PATTERN RECOGNITION RESEARCH

by

George Sebestyen

Jay Edie

Prepared by

Information Sciences Laboratory
Data Systems Division
LITTON SYSTEMS, INC.
335 Bear Hill Road
Waltham, Massachusetts 02154

Contract No. AF19(628)-1604
Project No. 5632
Task No. 563205

FINAL REPORT

14 June 1964

Prepared for

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS

FINAL REPORT

PATTERN RECOGNITION RESEARCH

Contract AF19(628)-1604

14 June 1964

Approved by

George Sebestyen
Technical Director

John W. Gerdes
Assistant Manager

Information Sciences Laboratory
Data Systems Division
LITTON SYSTEMS, INC.

Requests for additional copies by Agencies of the Department of Defense, their contractors, and other Government agencies should be directed to:

DEFENSE DOCUMENTATION CENTER (DDC)
CAMERON STATION
ALEXANDRIA, VIRGINIA

Department of Defense contractors must be established for DDC service or have their "need-to-know" certified by the cognizant military agency on their project or contract

All other persons and organizations should apply to the:

U. S. DEPARTMENT OF COMMERCE
OFFICE OF TECHNICAL SERVICES
WASHINGTON 25, D. C.

ABSTRACT

This report is concerned with the adaptive estimation of joint probability densities from a finite number of multi-dimensional vectors of known classification. An estimation procedure for the approximation of probability densities in the form of an n-dimensional histogram is described. The location and shape of the cells in the histogram are dependent on the data. The quality of the estimation procedure and its dependence on the order in which samples of known classification are introduced are described. Two quality measures are studied, one that estimates the probability that the decision is optimum and the other that the decision is correct. Techniques for analysis of data of unknown origin prior to the application of the adaptive pattern recognition techniques are studied. The measurement selection problem of pattern recognition is investigated and the mathematical and engineering problems are separated. Figures of merit to evaluate the usefulness of parameter sets are developed, and mathematical formulations of the parameter selection problem are given.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

1. INTRODUCTION

Even untrained human beings are credited with the uncanny ability to recognize a person's identity from his handwriting or from the sound of his voice, to recognize an author or singer from his style, or the sound of one musical instrument from another. Different words written by the same person have some common properties, they follow the same pattern. This pattern is different from that followed by the handwriting of another person. Early work in pattern recognition stemmed from an admiration of the facility of humans to learn with ease these common patterns, and from an admiration of their high degree of accuracy in recognizing the patterns in different handwriting and thereby to identify the authors.

From the vague notions of looking for the common patterns and from attempts to construct machines that can recognize patterns to approximate, in performance, the ability of humans to do the same grew the now scientific field of automatic Pattern Recognition. The two major problems "machine learning" (learning or discovering the common pattern of a class of things) and "recognition" or classification were distinguished early. One group of workers was concerned only with machines that did recognition automatically. They employed humans to learn the common pattern and to design the classifier*.

---

* "Shoebox" (a 16 word vocabulary spoken word recognizer) developed at IBM is an example that illustrates this point.

Members of a second group derived motivation from the argument that if humans (or even simple biological systems) can do pattern recognition with ease, then one ought to model biological systems by a partial simulation of their internal structures and perhaps similar performance might result*. The term Bionics is now applied to work based on this premise.

In the eyes of those who will be grouped together in the third category of workers in this report, "learning" and "recognition" problems of pattern recognition can be formulated in mathematical terms as problems of recognition of membership in classes, and some solutions can be obtained through the application of one of the mathematical disciplines such as group theory, set theory[1], Boolean algebra[2, 3], integral geometry[4], communication theory, statistical decision theory and others. The common starting point of each of these methods is to represent an input by a set of measurements, variously called features, receptors, parameters, coordinate dimensions, clues, properties or attributes. Each input that belongs to a given class can be regarded as a vector in a vector space and is located at a point defined by the set of measurements. A class is a collection of points scattered in some manner in the vector space (often referred to as observation or measurement space). Members of two different classes, A and B, are distributed, in general, in different manners in the space. Machine learning (or learning what the pattern is) is regarded by all of the above disciplines as the problem of determining the best shape and location of regions in the vector space so that A's and B's should become separated into regions called A and B. This is illustrated in Figure 1. Pattern recognition or classification is the act of naming the region (A or B) in which the measurements made on a new input are contained**.

---

* Early work on the Perceptron was based on such an argument.

** A more detailed exposition of the idea of vector representation and of the geometrical interpretation of "learning" and "recognition" can be found in many places in the literature. It will not be dealt with here.

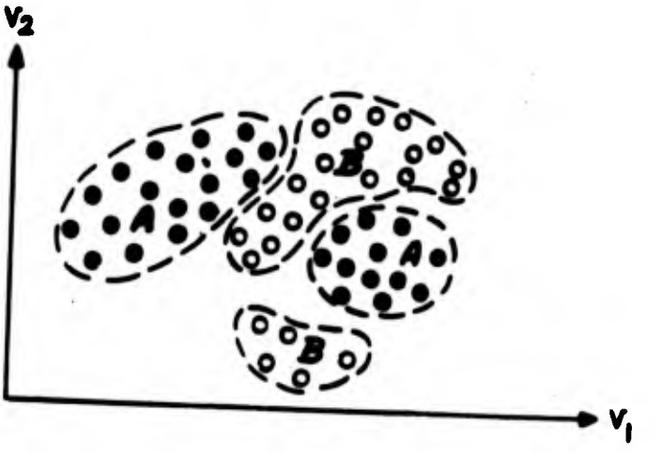


Figure 1. Machine Learning - Partitioning of Vector Space into Regions

The three parts of pattern recognition systems are illustrated by the block diagram of Figure 2. This shows the observation system that represents the input by a set of measurements. The choice of these measurements is an important problem that will be dealt with later in this report. The methods used to process inputs of known classification to discover their common pattern and thus to develop a good partition of the vector space is referred to as "learning". The act of evaluating a new input to decide in which partition of the space it is contained is performed by the classification or recognition system. It should be noted that in the final analysis all recognition systems can be regarded as table look-ups for they all associate a previously stored decision with each possible input and for the same input they always render the same decision. Of course, there are major differences in the manner in which different recognition systems store the decisions that should be made at any one of an infinite number of points in the vector space while they possess only a finite capacity of information storage. The important difference between different pattern recognition techniques, however, is not in the recognition system but in the learning system where the way in which partitions are obtained from the learning samples and where the restrictions on the type of obtainable partitions are determined.

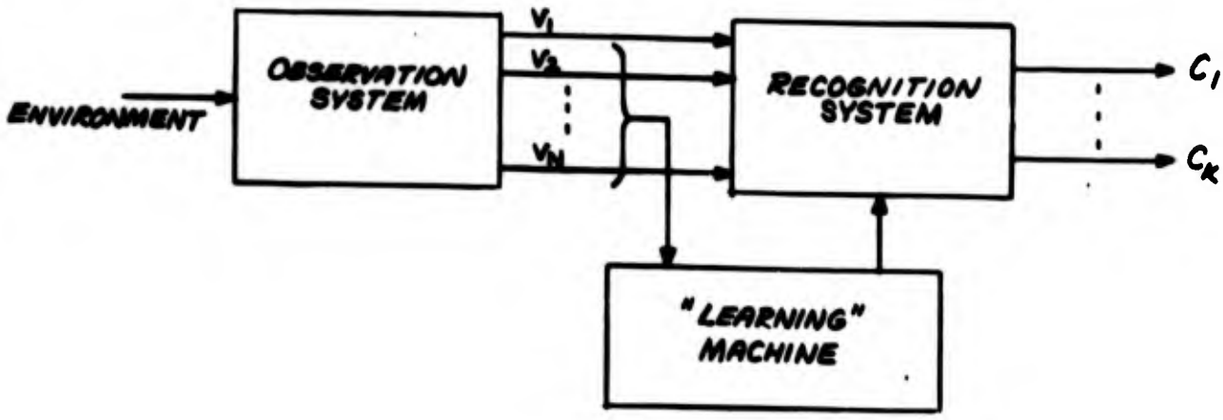


Figure 2. General Pattern Recognition System

There are many ways of partitioning the vector space into regions. In the last few years statistical methods (in particular, statistical decision theory) have emerged as leading contenders for effecting good partitions of the vector space. The applicability of decision theory in the design of pattern recognition systems is readily appreciated by considering its basic characteristics. Once input stimuli are expressed in terms of a set of measurements, we want to design a classification system with the best performance; i. e., one that makes the least number of mistakes. In addition, we recognize that the classification system will have to render decisions on inputs that are not identical to those from which classification was learned (although they will be similar, in general). It was shown by Wald, Middleton, Van Meter and others that if we wish to minimize the risk, the probability of error, or the maximum error due to the decision we make, then we should base our decision on the comparison of likelihood ratios with fixed constants. That is to say, if we must choose between two classes A and B as giving rise to the stimulus which we observe through a set of measured parameters, then we should base our decision on the comparison of the ratio of conditional probability densities with an appropriately chosen constant. In mathematical form this expresses the notion that if the set of measured parameter values is a more likely occurrence under the

assumption that the stimulus belongs to class A than under the assumption that it belongs to class B, then common sense (and statistical techniques) advise us to decide that probably A gave rise to our specific observations. Thus decision theory provides us with a design procedure that uses a figure of merit that reflects ultimate system performance as the basis for system design, and it also agrees with intuition.

There is a fundamental difference between the answers that are derivable from statistical techniques and the answers sought by pattern recognition. Decision theory assumes a state of knowledge by requiring knowledge of the relative frequency of occurrence of every observable set of measurements from all classes of interest. In pattern recognition problems, this state of knowledge is missing and estimates of the required quantities must be made from a finite number of class samples.

It is the purpose of this report to examine the major problems of pattern recognition from a statistically motivated point of view and to show the present solutions to these problems, where they are available, and to formulate problems so that they should be mathematically tractable in those cases where the present state-of-the-art has not yet given us solutions.

First we will examine the nature of the statistical solutions in the context of often-voiced objections against the use of a statistical approach. Next, we will describe an automatic technique to estimate arbitrary probability densities (needed by a statistical classification system) from a finite number of learning observations. Then the technique of efficiently storing an adequate approximation to these density functions will be described. Various properties of the technique will be discussed and some of the features of its hardware realization will be described. Some estimates of the quality of the approximation technique will be given. Methods of performing analyses on data prior to application of the automatic joint probability density estimating procedure will be discussed.

2. ON PARAMETER PROCESSING

While it may not be necessary to think of inputs as represented by a set of descriptors or parameters, once inputs are represented by vectors, the method of partitioning a vector space into regions must be considered. Statistical methods have shown that it is possible to construct a decision procedure (a partition of the vector space) which assigns a class label to each point of the space in such a way that the probability of making an error is minimized.

2.1 The Nature of Statistical Techniques

Many view with skepticism the statistical approach to decision making. They imply that an unreasonably large number of samples are needed to establish the required statistics, that decision making is not a statistical process for an object is definitely a member of one class or another, that there are foolproof clues of classification or decision making that (when present) must carry overwhelming weight, that "well-known" physical properties of the input classes are not exploited by statistical techniques, that a given set of measurements, depending on the context in which it is observed, may have to be assigned to different classes at different times, that statistics does not get at the heart of the matter, it does not tell us how the measurements should be selected.

It would be beyond the scope of this report to answer all these charges in detail. It is hoped, however, that the following brief discussion of the nature of statistical techniques will help to reduce some of the skepticism that may be based on an imperfect understanding of the nature of statistical techniques.

It is conceded at the outset that statistical techniques (or any mathematical technique for that matter) are inherently incapable to generate new methods of representing input stimuli or to improve on the parametric representation of the physical world with which it is confronted. The choice of suitable parameters is largely an engineering problem that requires our utmost inventiveness in every application to produce a useful method of representation. There are some aspects of the parameter selection problem, however, that admit to mathematical treatment. These will be the subject of later discussions. Let us now turn our attention to class separability by parameter processing. The two or more different classes into which we wish to divide all input stimuli either overlap in the vector space or they do not. If they do not (a case often assumed tacitly by those who do not propose statistical techniques), then identically the same combination of parameter values cannot be observed on both of two inputs where one belongs to one class while the other belongs to another. In this case, the classes are perfectly separable and a decision making system that never makes an error can be achieved. Let us see how statistical techniques behave in this situation.

The probability that a decision arrived at in this manner is incorrect can be expressed in terms of the equation below (where equal a priori probabilities of A and B are assumed so that the explanation should remain simple).

$$\text{Probability of decision being in error} = \frac{1}{1 + \ell(v_1 \ldots v_N)} \qquad (1a)$$

$$\text{where} \quad \ell(v_1 \ldots v_N) = \frac{P(v_1 \ldots v_N/A)}{P(v_1 \ldots v_N/B)} \quad \text{if } \ell > 1 \qquad (1b)$$

We note that if a particular combination of parameter values $v_1 \ldots v_N$ was only observed on stimuli belonging to A and never on stimuli belonging to B, then the denominator of the likelihood ratio becomes zero and the probability of error is $1/1 + \infty$ or zero. In other words, we are quite certain that we are correct in our decision and we have shown that statistical techniques lead to exactly the same decisions and with the same certainty as other techniques under tacit assumptions of perfect separability of the classes.

If two or more classes overlap in the vector space, that is, if identically the same set of observations have been made on members of class A at one time and on members of class B at a different time, it is impossible to make completely error-free decisions all of the time. It would then seem logical to invoke a criterion that requires that the "optimum" decision procedure should minimize the number of wrong decisions or should minimize the probability of error or the risk to the decision maker. Statistical decision-making systems provide us with a decision procedure with just such properties.

Thus, statistical techniques give us foolproof decisions when perfectly error-free decisions are feasible and give us the pragmatically best decisions when error-free decisions cannot be guaranteed.

Now let us examine the behavior of statistical techniques when so-called foolproof clues are present. A foolproof clue is an observation that is never encountered from any member of class B and is sometimes (but not necessarily always) observed on members of class A. When present (when it is observed), a foolproof clue is an observation that is a dead giveaway of the fact that we have encountered a member of class A and (in these cases) we can recognize the input as belonging to class A with certainty (with a probability of error equal to zero).

The fact that a specific clue (say parameter v_N) never assumes the value x, whereas members of class A sometimes do, renders $v_N = x$ a foolproof clue. By virtue of the same fact, the likelihood ratio (whenever $v_N = x$) will be infinity, since its denominator is zero regardless of the values assumed by the other parameters whenever the foolproof clue actually occurs. Thus not only will statistical techniques decide correctly that the input belongs to class A (when the foolproof clue occurs) but they will do so without having to make a special case to take into account the occurrence of foolproof clues.

In connection with language translation applications of pattern recognition, it is often desirable to translate a specific word one way at one time and differently at other times depending on context. Similarly in automatic speech transcription (conversion of speech into a phonetically spelled text), a given instantaneous spectrum representing a sound may have to be assigned different symbols in transcription, depending on context. The recognition of many consonants, for example, depends on the identity of the adjacent vowels.

Context is merely a different way of referring to conditional joint probabilities. If the distinction between two stimuli said to be different (because they are caused by different sources) but actually identical by measurement, can only be made upon the condition that the context in which they appear is known, then it would seem that decisions would have to be made on a more extended observation of stimuli. A good example illustrating the use of context in a decision theoretically correct manner is the improvement achieved in the performance of a faulty character recognizer through the utilization of letter digram or trigram frequencies.

The number of samples one must use in estimating probability densities for use in statistical techniques is a difficult problem that is probably least understood, and the excessive number of samples thought to be required is often cited as an important shortcoming of statistical techniques. Often this

charge is based on the tacit assumption that a probability density must be described (and is approximated) by its moments and that a very large number of samples are needed to approximate higher moments. While the latter is true, it is not true that hard to estimate moments must be used to describe the distribution of a set of observations. In response to the question, how many samples does one need of each of the stimulus classes to learn to recognize stimuli correctly, it is impossible to give a pat answer. "We must have a representative set of stimuli" is all that can be said without additional information about the classes in question. One sample per class may be all that is needed if it is known that classes are unimodal (such as Gaussian) and that the modes are well separated from one another and the different classes occupy largely non-overlapping regions of the vector space. More samples are needed as the densities become more overlapping and complex. Few samples, in general, are needed in those regions of the vector space where members of only one class can occur, while a higher density of samples are desirable near the boundaries of the decision regions. One of the fallacies surrounding the adequacy of a sample set for decision making purposes is that a good approximation of the probability densities is essential. It is not essential (although let us not refuse a good estimate if one can be obtained). It is more important that a good estimate of the decision region boundaries be obtained. This often can be obtained without a large sample size. The important point to note is that a good sample set is required for developing a good classification system and that the use of a statistically based "learning" system does not necessarily require a larger number of learning samples than a learning system of a different kind.

Now let us turn our attention to the utilization of known physical properties of members of the same class. For instance, we might define a class as any visual image of a generic type (like all vertically incident aerial photographs of airports of all types), and we wish to make the recognition of this class of images independent of the location, size, and orientation of the image in the visual field. Therefore, the class is really "airports under all possible translation, magnification or rotation" and the physical properties of insensitivity to translation, rotation or magnification should be utilized. There are several possible ways of dealing with a situation of this type. We may envoke some technique of "prenormalization", a many-to-one mapping of the input sensory space which reduces the undesired redundancies without destroying information needed for classification. This method, usually preferred for specific applications, yields a parameter space from the input sensor space. In the parameter space the input class is more simply distributed.

Another way of dealing with the situation is to employ techniques that can accommodate the more complex partitioning requirements imposed on the sensory space by the nature of the input classes. Usually a compromise must be struck between these approaches. It would be fallacious to believe that prenormalization techniques could always be invented to reduce the complexity of the input class distributions to the point where extremely simple-minded partitioning schemes could be invoked.

The need for complex partitions of the vector space has often been questioned and it is often said that successful prenormalization eliminates the need for complex partitioning schemes. Nothing could be further from the truth. Let us think of the target class "industrial complexes". We have steel mills with their long and narrow parallel buildings, oil refineries with their tremendous disarray of odd structures, pipes, and storage tanks, and we have the light industrial plant complexes with their almost residential suburban characteristics.

It is difficult to imagine that all of these varieties of industrial complexes should have the same type of parametric representations no matter what type of prenormalization may be employed. Thus, at least for the reason that members of a class will, in general, fall into many different subclasses, we must use techniques that can process input stimuli under much more general conditions than those afforded by the often invoked Gaussian assumption of class membership distribution in the parameter space.

To make the need for the ability to process multimodal class distributions even clearer, we must remember that just because we (as humans) lump a set of stimuli in the same subclass, we cannot assure that the techniques we employ will also automatically consider these stimuli to be members of the same subclass. In general, there are many more mathematical subclasses than can be explained physically.

2.2 The Economical Storage and Evaluation of Probability Densities

If the input observations, represented by the vector v are known to belong to one of K stimulus classes described by the set of probability density functions, $\{p_k(v)\}$ and have a priori probabilities of occurrence, $\{P_k\}$, then according to statistical decision theory, the optimum decision procedure is to compute K functions of the form given in Equation (2) and to decide that the input stimulus belongs to the class yielding the largest numerical value. This decision procedure minimizes the expected error. P(v, k) is the probability of the joint event that v and an input from class k are observed.

$$P(v, k) = P_k p_k(v) \tag{2}$$

Decision making consists of generating the value of the conditional joint probability density function for any given input vector v, computing the functional forms P(v, k) for each of the K probability densities, and choosing the stimulus

class which yields the largest numerical value. Often this is accomplished by comparing ratios of probability densities with constants. The ratios are called likelihood ratios.

In practice, however, this theory cannot be applied directly to classification problems because the conditional joint probability densities are generally unknown. The only data available on the statistical characteristics of the observations from each class consist of a finite number of labeled samples from each of the K classes or categories. Thus, the probability densities can never be known precisely and, therefore, the preceding decision rule will only approach optimality. Nevertheless, one approach to the solution of the classification problem is to estimate these probability densities using the available data samples and then use these estimated functions to evaluate the likelihood ratios.

The process of estimating the probability densities from labeled samples of known classification can be regarded as "learning" while the evaluation of likelihood ratios at points in the vector space corresponding to an input stimulus is called "recognition".

Probability densities, as any other functions, can be evaluated and approximated by a number of different procedures. In one of these, the function expressed in an analytical form is stored in memory and the numerical value of the function is computed for the specific set of observations represented by the vector v. In another method of evaluating a function we store its values at a sufficiently large number of points of the vector space, determine the stored point nearest to the point v, look up the value of the function at the nearest point and, perhaps, interpolate among stored values of the function near v.

An illustrative example using a function of a single variable may clarify these two methods of computing a function at point v. Suppose the function is

$p(v) = v^2 + \alpha v$. The function p(v) can be evaluated from its argument v according to the first method described by instrumenting the operation of squaring, addition, and multiplication by a constant. These operations can be arranged in the appropriate sequence so that the function p(v) is constructed as an operation to be performed on v. In this case the computer is the operator and the coefficients of the equation must be stored in memory.

By the second method of computing p(v), precomputed values of p(v) can be stored in a "look-up table" in a manner similar to tables of trigonometric functions. When p(v) for a specific value of v must be computed, we enter the table at the entries that straddle the specific value v, look up the stored value of p(v) at these points, and interpolate between them to obtain a sufficiently accurate estimate of the function. If stored values of p(v) are tabulated at sufficiently densely spaced values of v, interpolation is not necessary and one can look up the stored value of the function at the tabulated value "nearest" to the given argument, v.

Because of the complicated nature of the conditional joint probability densities, computations by the first method described above are not economical*. Since the region of the vector space of interest in many decision making problems is small in relation to the total volume of the vector space, the tabulation of values of a probability density at a relatively small number of stored points, judiciously selected for their representative nature, is a more economical method of approximation. Just as the one dimensional probability density p(v) shown in Figure 3 is approximated with a staircase approximation $\hat{p}(v)$ by use of a look-up table, similarly an N-dimensional probability density involving the joint probability of occurrence of N different numerical values can be approximated by the

*Methods of computing coefficients of polynomials in N-dimensions used in discriminating between different classes of data are described in Reference 5.

N-dimensional equivalent of a staircase approximation. Such an approximation of a probability density is a histogram in N dimensions, a generalization to which we will return later.

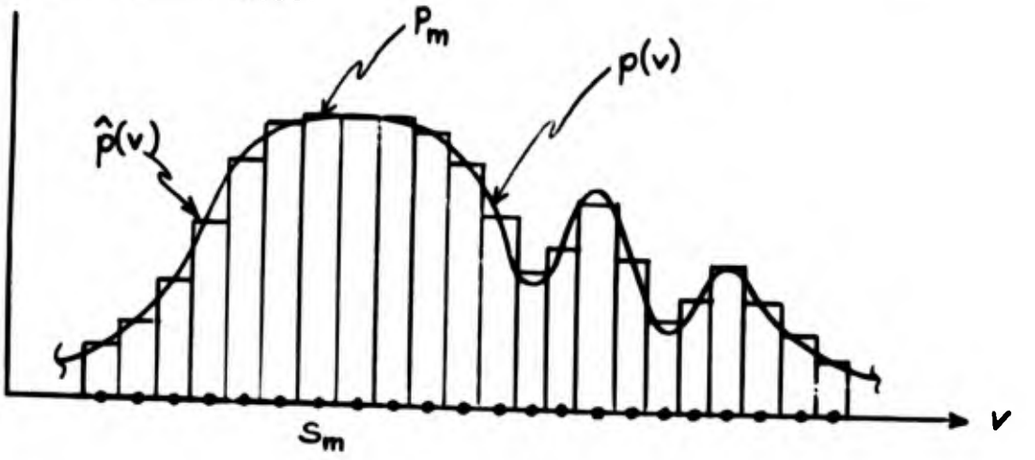


Figure 3. Approximation of a Function of a Single Variable

Since the function p(v) is approximated by a constant in each interval, it is obvious that only the boundaries of the quantiles and the values of the approximation in each interval must be stored. A simple method of evaluating a histogram approximation at an arbitrary point v can be devised. The procedure hinges on the ability to determine simply the identity of the cell or interval, m, in which the input to be classified is contained and then retrieving p_m the corresponding stored value of the approximation.

By storing the location of the centers of the cells as a set of points, $\{s_m\}$, where s_m is the stored center of the m^{th} cell, the interior of an arbitrary cell, i, is readily defined as the locus of points "nearer" to s_i than to any other stored point.

The classification procedure implied by the above argument consists of:

A. Determining the stored point s_i that is "nearer" to the input vector v than any other stored point s_m $(m \neq i)$

B. Retrieving the stored probability density $p(s_i)$ which is approximately equal to p(v).

C. Repeating this procedure for all classes and computing the necessary likelihood ratios, joint probabilities, etc.

From the point of view of minimizing the storage requirements of the recognition device, it is advantageous to minimize the number of stored points, s_m, necessary for the approximation of a given probability density. For instance, fewer points could be used to represent the function in a region where the function does not vary much, and a higher density of stored points could be used where the function varies rapidly. The histogram in Figure 4 illustrates a more economical manner of storing the probability density sketched in Figure 3. This procedure is equivalent to the construction of a histogram with unequal intervals (that is to say, the sizes of the histogram cells are tailored to better fit the distribution of the data).

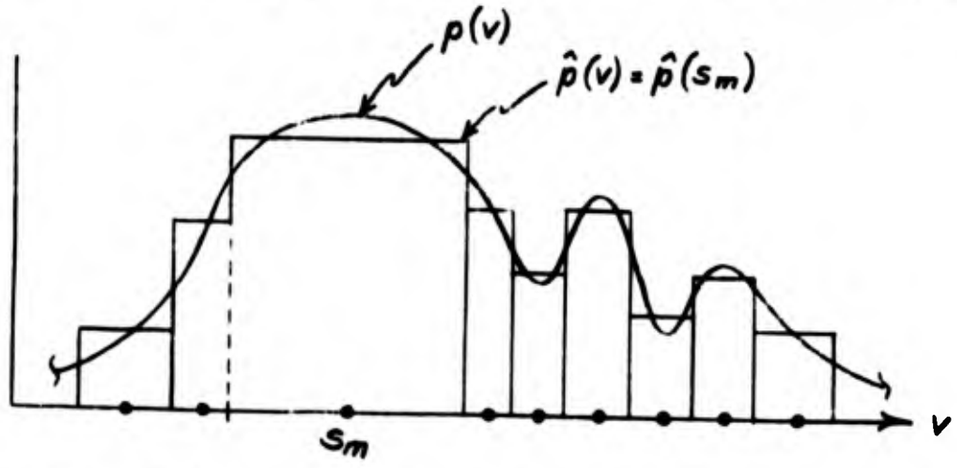


Figure 4. Approximation Requiring Less Storage

One can place the construction of storage-limited histograms with unequal cell sizes on an exact mathematical basis by asking (and solving) questions of the following type: "What is the optimum choice for the location, size, and height of M cells to minimize the expected error between p(v) and its approximation, and $\hat{p}(v)$?" The partial solution to a very similar method of stating the search for the optimum histogram of M cells is given in Appendix 1 of Reference 2.

Since in practice, p(v) is unknown and must be obtained from samples, it is more fruitful to tackle the problem of how to obtain a "good" histogram directly from samples. It is readily appreciated that cells representing the distribution of a set of known samples of class k must be located only in those regions of the vector space where members of class k are observed. In most problems the volume of the region wherein members of a class are contained is a very small fraction of the total volume of the vector space (truncated in accordance with the equipment-imposed limitation on the dynamic range of the variables); thus a significant reduction in the storage requirements can be achieved by having members of the class create and determine the locations and dimensions of the histogram cells. Since the cell centers thus obtained typify the distribution of class k, the stored points $\{s_m\}$ are called "typical samples" of the class.

The interior of an arbitrary cell i in this new histogram can still be defined as the locus of points "nearer" to s_i than to any other stored point s_m $(m \neq i)$. It is merely necessary to modify the distance measures used and to stretch the unit length of our yardstick when we measure "nearness" to a stored point s_m whose cell is wide, while we must shrink the unit length of our yardstick when we measure distance to the center of a narrow cell. A squared distance measure exhibiting the property that the length of a unit distance is dependent on the cell identity, m, and the specific dimension of the space under consideration, n, is expressed by the quadratic form $Q_m(v)$ given in Equation 3.

$$Q_m(v) = \sum_{n=1}^{N} \left(\frac{v_n - s_{mn}}{\sigma_{mn}} \right)^2 \qquad (3)$$

This quadratic form expresses not only the notion that the approximated function varies less in one neighborhood than in another (the location of the neighborhood is indicated by m), but it also expresses differences in the rate of variation of the function which depend on the coordinate direction. In mathematical terms, it is an expression of not only the location but also the shape of the cells of an N-dimensional histogram. In intuitive terms, it expresses the notion that the process of evaluating the probability density can be likened to a process of determining whether or not there is anything in our past experience (as represented by the set of "typical samples" $\{s_m\}$) that is "similar" (as measured by the quadratic form) to the present input that must be classified. Once we find something in storage that is similar to the input, we base the decision on the relative number of past observations of that type from all of the classes. The above quadratic form expresses the intuitive notion that our measure of "similarity" must depend on what we measure similarity to (it must depend on where the input is located in the vector space). A certain difference between parameter values of the input and a stored sample may be judged more significant in one neighborhood of the vector space than in another*.

To allow for the possibility that a new input vector v is not sufficiently near to any of the stored typical samples, we may make use of prior assumptions about the local behavior of the probability densities that are approximated. An often invoked assumption is that the probability density is "well behaved". While this is a qualitative expression, it is often a good assumption when dealing

*This is readily illustrated with an example in which we are interested in target recognition, and target speed is one of the measured parameters. If knowledge of whether or not the target is in motion is important for recognition and classification, then target speed should be a far more sensitive indicator of classification near zero speed than at higher speeds where the question of whether the target is moving or not is no longer in doubt.

with practical problems. Just as one often assumes, for computational reasons, that a function is locally linear (almost any function can be sufficiently well approximated locally by assuming that it is linear in a sufficiently small neighborhood), it is similarly convenient to assume that a probability density is locally Gaussian. It should be emphasized that the assumption is not made that the probability density is Gaussian; the only assumption made is that it is locally Gaussian in functional form.

Thus, the approximation of a probability density has been described above by a set of "typical samples" which serve to identify the neighborhoods where data of known classification has been observed by an estimate of the densities at these points, and by a locally Gaussian decay from those points, when the new input is not near enough to one of the typical samples. The manner of use of this assumption will be discussed later.

2.3 The Adaptive Approximation of Probability Densities from Limited Data

In the method of storage and evaluation of probability densities described in the preceding section, the approximated density was described and stored by means of a set of typical samples and cell shapes determined by quadratic forms specified by means, $\{s_m\}$, and "variances" $\{\sigma^2_{mn}\}$. In the following, an algorithmic technique is described for generating cells from data in an adaptive manner by accepting input samples of known classification sequentially. This algorithmic technique consists of three parts. First the data is analyzed to obtain an estimate of the minimum cell size and shape that will be required. This coarse analysis of the data is called Preanalysis and will be discussed later in this report. The second part of the procedure is to accept data of known classification and to construct an N-dimensional histogram cell structure where the locations and the shapes of the cells are generated in an economical fashion. The third part of the algorithmic procedure operates on the cell

structure thus created and attempts to reduce the number of cells by eliminating cells which contain too few input vectors and by enlarging other cells to contain cells so eliminated. In the description that follows, it is assumed that all of the sequentially introduced inputs are members of the same class and that the classification of these "learning" samples is known.

When the first learning sample is introduced, a cell of pre-chosen size and shape is created and is centered on the first learning sample. The chosen size and shape of the cell is determined by prior analysis of the data from which an estimate of the minimum desired cell size is obtained. The interior of the cell is defined by Equation (4a) (the equation of an ellipsoid in N dimensions) where the squared radii of the ellipsoid are expressed by $\sigma^2_{mn}(t)$ where σ^2_{mn} are the "variances" of the quadratic form and τ^2_N is a control parameter to be discussed presently In Equation (4a), the symbol t signifies the fact that the cell center and its shape are functions of the number of learning samples that have fallen into the m^{th} cell up to the present time. T will denote the total number of inputs to the present.

$$\Omega_m(v, t) = \sum_{n=1}^{N} \left(\frac{v_n - s_{mn}(t)}{\sigma_{mn}(t)} \right)^2 \leq \tau^2_N \qquad (4a)$$

The choice of the initial cell radii (their values for t = 0) is determined from a pre-analysis of the data in which the minimum desired cell size is estimated. Thus, the first input vector becomes the first "typical sample". In addition to the vector, the estimate of the density, given by Equation (4b), is also stored in memory. The density is estimated by the fraction of the total number of input vectors that fall in a cell, divided by the volume of that cell. Except for a constant, depending on the number of dimensions, the volume of the cell is expressed by the product of the "standard deviations" in the quadratic form used to define the boundaries of the cell.

$$p(s_m, t) \quad \frac{t}{T}\left[\prod_{n-1}^{N} \sigma_{nm}(t)\right]^{-1} \qquad (4b)$$

The second learning vector is used to generate a second cell similar to the first if it falls sufficiently outside the first cell. If the second vector falls inside the first cell, the center of that cell is shifted to the center of gravity or mean of the two learning vectors, the shape and size of the cell is adapted from a better knowledge of the local distribution of members of the class than that which was obtained from a prior analysis of members of the entire class, and the local estimate of the probability density is updated accordingly. If the second input vector falls sufficiently outside the first cell, it creates a new cell of size and shape obtained from prior estimates of the minimum desired cell size. If the second vector falls outside the first cell but not a very large amount, it is temporarily stored to be reused at a later time according to a procedure to be described in subsequent paragraphs.

The third and subsequent learning vectors are processed similarly, either generating new cells, updating old cells, or stored temporarily for use later. The cells so generated for each class are located only in the portion of the vector space where examples of the individual classes have been observed. Certain properties of the cell structure so generated will be discussed later.

As learning vectors are introduced sequentially, the cell in the immediate neighborhood of the input vector changes shape, location, and height. It is, therefore, important to examine the time dependent values of the cell's size, shape, and height. If we denote by j the identity of the input vector that fell within a specific cell (say, cell m) and was responsible for the j^{th} iteration of the cell shape updating procedure, and if t denotes the total number of inputs falling in the m^{th} cell to the present, then the "variances" that determine the cell shape are given by Equations (5) and (6).

$$\sigma^2_{mn}(t) = \max\left[\sigma^2_{mn}(0),\ a^2_{mn}(t)\right] \tag{5}$$

$$a^2_{mn}(t) = \frac{1}{t}\sum_{j=1}^{t}\left(v_{mn}(j) - s_{mn}(j)\right)^2 \tag{6}$$

Equation (5) expresses the manner in which the n^{th} coordinate cell radius, $\tau_N \sigma_{mn}(t)$, grows if the sample variance of the t vector in the cell, $a^2_{mn}(t)$, exceeds the initial "variance" $\sigma^2_{mn}(0)$. The cell radius is never allowed to shrink to less than the initial value, $\tau_N \sigma_{mn}(0)$, which is determined from a prior analysis of the data. The quantity $a^2_{mn}(t)$ is expressed in Equation (6) where $v_{mn}(j)$ is the n^{th} coordinate value of the j^{th} input vector falling in the m^{th} cell, $s_{mn}(j)$ is the n^{th} coordinate value of the m^{th} cell center after j contributions to the cell, and t is the total number of input vectors that fell into the m^{th} cell up to the present time. The reason for defining the cell in this way is to encourage the cells to increase in size as more inputs are received, thus keeping the total number of cells used in the approximation of the probability density small, while keeping the representation accuracy high.

To insure that each cell has enough room to grow and to reduce the chance for an overlapping coverage of the same region of the vector space by more than one cell, an "outer" control parameter $\theta(\geq 1)$ is introduced so that a vector v not falling within an existing cell (as defined by threshold τ_N) is used to generate a new cell only if it is outside a larger concentric cell defined by Equation (7).

$$Q_m(v, t) \leq (\theta\tau_N)^2 \tag{7}$$

It is seen that the quantity θ expresses the ratio of the outer to the inner diameter of a "guard ring" within which input vectors neither create new cells nor update old ones.

The input vectors which neither create nor update cells are stored temporarily for later use. As the cells grow in size, these stored vectors can be forced into the existing cell structure without the need to create new cells. The temporary storage lasts until the number of input vectors introduced reaches a specified factor, ω, times the number of cells generated up to this time. That is, if c_1 cells have been generated after t_1 samples have been received, and if the ratio of c_1 to t_1 reaches ω, then the temporarily stored vectors that have fallen into "guard zones" are forced into the cell structure existing at time t_1. The temporary storage process then starts again and continues between time t_1 and time t_2 are disposed of in the same way. This procedure is continued throughout the estimation phase, with

$$t_q = 2^{q-1} c_q \omega \tag{8}$$

Previous analytic and experimental studies[6] of random cell behavior in regions of constant and varying probability densities have indicated that evolution of a satisfactory cell structure is most likely to occur when values of the control parameters τ_N and ω are taken to be approximately $1.4\sqrt{N+2}$ and 4, respectively, with $\theta > 1$. This is further discussed in Section 2.5 below.

After a cell structure has been obtained by the algorithmic procedure described above, we may find that the number of cells so created is larger than the number we would like to have in the N-dimensional generalized histogram. We

may force the reduction of the number of cells created by altering the cell growth controlling parameters τ_N, θ and, to some extent, ω. In most cases, however, a significant percentage of the cells created contain very few input vectors which generally surround the more populous cells. This happens because each cell, after its initial creation, migrates in the vector space and generally tends toward the nearest mode of the probability density to be approximated. The fact that the typical samples acting as cell centers migrate toward the nearest mode (local peak of the probability density) is readily seen from the one-dimensional illustration shown in Figure 5.

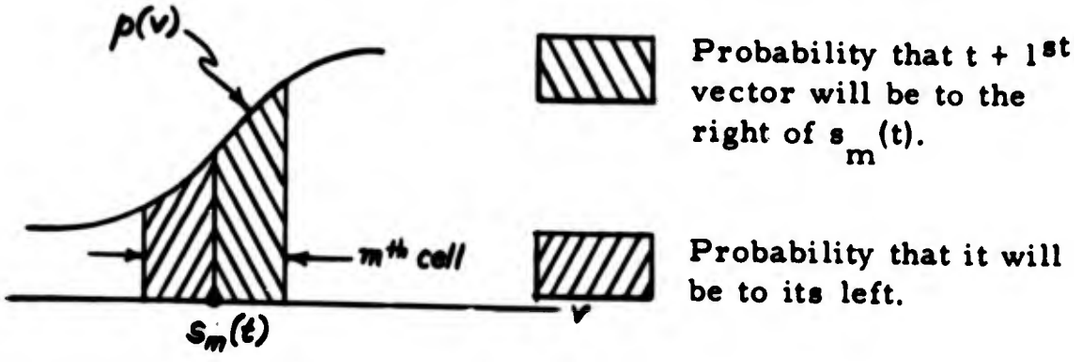


Figure 5. Mode Seeking Property of Cells

This figure shows a small range of the variable v and the probability density p(v) in that interval. The point $s_m(t)$ represents the cell center of the m^{th} cell after t members fell into the cell. The probability is greater that the next input that falls into cell m is to the right of the point $s_m(t)$ than the probability that it will be to the left of that point. This implies that the cell center will move to the right after the $t + 1^{st}$ input falling within the m^{th} cell has been introduced. It is thus seen that cells migrate in the direction of the nearest modes. As cells move toward modes and later inputs create cells at places from which older cells have migrated, there will always be cells which contain few members. The third part of the algorithmic procedure

to estimate probability densities reduces the number of such cells by forcing these cell locations (weighted by the number of contributing vectors) into cells whose contributing members exceed a predetermined number.

Having achieved the first sub-goal of estimating the probability densities of the data samples representing each class, the algorithm attempts to minimize the number of typical samples, subject to the constraints that the probability of error that results when these estimated densities are used in decision making should not be substantially increased. The underlying principles of this part of the algorithm are contained in facts that 1) the probability of error is not changed substantially if the density of a class is not well approximated in a region where densities of all other classes are very small, and 2) that a good approximation of the densities should be maintained near the boundaries of the decision regions which occur where probability densities of two or more classes have similar magnitudes.

First those typical samples or cells are identified which are also covered by cells generated by one or more of the other classes. The common coverage of a region by two cells is identified by the test given in Equation (9).

$$\sum_{n=1}^{N} \left(\frac{s_{an} - s_{bn}}{\sigma_{an}} \right)^2 \leq \text{CTHR} \quad \text{or}$$

$$\sum_{n=1}^{N} \left(\frac{s_{an} - s_{bn}}{\sigma_{bn}} \right)^2 \leq \text{CTHR} \qquad (9)$$

where s_a and s_b are two different typical samples (usually from two different classes) and CTHR is the Coverage THReshold which serves as a criterion of overlap between the two cells. Note that two comparisons must be made since the quadratic form used in measuring distance to cell a is different from that used in measuring distance from cell b.

Those cells from a class that cover regions already partially covered by cells of another class are identified and are not processed further. Those cells of a class, however, which participate in the common coverage of a region with only other cells of the same class [by the criterion of common coverage given in Equation (9)] are lumped into a single larger cell combining the populations of cells of the same class which multiply cover the same region. This procedure is iterated with increasing values of CTHR until the total number of typical samples for all classes is within a previously chosen upper bound. The algorithm is so arranged that various features can be invoked on an optional basis during any analysis.

2.4 Hardware Realization of the Pattern Recognition Computer (PARECOMPUTER)

In most practical applications of pattern recognition, the main requirement is the development of a device which can classify input vectors according to the stimulus class to which they belong. The device is to do so in a time period which is "real-time". Learning how to recognize patterns can usually be performed in the laboratory on collected data by means of a general purpose computer. For this reason the essential equipment that must be constructed for the practical application of pattern recognition techniques must include:

A. The "Observation System" (See Figure 2) which represents the input environment parametrically as a vector in N dimensions.

B. The "Recognition System" which operates on the input to determine the probability densities of each of the stimulus classes at the point in the vector space represented by the input.

C. Displays and controls necessary for the interpretation and control of the proper functioning of the equipment.

D A means for automatically collecting data, already represented as vectors, for use in an "off-li e" general purpose computer "Learning Machine"

E In addition, provisions for entering the adaptively computer joint probability densities of each of the stimulus classes into the "Recognition System" must be provided to facilitate the rapid updating of the "Recognition System" as new information becomes available through a continuing data collection program

Since the engineering aspects of the vector representation can only be discussed in the context of a specific application, and since the "Lear ... Machine" can be constructed adequately by programming the adaptive technique of approximating probability densities from limited data discussed earlier, only the practical implementation of the ...nition System" needs to be discussed here.

Figure 6 illustrates a digital realization of the recognition system that functions according to the pri ...es described in the preceding sections of this report. The Pattern Recognition Computer, PARECOMPUTER, is a special purpose digital computer that is capable of evaluating in just a few milliseconds the joint probability densities of an arbitrary number of stimulus classes at any point of the vector space of arbitrary number of dimensions. The only constraint on the machine's capabilities is that the product of the number of dimensions of the vector space and the number of "typical samples" should not exceed the machine's storage limitation, regardless of the distribution of "typical samples" among the various stimulus classes.

While a detailed discussion of the PAR COMPUTER is beyond the scope of this report, the discussion of certain of its design features is pertinent because it influences the nature of the approximation of joint probability densities implemented by the machine.

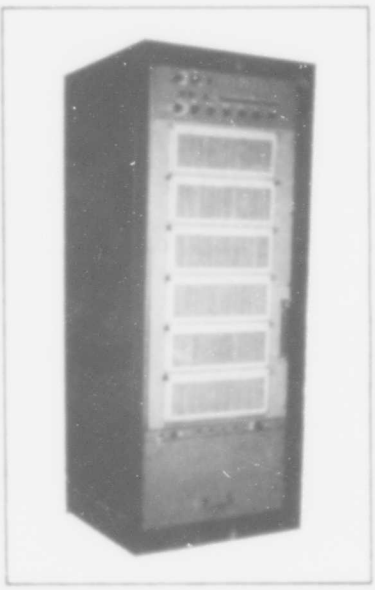


FIG. 6. The Pattern Recognition Computer (PARECOMPUTER)

It was stated before that the input vector is compared with each stored "typical sample" (by means of quadratic forms associated with each) in order to identify the "typical sample" nearest to the input vector and thus to identify the N-dimensional histogram cell which contains the input vector. After the identity of the histogram cell is established, the stored estimate of the probability density applicable in that cell for that stimulus class is retrieved from storage. If the quadratic form, $Q_i(v)$, (measuring the "distance" between the input and the "nearest", i^{th}, "typical sample") is too large, signaling the fact that the input is not near enough even to the nearest "typical sample" to permit the use of an estimate of the probability density corresponding to the i^{th} cell, the assumption that the density is locally Gaussian is invoked. To facilitate its implementation, a step-wise approximation of the locally Gaussian behavior is employed in the PARECOMPUTER. This is illustrated in Figure 7.

The locally Gaussian decay with "distance" from the nearest cell center, s_i, is illustrated in Figure 7a where $\hat{p}(s_i)$ is the density estimate at the center of the i^{th} cell. By working with the natural logarithms of probability densities in the PARECOMPUTER and by storing the logarithms of the local estimates of probability densities, we obtain the linear relationship between the logarithm of the estimated probability density and the quadratic form, shown in Figure 7b. A step-wise approximation of this linear relationship readily lends itself to implementation by the exclusive employment of shift operations of binary numbers. If the log probability density is decremented by $\log \beta$ for every α change in the numerical value of the quadratic form, the step-wise approximation of the locally Gaussian behavior illustrated in Figure 7c is obtained. With this method of approximation, the pre-computed estimate of the probability density is retrieved from memory whenever the input vector is within α of the cell center (if $Q_i(v) < \alpha$), and this estimate is Gaussianly decremented in a step-wise fashion for $Q_i(v) > \alpha$.

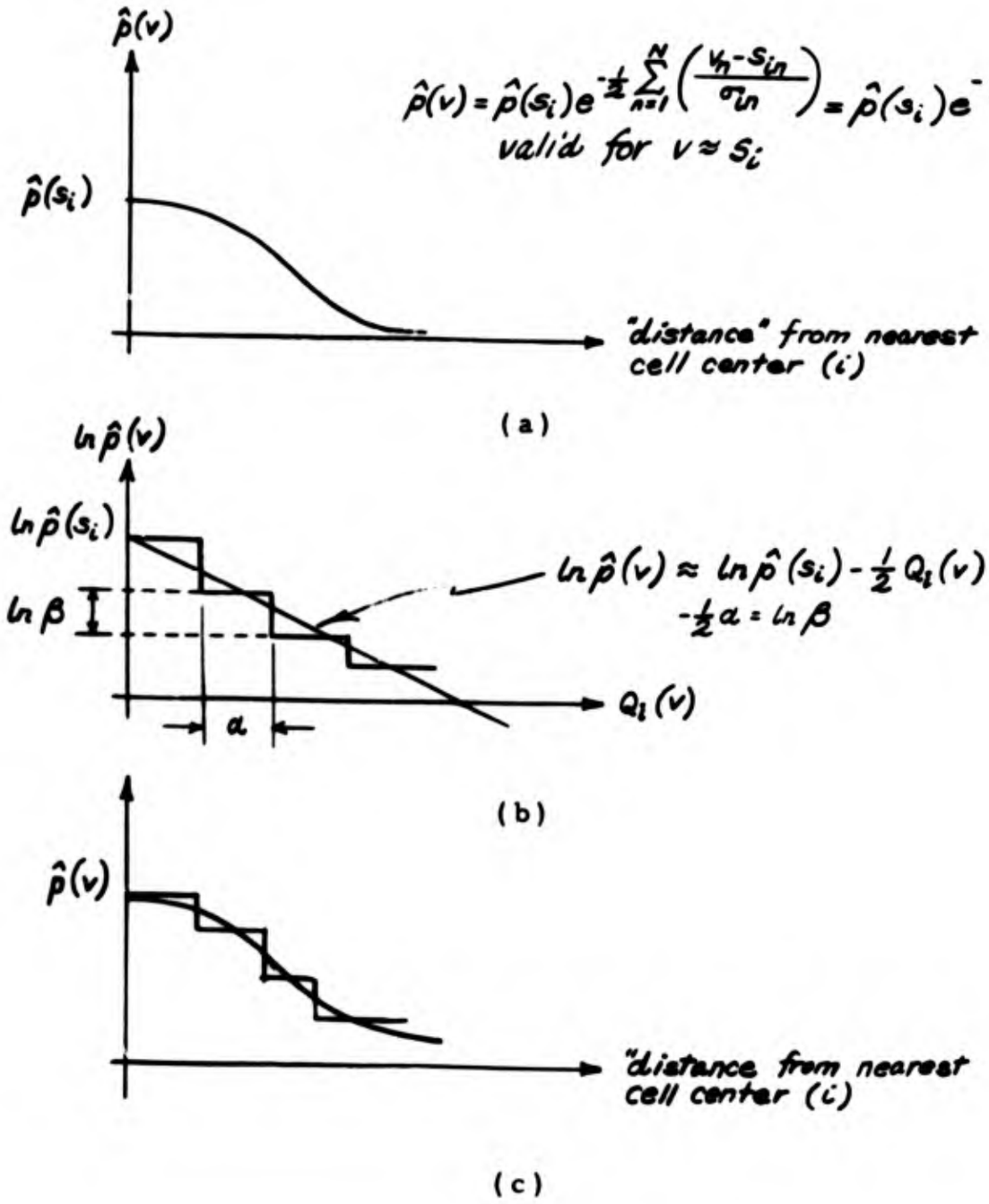


FIG. 7. Stepwise Approximation of Locally Gaussian Behavior in the PARECOMPUTER

Operationally, from the value of the density at s_i the quantity $\log \beta$ is subtracted as many times as α can be subtracted from $Q_i(v)$ without obtaining a negative number.

This method of approximation of a joint probability density of N-variables (for a two dimensional case) would take the shape of a terraced surface similar in appearance to a rice paddy, shown in Figure 8. That is to say, in different regions of the two dimensional plane representing the combination of two parameter values, the surface of the probability density approximation would be flat, but in each region the density would have a different value. . The locus of points where the approximated density has the same value is the region within which points are nearer to the "typical sample" contained within that region than to any of those samples located on the exterior of the region in question. The step-like decreases indicate the step-wise approximation of the locally Gaussian behavior of the approximated probability density at points distant from even the nearest typical sample. Of course, N-dimensional probability densities can no longer be pictured in the same way, but the mathematical representation of the approximation can be handled with equal ease.

The pattern recognition technique and its implementation, briefly described above, have been applied to numerous practical problems successfully. Hardware operating according to these principles has been in use since 1963.

2.5 Some Properties of the Learning and Recognition Techniques

A pattern recognition technique can be regarded as adaptive if, during the sequential introduction of learning samples, the Recognition System can be updated as a new sample is introduced without recourse to all preceding samples. One would expect that in an adaptive system the Recognition System (or the estimated probability densities) would depend not only on the set of learning

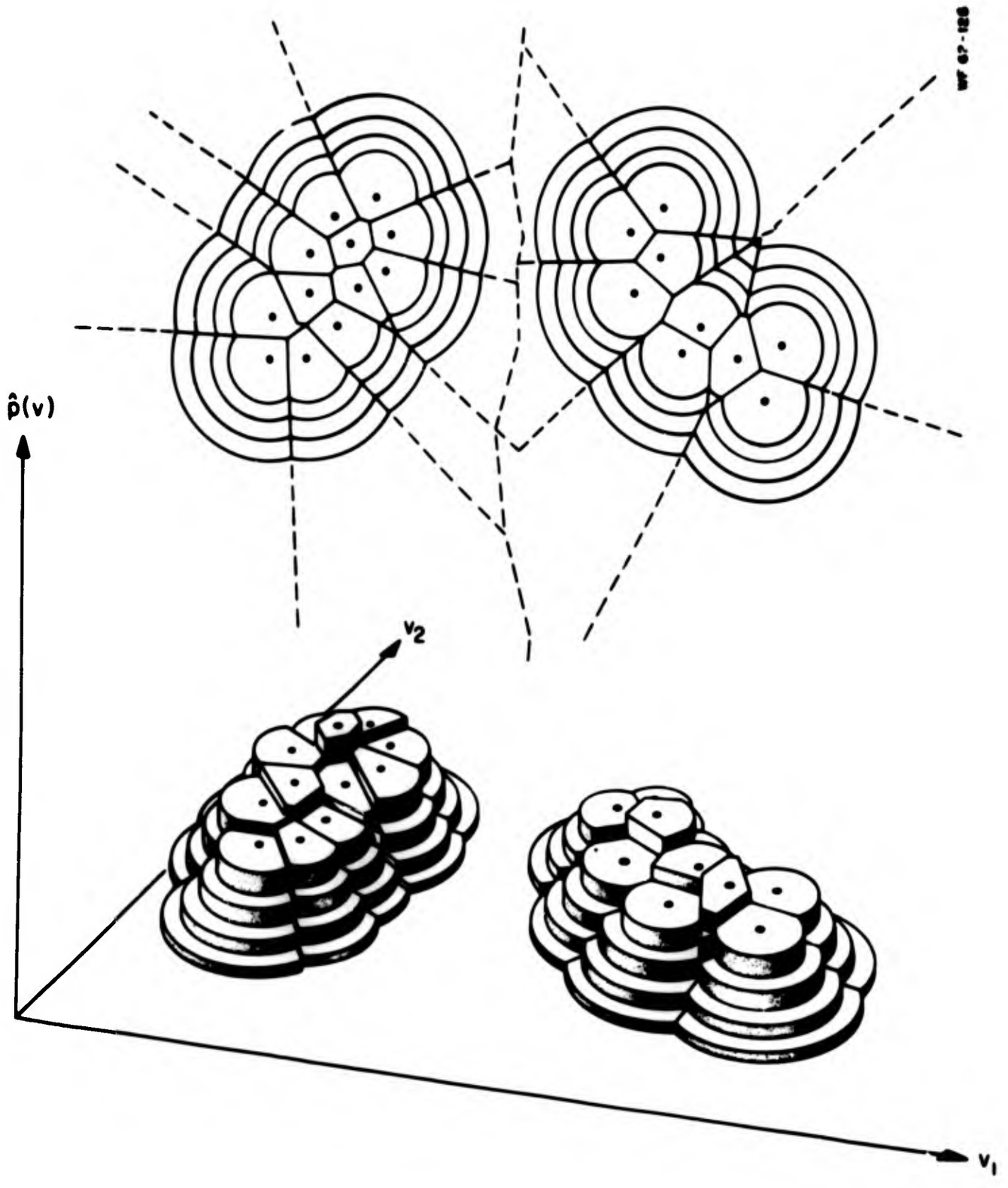


FIG. 8. Approximation of a Function of Two Variables by the PARECOMPUTER

samples but also on the order in whch they were introduced. The order of their introduction can be expected to influence the pattern recognition system's performance. Since the optimum system would depend only on the learning samples on which it is based and not on the order in which they are introduced, a good technique should be as insensitive to the order in which learning samples are introduced as possible.

It is also desirable that a quality measure be calculated to signal whether or not the estimation of the probability densities from the learning samples is reliable. Such a quality measure is discussed in Section 2.6 below.

While the initial cell size in the estimation procedure is due largely to conjecture and to engineering judgment, the control parameters, τ_N, θ and ω which govern the mechanism of cell shape adaptation can be determined from mathematical considerations. In the following paragraphs these three properties, cell growth control, estimation quality criteria, and learning sample order dependence will be discussed briefly.

The method of determining the cell growth controlling parameters, τ_N θ and ω is presented in Appendix II of Reference 6. Here only the key thoughts and the results will be recapitulated. Desirable properties of the cell growth mechanism are that:

A. If the cell is in a region of constant probability density, it should expand rapidly until it covers the region of constant density.

B. If the cell is in a region of non-constant probability density, it should not expand rapidly.

C. A cell's volume might be considered optimum if it is as large as possible and still estimates the probability density in a consistent manner with other estimates using smaller cells.

If an ellipsoidal cell contained a uniform distribution of points, the density of the projections of these points onto any vector dimension (normalized with respect to the radius of the ellipsoid in that dimension) is illustrated in Figure 9. The normalized variable is ν. It is seen that as the number of dimensions increases, the probability densities of the coordinate values become more and more Gaussian in appearance, (they are not Gaussian, however). The cell size will grow in the n^{th} dimension from its initial value, determined by $\tau_N \sigma_{mn}(0)$, if the inequality given in Equation (10) is satisfied; that is to say, if the cell sample variance exceeds the initial "variance".

$$a^2_{mn}(t) > \sigma^2_{mn}(0) \tag{10}$$

The probability that this should occur can be calculated with the result that cell growth will occur with probability 0.5 if Equation (11) is satisfied.

$$\tau_N = \sqrt{N+2} \tag{11}$$

The choice of τ_N also determines the average number of observations in a cell before cell growth can be expected to begin. Of course, if cell growth is to start after just a few inputs are contained in a given cell, then τ_N should be chosen larger than the value given in Equation (11). The larger τ_N the sooner cell growth will start. If we denote by β^* the factor by which τ_N should be chosen larger than the value given in Equation (11), we can calculate the value of β^* versus the number t^* of vectors in a cell before cell growth can be expected to commence. This curve, shown in Figure 10, indicates that β^* should be chosen on the order of 1.4. Thus $\tau_N \approx 1.4\sqrt{N+2}$. It is readily appreciated that the number of vectors in a cell should exceed the number required for the initiation of cell growth before those input vectors that fell in "guard rings" during learning are forced into the existing cell structure. Since cells

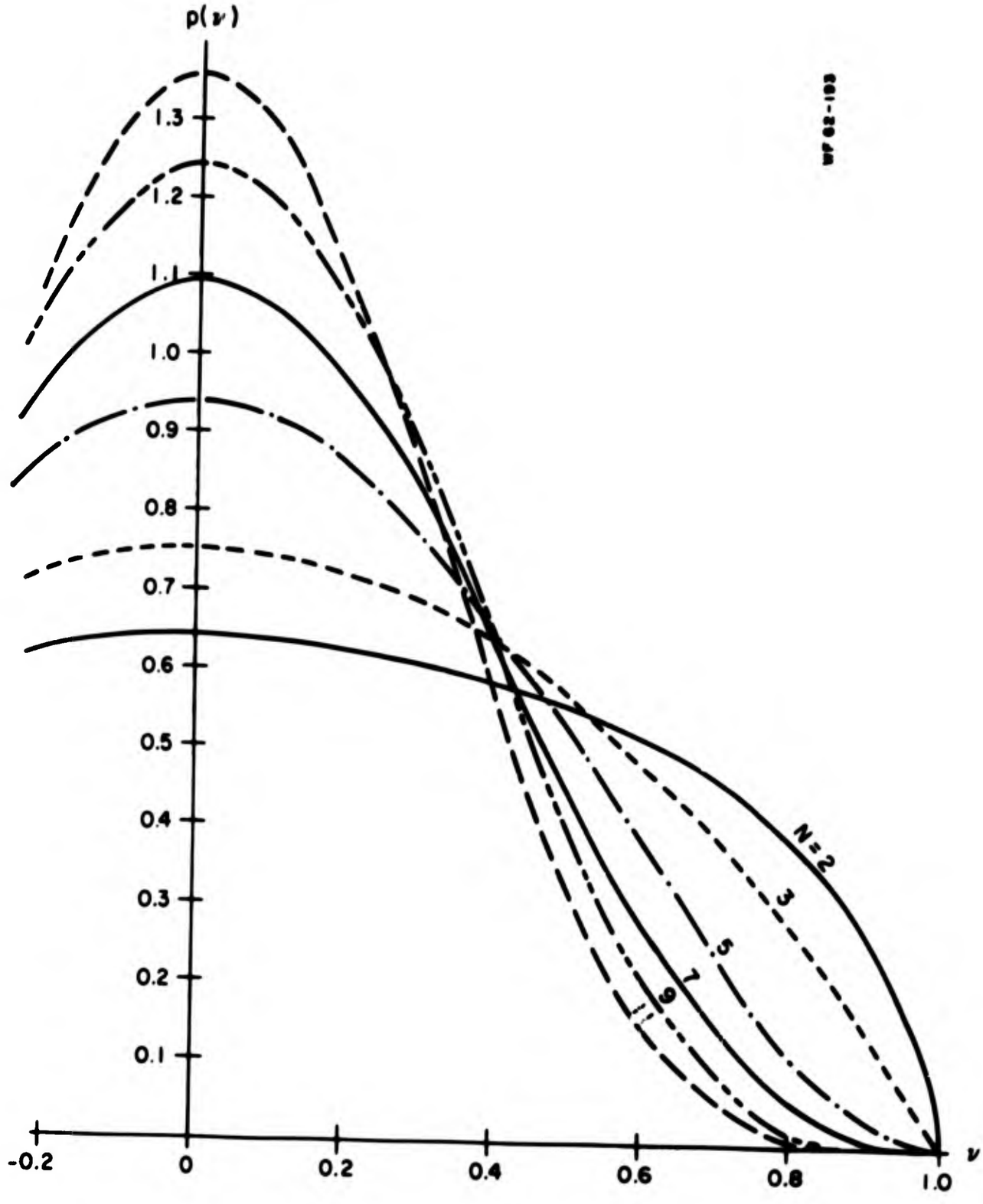


FIG. 9. The Probability Density of One Coordinate of a Uniform Distribution of Points in the Interior of an Ellipsoid of N Dimensions

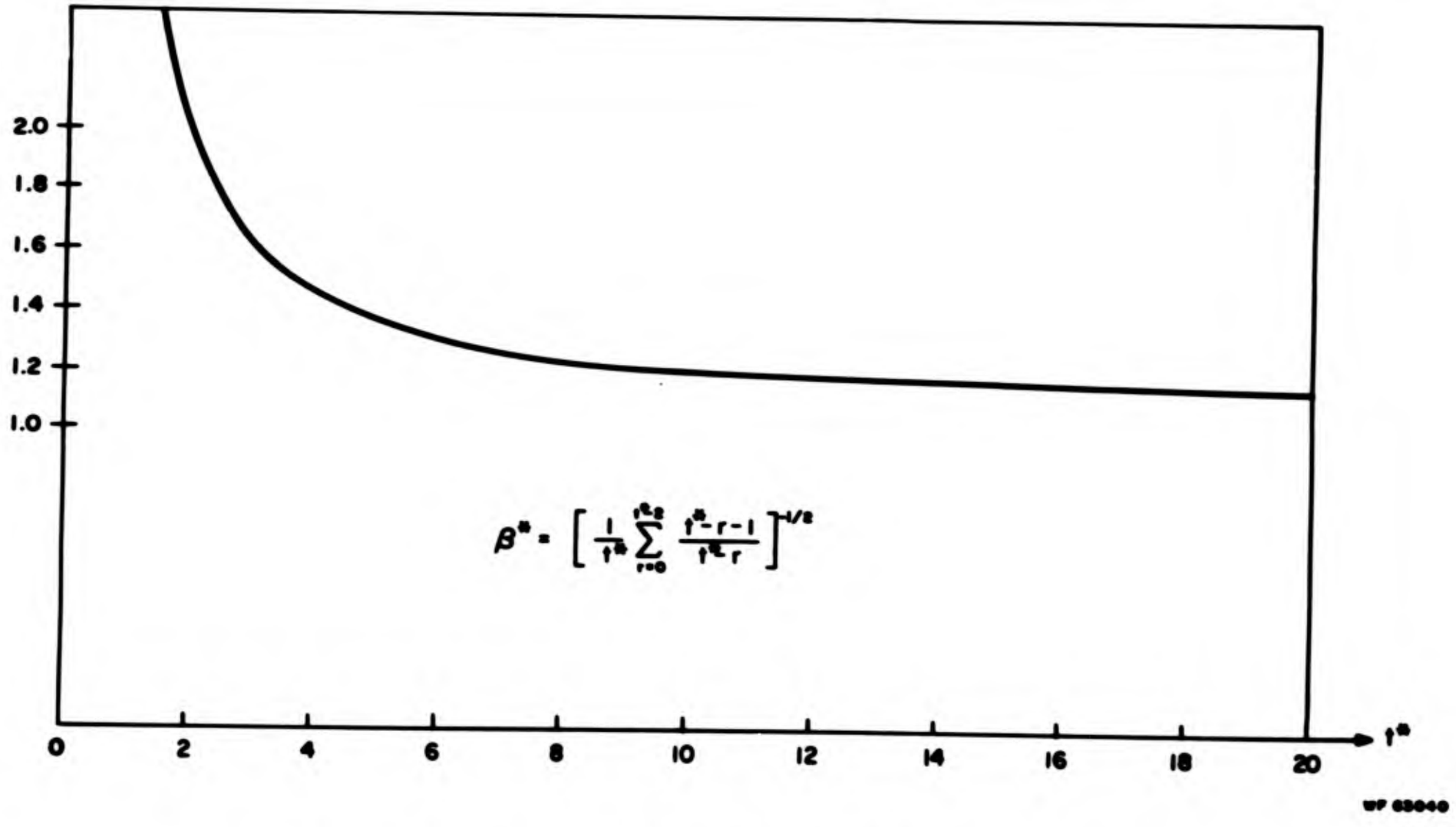


FIG 10. Curve Used in Selecting the Control Parameter τ_N

near the modes of the probability density are cells with a larger than average number of members, they will begin to grow before the majority of the cells have collected t* members. It is reasonable to expect that in many instances the cells located near the modes of the distribution will have grown to their maximum limit by the time an average of t* points have been processed for each of the cells in the entire cell structure. A reasonable choice of the control parameter ω is therefore $\omega = t^*$.

In any automatic decision making device the user must concern himself with the question of how much to rely on the decisions rendered by the machine. While the user of the machine will doubtless form his own opinions about the quality of the decisions rendered, it would be desirable for the machine to indicate the reliance the user should place on the quality of its decisions. A Decision Quality Indicator would be a useful diagnostic measure with which the decision making procedure could be analyzed so that improvements could be made either by increasing the number of learning samples or by altering the choices of minimum cell size and cell growth control parameters. Two useful decision quality indicating measures have been investigated. One of the quality indicators, I_1, is an indicator of the probability that the decision rendered is a correct one. The second indicator, I_2, is a measure of whether or not the decision is optimum. There is a rather subtle difference between these two measures. If the probability, after the vector observation is made, is higher that v is a member of class A than that it is a member of class B, the optimum decision is that v is most likely a member of class A. If the true probability densities are unknown to us and we use estimated densities to arrive at the same decision, our decision is optimum; however, it is not necessarily correct. The probability that the optimum decision is wrong may be high. For instance, if the probability that v is a member of A is 0.6, and that it is a member of B is 0.4, the optimum decision would be that v is a member of A,

but the probability that the optimum decision is wrong is 0.4, a rather high figure. The quality indicator I_1 would state, in the above example, that the machine decision is correct with probability 0.6.

If the estimate of the density of A is larger than the estimate of B (so that we would decide in favor of class A), we may inquire, "What is the probability that the actual density of A is larger than the actual density of B?" In this way we could determine the probability that the decision we make is an optimum one. The quality indicator, I_2, would state that the machine's decision is the "optimum" decision with probability (say) 0.9, but it would not say whether or not the decision is correct. A more detailed discussion of quality indicators is given later in this report.

The algorithmic procedure for estimating probability densities is obviously dependent on the order in which pattern samples are introduced. Since in practical problems the data from which "machine learning" must be accomplished will be ordered, it is important to investigate the order dependence of the approximation technique. In a problem of sonar target classification, for instance, the data is ordered by the methods used in data collection. If a sonar target is tracked, for example, the sequence of sample vectors will all be samples of a target seen at a given relative bearing, thus corresponding to only one of the conditions in which targets can be observed. In another instance the sequence of echoes may be due to a different target type or to a different target aspect. Thus a given sequence of sample vectors will not correspond to a random selection from the total target class population; it will, instead, correspond to only one subclass of the target class of interest.

The above data collection technique gives rise to the question of whether or not all observed samples should be mixed and randomized before application of the adaptive estimation technique. Alternately, the question arises

whether the order dependence of the technique is sufficiently weak to permit processing collected data as it is obtained.

To determine the nature of its order dependence and of its mode seeking features, a series of computer experiments were conducted. The probability density estimation method was that described in preceding sections of this report without the use of cell growth. The method of using the estimate so obtained was somewhat different in implementation, but this difference does not invalidate the results obtained. The experiments and results are described in Appendix I

As a result of a comparative experiment in which samples from a bi-modal distribution were introduced to the adaptive approximation technique completely randomized in one case and one mode at a time in the second case, (by separating members of the two modes prior to their introduction to the approximation technique), the tentative conclusion was reached that the order dependence of the "machine learning" technique is not as severe as originally expected. The probability densities of the samples and of the resulting approximation when the samples were mixed and when they were taken in the order in which they were generated are shown in Appendix I.

When the PARECOMPUTER is used to classify unknown observations, a meter on the console provides an indication of the reliability of the decisions made. The quantity displayed is directly related to the estimated probability of a wrong decision conditioned on each observation. The topic of such reliability indicators is discussed further in Section 2.6.

2.6 Measures of the Quality of Machine Learning and Machine Decisions

As discussed previously, machine learning may usually be interpreted as automatic estimation of the probability density functions (pdf's) and the a priori probabilities for each class. It is desirable while performing learning that one have some measure of the quality of the estimation job being done, if only to know when one can afford to stop sampling. And, while using a machine to make automatic classifications, it is desirable to have a measure of the quality of the decisions being made so that actions based on these decisions be taken with the proper degree of confidence.

There are two primary types of measures of performance quality that have been considered here. The first of these is the "sureness" one has that the decisions made will be (with present estimates in the case of learning) or are (in the case of operational classification) correct. The second type of quality measure is the "sureness" one has that the decisions made will be (learning) or are (classification) optimum. There are various quantities one may use as a measure of "sureness" such as probability, estimated probability, confidence level, etc. In this section a few such measures will be discussed and illustrated by examples.

We note in passing that the viewpoint taken here is possible only because one of the earliest obstacles to pattern recognition has largely been overcome, namely the problem of collecting large bodies of data on class members. Only recently somewhat analogous studies were conducted to devise criteria for judging when enough data had been processed during learning such that 90%, say, of all the reference vectors (typical samples), that would be generated if learning were continued indefinitely, had already been generated. For example, Van Meter[7] developed an occupancy theory model to help explain the learning curves encountered in certain speech processing. Similarly, an important part

of testing the SPEAR technique was the generation of such learning curves showing the number of typical samples generated as a function of the number of vectors processed[8]. In the infancy of the pattern recognition field, data collection was so expensive that one could not afford to do more than worry about how well the references generated represented the classes under consideration. This point is further brought out by Highleyman's consideration of partitioning a sample of fixed size between design (learning) and test phases of a pattern recognition machine[9], and by the small sample sizes used in some early pattern recognition experiments[10]

The most obvious performance measures of the first type are error probabilities. Although error probabilities are usually stated for average or long term performance they can also be conditioned on specific observations.[11] For example, suppose an observation x is made on an object known to belong to one of K classes, i.e., the k^{th} class has been selected with probability T_k $\left(\sum_{k=1}^{K} T_k = 1\right)$ and the (vector) observation x has been made according to a class probability density function p(x; k). Then the optimum processing of the observation x is the computation of the set of a posteriori probabilities for each class.

$$\Pr\{j \mid x\} = \frac{T_j p(x; j)}{\sum_{i=1}^{K} T_i p(x; i)}, \quad j = 1, 2, \ldots, K \tag{12}$$

and deciding in favor of the class with the largest a posteriori probability. This decision will be in error if an x has been drawn which yields a decision in favor of class j, $j \neq k$ where k is the class from which x was in fact taken. The probability of an error given the specific observation x is then

$$\Pr\{\text{error} \mid x\} = 1 - \max_{j=1,2,\ldots K} [\Pr\{j \mid x\}] \qquad (13)$$

An important fact may be deduced from Equation (13). This fact is that if the average error rate is to be small, there must be a (perhaps not simply connected) region in the observation space with a large probability measure such that at every point in the region one a posteriori probability dominates the sum of all other a posteriori probabilities. Or, to state this fact another way, for a low average error rate a large fraction of the observation space must be associable point-by-point with specific classes in an almost unique way.

Furthermore, it is apparent from Equation (13) that in order to compute such conditioned error probabilities one must have the complete set of a priori class probabilities $\{T_k\}$ and class pdf's $\{p(x; k)\}$. But it is intrinsic to the pattern recognition problem that some or all of these quantities are unknown and are often the quantities being estimated during the machine learning phase. Therefore, in practice, one can give only an estimate of the conditioned error probability as an indication of "sureness" that the decision made is correct. Although one might devise a separate estimation procedure for the conditioned error probabilities, it is much easier to simply substitute the estimates $\{\hat{T}_k\}$ and $\{\hat{p}(x; k)\}$ into Equation (12) where ever the corresponding true quantity is unknown. * Special cases of this type of decision quality measure are included in the output of the ASSC III recognition program**, and in the display of the PARECOMPUTER.

---

* Throughout this section the pointed circumflex symbol "^" over a quantity will mean an estimate of the true value of the quantity.

**ASSC III is a general purpose computer which incorporates the techniques described in previous sections. See 6 for descriptions of other programs useful in pattern recognition.

Of course, the precision of such an estimate will, in general, be high only for observations x such that one a posteriori probability dominates the sum of all others, i.e., for an x such that the conditioned error probability is, in fact, low. Therefore, the estimated value of the conditioned error probability is also a measure of the confidence one may have in the estimate, with low values implying high confidence.

Although in the classification phase (actual field operation) one is most interested in being sure that the decisions made are correct; while performing learning one is more interested in developing a classification system which will be optimum. Implementation of the (usual) optimum decision rule requires a complete knowledge of the set of the class a priori probabilities $\{T_k\}$ and class pdf's $\{p(x; k)\}$ and (ignoring varying costs of wrong decisions) consists of deciding in favor of the class j for which the a posteriori probability is maximum or, equivalently, deciding in favor of the class with maximum (weighted) likelihood L(k; x). L(k; x) is a function of the class index k and is numerically equal to $T_k p(x; k)$ for a given observation x. Since it is fundamental to the pattern recognition problem that the true value of L(k; x) is unknown, decisions must be based on estimates of the set of likelihood values. (For the purposes here, however, the term "optimum" will refer to the decision rule which assumes a knowledge of the true probabilities and pdf's).

Now at a particular point x_o in the observation space, the set of likelihood estimates $\{\hat{L}(k, x_o)\}$ lead to the optimum decision if $\hat{L}(j; x_o) = \max_{i=1, 2, \ldots, K} \hat{L}(i; x_o)$ when in fact $L(j; x_o) = \max_{i=1, 2, \ldots K} L(i; x_o)$. In that case, at the point x_o, the set of estimates are of high quality. If the set of points at which the estimates are of high quality has a sufficiently large probability measure, then the total estimation result can be judged to be of high quality. Note that one cannot expect

(or even hope) for high quality estimates at every point in the observation space since there will always be points, for some estimation or learning samples of finite size, where the maximum estimated likelihood will not correspond to the maximum true likelihood. In such a case the order of the estimates will be said to be interchanged and, at every point x in the observation space, since $\{L(k; x)\}$ are random variables taking on different values for different randomly chosen learning or estimation samples, the j^{th} value of the estimated likelihood function will be the maximum value with a probability $\pi_j(x)$. Hence, an acceptable measure of the quality of the estimation performed is the probability that, for a randomly selected test observation, an optimum decision will be made*.

To further develop a realistic criterion for judging the quality of machine learning by pdf estimation, let

P and π_o be selected numbers close to one (14a)

$$R_k = \{x \mid L(k \mid x) = \max_{i=1, 2, \ldots, K} L(i \mid x)\} \tag{14b}$$

$$N_k = \{x \mid x \in R_k;\ \pi_k(x) = Pr\{\hat{L}(k \mid x) = \max_{i=1, 2, \ldots K} \hat{L}(i \mid x)\} \geq \pi_o\} \tag{14c}$$

P_k = probability of an observation drawn from the k^{th} class falling

$$\text{in } N_k = \int_{N_k} p(x \mid k)\, dx \tag{14d}$$

$$p(x) = \text{pdf of a randomly selected observation } x = \sum_{k=1}^{K} T_k p(x; k) \tag{14e}$$

---

* In the first presentation of these ideas[12], an additional "generalization" was included to allow for $\nu \geq 1$ observations from the same class. This has been dropped here for simplicity.

Now the regions N_k, $k = 1, 2, \ldots, K$ depend upon the particular learning or estimation sample size n used and, in general, will increase as n increases. Hence, the sequence of sets of P_k will also be increasing with n.

Now the probability that a randomly selected observation will be classified according to the optimum decision rule, when the class pdf's and/or a priori probabilities have been estimated, may be expressed in terms of the sets and quantities defined in Equation (14).

$$\Pr\{\text{opt. decision}\} = \int \Pr\{\text{opt. decision} \mid x\}\, p(x)\, dx$$

$$= \sum_{k=1}^{K} \int_{R_k} \Pr\{\text{opt. decision} \mid x\}\, p(x)\, dx$$

$$= \sum_{k=1}^{K} \int_{R_k} \pi_k(x)\, p(x)\, dx \qquad (15a)$$

Although Equation (15a) is exact and might be used as a measure of estimation quality, it usually requires some fairly involved computation. It is much more useful to have a simple test that the learning machine can apply in a very short time. In this way the learning procedure suffers little interruption and little instrumentation or programming is required. Therefore, it is desirable to use an accurate approximation or lower bound to (15a). If the integrals in (15a) are taken only over the regions $N_k \subset R_k$, such a lower bound is obtained. This and the succeeding lower bounds below are close to (15a) if the classification error probabilities are low, i.e., if

$$\int_{R_k} [p(x) - T_k\, p(x;k)]\, dx << 1.$$

Thus,

$$\Pr\{\text{opt. decision}\} \geq \sum_{k=1}^{K} \int_{N_k} \pi_k(x)\, p(x)\, dx$$

$$\geq \sum_{k=1}^{K} \pi_o \int_{N_k} p(x)\, dx$$

$$\geq \sum_{k=1}^{K} \pi_o P_k T_k \tag{15b}$$

Further, if learning is continued, i.e., n is increased, until $P_k \geq P$ for k=1, 2, . . . , K, the probability of an optimum decision may be bounded below by:

$$\Pr\{\text{opt. decision}\} \geq \pi_o P \sum_{k=1}^{K} T_k = \pi_o P \tag{15c}$$

Of course, by use of the lower bound (15b) or requiring that $P_k \geq P$ so that (15c) holds, a somewhat larger sample size is normally required in order that a sufficiently high probability of an optimum decision is assured. However, for an application with a low error rate, the increase in required learning sample size should be moderate. The use of Equation (15c) has the sometimes useful characteristic of being uniform for all classes and for an application with a low error rate the expression (15c) is not significantly less than (15b).

Throughout the development of (15) as with Equation (13), the fact has been ignored that many or all of the required quantities are intrinsically unknown. But, since estimates of these unknown quantities are available, these estimates may

be used where necessary. When estimates are so used, the resulting expression is only an estimate of the probability of an optimum decision. If Equation (15a) were used as a measure of learning quality, with estimates used in place of the true quantities, then the precision of the estimate thus obtained would be subject to severe question. This is because it is usually difficult to obtain an accurate estimate of all the necessary quantities over the entire observation space.

For example, it is in general difficult to accurately estimate $\pi_k(x)$ over the entire region R_k with practical sample sizes because near the boundaries of R_k the relative magnitude of $\pi_k(x)$ can be quite small. And, in fact, the region R_k can only be estimated (for the goal of making decisions based on single observations, a determination of the set $\{R_k\}$ is really the whole game) so that evaluation of (15a) using estimates is doubly subject to error.

However, one can be much more confident that an inequality such as (15b) or (15c) is in fact satisfied if $\hat{\pi}_k(x)$ and $\hat{P}_k$ satisfy the necessary requirements on $\pi_k(x)$ and P_k. Specifically, let

$$\hat{R}_k = \{ x \mid \hat{L}(k \mid x) = \max_{i=1, 2, \ldots K} \hat{L}(i \mid x) \} \tag{16a}$$

$\hat{\pi}_k(x)$ = the estimate of $\pi_k(x)$ obtained by substituting the estimates $\{\hat{p}(x;k)\}$ in place of the true quantities $\{p(x;k)\}$ in the expression for $\pi_k(x)$ (16b)

$$\hat{N}_k = \{ x \mid x \in \hat{R}_k ; \hat{\pi}_k(x) \geq \pi_o \} \tag{16c}$$

$$\hat{P}_k = \int_{\hat{N}_k} \hat{p}(x;k)\, dx \tag{16d}$$

($\hat{P}_k$ is simultaneously an estimate of $P_k = \int_{N_k} p(x;k)\, dx$ and of $\int_{\hat{N}_k} p(x;k)\, dx$ = the

probability that an observation from the k^{th} class falls in the region $\hat{N}_k$). It must be kept in mind that the regions and quantities defined in (16) depend on the particular sample used for learning and that in general the precisions of these estimates increase with the learning sample size n.

One can compute an estimate of the probability of an optimum decision by simply substituting estimated quantities onto (15a) wherever necessary; thus,

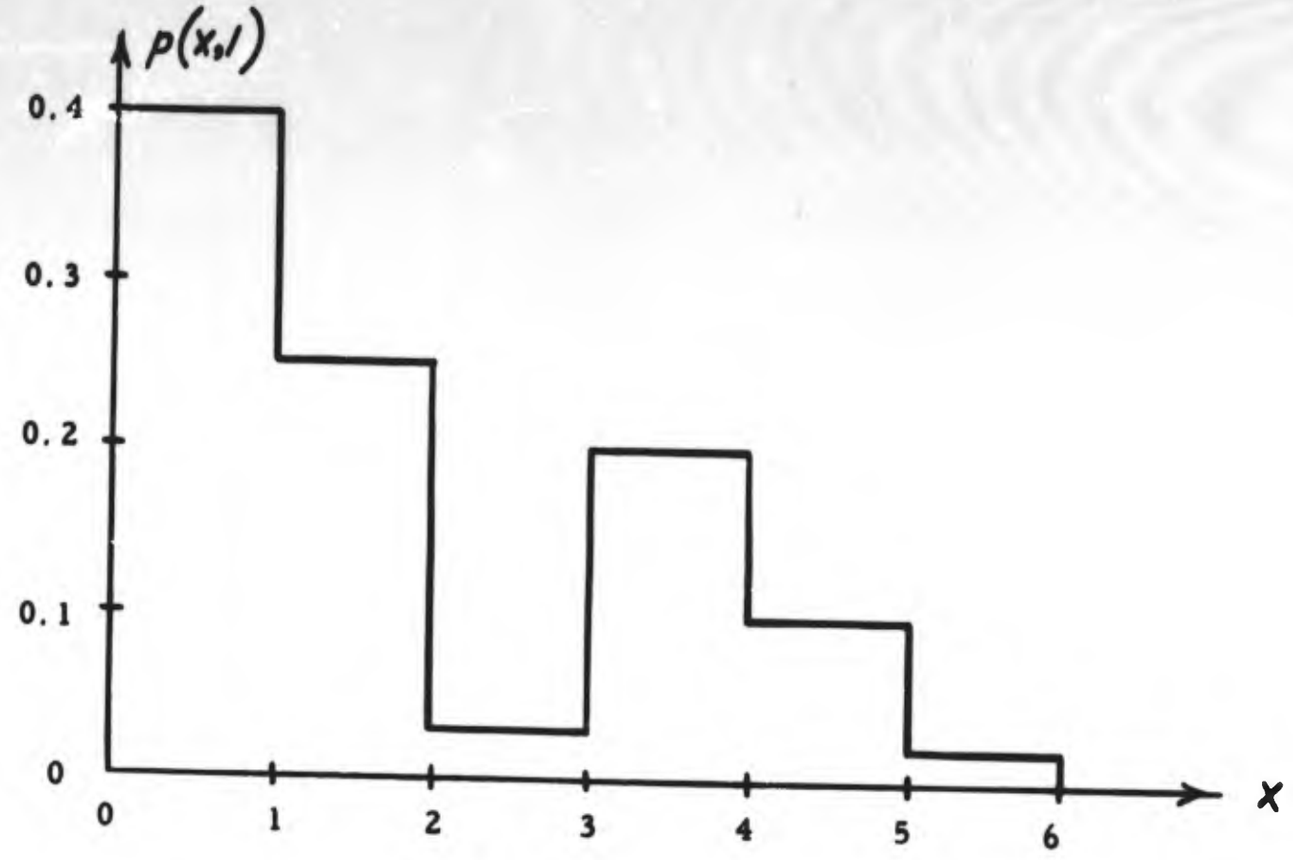
$$\hat{\Pr}\{\text{opt. decision}\} = \sum_{k=1}^{K} \int_{\hat{R}_k} \hat{\pi}_k(x)\, \hat{p}(x)\, dx \qquad (17)$$

This estimate may be biased and for practical learning sample sizes n may have an appreciable variance. But still, if the range of integration is restricted to $\hat{N}_k$ so that $\hat{\pi}_k(x) \geq \pi_o$ for k = 1, 2, . . . , K and if $\hat{P}_k \geq P$ for k = 1, 2, . . . , K, then $\hat{\Pr}\{\text{opt. decision}\} \geq \pi_o P$ and the lower bound will be rather loose so that by a confidence interval type of argument one can be fairly sure that in fact, $\Pr\{\text{opt. decision}\} \geq \pi_o P$. These comments will now be illustrated with a simple example.

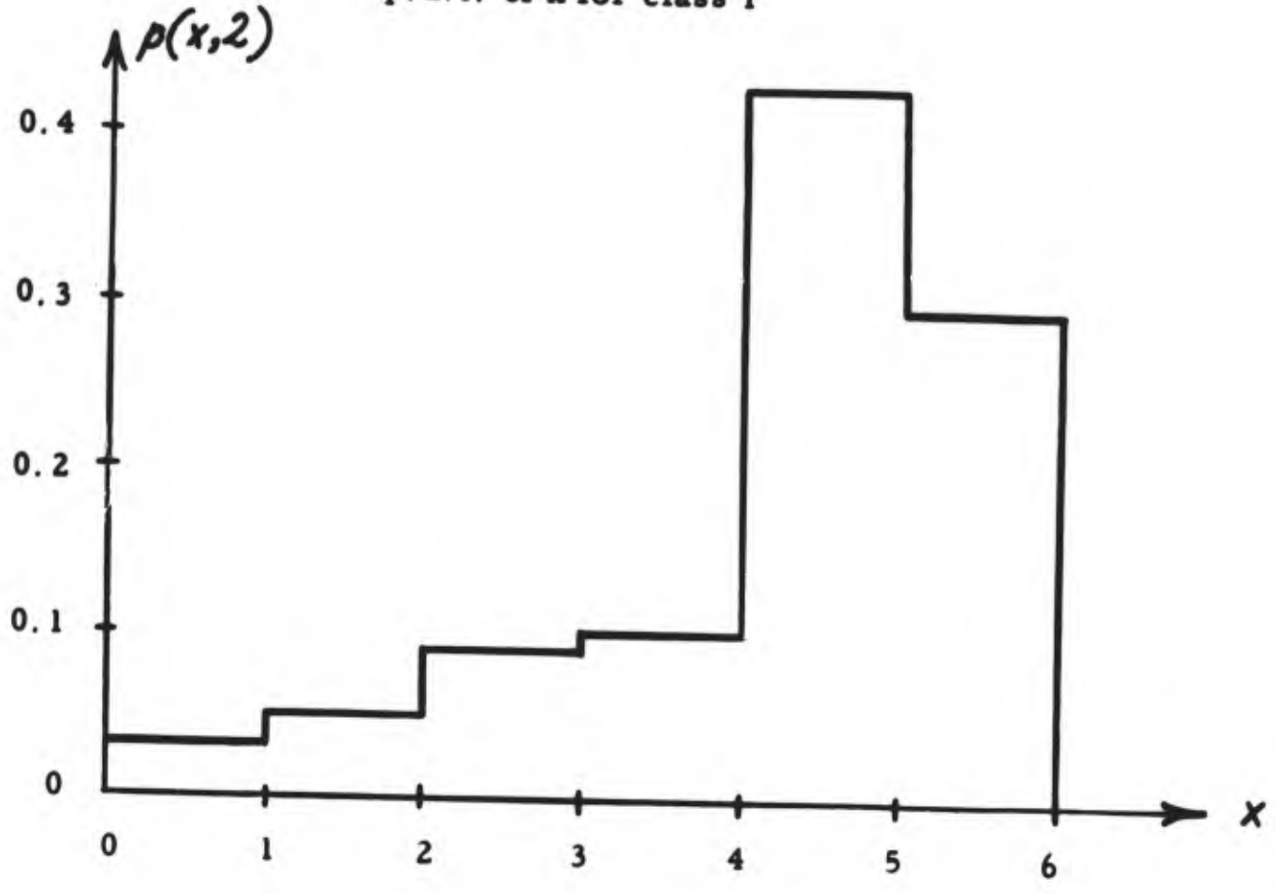
Suppose that the output of a transducer lies in the range 0 to 6 volts and is quantized at 1 volt intervals and that for two equally probable classes the true probabilities of falling in the quantization intervals are as shown in Figure 11. This example has relevance because SPEAR is a somewhat generalized histogram generator. (More generally, the quantization intervals in Figure 11 might just as well have been described as representing cells in a general N-space. This is not too unrealistic since often only a few cells will have significant associated probabilities and clearly the dimensionality of the observation space is unimportant here.) Data were generated by taking numbers from a table of random digits and classifying by the following intervals:

1st class - 0-39, 40-64, 65-67, 68-87, 88-97, 98-99

2nd class - 0-2, 3-7, 8-16, 17-26, 27-69, 70-99



p.d.f. of x for class 1



p.d.f. of x for class 2

FIG. 11. P.d.f.'s Used in Illustrating Measures of Estimation Quality

The cell probabilities were estimated for both classes with five different sample sizes. These estimates are given in Table I along with the true values. The estimates are unbiased and multinominally distributed random variables.

The estimates for an individual cell are binominally distributed and so $\pi_k(x)$ can be computed by:

$$\pi_k(x) = \sum_{\ell=0}^{n} \binom{n}{\ell} t_k^{\ell} \; (1 - t_k)^{n-\ell} \sum_{m=0}^{\ell-1} \binom{n}{m} t_j^{m} \; (1 - t_j)^{n-m}$$

where $t_k = t_k(x)$ is the probability of an observation from the k^{th} class falling in the cell containing the point x, k=1 and j=2 or k=2 and j=1. Similarly, $\hat{\pi}_k(x)$ can be computed by using the estimated values of t_k and t_j. However, it is much simpler to use an approximation which is adequate for this illustration and may, in fact, be used in the practical application of this test. If two random variables have pdf's such as are shown in Figure 12 with the uppermost ϵ quantile of the lower distribution less than the lowest ϵ quantile of the upper distribution, where $\epsilon << 1$, then the probability that the random variables will be in the "correct" order is approximately $(1 - \epsilon)^2$. Hence, for each true and estimated cell probability in Table I, a simple and approximate test of $\pi_k(x)$ or $\hat{\pi}_k(x)$ being greater than π_o may be performed by consulting a table or graph of confidence intervals.[13, 14]

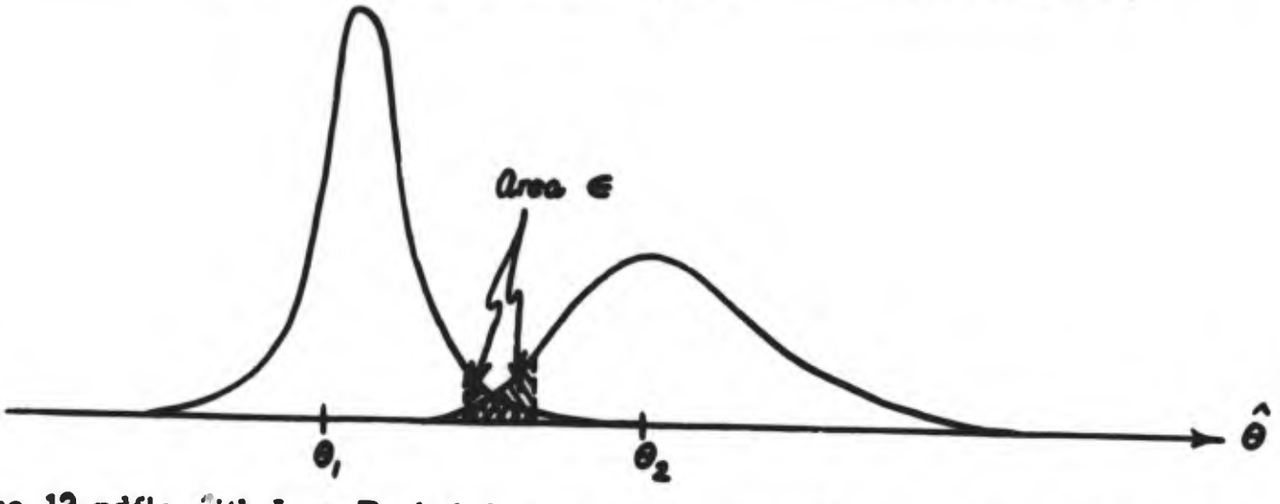


Figure 12. pdf's with Low Probability of Order Interchange of Estimates of θ_1 and θ_2

TABLE 2.6-I

1st class

cell	0-1^{v}	1-2^{v}	2-3^{v}	3-4^{v}	4-5^{v}	5-6^{v}
true	.40	.25	.03	.20	.10	.02
n = 30	.434	.267	0	.167	.133	0
50	.440	.260	.020	.220	.080	0
100	.440	.230	.040	.200	.080	.010
250	.416	.232	.032	.216	.088	.016

2nd class

cell	0-1^{v}	1-2^{v}	2-3^{v}	3-4^{v}	4-5^{v}	5-6^{v}
true	.03	.05	.09	.10	.43	.30
n = 30	0	.067	.033	.100	.433	.367
50	.020	.020	.100	.100	.400	.360
100	.020	.050	.120	.110	.360	.340
250	.016	.064	.104	.088	.440	.288

True and estimated cell probabilities for sample size n.

Selecting $\pi_o = .81$ and consulting the graph of 90% confidence belts shown in Figure 13, Table II was constructed to show the quantization cells for which $\pi_k(x) \geq \pi_o$ or $\hat{\pi}_k(x) \geq \pi_o$. These are the cells, therefore, which make up the regions N_k or $\hat{N}_k$. It so happened that the random samples used in this example displayed very average characteristics and only a very few estimates deviated significantly from the true values. As a result, for this simple example, there were no instances of estimates being in reversed order and the regions N_k and $\hat{N}_k$ were in complete agreement except for the exclusion of the $4 - 5^v$ cell from $\hat{N}_2$ for n = 30 (and this exclusion was marginal). Therefore, this example does not illustrate the need for a measure of pdf estimation quality. But, it is easy to see from an inspection of Figure 13 that it is possible, with reasonably large probability, to have estimates with order reversals. This is especially true if the cell probabilities for both classes are small (which is more nearly the usual situation in practice).

The lower bound (15b) to the probability of an optimum decision can be computed for the various sample sizes. This is shown in the next to last column in Table II and the maximum value of this lower bound is 0.676. That is, the probability of drawing samples from these two classes of size 250 or greater that yield estimates of the cell probabilities which will allow optimum decisions to be made is greater than 0.676.

For the particular random samples drawn and used here, when the estimated cell probabilities are used in generating the corresponding estimated regions, $\hat{N}_k$, of high quality estimation, the corresponding "estimated lower bounds" for the probability of an optimum decision are shown in the final column of Table II. Note that if one were to pick a value for P of 0.7 that $\hat{P}_1$ and $\hat{P}_2$ are greater than P for n = 50 and so, by Equation (15c), one would feel fairly safe in saying that the probability of an optimum decision is greater than $\pi_o P = 0.81 \times 0.7 = 0.567$.

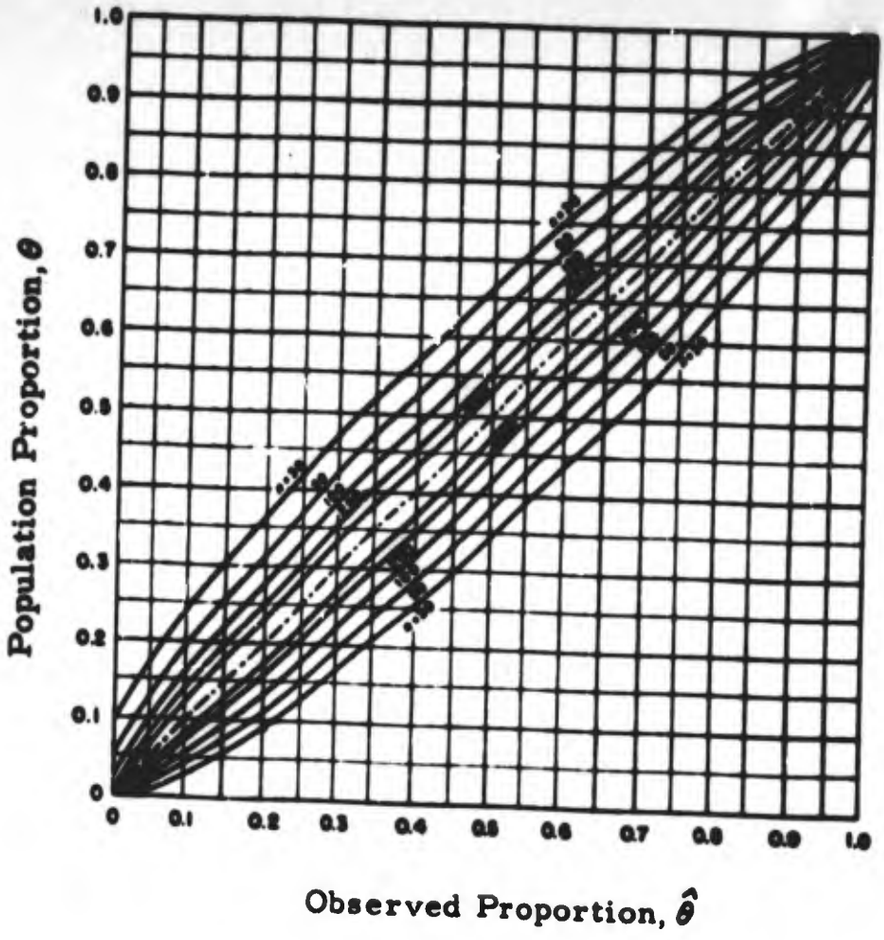


FIG. 13. 90% Confidence Belts for Proportions *

* This chart is taken from Reference [13] and was constructed by E. L. Crow, NOTS, following the method used by Clopper and Pearson [14]

TABLE 2.6-II

Cell		$0\text{-}1^v$	$1\text{-}2^v$	$2\text{-}3^v$	$3\text{-}4^v$	$4\text{-}5^v$	$5\text{-}6^v$		P_1	P_2	$\hat{P}_1$	$\hat{P}_2$	$\sum_{h=1}^{2} w_o P_k T_k$	$\sum_{1}^{12} w_o P_k T_k$
n = 30	N_1	✓												
	N_2					✓	✓		.40	.73			.458	
	N_1	✓												
	N_2						✓				.434	.367		.324
n = 50	N_1	✓	✓											
	N_2					✓	✓		.65	.73			.559	
	N_1	✓	✓											
	N_2					✓	✓				.70	.76		.59
n = 100	N_1	✓	✓											
	N_2					✓	✓		.65	.73			.559	
	N_3	✓	✓											
	N_4					✓	✓				.67	.70		.555
n = 250	N_1	✓	✓		✓									
	N_2			✓		✓	✓		.85	.82			.676	
	N_3	✓	✓		✓									
	N_4			✓		✓	✓				.864	.832		.676

True and estimated regions and probabilities

Whereas, if one applies the same test to the true probabilities P_1 and P_2, one would not be able to satisfy the same lower bound (15c) until a sample size of $n = 250$ had been reached. Hence, there definitiely is some danger in using the estimated cell probabilities in the computation of (15). However, this danger is of little practical importance since here one is not so much interested in making mathematically rigorous statements as one is in simply being reasonably sure that the machine learning job done is adequate for the application at hand.

This example has required several pages to present. However, it should be pointed out that the procedure followed can be programmed and is a relatively simple algorithm. This procedure provides a means for testing whether or not one can be reasonably sure that the pdf estimates one has obtained, at a particular stage of the sampling or learning process, will lead to an optimum decision rule.

Floyd has also performed studies on the topic of this section[15, 16]. In particular, he studied the case in which the two class pdf's are estimated by constructing generalized histograms, i.e., the quantities being estimated are the probabilities of an observation from the k^{th} class falling the the various cells of a fixed cell structure. This is a first step toward studying the type of pdf representation generated by the ASSC or SPEAR programs in which the cell structure varies in a random manner dependent on the data used for learning. Floyd's work will be summarized here for completeness.

Suppose N_1 independent random observations are made on class A and N_2 independent random observations are made on class B. These observations are then classified into c cells labeled by the numbers 1 to c, and the probabilities associated with these cells by each class are estimated by success counting. Then, m independent random observations are made on one of the classes and a decision is made as to which of the two classes the m observations came from. Let $\{k_i\} = \{k_1, \ldots, k_m\}$ denote the cells into which the m observations fall.

If the true cell probabilities $P_A(k_i)$ and P_B were known, than the maximum a posteriori decision rule could be expressed as:

decide in favor of class A if $$L = \frac{P_A}{P_B} \prod_{i=1}^{m} \frac{P_A(k_i)}{P_B(k_i)} > 1 \qquad (18a)$$

and

decide in favor of class B if $$L < 1 \qquad (18b)$$

However, it is assumed that $P_A(k_i)$ are unknown, and are instead estimated by $\hat{P}_A(k_i) = \frac{a_{k_i}}{N_1}$ and $\hat{P}_B(k_i) = \frac{b_{k_i}}{N_2}$, where a_{k_i} and b_{k_i} are the number of observations that fell in the k_i^{th} cell from class A and class B respectively during learning. Therefore, the decision rule that is used in place of (18) is

decide in favor of class A if $$\hat{L} > 1 \qquad (19a)$$

and

decide in favor of class B if $$\hat{L} < 1 \qquad (19b)$$

where

$$\hat{L} = \frac{P_A}{P_B} \prod_{i=1}^{m} \frac{\hat{P}_A(k_i)}{\hat{P}_B(k_i)} \qquad (20)$$

The first specific topic studied by Floyd was a comparison of the probability

$$Q' \equiv Pr\{(\hat{L}-1)(L-1) > 0\} \qquad (21)$$

= probability that a learning sample will be drawn which will yield an $\hat{L}$ leading to the same decision as L, for a given test sample of size m

with the estimated probability

$$Q \equiv \hat{P}r\{(\hat{L}-1)(L-1) > 0\} \qquad (22)$$

= estimate of Q' obtained by replacing $P_A(k_i)$ and $P_B(k_i)$ with $\hat{P}_A(k_i)$ and $\hat{P}_B(k_i)$

Specifically,

$$Q = \sum_{j}^{N_1} \sum_{r}^{N_2} \prod_{\ell=1}^{m} \binom{N_2}{j_\ell} \binom{N_2}{r_\ell} \left(\frac{a_{k_\ell}}{N_1}\right)^{j_\ell} \left(\frac{b_{k_\ell}}{N_2}\right)^{r_\ell} \left[1-\left(\frac{a_{k_\ell}}{N_1}\right)\right]^{N_1 - j_\ell} \left[1-\left(\frac{b_{k_\ell}}{N_2}\right)\right]^{N_2 - r_\ell}$$

$$\cdot \tfrac{1}{2} \left\{ \operatorname{sgn} \left[\left\{ \left(\frac{N_2}{N_1}\right)^m \prod_{i=1}^{m} \frac{j_i}{r_i} - 1 \right\} \left\{ \left(\frac{N_2}{N_1}\right)^m \prod_{i=1}^{m} \frac{a_{k_i}}{b_{k_i}} - 1 \right\} \right] + 1 \right\} \qquad (23)$$

where $\sum_{j}^{N_1}{}'$ and $\sum_{r}^{N_2}{}'$ indicate

$$\sum_{j_1=0}^{N_1} \sum_{j_2=0}^{N_1} \cdots \sum_{j_m=0}^{N_1} \quad \text{and} \quad \sum_{r_1=0}^{N_2} \sum_{r_2=0}^{N_2} \cdots \sum_{r_m=0}^{N_2}, \text{ respectively,}$$

and $\operatorname{sgn} x = \frac{x}{|x|}$ if $x \neq 0$ and $\operatorname{sgn} x = 0$ if $x = 0$, and wherever $j_i = r_\ell = 0$ in (23) the definition $\frac{j_i}{r_\ell} = \frac{0}{0} \equiv 1$ is used.

Two methods were used in comparing Q' and Q. The first of these was to try to evaluate the rms difference between Q' and Q for a single ($m=1$) test sample point in the k^{th} cell. Despite extensive effort applied toward this end, the formidable nature of expressions such as (23) was not overcome even though several fairly reasonable approximations were made. However, Floyd did succeed in showing that Q tends to an unbiased estimate of Q' for large learning sample sizes.

In view of the immense difficulty in obtaining theoretical results, an experiment was conducted which shed some light on the magnitude of the rms difference between Q' and Q^{15}. Specifically, 10 independent samples of size 40 ($N_1 = N_2 = N = 40$) were drawn from each of the populations with the pdf's given below:

$$q_A(x) = \frac{1}{\sqrt{1 - x^2}\,\pi} \quad \text{if } -1 < x < 1$$

$$= 0 \text{ if } |x| > 1$$

$$= \text{pdf of a randomly sampled sinewave} \tag{24a}$$

$$q_B(x) = \tfrac{1}{2} \text{ for } -1 \leq x \leq 1$$

$$= 0 \text{ if } |x| > 1 \tag{24b}$$

$$= \text{pdf of a randomly sampled saw tooth wave}$$

The data were classified into four equal sized cells (c = 4), and cell probabilities were estimated from each of the 10 pairs of samples. The estimated probability Q was obtained for each cell for all 10 examples of learning. The results of the experiment indicated that the rms difference between Q' and Q was quite large even for c/N=10. Furthermore, in each cell the sample average value of Q was greater than Q'. It would be dangerous to conclude, however, that Q has a significant bias for small learning sample sizes.

Also, and of considerable value, Floyd studied the effect of varying c, the number of cells. Specifically, assuming the class pdf's of (24) and using the approximation

$$\alpha \simeq \Phi\left(\frac{\sqrt{N_1 N_2}\,|P_A(k) - P_B(k)|}{\sqrt{N_1 P_B(k)[1 - P_B(k)] + N_2 P_A(k)[1 - P_A(k)]}}\right) \tag{25}$$

where $\Phi(x)$ is the normal cumulative distribution function, he obtained the graphs reproduced in Figures 14, 15, and 16. These graphs indicate (as might also be deduced from Figure 12) that the probability of an optimum decision remains low as $N = N_1 = N_2$ increases, only in the cells near the points $\pm (1 - \frac{4}{\pi})^{\frac{1}{2}}$ where $q_A(x) = q_B(x)$. Furthermore, these graphs indicate the enormous learning sample sizes required if one insists on having a high probability of an optimum decision for <u>every possible</u> test observation, i.e., at every point in the observation space or in every cell, and simultaneously insist on a fine cell structure, i.e., large c. A fine cell structure is desired in order to bring the probability of error down to what may be achieved if the class pdf's are known instead of only cell probabilities.

The graphs obtained by Floyd may now be used in conjunction with Equation (15a) to compute an accurate approximation to the (true) probability of an optimum decision for a randomly selected observation. However, since the class of pdf's given by Equation (24) are such that the error probability in making decisions is very high, the lower bound (15b) is much lower than the true probability. Furthermore, the values of P_k are so low that (15c) is of no use at all for the class pdf's (24).

The approximation used here is one that would be used whenever the pdf's are estimated by generalized histograms, and consists of using the quantities obtained from the average values of the class pdf's in every cell. That is, the "true histogram" of $q_A(x)$ is used in place of $p(x;1)$, etc. Thus, the approximation

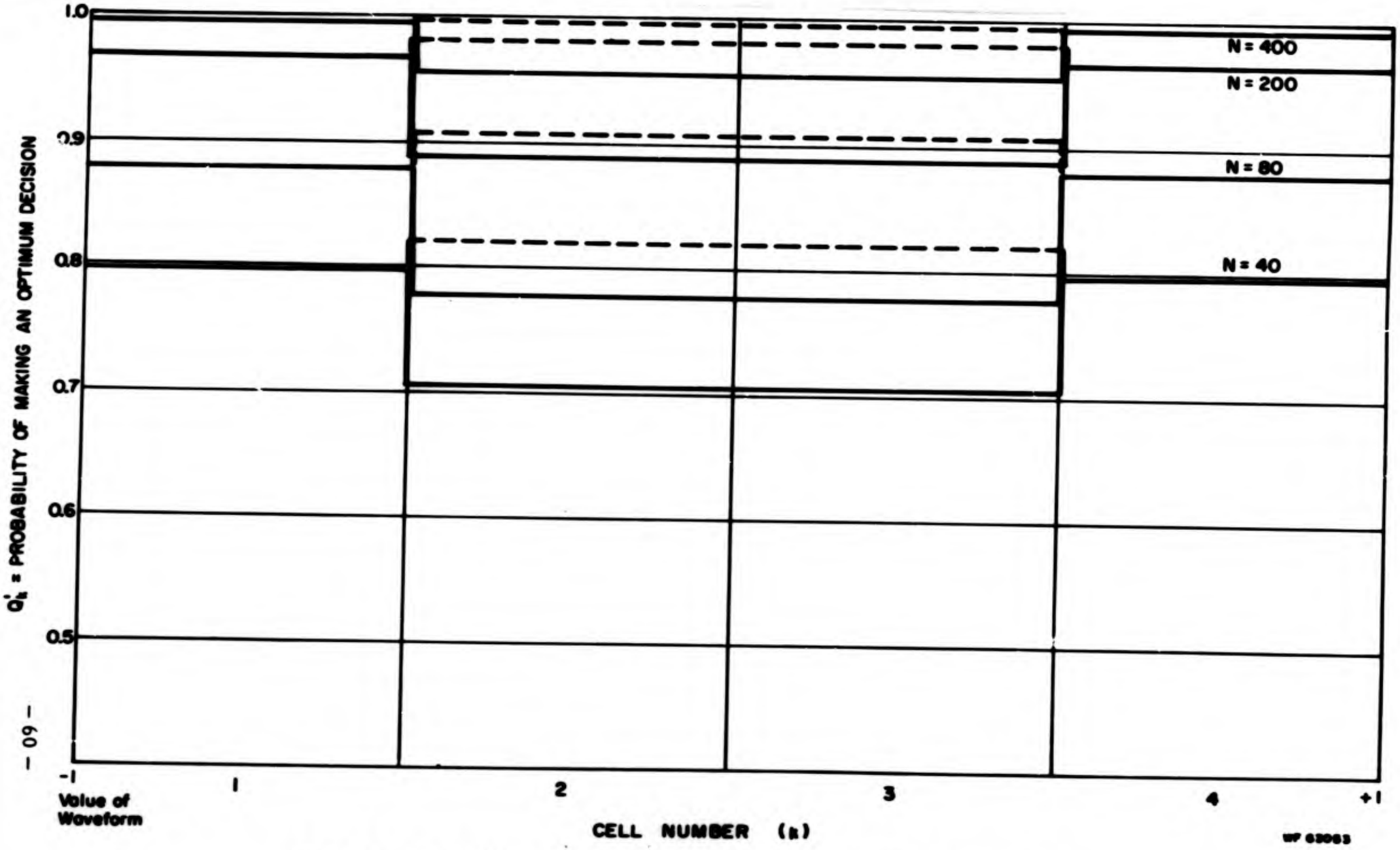


FIG. 14. Probability of Optimum Classification of a Sinewave and a Sawtooth Waveform (C = 4)

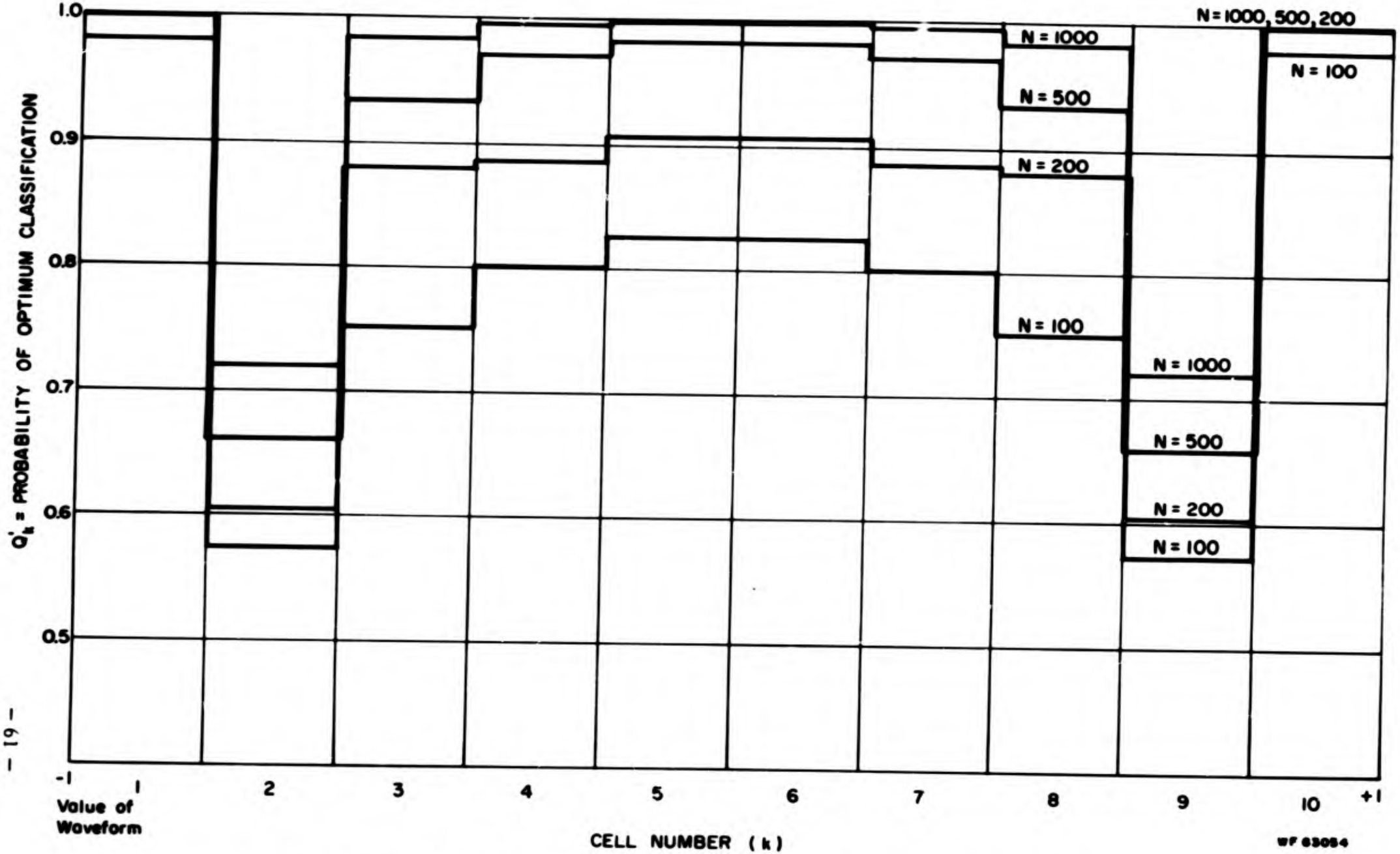


FIG. 15. Probability of Optimum Classification of a Sinewave and a Sawtooth Waveform (C = 10)

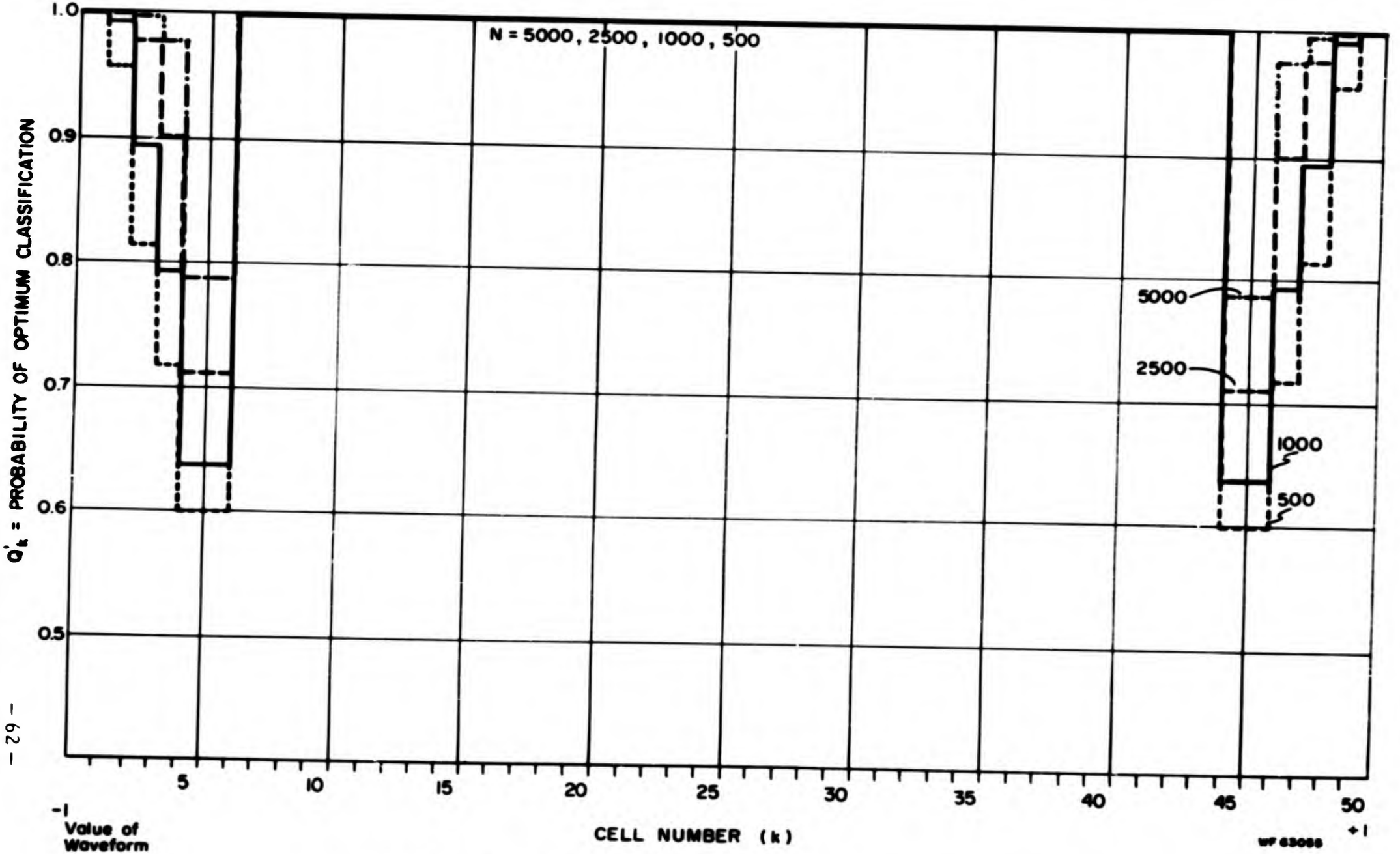


FIG. 16. Probability of Optimum Classification of a Sinewave and a Sawtooth Waveform (C = 50)

$$\pi_k(x) \tilde{=} Q' \text{ for } x \in R_k \tag{26}$$

can be made where Q' is given by (21). This approximation is accurate if the number of cells is reasonably large.

Inserting the values from Figure 15 into Equation (15a), Table III is obtained. The fact that the probability of an optimum decision is greater than 80 per cent for learning sample sizes of 100 or greater from each of the two distributions in (24) indicates that, under certain conditions, extremely large learning samples are not required. That is, if the cells used for pdf estimation are judiciously placed and the probability of observations falling in each of these cells is not extremely small, high quality learning can be achieved with thoroughly practical learning sample sizes. Of course, it is not always easy to satisfy these two conditions on the cell structure when little or no a priori knowledge of the class pdf's is available. (This is particularly true when the dimensionality of the observation space is greater than three.)

TABLE III

Size of learning sample from each class	100	200	500	1000
Pr {optimum decision}	.808	.870	.916	.942

Approximate evaluation of Equation (15a) for the class pdf's of Equation (24)

In conclusion, this section has presented simply, but useful methods of measuring the quality of machine learning by pdf estimation and the quality of machine decisions based on such learning. The usefulness of these measures of quality is greatest in those applications of greatest interest, i.e., those in which the error rate is low.

3. A COMPUTER PROGRAM FOR DATA PREANALYSIS

The techniques used for machine learning have been described in Section 2 of this report and in previous reports [6]. In order to apply these techniques, it is necessary to select values for certain program parameters. Some of these act as control parameters, and methods for selecting appropriate values for these control parameters have been developed elsewhere*. The results are summarized in Section 2. 5 of this report.

This section is concerned with one method of selecting values for the SPEAR learning program parameters which do not serve to control the learning process, but rather serve to specify the initial cell sizes. While inclusion of the cell-growth feature in SPEAR does reduce, to some extent, the precision required to specify the initial cell sizes (as well as making the final cell structure quite general), the need for judicious choices is not eliminated. For SPEAR to function at all well in its present form, the initial cell sizes must not be significantly greater or less than the "local spread" of the class probability density functions (pdf's) about the modes of the class distribution.

It must first be emphasized that one initial cell size may not be a best choice everywhere in the observation space. This is due to the fact that there are usually different "local spreads" around different modes. However, at present, SPEAR employs only one initial cell size. This restriction to only one initial cell size has not presented severe difficulties in any of the applications tried to date.

---

* See Ref. 12 and Appendix II of Ref. 6

With no a priori knowledge of the class "local spreads", the initial cell sizes for SPEAR must be chosen by an examination of data. A computer program has been written* which performs a simple analysis of data to obtain some idea of the magnitude of the class "local spreads". The use of this program precedes the use of SPEAR and would take the place of the "trial runs" now required in most practical applications. These trial runs have been required in order to make adjustments in the initial cell sizes if these proved to be unsatisfactory. This trail run method is expensive and too much of a "black art"; hence, the present "preanalysis" program was written to reduce the cost of applications and to make the overall learning process more systematic.

One of the simplest methods of obtaining a partial picture of the class pdf's is by making estimates of the coordinate marginal pdf's. That is, the observation is an N-dimensional vector $x = (\xi_1, \xi_2, \ldots, \xi_N)$ with a pdf for the k^{th} class, $p(x;k)$, which is a joint density function of the N coordinates. A knowledge of the coordinate marginal densities $\{p_i(\xi_i;k);\ i = 1, 2, \ldots, N\}$ conveys partial knowledge of the class pdf $p(x;k)$. The first part of the preanalysis program constructs estimates of the marginal densities $\{p_i(\xi_i;k);\ i = 1, 2, \ldots, N\}$ for each class k. Using these pdf estimates alone, a reasonable first choice for the SPEAR initial cell diameter in the i^{th} coordinate is the width of the narrowest "local spread" about a significant mode of $p_i(\xi_i;k)$ for any k. A significant mode is one with a reasonably large probability that a random observation will fall near that mode. While some judgment is necessary in applying such a rule, it is at least a partial systematizing of the job of choosing initial cell sizes.

---

* The program described in this section was written for use on the Computer Control Company's DDP-24 computer now at Litton's Information Sciences Laboratory.

3.1 THE BASIC PROCEDURE

The procedure for constructing estimates of the coordinate marginal pdf's is to construct a type of histogram. Specifically, a sample of size is taken from the class under investigation, say the k^{th} class, and the i^{th} coordinate values are arrayed in ascending order. The ordered coordinate sample values are denoted by $\xi_{i(1)}, \cdots, \xi_{i(n)}$. The lowest possible value $\xi_{i(0)}$ of the coordinate (assumed finite and determined by the measuring device) is taken as the lower limit of the first cell. The upper limit of the first cell, equal to the lower limit of the second cell is set equal to $\xi_{i(k)}$ where $m = (n + 1)/k$ is the number of cells to be generated. The upper limit of the second cell, equal to the lower limit of the third cell, is set equal to $\xi_{i(2k)}$. This procedure is continued, with the upper limit of the m^{th} or final cell, $\xi_{i(n+1)}$, equal to the maximum possible coordinate value (again assumed finite and determined by the measuring device). Then the histogram value of the l^{th} cell is given by

$$h_i(l;k) = 1/m\,(\xi_{i(lk)} - \xi_{i([l-1])}) \tag{27}$$

This value, $h_i(\ ;k)$, is an estimate of the average value of the class marginal coordinate density function over the cell. And, if k is large, the statement (as will be discussed presently)

$$h_i(l;k) \cong \int_{\xi_{i([l-1]k)}}^{\xi_{i(lk)}} p_i(\xi_i;k)\,d\xi_i \Big/ \left(\xi_{i(lk)} - \xi_{i([l-1])}\right) \tag{28}$$

is justified with high probability; $h_i(l;k)$ then provides an acceptable picture of the marginal density.

The next step in determining the width of the narrowest "local spread" about a significant mode of $p_i(\xi_i;k)$ is to find the smallest cell width of $h_i(\ell;k)$. The "local spread" may be defined as either this narrowest cell width, or in some similar suitable manner (which will depend on m).

The present program, in its simplest form, types out the values of $h_i(\ell;k)$ for $\ell = 1, 2, \ldots, m$; $i = 1, 2, \ldots, N$, and k ranges over all classes, as well as the corresponding cell limits. A flow chart of the main program is shown in FIG. 17. In addition to the histogram values, the program outputs the minimum cell length and the number of "modes" in $h_i(\ell;k)$. A "mode" is operationally defined here as a value of ℓ such that $h_i(\ell - 1; k) < h_i(\ell; k) > h_i(\ell + 1; k)$. To be significant, a "mode" must have a significant probability measure and there must be a significant departure from uniformity in $h_i(\ell; k)$ at the mode. In each case, however, the choice of what is significant is subject to the judgment of the experimenter and is closely connected to the choice of the number of cells, m.

3.2 CONSIDERATIONS IN THE CHOICE OF CONTROL PARAMETERS

The choice of m is affected primarily by the conflicting goals of obtaining a high resolution, accurate estimate of $p_i(\xi_i; k)$ and at the same time keeping the cost of preanalysis low. In view of this trade-off situation, the experimenter must attempt to specify the lowest resolution requirements for the preanalysis which is consistent with the goal of selecting reasonable initial cell sizes for the SPEAR automatic learning program. The average probability content of an individual cell is 1/m, so that the probability measure associated with a "mode" which just satisfies the operational definition used here may be reasonably defined as 1/m. However, if a "mode" exists such that $h_i(\ell - 2; k) < h_i(\ell - 1; k) < h_i(\ell; k) > h_i(\ell + 1; k) > h_i(\ell + 2; k)$, then the

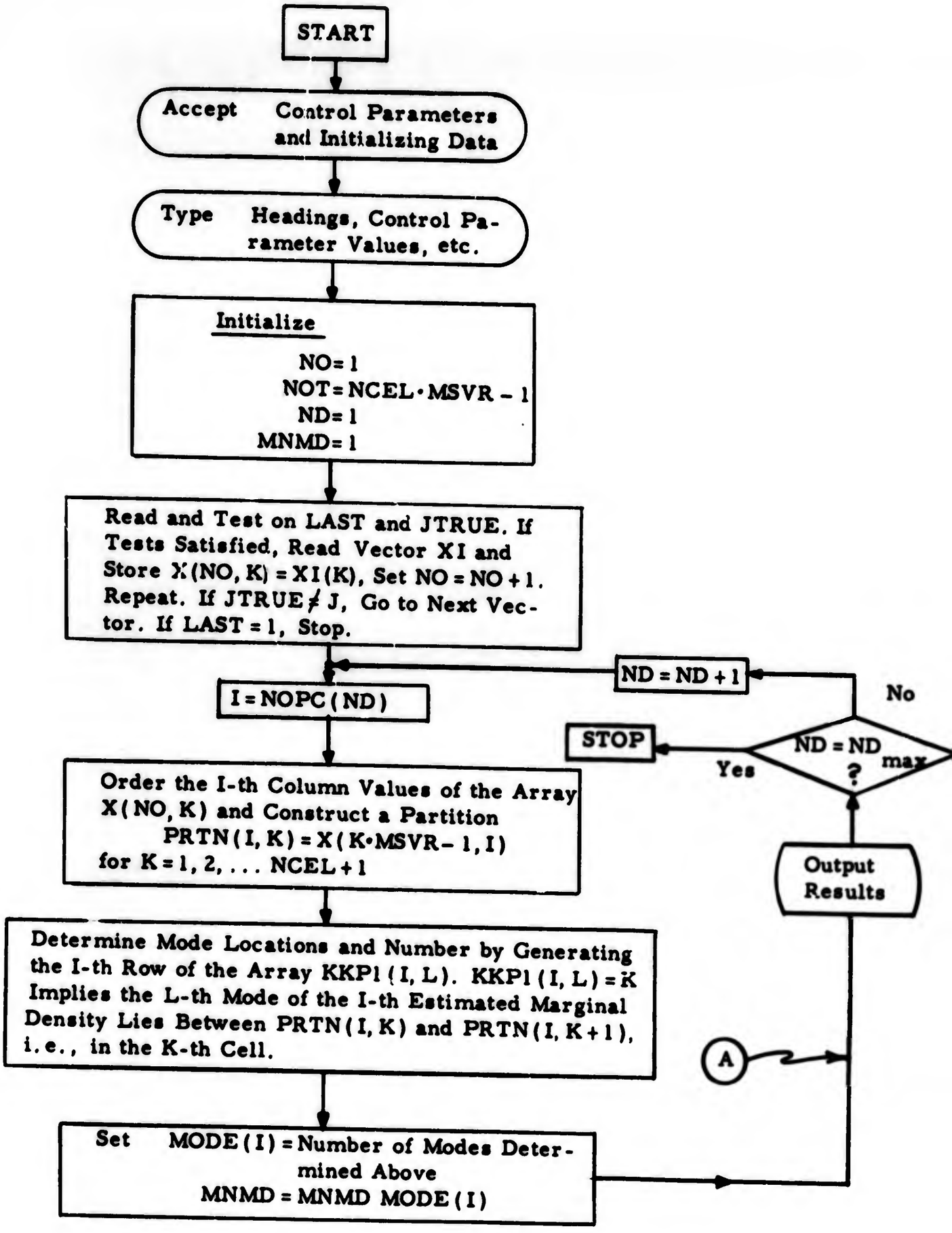


FIG. 17. Flow Chart of the Basic PREANALYSIS Program

probability measure associated with this "mode" might reasonably be defined as 3/m. Clearly, as m is increased, more cells will be generated near the modes of $p_i(\xi_i; k)$ giving a better picture of the behavior of the marginal density in that neighborhood. However, more cells are required to be associable with a "mode" as m is increased if the significance of the "mode" is not to be diminished.

The other program control parameter, which must be specified in order to utilize the program as described so far is k, the number of observations in each cell. The choice of k determines the accuracy, for given m, of the pdf estimate which is generated by using sample quantiles to specify the histogram cells in the above manner.

To demonstrate the role of k in determining the accuracy of such a "sample-quantile histogram", suppose that $x_{(1)} \leq x_{(2)} \cdots \leq x_{(n)}$ is an ordered sample from a univariate population with probability density function f(x). Let w be the area under f(x) between $x_{(r)}$ and $x_{(s)}$ where $r < s$. That is,

$$w = \int_{x_{(r)}}^{x_{(s)}} f(x)\, dx \tag{29}$$

Obviously w is a random variable. The pdf of w is

$$g(w) = \frac{n!}{(s - r - 1)!\,(n - s + r)!}\; w^{s-r-1}(1 - w)^{n-s+r}, \quad 0 \leq w \leq 1 \tag{30}$$

which is of the beta form with $\alpha = s - r - 1$ and $\beta = n - s + r$*. Therefore, the mean and variances of w are:

---

* See, for example Ref. 17

$$E(w) = \frac{n!\ (s-r)!}{(n+1)!\ (s-r-1)!} = \frac{s-r}{n+1} \tag{31}$$

and

$$\text{Var}(w) = \frac{n!\ (s-r+1)!}{(n+2)!\ (s-r-1)!} - \left[\frac{n!\ (s-r)!}{(n+1)!\ (s-r-1)!}\right]^2$$

$$= \frac{(s-r-1)(s-r)}{(n+2)(n+1)} - \left(\frac{s-r}{n+1}\right)^2$$

$$= \frac{s-r}{n+1}\left[\frac{(n+1)-(s-r)-2}{(n+1)(n+2)}\right] \tag{32}$$

Now, identifying (s-r) as k and (n+1)/(s-r) as m, the number of cells in the histogram, Eqs. (31) and (32) may be rewritten as

$$E(w) = \frac{1}{m} \tag{33}$$

and

$$\text{Var}(w) = \frac{1}{m}\left[\frac{mk-k-2}{(mk)(mk+1)}\right] \tag{34}$$

Then, the squared mean-to-variance ratio (MSVR) is

$$\frac{E^2(w)}{\text{Var}(w)} = k\frac{mk+1}{mk-k-2} \tag{35a}$$

$$< k\left(\frac{m}{m-1}\right) \tag{35b}$$

Thus, the approximation if (28) is valid with high probability if k is large.

Of course, validity of (28) is not sufficient to assure that the results of this type of analysis will be useful. Obviously the approximation of (28) is always exact if m = 1. Furthermore, since the chief objective here is a determination

of suitable initial cell sizes for use in the automatic learning technique described in Section 2 (the SPEAR program), some idea should be had of the manner in which the cell widths approximate the "local spreads" of the pdf.

If m is very large and $k \geq 1$, then the nature of the pdf will be accurately portrayed and the experimenter need only look at the resultant histogram to determine the "local spreads". However, this is usually impractical since it implies a very large sample size n = mk - 1*. Often in practice with m small, the experimenter will face estimated "modes" which are one cell in width. In this case, the use of cell width as an estimate of "local spread" somewhat resembles the use of the difference between maximum and minimum sample values as an estimate of the range of a uniform distribution. This rough analogy leads one to suspect that the cell width of a single cell "mode" will usually be somewhat less than what one would like to call the "local spread". This, indeed, turns out to be true as is illustrated in Section 3. 3

It appears that adherence to the following rule of thumb will yield useful results from this estimation technique; m and k should both be greater than 10 with k as large as possible.

---

* n = mk - 1 since the extreme limits of the distribution are assumed known.

3.3 A SIMPLE EXAMPLE

To illustrate the use of this preanalysis program and to indicate the type of information that it provides, data were taken from two simple bivariate populations. These populations or classes have p.d.f.'s which are constant over a set of rectangular areas in the plane shown in Figure 18. The probabilities of falling in each of the four rectangles for each class are equal. Two hundred sample vectors were generated with the aid of a table of random digits. The probability of an observation being made on class 1 was 0.6 and the probability of observing class 2 was 0.4, i.e., data from the two classes were mixed but labeled.

The coordinate marginal p.d.f.'s were estimated automatically with control parameter values m=6 and k=12. The sample-quantile histograms thus obtained are shown in Figure 19 along with the true coordinate marginal p.d.f.'s. It may be seen that in most of the cells generated, the approximation (28) is valid. Most of the estimated densities display the more important features of the true densities rather accurately considering the somewhat low values of m and k. As expected, the minimum cell widths are less than the "local spreads" in the neighborhoods of these smallest cells.

This estimation technique was repeated with m=6 and k=6 using the same data. Although the approximation (28) was only moderately less true in most of the generated cells, the point by point representation of the densities suffered badly. In particular, the minimum cell widths were unacceptably small in comparison to the "local spreads".

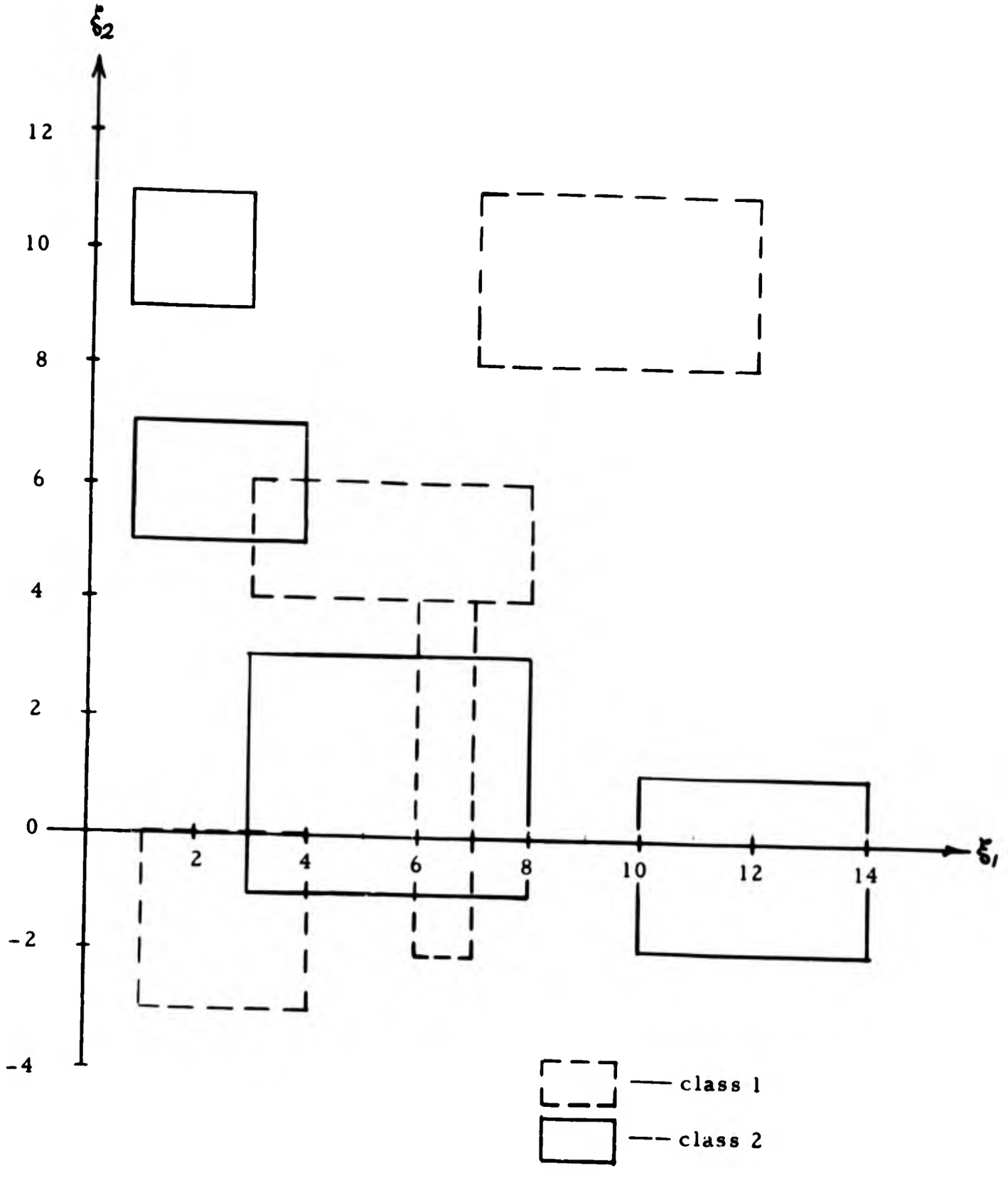


FIG. 18. Regions Associated with Each of Two Classes

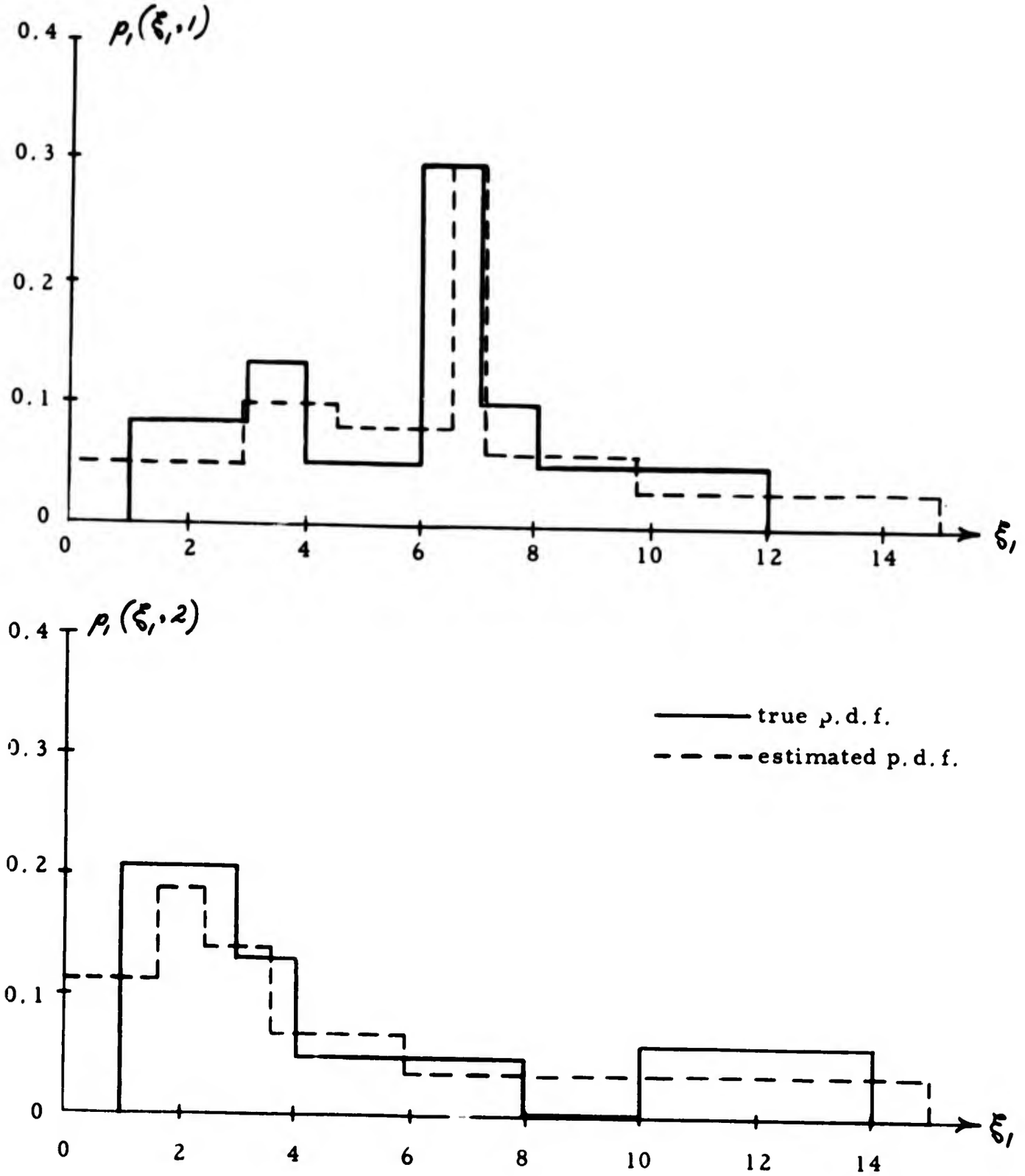


FIG. 19a. True and Estimated First Coordinate Marginal p.d.f.'s, for m = 6, k = 12

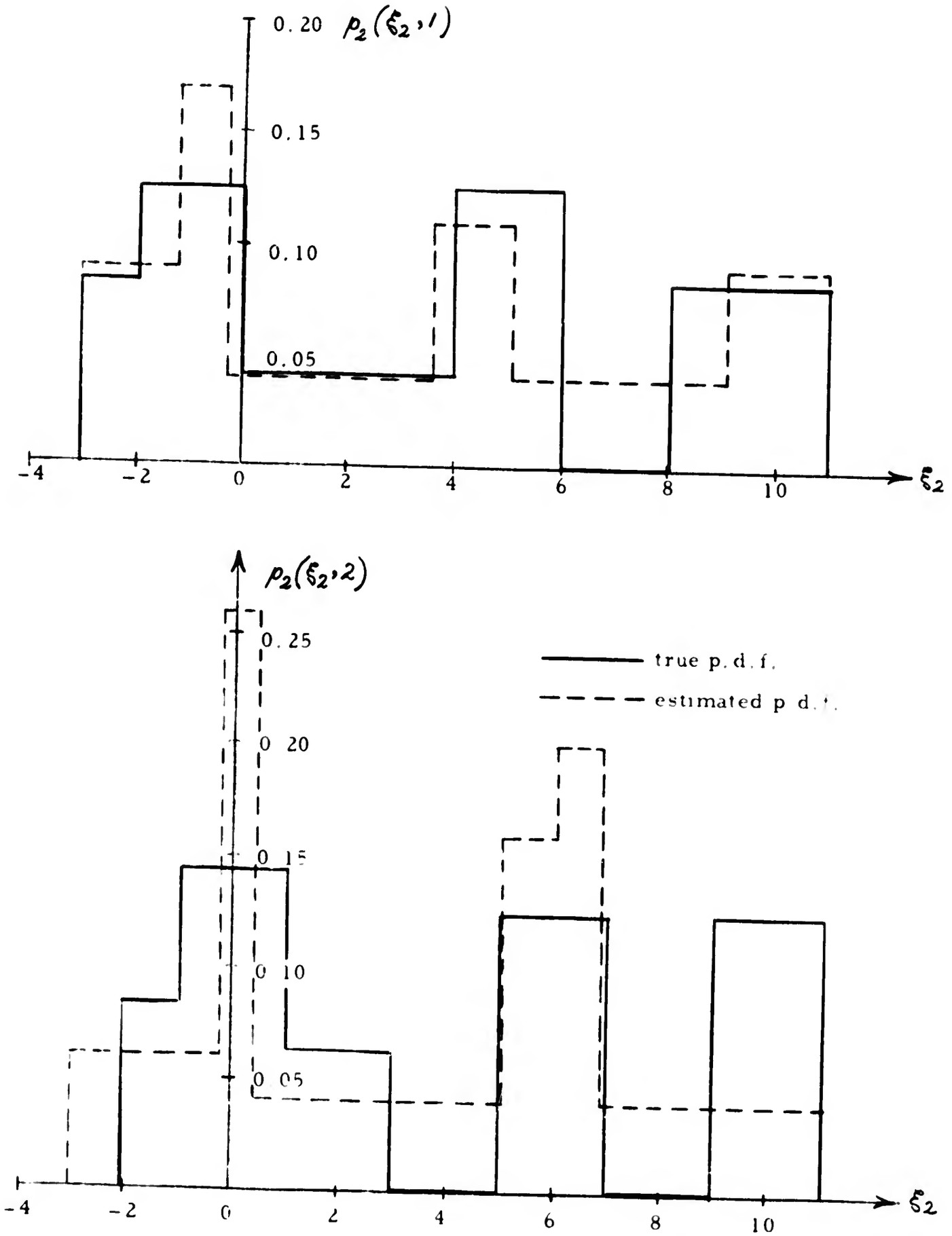


FIG. 19b True and Estimated Second Coordinate Marginal p.d.f.s, for m = 6, k = 12

3.4 HIGHER DIMENSIONAL ANALYSIS

Even if the univariate estimates generated by the technique described in Section 3.1 are highly accurate descriptions of the coordinate marginal densities, the information provided about ideal initial cell sizes is limited. What is desired is a rough idea of the "local spreads" in the N-dimensional observation space, and not just the individual coordinates. To obtain such information, the technique is extended to provide conditional marginal coordinate densities.

At the option of the experimenter the dimensionability of the observation space may be reduced as follows. The first coordinate is processed as described before and a number, μ_1, of "modes" are determined. Then, the program sorts through the data, collecting vectors whose first coordinates fall within the limits of the first "mode" determined above. These vectors are then analyzed on their second coordinate by the procedure in Section 3.1, and a number $\mu_{1,1}$ of modes are determined. The program then sorts through the (original) data collecting vectors whose first coordinates fall within the limits of the second "mode". These vectors are analyzed and a number, $\mu_{1,2}$, of "modes" are determined. This procedure is continued until the second coordinate marginal densities conditioned on the first coordinate falling in a "mode" have been determined. A total of $\mu_2 = \sum_{i=1}^{\mu_1} \mu_{1,i}$ second coordinate modes have thus been determined. The width of the smallest "mode" is recorded and may be used to determine the initial cell radius in the second coordinate.

This procedure is continued into the third and other coordinates with a total of $\mu_N = \sum_{i=1}^{\mu_{N-1}} \mu_{N-1,i}$ modes determined. The tree structure shown in FIG. 20 illustrates the general procedure. From each point labeled (i, j), and representing the j^{th} mode in the i^{th} coordinate, a number $\mu_{i,j}$ of lines lead to modes in the $(i+1)^{th}$ coordinate.

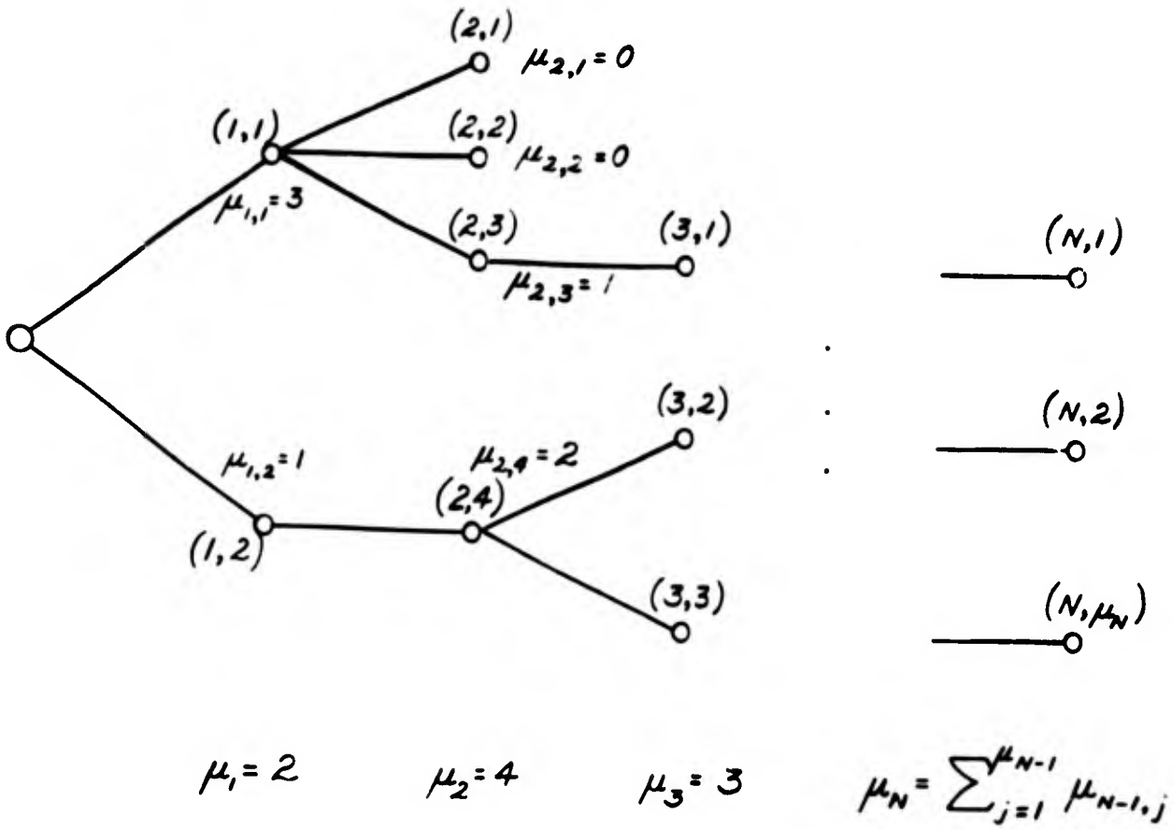


FIG. 20 A Tree Representation of the Dimensional-reducing "Mode" Seeking Proceedure

Clearly, the order in which the coordinates are processed affects the results obtained. Therefore, this order may be changed to any permutation desired by the experimenter. And, a limited number of the first few coordinates processed may be reprocessed subject to the conditions imposed by processing the latter coordinates.

This technique seems to work rather well in providing reasonable values for the distribution "local spreads". However, it has a major shortcoming. A total of $\sum_{j=1}^{\mu_{N-1}} \mu_{N,j}\, nm^N$ data vectors must be available in order to perform this analysis. This number can easily be excessive if either N or m is nontrivially large. No direct higher dimensional analog is known of the one dimensional procedure in which m cells each containing k observations are generated for each specified pair of numbers m and k.

4. ON THE MEASUREMENT SELECTION PROBLEM

As the state-of-the-art of parameter processing has grown out of its infancy, a number of questions that have remained suppressed in the minds of researchers for want of answers to more pressing problems have come to the fore front. The chief problem among these is the problem of how to select the measurements or parameters that should be processed. There are several reasons why the method of selecting the measurements is of great interest. The most important one is that a set of measurements must be found that contains enough information about all of the classes (or about the differences between classes) to permit decision making with a sufficiently low probability of error. The measurements so selected may contain sufficient information to permit a discrimination between classes; yet, for reasons of economy, a smaller set of measurements may be desired and these may be obtained by discarding the less useful ones and keeping the minimum number of those that prove most valuable in discriminating between the classes. Thus, one is led to consider the usefulness and economy of sets of measurements.

In other situations like those encountered in the recognition of shapes in two-dimensional visual patterns, it is desirable to obtain measurements that are substantially invariant with respect to the translation, rotation, and magnification of the pictorial input pattern. Parameters exhibiting such invariances may be processed by relatively simple recognition devices, such as simple correlators that can successfully compare input patterns with stored reference patterns only if the combination of input parameter values is substantially invariant over the set of inputs that belong to the same class.

The parameter selection problem has different meanings to different researchers. To some it means the desire to reduce the number of parameters that should be processed, to others it means the problem of selecting parameters that are invariant over members of the same class, while to still others it means the hope to learn something about the physical process giving rise to the input patterns by an analysis of the usefulness of the chosen parameters or the usefulness of the set of parameters derived from those initially chosen. In the discussion that follows, we will describe the various facets of the measurement selection problem. We will distinguish between those facets that must be solved by engineering reasoning and expertise in the problem being solved and those facets that admit to mathematical formulation and mathematical solution. Depending on the point of view of the system designer, the constraints he may place on the classification system or on the set of measurements, different mathematical formulations can be obtained. As in any mathematical approach to an engineering problem, mathematical models of the measurement selection problem can be set up, the system outputs can be expressed, and the achievement of the goals of the system designer can be measured by means of a figure of merit. Solutions of the measurement selection problem can then be formulated by the optimization of the figure of merit by which system performance, in terms of the achievement of the goals of the system designer, is judged. In our pursuance of this traditional route we will be concerned with models for measurement selection, design goals to be achieved, figures of merit for measuring the achievement of design goals, and the feasibility of obtaining solutions to the problems we thus formulate. Only few solutions to the measurement selection problem are presented; for the most part only the problems to be solved are formulated.

An electronic device perceives its environment only through a set of measurements and the numerical values of these measurements form a vector with which the device represents its environment. This is illustrated in the block

diagram of Figure 21 where the input is transduced by the "Measurement Subsystem" and is represented by a set of numerical values. The second block, the "Measurement Transformation Subsystem", modifies or transforms the original measurements to devise a set of "better" parameters on which the "Decision Rule" or "Classification Subsystem" operates to recognize the input pattern by distinguishing between members of different classes. We will be dealing with a design of the "Measurement Transformation Subsystem".

If an "optimum" Decision Rule or procedure were used on the original set of measurements, it would base decisions on the evaluation and comparison of the joint probability densities of the measurements computed for each of the classes of interest. These densities are either known or are "learned" from sets of samples of known classification. Theoretically, decisions with a minimum probability of error can be made on any chosen set of measurements no matter how these are selected. Even in practice, the state-of-the-art is rapidly approaching the point where nearly optimum decisions on arbitrary sets of measurements can be made. Since even the optimum Decision Rule can render classification decisions only with a generally non-zero probability of error, it follows that there is an inherent "irreducible error" associated with a given set of measurements and classes to be recognized. This irreducible error stems from the fact that the same measurement value combinations (vectors) are sometimes observed on members of several classes. Thus classes appear to overlap in the measurement space. Since any decision making system can associate only one decision (one classification) with a given input vector under conditions of overlap, any decision making system must at times be in error. Since mathematical methods can only operate, transform, or otherwise manipulate the original vector space, no mathematical technique of <u>any</u> kind can improve on the irreducible error inherent in the measurement

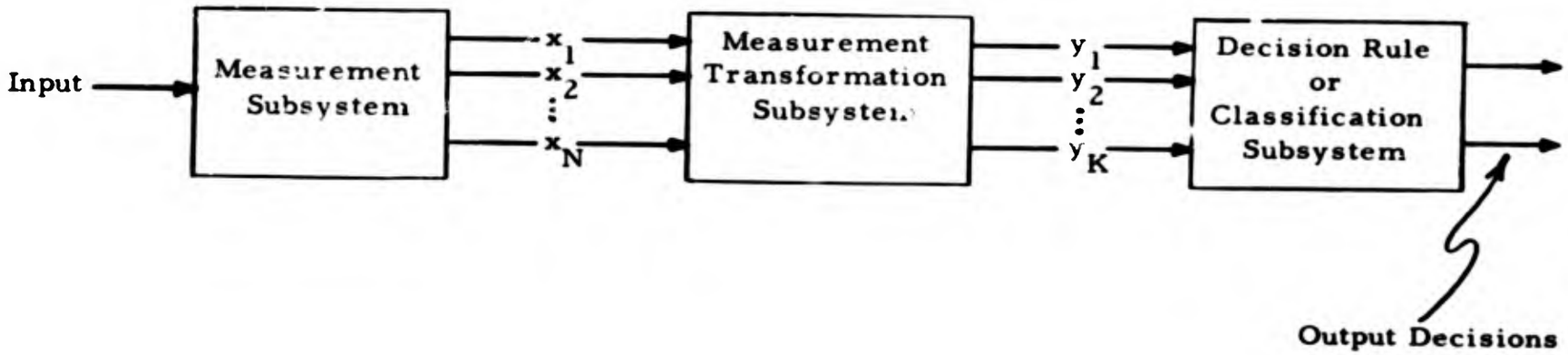


FIG. 21. Pattern Recognition System Exhibiting the Measurement Selection Problem

set by creating new functions from old measurements. Of course, the irreducible error associated with a different measurement set may be lower; however, this set cannot be derived from the original set by mathematical techniques.

The implication of the above argument is that the design of the "Measurement Subsystem" is an engineering task and this task cannot be aided by mathematics of any kind except inasmuch as it may be possible to draw inferences from mathematical results regarding the design of better sets of measurements. This limitation of mathematical methods is not necessarily a severe one in many instances. Often it is possible to choose an initial measurement set that contains all of the information about members of the classes of interest and assures that no information has been lost and, in fact, assures that the input can be reconstructed from the measurements completely. If the input is a waveform of bandwidth W and duration T, for example, a number of different measurement sets, each composed of 2 TW measurements can be devised in such a way as to assure the "completeness" of the method of representation. Of course, such an exhaustive measurement set is usually not necessary to classify the input, and we are thus led to consider the design of the "Measurement Transformation Subsystem" that should derive a smaller number of better measurements, that are still sufficient for the classification of inputs.

The design of the "Measurement Transformation Subsystem" cannot be divorced entirely from the design of the classification device. Since the figure of merit that serves as the design criterion for this subsystem is also in terms of system performance, and system performance can only be measured at the output of the decision device, the design of the "Measurement Transformation Subsystem" must include specification of the decision device that will use the new measurements. Since the proof of the pudding is in the eating, one may

use the probability of error as the figure of merit associated with the new set of measurements as the design criterion which one should seek to minimize by proper choice of the "Measurement Transformation Subsystem". It should be noted that several other figures of merit can also be envisioned. These will be discussed later. While, it was argued, the design of the original set of measurements is not a problem in mathematics, there are still two meaningful types of questions that can be asked in which mathematical methods can aid to reach a solution.

A. If there are N parameters with which to start and we insist on using only K (where K is less than N), which K measurements should be retained so that the probability of error should increase the least over the irreducible error probability associated with the N given measurements and classes to be recognized?

There are two basically different ways of formulating even this question. In one way no constraints are applied to the recognition system; that is to say, the recognition system may be assumed to be an "optimum" Decision Rule.

In an alternate way of formulating this question, and this is usually the more practical of the two, constraints are placed on the recognition system either by limiting the types of operations that it can perform or by limiting the associated storage implied by the recognition function. Different formulations of the parameter transformation problem falling within these two subclasses will be discussed later.

B. If the permissible machine complexity is constrained to a degree such that the classification system is less general than that which would be required to achieve the theoretically irreducible error probability inherent in the data and in the chosen set of parameters or measurements, what functions of the original vector space should be used to

derive a new vector space (we will call it the property space) from which the optimum machine of specified complexity should make decisions with the lowest probability of error?

This formulation of the measurement transformation problem places limitations on the allowable classification machine as well as on the allowable measurement transformations and then asks what operations on the original data will bring it into the highest degree of conformity with the machine limitations in order that the lowest probability of error should be achieved? It is important to recognize that, in this formulation, constraints are placed on the decision making device as well as on the Measurement Transformation Subsystem.

The first method of formulating the measurement transformation problem defines the problem as one of selecting a subset of the original measurements for use as a new set of properties. In the second method of formulating the subsystem, we define the property space as a set of transformations on the original measurement space and permit given classes of functions and a given type of classification function only.

Both of these methods of formulation can be expressed by Equation (36), where the N-dimensional vector $x = (x_1, x_2, \ldots, x_N)$ is the measurement vector and $y = (y_1, y_2, \ldots, y_K)$ is the property vector. In general, each of the coordinates of the property vector is some function of all N-measurement vector coordinates. In the equation, F is the figure of merit associated with a given class of classification functions operating on the property space y. Such a figure of merit, of course, includes a consideration of the functional form of the classification function. Illustrative examples will be given elsewhere:

$$F = f[y_1(x), y_2(x), \ldots y_K(x)] \tag{36}$$

4.1 Figures of Merit for a Measurement Transformation Subsystem

While in the preceding the probability of error was used to illustrate a pragmatic figure of merit which expresses the utility of measurement transformations directly in terms of overall system performance, a number of other figures of merit can be envisioned. Generally, these fall into two groups - one in which no assumption regarding the Decision Rule has to be made (other than that it is optimum), and the second where limitations of the classification function form an integral part of the figure of merit. For mathematical convenience, the nature of the classification function can be taken into account indirectly by specifying the desirable properties of the property space instead. A number of candidates for figures of merit in evaluating Measurement Transformation Subsystems are given below.

4.1.1 Risk and Average Probability of Error as Figures of Merit

A Decision Rule is a procedure for assigning a decision (or label) to every observation that consists of a set of property values represented by the vector y. The rule may state, for example, that when a specific vector y_o is observed, it is most likely a member of class C_i; hence it should be classified as a member of C_i. But if y_o (in reality) is a member of C_j instead, the decision is in error and the decision maker should pay the penalty by incurring a certain cost or loss, L. Since, once designed, the Decision Rule is fixed and it renders a specific classification decision upon observing each vector y, the penalty or loss, L, suffered by the decision maker upon making a decision depends on the observed vector (and thus the decision rendered) and on the actual class identity, C_i, of the input. The average loss (or "risk") of decision making is given in Equation (37), where loss has been averaged over all possible observations, y, and all classes, C_i. The a priori probability of class C_i is $P(C_i)$.

$$\text{Risk} = E[\, L(y,\ C_i\,] = \sum_{C_i} \int_{x} P(C_i)\ L\ (y,\ C_i)\ p(y \mid C_i)\ dy \tag{37}$$

Risk measures the discriminability of classes with a given property set and decision procedure. If we postulate an optimum decision procedure at all times, Risk measures the discriminability between classes achievable with a given set of properties and thus serves as a figure of merit of the property set.

Designing a Measurement Transformation Subsystem entails the choice of a set of properties that minimize Risk either by a choice of a suitable subset of the N original measurements or by the choice of the transformation of measurements to a set of properties.

If the loss suffered by a wrong decision is one and that of a correct decision is zero, the Risk becomes the Average Probability of Error, given in Equation (38) where $\overline{R_i}$ is the region of the property space where inputs are classified as other than C_i.

$$\text{Expected Probability of error} = \sum_{C_i} \int_{\overline{R_i}} P(c_i)\ p(y \mid C_i)\ dy \tag{38}$$

This region is the complement of R_i in which inputs are classified as members of C_i. Thus the Expected Probability of Correct Recognition, EPCR, is given in Equation (39) and is a useful figure of merit that increases monotonically with increasing performance.

$$\text{EPCR} = \sum_{C_i} \int_{R_i} P(C_i)\ p(y \mid C_i)\ dy \tag{39}$$

These equations can be checked readily by recognizing that each of the integrals over the R_i regions is unity if the probability densities are disjointed, causing EPCR to become unity, a fact consistent with our knowledge that no errors will ever be made if the classes occupy non-overlapping regions in property space. This figure of merit (EPCR) is illustrated in Appendix II for the simple case where an optimum classification function operates as a single property y that may belong either to class C_i or C_j with equal a priori probability. Each class is Gaussianly distributed with variance σ^2 and means μ_i and μ_j, respectively. It is shown that the figure of merit, ECPR, is a monotonically increasing function of the separation-to-spread ratio, $(\mu_i - \mu_j)/\sigma$, a result that agrees with intuition. This result can be extended to the N-dimensional case and has been done so by Anderson.

This apparently simple expression for Risk (or for ECPR) is, upon close examination, a very nasty expression indeed, and is one that does not lend itself readily to analytical or computational manipulations. Unfortunately, the regions of integration (the R_i's) can be determined only by evaluating the Decision Rule, a comparison of likelihood ratios with constants. Even if the conditional probability densities inside the integrals were known (they can be obtained by the adaptive approximation technique described previously), the search for better properties cannot be carried out analytically by application of this figure of merit except in certain special cases such as those where the different classes are Gaussian densities. This is the case treated in the scientific literature*.

---

*T. Marill and D. M. Green, "On the Effectiveness of Receptors in Recognition Systems", IEEE Transactions on Information Theory, January 1963.

P. M. Lewis, "The Characteristic Selection Problem in Recognition Systems" IEEE Transactions on Information Theory, February 1962.

4.1.2 Information-Theoretic Figures of Merit

Information theory deals with the quantitative measurement of changes in the state of knowledge of an observer that occur when an observation or measurement is made. In relation to the measurement transformation problem, information theory is applicable to the measurement or assessment of the amount of information contained in a set of properties about the classes (or about the differences between classes) we wish to recognize. We will discuss various quantities that can be used as figures of merit in the design of measurement transformation subsystems.

A. Entropy

Clusterability or "anticlusterability" (spread) can be measured directly by entropy. Entropy of a class of things distributed according to the probability distribution p(x), where x is a vector measurement, is given by Equation (40).

$$\text{Entropy} = H = -\int_{-\infty}^{\infty} p_i(x) \log p_i(x)\, dx \tag{40}$$

Note that all entropy is zero if all members of the class C_i have the same vector representations. For such perfect clustering, the spread and the entropy are zero. Note also that the entropy of a Gaussian distribution, given in Equation (41) is a monotonically increasing function of its variance. The more the spread, the higher the entropy.

$$H = \log (\sigma \sqrt{2\pi e}) \quad \text{for a Gaussian process} \tag{41}$$

A typical way in which entropy as a figure of merit could be applied to the measurement transformation problem is given below. Suppose we are given a very high-dimensional (N-dimensional) measurement space and samples of patterns distributed according to probability distributions $p_1(x)$, $p_2(x)$, . . . ,

$p_i(x)$, where $p_i(x)$ is the distribution of members of class C_i in the measurement space x. We wish to obtain a suitably small, K, number of transformations of the measurements $y_1(x)$, $y_2(x)$, . . . , $y_K(x)$, for the purpose of reducing the number of variables on which the classification function must operate ($K \leq N$), and for the purpose of increasing the clustering of members of all classes in the K-dimensional property space.

If we regard the smallness of the entropy of the class distributions in the K-dimensional property space, y, as a figure of merit that measures the fulfillment of the above stated desires, we could state the measurement transformation problem as follows.

Find a set of transformations $y_1(x) \ldots y_K(x)$ such that the sum of entropies of the densities $p_1(y)$, $p_2(y)$, $\ldots p_i(y)$ is minimum given $p_1(x)$, $\ldots$ $p_i(x)$, the densities in the measurement space. The class of functions $\{y(x)\}$ must be specified.

The solution of the above stated problem for the case where the $p_i(x)$ distributions are multivariate normal with arbitrary and different covariance matrices from class to class is derived in Appendix III when the constraint on y(x) transformations is that they be linear. This means that each y(x) is a linear transformation (a resistive network or a correlator) and the above minimization problem is used to design the best K such networks.

The solution presented in Appendix III states that we proceed in the following steps:

1. First form the covariance matrix of each class distribution U_i and form the matrix W by multiplying together all coveriance matrices in the x-space.

2. Then solve for the K smallest eigenvalues and vectors of the matrix W.

3. The set of eigenvectors thus derived specify the K linear transformations of the x-space that hield the best K-dimensional y-space in which the sum (or average) entropy of classes is minimized.

Appendix III contains a geometrical interpretation of this solution for a special case.

B. Information Gain

The Risk or Probability of Error, used as a figure of merit for evaluating the combined Measurement Transformation Subsystem and classification function, is a computationally difficult figure of merit to apply. For this reason a figure of merit operating on the interface between the classification function and the Measurement Transformation Subsystem was considered in the preceding subsection. Entropy (measured on the property space) is such a figure of merit and it is a measure of the degree of clustering of the classes.

Another somewhat similar information theoretic measure with which a set of properties can be graded is the amount of information a property set provides about the classes. The more informative a property set is, the more useful it is to the decision maker (although this does not assure in any way that classes will be more discriminable from one another in the property space). The information gained about a class C_i, when the set of property values (given by the vector) y are observed, is the difference between the information we had about C_i without the observation y and with the observation y. The information gained about C_i due to the observation y is given in Equation (42).

$$I_{C_i y} = \log \frac{p(C_i \mid y)}{p(C_i)} = \log \frac{p(y \mid C_i)}{p(y)} = \log \frac{p(C_i, y)}{p(C_i)\, p(y)} \qquad (42)$$

The various equivalent forms of information gain (or information transfer) are presented above to facilitate subsequent manipulations.

The information gained about C_i from the property set y, on the average, is given in Equation (43).

$$I_{C_i} = \int p(y \mid C_i) \log \frac{p(y \mid C_i)}{p(y)} \, dy \tag{43}$$

The average information gained about classes, in general, from the parameter set y, on the average, is given in Equation (44).

$$I = \sum_{i=1}^{C} P(C_i) \int p(y \mid C_i) \log \frac{p(y \mid C_i)}{p(y)} \, dy \tag{44}$$

If we did not consider the information gain about discrete classes but rather about a continuous variable C, we would have the familiar expression given in Equation (45),

$$I = \int\int p(C, y) \log \frac{p(C, y)}{p(C)\, p(y)} \, dy \tag{45}$$

where liberal use of (42) and the conventional manipulations on conditional and a priori probabilities was made.

Although obtained through different reasoning, P.M. Lewis in the above referenced article employs this figure of merit (measure of "goodness") to help to evaluate the relative utilities of different measurement subsets (for the case of Gaussian processes with independent variables).

The evaluation of I in a practical case for given data is not nearly as foreboding as the apparent complexity of the above equations would indicate.

C. Divergence

The greater the information gained about classes from a parameter set y, the better we are able to characterize classes. The more complete the characterization of classes, the better the classification system can become

to distinguish members of classes from one another. If our sole objective is to discriminate classes from one another, however, we do not necessarily have to be able to describe the classes. It is sufficient if we can describe the differences between classes. A generalized measure of the distance or differences between classes (their probability densities in the property space) is the "divergence" given by Kullback and written in Equation (46). The divergence is a measure of "discriminability" between two classes C_i and C_j.

$$J(C_i, C_j) = \int \left[p(y \mid C_i) - p(y \mid C_j) \right] \log \frac{p(y \mid C_i)}{p(y \mid C_j)} \, dy \qquad (46)$$

It can be shown that if we have two property sets, A and B, such that the "irreducible" error due to set A is lower than due to set B, then it follows that the discriminability between classes due to set A is greater than that due to set B. Unfortunately, the converse of this statement does not hold, in general. If it did, our minimization problems that attempt to maximize performance could be expressed by means of well-behaved quantities like those given in Equation (46) instead of the nasty expressions of probability of error.

Nevertheless, divergence is a useful figure of merit that can measure something closely related to the error rate.

Appendix IVillustrates (for two exponential probability densities) that in some instances a change in the property set (a change in corresponding probability densities) which increases the divergence also decreases the error probability. This is shown by showing that the algebraic sign of the derivative of the divergence (with respect to a parameter that varies the probability densities) is opposite to the algebraic sign of the derivative of the error probability (with respect to the same parameter).

4.2 Methods of Optimizing Measurement Transformations

In the following we will summarize the key statements made so far.

A. There is an irreducible error that is associated with the classes and with the choice of the measurement space.

B. The ultimate figure of performance, the probability of error, cannot be reduced by any kind of measurement transformation from that which could be obtained on the measurement space with an optimum decision procedure.

C. There are two legitimate types of problems that can be considered in deriving measurement transformations. One is to try to select the best subset of the given measurements as a set of properties with or without any restrictions being applied to the classification function that will utilize the properties. The second is to try to derive a set of transformations (with suitable restrictions on the classes of transformations) which, together with a classification function (on which there are also imposed some restrictions), will optimize the figure of merit we use as a measure of system performance.

D. There are two classes of figures of merit that can be used in the above optimization procedures. One operates at the output of the classification function and measures the worth of a property set by measuring overall system performance obtained with the property set. The second operates on the property space and measures characteristics of class distributions in the property space with a view toward discriminability of classes or with a view toward clustering of classes but without applying the figure of merit to overall system performance. From our point of view, the second type of figure of merit is the more convenient to utilize because it allows us to take into consideration the nature of the classification function in the characterization of the class distributions in property space without the computational difficulties that arise in calculating probability of error.

E. From a practical point of view, the behavior of the measurement transformation subsystem is of no relevence if the input classes are Gaussian. In fact, if the input classes are unimodal, then they are already clustered and it is no longer necessary to develop transformations that simplify the class distributions. Therefore, the situation where the densities in the measurement space are multimodal, a situation that can often be characterized by saying that an input class consists of subclasses, is of interest. If the input can be divided into subclasses then, usually, there are indeed several modes. It does not follow, however, that subclasses correspond to a single mode of the class distribution.

In the following, we will list some of the optimization problems with which the problem of deriving measurement transformations can be expressed.

Problem 1 - Find the linear transformation of the measurement space such that the average entropy of the class distributions in the property space is minimized. This optimization problem, already discussed for a special case, assures that a property space is derived in which, on the average, classes are maximally clustered. In Appendix III this optimization was carried out for the case where classes are multivariate distributions, and the linear transformation is orthonormal and maps the N-dimensional measurement space into a K-dimensional property space. The most serious practical shortcoming of the results obtained in Appendix III comes from the Gaussian assumption of the class distributions. As was pointed out above, if classes are already Gaussianly distributed, there is little need to develop transformations, except for the purpose of reducing the dimensionality of the measurement space or to reduce the correlation between variables.

Practical but more general constraints on the classes of transformations would include polynomial transformations or piece-wise linear transformations.

Polynomial transformations (where y(x) is the polynomial function of x) are relatively easily handled analytically because the functions one obtains are all expressible in terms of the moments of the probability densities. In addition, it is easy to derive algorithmic methods of obtaining these polynomials from a finite number of samples of the pattern classes. Piece-wise linear transformations, on the other hand, are easily instrumented with linear (resistive) networks coupled with amplitude comparators (Schmitt-trigger circuits).

Problem 2 - Derive a set of non-linear transformations (like polynomials of degree R) which minimize, on the average, the mean square distance between members of the same class in property space while they hold the distance between the means of classes in the property space constant. This optimization problem expresses the notion of clustering as a mean square distance between members of the classes, and attempts to minimize the cluster diameter while keeping the distance between clusters constant. *

Such a clustering transformation is given in Eq. (3.13) of the reference, and the solution for the coefficients of the polynomial transformation is derived in Eq. (3.14) which is given here as Equation (47).

$$a_n = \left(\frac{Y_{on}}{z_n [U^{-1}(n)] \, z_n^T} \right) Z_n [U^{-1}(n)] \tag{47}$$

The polynomial transformation can be expressed in Equation (48) where y_n is the n^{th} property and the x_n is the n^{th} measurement. Elements of the matrix U are defined by Equation (49) and z_n is the vector whose components

---

* This type of optimination problem is described in Chapter 3 of Reference 1.

are powers of an arbitrarily chosen point of the property space which is chosen to serve as a scale factor of the property space. These quantities are defined in more detail on pages 57-61 of the above reference. The vector a_n defines one row of the transforming matrix.

$$y_n = \sum_{p=1}^{R} a_{np} x^p_n \tag{48}$$

$$u_{rp}(n) = \overline{x^r_n x^p_n} - \overline{x^r_n}\ \overline{x^p_n} \quad \text{for } x \in C_i \tag{49}$$

As can be seen from the nature of the above solution, practical application of these transformations is severely limited by their relative complexity which is manifested by the implied computation time necessary to carry out the indicated operations. Certain algorithmic techniques for obtaining transformation of like character can, however, be obtained.

Problem 3 - Since a minimization of average entropy assures clustering of classes in property space but does not assure the retention of separability of classes, a combination of concepts of entropy and divergence into a single optimization problem seems warranted. A useful method of expressing an optimization problem that combines these two fundamental notions is given below.

Find a set of transformations that minimize, on the average, the entropies of class distributions in property space while the keep constant the average divergence (discriminability between classes). This optimization problem, subject to suitable constraints to permit its solution, is as nearly an optimum expression of the desirable characteristics of the measurement transformation problem as can be expected. Minimization of entropies while holding probability of error constants would be a better statement of our desired objectives. This statement, however, would involve computational steps for which an exact specification of the classification function must be used.

The above optimization problem is expressed in Equations (50) through (52).

$$\frac{\partial \{ E[H] \}}{\partial y_K(x)} + \lambda_K \frac{\partial \{ E[J(i,j)] \}}{\partial y_K(x)} = 0 \tag{50}$$

where

$$E[H] = \sum_i P(C_i) \int p(y \mid C_i) \log p(y \mid C_i) \, dy = \text{expected entropy of classes} \tag{51}$$

and

$$E]J(i,j)] = \sum \sum \int [p(y \mid C_j) - p(y \mid C_j)] \log \frac{p(y \mid C_i)}{p(y \mid C_j)} \, dy = \text{average divergence} \tag{52}$$

This type of problem statement is analogous to the type stated in Problem 2.

4.3 Approximate Solutions and Computational Steps

In the preceding sections, mathematical formulations and a few illustrative solutions of the measurement transformation problem were given. In many instances, however, it is not practical to strive for exact solutions of mathematical optimization problems. It may be more advantageous to develop simple algorithmic procedures that attempt to achieve the desired objectives qualitatively. In this section, a single illustrative example is given to demonstrate the type of technique that is motivated by the desired mathematical objectives but one that keeps computability and real-time instrumentability in front of the designer at all times to assure that a practical solution of the measurement transformation problem be obtained.

Suppose we are given three classes A, B, and C, shown in two dimensions in Figure 22. Here all three classes are bimodal and hence relatively complex.

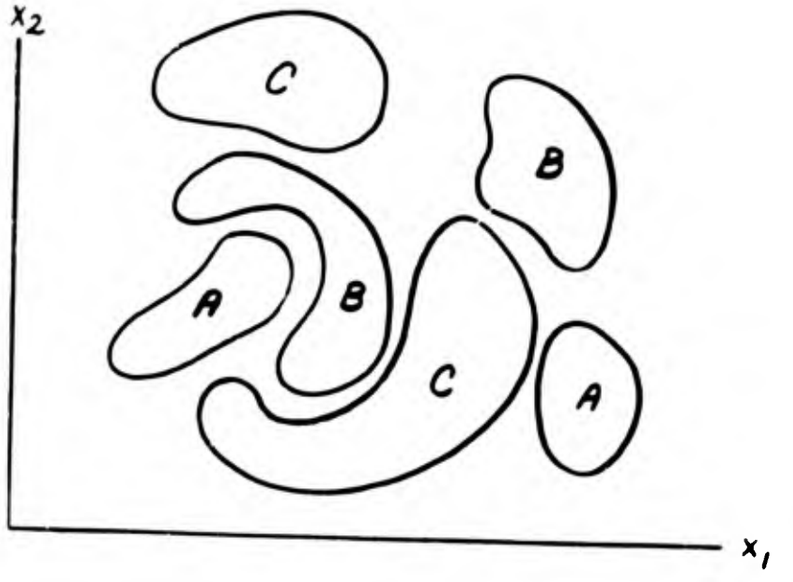
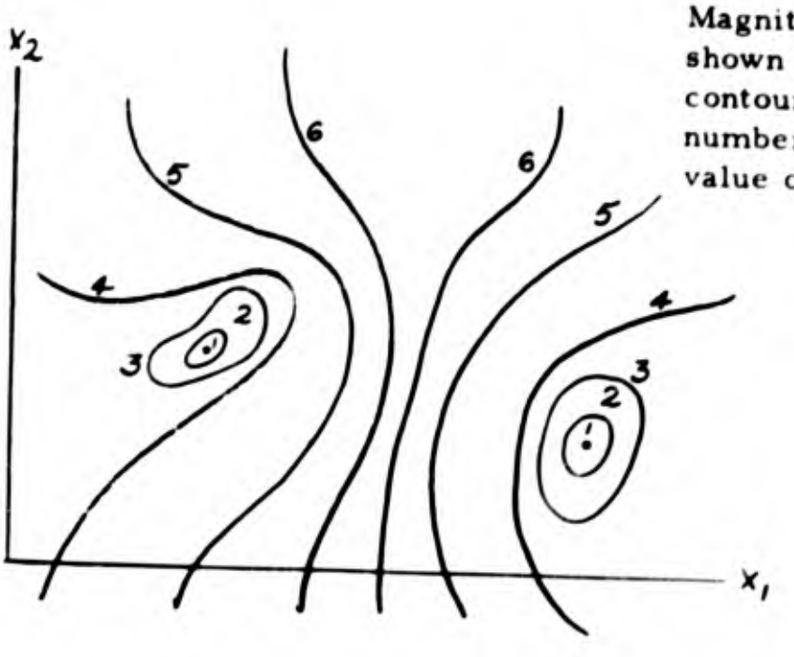


FIG. 22 Three Multiply Connected Classes, A, B, C



Magnitude of $y_1(x)$ is shown with numbered contour lines (small number indicates large value of $y_1(x)$

FIG. 23. $y_1(x)$ Shown as a Contour Map

We can develop a procedure for mapping the entire x space onto a line $y_1(x)$ so that samples of class A are maximally clustered along $y_1(x)$ and samples of not A (B and C) are also clustered along $y_1(x)$, but A and not A are separated from one another. This mapping is achieved with a nonlinear transformation to be described.

Similarly, members of class B can be made to cluster along $y_2(x)$ so that they are separated from members of not B. The transformation $y_3(x)$, similarly derived, will cluster members of class C on $y_3(x)$ and separate them from not C.

In the three-dimensional space formed from $y_1(x)$, $y_2(x)$, and $y_3(x)$ classes A, B, and C occupy disjointed and clustered (simply connected) regions.

It can be shown* that the optimum transformation $y_1(x)$ that clusters members of class A and separates them from not A is given by Equation (53) where $\bar{A}$ stands for "not A". **

$$y_1(x) = \frac{p(x|A) - p(x|\bar{A})}{p(x|A) + p(x|\bar{A})} = \frac{p(x|A) - p(x|\bar{A})}{p(x)} \tag{53}$$

$$y_1(x) = \frac{p(x|A)}{p(x)} - \frac{p(x|\bar{A})}{p(x)} \tag{54}$$

We thus need to know the probability density of members of A and of not A. A practical method of obtaining a good estimate of a probability density was described in Chapter 2 of this report.

---

* It is shown in Sections 3.2 and 3.3 of "Decision-Making Processes in Pattern Recognition" (pp. 61 - 68) Macmillan, 1962.

**Incidentally, note the similarity between $y_1(x)$, as written in (54) and the expressions for information gain given in Equation (42).

The nature of the function $y_1(x)$ can be illustrated graphically for the class distributions of Figure 22. The illustration, shown in Figure 23, shows $y_1(x)$ as a contour map, with the two highest points indicated by the two points of Figure 23, and decreasing values of $y_1(x)$ indicated by contour lines of increasing index. Qualitively, it is seen that $y_1(x)$ is large where members of class A occur in the measurement space and small where members of not A are observed. The transformations $y_2(x)$ and $y_3(x)$ can be interpreted similarly.

The description of the classes in the three-dimensional y space is shown in Figure 24. It is seen that each class is unimodal (simply connected) in the y space. This is readily appreciated by noting that, on $y_1(x)$, A is clustered and B union C, or not A, is clustered (but in a different interval of $y_1(x)$). Similar argument holds for y_2 and y_3.

The resulting transformation cannot be expressed analytically in a convenient manner. Algorithmically, however, these are readily obtained by simple extensions of the adaptive probability density approximation algorithm described earlier.

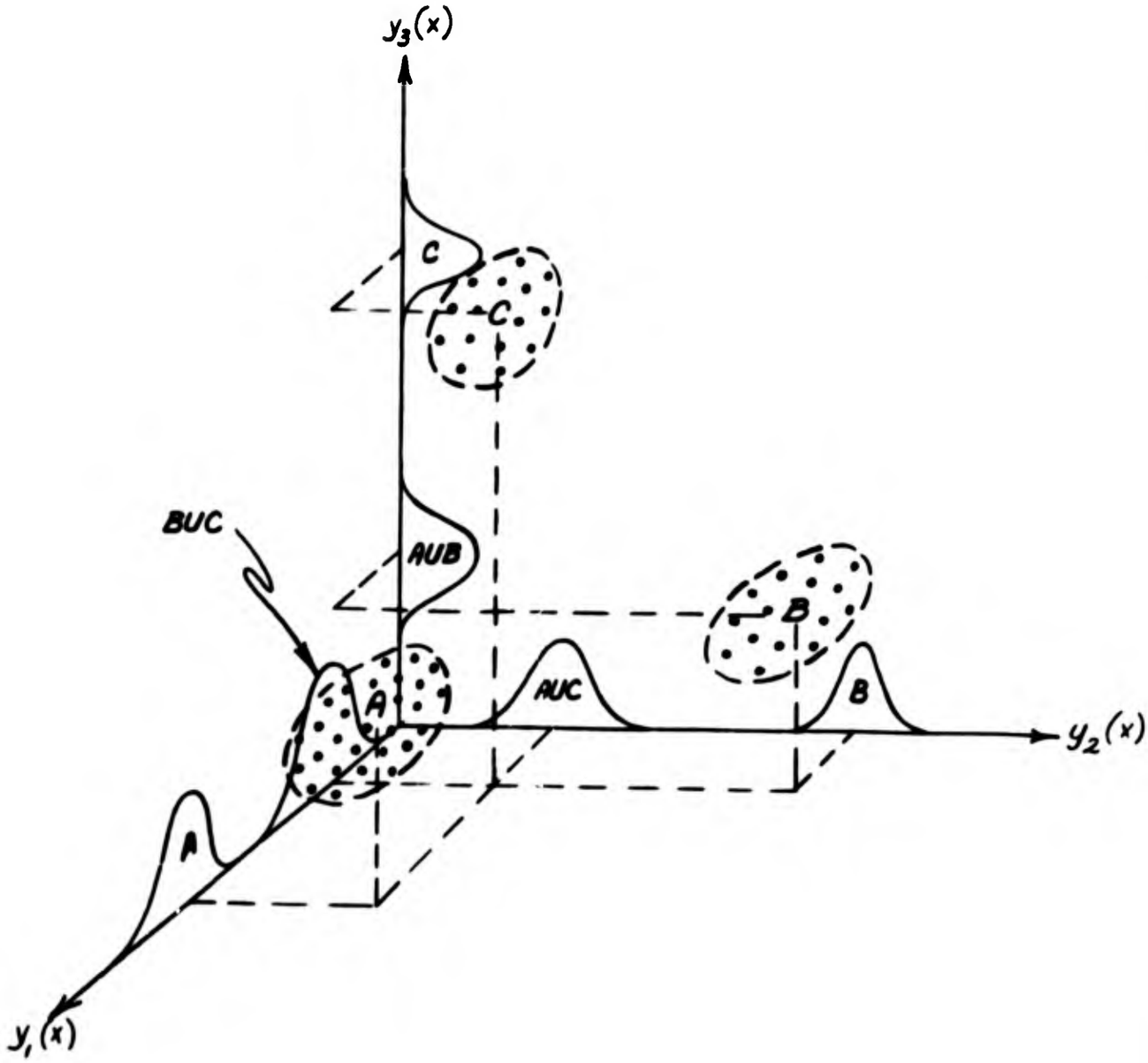


FIG. 24. A, B, and C are All Clustered in y Space

REFERENCES

1. Stram, O.B., "Arbitrary Boolean Functions of N Variables Realizable in Terms of Threshold Devices," Proc. of IRE, Vol. 49, (1961), pp. 210-220.

2. Mattson, R.L., "The Analysis and Synthesis of Adaptive Systems Containing Networks of Threshold Elements," Ph.D. Thesis, Stanford University, Stanford, Calif., June 1962.

3. Gabelman, I.J., "A Note on the Realization of Boolean Functions Using a Single Threshold Element," Proc. IRE (Correspondence), Vol. 50, No. 2, Feb 1962, pp. 225-226.

4. Ball, G.H., The Application of Integral Geometry to Machine Recognition of Visual Patterns, WESCON, 1962, (6.3).

5. Sebestyen, G.S., Decision Making Processes in Pattern Recognition, Macmillan, New York, 1962.

6. Litton Systems, Inc., Information Sciences Laboratory, Pattern Recognition Research, Scientific Report No. 1, by J. Edie, W. Floyd, and G.S. Sebestyen, Report AFCRI 63-548, Contract AF 19(628)-1604, Waltham, Mass., 1963.

7. Litton Systems, Inc., Information Sciences Laboratory, Investigation of Automation of Speech Processing for Voice Communications by G. Sebestyen and D. VanMeter, Final Report, Contract AF (604)-8828, Report No. AFCRL 62-946, Section 4, Waltham, Mass., Oct 1962.

8. Edie, J., W. Floyd, and G. Sebestyen, Op. cit (6), p. A VI-13.

9. Highleyman, W.H., "The Design and Analysis of Pattern Recognition Experiments," Bell Systems Technical Journal, Mar. 1962, pp. 723-744.

10. Highleyman, W.H., op. cit (9), Table I, p. 730.

11. Litton Systems, Inc., Information Sciences Laboratory, Error Probabilities Conditioned on Specific Observations (internal memorandum) by A.H. Nuttall, Waltham, Mass., 20 Dec. 1961.

12. Litton Systems, Inc., Information Sciences Laboratory, Pattern Recognition Research, Second Quarterly Status Report, by J.Edie, Contract AF 19(628)-1604, Waltham, Mass., Nov. 1962.

13. Clopper, C.J. and E.S. Pearson, "The Use of Confidence of Fiducial Limits Illustrated in the Case of the Binomial," Biometrika, Vol. 26, 1934, pp. 404-413.

14. Crow, E.L., F.A. Davis, and M.W. Maxfield, Statistics Manual, NAVORD Report 3369, NOTS 948, U.S. Naval Ordnance Test Station (and as reproduced by Dover Publications, Inc.).

15. Litton Systems, Inc., Information Sciences Laboratory, Pattern Recognition Research, Third Quarterly Status Report, by W. Floyd and J. Edie, Contract AF 19(628)-1604, Appendix B, Waltham, Mass., Feb. 1963.

16. Edie, J., W. Floyd and G. Sebestyen, op. cit (6), Appendix V.

17. Mood, A., Introduction to the Theory of Statistics, McGraw-Hill, New York, 1950, p. 387.

18. Marill, T. and D.M. Green, "On the Effectiveness of Receptors in Recognition Systems," IEEE Transactions on Information Theory, Vol. IT-9, No. 1 (Jan. 1963), pp. 11-17.

19. Lewis, P.M., "The Characteristic Selection Problem in Recognition Systems," IEEE Transactions on Information Theory, Vol. IT-8, No. 2, (Feb. 1962), pp. 171-178.

BLANK PAGE

APPENDIX I

INVESTIGATION OF ORDER DEPENDENCE OF THE ADAPTIVE PROBABILITY DENSITY ESTIMATION TECHNIQUE

An algorithmic procedure described in Section 2 of this report for estimating the joint probability densities of pattern classes from a finite number of samples is obviously dependent on the order in which pattern samples are introduced. This Appendix contains a description of a set of experiments performed to determine the degree to which the technique is dependent on the order in which samples are introduced. Instead of the approximation technique described in Section 2, these experiments were conducted with the aid of the "Adaptive Sample Set Construction" technique (ASSC*) which is a predecessor of the technique described in this report. The differences between the ASSC and the present technique are:

A. That in ASSC the cell size is not updated (it remains at its initial value)

B That the probability densities are approximated by a sum of Gaussian processes whose means and variances are the typical sample vectors and the pre-determined cell shape, respectively.

In the technique described in this report, instead of the sum of Gaussian densities we employ only the density whose mean is nearest to the input vector. The difference in the approximation technique is not too significant (except from

*Described in detail in "Pattern Recognition by an Adaptive Method of Sample Set Construction", PGIT, Vol IT-8, No. 5, September 1962

a computational point of view) for the contribution to the estimated probability density in ASSC is due mostly to a single term in the sum of Gaussian processes.

Four computer experiments were conducted. In the first experiment, a random number generating program was used to generate samples of a Gaussian process with 0 mean and unit standard deviation. Control parameters of the ASSC Learning Program were set at: THR = 1, $\sigma = 3/4$ and $\theta = 1$.

The means and variances of the typical samples created by the learning program are given below:

TABLE I

Sample Number	Mean	Variance	No. of Occurrences
1	-1.577	0.563 (σ=3/4)	23
2	0.368	0.563 (σ=3/4)	51
3	-0.652	0.563 (σ=3/4)	44
4	1.385	0.563 (σ=3/4)	29
5	2.268	0.563 (σ=3/4)	3

Five "typical samples" from 150 input samples were created. The corresponding Gaussian sub-populations are shown in Figure A-1 (labeled S_1 through S_5) and their sum (the result of approximating the probability density of the input variable) is compared with the known distribution of the input process (which is labeled G). The area between the two curves is 0.2605 and the largest percentage area in the region -1.5 to + 1.5 is 27 per cent.

It is interesting to note the order in which the typical samples are generated. The estimate of the mean is excellent while the standard deviation is somewhat large. It should be noted, however, that the initial choice of standard deviation

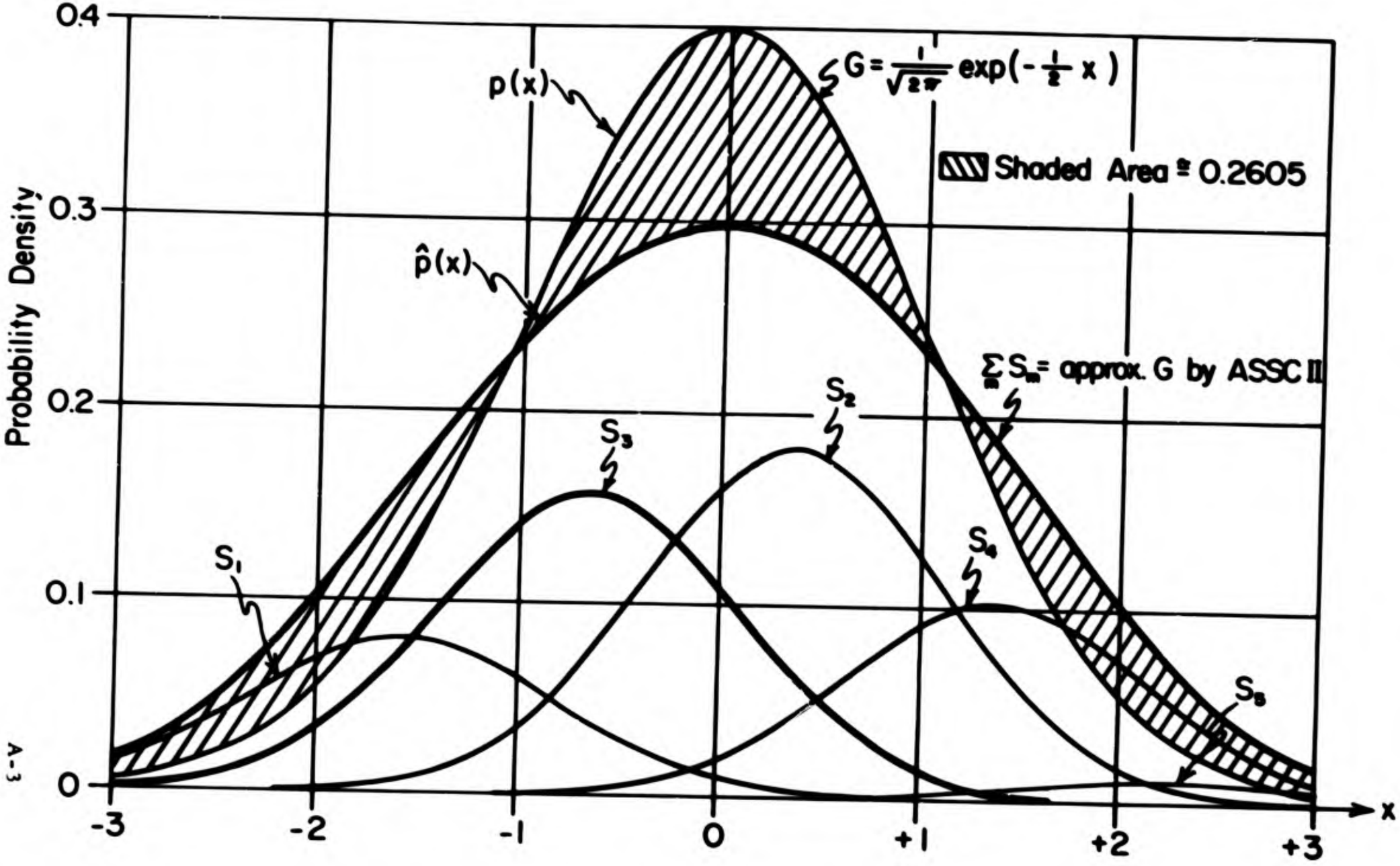


FIG. A-1. Approximation of G by ASSC II based on 150 samples of G
(Computer Run No. 1)

of 3/4 proved to be too large since the standard deviations of the sub-populations remained at its initial value. *

Identically, the same experiment was repeated with a different 150 sample vedtors of the 0 mean, unit standard deviation Gaussian process The results of this computer run are shown in FIG A-2 from which it is seen that 7 typical sample vectors, given in Table 2, were created.

TABLE 2

Sample Number	Mean	Variance	No. of Occurrences
1	- .7636	0 563	51
2	-1 704	0 563	12
3	.7311	0 563	24
4	.07689	0 563	45
5	2 537	0. 563	3
6	1 524	0 563	13
7	-2. 881	0 563	2

The area between the approximation (the sum curve) and the curve labeled "G", is 0. 1968 and seen to be considerably less than the corresponding error in the first computer run. The maximum percentage error between the actual and approximate probability density in the region -1 5 to + 1 5 is also 27 per cent but it occurs at the extreme of the interval Generally, the approximation is better than on Run No. 1. The best estimate of the mean of the approximation is -0. 2 and the standard deviation is in better agreement with that of curve "G"

* The program normally selects either the sample variances of the sub-population or the initial choice of the sub-population variance, whichever is greater.

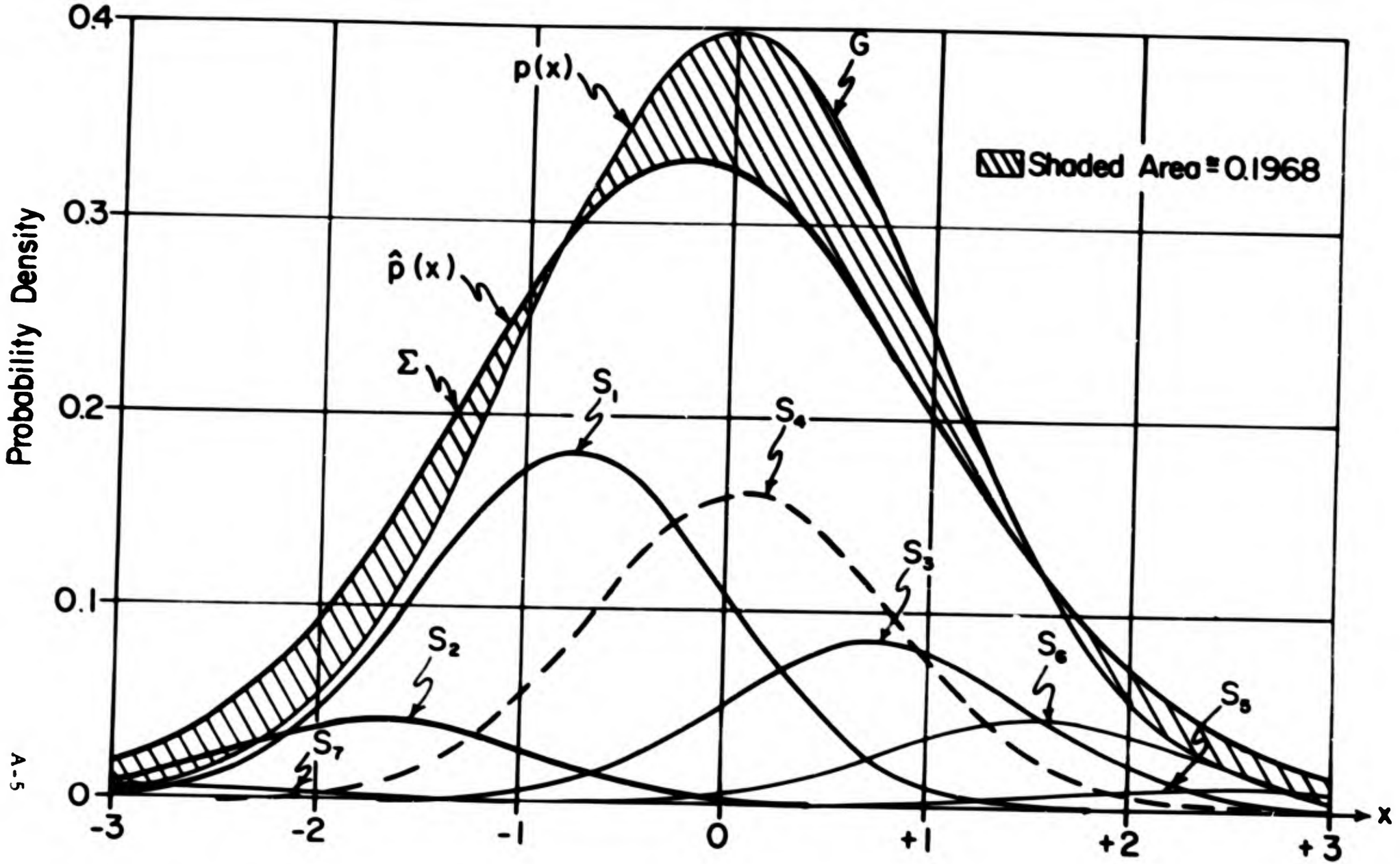


FIG. A-2. Approximation of G by ASSC II based on 150 samples of G

(Computer Run No. 2)

than in the first computer run. The approximation and the relative qualitative independence between actual density exceeds initial expectations.

A third computer run for which the data was drawn from a bimodal distribution was also performed. The 300 samples of the input process were obtained from the sum of two Gaussian densities; one with a mean of -2 and unit standard deviation, the other with a mean of +2 and unit standard deviation. Samples from these two distributions were mixed by taking samples from the two processes alternatingly. The resulting distribution is shown in FIG A-3 and is labeled "G". The learning program created ten typical vectors, shown in Table 3, whose locations and relative frequencies of occurrence are shown by the positions and magnitudes of the vertical bars of FIG A-4. The sum curve seems to be in good agreement with the distribution of the input process. The overall approximation is quite good with the exception that the variance of the approximation is somewhat high and the estimate of the second mode (the one located at +2 appears to be somewhat low). The area between the two curves is 0.1522.

TABLE 3

Sample Number	Mean	Variance	No. of Occurrences
1	-1.286	0.563	42
2	2.145	0.563	47
3	- .1963	0.563	22
4	-2.1	0.563	60
5	-2.993	0.563	28
6	1.196	0.563	55
7	2.969	0.563	33
8	4.706	0.563	4

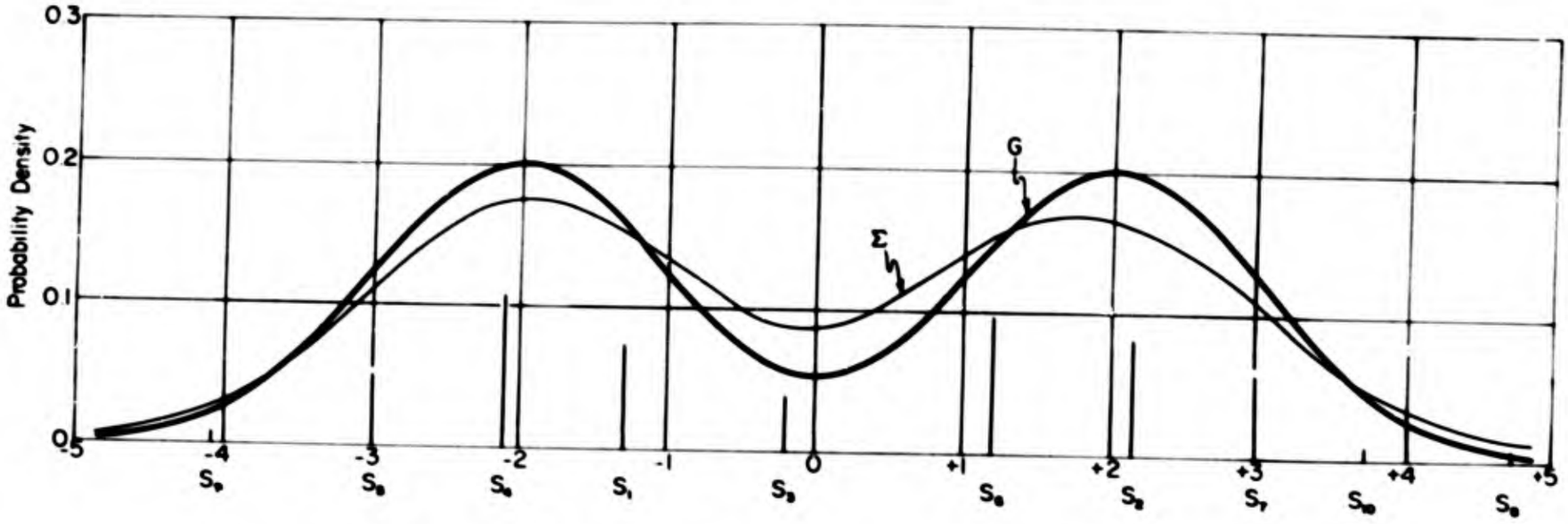


FIG. A-3. Approximation of G by ASSC II based on 300 samples of G

(Computer Run No. 3)

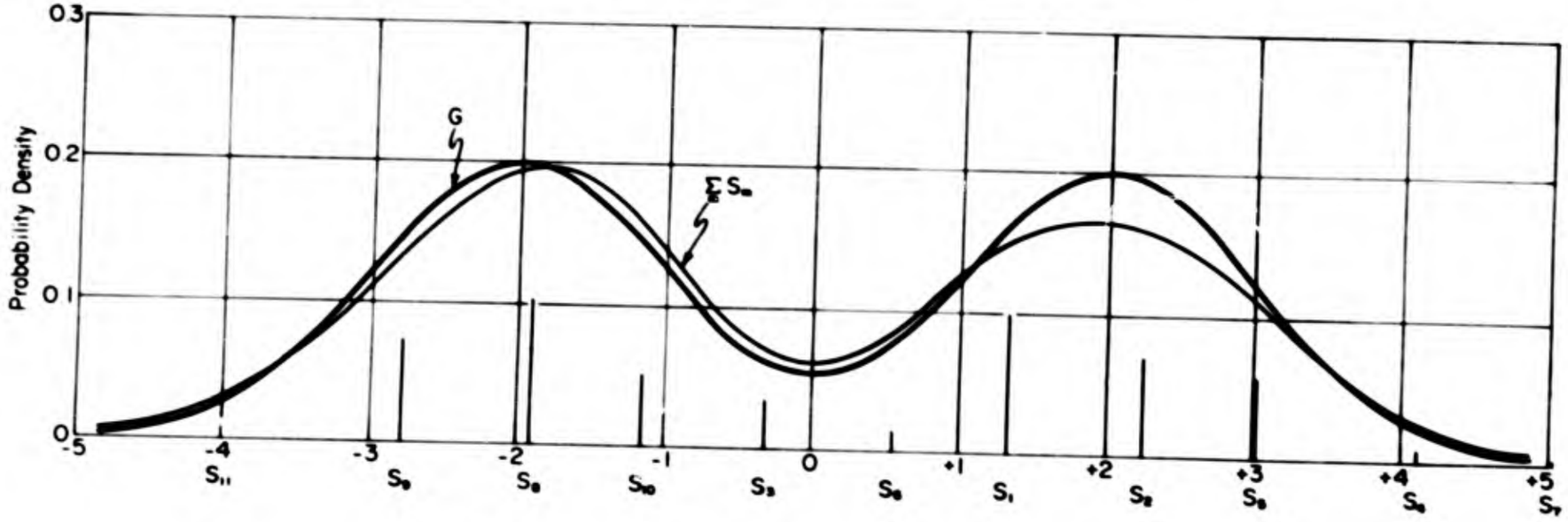


FIG. A-4. Approximation of G by ASSC II based on 300 samples of G

(Computer Run No. 4)

The fourth experiment was identical to the one just described with the exception that here members of the two Gaussian densities were introduced sequentially with samples drawn from the distribution of mean +2 introduced first. After 150 samples of this distribution were introduced, members of the second Gaussian process were introduced. In this manner, learning on bow submarine data followed by learning on, say, stern submarine data was simulated (as compared to mixing up the data before introduction to the program). No significant difference in performance was observed.

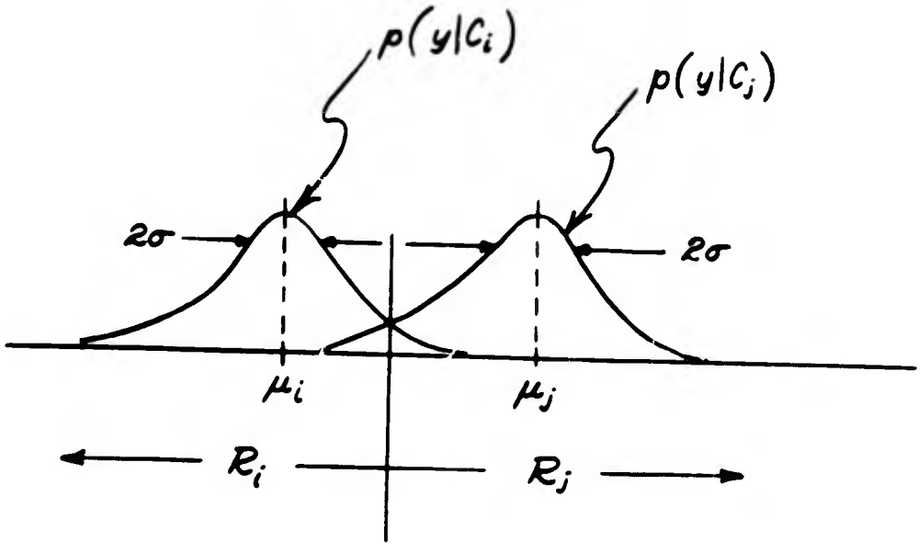
It is thus tentatively concluded that the order dependence of the learning programs is not as severe as originally expected. In fact, rather good agreement, more or less independent of the order in which samples were introduced was obtained throughout the four experiments.

APPENDIX II

THE RELATIONSHIP BETWEEN THE PROBABILITY OF CORRECT RECOGNITION AND CLASS CLUSTER TO SEPARATION RATIO (IN A SPECIAL CASE)

In this Appendix, we will show that the Expected Probability of Correct Recognition (EPCR) increases monotonically with the class separation to class cluster diameter ratio. This illustrates the intuitive results, at least in this example, that the optimum figure of merit (from a performance point of view) expresses desirable properties of the class distributions in the property space.

To illustrate this figure of merit (EPCR), consider a decision making system operating on a single variable y that may belong either to class C_i or to class C_j with equal a priori probability, each having a Gaussian density with variances σ^2 and means μ_i and μ_j respectively as shown in the figure below.



$$\text{EPCR} = \frac{1}{2}\int_{-\infty}^{\frac{1}{2}(\mu_i+\mu_j)} \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\tfrac{1}{2}\left(\frac{y-\mu_i}{\sigma}\right)^2\right]dy +$$

$$\tfrac{1}{2}\int_{\frac{1}{2}(\mu_i+\mu_j)}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\tfrac{1}{2}\left(\frac{y-\mu_j}{\sigma}\right)^2\right]dy$$

$$= \tfrac{1}{2}\left[\tfrac{1}{2} + \frac{1}{\sqrt{2\pi}\sigma}\int_0^{\frac{1}{2}(\mu_j-\mu_i)} \exp\left(-\tfrac{1}{2}\frac{y^2}{\sigma^2}\right)dy\right] +$$

$$\tfrac{1}{2}\left[\tfrac{1}{2} + \frac{1}{\sqrt{2\pi}\sigma}\int_0^{\frac{1}{2}(\mu_j-\mu_i)} \exp\left(-\tfrac{1}{2}\frac{y^2}{\sigma^2}\right)dy\right]$$

$$= \tfrac{1}{2} + \frac{1}{\sqrt{2\pi}\sigma}\int_0^{\frac{1}{2}(\mu_j-\mu_i)} \exp\left(-\tfrac{1}{2}\frac{y^2}{\sigma^2}\right)dy \qquad \text{(b-1)}$$

By a change of variables ($y = \sigma v$), EPCR can be written as shown:

$$\text{EPCR} = \tfrac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_0^{\frac{1}{2}(\mu_j-\mu_i)/\sigma} \exp\left(-\tfrac{1}{2}v^2\right)dv \qquad \text{(b-2)}$$

Since the integral is a monotonically increasing function of its upper limit, the figure of merit is a monotonically increasing function of $(\mu_j - \mu_i)/\sigma$ between $\frac{1}{2}$ and 1.

The usefulness of the property y increases as the ratio of the distance between means versus the standard deviation is increasing. A similar result can be obtained for N-dimensional Gaussian processes also. For Gaussian densities, therefore, the separation-to-spread ratio is a significant figure of merit.

BLANK PAGE

APPENDIX III

LINEAR TRANSFORMATIONS TO MINIMIZE ENTROPY IN PROPERTY SPACE

If samples of pattern class C_i of C-classes is multivariate normal in the N-dimensional measurement space x, and has an arbitrary covariance matrix U_i, find the set of K linear transformations ($K \leq N$) such that the sum of entropies of the C class distributions in the K-dimensional transformed y space is a minimum. This set of transformations minimizes the average entropy (or spread) of classes in y space and thus maximally clusters them (on the average).

Since the entropy of a class is a function only of the manner in which its members are distributed, obviously no complete representation of the classes in any different coordinate system (obtained by linear operations) can change the entropy of the class distributions. The only way to reduce the entropy is to select a manifold of the x space in which the spread of the classes is smaller. We will now derive the K optimum orthogonal directions in x space to be used as the coordinates of the Property Space.

We will first express the entropy of class C_i in the transformed space and then we will minimize the sum of class entropies in Property Space by finding the best K linear functions of x as the y-properties.

If x is an N-dimensional measurement vector (a column vector) and A is a K x N matrix expressing a linear transformation that expresses, as the coordinates of a vector y, the projections of x onto unit vectors pointing in the directions given by the rows of A, then the transformation y is given by Equation C-1.

$$y = A\,x \tag{C-1}$$

If class C_i has a multivariate normal distribution with mean μ_i and covariance matrix U_i, the distribution of C_i in the measurement space is given in Equation (C-2), and the distribution of y is given in Equation (C-3).

$$P_i(x) = \frac{1}{\sqrt{2\pi\,|\,U_i\,|}}\ \exp\left[\ -\tfrac{1}{2}(x-\mu_i)^T\ U_i^{-1}\ (x-\mu_i)\ \right] \tag{C-2}$$

$$P_i(y) = \frac{1}{(2\pi\,|\,A\,U_i\,A^T\,|\,)^{\frac{1}{2}}}\ \exp\left[-\tfrac{1}{2}(y-A\mu_i)^T\,[A\,U_i\,A^T]^{-1}(y-A\mu_i)\right] \tag{C-3}$$

The entropy of the density $p_i(y)$ is given in Equations (C-4) and (C-5) (obtained by substituting Equation (C-3) into (C-4)).

$$H_i(y) = -\int p_i(y)\ \log\ p_i(y)\ dy \tag{C-4}$$

$$H_i(y) = \tfrac{1}{2}\ \log|\,A\,U_i\,A^T\,|\ + \frac{K}{2}\ \log 2\pi e \tag{C-5}$$

The quantity we want to minimize is the sum of entropies of the classes in the y-coordinate system. The sum of entropies H(y) is expressed in Equation (C-6).

$$H_{(T)} = \sum_i H_i(y) = \tfrac{1}{2}\sum_i\ \log|\,A\,U_i\,A^T\,|\ + \frac{KC}{2}\ \log 2\pi e$$

$$= \tfrac{1}{2}\ \log\left(\prod_i |\,A\,U_i\,A^T\,|\right)\ + \frac{KC}{2}\ \log 2\pi e \tag{C-6}$$

But the product of determinants is the determinant of the product of the matrices. Using this relationship, we obtain Equation (C-7).

$$H(y) = \tfrac{1}{2} \log \left| \prod_i (AU_iA^T) \right| + \text{constant}$$

$$= \tfrac{1}{2} \log \left| AU_iA^T AU_2A^T AU_3A^T \ldots AU_cA^T \right| + \text{constant} \qquad \text{(C-7)}$$

If we let A be an orthogonal transformation (its K rows are orthonormal), $A^TA = I$, and H(y) can be written as in Equation (C-8).

$$H(y) = \tfrac{1}{2} \log \left| AU_1U_2 \ldots U_cA^T \right| + \text{constant}$$

$$= \tfrac{1}{2} \log \left| AWA^T \right| + \text{constant} \qquad \text{(C-8)}$$

where W is the product of covariance matrices of the different classes.

A minimization of H(y) can now be carried out to find the unknown transformation matrix A. It is known that the equality given in Equation (C-9) holds. *

$$d \log |B| = \text{tr}^{-1} dB \qquad \text{(C-9)}$$

Here B is a matrix and tr denotes the "trace" (sum of diagonal elements) of a matrix. Applying this to H(y), we get Equation (C-10) where we used the relationship given in Equation (C-11).

$$\frac{dH(y)}{d\,a_j} = \text{tr}\,(AWA^T)^{-1} \; W a_j^T = 0 \text{ for } j = 1, 2, \ldots, K \qquad \text{(C-10)}$$

$$\frac{d}{d\,a_j} a_j B a_j^T = 2 B a_j^T \qquad \text{(C-11)}$$

* Problem 10.3c on p. 207 in Information Theory and Statistics by Solomon Kullback (John Wiley and Sons, 1959).

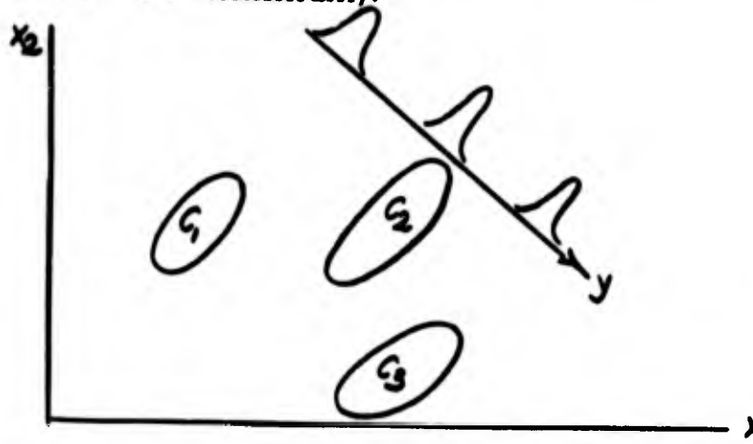
In Equation (C-10), a_j is the j^{th} row of the matrix A. We now impose the constraint on the minimization that $a_j I a_j^T = 1$, and, by using the method of Lagrange multipliers with the multiplier λ_j, we obtain Equation (C-12). Since $\text{tr}\,(AWA^T)^{-1}$ is a scalar, we can lump it with λ_j to obtain Equation (C-13).

$$\left| \text{tr}\,(AWA^T)^{-1} W - \lambda_j I \right| = 0 \quad \text{for } j = 1, 2, \ldots, K \tag{C-12}$$

$$\left| W - \lambda_j I \right| = 0 \quad \text{for } j = 1, 2, \ldots, K \tag{C-13}$$

It can now be shown that the K smallest eigenvalues of the matrix W will minimize the sum of entropies of the classes in the y-coordinate system, and that the K orthogonal directions (transformations of the measurement space) that will minimize the sum of entropies are given by the corresponding eigenvectors of W.

It is interesting to note that if the C covariance matrices are all equal, the W matrix is U^C and the eigenvectors of W are identical to the eigenvectors of U, while λ_j^C is the j^{th} eigenvalue of W (where λ_j is that of the covariance matrix U). This special case is easily interpreted geometrically in the figure below, which shows 3 bivariate normal densities with equal covariance matrices. The optimum linear transformation with K = 1 is the line y(along which the spread of all three classes is a minimum).

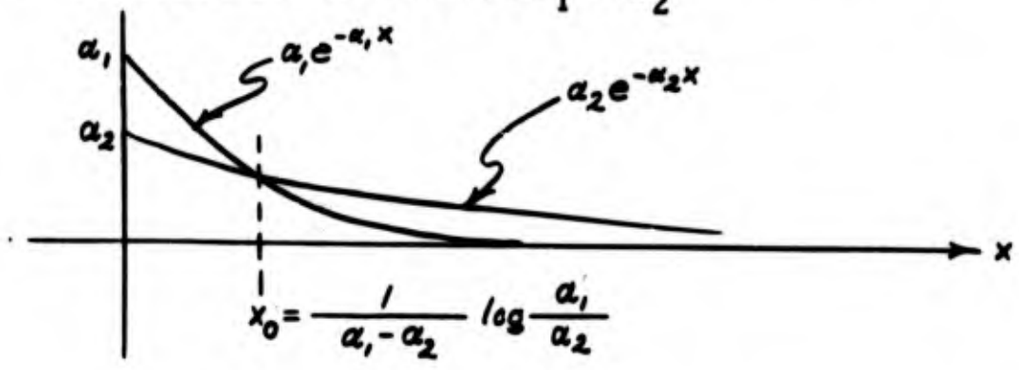


Minimizing Average Entropy for K =1, and $U_1 = U_2 = U_3$

APPENDIX IV

IF DIVERGENCE INCREASES, EXPECTED ERROR PROBABILITY DECREASES (FOR A SPECIFIC CASE)

Assume two exponential probability densities $p_1(x) = \alpha_1 e^{-\alpha_1 x}$ and $p_2(x) = \alpha_2 e^{-\alpha_2 x}$ between 0 and ∞. Assume $\alpha_1 > \alpha_2$



The divergence J(1, 2) is $\int_{\infty} [p_1(x) - p_2(x)] \log \frac{p_1(x)}{p_2(x)} \, dx$ (D-1)

and the Expected Error Probability (EEP) is given by Equation (D-2).

$$EEP = \frac{1}{2} \int_0^{x_0} p_2(x) \, dx + \frac{1}{2} \int_{x_0}^{\infty} p_1(x) \, dx \qquad \text{(D-2)}$$

These two are now evaluated as follows:

$$J(1, 2) = \int_0^{\infty} [\alpha_1 e^{-\alpha_1 x} - \alpha_2 (e)^{-\alpha_1 x}] (\log \alpha_1 - \alpha_1 x - \log \alpha_2 + \alpha_2 x) \, dx \qquad \text{(D-3)}$$

$$= \alpha_1(\alpha_2 - \alpha_1) \int_0^\infty e^{-\alpha_1 x} x\,dx + \alpha_2(\alpha_1 - \alpha_2) \int_0^\infty e^{-\alpha_2 x} x\,dx$$

$$= \frac{\alpha_2 - \alpha_1}{\alpha_1} e^{-\alpha_1 x}(-\alpha_1 x - 1)\Big]_0^\infty + \frac{\alpha_1 - \alpha_2}{\alpha_2} e^{-\alpha_2 x}(-\alpha_2 x - 1)\Big]_0^\infty$$

$$= \frac{\alpha_2 - \alpha_1}{\alpha_1} + \frac{\alpha_1 - \alpha_2}{\alpha_2}$$

$$\therefore J(1,2) = \frac{(\alpha_1 - \alpha_2)^2}{\alpha_1 \alpha_2} \qquad \text{(D-4)}$$

$$\text{EEP} = \frac{\alpha_2}{2} \int_0^{\frac{1}{\alpha_1 - \alpha_2}\log\frac{\alpha_1}{\alpha_2}} e^{-\alpha_2 x}\,dx + \frac{\alpha_1}{2} \int_{\frac{1}{\alpha_1 - \alpha_2}\log\frac{\alpha_1}{\alpha_2}}^\infty e^{-\alpha_1 x}\,dx$$

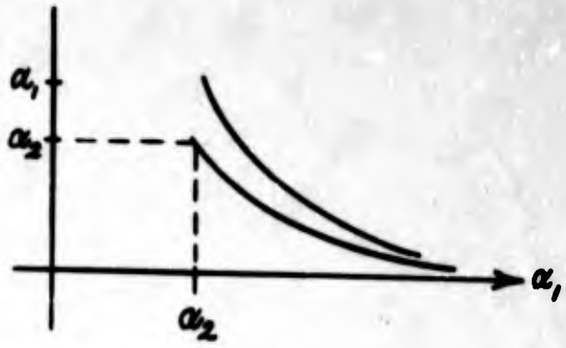
$$= -\tfrac{1}{2} e^{-\alpha_2 x}\Big]_0^{\frac{1}{\alpha_1 - \alpha_2}\log\frac{\alpha_1}{\alpha_2}} \quad -\tfrac{1}{2} e^{-\alpha_1 x}\Big]_{\frac{1}{\alpha_1 - \alpha_2}\log\frac{\alpha_1}{\alpha_2}}^\infty \qquad \text{(D-5)}$$

$$\text{EEP} = \tfrac{1}{2} + \tfrac{1}{2}\left[e^{-\frac{\alpha_1}{\alpha_1 - \alpha_2}\log\frac{\alpha_1}{\alpha_2}} - e^{-\frac{\alpha_2}{\alpha_1 - \alpha_2}\log\frac{\alpha_1}{\alpha_2}} \right] \qquad \text{(D-6)}$$

We thus obtained J(1, 2) in Equation (D-4) and EEP in Equation (D-6). Now differentiate J(1, 2) and EEP with respect to α_1 and show that the sign of the derivative of J(1, 2) and of EEP are opposite in the same regions of α_1. This proves that if $p_1(x)$ is perturbed so as to increase J(1, 2), the corresponding EEP decreases.

$$\frac{\partial J(1,2)}{\partial \alpha_1} = \frac{2\alpha_1\alpha_2(\alpha_1 - \alpha_2) - \alpha_2(\alpha_1 - \alpha_2)^2}{\alpha_1^2\,\alpha_2^2} = \frac{\alpha_1^2 - \alpha_2^2}{\alpha_1^2\,\alpha_2} \qquad \text{(D-7)}$$

which is positive if $\alpha_1 > \alpha_2$. The exponent of the second exponential in EEP as a function of α_1 for $\alpha_1 > \alpha_2$ is shown below. It decreases with increasing α_1. The exponent of the 1st exponential is just α_1/α_2 times that of the 2nd exponent. Hence, for $\alpha_1 > \alpha_2$, the 1st term in the bracket goes toward 0 faster than the 2nd term. But since the 1st term is always smaller than the 2nd, the derivative of EEP is negative. We have thus shown that derivatives of J(1, 2) and EEP have opposite signs, and thus that increasing divergence (in this case) implies lower error probability.



Negative of Value of Exponent

BLANK PAGE

Unclassified
Security Classification

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)*
Litton Industries
Information Sciences Laboratory
Waltham, Massachusetts

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
Pattern Recognition Research

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Final Report

5. AUTHOR(S) *(Last name, first name, initial)*
Sebestyen, George S.
Edie, Jay L.

6. REPORT DATE
14 June 1964

7a. TOTAL NO. OF PAGES
120

7b. NO. OF REFS
19

8a. CONTRACT OR GRANT NO.
AF 19(628)-1604

b. PROJECT NO.
5632

c. TASK
563205

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
DS-64-025

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
AFCRL-64-821

10. AVAILABILITY/LIMITATION NOTICES
U. S. Government Agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through U. S. Department of Commerce, Office of Technical Services, Washington, D. C. 20230.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Air Force Cambridge Research Labs.
Office of Aerospace Research
Hanscom Field, Bedford, Mass.

13. ABSTRACT

This report is concerned with the adaptive estimation of joint probability densities from a finite number of multidimensional vectors of known classification. An estimation procedure for the approximation of probability densities in the form of an n-dimensional histogram is described. The location and shape of the cells in the histogram are dependent on the data. The quality of the estimation procedure and its dependence on the order in which samples of known classification are introduced are described. Two quality measures are studied, one that estimates the probability that the decision is optimum and the other that the decision is correct. Techniques for analysis of data of unknown origin prior to the application of the adaptive pattern recognition is investigated and the mathematical and engineering problems are separated. Figures of merit to evaluate the usefulness of parameter sets are developed, and mathematical formulations of the parameter selection problem are given. (U) (Author)

DD FORM 1 JAN 64 1473

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Pattern Recognition **Automatic Learning Machines**						

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 context. The assignment of links, rules, and weights is optional.

BLANK PAGE