Linear Decision Functions, with Application to Pattern Recognition*

W. H. HIGHLEYMAN†, MEMBER IRE

Summary—Many pattern recognition machines may be considered to consist of two principal parts, a receptor and a categorizer. The receptor makes certain measurements on the unknown pattern to be recognized; the categorizer determines from these measurements the particular allowable pattern class to which the unknown pattern belongs. This paper is concerned with the study of a particular class of categorizers, the linear decision function. The optimum linear decision function is the best linear approximation to the optimum decision function in the following sense:

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

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

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

Various properties of linear decision functions are discussed. One such property is that a linear decision function is guaranteed to perform at least as well as a minimum distance categorizer. Procedures are then developed for the estimation (or design) of the optimum linear decision function based upon an appropriate sampling from the pattern classes to be categorized. Finally, the concepts and procedures thus developed are applied for illustrative purposes to the recognition of hand-printed numbers.

Introduction

THIS PAPER[1] is concerned with the practical design of a class of pattern recognition machines which is of interest for two reasons:

1) There is no need to make assumptions about the probability distributions of the various measurements made by the recognition machine.
2) This class of machines is amenable to an economic implementation.

Marill and Green [16] have described the general pattern recognition system in a very clear manner. They note that it consists of two principal parts, a *receptor* and a *categorizer:*

1) "The receptor has as its input a physical sample to be recognized, and as its output a set . . . of quantities which characterize the physical sample. These quantities will be called *measurements* of the sample . . . "
2) "The output . . . of the receptor constitutes the input to the categorizer. The categorizer is a device which assigns each of its . . . inputs to one of a finite number . . . of categories . . . "

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

The categorizer must apply some sort of decision criterion to the receptor output to decide to which of the allowable pattern classes, if any, the input pattern belongs. Or the categorizer may reject the pattern as being unrecognizable if the recognition decision is unreliable in some sense. If the machine attempts a recognition and is wrong, then it is said that the machine has made an *error.* Note that a rejection will not be considered as an error.

The Decision Theory Model of Pattern Recognition [5], [19]

Let the p allowable pattern classes be denoted s_i, $1 \leq i \leq p$, each having *a priori* probability of occurrence ω_i, where

$$\sum_{i=1}^{p} \omega_i = 1.$$

When an unknown pattern is presented to the receptor, the receptor makes certain measurements, n in number, upon it. The receptor output for a particular input pattern is the set of numbers $(m_1, m_2, \cdots, m_n) = m$. This set defines the coordinates of the point representing the input pattern in an n-dimensional *measurement space* M.

We assume the existence of a probability function (or density) over M, $\beta(M|S)$. Thus $\beta(m|s_i)$ is the conditional probability that a certain measurement m will be made, given a pattern from class i at the receptor.

* Received November 7, 1961; revised manuscript received, March 2, 1962. This paper is a summary of a doctoral dissertation submitted to the Polytechnic Institute of Brooklyn, N. Y.

† Bell Telephone Laboratories, Inc., Murray Hill, N. J.

[1] Since this paper is a summary of a Doctoral Dissertation [12], some of the topics and theorem proofs are omitted for the sake of brevity. Copies of the complete dissertation are available from University Microfilms, 313 N. First St., Ann Arbor, Mich.

Let there also exist a probability function (or density) $\delta(D|M)$, so that $\delta(d_j|m)$ is the probability that the categorizer will make the decision d_j, $0 \leq j \leq p$, given the measurement m. ($j=0$ corresponds to rejection; $1 \leq j \leq p$ corresponds to classification into one of the p pattern classes.) $\delta(D|M)$ is referred to as the *decision function* or *decision criterion. Note that the categorizer is nothing more than the implementation of the function* $\delta(D|M)$.

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

$$c_{ij} > c_{i0} > c_{ii},$$

where c_{i0} is the loss associated with rejection when the input state is s_i.

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

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

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

$$\bar{C}(s_i, \delta) = \sum_{j=0}^{p} c_{ij} \int_M \beta(m|s_i)\delta(d_j|m)dm. \quad (1)$$

Since the distribution of states is given by ω_i, $1 \leq i \leq p$, the expected loss for the pattern recognition system is

$$C(\delta) = \sum_{i=1}^{p} \sum_{j=0}^{p} c_{ij}\omega_i\beta(m|s_i)\delta(d_j|m)dm. \quad (2)$$

The optimum categorizer is defined as the implementation of that decision function δ^* which minimizes the expected loss $C(\delta)$ under the appropriate *a priori* distribution ω_i, $1 \leq i \leq p$, (Bayes strategy).

The general solution to this problem has been given by Chow [6]. He shows that (2) is minimized by using a certain nonrandomized decision criterion (*i.e.*, $\delta(D|m)$ is unity for one decision d_j, and zero for all others).

It has also been shown [10] that if all losses due to misrecognition are of value c, all losses due to rejection are of value c_0, and all losses due to correct recognition are zero, where

$$c > c_0 > 0;$$

then minimizing the loss is equivalent to minimizing the error rate for a given rejection rate.

Reference to Chow will show that the optimum decision function depends, aside from the loss function, only upon the quantities

$$\omega_i\beta(m|s_i), \qquad 1 < i \leq p.$$

Unfortunately, these probability functions, particularly $\beta(M|S)$, are usually unknown to the designer, and therefore categorizers based on the optimum decision function are not, in general, practically realizable. There are at least two ways around this difficulty:

1) Assume a certain form for the probability function $\beta(M|S)$. A common assumption is that of normality and independence: Given a certain pattern class, assume that the measurements made by the receptor are normally distributed, and that each measurement is independent of the others.
2) Make no assumptions about the particular distributions involved, but rather make certain restrictions on the structure of the categorizer. Then search through all possible structures of this type to find the categorizer which is optimum with respect to a sampling of patterns from the real world.

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

Linear Decision Functions

There is another practical advantage that is realized by the second approach, namely one of economic feasibility. Even if the optimum decision function were known, its implementation would require, in general, the use of a digital computer or other complex equipment. The cost of such equipment may, in many cases, outweigh the advantages of mechanized categorization. However, if the designer can limit his search to those structures which are economically feasible, and if the optimum structure in this class works well enough for the given purpose, then a technically feasible as well as an economically feasible solution has been found.

This paper is concerned with the study of just such a class of categorizers. To describe this class, consider a rephrasing of the optimum decision criterion. Note that every point in the measurement space M is preassigned to a particular pattern class or to the rejection class by the decision criterion, since it is nonrandomized. Thus, there is a subset M_j of M corresponding to each possible decision d_j, $0 \leq j \leq p$. Further, these subsets are nonoverlapping since the decision function is nonrandomized. The division of M into these subsets then uniquely identifies a certain decision function. We could equally well consider the decision function to be represented by the *boundaries* between the subsets. (Some liberty is taken here, since it will be assumed that a continuous boundary can be passed through a discrete space.) For instance, in Fig. 1 is shown a two-dimensional measurement space (the receptor makes only two measurements

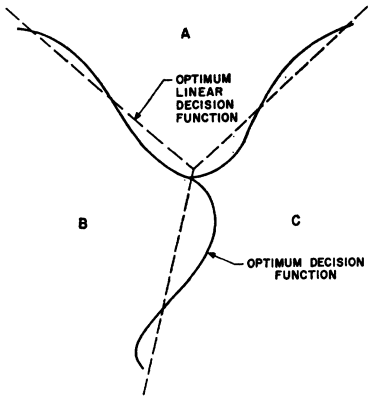

Fig. 1—Domains of three pattern classes in measurement space, as defined by optimum and optimum linear decision functions.

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

The class of categorizers to be discussed herein may be loosely described as the optimum linear approximation to the true boundaries, under the further constraint of only one boundary per pair of pattern classes (such as those shown dotted in Fig. 1). Optimum, as previously mentioned, is taken to mean minimum loss under the above constraints. Because of the linear properties of this decision criterion, a categorizer of this class will be said to implement a *linear decision function.* Although the primary purpose of the development is to study the synthesis of such a categorizer when the probability distributions are unknown, the problem of finding the optimum linear decision function when these distributions are known will also be discussed.

Implementation of a Linear Decision Function

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

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n + \alpha_0 = 0$$

for a given set of α_i's. The fact that the actual boundaries are only portions of hyperplanes, *i.e.*, each hyperplane usually terminates on other hyperplanes (Fig. 1), is of little consequence. As will be shown in the next section, the representation of each boundary by a full hyperplane is equivalent.

It will be shown later that, in order to classify a point m in M, it is only necessary to determine on which side of each hyperplane this point lies. This is determined by the sign of the quantity[2]

$$\sum_{i=1}^{n} \alpha_i m_i + \alpha_0. \tag{3}$$

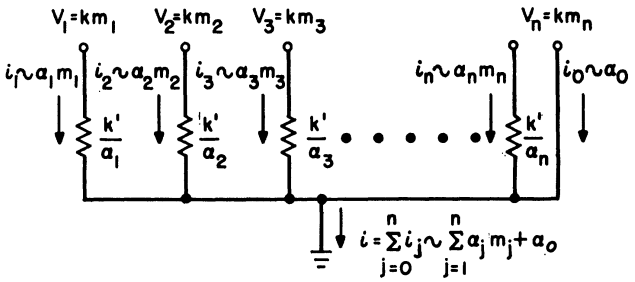

Fig. 2—Implementation of a hyperplane.

Consequently, in order to classify a point m (that is, recognize an input pattern), it is only necessary to evaluate a set of quantities like (3). But such a calculation can be done with several varieties of very inexpensive networks, such as the resistive adder shown in Fig. 2 (in which the voltages corresponding to the measurement values are inverted for negative α_i). This supports the statement of economy.

Some Properties of Linear Decision Functions

The Classifying Procedure

Before discussing some of the properties of linear decision functions, the classification procedure will first be discussed. Fig. 3 illustrates a measurement space in which the domains of three pattern classes are shown, as determined by a linear decision function. The boundaries, which are really truncated hyperplanes, will be represented by the complete hyperplanes as indicated by the dotted lines. It will be seen that the truncation is automatically taken into account by the classifying procedure. Since there is one and only one boundary per pair of pattern classes, Fig. 3 shows three boundaries separating the three classes. The boundary separating the ith and jth classes will be denoted B_{ij}. Further, in schematic representation as in Fig. 3, each hyperplane B_{ij} will be identified by the pair of numbers, i, j, placed in such a way as to show which side of B_{ij} corresponds to class i, and which to class j.

In order to classify a certain measurement, we note which side of each boundary it is on. If it is on the ith side of all boundaries B_{ik}, $1 \leq k \leq p$, $k \neq i$, then this pattern belongs to pattern class i. Using this criterion, point A in Fig. 3 is clearly identified as a member of pattern class 2.

Note that the point designated B cannot belong to

[2] Note that this linear form equated to zero defines a structure which is commonly found in the automata field. It goes by various names, such as artificial neuron [14], associative unit [21], [22], and Adaline [26], [27]. In this paper, it will simply be called by its already well-established name of "hyperplane."

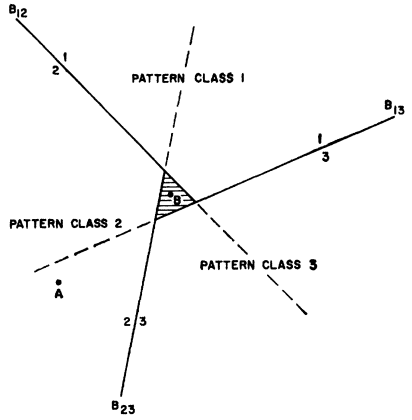

Fig. 3—Geometric representation of a linear decision function.

any of the three classes, and is hence rejected. This is not the normal sort of rejection due to an unreliable decision; rather it is a type of rejection inherent in a linear decision function.

One further comment is appropriate concerning the determination of which side of a hyperplane a point lies. Consider a hyperplane B represented by the set of points $\{x\}$ satisfying

$$\sum_{i=1}^{n} \alpha_i x_i + \alpha_0 = 0, \tag{4}$$

where

$$\sum_{i=1}^{n} \alpha_i^2 = 1. \tag{5}$$

The distance s from the hyperplane of a point with co-ordinates m_i, $1 \leq i \leq n$, is

$$s = \sum_{i=1}^{n} \alpha_i m_i + \alpha_0. \tag{6}$$

Hence, the distance of a point to the hyperplane (4) is simply given by substituting the coordinates of the point into the expression for the hyperplane (as in (6)), providing the expression is in a normalized form, that is, that (5) holds. The point is on one side of the hyperplane if (6) is positive, and on the other if (6) is negative. Which side of the hyperplane is to be positive or negative is completely arbitrary, since multiplication of (4) by -1 changes the sign of (6), but does not change the hyperplane.

Some Theorems Pertaining to Linear Decision Functions

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

Theorem 1: For any categorizer based upon minimizing a Euclidean distance[3] to a set of reference points, there exists a categorizer based on a linear decision function which is at least as good. This includes categorizers which maximize a normalized cross-correlation function, and those which minimize a Hamming distance.

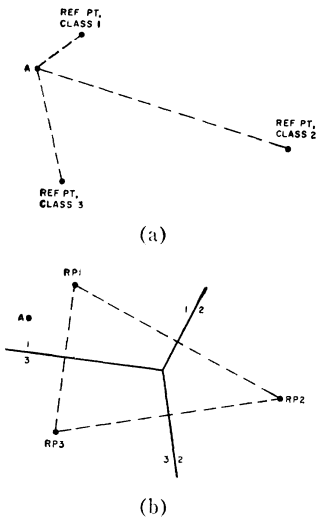

Fig. 4—The relation of a minimum distance categorizer to a linear decision function. (a) Minimum distance categorizer. (b) Linear decision function equivalent.

Proof: Fig. 4(a) illustrates a minimum distance categorizer. A measurement A is identified with the class represented by that reference point to which it is closest in a Euclidean sense. Consider reference points 1 and 2 (RP1 and RP2) and the hyperplane B_{12} which is the perpendicular bisector of the line segment joining RP1 and RP2 [Fig. 4(b)]. Then the statement that a point A is closer to RP1 than to RP2 is equivalent to the statement that the point lies on the 1 side of B_{12}. By constructing such a hyperplane for every pair of reference points, a linear decision function equivalent to the minimum distance decision function is obtained. Therefore, minimum Euclidean distance decision functions are a subclass of linear decision functions. (Sebestyen [23], [24] has considered non-Euclidean minimum distance decision functions, which are not a subclass of linear decision functions.)

It is well known that maximizing an appropriately normalized cross-correlation function or minimizing a Hamming distance is equivalent to minimizing a Euclidean distance.

The upper bound on the number of hyperplanes required for a linear decision function is determined by noting that, for every pattern class, there will be one hyperplane separating it from every other pattern

[3] If x and y are two points with co-ordinates x_i, y_i, $1 \leq i \leq n$, then the Euclidean distance s between x and y is

$$s = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{1/2}.$$

class. If there are n pattern classes, there will then be $n(n-1)$ such hyperplanes. But this has counted each hyperplane twice. Therefore,

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

It is shown in Highleyman [12] that not all the hyperplanes are always needed. Consequently, we will have occasion to refer to *complete linear decision functions*, in which all of the $n(n-1)/2$ hyperplanes are present, and *incomplete linear decision functions* in which some hyperplanes are not included.

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

Proof: Assume that a complete linear decision function has classified a measurement into both classes i and j. But because of the completeness criterion, this linear decision function contains a hyperplane B_{ij} which will indicate either that the point cannot belong to class i or that it cannot belong to class j (assuming that a point lying on a boundary is categorized according to some convention), thus contradicting the assumption. It has already been demonstrated that some measurements may not be classified into any of the allowable pattern classes by a linear decision function, complete or otherwise; these are the patterns which are rejected (see Fig. 3).

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

Proof: This is proven in the theory of linear algebra [9].

The suggestion is sometimes made that perhaps a linear transformation on the measurement space may group like patterns closer together and separate unlike patterns, so that a linear decision function may perform better under the transformation than otherwise. That this is an invalid suggestion is demonstrated by the next theorem which is not proven here; its proof may be found in Highleyman [12].

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

The Sequential Synthesis of a Linear Decision Function

Justification of Sequential Synthesis

The complete and accurate determination of a linear decision function requires the simultaneous determination of the several hyperplanes defining it. To see this more clearly, consider Fig. 5 in which a linear decision function categorizing three classes in a measurement space is illustrated. Let the closed curves shown in this figure represent, for purposes of discussion, the domains in measurement space of classes 1 and 2. In general, the losses associated with the various possibilities for misrecognition or rejection are different. Therefore, the boundary B_{12}, for instance, must be chosen so as to minimize the loss, given by (2), associated with various factors, such as:

1) The misclassification of members of class 1 into class 2 (the horizontally hatched area);
2) the misclassification of members of class 2 into class 1 (the vertically hatched area);
3) the misclassification of members of other classes into class 1;
4) the misclassification of members of other classes into class 2;
5) the rejection of members of various classes (the dotted area).

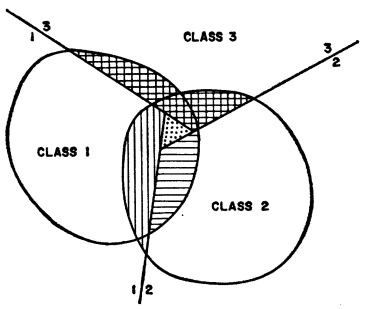

Fig. 5—Illustrating the requirement of simultaneous synthesis.

Note that the members of classes 1 and 2 which are already misclassified into other classes (in this case, into class 3 as illustrated by the cross-hatched area in Fig. 5) are not to be considered in the determination of the optimum B_{12}; these are members which are going to be misclassified anyway, regardless of the position of B_{12}. Therefore, in order to optimize B_{12}, the other boundaries, B_{13} and B_{23} in this case, must be known. But their determination also depends on B_{12}, by the same argument. Therefore, all of the boundaries comprising an optimum linear decision function must be determined simultaneously (simultaneous synthesis).

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

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

[5] A nonsingular affine transformation is a nonsingular linear transformation followed by a translation.

mum linear decision function is shown in Highleyman (12).

Even if it were deemed that this approximation is not good enough, the concept of sequential determination is still valid, for the approximation may be made better by an iterative process. First determine the hyperplanes independently, giving an initial linear decision function L_1. Then, only those members of each class which are correctly recognized by L_1 are used to recompute independently the hyperplanes, giving another linear decision function L_2. This process can be repeated until no significant improvement in performance is observed. Thus, to a better and better approximation, only those members of a particular pair of classes not misrecognized as belonging to some other class are used to determine the appropriate hyperplane. According to the previous argument, this then approaches the condition of simultaneous synthesis.

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

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

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

where

$$\int_{H_i(B_{ij})} \cdots dm$$

indicates integration over the half space which includes all points identified as class i by B_{ij}. It is of interest to relate the expected loss for the hyperplanes to the expected loss for the system; this relation is given by *Theorem 6.*

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

The proof to this theorem is given in Highleyman [12] and is burdensome and not particularly enlightening. In the hope that the theorem is easily accepted, the proof will not be repeated here.

A useful corollary follows:

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

Some Special Cases of Optimum Hyperplanes

It is shown in Highleyman [12] that, for the following two class problems, the optimum decision function is a linear decision function:

1) The two classes are equally probable *a priori*, have equal losses associated with misrecognition, and have probability distributions over the measurement space which are unimodal, spherically symmetrical, and identical except for a displacement of modes.
2) The two classes are equally probable *a priori*, have equal losses associated with misrecognition, and have probability distributions over the measurement space which are Gaussian and which have equal covariance matrices.
3) The convex hulls of the points in measurement space contained in each pattern class are nonintersecting.

The following theorem will be important later in the design and testing of practical machines.

Let us say that a set of q points in a space of n dimensions, where $q \leq n+1$, is *nondegenerate* if the points cannot be contained in a linear subspace of $q-1$ dimensions. In Fig. 6 are shown three nondegenerate points in two dimensions, and four nondegenerate points in three dimensions. Note that, in each case, the points can be separated into any two categories desired by an n-dimensional hyperplane. This is generalized in the next theorem, which is proven in Highleyman [12], but which should be intuitively acceptable from the above example.

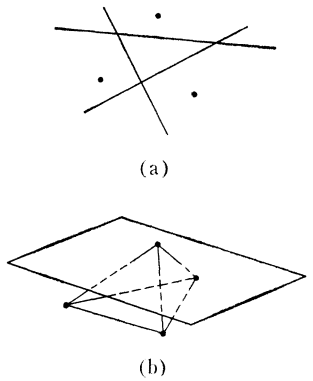

Fig. 6—Linear separability of nondegenerate points. (a) Two dimensions. (b) Three dimensions.

Theorem 7: Let S be a set of q nondegenerate points in an n-dimensional space, $q \leq n+1$. Let S_1 consist of any k of these points, and S_2 consist of the remaining $q-k$ points. Then S_1 and S_2 are linearly separable.

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

Corollary 1: Let S be a set of q points in an n-dimensional space such that any subset of S containing

no more than $n+1$ points is nondegenerate. Let $m \leq q \leq n+m-1$. Then S can be separated into m nonempty sets by a linear decision function.

Proof: By the corollary to *Theorem 6*, it is only necessary to show that each of the m sets is linearly separable. Let S be separated into the sets $S_1, \cdots, S_m$. Consider the case in which $q=n+m-1$, and the sets $S_1, \cdots, S_{m-1}$ each contain one point from S, leaving n points from S to comprise S_m. Then S_m and S_k, $1 \leq k \leq m-1$, are linearly separable by *Theorem 7*, since their union contains $n+1$ points. In any other possible case, the number of points contained in the union of any two sets S_i and S_j will be less than $n+1$, thus proving the corollary.

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

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

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

Then S can be separated into m subsets each of size no greater than $(n+1)/2$ by a linear decision function.

Proof: The union of any two subsets will contain at most n nondegenerate points if n is even, $n+1$ nondegenerate points if n is odd. Therefore, each pair of subsets is linearly separable by *Theorem 7*, and the corollary is then proved by invoking the corollary to *Theorem 6*.

Determination of the Optimum Linear Boundary Separating Two Classes

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

Three cases will be discussed. In the first, it is assumed that the pertinent conditional probability functions over the measurement space $\beta(m|s_i)$ are continuous, and that these probabilities and the *a priori* probabilities of occurrence ω_i are known. In the second case, it is assumed that nothing is known about the probabilities $\beta(m|s_i)$, and that the *a priori* probabilities ω_i may or may not be known. The determination of the optimum hyperplane is then based upon an appropriate sampling from the pattern classes. The third case is similar to the second case, but is applicable only when many independent measurements are made on the input pattern. Although it is a slightly more restrictive case, it leads to a better estimate of the hyperplane. The second and third cases are the cases of practical interest.

The Optimum Hyperplane for the Case of Known Distributions

Let $\beta(m|s_i)$ be the probability density function of class i over the measurement space, ω_i be the *a priori* probability of occurrence of class i, and c_{ij} be the loss associated with misidentifying a member of class i with class j. Denote a hyperplane which separates the classes i and j by B_{ij}, and let it be defined in the coordinate system $(x_1, \cdots, x_n)$ by the equation

$$x_1 = \sum_{k=2}^{n} \alpha_k x_k + \alpha_0. \tag{7}$$

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

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

$\int_{B_{ij}} \cdots dm'$ denotes integration over the boundary B_{ij}. Define the weighted conditional probability density function of class i over the boundary B_{ij} by

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

Theorem 8: The optimum linear boundary B_{ij}, separating two classes i and j which have weighted conditional proability density functions over B_{ij} given by

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

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

must satisfy the following conditions:

1) The integrals of τ_i and τ_j over B_{ij} must be equal.
2) The means of τ_i and τ_j must be equal.

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

$$x_1 < \sum_{k=2}^{n} \alpha_k x_k + \alpha_0.$$

The expected loss is then

$$C(B_{ij}) = c_{ij}\omega_i \int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{\infty} dx_2 \int_{\sum_{k=2}^{n} \alpha_k x_k + \alpha_0}^{\infty} \beta(m|s_i)dx_1$$

$$+ c_{ji}\omega_j \int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{\infty} dx_2 \int_{-\infty}^{\sum_{k=2}^{n} \alpha_k x_k + \alpha_0} \beta(m|s_j)dx_1. \tag{8}$$

We wish to find the coefficients of the hyperplane B_{ij} which correspond to extreme points of (8). First differentiate (8) with respect to α_0:

$$\frac{\partial C(B_{ij})}{\partial \alpha_0} = -\int_{B_{ij}} \tau_i dm + \int_{B_{ij}} \tau_j dm = 0,$$

which is condition 1 of the theorem. Next, differentiate (8) with respect to α_k, $2 \leq k \leq n$:

$$\frac{\partial C(B_{ij})}{\partial \alpha_k} = -\int_{B_{ij}} x_k \tau_i dm + \int_{B_{ij}} x_k \tau_j dm = 0,$$
$$2 \leq k \leq n.$$

A similar expression may be obtained for $k = 1$ by rewriting (7) and thus (8) in terms of some other coordinate. This set of conditions, *i.e.*, for $1 \leq k \leq n$, corresponds to condition 2 of the theorem.

In general, there will be several hyperplanes satisfying the conditions of *Theorem 8*. Some of these will correspond to maxima of $C(B_{ij})$, others to minima. These must then be searched to determine which corresponds to the absolute minimum of $C(B_{ij})$.

The Optimum Estimate for the Hyperplane for the Case of Unknown Distributions

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

If a hyperplane B_{ij} is passed through M such as to divide classes i and j in some fashion, then a certain portion of the members of classes i and j will be misidentified by B_{ij}. Let p_i be the probability of misidentification of a member from class i, given B_{ij} and a member of class i (p_i is the integral of $\beta(m|s_i)$ over the half-space on the j side of B_{ij}). Then the conditional loss associated with B_{ij} (see (1)) is

$$\begin{aligned}\bar{C}(B_{ij}) &= c_{ij}\omega_i' p_i + c_{ji}\omega_j' p_j \\ &= c_{ij}e_i + c_{ji}e_j,\end{aligned}$$

where

$$\omega_i' = \frac{\omega_i}{\omega_i + \omega_j}, \qquad \omega_j' = \frac{\omega_j}{\omega_i + \omega_j},$$

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

Theorem 9: Construct a hyperplane B_{ij} in the measurement space M which divides M into two half spaces, all the points in one being identified as class i, the points in the other being identified as class j. Consider two sampling procedures designed to estimate the conditional cost $\bar{C}(B_{ij})$:

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

$$\begin{aligned} m_i &= \omega_i' m \\ m_j &= \omega_j' m. \end{aligned} \tag{9}$$

(It will be assumed that ω_i' and ω_j' are such that (9) can be met exactly.) Identify each of these m samples according to B_{ij}. Let e_i be the number of samples from class i misidentified by B_{ij} as class j, and likewise for e_j.

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

$$\hat{\bar{C}}(B_{ij}) = \frac{c_{ij}e_i + c_{ji}e_j}{m}. \tag{10}$$

The proof is based on the fact that m_i and m_j are binomially distributed. Since the theorem is intuitively acceptable, the proof is not given here, but may be found in Highleyman [12].

If we take samples from a pair of classes according to either sampling procedure, there will be a set of hyperplanes (infinite in number) which will minimize the maximum likelihood estimate of the conditional loss (10). It is quite reasonable, then, to choose one of the hyperplanes from this set as the estimate of the optimum hyperplane separating the two classes. That is, it is clear from (10) that *we will search for a hyperplane which will minimize the loss associated with the sample points.* This is also intuitively quite reasonable.

Note that *Theorem 9* and the resulting procedure is independent of the probability functions over the measurement space. Hence, one need make no assumptions concerning the form of these functions, nor need one concern himself with the dependencies between the various measurements.

A Computation Algorithm for the Case of Unknown Distributions

In this section will be outlined an iteration algorithm which will be useful for determining that boundary which minimizes the maximum likelihood estimate of the conditional loss for the boundary. The detailed iteration equations are given in Highleyman [12]. There has been some work by others concerning similar boundaries when the measurement space is a binary space [17], [18], [26], [27], or when the classes are Gaussian distributed in measurement space (discriminant functions [2], [7], [25] yield a good approximation for this case).

Fig. 7(a) illustrates this problem for two classes, k and l. Samples from class k are shown by crosses, from class l by circles. A boundary B_{kl} is indicated. Let us number these samples from 1 to m, there being a total of m

samples, and define a weight T_j' for the jth sample point $1 \leq j \leq m$ such that

$T_j' = 0$ if the point is on the correct side of B_{kl};
$T_j' = c_{kl}$ if the point represents a sample from class k on the l side of B_{kl};
$T_j' = c_{lk}$ if the point represents a sample from class l on the k side of B_{kl}.

It is clear, then, that minimizing the estimate of the conditional loss (10) is equivalent to minimizing

$$T'(\alpha_i) = \sum_{j=1}^{m} T_j' \tag{11}$$

where the α_i, $0 \leq i \leq n$, are the coefficients of the hyperplane B_{kl} defined by (4).

$T'(\alpha_i)$ is an $(n+1)$-dimensional function for a system with n measurements. A convenient way to determine a minimum point of this function would be to use a gradient method, such as the method of steepest descent [1], [4], [20]. However, $T'(\alpha_i)$ is a discontinuous function of the α_i, and thus has no meaningful gradient.

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

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

where T_j is written as a function of the distance of the jth point from the hyperplane to emphasize this continuous dependence. Such a function is shown in Fig. 7(b), in which s_j is the distance of the jth point from the boundary B_{kl}, and will be considered to be positive if the point j is on the correct side of the boundary. The quantity c_j in Fig. 7(b) is equal to c_{kl} if the jth point represents a member from class k; $c_j = c_{lk}$ otherwise. If the function

$$T(\alpha_i, \lambda) = \sum_{j=1}^{m} T_j(s_j, \lambda) \tag{12}$$

were to be minimized for some finite λ with respect to the α_i, and then λ increased and (12) minimized again, and this process repeated, one would expect the hyperplane to converge to one of the set of hyperplanes minimizing (11). This minimization process can now make use of the method of steepest descent.

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

$$T_j(s_j, \lambda) = c_j[1 - G(\lambda s_j)], \tag{13}$$

where

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

The algorithm then consists of determining the direction of the gradient of (12), using some suitable function for $T_j(s_j, \lambda)$ such as (13). Some reasonable initial guess for the hyperplane and some value for λ are used as a starting point. The approximate minimum of (12) is determined by the method of steepest descent, and the process is then repeated with a larger value of λ. When the desired accuracy is achieved, the iteration is terminated.

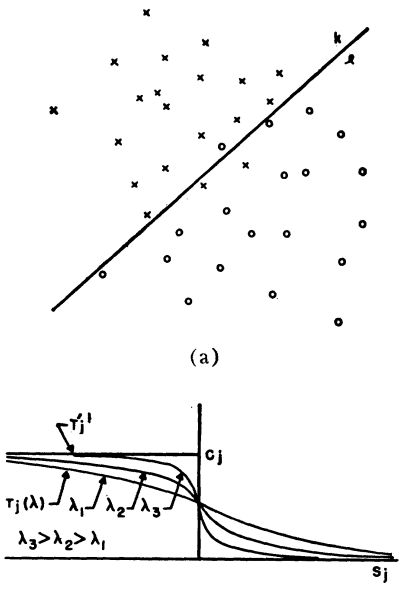

Fig. 7—Illustrating the iterative algorithm.

A Computational Algorithm for the Case of Many Independent Measurements with Unknown Distributions

We next consider the case of a large number of independent measurements whose distributions are otherwise unknown. The distances of members of pattern class i from a given hyperplane may be said to be distributed according to a probability density function $\eta_i(s)$. $\eta_i(s)$ is in general unknown. However, in many cases, one may estimate it quite accurately by the following argument.

Recall that the distance of a point m from a hyperplane is given by

$$s = \sum_{i=1}^{n} \alpha_i m_i + \alpha_0,$$

where m_i is the ith coordinate of the point (the ith measurement), and the α_i are the normalized coefficients of the hyperplane. But m_i is a random variable, and hence, if n is large, s is a weighted sum of a large number of random variables. If the dependencies between the random variables are weak, one may then reasonably expect from the Central Limit Theorem [8] that the distribution of s is approximated by a normal distribution. Of course, if the measurements m_i are independent and normally distributed, then the normality of s follows immediately for any n.

Consequently, $\eta_i(s)$ is, to a good approximation in many cases, a normal density function. Its mean and variance can be easily estimated from the samples which are to be used to design the linear decision function. Using this normality concept, one can develop an algorithm for estimating the optimum hyperplane separating the two classes. We are interested in choosing that hyperplane that minimizes the estimate of the expected error, or confusion, between the two classes. However, it is possible to estimate the error associated with a hyperplane by estimating the normal distribution of the distance of the members of each class from the hyperplane, and determining the area under the tails of these two distributions falling on the wrong side of the hyperplane. This might be expected to be a better estimate of the error than the proportion of points misrecognized, since more information is used in the estimate, *providing the assumption of normally distributed distances is valid.*

Consequently, it would be quite reasonable to choose, as an estimate of the optimum linear boundary, that hyperplane which minimizes the normal estimate of error rather than the estimate based on the proportion of misclassified samples, providing again that the assumption of normality holds. A computational algorithm based on minimizing this normal estimate of error, using the method of steepest descent, is developed in Highleyman [12], where confidence intervals for this and the preceding estimate are also developed. These confidence intervals illustrate that this latter estimate is indeed the better of the two. Note that the resulting hyperplane for each local minimum is unique, in contrast to the previous algorithm in which the hyperplane could be any one chosen from, in general, an infinite set.

An Example of Categorization

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

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

$$\begin{aligned} \text{Class 1:}\quad & \sigma_{1x} = 1 \qquad \mu_{1x} = 1 \\ & \sigma_{1y} = 0.5 \qquad \mu_{1y} = 1 \\ \text{Class 2:}\quad & \sigma_{2x} = 0.1 \qquad \mu_{2x} = 2 \\ & \sigma_{2y} = 2 \qquad \mu_{2y} = 0 \end{aligned}$$

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

In Fig. 8 are shown the 1δ contours of the classes 1 and 2. Also shown are the boundaries based on the previously mentioned approaches:

1) The optimum decision function for the two-class Gaussian case is well known [2], [16]. The result is the hyperbolic boundary shown, given by the equation

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

The region identified as class 2 is that between the two curves of the hyperbola.

2) The optimum linear decision function, given complete knowledge of the distributions, is worked out for this case in Highleyman [12], and is based on the use of *Theorem 8*. The result is

$$y = 1.04x - 1.32. \tag{14}$$

This is shown as the "theoretical" linear boundary in Fig. 8.

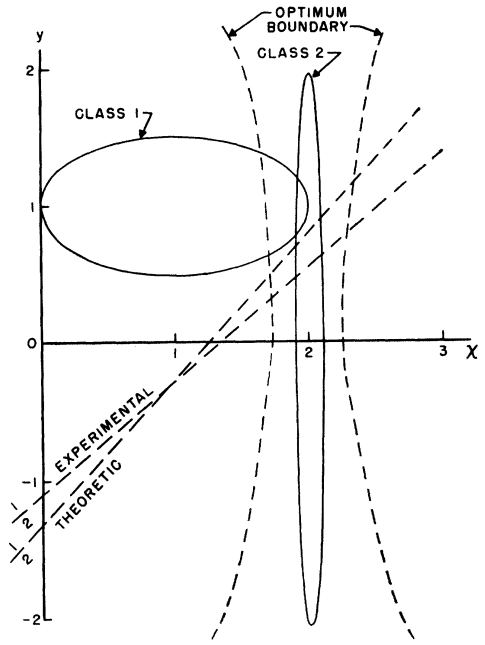

Fig. 8—An example of some of the approaches to categorization.

3) The optimum linear decision function based on sampling was determined using the first iteration algorithm (normally distributed distances were not assumed, although in this case they are normally distributed). A random number generator was programmed for the IBM 7090 digital computer which generated numbers according to the above distributions.

One hundred sample points were taken from each class, and various initial boundaries were tried:

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

Each of the final boundaries were slightly different, but the important point is that each one categorized the points in exactly the same manner. (Thirty-nine points were always misclassified.) An example of one of the final boundaries is

$$y = 0.816x - 1.11 \tag{15}$$

which is plotted in Fig. 8 as the linear boundary marked "experimental." Compare (14) to (15); the difference illustrates the sampling error.

Experimental Application—The Recognition of Hand-Printed Numbers

The recognition of hand-printed numbers was attempted with a linear decision function. The set of measurements which was used involved quantizing the number into a 12×12 binary matrix. A matrix element was given a weight of one if it contained a mark and a weight of zero if it contained no mark. The quantized number was then positioned in the matrix by aligning its center of gravity with the center of the matrix. Hence, a 144-dimensional binary measurement space was used. This set of measurements is a rather unsophisticated set in that the measures are not at all invariant within a particular class; thus, one would not be too surprised if a linear decision function did not perform very well.[6] However, the attempt is still interesting since it will allow the testing of the preceding ideas in some detail.

Estimating the Linear Decision Function

The data used to estimate the optimum linear decision function was gathered in the following manner. A subject was asked to print neatly the ten numbers on a piece of quadruled paper at a size approximating the ruled boxes. Fifty different people were asked, resulting in a sample size of 50 for each of the ten pattern classes. These data were then automatically reduced to a 12×12 matrix (encoded on IBM punched cards) by an optical matrix scanner constructed by the author.[7]

In Fig. 9 is shown an example of some of this design data, illustrating approximately the range of size and neatness obtained. In Fig. 10 are shown examples of some of the quantized numbers.

Forty-five hyperplanes are required in the complete linear decision function categorizing the ten numbers. It was assumed that all losses due to misrecognition are equal (minimum error), and that all *a priori* probabilities are equal. Each hyperplane was determined by first finding a hyperplane which correctly categorized the maximum number of sample points (according to the first computational algorithm), and then by "trimming up" that hyperplane so that the normal estimate of error was minimized (according to the second computational algorithm). About 35 seconds, on the average, was required to determine a hyperplane, given an initial position.

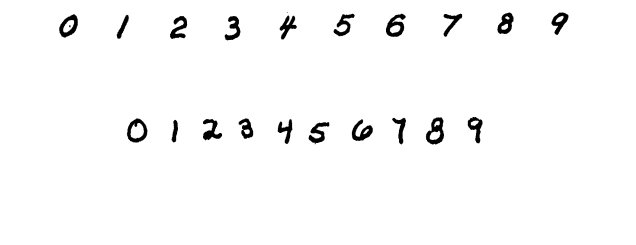

Fig. 9—Some examples of the hand-printing design data.

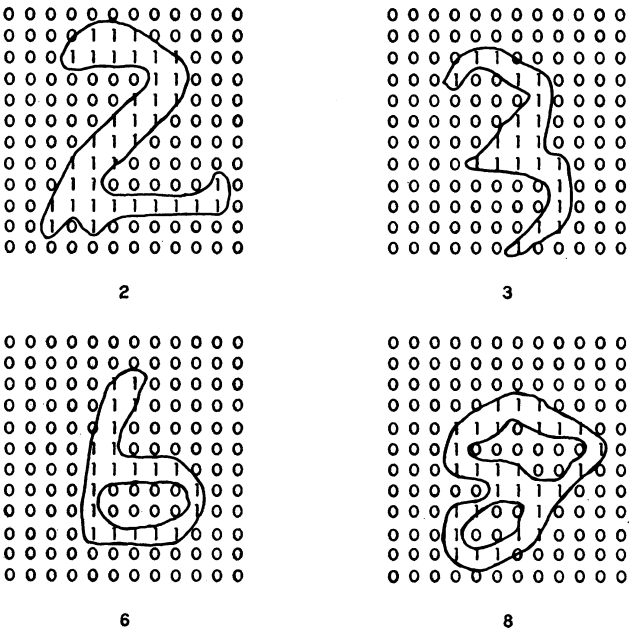

Fig. 10—Examples of quantized forms of the hand-printed numbers

For each pair of pattern classes, four initial hyperplanes were tried. One of these was that hyperplane which was the perpendicular bisector of the line segment joining the means of the two classes. The other three initial hyperplanes were parallel to this one (*i.e.*, the direction cosines were the same) and corresponded to an α_0 of 0, −5 and +5. Each of these initial conditions led to a hyperplane better than any of the other initial conditions in at least one of the 45 cases, thus illustrating the importance of trying several initial hyperplanes.

In Fig. 11 is shown the estimated optimum hyperplane B_{21} which separates the numbers 2 and 1. The coefficients α_i, $1 \leq i \leq n$, are shown arranged in a matrix corresponding to the receptor matrix. The positive side of B_{21} corresponds to the number 2. One

[6] A very effective set of measurements has been proposed by Kamentsky [15] for the recognition of hand-printed numbers. This involves using a "flying-polar" scan which is capable of determining the number of closures and cusps (partial closures) and the orientation of cusps in a character.

[7] These are the same data used in the Bledsoe-Browning comparison, reported in Bledsoe [3], and Highleyman and Kamentsky [11].

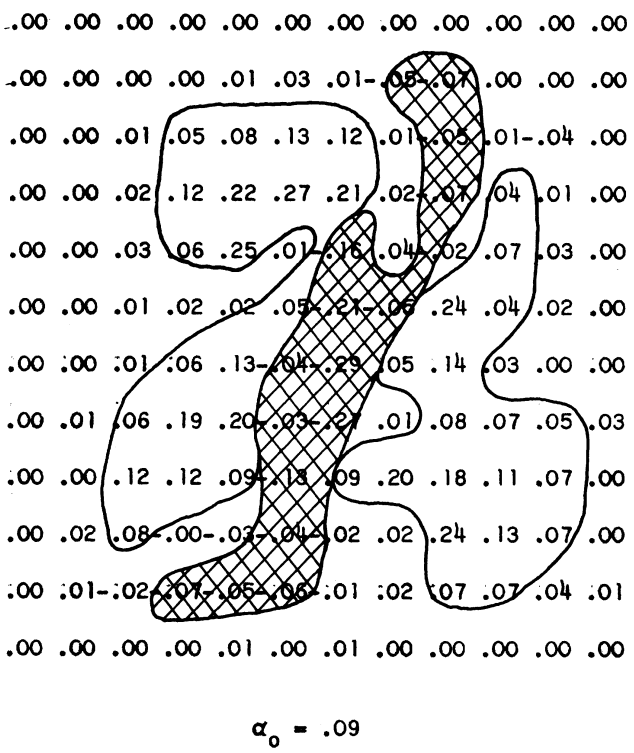

Fig. 11—The hyperplane B_{12}.

would then expect that those coefficients which corresponded to matrix elements in which a mark from a two was likely to occur and a mark from a one was not likely to occur would be weighted positively, and vice versa for those elements in which a mark from a one is more likely to occur. Contours are drawn around regions of large positive and negative weight in Fig. 11, and the negative regions are shaded. One sees that the above intuitive observation does indeed hold.

The resulting linear decision function miscategorized five patterns (1.0 per cent) and rejected one pattern (0.2 per cent) of the total design sample of 500, as shown in the confusion matrix of Table I (the *R* column indicates the input patterns rejected by the inherent rejection of the linear decision function). Also shown in the table are the values of the error-rate estimates based on the normality assumption. However, one cannot conclude that these percentages are any sort of valid estimate for the performance of the system, since they are based on the samples used to design the system. In fact, since only 100 points are being separated in 144 dimensions by each hyperplane, one might expect from *Theorem 7* that the linear decision function ought to do well on the design data. The saving grace here is the fact that the measurement space is binary, and therefore the sample points are highly degenerate in the sense of *Theorem 7*. It is therefore not to be expected

TABLE I

Input Class \ Recognized As	0	9	8	7	6	5	4	3	2	1	*R*
0	50 —	0.004	0.000	0.014	0.021	0.005	0.010	0.009	0.005	0.000	—
9	0.004	48 —	1 0.030	1 1.118	0.003	0.001	1.686	0.001	0.044	0.000	—
8	0.000	0.184	50 —	0.005	0.038	0.033	0.133	0.296	0.034	0.007	—
7	0.026	0.236	0.001	50 —	0.000	0.000	0.003	0.022	0.111	0.061	—
6	0.024	0.005	0.052	0.000	50 —	0.018	0.024	0.029	0.136	0.000	—
5	0.002	0.002	0.033	0.000	0.067	50 —	0.011	0.068	0.002	0.000	—
4	0.004	1.771	0.039	0.011	0.013	0.009	50 —	0.002	0.032	0.000	—
3	0.003	0.001	2 0.694	0.042	0.026	0.152	0.001	47 —	1 0.894	0.010	—
2	0.017	0.041	0.034	0.037	0.157	0.003	0.027	1.990	49 —	0.002	1 —
1	0.000	0.000	0.002	0.001	0.000	0.000	0.000	0.003	0.003	50 —	—

Correct 494 (98.8%)
Error 5 (1.0%)
Reject 1 (0.2%)

Confusion matrix for the design sample: Upper numbers give the categorization of the design data; lower numbers give the normal estimates of error, in per cent.

that any set of points, no greater in number than $n+1$ (145 in this case), will be linearly separable in general in this measurement space.

Testing the Linear Decision Function

The resulting system was tested with 120 additional samples (12 samples of each number) gathered in the same manner as the design data. Fig. 12 shows this test sample. The confusion matrix of Table II represents the categorization of these samples.

The resulting estimate of the system error rate, rejection rate, and correct recognition rate, from the results shown in Table II, are 19.2 per cent (23 points), 19.2 per cent (23 points) and 61.6 per cent (74 points), respectively. From confidence intervals given in Highleyman [12] covering this sort of test, one can then state that, with probability 0.95, the intervals 0.13–0.28, 0.13–0.28, and 0.52–0.70 include the system error probability, rejection probability, and correct recognition probability, respectively.

It is not surprising to find the estimated performance of this linear decision function to be so poor. This can be blamed on two factors: 1) a poor choice of measurements, inthat the measurements used were very dependent upon the distortions and various noise effects (smudging, etc.) which might occur, and 2) a design sample size which is too small, leading to a poor estimate of the optimum hyperplanes. This latter point is emphasized by the difference in the results obtained with the design sample and with the test sample (98.8 per cent recognition vs 61.8 per cent). If the design sample were sufficiently large, one would expect the results based on the two samples to be comparable; hence, one might expect that, for a large design sample size, the performance of the resulting machine would be somewhere between the two results obtained herein. However, this is not so important, since this experiment was not meant to result in the design of a practical character recognition machine, but was rather meant to test certain aspects of the theory previously developed.

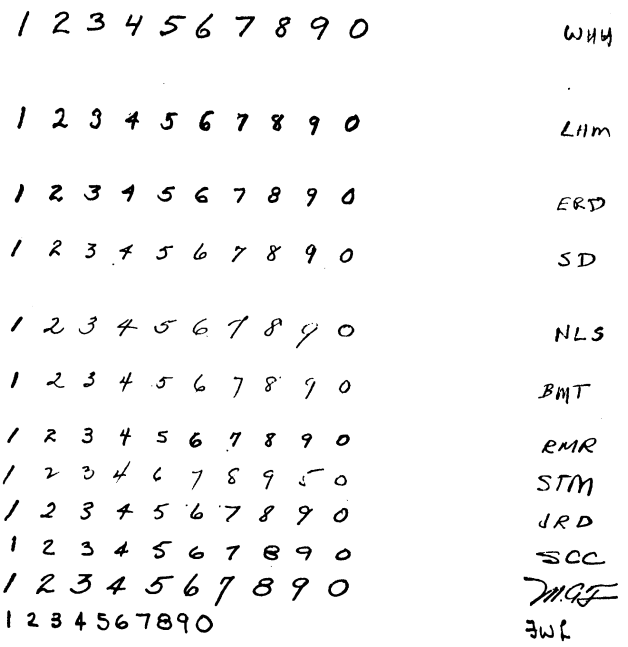

Fig. 12—The test sample.

TABLE II

Input Class	Recognized As: 0	9	8	7	6	5	4	3	3	1	R
0	9										3
9		6		1			1				4
8	1		5					2		1	3
7		1		8			2				1
6					8	1					3
5	1					8		2			1
4		3			2		4				3
3						2		9			1
2				2	1				6		3
1										11	1

Correct 74 (61.6%)
Error 23 (19.2%)
Reject 23 (19.2%)

Confusion matrix for the test sample.

Conclusion

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

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

It is not guaranteed that a linear decision function will always perform well, although it is guaranteed that it will perform better than (or at least as well as) the minimum distance categorizer which is popular in the present-day art. Nor is it a simple matter to predict in advance whether a linear decision function has a chance of working (this problem is discussed in Highleyman [13]).

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

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

One area which has not been discussed in this paper is the problem of minimizing a linear decision function. It often happens that not all of the hyperplanes are needed, *i.e.*, some may fall outside of the convex regions determined by the others. The linear decision function may also be used to detect redundancy in the measurements. This problem is discussed further in Highleyman [12], where it is experimentally shown that the linear decision function for the hand-printing case discussed herein may be reduced from 144 measurements and 45 boundaries to 110 measurements and 39 boundaries.

Acknowledgment

The author would like to thank Professor A. E. Laemmel of the Polytechnic Institute of Brooklyn for his guidance in this work. He would also like to thank L. A. Kamentsky, W. H. Williams, F. W. Sinden, and R. Gnanadesikan for their helpful discussions and comments. He is especially grateful to W. R. Cowell and E. Wolman for reviewing portions of the manuscript, as well as for the many other discussions to which they so willingly contributed.

Bibliography

[1] S. Agmon, "The relaxation method for linear inequalities," *Can. J. Math.*, vol. 6, pp. 382–392; 1954.
[2] T. W. Anderson, "An Introduction to Multivariate Statistics," John Wiley and Sons, Inc., New York, N. Y.; 1958.
[3] W. W. Bledsoe, "Further results on the *n*-tuple pattern recognition method," IRE Trans. on Electronic Computers, vol. EC-10, p. 96; March, 1961.
[4] S. H. Brooks, "Comparison of maximum-seeking methods" *Operations Res.*, vol. 7, pp. 430–457; July, 1959.
[5] H. Chernoff and L. E. Moses, "Elementary Decision Theory," John Wiley and Sons, Inc., New York, N. Y.; 1953.
[6] C. K. Chow, "An optimum character recognition system using decision functions," IRE Trans. on Electronic Computers, vol. EC-6, pp. 247–254; December, 1957.
[7] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, p. 179; 1936.
[8] D. A. S. Fraser, "Nonparametric Methods in Statistics," John Wiley and Sons, Inc., New York, N. Y.; 1957.
[9] S. I. Gass, "Linear Programming," McGraw-Hill Book Co., Inc. New York, N. Y.; 1958.
[10] W. H. Highleyman, "A note on optimum pattern recognition systems," IRE Trans. on Electronic Computers, vol. EC-10, pp. 287–288; June, 1961.
[11] W. H. Highleyman and L. A. Kamentsky, "Comments on a character recognition method of Bledsoe and Browning," IRE Trans. on Electronic Computers, vol. EC-9, p. 263; June, 1960.
[12] W. H. Highleyman, "Linear Decision Functions, with Application to Pattern Recognition," Ph.D. dissertation, Elec. Engrg. Dept., Polytechnic Inst. Brooklyn, N. Y.; June, 1961.
[13] W. H. Highleyman, "A note on linear separation," IRE Trans. on Electronic Computers, vol. EC-10, pp. 777–778; December, 1961.
[14] L. A. Kamentsky, "Pattern and character recognition systems—picture processing by nets of neuron-like elements," *Proc. Western Joint Computer Conf.*, San Francisco, Calif., March 3–5, 1959; pp. 304–309.
[15] L. A. Kamentsky, "Simulation of three machines which read rows on handwritten arabic numbers," IRE Trans. on Electronic Computers, vol. EC-10, pp. 489–501; September, 1961.
[16] T. Marill and D. M. Green, "Statistical recognition functions and the design of pattern recognizers," IRE Trans. on Electronic Computers, vol. EC-9, pp. 472–477; December, 1960.
[17] R. L. Mattson, "The Design and Analysis of an Adaptive System for Statistical Classification," M.S. thesis, Elec. Engrg. Dept., Mass. Inst. Tech.; May, 1959.
[18] R. L. Mattson, "A self-organizing binary system," *Proc. Eastern Joint Computer Conf.*, Boston, Mass., December 1–3, 1959; pp. 212–217.
[19] D. Middleton and D. Van Meter, "Detection and extraction of signals in noise from the point of view of statistical decision theory," *J. Soc. Ind. and Appl. Math.*, vol. 3, pp. 192–235, September, 1955; and vol. 4, pp. 86–119, June, 1956.
[20] T. S. Motzkin and I. J. Schenberg, "The relaxation method for linear inequalities," *Can. J. Math.*, vol. 6, pp. 393–404; 1954.
[21] F. Rosenblatt, "A Theory of Statistical-Separability in Cognitive Systems," Cornell Aeronautical Lab., Inc., Report No. VG-1196-G-1, Buffalo, N. Y.; January, 1958.
[22] F. Rosenblatt, "Perceptron simulation experiments," Proc. IRE, vol. 48, pp. 301–309; March, 1960.
[23] G. S. Sebestyen, "Categorization in Pattern Recognition," Ph.D. dissertation, Elec. Engrg. Dept., Mass. Inst. Tech.; April, 1960.
[24] G. S. Sebestyen, "Recognition of membership in classes," IRE Trans. on Information Theory, vol. IT-7, pp. 44–50; January, 1961.
[25] G. Tintner, "Econometrics," John Wiley and Sons, Inc., New York, N. Y.; 1952.
[26] B. Widrow and M. E. Hoff, "Adaptive Switching Circuits," Stanford Electronics Lab., Stanford, Calif., Tech. Rept. No. 1533-1; June, 1960.
[27] B. Widrow, "Adaptive Sampled Data Systems," Stanford Electronics Lab., Stanford, Calif., Tech. Rept. No. 2104-1; July, 1960.