UNCLASSIFIED

AD 426426

DEFENSE DOCUMENTATION CENTER

FOR

SCIENTIFIC AND TECHNICAL INFORMATION

CAMERON STATION, ALEXANDRIA, VIRGINIA

UNCLASSIFIED

NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

426426

CATALOGED BY DDC
AS AD No.

AFCRL-63-548

PATTERN RECOGNITION RESEARCH

by

Jay Edie

William Floyd

George Sebestyen

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

Scientific Report No. 1

14 June 1963

DDC
JAN 10 1964
TISIA B

Prepared for

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

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 services or have their 'need-to-know' certified by the cognizant military agency of 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.

Scientific Report

PATTERN RECOGNITION RESEARCH

Contract AF19(628)-1604

14 JUNE 1963

Approved by

George Sebestyen
Technical Director

John W. Gerdes
Assistant Manager

Communication Sciences Laboratory
Data Systems Division
LITTON SYSTEMS, INC.

PREFACE

This report was prepared by J. Edie, W. Floyd, and G. Sebestyen of the Communication Sciences Laboratory, Data Systems Division, Litton Systems, Inc. In addition to those listed above, the significant contributions made by P. Connoly, V. Maglione, and H. O'Shea of the Computation and Analysis Group are gratefully acknowledged.

ABSTRACT

Machine Learning and Pattern Recognition is treated as the problem of adaptively constructing approximations to the joint probability densities of the N-variables with which members of classes are represented. The adaptive techniques studied construct approximations to the joint probability densities in the form of generalized N-dimensional histograms in which the locations, shapes and sizes of the histogram cells are generated by the known samples of the pattern classes. To economize on the number of cells constructed, a cell growth mechanism was devised to adapt the size and shape of the cells to best represent the probability densities. The accuracy of this method of representation was tested with the aid of a digital computer on large quantities of pattern samples of known probability distribution. The experimental results were compared with those that could be predicted theoretically. Quality criteria to assess the reliability of the decision rendered by a classification device and to influence the mechanism of machine learning were considered.

CONTENTS

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (Cont.)

1. INTRODUCTION

In today's technology, we are used to the idea that machines can be made to do most anything. We are not used to the idea, however, that machines could be made to learn to do tasks that we ourselves do not know explicitly how to perform. The last decade, but particularly the last few years, has seen the emergence of a number of fields of related activities. Artificial intelligence, pattern recognition and bionices have gaine prominence in the scientific literature. These fields of activity, in one way or another, deal with methods of examining input stimuli for the purpose of gaining information by eliminating certain redundancies inherent in the input. The information obtained is used to describe the input, to draw conclusions from it, or to perform other tasks normally considered to lie in the domain of human activity. Claims or achievements in these new fields of activity have persuaded us to look upon machines built by man as tools that no longer merely perform explicitly instructed tasks, but can serve as useful aids in the performance of normally human functions.

While this belief has been gaining in general acceptance, and the applicability of machines to perform more or less human functions is believed to be perhaps just around the corner, there has been a tendency to associate human qualities with machines and to extrapolate actual performance capabilities to implied capability no longer based on supportable facts. The purpose of this Introduction is, in part, to strip some of the mystery and vagueness from the pattern recognition field by outlining briefly the approximate present state of the art and by touching on the type of problems that present methods can solve. In the process of doing so, the important problems in pattern recognition will come to light and this will enable us to put in proper context the work reported in this Scientific Report on Pattern Recognition Research.

Systems that examine the physical world through a set of sensors and attempt to select a course of action or attempt to make a decision depending on what they see must be able to describe the world first. Systems that can provide such a description in a language of their own are not going to be possible in the foreseeable future. Another approach to describing the world and constructing its model is for the pattern recognition system to describe the world in terms prescribed for the system instead of using terms invented by the system.

This approach is analogous to preparing a questionnaire where each question in effect asks the system to test its environment in a certain manner and report its observation (the result of the test on the environment) by filling out the questionnaire. The question may be one that requires a numerical answer, or may be one that requires a binary yes or no answer. In either case, the questions stated by the questionnaire can be thought of as descriptors of the environment which can be treated as a set of parameters in a parametric description of the physical world. This notion leads to a vector representation of the environment which is simply another way of expressing a set of responses that the system would make to a questionnaire prepared for it in advance.

Two important questions arise in connection with the above notion of considering the model of the physical world for use in a pattern recognition system as a means for filling in a questionnaire. "What should the questions ask of the physical world?" and "What should we do with the answers in the questionnaire to obtain the information we are really seeking?" The first problem, that of formulating a set of questions to be answered, is the problem of selecting the parameters in a parametric description of the environment, while the second problem (that of determining how to process the answers or parameter values) is a question of data processing.

Regarding the question of data processing, one method is to predigest all the possible answers to the set of questions and determine, by some means, the conclusion that could be drawn from the responses in the questionnaire. This

is the type of approach used in decision-making systems that are of the "table lookup" or "truth table" variety. Here, in one manner or another, essentially all answers or possible combination of answers to the questions in the questionnaire have been preprocessed by the designer of the data processing system, and the conclusions have been preformulated and have been built into a machine. A second approach to the question of utilization of the data is to build into the machine not the answers, that is, the conclusions to be drawn from a set of answers, but instead to build into the machine the goal of the data processing system. In a decision making application, this might take the form of partitioning the large sets of possible answers into subsets, wherein each subset requires that the same conclusion be drawn from it. As an example, suppose that there are twenty different personality types that we wish to distinguish from one another. One of our tasks may be to decide on the basis of information to be gained from a questionnaire which of these personality types best describes a given candidate we interview. We can treat each personality type as a class and a specific combination of answers to the questionnaire generated by a single individual as a member of one of the classes. We are thus led to a problem where we wish to decide membership in classes. That is to say, we wish to decide that the particular combination of answers generated by the candidate in question is characteristic of one or another class of personality types.

Instead of listing all possible answer combinations that may occur during an interview and deciding ahead of time which personality type is the best fit to each answer combination, we may wish to construct a machine (or at least an automatic technique) to classify the interviewed candidates. In this case we must give the machine some information about the personality types, either by defining them or by giving it some examples of each. If a definition is given to the machine, it must merely apply the definitions to the particular set of answers generated by a candidate to determine which definition fits best. As a result, the machine operates in a manner where the specific operations it performs on the input to be classified are programmed in advance. This is of no interest in pattern recognition research.

If instead, we provide the machine with examples of each class and ask it to come up with its own definition of personality types, definitions which it may later apply to classify new candidates, we are dealing with the typical problem of concern in pattern recognition research. The pattern recognition system must abstract class definitions from samples of classes so that it should be able to recognize membership in classes at a later time. The pattern recognition system and its environment are illustrated in Figure 1.

It is possible, of course, that the requirements on data processing are other than those of decision making. It may be that the set of questions asked by the questionnaire in a set of responses that were received do not allow for conclusive determination of the type of personality. It may then be necessary to initiate new action, perhaps in the form of asking additional questions. What kind of questions should we ask, and how can such new questions be formulated automatically from the examination of the responses that appeared on the preceding questionnaire?

A problem of a different type that may occur is that the number of questions that appear on the questionnaire are so large that it would be next to impossible to try to utilize all answers to render a decision. We may then be forced to select a subset of the answers and base our decision on the subset. The question then occurs, how should we select the subset of the answers on which we should base our decision?

Suppose that we have formed a tentative conclusion from a subset of questions already examined. How should we firm up and increase the certainty of our conclusion by asking for additional information; that is, which of the remaining answers shall we seek out as a basis for a more accurate and certain decision?

Suppose that we are accustomed to making decisions based on questionnaires that are fully completed. What shall we do with, and how should we base a decision on incomplete questionnairs, i.e., questionnaires that have not been filled in completely by the candidate?

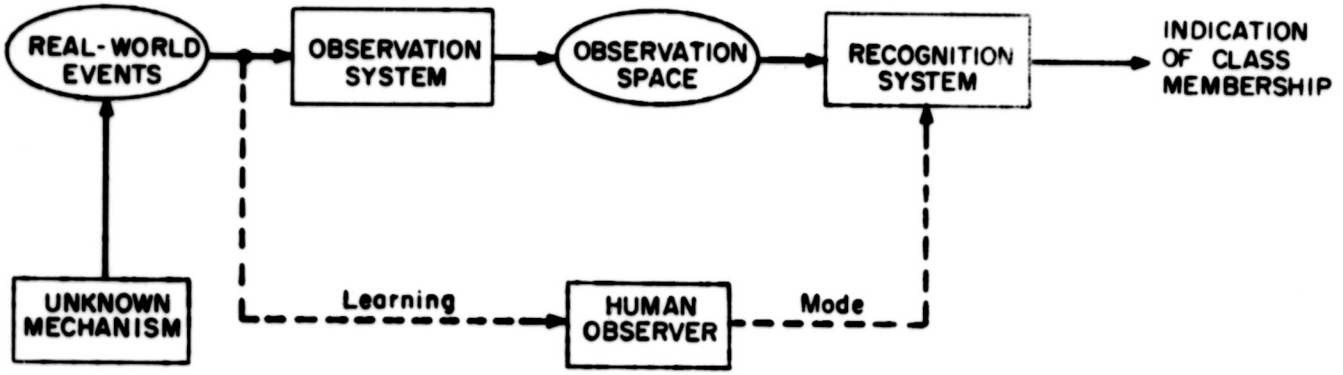


Fig. 1. The Classification Problem Model . . . . . .

How shall we evaluate trends? It may be that not enough information is available to make a decision, but we can assess the trend (perhaps in the form of measuring the derivative with respect to time of the probability that a specific decision is correct).

Suppose that we send out several interviewers, each with the same questionnaire to be filled in, and they return with answers in their questionnaires that are not identical. How shall we evaluate the set of questionnaires all answering the same questions asked of the same environment? Which one should we believe, or how should we process them to arrive at the best possible decision?

These are but a few of the types of problems that can be encountered. Problems of this type can be encountered at all levels, in speech recognition, character recognition, and even in business.

Artificial intelligence, bionics, pattern recognition, all intend to design systems that can eventually answer qeustions of this type. What has actually been achieved to date? For the most part we can say that the automation of decision-making problems is reaching the point where reliable and sophisticated decisions can be made on already parametrically represented inputs. Furthermore, the automation of the construction of the decision-making procedure has reached the point that useful devices can be constructed that have a built-in goal, a goal of constructing decision-making procedures from known samples of the decision classes. The subject area of pattern recognition deals with techniques of this type. Progress in the automatic formulation of the questions to be asked of the physical word, however, has not reached a satisfactory state, as yet; but this subject is currently under exploration by an increasing number of researchers, and it can be expected that the mathematically treatable segments of the parameter selection problem will reach a fairly satisfactory state within 1 or 2 years. Unfortunately, the same cannot be said of the nonmathematical aspect of the problem of parameter selection. More will be said about this later.

The research reported on in this Scientific Report is concerned, almost exclusively, with the problems that arise once the parametric representation of the environment has been decided upon and once the class samples, from which the pattern recognition system must derive its class definitions, have been designated. To put this year's work in proper perspective, a brief review of the chronological developments is in order.

Up to the end of 1961 on Contract AF19(604)-8024 and other related work the pattern recognition problem was formulated as a problem of vector representation of the environment followed by the automatic partitioning of the vector space into regions from known members of the class populations. At first, partitions were created by considering various transformations of the vector space that would tend to cluster known members of the same class while they would tend to separate known members of different classes. Depending on the classes of transformations that were considered, the resulting partitions of the vector space were by means of hyperplanes or by means of the intersections of quadratic forms. The former (those that use hyperplanes) have been referred to as linear methods and discriminant methods. These methods, in a different context and with different motivations and derivations, have been employed by anthropologists earlier. The rediscovery and application of these methods, however, has occurred in this time period. During this time period (prior to the end of 1961) the relationship between the clustering ideas (using linear transformations, and quadratic forms operating on linear transformations) and statistical decision theory was shown. In the same time period the idea of using transformations that cluster members of the same class while they separate members of different classes was extended to the use of nonlinear transformations as well. It was shown that decisions that are optimum in a decision theoretic sense and optimum with regard to clustering notions are one and the same, and procedures for constructing such decision making systems have been devised. These methods have been programmed, tested, and found working correctly. In the process of demonstrating the method of operation of these

nonlinear techniques, the computational difficulties were explored and the conclusion was reached that one has to pay a high price in computer time and complexity for strict optimality.

It was for this reason that approximate techniques and adaptive techniques were developed that did not only behave in an adaptive manner but were designed specifically to approximate the optimum nonlinear techniques with which excellent results can be obtained (at a high cost). That is to say, the adaptive and approximate techniques developed on Contract AF19(604)-8024 attempted the approximation of the optimum nonlinear techniques. At the same time adaptive techniques developed elsewhere and intended to be optimum were still restricted to operations no more general than those performed by the linear and discriminant techniques. Toward the latter part of 1961 the first generalized adaptive techniques were tested successfully. The Adaptive Sample Set Construction (ASSC) techniques approximate the joint probability density of the parameters for each of the sample populations and base statistically optimum decisions on the approximate probability densities. They can also make decisions on incomplete parameter sets and on repeated observations of the same input. Certain peripheral problems such as recognition in noise were also considered.

During the research program reported in this Scientific Report, the adaptive methods of approximating optimum decisions and constructing optimum decision making systems have been refined to such an extent that they now process parametric data in a nearly optimum manner. Great care has been taken to develop the techniques along lines that retain the simple program and the simple hardward realization aspects of the early ASSC techniques while increasing the generality of the problems they can solve successfully. A hierarchy of methods based on the early ASSC techniques have been developed and the estimation of the joint probability densities of the parameters for an arbitrary distribution of class populations has been studied. The reliability of the estimation procedure and the trustworthiness of a decision can be calculated. We thus have developed a figure of merit with which the quality of

the pattern recognition system can be measured. It would now seem to be an obvious step to attempt to feed back the figure of merit in order to improve the classification system.

One of the directions in which current efforts are concentrated attempt to utilize the figure of merit (and other diagnostic observations) to improve the classification system; that is to say, to improve the method of estimating probability densities. It is believed that the work of the last year has brought us much closer to the realization of nearly optimum decision making systems. It now remains to couple these systems with automated diagnostic examination of the data and automated selection of all of the system variables to realize a completely automated machine learning system.

Once the system is automatic, the human observer loses insight of the detailed processes that occur and can no longer interpret the data. For this reason a study must be made of the human requirements when using automatic systems. We must decide what to output to the human so that he will not feel "left out" and will profit from the machine's experience with the problem being solved.

Work on the automatic analysis of the chosen parametric representation has also been undertaken. This work is not reported in this Scientific Report. The problem of parameter selection can be shown to consist of a facet that can never be attacked with <u>any</u> kind of mathematical, numerical, algorithmic or (for that matter) any systematic procedure. The second facet of the parameter selection problem (that which can be attacked systematically) can also be stated and solved mathematically. During the coming period of time, these problem statements will be worked out and easily implemented approximations of the solutions will be devised to permit not only the automatic processing of parametric data, but also the automatic modification of the choice of parameters to achieve the overall desired result of realizing improved classification systems within the allowable system constraints.

In conclusion a very brief survey of the last few years of pattern recognition developments will be given.*

The pattern recognition problem can be conveniently divided into two parts — first, selection of the measurements, of the space to be partitioned, and second, finding the method of partitioning this space. Generally speaking, the problem of finding a set of parameters and thus defining the vector space is an engineering problem, while the method of finding a partition of the vector space once the parameters are defined is a problem that can be treated mathematically in any one of a large number of different ways. We will thus refer to the two parts of the problem as the parameter selection problem and the problem of partitioning the measurement space. In addition to these two main problems, there are a number of important practical considerations that have received theoretical treatment. Among these must be counted the method of decision-making when not all observable and measurable parameters are available at a given time, decision-making when multiple observations of the same stimulus can be obtained, recognition in noise, recognition when the measurements are known to have been taken at the time more than one input stimulus class was present (multiple target problem), and the subdivision of the finite number of samples into "learning sample" and "testing sample" subsets.

In theoretical advances, the trend has been toward the increased application of statistical methods. A few years ago, besides exact match techniques, correlation with stored references was about the most sophisticated technique employed. This is not too surprising for correlation has just recently become a household word and has just recently reached wide-spread laboratory use

*The following is based on a portion of "Pattern Recognition and Machine Learning" by N. Abramson, D. Braverman, and G. Sebestyen.

despite the fact that the peak of the flood of correlation literature was reached eight or ten years ago. In addition to this fact, the mathematically more sophisticated scientific community was somewhat repulsed by the unquantitative and unscientific outward appearances of pattern recognition problems. This is often the case when the problem we wish to solve cannot be stated clearly.

Since a few years ago the formulation of pattern recognition problems as problems in hypothesis testing on multivariable inputs has gained widespread acceptance. Inputs have come to be thought of as vectors in a multidimensional space or member functions of a random process.

The vector representation has caused many authors to view the pattern recognition problem as composed of two subsidiary problems — first, selection of the measurements, or the space to be partitioned, and second, finding a method of partitioning the space. Marill and Green[1] have stressed this point, calling that part of the system dealing with the first problem the "receptor" and that part dealing with the second problem the "categorizer." Most of the progress in the theory of pattern recognition during the past three years has been in determining methods of partitioning the measurement space, once the measurements have been specified.

Probably the most investigated method of partitioning the measurement space is with hyperplanes.[2, 3, 4] A number of authors have pointed out that the linear discriminant methods, correlation methods and matched filter methods are, in fact, but different realizations of identical operations. Highleyman[5, 6] has presented an exhaustive treatment of linear discriminant techniques and their uses; among other topics, he has discussed general properties of hyperplane partitions and heuristic methods of constructing separating hyperplanes from empirical data. Linear methods, although they have been studied most intensively, are not the only methods used to partition the measurement space in pattern recognition. Polynomial discriminants,[4] hyperspheres[7] and quadratic forms[8] have also been investigated in some detail.

Much work has been done in determining optimum partitions of the measurement space when the measurements from each class are Gaussian.[1, 2, 9, 10] In general, the Gaussian assumption leads to linear and quadratic data processing;[3] in certain important special cases, the quadratic terms need not be included, and the Gaussian assumption leads to the linear methods discussed above. It has been shown[4, 11, 12, 13] that geometrical distortions (transformations) of the vector space to increase the separability of vectors belonging to different classes lead to statistically optimum (Bayes) decisions under quite general assumptions about the class probability densities. Thus, the restrictive Gaussian assumption can be removed.

Several studies have been concerned primarily with the use of learning observations in pattern recognition. Two different sets of assumptions about the nature of the learning observations are possible. We may assume that each learning observation is presented to the machine together with a label specifying the class of patterns to which the observation should be assigned or we may assume that each observation is presented without such a label. The first problem is called "learning with a teacher," the second, "learning without a teacher." Learning without a teacher is undoubtedly the harder of the two problems. Learning without a teacher has been investigated by Daly,[14] who has been able to prove certain convergence properties when there are only two classes to be recognized. The data processing necessary for an optimum solution, however, grows exponentially, so that Daly[15] has also investigated a suboptimum linear system for learning without a teacher. Jakowatz, Shuey and White[16] have investigated learning without a teacher in another form. They have analyzed and built a system for detecting fixed pulses occurring at random intervals in noise. Initially, the system does not know the shape of the pulse it is to detect, but as it observes the sum of the pulses and noise, it "learns" the pulse shape and "evolves" into a matched filter. Some analysis related to this system has been done by Roe and White[17] and by Hinich,[18] but the theory is by no means complete. A similar system has been treated by

0426426

0426426

AF FORM 993A 1 AUG 54

GENERAL PURPOSE SUMMARY CARD

CATALOGED BY DDC
AS AD No.

A B C D E 1 2 3 4 5 6 7 8 9

414788

Glaser.[19] Glaser's system consists of a linear weighting of an energy detector and a matched filter detector; as the pulse form is learned, the weight assigned to the energy detector output is decreased in favor of the matched filter output.

The theoretical problems involved in learning with a teacher have also received attention. Braverman[3, 9] has derived optimum data processing for learning the mean of Gaussian patterns. In addition, he has shown that the difference between the error probability of his filters and the error probability after learning is inversely proportional to the number of learning samples. The concept of a reproducing density is used by Abramson and Braverman[20] to obtain an iterative solution to the problem of optimal learning of the mean of Gaussian patterns. Keehn[21] has used the idea of the reproducing density to derive optimum methods of learning the covariance matrix of Gaussian patterns. When little a priori information is available, Keehn's solution reduces to that of Sebestyen's clustering transformation.[11]

In a simple case where the measurement space has been reduced to a single dimension, and a threshold decision rule is optimum, Kac[22] has investigated an adaptive method of adjusting the threshold. The threshold is given a small increment after each incorrect decision; this simple procedure is shown to bring the threshold to the point of equal errors of the first and second kind.

Those researchers who had character recognition in mind or those who were dedicated to a digital approach by inclination have restricted themselves to binary measurement spaces. When the measurement space is binary and the number of classes to be recognized is just two, the pattern recognition problem becomes one of selecting a truth function of n variables. Almost all of the work done in this area has restricted the set of allowable functions to the linearly separable truth functions. This has been done both because of the relative ease with which search procedures can be described in this restricted class, and because of the ease of instrumenting these procedures. Widrow and

Hoff[23] have described an iterative adaption procedure for searching the class of linearly separable functions which is equivalent to mean square error minimization. The properties of linearly separable truth functions have also been treated by Gableman,[24] Stearns[25] and Highleyman.[26] Coates, Kirchner and Lewis[27] have provided a simplified method of realizing and synthesizing linearly separable truth functions. Ridgway[28] has looked into the synthesis of truth functions by using a parallel combination of linearly separable functions followed by a majority logic circuit, and has investigated an adaption procedure useful in such a system. A novel method of analyzing linearly separable functions has been used by Hoff.[29] Linearly separable functions — mappings of the binary n cube into two points — are approximated by mappings of the (non-binary) n sphere into two points. Hoff also discusses functions of this type which are invariant under certain transformations of the binary input space.

The primary unsolved question in pattern recognition is undoubtedly that of selecting the measurement space. This problem is analagous to the classical statistical problem of design of experiments and suffers from many of the mathematical difficulties of that field. Ball[30] has used integral geometry to suggest measurements for the character recognition problem which are invariant with respect to rotations, translations and scale changes. He has also attempted to use some simple results in comparison of statistical experiments to evaluate the usefulness of his measurements. Another study dealing with the selection of measurements is provided by Lewis.[31] Lewis examines the notion of a single number statistic for each coordinate of the measurement space which would have certain desirable properties related to the "goodness" of the coordinate. He shows that, in general, no such statistic exists, but he does show an alternate statistic with some merit to recommend it. It is clear that a good deal of work will be done in the area of measurement selection in the years to come. It is one of the principle areas where our own work is also concentrated.

REFERENCES FOR INTRODUCTION

1. Marill, T., and Green, D.M., "Statistical Recognition Functions and the Design of Pattern Recognizers," IRE Trans. on Elect. Computers, Vol. EC-9, No. 4, p. 472, December 1960.

2. Cooper, P.W., "Classification by Statistical Methods," Melpar Technical Note 61/2, April 1961.

3. Andrew, A.M., "An Experimental Comparison of Some Algorithms for Self-Organizing Systems," IRE Trans. on Info. Th., Vol. IT-8, No. 5, pp. 163-168, September 1962.

4. Sebestyen, G.S., "Decision Making Processes in Pattern Recognition," Macmillan Co., New York, 1962.

5. Highleyman, W.H., "The Design and Analysis of Pattern Recognition Experiments," The Bell System Technical Journal, Vol. XLI, No. 2, p. 723, March 1962.

6. Highleyman, W.H., "Linear Decision Functions with Application to Pattern Recognition," Proc. IRE, pp. 1501-1514, June 1962.

7. Cooper, P.W., "The Hyper-Sphere in Pattern Recognition," Melpar Technical Note 62/1, February 1962.

8. Cooper, P.W., "Statistical Pattern Recognition with Quadratic Forms," Melpar Technical Note 62/4, June 1962.

9. Braverman, D.J., "Machine Learning and Automatic Pattern Recognition," Stanford Electronics Laboratories Technical Report No. 2003-1, February 17, 1961.

10. Smith, J.E., and Klem, L., "Vowel Recognition Using a Multiple Discriminant Function," JASA, Vol. 33, No. 3, p. 358, March 1961.

11. Sebestyen, G.S., "Recognition of Membership in Classes," IRE Trans. on Info. Th., Vol. IT-7, No. 1, p. 44, January 1961.

12. Sebestyen, G.S., and Hartley, A.K., "Study Program of Pattern Recognition Research," AFCRL 62-65, December 1961.

13. Sebestyen, G.S., "Classification Decisions in Pattern Recognition," Technical Report 381, Res. Lab. of Electronics, MIT, April 1960.

14. Daly, R.F., "Adaptive Binary Detectors," Stanford Electronics Laboratories Technical Report No. 2003-2, June 26, 1961.

15. Daly, R.F., "The Adaptive Binary-Detection Problem of the Real Line," Stanford Electronics Laboratories Technical Report No. 2003-3, February 1962.

16. Jakowatz, C.V., Shuey, R.L. and White, G.M., "Adaptive Waveform Recognition," Proc. of the Fourth London Symposium on Information Theory, Colin Cherry, editor, Butterworths, 1961.

17. White, G.M., and Roe, G.M., "Probability Density Function for Correlators with Noisy Reference Signals," IRE Trans. on Info. Th., Vol. IT-7, pp. 13-19, January 1961.

18. Hinich, M.J., "A Model for a Self-Adapting Filter," Information and Control, Vol. 5, No. 3, Sept. 1962, pp. 185-203.

19. Glaser, E., "Signal Detection by Adaptive Filters," IRE Trans. PGIT, Vol. IT-7, No. 2, pp. 87-98, April 1961.

20. Abramson, N. and Braverman, D., "Learning to Recognize Patterns in a Random Environment," IRE Trans. on Info. Th., Vol. IT-8, No. 5, pp. 58-63, September 1962.

21. Keehn, D.G., "Learning the Mean Vector and Covariance Matrix of Gaussian Signals in Pattern Recognition," Stanford Electronics Laboratories Technical Report No. 2003-6, February 1963.

22. Kac, M., "A Note on Learning Signal Detection," Special Issue on Sensory Information Processing, Trans. IRE PGIT, Vol. IT-8, No. 2, February 1962, pp. 126-128.

23. Widrow, B. and Hoff, M.E., Jr., "Adaptive Switching Circuits," 1960 WESCON Convention Record, Pt. IV, pp. 96-104, August 23, 1960.

24. Gableman, I.J., "The Synthesis of Boolean Functions Using a Single Threshold Element," IRE Trans. on Elect. Computers, Vol. EC-11, No. 5, p. 639, October 1962.

25. Stearns, S.D., "A Method for the Design of Pattern Recognition Logic," IRE Trans. on Elect. Computers, Vol. EC-9, No. 1, March 1960.

26. Highleyman, W.H., "A Note on Linear Separation," IRE Trans. on Elect. Computers, Vol. EC-10, No. 4, p. 777, Dec. 1961.

27. Coates, G.L., Kirchner, R.B., and Lewis, P.M., "A Simplified Procedure for the Realization of Linearly-Separable Switching Functions," IRE Trans. On Elect. Computers, Vol. EC-11, No. 4, p. 447, August 1962.

28. Ridgway, W.C., III, "An Adaptive Logic System with Generalizing Properties," Technical Report No. 1556-1, Stanford Electronic Laboratories, April 1962.

29. Hoff, M.E., Jr., "Learning Phenomena in Networks of Adaptive Switching Circuits," Stanford Electronics Laboratories Technical Report No. 1554-1, July 1962.

30. Ball, G.H., "The Application of Internal Geometry to Machine Recognition of Visual Patterns," WESCON, 1962. (6.3)

31. Lewis, P.M., "Characteristic Selection Problem in Recognition Systems," IRE Trans. on Info. Th., Vol. IT-i, No. 2, p. 171, February 1962.

2. CLASSIFICATION BY LIKELIHOOD FUNCTION ESTIMATION

Consider the problem of deciding which of M classes has given rise to an observed event, $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$, and suppose that the statistics of events and classes are known, i.e., the joint probability density function of $\underset{\sim}{x}$ and m is known, where m denotes the class label: m = 1, 2, . . . , M. The decision-theoretical optimum method of processing a measured event $\underset{\sim}{x}$ to render the classification is well known. Specifically, $\underset{\sim}{x}$ should be regarded as a member of the k-th class (Eq. (18), Reference 1) if the cost of deciding in favor of the k-th class is less than that of deciding in favor of any of the other classes. This is stated in Eq. (2.1).

$$\sum_{m=1}^{M} P_m p_m(\underset{\sim}{x}) \left[C_k^{(m)} - C_j^{(m)} \right] \leq 0 \text{ for all } j \neq k, \; j = 1, 2, \ldots, M, \qquad (2.1)$$

where

$C_j^{(m)} \equiv$ the cost (i.e., loss) associated with deciding that $\underset{\sim}{x}$ belongs to the j-th class when in fact $\underset{\sim}{x}$ belongs to the m-th class,

$P_m \equiv$ the a priori probability that an event from class m will occur,

and $p_m(\underset{\sim}{x}) \equiv$ the conditional probability density function of $\underset{\sim}{x}$, given that $\underset{\sim}{x}$ belongs to the m-th class.

This method of decision-making minimizes the average risk[2] associated with the classifications.* If, as is appropriate with many practical classification

*Evidently the basic form of the procedure is the same for other optimization criteria. See, for instance [6] for the binary decision case.

problems, the cost or loss is the same for all misclassifications, then Eq. (2.1) reduces to the following decision rule: decide $\underset{\sim}{x}$ is a member of the k-th class if

$$P_k p_k(\underset{\sim}{x}) \geq P_j p_j(\underset{\sim}{x}) \text{ for all } j \neq k,\ j = 1,\ 2, \ldots,\ M \tag{2.2}$$

Further, if the a priori probabilities are the same for all classes ($P_m = 1/M$ for all m), then Eq. (2.2) becomes: decide $\underset{\sim}{x}$ is a member of the k-th class if

$$L_{\underset{\sim}{x}}(k) \geq L_{\underset{\sim}{x}}(j) \text{ for all } j \neq k,\ j = 1,\ 2, \ldots,\ M, \tag{2.3}$$

where $L_{\underset{\sim}{x}}(m) \equiv p_m(\underset{\sim}{x})$ is commonly called the likelihood function of m given the event, $\underset{\sim}{x}$. When class a priori probabilities are the same, the likelihood function is equal to the a posteriori probability of class occurrence; i.e., $L_{\underset{\sim}{x}}(m) \equiv p_m(\underset{\sim}{x}) = p_{\underset{\sim}{x}}(m)$.

Thus, we see that if the statistics of events and classes are known, then an optimum (from the standpoint of minimizing risk) method of establishing classification decision boundaries in observation space is known, and the only hurdle which remains is implementation of this procedure. Unfortunately, however, this result can only be used as a guide to solving any practical classification problem, because the statistics of events and classes are usually not known precisely.

In most practical problems, all the information available on the statistics of events is contained in the values of a finite number, N, of the sample events processed in the learning mode of operation of a recognition system. A reasonable way to proceed in this situation is to generate an estimate of the likelihood function (or equivalently, the probability density function) of the different classes, over the observation space, and render classification decisions in the manner

dictated by decision theory using the estimated quantity in lieu of the "true" function. This is the basis for all of the classification methods discussed in later sections of this report.

With this approach to establishing classification decision boundaries in observation space, the method of estimating probability density functions plays the key role. The degree to which the estimate corresponds to the true function determines the similarity between the decision boundaries actually utilized and those which would minimize the misclassification probability. In addition, and perhaps equally important for advancing the development of automatic recognition systems, the form of the estimate should be selected to minimize the equipment complexity (primarily the storage requirements and operating speeds) associated with its implementation.

If the probability density function is known, an approximation of the function in a form economical from the point of view of the storage requirements can be obtained in the form of a histogram. If we are willing to devote a given number of storage locations to the storage of an approximate probability density, we would like to assure that the approximation we store is the "best" under the given constraints imposed by our storage limitations.

As an illustrative example of how imposed storage limitations on the construction of an optimum histogram approximation of a probability density can be utilized to derive the optimum histogram, the following problem was considered.

Given a probability density $p_m(\underset{\sim}{x})$ we would like to construct a histogram $h_m(\underset{\sim}{x})$ such that the mean-squared error between $p_m(\underset{\sim}{x})$ and $h_m(\underset{\sim}{x})$ is minimized. We impose the constraint that the histogram should be composed of exactly K bars but neither the location, widths, nor heights of the bars are in any way constrained.

The above problem is solved in Appendix I, where a set of simultaneous equations are derived that relate the locations of the left and right hand boundaries of the histogram bars in terms of the function $p_m(\underset{\sim}{x})$ to be approximated.

The solution of the simultaneous equations could be obtained only under rather limited conditions by an iterative method of numerical solution. This solution was programed on a computer and the solution was demonstrated on a number of examples given in Appendix I.

In most practical cases in pattern recognition problems the probability density to be approximated and stored is not known. Only a few sample observations are available from which the density must be estimated. Although there are many methods of estimating probability density functions,* two approaches to the problem stand out as most suitable for consideration in a recognition system. The first consists of estimation through histogram construction by counting the number of occurrences of events in pre-specified regions (cells) in the observation space. Such an estimate is illustrated in Figure 2(a) for a one-dimensional observation space, and N = 20 samples in the learning set of data. The area of each vertical bar is an estimate of the probability that $\underset{\sim}{x}$ will occur within the range of values defined by the boundaries of the (in this case, one-dimensional) cell. This probability estimate for any cell is provided by the ratio of the number of learning samples (successes) which occur within the cell, N_j, to the total number of learning samples, N. In general, the probability density function $p(\underset{\sim}{x})$, of a multidimensional random variable, $\underset{\sim}{x}$, is assumed to be constant over the cell, and equal to the ratio of the estimated probability of obtaining a sample within the cell to the hypervolume of the cell. In symbols,

$$p(\underset{\sim}{x}) = \frac{N_j}{N} \cdot \frac{1}{V_j} \tag{2.4}$$

where

$N_j \equiv$ the number of learning samples which occur within the j-th cell,

and $V_j \equiv$ the volume of the j-th cell.

*See, for instance, Reference [3] for descriptions of several methods.

The caret symbol indicates that an estimate, rather than a true probability density function is obtained.

Straightforward application of this method of estimation requires a priori specification of the cell structure (size, shape and number in observation space) over which the histogram is to be constructed. To reduce storage requirements it would be desirable to keep the number of cells small. However, to represent the probability density function accurately in regions where this function is sensitive to small changes in $\underset{\sim}{x}$, the cells should be small, which would make the number of cells large. A third factor which must be considered in selecting a cell structure is that the accuracy of estimation of the probability that $\underset{\sim}{x}$ will occur in a given cell, is proportional to the number of learning samples which occur within the cell.[3] Thus, the minimum resolution which should be attempted with a cell structure is limited not only by the (unknown) character of the true probability density function, but also by size of the learning set.

Since the character of the (unknown) probability density function plays such an important role in determining the appropriateness of a given cell structure, it is reasonable to utilize the only information available on this function (the learning set) to select the cell structure. This could be accomplished during the learning mode of operation of a recognition system by adjusting the cell structure (according to a pre-established criterion) with each exposure of the system to a new learning sample. There are many ways of implementing this procedure. One would be to start with a coarse cell structure consisting of a few hypercubes, and then increase the cell structure resolution by subdividing existing cells to avoid a violation of the constraint that no more than 4 (say) samples should be allowed in a single cell. This method of adjusting cell structure is illustrated in Figure 2(b) for the same 25 one-dimensional samples used to construct the (uniform cell structure) histogram in Figure 2(a). Even though the modified cell structure involves two less cells than the uniform structure, considerably greater resolution is attained with the modified structure. If the range of possible values of the observation variable (x) is partitioned

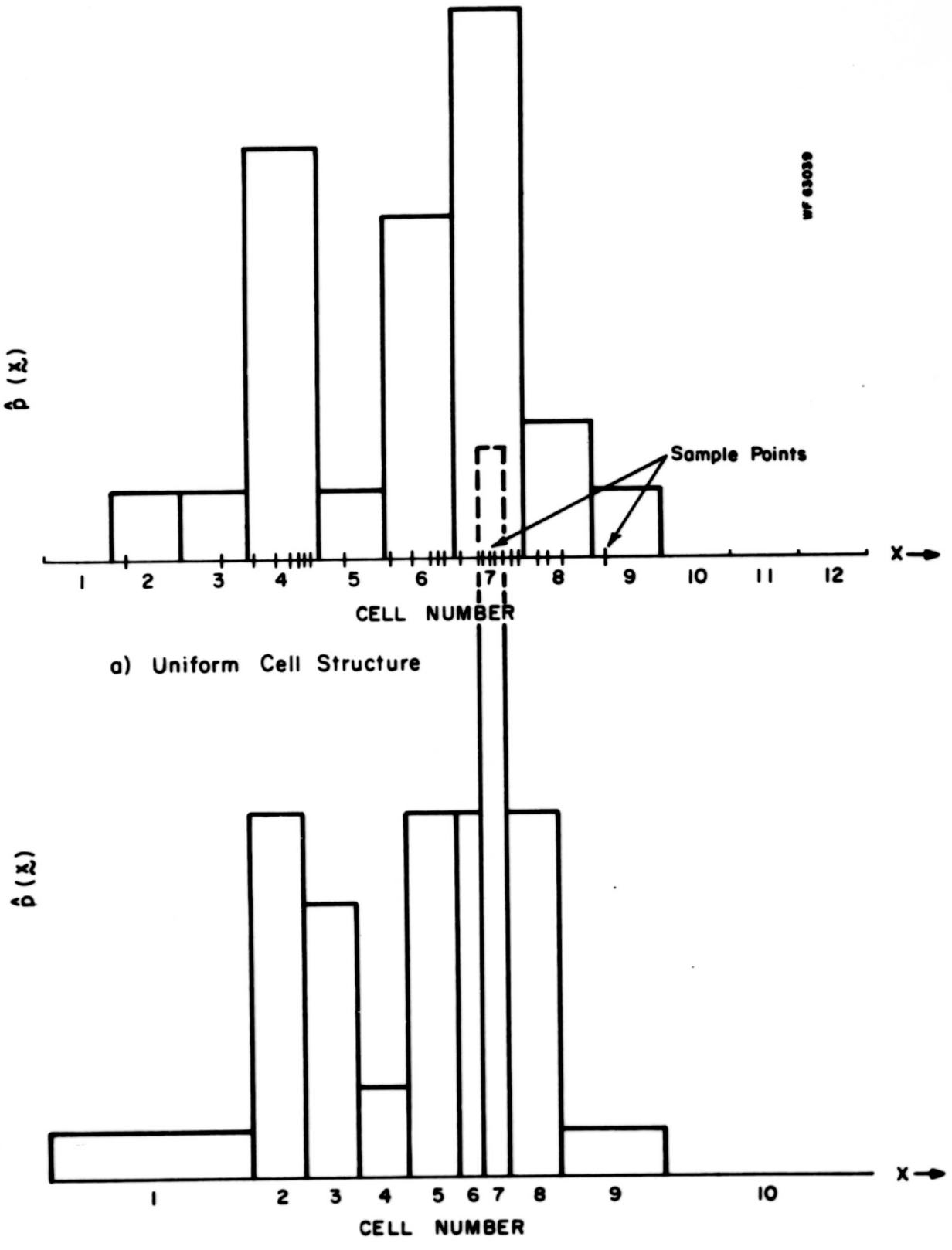


Fig. 2. Histogram Estimates of the Probability Density Function of a One-Dimensional Random Variable

into segments corresponding to unchanging values of the probability density function, then both the uniform cell structure and the adapted structure would require 9 quantities to represent their corresponding histograms, although the adapted structure attains a higher resolution.

Of course, the accuracy of an estimation based on rules for adapting the histogram cell structure to the learning samples must be evaluated before the utility of such a procedure can be ascertained. The purpose of this histogram illustration is to point out the possibility of using an adaptive procedure for estimating probability density functions (and therefore, decision boundaries). The significant difference between this approach to adaptation of decision boundaries in observation space and most of the other methods which have been proposed in the past few years is that this approach makes a conscious attempt to approximate the class probability densities without any prior assumptions about the distribution of events in the observation space. Having estimated the densities, the procedure known to be "optimum" is used to construct the decision boundaries. While constraints on the number and type of boundaries which can thus be generated do exist with this approach, these constraints impose no serious limitations on the distribution of events in observation space for a successful separation of classes.

The second important aspect to estimation of probability density functions, called typical sample representation, takes the adaptation procedure outlined above for histogram estimates as a point of departure, but implements this approach in a somewhat different manner. As before, the observation space is partitioned adaptively into regions called cells; however, the role of the estimation process and the geometrical disposition of these cells are not necessarily the same as for histogram construction. First of all, cells are created only in those portions of the observation space where learning samples have been observed. Since it is expected that in most practical problems a very high percentage of the volume of the observation space is empty, this serves to reduce significantly the storage requirements. Secondly, the size, shape

and height of a cell is determined from an examination of the local behavior of the learning samples in the neighborhood of the cell in question. From the local behavior of the learning samples a component function is generated which represents the learning samples in the immediate neighborhood of the cell.

The entire process of typical sample representation can be regarded as an adaptive method of approximating the probability density by expanding it in a set of not a priori specified component functions. The component functions represent and typify each of the different significant manifestations of members of the class by creating a cell corresponding to each of the different manifestations. The component functions also describe the local characteristics of each "typical" concentration of learning samples and they shape the cells.

Figure 3 illustrates the behavior of these component functions for a one-dimensional random variable. The process by which such an estimate of the probability density function is constructed encompasses three basic steps:

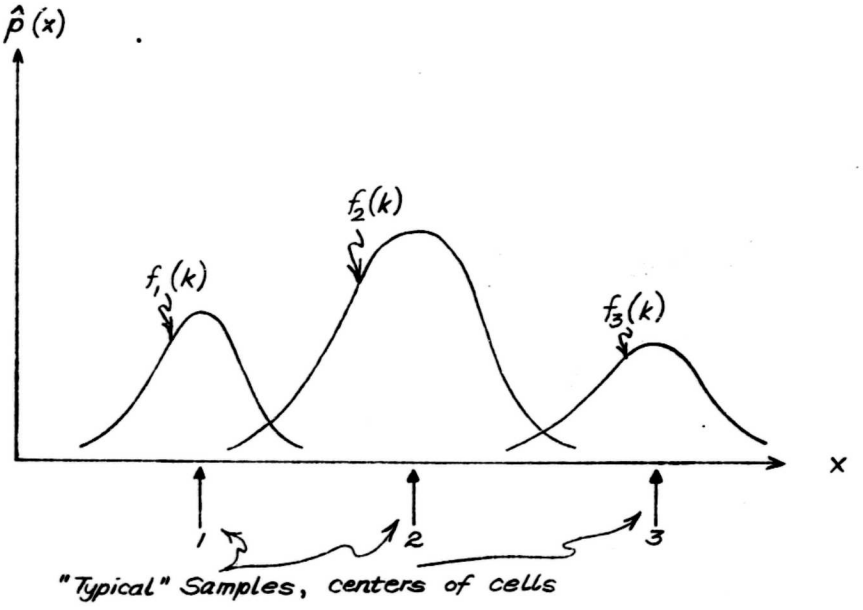


Figure 3. Typical Samples of a Probability Density

- A cell structure consisting of c cells is generated by the learning data.
- Corresponding to each cell, one of a class of functions, $\{f(\underset{\sim}{x})\}$, is selected according to values of learning data samples occurring within that cell.
- The probability density function is estimated by some sort of combination of the selected functions, $\left\{f_j(\underset{\sim}{x})\right\}$, $j = 1, 2, \ldots, c$, corresponding to the c cells.

Thus, the probability density function is represented by a set of "typical samples", where each typical sample consists of a reference point, a component function, and a cell boundary.

The cell structure is established by adaptive adjustments controlled by the sequence of samples contained in the learning set. Of the many ways in which rules for the adaptation can be established, * the following has been studied most extensively. The first learning data sample is established as the "reference point" for the first cell. The second sample is then compared with this reference point according to a criterion which indicates whether this sample should be used to modify the first cell (by adjusting its reference point), or be established as a new reference point for 9 second cell. If used to modify the first cell, then the criterion by which future learning samples will be compared with the reference point for that cell may also be modified. If the second sample is established as the initial reference point for the second cell, then a second criterion is also assigned to that cell. The third and succeeding learning data samples are compared with each of the established cell reference points (according to their respective criteria) and used to either modify one of these cells, or establish a new one.

The criterion by which new learning data samples are compared with an established cell is constrained to reflect a notion of similarity between the cell reference point and a new sample. The provision for adjustment of the criterion

*See Chapter 4 of Reference [4] for a general discussion of adaptive methods.

according to the value of a new sample, allows for development of different measures of similarity between events in observation space, according to location of the events in that space, as well as class membership. The criterion associated with a given cell may be regarded as a maximum allowable distance between its reference point and any other point in observation space to be associated with that cell, where "distance" is measured in an adjustable way. Thus, modification of the criterion for a cell changes the cell boundary which consists of all points in observation space equi-distant (at a specified value) from the cell reference point.

Either during or at the end of the process of cell formulation, the samples occurring in, say, the ν-th cell are used to select the component function $f_\nu(\underset{\sim}{x})$, from a pre-established class of functions, $\{f(\underset{\sim}{x})\}$. This set of functions may or may not allow for non-zero values of $f_\nu(\underset{\sim}{x})$ outside of the ν-th cell.

In practice, it is convenient to relate the class of component functions to the way in which distances are measured between points in the observation space and cell reference points. In particular, if the component function for a cell decreases monotonically as the distance between $\underset{\sim}{x}$ and the cell reference point increases, then the process of computing probability density function values at $\underset{\sim}{x}$ may be reduced to the calculation of distances between $\underset{\sim}{x}$ and the cell reference points. An illustration of this relationship is the use of quadratic forms for measuring distances, and Gaussian forms for the component functions.

The last step in the process of estimating probability density functions with typical samples consists of combining the component functions over the entire observation space. One way is to consider the probability density function to be the sum of the component functions:

$$p(\underset{\sim}{x}) = \sum_{\nu=1}^{c} f_\nu(\underset{\sim}{x}) \qquad (2.5)$$

For uncorrelated Gaussian component functions, this method allows for convenient processing of recognition data samples in which some parameter values are missing.[5] Another method of combining component functions is to use the function whose cell reference point is closest to the point at which the probability density function is to be estimated.

In the next section, recognition systems embodying the likelihood estimation approach to the classification problem are described. Each of these systems employs a version of the typical sample method of estimating and representing probability density functions. The adaptation, construction, and utilization of classification decision boundaries are illustrated for several methods involving different levels of equipment complexity for implementation.

3. MACHINE LEARNING BY PROBABILITY DENSITY FUNCTION ESTIMATION

Having phrased the problem of "machine learning" in terms of estimating class probability density functions, we may now turn, in Section 3.1, to the techniques which have been developed to implement and which operate on the principle of p.d.f. (probability density function) estimation. In Section 3.2 (and Appendix II) a theory is developed for selecting the control parameters which determine the performance of the techniques for any given set of data. Experiments are described in Section 3.2 (and Appendix III) which have substantially confirmed the theoretical predictions of Section 3.2.

3.1 ADAPTIVE SAMPLE SET CONSTRUCTION TECHNIQUES

There are two aspects of the problem of p.d.f. estimation in pattern recognition that must be emphasized. The method of estimation must be specified; the method of characterizing the estimated p.d.f. over the observation space, both at points where known examples have occurred and at points where past experience does not provide explicit guidance, must be determined. The first of these problems can be attacked in a straightforward manner with the aid of the theory of statistics. The second problem stems from the fact that a finite sample size (usually small) is used for estimation, so that the p.d.f. cannot be determined accurately over large portions of the observation space, and even if it could, consideration of the practical implementation schemes available must be taken into account. To illustrate these problems and a possible solution consider the computer program known as the Proximity

Algorithm.* This program, which has been used with some success in practical applications, does not explicitly consider the estimation problem as such but provides an immediate simple solution to the second problem. In the Proximity Algorithm "machine learning" consists simply of storing all the known data vectors in the computer memory. To perform recognition, the Proximity Algorithm computes the Euclidean distance between the unknown or test sample vector and all of the stored known vectors. The unknown vector is then assigned to (or decided to be a member of) the same class as the nearest stored vector. In effect, an input is considered a member of class A if it is closer to any known member of class A than to any known member of any other class. By this procedure the class p.d.f.'s are "estimated" to be "high" (only relative magnitudes matter) near known vectors of the class and to fall off in a manner inversely related to the distance from the nearest known vector. The estimated p.d.f. is therefore characterized by this rule over the entire observation space. The recognition machine consists simply of a memory and a distance computer. The Proximity Algorithm uniquely determines a set of decision regions in the one-observation sample space (but no more), and may be used to make decisions based on one observation.

The class of "learning" techniques which develop only a set of decision regions in an observation space have a fundamental limitation which is important in many applications. If the application is such that repeated observations may be made on an individual class member (or on different members of the same class) then it may well be desirable to base decisions on a sequence of observations and to change the number of observations used from decision to decision in a sequential decision rule. However, this is not convenient if the decision rule is tied to a fixed number of observations, and, in fact, an optimum decision

*Page 91 of Reference [4]

rule cannot be implemented. For this reason it is important to develop techniques which yield an absolute estimate of the class p.d.f.'s even if the single observation error probabilities are not significantly reduced. It is fortunate that the more sophisticated techniques to be discussed next are clearly not significantly more difficult to implement than the Proximity Algorithm.

A program which explicitly considers the p.d.f. estimation problem, and which simultaneously introduces an important concept for what is to follow, is known as the Adaptive Sample Set Construction Program, ASSC II.* The concept to be introduced is that of a "floating" cell structure determined by the data. The program operates by processing the known vectors sequentially in the following manner. A cell of prechosen size and shape is constructed with center at the first known vector. The defining relation for the cell is that all points in the observation space that are within a prechosen distance of the cell center belong to the cell. Thus, only two quantities are required to specify the cell: The (vector) cell center and the (vector) cell radius.**

The second known or "learning" sample vector is used to generate a second cell similar to the first if it falls outside the first cell. If the second "learning" vector falls inside the first cell, the center of the cell is shifted to the center of gravity or mean of the two "learning" vectors. The third and subsequent "learning" sample vectors are processed similarly, either generating new cells

*This program has been described in more detail and with some results of practical applications in the literature [7, 5].

**The coordinates of observation space may be normalized before distance is computed, thus allowing elipsoidal cells in the unnormalized space.

or updating old cells. Thus, the cells so generated for each class are located in the portion of the observation space where examples of the individual classes have been observed. Furthermore, the cells tend to move toward local peaks or modes of the class p.d.f.'s.

Once a cell structure over the observation space has been established, the estimation of the class p.d.f. proceeds in a manner similar to that in constructing a histogram. The quantities to be estimated are the probabilities of a sample vector falling in each cell. The estimator used, of course, is the fraction of the vectors in the learning sample that fall in each cell (this estimator is binomially distributed and has the desirable statistical properties of unbiasedness and sufficiency). In an endeavor to have as efficient a sampling procedure as possible, the estimate of the i-th cell probability is taken as $\hat{p}_i = n_i/n$ where n_i is the total number of vectors that have fallen in the cell throughout its history and n is the total number of known observations from the class in question (n is the size of the learning sample). That is, no account is taken of the fact that the cell moves about during the learning procedure.

In a conventional histogram the estimated p.d.f. is represented as a "staircase" function (constant over each cell). However, this is an undesirable form of p.d.f. representation for the present purposes since the cell structure as defined here does not (necessarily) cover the entire observation space. Therefore, regions in which no observations have occurred in the particular finite learning sample used would be assigned a p.d.f. value of zero.

The assumed form for characterizing the class p.d.f.'s once the cell structure has been determined, is to fit a function of the form $P_k \exp\left[-\frac{1}{2}\sum_i\left(\frac{x_i - m_{ik}}{b_i}\right)^2\right]$ to the k-th cell (for each class) where $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$ is an arbitrary point in the observation space, m_{ik} is the i-th coordinate of the center of the k-th cell, b_i is the prechosen "radius" of the cells in the i-th coordinate direction, and P_k is the proportion of the known

vectors (from that class) which fell in the k-th cell. The total p.d.f. estimate is then taken as the sum of the functions fitted to the individual cells, so that the assumed form of the class p.d.f. is that of a union of normal p.d.f.'s. Like the Proximity Algorithm, ASSC II assumes the class p.d.f. to fall off inversely with the distance from the known vectors.

In ASSC II the generated cells were made to depend on the known vectors both in position and number. An immediate generalization of this is to make the size and shape of each individual cell also depend on the data. This concept has been embodied in the computer program known as SPEAR (Statistical Property Estimation And Regeneration). In addition to updating the cell locations, the data is used to update the cell radii in each coordinate direction by an arbitrary rule. (The rule that has been used so far is discussed in detail in the next subsection.) Furthermore, the distance measure used in defining the cells is arbitrary although only Euclidean distance has been used so far.

The purpose of allowing the cell structure generation mechanism to depend so much on the data is twofold. First, a distinct possibility exists for reducing the number of required cells from that generated by, say, ASSC II. This, of course, would result in a reduction in the storage needed to specify the class p.d.f.'s. Second, making the cell structure depend so completely on the data relieves the experimenter of the difficult task of choosing the cell sizes or radii. This problem becomes particularly difficult when the observation space has many dimensions. Using SPEAR, the experimenter exerts an influence on the p.d.f. estimation procedure only through certain control parameters. Rules for choosing these control parameters can be theoretically derived from rather simple models; this is discussed in detail in the next subsection.

The cell structure generated by SPEAR is more general than that generated by ASSC II in that different cells may have different sizes and shapes. In order to perform recognition using the "learning" results of SPEAR, a program called ASSC III has been written. ASSC III is similar to the recognition part

of ASSC II. For reasons of computational economy, however, ASSC III uses a quantized form of Gaussian function about each cell center (or "typical sample") and the p.d.f. estimate at a point is determined by only the nearest "typical sample" rather than by the sum of Gaussian functions. Experimental evidence has substantiated the expectation that ASSC III and ASSC II will have substantially the same performance.

The programs described above and related programs are listed in Appendix IV with a short description of each.

In order to illustrate certain features of the above techniques, a series of three experiments using two-dimensional data (so that the results might be plotted) was designed. A description of these illustrative experiments and the results obtained is contained in Appendix III.

3.2 METHOD OF SELECTING CONTROL PARAMETERS

Most automatic machine learning techniques are not completely automatic. Thresholds and initial settings must be determined and preliminary analyses must be performed on the data before the so-called automatic pattern recognition techniques can commence operations. These thresholds or initial constants are usually determined by human inspection of the data in a particular experimental application. The purpose of this section of the Report is to describe attempts to automate, at least partly, the process of initial control parameter value selection. The rationale behind this attempt is that whatever human inspection of the data (or of the preliminary analyses) reveals, and whatever properties of the data prompt us to determine these starting constants, these very same properties can be examined by machine to automate the initial control parameter value selection process.

The choice of these control parameters can be determined theoretically. In order to use SPEAR for class probability density estimation, it is necessary to specify the initial cell size.

In Appendix II the expected behavior of a cell is derived as a function of the control parameter τ_n under the assumption that the cell is isolated and located in a region of the observation space over which the class p.d.f. is constant. Since the desired behavior of the cell in such circumstances is that it should grow in size, the value of τ_n can be chosen to assure that this should happen at least statistically.

The control parameters chosen in accordance with the conclusions derived in Appendix II have been used in a series of experiments that have substantially confirmed theoretical predictions.

The conclusion reached in Appendix II is that, for effective utilization of the cell growth feature of SPEAR, τ_n should lie in the range $1.1\sqrt{N+2}$ to $1.5\sqrt{N+2}$. For $\tau_n < 1.1\sqrt{N+2}$, very little cell growth will take place. For $\tau_n > 1.5\sqrt{N+2}$, the cells grow too rapidly. The particular choice of τ_N should, in part, be influenced by the sample size available for "learning". It is thus seen that the choice of this control parameter does not depend upon the data but only upon the number of dimensions. The initial cell size must, however, still be determined from a consideration of the data.

3.3 EXPERIMENTAL RESULTS OF SPEAR "LEARNING" WITH THEORETICALLY DETERMINED CONTROL PARAMETERS

The technique presented in Appendix 2 indicates that the controlled cell-growth mechanism obtains the best approximations of the joint probability densities when the numerical value of the control parameter τ_N is between $1.2\sqrt{N+2}$ and $1.5\sqrt{N+2}$ where N is the number of dimensions of the vector space. A series of experiments were conducted to determine how well SPEAR can estimate class probability densities. In order to facilitate the comparison of results with results that would be obtained if the approximation were perfect, data of complicated but known probability density was used in the tests.

The data was composed of two distributions, each containing four modes in a four-dimensional space. The probability densities were estimated by employing SPEAR from about 1000 samples of the dictributions. A number of different values of the control parameter τ_N were selected; best results were obtained with $\tau_N = 1.3\sqrt{N+2}$. For values of $\tau_N < 1.2\sqrt{N+2}$, the cell-growth mechanism was ineffective. For values of $\tau_N > 1.5\sqrt{N+2}$, cell growth was excessive and no discriminability between classes could be achieved. These results are in good agreement with predictions presented in Appendix 2, Figure B-3.

In addition to a point-by-point comparison of pdf estimates with the true values, independent data from the same distributions also was used to confirm the quality of the approximation technique. An additional 1000 vectors from each class were tested against the known true probability densities and against those obtained through SPEAR learning with $\tau_N = 1.3\sqrt{N+2}$. An error rate of 7.55% was obtained when the known probability densities were used, while an error rate of 7.8% was obtained using the results of machine learning. This result is especially significant since the data distribution was deliberately chosen to make class separability difficult. The two distributions may be described as two intertwined spirals in four-space.

The good agreement between error rates achieved with the true (0.0755) and the approximate (0.078) probability densities is strong evidence that the true and estimated densities are close almost everywhere, and that the SPEAR - ASSC II technique of machine learning and recognition is close to optimum in a wide class of recognition problems.

The data, results, and the experiments are described more fully in Appendix 6.

4. IDENTIFICATION OF HIGH QUALITY DECISION REGIONS

With any method of automatically making classification decisions, the question of when or how much to rely on the decisions rendered by the method usually is of operational significance. Particularly in applications requiring decisions with which a human may not have had prior experience it would be desirable to have an automatic decision-making machine emit an indication of the confidence which may be placed in a decision, as well as the decision (class indication) itself. Although the user of any automatic device will doubtless form his own opinion of the overall quality of a classification device, there are at least two practical reasons for building into the machine a criterion for judging quantitatively the quality of its own decision for any given event. First, if the decision is judged to be of low quality, then the user may be in a position to either ignore the decision or repeat an experiment to provide data for another decision. Secondly, the decision quality indicator may be used as a score or figure of merit marking the automatic decision making machine. It is conceivable that procedures could be found to alter the nature of the cell structure, sampling procedure, sample size, or other controllable variables to maximize the decision quality indicator and thus improve the classification device.

For instance, in the method of estimating class probability densities described in the preceding section and in the appendixes the initial cell sizes and cell dimensions are controllable quantities which have an important influence on the quality of the density function estimate on which the decisions are based. In the course of this study, geometrical considerations were employed to determine the numerical values of cell sizes, thresholds, and initial settings of the

program variables.* The use of procedures to maximize the decision quality indicator subject to variation of the program variables may lead to better classification devices. Mainly, the decision quality indicator would serve as a diagnostic tool with which to analyze the decision making procedure in order that improvements could be made.

While several measures may suggest themselves as suitable indicators of the quality of a decision (or of the decision making system), two of these seem to be particularly powerful candidates. A mathematical formulation of these two measures is given in the next subsection; here only the motivating notions are explained.

One of the proposed measures of the quality of a decision is the probability that the decision is correct. The proposed quality indicator thus measures whether or not the decision is correct.

The second is a measure of whether or not the decision is optimum. An optimum decision decides in favor of the class having the largest probability density at the point in the vector space in question. Since we only have estimates of each probability density, we would have to decide in favor of the class having the larger estimated probability density. If the estimate of the density of class A is larger than the estimate of the density of class B (so that we would decide in favor of A), the question is, "What is the probability that the actual density of A is larger than the actual density of B?" That is to say, "What is the probability that the decision we make is optimum?"

In each of these two cases, it is desirable to define a measure of quality, $Q(\underset{\sim}{x})$ and then invent procedures that maximize the volume of the region of the vector space over which $Q(\underset{\sim}{x}) \geq \eta$, where η is a quality threshold level.

*See Appendix II

4.1 TWO DECISION QUALITY CRITERIA

We have considered two different notions of what the attribute of high quality should reflect when applied to a classification decision. The first is simply that the decision should be correct. This suggests that an acceptable quality indicator would be an estimate of the probability that a decision based on the event $\underset{\sim}{x}$ is correct. When using likelihood estimation for constructing classification boundaries, this means:

$$Q_1(\underset{\sim}{x}) = \hat{\Pr}\left\{\underset{\sim}{x} \text{ belongs to the k-th class} \,\middle|\, \hat{L}_{\underset{\sim}{x}}(k) = \max_j \hat{L}_{\underset{\sim}{x}}(j)\right\} \tag{4.1}$$

The true probability of correct decision for the classification is, of course, unknown. The estimated probability can be obtained by substituting estimated probability density functions wherever true probability density functions would be involved in calculating the true probability of correct decision for the (hypothetical) optimum classifications device. This method of estimation produces the following result: *

$$Q_1(\underset{\sim}{x}) = \frac{P_k \hat{L}_{\underset{\sim}{x}}(k)}{\sum_{j=1}^{M} P_j \hat{L}_{\underset{\sim}{x}}(j)}, \tag{4.2}$$

where

$$\hat{L}_{\underset{\sim}{x}}(k) = \max_j \hat{L}_{\underset{\sim}{x}}(j)$$

*The true counterpart to this quality indicator has been discussed by A. H. Nuttall in a Litton Internal Memorandum entitled, "Error Probabilities Conditioned on Specific Observations," December 20, 1961.

With equal a priori class probabilities,

$$Q_1(\underset{\sim}{x}) = \frac{\hat{L}_{\underset{\sim}{x}}(k)}{\sum_{j=1}^{M} \hat{L}_{\underset{\sim}{x}}(j)} \tag{4.3}$$

Since values of the estimated likelihood functions have to be calculated anyway, Eq. (4.3) constitutes an extremely simple additional computation to obtain a quality indication for each point in observation space, and each learning set of data.

The second quality indicator measure, $Q_2(\underset{\sim}{x})$, reflects the degree to which the decision associated with a given point in observation space can be expected to be the same as that which would have been rendered by an (hypothetical) optimum machine. Thus, the criterion is not accuracy of decision per se, but rather accuracy with respect to that attainable by any means. The measure indicates whether or not machine learning has been adequate and whether it was based on a sufficient number of learning samples.

The "optimum" device which serves as a standard with this criterion establishes classification decision regions by computing (true) likelihood functions, and choosing the class for which the likelihood function is largest. Explicitly, the quality indicator is defined as

$$Q_2(\underset{\sim}{x}) \equiv \hat{\Pr}\left\{ \hat{L}_{\underset{\sim}{x}}(k) = \max_j \hat{L}_{\underset{\sim}{x}}(j) \,\middle|\, L_{\underset{\sim}{x}}(k) = \max_j L_{\underset{\sim}{x}}(j) \right\} \tag{4.4}$$

Since the estimated likelihood function $\hat{L}_{\underset{\sim}{x}}(j)$ depends on the particular set of learning samples used in the estimation, repeated machine learning experiments with new learning samples would, in general, result in different estimates

of $L_{\underset{\sim}{x}}(j)$. Therefore we may regard the estimate of the likelihood function to be a random variable. This is portrayed in Figure 4 which shows the probability density of the estimated likelihood function of classes j and k at point $\underset{\sim}{x}$ when a limited number of samples are used in the estimation. If the number of samples on which $\hat{L}_{\underset{\sim}{x}}(j)$ and $\hat{L}_{\underset{\sim}{x}}(k)$ are based is increased, the variances of the distributions would decrease.

It is seen in Figure 4 that there is some probability that the estimated likelihood functions are in error and that the estimate indicates that j is more likely than k when indeed the reverse is true. The measure $Q_2(\underset{\sim}{x})$ indicates an estimate of the probability that decisions based on estimates are the same as decisions if the true densities were known. The quality of a decision depends on the number of samples and on how different the true likelihood functions are

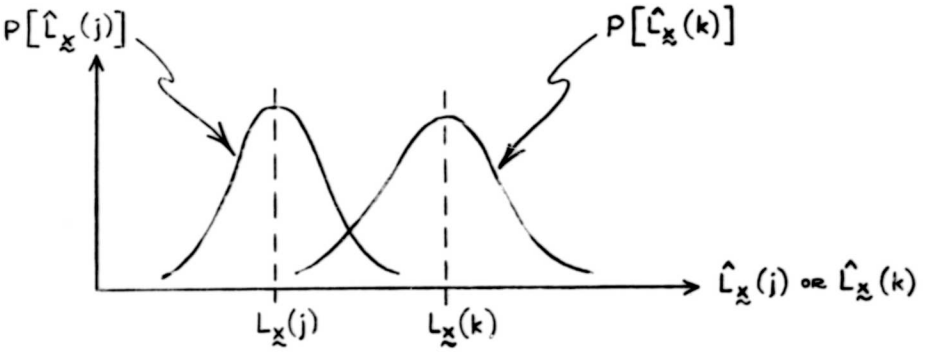


Figure 4. Distribution of Estimates of Likelihood Functions

at $\underset{\sim}{x}$. The explicit relationship between $Q_2(\underset{\sim}{x})$ and the $\left\{\hat{L}_{\underset{\sim}{x}}(j)\right\}$ involves the multinomial distribution. For simplicity only the form of $Q_2(\underset{\sim}{x})$ for the binary classification case (M = 2) is presented here. In the ν-th cell, it can be shown that*

*Appendix V

$$Q_{2\nu} = \sum_{r_1=0}^{N_1} \sum_{r_2=0}^{N_2} \prod_{j=1}^{2} \binom{N_j}{r_j} \left[\frac{N_{j\nu}}{N_j}\right]^{r_j} \left[1 - \frac{N_{j\nu}}{N_j}\right]^{N_j - r_j}$$

$$\cdot \operatorname{sgn}\left\{\left[\left(\frac{N_2 r_1}{N_1 r_2} - 1\right)\left(\frac{N_2 N_{1\nu}}{N_1 N_{2\nu}} - 1\right)\right] + 1\right\} - 1 \qquad (4.5)$$

where

$$\operatorname{sgn} x = \frac{x}{|x|} \text{ if } x \neq 0$$

$$= 0 \text{ if } x = 0$$

Note that dependence of $Q_{2\nu}$ on the learning set, $\{N_{j\nu}\}$, is indicated explicitly in Eq. (4.5)

4.2 EVALUATION OF DECISION QUALITY INDICATORS

Involved in each of the decision quality indicators introduced above is an estimate of a probability. If the true probability could be calculated, then there should be little reluctance to using each of these indicators in its intended role. However, since only estimates can be obtained, the validity of their use may be questioned. To obtain some idea of the degree to which the estimated quantities produce indications comparable to those which would be produced if the true probabilities were known and utilized, both analytical and numerical comparisons have been made. To simplify matters we have confined the evaluation to a rudimentary method of estimating probability density functions from the learning data: histogram construction. The same cell structure (uniformly spaced hypercubes) is assumed for all classes, and each class is assumed equally likely to be selected. Again for simplicity, only situations in which a classification decision is desired for a single test sample will be considered.* With

*Decisions based on multiple test samples are treated for binary classifications in Appendix V.

these restrictions, within the ν-th cell the likelihood function of the j-th class is assumed to be constant and is estimated by

$$L_{j\nu} \equiv L_{\underset{\sim}{x}}(j) = \frac{N_{j\nu}}{N_j V_\nu} \text{ for all } \underset{\sim}{x} \text{ in the } \nu\text{-th cell,} \tag{4.6}$$

where

$N_{j\nu}$ = the number of learning samples of the j-th class which occur within the ν-th cell

N_j = the total number of learning samples of the j-th class

and V_ν = the hypervolume of the ν-th cell.

Consider the indicators, Q_ℓ, $\ell = 1, 2$, defined in Eq. (4.1) and (4.4), respectively. Either indicator is useful only to the degree that it corresponds to the true probability (of correct or optimum decision) at a given point in observation space. Let $Q'_\ell(\underset{\sim}{x})$ denote the latter quantity. Now, for convenience, define

$$L_{j\nu} \equiv \int_{R_\nu} L_{\underset{\sim}{x}}(j)\, d\underset{\sim}{x} \tag{4.7}$$

where R_ν denotes the region comprising the ν-th cell. The quantity $L_{j\nu}$ corresponds to the average value of the likelihood function of the j-th class within the ν-th cell.* The lack of correspondence between the indicator and its true counterpart can be measured by the squared difference between the estimated and true quantities. Since $Q_\ell(\underset{\sim}{x})$ is a constant within a given cell (by histogram

*Also, $L_{j\nu}$ = the probability that a sample chosen from the j-th class gives rise to an event in the ν-th cell.

estimation), it is convenient to use $L_{j\nu}$ in lieu of $L_{\underset{\sim}{x}}(j)$ in calculating $Q'_{\ell}(\underset{\sim}{x})$ within that cell. This leads to the following measure of difference between the decision quality indication, Q_{ℓ} and its true counterpart:

$$\epsilon^2_{\ell\nu} = \left[Q_{\ell\nu} - Q'_{\ell\nu} \right]^2 \qquad = 1,\ 2;\ \nu = 1,\ 2, \ldots, c \tag{4.8}$$

for all $\underset{\sim}{x}$ in the ν-th cell.

For some learning sets, $\epsilon^2_{\ell\nu}$ will be large at a given point in observation space, and for other learning sets $\epsilon^2_{\ell\nu}$ will be small at that point. A reasonable way to obtain an overall measure of correspondence between $Q_{\ell\nu}$ and $Q'_{\ell\nu}$ is to average the quantity $\epsilon^2_{\ell\nu}$ over a class of learning sets to obtain a rms difference between the two. Of particular interest in most applications is the class of learning sets consisting of a specified number of samples. Thus, the lack of correspondence between $Q_{\ell\nu}$ and $Q'_{\ell\nu}$ is defined quantitatively as

$$E^2_{\ell\nu} \equiv \left\{ \overline{\epsilon^2_{\ell\nu}} \right\}^{\frac{1}{2}} \tag{4.9}$$

where the horizontal bar above a quantity indicates that an average of the quantity is taken over all learning sets consisting of a specified number of samples.

Of course, Eq. (4.9) can be calculated only by using the histogram estimation and maximum likelihood classification procedures for classes and events with known statistics. For evaluation, we must set up the estimate $Q_{\ell\nu}$ for each cell by generating samples for many learning sets according to these known statistics, and then proceed to average the values obtained for $\epsilon^2_{\ell\nu}$ from these learning sets. Ideally, to obtain an analytical evaluation of $E_{\ell\nu}$, the joint probability density function of $Q_{\ell\nu}$ and $Q'_{\ell\nu}$ would be determined for prescribed numbers of learning samples for the M classes, and $E_{\ell\nu}$ calculated by

$$E_{\ell\nu} = \int_0^1 dQ_{\ell\nu} \int_0^1 dQ'_{\ell\nu} \, (Q_{\ell\nu} - Q'_{\ell\nu})^2 \, p(Q_{\ell\nu}, Q'_{\ell\nu}) \tag{4.10}$$

In general this proves to be a formidable task.

An analytical evaluation of $E_{2\nu}$ has been performed for the binary classification problem, details of which are presented in Appendix V. An example is worked out to illustrate the relationship between the cell structure resolution (as represented by the number of cells, c) and the size of the learning set (as represented by N_1 and N_2). Results of the preliminary calculations reported in that appendix indicate that for a given learning set size the rms difference between $Q'_1(\underset{\sim}{x})$ and $Q(\underset{\sim}{x})$ actually decreases as the cell structure resolution increases, at least for values $c \leq N = N_1 = N_2$.

During the next phase of this study program, further calculations will be conducted to assess the validity of both quality indicators discussed here, for several class distributions.

5. CONCLUSION

The material discussed in this report is concerned with decision making and classificatory problems in cases where a parametric representation of the machine's environment is available. Machine learning and decision making are considered as the task of automatically partitioning the parameter space where, in each region, only members of one class are contained. Partitioning of the parameter space is accomplished by estimating the joint probability densities of the parameters for each of the input classes in question and by performing maximum likelihood ratio decisions on the estimated joint probability densities. Thus "machine learning" is the process of estimating joint probability densities of the parameters of the input classes, while recognition is a process of evaluating and comparing the probability densities obtained during learning.

This report considers adaptive methods of estimating joint probability densities by use of a generalized histogram construction procedure where the cells of the histogram are obtained adaptively and in a manner dependent upon the input data. In essence, cells for histogram construction are obtained in only those regions of the vector space where input on the number of cells so constructed by the use of a cell-growth procedure are described and tested both theoretically and experimentally. The results obtained indicate the validity and the practical utility of achieving a reduction in machine storage requirements by use of the mechanism for growing and adapting "histogram" cells. The methods developed are intended to be readily implementable on both general and special purpose computers; in fact, a special purpose computer implementing the recognition functions created by the methods described in this report has been constructed on another program.

The decision procedures studied in this report are based on estimates of probability densities. It is important to know, therefore, the quality of the estimating procedure, both for the purpose of determining how reliable the decision rendered in any one instance is and for the purpose of modifying the learning procedure to yield decisions with a lower probability of error.

Future work will be directed toward increasing the automaticity of the methods described above by studying automated methods of selecting input parameters, and by developing techniques that make use of quality estimation criteria in influencing machine learning.

REFERENCES

[1] Middleton, D. and Van Meter, D., "On Optimum Multiple-Alternative Detection of Signals in Noise" IRE Trans. PGIT, Vol. IT-1, No. 2 September 1955.

[2] Middleton, D. and Van Meter, D., "Detection and Extraction of Signals in Noise from the Point of View of Statistical Decision Theory", Jour. Sor. Industrial and Applied Math., Part I: 3: 192-253.

[3] Crooke, A., Floyd, W., and Nuttall, A., "Statistical Instrumentation Study", Final Report on Contract AF30(602)-2663, 14 March 1963.

[4] Sebestyen, G.S., "Decision Making Processes in Pattern Recognition", The Macmillan Company, New York, 1962.

[5] Sebestyen, G.S., "Pattern Recognition by an Adaptive Process of Sample Set Construction", Trans. PGIT, Vol. IT-8, No. 5, September 1962.

[6] Peterson, W.W., Birdsall, T.G., and Fox, W.C., "The Theory of Signal Detectability", Trans. PGIT, Vol. IT-1, No. 3, 1955.

[7] Edie, J. and Sebestyen, G., "Voice Identification, General Criteria", Final Report on Contract AF30(602)-2499, 16 May 1962.

[8] Sommerville, D.W.Y., "An Introduction to the Geometry of N Dimensions", Ch. VIII, Dover Publications, Inc., New York, N.Y., 1958.

APPENDIX I

OPTIMUM HISTOGRAM APPROXIMATION OF A PROBABILITY DENSITY

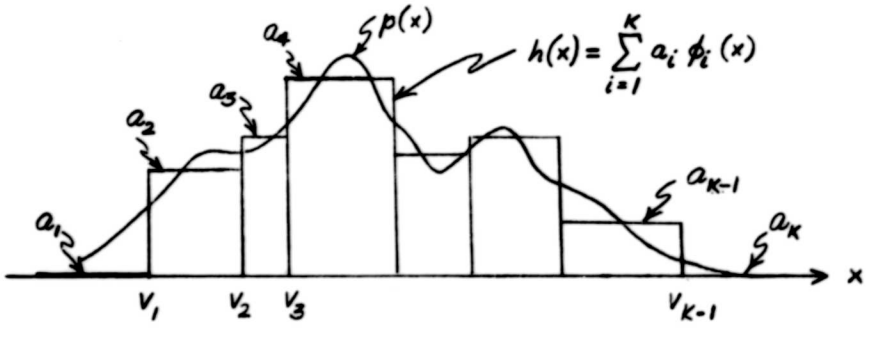
Given the probability density $p_m(\underset{\sim}{x})$, or just $p(x)$, we want to construct a histogram $h_m(\underset{\sim}{x})$, or just $h(x)$, composed of exactly K bars, which minimizes the mean-square error between $p(x)$ and $h(x)$. The mean-square error, E, is given in Eq. (A-1) where $\phi_i(x) = 1$ for $v_{i-1} < x < v_i$ and is 0 elsewhere.

$$E = \int_{-\infty}^{\infty} [p(x) - h(x)]^2 \, dx = \int_{-\infty}^{\infty} \left[p(x) - \sum_{i=1}^{K} a_i \phi_i(x) \right]^2 dx \qquad \text{(A-1)}$$

We may or may not want to impose the constraint that the area under $h(x)$ should be unity. If this constraint is imposed, as in Eq. (A-2), it is obvious that $a_1 = a_K = 0$.

$$\int_{-\infty}^{\infty} \sum_{i=1}^{K} a_i \phi_i(x) dx = 1 = \int_{-\infty}^{v_1} a_1 \, dx + \int_{v_{K-1}}^{\infty} a_K dx + \sum_{i=2}^{K=1} a_i (v_i - v_{i-1}) \qquad \text{(A-2)}$$

We now wish to minimize E by proper choice of the a_i's and the $\phi_i(x)$'s.



We will first minimize the error E, subject to the constraint stated in Eq. (A-2), with respect to the choice of a_i's. We will thus obtain a solution which expresses the optimum choice of the magnitude of each bar in the bar graph for any given assumption regarding the choice of bar boundaries. The a_i's will be expressed in terms of p(x) and the $\phi_i(x)$'s. Then, substituting these results into the expression of the error, E, we obtain E in terms of p(x) and the $\phi_i(x)$'s. Minimizing E with respect to the choice of $\phi_i(x)$'s should then obtain a solution of the optimum bar graph that minimizes the squared error between p(x) and h(x).

Carrying out the indicated operations by employing the methods of the calculus of variations, we solve for the variation of H in Eq. (A-3) from which, by noting that $a_i = a_K = 0$, we can solve for a_i.

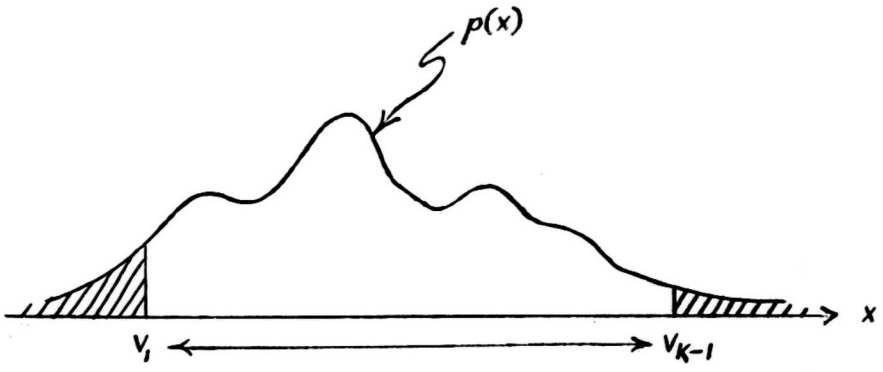
$$\delta H = \delta \int_{-\infty}^{\infty} \left(\left[p(x) - \sum_{i=1}^{K} a_i \phi_i(x) \right]^2 - \lambda \sum_{i=1}^{K} a_i \phi_i(x) \right) dx = 0 \qquad \text{(A-3)}$$

$$a_i = \left[\left(\frac{1}{x_i - v_{i-1}} \right) \int_{v_{i-1}}^{v_i} p(x)dx - \lambda \right] \qquad \text{(A-4)}$$

$$a_i = \left(\frac{1}{v_i - v_{i-1}} \right) \int_{v_{i-1}}^{v_i} p(x)dx + \left(\frac{1}{v_{K-1} - v_1} \right) \left[\int_{-\infty}^{v_1} p(x)dx + \int_{v_{K-1}}^{\infty} p(x)dx \right] \qquad \text{(A-5)}$$

It is thus seen that a_i is somewhat <u>larger</u> than the average p(x) in the interval defined by the i^{th} histogram bar. The amount by which it is larger is

the shaded area divided by the distance between shaded areas. This result is a simple statement of the fact that the area of the extreme quantiles is spread evenly over the rest of the quantiles.



In most cases this term will be small. For computational reasons it will be omitted and we will consider a_i to be the average value of p(x) in the i^{th} quantile.

Substituting a_i into E and partially differentiating it with respect to v_n, we obtain Eq. (A-6).

$$\frac{\partial E}{\partial v_n} = \frac{\partial}{\partial v_n} \sum_{i=1}^{K} \int_{v_{i-1}}^{v_i} \left[p(x) - \left(\frac{1}{v_i - v_{i-1}} \right) \int_{v_{i-1}}^{v_i} p(\zeta) d\zeta \right]^2 dx \qquad \text{(A-6)}$$

$$= \frac{\partial}{\partial v_n} \left\{ \int_{v_{n-1}}^{v_n} \left[p(x) - \left(\frac{1}{v_n - v_{n-1}} \right) \int_{v_{n-1}}^{v_n} p(\zeta) d\zeta \right]^2 dx \right.$$

$$\left. + \int_{v_n}^{v_{n+1}} \left[p(x) - \left(\frac{1}{v_{n+1} - v_n} \right) \int_{v_n}^{v_{n+1}} p(\zeta) d\zeta \right]^2 dx \right\} \qquad \text{(A-7)}$$

$$= \frac{\partial}{\partial v_n} \left\{ I + J \right\} \qquad \text{(A-8)}$$

But $\partial I/\partial v_n$ can be expressed as in Eq. (A-9) and $\partial J/\partial v_n$ as in Eq. (A-10).

$$\frac{\partial I}{\partial v_n} = -2\int_{v_{n-1}}^{v_n}\left[p(x) - \left(\frac{1}{v_n - v_{n-1}}\right)\int_{v_{n-1}}^{v_n} p(\zeta)\,d\zeta\right]\left[\frac{(v_n - v_{n-1})\,p(v_n) - \int_{v_{n-1}}^{v_n} p(\zeta)\,d\zeta}{(v_n - v_{n-1})^2}\right]dx$$

$$+\left[p(v_n) - \left(\frac{1}{v_n - v_{n-1}}\right)\int_{v_{n-1}}^{v_n} p(\zeta)\,d\zeta\right]^2$$

$$\frac{\partial I}{\partial v_n} = \left[p(v_n) - \left(\frac{1}{v_n - v_{n-1}}\right)\int_{v_{n-1}}^{v_n} p(\zeta)\,d\zeta\right]^2, \qquad \text{(A-9)}$$

since the second bracket in the first integral is identically zero. Similarly,

$$\frac{\partial J}{\partial v_n} = -\left[p(v_n) - \left(\frac{1}{v_{n+1} - v_n}\right)\int_{v_n}^{v_{n+1}} p(\zeta)\,d\zeta\right]^2 \qquad \text{(A-10)}$$

Therefore:

$\frac{\partial E}{\partial v_n} = 0$ implies Eq. (A-11) which is illustrated in Fig. A.1.

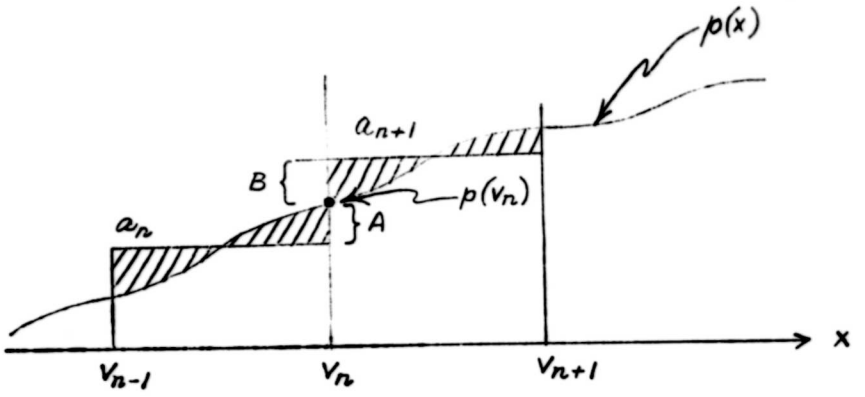
$$\left[p(v_n) - \left(\frac{1}{v_n - v_{n-1}}\right)\int_{v_{n-1}}^{v_n} p(\zeta)\,d\zeta\right]^2 = \left[p(v_n) - \left(\frac{1}{v_{n+1} - v_n}\right)\int_{v_n}^{v_{n+1}} p(\zeta)\,d\zeta\right]^2 \qquad \text{(A-11)}$$

$$A = B$$

for $n = 1, 2, \ldots, K-1$

$v_o = -\infty$

$v_K = +\infty$



A special care of interest is $a_2 = 2\,p(v_1)$. It is readily seen that if p(x) is monotonic, as shown in the sketch, the solution is completely determined from the choice of v_1. By iterative trials of v_1, it is possible to converge to the correct solution (if p(x) is monotonic). In general, the solution of Eq. (A-11) cannot be obtained.

In an attempt to obtain a general solution to the set of simultaneous equations, Eq. (A-11), several ideas have been tested. These will be listed below.

1) Since an iterative solution could be obtained to determine the locations of the histogram quantile boundaries, the v_n's, if p(x) were monotonic on an interval of x, it was thought that perhaps approximating the cumulative function $P(\eta) = \int_{-\infty}^{\eta} p(x)\,dx$ with a bar graph $H(\eta)$ of exactly K bars would yield quantile boundaries that have some simple relationship to boundaries of h(x) on p(x). In particular it was thought that the optimum quantiles on $H(\eta)$ may be identical to those on h(x). If this were the case, a general solution could always be obtained, since $P(\eta)$ is always monotonic. Unfortunately, a counter example proved that this hope was optimistic.

2) A second thought for obtaining a general solution was based on the idea that if the solution were known for some density function $p_1(x)$ then the solution for the optimum histogram on the given density p(x) could be obtained by finding the transformation that maps $p_1(x)$ into p(x), and thus the v_n's obtained for $p_1(x)$ into the v_n's valid for p(x). Since the Gaussian density (and several others) are symmetric (and monotonic over the two symmetric halves), the iterative numerical solution of Eq. (A-11) can be obtained for the Gaussian density (at least for K = even). Since it is possible to transform an arbitrary density p(x) to a Gaussian density, and similarly, a Gaussian density to an arbitrary density p(k),* it is therefore possible to map the quantile boundaries on the Gaussian density into quantile boundaries on p(x). It was hoped that boundaries obtained in this manner would be optimum.

Since theoretically the above hopes could not be readily proved either valid or false, a set of computer programs was written to carry out the iterative solution of h(x) on monotonic functions p(x) and for mapping one density into another. The following experimental procedure to prove or disprove the second idea was devised.

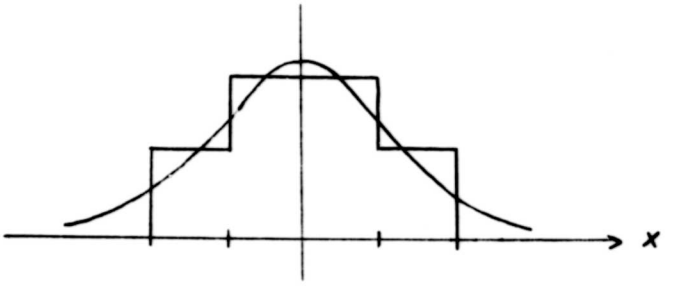
The optimum histograms on two different densities $p_1(x)$ and $p_2(x)$ were obtained by use of the iterative numerical procedure. Both $p_1(x)$ and $p_2(x)$ were monotonic (at least over half of the symmetric functions). The optimum histogram quantile boundaries of $p_1(x)$ thus obtained were mapped onto the second density $p_2(x)$. The location of quantile boundaries so obtained were then compared with those obtained from the direct iterative solution of the optimum histogram approximation pf $p_2(x)$. The magnitude of the squared error of approximation were also compared as obtained by the two different procedures.

---

*See pages 135-138, "Decision-Making Processes in Pattern Recognition", G. Sebestyen, The Macmillan Company, 1962.

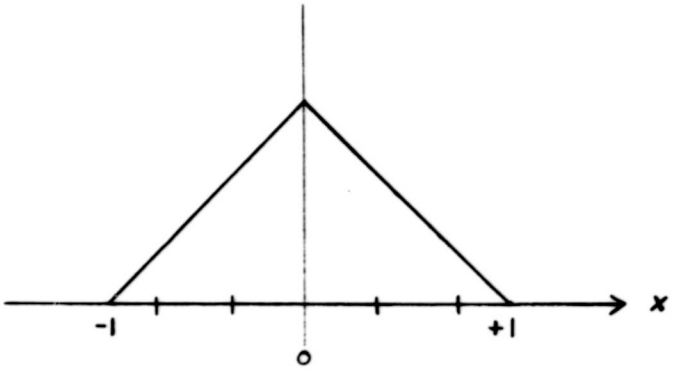
A summary of experimental results is given below. "Optimum" histograms of 6 bars were obtained for three densities.

a) **Gaussian:** $p(x) = \frac{1}{\sqrt{2\pi}} \exp(-1/2\, x^2)$. **The locations of quantile boundaries were $-\infty$, -1.796, -0.957, 0, $+0.957$, $+1.796$, $+\infty$. The error (area between p(x) and h(x) was 0.222.**

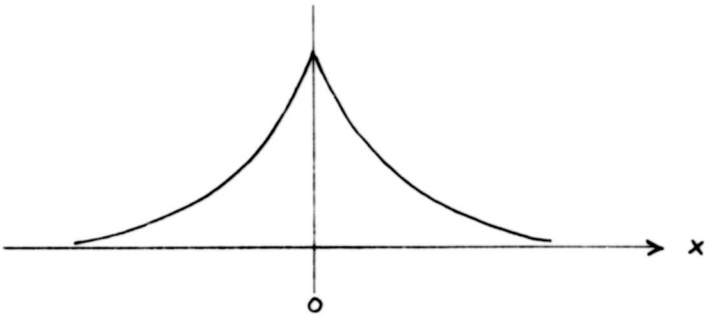


b) Triangular: $p(x) = 1 + x$ for $-< x < 0$ and is 0 elsewhere.
$p(x) = 1 - x$ for $0 < x < +1$

The quantile boundaries were at -0.8, -0.4, 0, $+0.4$, $+0.8$. The error due to the approximation was 0.2000.



c) Back-to-Back Exponential: $p(x) = 1/2\, e^{x}$ for $x < 0$ and $p(x) = 1/2\, e^{-x}$ for $x > 0$. Quantile boundaries were at -1.933, -0.677, 0, $+0.677$, $+1.933$. The error due to the approximation was 0.3400.



When the quantile boundaries from the Gaussian density were mapped onto each of the other two densities and the error between each p(x) and corresponding h(x) was compared with the error between p(x) and the histogram obtained by the mapping, the following results were obtained.

Error Between p(x) and h(x) for	Optimum h(x)	Mapped h(x) from Gaussian Solution	Difference in Two Errors	% Difference Between Two Errors
triangular p(x)	0.2000	0.2098	0.0098	4.9
exponential p(x)	0.3400	0.3482	0.0082	2.4

The analysis of these results did not conclusively reveal whether or not these small differences between the two methods of obtaining histograms for arbitrary densities were due to round-off errors in the digital computer or whether they could be regarded as counter examples proving that the second assumption was not valid. It is believed, however, that the above method will not yield a general solution.

APPENDIX II

CONTROL PARAMETER SELECTION THEORY

In this appendix the SPEAR computer program will be described in some detail and rules for the selection of the parameters which control the p.d.f. estimation procedure will be derived. It is desirable that the individual cells be adjusted by the data so that a good approximation to the class p.d.f. should be obtained with a minimum number of cells. Furthermore, the size and shape of the individual cells should be determined by a reasonable and automatic procedure from the data in order to relieve the experimenter of the almost impossible task of picking appropriate cell sizes.

For simplicity consider an isolated cell and let $\underset{\sim}{x}(t)$ be the t-th observation point (known class member) that falls in the cell, let $\underset{\sim}{m}(t)$ be the same mean of the first t observations that lie in the cell — $\underset{\sim}{m}(t)$ is the center of the cell at the t-th step — let $\underset{\sim}{a}(t)$ be a vector weighting parameter determined according to a given rule and indicating the cell shape, and finally, let τ_N be a scaler constant. τ_N is the control parameter being studied here. Then, the cell is defined at the t-th step to be the set of points in the observation space

$$C(t) = \left[\underset{\sim}{x} \left| \sum_{s=1}^{N} \left(\frac{x_s - m_s(t)}{a_s(t)} \right)^2 \leq \tau_N^2 \right.\right] \qquad \text{(B-1)}$$

Thus the cell is the (elipsoidal) locus of points "closer" to the cell mean, $m_s(t)$, than $\tau_N a_s(t)$ in the s-th direction. It should be noted that such a cell is "mode seeking" in that it will move (as a function of t) in the direction of the greatest concentration of data

APPENDIX II

CONTROL PARAMETER SELECTION THEORY

In this appendix the SPEAR computer program will be described in some detail and rules for the selection of the parameters which control the p.d.f. estimation procedure will be derived. It is desirable that the individual cells be adjusted by the data so that a good approximation to the class p.d.f. should be obtained with a minimum number of cells. Furthermore, the size and shape of the individual cells should be determined by a reasonable and automatic procedure from the data in order to relieve the experimenter of the almost impossible task of picking appropriate cell sizes.

For simplicity consider an isolated cell and let $\underset{\sim}{x}(t)$ be the t-th observation point (known class member) that falls in the cell, let $\underset{\sim}{m}(t)$ be the same mean of the first t observations that lie in the cell — $\underset{\sim}{m}(t)$ is the center of the cell at the t-th step — let $\underset{\sim}{a}(t)$ be a vector weighting parameter determined according to a given rule and indicating the cell shape, and finally, let τ_N be a scaler constant. τ_N is the control parameter being studied here. Then, the cell is defined at the t-th step to be the set of points in the observation space

$$C(t) = \left[\underset{\sim}{x} \left| \sum_{s=1}^{N} \left(\frac{x_s - m_s(t)}{a_s(t)} \right)^2 \leq \tau_N^2 \right.\right] \qquad \text{(B-1)}$$

Thus the cell is the (elipsoidal) locus of points "closer" to the cell mean, $m_s(t)$, than $\tau_N a_s(t)$ in the s-th direction. It should be noted that such a cell is "mode seeking" in that it will move (as a function of t) in the direction of the greatest concentration of data

points. This is a very desirable feature. The cell is first established according to some rule by a data point which does not fall in any other cell so that $\underset{\sim}{m}(1) = \underset{\sim}{x}(1)$, i.e., the cell is initially centered about the first or establishing data point. If $\underset{\sim}{a}(t) \equiv \underset{\sim}{a}(0)$ for all t, the cell size and shape remains the same throughout the estimation process as in the ASSC II program. Then the choice of $\underset{\sim}{a}(0)$, which is based largely on physical considerations and intuition, is very critical and an intelligent choice is very difficult. But, if $\underset{\sim}{a}(t)$ is made to depend on the data sample the volume of the cell may be made to grow to an "optimum" size by proper choice of the constant τ_N. Although the cell might alternatively be made to shrink if the data indicated this were desirable, it is assumed here that the initial cell size is small compared to intervals in which the class p.d.f. changes greatly and, hence, only cell expansion is discussed below.

The rule in SPEAR for updating the vector $\underset{\sim}{a}(t) = (a_1(t), a_2(t), \ldots, a_N(t))$ is given in Eq. (B-2).

$$a_s^2(t) = \max \left[a_s^2(0), \zeta_s(t) \equiv \frac{1}{t} \sum_{r=1}^{t} (x_s(r) - m_s(t))^2 \right] \qquad \text{(B-2)}$$

Thus, $a_s(t)$ begins at a preset value and normally grows to be the sample standard deviation of the sample vectors in the cell neighborhood.

The radius of the cell defined by Eq. (B-1) in the j-th coordinate direction is $a_j(t)\tau_N$. τ_N is chosen according to the theory to be developed here, but the initial cell radii $a_j(0)\tau_N$ must still be picked on the basis of physical considerations.

In order that the cells be assured room to grow, an additional feature was incorporated into SPEAR. A (known) vector $x(t) = (x_1(t), x_2(t), \ldots, x_N(t))$ is used to generate a new cell if it does not fall within a given distance of any of the existing cell centers; i.e., if

$$\sum_{s=1}^{N} \left(\frac{x_s(t) - m_s(t)}{a_s(t)} \right)^2 > (\theta_N \tau_N)^2 \qquad \text{(B-3)}$$

where

θ_N is a prechosen parameter, $\theta_N \geq 1$.

If

$$\tau_N^2 < \sum_{s=1}^{N} \left(\frac{x_s(t) - m_s(t)}{a_s(t)} \right)^2 \leq (\theta_N \tau_N)^2 \qquad \text{(B-4)}$$

for all existing cells ($\underset{\sim}{m}(t)$ and $\underset{\sim}{a}(t)$), then $\underset{\sim}{x}(t)$ is temporarily discarded. This procedure continues until $t = t_1$, where t_1 satisfies $t_1 = c_1\omega$, where c_1 is the number of cells generated up to that time and ω is a control parameter. After $\underset{\sim}{x}(t_1)$ has been processed the temporarily discarded vectors are forced into the then existing cell structure — each vector going into the nearest cell — and used to update the individual cells. The next incoming (known) vector is put into one of the existing cells, temporarily discarded, or used to generate a new cell just as were the earlier vectors. This process continues until the number of vectors is equal to t_2 where $t_2 = 2c_2\omega$ with c_2 equal to the number of cells at $t = t_2$. At this time the second group of temporarily discarded vectors is forced into the then existing cell structure. This procedure is continued throughout the estimation, in general $t_k = 2^{k-1} c_k \omega$.

The cell volume might be considered optimum if it is as large as possible and still have the estimated p.d.f. over the large cell consistent with that obtained by estimating over smaller cells. It may be possible to show a trend to a desirable cell size only for special cases of local p.d.f. behavior. This, however, should be all that is needed to generate practical rules for updating $\underset{\sim}{a}(t)$ and for choosing τ_N.

In particular, if a cell is located in a region of the observation space over which the class p.d.f. is a constant, the cell size should expand until it covers the region of uniform distribution. Furthermore, once the cell is "firmly established" in the sense that a number of observations have fallen in the cell, the rate of expansion should be fairly rapid provided it does not grow substantially beyond the region of constant p.d.f. On the other hand, if the cell is initially located in a region over which the class p.d.f. is changing, the cell should not expand rapidly. Therefore, the rule for updating $\underset{\sim}{a}(t)$ and the choice of τ_N should be such that the expected cell behavior obeys these two intuitive rules.

To construct a model of the "cell growth" mechanism, a cell is assumed to lie in a region of uniform class p.d.f. and the random behavior of the cell is studied. For simplicity, in the following discussion only one cell is assumed. This is not too unrealistic since the cells are kept isolated, as discussed previously, at least for the critical early stages of cell structure generation.

The volume of an N-ellipsoid is $V_N = K_N \prod_{i=1}^{N} b_i$ when $\sum_{i=1}^{N} \frac{x_i^2}{b_i^2} = 1$ specifies the N-ellipsoid and $K_N = \frac{\pi^{N/2}}{\Gamma(\frac{N}{2})+1}$, [8]. A short table of K_N is given below.

N	1	2	3	4	5	6	7	8	9
K_N	2	π	$\frac{4\pi}{3}$	$\frac{\pi^2}{2}$	$\frac{8\pi^2}{15}$	$\frac{\pi^3}{6}$	$\frac{16\pi^3}{105}$	$\frac{\pi^4}{24}$	$\frac{32\pi^4}{945}$

A slice perpendicular to the x_j-axis at x_j is an (N-1)-ellipsoid specified by

$$\sum_{\substack{i=1 \\ i \neq j}}^{N} \frac{x_i^2}{b_i^2\left(1 - \frac{x_i^2}{b_j^2}\right)} = 1 \qquad \text{(B-5)}$$

and of volume

$$K_{N-1}\left(1-\frac{x_j^2}{b_j^2}\right)^{\frac{N-1}{2}}\prod_{i\neq j} b_i \qquad \text{(B-6)}$$

(Therefore, the volume V_N is obtained by integrating over x_j, thus,

$$V_N = K_N\prod_{i=1}^{N} b_i = 2\int_0^{b_j} K_{N-1}\left(1-\frac{x_j^2}{b_j^2}\right)^{\frac{N-1}{2}}\prod_{i\neq j} b_i\,dx_j$$

$$= K_{N-1}\frac{\Gamma\left(\frac{N+1}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{N}{2}+1\right)}\prod_{i=1}^{N} b_i .)$$

Assume a uniform probability distribution over the N-ellipsoid specified by $\sum_{i=1}^{N}\frac{x_i^2}{b_i^2} = 1$. Then, the p.d.f. of the x_j coordinate is

$$g_N(x_j) = \frac{K_{N-1}\left(1-\frac{x_j^2}{b_j^2}\right)^{\frac{N-1}{2}}\prod_{i\neq j} b_i}{K_N\prod_{i=1}^{N} b_i}$$

$$= \frac{\Gamma\left(\frac{N}{2}+1\right)}{\Gamma\left(\frac{N+1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \frac{\left(b_j^2 - x_j^2\right)^{\frac{N-1}{2}}}{b_j^N}, \quad -b_j \leq x_j \leq b_j \tag{B-7}$$

$$= 0, \text{ if } |x_j| > b_j .$$

Making the transformation of variables $x_j = \lambda_j b_j$, the p.d.f. of λ_j is

$$h_N(\lambda_j) = \frac{\Gamma\left(\frac{N}{2}+1\right)}{\Gamma\left(\frac{N+1}{2}\right)\Gamma\left(\frac{1}{2}\right)} \left(1-\lambda_j^2\right)^{\frac{N-1}{2}}, \text{ for } |\lambda_j| \leq 1 \tag{B-8}$$

$$= 0, \text{ for } |\lambda_j| > 1 .$$

The maximum value of $g_N(x_j)$ is $\frac{K_{N-1}}{K_N b_j}$ and of $h_N(\lambda_j)$ is $\frac{K_{N-1}}{K_N}$. Examples of $h_N(\lambda_j)$ for several values of N are shown in Figure B-1.

The mean and variance of λ_j are easily found to be

$$\overline{\lambda_j} = 0, \tag{B-9}$$

and

$$\text{Var } \lambda_j \equiv E(\lambda_j - \overline{\lambda_j})^2 = \frac{K_{N-1}}{K_N} \int_{-1}^{1} \lambda_j^2 \left(1-\lambda_j^2\right)^{\frac{N-1}{2}} d\lambda_j$$

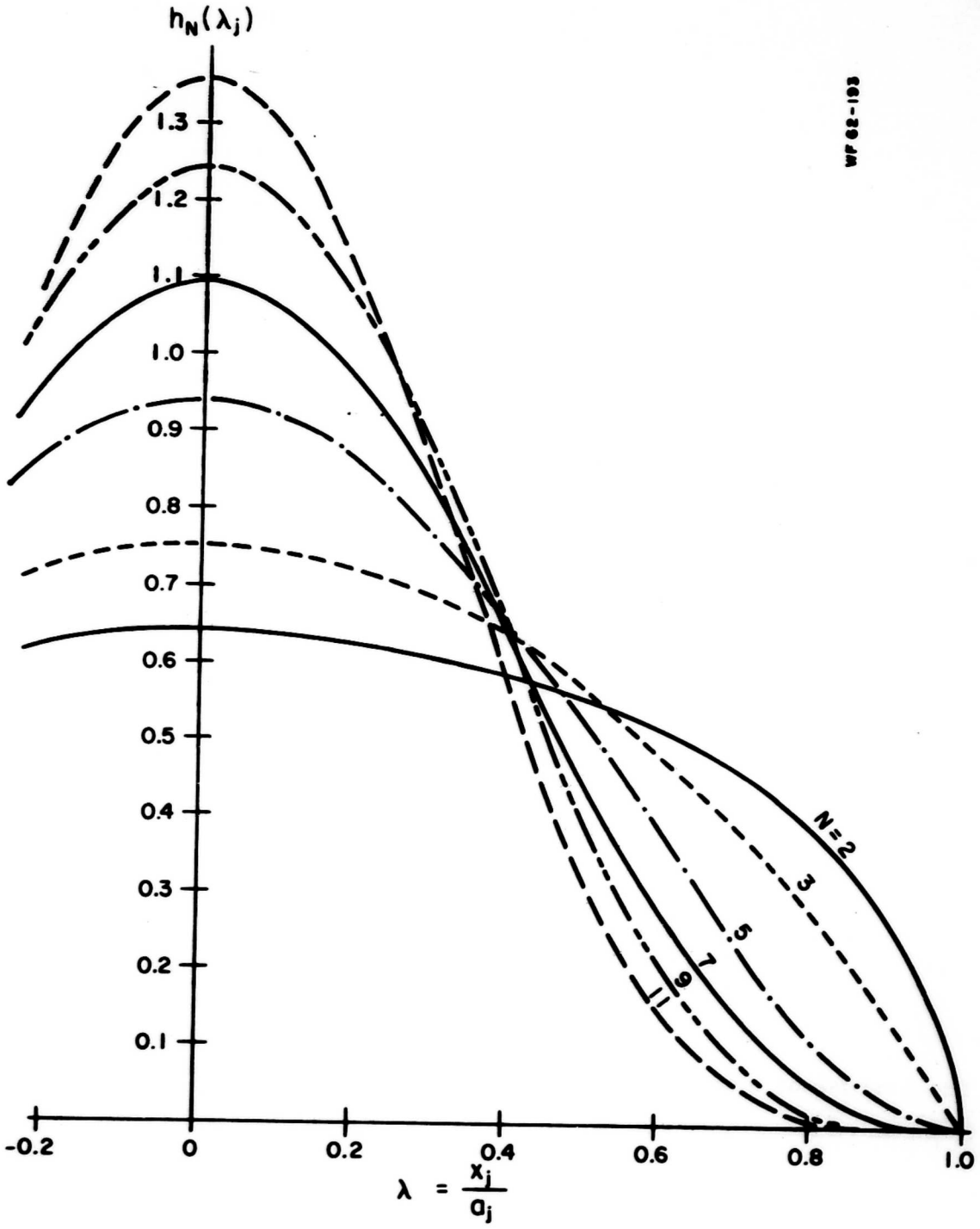


Figure B-1. p. d. f. of one coordinate of a point (uniform) randomly distributed over an ellipsoid of N dimensions.

$$= \frac{K_{N-1}}{K_N} \int_{-\pi/2}^{\pi/2} \sin^2\theta \cos^N\theta \, d\theta$$

$$= \frac{K_{N-1}}{K_N} \left[\frac{-1}{N+2} \sin\theta \cos^{N+1}\theta \Big|_{-\pi/2}^{\pi/2} + \frac{1}{N+2} \int_{-\pi/2}^{\pi/2} \cos^N\theta \, d\theta \right]$$

$$= \frac{K_{N-1}}{K_N} \frac{1}{N+2} \int_{-\pi/2}^{\pi/2} \cos^N\theta \, d\theta .$$

Therefore Var $\lambda_j = \frac{1}{N+2} \frac{K_{N-1}}{K_N} \frac{\sqrt{\pi}\, \Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{N}{2}+1\right)}$

$$= \frac{1}{N+2} . \qquad \text{(B-10)}$$

The p.d.f. of λ_j may be written as

$$h_N(\lambda_j) = \frac{\Gamma\left(\frac{N}{2}+1\right)}{\sqrt{\frac{N-1}{2}}\, \Gamma\left(\frac{N+1}{2}\right)} \cdot \sqrt{\frac{N-1}{2\pi}} \left(1-\lambda_j^2\right)^{\frac{N-1}{2}} \qquad \text{(B-11)}$$

The first factor tends to unity as $N \to \infty$, by Sterling's formula. The logarithm of the third factor is

$$\frac{N-1}{2} \ell n \left(1-\lambda_j^2\right) = \frac{N-1}{2} \left(-\lambda_j^2 - \frac{\lambda_j^4}{2} - \frac{\lambda_j^6}{3} - \cdots\right)$$

$$= - \frac{\lambda_j^2}{2 \frac{1}{N-1}} \left(1 + \frac{\lambda_j^2}{2} + \frac{\lambda_j^4}{3} + \cdots\right).$$

Hence, for any fixed λ_j such that $1 >> \left(\frac{\lambda_j^2}{2} + \frac{\lambda_j^4}{3} + \cdots\right)$,

$$\left(1-\lambda_j^2\right)^{\frac{N-1}{2}} \simeq \exp\left[-\frac{\lambda_j^2 (N-1)}{2}\right]. \qquad \text{(B-12)}$$

In this region, $h_N(\lambda_j)$ is approximately normal with mean zero and variance $\frac{1}{N-1}$. The tails of the distribution of λ_j decrease much faster than for the normal distribution.

Using the fact that the p.d.f. of λ_j is approximately normal over the range, say, $|\lambda_j| < 1/\sqrt{5}$ (but for the factor of

$$\frac{\Gamma\left(\frac{N}{2}+1\right)}{\sqrt{\frac{N-1}{2}}\,\Gamma\left(\frac{N+1}{2}\right)}$$

which is only slightly greater than unity for $N > 3$), an indication of the probability of cell growth may be easily obtained. In particular, the probability that the j-th component of the difference between the t-th sample point to fall in the cell and the (t-1)th sample mean,

$$\delta_j^2(t) = [x_j(t) - m_j(t-1)]^2 > a_j^2(t-1) \qquad \text{(B-13)}$$

is plotted in Figure B-2 as a function of τ_N for several values of N. These curves are sufficiently accurate for practical use, particularly for $N > 3$. The t-th observation may be said to contribute to cell growth if

$$[x_j(t) - m_j(t)]^2 = \delta_j^2(t) \left(1 - \frac{1}{t}\right)^2 > a_j^2(t-1). \qquad \text{(B-14)}$$

Therefore, the curves in Figure B-2 may be interpreted as the probability of ultimate cell growth as t becomes large.

Letting

$$\delta_j(t) \equiv \begin{cases} 0, & t = 1 \\ x_j(t) - m_j(t-1), & t > 1, \end{cases} \qquad \text{(B-15)}$$

the t-th cell center is located at

$$m_j(t) = m_j(t-1) + \frac{1}{t}\,\delta_j(t). \qquad \text{(B-16)}$$

Therefore,

$$\zeta_j(t) \equiv \frac{1}{t} \sum_{r=1}^{t} [x_j(r) - m_j(t)]^2$$

$$= \frac{1}{t}\left\{ \sum_{r=1}^{t-1} [x_j(r) - (m_j(t-1) + \frac{1}{t}\,\delta_j(t))]^2 + \frac{t-1}{t}\,\delta_j(t)^2 \right\}$$

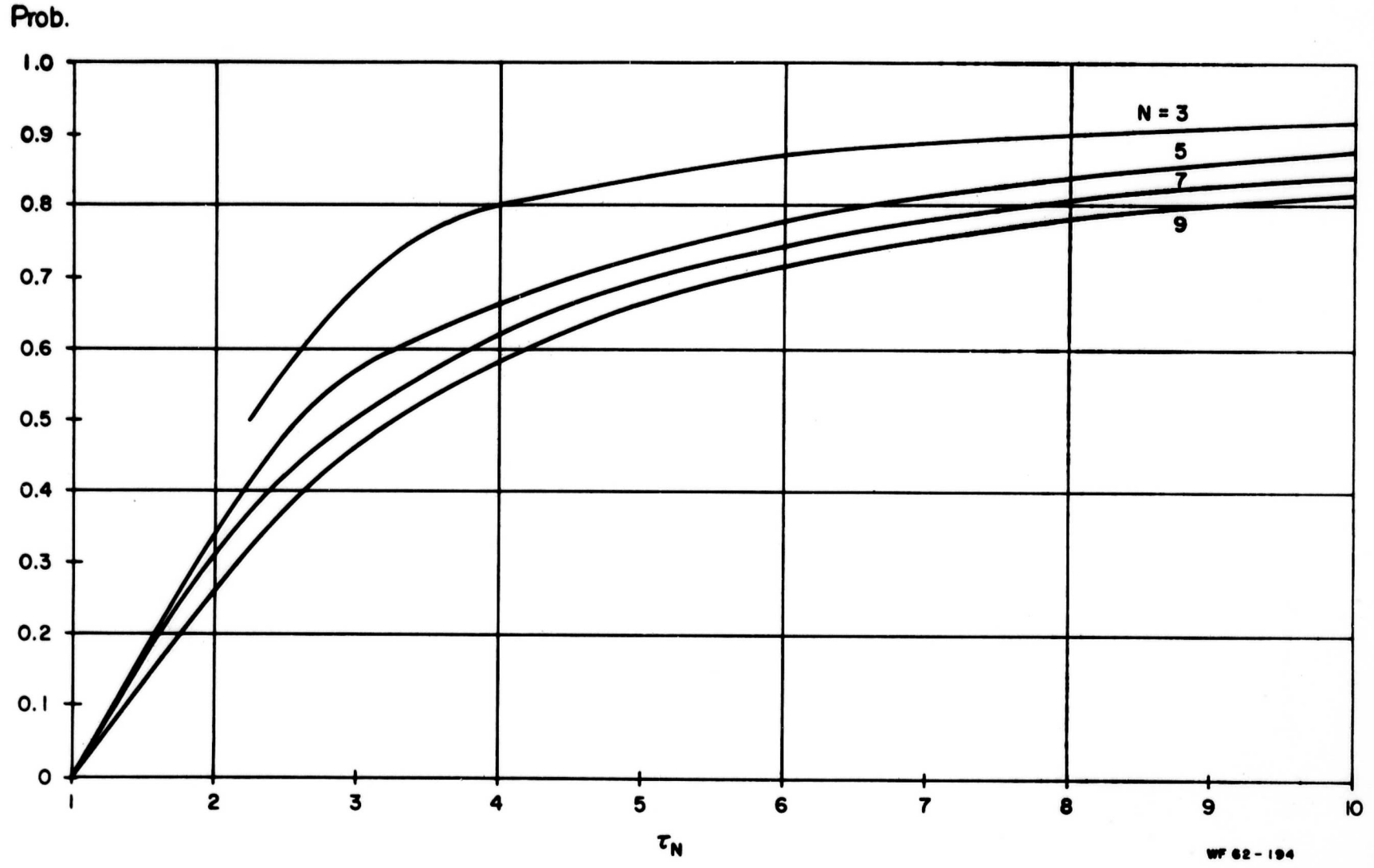


Figure B-2. Probability that $\delta_j^2(t) > a_j^2(t-1)$ as a function of τ_N, for a cell located in a region of uniform (class) probability density.

$$= \frac{1}{t}\left\{\sum_{r=1}^{t-1}\left[(x_j(r) - m_j(t)) - \frac{1}{t}\delta_j(t)\right]^2 + \left(\frac{t-1}{t}\delta_j(t)\right)^2\right\}$$

$$= \frac{1}{t}\left\{(t-1)\zeta_j(t-1) + \frac{(t-1)}{t^2}\delta_j^2(t) + \left(\frac{t-1}{t}\delta_j(t)\right)^2\right\}$$

$$= \frac{t-1}{t}\left\{\zeta_j(t-1) + \frac{\delta_j^2(t)}{t}\right\}$$

$$= \frac{1}{t}\sum_{r=0}^{t-2}\frac{t-r-1}{t-r}\,\delta_j^2(t-r). \qquad \text{(B-17)}$$

Let $t = t'$ be the index of the first sample point for which $a_j(t') > a_j(0)$, i.e., the first time cell growth occurs. Then, for $t \leq t'$, by Eq. (B-1) and (B-10) the expected value of $\zeta_j(t)$ is

$$\overline{\zeta_j(t)} = \frac{1}{t}\sum_{r=0}^{t-2}\frac{t-r-1}{t-r}\,\overline{\delta_j^2(t-r)}$$

$$= \overline{\delta_j^2(t-r)} \cdot \frac{1}{t}\sum_{r=0}^{t-2}\frac{t-r-1}{t-r}$$

$$= \frac{a_j^2(0)\,\tau_N^2}{N+2} \cdot \frac{1}{t}\sum_{r=0}^{t-2}\frac{t-r-1}{t-r}, \quad t \leq t', \qquad \text{(B-18)}$$

$$\longrightarrow \frac{\tau_N^2}{N+2}\, a_j^2(0) \text{ as } t \to \infty \qquad \text{(B-19)}$$

The sum on the right is conveniently approximated for the range of principal interest, say, $4 \leq t \leq t' \leq 20$, by

$$\sum_{r=0}^{t-2} \frac{t-r-1}{t-r} \simeq \frac{1}{6}(5t-9).$$

Then,

$$\overline{\zeta_j(t)} \simeq \frac{\tau_N^2}{N+2}\, a_j^2(0) \cdot \frac{(5t-9)}{6t} \text{ for } 4 \leq t \leq t' \leq 20. \qquad \text{(B-20)}$$

From Eq. (B-2) and (B-18) or (B-19), it is easily seen that a necessary condition for $\overline{a_j(t)}$ to be greater than $a_j(0)$ so that cell growth may be expected to begin is

$$\tau_N^2 > N+2.$$

It may be observed from the curves of Figure B-2 that the probability that $\delta_j^2(t) > a_j^2(t-1)$ is equal to one-half for $\tau_N = \sqrt{N+2}$.

The choice of τ_N determines not only whether the cell may be expected to grow, but also the number t* of observations that must fall in the cell before cell growth can be expected to begin. It is desirable that t* be chosen sufficiently large to establish a firm cell location and that a preliminary estimate of the cell probability has been made before the cell may be expected to grow.

This implies that t* should be at least four or greater. On the other hand, since the amount of data available for p.d.f. estimation is always limited in practice, t* must not be too large. For most applications t* must be chosen less than, say, 20 or else the learning sample will be exhausted before a significant number of cells have had a chance to grow to the size and shape they would ultimately reach if the amount of data was unlimited.

Having chosen a value for t* that is consistent with the cost of sampling (normally increasing at least linearly with the sample size n) and with the desired accuracy of the quantities to be estimated (the standard deviations of the extimates of these quantities decrease as $n^{-1/2}$) the choice of the control parameter τ_N becomes automatic. Writing $\tau_N = \beta\sqrt{N+2}$, and considering β as an unknown, Equation (B-18) is easily solved for β.

$$\beta = \left[\frac{\overline{\zeta_j(t)}}{a_j^2(0)} \cdot \frac{1}{\frac{1}{t}\sum_{r=0}^{t-2} \frac{t-r-1}{t-r}} \right]^{1/2}, \quad t \leq t \tag{B-21}$$

Now t* is the index of the first sample point for which $\overline{a_j^2(t^*)} > a_j(0)$ or $\overline{\zeta_j(t^*)} > a_j^2(0)$. Setting

$$\beta^* = \left[\frac{1}{t^*} \sum_{r=0}^{t^*-2} \frac{t^*-r-1}{t^*-r} \right]^{-1/2} \tag{B-22}$$

the choice

$$\tau_N = \beta^* \sqrt{N+2} \tag{B-23}$$

will result in a beginning of cell growth after an average of t* sample points fall in the cell when the true class p.d.f. is uniform over the cell.

A curve of β^* as a function of t* is plotted in Figure B-3. For the particular choice of $\tau_N = 1.4\sqrt{N+2}$, in a long series of repeated trials of placing an isolated cell in a region of uniform class p.d.f. (and sampling from the class) the curve in Figure B-3 indicates that the average value of t will be approximately 4.7.

Since the p.d.f.'s of greatest interest are distinctly non-uniform over the entire space, there will normally be a wide spread in the range of cell probabilities. Therefore, the cells with high probabilities will normally begin to grow before the majority of the cells have collected t* observations. 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. (The growth of an individual cell is limited by the presence of surrounding cells as well as by the non-uniform local nature of the class p.d.f.) Therefore, a reasonable choice of the control parameter ω is $\omega = t^*$.

An investigation of the dynamics of the growth mechanism (which is currently underway) may be expected to shed more light on the method of choosing the control parameter discussed here. Experimentation should also be of value in substantiating the theory presented here and should indicate if modifications to the viewpoint taken here are necessary.

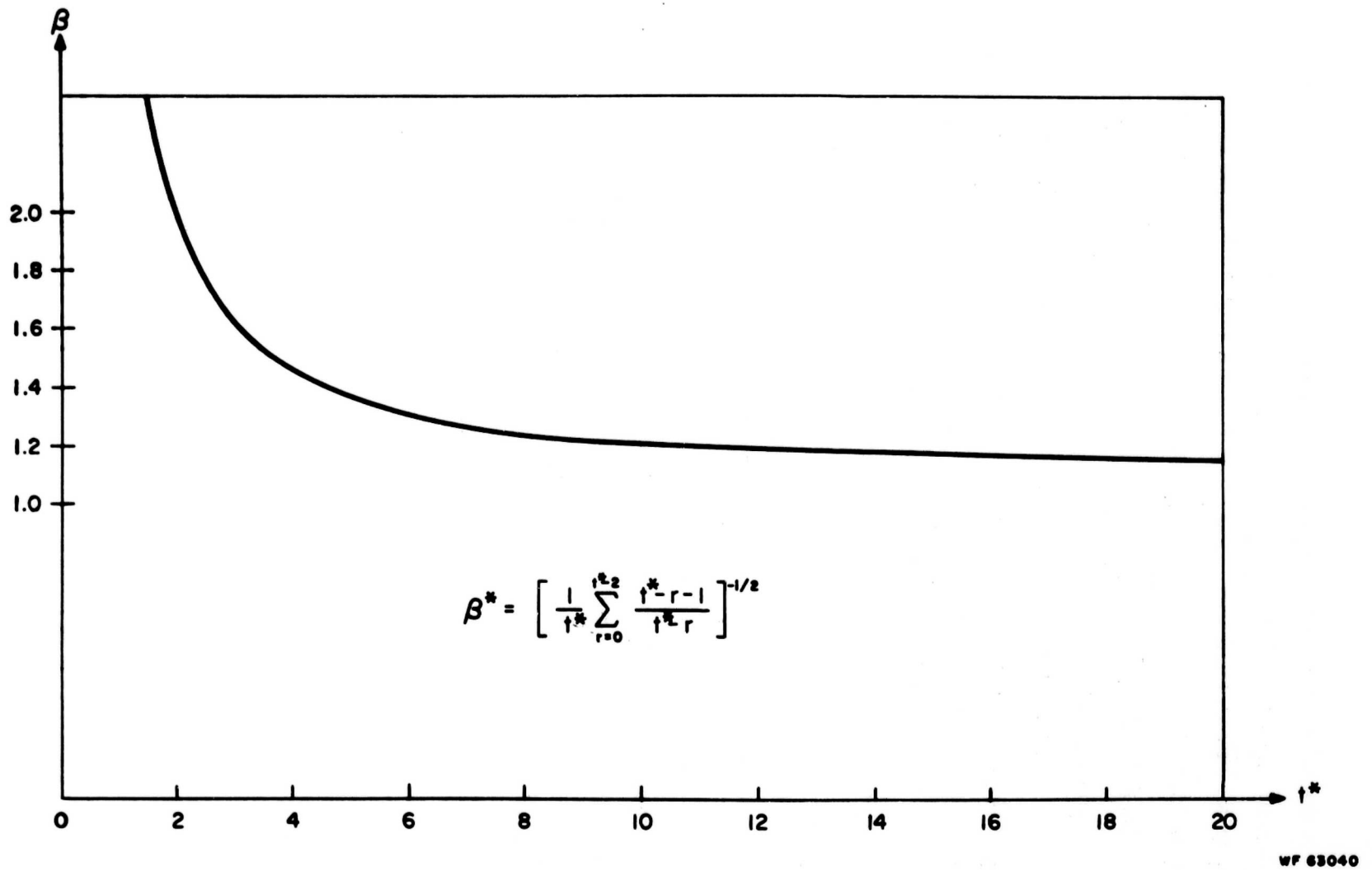


Fig. B-3 Curve Used in Selecting the Control Parameter τ_n

APPENDIX III

COMPARISON OF THREE PATTERN RECOGNITION TECHNIQUES

In this appendix the three (related) pattern recognition techniques described in Section 3.1 will be illustrated on the same set of two-dimensional data. The data used here was generated artificially in such a way that it should be impossible to separate classes with simple techniques such as correlation techniques. Furthermore, the classes overlap sufficiently to preclude perfect separation by any technique. Thus some indication may be obtained of the power of any technique tested with this data.

The first class is represented by 200 examples shown as dots in Figure C-1. The second class is represented by 150 examples shown as crosses in Figure C-1. For "learning" purposes one hundred examples were selected at random for each class; these are shown in Figure C-2 with the one-sample decision boundary obtained by use of the Proximity Algorithm, ASSC II, and SPEAR II* computer programs. Note the high degree of similarity of the class separation boundaries obtained with the three different methods.

With the Proximity Algorithm, "learning" consists wholly of determining the decision boundary shown in Figure C-2. It should be noted that the decision boundary in Figure C-2 is determined by pairs of points, one from each class, such that any point on the boundary lies on the perpendicular bisector of the line joining the nearest pair. All data points not included in one of these boundary

---

*The decision boundaries for ASSC II and SPEAR II are dependent on the representing functions as well as on the cell structures to which these functions are fitted.

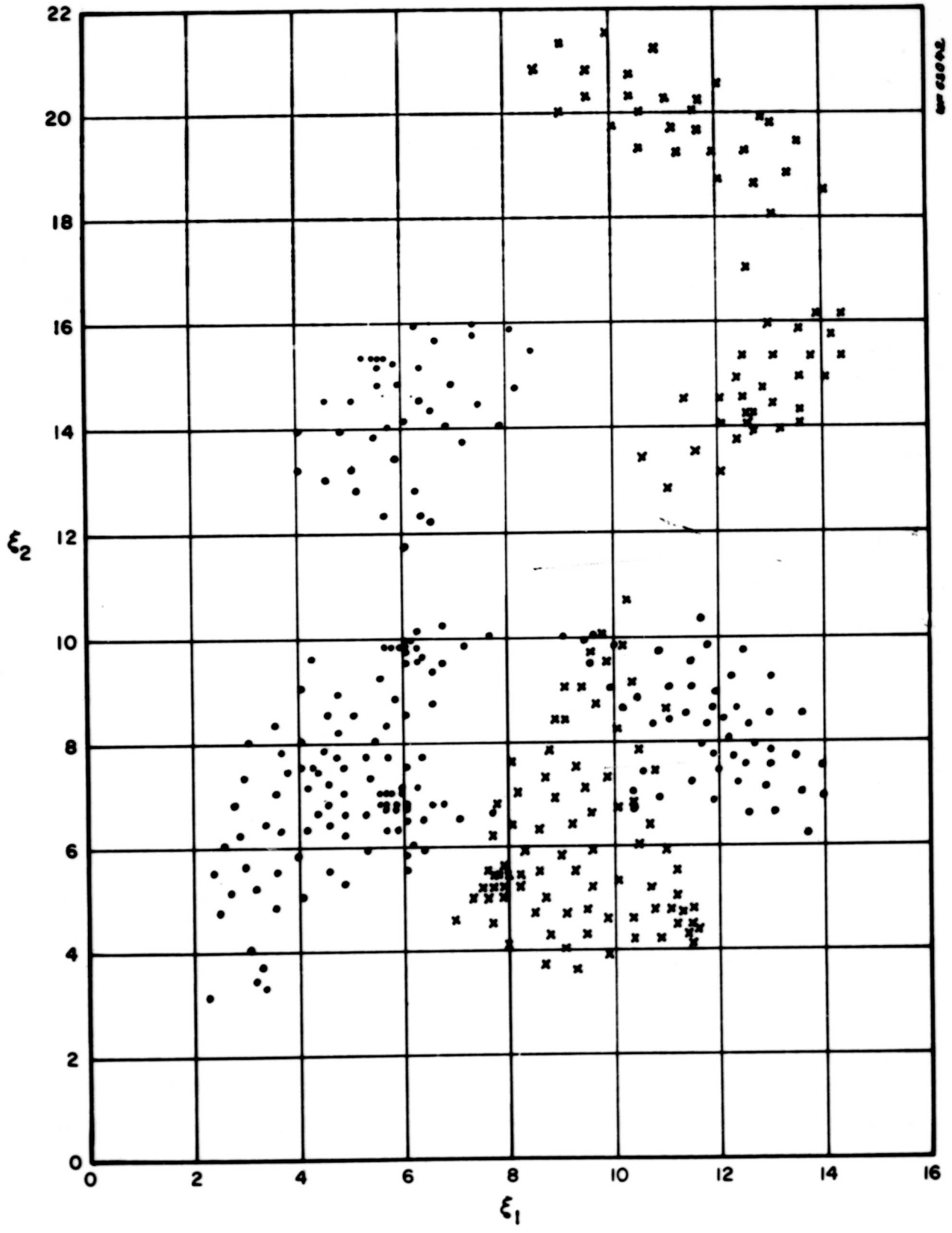


Fig. C-1 Bivariate Data Representing Two Classes to be Separated by Machine.

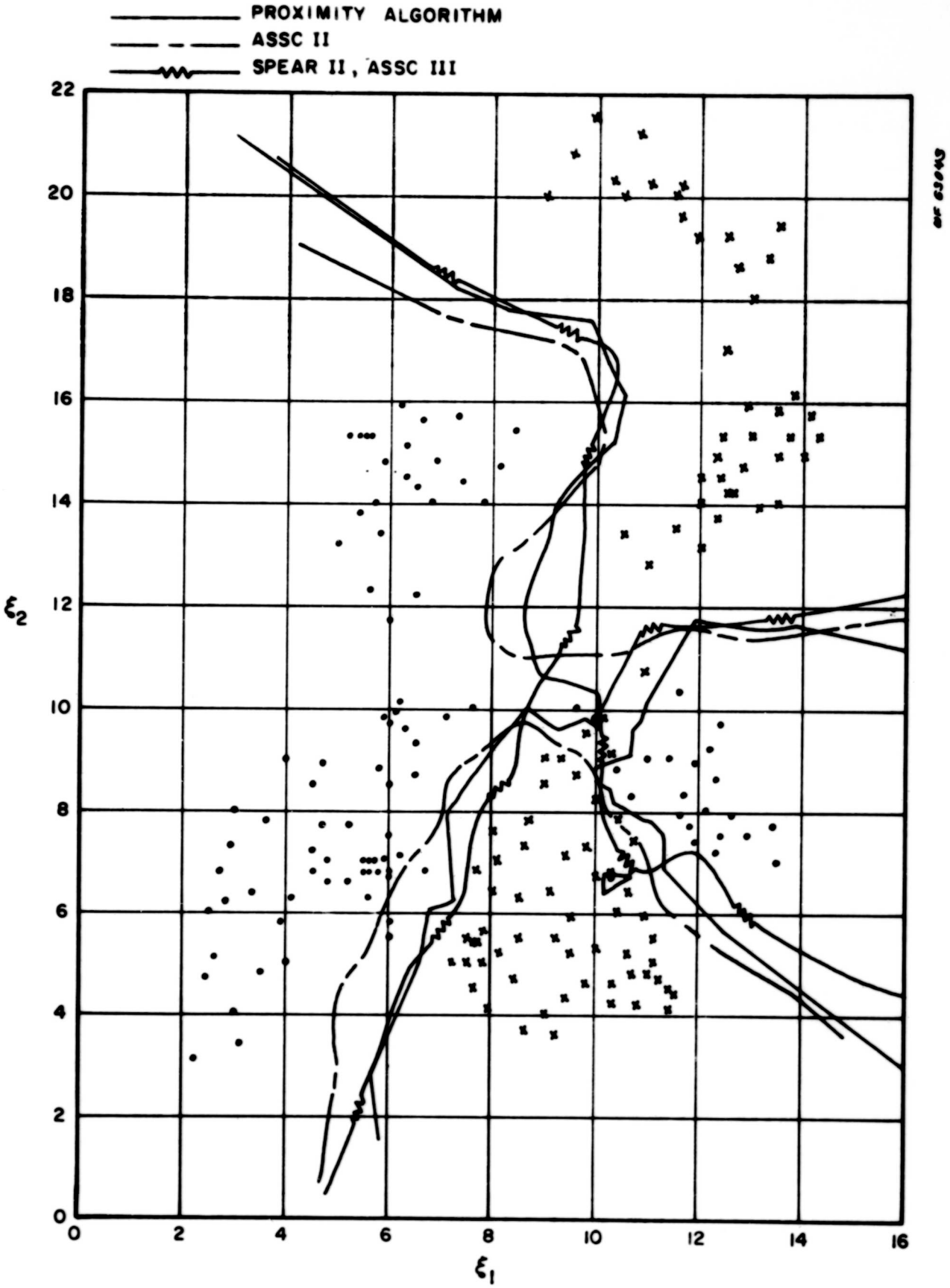


Fig. C-2 200 Data Points Used for Machine "Learning" with Class Decision Boundary Generated by the Proximity Algorithm, ASSC II and SPEAR II. . .

determining pairs might be considered as "wasted". For example, the data point (from class 1) at (4, 9) has no effect on the boundary and really need not be stored.

The ASSC II program generates a cell structure for each class in which the cell locations are determined from the data but the cell sizes and shapes are prechosen and remain fixed throughout the "learning" procedure (except for overlapping of the cell defining circles). The cell structure generated by ASSC II for one particular ordering of the data in Figure C-1 is shown in Figures C-3 and C-4. The number in each cell is (in percent) the estimate of the probability that a point from the generating class will lie in that cell; i.e., the fraction of the "learning" sample that fell in the cell. Note that there is sufficient separation between these two classes so that only a few cells of each class overlap cells of the other class. There are 32 cells in Figure C-3 and 31 in Figure C-4.

The cells generated for each class by SPEAR II depend in size, shape and in position on the "learning" samples. Furthermore, individual cells are kept isolated during the early stages of the "learning" procedure. The cell structure generated by SPEAR from the data of Figure C-2 is shown in Figures C-5 and C-6. The numbers of cells for the first and second classes are 10 and 8, respectively. This represents a reduction of 69 percent for class 1 and 74 percent for class 2 from the number of cells generated by ASSC II. The numbers at the cell centers in Figures C-5 and C-6 are the percentage of class members that fall in each cell. Note that while the true shape of the class distributions are not portrayed as faithfully by SPEAR than by ASSC II, the decision boundaries are substantially the same.

A total of 15 or 83 percent of the cells changed in size and shape from the initial setting. Since the control parameters were chosen so that the probability of cell growth was appreciably higher than one-half, this figure is consistent with theoretical predictions. In this experiment, the cell radius in the coordinate

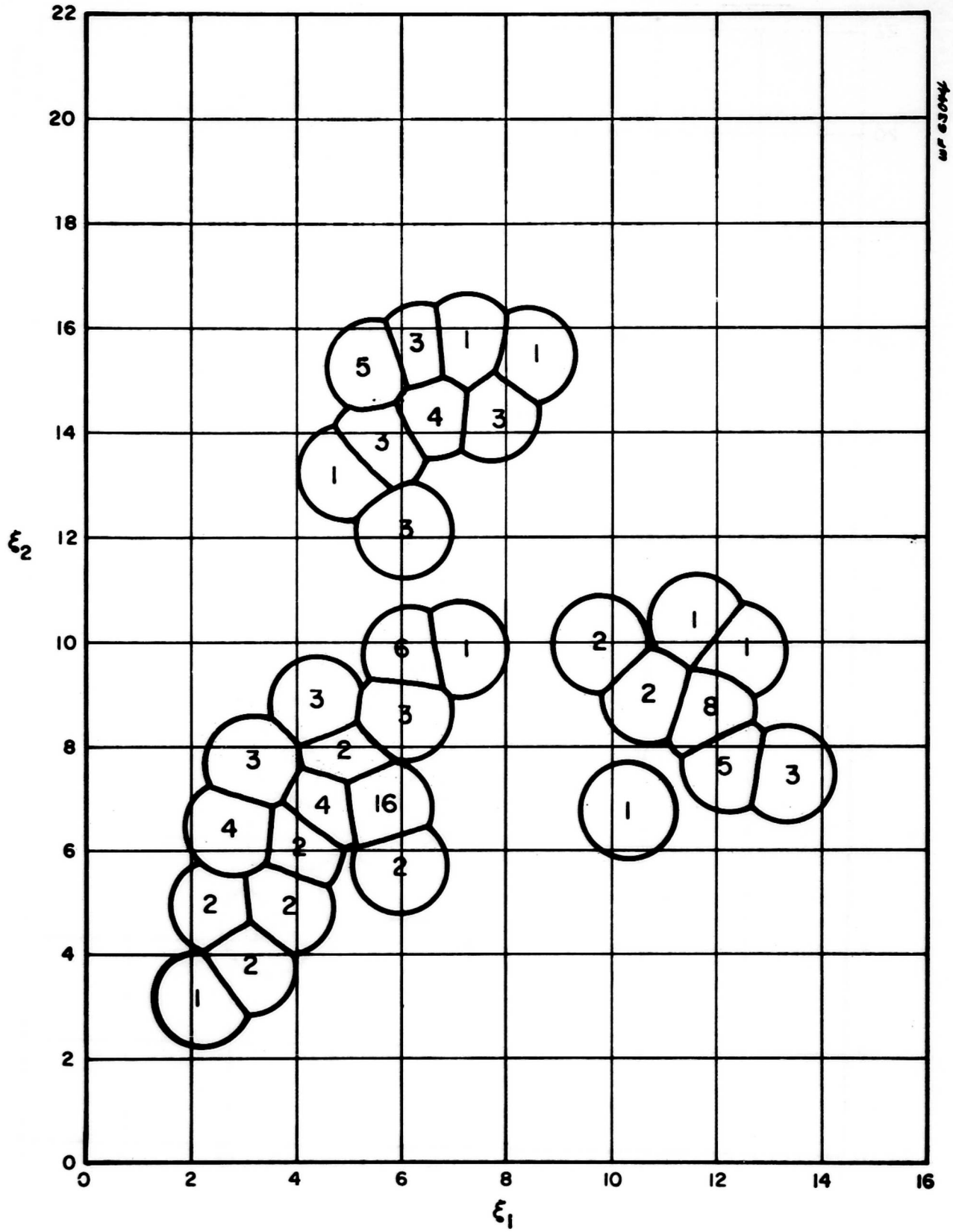


Fig. C-3 Cell Structure Generated by ASSC II for Class 1.

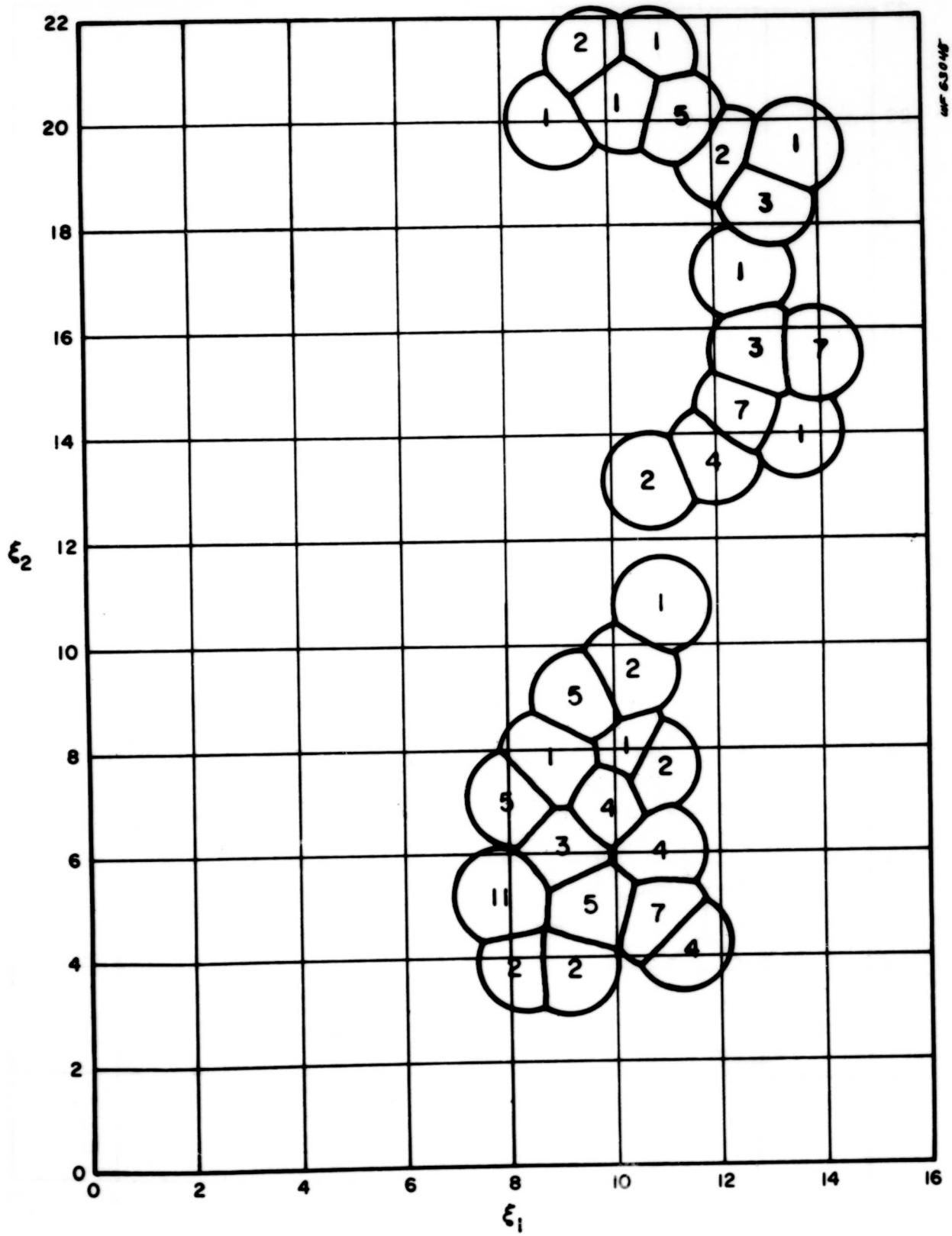


Fig. C-4 Cell Structure Generated by ASSC II for Class 2

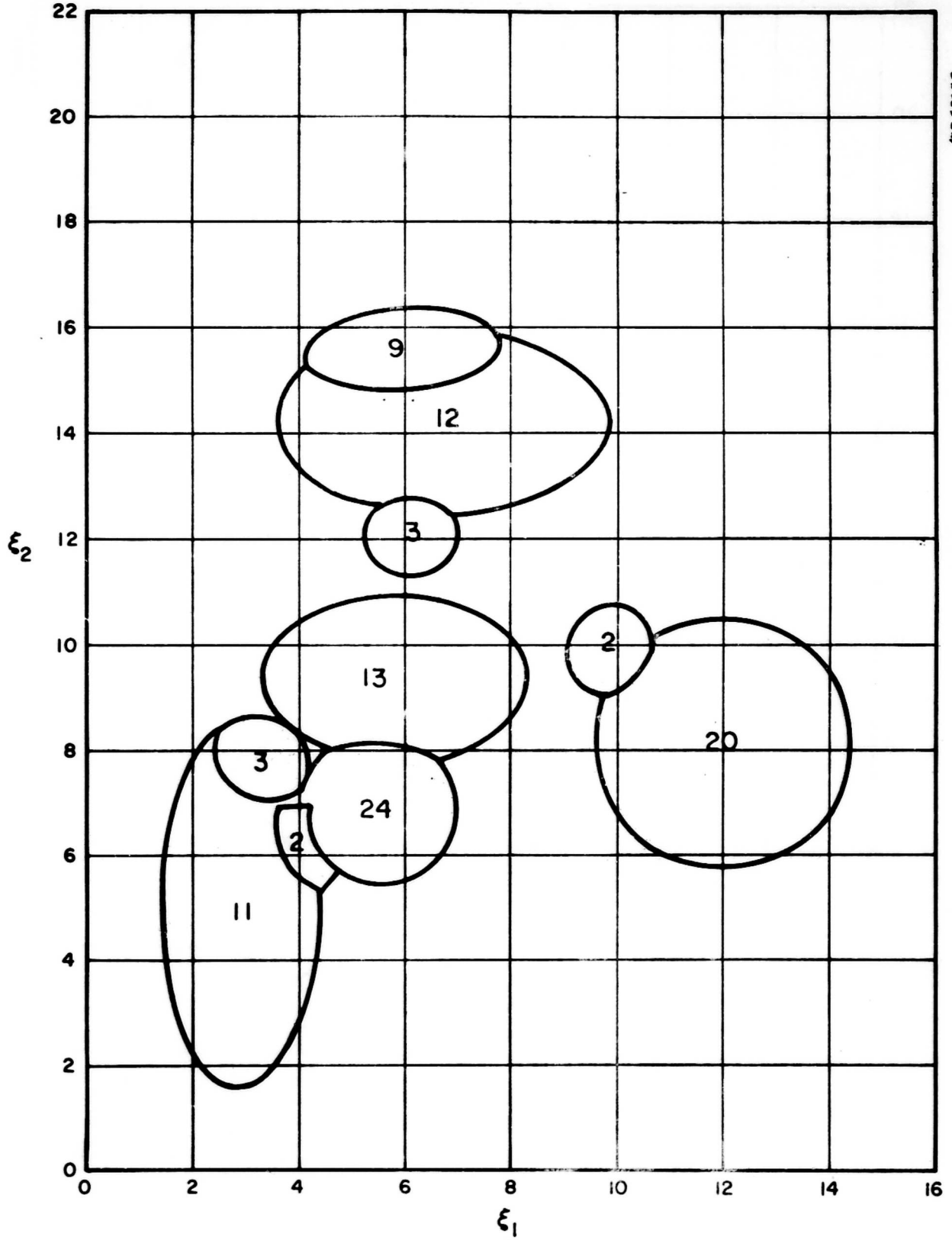


Fig. C-5 Cell Structure Generated by SPEAR II for Class 1

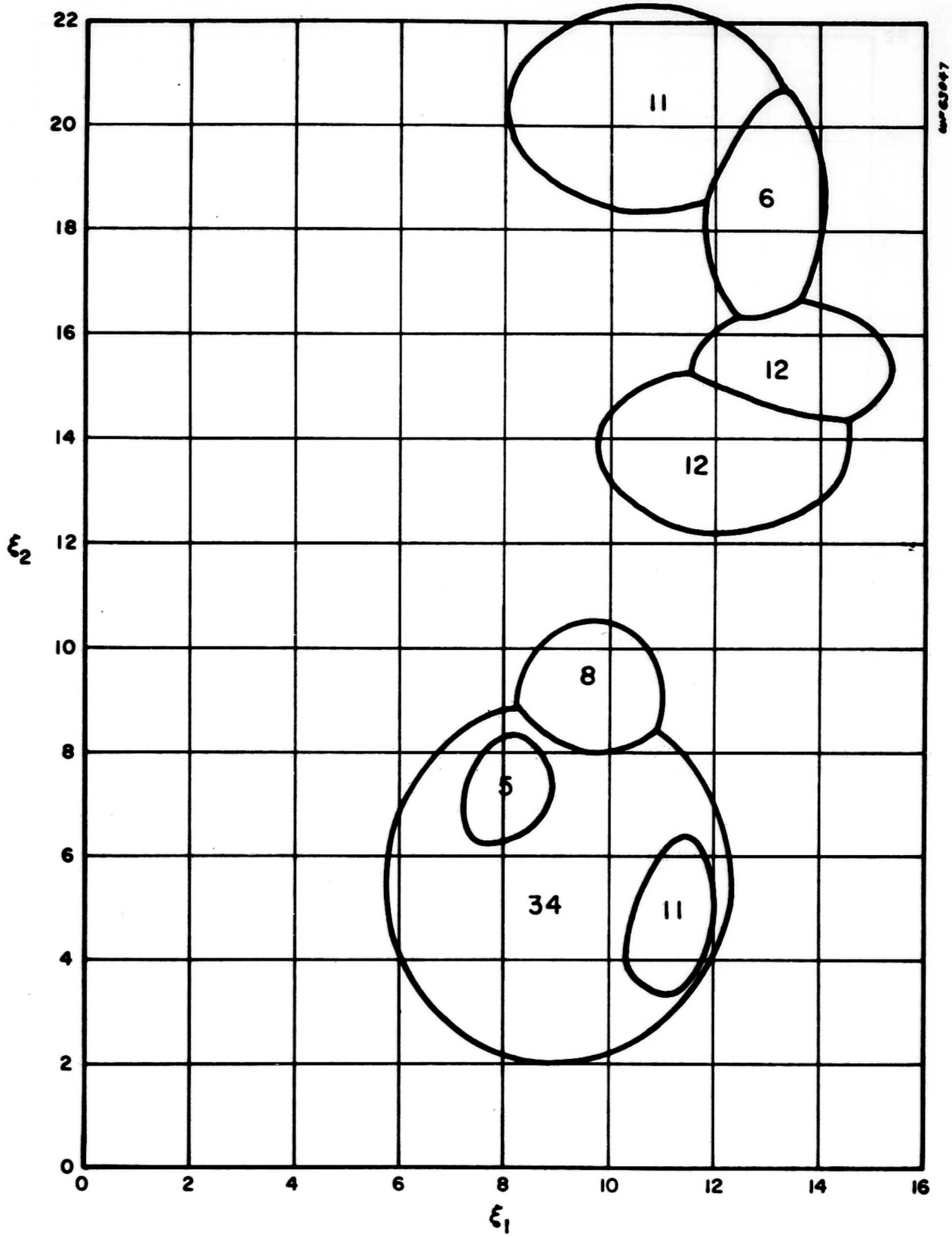


Fig. C-6 Cell Structure Generated by SPEAR II for Class

direction is three times the sample (experimental) standard deviation of the k-th coordinate of points falling in the cell. (See Appendix II.)

After performing the "learning" operations, the remaining 100 points from class 1 and 50 points from class 2 were tested against the three decision rules. The results of this test are shown in Figure C-7 which shows the test sample from each class and the Proximity Algorithm decision boundary (as in Figure C-2). Three examples from the first class are on the wrong side (or class 2 side) of the decision boundary. Therefore, the probability of deciding with the Proximity Algorithm in favor of class 2 when in fact the example is in class 1 is approximately $\hat{Pr}\{2|1\} \cong .03$. (The circumflex over the probability symbol indicates that the quantity is an estimate, i.e., $\hat{Pr}\{i|j\}$ is read the estimated probability of deciding in favor of the i-th class given that the example is from the j-th class .) The results of testing each of the three schemes are summarized in Table C-1. No significance should be attached to the relative magnitudes of these error probabilities since the sample sizes used were small.

Table C-1. Error Probabilities For Three Decision Rules

Learning and Decision Rule	Pr {2|1}	Pr {1|2}
Proximity Algorithm	.03	.10
ASSC II	.08	.06
SPEAR II - ASSC III	.07	.06

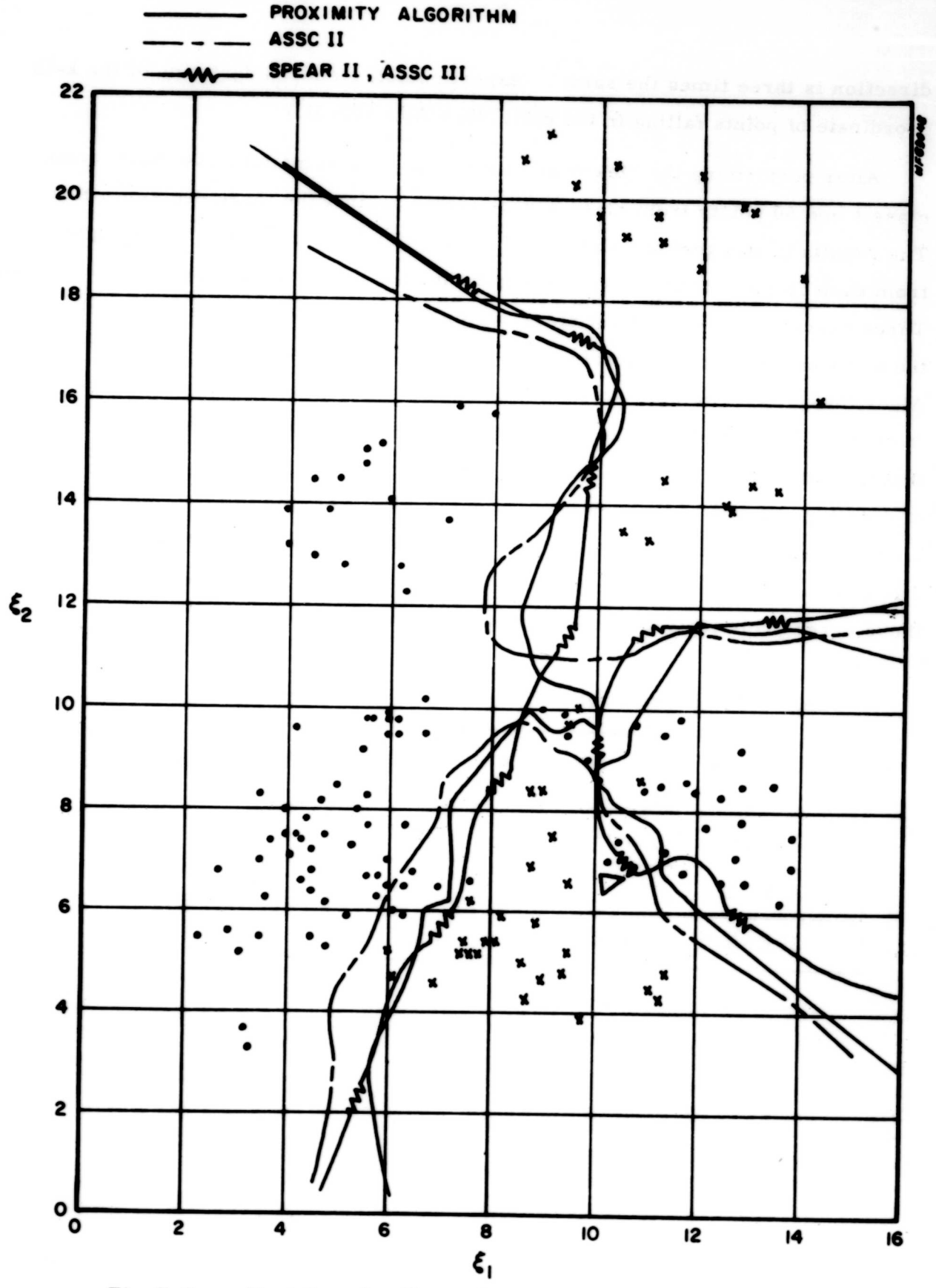


Fig. C-7 Test Samples From Two Classes With Three Decision Boundaries Formed From Independent Samples .

APPENDIX IV

LIST OF COMPUTER PROGRAMS RELATED TO PROBABILITY DENSITY ESTIMATION

1. Name: Proximity Algorithm

Description: This program finds that vector in a set of stored vectors which is nearest (by Euclidean distance) to an input test vector. The known classification of the nearest stored vector determines the classification decision associated with the input.

Computer: Recomp II

2. Name: ASSC II (Adaptive Sample Set Constructor)

Description: This program does "learning", recognition, and "unsupervised learning" using the "typical sample" technique. Cells for p.d.f. estimation of fixed size and shape are generated and updated in location by the known vectors learning samples. In recognition the p.d.f. is represented as a sum of Gaussian functions, each fitted to an individual cell.

Computer: IBM 7090

3. Name: ASSC III

Description: This subroutine, in conjunction with program GSREC II, utilizes the results of either ASSC II or SPEAR to perform recognition by representing the class p.d.f. by a quantized

Gaussian function fitted to the nearest individual cell (cells not restricted to be of equal size and shape).

Computer: IBM 7090

4. Name: ATT

Description: This program transforms the results of the ASSC II program "learning" into a form acceptable to the ASSC III program.

Computer: IBM 7090

5. Name: GSREC

Description: This program performs recognition on single sample vectors as well as on a sequence of vectors from the same class, assuming them to be independent.

Computer: IBM 7090

6. Name: SPEAR

Description: Generalized form of ASSC II learning program. Cells are updated by the data in size, shape, and in location. This program includes a "cell collapsing" subroutine in which redundancy is eliminated by combining cells whenever possible.

Computer: IBM 7090

7. Name: RATIO

Description: Two arrays of single sample vector recognition probabilities are examined by searching sequentially with a predetermined "window".

Computer: Recomp II.

8. Name: GENSPR

Description: This program generates multivariate data of known statistics for testing the class p.d.f. estimation programs. The possible form of the data distribution is of a union of Gaussian modes with the mode a priori probabilities, means, and variances as selectable parameters.

The computer program GENSPR was written in order to divorce the problem of testing the various p.d.f. estimation techniques from the problem of collecting a sufficient body of representative data on a set of physical classes. With GENSPR random vectors (up to 16 dimensional) are generated with specified statistics. The form of the distribution is that of a sum of Gaussian modes, i.e., the p.d.f. $q(\underset{\sim}{x})$ is given by

$$q(\underset{\sim}{x}) = \sum_{r=1}^{M} Q_r \phi_r(\underset{\sim}{x}),$$

where $M \leq 10$ is the number of modes, Q_r is the (prechosen) a priori probability of $\underset{\sim}{x}$ be associated with the r-th mode, and $\phi_r(\underset{\sim}{x})$ is an N-variate Gaussian p.d.f. ($N \leq 16$) with means and variances selectable by the experimenter (covariances equal zero).

$$\phi_r(\underset{\sim}{x}) = \prod_{s=1}^{N} \frac{1}{(2\pi)^{N/2}\sigma_{rs}} \exp\left[-\frac{1}{2}\left(\frac{x_s - m_{rs}}{\sigma_{rs}}\right)^2\right],$$

with σ_{rs} = s-th coordinate standard deviation of the r-th mode, and m_{rs} = s-th coordinate mean of the r-th mode. The selectable parameters for using GENSPR are, therefore:

N, M, Q_r, m_{rs}, and σ_{rs}. A flow chart of GENSPR is shown on the next page.

Computer: IBM 7090

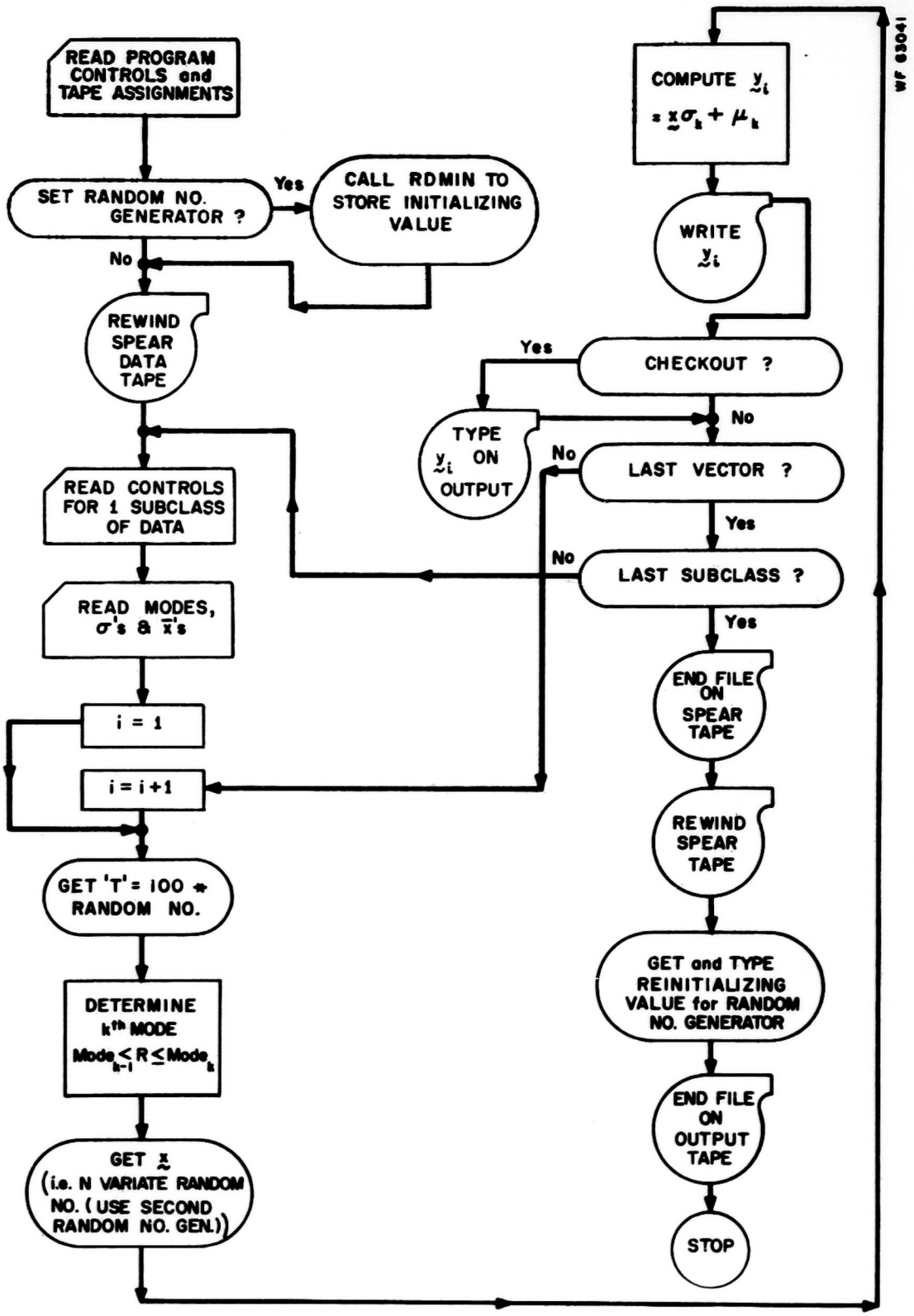


Fig. D-1 Flow Chart of GENSPR

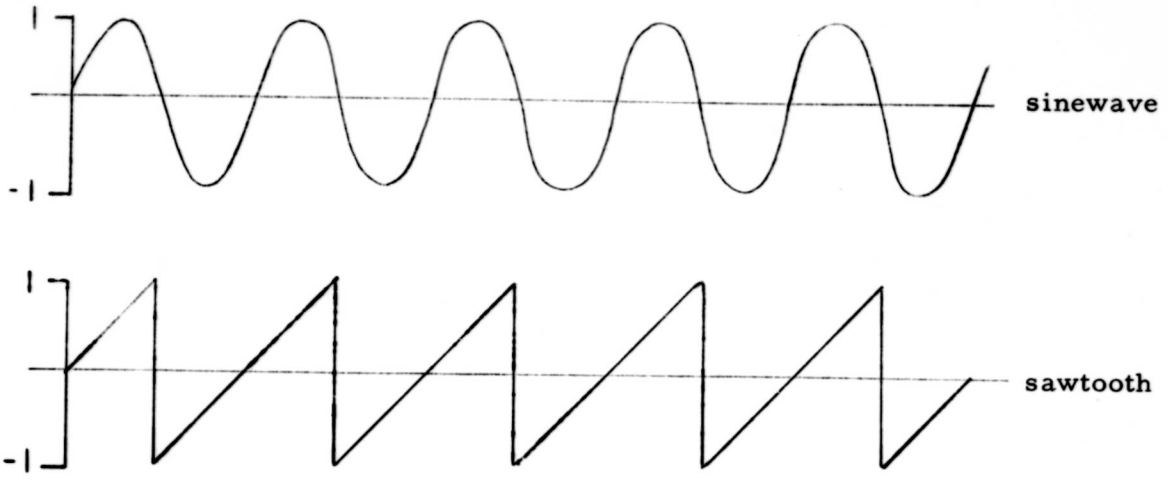
APPENDIX V

A POSTERIORI IDENTIFICATION OF HIGH QUALITY BINARY CLASSIFICATION DECISIONS

In Section 4.2, two methods of identifying sample events for which a high quality decision could be rendered, were outlined. In the second method the quality of a classification decision is equated to the degree to which the classification can be expected to correspond to the optimum decision which could have been rendered if the class statistics were fully known. With this method it is possible to partition the observation space into high and low quality decision regions. The utility of such a capability lies in the fact that different cell structures (for constructing estimates of probability density functions) can be evaluated. A good cell structure can be identified as one which produces a large, high-quality decision region.

With a view toward the establishment of bounds within which this indicator can be relied upon to provide a satisfactory indication of high quality decision regions, its application to the binary classification problem has been examined. Although couched in terms of the artificial problem of automatically distinquishing between two waveforms, based on a finite number of samples taken at randomly spaced intervals, the following analysis applies to any binary classification problem with reservations as indicated.

As an illustration of the binary classification problem, consider the problem of distinguishing between periodic sawtooth and sinewave waveforms with identical periods and amplitude ranges, as indicated in the diagram.



Suppose that the only information available for learning how to distinguish between these waveforms are N_1 samples of a sinewave taken at randomly selected instants and N_2 samples of a sawtooth waveform taken at randomly selected instants. These $N_1 + N_2$ samples comprise the learning set. Suppose further that classification decisions must be rendered solely on the basis of m new samples selected at random time instants from either the sinewave or sawtooth. These m samples comprise the recognition set.

Since we have assumed that the two waveforms are known only in terms of a set of sample values, a procedure for constructing a decision mechanism from these samples must be selected. The procedure investigated here consists of calculating a likelihood ratio using histograms (constructed from the $N_1 + N_2$ learning samples) as estimates of the probability density functions of the two waveform amplitudes.

The classification method under study requires that the classification decision be rendered as follows. Let c denote the number of cells in the cell structure, and let the cells be labeled by the numbers 1 to c. Further, let $\{k_i\} = \{k_1, \ldots, k_m\}$ denote the cells into which the m samples of the recognition set fall.

The recognition set is classified according to the rule:

If $\hat{L} > 1$, decide in favor of sinewave

If $\hat{L} < 1$, decide in favor of sawtooth (E-1)

where $\hat{L}$ is calculated by*

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

and

$\hat{P}_A(k_i)$ = an estimate of the probability that the i-th sample in the test set would fall into the k_i-th cell, assuming that the sample is drawn from the sinewave.

$\hat{P}_B(k_i)$ = the corresponding quantity assuming that the sample is drawn from the sawtooth.

and finally,

P_A = the a priori probability that the recognition set is drawn from the sinewave

P_B = the corresponding quantity assuming that the recognition set is drawn from the sawtooth waveform.

The caret symbol is used to indicate that estimates, rather than "true" quantities are used in calculating the likelihood ratio. The specific forms of the estimates implied by the use of histograms in lieu of probability density function are

$$\hat{P}_A(k) = \frac{a_k}{N_1}$$

*This is equivalent to Equation (2.2).

and

$$\hat{P}_B(k) = \frac{b_k}{N_2} \qquad \text{(E-2)}$$

where a_k is the number of sinewave samples in the learning set which fall into the k-th cell, and b_k is the number of sawtooth waveform samples in the learning set which fall into the k-th cell. For simplicity, we shall assume hereafter that each waveform is known to be equally likely to occur; i.e.,

$$P_A = P_B = \frac{1}{2} .$$

For a given learning set, we would like to be able to specify those recognition sets for which we are confident that an optimum classification decision will be rendered. The histogram cell structure effectively quantizes the space consisting of all possible recognition sets into c^m points. Since each recognition set corresponds to the m-dimensional vector $\{k_1, k_2, \ldots, k_m\}$, and each coordinate, k_i, can take on c values, the histogram cell structure effectively quantizes the space consisting of all possible recognition sets into c^m points. Therefore, the problem of specifying those recognition sets for which we are confident that an optimum decision will be rendered, becomes one of specifying a subset of these c^m points. Let H denote the subset of these c^m recognition points at which we are willing to state that a high quality decision will be rendered. The proposed method of determing H is to let H consist of those test points $\{k_i\}$ for which $Q\{k_i\} > \eta$,

where

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

= an estimate of the probability that the estimated likelihood ratio, $\hat{L}$, will produce the same classification as the true likelihood ratio, L, would produce if it were known.

The "true" probability referred to in the definition of Q is that which can be calculated knowing the statistics of the classes, the method of estimating likelihood functions, the size of the learning set (N_1 and N_2), and the cell structure (c). The estimate, Q, is determined as follows. Before the learning set is selected, the true probability, Q', of making the same classification with $\hat{L}$ as with L at a given test point may be written as a function of the quantities: $P_A(k)$, $P_B(k)$, $k = 1, 2, \ldots, c$; N_1; N_2; and the recognition point, $\{k_1, k_2, \ldots, k_m\}$.

Specifically,

$$Q' = \Pr\{(\hat{L}-1)(L-1) > 0\}$$

$$= \Pr\left\{\left[\left(\frac{N_2}{N_1}\right)^m \prod_{i=1}^{m} \frac{a_{k_i}}{b_{k_i}} - 1\right]\left[\prod_{i=1}^{m} \frac{P_A(k_i)}{P_B(k_i)} - 1\right] > 0\right\} \qquad \text{(E-4)}$$

$$= \sum_{j}^{N_1}{}' \sum_{r}^{N_2}{}' \prod_{\ell=1}^{m} \binom{N_1}{j}\binom{N_2}{r} P_A(k_\ell)^{j_\ell} P_B(k_\ell)^{r_\ell} \left[1 - P_A(k_\ell)\right]^{N_1 - j_\ell} \left[1 - P_B(k_\ell)\right]^{N_2 - r_\ell}$$

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

$$= Q'\left[P_A(k_1), P_A(k_2), \ldots, P_A(k_m), P_B(k_1), \ldots, P_B(k_m), k_1, \ldots, k_m, N_1, N_2\right]$$

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 $\operatorname{sgn} x = \dfrac{x}{|x|}$ if $x \neq 0$

$= 0$ if $x = 0$

For terms in the expression for Q' involving a $j_i = 0$ and $r_\ell = 0$, we define $\dfrac{j_i}{r_\ell} = \dfrac{0}{0} \equiv 1$.

The estimate, Q, is obtained by substituting $\hat{P}_A(k_i) = \dfrac{a_{k_i}}{N_1}$ for $P_A(k_i)$, and $\hat{P}_B(k_i) = \dfrac{b_{k_i}}{N_2}$ for $P_B(k_i)$, in the expression for Q'. Thus,

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

$$\cdot\frac{1}{2}\left\{\operatorname{sgn}\left[\left\{\left(\frac{N_2}{N_1}\right)^m \prod_{i=1}^{m}\frac{j_i}{r_i}-1\right\}\left\{\left(\frac{N_2}{N_1}\right)^m \prod_{i=1}^{m}\frac{a_{k_i}}{b_{k_i}}-1\right\}\right]+1\right\} \qquad \text{(E-5)}$$

Note that the estimate Q is a function not only of the recognition point $\{k_1, k_2, \ldots, k_m\}$, but also the learning set, through the $\left\{a_{k_i}\right\}$ and $\left\{b_{k_i}\right\}$.

Having set up a basis for estimating the likelihood that a given recognition point will be classified in an optimum manner, several questions arise. Perhaps the most important questions are:

(1) Given a recognition point, for what values of N_1 and N_2 can Q be expected to provide a close estimate of Q' (not knowing the learning set)?

(2) Given values of N_1, N_2, and c, for prescribed classes with known statistics what is the minimum (true) probability of rendering an optimum decision in any cell?

(3) For a given value of η, what values of N_1, N_2, c, and m will provide values of P_Q greater than P_0, where

$$P_Q \equiv \frac{1}{2} \sum_{\text{All } \{k_\ell\} \epsilon H} \left\{ \prod_{i=1}^{m} \hat{P}_A(k_i) + \prod_{i=1}^{m} \hat{P}_B(k_i) \right\}$$

$$= \frac{1}{2} \frac{1}{(N_1 N_2)^m} \sum_{\text{All } \{k_\ell\} \epsilon H} \left\{ N_2{}^m \prod_{i=1}^{m} a_{k_i} + N_1{}^m \prod_{i=1}^{m} b_{k_i} \right\} \qquad \text{(E-6)}$$

The remainder of this appendix is devoted to the first and second of these questions. To obtain explicit answers to these questions, it will be convenient to restrict our attention to recognition sets consisting of a single sample, i.e., m = 1. Introducing this restriction into the equations for Q and Q', we obtain

$$Q'_k = Q'_k\left(N_1,\ N_2,\ P_A(k),\ P_B(k)\right)$$

$$= \sum_{j=0}^{N_1} \sum_{r=0}^{N_2} \binom{N_1}{j} \binom{N_2}{r} P_A(k)^j P_B(k)^r [1 - P_A(k)]^{N_1 - j} [1 - P_B(k)]^{N_2 - r}$$

$$\cdot \frac{1}{2} \left\{ \operatorname{sgn} \left[\left(\frac{N_2 j}{N_1 r} - 1 \right) \left(\frac{P_A(k)}{P_B(k)} - 1 \right) \right] + 1 \right\} \qquad \text{(E-7)}$$

and

$$Q_k = Q_k\left(N_1,\ N_2,\ a_k,\ b_k\right)$$

$$= Q'_k\left(N_1,\ N_2,\ \frac{a_k}{N_1},\ \frac{b_k}{N_2}\right)$$

$$= \sum_{j=0}^{N_1} \sum_{r=0}^{N_2} \binom{N_1}{j}\binom{N_2}{r}\left(\frac{a_k}{N_1}\right)^j \left(\frac{b_k}{N_2}\right)^r \left(1 - \frac{a_k}{N_1}\right)^{N_1 - j} \left(1 - \frac{b_k}{N_2}\right)^{N_2 - r}$$

$$\cdot\left\{\frac{1}{2}\ \mathrm{sgn}\left[\left(\frac{N_2 j}{N_1 r} - 1\right)\left(\frac{N_2 a_k}{N_1 b_k} - 1\right)\right] + 1\right\} \qquad \text{(E-8)}$$

In this case, the k-th recognition point corresponds to the single test sample falling into the k-th cell.

To answer the first question posed above, we wish to evaluate the rms difference between Q_k and Q'_k, where the averaging is performed over all learning sets generated by known class statistics; i.e.,

$$E_k^2 = \overline{(Q_k - Q'_k)^2} = \overline{Q_k^2} - 2\,\overline{Q_k Q'_k} + \overline{(Q'_k)^2} \qquad \text{(E-9)}$$

The random variables involved in the averages indicated in Eq. (E-9) are a_k and b_k. The probability density functions for these variables are

$$p_A(a_k) = \sum_{j=0}^{N_1} \binom{N_1}{j}\left[P_A(k)\right]^j \left[1 - P_A(k)\right]^{N_1 - j} \delta(a_k - j)$$

$$p_B(b_k) = \sum_{j=0}^{N_2} \binom{N_2}{j} \left[P_B(k)\right]^j \left[1 - P_B(k)\right]^{N_2 - j} \delta(b_k - j) \qquad \text{(E-10)}$$

where $\delta(x)$ is the Dirac delta function.

Using Eq. (E-7) — (E-10), the mean squared difference between Q_k and Q'_k may be written:

$$E_k^2 = \int_{-\infty}^{\infty} da_k \int_{-\infty}^{\infty} db_k p_A(a_k) p_B(b_k) \left[Q'_k\left(N_1, N_2, \frac{a_k}{N_1}, \frac{b_k}{N_2}\right) - Q'_k\left(N_1, N_2, P_A(k), P_B(k)\right)\right]^2$$

$$= \sum_{r_1=0}^{N_1} \sum_{r_2=0}^{N_2} Q'_k\left(N_1, N_2, \frac{r_1}{N_1}, \frac{r_2}{N_2}\right)\left[Q'_k\left(N_1, N_2, \frac{r_1}{N_1}, \frac{r_2}{N_2}\right) - 2Q'\left(N_1, N_2, P_A(k), P_B(\kappa)\right)\right] \cdot F_2\left(N_1, N_2; P_A(k), P_B(k); r_1, r_2\right) - \left[Q_k\left(N_1, N_2, P_A(k), P_B(k)\right)\right]^2 \qquad \text{(E-11)}$$

where

$$F_M(\mu_1, \mu_2, \ldots, \mu_M; Y_1, Y_2, \ldots, Y_M; \tau_1, \tau_2, \ldots, \tau_M) \equiv \prod_{j=1}^{M} \binom{\mu_j}{\tau_j} \left(Y_j\right)^{\tau_j} \left(1 - Y_j\right)^{\mu_j - \tau_j} \qquad \text{(E-12)}$$

The deviation E_k is a function of the size of the learning set in terms of N_1, and N_2, and a function of the cell structure and individual class statistics in terms of $P_A(k)$ and $P_B(k)$, $k = 1, 2, \ldots, c$. To obtain a quantitative indication of how E_k depends on these quantities, Eq. (E-11) has been programmed for evaluation on the Recomp II computer. As in illustration, the sinewave and sawtooth waveforms were taken to be the two classes, and a uniform cell structure was assumed. In this case, the k-th cell consists of waveform sample values in the interval $\left(-1 + \frac{2(k-1)}{c},\ -1 + \frac{2k}{c}\right)$, $k = 1, 2, \ldots, c$, and the true probabilities of a sample value occurring within the k-th cell are

$$P_A(k) = \frac{1}{c}$$

$$P_B(k) = \frac{1}{\pi}\left[\sin^{-1}\left(1 - \frac{2(k-1)}{c}\right) - \sin^{-1}\left(1 - \frac{2k}{c}\right)\right], \qquad \text{(E-13)}$$

for $k = 1, 2, \ldots, c$.

The curves in Figure E-1 indicate the average value, E, of E_k, that is, $E \equiv \frac{1}{c}\sum_{k=1}^{c} E_k$, as a function of the size of the learning set, $N \equiv N_1 = N_2$, for $c = 2$ and $c = 4$. For these two classes it appears that increasing resolution of the cell structure serves to <u>increase</u> the accuracy of the quality indicator, for a given learning set size. Below some value of N, however (probably near N = C), this effect will not be sustained. In fact, we conjecture that for a given value of N, there may be an optimum value of c which minimizes E. Further calculations will be performed to study this effect.

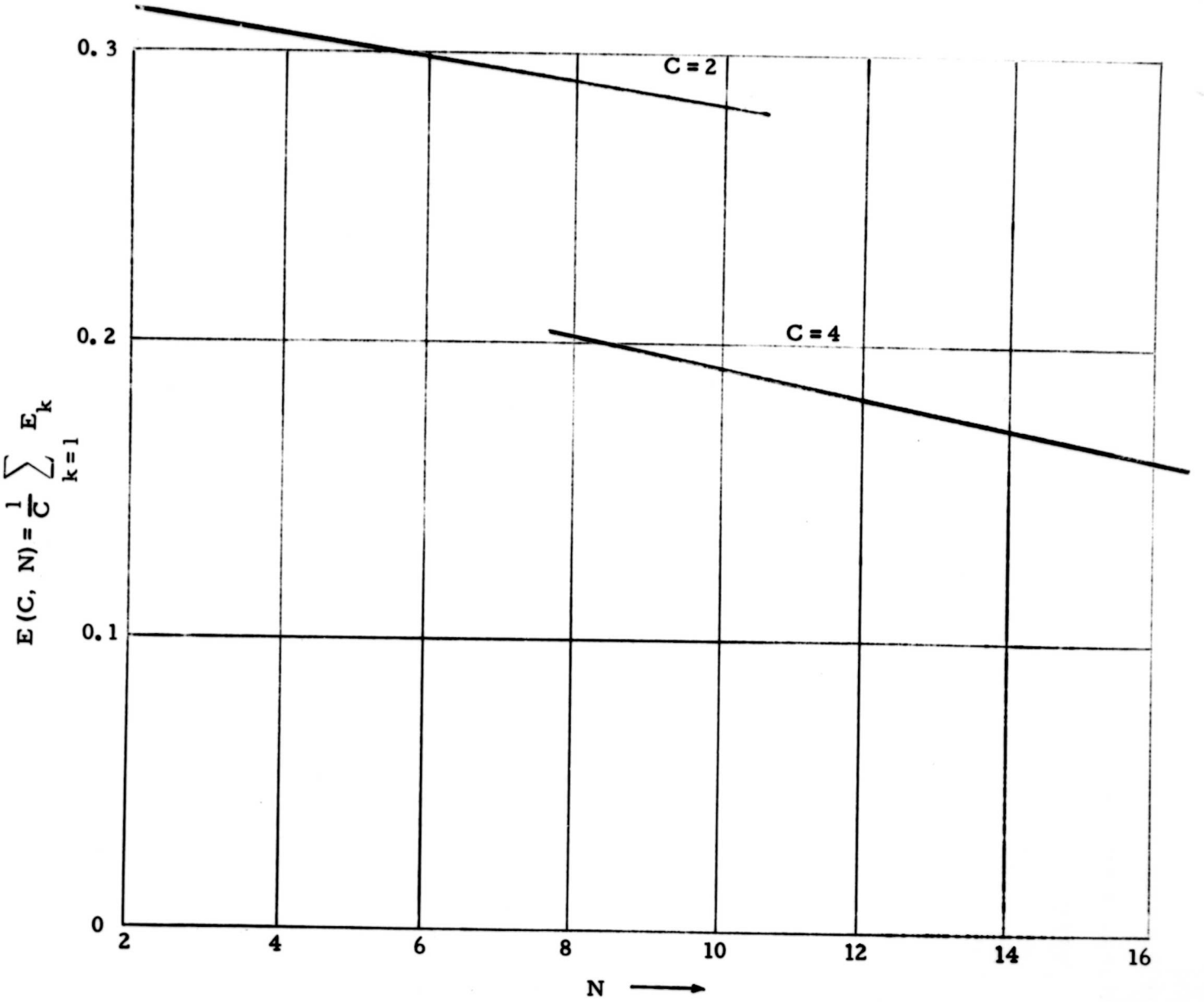


Figure E-1. Average RMS Difference Between Q_k and Q^1_k as a Function of N, for the Sinewave and Sawtooth Waveform Classes.

At this point it is convenient for analytical evaluation to introduce the Gaussian approximation to the binomial distribution. This may* produce a close approximation for Q' if $\min\left\{N_1 P_A(k),\ N_2 P_B(k)\right\} \geq 5$ (say), and a close approximation for Q if $\min\left\{a_k,\ b_k\right\} \geq 5$. The former condition will be satisfied if $\min(N_1,\ N_2) \geq 10\ c$. Assuming that both conditions are satisfied, it is easily shown that the Gaussian approximation produces expressions for Q' and Q as follows:

$$Q'_k \tilde{=} \Phi\left(\frac{\sqrt{N_1 N_2}\left|P_A(k) - P_B(k)\right|}{\sqrt{N_1 P_B(k)[1 - P_B(k)] + N_2 P_A(k)[1 - P_A(k)]}}\right)$$

and

$$Q_k \tilde{=} \Phi\left(\frac{\left|\sqrt{\frac{N_2}{N_1}}\, a_k - \sqrt{\frac{N_1}{N_2}}\, b_k\right|}{\sqrt{\frac{N_1}{N_2}\, b_k\left(1 - \frac{b_k}{N_2}\right) + \frac{N_2}{N_1}\, a_k\left(1 - \frac{a_k}{N_1}\right)}}\right) \qquad \text{(E-14)}$$

where

$$\Phi(x) \equiv \int_{-\infty}^{x} \phi(u),\ du,$$

*It has not been proven that the Gaussian approximation to the binomial is valid. The conditions stated simply ensure that the central term in the binomial distribution is more than 2.5 standard deviation units away from the origin.

and

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} . \qquad \text{(E-15)}$$

Now we wish to evaluate the mean squared difference between Q_k and Q'_k, where the averaging is carried out over all learning sets, for a given recognition point (cell) k. Since a_k and b_k are binomially distributed, we may write for the mean value of Q,

$$\overline{Q}_k = \sum_{j=0}^{N_1} \sum_{r=0}^{N_2} \binom{N_1}{j} \binom{N_2}{r} P_A(k)^j P_B(k)^r \left[1 - P_A(k)\right]^{N_1 - j} \left[1 - P_B(k)\right]^{N_2 - r}$$

$$\cdot \Phi\left(\frac{\left| \sqrt{\frac{N_2}{N_1}}\, j - \sqrt{\frac{N_1}{N_2}}\, r \right|}{\sqrt{\frac{N_1}{N_2}\, r\left(1 - \frac{r}{N_2}\right) + \frac{N_2}{N_1}\, j\left(1 - \frac{j}{N_1}\right)}} \right) \qquad \text{(E-16)}$$

Again we make use of the Gaussian approximation to the binomial distribution. This may* be an accurate approximation if $\min\left\{N_1 P_A(k),\ N_2 P_B(k)\right\} \geq 5$, or $\min\left\{N_1,\ N_2\right\} \geq 10\ c$. With this approximation $\overline{Q}_k$ becomes

*Ibid

$$\bar{Q}_k = \int_{-\infty}^{\infty}\int dx\, dy\, \phi(x)\, \phi(y)$$

$$\Phi\left(\sqrt{N_1 N_2}\left|\frac{P_A(k)\left[1-P_A(k)\right]x}{\sqrt{N_1}} - \frac{P_B(k)\left[1-P_B(k)\right]y}{\sqrt{N_2}} + \left[P_A(k)-P_B(k)\right]\right| \Big/ \right.$$

$$\sqrt{N_2\left\{\frac{P_A(k)\left[1-P_A(k)\right]x}{\sqrt{N_1}} + P_A(k)\right\}\left\{1 - \frac{P_A(k)\left[1-P_A(k)\right]x}{\sqrt{N_1}} - P_A(k)\right\} + \cdots}$$

$$\left. \cdots + N_1\left\{\frac{P_B(k)\left[1-P_B(k)\right]y}{\sqrt{N_2}} + P_B(k)\right\}\left\{1 - \frac{P_B(k)\left[1-P_B(k)\right]y}{\sqrt{N_2}} - P_B(k)\right\}}\right) \quad \text{(E-17)}$$

For suitably large values of N_1 and N_2,* this expression can be simplified to

$$\bar{Q}_k \tilde{=} \int_{-\infty}^{\infty}\int dx\, dy\, \phi(x)\, \phi(y)\ \Phi\left(\frac{\sqrt{N_1 N_2}\left|P_A(k)-P_B(k)\right|}{\sqrt{N_2 P_A(k)\left[1-P_A(k)\right] + N_1 P_B(k)\left[1-P_B(k)\right]}}\right) \quad \text{(E-18)}$$

$$= Q'_k$$

Thus, at least for large learning sets, Q_k provides an "unbiased" estimate of Q'_k.

*These would depend on the values of $\left\{P_A(k)\right\}$ and $\left\{P_B(k)\right\}$

To complete an approximate evaluation of the mean squared error of estimation, it is necessary to obtain a comparable estimate of $\overline{Q_k^2}$ as a function of $P_A(k)$, $P_B(k)$, N_1, and N_2. Unfortunately all attempts to derive an expression for this quantity have failed; however, further effort may yield an approximate expression.

In the absence of such an expression it is only possible to conduct experiments and compare the resulting values of Q_k' and Q_k. For classification of the sinewave and sawtooth waveforms, values of Q_k' are plotted in Figures E-2, E-3, and E-4 for several values of $N = N_1 = N_2$, and $c = 4$, 10, and 50 respectively. Equation (E-14) has been used to obtain these plots and therefore they must be regarded as approximations.

The graphs in Figures E-2, E-3, and E-4 indicate (as expected) that the true probability of making an optimum classification is higher in the cells for which the differences between $P_A(k)$ and $P_B(k)$ are large, and this probability increases as the number of learning samples, $N = N_1 = N_2$ is increased. The probability of making an optimum decision remains low (as N is increased) only near the points $\left[\pm\left(1 - \frac{4}{\pi^2}\right)\right]$ at which $P_A(k)$ is approximately equal to $P_B(k)$. These figures also indicate that for these two classes (sinewave and sawtooth waveforms) the minimum number of learning samples required to produce high quality decisions for all sample events in the observation space is critically dependent on the resolution of the cell structure. For a structure consisting of 4 cells, for instance, 40 learning data samples are sufficient to assure an approximate 80 percent probability of rendering an optimum decision within each cell. For a structure of 50 cells, however, approximately 500 learning data samples are required to assure this level for all cells. Thus it appears that to be confident of rendering an optimum decision at all points in this observation space, if the number of cells is increased by a factor $F (\geq 1)$, the size of the learning set must be increased by the factor F_ν, where $\nu > 1$.

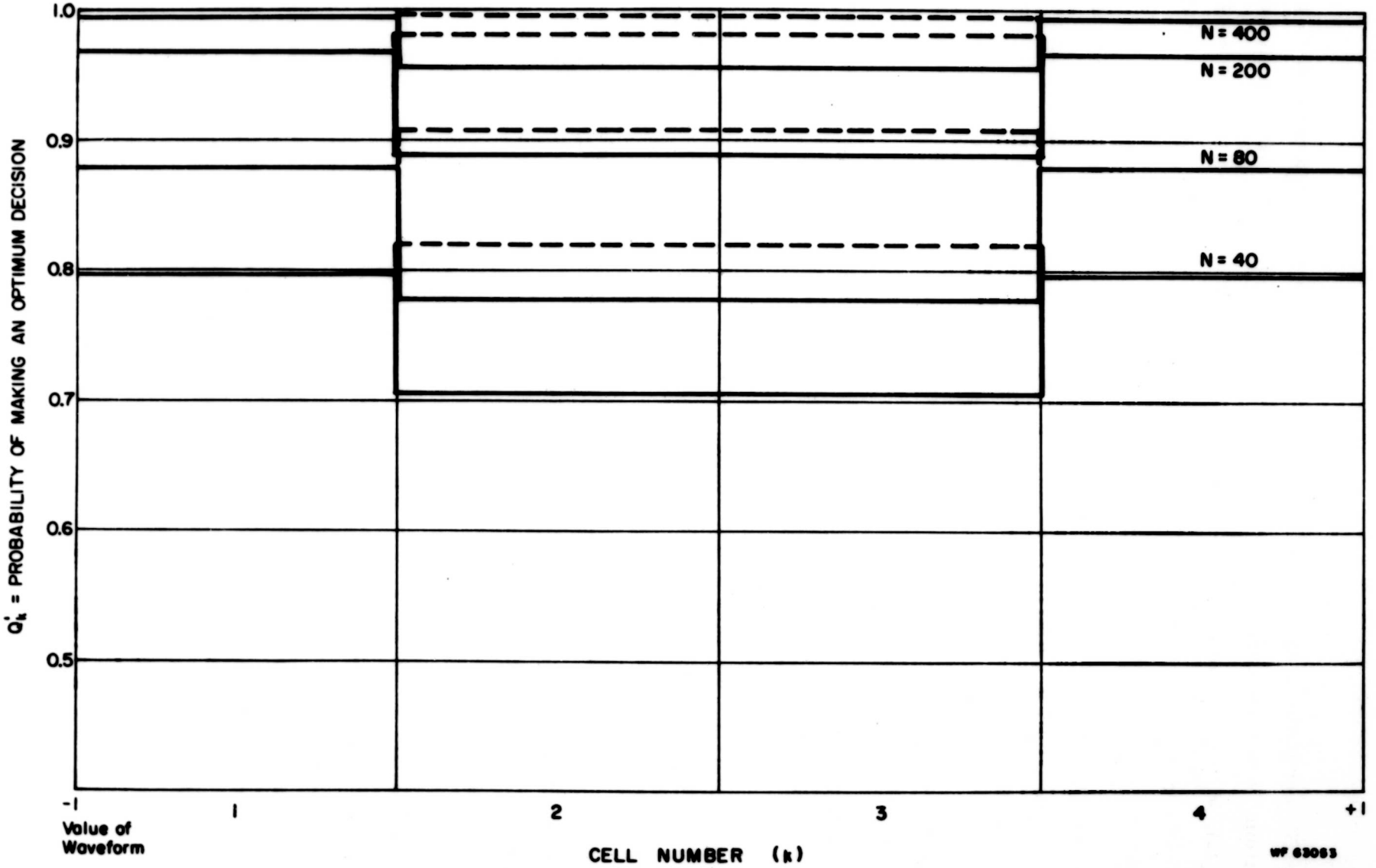


Fig. E-2 Probability of Optimum Classification of a Sinewave and a Sawtooth Waveform (C = 4)

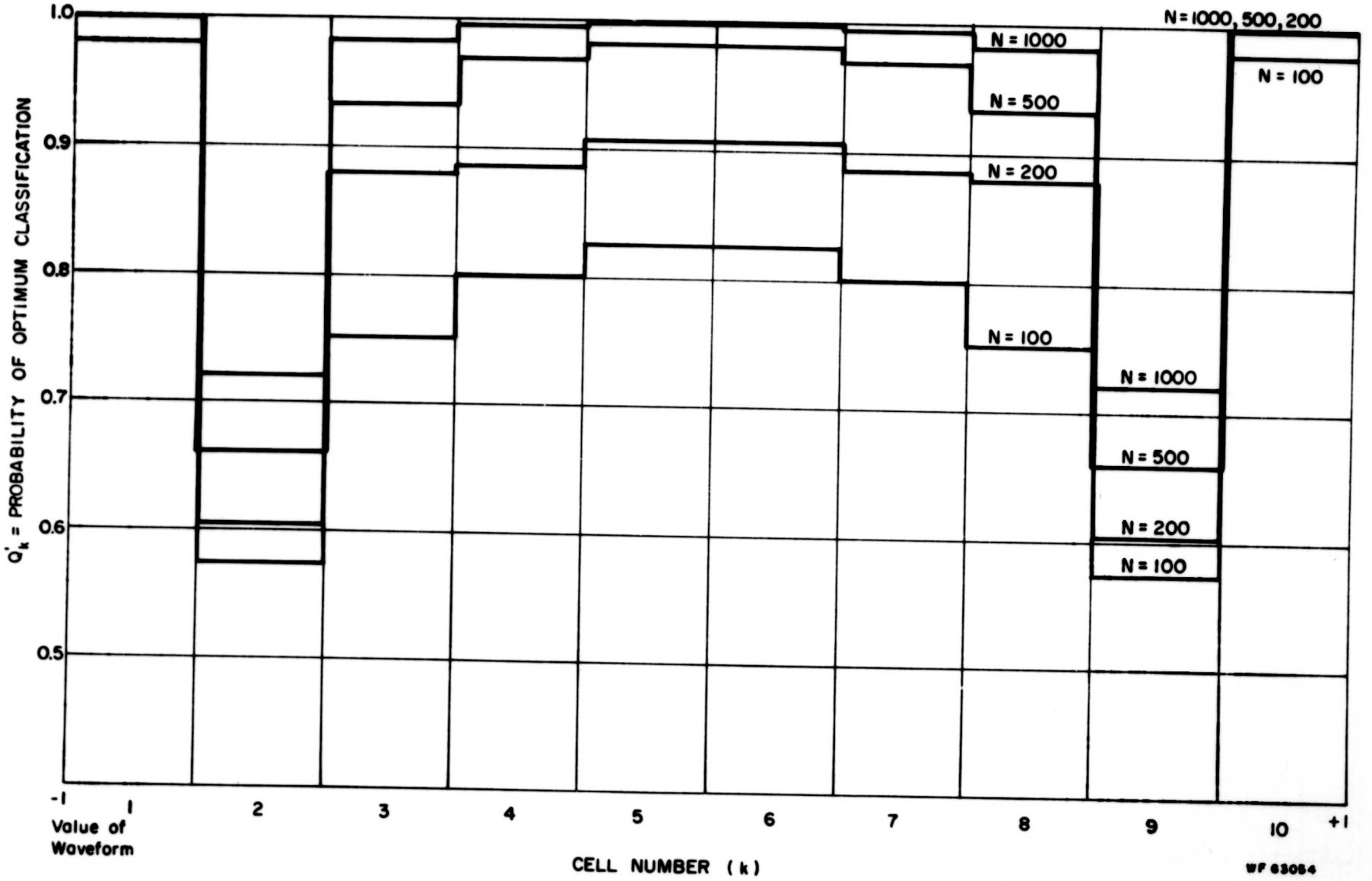


Fig. E-3 Probability of Optimum Classification of a Sinewave and Sawtooth Waveform (C = 10).

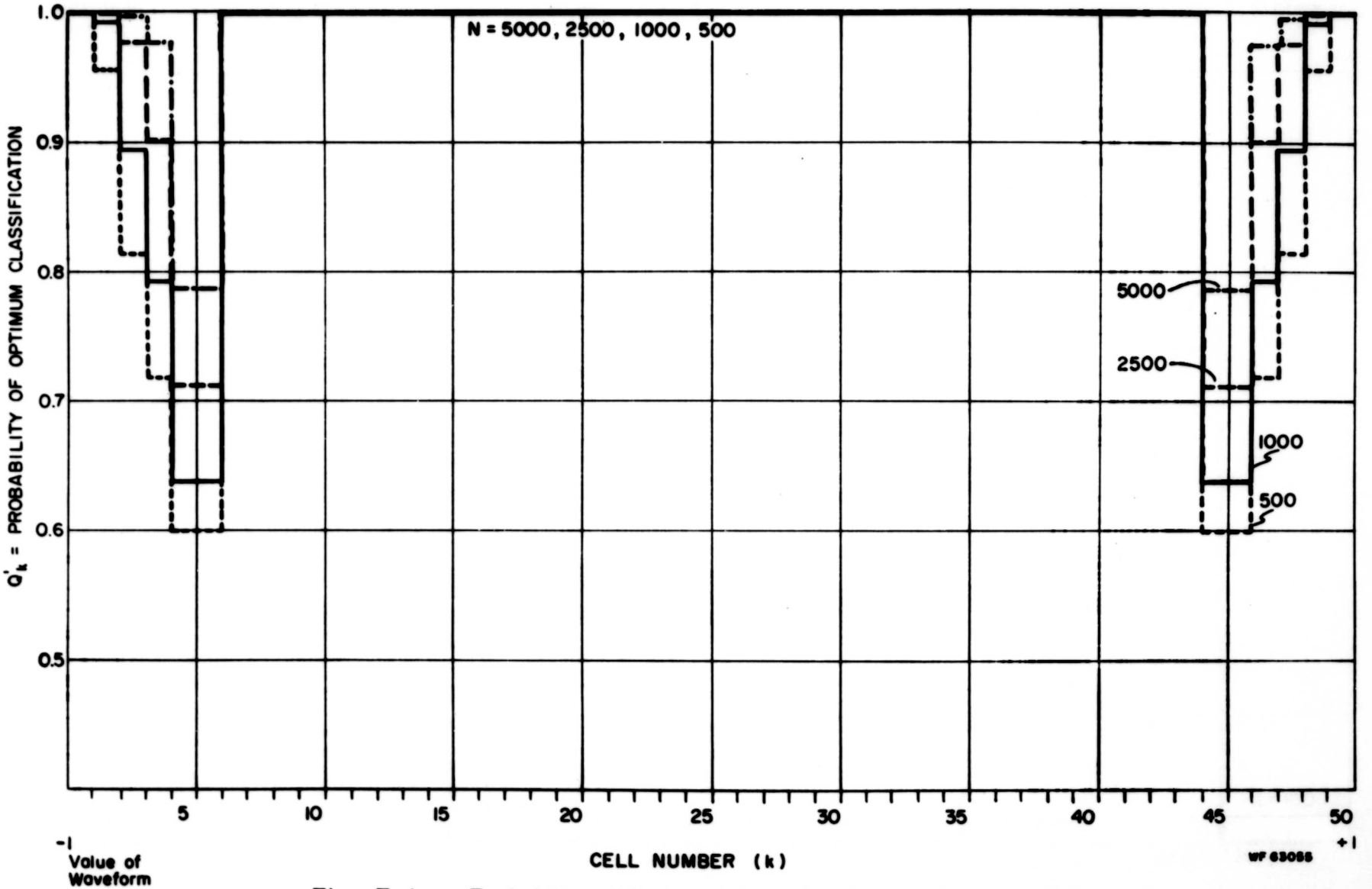


Fig. E-4 Probability of Optimum Classification of a Sinewave and Sawtooth Waveform (C = 50).

APPENDIX VI

EXPERIMENTAL STUDY OF SPEAR LEARNING AND CONTROL PARAMETER SELECTION

A series of experiments were conducted to study the utility of the SPEAR program feature which makes the generated cell shapes and sizes dependent upon the known class examples. The theory of control parameter selection presented in Appendix II was tested experimentally by repeated trials made with a variety of control parameter values. This series of experiments provided a good picture of the usefulness of the technique generally and, specifically, the "cell growth" feature of the SPEAR program. These experiments have shown that this technique can yield a good approximation to a general multimodal, multivariate probability density function (pdf) with a very reasonable number of stored quantities and processing time. In addition, the method of selecting control parameters outlined in Appendix II has given satisfactory results. A qualitative indication of the necessary precision in parameter selection also has been obtained.

The data used in these experiments was generated by the GENSPR computer program described in Appendix II. This data was used because its statistics could be specified in advance and thus the problem of collecting statistically representative data could be avoided. Knowledge of the class statistics enabled a precise determination of the lowest achievable error probabilities.

The distribution of each of the two classes considered had four modes in a four-dimensional observation space. Specifically, the pdf for the first class at the point $\underset{\sim}{X} = (\xi_1, \xi_2, \xi_3, \xi_4)$ is

$$q_1(\underset{\sim}{X}) = 0.2 \frac{1}{(2\pi)^2 \cdot 16} \exp\left\{-1/2\left[\left(\frac{\xi_1 - 21}{4}\right)^2 + \left(\frac{\xi_2 - 10}{2}\right)^2 + \left(\frac{\xi_3 - 9}{2}\right)^2 + \left(\frac{\xi_4 - 9}{1}\right)^2\right]\right\} +$$

$$0.25 \frac{1}{(2\pi)^2 \cdot 16} \exp\left\{-1/2\left[\left(\frac{\xi_1 - 25}{2}\right)^2 + \left(\frac{\xi_2 - 12}{2}\right)^2 + \left(\frac{\xi_3 - 10}{2}\right)^2 + \left(\frac{\xi_4 - 11}{2}\right)^2\right]\right\} +$$

$$0.35 \frac{1}{(2\pi)^2 \cdot 32} \exp\left\{-1/2\left[\left(\frac{\xi_1 - 25}{2}\right)^2 + \left(\frac{\xi_2 - 18}{4}\right)^2 + \left(\frac{\xi_3 - 15}{4}\right)^2 + \left(\frac{\xi_4 - 13}{1}\right)^2\right]\right\} +$$

$$0.2 \frac{1}{(2\pi)^2 \cdot 40} \exp\left\{-1/2\left[\left(\frac{\xi_1 - 20}{5}\right)^2 + \left(\frac{\xi_2 - 20}{2}\right) + \left(\frac{\xi_3 - 16}{2}\right)^2 + \left(\frac{\xi_4 - 15}{2}\right)^2\right]\right\}$$

(F-1)

and the pdf for the second class is

$$q_2(\underset{\sim}{X}) = 0.15 \frac{1}{(2\pi)^2 \cdot 40} \exp\left\{-1/2\left[\left(\frac{\xi_1 - 29}{5}\right)^2 + \left(\frac{\xi_2 - 20}{2}\right)^2 + \left(\frac{\xi_3 - 9}{2}\right)^2 + \left(\frac{\xi_4 - 7}{2}\right)^2\right]\right\} +$$

$$0.3 \frac{1}{(2\pi)^2 \cdot 32} \exp\left\{-1/2\left[\left(\frac{\xi_1 - 19}{2}\right)^2 + \left(\frac{\xi_2 - 16}{4}\right)^2 + \left(\frac{\xi_3 - 14}{2}\right)^2 + \left(\frac{\xi_4 - 11}{2}\right)^2\right]\right\} +$$

$$0.25 \frac{1}{(2\pi)^2 \cdot 16} \exp\left\{-1/2\left[\left(\frac{\xi_1 - 23}{2}\right)^2 + \left(\frac{\xi_2 - 11}{2}\right)^2 + \left(\frac{\xi_3 - 15}{2}\right)^2 + \left(\frac{\xi_4 - 15}{2}\right)^2\right]\right\} +$$

$$0.3 \frac{1}{(2\pi)^2 \cdot 16} \exp\left\{-1/2\left[\left(\frac{\xi_1 - 28}{2}\right)^2 + \left(\frac{\xi_2 - 13}{2}\right)^2 + \left(\frac{\xi_3 - 10}{2}\right)^2 + \left(\frac{\xi_4 - 17}{2}\right)^2\right]\right\}. \quad \text{(F-2)}$$

Samples taken from these two populations may be expected to form two interlocking spirals in four-space. Visualization of this is aided by the projections of these two density functions on the six coordinate planes as is shown in Figure F-1 where each ellipse represents a mode of the distribution with the coordinate radii equal to one (mode) standard deviation.

To study the effectiveness of "cell growth" as a technique for reducing the required number of cells (or "typical samples") for good pdf approximation and to study the precision needed in specifying the control parameters eight computer runs with SPEAR ("learning" only) were made on 1000 vectors from each of the populations with pdf's (F-1) and (F-2). The series of experiments is specified by the diagram in Figure F-2 (here N = 4) in which each choice of the control parameters τ_N, θ_N, ω, and the initial cell radii, $\tau_N \alpha_S(0)$, represents an experiment.

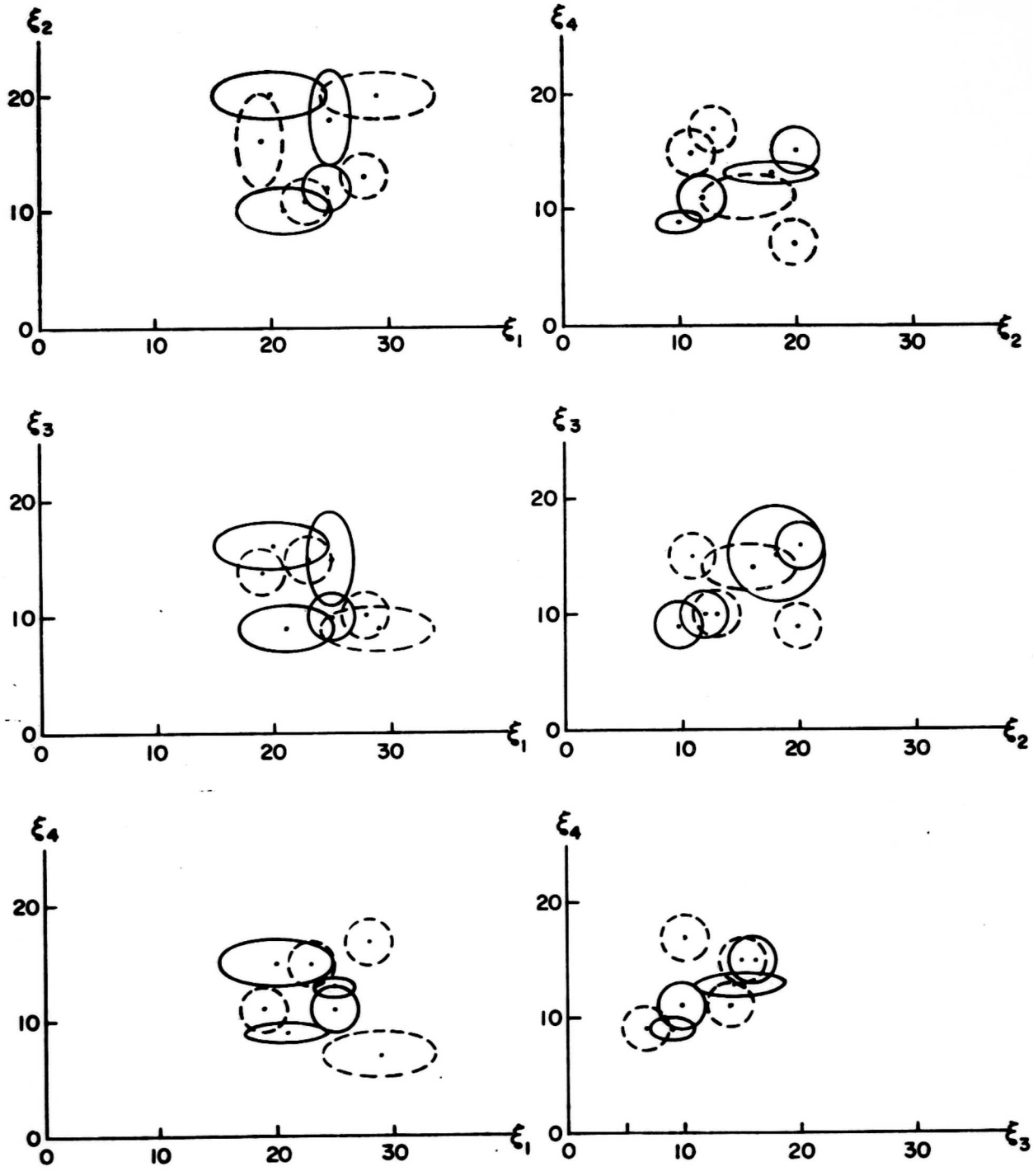


Fig. F-1 Bivariate Representation of $q_1(\underset{\sim}{x})$ and $q_2(\underset{\sim}{x})$.

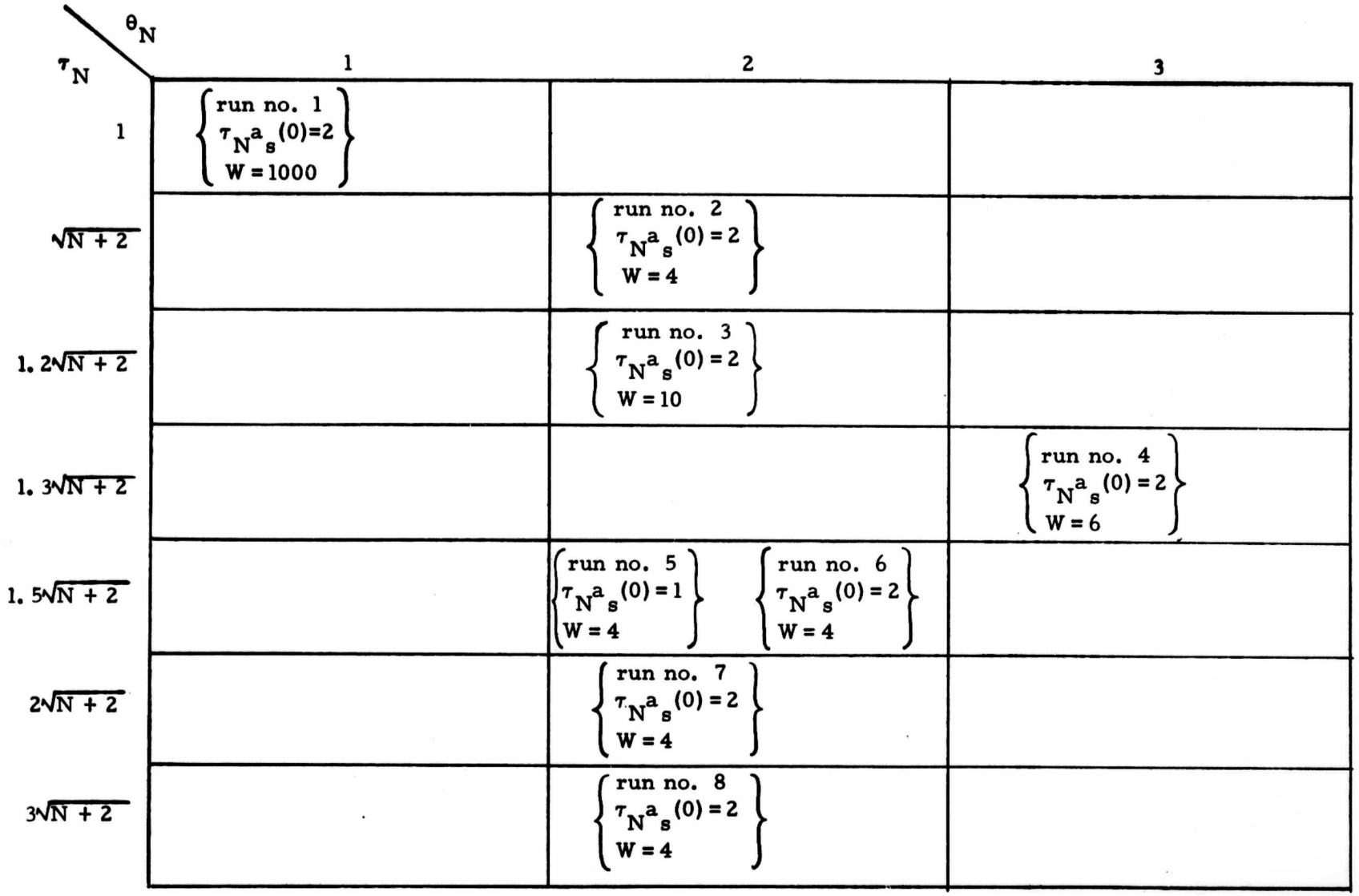


Figure F-2. Diagram specifying seven "learning" experiments on SPEAR. (N = 4)

Run No. 1, with $\tau_N = \theta_N = 1$, was included as a partial simulation of the older ASSC II program (see Appendix IV and Reference [5]). The cell structure generation for Run No. 1 was essentially the same as would have been obtained with ASSC II except that SPEAR is not written to handle the large number of cells ASSC II would have generated and so the run was not actually completed. (See Table F-1). It should be noted that the original ASSC II program was written in such a way that the number of cells or references generated could be controlled. Usually an upper limit for the number of cells created is specified and the program sequences through successive "learning" phases with successively increasing values of τ_N until the number of cells created drops below the specified upper limit. Since the purpose of these experiments was to determine the effect of the choice of τ_N on cell growth, two more runs on ASSC II were performed.

Most of the other computer runs were made with $\theta_N = 2$, so that, by equation B-4, a new cell could not be generated within two cell radii of any existing cell; i.e., a new cell was not allowed to "overlap" any existing cell.

All runs but No. 5 had an initial cell radius (in all directions) of $\tau_N \alpha_S(0) = 2$. This is the minimum standard deviation for all but two modes in equations F-1 and F-2. Furthermore and more to the point of designing experiments with complicated data of unknown distribution, a plot of the data points in one or two dimensions will show characteristic spreads of one to two units about local modes. The initial cell sizes including the annular region about the cell, must not, of course, be so large that sample points from more than one clearly distinguishable mode will fall in a single cell causing the cell to grow excessively. For the larger values of τ_N, the initial trial standard deviations, $\alpha_S(0)$, are considerably smaller than the spreads observable by univariate data analysis. Run No. 5 was included to determine the effect of very small initial cell sizes on the rate of cell growth. As shown in the table of results (Table F-1), an initial cell radius of one unit proved to be so small that no

Table F-1. Results of "Learning" Experiments on SPEAR

Run No.	Initial Cell Radius	τ_N	$\alpha_S(0)$	θ_N	ω	No. of Generated Cells ** Class 1 >0.1%	Class 1 ≥2.5%	Class 2 >0.1%	Class 2 ≥2.5%	Comments
1	2	1	2	1	1000	160*	0	147*	0	222 vectors from class 1 and 241 vectors from class 2 were not included in any cell. * The program could not handle more than 300 cells as the run was not completed.
2	2	$\sqrt{N+2}$	0.816	2	4	75	15	80	10	The cells grew little in size.
3	2	$1.2\sqrt{N+2}$	0.673	2	10	63	11	68	11	The cells grew little in size.
4	2	$1.3\sqrt{N+2}$	0.629	3	6	17	9	19	4	The pdf for class 1 is approximated well, and for class 2 the approximation is very good.
5	1	$1.5\sqrt{N+2}$	0.272	2	4	91	3	--	--	Not enough vectors fell in the same cell to start growth.
6	2	$1.5\sqrt{N+2}$	0.543	2	4	30	3	32	5	The cell structure seems to be only fair. The "cell growth" appears to have been too rapid.

Table F-1. Results of "Learning" Experiments on SPEAR (cont'd)

Run No.	Initial Cell Radius	τ_N	$\alpha_S(0)$	θ_N	ω	No. of Generated Cells** Class 1 >0.1%	Class 1 ≥2.5%	Class 2 >0.1%	Class 2 ≥2.5%	Comments
7	2	$2\sqrt{N+2}$	0.409	2	4	16	2	17	2	926 vectors from class 1 fell in two cells. 871 vectors from class 2 fell in one cell.
8	2	$3\sqrt{N+2}$	0.272	2	4	8	1	9	1	956 vectors from class 1 fell in one cell. 953 vectors from class 2 fell in one cell. The cells for both classes were placed very badly.

* The older ASSC II program would have generated approximately 320 cells for each of the two classes; approximately half of these would have had only one vector in them. The ASSC II program would then have repeated the "learning" procedure with a larger τ_N while a sufficiently low number of cells had been created. The SPEAR program does not have this capability.

** Shown are the numbers of cells that have greater than 0.1% and 2.5% of the data in them.

significant cell growth took place even after 1000 vectors from the first class had been processed.

For $\tau_N = \sqrt{N+2}$ and $1.2\sqrt{N+2}$, little cell growth occurred. This is predicted by Figure B-3. By contrast, excessive cell growth occurred for large values of $\tau_N (> 1.5\sqrt{N+2})$. From Figure B-3 rapid but controlled growth can be expected for values of τ_N between $1.2\sqrt{N+2}$ and $1.5\sqrt{N+2}$. This is supported by the experimental evidence in Tables F-1 and F-2. In Run No. 4 ($\tau_N = 1.3\sqrt{N+2}$, $\theta_N = 3$), only four cells were generated for each class which contained at least 10% of the "learning" sample. These four cells were located at or very near the four distribution modes. This indicates that the pdf's were approximated very well.

The results shown in Table F-2 give an indication of the accuracy of estimation for various values of τ_N. Table F-2 shows a comparison of the true pdf value, $q(\underset{\sim}{X})$, to the estimate, $\hat{q}(\underset{\sim}{X})$, developed in each of four SPEAR experiments (using the same data in each) for $\underset{\sim}{X}$ falling in three distinct regions.

In the first set of comparisons, the ratio $q_1(\underset{\sim}{X})/\hat{q}_1(X)$ is computed for points $\underset{\sim}{X}$ near (25, 12, 10, 11) the second mode in Equation F-1. (For convenience the computations were made at cell centers or "typical samples" rather than at a single common point.) The second "subdistribution" about (25, 12, 10, 11) in Equation F-1 lies near the first and third "subdistributions" of Equation F-1 as may be seen by an inspection of Figure F-1. Therefore, it would not be too surprising to find cells scattered between these distribution modes as well as, or instead of, at the true modes. (This would not necessarily mean a poorer approximation to the pdf). The fact that cells were located near actual modes is an indication of the mode separating capability of SPEAR.

The second set of comparisons in Table F-2 shows the ratio $q_2(\underset{\sim}{X})/\hat{q}_2(\underset{\sim}{X})$, computed at points $\underset{\sim}{X}$ near (19, 16, 14, 11), the second mode in Equation F-2.

Table F-2. Comparison of the Accuracy of PDF Estimation Between Runs with Various Control Parameter Values

Run No.	3/4 in. τ_N	θ_N	Point Where Estimate is Computed	$q(\underset{\sim}{X})/\hat{q}(\underset{\sim}{X})$	Comments
2	$\sqrt{N+2}$	2	(26.18, 12.52, 9.22, 12.51)	3.99	$\underset{\sim}{X}$ is from the first class and is near the second mode. (This mode has standard deviations equal to the initial cell radii and so could be fitted perfectly with one cell that did not grow.
4	$1.3\sqrt{N+2}$	3	(25.54, 12.65, 9.85, 11.60)	1.02	
6	$1.5\sqrt{N+2}$	2	(26.15, 12.37, 9.20, 12.17)	4.70	
2	$\sqrt{N+2}$	2	(18.39, 15.55, 13.99, 10.21)	0.209	$\underset{\sim}{X}$ is from the second class and is near the second mode. (This mode requires cell growth before it can be fitted with only one cell.)
4	$1.3\sqrt{N+2}$	3	(19.03, 16.08, 13.85, 10.82)	1.01	
6	$1.5\sqrt{N+2}$	2	(18.88, 16.64, 13.58, 10.08)	0.89	
2	$\sqrt{N+2}$	2	(17.54, 9.91, 7.96, 9.21)	0.202	$\underset{\sim}{X}$ is from the first class and is on a shoulder of the first subdistribution, i.e., $\xi \cong$ 1 s.d. less than the first mode.
4	$1.3\sqrt{N+2}$	3	(16.70, 9.97, 7.62, 9.15)	0.224	
6	$1.5\sqrt{N+2}$	2	(17.19, 9.23, 8.24, 8.99)	0.131	

The second "subdistribution" in Equation F-2 is well separated from the other "subdistributions" of $q_2(X)$. Therefore, it was expected that a cell would locate itself about (19, 16, 14, 11) and that this "subdistribution" would be approximated rather well. It is apparent from the table that, for Run No. 4, the ratio of estimated to actual density values is near unity.

The third set of comparisons in Table F-2 shows the ratio $q_1(\underset{\sim}{X})/\hat{q}_1(\underset{\sim}{X})$, computed at points $\underset{\sim}{X}$ near (17, 10, 8, 9), i.e., approximately one standard deviation off in the first coordinate from the first mode in Equation F-1. An unusually large number of vectors fell in a small neighborhood about (17, 10, 8, 9). This is responsible for the fact that every run produced a cell near this point and is at least partly responsible for the estimates deviating from the true pdf values. Since there are more than the expected number of sample vectors in a small neighborhood of (17, 10, 8, 9), the estimates $\hat{q}_1(\underset{\sim}{X})$ in that neighborhood are higher than $q_1(\underset{\sim}{X})$ so that the ratio $q_1(\underset{\sim}{X})/\hat{q}_1(\underset{\sim}{X})$ is less than one.

Certainly in both the first and second set of comparisons an accurate estimate of the appropriate pdf has been obtained, for the points at which the computations were made, with a value of $\tau_n = 1.3\sqrt{N+2}$. Although detailed computations were not made at the other modes of $q_1(\underset{\sim}{X})$ and $q_2(\underset{\sim}{X})$, the accuracy of the estimate obtained with $\tau_N = 1.3\sqrt{N+2}$ appears to be uniformly good. Since the "subdistribution" variances were well estimated with $\tau_N = 1.3\sqrt{N+2}$, the total pdf's were estimated well over the entire space with the exception of a few local discrepancies caused by cells not located at the true modes. It must be remembered that the special form of the pdf's in Equations F-1 and F-2 allows an almost perfect approximation with the type of representation used in ASSC III.

The third set of comparisons corroborates the conclusion that the best accuracy in pdf estimation is obtained for values of τ_N near $1.3\sqrt{N+2}$. As

described above, at least part of the deviation from the true value is the result of sampling errors and not a characteristic of cells located away from the true modes.

Figure F-3 shows the total number of cells generated (including those with only one vector in them) for the first class versus τ_N (and $\beta = \tau_N/\sqrt{N+2}$) for the computer runs in which the initial cell radius was two units. The ASSC II simulation produced approximately 320 cells. In large part the fact that so many more cells were generated on this run than with larger values of τ_N and θ is due to the absence of an annular zone about the cell in which no new cell centers can be generated (see Equation B-4). That is holding the cell centers (or "typical samples") apart so that the cell defining ellipsoids do not overlap (see Figure C-4) is a more powerful method of reducing the number of generated cells than is the cell growth mechanism alone. However, by eliminating those cells which can be considered as "singular events" because they contain only a small percentage of the learning sample after the cell growth mechanism has been employed, a significant reduction in the number of useful cells is possible. For example, as seen in Figure F-3, Runs No. 2 and 3 both generated approximately 90 cells. In Table F-1, however, it is seen that 75 cells for Run No. 2 (class 1) contained more than one vector (0.1% of the "learning" sample) while only 63 of the cells generated in Run No. 3 contained more than one vector and thus were retained by the present SPEAR program. This represents a reduction in the number of useful cells by a factor of 12/75 or 16% due solely to cell growth for values of τ_N such that little growth occurred. For higher values of τ_N this percentage reduction is even more significant. For $\tau_N = 1.5\sqrt{N+2}$ (Run No. 6), only 30 cells had more than one vector in them. This is a reduction of 67% due to cell growth and a reduction by approximately 90% from ASSC II with an equal initial cell size.

In Figure F-4, the cumulative total number of cells generated (including those containing only one vector) is shown as a function of the number of

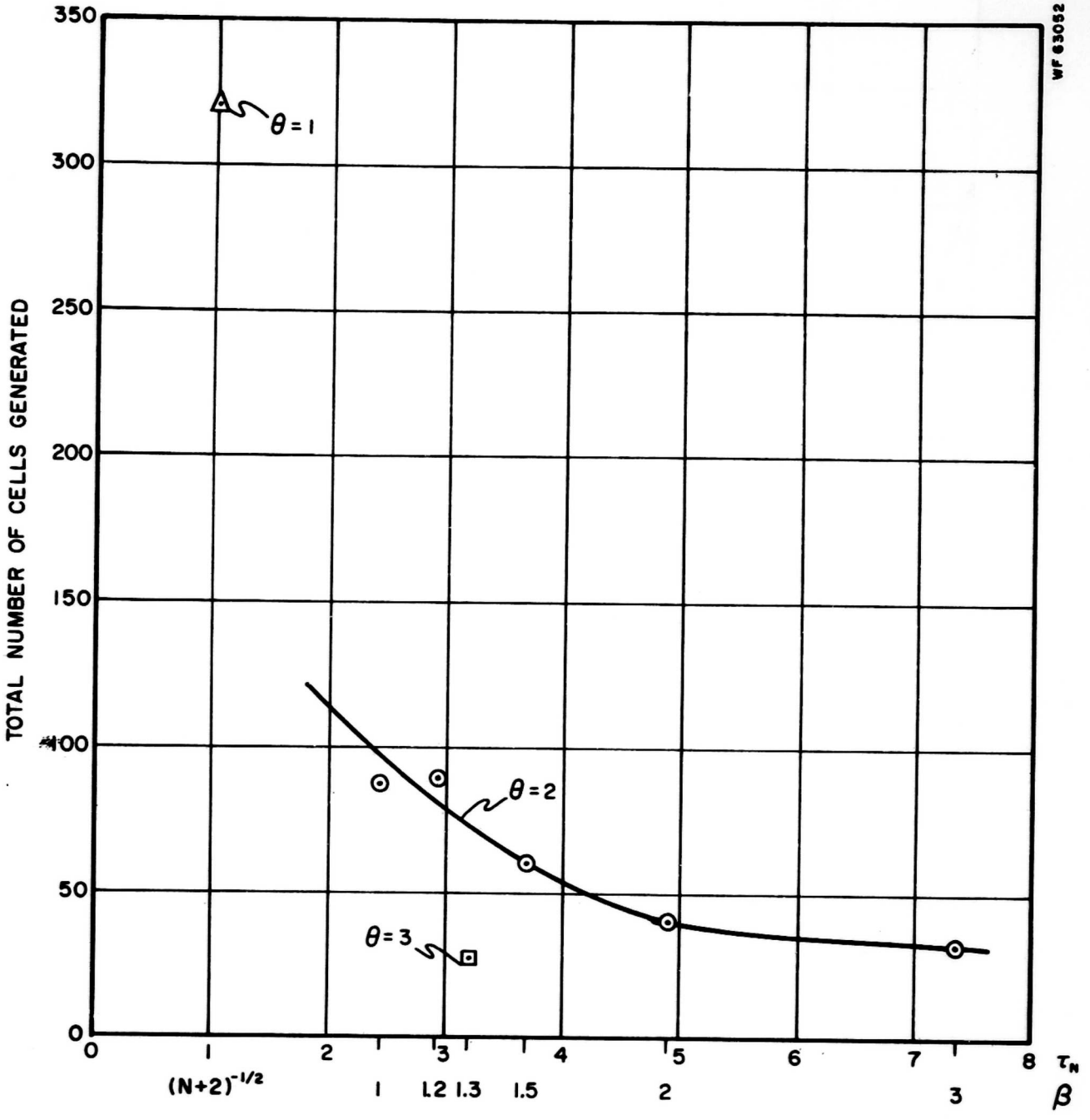


Fig. F-3 Indication of the Effectiveness of τ_N and θ as Storage Reducing Control Parameters.

Fig. F-4 Indication of the "Learning Rate" (Rate of Cell Generation) for Various Control Parameter Settings

known vectors processed for various settings of the control parameters. In each of the runs for which these results are displayed the initial cell radius equals two units. Observe that for fixed θ and initial cell radius, the curves in Figure F-4 have a common initial behavior. For low values of τ_N, new cells were generated throughout the learning procedure, i.e., until all 1000 known vectors had been treated. For large values of τ_N, the rate of cell generation proceeded normally until one or more cells began to expand explosively, taking in all remaining vectors in the "learning" sample, so that generation of new cells ceased. Figure F-5 shows the number of vectors that were processed before 80% of the final number of cells were generated. The major cause of the difference in results shown in Figures F-3, -4, and -5, between the runs with $\tau_N = \sqrt{N+2}$ and $\tau_N = 1.2\sqrt{N+2}$ (with $\theta = 2$) seems to be that for Run No. 3 ($\tau_N = 1.2\sqrt{N+2}$ the initial cell radii were equal to 1.98 units as a result of round-off error.

The results of these "learning" experiments may be summarized by stating that strong evidence has been obtained which indicates that the method of control parameter selection set forth in Appendix II is valid, and that SPEAR can (with proper control parameters and initial cell sizes) produce satisfactory estimates of the class pdf's with efficient use of available storage capacity. Thus, classification error rates approximating the optimum can be achieved with SPEAR pdf estimation or "learning", and an improvement has been achieved in performance as well as generality over the earlier ASSC II program.

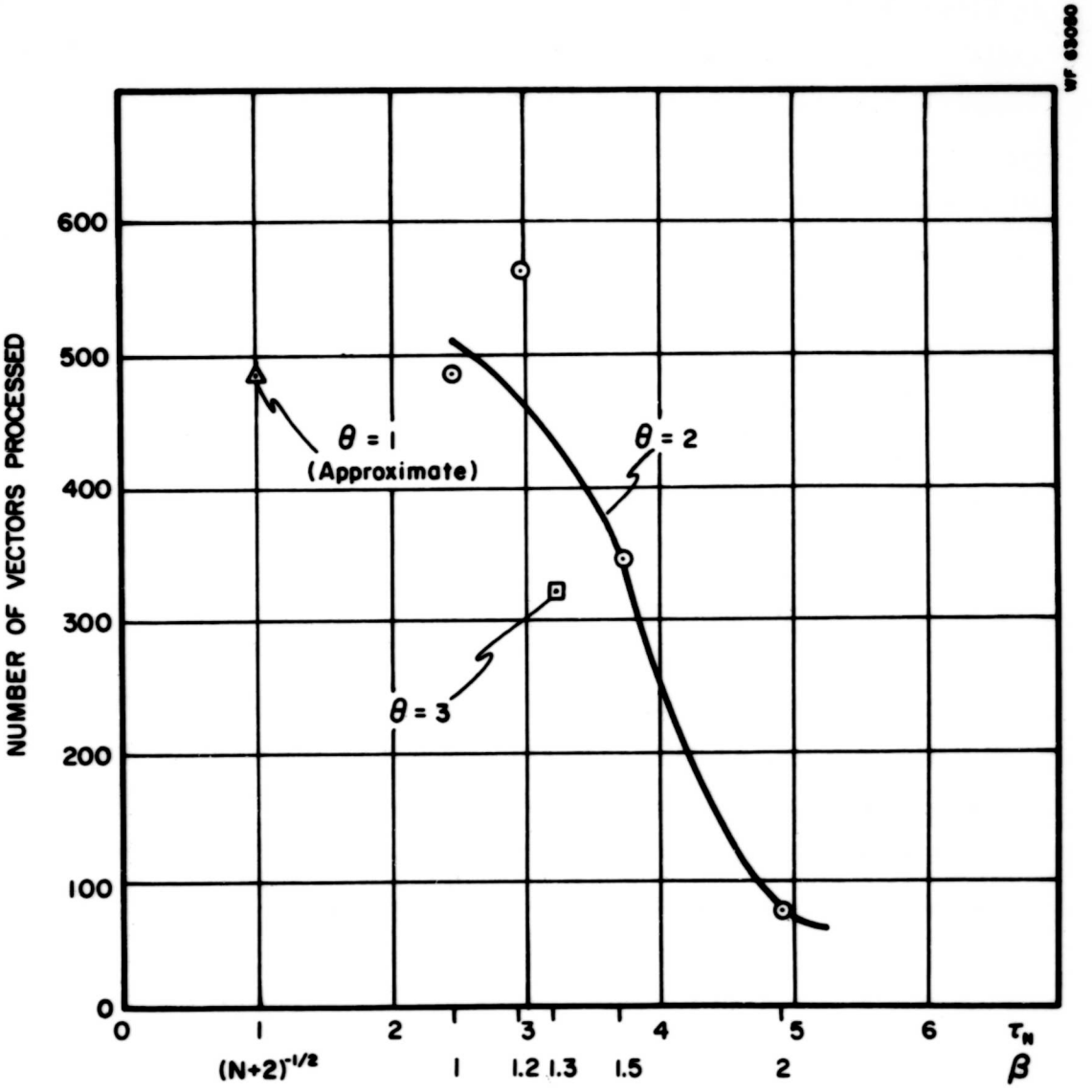


Fig. F-5 Number of Vectors Processed Before 80% of the Final Number of Cells Were Generated

UNCLASSIFIED

UNCLASSIFIED