

Technical Report No. 168

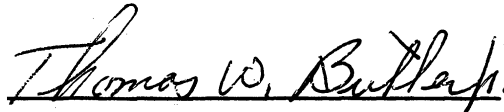
6137-14-T

A FINITE-MEMORY ADAPTIVE
PATTERN RECOGNIZER

by

K. B. Irani

Approved by:



T. W. Butler, Jr.

for

COOLEY ELECTRONICS LABORATORY

Department of Electrical Engineering
The University of Michigan
Ann Arbor, Michigan

Contract No. DA-36-039 AMC-03733(E)
Signal Corps, Department of the Army
Department of the Army Project No. 1PO21101A0420102

September 1965

TABLE OF CONTENTS

	<u>Page</u>
ACKNOWLEDGEMENTS	iv
LIST OF ILLUSTRATIONS	v
SUMMARY	vi
1. INTRODUCTION	1
2. BACKGROUND	3
3. THE MATHEMATICAL STRUCTURE	5
4. THEOREMS AND USEFUL RESULTS	10
5. DESCRIPTION OF THE PROBLEM	25
6. RESULT OF THE INVESTIGATION	28
6.1 The Adaptive Procedure and the Proof of the Theorem	29
7. ILLUSTRATIVE EXAMPLE	36
8. CONCLUSIONS	43
REFERENCES	44
DISTRIBUTION LIST	45

ACKNOWLEDGEMENTS

The author wishes to acknowledge the valuable assistance he received from several co-workers at the Cooley Electronics Laboratory.

He is indebted to Professor A. W. Naylor for many useful suggestions including the application of one-dimensional example to signal detection problem.

Professor F. M. Waltz wrote the computer program for the example that is worked out. He was assisted in getting numerical data and plotting graphs by Mr. R. L. Haken.

The author has also had some useful discussions with Professor F. N. Bailey.

Messrs. J. N. Gittleman, J. A. Colligan, and D. G. Mueller read the manuscript and offered construction criticisms.

LIST OF ILLUSTRATIONS

<u>Figure</u>	<u>Title</u>	<u>Page</u>
1(a)	Loss vs. Discriminant (odd numbers) used in the illustrative example of signal detection.	38
1(b)	Average loss over twenty runs at each stage of the adaptive process.	38
2(a)	Loss vs. Discriminant (odd numbers) used in the illustrative example of signal detection.	39
2(b)	Average loss over twenty runs at each stage of the adaptive process.	39
3(a)	Loss vs. Discriminant (odd numbers) used in the illustrative example of signal detection.	40
3(b)	Average loss over twenty runs at each stage of the adaptive process.	40
4(a)	Loss vs. Discriminant (odd numbers) used in the illustrative example of signal detection.	41
4(b)	Average loss over twenty runs at each stage of the adaptive process.	41
5(a)	Loss vs. Discriminant (odd numbers) used in the illustrative example of signal detection.	42
5(b)	Average loss over twenty runs at each stage of the adaptive process.	42

SUMMARY

This report gives an adaptive procedure for selecting a discriminant for a pattern recognizer. No a priori knowledge of the probability density on the observation-space is assumed. Moreover the pattern recognizer is assumed to have a finite memory.

A mathematical model of the problem of pattern recognition is constructed and several theorems are proved. With the help of these theorems, the adaptive procedure is developed. This adaptive procedure is, in effect, a method of using the finite memory efficiently in "training" the pattern recognizer.

1. INTRODUCTION

The problem of pattern recognition is a problem of separating objects into two or more classes. Given an object, a set of measurements is made on this object. Based on the results of these measurements, the object is assigned to one of two or more classes.

Most of the problems of pattern recognition have been reduced to a common mathematical model in which an object is represented by a point in an n-dimensional space (the measurement-space). Each coordinate in the space corresponds to one of the measurements made on an object. By a "pattern" we shall mean a point in the measurement-space which represents an object to be classified. We shall sometimes refer to the measurement-space as the pattern-space.

The problem of classification is then reduced to the problem of defining simply-connected or multiply-connected boundaries (discriminants) in the pattern-space. The region on one side of each boundary is associated with one class of patterns, i. e. , points in the pattern-space representing one class of objects. If a received pattern falls on the side of a boundary which is associated with a class, then the pattern (with the associated object) is declared to belong to that class; otherwise it is declared not to belong to that class. The boundary may be included with one region or the other. The boundaries are said to be optimally selected if the "cost" (to be defined later) resulting from misclassifying patterns is minimum.

It becomes obvious that the problem of defining boundaries is greatly simplified if the patterns belonging to each class are clustered together in the measurement-space. This simplification can be achieved if the distinguishing features of different classes of objects are recognized and if the measurements made on an object are the measurements of these features.

Unfortunately, it is not always possible to directly measure the distinguishing

features of an object. This may be due to the fact that either the distinguishing features of the different classes cannot be recognized, or direct measurements of these features are not possible. Under these circumstances one does the "best" one can in selecting features to be measured.

In this report we shall not concern ourselves with the method of selecting the features. We shall assume that the features to be measured have been selected, and objects are represented by patterns or points in the n -dimensional space, n being the number of features being measured. The problem we are concerned with is the problem of selecting the optimum boundaries which separate different classes of patterns.

2. BACKGROUND

Considerable work has been done in determining the optimum boundaries in the measurement-space. A large percentage of this work is devoted to linear discriminants (hyperplane boundaries) in the measurement-space. Hu (Refs. 1, 2, 3, 4) and Singleton (Ref. 5) determine conditions under which two classes of binary numbers are linearly separable (i. e. , can be separated completely by a linear discriminant). Highleyman (Ref. 6) determines the optimum linear discriminants for the case of not-necessarily linearly separable classes. Cooper (Ref. 7) investigates the problems in which a linear discriminant is optimum. Sebestyen (Ref. 8) shows that the correlation and regression line techniques are also examples of linear discriminants.

Discriminants other than linear are also considered in the literature. For example, Cooper (Ref. 9) considers problems for which a hypersphere forms the optimum boundary of separation between two classes of patterns.

Transformations are sometimes applied to the measurement-space in such a way that the patterns belonging to each class are brought together or in such a way that in the new space the boundaries have simple geometric forms. Sebestyen (Ref. 8) considers several linear and nonlinear transformations. In some cases of linear transformations, he identifies the eigenvectors of the transformations with the features of the objects and the eigenvalues with the weights to be attached to these features.

For the cases where the occurrence of the patterns is stochastic, the theory of pattern recognition falls within the framework of statistical inference and decision theory. As Cooper (Ref. 10) shows, if the probability densities are known, classifications can be done on the basis of one of the four classical criteria, viz, Bayes, maximum-likelihood, minimax, and Neyman-Pearson. All four criteria are optimally satisfied by a decision rule in which the test statistic is a likelihood ratio which is compared to a threshold k . Chow (Ref. 11) describes a functional diagram of such an optimum system.

In actual practice, however, the exact probability densities, more often than not, are not known a priori. If the probability densities are known except for a few parameters, the optimum discriminant can be determined by an adaptive process. See, for example, Cooper (Ref. 9).

In many problems of practical importance, however, probability densities are completely unknown. Discriminants have to be determined from a few typical samples with known classifications. This is called the "learning mode" of pattern recognition. In such cases, it is natural to select discriminants which are optimum with respect to the given samples. As illustrated by Sebestyen (Ref. 8), these "optimum" discriminants may turn out to be very poor as far as the unknown patterns are concerned. The remedy is to continue with the learning mode even after the "recognition mode" (i. e. , the mode of operation in which a discriminant is used to classify unknown patterns) is started. The process is thus made adaptive by continuously improving upon the discriminant employed. This is possible if the correct classification of a pattern is made known by some outside agency after a classification decision is made about it. The process converges to the optimum discriminant or discriminants (if there are more than two classes of pattern), if the system has infinite storage capacity (infinite memory).

Infinite memory is not required if the patterns are known to be linearly separable. This has been demonstrated by Widrow and Hoff (Ref. 13) with the Adeline and Rosenblatt (Ref. 14) with the Perceptron. The convergence is proved mathematically, among others, by Novikoff (Ref. 15) and Albert (Ref. 16).

In this report, we shall consider a problem in which no prior knowledge of probability densities is assumed, the patterns are not necessarily linearly separable, the discriminants are not necessarily linear, and only a finite storage space is available. We give an adaptive procedure for obtaining the optimum discriminant at any time, and for sequentially "improving" the discriminant. For simplicity, we restrict ourselves to the case in which a pattern is to be classified into one of two classes.

3. THE MATHEMATICAL STRUCTURE

Assuming that there are n measurements made on an object and that the values of these measurements are real numbers, a point ω in the real n -dimensional space R^n is a pattern, and the space R^n is the pattern-space. Objects represented by these patterns belong either to a class A or to a class B. Due to noise, either in the measurements or in the system through which these measurements are transmitted, a point in R^n sometimes actually corresponds to an object of Class A and at other times an object of Class B. We shall refer to the product space $R^n \times \{A, B\}$ as the observation-space, and a point $(\omega, .)$, which represents an object and its true classification, as an observation. Thus a pattern is the first of the two elements of the ordered set which we call an observation.

The probability density on the observation-space will be denoted by $p(\omega, .)$. Sometimes, for convenience, we shall drop the argument $(\omega, .)$ and denote this probability density by p . At this time we want to point out that in the problem we shall be considering, we shall assume that $p(\omega, .)$ is not known.

Let $\Gamma = \{c_1, c_2, \dots, c_M\}$ be the set of all discriminants (boundaries in the pattern-space R^n) under consideration. We assume that M is finite. Each discriminant c_i divides the space R^n into two disjoint sets Ω_A^i and Ω_B^i . The points in Ω_A^i are classified as patterns belonging to Class A, and those in Ω_B^i are classified as patterns belonging to Class B.

If the probability density p on the observation space were known, one plausible way to define the "best" discriminant is to say that it is a discriminant which minimizes a given loss function $L(c_i, p)$. One such loss function is given by

$$L(c_i, p) = \int_{\Omega_B^i} p(\omega, A) d\omega + \int_{\Omega_A^i} p(\omega, B) d\omega .$$

This loss function assumes that the loss is one if a pattern is incorrectly classified, and is

zero if it is correctly classified. If we define E_i as the event that the discriminant c_i classifies an unknown pattern incorrectly, then $L(c_i, p)$, defined above, is the probability of the event E_i . In the future when we refer to the loss function, we shall mean the loss function defined above.

Since $p(\omega, \cdot)$ is not known, we cannot use the simple criterion of minimizing the loss function $L(c_i, p)$ to select the best discriminant. We now give an intuitive explanation of the criterion we actually use for our case.

It is not necessary to know the probability density $p(\omega, \cdot)$ completely in order to determine the best discriminant defined above. It is not even necessary to know the exact values of $L(c_i, p)$ of the loss function. It is sufficient to know the correct ordering of the discriminants according to the values of the loss function, or, what amounts to the same thing, the ordering of the events E_i 's according to their probabilities.

In the absence of any information, all possible orderings (which are $M!$ in number)¹ are assumed equally probable. The absence of any information about $p(\omega, \cdot)$ implies that the probability distribution over the space of all possible probability density functions on the observation-space is uniform. We consider a more general situation in which the probability distribution over the space of all possible probability density functions on the observation-space is not necessarily uniform, yet the $M!$ orderings of the events E_i 's remain equally probable, in absence of any other information.

To describe how such a situation may arise, we first define a class of sets A_k , $k = 1, 2, \dots, M$. Each set A_k is a set of $\frac{M(M-1)\dots(M-k+1)}{k!}$ positive numbers in the range $[0, 1]$. There is a one-to-one correspondence between the elements of A_k and subsets of k distinct elements of the set A_1 . The numbers in the various sets are compatible in the sense that if the numbers in A_1 represent the probabilities of M events, each number in the set A_k represents the probability of the intersection of the corresponding k events. With this definition of the Class $\{A_k\}$ we describe the general situation we referred to before as follows:

¹We consider $M!$ distinct orderings, even though it may happen that the probabilities of two or more events are equal.

A one-to-one mapping is known to exist between $\{E_i\}$ and A_1 such that if a_j in A_1 corresponds to E_i , then a_j is the probability of E_i . Though such a mapping is known to exist, the actual mapping is unknown. In other words, knowing $\{A_k\}$, one knows the numbers which are the probabilities of various events E_i 's and their intersections, but one does not know which number in A_1 is the probability of which event. As such, all the $M!$ orderings of the events E_i 's still remain equally probable. What makes the situation more general than that implied by the complete lack of information about the probability density $p(\omega, .)$ is that we assume that the probability distribution over all possible values of $\{A_k\}$ is not-necessarily uniform.

In such a situation, one way of selecting a discriminant would be to select a discriminant conditional to knowing $\{A_k\}$ and then somehow weigh the discriminants for various values of $\{A_k\}$ according to the probability distribution on $\{A_k\}$. Fortunately, it turns out this does not become necessary, because the way we select a discriminant conditional to knowing $\{A_k\}$ is independent of the assumed value of $\{A_k\}$. From now on we shall assume that $\{A_k\}$ is known.

Corresponding to a given $\{A_k\}$ there is a set \mathcal{E} of the probability densities on the observation space which lead to the probabilities of various events E_i 's and probabilities of the intersections of these events as given by $\{A_k\}$. This set \mathcal{E} is divided into $M!$ subsets $B_1, B_2, \dots, B_{M!}$. Each B_j is a set of probability densities on the observation-space all of which correspond to one ordering of events E_i 's. The probability of the actual probability density $p(\omega, .)$ belonging to any subset B_j conditional to knowing $\{A_k\}$ is $\frac{1}{M!}$, when no other information is available. At the beginning, therefore, there is no reason to consider that one discriminant is "better" than any other.

At any future time in the process when some information about the observation-space is available, a discriminant (to classify unknown patterns) is selected on the basis of the available information. The selection rule for the discriminant is called a decision function. A decision function is a "mapping"² of the space of available information into the set Γ of the discriminants.

²In the entire text, the term "mapping" as applied to a decision function is intended to include the case when the values of a function are not deterministic but probabilistic.

Since, for our problem, we are assuming that we have a limited amount of memory available, we can store only a limited amount of available information. (We shall assume that the amount of memory available is less than that required to store M numbers. This is necessary because otherwise, as will become apparent later, the process becomes trivial.) The available information can be in the form of observations (i. e. , patterns with their true classifications) or the discriminants that were selected in the past with the number of errors these discriminants made in classifying patterns (or the number of patterns correctly classified by these discriminants). For convenience, we break up the space of information into subspaces and shall refer to the restrictions of a decision function on different subspaces as decision functions of different types. What these different subspaces are and what the corresponding different types of decision functions are will become clear in the next section. For the time being we shall concentrate on one of the subspaces which we shall denote by Z . A decision function δ of the corresponding type "maps" the subspace Z into the set Γ of the discriminants.

Instead of defining the best discriminant as we do when the probability density on the observation-space is known, we now define the best or the optimum decision function of Type Z . Towards that end we first define a risk function R , whose values for a given decision function δ and a given subset B_i is given by

$$R(\delta, B_i) = \sum_{j=1}^M L(c_j, p) P(Z_j | p \in B_i) .$$

Here Z_j is the set of all elements of the information subspace Z which are mapped by δ into the element c_j of the set Γ , and $P(Z_j | p \in B_i)$ is the probability density $p(\omega, .)$ on the observation-space is an element of the subset B_i . It is well to remember that the value of $R(\delta, B_i)$ actually depends on which element of B_i is the probability density $p(\omega, .)$. However, we have chosen not to depict this fact in the symbol for the risk function in order to avoid awkwardness in notation.

Since initially the probability density $p(\omega, .)$ on the observation-space is by assumption equally likely to come from any one subset B_i of the set \mathcal{B} , we define the average risk $\bar{R}(\delta)$ as

$$\bar{R}(\delta) = \frac{1}{M!} \sum_{i=1}^{M!} R(\delta, B_i) .$$

Here again we would like to point out that, though we have chosen not to show it symbolically, the fact is that $\bar{R}(\delta)$ is a function of the particular elements of the subsets B_i , $i = 1, 2, \dots, M!$, which are assumed to be the probability density $p(\omega, \cdot)$. For a fixed set of these elements, we define an optimum decision function δ^* by the following inequality:

$$\bar{R}(\delta^*) \leq \bar{R}(\delta) ,$$

where δ is any decision function of Type Z which maps the subspace of information Z into the set Γ .

Though we have defined the optimum decision function for a fixed set of $M!$ elements, one from each subset B_i , $i = 1, 2, \dots, M!$, it turns out that $\bar{R}(\delta^*)$ does not depend on this set but only on the Class $\{A_k\}$, while the optimum decision function of the Type Z is independent even of $\{A_k\}$.

4. THEOREMS AND USEFUL RESULTS

In this section we consider decision functions of various types, i. e. , the restrictions of decision functions which map various subspaces of the information space into the set Γ of discriminants. We also state and prove theorems concerning optimum decision functions of various types.

Recall that by an observation $(\omega, .)$ we mean a pattern and its correct classification. A set of m observations will be called a sample of m observations, and will be denoted by S^m . We denote the cardinality of an arbitrary set T by $\rho(T)$. Thus $\rho(S^m) = m$. The symbol $\rho(S^m \cap E_i)$ then denotes the number of errors that the discriminant c_i makes in classifying the m patterns of the sample S^m . We also define $\rho^o(S^m)$ and $\mathcal{C}^o(S^m)$ as follows:

$$\rho^o(S^m) = \min_{E_i} \rho(S^m \cap E_i)$$

and

$$\mathcal{C}^o(S^m) = \{c_i \mid \rho(S^m \cap E_i) = \rho^o(S^m)\} .$$

Thus $\rho^o(S^m)$ is the minimum number of errors made by a discriminant in classifying m patterns of a sample S^m , and $\mathcal{C}^o(S^m)$ is the set of all the discriminants which make the least number of errors in classifying the m patterns of the sample S^m . An arbitrary element of $\mathcal{C}^o(S^m)$ will be denoted by $c^o(S^m)$.

For a fixed m , a decision function δ_I^m is a mapping of the space \mathcal{S}^m of all S^m 's into Γ . Here we have defined one type of decision function.

Theorem 1

$$\delta_I^{m*}(S^m) = c^o(S^m)$$

for $S^m \in \mathcal{S}^m$

Remark: To put it in words, the theorem states that given a sample of m observations, the best decision that can be made about selecting a discriminant is to select any one of the discriminants which make the least number of errors in classifying the m patterns of the sample S^m .

Proof: We shall prove this theorem first for two special cases before proving it for the most general case.

Case (i):

$$\Gamma = \{c_1, c_2\}$$

$$= \{B_1, B_2\}$$

$$P(E_1 | p \in B_1) = P(E_2 | p \in B_2) = a$$

$$P(E_2 | p \in B_1) = P(E_1 | p \in B_2) = 1 - a$$

$$P(E_1 \cap E_2 | p \in B_1) = P(E_1 \cap E_2 | p \in B_2) = 0$$

Thus, for this case, $A_1 = \{a, 1 - a\}$ and $A_2 = \{0\}$. This is the case of only two discriminants, and these two discriminants are such that if one classifies a pattern correctly, the other does not.

Let a decision function δ (for convenience we shall drop the superscript m and the subscript I for this proof) be such that $F_{11} \triangleq P[\delta(\cdot) = c_1 | p \in B_1]$ = probability that the value of the δ is c_1 if the probability density $p(\omega, \cdot)$, on the observation-space, is some element of subset B_1 ; $F_{22} \triangleq P[\delta(\cdot) = c_2 | p \in B_2]$ = probability that the value of the δ is c_2 if the probability density $p(\omega, \cdot)$, on the observation-space, is some element of the subset B_2 .

$$R(\delta, B_1) = a F_{11} + (1 - a)(1 - F_{11})$$

and

$$R(\delta, B_2) = (1 - a)(1 - F_{22}) + a F_{22} .$$

Consequently, since for this case

$$\bar{R}(\delta) = \frac{1}{2} R(\delta, B_1) + \frac{1}{2} R(\delta, B_2),$$

we have

$$\bar{R}(\delta) = (1-a) + \frac{1}{2} (2a-1) (F_{11} + F_{22}).$$

If $a = 1-a$, any decision function is optimum and the theorem is proved. If $a \neq 1-a$, the optimum decision function minimizes $F_{11} + F_{22}$. Let us assume that $a > (1-a)$. The proof for the case $(1-a) > a$ should follow similarly.

Let us look at this problem from the point of view of hypothesis testing (Ref. 17). Let the hypothesis to be tested by H: Probability of $E_2 >$ probability of E_1 , against the alternate K: Probability of $E_1 >$ probability of E_2 . If the hypothesis is accepted when $\delta(S^m) = c_1$, and rejected when $\delta(S^m) = c_2$, then F_{11} is the level of significance of the test, and $F_{22} = 1 - \beta$, where β is the power of the test.

According to the fundamental lemma of Neyman and Pearson (Ref. 17), there is a most powerful test for which F_{11} is minimum for a given F_{22} . If the result of that lemma is applied to the case of the testing of the above hypothesis, then the most powerful test is given by the following:

For a fixed F_{22} , F_{11} is minimum if

$$\begin{aligned} \delta(S^m) &= c_1 \text{ for } \rho(S^m \cap E_2) > k \rho(S^m \cap E_1) \\ &= c_2 \text{ for } \rho(S^m \cap E_2) < k \rho(S^m \cap E_1) \\ &= c_1 \text{ with probability } \gamma \\ &= c_2 \text{ with probability } (1-\gamma) \end{aligned} \left. \vphantom{\begin{aligned} \delta(S^m) &= c_1 \text{ for } \rho(S^m \cap E_2) > k \rho(S^m \cap E_1) \\ &= c_2 \text{ for } \rho(S^m \cap E_2) < k \rho(S^m \cap E_1) \\ &= c_1 \text{ with probability } \gamma \\ &= c_2 \text{ with probability } (1-\gamma) \end{aligned}} \right\} \text{ for } \rho(S^m \cap E_1) = \rho(S^m \cap E_2).$$

Here $k \geq 0$ and $0 \leq \gamma \leq 1$. The dependence of F_{22} on k and γ are given by:

$$\begin{aligned} F_{22} &= \{P[\rho(S^m \cap E_2) < k \rho(S^m \cap E_1) \mid p \in B_2]\} \\ &+ (1-\gamma) \{P[\rho(S^m \cap E_2) = k \rho(S^m \cap E_1) \mid p \in B_2]\} \end{aligned}$$

if k is such that $r \triangleq \frac{km}{k+1}$ is an integer; otherwise, $\rho(S^m \cap E_2)$ cannot be equal to $k\rho(S^m \cap E_1)$, and so

$$F_{22} = P[\rho(S^m \cap E_2) < k\rho(S^m \cap E_1) | p \in B_2] .$$

The various probabilities involved in the equations for F_{22} can be easily calculated by applying binomial distributions. Thus

$$F_{22} = \sum_{i=0}^{i=r-1} \binom{m}{m-i} a^i (1-a)^{m-i} + (1-\gamma) \binom{m}{m-r} a^r (1-a)^{m-r} ,$$

if r is an integer; otherwise, the second term is zero and the summation in the first term is taken up to the largest integer less than r . For a fixed value of F_{22} , the minimum value of F_{11} which is obtained by using the most powerful test described above is given by:

$$\begin{aligned} F_{11} &= P[\rho(S^m \cap E_2) < k\rho(S^m \cap E_1) | p \in B_1] \\ &+ \gamma P[\rho(S^m \cap E_2) = k\rho(S^m \cap E_1) | p \in B_1] \\ &= \sum_{i=r+1}^m \binom{m}{m-i} (1-a)^i a^{m-i} + \gamma \binom{m}{m-r} (1-a)^r a^{m-r} \end{aligned}$$

if r is an integer; otherwise the second term in both the above equations is zero, and the summation in the first term of the second equation starts with the smallest integer that is larger than r .

Using the above equations for F_{11} and F_{22} , it can be easily deduced that $F_{11} + F_{22}$ is minimum, and the resulting decision function is optimum, if $k = 1$ and γ takes any value in the closed interval $[0, 1]$. Thus for the case (i)

$$\begin{aligned} \delta^*(S^m) &= c_1 \text{ if } \rho(S^m \cap E_2) > \rho(S^m \cap E_1) \\ &= c_2 \text{ if } \rho(S^m \cap E_1) > \rho(S^m \cap E_2) \\ &= c_1 \text{ with probability } \gamma \\ &= c_2 \text{ with probability } (1-\gamma) \end{aligned} \left. \vphantom{\begin{aligned} \delta^*(S^m) &= c_1 \text{ if } \rho(S^m \cap E_2) > \rho(S^m \cap E_1) \\ &= c_2 \text{ if } \rho(S^m \cap E_1) > \rho(S^m \cap E_2) \\ &= c_1 \text{ with probability } \gamma \\ &= c_2 \text{ with probability } (1-\gamma) \end{aligned}} \right\} \text{ if } \rho(S^m \cap E_1) = \rho(S^m \cap E_2)$$

where γ is any number in the closed interval $[0, 1]$. In other words $\delta^*(S^m) = c^0(S^m)$. This completes the proof for the case (i).

Case (ii):

$$\Gamma = \{c_1, c_2\}$$

$$= \{B_1, B_2\}$$

$$P(E_1 | p \in B_1) = P(E_2 | p \in B_2) = a + c$$

$$P(E_2 | p \in B_1) = P(E_1 | p \in B_2) = b + c$$

$$P(E_1 \cap E_2 | p \in B_1) = P(E_1 \cap E_2 | p \in B_2) = c$$

$$a + b + c < 1$$

Thus, for this case, $A_1 = \{a, b\}$ and $A_2 = \{c\}$. This is the case when the two events E_1 and E_2 are not necessarily disjoint and all-inclusive, i. e. , the two discriminants c_1 and c_2 are such that there are patterns that can be correctly classified by both the discriminants, and there are patterns which are incorrectly classified by both the discriminants.

For a decision function δ , let F_{ij}^{rt} be the probability that a sample S^m which satisfies the conditions

$$\rho[(S^m \cap E_1) \cap (S^m \cap E_2)] = r$$

$$\rho[(S^m \cap E_1) \cup (S^m \cap E_2)] = m - t ,$$

is mapped into c_j when the probability density $p(\omega, \cdot)$ on the observation-space is an element of the subset B_i , $i = 1, 2, \dots$. Notice that r is the number of patterns in the sample S^m which are incorrectly classified by both the discriminants c_1 and c_2 , and t is the number of patterns which are correctly classified by both c_1 and c_2 . For this case

$$\begin{aligned} \bar{R}(\delta) &= \frac{1}{2} [R(\delta, B_1) + R(\delta, B_2)] \\ &= \frac{1}{2} \sum_t \sum_r \left(a F_{11}^{rt} + a F_{22}^{rt} + b F_{21}^{rt} + b F_{12}^{rt} \right). \end{aligned}$$

The summations are taken over all possible values of r and t . But

$$F^{rt} \triangleq F_{11}^{rt} + F_{12}^{rt} = F_{21}^{rt} + F_{22}^{rt} .$$

So

$$\bar{R}(\delta) = \frac{1}{2} \sum_t \sum_r F^{rt} \left[(b + c) + (a - b) \left(\frac{F_{11}^{rt}}{F^{rt}} + \frac{F_{22}^{rt}}{F^{rt}} \right) \right] .$$

If $a = b$, any decision function is optimum, and the theorem is proved for Case (ii). If $a \neq b$, the optimum decision function minimizes $[(F_{11}^{rt}/F^{rt}) + (F_{22}^{rt}/F^{rt})]$, for every r and t . Minimizing $[(F_{11}^{rt}/F^{rt}) + (F_{22}^{rt}/F^{rt})]$, for fixed r and t , is similar to minimizing $F_{11} + F_{22}$ in the case (i), where a sample consists of $m - (r + t)$ observations, each pattern being correctly identifiable by one and only one of the two discriminants c_1 and c_2 . This is also evident from the fact that the patterns which are either correctly classifiable, or incorrectly classifiable by both the discriminants, do not help in the selection of one discriminant over the other.

The decision function which minimizes $[(F_{11}^{rt}/F^{rt}) + (F_{22}^{rt}/F^{rt})]$, for fixed r and t , is the optimum decision function of the case (i) when applied to the sample of $m - (r + t)$ observations which are correctly classifiable by one and only one discriminant. However, since adding the remaining $(r + t)$ observations does not change the value of the decision function, we conclude that for fixed r and t , the optimum decision function of the case (i), when applied to the complete sample S^m of the case (ii), minimizes $[(F_{11}^{rt}/F^{rt}) + (F_{22}^{rt}/F^{rt})]$. Since this is true for each r and t , the optimum decision function of the case (i), when applied to the complete sample S^m of the case (ii), minimizes $\bar{R}(\delta)$ for the case (ii). Thus the optimum decision function of case (i) is also the optimum decision function for the case (ii). This completes the proof of the theorem for the case (ii).

Case (iii): This is the most general case.

$$\Gamma = \{c_1, c_2, \dots, c_M\} .$$

$$\mathcal{B} = \{B_1, B_2, \dots, B_M\} .$$

For an arbitrary but fixed class $\{A_k\}$, B_i 's are the subsets of probability densities on the observation-space as defined before.

For this case, the theorem can be proved by applying the result of Case (ii) successively. Thus according to the case (ii), given a sample S^m , if the choice is between c_i and c_j , the optimum decision function has the value c_i

with probability 1 if $\rho(S^m \cap E_i) < \rho(S^m \cap E_j)$

with probability 0 if $\rho(S^m \cap E_i) > \rho(S^m \cap E_j)$

with probability γ if $\rho(S^m \cap E_i) = \rho(S^m \cap E_j)$

where $0 \leq \gamma \leq 1$. Notice that the last part of the above statement is equivalent to saying: Select c_i or c_j arbitrarily if $\rho(S^m \cap E_i) = \rho(S^m \cap E_j)$.

This optimum selection between c_i and c_j is true for all pairs of i and j . Combining the results of all such pairs of i and j , we conclude that, given a sample S^m , the optimum decision function selects c_i from the set Γ

with probability 1 if $\rho(S^m \cap E_i) < \rho(S^m \cap E_j)$ for all $j \neq i$,

with probability 0 if $\rho(S^m \cap E_i) > \rho(S^m \cap E_j)$ for any $j \neq i$,

with probability γ if $\rho(S^m \cap E_i) < \rho(S^m \cap E_j)$ for some $j \neq i$,

and $\rho(S^m \cap E_i) = \rho(S^m \cap E_j)$ for rest of $j \neq i$,

where $0 \leq \gamma \leq 1$. This is equivalent to saying that the value of the optimum decision function for a given sample S^m is $c^0(S^m)$. This proves Theorem 1.

Corollary:

$$\bar{R}(\delta_I^{m*}) \geq \bar{R}[\delta_I^{(m+i)*}], \text{ for } i > 0.$$

The proof of the corollary follows by observing that, given a sample S^{m+i} of $m+i$ observations, one way of selecting a discriminant is to disregard i of the observations

and use the decision function δ_I^{m*} on the remaining m observations. But this is not necessarily the best decision function $\delta_I^{(m+i)*}$. The equality holds for the trivial case when all E_i 's are equally probable.

We now define another type of decision function. For this decision function, the domain is $(\alpha, c_\beta) \times \mathcal{S}^m$, where α is a fixed positive integer less than or equal to m , and c_β is a fixed discriminant. The number α and the discriminant c_β indicate that sometime in the past there existed a sample S^m for which $\rho^0(S^m) = \alpha$ and $c^0(S^m) = c_\beta$, i. e., the minimum number of errors made by a discriminant in classifying m patterns of that sample was α , and one such discriminant which made α errors was c_β . This type of decision function will be denoted by $\delta_{II}^{m/\alpha}$. We shall further assume that for a given pair (α, c_β) , the decision function $\delta_{II}^{m/\alpha}$ first maps \mathcal{S}^m into $\{[\alpha_i(S^m), c_i]\}$, where $\alpha_i(S^m)$ denotes the number of errors made by a discriminant c_i in classifying the m patterns of a sample S^m , and then the decision function maps $(\alpha, c_\beta) \times \{[\alpha_i(S^m), c_i]\}$ into Γ . With the definition of this new type of decision function we prove the following:

Theorem 2:

$$\begin{aligned} \delta_{II}^{m/\alpha^*}(S^m) &= c_\beta && \text{if } \rho^0(S^m) > \alpha \\ &= c^0(S^m) && \text{if } \rho^0(S^m) < \alpha \\ &= c_\beta \text{ with probability } \gamma && \left. \vphantom{\begin{aligned} \delta_{II}^{m/\alpha^*}(S^m) \\ = c^0(S^m) \end{aligned}} \right\} \text{if } \rho^0(S^m) = \alpha . \\ &= c^0(S^m) \text{ with probability } (1 - \gamma) && \end{aligned}$$

Here $0 \leq \gamma \leq 1$.

Remark: To put it in words, the theorem states that if the minimum number of errors that a discriminant can make in classifying m patterns of a given sample S^m is greater than α , then it is better to select the discriminant c_β which made only α errors while classifying m patterns of some sample S^m . If the minimum number of such errors is less than α , then it is better to select any discriminant which makes the minimum number of errors while classifying m patterns of the present sample S^m . If, however, the minimum

number of errors is equal to α , any one of the two discriminants is selected arbitrarily.

Before we prove this theorem, we shall prove three lemmas.

Lemma 1:

Given two disjoint subspaces ζ_1 and ζ_2 of the total information space, a decision function δ which maps $\zeta_1 \cup \zeta_2$ into Γ is optimum if and only if the restriction of δ on ζ_i is optimum for $i = 1, 2$.

Remark: There are three types of decision functions involved in the above lemma—one that maps the subspace of information ζ_1 into Γ , the second that maps the subspace of information ζ_2 into Γ , and a third one which maps the union $\zeta_1 \cup \zeta_2$ of the two subspaces ζ_1 and ζ_2 into Γ . The lemma asserts that a decision function of this third type is optimum if and only if its restriction on the subspace ζ_1 is an optimum decision of the first type and its restriction on the space ζ_2 is an optimum decision function of the second type.

Proof: Let $R_i(\delta)$, $i = 1, 2$, be the average risk if the decision function δ is restricted to the subspace ζ_i . Let F_i , $i = 1, 2$, be the probability that the available information z is an element of the subspace ζ_i when the probability density $p(\omega, \cdot)$ on the observation-space is any one of the $M!$ (equally probable) possibilities used in evaluating $\bar{R}_i(\delta)$. Then the average risk $\bar{R}(\delta)$ of the decision function δ (unrestricted either to subspace ζ_1 or to ζ_2) is given by

$$\bar{R}(\delta) = F_1 \bar{R}_1(\delta) + F_2 \bar{R}_2(\delta)$$

The proof of the theorem follows immediately from the fact that a decision function δ , which maps $\zeta_1 \cup \zeta_2$ into Γ , is optimum if and only if it minimizes $R_i(\delta)$ for $i = 1, 2$.

Corollary:

Let $\mathcal{S}^{m/\alpha} \triangleq \{S^m \mid \rho^0(S^m) = \alpha\}$. Let $\delta_I^{m/\alpha}$ denote a type of decision function that maps $\mathcal{S}^{m/\alpha}$ into Γ . Then $\delta_I^{m/\alpha*}$ is the restriction of δ_I^{m*} on $\mathcal{S}^{m/\alpha}$.

Distinction between the two types of decision functions, $\delta_I^{m/\alpha}$ and $\delta_{II}^{m/\alpha}$, should be recognized. The domain of $\delta_I^{m/\alpha}$ is the set $\mathcal{S}^{m/\alpha}$, i. e., the set of all samples S^m for which the minimum number of errors a discriminant can make is α . The domain of the decision function $\delta_{II}^{m/\alpha}$ is the product space $(\alpha, c_\beta) \times \mathcal{S}^m$, where α and c_β are fixed.

Lemma 2:

$$\bar{R}(\delta_I^{m/\alpha^*}) \geq \bar{R}(\delta_I^{m/\alpha-1^*})$$

where $m > \alpha > 1$.

The equality sign holds if and only if all the events E_i 's are equally probable, i. e. , all the M numbers in the set A_1 are equal.

Remark: This lemma suggests that if there is a choice between two discriminants, one of which makes $(\alpha - 1)$ errors in classifying m patterns of a sample S_1^m for which the least number of errors a discriminant can make is $(\alpha - 1)$, and another discriminant which makes α errors in classifying m patterns of another sample S_2^m for which the least number of errors a discriminant can make is α , then it is "better" to select the first of the two discriminants for the nontrivial case when all the events are known to be not equally probable. In the trivial case any one of the two discriminants can be selected arbitrarily.

Proof: As in the case of Theorem 1, we shall prove this lemma for two special cases before proceeding to the proof of the third and the most general case. For this proof, we shall drop the subscript I and the superscript $*$ from the optimum decision functions.

Case (i):

$$\Gamma = \{c_1, c_2\}$$

$$B = \{B_1, B_2\}$$

$$P(E_1 | p \in B_1) = P(E_2 | p \in B_2) = a$$

$$P(E_2 | p \in B_1) = P(E_1 | p \in B_2) = 1 - a$$

$$P(E_1 \cap E_2 | p \in B_1) = P(E_1 \cap E_2 | p \in B_2) = 0$$

For this case

$$\begin{aligned}
R(\delta^{m/\alpha}, B_1) &= R(\delta^{m/\alpha}, B_2) = a \frac{\binom{m}{\alpha} a^\alpha (1-a)^{m-\alpha}}{\binom{m}{\alpha} a^\alpha (1-a)^{m-\alpha} + \binom{m}{\alpha} (1-a)^\alpha a^{m-\alpha}} \\
&+ (1-a) \frac{\binom{m}{\alpha} a^{m-\alpha} (1-a)^\alpha}{\binom{m}{\alpha} a^\alpha (1-a)^{m-\alpha} + \binom{m}{\alpha} (1-a)^\alpha a^{m-\alpha}} \\
&= a \frac{1 + \frac{a}{1-a}^{m-2\alpha-1}}{1 + \frac{a}{1-a}^{m-2\alpha}}
\end{aligned}$$

The lemma is proved for this case, if we prove that

$$R(\delta^{m/\alpha}, B_1) \geq R(\delta^{m/\alpha-1}, B_2).$$

Now

$$\begin{aligned}
R(\delta^{m/\alpha}, B_1) - R(\delta^{m/\alpha-1}, B_2) &= a \frac{1 + \left(\frac{a}{1-a}\right)^{m-2\alpha-1}}{1 + \left(\frac{a}{1-a}\right)^{m-2\alpha}} - a \frac{1 + \left(\frac{a}{1-a}\right)^{m-2\alpha+1}}{1 + \left(\frac{a}{1-a}\right)^{m-2\alpha+2}} \\
&= a \frac{\left(\frac{a}{1-a}\right)^{m-2\alpha-1} \left(1 - \frac{a}{1-a}\right) \left[1 - \left(\frac{a}{1-a}\right)\right]^2}{\left[1 + \left(\frac{a}{1-a}\right)^{m-2\alpha}\right] \left[1 + \left(\frac{a}{1-a}\right)^{m-2\alpha+2}\right]} \\
&\geq 0
\end{aligned}$$

The equality sign holds if and only if $a = 1 - a$. Notice that a or $(1 - a)$ cannot be equal to zero, because otherwise $\mathcal{S}^{m/\alpha}$, for $\alpha \neq 0$, is empty. This completes the proof for Case (i).

Case (ii):

$$\Gamma = \{c_1, c_2\}$$

$$B = \{B_1, B_2\}$$

$$P(E_1 | p \in B_1) = P(E_2 | p \in B_2) = a + c$$

$$P(E_2 | p \in B_2) = P(E_1 | p \in B_1) = b + c$$

$$P(E_1 \cap E_2 | p \in B_1) = P(E_1 \cap E_2 | p \in B_2) = c$$

$$a + b + c < 1$$

The lemma is proved for this case if we prove that $R(\delta^{m/\alpha}, B_1) + R(\delta^{m/\alpha}, B_2) \geq R(\delta^{m/\alpha-1}, B_1) + R(\delta^{m/\alpha-1}, B_2)$. The risk $R(\delta^{m/\alpha}, B_i)$, for $i = 1, 2$, can be written as

$$R(\delta^{m/\alpha}, B_i) = \sum_s \sum_t R_{st}(\delta^{m/\alpha}, B_i),$$

where R_{st} is that portion of the risk which corresponds to all the elements of $\mathcal{S}^{m/\alpha}$ which have s patterns correctly classifiable by both c_1 and c_2 , and t patterns incorrectly classifiable by both c_1 and c_2 . Using the procedure similar to the one used in the proof of Case (i), it can be shown that

$$R_{st}(\delta^{m/\alpha}, B_1) + R_{st}(\delta^{m/\alpha}, B_2) \geq R_{st}(\delta^{m/\alpha-1}, B_1) + R_{st}(\delta^{m/\alpha-1}, B_2)$$

for every possible value of s and t . The equality holds if and only if $a + c = b + c$, i. e., the two events E_1 and E_2 are equally probable. This completes the proof for Case (ii).

Case (iii):

$$\begin{aligned} \Gamma &= \{c_1, c_2, \dots, c_M\} \\ &= \{B_1, B_2, \dots, B_{M!}\} \end{aligned}$$

where the sets B_i 's have been defined before. For this case

$$\bar{R}(\delta^{m/\alpha}) = \frac{1}{M!} \sum_{i=1}^{M!} R(\delta^{m/\alpha}, B_i).$$

Because of their properties mentioned before, the subsets B_i 's of the set \mathcal{B} can be paired off. Each such pair (B_{i_1}, B_{i_2}) has the following properties:

$$\begin{aligned} P(E_{j_1} | p \in B_{i_1}) &= P(E_{j_2} | p \in B_{i_2}) \\ P(E_{j_2} | p \in B_{i_1}) &= P(E_{j_1} | p \in B_{i_2}) \\ P(E_{j_1} \cap E_{j_2} | p \in B_{i_1}) &= P(E_{j_1} \cap E_{j_2} | p \in B_{i_2}), \end{aligned}$$

Also $P(E_k | p \in B_{i_1}) = P(E_k | p \in B_{i_2})$ for $k = 1, 2, \dots, M$ except j_1 and j_2 .

We can, therefore, write

$$\bar{R}(\delta^{m/\alpha}) = \frac{1}{M!} \sum R(\delta^{m/\alpha}, B_{i_1}) + R(\delta^{m/\alpha}, B_{i_2})$$

where the summation is taken over distinct pairs (i_1, i_2) . Each pair of terms under the summation represents Case (ii). Thus we consider the case (iii), and hence the lemma, proved, with the observation that the equality in the lemma for this case will hold if and only if

$$\begin{aligned} P(E_k | p \in B_{i_1}) &= P(E_k | p \in B_{i_2}) \\ \text{for } i_1, i_2 &= 1, 2, \dots, M! \\ \text{and } k &= 1, 2, \dots, M. \end{aligned}$$

Corollary to Lemma 2:

$$\begin{aligned} \bar{R}(\delta_I^{m/\alpha^*}) &\geq \bar{R}(\delta_I^{m/\alpha-k^*}) \\ \text{for } m &\geq \alpha \geq k > 0. \end{aligned}$$

The equality holds if and only if the probabilities of all the events E_i 's are the same.

This is a generalization of Lemma 2 and can be proved by successive application of the lemma.

Lemma 3:

Let $\zeta_1, \zeta_2, \dots, \zeta_\ell$ be ℓ disjoint information subspaces. Let δ_i denote a decision function of the i th type which maps ζ_i into Γ , for $i = 1, 2, \dots, \ell$. Let δ denote a decision function of the type which maps $\left(\bigcup_{i=1}^{\ell} \zeta_i\right) \times \left(\bigcup_{i=1}^{\ell} \zeta_i\right)$ first into $\{i, \delta_i(z_i)\} \times \{j, \delta_j(z_j)\}$ and then maps $\{i, \delta_i(z_i)\} \times \{j, \delta_j(z_j)\}$ into Γ , where (z_i, z_j) is an element of $\left(\bigcup_{i=1}^{\ell} \zeta_i\right) \times \left(\bigcup_{i=1}^{\ell} \zeta_i\right)$ with $z_i \in \zeta_i$ and $z_j \in \zeta_j$. Then

$$\begin{aligned} \delta^*(z_i, z_j) &= \delta_j^*(z_i) & \text{if } \bar{R}(\delta_i^*) < \bar{R}(\delta_j^*) \\ &= \delta_i^*(z_j) & \text{if } \bar{R}(\delta_i^*) > \bar{R}(\delta_j^*) \\ &= \text{either } \delta_i^*(z_i) \text{ or } \delta_j^*(z_j) \text{ arbitrarily if} \end{aligned}$$

$$\bar{R}(\delta_i^*) = \bar{R}(\delta_j^*) ,$$

for $i, j = 1, 2, \dots, \ell$.

Proof: By Lemma 1, it is sufficient to prove this lemma for every restriction δ_{ij} of the decision function δ on the subspace $\zeta_i \times \zeta_j$, $i, j = 1, 2, \dots, \ell$. Let the value of a decision function δ_{ij} be $\delta_i(z_i)$ with probability a and $\delta_j(z_j)$ with probability $(1-a)$. Then the average risk $\bar{R}(\delta_{ij})$ is given by

$$\bar{R}(\delta_{ij}) = a\bar{R}(\delta_i) + (1-a)\bar{R}(\delta_j) .$$

It now becomes evident that $\bar{R}(\delta_{ij})$ becomes minimum, and consequently δ_{ij} becomes optimum, for $a = 1$ and $\delta_i = \delta_i^*$ if $\bar{R}(\delta_i^*) < \bar{R}(\delta_j^*)$; $a = 0$ and $\delta_j = \delta_j^*$ if $\bar{R}(\delta_j^*) < \bar{R}(\delta_i^*)$; a arbitrary and $\delta_i = \delta_i^*$ and $\delta_j = \delta_j^*$ if $\bar{R}(\delta_i^*) = \bar{R}(\delta_j^*)$. Thus the lemma is proved.

Proof of Theorem 2:

The decision function $\delta_{II}^{m/a}$ can be looked upon as a part of a mapping of the space $\left(\bigcup_{i=0}^m \mathcal{L}^{m/i}\right) \times \left(\bigcup_{i=0}^m \mathcal{L}^{m/i}\right)$ first into $\{i, \delta_I^{m/i}(s^{m/i})\} \times \{j, \delta_I^{m/j}(s^{m/j})\}$ and then into the set Γ . According to the lemma 2, $\bar{R}(\delta_I^{m/i^*}) \geq \bar{R}(\delta_I^{m/j^*})$ for $i > j$, where the equality sign holds for the case when the probabilities of all the events E_i 's are equal. The theorem 2 follows from Lemma 3.

Corollary 1:

$$\bar{R}\left(\delta_{\text{II}}^{\left(m/\alpha_1^*\right)}\right) \leq \bar{R}\left(\delta_{\text{II}}^{\left(m/\alpha_2^*\right)}\right)$$

if $\alpha_1 < \alpha_2$. The equality sign holds only for the case when the events E_i 's are equally probable.

Corollary 2:

$$\bar{R}\left(\delta_{\text{II}}^{m/\alpha^*}\right) \leq \bar{R}\left(\delta_{\text{I}}^{m/\alpha^*}\right).$$

The equality sign holds only for the case when the events E_i 's are equally probable.

Theorem 3:

$$R\left(\delta_{\text{I}}^{m/\alpha^*}\right) = R\left(\delta_{\text{I}}^{(m-\alpha)/0^*}\right) \quad \text{for } m > \alpha.$$

Proof:

Given a $bs^{m/\alpha}$ (i. e. , a sample of m observations such that the minimum number of errors a discriminant can make in classifying its patterns is α), one can always pick a sample $bs^{(m-\alpha)/0}$ from it, if $m > \alpha$.

The discriminant $\delta_{\text{I}}^{m/\alpha^*}(bs^{m/\alpha})$ is any element of the set $\zeta^0(bs^{m/\alpha})$, the set of all discriminants which make α errors in classifying patterns of the sample $bs^{m/\alpha}$. One can choose to select this element of $\zeta^0(bs^{m/\alpha})$ by first selecting a sample $s^{(m-\alpha)/0}$ from the sample $bs^{m/\alpha}$ and then taking the value $\delta_{\text{I}}^{(m-\alpha)/0^*}[bs^{(m-\alpha)/0}]$. This proves the theorem.

Corollary:

$$\bar{R}\left(\delta_{\text{I}}^{\left(m_1/\alpha_1^*\right)}\right) \leq \bar{R}\left(\delta_{\text{I}}^{\left(m_2/\alpha_2^*\right)}\right)$$

if $(m_1 - \alpha_1) > (m_2 - \alpha_2)$, with the equality holding only for the case when all the events E_i 's have equal probabilities.

Before going on to the next section we would like to point out that though we assumed the knowledge of the class $\{A_k\}$, none of the optimum decision functions depend on this class.

5. DESCRIPTION OF THE PROBLEM

In this section we describe the pattern-recognition problem we have considered. The problem is described below in the form of a list of assumptions and related remarks:

A. We assume that the observation-space is known to be $R^n \times \{A, B\}$. This means that an object is known to be represented by a set of n measurements and that it is known to belong to one of the two classes A and B .

B. The probability density $p(\omega, \cdot)$ on the observation-space is assumed to be unknown a priori.

C. We assume that observations are made available as a sequence in real time, not necessarily at fixed intervals. Note again that by an observation (ω, \cdot) we mean a pattern and its correct classification.

D. The instant of time immediately after a new observation is available will be called a "stage" of the "process." At every stage a fresh discriminant is selected to classify unknown patterns. The discriminant is selected from a given set $\Gamma = \{c_1, c_2, \dots, c_M\}$ of M discriminants, where M is finite. In some special cases where the discriminants are simple geometrical boundaries, it may not be necessary to require M to be finite. In any case, in order that the problem does not become trivial, M is required to be larger than the amount of available memory, so that it is not possible to count the number of errors made by each discriminant in classifying known patterns and to select a discriminant on the basis of those numbers by the application of Theorem 1.

In general, we make no assumptions about the relative geometric properties of the discriminants. However, we shall point out wherever advantage can be taken of simple relations, if they exist, between the elements of the set Γ . For example, if all the elements of Γ are completely arbitrary, each one has to be stored separately. However, if

the elements of Γ are, say, equally-spaced concentric hyperspheres, or hyperplanes inclined at equally-spaced angles to a fixed hyperplane, then it is sufficient to have an algorithm that would generate these discriminants when required.

E. Storage space is available to keep a record of the past observations. However, this storage space is not infinite; therefore, there comes a time after which all the past observations cannot be stored. To be very specific, we shall assume that there are λ units of storage space available to store information about past observations, one unit equivalent to the space required to store one observation. We shall refer to this storage space as the "information memory." Note that it is not necessary that all the λ units be used up in storing past observations. However, the information memory is to be used only for the purpose of storing information about the past observations. The information we shall be concerned with storing will be old observations and discriminants with known numbers of errors these discriminants have made, or known numbers of patterns these discriminants have correctly classified.

F. At every stage, after making the decision about the discriminant to be selected, there is another decision to be made if the information memory is full. This concerns the question of how the presently available information in the information memory is to be rearranged, and how much of that information should be destroyed in order to make a unit available for a latest observation.

G. If a new observation arrives before the step of making room for it is completed, that observation is not taken into account.

H. The "process," therefore, consists of repeating the cycle of storing a new observation, selecting the "optimum" discriminant from the information available in the information memory, and making room in the information memory for a fresh observation in such a way that the "optimum" discriminant can be "improved" in real time. The cycle is repeated as long as new observations are available.

I. We shall mention one more constraint on our system. In the adaptive procedure to be described later, very often before the "optimum" discriminant is selected at a stage, we want to find a discriminant which makes the least number of errors in classifying patterns associated with the observations which are stored in the information

memory at that stage. It is possible that there is more than one such discriminant. In fact, let us say that there is a subset Γ_0 of discriminants which satisfy this requirement. If Γ_0 contains more than one element, then the adaptive procedure requires an arbitrary selection of one of the elements of Γ_0 .

We assume that there is no storage space available to store the subset Γ_0 . We assume that there is enough storage space for two discriminants and two numbers in addition to information memory to compare two discriminants according to the number of errors they make in classifying the patterns in the information memory. Thus we have the constraint that only one element of the subset Γ_0 can be known at a time.

If we wish to make the selection of this element of Γ_0 arbitrary, we start testing discriminants at an arbitrary point in the sequence $\Gamma = \{c_1, c_2, \dots, c_M\}$ and go through the set Γ systematically. We arrange to select the first discriminant that made the least number of errors.

It is not necessary to go through the entire set Γ , if the elements of Γ are geometrically related. For example, suppose the elements of Γ are concentric hyperspheres with the points in the inside of each hypersphere representing the patterns of one fixed class. The procedure in such a case would be to select a hypersphere arbitrarily and determine the number of errors it makes and to store them. Then test the next larger hyper-sphere and determine the number of errors it makes. If the latter makes fewer errors than the former, store it and the number of errors it has made in place of the former; otherwise retain the former. Then test the next larger sphere and continue the process until the number of errors made by a sphere is twice as many as the one that is stored. It can be easily established that there is no larger sphere that can do better than one that is stored. At that stage start testing spheres smaller than the one that was first tested, and repeat the procedure for the smaller spheres. It can be shown that this procedure leads to the selection of an arbitrary element of Γ_0 without necessarily testing all the elements of the set Γ .

6. RESULT OF THE INVESTIGATION

We now give the result of our investigation in the form of the following theorem:

Theorem 4:

Let $x_1, x_2, \dots, x_k, \dots$ be a sequence of independent observations in real time (i. e. , $t_i < t_j$ for $i < j$, where t_i is the time of arrival of the observation x_i). Let y_k be the discriminant selected from the set Γ (to classify unknown patterns) at a stage k (i. e. , at the instant of time immediately after the obseravtion x_k is stored). Then if y_k is selected according to the adaptive procedure given below, it is "adaptively optimum" in the following sense:

(i) For the content of the information memory at the k th stage, the decision function used to determine y_k is the optimum.

(ii) The information retained in the information memory after the discriminant y_k is selected is such that the type of decision function anticipated to be used at $(k+1)$ st-stage is better than (i. e. , its average risk is less than that of) the type of decision function actually used at the k th-stage, unless there is not enough information in the information memory at the k th-stage. In the latter case the information retained after the discriminant y_k is selected is such that the anticipated type of decision function to be used at the $(k+i)$ th-stage is better than the one used at the k th-stage, when $i > 1$; and the decision function to be used at $(k+1)$ st-stage is as good as the one used at the k th-stage.

Remarks:

1. The difference between the "type of decision function anticipated to be used" and the "type of decision function actually used," should be noted. One anticipates using a certain type of decision function before one knows the actual value of the newly arrived observation. One knows what decision function was actually used, after the decision

has been made.

2. The adaptive procedure can be described roughly in words as follows: There are three modes of operation. During the first mode space is available in the information memory to store new observations. In this mode any one of the discriminants which make the least number of errors in classifying stored patterns is selected. During the second mode of operation the information memory is filled with observations and a number which shows the errors made by the discriminant in classifying patterns in the memory at the time it was selected. A new observation is accommodated by removing the oldest observation from the memory. A comparison is made between the number of errors of the present discriminant and the least number of errors that can be made in classifying patterns in the memory by an element of the set Γ . If the former is larger than the latter, a new discriminant is selected which makes the least number of errors in classifying the patterns in the information memory, and the number of errors it makes is stored in the memory instead of the number that is presently stored. Otherwise, no change is made. This mode is continued until all patterns in the memory are correctly classified by a discriminant. At that stage the adaptive procedure goes in the third mode of operation. To describe the third mode of operation roughly, a comparison is made between the number of "successive" patterns correctly classified by the present discriminant and the number of "successive" patterns correctly classified by an "appropriately" selected discriminant. If the former is larger than the latter, the present discriminant is retained; otherwise it is replaced by the other discriminant. In the latter case, another discriminant is "appropriately" selected for future comparisons. The process is continued ad infinitum.

6.1 The Adaptive Procedure and the Proof of the Theorem

It will be assumed that no element of the sequence $\{x_i\}$ is either correctly classifiable or incorrectly classifiable by all the elements of the set Γ . If such elements exist, they are taken out of consideration, because such elements do not contribute any information towards the selection of a discriminant but use up the information memory.

The adaptive procedure is broken down into three modes of operation and into six steps. The first mode extends from Step 1 through Step 3, the second mode consists of Step 4 and Step 5, and the third mode consists of Step 6.

We now give the adaptive procedure step by step and simultaneously the step by step proof of Theorem 4.

Mode 1

1. Initially, when no observations are available and when the information memory is empty, the discriminant y_0 is any arbitrary element of the set Γ .

Proof: This is obvious because in the absence of any available observations, all the elements of Γ are equally likely to be the best discriminant.

2. As long as $k \leq \lambda$, where λ is the number of units of available information memory, the new observation x_k is stored in the information memory and

$$y_k = \delta_I^{k*}(x_1, x_2, \dots, x_k).$$

Note: As mentioned in Point I of Section 5, we have, in addition to the λ units of information memory, memory units available to store two discriminants and two numbers, to assist us in the selection of a discriminant which makes the least number of errors in classifying patterns in the information memory.

Proof: As long as there is room for a new observation in the information memory, there is no need to destroy any of the old observations. Also, since all the old observations are preserved, there is no need to preserve any discriminant which may have been tested or selected in the past. The choice of y_k follows directly from Theorem 1.

3. For $k = \lambda$, after y_λ is selected, the need arises to make a place for a new observation. The course of action depends on the value of $\rho^0(x_1, x_2, \dots, x_\lambda)$, i. e., the minimum of the numbers of errors made by the discriminants in the set Γ , in classifying the patterns of the observations $x_1, x_2, \dots, x_\lambda$. If the value of $\rho^0(x_1, x_2, \dots, x_\lambda)$ is 0 or 1, the process goes into Mode 3 which consists of Step 6; otherwise the process goes to Mode 2 which begins with Step 4.

Mode 2

4. The first thing to do is to reorganize the information in the information

memory. An arbitrary x_i which is not correctly classifiable by the discriminant y_λ is removed from the information memory. The observations x_j , $j > i$, are moved one step upwards in the information memory. The last unit of the information memory is thus made empty. From henceforth, the content of this memory unit, at any stage k , will be called α_k . For $k = \lambda$, α_k is made equal to $\rho^0(x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_\lambda)$.

Note: 1 The only reason for eliminating an observation this way, for making room for the number α_k , is to show that in the process of making room for α_k the optimum decision function is not "degraded."

Note: 2 We recognize here that what we called a unit of information memory (viz. the amount of space required to store an n -dimensional vector i. e., a pattern, and its correct classification) is a lot more storage space than required to store the number $\rho^0(x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_\lambda)$. One could use the remaining space to store some other information. However, for the sake of keeping the procedure simple we will not use the remaining fraction of the unit of information memory, for storing information.

At any future stage, a newly arriving observation is stored at the end of the sequence of observations in the information memory. This place at the end of the sequence is created by destroying the observation at the head of the sequence and moving the rest of the sequence up by one step. We shall denote the sequence of observations in the information memory at the k th-stage by S_k . The discriminant $y_{\lambda+1}$ is selected by $y_{\lambda+1} = \delta_{II}^{\lambda-1/\alpha_\lambda^*}(S_{\lambda+1})$, i. e.,

$$\begin{aligned}
 y_{\lambda+1} &= y_\lambda \text{ if } \alpha_\lambda < \rho^0(S_{\lambda+1}) \\
 &= c^0(S_{\lambda+1}) \text{ if } \alpha_\lambda > \rho^0(S_{\lambda+1}) \\
 &= y_\lambda \text{ with probability } \gamma \\
 &= c^0(S_{\lambda+1}) \text{ with probability } (1-\gamma)
 \end{aligned}
 \left. \vphantom{\begin{aligned} y_{\lambda+1} \\ &= c^0(S_{\lambda+1}) \\ &= y_\lambda \\ &= c^0(S_{\lambda+1}) \end{aligned}} \right\} \text{ if } \rho^0(S_{\lambda+1}) = \alpha_\lambda$$

where $0 \leq \gamma \leq 1$.

Proof: The decision function used at the λ th stage was $\delta_I^{\lambda/\alpha_{\lambda+1}^*}$, which is as good as the decision function $\delta_I^{\lambda-1/\alpha_{\lambda}^*}$, according to Theorem 3. The decision function anticipated to be used at $\lambda + 1$ stage is $\delta_{II}^{\lambda-1/\alpha_{\lambda}^*}$, which by Corollary 2 of Theorem 2 is better than $\delta_I^{\lambda-1/\alpha_{\lambda}^*}$, which was the decision function actually used at the λ th stage.

At the stage $\lambda + 1$, the decision function $\delta_{II}^{\lambda-1/\alpha_{\lambda}^*}$ operates on a new sample of $\lambda - 1$ observations which is formed from the old $\lambda - 1$ observations and newly arriving observations, by destroying an arbitrary observation from the $\lambda - 1$ old observations. We choose to destroy the oldest one.

5. At any future stage $k \geq \lambda + 1$, after selecting the discriminant y_k , if $\rho^0(S_k) = 0$, the process goes to the third mode and Step 6; otherwise the process runs as follows:

$$\begin{aligned} \alpha_k &= \alpha_{k-1} & \text{if } \alpha_{k-1} \leq \rho^0(S_k) \\ &= \rho^0(S_k) & \text{otherwise.} \end{aligned}$$

The new discriminant y_{k+1} selected at the $(k+1)$ st-stage (i. e. , after the newly arrived observation is stored at the end of the sequence) is

$$y_{k+1} = \delta_{II}^{\lambda-1/\alpha_k^*}(S_{k+1}).$$

After that Step 5 is repeated.

Proof: If $\rho^0(S_k) \geq \alpha_{k-1}$, then the decision function actually used at the stage k is $\delta_I^{\alpha-1/\alpha_{k-1}^*}$.

With $\alpha_k = \alpha_{k-1}$, the decision function that is anticipated to be used at the $(k+1)$ st-stage is $\delta_{II}^{\alpha-1/\alpha_k^*}$ which is better than the decision function $\delta_I^{\alpha-1/\alpha_{k-1}^*}$, according to Corollary 2 of Theorem 2.

If $\rho^0(S_k) \neq 0$ and $\rho^0(S_k) < \alpha_{k-1}$, and $\alpha_k = \rho^0(S_k)$, then the decision function that is anticipated to be used at $(k+1)$ st-stage is $\delta_{II}^{\alpha-1/\rho^0(S_k)^*}$, which is better than the decision function $\delta_I^{\alpha-1/\rho^0(S_k)^*}$ which was actually used at the k th-stage.

If $\rho^0(S_k) = 0$, there is a perfectly separable sample (i. e. , a sample of observation for which the least number of errors a discriminant can make in classifying its

patterns is zero), the third mode of operation is followed. The proof for Step 5 is complete.

Note: (i) Consistant with our policy of treating the case where there is no geometrical relation between the elements of the set Γ , we are giving the third mode of the adaptive procedure which does not take advantage of the geometrical relation, should it exist. It must be pointed out, however, that special procedures which take such geometrical relations into account turn out to be much more efficient.

(ii) In the third mode it is not necessary at every stage to find a discriminant which makes the least number of errors in classifying patterns in the memory. As such the storage space of two discriminants and two numbers (mentioned in the note in Step 2) reserved for that purpose is free to be used in the intermediate stages. We shall use the storage spaces for one of the discriminants and one of the numbers. The content of the storage space for the discriminant will be called z_k at the kth stage, and the content of the storage space for the number will be called γ_k .

6. In coming from Step 3, make room in the information memory for a number whose value at any future stage k will be called α_k . This is done by removing the observation in the information memory whose pattern is incorrectly classified by the discriminant y_λ , and moving the subsequent other observations in the sequences one step upwards in the information memory, if $\rho^0(x_1, x_2, \dots, x_\lambda) = 1$. If $\rho^0(x_1, x_2, \dots, x_\lambda) = 0$, the spaces for α_k can be made by removing any observation arbitrarily, and moving the subsequent observations one step upwards. For $k = \lambda$, α_k is made equal to $\lambda - 1$ if $\rho^0(x_1, x_2, \dots, x_\lambda) = 1$, and equal to λ if $\rho^0(x_1, x_2, \dots, x_\lambda) = 0$. In otherwords, α_k will represent the number of patterns correctly classified by the discriminant y_k . z_λ is made equal to y_λ , and γ_λ is made equal to α_λ .

If coming from Step 5, after, say k_0 th-stage, α_{k_0} is set equal to $\lambda - 1$. z_{k_0} is made equal to y_{k_0} and γ_{k_0} is made equal to α_{k_0}

At any future stage $k > \lambda$, if coming from Step 3, and $k > k_0$ if coming from Step 5, the following procedure is followed:

$$\begin{aligned}
y_{k+1} &= y_k \quad \text{if } \gamma_k < \alpha_k \\
&= y_k \quad \text{if } \gamma_k = \alpha_k \text{ and the pattern in the new observation} \\
&\quad \text{is not correctly classifiable by } z_k \\
&= z_k \quad \text{if } \gamma_k = \alpha_k \text{ and the pattern in the new observation} \\
&\quad \text{is correctly classifiable by } z_k.
\end{aligned}$$

If the pattern in the new observation is correctly classifiable by z_k , then

$$\gamma_{k+1} = \gamma_k + 1$$

and

$$z_{k+1} = z_k .$$

Also, if $\alpha_k = \gamma_k$, then $\alpha_{k+1} = \gamma_{k+1}$.

If the pattern in the new observation is incorrectly classified by z_k , then

$$\begin{aligned}
\alpha_{k+1} &= \alpha_k \\
z_{k+1} &= c^0(S_k) \\
\gamma_{k+1} &= \lambda - 1 - \rho^0(S_k) .
\end{aligned}$$

Remark: z_k is the discriminant which we previously referred to as the "appropriately" selected discriminant. γ_k is the number of "successive" patterns correctly classified by z_k up to the stage k , and α_k is the number of "successive" patterns correctly classified by y_k . In Mode 3, y_{k+1} is made equal to one of the two discriminants y_k and z_k depending on whether the number of "successive" patterns correctly classified by y_k or z_k is larger.

Proof: We notice that

$$\begin{aligned}
y_k &= \delta_I^{\alpha_k/0^*} \binom{\alpha_k}{S^k} \\
z_k &= \delta_I^{\gamma_k/0^*} \binom{\gamma_k}{S^k} .
\end{aligned}$$

The decision function used in Mode 3 is of the type mentioned in Corollary 3 of Theorem 2. Here at the $(k+1)$ st-stage the choice is between the values of two types of decision functions; one is $\delta_I^{\alpha_k/0^*}$ and the other is equivalent to $\delta_{II}^{\gamma_{k+1}/1^*}$. If $\alpha_k = \gamma_k$, then the anticipated decision function is equivalent to $\delta_{II}^{\gamma_{k+1}/1^*}$ which, by Corollaries 2 and 3 of Theorem 2, is better than $\delta_I^{\alpha_k/0^*}$ which was the decision function used in the k th stage. If $\alpha_k > \gamma_k$, then the decision function at the $(k+1)$ st-stage is equivalent to $\delta_I^{\alpha_k/0^*}$ which is the same as the one used at the k th stage. Because there is not enough information the anticipated decision function at the $(k+1)$ st-stage cannot be made better than the one used at the k th-stage in the case $\gamma_k < \alpha_k$. However, the anticipated decision function at the $(k+i)$ th-stage, where $i = \alpha_k - \gamma_k$, is again of the type mentioned in Corollary 3 of Theorem 2. This time the choice is between the values of two decision functions $\delta_I^{\alpha_k/0^*}$ and a decision function which is equivalent to $\delta_{II}^{\gamma_{k+i}/i^*}$. This decision function at $(k+i)$ th-stage is better than $\delta_I^{\alpha_k/0^*}$. This completes the proof of Theorem 4.

7. ILLUSTRATIVE EXAMPLE

A pattern recognizer, which uses the adaptive procedure given in the last section, was simulated on the IBM 7090 digital computer to work out a problem of signal detection in the presence of additive noise.

The usual problem of signal detection is the following: Given a waveform $x(t)$, it is required to determine whether it is pure noise $n(t)$, or signal-plus-noise, $s(t) + n(t)$, for $0 \leq t \leq T$. It is well known (Ref. 12) that if the signal $s(t)$, with energy E , is known exactly, and if the noise is white Gaussian noise with power per bandwidth known to be equal to N_0 , and if the probability $P(\text{SN})$ of signal occurring with noise is known, then the decision rule which minimizes the average "cost" is the following: If $\frac{2}{N_0} \int_0^T x(t) s(t) dt \geq c$, then the hypothesis that $x(t)$ contains the signal $s(t)$ is accepted; otherwise the alternate hypothesis that $x(t)$ is pure noise is accepted. The value of the threshold level c is given by

$$c = \frac{E}{N_0} + \ln \beta ,$$

where β depends on $P(\text{SN})$, and on the cost. In this report the cost has been taken equal to unity if there is an incorrect classification, and zero if there is a correct classification.

For such a cost, β is given by

$$\beta = \frac{1 - P(\text{SN})}{P(\text{SN})} .$$

With known E , N_0 , and $P(\text{SN})$, the value of the optimum threshold c can be calculated.

For this illustrative example we are assuming that the knowledge about E , N_0 , and $P(\text{SN})$ is not available. It is the job of the pattern recognizer to select a threshold (discriminant), according to the adaptive procedure given in Section 6, from a given set Γ of thresholds. The set Γ actually used is the set of all 41 odd numbers from 1 through 81.

The input to the pattern recognizer is a sequence of observations. Each observation, for this example, is the value of the integral I and the information that the corresponding $x(t)$ is $s(t) + n(t)$ or $n(t)$. In the language of this report, I is the one-dimensional

pattern ω , and the information about $x(t)$ is the classification of the pattern. Thus we may say that the pattern ω represented by a value of I belongs to Class A if the corresponding $x(t)$ is $n(t)$; otherwise it belongs to Class B. The unknown probability density on the observation-space is given by

$$p(\omega, A) = P(SN) p[I | x(t) = n(t)]$$

$$p(\omega, B) = [1 - P(SN) p[I | x(t) = n(t) + s(t)]] .$$

For white Gaussian noise, the probability densities $p[I | x(t) = n(t)]$ and $p[I | x(t) = n(t) + s(t)]$ are normal, each having a variance equal to $\frac{2E}{N_0}$, and means differing by $\frac{2E}{N_0}$.

The simulated pattern recognizer has an information memory of 10 units, i. e., the pattern recognizer can store, at the most, 10 observations at a time.

For each set of values of $\frac{2E}{N_0}$ and $P(SN)$, twenty runs of the adaptive process were made. Each run was carried up to 100 stages. Five sets of values of $\frac{2E}{N_0}$ and $P(SN)$ were considered.

The (a) parts of the five graphs show the values of the losses (as defined in Section 3) for the 41 discriminants. The number with the arrow, in each graph, shows the best discriminant. In some cases there are two "best" discriminants. The (b) part of each of the five graphs shows the average over twenty runs of the loss of the discriminant selected at each stage. These graphs clearly show a decreasing trend in the average value of the loss of the discriminant as the number of the stage increases.

This decreasing trend indicates the fact that the pattern recognizer is learning. The learning is very rapid in the beginning. However, the graphs clearly show that the process of learning continues even after the tenth stage. This is especially true when the signal-to-noise ratio is small. This result clearly indicates that the adaptive procedure leads to a better discriminant than the one that would be selected on the basis of the first ten observations. The latter would have been selected if the observations beyond the first ten were ignored because the pattern recognizer has memory to store ten observations at a time.

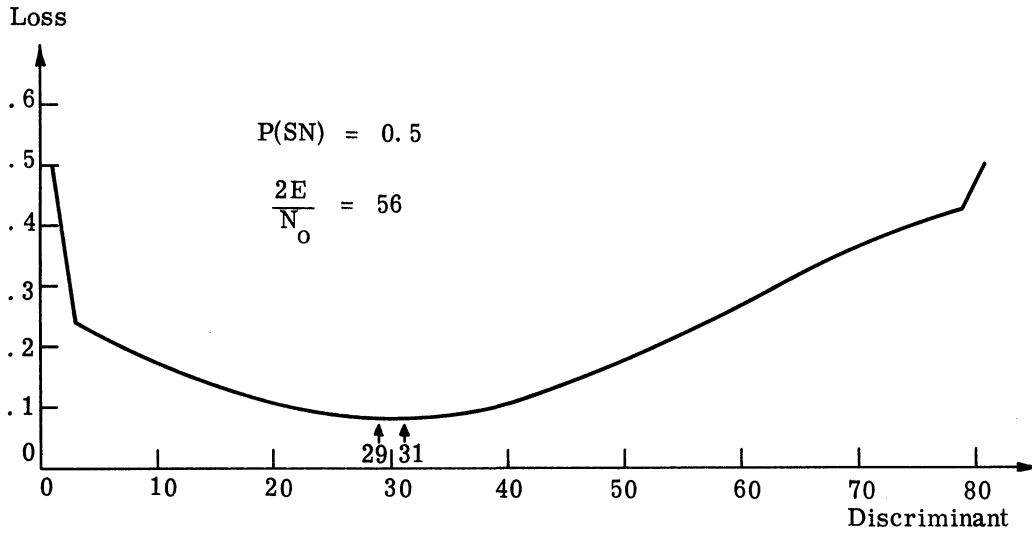


Fig. 1(a). Loss vs. Discriminant (odd numbers) used in the illustrative example of signal detection.

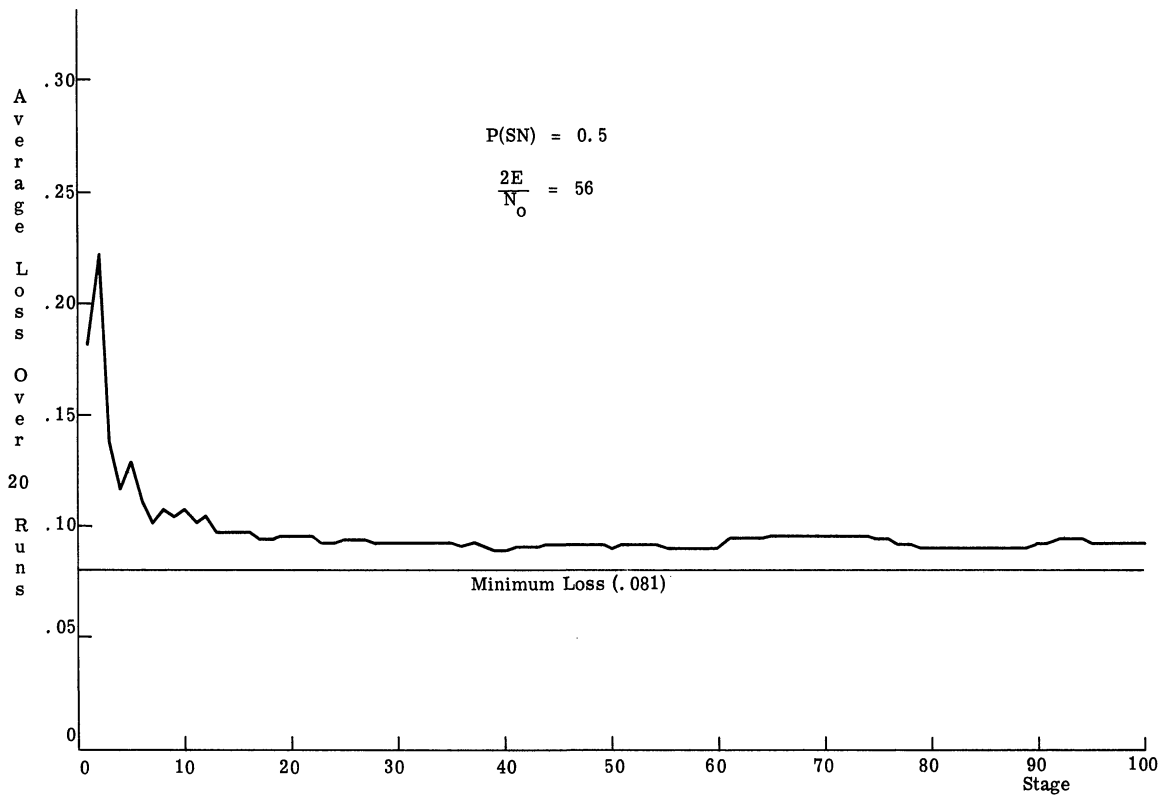


Fig. 1(b). Average loss over twenty runs at each stage of the adaptive process.

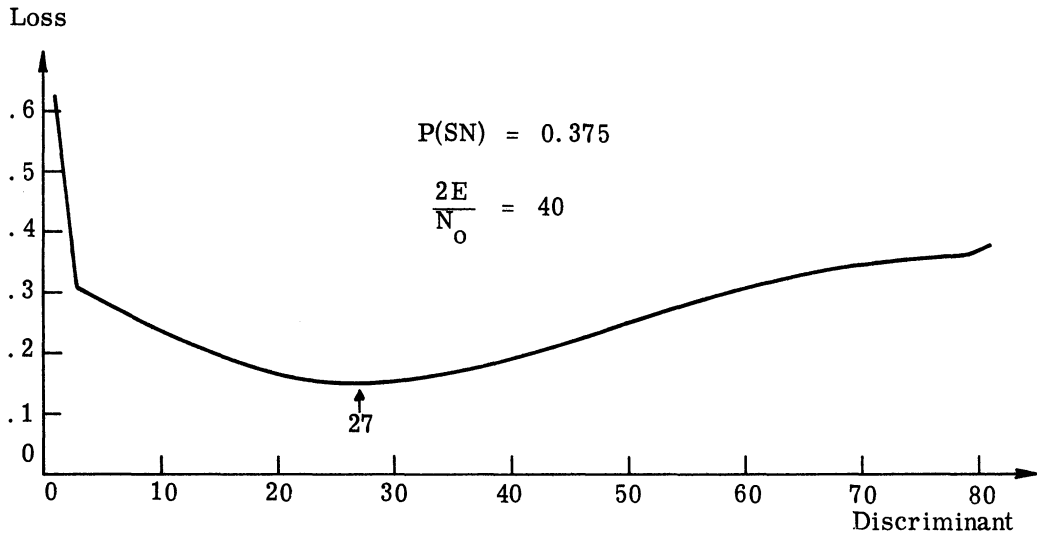


Fig. 2(a). Loss vs. Discriminant (odd numbers) used in the illustrative example of signal detection.

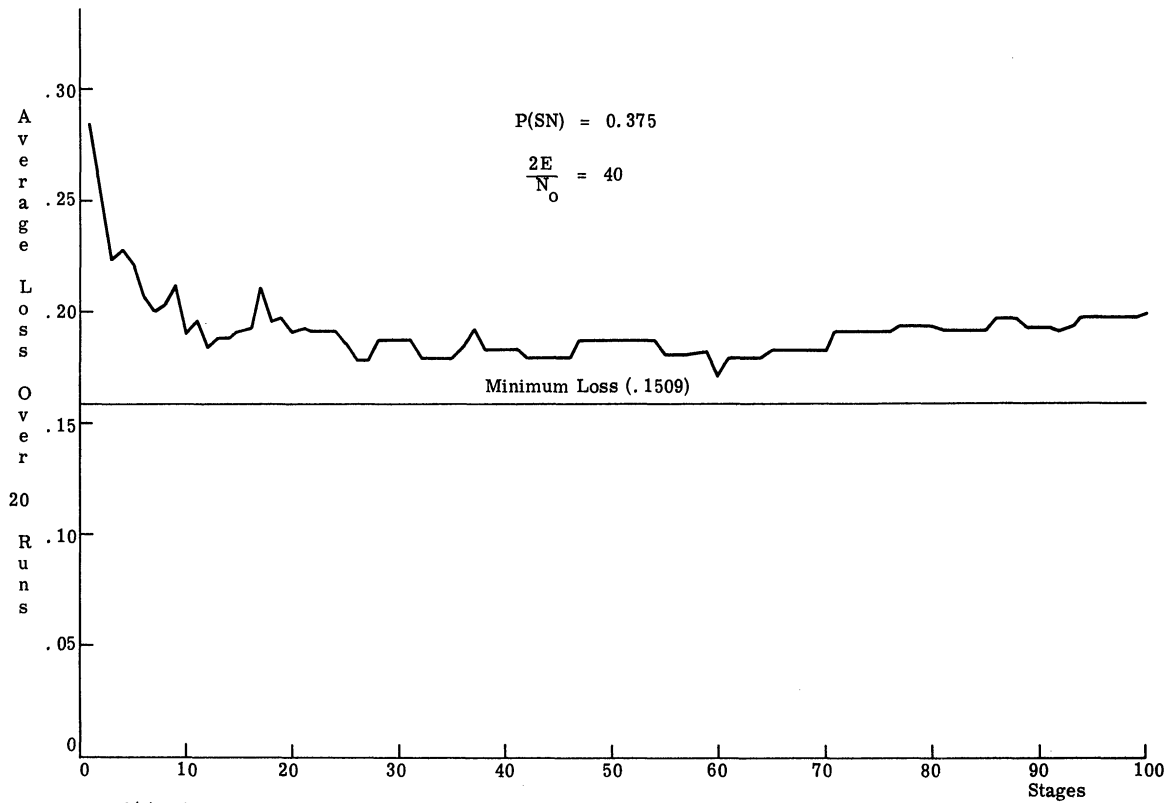


Fig. 2(b). Average loss over twenty runs at each stage of the adaptive process.

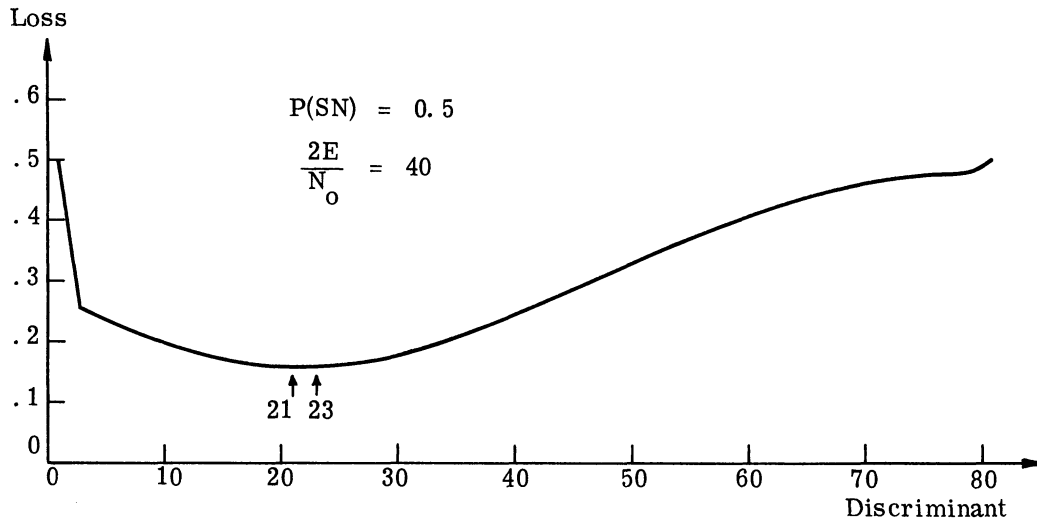


Fig. 3(a). Loss vs. Discriminant (odd numbers) used in the illustrative example of signal detection.

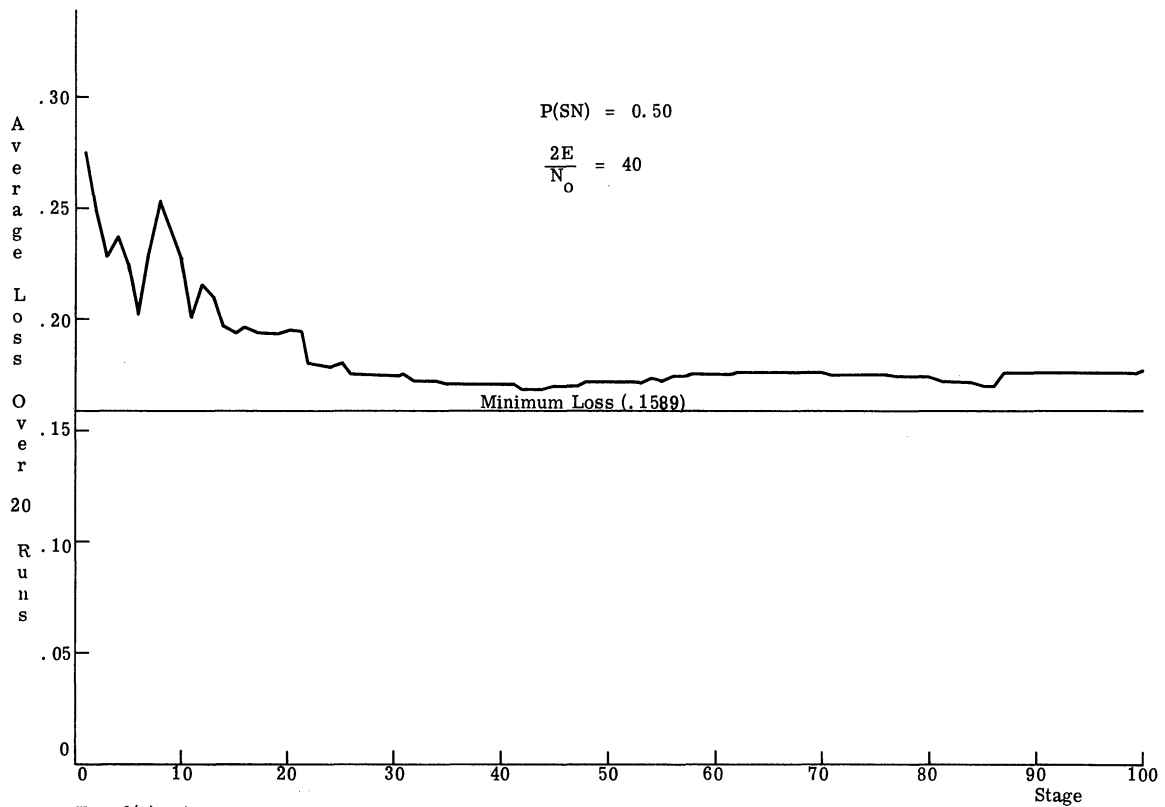


Fig. 3(b). Average loss over twenty runs at each stage of the adaptive process.

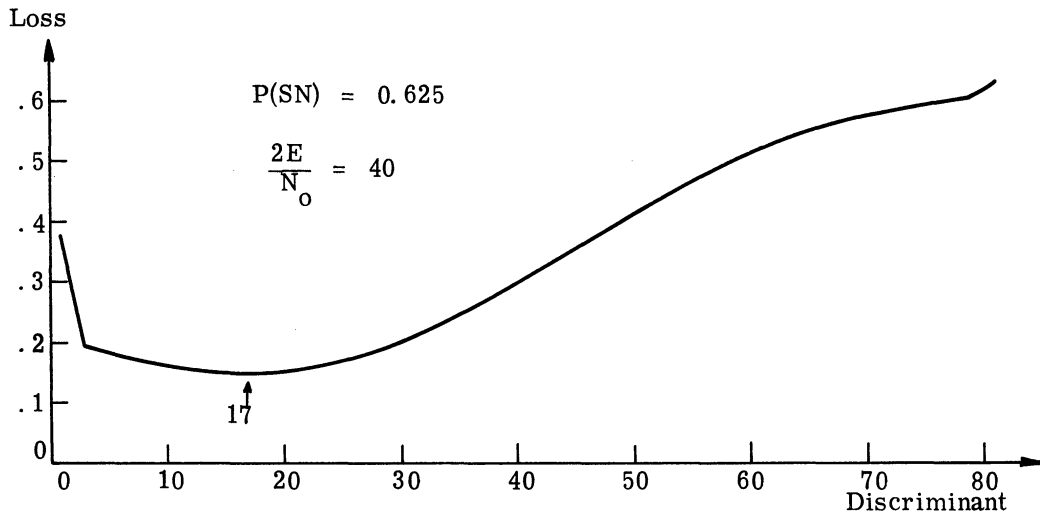


Fig. 4(a). Loss vs. Discriminant (odd numbers) used in the illustrative example of signal detection.

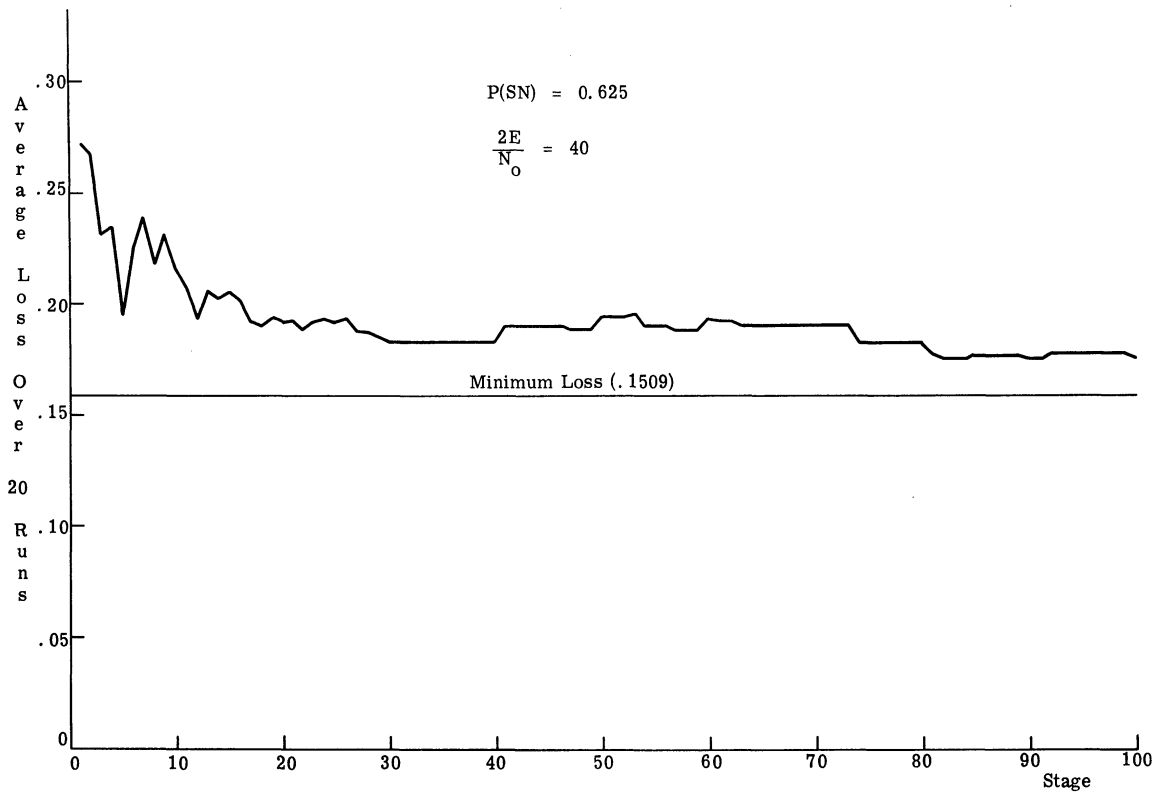


Fig. 4(b). Average loss over twenty runs at each stage of the adaptive process.

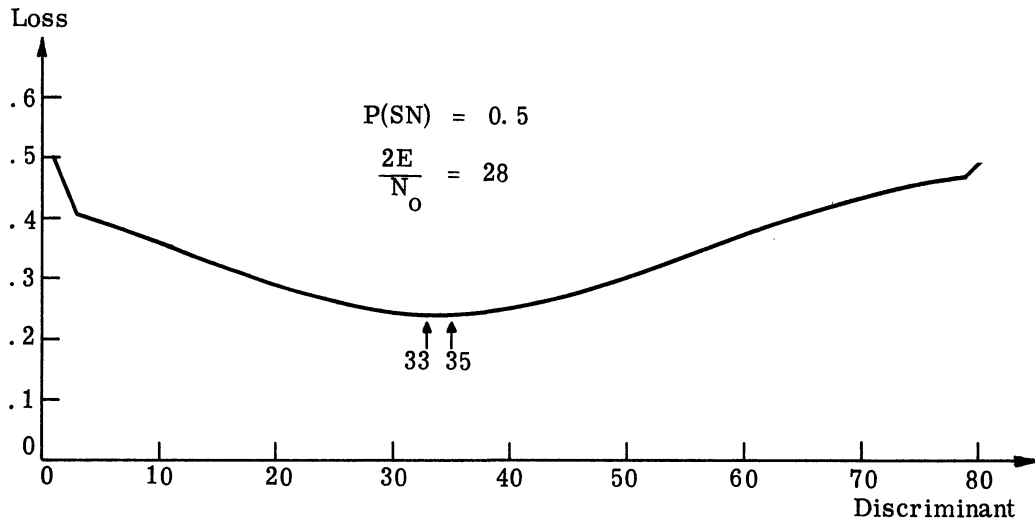


Fig. 5(a). Loss vs. Discriminant (odd numbers) used in the illustrative example of signal detection.

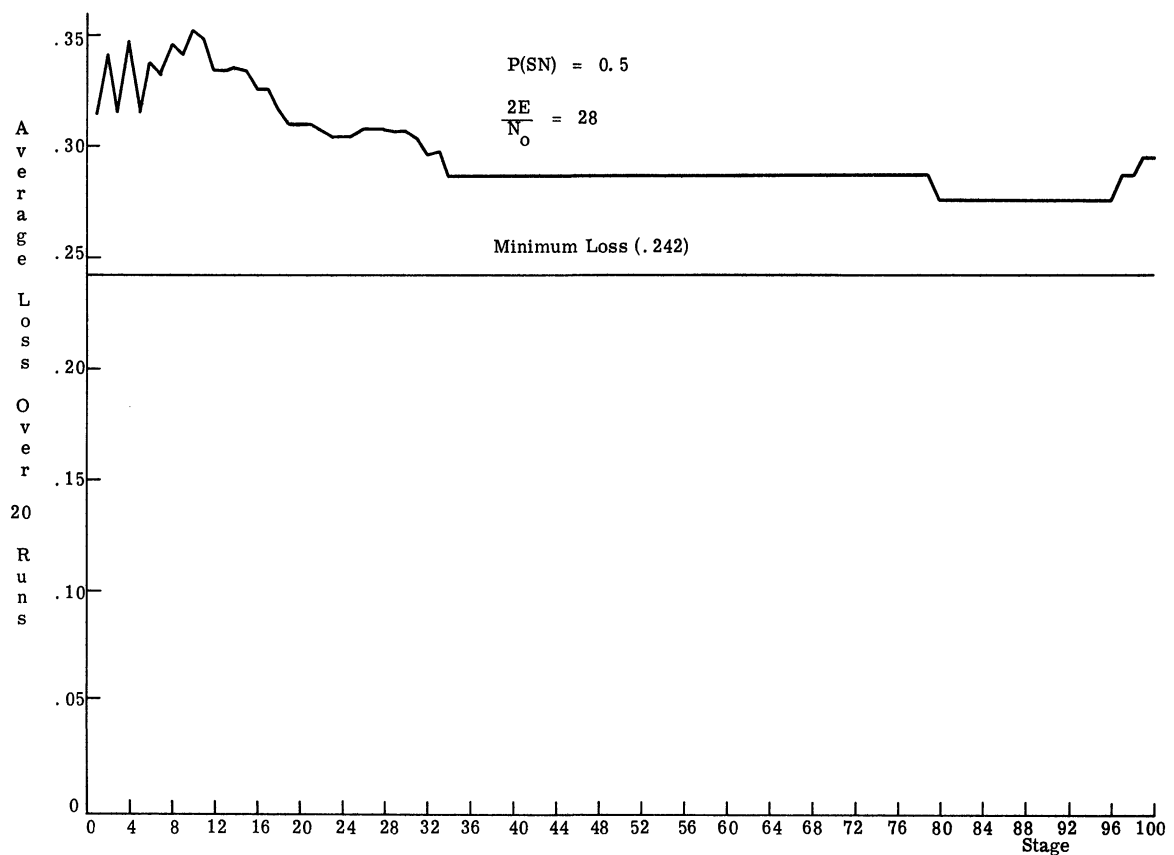


Fig. 5(b). Average loss over twenty runs at even stage of the adaptive process.

8. CONCLUSIONS

An adaptive procedure is given whereby a pattern recognizer continually "learns" to recognize patterns. The learning is demonstrated by the continual "improvement" of the discriminant selected for recognizing unknown patterns, as long as new observations (patterns and their respective correct classifications) are made available to the pattern recognizer. No a priori knowledge of the probability density on the observation-space is assumed. Moreover the pattern recognizer is assumed to have a finite-memory. The adaptive procedure not only makes efficient use of the finite memory in storing past information, but uses "optimum" decision rules to select a discriminant on the basis of the stored information. At any stage of the process, a discriminant is selected from a prescribed set of discriminants. The factors affecting the choice of this prescribed set of discriminants have not been discussed.

REFERENCES

1. Hu, S. T. , "On the Willis Method of Finding Separating Systems," TR 6-90-61-44, Lockheed Missiles and Space Division, December 1960.
2. Hu, S. T. , "The Synthesis Problem of Linear Separability," TR 6-90-62-17, Lockheed Missiles and Space Division, September 1961.
3. Hu, S. T. , "An Elimination Process for Willis Synthesis Method," TR 6-90-62-18, Lockheed Missiles and Space Division, December 1961.
4. Hu, S. T. , "Successive Approximation Applied to the Synthesis of Linear Separability," TR 6-90-62-99, Lockheed Missiles and Space Division, January 1962.
5. Singleton, R. C. , "A Test for Linear Separability as Applied to Self-Organizing Machines," Stanford Research Institute, Project No. 3605, May 1962.
6. Highleyman, W. H. , "Linear Decision Functions, with Application to Pattern Recognition," Proceedings of the IRE, June 1962.
7. Cooper, P. W. , "The Hyperplane in Pattern Recognition," Sylvania Applied Research Laboratory, 1962.
8. Sebestyen, G. S. , "Decision-Making Processes in Pattern Recognition," ACM Monograph Series, McMillan Company, 1962.
9. Cooper, P. W. , "The Hypersphere in Pattern Recognition," Information and Control, Vol. 5, No. 4, December 1962.
10. Cooper, P. W. , "Classification by Statistical Methods," Melpar Technical Note, April 1961.
11. Chow, C. K. , "An Optimum Character Recognition System Using Decision Functions," IRE Transactions on Electronic Computers, December 1957.
12. Peterson, W. W. , Birdsall, T. G. , "The Theory of Signal Detectability," Part I "Applications with Gaussian Noise," Part II, Cooley Electronics Laboratory Technical Report No. 13, June 1953, July 1953.
13. Widrow, B. , Hoff, M. E. , "Adaptive Switching Circuits," Stanford Electronic Laboratory Technical Report No. 1533-1, June 1960.
14. Rosenblatt, F. , "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms," Spartan Books, 1961.
15. Novikoff, A. B. , "On Convergence Proofs for Perceptrons," Stanford Research Institute Technical Report No. 3605, January 1963.
16. Albert, A. , "A Mathematical Theory of Pattern Recognition," The Annals of Mathematical Statistics, Vol. 34, No. 1, March 1963.

DISTRIBUTION LIST

<u>No. of Copies</u>		<u>No. of Copies</u>	
2	Commanding Officer U. S. Army Electronics Command U. S. Army Electronics Laboratories Fort Monmouth, New Jersey ATTN: Senior Scientist Electronic Warfare Division	1	Headquarters USAF Washington 25, D. C. ATTN: AFRDR
		1	AFAL (AVWW/ECM Technology) Wright-Patterson Air Force Base Ohio 45433
1	Commanding General U. S. Army Electronic Proving Ground Fort Huachuca, Arizona ATTN: Director Electronic Warfare Division	1	Commander Aeronautical Systems Division Wright-Patterson Air Force Base Ohio ATTN: ASAPRD
1	Commanding General U. S. Army Materiel Command Bldg. T-7 Washington 25, D. C. ATTN: AMCRD-DE-E	1	Commander Aeronautical Systems Division Wright-Patterson Air Force Base Ohio ATTN: ASNP
1	Commanding General U. S. Army Materiel Command Bldg. T-7 Washington 25, D. C. ATTN: AMCRD-RP-E	1	ESD (ESTI) L. G. Hanscom Field Bedford, Massachusetts
1	Commanding Officer Signal Corps Electronics Research Unit 9560th USASRU P. O. Box 205 Mountain View, California	1	Commander Rome Air Development Center Griffiss Air Force Base New York ATTN: RAYLD
1	U. S. Atomic Energy Commission 1901 Constitution Avenue, N. W. Washington 25, D. C. ATTN: Chief Librarian	1	Commander Air Proving Ground Center Eglin Air Force Base Florida ATTN: ADJ/Technical Report Branch
1	Director Central Intelligence Agency 2430 E Street, N. W. Washington 25, D. C.	1	Chief of Naval Operations EW Systems Branch OP-35, Department of the Navy Washington 25, D. C.
1	U. S. Army Research Liaison Officer MIT-Lincoln Laboratory Lexington 73, Massachusetts	1	Chief, Bureau of Ships Code 691C Department of the Navy Washington 25, D. C.
1	Commander Air Force Systems Command Andrews Air Force Base New Jersey		

DISTRIBUTION LIST (Cont.)

<u>No. of Copies</u>		<u>No. of Copies</u>	
1	Commander Bu Naval Weapons Code RRRE-20 Department of the Navy Washington 25, D. C.	1	President U. S. Army Airborne and Electronics Board Fort Bragg, North Carolina
1	Commander Naval Ordnance Test Station Inyokern China Lake, California ATTN: Test Director - Code 20	1	U. S. Army Anti-Aircraft Artillery and Guided Missile School Fort Bliss, Texas ATTN: ORL
1	Commander Naval Air Missile Test Center Point Mugu, California	1	Commander USAF Security Service San Antonio, Texas ATTN: CLR
1	Director Naval Research Laboratory Countermeasures Branch, Code 5430 Washington 25, D. C.	1	Chief of Naval Research Department of the Navy Washington 25, D. C. ATTN: Code 427
1	Director Naval Research Laboratory Washington 25, D. C. ATTN: Code 2021	1	Commanding Officer 52d U. S. Army Security Agency Special Operations Command Fort Huachuca, Arizona
1	Director Air University Library Maxwell Air Force Base Alabama ATTN: CR-4987	1	President U. S. Army Security Agency Board Arlington Hall Station Arlington 12, Virginia
1	Commanding Officer-Director U. S. Navy Electronics Laboratory San Diego 52, California	1	The Research Analysis Corporation McLean, Virginia 22101 ATTN: Document Control Officer
1	Commanding Officer U. S. Naval Ordnance Laboratory Silver Spring 19, Maryland	10	Headquarters Defense Documentation Center Cameron Station Alexandria, Virginia
3	Chief U. S. Army Security Agency Arlington Hall Station Arlington 12, Virginia 22212 ATTN: 2 Cyps - IADEV 1 Copy - EW Div. IATOP	1	Commanding Officer U. S. Army Electronics Research and Development Laboratory Fort Monmouth, New Jersey ATTN: U. S. Marine Corps Liaison Office, Code: SIGRA/SL-LNR
1	President U. S. Army Defense Board Headquarters Fort Bliss, Texas	1	Director Fort Monmouth Office Communications-Electronics Combat Developments Agency Building 410 Fort Monmouth, New Jersey

DISTRIBUTION LIST (Cont.)

<u>No. of Copies</u>		<u>No. of Copies</u>	
15	Commanding Officer U. S. Army Electronics Command U. S. Army Electronics Laboratories Fort Monmouth, New Jersey ATTN: AMSEL-RD-DR AMSEL-RD-NSR AMSEL-RD-SM AMSEL-RD-SA AMSEL-RD-SEA AMSEL-RD-SEJ AMSEL-RD-SES AMSEL-RD-SEE AMSEL-RD-ADO AMSEL-RD-SR AMSEL-RD-SE AMSEL-RD-ADT AMSEL-RD-GFR AMSEL-RD-PRM AMSEL-RD-RHA	1	U. S. A. F. Project Rand The Rand Corporation 1700 Main Street Santa Monica, California
		1	Stanford Electronics Laboratories Stanford University Stanford, California
		1	Director National Security Agency Fort George G. Meade, Maryland ATTN: RADE-1
		1	Bureau of Naval Weapons Representative Lockheed Missiles and Space Company P. O. Box 504 Sunnyvale, California
1	Commanding Officer U. S. Army Signal Missile Support Agency White Sands Missile Range White Sands, New Mexico ATTN: SIGWS-MEW	1	Dr. T. W. Butler, Jr., Director Cooley Electronics Laboratory The University of Michigan Ann Arbor, Michigan
		11	Cooley Electronics Laboratory The University of Michigan Ann Arbor, Michigan
1	Commanding Officer U. S. Naval Air Development Center Johnsville, Pennsylvania ATTN: Naval Air Development Center Library		

Above distribution is effected by Electronic Warfare Division, Surveillance Department, USAEL, Evans Area, Belmar, New Jersey. For further information contact Mr. I. O. Myers, Senior Scientist, Telephone 59-61252.

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) The University of Michigan		2 a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	
		2 b. GROUP - - - - -	
3. REPORT TITLE A Finite-Memory Adaptive Pattern Recognizer			
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) Technical Report No. 168			
5. AUTHOR(S) (Last name, first name, initial) Irani, K. B.			
6. REPORT DATE September 1965		7 a. TOTAL NO. OF PAGES 52	7 b. NO. OF REFS 16
8 a. CONTRACT OR GRANT NO. DA 36-039-AMC-03733(E)		9 a. ORIGINATOR'S REPORT NUMBER(S) 6137-14-T	
b. PROJECT NO. 1P0 21101 A042-01-02		9 b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) - - - - -	
c. 			
d. 			
10. AVAILABILITY/LIMITATION NOTICES Qualified requesters may obtain copies of this report from DDC Foreign announcement and dissemination by DDC is limited.			
11. SUPPLEMENTARY NOTES - - - - -		12. SPONSORING MILITARY ACTIVITY U. S. Army Electronics Command ATTN: AMSEL-WL-S Fort Monmouth, N. J.	
13. ABSTRACT This report gives an adaptive procedure for selecting a discriminant for a pattern recognizer. No a priori knowledge of the probability density on the observation-space is assumed. Moreover the pattern recognizer is assumed to have a finite memory. A mathematical model of the problem of pattern recognition is constructed and several theorems are proved. With the help of these theorems, the adaptive procedure is developed. This adaptive procedure is, in effect, a method of using the finite memory efficiently in "training" the pattern recognizer.			

14. KEY WORDS Pattern recognition Adaptive procedure	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT

INSTRUCTIONS

1. **ORIGINATING ACTIVITY:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization (*corporate author*) issuing the report.

2a. **REPORT SECURITY CLASSIFICATION:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. **GROUP:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. **REPORT TITLE:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. **DESCRIPTIVE NOTES:** If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. **AUTHOR(S):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. **REPORT DATE:** Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. **TOTAL NUMBER OF PAGES:** The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. **NUMBER OF REFERENCES:** Enter the total number of references cited in the report.

8a. **CONTRACT OR GRANT NUMBER:** If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. **PROJECT NUMBER:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. **ORIGINATOR'S REPORT NUMBER(S):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. **OTHER REPORT NUMBER(S):** If the report has been assigned any other report numbers (*either by the originator or by the sponsor*), also enter this number(s).

10. **AVAILABILITY/LIMITATION NOTICES:** Enter any limitations on further dissemination of the report, other than those

imposed by security classification, using standard statements such as:

- (1) "Qualified requesters may obtain copies of this report from DDC."
- (2) "Foreign announcement and dissemination of this report by DDC is not authorized."
- (3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."
- (4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."
- (5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. **SUPPLEMENTARY NOTES:** Use for additional explanatory notes.

12. **SPONSORING MILITARY ACTIVITY:** Enter the name of the departmental project office or laboratory sponsoring (*paying for*) the research and development. Include address.

13. **ABSTRACT:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. **KEY WORDS:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical content. The assignment of links, rules, and weights is optional.