NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Technical Memorandum 33-482

A Multiclass Sequential Hypothesis Test With Applications in Pattern Recognition

Jung Pyo Hong

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

June 15, 1971

JET PROPULSION LABORATORY *California Institute of Technology • 4800 Oak Grove Drive, Pasadena, California 91103*

June 11, 1971

Recipients of Jet Propulsion Laboratory
Technical Memorandum 33-482

Subject: Erratum

Gentlemen:

Please note the following corrections to Technical Memorandum 33-482, A Multiclass Sequential Hypothesis Test With Applications in Pattern Recognition, by Jung Pyo Hong, dated June 15, 1971:

Page v, TABLE OF CONTENTS, should read:

Delete the following entry:

Very truly yours,

John Kempton, Manager
Publications Section

Telephone 354-4321

Twx 213-449-2451

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Technical Memorandum 33-482

A Multiclass Sequential Hypothesis Test With Applications in Pattern Recognition

Jung Pyo Hong

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

June 15, 1971

Prepared Under Contract No. NAS 7-100
National Aeronautics and Space Administration

PREFACE

The work described in this report was performed by the Astrionics Division of the Jet Propulsion Laboratory.

The research reported in this Technical Memorandum is a dissertation presented to and accepted by the Faculty of the Graduate School, University of Southern California, in partial fulfillment of the requirements for the Degree Doctor of Philosophy (Electrical Engineering).

The examples in this report pertain to pattern recognition of characters. However, the theory of multiclass sequential hypothesis test can be applied in other disciplines. The theory is useful in signal detection as well as in detection of objects by a robot, for instance.

ACKNOWLEDGEMENT

The author wishes to acknowledge the help provided by Professors R. M. Gagliardi and T. E. Harris. The author particularly wishes to thank Professor L. D. Davisson, whose interest and guidance made this work possible.

The research reported herein was supported to a significant degree by NASA Contract NAS 7-100.

This research was also supported in part by a NASA Traineeship NGT 05-018-127-Amendment 1 and by a NASA Research Grant NGL 05-018-118. The author wishes to acknowledge this generous assistance.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

In recent years there has been a sharp rise in the need and interest for pattern recognition. In particular much work has been done on the problems of machine reading. There are algorithms which partially solve the problem of reading impact printed material. This dissertation presents an algorithm which can be used to build a reading machine that will read impact printed characters and handwritten letters.

Invariant features are extracted by random lines. The number of intersections and also the total length of intersection that these lines produce are the random variable observations used as inputs to a hypothesis test. This method allows the pattern to be anywhere in the retina. It eliminates the cost of fine alignment of the pattern before taking samples. Many previous users of these features utilized only the mean of the random variable. Here the whole probability distribution of the random variable is used. This allows the introduction of size invariant methods.

The sequential multiclass hypothesis test presented in this dissertation is in such a form as to allow rapid computation of the errors of the first and second kinds

for each possible decision. This is useful because in any practical system the user desires to have easy access to the parameters which control the performance of the machine.

One interpretation of this test is that it computes the ratio of the likelihood of an observation coming from a class to the likelihood of an observation coming from any other class. When this ratio exceeds a threshold a decision in favor of the class is made. For each sample, there are as many such comparisons as there are classes.

The sequential multiclass hypothesis test proposed in this dissertation is a Bayes test at each step. The proposed test is Wald's sequential probability ratio test for the two-class problem. It is not like the generalized Wald's test which tests all combinations of two hypotheses, nor is it like the M-hypothesis test which also tests the same number of combinations. The number of comparisons these tests make is (1/2)M(M-1), where M is the number of classes. They require far more computations than the proposed test.

Extensive experiments with block letters and handwritten numerals are reported. These experiments verify the usefulness of the proposed multiclass hypothesis test. These experiments show that the error rates are under the

control of the user and that the average length of the test can be predicted.

A survey of the methods in pattern recognition is presented to put the author's contribution in perspective.

CHAPTER I

INTRODUCTION

A. THE PROBLEM

There are two aspects to pattern recognition. In one form of the problem a field of data is given to a recognition machine and it is asked to state whether there are patterns. In the second form, the algorithm is required to decide under certain criteria which of the known patterns the data represent. (Often the null pattern or the reject option is included as a possible decision.) In this work the emphasis shall be on the solution of the second problem.

Pattern recognition is a two step process. First, observations are made, then an algorithm uses these observations to arrive at a conclusion. Observations include all forms of measurement, filtering, and digitizing. The decision algorithms may be linear, nonlinear, or statistical functions of the observations. There are abundant examples of such processes in nature. One first hears sounds of speech, then understands their properties. One must see the printed page before one can read the words.

It is not the object here to study how these processes operate in nature. But these examples clearly point out the important interrelationship between the observation and the algorithm. Hence to obtain a good pattern recognition machine, it is required that the observation and the decision algorithm be studied together.

Often investigators in pattern recognition have taken the observation phase of the process in an expedient manner. By arbitrarily limiting the type of observation, one severely narrows the possible class of compatible or feasible decision algorithms. As an example consider the early investigators who used the time signal from a television-like scanner. The two dimensional region of interest is divided in a checkerboard manner and each square is assigned a gray level according to the image. The choice of such a set of n by m samples as the observation features is unfortunate. Computational requirements on the large set of numbers limit the types of algorithms.

The requirements of the problem often suggest a class of decision algorithms. Then one must know how to choose the best features: For instance, when the requirements of the problem are stated in terms of minimizing the average risk or in terms of the probability

of the recognition, certain statistical tests come to mind. What remains is to find the feature extraction scheme which will meet the needs of the statistical methods consistent with computational and other requirements.

An example of a practical problem in pattern recognition with some of its requirements would be the design of a machine which could read the address of a letter and sort it according to the postman's route, with assignable probability of the correctness of the sort. It would reject as few as possible and sort at the highest possible speed. So far the only "machine" that comes close to meeting these requirements is man.

Exactly what features man extracts from the address label is not known nor is it known what algorithm he uses to read written material. The motivation behind developing a machine which will perform reading is that the machine may be faster for a subset of "easy" problems. It seems that the speed of the algorithm can be enhanced if the algorithm is based on some random sampling of the data rather than on some fixed extraction such as contour tracing which takes more effort.

The pattern recognition system presented in this thesis will use a statistical hypothesis test. The method used in the observation phase is carefully chosen

to assure the stochastic nature of the input.

B. THE DISSERTATION

This dissertation explores a solution to the pattern recognition problem which attains a requested performance level and which optimizes speed (amount of computation) and storage requirements. One may observe that a probabilistic decision machine is most natural to the requirements of certain problems. Then suitable features are chosen as the input to the algorithm.

This dissertation relies heavily on the problems of character recognition for examples and illustrations. Let it be noted that the ideas of randomized feature extraction may be used for other types of problems. For instance they may be used for feature extraction of phonemes in audio signals.

Chapter II contains a survey of pattern recognition. A few of the important tools used in pattern recognition are presented to put this dissertation in perspective. The works of certain investigators are discussed so that the two steps in pattern recognition can be illustrated. Multiclass hypothesis testing is discussed in Chapter II. Maximum likelihood and Bayes procedures are reviewed.

In general it is difficult to compute the significance of a test. That is, it is difficult to compute how many samples are needed for a level of performance

because n-fold integrals are involved. Chapter III describes a method of approximating the significance of a test when the test is of a special form. Since the significance of the test can be monitored easily for each sample as it is observed, a sequential multiclas hypothesis test results.

The requirements of the problem demand a stochastic decision algorithm. Line intersection length and the number of intersections of a random line with the figure are presented as two invariant feature extraction techniques. The properties of these features relevant to font, size, and noise are discussed in Chapter IV.

Chapter V presents experimental results using the features of Chapter IV and the algorithms of Chapter III. Chapter V also includes the results of a recognition experiment of hand printed digits.

C. NOTATIONS AND DEFINITIONS

The notations and definitions used in this dissertation are consistent throughout. A glossary is included at the beginning of this dissertation.

In the problems considered in this work it is assumed that there are $M = 2,3,\ldots$ hypotheses. Only one hypothesis is actually true. The ith hypothesis, denoted H_i, shall be the proposition that the observations $v_i = (v_1, v_2, \ldots, v_n)$ are taken from the ith class of

distribution $\tilde{F}_i$. The symbol above the function name (~) allows the same name F_i to be given to a family of functions associated with a hypothesis. Strictly, the functions $\tilde{F}_i(v_1,\ldots,v_{10})$ and $\tilde{F}_i(v_1,\ldots,v_{55})$ are not the same thing. F_i will be used to denote the distribution function of one variable $F_i(v_4)$.

It is assumed that the distributions F_i are distinct. That is, $F_i \neq F_j$ if $i \neq j$. If the densities exist, then dF_i is the probability density function.

The a priori probability that H_i is true is

$$P_i = \text{Prob}\{H_i \text{ is really true}\} \tag{1.1}$$

Clearly

$$P_i d\tilde{F}_i(v) = \text{Prob}\{H_i \text{ is true and } v = (v_1,v_2,\ldots,v_n) \text{ is observed}\} \tag{1.2}$$

Or

$$d\tilde{F}_i(v) = \text{Prob}\{v = (v_1,v_2,\ldots,v_n) \text{ is observed given } H_i \text{ is true}\} \tag{1.3}$$

The algorithms considered here will be allowed to verify one of the hypotheses or none at all. This last decision is often called a reject.

$$D_0 = \text{reject} \tag{1.4}$$

$$D_i = \text{accept } H_i \quad i = 1,2,\ldots,M \tag{1.5}$$

The ratio of two probability densities will be named $Z_n(i,j)$

$$Z_n(i,j) = \frac{d\tilde{F}_j(v_1,v_2,\ldots,v_n)}{d\tilde{F}_i(v_1,v_2,\ldots,v_n)} \tag{1.6}$$

When the samples are independent,

$$Z_n(i,j) = \prod_{m=1}^{n} \frac{dF_j(v_m)}{dF_i(v_m)} \tag{1.7}$$

$$= \prod_{m=1}^{n} z_m(i,j) \tag{1.8}$$

where

$$z_m(i,j) = \frac{dF_j(v_m)}{dF_i(v_m)}$$

Often the logarithm of Ratios 1.6 and 1.9 are useful.

$$\tilde{Z}_n(i,j) = \ln Z_n(i,j) \tag{1.10}$$

and

$$\tilde{z}_m(i,j) = \ln z_m(i,j) \tag{1.11}$$

Because the logarithm of a product is a sum the logarithm,

$$\tilde{Z}_n(i,j) = \sum_{m=1}^{n} \tilde{z}_m(i,j) \tag{1.12}$$

Any statistical decision algorithm is subject to errors. The probabilities of the errors are given the names e_{ij}.

$$e_{ij} = \text{Prob}\{\text{accepting } H_i \text{ when } H_j \text{ is true}\}$$

$$= \text{Prob}\{D_i | H_j \text{ true}\} \tag{1.13}$$

More precisely, let $\psi_i^{(n)}$ be the region in $v = (v_1, v_2, \ldots)$ such that a decision D_i is made at the nth stage.

$$\{v \in \psi_i^{(n)}\} \Rightarrow \{\text{decision } D_i \text{ is made exactly when } n \text{ components of } v \text{ are observed}\} \tag{1.14}$$

Also let $e_{ij}^{(n)}$ be the probabilities of error for decisions made with n samples. Then

$$\overline{e_{ij}}^{(n)} = \int_{\psi_i^{(n)}} d\tilde{F}_j(v_1, v_2, \ldots, v_n) \tag{1.15}$$

The superscript is used to stress that there are n components in the vector v. This is necessary to compute the error probabilities for the sequential tests.

Let p(n) be the probability that the test ends at the nth stage.

$$p(n) = \text{Prob}\{\text{sequential test ends at the nth stage}\} \tag{1.16}$$

The total error rates are

$$e_{ij} = \sum_{n} e_{ij}^{(n)} \tag{1.17}$$

$$= \sum_{n} \int_{\psi_i(n)} d\tilde{F}_j(v) \tag{1.18}$$

where the last equality is by Definition 1.15.

A table of $[e_{ij}]$ is called a confusion matrix. The probability of correctly accepting the ith hypothesis given that H_i is true is

$$e_{ii} = \text{Prob}\{D_i | H_i \text{ true}\} \tag{1.19}$$

Of course this is not an error, but the letter "e" is used for consistency with the other entries of this table. The probability that a decision algorithm will correctly choose a hypothesis is

$$P_i e_{ii} = \text{Prob}\{D_i \text{ and } H_i\} \tag{1.20}$$

This term appears frequently in subsequent chapters. It will be called the probability of <u>detection</u> and given the notation

$$\gamma_i = P_i e_{ii} \tag{1.21}$$

Two types of errors are of particular importance in pattern recognition. The first is the probability that the result of a classification is incorrect. The second

is the probability that, given a known pattern, the algorithm will not correctly detect it.

An example of an application of a pattern recognition algorithm will clarify this point. Suppose a reading machine is scanning a typewritten page. If it reports that the next letter is "Q" it is desirable to know the probability that such a report is incorrect, i.e., the machine is really observing another letter. The probability of such an event is called error probability of the first kind, α_Q. On the other hand, the reading machine may be positioned over a known letter, "B". The probability that the machine will correctly identify a letter is the probability of detection, γ_B. If there is a misclassification then there has been an error of the second type. Its probability is β_B and

$$\beta_B = P_B - \gamma_B \tag{1.22}$$

An algorithm may classify a given test pattern into an incorrect class.

$$\alpha_i = \text{Prob}\{D_i \text{ is incorrect}\}$$

$$= \sum_{j \neq i} e_{ij} P_j \tag{1.23}$$

This is the probability of <u>false declaration</u>.

The $\{\alpha_i\}$ are defined for all $i = 0,1,2,\ldots,M$. There are M classes and $i = 0$ corresponds to the null pattern or the reject option. That is α_0 is the probability of a reject occuring.

Another type of error is the probability of a miss. That is, there is the probability that H_i is true and an incorrect decision D_j, where $j \neq i$, is made.

$$\beta_i = P_i \sum_{j \neq i} e_{ij} \tag{1.24}$$

The $\{\beta_i\}$ are meaningfully defined for all $i = 1,2,\ldots,M$ but $\beta_0 = 0$.

It seems reasonable to characterize a pattern recognition system in terms of $\{\alpha_i\}$ and $\{\gamma_i\}$. It is useful to be able to find an algorithm at a specified level of $\{\alpha_i\}$ and $\{\gamma_i\}$.

CHAPTER II

SURVEY

A survey of the mathematical tools often used in pattern recognition and a few experiments in pattern recognition are presented in this chapter. The purpose is to put this dissertation in perspective. For other examples in this field the reader is directed to Nagy [Ref. 1] and to Pattern Recognition [Ref. 2].

A. MATHEMATICAL TOOLS

1. Linguistic Approach

In the linguistic approach, the input features are the strokes and the stroke locations. Without becoming too involved, an example will be given.

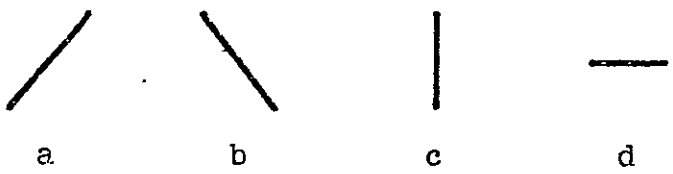

Fig. 2-1. Basic Elements of a Two Dimensional Field

The set of all basic elements is called the alphabet. Suppose the alphabet is as displayed in Figure 2-1. A few of the possible inputs are:

a) c d c which represents H (2.1)

b) a d b which represents A (2.2)

c) c b a c which represents M (2.3)

d) c b c which represents N (2.4)

The action of the decision algorithm is much like a compiler. It checks to see if the combination of the input elements forms a pattern in a dictionary. To perform this chore efficiently one uses all the mathematics of context free language, graph theory, and compiler theory [Ref. 3 and 4].

2. Linear Operations

Many methods look upon the input x as a matrix or a vector. Nilsson [Ref. 5, p. 79] discusses partitioning of the observation space into classes. Andrews [Ref. 6] on the other hand uses transform methods on the input.

The input is sometimes looked upon as a matrix $X = [X_{ij}]$ and transformations upon X are performed.

$$Z = P\ X\ Q \tag{2.5}$$

Functions of Z are used in the decision algorithm.

Andrews [Ref. 6] uses cross-correlation of a letter prototype against a field of letters to find the matching

letters. To optimize on speed of computation the fourier transform of $[X_{ij}]$ is used. The transform of the input and the transform of the reference are multiplied together, thus giving a matched filter operation. This method requires huge amounts of computation and is sensitive to rotation as well as to scale variations.

Often a class is defined by several prototypes $G_1^{(i)}$, $G_2^{(i)}$,... The superscript (i) says that the prototype belongs to class i. A prototype is described by a vector of measurements or features

$$G_1^{(i)} = (g_{11}^{(i)}, g_{12}^{(i)}, \ldots, g_{1n}^{(i)}) \qquad (2.6)$$

Therefore the prototypes are points in the n-space of features.

One method of recognizing an observed sample $X = (x_1, x_2, \ldots, x_n)$ is to classify it into the class of the "closest" prototype.

Many functions have been used to measure the closeness of two points in the n-space. The Euclidean distance

$$d^2(G_j, X) = (G_j - X) \cdot (G_j - X) \qquad (2.7)$$

ample.

The shortcomings of such a method are three-fold. First when there are many prototypes a large number of

computations are required. Also the error rates are difficult to predict or control. Furthermore $d^2(G_j,X)$ depends upon the units chosen for the individual features: X = (5 dollars, 4 inches, 6 volts) versus X = (500 cents, 10 cm, 6000 mv).

One approach often used to "normalize" the n-space of features is to find weight vectors

$$w^{(i)} = (w_1{}^{(i)}, w_2{}^{(i)}, \ldots, w_n{}^{(i)}) \tag{2.8}$$

constrained by the product

$$\prod_{j=1}^{n} w_j{}^{(i)} = 1 \tag{2.9}$$

or by the sum

$$\sum_{j=1}^{n} w_j{}^{(i)} = 1 \tag{2.10}$$

so that the intra-class distances $d^2(G_m{}^{(i)}, G_k{}^{(i)})$ are minimized and the inter-class distances $d^2(G_m{}^{(i)}, G_k{}^{(j)})$ are maximized. The reason for doing this is that the prototypes of one class ought to be "close" whereas prototypes from different classes ought to be "distant".

Some investigators have attempted to measure the distance between classes [Ref. 7]. One distance is called the divergence and another the Bhatlacharyya. They are defined, respectively, for the two-class problem as

$$J(H_1,H_2) = \underset{H_1}{\mathcal{E}}\left[\ln \frac{d\tilde{F}_1(x)}{d\tilde{F}_2(x)}\right] - \underset{H_2}{\mathcal{E}}\left[\ln \frac{d\tilde{F}_1(x)}{d\tilde{F}_2(x)}\right] \tag{2.11}$$

and

$$B(H_1,H_2) = -\ln\left[\int_{-\infty}^{\infty}[d\tilde{F}_1(x)d\tilde{F}_2(x)]^{1/2}dx\right] \tag{2.12}$$

where the probability distribution of a sample X over the ith hypothesis is $\tilde{F}_i(X)$. These distances are not metric since the triangular inequality does not hold.

A tremendous amount of computation is involved in the determination of w, the weight vectors. And yet, such a method still leaves open the question of predicting the performance of the classifier in terms of error probabilities.

Nilsson uses hyperplanes to separate the classes in the n-space of measurements. A linear discriminant function for the ith class is formed by taking a dot product of the input and a weight vector for each class. This gives the discriminant function

$$d_i = X \cdot w^{(i)} \tag{2.13}$$

where the weight vector for the ith class is

$$w^{(i)} = (w_1(i), w_2(i), \ldots, w_n(i)) \tag{2.14}$$

The i for which d_i is the largest is chosen as the class. A recursive method of choosing $w^{(i)}$ so that linearly

separable patterns can be partitioned is given in Nilsson [Ref. 5, p. 79], under trainable linear classifiers. A theorem given by Nilsson assures a partition of the training set if the patterns are linearly separable.

The drawback to this approach is that often classes are not linearly separable. Also error rates are extremely difficult to compute.

3. Potential Functions

Another approach to the assignment of points of a finite-dimensional vector space to one of a family of classes on the basis of prototypes of those classes is called the method of potential functions [Ref. 8]. This method reduces to the construction of functions $q_i(X)$, one for each class, so that if

$$q_j(X) \geq q_i(X) \quad \text{for all } i \neq j \tag{2.15}$$

then $X = (x_1, x_2, \ldots, x_n)$ is classified as a member of class j, and where these functions are constructed as superpositions of potential functions f(X,G)

$$q_i = \frac{1}{m} \sum_{j=1}^{m} f(X, G_j^{(i)}) \tag{2.16}$$

The sum is over the prototypes of class i.

A reasonable set of restrictions on the potential

functions can be phrased in intuitive terms as:

a) $f(X,Y)$ should be maximum for $X = Y$.

b) $f(X,Y)$ should go to zero for X "distant" from Y.

c) $f(X,Y)$ should be smooth for easy analytic manipulations and decrease monotonically with the "distance" between X and Y.

d) $f(X,G^{(i)}) = f(X,G^{(j)})$ should imply that X is equally "similar" to the prototypes of class i and j.

A function often used for $f(X,Y)$ is

$$f(X,Y) = \frac{1}{1 + Ad^2(X,Y)} \tag{2.17}$$

where A is some constant and $d^2(X,Y)$ is some distance function. A form also used for the potential function is

$$f(X,Y) = A \exp \frac{-||X-Y||^2}{2\sigma^2} \tag{2.18}$$

where A and σ are constants and $||X-Y||^2$ is the norm square of the difference vector.

Clearly this function determines the way the space is partitioned. For instance if σ approaches 0, only the prototypes will be defined to belong to the classes, whereas when σ approaches ∞, increasing portions of the feature space will be defined.

It is not clear how the error probabilities are computed. Also the amount of computation one must carry out for each classification is very large.

4. Statistical Methods

The observations are often looked upon as random variables. This allows statistical methods to be applied to the classification problem. Usually some form of a Bayes test is used, as in the maximum likelihood classification technique. Such tests are optimum with respect to certain loss functions. However, some authors modify well known methods for computational or experimental expediency. In so doing they lose the optimality of the test, Reed's work described below being an example.

Both fixed length sample tests and sequential techniques have been used in pattern recognition. In this section many methods are discussed in detail with comments as to the special needs of each technique. Where appropriate, comments are made as to the inadequacy of the method.

a. Definition of Bayes Decision Rule

Bayes rule minimizes the average cost of making decisions [Ref. 9, p. 24]. The average cost r is computed as

$$r = \int_{\Gamma} \sum_{i=0}^{M} \delta(D_i|v) \left[\sum_{j=1}^{M} L_{ij} d\tilde{F}_j(v) P_j \right] dv \qquad (2.19)$$

where the decision rules are

$$\delta(D_i|v) = \text{Prob } \{\text{deciding } D_i \text{ when observing } v\} \qquad (2.20)$$

and $v \in \Gamma$. Clearly this integral is minimized if $\delta(D_i|v) = 1$ for the i which gives the minimum

$$\sum_{j=1}^{M} L_{ij} d\tilde{F}_j(v) P_j \leq \min_i \left[\sum_{j=1}^{M} L_{ij} d\tilde{F}_j(v) P_j \right] \qquad (2.21)$$

and

$$\delta(D_k|v) = 0 \quad \text{for } k \neq i \qquad (2.22)$$

This formulation is general enough to include many useful tests. The difficulty arises in choosing meaningful values for the loss functions. In pattern recognition applications a further difficulty is due to the complexity of computing the error rates for various loss functions. In Chapter III a method of choosing one meaningful form of the loss function that allows easy estimates of the errors is given.

b. Maximum Likelihood Decision

A special form of the Bayes test is the maximum likelihood decision rule. The criterion for decisions

is that the probability of the observation coming from the class be maximized. This is not necessarily the best thing to do, as an example will illustrate. However, under this criterion one chooses D_i such that

$$\frac{P_i d\tilde{F}_i(x)}{\sum_{j=1}^{k} P_j d\tilde{F}_j(x)} = \max_{l} \frac{P_l d\tilde{F}_l(x)}{\sum_{j=1}^{k} P_j d\tilde{F}_j(x)} \tag{2.23}$$

Clearly the denominator is constant for a given x. The rule is equivalent to choosing D_i such that

$$P_i d\tilde{F}_i(x) = \max_{j}\{P_j d\tilde{F}_j(x)\} \tag{2.24}$$

It can be shown that this is the Bayes rule with $L_{ii}=0$, $L_{ij}=1$, $i\neq j$. This rule classifies the observation without regard to the type of error that it is making. Consequently there is little control over the operating characteristic of the algorithm.

As an example consider three hypotheses H_1, H_2 and H_3 with dF_1, dF_2 and dF_3 as illustrated in Figure 2-2.

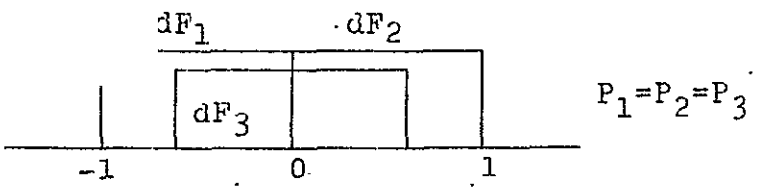

Fig. 2-2. Three Probability Densities

It is obvious that D_3 will never be chosen with this algorithm, even when x has come from H_3.

c. Tradeoff of Recognition Probability against Reject Rate

This section studies a method proposed by Chow [Ref. 10]. It concludes that both the definition of optimum and the proposed optimum rule are deficient. The total error probability and the reject rate are often used to characterize the performance of a pattern recognition system. Chow describes a classification and rejection rule based on these parameters.

Chow [Ref. 10 and 11] modifies the maximum likelihood classification technique. Optimum here means that a rule minimizes the reject probability for a given level of total misclassification. The rule rejects the pattern if the maximum of the likelihood function is less than a threshold. The rule is defined as

$$\delta(D_i|x) = \begin{cases} 1 \text{ if}^{1} \; P_i d\tilde{F}_i(x) \geq P_j d\tilde{F}_j(x) \text{ for all } j=1,2,\ldots,k \\ \qquad \text{and} \\ \qquad P_i d\tilde{F}_i(x) \geq (1-t) \sum_{i=1}^{n} P_i dF_i(x) \\ 0 \text{ otherwise} \end{cases} \tag{2.25}$$

---

[1]There is ambiguity when $P_i dF_i(x) = P_j dF_j(x)$, $i \neq j$. When this happens a decision can be made randomly.

$$\delta(D_0|x) = \begin{cases} 1 \text{ if } \delta(D_i|x) = 0 \text{ for all } i=1,2,\ldots,k \\ 0 \text{ if } \delta(D_i|x) = 1 \text{ for any } i=1,2,\ldots,k \end{cases} \tag{2.26}$$

The parameters $t \in [0,1]$ and controls the reject region. The error rate, reject rate and the probability of correct recognition are defined, respectively, as

$$E(t) = \int_{R_n} \sum_{i=1}^{k} \sum_{j=1}^{k} \delta(D_j|x) P_i d\tilde{F}_i(x) \tag{2.27}$$

$$R(t) = \int_{R_n} \delta(D_0|x) \sum_{i=1} P_i d\tilde{F}_i(x) \tag{2.28}$$

$$C(t) = 1 - E(t) - R(t) \tag{2.29}$$

Now two useful functions can be defined as

$$m(x) = \frac{\max_i \left[P_i d\tilde{F}_i(x)\right]}{dP(x)} \tag{2.30}$$

and

$$dP(x) = \sum_{i=1}^{k} P_i d\tilde{F}_i(x) \tag{2.31}$$

The decision rule can be restated in terms of these functions:

1) accept a pattern whenever $m(x) \geq 1-t$ (2.32)

2) reject the pattern whenever $m(x) < 1-t$ (2.33)

The region of acceptance ψ_A and the region of rejection

ψ_R can be defined in terms of these definitions:

$$\psi_A = \{x: m(x) \geq 1-t\} \tag{2.34}$$

$$\psi_R = \{x: m(x) < 1-t\} \tag{2.35}$$

Clearly the rejection and acceptance probabilities are

$$R(t) = \int_{\psi_R} dP(x) \tag{2.36}$$

$$A(t) = \int_{\psi_A} dP(x) \tag{2.37}$$

A few simple properties of the rejection threshold t are:

1) Both the error and reject rates are monotonic in t, one decreasing and the other increasing.

2) The reject threshold t is an upper bound on the error rate E(t). Let x be in ψ_A, the region where $\delta(D_0|x) = 0$. That is, $m(x) \geq (1-t)$ and

$$E(t) = A(t) - C(t) = \int_{\psi_A} [1-m(x)]dP(x)$$

$$\leq t\int_{\psi_A} dP(x) \leq tA(t) \leq t \tag{2.38}$$

3) The reject threshold t is a differential ratio of error rate and reject rate when R(t) can be differentiated.

$$\frac{dE}{dR} = -t \tag{2.39}$$

4) The probability of acceptance A(t) has the property $0 \leq A(t) \leq 1$ for $t \in [0,1]$. $A(0) \geq 0$, $A(1) = 1$.

There is a question concerning optimality. This method says, in effect, to raise a threshold (1-t) from zero to an appropriate value so that the total probability of error,

$$E(t) = \int \sum_{i=1}^{k} \sum_{j \neq i}^{k} (D_j|x) P_i dF_i(x)$$

$$= \sum_{i=1}^{k} \sum_{j \neq i}^{k} e_{ij} P_j \qquad (2.40)$$

equals a design criterion. One question that is unanswered is whether there are many decision rules that will give the same "optimality". This question arises since the total error rate in the definition relies only on the total of the probabilities. There are many sets $\{e_{ij}\}$ which give the same sum. Clearly not all such tests are optimal, from the user's point of view. This may mean that wrong costs were chosen. Here is an example that will point out the weakness of the above method. Consider the three hypotheses as illustrated in Figure 2-3. Assume that each class is equally likely. According to the definition $3dP = dF_1 + dF_2 + dF_3$. This is Equation 2.31, illustrated by Figure 2-4.

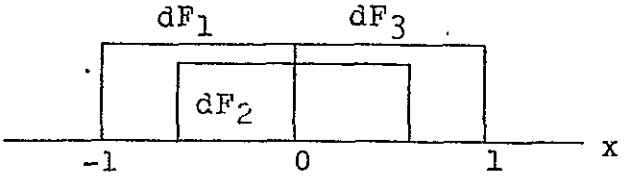

Fig. 2-3. Example for Chow's Method

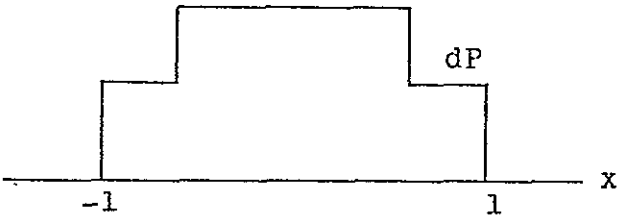

Fig. 2-4. Illustration of Eq. 2.31

Since $\max_i P_i d\tilde{F}_i(x) = 1$, for $-1 < x < 1$, then $m(x) = \frac{1}{dP(x)}$.

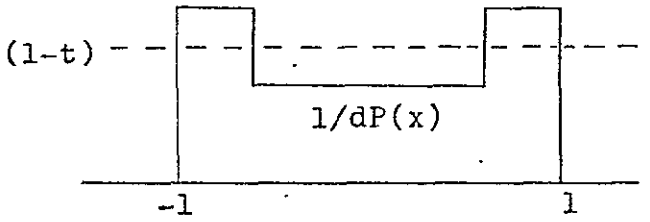

Fig. 2-5. Chow's Rejection Region

Figure 2-5 illustrates m(x), and (1-t) corresponds to some value of the total probability of error. The reject region is the interval in R_1 for which $m(x) < (1-t)$. The reject region is the whole of the region over which class 2 is defined. From the Bayes point of view this is a result of minimizing the average cost with respect to some cost. In particular it is like assigning a cost of

zero for misclassifying class 2. The unfortunate conclusion is that class 2 will never be detected.

Thus there are two flaws to Chow's method. First "optimum" is ambiguously defined in terms of total probabilities. Second the "optimum" rule may never detect certain classes, even when requested to do so. Consequently it is important to ask whether it is possible to devise a test which is defined in terms of the natural parameters of the pattern recognition problem, probability of false declarations $\{\alpha_i\}$ and probability of detection $\{\gamma_i\}$. Are there tests that give results which approximate the design criteria? In the next chapter this question is answered in precise terms.

d. Wald's Sequential Probability Ratio Test

Many applications use Wald's sequential probability ratio test. This test assumes the samples are from one of two classes. Samples from class i have a distribution $\tilde{F}_i$ and samples from class k have a distribution $\tilde{F}_k$.

Samples are taken one after the other. It is not necessary to assume that they are independent samples. However, such an assumption clearly reduces the requirements for computation. After n samples are taken, $x = (x_1, x_2, \ldots, x_n)$. A rejection takes place if

$$A_k < \frac{P_i d\tilde{F}_i(x)}{P_k d\tilde{F}_k(x)} < A_i \tag{2.41}$$

where A_k and A_i are limits which will be defined.

$$P_i d\tilde{F}_i(x) \geq A_i P_k d\tilde{F}_k(x) \tag{2.42}$$

then it is decided that the sample must have come from the class whose distribution is $F_i(x)$. The quantity A_i must satisfy Equation 2.42. Integrating over ψ_i the region of R_n, which gives the decision D_i, gives

$$\int_{\psi_i} P_i d\tilde{F}_i(x) \geq A_i \int_{\psi_i} P_k d\tilde{F}_k(x) \tag{2.43}$$

The left side is γ_i, the probability of correctly deciding class i, whereas the right integral is α_i, the probability of deciding i when decision k is correct.

This is the traditional presentation. It was pointed out to the author that a more careful study must be taken. In Chapter III ψ_i is defined more specifically.

$$\gamma_i \geq A_i \alpha_i \tag{2.44}$$

By neglecting the excess over the thresholds,

$$\gamma_i \approx A_i \alpha_i \tag{2.45}$$

or an approximation to the threshold A_i is

$$A_i \approx \gamma_i / \alpha_i \tag{2.46}$$

Hence the quantities A_i are chosen as the functions of α's and γ's.

For a discussion on termination of Wald's method see [Ref. 12], Wilks [Ref. 13, pp. 482-497], and Selin [Ref. 14, pp. 90-95]. In the above references and in Appendix 3 the average sample number at termination is computed.

Most techniques using this test take the data in a fixed manner. For geometrical data, quite often, the measurements are taken using scan-by-predetermined-lines or scan-by-matrix-digitization or edge followers.

Unfortunately in many cases the inputs to the algorithm do not take the assumed statistical form. Often the Gaussian assumptions are made for analytical convenience when little is really known about the inputs.

e. Extension of Wald's Sequential Probability Ratio Test

Wald and Sobel [Ref. 15] extended the hypothesis test to the three-class problem. However, as the title of their paper, "A Sequential Decision Procedure for Choosing One of Three Hypotheses concerning the Unknown Mean of a Normal Distribution," suggests, the problem they solved is related to the normal distribution.

Barnard [Ref. 16] and Armitage [Ref. 17] have extended Wald's sequential probability ratio test beyond

the two class problem in a more general way. Armitage is more concise, his work being an outline of Barnard's studies.

In their studies there are k hypotheses $H_1, H_2, \ldots, H_k$. Applying Wald's test to each pair of hypotheses, there are (1/2)k(k-1) likelihood ratios.

$$R_{ij} = \frac{P_i d\tilde{F}_i(x)}{P_j d\tilde{F}_j(x)} \text{ for all } i \neq j \tag{2.47}$$

And there are k ratios of the form

$$R_{ii} = 1 \tag{2.48}$$

The observations are taken sequentially until all the inequalities in one of the k sets are simultaneously satisfied. Accept hypothesis i if $R_{ij} > A_{ij}$ for each $j=1,2,\ldots,k$, where A_{ii} is made less than one. Two hypotheses cannot be accepted simultaneously when A_{ij} are chosen meaningfully.

This test terminates with probability one if the variance of the distribution of R_{ij} is finite. The proof of this is in Barnard and Armitage.

Rewriting the condition for accepting the ith hypothesis and neglecting the excess over the boundary

$$P_i d\tilde{F}_i(x) = A_{ij} P_j d\tilde{F}_j(x) \tag{2.49}$$

Integrating over the correct decision region of the ith hypothesis gives

$$A_{ij}P_j e_{ij} \simeq P_i e_{ii} < P_i \tag{2.50}$$

where e_{ij} is the probability of deciding i given H_j, and where the right inequality is noted for convenient over bounding of A_{ij}. Similar caution as in Equation 2.43 applies here. Hence

$$A_{ij} \simeq \frac{P_i}{P_j e_{ij}} \tag{2.51}$$

is a rule for choosing the boundaries.

Recalling that H_i is accepted when each $R_{ij} > A_{ij}$, it is clear that if the ith hypothesis is accepted then

$$e_{ij} > e'_{ij} \text{ for all } j \neq i \tag{2.52}$$

where e'_{ij} is the true error rate and e_{ij} is the desired error rate.

The desired false alarm rate α_i is

$$\alpha_i = \sum_{j \neq i} P_j e_{ij} \tag{2.53}$$

An estimate of the actual false alarm rate is

$$\sum_{j \neq i} P_j e'_{ij} = \alpha'_i < \alpha_i . \tag{2.54}$$

where α' is used to mean the probability of this test result being false.

There are two difficulties with this procedure. The first has to do with computational requirements and the second concerns a priori knowledge of $\{e_{ij}\}$.

After each sample is taken, (1/2)k(k-1) ratios are formed. Each ratio is compared to a level. Again, there are (1/2)k(k-1) tests. This algorithm requires an amount of computation which grows as the square of the number of classes grows. For 10 classes, 45 steps are required. For 64 classes, 2016 steps are needed for each sample! Such requirements proscribe real-time computation.

The second difficulty with this method is that often not all $\{e_{ij}\}$ are known. Sometimes it is of no concern what the individual e_{ij}, error rate, is. An example illustrates this point. In character recognition, it really does not matter what the probability of misclassifying "Q" into "B" is. What matters is, that once "B" is announced, that it be true with high probability. Next, when "Q" is given to a machine it is desired that the probability of it announcing "Q" be high. How the misprobability is distributed is immaterial. Again, α_i and γ_i are the fundamental quantitie of pattern recognition.

f. Reed's Generalized Sequential Probability Ratio Test

Reed [Ref. 18] proposed a ratio test for multi-class problems. But Fu [Ref. 19, p. 176] points out that, for more than two classes, it has not been shown that the procedure is justified. The only grace of the method is that if the number of classes is two, then the method coincides with Wald's method.

In Reed's method a ratio,

$$U_i(x) = \frac{P_i d\tilde{F}_i(x)}{\left[\prod_{j=1}^{k} P_j d\tilde{F}_j(x)\right]^{1/k}} \qquad i=1,2,\ldots,k \tag{2.55}$$

is formed at each stage of a sample. The notation $x = (x_1, x_2, \ldots, x_n)$ is used. The stopping boundaries are A_i,

$$A_i = \frac{P_i(1-e_{ii})}{\left[\prod_{j=1}^{k} P_j(1-e_{ij})\right]^{1/k}} \tag{2.56}$$

U_i is compared to A_i for every i, and H_i is rejected if $U_i < A_i$ for any such i. The number k is reduced by an appropriate amount and the U_i recomputed.

Analysis of this test behavior is not available except in the two-class problem.

5. Geometric Probability

Geometric probability [Ref. 20] has to do with the probabilities of certain basic geometric events such as a line intersecting a convex figure. First an appropriate measure is given to the basic elements. The basic element of measure for a random line $s(P,\theta)$ is assigned in Appendix 4. A uniform random line $s(P,\theta)$ is described by P and θ and the probability of such a line is proportional to $dPd\theta$. Then the probability of these events can be described as the integrals of the measure over the event. Many results relate only to convex figures. Of course in the real world, figures are more often nonconvex. However when the figures are convex many such probabilities are related in a simple manner to the basic features of the figure. If C represents the set of all random lines intersecting a convex figure with total perimeter L and if $dPd\theta$ is the element of measure for a random line then

$$\iint_C dPd\theta = L(\text{meters}) \tag{2.57}$$

In order for this to be a probability it must be normalized, usually by the perimeter of the retina. When one assigns zero or one to N, a random variable is formed which indicates whether there is an intersection. This equation becomes just

$$\mathcal{E}N = L(\text{meters}) \tag{2.58}$$

$$N = \begin{cases} 1 \text{ if } s(P,\theta) \text{ intersects } C \\ 0 \text{ if } s(P,\theta) \text{ does not intersect } C \end{cases} \tag{2.59}$$

For a complete proof of this see Kendall and Moran [Ref. 20, pp. 58-59].

Another interesting result is that if the basic element is taken as a point in the plane, and its measure is dxdy then

$$\iint_C dxdy = A(\text{meters}^2) \tag{2.60}$$

where A is the area. Again this must be normalized by the retina area or by some other constant. Let P, a random variable, equal zero when the point is outside of C. It is equal to one when it is inside the figure. The above integral becomes

$$\mathcal{E}P = A(\text{meters}^2) \tag{2.61}$$

Ball [Ref. 21] uses the moments of such random variables to perform classification. Scale invariance is obtained by raising such moments to appropriate powers, then taking ratios. For instance the feature,

$$\frac{\left[\iint_C dPd\theta\right]^2 (\text{meters})^2}{\iint_C dxdy\ (\text{meters}^2)} \tag{2.62}$$

is dimensionless. Cor / this feature is size invariant. It is also suggested that moments of functions of such basic random events be used as an input to a recognition machine. Integrals like

$$\int_\Omega f(s(P,\theta))dPd\theta,\ (P,\theta)\ \epsilon\ \Omega \tag{2.63}$$

are considered, where $s(P,\theta)$ is a random line and $dpd\theta$ its measure.

The moments[1] are estimated with the aid of the weak law of large numbers. That is,

$$\lim_{n\to\infty} \text{Prob}\left[\left|\frac{1}{n}\sum^{n}(\text{integrand}) - \mathcal{E}(\text{integrand})\right| > \epsilon\right] = 0 \tag{2.64}$$

where $\epsilon > 0$. This convergence may be slow causing errors in the estimate of the moment. When these numbers are raised to powers, so that the dimensions will cancel, uncertainties become greater[2]. The conclusion is that

---

[1]Ball does not use the random variables directly. He uses the term integral geometry. Moments are integrals.

[2]The relative maximum absolute error of a product is the sum of the relative maximum absolute errors of each factor. Hence the powers of uncertain numbers become more uncertain.

obtaining the moments by random samples is unsatisfactory. Other methods similar to quadrature in numeric integration are proposed.

B. EXPERIMENTS IN PATTERN RECOGNITION

These experiments are presented here to illustrate the measuring and classifying techniques that are often used.

1. Random Features

One of the earliest suggestions for the use of random lines in pattern recognition was made by Rubinstein [Ref. 22]. He used the average number of intersections that a random line makes with open angular figures to attempt the recognition of the type of intersection.

Ball [Ref. 21] introduces geometric probability to give the above method a firm foundation. Yet he too uses only the estimate of the various means as the input classification. On the other hand Wong uses the distribution of the random variables.

Wong [Ref. 23, pp. 535-546] uses random lines thrown against geometric shapes such as squares, circles and polygons to find the total length of intersection. This feature is used in Wald's sequential probability ratio test. He considers shapes that are similar in convex area. A set of fifteen simple basic figures are used.

Results of five pair-wise tests are reported.

The first experiment reported in Chapter V is an extension of this work. The figures considered were complex--block letters H and U.

The second experiment in Chapter V is a problem in multiclass classification. Pair-wise tests are avoided. All classes are considered at once using the algorithm of Chapter III. Furthermore the feature used in the experiment is the number of intersections a line makes with a pattern. The complete distribution of this feature is used.

2. Handwritten Character Classification

Demonstrations have shown that even humans perform rather poorly in recognizing handwritten characters out of context. The general problem is very difficult. Work has been done by Brain and Hart [Ref. 24] on special types of handwritten characters--printing in confined squares as on FORTRAN coding sheets [Fig. 2-6]. Their feature extraction is in two steps. First each character is quantized into a 24 x 24 matrix. The matrix is compared with 84 edges in 9 translated positions. The results of this edge detection are fed into a trainable linear classifier[1]. The training set consisted of 8000

[1]Trainable linear classifiers are discussed in Nilsson [Ref. 5, p. 79].

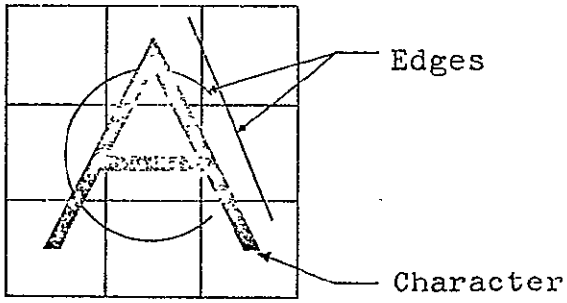

Fig. 2-6. Edge Detection of Geometric Figures

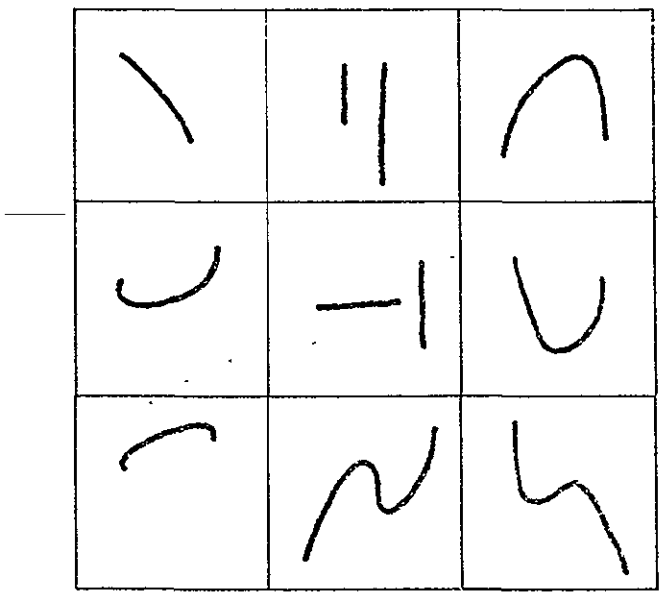

Fig. 2-7. Elementary Segments of Letters

characters from 10 writers. The reported error rate on material obtained from writers not included in the training set was about 20%. This is a rather poor result considering the number of computations--84 x 9 = 756 edge detections!

Segment analysis has been quite successful in pattern recognition of handwritten letters. Mori, et al. [Ref. 25] as well as Sheinberg [Ref. 26] have used elementary strokes as input to a linguistic pattern recognizer. Some of the elements used are pictured in Figure 2-7. Both groups of investigators have been successful and have machines on the market. Their operating characteristics are not quoted here due to the unknown experimental standards. These characteristics are sensitive to the source of the experimental data.

Fu [Ref. 19, p. 36] reports a handwritten character classification system that makes 7% error. The number of computations that it needs is far less than required above.

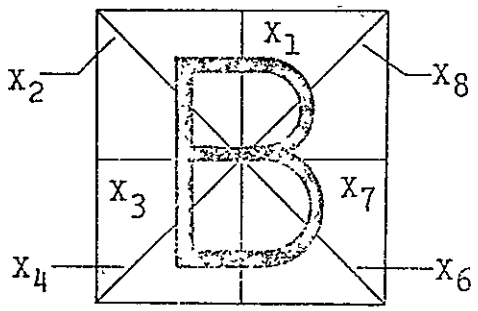

Fig. 2-8. Preselected Line Intersection

The feature which is extracted is the length that the predetermined lines 1,2,...,8 make with the figure [Fig. 2-8]. Therefore the measurement is $X_1, X_2, \ldots, X_8$. To get the test it is assumed that $X_1, X_2, \ldots, X_8$ are Gaussian, with a mean and variance depending on the figure. The sequential probability ratio test [Ref. 12] is employed. On a set of two characters, A and B, decisions are made on the average after six measurements. On a set of four characters, a b c d, a similar experiment shows that the average number of measurements needed to make decisions with 7% error rates is less than ten. [Ref. 19, p. 39].

It shall be noted that this system is sensitive to alignment and the algorithm falls apart if the centering of the figure exceeds a tolerance. The foundation for the Gaussian assumption is weak and the prediction of the probability of error depends on this important assumption.

3. Machine Produced Impact Printing

A method widely used for feature extraction of machine produced impact printing is to scan the page in a zig-zag pattern [Fig. 2-9] or by a group of parallel lines [Fig. 2-10].

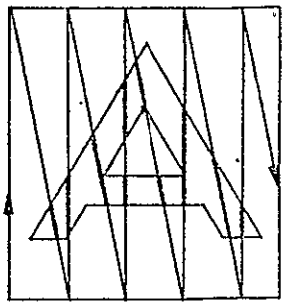

Fig. 2-9. Zig-Zag Scan

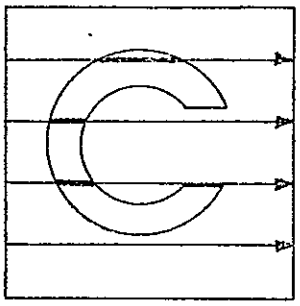

Fig. 2-10. Parallel Line Scan

The continuous data stream is matched-filtered (usually digitally) with patterns already stored in a machine. The most sophisticated machine on the market is the IBM 1975 Optical Page Reader [Ref. 27], which operates in the one-error-per-million region. This performance is attained by checking for context when it is "uncertain" about a character's classification. Also it has a set of stored patterns for each font.

The IBM 1975 Optical Page Reader is operational and is used by the Social Security Administration to digitize quarterly employers' reports. It checks context by looking through a dictionary of names. The IBM 1975 is a specialized machine which is prohibitively expensive for most applications.

Other similar experiments have used matrix digitization or edge followers to quantize the visual data. The matrix scan or the scan-by-predetermined-lines techniques are subject to alignment constraints. Character

registration also plays an important part in the machine's performance. Skewed or smudged characters play a critical role. It would be desirable if one could find a feature extraction that is invariant to displacement, rotation, and size.

CHAPTER III

PROPOSED SEQUENTIAL MULTICLASS HYPOTHESIS TEST

In this chapter a sequential hypothesis test is proposed. It is one solution to the multiclass classification problem. Its attributes are:

a) Its performance (in terms of the error rate of the first kind and the probability of detection) can be controlled.

b) At each step it is a Bayes test with a special structure for the loss function.

c) It terminates almost surely.

d) For the two-class problem it is a Wald's test.

e) The average sample number required for a test can be estimated.

The notation used throughout this chapter is defined in Chapter I, Section C. In the subsequent sections each of the properties listed above is derived or proven.

The proposed sequential multiclass test will have two forms well suited for rapid computation. But before the algorithm is presented a few words concerning the motivations for forming such a test will be given. Also some interpretations of the algorithm will be given.

It was pointed out earlier that in pattern recognition the important parameters are the error rates of the first and second kinds, α_i and β_i, respectively. That is, if a machine announces a decision, for instance that a letter on a page is Q, the user wishes to have an estimate on the probability of that declaration being incorrect. In many applications it matters not what letter is really there, if indeed an error has been made. Hence, it may be reasonable to formulate the decision algorithm on the basis of the likelihood of an observation compared to some acceptable threshold. If there are M classes, M successive comparisons may be made. In other words, for each possible decision one can ask whether the likelihood of the observation coming from a class is sufficiently greater than the average likelihood of the observation coming from any other class. More precisely, decide D_i if

$$\frac{\sum_{j\neq i} P_j d\tilde{F}_j(v_1,v_2,\ldots,v_n)}{P_i d\tilde{F}_i(v_1,v_2,\ldots,v_n)} < C_i \qquad (3.1)$$

otherwise take more samples.

It turns out that this test is in a very convenient form for the computation of α_i and β_i as will be shown in subsequent sections.

Clearly Equation 3.1 looks familiar. It is shown

that this equation collapses into Wald's test for the two-class problem. Also, this form of the test looks like the generalized M-hypothesis test in Van Trees [Ref. 9, p. 48]. But unlike the M-hypothesis test for which the error rates are difficult to compute, as stated by Van Trees, the error rates for the proposed test are simple to compute.

The proposed sequential multiclass hypothesis test will have two forms. The observations $v_1, v_2, \ldots, v_{n-1}$ have been observed and the test has not yet terminated. That is, the decision so far is D_0. The proposed algorithm, first form, is as follows:

1) Take the nth sample v_n and let $v = (v_1, v_2, \ldots, v_n)$

2) Take another observation (declare D_0, $n \leftarrow n+1$, go to Step 1), if for every $i = 1,2,\ldots,M$

$$\sum_{j \neq i} P_j d\tilde{F}_j(v) \geq C_i P_i d\tilde{F}_i(v) \tag{3.2}$$

3) The test terminates (go to Step 4) if for any $i = 1,2,\ldots,M$

$$\sum_{j \neq i} P_j d\tilde{F}_j(v) < C_i P_i d\tilde{F}_i(v) \tag{3.3}$$

4) Choose to verify H_i which minimizes

$$\sum_{j \neq i} P_j d\tilde{F}_j(v) - C_i P_i d\tilde{F}_i(v) \tag{3.4}$$

5) Declare D_i. Terminate.

Step 4 above is intuitively appealing. It will be shown that the ratio of the false alarm rate α_i and the probability of detection γ_i is less than the thresholds used in the test.

$$C_i \leq \frac{\alpha_i}{\gamma_i} \tag{3.5}$$

Some manipulation will show that Step 4 directs the algorithm to choose a class in the quickest way while not overstepping that requirement.

In Appendix 1 an equivalent form of the test is derived. The second form is more convenient for computation. The proposed algorithm, second form, is as follows:

1) Take the nth observation v_n and let $v = (v_1, v_2, \ldots, v_n)$.

2) If

$$\sum_{j=1}^{M} P_j d\tilde{F}_j(v) \geq \max_i \{(C_i+1) P_i d\tilde{F}_i(v)\} \tag{3.6}$$

take another sample (declare D_0, $n \leftarrow n+1$, go to Step 1).

3) Otherwise choose to verify H_i so that

$$(C_i+1) P_i d\tilde{F}_i(v) = \max_j \{(C_j+1) P_j d\tilde{F}_j(v)\} \tag{3.7}$$

A. DETERMINATION OF $\{\alpha_j\}$ AND $\{\gamma_i\}$ FOR A TEST WITH CONSTANTS $\{C_i\}$

Here it is supposed that a test as described in Equations 3.2 to 3.7 has the constants $\{C_i\}$ fixed at some known numbers.

When a test terminates at the nth stage with D_i,

$$\sum_{j\neq i} P_j d\tilde{F}_j(v) < C_i P_i d\tilde{F}_i(v) \tag{3.8}$$

Neglecting the excess over the boundary allows Inequality 3.8 to be written as an approximation,

$$\sum_{j\neq i} P_j d\tilde{F}_j(v) \simeq C_i P_i d\tilde{F}_i(v) \tag{3.9}$$

Now integrating over the region of v, $\psi_i^{(n)}$, such that the test terminates at stage n with D_i, gives

$$\int_{\overline{\psi_i(n)}} \sum_{j\neq i} P_j d\tilde{F}_j(v) \simeq \int_{\psi_i(n)} C_i P_i d\tilde{F}_i(v) \tag{3.10}$$

which is the same as

$$\sum_{j\neq i} P_j \int_{\psi_i(n)} d\tilde{F}_j(v) \simeq C_i P_i \int_{\psi_i(n)} d\tilde{F}_i(v) \tag{3.11}$$

But by Equation 1.15, the definition of $e_{ij}^{(n)}$, the above becomes

$$\sum_{j\neq i} P_j e_{ij}^{(n)} \simeq C_i P_i e_{ii}^{(n)} \tag{3.12}$$

Using the error probabilities when the test stops at the nth stage, $e_{ij}^{(n)}$, the total error probabilities are

$$\sum_{n} \sum_{j \neq i} P_j e_{ij}^{(n)} \simeq \sum_{n} C_i P_i e_{ii}^{(n)} \tag{3.13}$$

Interchange of summations on the left gives

$$\sum_{j \neq i} P_j \sum_{n} e_{ij}^{(n)} \simeq C_i P_i \sum_{n} e_{ii}^{(n)} \tag{3.14}$$

Using the definition of the error probabilities e_{ij}, Equation 1.17, this can be rewritten as

$$\sum_{j \neq i} P_j e_{ij} \simeq C_i P_i e_{ii} \tag{3.15}$$

Making use of the definitions of the error of the first kind α_i and the detection probability γ_i Equation 3.15 can be written as

$$\alpha_i \simeq C_i \gamma_i \tag{3.16}$$

The reasoning used so far gives an approximation of C_i.

$$C_i \simeq \frac{\alpha_i}{\gamma_i} \tag{3.17}$$

One further condition is made to obtain still a simpler estimate of the threshold C_i. If the detection rate is high, it is approximately equal to the a priori probability of that class.

$$\gamma_i \approx P_i \tag{3.18}$$

This suggests the following important point. If one desires a recognition system with false declaration probabilities $\{\alpha_i\}$ then one chooses

$$C_i \approx \frac{\alpha_i}{P_i} \tag{3.19}$$

In Appendix 2 the Chernoff Bound is used to show that $\gamma_i \to P_i$ as $n \to \infty$. This will establish more firmly the value of this approximation method.

This is a flexible algorithm. Output error rates are under the control of the user.

Extensive experimental results reported in Chapter V verify the usefulness of Approximation 3.19.

B. RELATION TO THE BAYES TEST

This section shows how the proposed test relates to the Bayes test for any fixed number of samples n. It is shown first that the reject region is the same as a Bayes test with particular loss functions. Next it is shown that the decision regions are the same. In fact, the proposed test is a Bayes test with a special cost structure which permits rapid computation of the error rates. This last computation is not generally feasible for the Bayes test. But as shown in Section A of this chapter,

the error rates for the proposed test can be readily computed and controlled.

The Bayes rule minimizes the average risk. In Chapter II, Section A it was shown that D_i is chosen to minimize

$$\sum_{j=1}^{M} L_{ij} P_j d\tilde{F}_j(v) \tag{3.20}$$

over all $i = 0,1,\ldots,M$

Suppose that the Bayes rule rejects all hypotheses after taking n samples. Then for every $k = 1,2,\ldots,M$,

$$\sum_{j=1}^{M} L_{0j} P_j d\tilde{F}_j(v) \leq \sum_{j=1}^{M} L_{kj} P_j d\tilde{F}_j(v) \tag{3.21}$$

which can be rewritten as (for a particular k=i)

$$\sum_{j\neq i}^{M} (L_{ij}-L_{0j}) P_j d\tilde{F}_j(x) \geq (L_{0i}-L_{ii}) P_i d\tilde{F}_i(v) \tag{3.22}$$

For the proposed test a rejection occurs at the nth stage whenever, for every $i = 1,2,\ldots,M$,

$$\sum_{j\neq i}^{M} P_j d\tilde{F}_j(v) \geq C_i P_i d\tilde{F}_i(v) \tag{3.23}$$

Clearly the proposed test and the Bayes rule have the same rejection criteria if, for all $k = 1,2,\ldots,M$,

$$C_i = \frac{L_{0i} - L_{ii}}{L_{ik} - L_{0k}} \quad k \neq i \tag{3.24}$$

The division implies that $(L_{ik} - L_{0k}) \neq 0$.

A solution to these equations can be obtained by observation. They are, for $i \neq j \neq 0$,

$$L_{0i} = 0$$

$$L_{ij} = 1$$

$$L_{ii} = -C_i \tag{3.25}$$

Clearly these loss functions satisfy

$$C_i = \frac{L_{0i} - L_{ii}}{L_{ik} - L_{0k}} = \frac{C_i}{1} = C_i \tag{3.26}$$

and

$$(L_{ik} - L_{0k}) = 1 \text{ for all } k \neq i \tag{3.27}$$

These assignments also satisfy one's intuition. If there is a reject at the nth stage there is no loss because another sample is taken and the test is continued. However, if there is a misclassification, then a penalty of one is assigned. On the other hand, when the correct decision is made, a reward is given. If the permitted error type of the first kind α_i is large, so is C_i.

Thus, when a correct answer is given under conditions allowing more errors, the reward is also larger.

Now suppose that a Bayes rule makes decision D_i, $i \neq 0$, after n observations. Then,

$$\sum_{j=1}^{M} L_{ij} P_j d\tilde{F}_j(v) \leq \sum_{j=1}^{M} L_{kj} P_j d\tilde{F}_j(v) \tag{3.28}$$

for every $k = 1,2,\ldots,M$.

Substitution for the loss functions yields,

$$\sum_{j \neq i} P_j d\tilde{F}_j(v) - C_i P_i d\tilde{F}_i(v) \leq \sum_{j \neq k} P_j d\tilde{F}_j(v) - C_k P_k d\tilde{F}_k(v) \tag{3.29}$$

Eliminating the common terms from both sides gives

$$-(C_i+1) P_i d\tilde{F}_i(v) \leq -(C_k+1) P_k d\tilde{F}_k(v) \tag{3.30}$$

Hence the Bayes rule is

$$\max_{i \neq 0} \{(C_i+1) P_i d\tilde{F}_i(v)\} \tag{3.31}$$

It can be seen that the proposed test, Equation 3.7, is indeed a special form of a Bayes rule.

C. TERMINATION

It will be shown that the proposed test terminates almost surely. Doob [Ref. 28, p. 349] shows the

convergence of the probability ratio

$$\frac{d\tilde{F}_k(v)}{d\tilde{F}_i(v)} \rightarrow 0 \quad \text{a.s.} \tag{3.32}$$

given H_i. This fact will be used below.

Now suppose that some hypothesis H_i is given. Recall from Equation 3.3 that the test terminates whenever

$$C_i > \frac{\sum_{j\neq i} P_j d\tilde{F}_j(v)}{P_i d\tilde{F}_i(v)} \tag{3.33}$$

That is whenever the ratio is less than C_i the test ends. It will be shown that the ratio in Equation 3.33 will approach zero with probability one. The right side of Equation 3.33 can be written as

$$\sum_{j\neq i} \frac{P_j d\tilde{F}_j(v)}{P_i d\tilde{F}_i(v)} = \sum_{j\neq i} \frac{P_j}{P_i} Z_n(i,j) \tag{3.34}$$

where by Definition 1.6

$$Z_n(i,j) = \frac{d\tilde{F}_j(v_1,v_2,\ldots,v_n)}{d\tilde{F}_i(v_1,v_2,\ldots,v_n)} \tag{3.35}$$

But by Equation 3.32

$$Z_n(i,j) \rightarrow 0 \quad \text{a.s.} \quad \text{for all } j\neq i$$

Hence the test terminates almost surely.

In any practical application of a sequential test, one must consider a large number such that when the sample number reaches this number the test is arbitrarily terminated. Because the error rates go to zero as n becomes large, this safeguard should not affect the performance significantly. In the tests reported in Chapter V the arbitrary termination point is placed at about ten times the average sample number. In more than five thousand cases which are run, not one test has a sample number this large.

D. RELATION TO WALD'S TEST

The proposed test for the two-class problem is clearly Wald's test. The proposed test takes another sample if

$$P_2 d\tilde{F}_2(v) > C_1 P_1 d\tilde{F}_1(v) \tag{3.36}$$

and

$$P_1 d\tilde{F}_1(v) > C_2 P_2 d\tilde{F}_2(v) \tag{3.37}$$

But rewriting these equations we get

$$C_1 < \frac{P_2 d\tilde{F}_2(v)}{P_1 d\tilde{F}_1(v)} < \frac{1}{C_2} \tag{3.38}$$

This shows that the proposed test takes another sample if the ratio of probability has not crossed either of the two boundaries.

Equation 3.38 is the same as Equation 2.41 with the lower boundary

$$A_k = C_1 \tag{3.39}$$

and the upper boundary

$$A_i = C_2^{-1} \tag{3.40}$$

E. APPROXIMATION OF THE AVERAGE SAMPLE NUMBER

In Section C the termination of the proposed test with probability one was shown. This section addresses the problem of computing an approximation of the average number of samples at termination under H_i.

Two assumptions will be made to assist in this approximation. First, assume that some one probability distribution F_k causes the delay of a decision or the incorrect classification. Second, assume that when any decision is made $m = 1,2,\ldots,M$,

$$\sum_{j \neq i} P_j d\tilde{F}_j(v) \simeq C_m P_m d\tilde{F}_m(v) \tag{3.41}$$

This is sometimes known as neglecting the excess over the threshold. It is similar to assuming that each step toward a goal is small and that the goal is far. Hence, when the goal is crossed, the position is near the goal.

With these assumptions, Equation 3.3 becomes

$$\sum_{j \neq i} P_j d\tilde{F}_j(v) \simeq C_i P_i d\tilde{F}_i(v) \tag{3.42}$$

when the correct decision is made, and

$$\sum_{j \neq k} P_j d\tilde{F}_j(v) \simeq C_k P_k d\tilde{F}_k(v) \tag{3.43}$$

when an incorrect decision is made.

By invoking the first assumption that F_k causes the delay, the sum is approximated by one term. Hence,

$$\frac{d\tilde{F}_k(v)}{d\tilde{F}_i(v)} \simeq C_i \frac{P_i}{P_k} \tag{3.44}$$

when the correct decision is made, and

$$\frac{d\tilde{F}_k(v)}{d\tilde{F}_i(v)} \simeq \frac{P_i}{C_k P_k} \tag{3.45}$$

when an incorrect decision is made.

N is the sample number at termination. Taking the logarithm of the ratio of probabilities at termination and denoting it $\tilde{Z}_N(i,k)$ gives

$$\tilde{Z}_N(i,k) = \ln \frac{d\tilde{F}_k(v)}{d\tilde{F}_i(v)} \simeq \ln \frac{C_i P_i}{P_k} \tag{3.46}$$

when the correct decision is made, and

$$\tilde{Z}_N(i,k) = \ln \frac{d\tilde{F}_k(v)}{d\tilde{F}_i(v)} \simeq \ln \frac{P_i}{C_k P_k} \tag{3.47}$$

when the incorrect decision is made.

Under the ith hypothesis, decision D_i is made with probability

$$\text{Prob}\{D_i | H_i \text{ true}\} = e_{ii} \tag{3.48}$$

The probability of an error is

$$\text{Prob}\{D_k | H_i \text{ true}\} = 1 - e_{ii} \tag{3.49}$$

Hence, the conditional expectation of the logarithm of the ratio of the probabilities at termination $\underset{H_i}{\mathcal{E}}(\tilde{Z}_N(i,k))$ can be computed,

$$\underset{H_i}{\mathcal{E}}(\tilde{Z}_N(i,k)) \simeq e_{ii} \ln \frac{C_i P_i}{P_k} + (1-e_{ii}) \ln \frac{P_i}{C_k P_k} \tag{3.50}$$

But from Appendix 3,

$$\underset{H_i}{\mathcal{E}}(\tilde{Z}_N(i,k)) = \underset{H_i}{\mathcal{E}}(\tilde{z}(i,k)) \underset{H_i}{\mathcal{E}}(N) \tag{3.51}$$

Therefore, the average sample number given H_i is approximately,

$$\underset{H_i}{\mathcal{E}}(N) \simeq \frac{e_{ii} \ln \frac{C_i P_i}{P_k} + (1-e_{ii}) \ln \frac{P_i}{C_k P_k}}{\underset{H_i}{\mathcal{E}}(\tilde{z}(i,k))} \tag{3.52}$$

In some applications a portion of the observations is not used. In particular, when the number of intersections that a random line makes with a pattern is used as a feature, neglecting those observations with zero intersections and using the conditional probability distributions provide for size invariance. This is more carefully presented in the following chapters.

The estimate of the average sample number, Equation 3.52, is an approximation of the average number of samples used per test. To obtain the average number of samples observed one must modify Equation 3.52 to reflect the fraction of the observations which are neglected.

Extensive experiments reported in Chapter V show that this estimate is a good one.

CHAPTER IV

INVARIANT FEATURE EXTRACTION

A. HEURISTIC DISCUSSION

In Chapter II some methods of taking data were mentioned. Many of these methods are sensitive to the location and orientation of the object under study. In this chapter methods which are not dependent on these factors are developed. The cost of aligning the patterns to the transducer motivates such a development.

In hand-printed material, the reasons for variation in location and orientation are clear. There are similar reasons why impact printed characters also have alignment irregularities. For example in high speed computer printouts the characters may be misplaced horizontally or vertically. This is due to variations in the way the hammer strikes the moving letter form. In aerial photographic reconnaissance, the significance of the data may be unrelated to either the precise location or the orientation of the object being sought. If one is scanning a picture for airfields, its recognition should not be affected by its whereabouts.

There is a need for an observation scheme which is

invariant to certain features. Precise location and orientation are two aspects of geometric subjects which do not contribute to their classification. A square and a rectangle ought to be different no matter where they are in the field of view. Another feature which is often unimportant from one class to another is size.

This is not to say that location, orientation, and size are unimportant. Often these features give different meanings to the same symbol. Arrows are good examples, "←" and "→" having opposite meanings. Also observe how these same[1] shapes, b d p q, are used to represent quite different things. The search for an invariant feature extraction method which will classify these letters in the same class is still fruitful. There are other methods which can subclassify them.

B. INVARIANCE

Invariance of decision rules is discussed in detail in Ferguson [Ref. 30, p. 144]. It is defined for a group[2] of measurable[3] transformations over the space upon which decision theory is founded. Here the concern is over

---

[1]These shapes are rotational and mirror-image transformations of one another.

[2]For the exact definition of group see Birkoff [Ref. 31, p. 117].

[3]It must be measurable to assure that a random variable X is transformed into a random variable g(x). See Breiman [Ref. 29, p. 106].

the invariance of the observations when the transformations are applied to the patterns in the retina. The decision algorithm is constant.

What features are insensitive to particular transformations? How can observations be taken so that they are independent of the transformations? It is shown in Section D of this chapter that randomizing answers these questions.

C. FEATURES

Any aspect or quantity derived from a pattern is a feature. The word feature is used to mean a scalar, vector or matrix quantity. The area, the perimeter, the convex hull perimeter and the "convexity" of a geometric shape are four examples of features. The gray level of a matrix scan is an example of a widely used feature.

It is tempting to try to measure the usefulness of a feature. However it is quite difficult to assign a numerical quantity to the usefulness of a feature. Investigators have used the entropy[1] of features as a measure. Others have used variance. It is unclear how either of these quantities relate to the fundamental quantities of pattern recognition (error and reject rates). Since the object of pattern recognition is to

---

[1]For a definition see Ash [Ref. 32, p. 24].

give the best classification as quickly as possible, a feature is best if a decision algorithm requires the least amount of such features.

A good feature is one which requires a minimum number of samples for a given decision algorithm operating at certain error and reject rates, and which can be extracted with a minimum of effort.

D. RANDOM EXTRACTION

One way to take data is by cross-correlating the pattern with certain reference elements. These reference elements may be the most primary elements of geometry or they may be as complex as the prototypes of the patterns. Here the basic elements of geometry are chosen as the reference elements due to the ease with which they can be generated. They are taken to be appropriately distributed within the retina. The reference elements are taken at random. This is done so that the features will no longer be dependent upon the location and orientation.

As an illustration of feature extraction (see Figure 4-1) let the reference set be the points in the retina. The elements are chosen one by one, at random and uniformly. These are correlated with the pattern. The result is a random variable,

$$X(x,y) = \begin{cases} 1 \text{ if } (x,y) \in \{\text{Pattern Interior}\} \\ 0 \text{ if } (x,y) \in \{\text{Pattern Exterior}\} \end{cases} \qquad (4.1)$$

Clearly X is a feature of the pattern. It is independent of where the pattern is in the retina. Also functions of X are features of the pattern.

Observe that the mean of X is proportional to the area of the pattern.

$$\frac{\text{Area Pattern}}{\text{Area Retina}} = (X) \qquad (4.2)$$

Unfortunately in most geometric pattern recognition problems this data is insufficient for classification because many different shapes may have the same area.

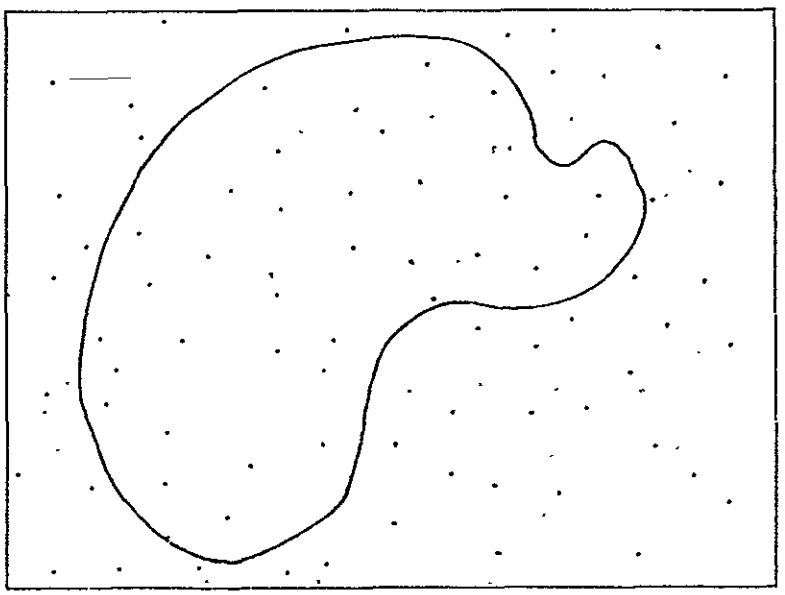

Fig. 4-1. Random Points Used as a Feature Extraction

Next consider the set of all lines intersecting the retina. These lines may be parameterized by the polar coordinates (P,θ) of the point closest to the origin. Random lines uniformly distributed over the retina may be obtained by choosing P and θ uniformly in $0 < P \leq R$, $0 < \theta \leq 2\pi$, respectively, where R is the retina radius. (See Appendix 4.)

Features of the pattern may be obtained by observing the cross-correlation of such lines with the pattern. Let X be equal to the total length of the intersection of the line with the pattern. It is clear that X is independent of where the pattern is situated. X is a random variable which depends only on the pattern itself. The properties of this random variable and other random variables derived from random lines that intersect the pattern are discussed fully in the next section.

Other geometric elements can be used as the reference set (Figure 4-2). However, when the elements become complex, the process of making their measure not dependent on orientation and location also becomes complex. Random ellipses can be used as a basis for feature extraction. Random variables can be defined in terms of the length of intersection, number of intersections, etc. Circles, lines, and points are degenerate forms of ellipses. It is questionable whether these random variables can be

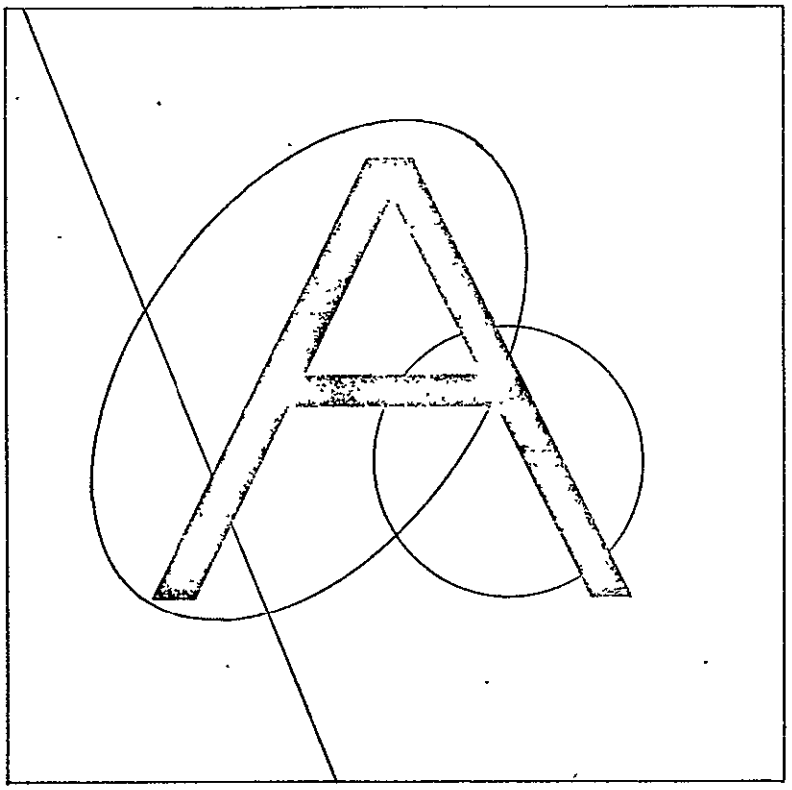

Fig. 4-2. Intersection of Ellipses with a Pattern

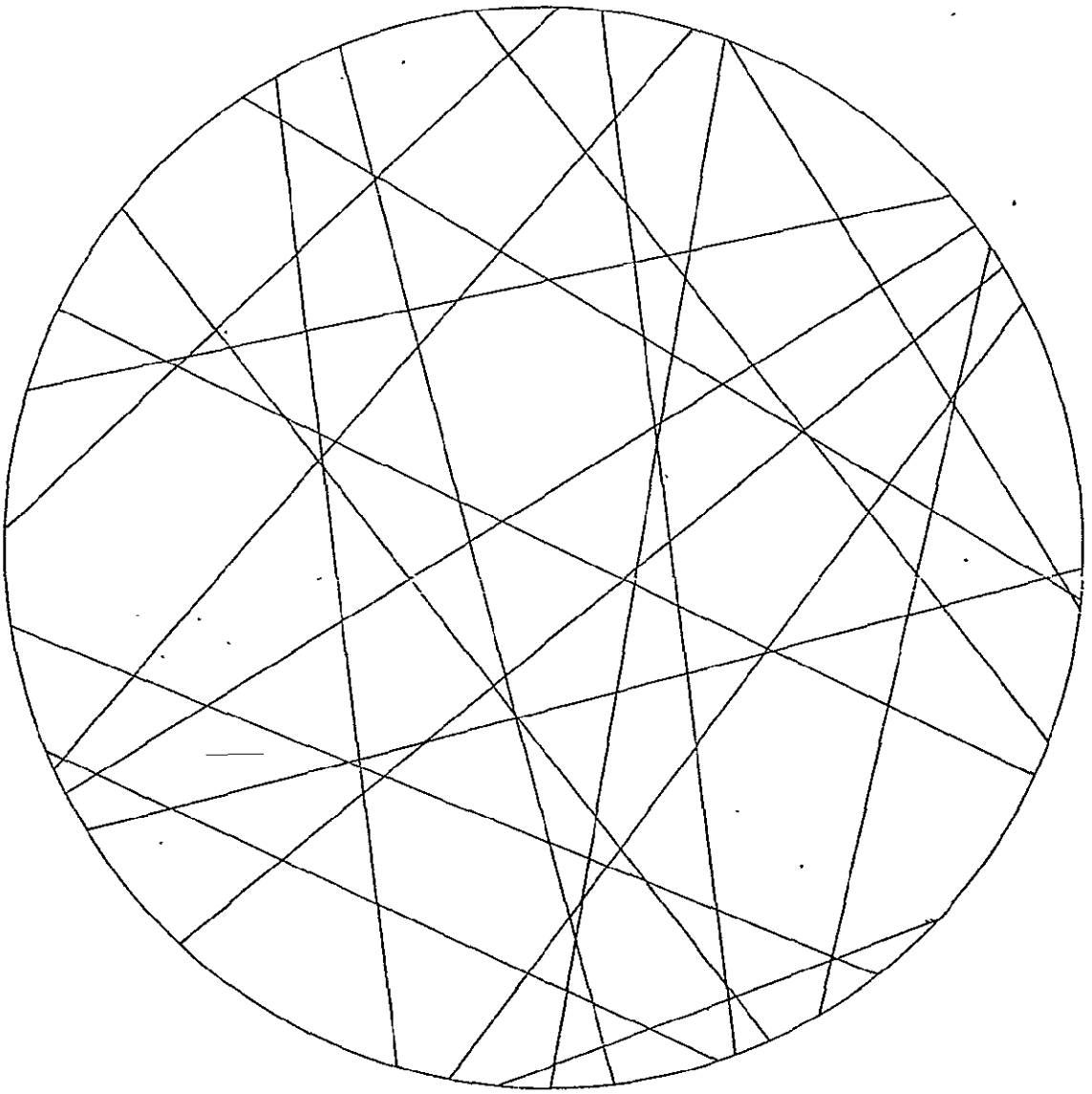

Fig. 4-3. Uniform Random Lines in a Retina

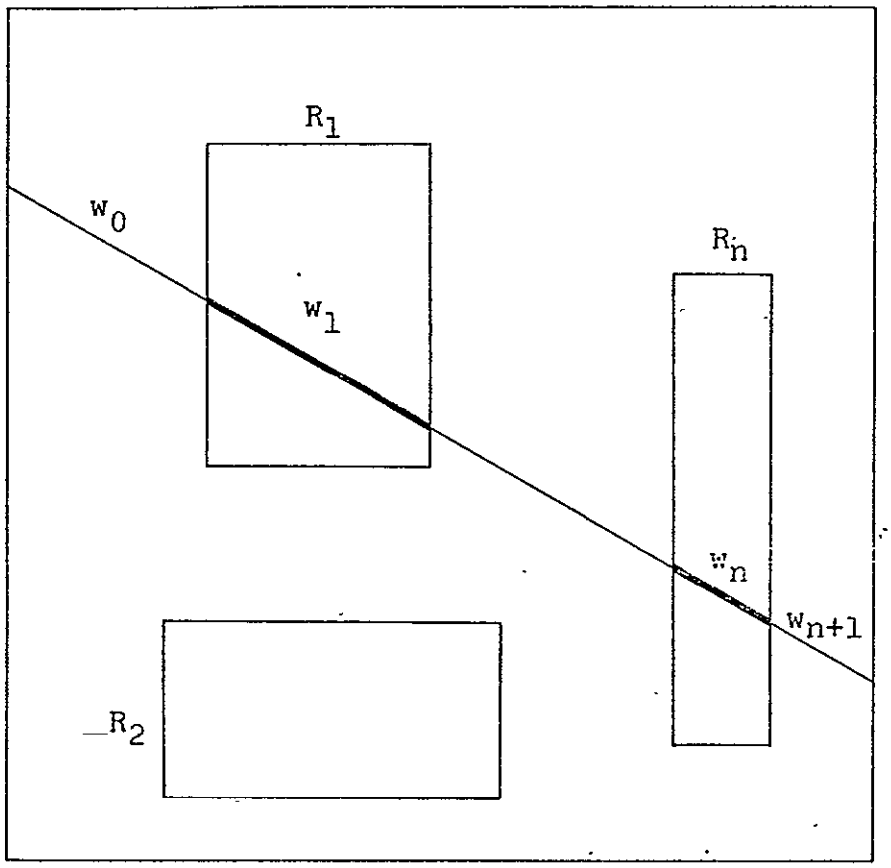

Fig. 4-4. Detail of a Line Intersecting a Pattern

made independent of the location and orientation of the figures. Also the computational disadvantages of using higher order elements limit the scope of this dissertation to the use of random lines.

E. PATTERNS AND RANDOM LINES

Random lines have been defined earlier. (Also see Appendix 4.) Figure 4-3 shows a field of random lines. Patterns have not been given a formal definition and it will be defined only implicitly by examples. In Figure 4-4 a pattern is represented by n rectangles $R_1, R_2, \ldots, R_n$. These n rectangles jointly form a pattern. They are disjointed so that the concept of a line intersecting a pattern can be clearly illustrated. The two segments, w_0 and w_{n+1}, are dependent upon the position and orientation of the pattern. The w_i represents the intersection of the line with the ith region of the pattern. There are many functions that can be formed from the w_i. A few of them are:

a) a multivariate random variable

$$W = (w_1, w_2, \ldots, w_n) \tag{4.3}$$

b) the largest intersection segment

$$U = \max(w_1, w_2, \ldots, w_n) \tag{4.4}$$

c) the smallest non-zero segment

$$V = \min(w_1, w_2, \ldots, w_n) \tag{4.5}$$

d) the number of intersections

$$N = \sum_{i=1}^{n} \operatorname{sgn}(w_i) \tag{4.6}$$

where sgn(·) is the sign function, +1 when the argument is positive, -1 when negative

e) the total length of intersection

$$X = \sum_{i=1}^{n} w_i \tag{4.7}$$

f) the joint random variable

$$Z = (N, X) \tag{4.8}$$

The multivariate random variable W has all the information contained in the other random variables. But the computational requirements to estimate, store, and use multivariate random variables are severe. Hence for these reasons and not on the basis of theoretics, the multivariate random variables are no longer considered.

The random variable U swamps the small contributions of the lesser w_i. Yet it may be these small quantities which make the pattern different. V is formed by the

smallest w_i, and it is susceptible to noise. These two random variables will no longer be considered.

The number of intersections N has interesting properties. It was pointed out that its average was related to the perimeter if the pattern is convex. (See Chapter II.)

$$\mathcal{E}N = \text{Perimeter} \tag{4.9}$$

For nonconvex figures $\mathcal{E}N$ can be used as an indicator of its convexity[1].

$$\text{Perimeter of convex hull} = \int_{\text{Pattern}} dP d\theta$$

$$< \text{Perimeter} = \int_{\text{Pattern}} N dP d\theta = \mathcal{E}N \tag{4.10}$$

The probability density function of N is more interesting. Clearly the probability of intersection is determined by the convex hull of the pattern. For a convex pattern there is one intersection. But an arbitrary shape has a probability density function which depends on the figure. As will be illustrated in Chapter V, this feature is useful when the figure is narrow or when the width of a figure is of no consequence.

Figure 4-5 displays the probability density function

---

[1]See Ball [Ref. 21, p. 38].

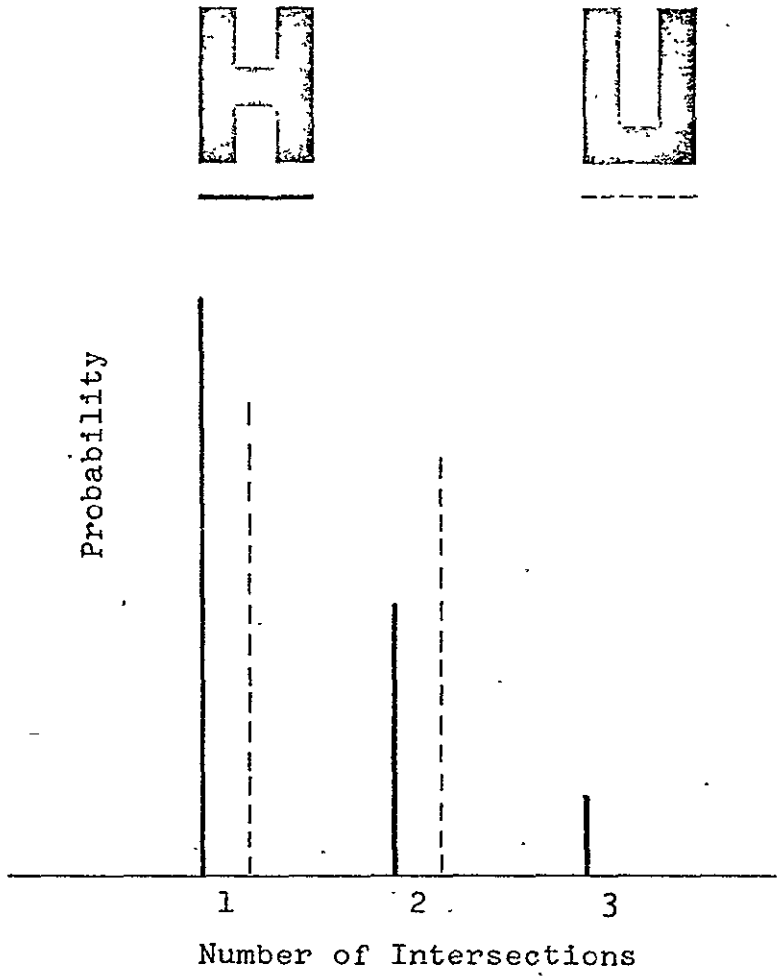

Fig. 4-5. Probability of Number of Intersections of Random Lines against Block H and U

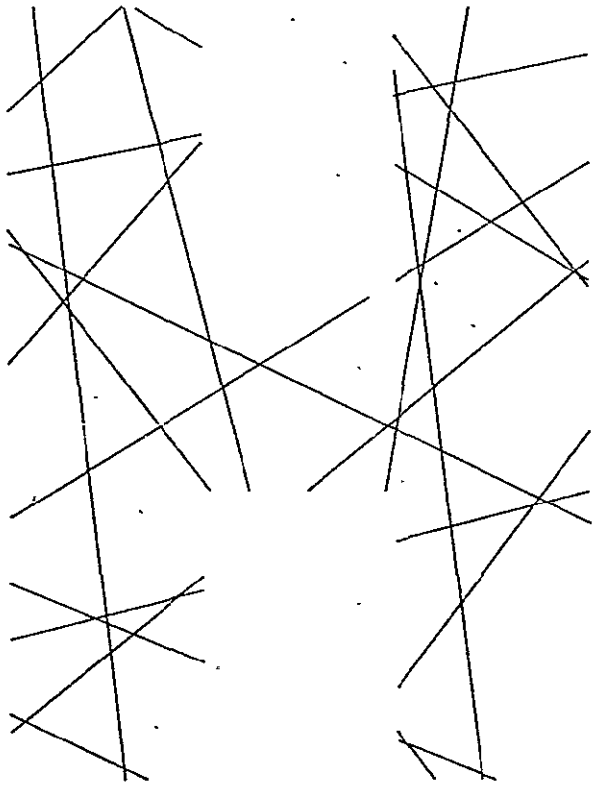

Fig. 4-6. Detail View of Random Lines Intersecting the Letter H

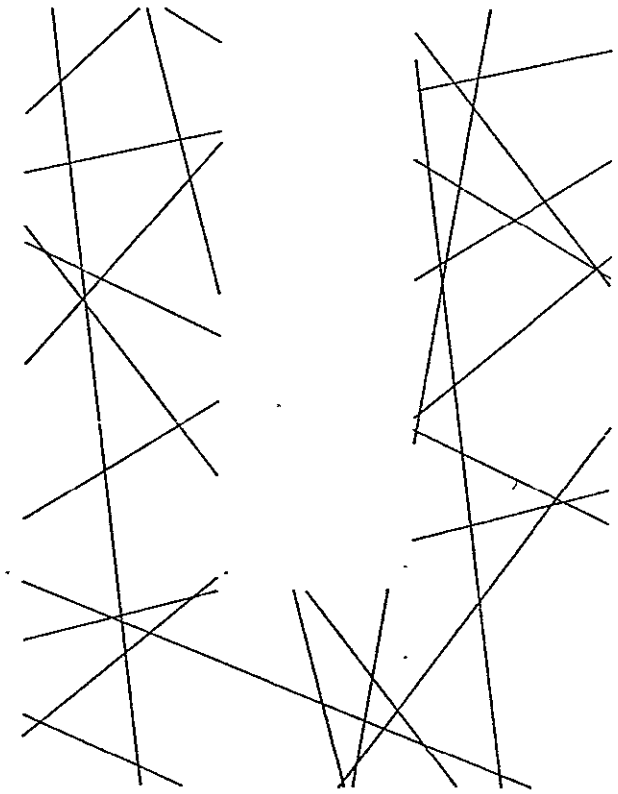

Fig. 4-7. Detail View of Random Lines Intersecting the Letter U

of N. Random lines are thrown against block letters.

The total length of intersection, random variable X, is also a useful quantity because this scalar number is easy to extract. A movable spot scanner can be used, for instance. Figure 4-6 shows how random lines may intersect a block H and Figure 4-7 illustrates random lines intersecting a block U. The outlines of the block letters were omitted to stress the point that not many lines are needed for a person to decide what the pattern is. X seems to be promising as an input to the decision algorithm of Chapter III. It will be shown in the next chapter that indeed quick decisions can be obtained by using X as an input.

Z is the joint random variable. Its two components are X and N. One may need to use Z when either N or X alone produces unsatisfactory results due to noise, font, or style changes.

F. NOISE, SIZE, FONT, AND STYLE

A few factors which affect the random variables N and X are noise, size, font, and style. Noise refers to smudges, distortions, or breaks in the pattern due to the printing, the paper, or the photographic process. Also the texture of the background is considered noise. Font refers to the various printing faces. There are so many fonts that even the best reading machines available

today can handle only a small percentage of them. The problems involved in reading handwritten material are obvious.

Noise affects N more than X. This is due to the fact that X is an "integral" of the overlap. A small extraneous blob affects X only slightly. On the other hand, whenever a random line intersects such a blob, N is made to differ by one, which is a significant change. In most character patterns N is most likely to be less than four.

Size does not affect the conditional probability density function of N, for N not equal to zero. Size changes X proportionally. However changes in size can be dealt with if those changes occur "slowly" or "infrequently" by putting X through an automatic gain control.

How font changes affect X and N is a question that can be answered experimentally. The variations in the fonts are subtle and cannot be handled analytically.

All the questions associated with font and style are complex. Further experiments are needed to find cross-font invariant features. X and N seem to be good random features.

CHAPTER V

EXPERIMENTAL RESULTS

In Chapter III a multiclass hypothesis test which performs at the desired error rates is developed. It allows rejects to occur. Upon a reject, the test continues by adjoining an additional sample to the observation. It is shown that this test is Bayes. In the case where there are only two classes this test is the same as Wald's sequential probability ratio test.

Feature extraction is discussed in Chapter IV. Two random variable features X and N which are invariant to translation and rotation are found. They meet with the needs of the multiclass hypothesis test. $\{X_i\}$ are independent and may be obtained quickly. The same comments hold true for $\{N_i\}$. For a given figure in a retina, the sequence of $\{N_i\}$ or $\{X_i\}$ is virtually limitless.

Two experiments are described in this chapter. The first experiment has to do with block letters, and the random variable X, the total length of a random line intersection. In the second experiment, handwritten numerals are classified using N, the number of intersections a random variable makes with the numerals.

A. CLASSIFICATION OF BLOCK LETTERS

Block letters are used in this experiment. Their shapes are illustrated in Figures 4-6 and 4-7. Two similar letters were chosen. These two letters have equal areas and the same convex hull areas. They differ in 22% of the area. Certainly if such letters can be classified, there is hope for more differing letters.

Random observations X are made. X is the total length of intersection that a uniformly distributed random line makes with a block letter. The random lines are taken one at a time independent of each other. Appendix 4 describes the theory of choosing uniform independent random lines. Hence, $\{X_i\}$ are clearly independent.

In the proposed test the probability distribution functions of X, given each letter, are prerequisites. Hence the first step is to learn these distributions. This is done empirically because the mathematics available today (such as geometric probability) allow the direct computation of only a few of the simple moments of the random variables.

Experimentally it is noticed that the p.d.f. changes hardly at all after 5,000 samples are tabulated. The p.d.f.'s used in this experiment are estimated by 50,000 samples of X.

Figure 5-1 illustrates the empirical p.d.f. The peaks at 50, 100, 150, and 200 units correspond to the dimensions of the block letter H. The peak at 140 for the letter U is due to the horizontal area. For the letter U it is on the bottom and for the letter H it is in the middle of the letter. The p.d.f.'s at zero are omitted from the diagram because they are the same for all convex hulls of the same perimeter. In fact, the p.d.f.'s are proportional to the ratio of the convex hull perimeter and the perimeter of the retina. See Kendall and Moran [Ref. 20, p. 58].

The difference between the two conditional random variables is more apparent in Figure 5-2 where the cumulative distribution functions are displayed.

The average number of samples needed to come to a decision is a function of the error probability which one desires (Chapter III). Figure 5-3 displays the average sample number as a function of the significance of the test. Figure 5-4 shows four decades of this relationship. The errors,

$$e_{12} = e_{21} \tag{5.1}$$

are held constant with respect to each other.

The samples $X = (x_1, x_2, \ldots, x_n)$ are independent. This makes the computation of $P_i d\tilde{F}_i(x)$ extremely simple.

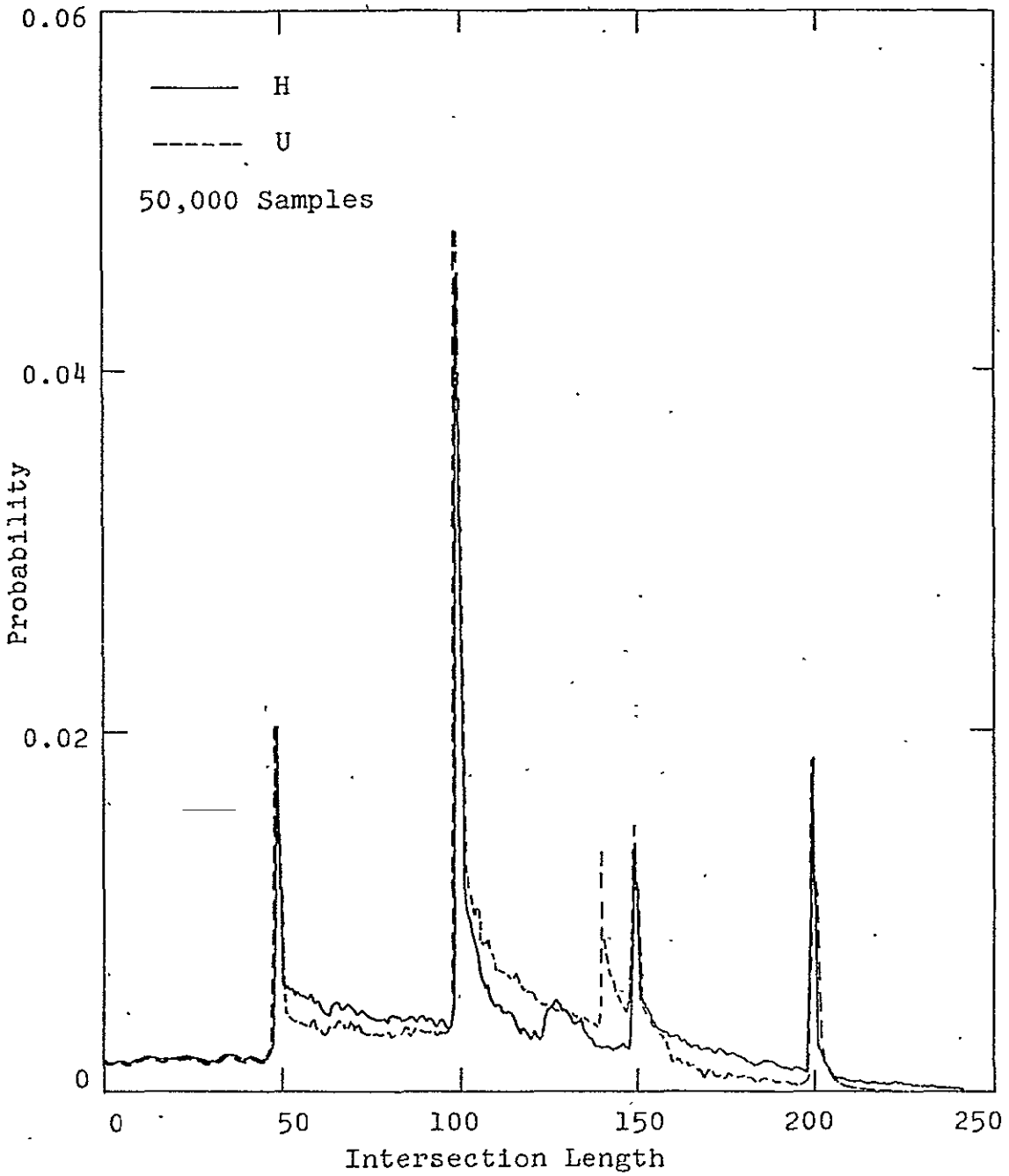

Fig. 5-1. Probability Density Functions of the Overlap Length Given an Intersection with Block Letters H and U

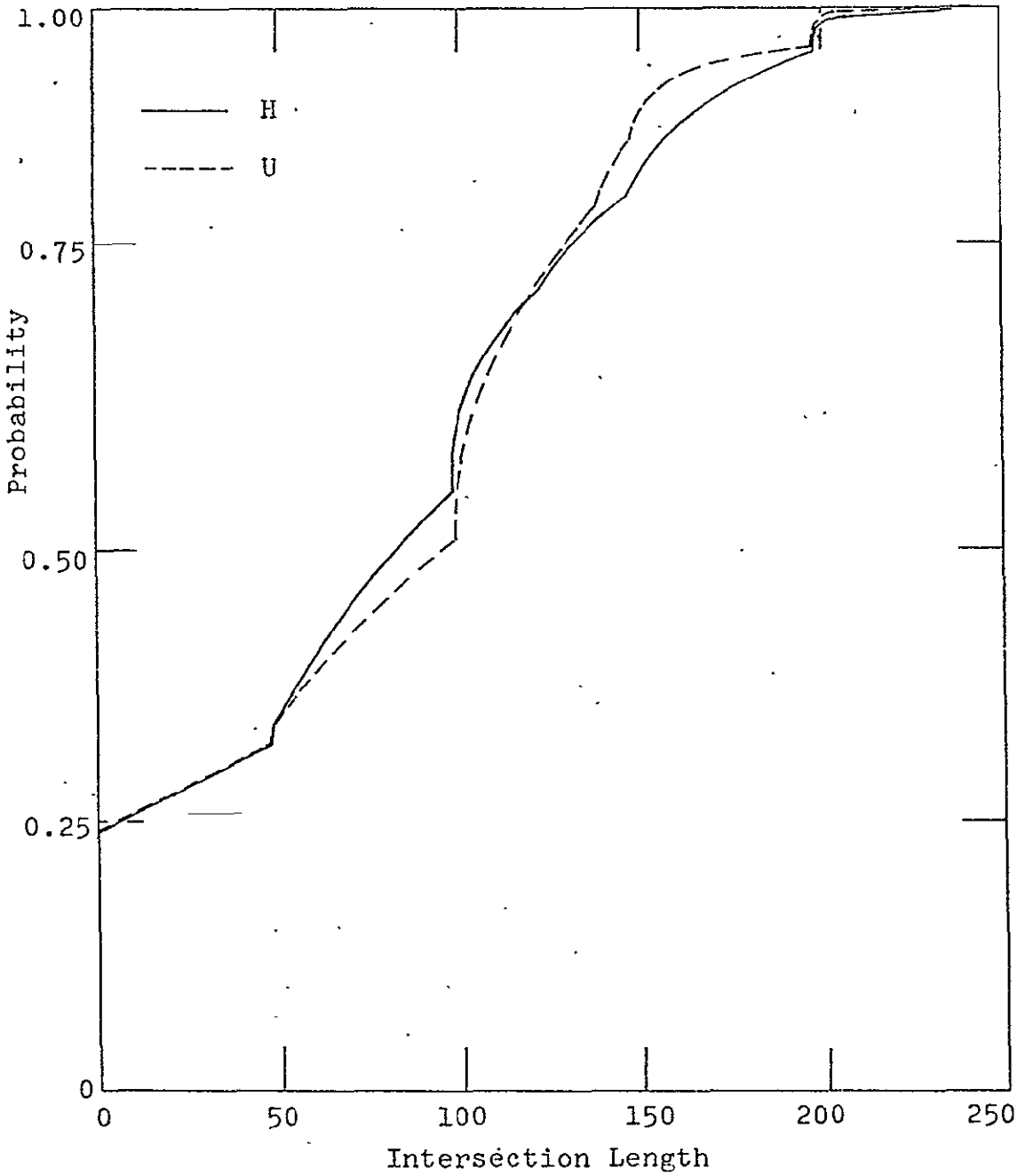

Fig. 5-2. Cumulative Distribution Functions of the Overlap Length Given an Intersection with Block Letters H and U

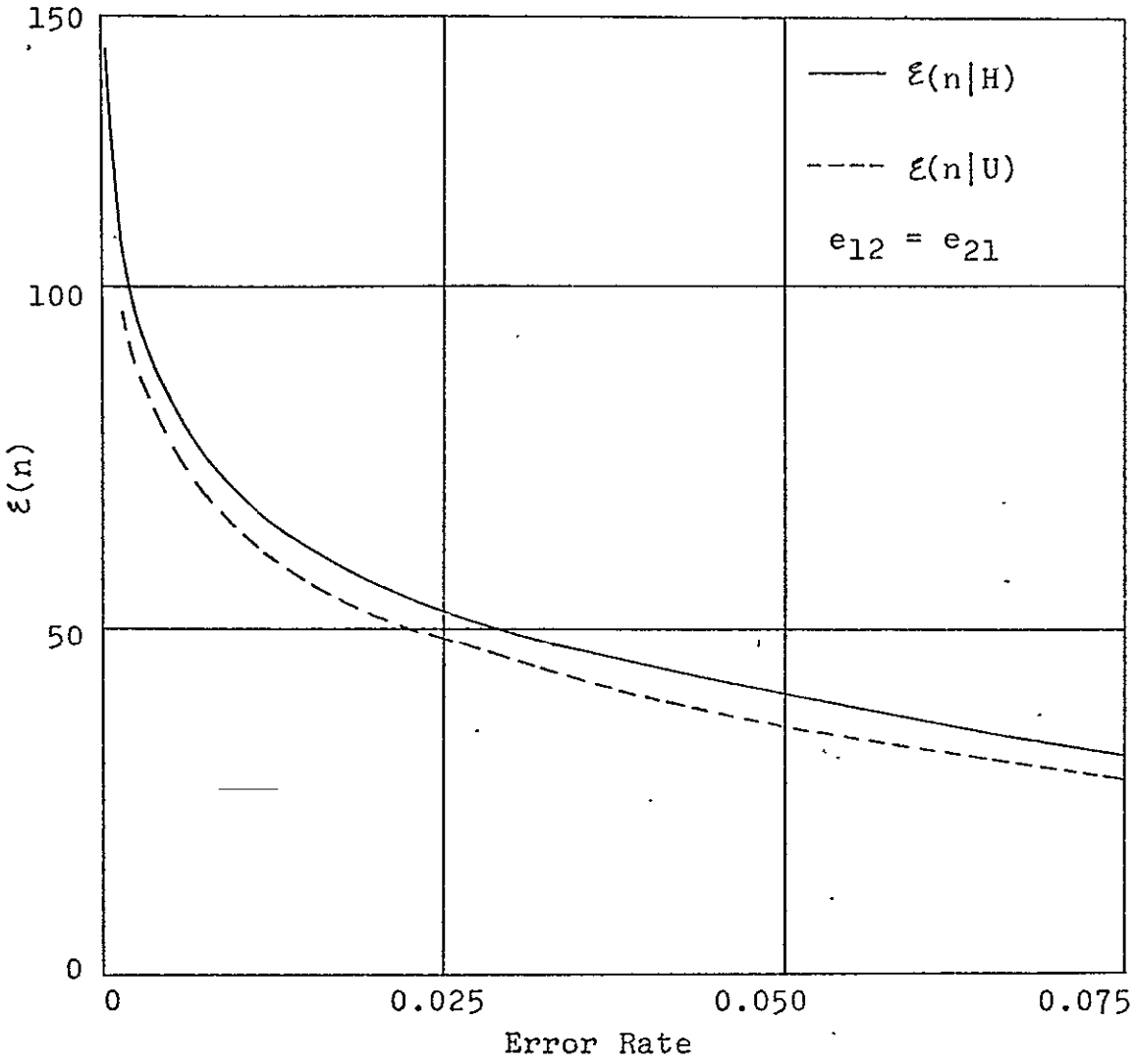

Fig. 5-3. Average Sample Number at Termination for the Block Letters H and U

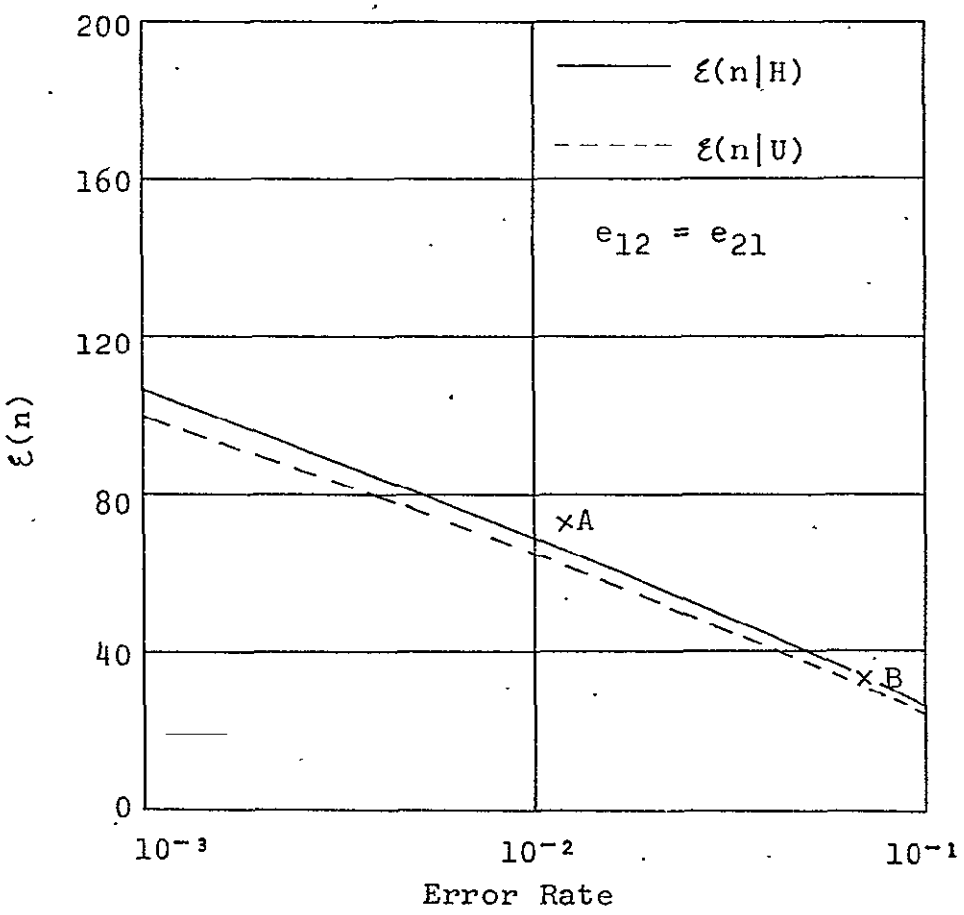

Fig. 5-4. Average Sample Number at Termination for Block Letters H and U with Experimental Points

When n samples are used,

$$P_i d\tilde{F}_i(x) = P_i \prod_{j=1}^{n} dF_i(x_j) \tag{5.2}$$

The evaluation of this function after another observation becomes

$$P_i d\tilde{F}_i(x) = P_i \prod_{j=1}^{n+1} dF_i(x_j) \tag{5.3}$$

$$= dF_i(x_{n+1}) P_i \prod_{j=1}^{n} dF_i(x_j) \tag{5.4}$$

The probability ratios are tested against thresholds as indicated by Equations 3.38 to 3.40.

$$A_2 = C_2 < \frac{P_1 d\tilde{F}_1(x)}{P_2 d\tilde{F}_2(x)} < \frac{1}{C_1} = A_1 \tag{5.5}$$

If the boundaries are exceeded a decision is made, whereas if neither boundary is crossed, further samples are taken.

In the following tests the classes are equally likely.

$$P_1 = 1/2 \tag{5.6}$$

$$P_2 = 1/2 \tag{5.7}$$

In these experiments, no rejects are allowed after

200 samples. They are then classified using the maximum likelihood test. This is done to limit the computing time.

Two runs of experiments are made. In the first run the algorithm is requested to make a 1% test, $e_{21} = e_{12} = 0.01$. Point A of Figure 5-4 is obtained from 7600 tests. The average number of samples used for a test is 76. The point is shifted to the right from the predicted position. This is due, in part, to the arbitrary truncation of the test at 200 samples.

In the second run the algorithm is requested to make a 10% test. Its actual performance is Point B of Figure 5-4. It requires 34 samples on the average. It uses more samples than predicted, but the decision is better.

It is interesting to observe the behavior of the likelihood ratio. It is a random walk biased upwards by

$$\varepsilon \frac{P_1 d\tilde{F}_1}{P_2 d\tilde{F}_2} \tag{5.8}$$

The logarithm of the ratio is displayed in Figure 5-5 along with the logarithm of the limits A_1 and A_2. Four tests are detailed, step by step. Tests 1 and 2 terminate well below the expected average sample number and Tests 3 and 4 terminate above it.

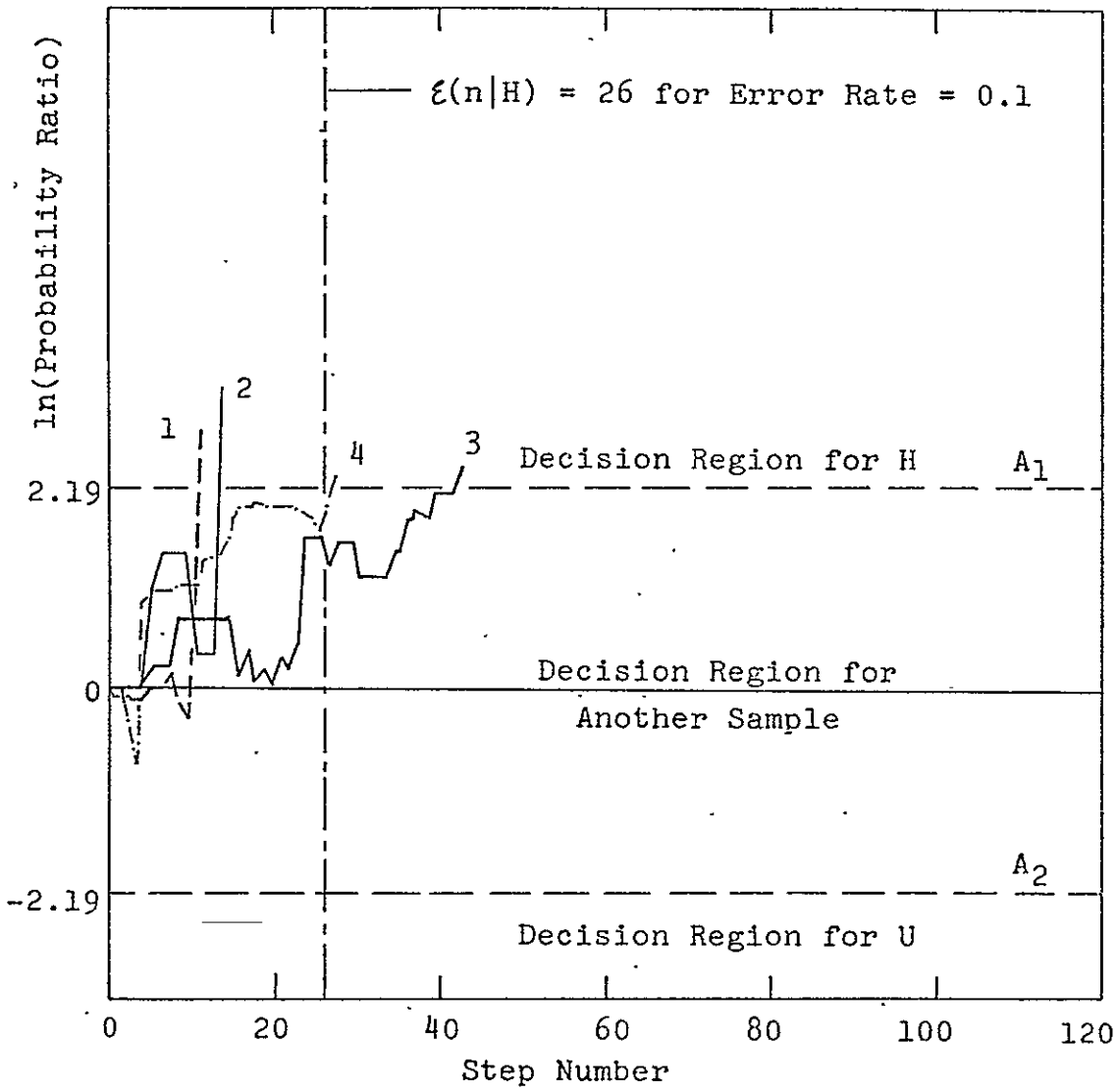

Fig. 5-5. Random Walk of the Logarithm of the Probability Ratio

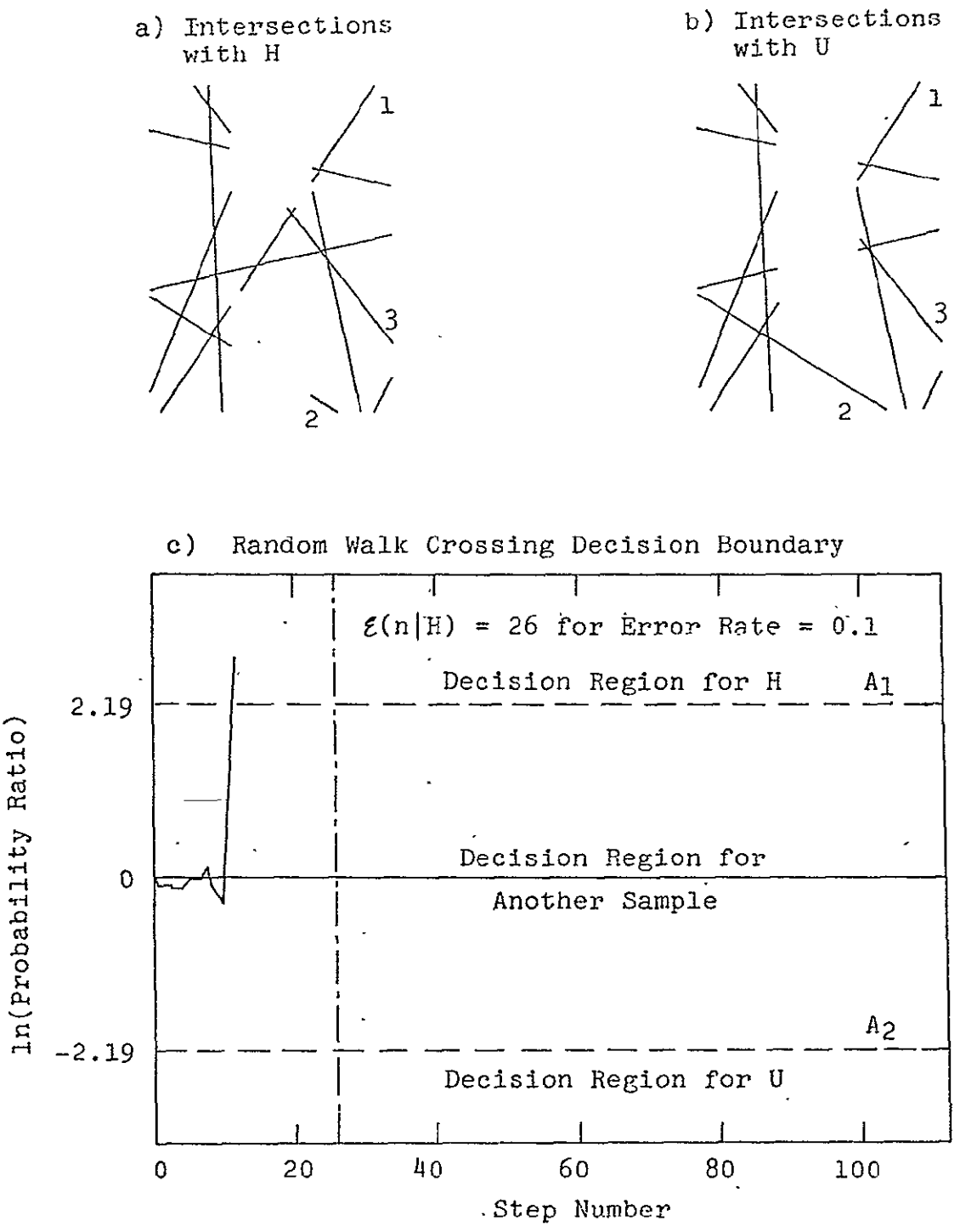

Fig. 5-6. Detail of One Classification

Figure 5-6 displays the actual lines which are used in Test 2 of Figure 5-5. Lines 1 and 2 of Figure 5-6, Parts a and b are different for each letter, whereas Line 3 is the same for each figure. Lines 1 and 2 give information about H and U but Line 3 does not.

B. HANDWRITTEN NUMERALS

Four digits illustrated in Figure 5-7 are used to learn the probability density functions of N, the number of times a random line intersects a given figure. These p.d.f.'s are learned by using many random lines. Figure 5-8 illustrates their nature. Figure 5-9 shows the cumulative distribution functions for the numbers 2, 3, 4, and 5. They are superimposed on one drawing so that the differences will be apparent. For these experiments 20,000 lines are used to estimate each p.d.f.

The Prob{N=0} is used to normalize the p.d.f. This is the same as normalizing by size since the probability of intersection is directly proportional to the convex hull of the character.

Using the formulas which are developed in Chapter III the average number of samples needed can be computed for any given error rate. Figure 5-9 displays such a relationship for $\alpha_i = \alpha_j$.

Tests are run for Points A and B of Figure 5-10. Point A is due to 782 tests for the class. The $\{\alpha_i\}$

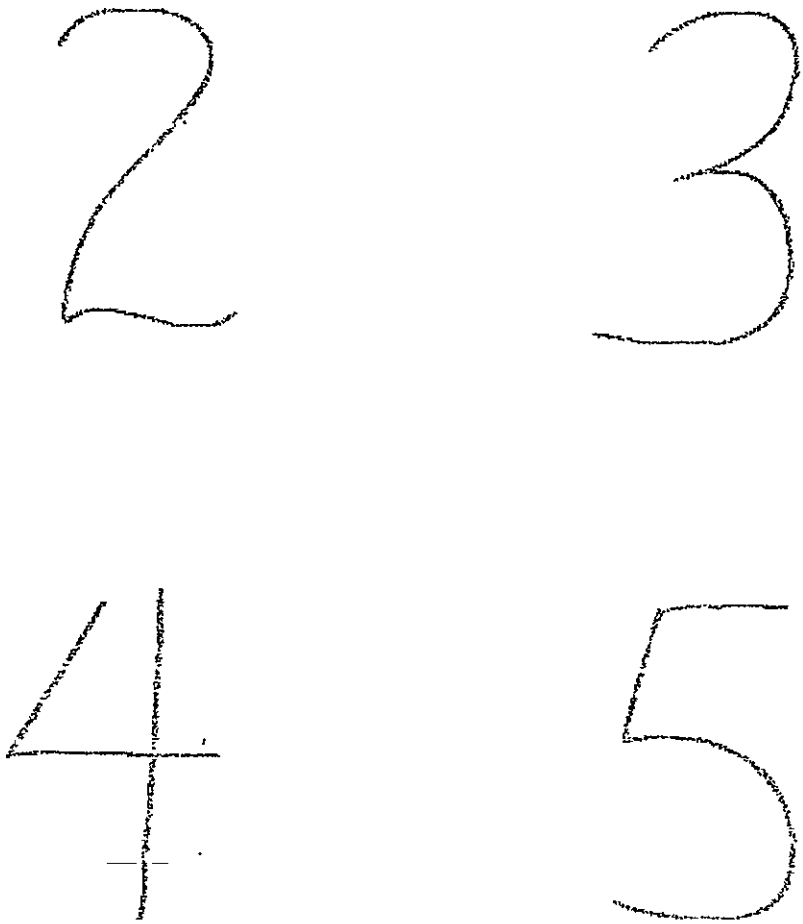

Fig. 5-7. Handwritten Numerals Used in the Experiment

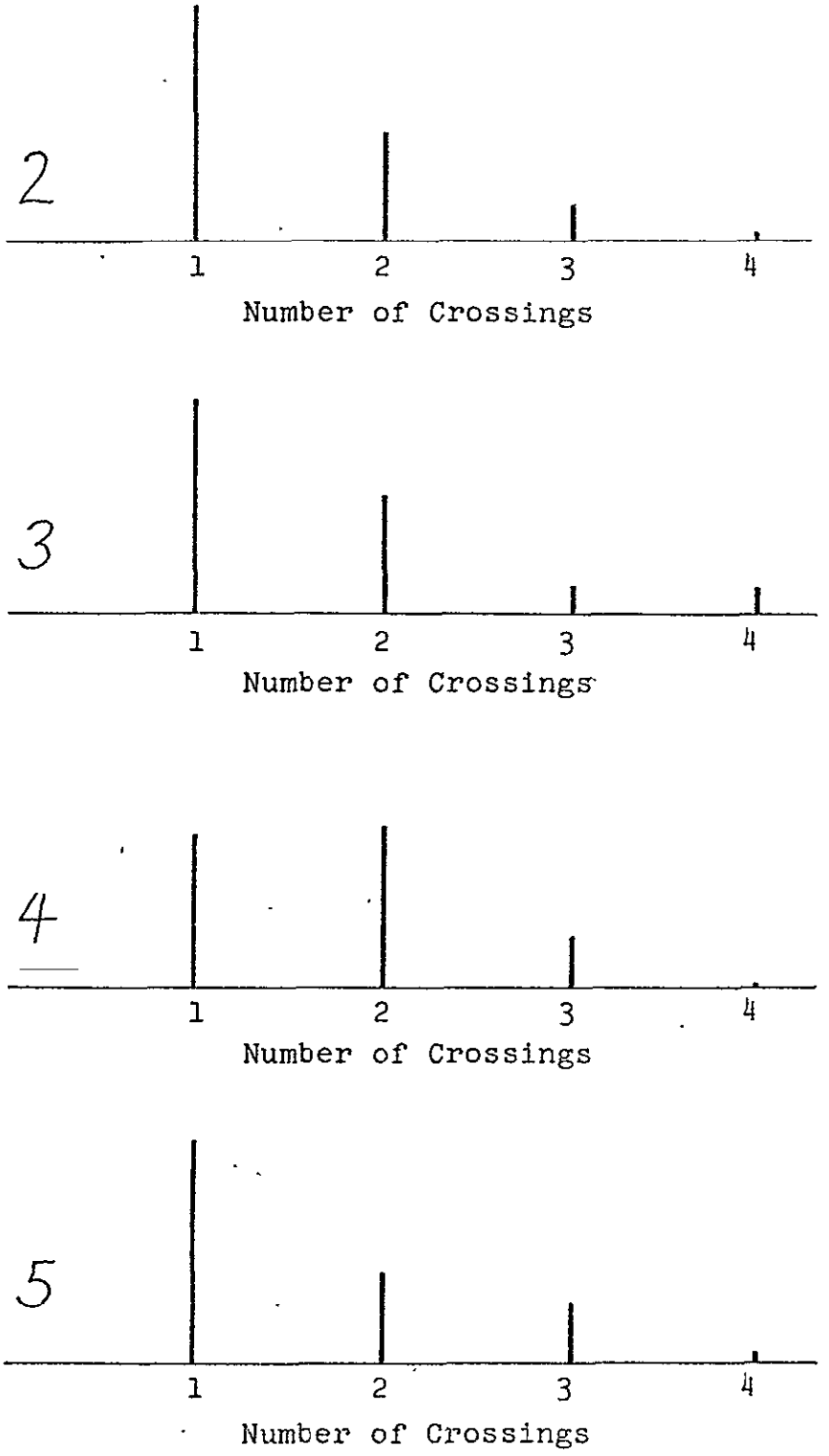

Fig. 5-8. Probability of n Crossings Given an Intersection for Each Number

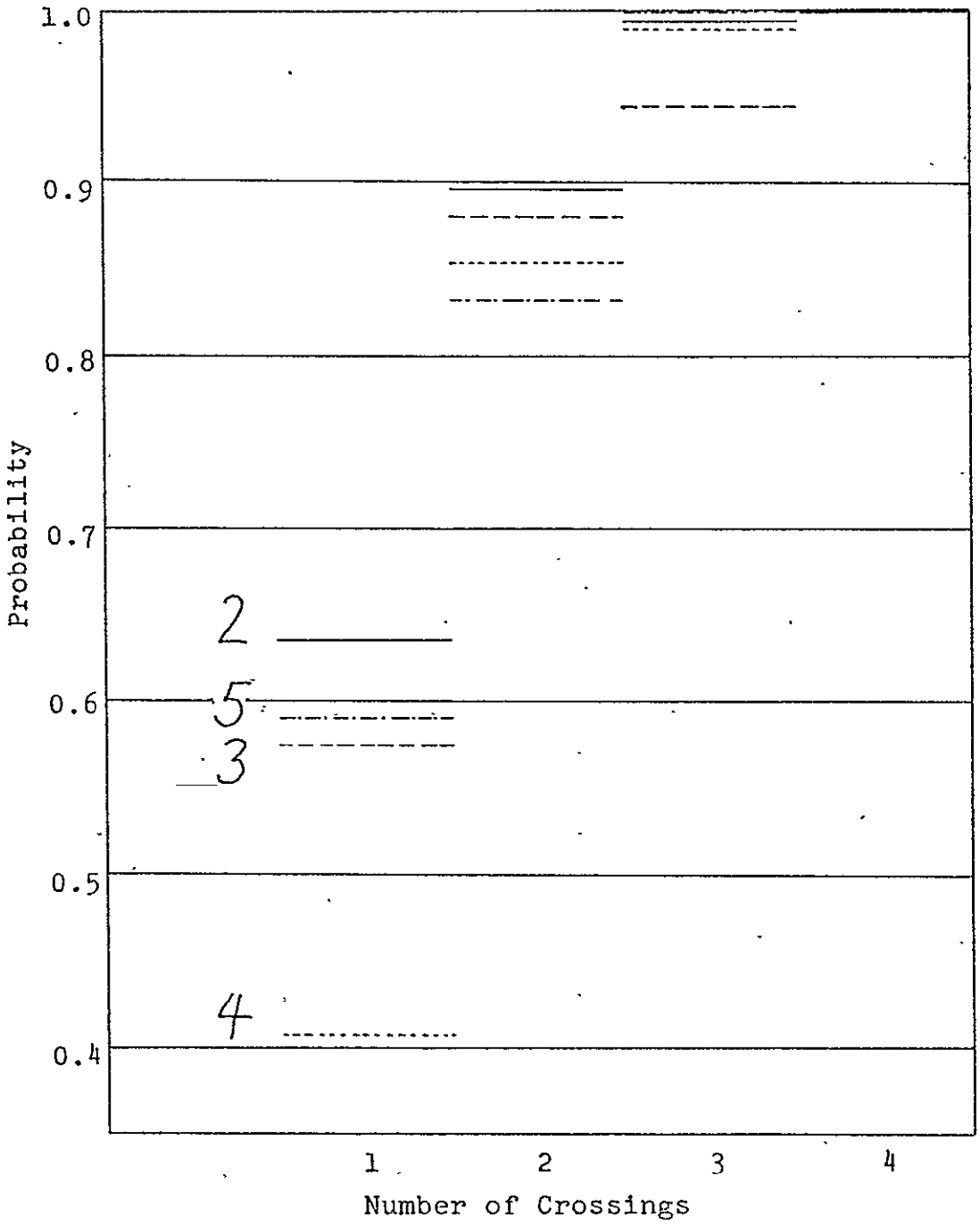

Fig. 5-9. Cumulative Distribution Functions of n Crossings Given an Intersection for Each Number

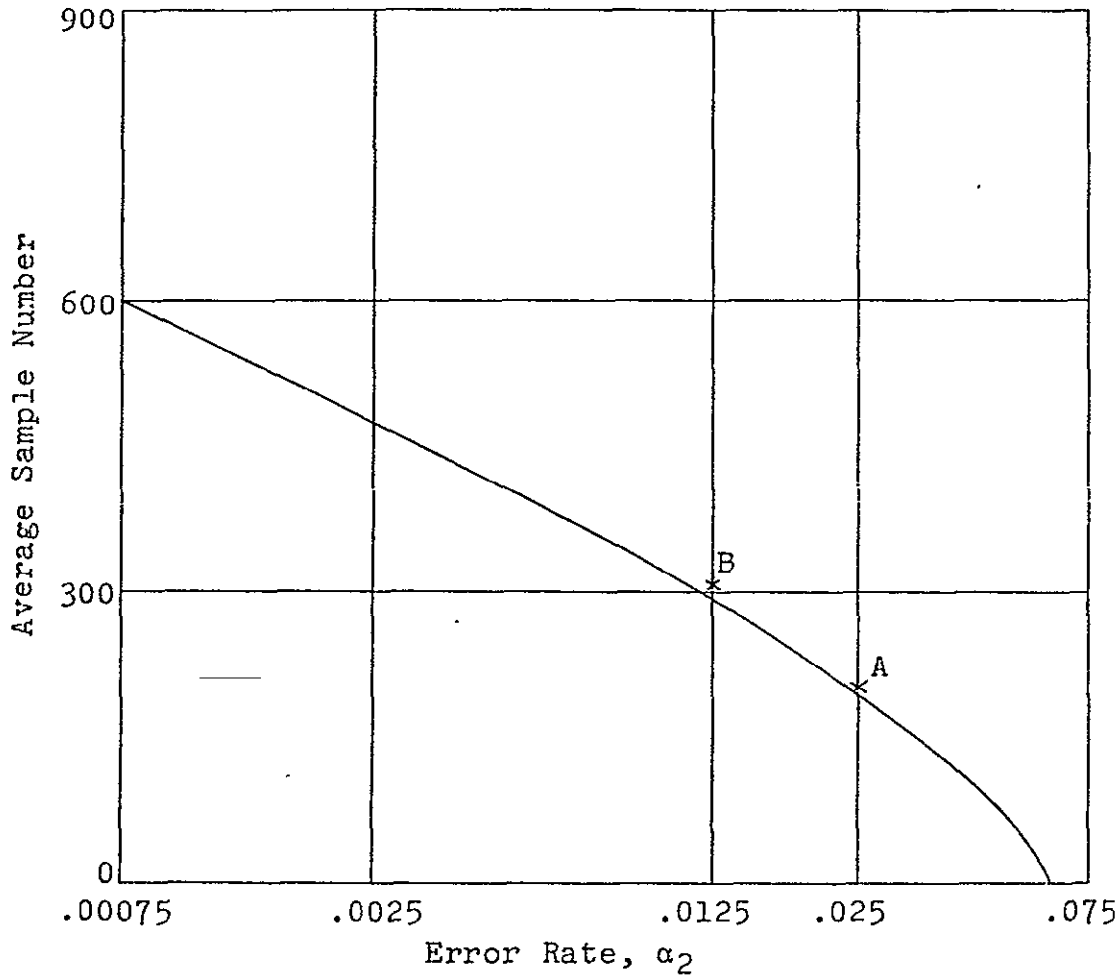

Fig. 5-10. Average Sample Number at Termination for Class 2 with Experimental Points

are specified at .025. Point B is due to 637 tests for that class with $\{\alpha_i\}$ specified at .0125. The four hypotheses are assumed equally likely. The results of the test are listed in Tables 5-3 to 5-6.

Recall from Equation 1.23 that

$$\alpha_i = \sum_{j \neq i} P_j e_{ij} \tag{5.9}$$

The error rates can be easily computed for each experiment as shown in Tables 5-1 and 5-2. These tables show that the performance of the proposed test can be controlled by the experimenter.

Tables 5-7 and 5-8 show the average number of samples taken for the various results, D_i given H_j. These numbers include the samples for which there is zero intersection. An estimate of the average sample number can be computed using Equation 3.52. The plot in Figure 5-10 reflects these estimates. The curve should be below the experimental points as a result of the estimates made in Equations 3.41 to 3.52.

The experimental results tabulated in Tables 5-3 and 5-5 show that 2 and 5 are similar. The majority of the errors made when 5 is true is the decision 2. This can be anticipated by observing the probability distribution function of N for these two letters, Figure 5-7, or by simply noting that 2 is quite like the upside-down

α_1	$\approx$	0.017	$<$	.025
α_2	$\approx$	0.001	$<$	.025
α_3	$\approx$	0.016	$<$	.025
α_4	$\approx$	0.027	$\approx$	.025

Table 5-1. Error Rates for the First Experiment

α_1	$\approx$	0.0114	$<$	.0125
α_2	$\approx$	0.0016	$<$	.0125
α_3	$\approx$	0.0095	$<$	.0125
α_4	$\approx$	0.0142	$\approx$	.0125

Table 5-2. Error Rates for the Second Experiment

	Given			
Decide	H_1	H_2	H_3	H_4
D_1	654	14	4	30
D_2	20	758	5	4
D_3	28	9	766	18
D_4	80	1	7	730

Table 5-3. Confusion Matrix $[e_{ij}]$; Experimental Data from 782 Tests for Each Class α_i Set Equal to .025 for All i=1,2,3,4

	Given			
Decide	H_1	H_2	H_3	H_4
D_1	834	18	5	38
D_2	26	970	6	5
D_3	37	11	980	23
D_4	103	1	9	934

Table 5-4. Confusion Matrix $[e_{ij}]$; Experimental Data Normalized to 1000

	Given			
Decide	H_1	H_2	H_3	H_4
D_1	585	10	1	16
D_2	3	623	1	0
D_3	12	4	637	9
D_4	37	0	0	612

Table 5-5. Confusion Matrix $[e_{ij}]$; Experimental Data from 637 Tests α_i Set Equal to .0125 for All i=1,2,3,4

Decide	H_1	H_2	H_3	H_4
D_1	928	16	2	25
D_2	5	978	2	0
D_3	19	6	996	15
D_4	58	0	0	960

Table 5-6. Confusion Matrix $[e_{ij}]$; Experimental Data Normalized to 1000

Decide	Given H_1	H_2	H_3	H_4
D_1	232	161	241	214
D_2	30	48	16	15
D_3	94	44	72	54
D_4	187	80	114	156

Table 5-7. Conditional Average Sample Number $\alpha_i = 0.025$ Including the Samples with Zero Intersection

Decide	Given H_1	H_2	H_3	H_4
D_1	314	192	233	323
D_2	55	67	21	0
D_3	88	42	88	99
D_4	290	0	0	214

Table 5-8. Conditional Average Sample Number $\alpha_i = 0.0125$ Including the Samples with Zero Intersection

image of 5. The result is the large average sample number at termination.

There are other pairs of numbers, such as 6 and 9, which this algorithm will have difficulty classifying. It is suggested that these similar shapes be classed into common families and other algorithms which are not invariant to rotation and mirror-imaging be used to subclassify within each family. Application of the author's algorithm first may simplify the subclassification process.

CHAPTER VI

CONCLUSION

A. RESULTS

The fundamental characteristics of a classification algorithm are identified as the false declaration rate and the probability of detection. A sequential multiclass hypothesis test is proposed in Chapter III. A detailed study of the test shows that it terminates almost surely, and its performance can be readily controlled.

The sequential machine becomes a threshold tester of certain functions of probability densities. The threshold levels which determine the operating characteristics of the pattern recognition machine are under the control of the experimenter.

The input to this machine must necessarily be random quantities. In Chapter V invariant feature extraction is developed. A few features extracted by random lines are presented. They are used in Chapter V in experiments that simulate pattern recognition machines. Results of recognizing block letters and handwritten numerals are presented.

B. APPLICATIONS AND FURTHER RESEARCH

It is anticipated that the new ideas and results presented here will be the forerunner of broader research and development in applications of multiclass hypothesis testing to pattern recognition using randomized features.

In particular the results of the experiment with handwritten numerals indicate that inexpensive and fast sorters can be built. Also the preliminary findings indicate that the features used in the experiments may be insensitive to changes in style. The applications of such a system in the postal service or in business may relieve much of the burden of hand sorting.

This method could be fruitfully applied in understanding and solving the problems of machine reading multifont and handwritten (script) matter.

Multiclass sequential tests that use random features may fit well with techniques which use context information. A reader that correlates at the word level, for instance, does not need complete accuracy on each individual letter. If a word like "California" is anticipated, even a ten per cent error rate on each individual letter can give highly accurate results.

GLOSSARY

A_i	threshold used in Wald's sequential probability ratio test; also used by Armitage
α_i	false alarm probability
β_i	probability of incorrectly deciding D_i $(P_i - \gamma_i)$
C_i	thresholds used in the proposed test
$\delta(D_i \mid v)$	probability of verifying H_i when the observation is v
D_0	decision to reject
D_i	decision i
d_i	discriminant function for the ith class
$\mathcal{E}$	expectation
e_{ij}	probability of deciding D_i given H_j is true; entries to the confusion matrix
F_i	probability distribution function of v_1 given H_i
$\tilde{F}_i$	probability distribution function of $v = (v_1, v_2, \ldots, v_n)$ given H_i
$d\tilde{F}_i$	probability density function, if it exists, of the observation v given H_i
Γ	observation space $v \in \Gamma$
γ_i	probability of correctly deciding D_i (detection probability)
H_i	hypothesis that the observation v is from the probability distribution $\tilde{F}_i$
i,j,k	often used as dummy index

L_{ij}	loss incurred by deciding D_i when H_j is true
M	number of hypotheses possible $H_1, H_2, \ldots, H_M$
P_i	probability of H_i being true
ψ	space of observation $v \in \psi$
ψ_i	space of v in which decision D_i is made $v \in \psi_i \Rightarrow D_i$
R_{ij}	probability ratio used by Armitage
r	average loss, risk
$s(P,\theta)$	random line specified by P and θ
t	$t \in [0,1]$, parameter used by Chow to control the rejection rate
U_i	ratio used in Reed's test
v	observation, sample $v = (v_1, v_2, \ldots, v_n)$
$w^{(i)}$	a vector used to weigh $v = (v_1, v_2, \ldots, v_n)$ for the ith class
x	$x = (x_1, x_2, \ldots, x_n)$ an observation
$Z_n(i,j)$	product of probability ratios $\prod_{m=1}^{n} z_m(i,j)$
$\tilde{Z}_n(i,j)$	logarithm of $Z_n(i,j)$
$z_m(i,j)$	ratio of probabilities $dF_j(v_m)/dF_i(v_m)$
$\tilde{z}_m(i,j)$	logarithm of $z_m(i,j)$

REFERENCES

1. G. Nagy, "State of the Art in Pattern Recognition," Proceedings IEEE, Vol. 56, No. 5, pp. 836-862, May 1968.

2. Pattern Recognition, Vol. 2, No. 3, September 1970.

3. W. F. Miller and A. C. Shaw, "Linguistic Methods in Picture Processing - A Survey," Proceedings of the Fall Joint Computer Conference, pp. 279-290, 1968.

4. R. H. Anderson, "Syntax-Directed Recognition of Hand-Printed Two-Dimensional Mathematics," Interactive Systems for Experimental Applied Mathematics. New York: Academic Press, Inc., 1968.

5. N. J. Nilsson, Learning Machines. New York: McGraw-Hill, 1965.

6. H. C. Andrews, "Fourier Coding Images," University of Southern California Electrical Engineering Report, No. 271, pp. 55-59, June 1968.

7. T. Kailath, "The Divergence and Bhatlacharyya Distance Measures in Signal Detection," IEEE Transactions on Communications Technology, Vol. 15, No. 1, pp. 52-60, 1967.

8. W. S. Meisel, "Notes on Mathematical Pattern Recognition," Computer-Oriented Techniques in Pattern Recognition, University of Southern California Summer Seminar, July 1969.

9. H. L. Van Trees, Detection, Estimation, and Modulation Theory. New York: John Wiley and Sons, Inc., 1968.

10. C. K. Chow, "An Optimum Character Recognition System Using Decision Functions," IRE Transactions of Electronic Computers, Vol. EC-6, pp. 247-254, December 1957.

11. C. K. Chow, "On Optimum Recognition Error and Reject Tradeoff," IEEE Transactions on Information Theory, Vol. IT-16, No. 1, pp. 41-46, June 1969.

12. A. Wald, Sequential Analysis. New York: John Wiley and Sons, Inc., 1947.

13. S. S. Wilks, Mathematical Statistics. New York: John Wiley and Sons, Inc., 1962.

14. I. Selin, Detection Theory. Princeton, New Jersey: Princeton University Press, 1965.

15. M. Sobel and A. Wald, "A Sequential Decision Procedure for Choosing One of Three Hypotheses concerning the Unknown Mean of a Normal Distribution," Annals of Mathematical Statistics, Vol. 20, pp. 502-522, November 1949.

16. G. A. Barnard, "Review of 'Sequential Analysis' by A. Wald," Journal of the American Statistical Association, Vol. 42, p. 658, 1947.

17. P. Armitage, "Sequential Analysis with More than Two Alternative Hypotheses and its Relation to Discriminant Function Analysis," Journal of the Royal Statistical Society, Series B, Vol. 12, pp. 137-144, 1950.

18. F. C. Reed, "A Sequential Multi-Decision Procedure," Proceedings of the Symposium on Decision Theory and Applications to Electronic Equipment Development, USAF Development Center, Rome, New York, April 1960.

19. K. S. Fu, Sequential Methods in Pattern Recognition and Machine Learning. New York: Academic Press, Inc., 1968.

20. M. G. Kendall and P. A. P. Moran, Geometric Probability. London: Charles Griffin and Company Limited, 1963.

21. G. Ball, "An Invariant Input for a Pattern Recognition Machine," Ph.D. Dissertation, Stanford University, 1962.

22. D. Rubinstein, "Pattern Recognition by Random Lines," AIEE Conference, Paper No. CP, pp. 62-64, 1961.

23. E. Wong and A. Steppe, "Invariant Recognition of Geometric Shapes," Methodologies of Pattern Recognition, Edited by S. Watanabe. London: Academic Press, 1969.

24. A. E. Brain and P. E. Hart, "Graphical Data Processing Research Study and Experimental Investigation," Technical Report, No. ECOM-01901-25, Stanford Research Institute, Menlo Park, California, December 1966.

25. K. Mori, et al., "Micro-Programming Controlled Pattern Processing in a Handwritten Mail Reader-Sorter," Pattern Recognition, Vol. 2, No. 3, pp. 175-186, September 1970.

26. I. Sheinberg, "The INPUT 2 Document Reader," Pattern Recognition, Vol. 2, No. 3, pp. 167-173, September 1970.

27. R. B. Hennis, et al., "IBM 1975 Optical Page Reader," IBM Journal of Research Development, Vol 12, No. 5, pp. 346-371, September 1968.

28. J. L. Doob, Stochastic Processes. New York: John Wiley and Sons, Inc., 1953.

29. L. Breiman, Probability. Reading, Massachusetts: Addison-Wesley Publishing Company, 1968.

30. T. S. Ferguson, Mathematical Statistics: A Decision Theoretic Approach. New York: Academic Press, Inc., 1967.

31. G. Birkhoff and S. MacLane, A Survey of Modern Algebra. New York: The MacMillan Company, 1953.

32. R. Ash, Information Theory. New York: John Wiley and Sons, Inc., 1965.

APPENDIX 1

THE EQUIVALENCE OF THE TWO FORMS OF THE PROPOSED TEST

The two forms of the test in Chapter III are really the same test. For any $i = 1,2,\ldots,M$

$$\sum_{j\neq i} P_j d\tilde{F}_j(v) \geq C_i P_i d\tilde{F}_i(v) \tag{1}$$

if and only if

$$\sum_{j=1}^{M} P_j d\tilde{F}_j(v) \geq (C_i+1) P_i d\tilde{F}_i(v) \tag{2}$$

The statement of Equation 1 is the same as

$$\sum_{j=1}^{M} P_j d\tilde{F}_j(v) \geq \max_i \{(C_i+1) P_i d\tilde{F}_i(v)\} \tag{3}$$

because $\{C_i\}$, $\{P_i\}$, $\{dF_i(v)\}$ are all positive numbers. Now consider

$$\min_i \left[\sum_{j\neq i} P_j d\tilde{F}_j(v) - C_i P_i d\tilde{F}_i(v) \right] \tag{4}$$

under the condition of Equation 1. Then for any $k = 1,2,\ldots,M$

$$\sum_{j\neq i} P_j d\tilde{F}_j(v) - C_i P_i d\tilde{F}_i(v)$$

$$\leq \sum_{j\neq k} P_j d\tilde{F}_j(v) - C_k P_k d\tilde{F}_k(v) \qquad (5)$$

Now cancelling equal terms and rearranging the terms gives

$$-(C_i+1)P_i d\tilde{F}_i(v) \leq -(C_k+1)P_k d\tilde{F}_k(v) \qquad (6)$$

or

$$(C_i+1)P_i d\tilde{F}_i(v) = \max_j\{(C_j+1)P_j d\tilde{F}_j(v)\} \qquad (7)$$

Hence, the two forms of the test differ only in the computations which are performed.

APPENDIX 2

THE CHERNOFF BOUND

The Chernoff bound uses the fact that an exponential curve bounds a step function.

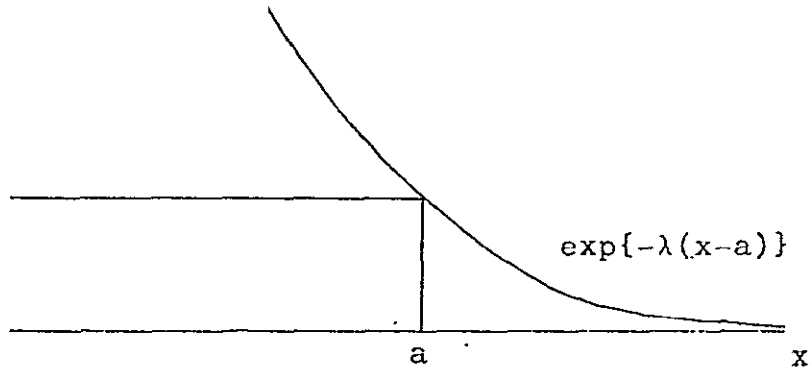

Fig. A-1. Chernoff Bound

$$\frac{1 - \text{sign}(x-a)}{2} \leq \exp\{-\lambda(x-a)\}, \quad \lambda \geq 0 \tag{1}$$

$$\text{Prob}\{x<a\} = \int \frac{1 - \text{sign}(x-a)}{2} dF(x) \tag{2}$$

$$\leq \int \exp\{-\lambda(x-a)\} dF(x) \tag{3}$$

$$= \mathcal{E}\{\exp(-\lambda(x-a))\} \tag{4}$$

An error occurs if decision D_k is made when H_i is true, $i \neq k$. Then

$$\sum_{j\neq k} P_j d\tilde{F}_j(v) < C_k P_k d\tilde{F}_k(v), \quad k\neq i \tag{5}$$

and in particular

$$P_i d\tilde{F}_i(v) < C_k P_k d\tilde{F}_k(v), \quad k\neq i \tag{6}$$

or

$$\frac{d\tilde{F}_i(v)}{d\tilde{F}_k(v)} < \frac{C_k P_k}{P_i} \tag{7}$$

Hence the probability of error is bounded by

$$\text{Prob}\{D_k|H_i \text{ true}\} = \text{Prob}\left(\sum_{j\neq k} P_j d\tilde{F}_j < C_k P_k d\tilde{F}_k(v)\right) \tag{8}$$

$$\leq \text{Prob}\{P_i d\tilde{F}_i(v) < C_k P_k d\tilde{F}_k(v)\} \tag{9}$$

$$= \text{Prob}\left(\frac{d\tilde{F}_i(v)}{d\tilde{F}_k(v)} < \frac{C_k P_k}{P_i}\right) \tag{10}$$

$$= \text{Prob}\left(\ln \frac{d\tilde{F}_i(v)}{d\tilde{F}_k(v)} < \ln \frac{C_k P_k}{P_i}\right) \tag{11}$$

$$= \mathcal{E}_{H_i}\left(\exp\left[-\lambda \ln \frac{d\tilde{F}_i(v)}{d\tilde{F}_k(v)} + \lambda \ln \frac{C_k P_k}{P_i}\right]\right) \tag{12}$$

$$= \lambda \ln\left(\frac{C_k P_k}{P_i}\right) \mathcal{E}_{H_i}\left(\frac{d\tilde{F}_k(v)}{d\tilde{F}_i(v)}\right) \tag{13}$$

or where a is some constant,

$$\text{Prob}\{D_k | H_i \text{ true}\} \leq a \underset{H_i}{\mathcal{E}} \left(\frac{d\tilde{F}_k(v)}{d\tilde{F}_i(v)} \right) \tag{14}$$

But by the arguments of Chapter III, Section C,

$$\frac{d\tilde{F}_k(v)}{d\tilde{F}_i(v)} \Bigg|_{H_i \text{ true}} \longrightarrow 0 \quad \text{a.s.} \tag{15}$$

The conclusion is that as n becomes large, errors of the type e_{ki}, $k \neq i$, when H_i is true, approach 0.

$$\text{Prob}\{D_k | H_i \text{ true}\} \to 0 \tag{16}$$

APPENDIX 3

AVERAGE SAMPLE NUMBER

The computation of the expected number of samples before a test ends, for a two-class test, is well known. The theorem given here is after Theorem 9.1 of Selin [Ref. 4].

Two computational devices are used. The first is

$$\sum_{i=1}^{\infty} \sum_{j=1}^{i} \equiv \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} \tag{1}$$

The second is

$$\sum_{i=1}^{\infty} P(N \geq i) = \mathcal{E}(N) \tag{2}$$

where N is a positive valued integer random variable. This can easily be seen by writing out a few terms of the expectation.

$$\mathcal{E}(N) = P(N=0) \cdot 0 + P(N=1) \cdot 1 + P(N=2) \cdot 2 \ldots \tag{3}$$

$$= P(N=1) + P(N=2) + P(N=3) + \ldots$$

$$+ P(N=2) + P(N=3) + \ldots$$

$$+ P(N=3) + \ldots \tag{4}$$

Summing the terms row by row,

$$\mathcal{E}(N) = P(N \geq 1) + P(N \geq 2) + \ldots \tag{5}$$

$$= \sum_{i=1}^{\infty} P(N \geq i) \tag{6}$$

Theorem: If the test ends at the Nth observation and $\mathcal{E}(N) < \infty$, then

$$\mathcal{E}(N) = \frac{\mathcal{E}(\tilde{Z}_N(i,j))}{\mathcal{E}(\tilde{z}(i,j))} \tag{7}$$

where

$$\tilde{Z}_N(i,j) = \sum_{n=1}^{N} \ln \frac{dF_j(v_n)}{dF_i(v_n)} = \sum_{n=1}^{N} \tilde{z}_n(i,j) \tag{8}$$

Proof: Since N is itself a random variable,

$$\mathcal{E}(\tilde{Z}_N(i,j)) = \sum_{k=1}^{\infty} P(N=k)\mathcal{E}(\tilde{Z}_N(i,j)|N=k) \tag{9}$$

$$= \sum_{k=1}^{\infty} P(N=k) \sum_{m=1}^{k} \mathcal{E}(\tilde{z}_m(i,j)|N=k) \tag{10}$$

$$= \sum_{m=1}^{\infty} \sum_{k=m}^{\infty} P(N=k)\mathcal{E}(\tilde{z}_m(i,j)|N=k) \tag{11}$$

$$= \sum_{m=1}^{\infty} P(N \geq m)\mathcal{E}(\tilde{z}_m(i,j)|N \geq m) \tag{12}$$

The event $N \geq m$ can occur only if the test has not ended by the (m-1)th observation, and hence this event is independent of $\tilde{z}_m(i,j)$. Therefore,

$$\mathcal{E}(\tilde{Z}_N(i,j)) = \sum_{m=1}^{\infty} P(N \geq m)\mathcal{E}(\tilde{z}(i,j)) \tag{13}$$

$$= \mathcal{E}(\tilde{z}(i,j)) \sum_{m=1}^{\infty} \sum_{k=m}^{\infty} P(N=k) \tag{14}$$

$$= \mathcal{E}(\tilde{z}(i,j)) \sum_{k=1}^{\infty} \sum_{m=1}^{k} P(N=k) \tag{15}$$

$$= \mathcal{E}(\tilde{z}(i,j)) \sum_{k=1}^{\infty} kP(N=k) \tag{16}$$

$$= \mathcal{E}(\tilde{z}(i,j))\mathcal{E}(N) \tag{17}$$

APPENDIX 4

UNIFORM RANDOM LINES

A line can be named uniquely [Ref. 20, p. 14] by u and v and the equation,

$$ux + vy + 1 = 0 \qquad (1)$$

This representation excludes lines which pass through the origin. This is of little concern since the measure of all such lines is zero.

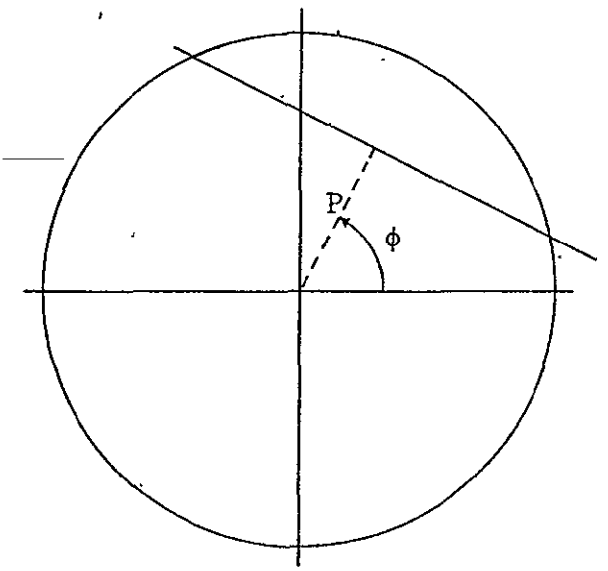

Fig. A-2. Definition of a Random Line

Another representation of a line may use the polar coordinates of the point on the line nearest the origin. The equation of the line is

$$x \frac{\cos \phi}{-P} + y \frac{\sin \phi}{-P} + 1 = 0 \qquad (2)$$

A rotation and translation (a,b) in the two dimensional plane can be represented by

$$X = a + x \cos \theta - y \sin \theta$$

$$Y = b + x \sin \theta + y \cos \theta \qquad (3)$$

where $0 \leq \theta < 2\pi$.

The parameters change and the new line is

$$x \frac{\cos \phi'}{P'} + y \frac{\sin \phi'}{P'} + 1 = 0 \qquad (4)$$

where

$$P' = P - \cos \phi - b \sin \phi$$

$$\phi' = \phi - \theta \qquad (5)$$

This can be shown by substituting Equation 3 into Equation 2 and collecting terms to obtain Equation 4.

Let a set of random lines be denoted by E. A transformation places these lines in E'. Random lines are uniformly distributed when P(E) = P(E'). It is shown

in Kendall and Moran [Ref. 20] that uniform random lines are possible if P and ϕ are each uniform.

A uniformly distributed random line may be generated by choosing P and ϕ uniformly. The line so chosen is Equation 2. These lines will intersect a circular retina of radius R. The circular retina is the area in which all observations are confined if $0 < P \leq R$ and $0 < \phi \leq 2\pi$.

NASA — JPL — Coml., L.A., Calif.