2 - \mu_i^2\sigma_j^2}{\sigma_i^2\sigma_j^2},$$

$$\delta_i = 2\ln\left(\frac{\omega_j\sigma_i}{\omega_i\sigma_j}\,k\right),$$

$$\delta_j = 2\ln\left(\frac{\omega_j\sigma_i}{\omega_i\sigma_j k}\right).$$

The solution of (7.3) for m = i corresponds to the rejection hyperplane on the negative side of the optimum boundary; for m = j, the positive side (since $\mu_j < \mu_i$). Note, however, that if the variances are different (as is usually the case), there are two values of s satisfying the quadratic of (7.3). This is illustrated in Figure 22 (the superfluous boundaries) and arises because there are two boundaries satisfying $L = c_{ji}/c_{ij}$. Since the optimum boundary used herein is that

one lying between the means* μ_i and μ_j, the rejection hyperplanes which are chosen are those associated with this boundary, and not with the boundary which is on the far side of the class with the smaller variance.

In summary, then, a rejection region parallel to the hyperplanes comprising a linear decision function can be easily incorporated. The hyperplanes are considered independently, and the distribution of distances of members of a pattern class to the hyperplane is approximately normal in many practical cases. The results of decision theory can then be applied directly to determine the rejection region on either side of the hyperplane, resulting in the conditions (7.1) or (7.2). This sort of rejection region is easily implemented in the equipment synthesizing a linear decision function.

---

*This is not quite an accurate statement. If the means are close enough together, neither optimum boundary may lie between them. However, a sketch of the distributions will make the choices clear.

CHAPTER VIII

THE DESIGN AND ANALYSIS OF PATTERN RECOGNITION EXPERIMENTS

There are two distinct and consecutive processes usually involved in the feasibility study of a pattern recognition method or machine. The first process is the actual design of the machine. This might be based upon a set of sample patterns which the experimenter has gathered, from which he estimates the parameters of the machine [4,13, 31,32]. Alternatively, the experimenter may base his design on some *a priori* knowledge concerning the pertinent characteristics of the pattern classes under study [5,21]. The second process is then the testing of this machine in either its hardware form or by its simulation on a general purpose computer. A different set of sample patterns from that used in the design are usually used for this test.

These two processes will be discussed in this chapter. All but the second section are generally applicable to pattern recognition studies; Section 8.2 applies only to linear decision functions. The general loss formulation will now be dropped for the rest of this paper in favor of considering error rates and rejection rates. Although this is not as general as considering the loss, it complies more closely with popular practices.

The first section will deal with the interpretation of test results when a pattern recognition machine is tested with samples which were not used to design the machine. The second section deals with a method of testing a linear decision function which gives an estimate of an upper bound on the error-plus-natural-rejection rate. Although this estimate is not as desirable as those discussed in the first section, it is applicable to the same sample set which was used to design the linear decision function.

The third section deals with the following problem. An experimenter finds that his sample size from the real world of patterns is fixed (for instance, due to economy reasons). He wants to use some of these patterns to design a categorizer, and the rest to test it. His machine will more closely approximate the optimum machine if he uses a larger sample size in the design stage. Likewise, the estimate of the machine's performance will become better as the test sample size increases. Consequently the experimenter is faced with the problem of deciding how to split his fixed sample set between a design sample set and a test sample set. This problem is not completely solved, but an approach to it is discussed, and some results are given.

8.1 Performance Estimation for Pattern Recognition Machines

Usually, a pattern recognition machine should be tested with a set of samples not used in its design. The popular procedure for interpreting these test results is to

take the proportion of patterns in the test data which have been misrecognized or rejected by the machine as the estimates of the error probability and rejection probability, respectively, for the machine. There are several questions which might be raised concerning this testing procedure, such as:

1. Are these estimates the best estimates?
2. If so, how good are these estimates?
3. How does the estimate improve as the sample size is increased?

Questions such as these are discussed in this section. Two cases are considered; one is the case in which the a priori probabilities of class occurrence are unknown, and the other case assumes full knowledge of the a priori probabilities.

8.1.1 Unknown a priori Probabilities - Random Sampling

Let the number of allowable pattern classes be p. It will be assumed that, for each allowable class i, there exists an a priori probability of occurrence ω_i, a probability of error e_i, and a probability of rejection r_i. The term "error" will refer to an undetected error; all detected errors will be assumed to be rejected. These probabilities are unknown to the experimenter, who is interested in estimating the over-all probability of error for the machine,

$$e = \sum_{i=1}^{p} \omega_i e_i \ , \qquad (8.1)$$

and the over-all probability of rejection,

$$r = \sum_{i=i}^{p} \omega_i r_i \ .$$

Let him perform the following experiment, which will be called random sampling. Consider the patterns to be randomly generated by a "pattern source" according to the a priori probabilities of occurrence. He takes a pattern from the source, identifies it, and then lets his pattern recognition machine attempt identification. He notes which of the three possible outcomes occur: correct recognition, misrecognition, or rejection. This experiment is repeated n times, resulting in m_e samples which have been misrecognized and m_r samples which have been rejected.

Since each of these outcomes are mutually exclusive, and each experiment independent, then the resulting random variables, m_e and m_r, clearly are distributed according to the multinomial probability distribution. That is, the joint probability distribution of m_e and m_r, $P(m_e,m_r)$, is given by

$$P(m_e,m_r) = \binom{n}{m_e\ m_r} e^{m_e} r^{m_r} (1-e-r)^{n-m_e-m_r} \ .$$

The maximum likelihood estimates for e and r, denoted by $\hat{e}$ and $\hat{r}$, are then [20]

$$\hat{e} = \frac{m_e}{n} ,$$

$$\hat{r} = \frac{m_r}{n} ,$$

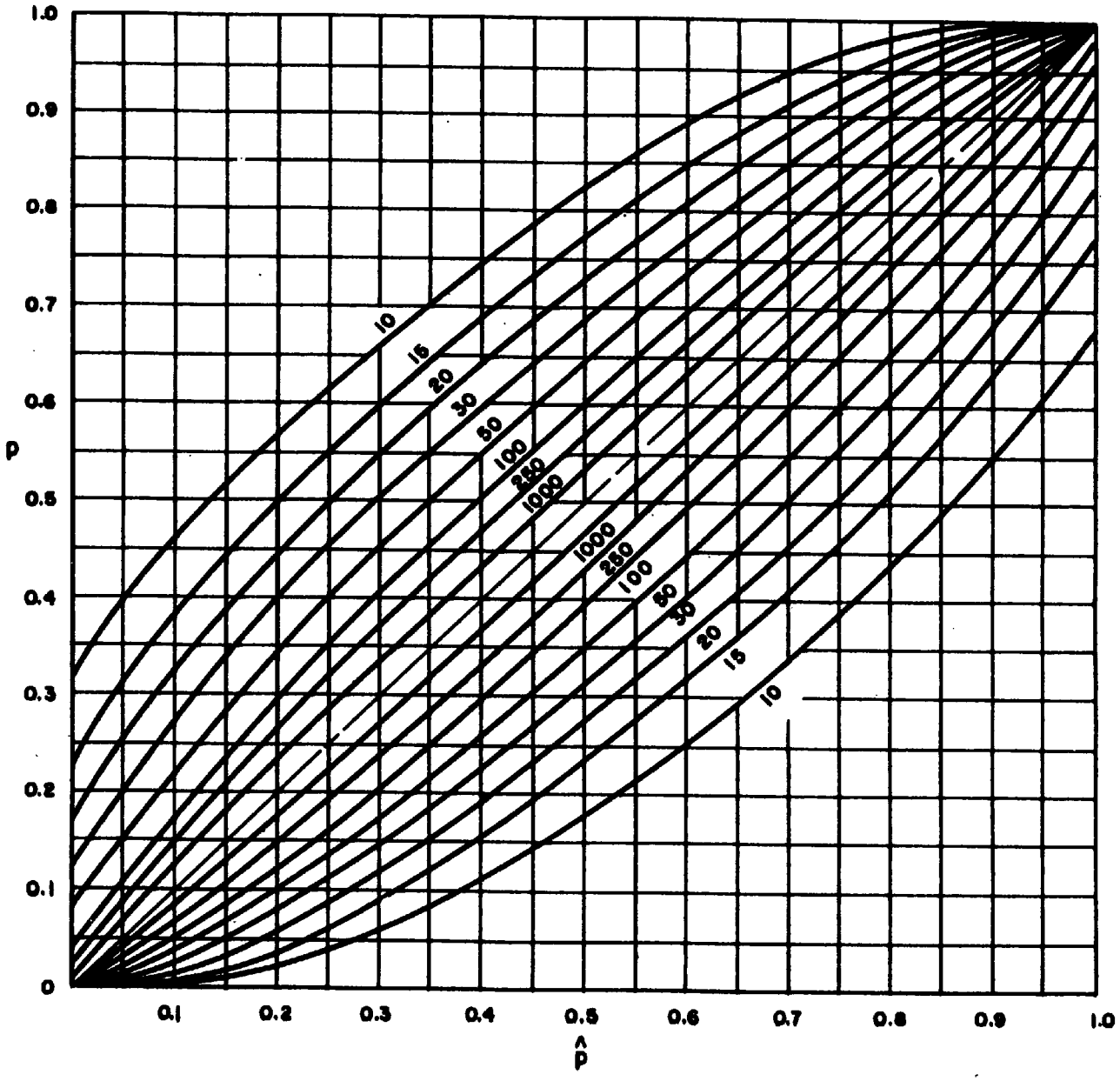
which are the estimates in common use. Further, each of these estimates is proportional to a single random variable having a binomial distribution; therefore, $n\hat{e}$ and $n\hat{r}$ are themselves binomially distributed. The mean value of each estimate is the parameter for which it is an estimate; the variance of each is [20]

$$\sigma^2_{\hat{e}} = \frac{1}{n^2}\,\sigma^2_{m_e} = \frac{e(1-e)}{n} \qquad (8.2)$$

$$\sigma^2_{\hat{r}} = \frac{r(1-r)}{n} .$$

Because it is known that $n\hat{e}$ and $n\hat{r}$ are binomially distributed, confidence intervals can be applied to these estimates.* These confidence intervals require rather involved computations, but fortunately have been plotted for several values of n by various people.[8,12,39] In Figure 23 is shown such a plot of intervals for a 95% confidence level computed by C. S. Clopper and E. S. Pearson. The use of this graph is fairly simple. A vertical line extended upward from the observed value of the estimate given on the abscissa will intersect the pair of curves pertaining to the particular

*Mattson [33] has used a similar argument for determining convergence of an adaptive system. However, he used Tchebycheff's inequality to obtain confidence intervals which are necessarily larger than if he had used such intervals pertaining to the binomial distribution.



95% CONFIDENCE INTERVALS FOR A BINOMIALLY DISTRIBUTED VARIABLE

FIGURE 23

sample size used. Projecting these two intersections horizontally onto the ordinate axis gives an interval for the parameter being estimated. The probability is .95 that the actual value of the parameter lies within this interval. For instance, if a sample size of n = 250 yielded 50 errors, then the estimate of the probability of error is .20. Using Figure 23, it can be stated that, with probability .95, the interval from .15 to .26 contains the true probability of error.

8.1.2 Known a priori Probabilities - Selective Sampling

It is now assumed that the a priori probability of occurrence for each class, ω_i, is known. To take advantage of this knowledge, the experimenter takes n_i samples from each class i such that

$$\frac{n_i}{n} = \omega_i \ , \qquad (8.3)$$

where n is the total number of samples. This process will be referred to as selective sampling.* (It will be assumed that the ω_i are such that equation (8.3) can be fulfilled with the desired sample size, n.)

The machine is again allowed to attempt recognition of these patterns, resulting in m_{e_i} samples from class i being misrecognized, and m_{r_i} samples from class i being rejected.

---

*This sort of sampling dichotomy has been noted by others. For instance, Bowley [6] and Neyman [38] have referred to these two methods as "unrestricted" and "stratified" sampling.

For any class i, the joint probability distribution for m_{e_i} and m_{r_i} again is multinomial:

$$P\left(m_{e_i}, m_{r_i}\right) = \binom{n_i}{m_{e_i}\ m_{r_i}} e_i^{m_{e_i}} r_i^{m_{r_i}} (1-e_i-r_i)^{n_i-m_{e_i}-m_{r_i}} . \tag{8.4}$$

Since each of these distributions is independent of the others in this experiment, then the joint probability of the outcome for all p classes is the product of the individual probabilities (8.4):

$$P\left(m_{e_1},\ldots,m_{e_p},m_{r_1},\ldots,m_{r_p}\right) = \prod_{i=1}^{p} \binom{n_i}{m_{e_i}\ m_{r_i}} e_i^{m_{e_i}} r_i^{m_{r_i}} (1-e_i-r_i)^{n_i-m_{e_i}-m_{r_i}} .$$

This is no longer a multinomial probability distribution. However, since the maximum likelihood estimate of a sum of independent variables is the sum of the maximum likelihood estimates, then these estimates for e and r are

$$\hat{e} = \frac{\sum_{i=1}^{p} m_{e_i}}{n} \tag{8.5}$$

$$\hat{r} = \frac{\sum_{i=1}^{p} m_{r_i}}{n} , \tag{8.6}$$

which again agree with the popular practice of using the proportions as estimates. The random variables of which $n\hat{e}$ and $n\hat{r}$ are values are not now binomially distributed, since a sum of binomially distributed variables is not itself a binomial distribution in general.

The mean of each estimate is again the particular parameter being estimated. The variance of each of these estimates can be computed:

$$\sigma'^2_{\hat{e}} = \frac{1}{n^2} \sum_{i=1}^{p} \sigma'^2_{m_{e_i}} = \frac{1}{n^2} \sum_{i=1}^{p} n_i e_i (1-e_i)$$

$$= \frac{1}{n} \sum_{i=1}^{p} \omega_i e_i (1-e_i) \ , \tag{8.7}$$

in which use of equation (8.3) is made, and the prime distinguishes this variance from that for random sampling. Similarly,

$$\sigma'^2_{\hat{r}} = \frac{1}{n} \sum_{i=1}^{p} \omega_i r_i (1-r_i) \ .$$

It is of interest to compare these variances for selective sampling with those obtained for the case of random sampling. Since the variance for $\hat{r}$ has the same form as $\hat{e}$ in both cases, it is necessary to consider only one of them, say $\hat{e}$. First note that $\sigma^2_{\hat{e}}$ can be written, using (8.1) and (8.2), as

$$\sigma^2_{\hat{e}} = \frac{1}{n} \left(\sum_{i=1}^{p} \omega_i e_i \right) \left(1 - \sum_{k=1}^{p} \omega_k e_k \right)$$

From (8.7)

$$\sigma^2_{\hat{e}} - \sigma'^2_{\hat{e}} = \left[\frac{1}{n}\sum_{i=1}^{p}\omega_i e_i - \frac{1}{n}\left(\sum_{i=1}^{p}\omega_i e_i\right)^2\right]$$

$$- \left[\frac{1}{n}\sum_{i=1}^{p}\omega_i e_i - \frac{1}{n}\sum_{i=1}^{p}\omega_i e_i^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{p}\omega_i e_i^2 - \frac{1}{n}\left(\sum_{i=1}^{p}\omega_i e_i\right)^2 . \qquad (8.8)$$

Noting that $\sum_{i=1}^{p}\omega_i = 1$, we let

$$\left(\sum_{i=1}^{p}\omega_i e_i\right)^2 = 2\sum_{i=1}^{p}\omega_i e_i \sum_{k=1}^{p}\omega_k e_k - \sum_{i=1}^{p}\omega_i \left(\sum_{k=1}^{p}\omega_k e_k\right)^2 .$$

Then (8.8) can be written

$$\sigma^2_{\hat{e}} - \sigma'^2_{\hat{e}} = \frac{1}{n}\sum_{i=1}^{p}\left[\omega_i e_i^2 - 2\omega_i e_i \sum_{k=1}^{p}\omega_k e_k + \omega_i \left(\sum_{k=1}^{p}\omega_k e_k\right)^2\right]$$

$$\sigma^2_{\hat{e}} - \sigma'^2_{\hat{e}} = \frac{1}{n}\sum_{i=1}^{p}\omega_i\left(e_i - \sum_{k=1}^{p}\omega_k e_k\right)^2 = \frac{1}{n}\sum_{i=1}^{p}\omega_i (e_i - e)^2 = \sigma_e^2 \geq 0 .$$

(8.9)

Hence, the variance in the case of random sampling is greater than the variance in the case of selective sampling, the difference being what might be interpreted as the variance of the class errors. That is, if e_i is treated as a random variable with probability distribution ω_i, then σ_e^2 is the variance of e_i. (A similar derivation holds for the variance of the rejection probability estimates.) That the selective sampling variance should be smaller than the random sampling variance might be expected, since in selective sampling more information is used, namely the a priori probabilities.

Although statements have been made concerning the mean and variance of the estimates in the selective sampling case, nothing has been said yet concerning confidence intervals. This is a much more complicated problem than in the case of random sampling, since the estimates do not have a simple distribution function. In fact, the confidence intervals will in general depend on the particular set of e_i's (or r_i's) pertaining to the machine, and not simply on e (or r).

However, for small probabilities, the binomial distribution is quite closely approximated by the Poisson distribution, the fit becoming perfect as the probability approaches zero.[20] For any reasonable recognition machine, one would expect the probabilities of error and rejection to be small; consequently, the marginal form of (8.4) for m_{e_i} or m_{r_i} may be approximated by a Poisson distribution. The estimates given by (8.5) and (8.6) are now sums of random variables

with Poisson distributions (approximately) which are then themselves Poisson distributed. If the over-all error is also small, as is usually the case, the binomial-Poisson approximation can now be used in reverse, and one may state that, for small error rates, the error and rejection estimates (8.5) and (8.6) are approximately binomially distributed. Consequently, one can use Figure 23 to obtain 95% confidence intervals for the error and rejection probabilities. Further, from (8.9), we would expect this confidence interval to be on the safe side, that is, the actual 95% confidence interval should be slightly smaller than this.

8.1.3 Application to Published Results

To illustrate the ease of determining these confidence intervals, some published results in pattern recognition are listed in Table 1 along with the 95% confidence intervals as determined from Figure 23. Three points of caution should be noted concerning the validity of the confidence intervals in this table. First, the author is not positive that the test data is different from the design data in every case. Second, to the best of the author's knowledge, in every case the number of samples taken from each allowable pattern class was predetermined. This is selective sampling; therefore, it is assumed that the proportion of samples taken from each class represents its <u>a priori</u> probability of occurrence. The third assumption is that the patterns used to test the machine

TABLE 1

95% CONFIDENCE INTERVALS FOR SOME

PUBLISHED RESULTS

AUTHOR	PATTERN CLASSES	MEASURED CHARACTERISTICS	RECOGNITION CRITERIA	SAMPLE SIZE	ERROR	95% CONFIDENCE INTERVAL
Baran, Estrin [3]	Machine Printed Numbers	Presence of ink in elements of 30 x 32 matrix	Maximize a posteriori probability (Bayes' Equation)	480	9%	7% - 12%
Bledsoe, Browning [4]	Hand-Printed Alpha-Numerics	Presence of mark in elements of 10 x 15 matrix	Matching 2-tuples of matrix elements against table	180	21.6%	13% - 29%
Bomba [5]	Hand-Printed Alpha-Numerics	Topological features (orientation of straight lines, intersections, etc.)	Decision tree	112	0%	0% - 4%
Doyle [14]	Hand-Printed AEILMNORST	Simply measured topological features	Maximize a posteriori probability (Bayes' Equation)	~450	12.5%	10% - 16%
Frishkopf [21]	Handwritten words	Extremes, and interconnections between extremes	Cross-correlation against dictionary	160	68%	57% - 77%
Harmon [21]	Unsegmented Handwritten Letters	Topological features (cusps, closures, special marks, etc.)	Decision tree	412	41.1%	37% - 46%
Highleyman [26]	Machine-printed numbers	Presence of ink in elements of 12 x 12 matrix	Cross-correlation against probability matrices	500	error 0% rejection 0.6%	0% - 2% 0% - 2.5%
Mathews, Denes [32]	Spoken digits	Frequency vs time spectra	Cross-correlation against previous averaged spectra from same speaker	99	6%	2% - 12%

TABLE 1 (Cont'd)

95% CONFIDENCE INTERVALS FOR SOME

PUBLISHED RESULTS

AUTHOR	PATTERN CLASSES	MEASURED CHARACTERISTICS	RECOGNITION CRITERIA	SAMPLE SIZE	ERROR	95% CONFIDENCE INTERVAL
Marill, Green [31]	Handwritten A,B,C (done as example only)	Distance of character from field edge along eight different line segments	Likelihood function assuming independent normal distribution of measures	90	3%	1% - 10%
Sebestyen [46,47]	Spoken digits	Frequency vs time spectra	Minimization of non-Euclidean distance measure to average spectra	20	0%	0% - 18%

are a reasonable sampling from the real-life world of patterns, and are not biased toward either well-formed or poorly-formed (noisy) patterns.

8.1.4 Summary

Two important cases concerning the testing of pattern recognition methods or machines have been considered: random sampling for the case of unknown a priori probabilities of class occurrence, and selective sampling for the case of known a priori probabilities.* The most predominant form of testing in the present day art is to assume that the pattern classes have equal a priori probabilities of occurrence, and consequently to use equal sample sizes for each class; this is a special case of selective sampling.

It has been shown that, for both cases, the maximum likelihood estimate for the error probability or rejection probability is simply the proportion of samples misrecognized or rejected. In the case of random sampling, the estimates are binomially distributed, and accurate confidence intervals can be obtained. In the case of selective sampling, tighter estimates are obtained which are approximately binomially distributed for small error rates. Conservative confidence limits may then be obtained for these estimates.

Using these notions, the experimenter can determine the sample size required to obtain results which he deems significant. Alternatively, if he has a limited sample

---

*A more general form of sampling is discussed in Appendix II.

size, he can determine the significance of his results. Note that in both cases considered, the variance is inversely proportional to the sample size. This does not mean that the confidence interval is inversely proportional to the square root of the sample size, however, since a binomial rather than a normal distribution pertains. However, perusal of Figure 23 seems to indicate that this is a good rule of thumb. Note also that the total number of samples required to obtain a certain confidence in the results seems to be independent of the number of allowable pattern classes. This is an interesting philosophical point to ponder.

8.2 Performance Estimation for a Linear Decision Function

It has been previously stated that the sample set used to test a pattern recognition machine should not include samples which were used in the design of that machine. In the case of linear decision functions, the reason for this is demonstrated by Theorem 10 and its corollaries. As was proved there, if the number of points used in designing a linear decision function is less than a certain threshold value, assuming no degeneracies among the sample points, then the optimum linear decision function would be expected to separate these points perfectly. However, the actual error and rejection rates may very well be quite large; thus it clearly is not valid to claim that the estimate of the error rate, for instance, is 0% based on this test in this case.

However, it is possible to estimate an upper bound on the total error-plus-natural-rejection rate, (i.e., a lower bound on the recognition rate) for a linear decision function when rejection criteria such as discussed in Chapter VII are not incorporated. This procedure is based on Theorem 6 and the normality argument of Section 7.2. By this argument, under certain conditions, the distances of a set of sample points from a hyperplane are approximately normally distributed. This is true regardless of the hyperplane, and will hold for any set of samples, including those used to design the linear decision function. The parameters of this normal distribution can be estimated by computing the sample mean $\bar{s}$ and the sample variance v of the set of n points in question:

$$\bar{s} = \frac{1}{n} \sum_{j=1}^{n} s_j \ ,$$

$$v = \frac{1}{n-1} \sum_{j=1}^{n} (s_j - \bar{s})^2 \ . \qquad (8.10)$$

Let these points represent the samples from a particular pattern class i, and let the hyperplane be one which separates this pattern class from some other pattern class j. Then the probability of misrecognizing a member of class i as belonging to class j can be estimated by determining the area under that part of the normal curve with

parameters (8.10) which falls on the j side of B_{ij}. Likewise, the probability of misrecognizing a member of class j as belonging to class i can be estimated. The sum of these two probabilities, after weighting them according to the a priori probabilities of occurrence of the respective pattern classes, is an estimate of the probability of error for the hyperplane B_{ij}.

The probability of error for each of the hyperplanes in the linear decision function can be estimated in this manner. Then Theorem 6 can be interpreted as stating that the probability of error plus the probability of (natural) rejection for the linear decision function is equal to or less than the sum of the probabilities of error for the constituent hyperplanes. Hence, an upper bound on the total error-plus-natural-rejection rate can be determined. This estimate is valid even for the sample set used to design the linear decision function.

It is possible to obtain a confidence interval for this estimate of the probability that one class will be misrecognized as another (i.e., a confidence interval for the estimate of the area on the wrong side of the hyperplane). Consider a normal distribution with positive mean μ, variance σ, and variate s. The area under this curve for $s < 0$ (which corresponds to the above probability of error) is

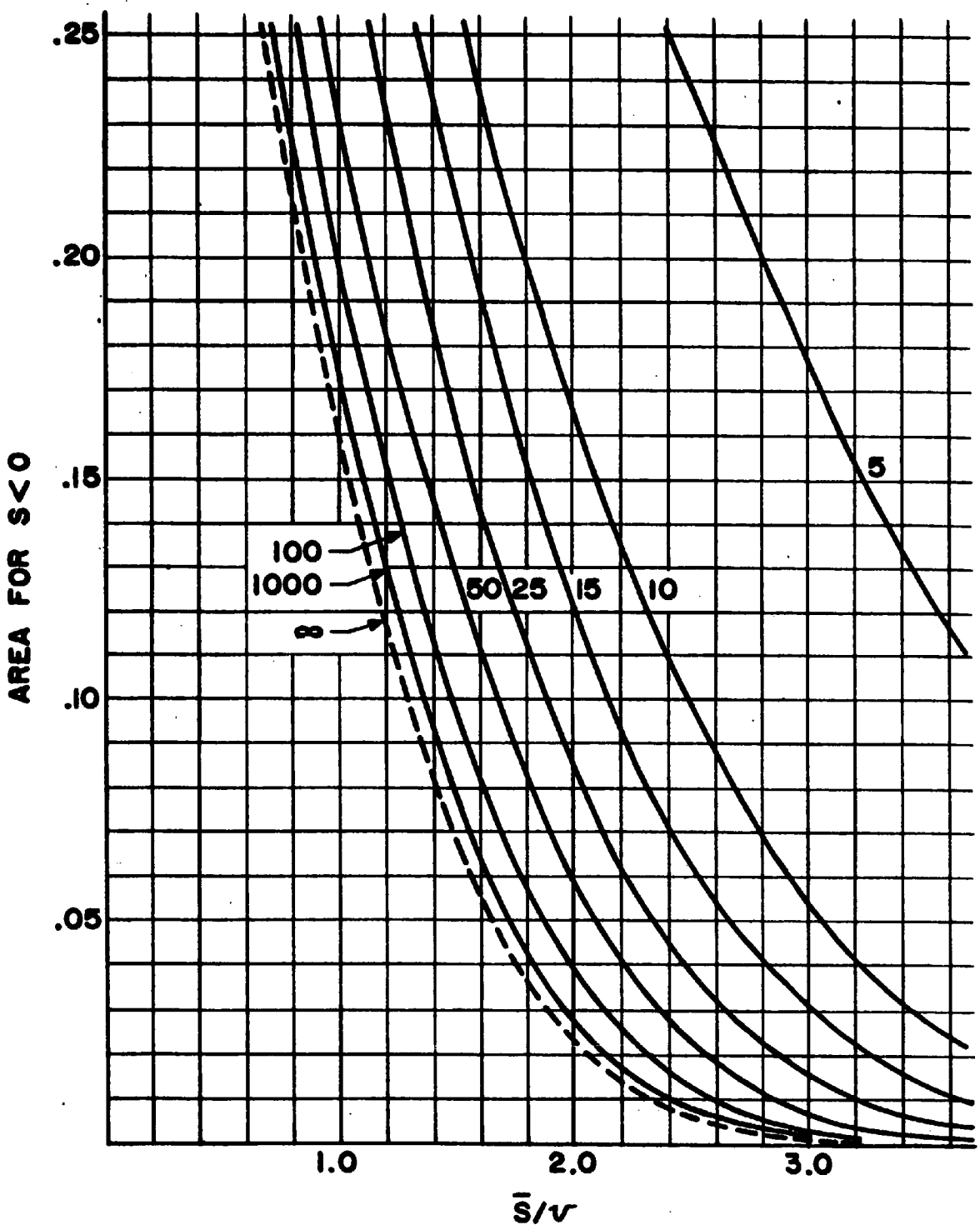
$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{0} e^{-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2} ds = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-\frac{\mu}{\sigma}} e^{-\frac{1}{2}x^2} dx ,$$

where $x = \frac{s-\mu}{\sigma}$. Consequently, this area depends only upon the ratio μ/σ, which is estimated by the ratio $\bar{s}/v$. But the random variable

$$t = \sqrt{n}\,\frac{\bar{s}}{v}$$

is distributed according to the noncentral t distribution with noncentrality parameter $\sqrt{n}\,\mu/\sigma$.[48] The parameters of this distribution are only the noncentrality parameter (a function of n and μ/σ), and the number of degrees of freedom, n-1 (a function of n). Consequently, confidence intervals for μ/σ, and hence, the area of interest, can be computed using this distribution, and will be a function of the sample size n and the estimate $\bar{s}/v$.

This distribution has been tabulated.[27,40] The curves in Figure 24 for sample sizes up to and including 50 have been plotted from tables given in reference [40]. The curves for larger sample sizes were computed from the normal approximation to the noncentral t distribution [27] ($\bar{s}/v$, for large sample sizes, is approximately normally distributed with mean μ/σ and variance $[1 + \frac{1}{2}(\mu/\sigma)^2]/n$). The use of Figure 24 is similar to the use of the graph in Figure 23,



95% CONFIDENCE INTERVALS FOR THAT AREA UNDER A NORMAL DENSITY FUNCTION CORRESPONDING TO NEGATIVE VARIATE, FOR VARIOUS SAMPLE SIZES

FIGURE 24

except that only the upper bound has been plotted. The reason for this is that we are estimating an upper bound of a quantity, and a lower bound on this estimate is not very useful. To use Figure 24, compute the quantity $\bar{s}/v$, and extend a vertical line upward from this value on the abscissa axis. Determine the projection onto the ordinate axis of the intersection of this vertical line with the curve corresponding to the sample size. It can then be stated that this value is greater than the true value of the area with probability .95.

Using the estimated normal distribution, one may also compute similar estimates for the probability of rejection for each hyperplane if the rejection criterion of Chapter VII is used. Consequently one may obtain an estimate of the upper bound for the rejection probability for the linear decision function. However, the computation of confidence intervals as described here will not apply.

8.3 Partitioning a Sample for Design and Test Purposes

Section 8.1 was concerned with the estimation of the performance of a given pattern recognition machine. There it was shown how confidence intervals could be found for these estimates. Two types of sampling from the real world of patterns were discussed: A procedure called random sampling was used when the *a priori* probabilities of pattern class occurrence were unknown, and a somewhat different procedure called selective sampling was used when the *a priori*

probabilities were known. It was shown for both cases that the maximum likelihood estimate of the error rate (or rejection rate) is simply the proportion of samples misrecognized (or rejected), in agreement with popular practice. It was further shown that the estimates in the case of random sampling obey a binomial distribution, and that the estimates in the case of selective sampling are approximately binomially distributed with a somewhat smaller variance than in the random sampling case. Consequently, confidence intervals may be applied to the estimates. These results are non-parametric in that they hold for any categorization machine (or procedure), regardless of its structure.

We now consider the following problem. An experimenter desires to solve a particular pattern recognition problem. He has at his disposal a set of different methods for solving this problem, but it is not clear to him which is the best to use. Consequently he desires to estimate the performance of each method when applied to his problem, and choose the best. Let us assume that each method is characterized by certain key parameters which, when known, completely determine the recognition machine. To evaluate any particular recognition method, the experimenter plans to estimate its parameters on the basis of one sampling from the real world of patterns, and then to test this machine based on another sampling.

However, in many practical applications, the total sample size available to the experimenter for design and test purposes is limited. For instance, he may be interested in building a machine to read hand-printed numbers, but he may not have an automatic scanner available to him. Since simulating a scanner by hand is very tedious, he may not be willing to scan more than a certain number of samples.

Or he may be interested in distinguishing between radar returns caused by missiles and those caused by decoys. Since it is expensive to actually run the sort of experiment required to gather data for this problem, budget limitations will certainly place a limit on the number of available samples.

Another example is in the field of automatic diagnosis of diseases. The experimenter may, for instance, be interested in building a machine which would determine the presence of cancer based on a list of symptoms. However, records have been maintained for only a certain number of people who have contracted this disease, and the sample size is thus definitely limited.

The following problem then arises. If the total sample size is fixed, what is the optimum partitioning of this sample between the design and test phases? This is a rather loose, but concise, statement of the problem. A more accurate one follows.

Assume that the experimenter is concerned with the study of a particular pattern recognition method as applied to some particular problem. The optimum pattern recognition machine based upon this method would have an error probability e_o. The experimenter is interested in estimating e_o so that he can decide whether the particular method under study is adequate for the solution of his problem, or alternately whether it is better than another method. To do this, he takes a sample of a certain size t from the real-life world of patterns. He desires to use part of this sample to design a machine according to the particular method under study. The machine which he thus designs will have an actual error probability $e > e_o$ (both quantities are unknown to the experimenter). He then uses the remaining part of his original sample to test the machine (according to the procedures of section 8.1). He thus obtains an estimate of e, which will be denoted by $\hat{e}$. It will be shown that $\hat{e}$ is a biased estimate of e_o, and that the bias can be computed. Consequently $\hat{e}$ can be adjusted so that it gives an unbiased estimate, $\hat{e}_o$, of e_o. The optimum partitioning of the total sample will be defined as that partitioning which minimizes the variance of $\hat{e}_o$. Thus, if the experimenter follows this procedure, he will obtain an unbiased minimum variance estimate of e_o, the optimum error probability. Of course, if he decides that a particular method is applicable, he can then redesign the corresponding machine with the entire sample size.

We are interested, then, in minimizing the quantity

$$\sigma^2_{\hat{e}_o} = E\left[(\hat{e}_o - e_o)^2\right] = E\left[\hat{e}_o^2\right] - e_o^2 , \qquad (8.11)$$

where $E[x]$ and σ_x^2 denotes the expected value and variance of x, respectively.

Let us first digress and consider the biased estimate $\hat{e}$. Since $\hat{e}$ is discrete (it is the proportion of test samples misrecognized), its expected value can be written

$$E[\hat{e}] = \sum \hat{e} p(\hat{e}) ,$$

where the summation is over all values of $\hat{e}$, and $p(x)$ denotes the probability of x. But

$$p(\hat{e}) = \int p(\hat{e} \mid e) p(e) de ,$$

where $p(\hat{e} \mid e)$ is the probability of $\hat{e}$ given e and the integral is over all (continuous) values of e (by definition $e_o \leq e \leq 1$). Hence

$$E[\hat{e}] = \sum \hat{e} \int p(\hat{e} \mid e) p(e) de$$

$$= \int \left[\sum \hat{e} p(\hat{e} \mid e) \right] p(e) de .$$

Let us hence forth consider only the case of random sampling. Then $\hat{e}$ is binomially distributed with parameter e. Therefore

the term in brackets, which is the expected value of $\hat{e}$ given the parameter e, is just e. Then

$$E[\hat{e}] = \int ep(e)de = E[e] \ . \qquad (8.12)$$

E[e]is a function only of the parameters of the problem and the design sample size; it is not a random variable.

We next determine $E[\hat{e}^2]$. By going through a process analogous to the above, and by making use of (8.12), we obtain

$$\sigma^2_{\hat{e}} = E[(\hat{e}-E[e])^2] = E[\hat{e}^2] - (E[e])^2 = \frac{E[e(1-e)]}{n} \ ,$$

where n is the size of the test sample. Hence

$$E[\hat{e}^2] = \frac{E[e(1-e)]}{n} + (E[e])^2 \ . \qquad (8.13)$$

We now determine E[e]. Let the optimum machine be described by c different parameters δ_{oi}, $1 \leq i \leq c$. The design of the machine consists of estimating the parameters δ_{oi} by making measurements on a set of sample patterns (the design sample). Let the estimated parameters be denoted δ_i, $1 \leq i \leq c$. Then the error probability e of the resulting machine is a function of the estimates of the true parameters:

$$e = e(\delta_1, \delta_2, \ldots, \delta_c) \ .$$

One can expand e in a Taylor series expansion about its minimum point, e_o. Since this is a minimum point, all the

coefficients of the linear terms will be zero. If the error deviation, $\Delta e = e - e_o$, is small, terms above the second order term may be neglected:

$$e \approx e_o + \frac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{c} \frac{\partial^2 e}{\partial \delta_i \partial \delta_j} \Bigg|_{\delta_o} (\delta_i - \delta_{oi})(\delta_j - \delta_{oj}) \ .$$

The expected value of the error for the resulting machine is then

$$E[e] = e_o + \frac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{c} \frac{\partial^2 e}{\partial \delta_i \partial \delta_j} \Bigg|_{\delta_o} E[(\delta_i - \delta_{oi})(\delta_j - \delta_{oj})] \ ,$$

or

$$E[e] = e_o + \frac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{c} a_{ij} \sigma_{ij} \ , \qquad (8.14)$$

where

$$a_{ij} = a_{ji} = \frac{\partial^2 e}{\partial \delta_i \partial \delta_j} \Bigg|_{\delta_o} \ ,$$

σ_{ij} is the covariance of the estimates for δ_{oi} and δ_{oj}, and $\sigma_{ii} = \sigma_i^2$ is the variance of the estimate for δ_{oi}. (8.14) is valid for small Δe.

Let each parameter be estimated with m samples. If each of these estimates is an efficient estimate, and if the estimates are independent (either because the estimates are

statistically independent, or because different samples are-- used to estimate each), then all $\sigma_{ij} = 0$, $i \neq j$, and all σ_i^2 will be proportional to 1/m. Hence one can rewrite (8.14) as

$$E[e] = e_o + \frac{b}{m} , \tag{8.15}$$

where b is some constant calculated from (8.14). (Often, E[e] is in the form (8.15) even if the estimates are not independent.)

Let t be the total sample size, and p be the number of sets of m samples used to design the machine. p is chosen to be the smallest number which insures that E[e] is of the form (8.15). It is often simply the number of allowable pattern classes, since, of course, parameters of different classes must be estimated with different samples. If n is the test sample size, then

$$t = n + pm . \tag{8.16}$$

From (8.12) and (8.15),

$$E[\hat{e}] = E[e] = e_o + \frac{b}{m} . \tag{8.17}$$

Consequently, $\hat{e}$ is a biased estimate of e_o. The adjusted estimate, $\hat{e}_o$, given by

$$\hat{e}_o = \hat{e} - \frac{b}{m} , \tag{8.18}$$

is an unbiased estimate of e_o, with variance given by (8.11). This variance can now be rewritten using (8.18):

$$\sigma^2_{\hat{e}_o} = E\left[\hat{e}_o^2\right] - e_o^2 = E\left[\left(\hat{e} - \frac{b}{m}\right)^2\right] - e_o^2$$

$$= E[\hat{e}^2] - 2\,\frac{b}{m}\,E[\hat{e}] + \left(\frac{b}{m}\right)^2 - e_o^2 .$$

From (8.13) and (8.17),

$$\sigma^2_{\hat{e}_o} = \frac{E[e(1-e)]}{n} + (E[e])^2 - 2\,\frac{b}{m}\,e_o - \left(\frac{b}{m}\right)^2 - e_o^2$$

$$= \frac{E[e(1-e)]}{n} + (E[e])^2 - \left(e_o + \frac{b}{m}\right)^2 .$$

Thus, from (8.17)

$$\sigma^2_{\hat{e}_o} = \frac{E[e(1-e)]}{n} . \tag{8.19}$$

If $\frac{b}{m} \ll 1$ (which will certainly be true for any reasonable design), then

$$\sigma^2_{\hat{e}_o} \approx \frac{E[e(1-e_o)]}{n} = (1-e_o)\,\frac{e_o + \frac{b}{m}}{n}$$

$$= (1-e_o)\,\frac{e_o + \frac{pb}{t-n}}{n} \tag{8.20}$$

where the relation (8.16) was used.

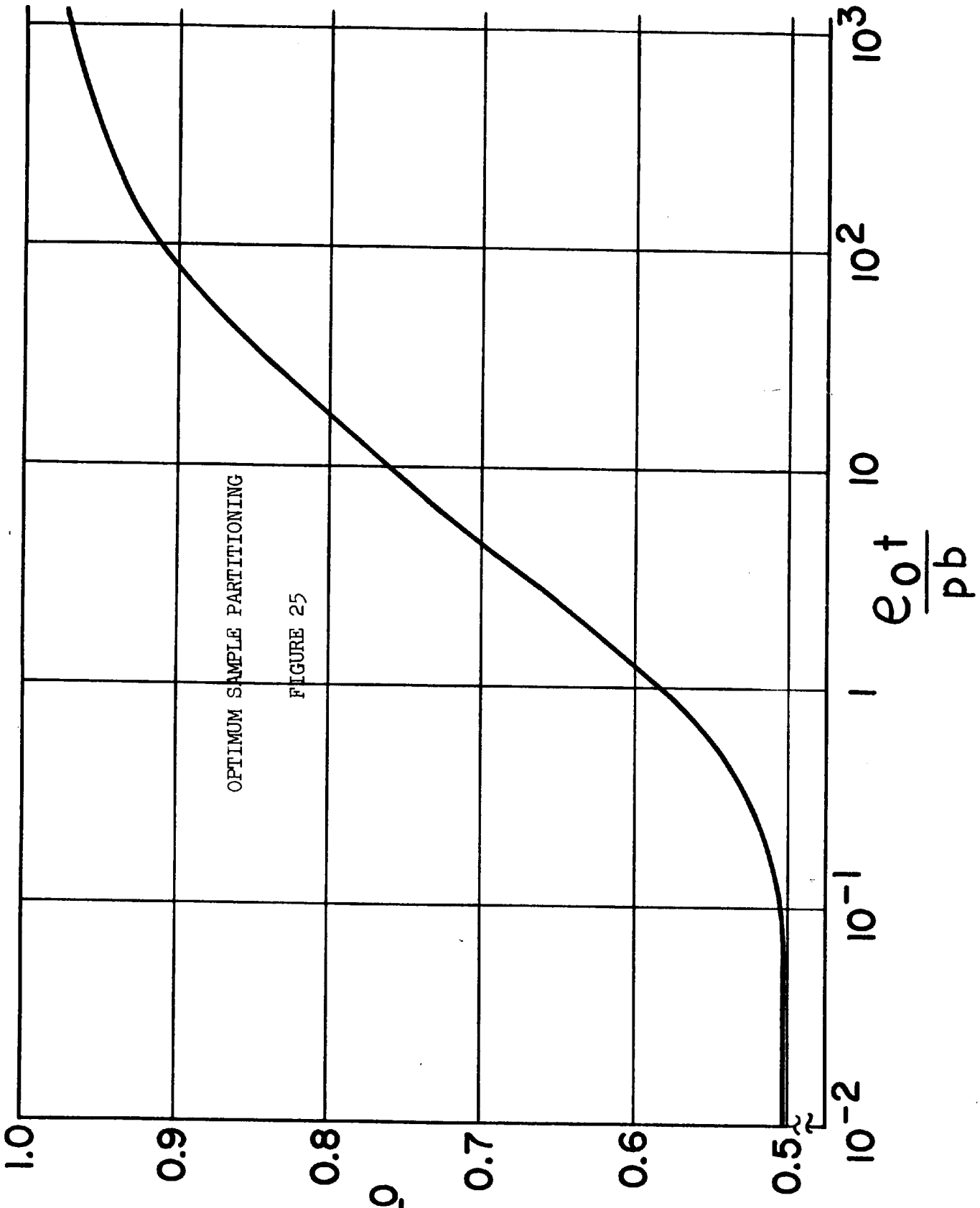
We wish to choose n so that (8.20) is minimized. Differentiating (8.20) and equating to zero, one obtains

$$\frac{e_o t}{pb} = \frac{2\frac{n_o}{t} - 1}{\left(1 - \frac{n_o}{t}\right)^2}, \qquad (8.21)$$

where n_o is that value of n satisfying (8.21); it is the optimum test sample size in the sense previously discussed. $\frac{n_o}{t}$ is of course the proportion of the total sample used for the test. One interesting result is immediately obvious: $\frac{n_o}{t}$ must be greater than .5 for all cases. The equation (8.21) is plotted in Figure 25, from which the following general statements can be made.

1. The proportion of the total sample that should be used to test the machine should never be less than 50%.
2. If $e_o t/pb < 0.1$, then the proportion used for design should be about 50%.
3. The proportion of the total sample that should be used to test the machine becomes larger as:
   a. The total sample size increases,
   b. the error of the optimum machine increases,
   c. the effectiveness of the design increases (pb decreases).

Here 1/pb is taken as a measure of the effectiveness of the design, since pb is the product of the expected deviation from optimum, $E[e-e_o]$, and the design sample size, pm.



OPTIMUM SAMPLE PARTITIONING

FIGURE 25

These results indicate just how a sample should be split between the design and test stages of a feasibility study of a pattern recognition method. If the experimenter follows this procedure, he will obtain an estimate $\hat{e}_o$ of e_o which is unbiased and has minimum variance.

The value of this minimum variance can be expressed as

$$\sigma^2_{\hat{e}_{o_{min}}} = \frac{e_o(1-e_o)}{n}\left(1 + \frac{1 - \frac{n_o}{t}}{2\frac{n_o}{t} - 1}\right),$$

which was obtained by eliminating pb between (8.20) and (8.21). Note that this is the variance that would have been obtained if the optimum machine were tested with n samples, increased by a factor which accounts for the design error.

As an illustration of these ideas, consider the following example (perhaps the simplest of the n-dimensional problems). A pattern recognition machine is to be designed using the optimum decision function (see Chapter II) which will distinguish between q classes. The occurrence of each class is equally probably a priori, and all costs of misrecognition are the same. The receptor makes a set of k measurements m_j, $1 \leq j \leq k$, on each input pattern. It is known that each measurement is normally distributed with variance σ, and that all measurements are independent. Further, it is known that the distances between the mean vectors in

measurement space are all equal. (Consequently, there can be no more than k+1 pattern classes. The tips of the mean vectors are the vertices of a regular polytope.)

Consequently, the distribution of the measurements for each of the classes is spherically symmetric and unimodal. We know then, from Theorem 7, that the optimum decision function is a linear decision function comprised of those hyperplanes which are the perpendicular bisectors of the line segments joining all pairs of means. (This is true even for the multiple class case, providing no rejection decision is required. There will also be no natural rejection regions, since this linear decision function is also an optimum decision function with no rejection decision.) The hyperplane separating two classes, say classes 1 and 2, is given by Theorem 8, and is the set of all points $\overline{X}$ which satisfy

$$\overline{X}\cdot(\overline{\mu}_1-\overline{\mu}_2) = \frac{1}{2}(\overline{\mu}_1\cdot\overline{\mu}_1-\overline{\mu}_2\cdot\overline{\mu}_2) , \qquad (8.22)$$

where $\overline{\mu}_i$ is the mean vector of class i.

The design procedure consists of estimating each mean vector from a sampling; denote the estimated mean vector for class i by $\overline{x}_i$. The distribution of the estimate of a mean vector from a normal distribution with covariance matrix [V] is also normal with covariance matrix $\frac{1}{m}$[V], where m is the sample size used in the estimate.[2] Since the measurements are independent in this case, then so will be the

estimates of the means of the various measurements. Furthermore, each estimate will have a variance of σ^2/m. Consequently, only one set of samples of size m from each pattern class is required to insure that the form (8.15) is valid, and p is hence equal to the number of allowable pattern classes, q.

We now determine the coefficient b in equation (8.15). If the mean vectors are more than about 3σ apart, then only a small error is made if the total error is approximated by adding the errors of each hyperplane taken alone. That is, the integrals on the wrong side of the hyperplane that are counted more than once will be quite small compared to the integrals counted only once (this is discussed in more detail in the proof of Theorem 6).

Due to the symmetry of the problem, the error associated with each hyperplane for the optimum decision function is identical, and the derivatives of (8.14) will also be identical for each hyperplane. Since there are $q(q-1)/2$ hyperplanes, b may be expressed (from (8.14) and (8.15)) as

$$\frac{b}{m} = \frac{q(q-1)}{2} \frac{1}{2} \sum_{i=1}^{k} \left[\frac{\partial^2 e_{12}}{\partial \bar{x}_{i1}^2} \bigg|_{\mu_1,\mu_2} + \frac{\partial^2 e_{12}}{\partial x_{i2}^2} \bigg|_{\mu_1,\mu_2} \right] \frac{\sigma^2}{m} , \tag{8.23}$$

where the hyperplane separating classes 1 and 2 is taken as typical, and the independence of the estimates is used. e_{12} is the error associated with this hyperplane, μ_1 and μ_2 are the mean vectors of these classes, and $\bar{x}_1$ and $\bar{x}_2$ are the estimates of the mean vectors.

There is no loss in generality if μ_1 is taken as zero, and all the components of $\mu_2(\mu_{12},\ldots,\mu_{k2})$ are taken as zero except for μ_{12}. That is,

$$\mu_1 = (0,0,\ldots,0)$$

$$\mu_2 = (\mu,0,\ldots,0) ,$$

where μ_{12} is denoted μ, $\mu > 0$. Consequently, the optimum boundary is given by

$$x_1 = \mu/2 .$$

A sampling of size m is taken from each class, and the mean vectors are estimated, giving

$$\bar{x}_1 = (\bar{x}_{11},\bar{x}_{21},\ldots,\bar{x}_{k1})$$

$$\bar{x}_2 = (\bar{x}_{12},\bar{x}_{22},\ldots,\bar{x}_{k2}) .$$

A boundary given by (8.22) is computed based on the above estimates, and this, together with the other estimated boundaries, determines the structure of the machine.

The error e_1 associated with this particular boundary for class 1 is

$$e_1 = \prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2} dx_j \int_{\xi_1(x_2,\ldots,x_k)}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_1}{\sigma}\right)^2} dx_1 ,$$

where $\xi_1(x_2,\ldots,x_k)$ is the value of x_1 on the boundary, and is given by (from (8.22))

$$\xi_1(x_2,\ldots,x_n) = -\sum_{i=2}^{k} \frac{\bar{x}_{i1}-\bar{x}_{i2}}{\bar{x}_{11}-\bar{x}_{12}} x_i + \frac{1}{2} \sum_{i=1}^{k} \frac{\left(\bar{x}_{i1}^2-\bar{x}_{i2}^2\right)}{\bar{x}_{11}-\bar{x}_{12}}$$

$$= \frac{\bar{x}_{11}+\bar{x}_{12}}{2} - \frac{1}{2} \sum_{i=2}^{k} \frac{2\left(\bar{x}_{i1}-\bar{x}_{i2}\right)x_i - \left(\bar{x}_{i1}^2-\bar{x}_{i2}^2\right)}{\bar{x}_{11} - \bar{x}_{12}} .$$

Then

$$\frac{\partial e_1}{\partial \bar{x}_{i1}} = \prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2} dx_j \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\xi_1}{\sigma}\right)^2} \right) \left(\frac{x_1-\bar{x}_{i1}}{\bar{x}_{11}-\bar{x}_{12}} \right) ,$$

$$2 \le i \le n .$$

$$\frac{\partial^2 e_1}{\partial \bar{x}_{i1}^2} = \prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2} dx_j \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\xi_1}{\sigma}\right)^2} \right)$$

$$\left[\frac{\xi_1}{\sigma^2} \left(\frac{x_i-\bar{x}_{i1}}{\bar{x}_{11}-\bar{x}_{12}} \right)^2 - \left(\frac{1}{\bar{x}_{11}-\bar{x}_{12}} \right) \right] ,$$

$$2 \le i \le n .$$

$$\left.\frac{\partial^2 e_1}{\partial \bar{x}_{11}^2}\right|_{\mu_1,\mu_2} = \left(\frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{1}{2}\left(\frac{\mu/2}{\sigma}\right)^2}\right) \prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2} dx_j \left[-\frac{x_1}{2\sigma^2} + \frac{1}{\mu}\right]$$

$$= \frac{1}{\sigma} N\left(\frac{\mu}{2\sigma}\right)\left[-\frac{1}{2\sigma^2} E[x_i] + \frac{1}{\mu}\right] = \frac{1}{\mu\sigma} N\left(\frac{\mu}{2\sigma}\right),$$

$$2 \le i \le n,$$

where $N\left(\frac{\mu}{2\sigma}\right)$ is the value of the standard normal density function for the variate $\mu/2\sigma$. In a like manner,

$$\left.\frac{\partial^2 e_2}{\partial \bar{x}_{11}^2}\right|_{\mu_1,\mu_2} = -\frac{1}{\mu\sigma} N\left(-\frac{\mu}{2\sigma}\right) = -\frac{1}{\mu\sigma} N\left(\frac{\mu}{2\sigma}\right), \qquad 2 \le i \le n,$$

where e_2 is the error associated with this boundary for class 2. Since the total error for this boundary is $e_{12} = e_1 + e_2$, then

$$\left.\frac{\partial^2 e_{12}}{\partial \bar{x}_{11}^2}\right|_{\mu_1,\mu_2} = \left.\frac{\partial^2 e_1}{\partial \bar{x}_{11}^2}\right|_{\mu_1,\mu_2} + \left.\frac{\partial^2 e_2}{\partial \bar{x}_{11}^2}\right|_{\mu_1,\mu_2} = 0, \qquad 2 \le i \le n.$$

A like result holds for $\frac{\partial^2 e}{\partial \bar{x}_{12}^2}$, $2 \le i \le n$. Going through this same procedure for $\bar{x}_{11}$,

$$\frac{\partial e_1}{\partial \bar{x}_{11}} = -\prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2} dx_j \left[\frac{1}{2\sigma} N\left(\frac{\xi_1}{\sigma}\right)\right].$$

$$\frac{\partial^2 e_1}{\partial \bar{x}_{11}^2} = - \prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2} dx_j \left[- \frac{1}{4\sigma}\left(\frac{\xi_1}{\sigma^2}\right) N\left(\frac{\xi_1}{\sigma}\right) \right] .$$

$$\left. \frac{\partial^2 e_1}{\partial \bar{x}_{11}^2} \right|_{\mu_1, \mu_2} = \frac{1}{8} \frac{\mu}{\sigma^3} N\left(\frac{\mu}{2\sigma}\right) .$$

Similarly

$$\left. \frac{\partial^2 e_2}{\partial \bar{x}_{11}^2} \right|_{\mu_1, \mu_2} = \frac{1}{8} \frac{\mu}{\sigma^3} N\left(\frac{\mu}{2\sigma}\right) .$$

Hence

$$\left. \frac{\partial^2 e_{12}}{\partial \bar{x}_{11}^2} \right|_{\mu_1, \mu_2} = \frac{1}{4} \frac{\mu}{\sigma^3} N\left(\frac{\mu}{2\sigma}\right) .$$

It would also be found that

$$\left. \frac{\partial^2 e_{12}}{\partial \bar{x}_{22}^2} \right|_{\mu_1, \mu_2} = \frac{1}{4} \frac{\mu}{\sigma^3} N\left(\frac{\mu}{2\sigma}\right) .$$

This analysis is perfectly general for arbitrary mean vectors, providing that μ is merely interpreted as the distance between a pair of mean vectors (all such distances being assumed identical). This distance will henceforth be written $\Delta\mu$ to indicate that it is a difference of means.

Therefore, from (8.23), we find that

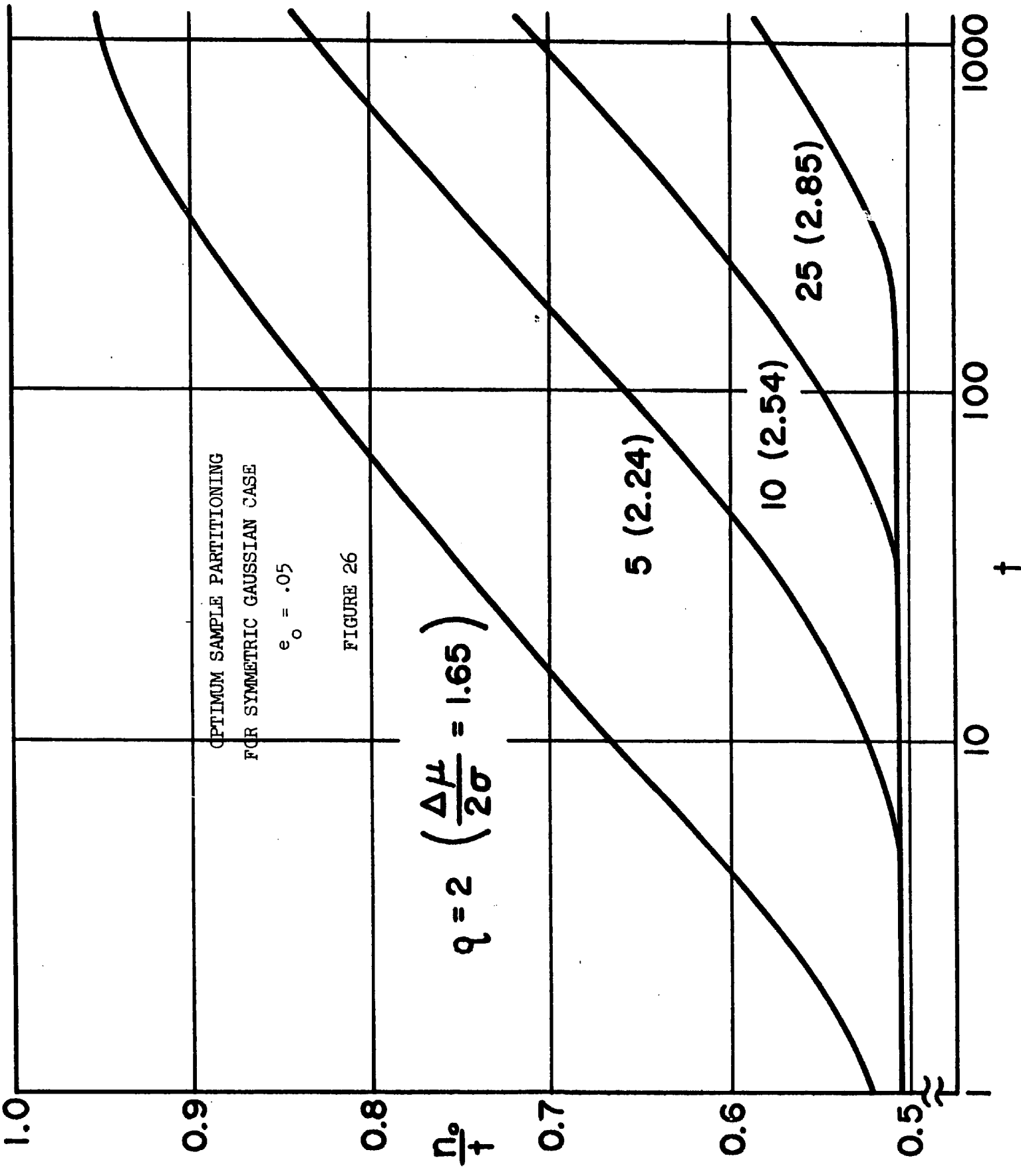
$$b = \frac{q(q-1)}{4} \frac{\Delta\mu}{2\sigma} N\left(\frac{\Delta\mu}{2\sigma}\right) .$$

The equation (8.21) becomes

$$\frac{4e_o t}{q^2(q-1) \frac{\Delta\mu}{2\sigma} N\left(\frac{\Delta\mu}{2\sigma}\right)} = \frac{2\frac{n_o}{t} - 1}{\left(1 - \frac{n_o}{t}\right)^2} . \qquad (8.24)$$

Some curves representing (8.24) are plotted in Figure 26 in which the proportion of the total sample to be used in the test, n_o/t, is shown as a function of t, the total sample size, with the number of allowable pattern classes, q, as a parameter. e_o was held constant at .05, which involves the choosing of the proper value of $\Delta\mu/2\sigma$ for each q. This is done as follows. The conditional error for each class, if the a priori probabilities are equal, is e_o. If there are q classes, then the conditional error associated with this class being categorized as any of the other (q-1) classes must be $e_o/q-1$, since all of these conditional errors are identical. Let us consider class 1, with mean μ_1. Let x be in the direction of the line segment joining μ_1 and μ_2 of class 2. Then the probability of miscategorizing a member of class 1 as class 2 is (approximately)

$$\frac{e_o}{p-1} = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu_1 + \frac{\Delta\mu}{2}}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} dx = \frac{1}{\sqrt{2\pi}} \int_{\frac{\Delta\mu}{2\sigma}}^{\infty} e^{-\frac{1}{2}y^2} dy ,$$



OPTIMUM SAMPLE PARTITIONING
FOR SYMMETRIC GAUSSIAN CASE
$e_o = .05$

FIGURE 26

where $y = \frac{x-\mu_1}{\sigma}$. From this relation, $\Delta\mu/2\sigma$ can be determined, given values for e_o and p, from tables of the standard normal distribution.

From Figure 26 it is seen that, for many cases, the sample should be split evenly between design and test, as one might intuitively suspect. However, there are some drastic deviations from this. For instance, if the categorizer is to separate only two classes, and 1000 samples are available, then only 50 of these should be used to design the machine, and 950 should be used to test it. Consequently, it is seen that intuition may go wrong in some cases.

This section has not solved the problem of sample partitioning. One problem is the determination of b, which in many cases will be very difficult to find. It would also be interesting to consider the case in which there is an overlap between the design and test samples. This discussion has, however, developed one approach to the problem and illustrated that intuition is often, but not always, a good guide.

CHAPTER IX

EXPERIMENTAL APPLICATION - DETERMINATION OF THE GEOGRAPHICAL SOURCE OF RADIO SIGNALS

This chapter and the next will describe two experimental applications of linear decision functions to categorization problems. The experiment described in this chapter is small enough so that the data and results can be described in detail; Chapter X discusses an application much larger in scope, and hence only gross results are given.

Between these two experiments, most of the concepts and procedures as discussed previously are applied. In the application described in this chapter, the estimate to the optimum linear decision function is found, and the upper bound for the probability of error-plus-natural-rejection is determined by the method given in Section 8.2. Also, a linear rejection criterion is applied. The results of this experiment are compared to one in which the same data was categorized by using the optimum decision function based upon an assumption of normally distributed, independent measurements.

In the experiment described in the next chapter, the optimum linear decision function is also estimated, and is minimized by eliminating redundant boundaries and redundant measurements. The resulting incomplete linear decision function is then tested in two ways:

1. by using an additional test sample different from the design sample, and
2. by estimating the upper bound for the error-plus-natural-rejection probability.

9.1 Description of the Application

The problem to be described here is one studied by Professor A. E. Laemmel of the Polytechnic Institute of Brooklyn. A radio signal is received at a monitoring station, and it is desired to determine from which geographical location this signal originated. It is assumed that there are a finite number of sources whose geographical locations are known. Certain measurements are made on a sampling from these stations. The problem is to design a categorizer based on these samples.

The measurements chosen (by Professor Laemmel) are based on measuring certain fading characteristics of the received wave. The output of the automatic gain control (agc) of the receiver is monitored for a 50 second interval, and the following measurements are made relative to the peak output during this interval:

m_1 - number of seconds during which the agc output is greater than one-half of its peak value;

m_2 - number of crossings of the half-peak level by the agc output.

m_3 - number of seconds during which the agc output is greater than three-quarters of its peak value.

m_4 - number of seconds during which the agc output is less than one-quarter of its peak value.

m_5 - duration (in seconds) of the maximum interval during which the agc output is greater than one-half of its peak value.

m_6 - number of crossings of the three-quarter level during the maximum interval measured by m_5.

Note that this is an example of a receptor which makes both continuous (m_1, m_3, m_4, m_5) and discrete (m_2, m_6) measurements.

9.2 Results

These measurements were made on five geographical locations (the allowable pattern classes): Ohio, Canada, Quito (Ecuador), London, and Lisbon (Portugal). A sample size of five was taken from each source over a period of time, and the resulting measurements (furnished by Professor Laemmel) are shown in Table 2.

9.2.1 The Estimated Linear Decision Function

The linear decision function was determined by using the algorithm of section 5.3 to estimate each of the constituent hyperplanes. In all cases, the initial hyperplane was the perpendicular bisector of the line segment joining the means of the two classes. The ten resulting 6-dimensional hyperplanes are shown in Table 3. x_i is the coordinate representing the measurement m_i, $1 \leq i \leq 6$. Each hyperplane is identified by using the numbering arrangement of Table 2,

			m_1	m_2	m_3	m_4	m_5	m_6
5.	Ohio	1	50.0	0	50.0	0.0	50.0	0
		2	50.0	0	50.0	0.0	50.0	0
		3	50.0	0	43.0	0.0	48.0	1
		4	50.0	0	50.0	0.0	50.0	0
		5	24.6	9	19.3	10.4	12.6	2
4.	Canada	1	47.4	19	31.8	0.0	16.7	12
		2	49.5	2	32.7	0.0	37.7	39
		3	44.8	36	23.6	0.0	11.6	16
		4	50.0	0	39.7	0.0	50.0	7
		5	44.4	34	26.3	0.6	25.0	25
3.	Quito	1	22.6	5	8.6	12.6	10.8	2
		2	23.2	10	8.1	5.2	11.9	2
		3	24.8	9	15.1	1.3	9.0	3
		4	29.3	7	10.8	5.4	16.6	5
		5	15.9	26	4.8	18.3	4.0	4
2.	London	1	9.3	114	0.8	17.9	0.5	2
		2	38.0	51	17.7	3.6	4.3	2
		*3	24.7	29	10.0	9.3	4.7	2
		4	39.3	41	15.8	2.3	14.7	20
		5	38.5	68	15.6	2.7	3.5	4
1.	Lisbon	1	22.8	58	7.6	10.3	2.2	2
		2	40.2	39	17.5	1.0	5.6	2
		3	36.3	36	16.4	3.7	13.4	10
		4	23.0	60	8.7	13.2	3.7	6
		5	32.0	41	16.8	6.1	8.7	8

THE SAMPLE UPON WHICH THE LINEAR DECISION FUNCTION OF CHAPTER IX IS BASED

TABLE 2

B_{54}: $-.22x_1 - .22x_2 + .27x_3 + .53x_4 + .06x_5 - .73x_6 + .03 = 0$

B_{53}: $-.54x_1 - .21x_2 + .75x_3 + .19x_4 + .16x_5 - .20x_6 - .11 = 0$

B_{52}: $.13x_1 - .76x_2 + .46x_3 + .12x_4 + .41x_5 - .14x_6 + .01 = 0$

B_{51}: $.10x_1 - .76x_2 + .43x_3 + .29x_4 + .36x_5 - .13x_6 + .02 = 0$

B_{43}: $-.31x_1 + .02x_2 + .33x_3 - .75x_4 + .04x_5 + .48x_6 - .14 = 0$

B_{42}: $.32x_1 - .71x_2 + .53x_3 - .28x_4 + .16x_5 + .11x_6 - .01 = 0$

B_{41}: $.25x_1 - .72x_2 + .36x_3 - .21x_4 + .31x_5 + .38x_6 - .01 = 0$

B_{32}: $.10x_1 - .54x_2 + .10x_3 + .74x_4 + .36x_5 + .08x_6 + .22 = 0$

B_{31}: $.22x_1 - .53x_2 + .04x_3 + .77x_4 + .27x_5 - .09x_6 + .13 = 0$

B_{21}: $-.13x_1 + .22x_2 - .12x_3 - .90x_4 - .19x_5 + .29x_6 - .12 = 0$

THE ESTIMATE OF THE OPTIMUM LINEAR DECISION FUNCTION, BASED ON THE DATA OF TABLE 2

TABLE 3

i.e., B_{42} is the hyperplane which separates the sample points for Canada from those for London. The first number in the subscript of B_{ij} corresponds to the plus side of the hyperplane. That is, the half-space consisting of all those points which are a positive distance from B_{42} are classified as belonging to Canada by B_{42}; the other half-space (for negative distances) contains all points classified as London by B_{42}.

No attempt was made to determine whether any of these boundaries were redundant, nor whether any of the measurements were redundant. Note, however, that the maximum absolute value of the direction cosine associated with each coordinate is large (>0.4). Therefore, one might conclude that all of the measurements are significant (see section 6.1).

9.2.2 Error Estimates

The complete linear decision function of Table 3 categorized all but one of the sample points from Table 2 correctly. The misclassified point was one representing London, and was classified as coming from Lisbon. One cannot claim, however, that this experimental error rate, i.e., 4%, represents in any way the actual expected error rate for the linear decision function. Theorem 10 states that, in 6 dimensions, 7 nondegenerate points can always be separated by a hyperplane. In this example, only 10 points were being separated by each hyperplane. Since

this is quite close to the "threshold" predicted by Theorem 10, one is not surprised to find that the linear decision function works well in this case.

A better estimate of the expected error probability can be obtained by estimating the upper bound as described in Section 8.2. These results are shown in the form of a confusion matrix in Table 4. Each entry corresponds to the area under the estimated normal density function $\eta_i(s)$ (for distances of the points representing class i from the hyperplane separating class i from class j) which lies on the j side of B_{ij}. Note that the one error (London categorized as Lisbon) corresponds to the largest estimated upper bound in Table 4. If each geographical source is considered equally probable a priori, then the estimated upper bound for the error-plus-natural-rejection probability is the average of the total probabilities of error for each class. This upper bound is .30. (Conversely, one may say that the estimated lower bound for the recognition rate is 70%.) Unfortunately, reference to Figure 24 shows that this estimate is not a very reliable one. For instance, if $\bar{x}/s$ were 3.0, one would say that the estimate of the error rate (for that class and hyperplane) would be .13%. However, the 95% confidence interval for this estimate, from Figure 24, is 0%-17.7%! The extremely small sample size (five per class) used yields very poor estimates.

RADIO SIGNAL FROM	IDENTIFIED AS: OHIO	CANADA	QUITO	LONDON	LISBON
OHIO		.06	.11	.71	.55
CANADA	4.01		14.0	8.69	9.18
QUITO	.51	4.65		4.95	2.94
LONDON	11.9	19.1	11.7		36.7
LISBON	1.92	10.4	2.22	4.95	

Estimated Upper Bound on System Error = 30%

ESTIMATED UPPER BOUNDS FOR THE VARIOUS CLASSIFICATION ERRORS IN THE GEOGRAPHICAL LOCATION OF RADIO SIGNALS, IN PER CENT

TABLE 4

It is interesting to compare these results with those of a trial by Professor Laemmel, in which the distribution of a particular measurement for a particular class was assumed to be normally distributed, and each measure was assumed independent of the others. The various means and variances were estimated from the sample of Table 2, and the a posteriori probability of each point coming from each class was calculated. Categorization was based on maximizing this probability over the set of classes (the optimum decision rule under the above assumptions if it is also assumed that the a priori probabilities of occurrence and the misrecognition costs are equal). Using this procedure, four points from the sample were misclassified.

9.2.3 Application of a Rejection Criterion

The linear rejection criterion as described in Section 7.2 was applied to this linear decision function. The rejection hyperplanes were determined by the condition that the loss associated with an error was ten times as large as the loss associated with a rejection:

$$c_{ij} = c = 10$$
$$c_{io} = c_o = 1 .$$

According to equation (7.2), (assuming equal a priori probabilities ω_i), the rejection region around the hyperplane B_{ij} corresponds to that region for which

$$\frac{1}{9} \leq \frac{\eta_1(s)}{\eta_j(s)} \leq 9 \ . \qquad (9.1)$$

The equations (9.1) were solved for this case. The resulting separation of the rejection hyperplanes from the estimated optimum hyperplanes are shown in Table 5. $\Delta\alpha_{o-}$ is the distance of the rejection hyperplane on the negative (j) side of B_{ij}; $\Delta\alpha_{o+}$ is the corresponding distance on the positive (i) side. The fact that, in some cases, both rejection hyperplanes are on the same side of B_{ij} (i.e., those with the same sign in Table 5) can be attributed to either sampling error or deviations from normality of the distributions of the distances.

Note that one rejection hyperplane ($\Delta\alpha_{o-}$ for B_{21}) is at infinity. This means physically that any attempted categorization of a signal as coming from London would always be rejected, since the likelihood of its originating in London rather than in Lisbon will always be less than 9 (the rejection condition (9.1)).

This rejection boundary was replaced by an arbitrary (hence suboptimum) rejection hyperplane at $\Delta\alpha_{o-} = -.7$ (which seemed not too unreasonable from a plot of the distributions). Based on this modified set of rejection hyperplanes, the rejection and error probabilities for each class-hyperplane combination were computed. These are shown in the confusion matrix of Table 6, where the upper number represents the

Boundary	$\Delta\alpha_{o-}$	$\Delta\alpha_{o+}$
B_{54}	- .85	+ 1.4
B_{53}	+ .10	+ 1.9
B_{52}	+ 2.2	+19
B_{51}	+ 2.0	+ 9.0
B_{43}	- 5.7	+ 2.8
B_{42}	-17.9	+31.4
B_{41}	- 7.2	+19.0
B_{32}	- 5.3	+ 4.7
B_{31}	- 2.5	+ 3.0
B_{21}	- ∞	+ 1.4

SEPARATION OF REJECTION HYPERPLANES
FROM ESTIMATED OPTIMUM HYPERPLANES

TABLE 5

RADIO SIGNAL FROM	(IDENTIFIED AS) (REJECTED AS POSSIBLY BEING) OHIO	CANADA	QUITO	LONDON	LISBON
OHIO		0.0 1.5	.11 .28	1.1 8.4	.80 1.9
CANADA	2.9 3.4		1.9 25.5	.99 63.4	5.9 31.1
QUITO	.02 .42	1.3 27.5		2.6 25.9	.73 10.0
LONDON	3.2 8.1	3.8 37.5	7.4 11.6		20.6 28.6
LISBON	.20 1.0	3.6 18.8	.49 5.6	.05 20.5	

Estimated Upper Bound on System Error = 10.4%
Estimated Upper Bound on System Rejection = 66.0%

RESULTS OF APPLICATION OF LINEAR REJECTION CRITERION TO GEOGRAPHICAL LOCATION OF RADIO SIGNALS, WITH $c/c_o = 10$.
UPPER NUMBERS ARE ERROR PROBABILITIES;
LOWER NUMBERS ARE REJECTION PROBABILITIES,
IN PER CENT

TABLE 6

per cent error and the lower number the per cent rejection. Averaging these (assuming equal a priori probabilities) gives an estimated upper bound for the system error probability of .10, and an estimated upper bound for the system rejection probability of .66.

If one wanted to determine the optimum linear rejection criterion for a fixed error rate or fixed rejection rate (for instance, design for the minimum rejection rate which will yield an error rate of 1%), then he would have to try several values of the loss ratio c/c_o and obtain plots of the error rate and rejection rate versus c/c_o. From these plots, the appropriate loss ratio could be determined.

CHAPTER X

EXPERIMENTAL APPLICATION - THE RECOGNITION OF HAND-PRINTED NUMBERS

The recognition of hand-printed numbers was attempted with a linear decision function. The set of measurements which was used involved quantizing the number into a 12 x 12 binary matrix. A matrix element was given a weight of one if it contained a mark and weight of zero if it contained no mark. The quantized number was then positioned in the matrix by aligning its center-of-gravity with the center of the matrix.

Hence, a 144-dimensional binary measurement space was used. This set of measurements is a rather unsophisticated set in that the measures are not at all invariant within a particular class. That is, in order for a linear decision function to perform well, those points in measurement space corresponding to a particular class ought to be grouped together with respect to points representing other classes. This will occur if the measurements (or at least some of the measurements) are somewhat invariant under the various distortions and noises that might affect a real life pattern. In the case of hand-printing, these perturbations from ideal include size variations, tilt, varying pencil width and density, the various forms that people use to form a character,

and the effects of sloppiness. Clearly the measures used here are in no way invariant under such perturbations, and one would not be too surprised if a linear decision function did not perform very well.* However, the attempt is still interesting since it represents a more complex problem than that discussed in the previous chapter, and will consequently allow the testing of some of the preceding ideas in more detail.

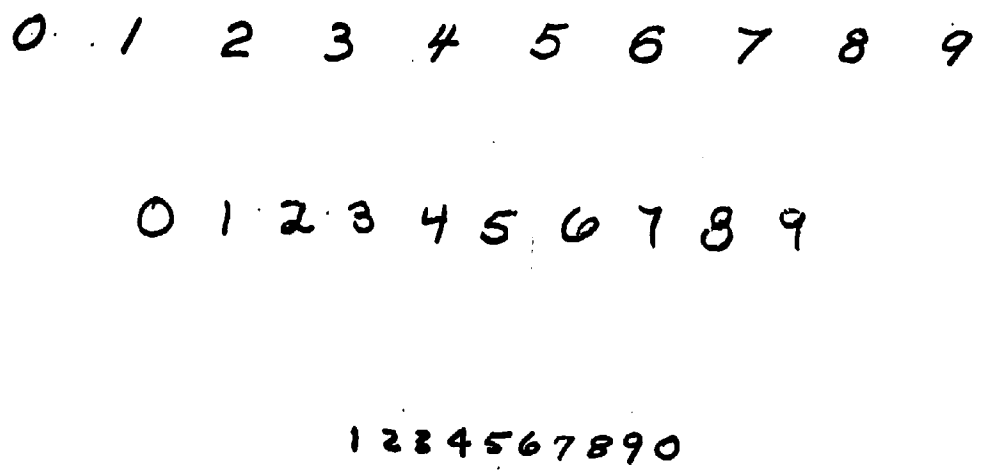
10.1 Estimating the Linear Decision Function

The data used to estimate the optimum linear decision function was gathered in the following manner. A subject was asked to neatly print the ten numbers on a piece of quad-ruled paper at a size approximating the ruled boxes. Fifty different people were so asked, resulting in a sample size of 50 for each of the ten pattern classes. This data was then automatically reduced to a 12 x 12 matrix (encoded on IBM punched cards) by an optical matrix scanner constructed by the author.

In Figure 27 is shown an example of some of this design data, illustrating approximately the range of size and neatness obtained. In Figure 28 are shown examples of some of the quantized numbers.

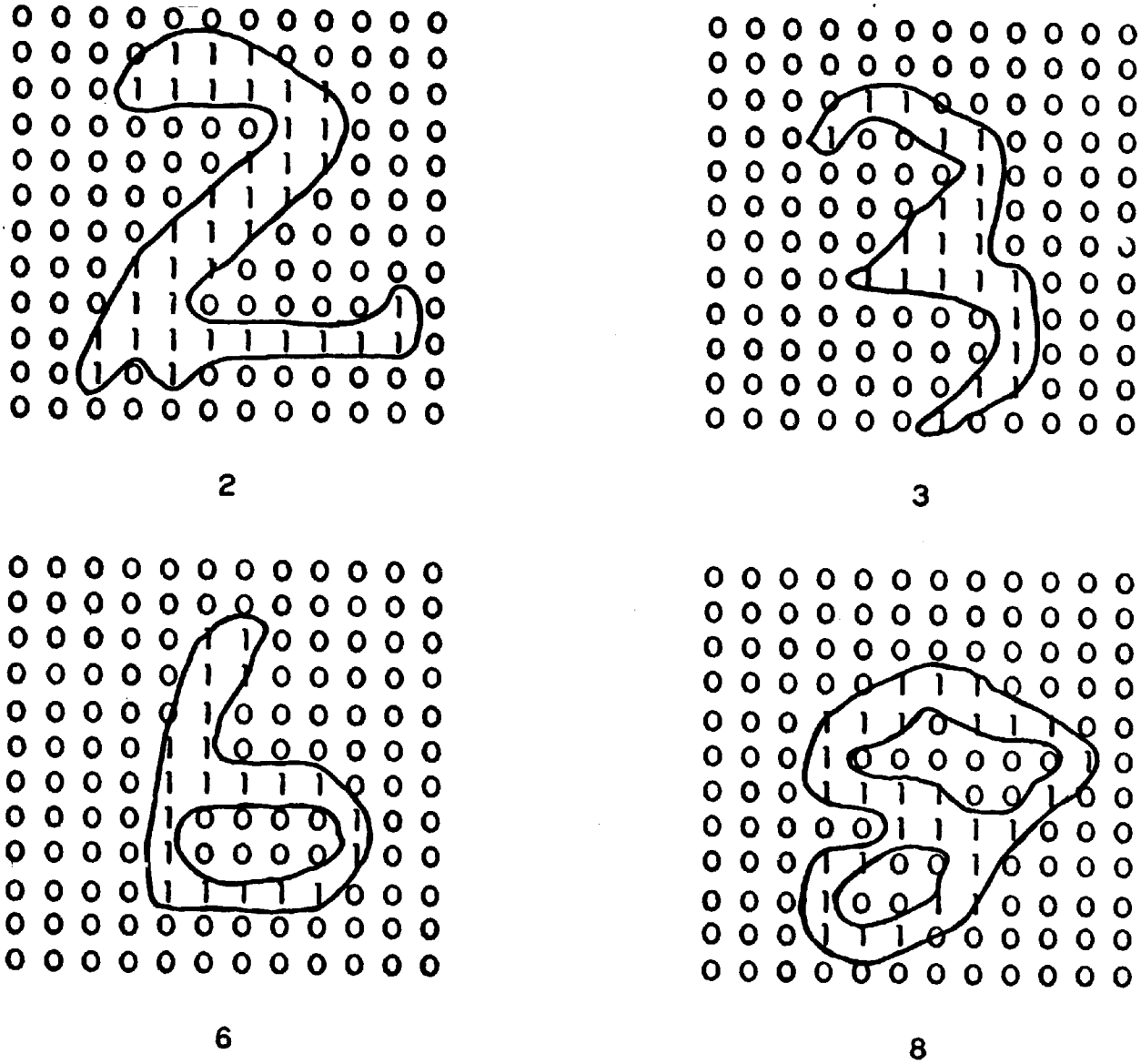
- -

*A very effective set of measurements has been proposed by Kamentsky [29] for the recognition of hand-printed numbers. This involves using a "flying-polar" scan which is capable of determining the number of closures and cusps (partial closures) and the orientation of cusps in a character.



SOME EXAMPLES OF THE
HAND-PRINTING DESIGN DATA

FIGURE 27



EXAMPLES OF QUANTIZED FORMS
OF THE HAND-PRINTED NUMBERS

FIGURE 28

Forty-five hyperplanes are required in the complete linear decision function categorizing the ten numbers. These were determined by the computation algorithm (Section 5.3) on the IBM 7090 digital computer. About 25 seconds, on the average, was required to determine a hyperplane, given an initial position.

For each pair of pattern classes, four initial hyperplanes were tried. One of these was that hyperplane which was the perpendicular bisector of the line segment joining the means of the two classes. The other three initial hyperplanes were parallel to this one (i.e., the direction cosines were the same), and corresponded to an α_o of 0, -5, and +5. The number of successes* after iteration of each of these initial conditions is shown in Table 7, along with the number of trials in which each initial condition produced a unique minimum (that is, it separated the points better than the other three initial hyperplanes, after iterating to its minimum).

It is seen from this table that each of the initial conditions was often successful in reaching at least that absolute minimum determined by the set of initial conditions. More important, however, is the fact that each initial condition was the most successful in at least one trial; therefore, benefit was certainly derived from trying various initial conditions.

---

*An initial condition is successful if the performance of the resulting hyperplane is at least as good as the performance of the hyperplanes corresponding to the other initial conditions.

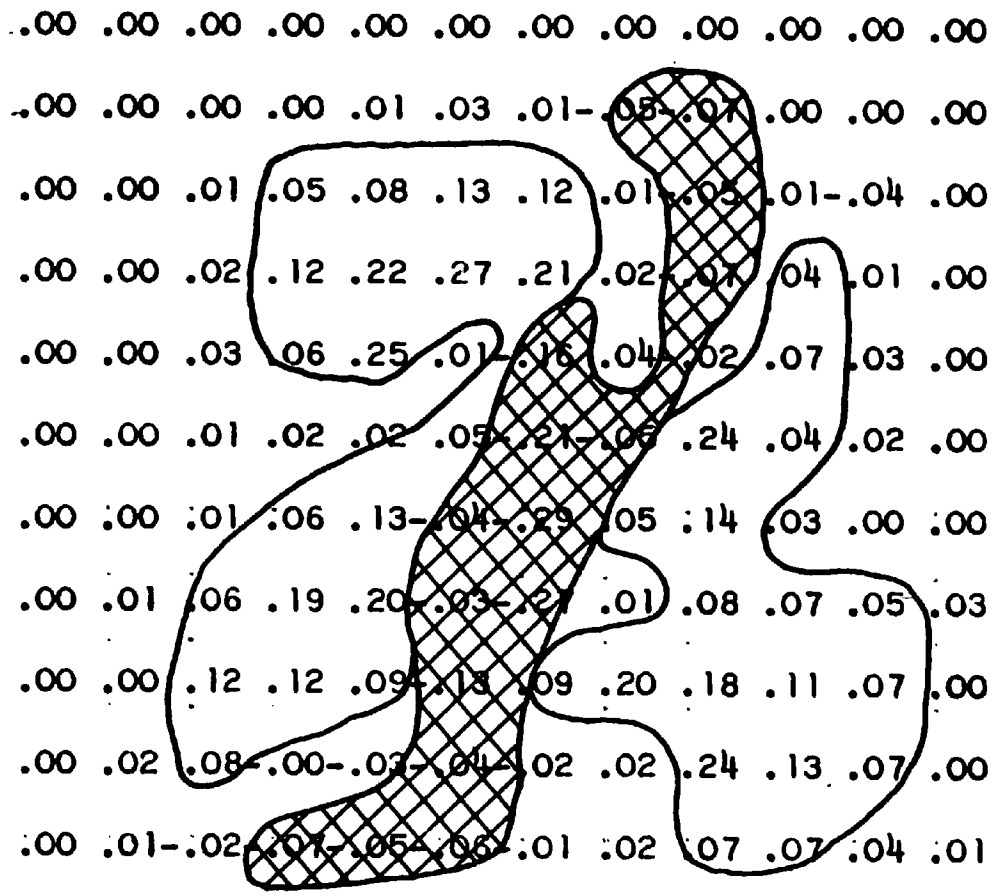
Initial Condition	No. of Successes	No. of Unique Successes
Perp. Bis.	26	4
$\alpha_o = 0$	32	5
$\alpha_o = +5$	29	6
$\alpha_o = -5$	17	1

THE SUCCESS RECORD OF
VARIOUS INITIAL HYPERPLANES

TABLE 7

In Figure 29 is shown the estimated optimum hyperplane, B_{21}, which separates the numbers 2 and 1. The coefficients α_i, $1 \leq i \leq n$, are shown arranged in a matrix corresponding to the receptor matrix. The positive side of B_{21} corresponds to the number 2. One would then expect that those coefficients which corresponded to matrix elements in which a mark from a two was likely to occur and a mark from a one was not likely to occur would be weighted positively, and vice versa for those elements in which a mark from a one is more likely to occur. Contours are drawn around regions of large positive and negative weight in Figure 29, and the negative regions are shaded. One sees that the above intuitive observation does indeed hold.

The resulting linear decision function mis-categorized 21 patterns (4.2%) and rejected 9 patterns (1.8%) of the total design sample of 500, as shown in the confusion matrix of Table 8 (the R column indicates the number of test patterns rejected by the inherent rejection). However, one cannot conclude that these percentages are any sort of valid estimate for the performance of the system, as discussed previously. In fact, since only 100 points are being separated in 144 dimensions by each hyperplane, one might expect from Theorem 11 that the linear decision function ought to do well on the design data; in fact, 50 samples from each class might be too small a sample size for design purposes for this reason. The saving grace here is the fact that the measurement space



$\alpha_o = .09$

THE HYPERPLANE B_{12}

FIGURE 29

RECOGNIZED AS

Input Class	0	9	8	7	6	5	4	3	2	1	R
0	48		1								1
9		41	2	1			3				3
8		1	45				2	1			1
7				47						2	1
6					49						1
5						50					
4	1	2	1				45				1
3			2					47	1		
2					1				49		
1										49	1

CORRECT	470	(94 %)
ERROR	21	(4.2%)
REJECT	9	(1.8%)

CONFUSION MATRIX FOR THE DESIGN DATA (HAND-PRINTING RECOGNITION PROBLEM). THE ENTRIES CORRESPOND TO THE NUMBER OF SAMPLES RECOGNIZED CORRECTLY, MISRECOGNIZED, OR REJECTED.

TABLE 8

is binary, and therefore the sample points are highly degenerate in the sense of Theorem 10. It is therefore not to be expected that any set of points, no greater in number than n+1 (145 in this case), will be linearly separable in general in this measurement space.

10.2 Minimizing the Linear Decision Function

The linear decision function thus determined was minimized by determining the redundant boundaries and the redundant measurements.

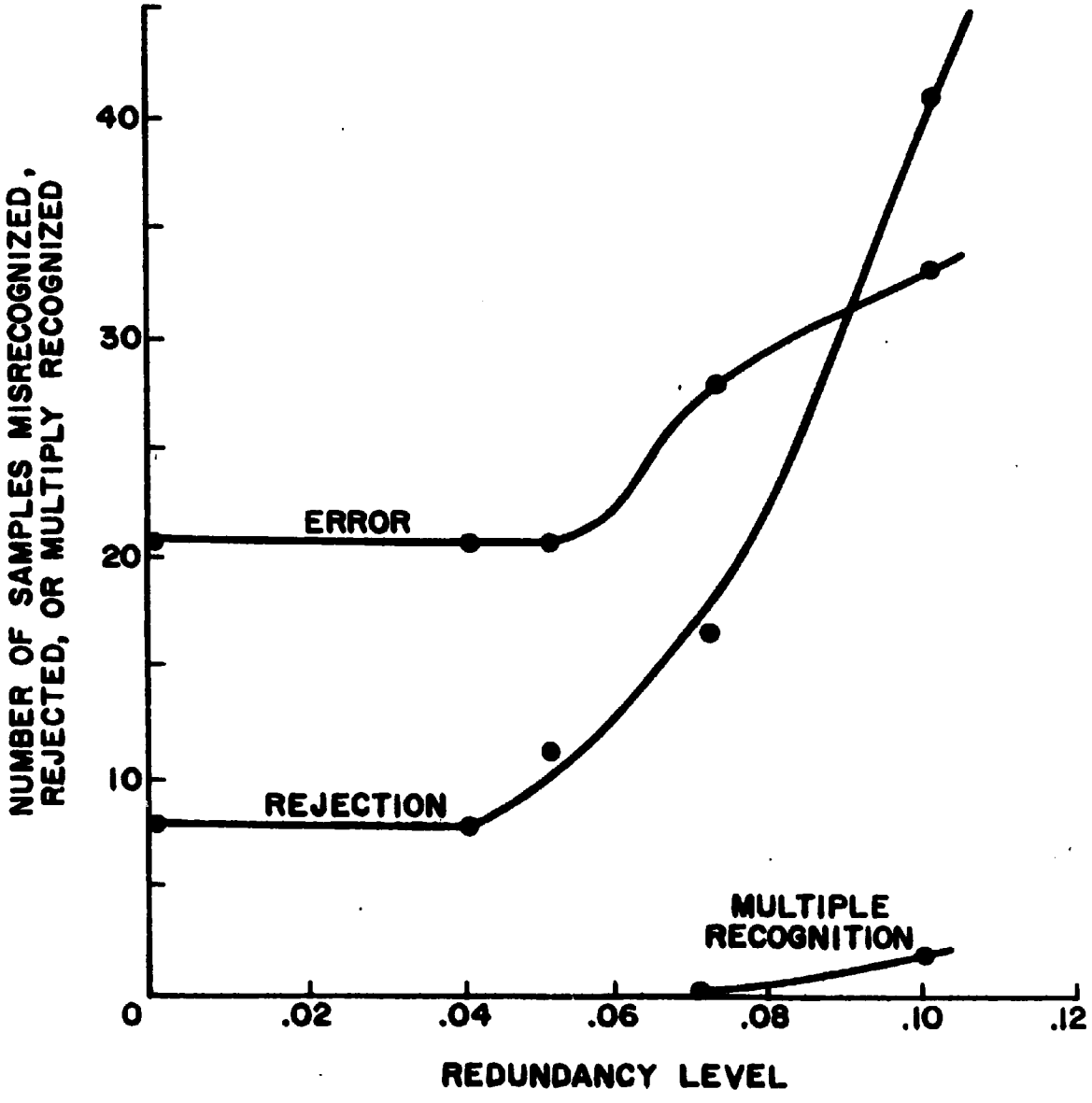
The procedure used for determining the redundant boundaries was the algorithm given in Section 6.2.2 based on the definition of redundancy in a sample sense. According to this algorithm, the hyperplanes were removed one at a time, and the conditionally redundant hyperplanes were determined (those whose removal caused no change in the categorization of the samples). Six hyperplanes were found to be conditionally redundant; they were B_{01}, B_{92}, B_{81}, B_{76}, B_{43}, and B_{41}. (Actually, the removal of B_{81} caused a "one" which was originally rejected to be correctly categorized. Since this was an improvement, it was decided to treat this hyperplane as conditionally redundant.)

It is seen, then, that the hyperplanes B_{92} and B_{76} are unconditionally redundant boundaries of the first kind, and may be definitely removed. These and the remaining four conditionally redundant hyperplanes were removed simultaneously, and the design samples were recategorized with the remaining

39 hyperplanes. Again there was no change in categorization (except the one good change due to B_{81}); therefore, B_{01}, B_{81}, B_{43}, and B_{41} are unconditionally redundant boundaries of the second kind. All six boundaries can then be removed, leaving a reduced (incomplete) linear decision function comprised of 39 boundaries.

In order to determine the redundant measurements, recall that it is those measurements which have small magnitudes of the associated direction cosines for all the constituent hyperplanes which are most likely to be redundant. This concept was used in the following manner. All direction cosines with magnitude less than a certain "redundancy level" were set to zero, and the sample points were recategorized by this modified linear decision function (the problem of renormalizing the coefficients of the hyperplanes was ignored, since the correction would be small).

A plot of the error rate, rejection rate, and multiple recognition rate (since the linear decision function is now incomplete) versus the redundancy level is shown in Figure 30, from which it is seen that measurements with direction cosines of magnitude less than .04 are redundant, i.e., their removal will cause no reclassification of sample points. Thus, if a particular measurement has all 39 of its direction cosines less than .04, then it may be removed from the receptor. The resulting receptor, minimized by removing



DETERMINATION OF REDUNDANT MEASUREMENTS

FIGURE 30

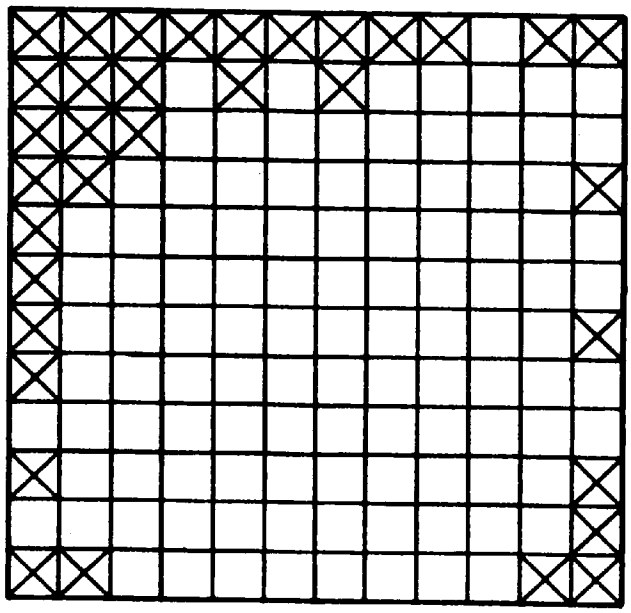
these redundant measurements, consists of the 110 clear elements of the 12 x 12 matrix shown in Figure 31. (It is interesting to compare this procedure to that of Gill's [23] for determining redundant binary measurements.)

Consequently, the original recognition machine has been reduced from 144 measurements and 45 boundaries to 110 measurements and 39 boundaries. The reduced machine categorizes the design sample patterns exactly as the original complete machine (save for one improvement). It remains to be seen whether this correspondence holds for further sample patterns.

10.3 Testing the Linear Decision Function

The reduced linear decision function was tested using both techniques discussed in Chapter VIII (Sections 8.1 and 8.2). The upper bound on the error-plus-natural-rejection rate was estimated to be 21.5% (assuming equal a priori probabilities). This can also be interpreted as a lower bound on the recognition rate of 78.5%. The breakdown of this estimate is shown as the lower numbers in the confusion matrix of Table 9.

Both the complete and reduced systems were also tested with 120 additional samples (12 samples of each number) gathered in the same manner as the design data. Figure 30 shows this test sample. The upper two numbers in the confusion matrix of Table 9 represent the categorization of these points for the complete machine (complete linear



MINIMIZED RECEPTOR FOR
THE HAND-PRINTING EXAMPLE

FIGURE 31

RECOGNIZED AS

INPUT PATTERN

	0	1	2	3	4	5	6	7	8	9	R
0	8 9 .74	--	2.33	1 1 1.33	.26	3.10	3.29	2.38	.53	1 1 2.68	2 1
1	12 12 --		.37	.39	--	.01	.00	.57	--	.59	
2	6 6 .74	1.39		5.94	2.22	.87	4 4 7.08	1 1 5.48	5.27	--	1 1
3	8 10 2.28	.99	10.20		--	3 2 1.97	2.56	3.00	4.56	.40	1
4	1 5.83	--	4.01	1 --	3 3	1 1 2.87	4 5 2.68	.92	3.59	1 1 9.68	2 2
5	2 1 .75	.28	1.70	1 1 2.22	.16	8 7	.84	1 1 .50	1.70	.36	1
6	1 1.02	.41	.94	1.22	.92	1 1 4.37	7 7	--	1 1.08	.34	2 3
7	1 1 .91	4.27	1.88	2.68	1 1 3.22	.44	--	7 7	2.07	7.08	3 3
8	0 1 .53	--	3.68	3 3 5.71	2.81	1 1 1.97	1 1 3.59	1 1 2.50	6 6	2.94	
9	2 2 .54	.30	--	1.83	1 13.35	1.62	1.97	1 1 7.08	1 1 9.51	8 7	

For minimum machine:

Estimated Recognition Rate
Using Test Sample (95% Conf. Int.) = 52%-69%

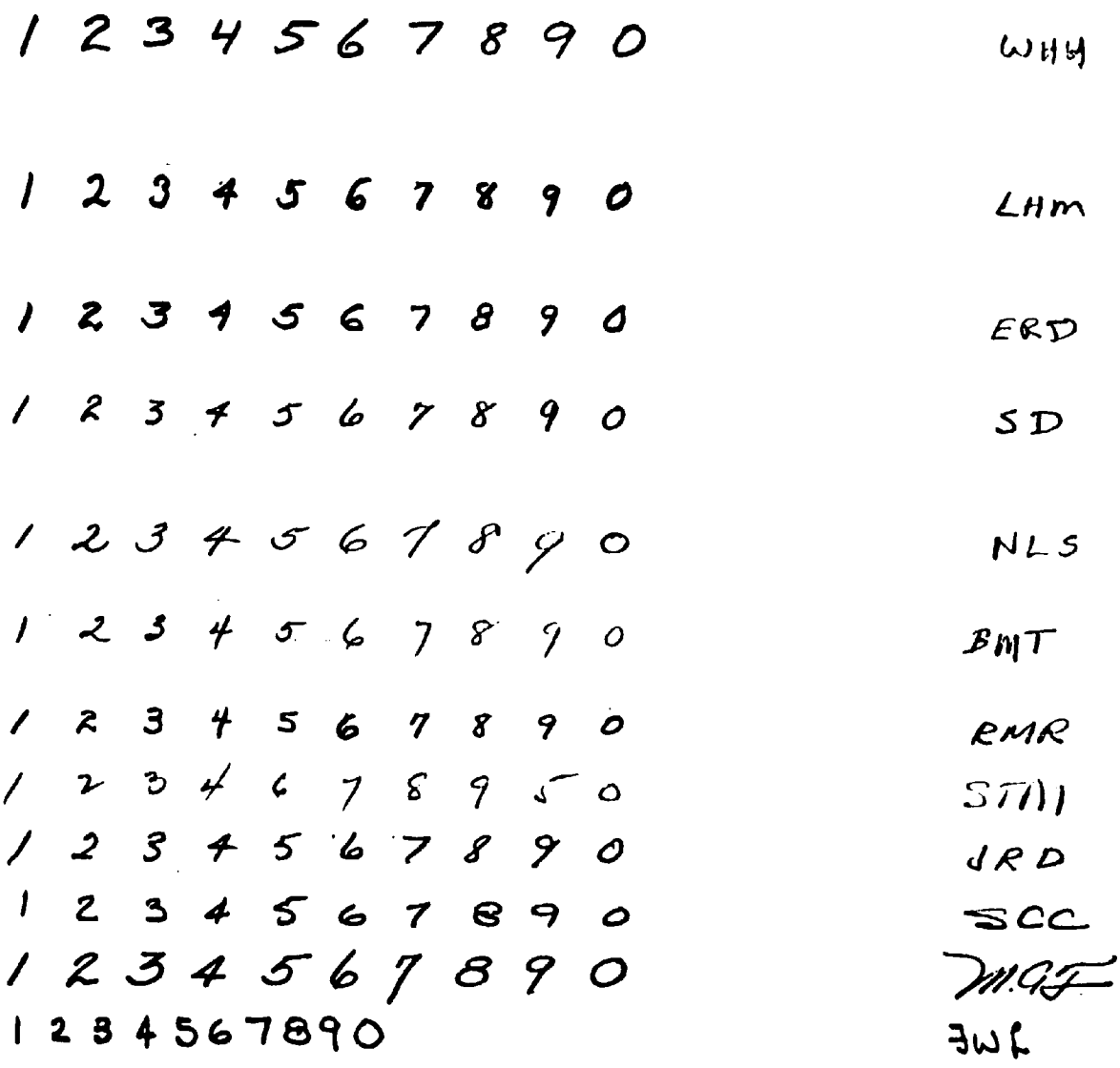
Estimated Lower Bound on
Recognition Rate = 78.5%

CONFUSION MATRIX FOR TEST SAMPLE. R IS REJECTION COLUMN. UPPER NUMBER IS FOR COMPLETE MACHINE; MIDDLE NUMBER IS FOR MINIMIZED MACHINE. (BOTH IN NO. OF SAMPLES.) LOWER NUMBER IS ESTIMATED PERFORMANCE BOUND (PER CENT) FOR THE MINIMUM MACHINE.

TABLE 9

decision function and complete receptor) and the minimum machine. Note that there are a few differences, but that the performance is almost exactly the same (in fact, the minimum machine correctly recognized one more sample than the complete machine). Therefore, the minimization process seems to give reasonable results. The one point that should be noted however, is that, in the minimum machine, one point was multiply recognized (an 8 as an 8 and a 1). This is an indication that perhaps the boundary B_{81} should not have been removed, and illustrates the possibility of failure of the definition of redundancy in a sample sense. Recall, however, that the removal of B_{81} actually did cause one recategorization, although it was a favorable one. Thus B_{81} is a face of the polytope enclosing the class 1 and hence is not geometrically redundant.

The resulting estimates of the minimized system error rate, rejection rate, and correct recognition rate, from the results shown in Table 9, are 30.0% (36 points), 9.2% (11 points) and 60.8% (73 points) respectively. From Figure 32, one can then state that, with probability 0.95, the intervals .20-.40, .03-.16, and .52-.69 include the system error probability, rejection probability, and correct recognition probability respectively. The agreement between this performance test and the estimated upper bound is not all that might be desired, the estimated bound indicating somewhat better performance than that attained in the test.



THE TEST SAMPLE

FIGURE 32

It is not surprising to find the estimated performance of this linear decision function to be so poor. This can be blamed on two factors; 1) a poor choice of measurements, and 2) a design sample size which might have been too small, leading to a poor estimate of the optimum hyperplanes. However, this is not so important, since this experiment was not meant to result in the design of a practical character recognition machine, but was rather meant to test certain aspects of the theory previously developed.

CHAPTER XI

CONCLUSION

11.1 Summary

This paper has discussed the properties and design of a particular class of categorizer, the linear decision function, which is of practical interest for two reasons:

1. It can be empirically designed without making any assumptions whatsoever about either the distribution of the receptor measurements or the *a priori* probabilities of occurrence of the pattern classes, providing an appropriate pattern source is available.
2. Its hardware realization is quite economic.

It is not guaranteed that a linear decision function will always perform well, although it is guaranteed that it will perform better than (or at least as well as) the minimum distance categorizer which is popular in the present day art. Nor is it a simple matter to predict in advance whether a linear decision function has a chance of working. The corollary to Theorem 9 may be used to set up a straightforward procedure to determine whether a set of points is linearly separable, but this will usually involve a good deal of computation. Besides, if it is determined that such a set is not linearly separable, there is no way of telling

to what degree this is so; the classes may still be separable with small probability of error.

Consequently, if one is interested in a linear decision function type of categorizer, his best approach is to actually design the categorizer and estimate its performance. If the estimated performance is good enough, then the designer has succeeded in designing an economic categorizer. If the performance is not good enough, the designer has two choices:

1. Search for a better set of measurements, a set which is more invariant to the natural perturbations of patterns contained within a class (the results of the experiment on hand-printing illustrate the importance of invariant measurements); or
2. go to a different type (usually a more complicated type) of categorizer.

A linear decision function has an interesting property which may be used even if the performance of such a categorizer is not all that is desired. This is its ability to help detect redundant measurements. For instance, in the example of the hand-printing, the designer may be required to use the matrix representation of the hand-printed characters. Consequently, he would have to go to a more sophisticated sort of categorizer. However, he can do so with the reduced receptor (the partial matrix) of

Figure 31. Although the linear decision function type of categorizer may not be usable in a given situation, it can in this way help to simplify a more complicated categorizer by simplifying the receptor.

A very general discussion with quite practical results was given concerning the testing of pattern recognition machines regardless of their structure. In particular, it was shown how to obtain confidence intervals for such results in a very simple fashion. It appears to the author that published results for pattern recognition tests would be greatly enhanced by the inclusion of such confidence intervals. Although a pattern recognition machine ought to be tested with a different set of samples than those used in the design, it is shown that it is possible to estimate a bound on the performance of a linear decision function with the design data. This can be very useful if the data is limited for some reason, but does not give as desirable an estimate as the use of further data.

11.2 Areas of Further Work

This effort has by no means completed the study of linear decision functions and related topics. One interesting problem is the study and design of linear decision functions in which the constituent hyperplanes are required to do either more or less work than those discussed herein. For instance, a hyperplane may be required to separate more than two classes; in this manner the lower limit of $\log_2 m$ hyperplanes for m

pattern classes may be approached. The computation algorithm developed in this paper for determining an optimum hyperplane is applicable here.

On the other hand, one may want to use more than one hyperplane per pair of pattern classes. In this way, nonlinear optimum boundaries may be more closely approximated. (Ridgway [41] is studying this problem for a binary measurement space. A possible approach is also given by Theorem 11.) As one makes a hyperplane do less work in the categorization process by using more of them, the categorizer will more closely approach the optimum categorizer, and also become more expensive. Consequently, the entire spectrum is of interest, since performance is traded for cost.

Of course, the study of quadratic and higher order decision functions has hardly been started. Mattson [33] gives a brief but enlightening discussion of this problem.

The problem of rejection criteria requires a good deal more study. The simplest form was analyzed in this paper; however, two other methods for rejection which are more general were mentioned but not analyzed. Both of these methods would do a better job than that one analyzed; one of these would have the same cost of implementation; the other would require twice as many hyperplane implementations but is the most general of the three discussed. There are probably other rejection criteria which are compatible with a linear decision function which haven't even been metnioned.

Using the normality concept of Section 7.2, which states that in many cases the distribution of distances of members of a class from a hyperplane is normal, one can develop another algorithm for estimating the optimum hyperplane separating the two classes. We are interested in choosing that hyperplane that minimizes the estimate of the expected error, or confusion, between the two classes. However, it is possible to estimate the error associated with a hyperplane by estimating the normal distribution of the distance of the members of each class from the hyperplane, and determining the area under the tails of these two distributions falling on the wrong side of the hyperplane (as discussed in Section 8.2). This might be expected to be a better estimate of the error than the proportion of points misrecognized, since more information is used in the estimate, providing the assumption of normally distributed distances is valid.

This improvement can be seen from the confidence interval curves of Figures 23 and 24 (keeping in mind that one is two-sided, the other single-sided). For instance, consider one class and a hyperplane. Let the sample size be 50, and the error estimate in either case be 5%. From Figure 23, a 95% confidence interval for the estimate based on the proportion of samples misrecognized is 1%-15%. From Figure 24, a 95% confidence interval for the normal

distribution estimate (obtained by moving horizontally from the .05 value of the ordinate to the "estimated area" curve, then up to the 50 sample curve, and back to the ordinate) is approximately 0%-8.5%. The latter estimate is obviously significantly better in this case.

Consequently, it would be quite reasonable to choose as an estimate of the optimum linear boundary that hyperplane which minimizes the normal estimate of error rather than the estimate based on the proportion of misclassified samples, providing again that the assumption of normality holds. An algorithm based on minimizing this normal estimate of error, using the method of steepest descent, is developed in Appendix III. Note that the resulting hyperplane for each local minimum is unique, in contrast to the previous algorithm in which the hyperplane could be any one chosen from, in general, an infinite set.

With regard to the experiments of Chapters IX and X, it appears that the assumption of normality is a good approximation in either case, and consequently that this algorithm would have been useful. In Chapter IX, the measurements are probably not too far from being normally distributed and independent, which allows for the small dimensionality. In Chapter X, the high dimensionality ought to make the approximation good. At any rate, this assumption was made in both chapters when the bounds on the performance were estimated.

A discussion of an unsolved problem, which deals with pattern recognition in general, is given in Section 8.3. This is the problem of partitioning a sample between the design and testing phases of a pattern recognition study when the sample size is limited. The case in which some of the samples are used to design the machine, and only those remaining are used to test the machine, seems to be reasonably solved in this section providing the deviation of the resulting machine from the optimum is small. The value of b, however, is in general difficult to calculate, and methods for estimating it warrant further study.

This sample partitioning is only one possibility, however. Perhaps more efficient use could be made of the total sample if some overlap in the design and test data were allowed. There may be an even better technique based on some sort of sequential procedure. It would also be advantageous to remove the restriction of small deviation of the actual error from the optimum (minimum) error. These various problems have yet to be investigated.

The above discussion has been intended to point out some of the areas in linear decision functions in particular and pattern recognition systems in general in which some strong theoretical attack can be made. It appears to the author that the state of the pattern recognition art has come to a point where less emphasis ought to be placed on gadgetry

(an emphasis that is certainly required in the early stages of a problem such as this), and more emphasis put on good theoretical work aimed at practical results. It is time for pattern recognition to grow from an art to a science.

APPENDIX I

EXTENSION OF CHOW'S RESULTS TO THE CASE OF CONSTANT COSTS

Chow [10] has shown that, for a given rejection rate, the error rate in a recognition system is minimized if the following decision criterion is used:

Choose class k if

$$\omega_k \beta(m \mid s_k) \geq \omega_j \beta(m \mid s_j) \qquad \text{for all } j \neq k$$

and

$$\omega_k \beta(m \mid s_k) \geq \gamma \sum_{i=1}^{p} \omega_i \beta(m \mid s_i) , \qquad 0 \leq \gamma \leq 1 ;$$

reject the pattern if

$$\omega_j \beta(m \mid s_j) < \gamma \sum_{i=1}^{p} \omega_i \beta(m \mid s_i) \qquad \text{for all } 1 \leq j \leq p .$$

Here ω_i is the a priori probability of the occurrence of class i, $\beta(m \mid s_i)$ is the conditional probability of making the measurement m given that a member of class i is present, p is the number of pattern classes, and γ is a constant chosen to force the system to meet the given rejection rate.

The proper value for γ is generally difficult to determine, and an empirical approach may often be necessary. However, there is one important case in which γ may be determined analytically, the discussion of which follows.

Let the cost of misrecognizing a pattern, of rejecting a pattern, and of correctly recognizing a pattern be independent of the pattern class. In particular let

$$c_{ij} = c \quad = \text{cost of misrecognition,}$$
$$c_{io} = c_o \quad = \text{cost of rejection,}$$
$$c_{ii} = 0 \quad = \text{cost of recognition,}$$

where

$$c > c_o > 0 .$$

(Since a Bayes criterion is being used, no generality is lost by setting $c_{ii} = 0$.[9]) The general loss function is given by (2.2):

$$C(\delta) = \sum_{i=1}^{p} \sum_{j=0}^{p} \int_M c_{ij}\omega_i\beta(m \mid s_i)\delta(d_j \mid m)dm$$

where $\delta(d_j \mid m)$ is the probability that class j will be decided given the measurement m (d_o is the rejection decision), and $C(\delta)$ is the loss associated with the decision function δ.

Using the above cost schedule, the loss function may be written

$$C(\delta) = \int_M \sum_{j=0}^{p} \delta(d_j \mid m) \sum_{i=1}^{p} c\omega_i \beta(m \mid s_i) dm$$

$$- \int_M \sum_{i=1}^{p} \delta(d_i \mid m) c\omega_i \beta(m \mid s_i) dm$$

$$- \int_M \delta(d_o \mid m) \sum_{i=1}^{p} (c-c_o)\omega_i \beta(m \mid s_i) dm \ .$$

Noting that

$$\sum_{j=0}^{p} \delta(d_j \mid m) = 1 \ ,$$

$$\sum_{i=1}^{p} \omega_i = 1 \ ,$$

$$\int_M \beta(m \mid s_i) dm = 1 \ ,$$

the first integral can be reduced, allowing the cost function to be written

$$C(\delta) = c - c \int_M \sum_{i=1}^{p} \delta(d_i \mid m) \omega_i \beta(m \mid s_i) dm$$

$$- (c-c_o) \int_M \delta(d_o \mid m) \sum_{i=1}^{p} \omega_i \beta(m \mid s_i) dm \ .$$

To minimize $C(\delta)$, $\delta(d_i \mid m)$ is chosen as follows:

$$\delta(d_k \mid m) = 1 , \qquad k \neq 0 ,$$

if

$$\omega_k \beta(m \mid s_k) \geq \omega_j \beta(m \mid s_j)$$

and

$$\omega_k \beta(m \mid s_k) \geq \left(\frac{c-c_o}{c}\right) \sum_{i=1}^{p} \omega_i \beta(m \mid s_i) ;$$

$$\delta(d_o \mid m) = 1$$

if

$$\omega_j \beta(m \mid s_j) < \frac{c-c_o}{c} \sum_{i=1}^{p} \omega_i \beta(m \mid s_i) \qquad \text{for all } 1 \leq j \leq p .$$

But this decision criterion is of the same form as that derived by Chow for the case of minimum error rate given a fixed rejection rate, with

$$\gamma = \frac{c-c_o}{c} .$$

Therefore, minimizing the cost in the case of constant costs also minimizes the error rate for the rejection rate which corresponds to the above γ.

APPENDIX II

THE OPTIMUM SAMPLE STRATIFICATION FOR ESTIMATING THE PERFORMANCE OF A PATTERN RECOGNITION MACHINE*

The sample stratification for selective sampling which gives the minimum variance for a single estimate is derived. When the a priori probabilities of occurrence ω_i for each class are known, the maximum likelihood estimate for, say, the total probability of error, as derived in Chapter VIII, may be written

$$\hat{e} = \sum_{i=1}^{p} \omega_i \frac{m_{e_i}}{n_i} .$$

The variance of $\hat{e}$ may be written

$$\sigma'^2_{\hat{e}} = \sum_{i=1}^{p} \omega_i^2 \frac{e_i(1-e_i)}{n_i} .$$

Maximize this with respect to n_i under the constraint

$$\sum_{i=1}^{p} n_i = n . \qquad \text{(A3.1)}$$

- -

*Suggested by W. H. Williams.

Then

$$-\frac{\partial \sigma'^2_{\hat{e}}}{\partial n_i} = \frac{\partial}{\partial n_i}\left[\sum_{i=1}^{p} \omega_i^2 \frac{e_i(1-e_i)}{n_i} + \lambda^2\left(\sum_{i=1}^{p} n_i - n\right)\right]$$

$$= -\frac{\omega_i^2 e_i(1-e_i)}{n_i^2} + \lambda^2 = 0 \ ,$$

where λ is the Lagrange multiplier. Hence

$$n_i = \frac{\omega_i \sqrt{e_i(1-e_i)}}{\lambda} \ .$$

λ is chosen to satisfy the condition (A3.1). Therefore, if one has any knowledge at all of the error rates for each class, he will get a better estimate if he adjusts the size of the sample taken from each class according to (A3.2).

APPENDIX III

A COMPUTATION ALGORITHM FOR FINDING THAT HYPERPLANE WHICH MINIMIZES THE NORMAL ESTIMATE OF THE ERROR

If the approximation of normally distributed distances as indicated in Section 7.2 is valid, then the estimate of the expected loss $C_{k\ell}$ for a given hyperplane $B_{k\ell}$ based on this assumption can be written (as in Section 8.2)

$$C_{k\ell} = \frac{\omega_k c_{k\ell}}{\sqrt{2\pi}} \int_{-\infty}^{-\frac{\bar{s}_k}{v_k}} e^{-\frac{1}{2}x^2} dx + \frac{\omega_\ell c_{\ell k}}{\sqrt{2\pi}} \int_{\frac{\bar{s}_\ell}{v_\ell}}^{} e^{-\frac{1}{2}x^2} dx , \quad (A4.1)$$

where ω_k is the a priori probability of the occurrence of class k, $c_{k\ell}$ is the loss associated with misrecognizing a member of class k as belonging to class ℓ, $c_{kk} = 0$, $\bar{s}_k$ is the mean of the distances of the sample points of class k from $B_{k\ell}$, and v_k is the sample standard deviation of these distances. Class k is assumed to be on the positive side of $B_{k\ell}$, and class ℓ on the negative side.

We wish to determine that hyperplane which minimizes $C_{k\ell}$. To do this, we use the method of steepest descent, in which we choose an initial hyperplane, and continually adjust its coefficients along the direction of negative gradient until a local minimum is reached. Consequently, we need to determine the gradient of (A4.1).

Let the hyperplane be described by

$$\sum_{i=1}^{n} \alpha_i x_i + \alpha_o = \sum_{i=0}^{n} \alpha_i x_i = 0 \ ,$$

where x_o is a dummy coordinate equal to +1 (as in Section 5.3). Then,

$$\frac{\partial C_k}{\partial \alpha_i} = \omega_k c_{k\ell} N\left(\frac{\bar{s}_k}{v_k}\right) \frac{\partial\left(-\frac{\bar{s}_k}{v_k}\right)}{\partial \alpha_i}$$

$$- \omega_\ell c_{\ell k} N\left(\frac{\bar{s}_\ell}{v_\ell}\right) \frac{\partial\left(\frac{\bar{s}_\ell}{v_\ell}\right)}{\partial \alpha_i} \ , \qquad 0 \le i \le n \ , \qquad \text{(A4.2)}$$

where $N\left(\frac{\bar{s}}{v}\right)$ is the ordinate of the standard normal density function for variate equal to $\bar{s}/v$.

Consider $\bar{s}_k$ and v_k:

$$\bar{s}_k = \frac{1}{K} \sum_{j=1}^{K} \sum_{i=0}^{n} \alpha_i m_{ijk} = \frac{1}{K} \sum_{i=0}^{n} \alpha_i \sum_{j=1}^{K} m_{ijk} \ ,$$

where m_{ijk} is the i^{th} coordinate of the j^{th} point of the k^{th} class, there being K sample points from class k, and $m_{ojk} = +1$. Let

$$\frac{1}{K} \sum_{j=1}^{K} m_{ijk} = \bar{m}_{ik} \ .$$

$\bar{m}_{ik}$ is the average of the i^{th} coordinate over all sample points from class k. Then

$$\bar{s}_k = \sum_{i=1}^{n} \alpha_i \bar{m}_{ik} \ . \qquad \text{(A4.3)}$$

For v_k, we write

$$v_k = \left[\frac{1}{K-1} \sum_{j=1}^{K} \left(\sum_{i=1}^{n} \alpha_i m_{ijk} - \bar{s}_k \right)^2 \right]^{1/2} \ . \qquad \text{(A4.4)}$$

When taking derivatives, it must be insured that the coefficients remain normalized, i.e.,

$$\sum_{i=1}^{n} \alpha_i^2 = 1 \ .$$

A procedure similar to that in Section 5.3 can be used whereby every α_i, $1 \leq i \leq n$, is divided by

$$\sqrt{\sum_{i=1}^{n} \alpha_i^2} \ .$$

However, perusal of (A4.3) and (A4.4) shows that when the ratio $\bar{s}_k / v_k$ is formed, these terms will cancel. Therefore, we need not concern ourselves with the normalization problem in this algorithm (the same argument holds for $N(\bar{s}/v)$). Then

$$\frac{\partial}{\partial \alpha_i}\left(\frac{\bar{s}_k}{v_k}\right) = \frac{1}{v_k}\frac{\partial}{\partial \alpha_i}(\bar{s}_k) - \frac{\bar{s}_k}{v_k^2}\frac{\partial}{\partial \alpha_i}(v_k), \qquad 0 \le i \le n . \tag{A4.5}$$

From (A4.3),

$$\frac{\partial}{\partial \alpha_i}(\bar{s}_k) = \bar{m}_{ik} . \tag{A4.6}$$

From (A4.4)

$$\frac{\partial}{\partial \alpha_i}(v_k) = \frac{K}{(K-1)v_k}\left[\frac{1}{K}\sum_{j=1}^{K} s_{jk}m_{ijk} - \bar{s}_k\bar{m}_{ik}\right], \qquad 0 \le i \le n , \tag{A4.7}$$

where s_{jk} is the distance of the j^{th} point of the k^{th} class from $B_{k\ell}$. Similar expressions hold for $\frac{\partial}{\partial \alpha_i}\left(\frac{\bar{s}_\ell}{v_\ell}\right)$.

Substituting (A4.6) and (A4.7) into (A4.5), thence into (A4.2), one obtains the gradient:

$$\frac{\partial C_{k\ell}}{\partial \alpha_i} = \frac{\omega_k c_{k\ell}}{v_k} N\left(\frac{\bar{s}_k}{v_k}\right)\left[\bar{m}_{ik} - \frac{K}{K-1}\frac{\bar{s}_k}{v_k^2}\left(\frac{1}{K}\sum_{j=1}^{K} s_{jk}m_{ijk} - \bar{s}_k\bar{m}_{ik}\right)\right]$$

$$- \frac{\omega_\ell c_{\ell k}}{v_\ell} N\left(\frac{\bar{s}_\ell}{v_\ell}\right)\left[\bar{m}_{i\ell} - \frac{L}{L-1}\frac{\bar{s}_\ell}{v_\ell^2}\left(\frac{1}{L}\sum_{j=1}^{L} s_{j\ell}m_{ij\ell} - \bar{s}_\ell\bar{m}_{i\ell}\right)\right],$$

$$0 \le i \le n .$$

New coefficients α_i' are chosen so that

$$\alpha_i' = \alpha_i - \theta \frac{\partial C_{k\ell}}{\partial \alpha_i} ,$$

where θ is some arbitrary adjustable constant chosen to afford a compromise between the number of iterations required and stability.

The passing of a minimum is simply detected by computing (A4.1) after each iteration and watching for an increase in $C_{k\ell}$ (subroutines for evaluating the integrals of (A4.1) are available for many computers). This minimum can be estimated as closely as desired by taking successively smaller values for θ. This minimum, when determined, may not be the absolute minimum. Several initial hyperplanes should be tried, and the best result used.

BIBLIOGRAPHY

1. S. Agmon, "The Relaxation Method for Linear Inequalities," Canad. Jour. Math., Vol. 6, pp. 382-392; 1954.

2. T. W. Anderson, An Introduction to Multivariate Statistics, John Wiley and Sons, New York; 1958.

3. P. Baran, G. Estrin, "An Adaptive Character Reader," IRE Wescon Record, Part 4, pp. 29-41; 1960.

4. W. W. Bledsoe, I. Browning, "Pattern Recognition and Reading by Machine," Proc. EJCC, pp. 225-232; December 1959.

5. J. S. Bomba, "Alphanumeric Character Recognition Using Local Operations," Proc. EJCC; December 1959.

6. A. L. Bowley, "Measurement of the Precision Attained in Sampling," Bull. Int. Stat. Inst., Vol. XXII, pp. 1-62; 1926.

7. S. H. Brooks, "Comparison of Maximum-Seeking Methods," Operations Research, Vol. 7, No. 4, pp. 430-457; July 1959.

8. R. S. Burington, D. C. May, Jr., Handbook of Probability and Statistics with Tables, Handbook Publishers, Sandusky, Ohio; 1958.

9. H. Chernoff, L. E. Moses, Elementary Decision Theory, John Wiley and Sons, New York; 1953.

10. C. K. Chow, "An Optimum Character Recognition System Using Decision Functions," IRE Trans. on Electronic Computers, Vol. EC-6, pp. 247-254; December 1957.

11. W. A. Clark, B. G. Farley, "Generalization of Pattern Recognition in Self-Organizing Systems," Proc. WJCC, pp. 86-91; 1955.

12. C. S. Clopper, E. S. Pearson, "The Use of Confidence or Fiducial Limits, Illustrated in the Case of the Binomial," Biometrika, Vol. 26, pp. 404-413; 1934.

13. C. B. Crumb, C. E. Rupe, "The Automatic Digital Computer as an Aid in Medical Diagnosis," Proc. EJCC; December 1959.

14. W. Doyle, "Recognition of Sloppy, Hand-Printed Characters," Proc. WJCC, pp. 133-142; 1960.

15. K. R. Eldredge, F. J. Kamphoefner, P. H. Werdt, "Automatic Input for Data Processing Systems," Proc. EJCC, p. 60; December 1956.

16. B. G. Farley, W. A. Clark, "Simulation of Self-Organizing Systems by Digital Computer," IRE Trans. on Information Theory, Vol. IT-4, pp. 76-84; September 1954.

17. R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, Vol. 7, p. 179; 1936.

18. I. Flores, L. Grey, "Optimization of Reference Signals for Character Recognition Systems," IRE Trans. on Elec. Comp., Vol. EC-9, No. 1, pp. 54-61; March 1960.

19. D. A. S. Fraser, Nonparametric Methods in Statistics, John Wiley and Sons, New York; 1957.

20. D. A. S. Fraser, Statistics: An Introduction, John Wiley and Sons, New York; 1960.

21. L. S. Frishkopf, L. D. Harmon, "Machine Reading of Cursive Script," Proc. 4th London Symposium on Information Theory; 1960.

22. S. I. Gass, Linear Programming, McGraw-Hill, New York; 1958.

23. A. Gill, "Minimum Scan Pattern Recognition," IRE, Trans. on Info. Theory, Vol. IT-5, No. 2, pp. 52-58; June 1959.

24. R. L. Grimsdale, F. H. Sumner, C. J. Tunis, T. Kilburn, "A System for the Automatic Recognition of Patterns," Proc. IEE, Vol. 106, Part B, No. 26, p. 210; March 1959.

25. W. H. Highleyman, "A Note on Optimum Pattern Recognition Systems," to be published in IRE Trans. on Elec. Comp.

26. W. H. Highleyman, "An Analog Method for Character Recognition," to be published in IRE Trans. on Elec. Comp.

27. N. L. Johnson, B. L. Welsh, "Applications of the Noncentral t-Distribution," Biometrika, Vol. 31, pp. 362-389; 1940.

28. L. A. Kamentsky, "Pattern and Character Recognition Systems - Picture Processing by Nets of Neuron-Like Elements," Proc. EJCC, pp. 304-309; 1959.

29. L. A. Kamentsky, "Simulation of Three Machines which Read Rows of Handwritten Arabic Numbers," to be published in IRE Trans. on Elec. Comp.

30. A. E. Laemmel, private communication.

31. T. Marill, D. M. Green, "Statistical Recognition Functions and the Design of Pattern Recognizers," IRE Trans. on Elec. Comp. Vol. EC-9, No. 4, pp. 472-477; December 1960.

32. M. V. Mathews, P. Denes, "Spoken Digit Recognition Using Time-Frequency Pattern Matching," Jour. Acous. Soc. Am., Vol. 32, No. 11, p. 1450; November 1960.

33. R. L. Mattson, "The Design and Analysis of an Adaptive System for Statistical Classification," Masters Thesis, E.E. Dept., M.I.T.; May 1959.

34. R. L. Mattson, "A Self-Organizing Logical System," Proc. EJCC; December 1959.

35. R. McNaughton, "Unate Truth Functions," IRE Trans. on Elec. Comp., Vol. EC-10, No. 1. pp. 1-6; March 1961.

36. D. Middleton, D. Van Meter, "Detection and Extraction of Signals in Noise from the Point of View of Statistical Decision Theory," Jour. Soc. Ind. and App. Math., Vol. 3, pp. 192-253, September 1955; and Vol. 4, pp. 86-119, June 1956.

37. T. S. Motzkin, I. J. Schoenberg, "The Relaxation Method for Linear Inequalities," Canad. Jour. Math., Vol. 6, pp. 393-404; 1954.

38. J. Neyman, "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Purposive Selection," Jour. Royal Stat. Soc., Vol. 97, Part IV, pp. 558-625; 1939.

39. E. J. Pearson, H. O. Hartley, Biometrika Tables for Statisticians, pp. 204-205, The University Press, Cambridge; 1954.

40. G. J. Resnikoff, G. J. Lieberman, Tables of the Noncentral t-Distribution, Stanford University Press, Stanford, Calif.; 1957.

41. W. C. Ridgway, private communication.

42. L. G. Roberts, "Pattern Recognition with an Adaptive Network," IRE Internat. Conv. Rec., Part 2, pp. 66-70; 1960.

43. F. Rosenblatt, "A Theory of Statistical Separability in Cognitive Systems," Report No. VG-1196-G-1, Cornell Aeronautical Laboratory, Inc.; January 1958.

44. F. Rosenblatt, "Perceptron Simulation Experiments," Proc. IRE, Vol. 48, No. 3, pp. 301-309; March 1960.

45. B. Scott, P. A. M. Curry, "Automatic Printed Character Reading," Jour. SMPTE, Vol. 68, No. 4, p. 240; April 1959.

46. G. S. Sebestyen, "Categorization in Pattern Recognition," Doctors Thesis, E.E. Dept. M.I.T.; April 1960.

47. G. S. Sebestyen, "Recognition of Membership in Classes," IRE Trans. on Info. Th., Vol. IT-7, No. 1, pp. 44-50; January 1961.

48. D. H. Shepard, P. F. Bargh, C. C. Heasly, "A Reliable Character Sensing System for Documents Prepared on Conventional Business Devices," Wescon Conv. Rec., Part 4, p. 111; 1957.

49. A. E. Taylor, Introduction to Functional Analysis, John Wiley and Sons, New York; 1958.

50. G. Tintner, Econometrics, John Wiley and Sons, New York; 1952.

51. G. P. Wadsworth, J. G. Bryan, Introduction to Probability and Random Variables, McGraw-Hill, New York; 1960.

52. B. Widrow, M. E. Hoff, "Adaptive Switching Circuits," Stanford Electronics Lab., Stanford, Calif., Tech. Rept. No. 1533-1; June 1960.

53. B. Widrow, "Adaptive Sampled Data Systems," Stanford Electronics Lab., Stanford, Calif., Tech. Rept. No. 2104-1; July 1960.